EXTENSIONS OF THE LANGEVIN EQUATION FOR PROTEIN DYNAMICS FOR MODELLING EQUILIBRIUM FLUCTUATIONS OF PROTEINS

by

ERIC BEYERLE

A DISSERTATION

Presented to the Department of Chemistry and Biochemistry
and the Division of Graduate Studies of the University of Oregon
in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy

September 2021

DISSERTATION APPROVAL PAGE

Student: Eric Beyerle

Title: Extensions of the Langevin Equation for Protein Dynamics for Modelling Equilibrium Fluctuations of Proteins

This dissertation has been accepted and approved in partial fulfillment of the requirements for the Doctor of Philosophy degree in the Department of Chemistry and Biochemistry by:

Jim Prell	Chair
Marina Guenza	Advisor
Andy Marcus	Core Member
John Toner	Institutional Representative

and

Andy Karduna	Interim Vice Provost for Graduate Studies

Original approval signatures are on file with the University of Oregon Division of Graduate Studies.

Degree awarded September 2021

DISSERTATION ABSTRACT

Eric Beyerle

Doctor of Philosophy

Department of Chemistry and Biochemistry

September 2021

Title: Extensions of the Langevin Equation for Protein Dynamics for Modelling Equilibrium Fluctuations of Proteins

Proteins are not static structures; they must undergo conformational fluctuations about their folded state to function. Typically, the slow, near-equilibrium conformational dynamics of proteins encode the functional motions; an accurate description of these dynamics is useful for elucidating the functional motions of proteins. Use of molecular dynamics (MD) simulations gives a physical model of proteins' motions, but the dynamics are too high dimensional and coupled to determine the functional motions purely from observation of the MD trajectory; thus, methods to efficiently extract the slow conformational dynamics of proteins from atomistic models are valuable.

This dissertation advances the Langevin equation for protein dynamics (LE4PD), a diffusive, coarse-grained equation of motion for modeling protein dynamics adapted from the field of polymer physics. The LE4PD is solved by an eigenvalue decomposition into a set of normal mode coordinates, each of which encodes dynamics on a specific time- and lengthscale. A discrete-state master equation approach, Markov state modeling, is used to precisely determine the dynamics and kinetics

of conformational dynamics described by the slow LE4PD modes by analyzing a 1-microsecond, folded simulation of the protein ubiquitin. The approach is able to extract slow dynamics in important binding regions of ubiquitin. In chapter III, Markov state models are used to determine the contributions of metastable states to the circular dichroism spectrum of a dinucleotide system.

Because protein dynamics is inherently anisotropic, we develop an anisotropic version of the LE4PD. When both hydrodynamic effects and free-energy barriers are neglected, the model reduces to a principal component analysis of the alpha-carbon coordinates; including both these effects are important for quantitatively modelling the decay of simulated autocorrelation functions.

Finally, we compare the LE4PD predictions from the ubiquitin simulation to the slow modes extracted by a time-lagged independent component analysis of the trajectory. We find both methods are able to extract the slow dynamics of the protein, but the tICA compresses the information into a smaller number of modes; however, for ubiquitin, the tICA modes cannot model the simulated autocorrelation functions as effectively as the anisotropic LE4PD model.

This dissertation includes previously published and unpublished co-authored material.

CURRICULUM VITAE

NAME OF AUTHOR: Eric Beyerle

GRADUATE AND UNDERGRADUATE SCHOOLS ATTENDED:

University of Oregon
Centre College

DEGREES AWARDED:

Doctor of Philosophy, Physical Chemistry, 2021, University of Oregon
Bachelor of Science, Chemical Physics, 2015, Centre College

AREAS OF SPECIAL INTEREST:

Statistical mechanics, coarse graining, stochastic processes, cross country

PROFESSIONAL EXPERIENCE:

Graduate Researcher, Guenza Lab, University of Oregon, 2016-2021

Gradaute Teaching Fellow, Department of Chemistry and Biochemistry, University of Oregon, 2015-2021

Teaching Assistant, Department of Chemistry, Centre College, 2014-2015

GRANTS, AWARDS AND HONORS:

John Keana Fellowship, University of Oregon Department of Chemistry and Biochemistry, 2019-2020

Valedictorian Prize, Centre College, 2015

Max P. Cavnes Award, Centre College, 2012

PUBLICATIONS:

E. R. Beyerle and M. G. Guenza. Identifying the leading dynamics of ubiquitin: a comparison between the tICA and the LE4PD slow fluctuations in amino acids' position. Manuscript submitted to *The Journal of Chemical Physics*.

E. R. Beyerle and M. G. Guenza. Comparison between slow, anisotropic LE4PD fluctuations and the Principal Component Analysis modes of Ubiquitin. *J. Chem. Phys.*, **154**, 12411 (2021). Paper published as part of the special topic on Special Collection in Honor of Women in Chemical Physics and Physical Chemistry.

E. R. Beyerle, M. Dinpajooh, H. Ji, P. von Hippel, A. H. Marcus, and M. G. Guenza. Dinucleotides as simple models of the base stacking-unstacking component of DNA 'breathing' mechanisms. *NAR*, **49**, (2021), 1872 - 1885.

E. R. Beyerle and M. G. Guenza. Kinetic Analysis of Ubiquitin Local Fluctuations with Markov State Modeling of the LE4PD Normal Modes. J. Chem. Phys., **151**(16):164119, 2019. Paper published as part of the special topic on Markov Models in Molecular Kinetics.

J. Copperman, M. Dinpajooh, E. R. Beyerle, and M. G. Guenza Universality and specificity in protein fluctuation dynamics. *Phys. Rev. Lett.*, **119**, 158101 (2017).

S. Asmus, M. Raghanti, E. Beyerle, J. Fleming-Beattie, S. Hawkins, C. McKernan, and N. Rauh. Tyrosine hydroxylase-producing neurons in the human cerebral cortex do not colocalize with calcium-binding proteins or the serotonin 3A receptor. *J. Chem. Neuroanat.*, **78** (2016), 1-9.

ACKNOWLEDGEMENTS

This dissertation would not exist without outside assistance. First, many thanks to my adviser Dr. Marina Guenza for the academic and financial support, intellectual rigor, and motivation. To my thesis committee, Drs. Jim Prell, Marina Guenza, Andy Marcus, and John Toner: thanks for the questions and insights you afforded me during my oral exams and annual reviews – I apologize they were of such long duration. To the other members of the Guenza Lab, especially Jeremy Copperman, who was (briefly) my graduate mentor; Hadi Dinpajooh, who I consider a postdoctoral role model; Pablo Romano, who introduced me to the Python programming language and assisted me with various practical research problems during my early graduate career; and Tomas Fencl for his friendship.

Thanks to the fellow graduate students in the Chemistry and Biochemisty department at the university for their support over the years, especially Alexis Kiessling, Phil Lotshaw, Phil Kovac, Brett Israels, Andrew Carpenter, Emma Tran, Jack Maurer, and Dylan Heussman.

Thanks to Drs. Tom Greenbowe and Deb Exton and the rest of the general chemistry lab staff (including the head teaching assistants!)– I learned quite a bit about chemistry, science, and teaching during my rounds of TA-ing general chemistry lab, which was a pleasant surprise.

A massive thanks to the running community of Eugene, especially the University of Oregon running club. In particular: Tom Heinonen, for your organization, weekend emails, insight, advice, and so much more; to Josh Gordon for all the miles, mentorship, humor, and so much else besides; to Magda van Leeuwen, for putting up with me as fellow coordinator for two years...I feel I was never at my best; to Ryan

Jones, for being Ryan Jones; and to Robert Pedersen, Jr., for all the miles, motivation, and companionship. I would also like to thank additional run clubbers not previously mentioned who I think have made the club great in my time here: Dana (Fry) Hunter, Taylor Howat, Andrew Wagner, August Howell, Jake Willard, Derrick Marshall, Seth Berdahl, Carter Christman, Jake Brohman, Jack Rising, Evan Kwiecien, and Wolf Seifer, among others. Thanks also to Justin Banks, Matt Barnhart, Orin Schumacher, Kevin Cave, and Emmett Saulnier for giving me the ability to have daily social interactions during the summer, fall, and winter of 2020.

Finally, I would like to thank my parents, Jude and Judy Beyerle. Without their unyielding support, none of anything I've done here would have been possible. Thank you, Mom and Dad...for everything.

"The one thing I was good at was winning scholarships and prizes, and that era was coming to an end. I felt like a racehorse in a world without racetracks or a champion college footballer suddenly confronted by Wall Street and a business suit, his days of glory shrunk to a little gold cup on his mantel with a date engraved on it like the date on a tombstone." – *The Bell Jar*

"The soul? There's nothing but chemistry here." – *Breaking Bad*

TABLE OF CONTENTS

LIST OF FIGURES

Figure Page

LIST OF TABLES

Table Page

CHAPTER I

INTRODUCTION

Proteins are linear chains of amino acids that carry out many activities crucial to sustaining biological life [6]. While techniques for determining the three-dimensional structure of proteins have been known and used for many years [7, 8], with the advent of various models for describing the conformational dynamics of proteins, such as the induced fit [9] and the Monod-Wyman-Changeaux allosteric model [10], which would eventually be developed into the conformational selection model [11–13], it has become recognized that not only a protein's primary sequence and tertiary structure, but also the protein's ability to sample different conformational states, or its dynamics, is important to describing the protein's function.

A technique known as molecular dynamics (MD), which evolves the atomistic coordinates of a physical system in time using a variant of Newton's equations of motion according to a defined energy function [14], is frequently used to model protein motions at the atomistic level of resolution. First applied to the bovine pancreatic trypsin inhibitor protein at the picosecond timescale, [15] state-of-the-art MD simulations of proteins have reached the millisecond and longer timescales. [16–19] However, even with this impressive increase in performance, there is still a gap between the time- and lengthscales accessible through MD and those accessible through experiment; furthermore, even with long simulations, there can still be a lack of statistics for the slowest dynamic motions observed in the 'long' simulations (e.g. in a 1-millisecond simulation, an event lasting 0.5 milliseconds is only sampled twice, once in the forward direction and once in the reverse direction, meaning an estimation of the kinetics of this event is plagued by statistical uncertainty). However, MD has

the advantage of resolving the motions of the entire protein and surrounding solvent at an atomistic level, which is not possible experimentally.

The output of the MD simulation, the phase space trajectory of all the atoms in the system, can be challenging to process and interpret. For example, in a modestly sized protein containing $N = 1000$ atoms, the phase space trajectory output from the simulation contains $6 \times N = 6000$ degrees of freedom. Simply observing the trajectory using computational tools [20, 21] is useful to see what is actually occurring over the course of the simulation, but given thousands or more degrees of freedom, it is nearly impossible to sort out what motions are important 'by eye.' For example, if one is interested in the long timescale structural rearrangements of the protein, then the fast, femto- to picosecond vibrations of the protein's hydrogen atoms are of little interest, and the interested party should find a way to safely discard these and other fast motions irrelevant to describing the protein's functional dynamics, which tend to be controlled by its slowest degrees of freedom.[22]

Thus, there is a need to extract the collective motions or 'leading fluctuations' of proteins that are observed on the time- and lengthscales probed by MD simulations. These slow dynamics involve barrier crossing on the protein's free-energy surface, and thus describe conformational changes between two folded states of the protein. [23] These motions can serve as a 'bridge' to connect the motions in the MD simulation and those observed experimentally by combining the extracted collective coordinates with an advanced sampling technique such as metadynamics. [24, 25] Furthermore, due to the intrinsic high dimensionality of a protein's conformational space, there is also great value in discovering a low-dimensional, interpretable, and tractable set of collective coordinates that captures all the essential dynamics of the protein.

So, a major, current goal in the field of protein dynamics is discovering methods that greatly simplify the observed dynamics while still capturing the important, functional dynamics of the protein. Many techniques have been developed previously to determine the most important conformational dynamics of proteins: principal component analysis (PCA) or essential dynamics [26], network models [27–29], independent component analysis [30–32], discrete master equations [33, 34], and, more recently, deep learning approaches [35, 36]. These methods ideally provide the desired low-dimensional set of collective coordinates encoding the important dynamics of the protein, and estimates of the localization, amplitude, and timescales of these important motions.

This dissertation advances a method for accurately describing protein dynamics around the folded state. The Langevin equation for protein dynamics (LE4PD) [37, 38] is a coarse-grained, diffusive, Langevin approach for extracting a set of Langevin modes describing the collective motions of proteins over a range of time- and lengthscales. The LE4PD is itself an extension of the optimized Rouse-Zimm approach for describing the dynamics of unstructured polymers [39], with the coarse grained sites selected as the alpha-carbons of each amino acid in the protein chain. Solving the LE4PD via diagonalization yields a set of modes [40] or collective coordinates that order the dynamics described by the modes according to their diffusive timescales, with the slowest modes being accorded the lowest mode index. The first few modes should describe high-amplitude, slow collective motions encoding the functional dynamics of the protein.

In the LE4PD, the friction coefficients are found by calculating the solvent accessible surface area of each amino acid in the protein using an extension of Stokes' law that accounts for dissipation of energy through both the hydrophobic interior

('internal friction' or 'internal viscosity' [41, 42]) of the protein and the surrounding solvent, where it is assumed the internal friction is coupled to the solvent viscosity. [38, 43, 44] Because the LE4PD is a coarse-grained approach, the timescales described by each mode are accelerated due to the effective reduction of the friction coefficient along each mode coordinate due to the coarse graining. [45, 46] To account for this smoothing effect, the LE4PD calculates a free-energy surface for each mode, and finds the average barrier to transport along each mode by finding the median absolute deviation [47] from the energetic minimum of each mode. [48]

The main theme of this dissertation is extending the previously developed LE4PD theory by more precisely accounting for barriers along each LE4PD mode and accounting explicitly for the anisotropy in protein dynamics. To more precisely account for the mode-dependent barriers, a discrete master equation approach known as Markov state modeling[49] is used to find the transfer operator[33, 50] on the slow LE4PD free-energy surfaces and use its spectral decomposition to estimate the location of the barrier on the surface as well as the timescale required to move between the wells on either side of the barrier. The anisotropy in the dynamics of the protein is taken into account by to switching the basis of the residue fluctuations, which also allows for a direct comparison to the results generated by PCA. These two extensions can be combined to determine the timescales, lengthscales, and localization along the primary sequence of the protein's slow fluctuations, which should encode the functional dynamics of the protein. When the approach is compared to a tICA of the same set of input features, the fluctuations of the alpha-carbons in each residue of ubiquitin, we find that the two approaches predict similar slow dynamics over timescales that are within an order of magnitude of each other. Combined with the LE4PD's ability to reproduce time correlation functions from the base simulation,

[51] the extended LE4PD appears to be a acceptable method for determining the slow, 'essential' dynamics [26] of proteins, based on the results found from thoroughly analyzing molecular dynamics simulations of a small, regulatory protein: ubiquitin. [52]

Ubiquitin is used as a model system to test the extensions to the LE4PD approach because it is a small, but biologically important, well-folded protein possessing a variety of secondary structures with regions of β-sheets, α-helices, and flexible, intrinsically disordered regions. [53] Furthermore, there is a wealth of experimental data available for ubiquitin, especially NMR data,[54–56] so that the predictions of observables predicted from the theory can be compared to their experimentally measured values. [37, 38, 48, 57] Since ubiquitin is highly stable protein with a suspected unfolding temperature at or above 100° C at neutral pH [58], it is often referred to as 'a rock' due to this stability. However, despite a high-proportion of stable secondary structure, ubiquitin also contains several flexible loops and intrinsically disordered regions, most notably the C-terminal tail [52], whose dynamics occurs over timescales of the order of 10 ns. As will elucidated in detail throughout the dissertation, we postulate the slow dynamics in these flexible loops are related to ubiquitin sampling binding conformations, since these loops are known to bind to multiple proteins [52, 59–64]. So, although it is true that ubiquitin is highly stable and most of its structure fluctuates little from its equilibrium, folded state, ubiquitin possesses several flexible regions that undergo large amplitude fluctuations, sampling the many local minima at the bottom of the global energetic minimum corresponding to the folded state. These sampled conformations should overlap with the observed binding configurations of ubiquitin, [55] supporting the conformational selection hypothesis. [10, 11]

The dissertation is organized as follows: in Chapter II another, more rigorous method for determining the timescales of the slowest LE4PD modes using Markov state models (MSMs) [4]; in Chapter IV an extension of the original, isotropic LE4PD approach to an anisotropic Langevin equation for protein dynamics called the LE4PD-XYZ model, [51] which describes the anisotropic fluctuations of the protein and completely eliminates the global motions of translation and rotation from the analysis, allowing for a comparison to the analogous PCA of the same set of coarse-grained sites; in Chapter V, the MSM approach is applied to the LE4PD-XYZ model, and the results are compared to the analogous time-lagged independent component analysis (tICA) for the same trajectory of the protein ubiquitin [65], where it is seen that the LE4PD-XYZ and tICA predict similar slow motions in the protein; and, for completion, Chapter III is an aside illustrating another application of MSMs to describing the conformational dynamics of a simple single-stranded DNA model, deoxyadenine dinucleotide (dApdA). [66]

The research presented in Chapters II and IV is co-authored with Dr. Marina Guenza, have been published in peer-reviewed journals [4, 51] while the research presented in Chapter V is currently undergoing peer-review. [65] Finally, the research presented in Chapter III is co-authored with Hadi Dinpajooh, Huying Ji, Dr. Pete von Hippel, Dr. Andy Marcus, and Dr. Marina Guenza and was also published in a peer-reviewed journal. [66]

CHAPTER II

DESCRIBING THE KINETICS AND DYNAMICS OF THE SLOW LE4PD MODES OF UBIQUITIN WITH MARKOV STATE MODELS

From Beyerle, E. R. and Guenza, M. G. Kinetic Analysis of Ubiquitin Local Fluctuations with Markov State Modeling of the LE4PD Normal Modes. *J. Chem. Phys.*, **151**(16):164119, 2019.

Fluctuation dynamics allow proteins to modify their shape and efficiently sample conformational states that are potentially useful to perform their biological function.[11] Large scale fluctuations can be precursors to unfolding [67] because these fluctuations involve slow cooperative rearrangements of large portions of the protein; these cooperative motions are thought to guide and define the most relevant kinetic pathways of a protein, identifying reaction coordinates that can be important, for example, in substrate binding,[11, 55, 68] product release,[69] regulation,[11] and allostery.[70–72]

Local fluctuations, occurring at a precise lengthscale, are supposed to be relevant in identifying regions of the protein likely involved in molecular recognition.[55, 73] Following the hypothesis of the conformation-selection model by Monod-Wyman-Changeux (or MWC model),[10] local fluctuations along the primary sequence of a protein provide information on the propensity to bind other molecules at the given segment of the protein's primary sequence.

This hypothesized correlation between local fluctuations and binding lies at the foundation of the MWC model, where a substrate selects among a large ensemble of conformations the one that is geometrically and energetically most favorable to

binding.[12] Thus, spontaneous fluctuations are expected to occur even in the absence of a binding partner,[55, 74] so that the modeling of spontaneous local fluctuations of an isolated protein may provide essential information on the kinetic mechanisms of protein binding.

These local fluctuations involve internal deformations of the protein, which require surmounting a free-energy barrier. A dynamical study of spontaneous fluctuations will likely uncover the length- and timescales over which these fluctuations occur, thus potentially highlighting the characteristic spatial and temporal parameters that set the limits for the binding process. Sometimes, internal fluctuations occur on a timescale that is comparable in magnitude with the slowest timescale of protein relaxation (for example rotational diffusion),[38] indicating that those internal motions can be important participants in the mechanisms of molecular recognition.

This chapter shows a study of the emergence of local fluctuations along a protein's primary sequence, and the length- and timescales associated with them, starting from the coarse-grained Langevin Equation for Protein Dynamics (LE4PD). The protein investigated is a regulatory protein in eukaryotic cells, ubiquitin, which is most notable for its ability of post-translationally modifying other proteins through the process of mono- or poly-ubiquitination,[75, 76] a necessary event for a number of important biological functions.[60, 61, 64, 75, 77]

In the process of ubiquitination two ubiquitin molecules bind by forming an amide bond between the carboxyl group at the C-terminus and the ϵ-amino group of a lysine amino acid. The reaction is catalyzed by a number of enzymes called ubiquitin ligases.[78] Ubiquitin has seven lysines, and the length and shape of the chain of ubiquitins depend on which lysine in the protein participates to the binding of the

C-terminus of another ubiquitin. Other parts of ubiquitin are also involved in other binding processes.[52, 61, 75] Following the hypothesis of the MWC model of binding, local fluctuations along the primary sequence of ubiquitin provide information on the propensity of the protein to form bonds at a specific amino acid site; the height of the barriers and the kinetics of crossing these barriers will provide information on the timescale of binding. The different time- and length scales related to fluctuations in different parts of the protein may be useful in the reaction mechanism to discriminate among the different binding sites. Because ubiquitin has a highly conserved primary sequence in the family of proteins that have similar functions (for example, the primary sequence of ubiquitin has only a few residues that are different in animals, yeasts, and plants), we expect the mechanisms that guide these processes to be kinetically and thermodynamically robust.

The LE4PD approach effectively projects the dynamics of a protein onto a coarse-grained (CG) description where the protein is represented by a collection of vectors connecting pairs of α-carbons (C_α). The method starts with a molecular dynamics (MD) simulation of a protein in physiological conditions in the canonical (NVT) ensemble, where data were collected from a 1-μs equilibrium simulation. Then, it decomposes the MD dynamics by projecting the trajectory, which represents the complex dynamics of a protein with motion coupled across multiple length- and timescales, onto quasi-linearly-independent LE4PD normal mode coordinates derived from the CG description. While here we use MD simulations in the canonical ensemble, the LE4PD equation is equally useful when starting from other ensembles, statistical averages derived experimentally (e.g. NMR conformational ensembles[48, 57] or a set of X-ray crystal structures), or by Monte Carlo simulations. The conversion into normal mode coordinates yields a description of the local

fluctuations that is largely uncoupled, and can be analyzed independently: each normal-mode trajectory encapsulates the dynamics occurring at a selected length- and timescale in the simulation. The real-space dynamics can be reconstructed *a posteriori* from a linear combination of the normal modes.

When the simulated dynamics is projected onto the LE4PD normal modes, a free-energy map of the conformational space is generated for each mode. The mode-dependent free-energy maps display complex landscapes with energy barriers and unique pathways between minima on the surface. To calculate the timescale of transition between minima in the Free-energy Surfaces (FES), we combine, here for the first time, the LE4PD normal mode description with a mode-dependent Markov State Model (MSM) analysis of the dynamics. MSMs have been applied to the study of the kinetics of a wide range of biologically relevant systems, providing a reliable analysis of the dynamical pathways.[33, 79–83] Here we propose a refined MSM method for the determination of the slow kinetic transitions between minima in each free-energy map. Using this approach we evaluate the characteristic time of transition between two well-defined energy minima in LE4PD mode-dependent FES of ubiquitin.

Decoupling the real dynamics by decomposition into independent normal modes is similar in purpose to the Principal Component Analysis (PCA) and the time-lagged Independent Component Analysis (tICA) approaches, which have been previously used in conjunction with MSM.[84–86] Interestingly, because of their direct connection with the physical picture of the system, the LE4PD modes identify the contributions that arise from the type of amino acids and their local flexibility, hydrodynamics, and friction within the protein.[38, 48, 57, 67] Thus the projection onto the LE4PD diffusive normal modes provides a detailed physical interpretation of the dynamics of

the protein: in a given time window and at a given spatial scale, fluctuations occur on well-defined fragments of the protein primary sequence, or, equivalently, one can see how different parts of the protein become dynamically active (fluctuate) on different time scales. This study shows that the newly proposed LE4PD-MSM method allows for the careful and accurate evaluation of mode-dependent local fluctuations and kinetic pathways.

From Molecular Dynamics Simulations to a LE4PD Normal Mode description

The LE4PD is a linear Langevin equation of motion for a set of coarse-grained units (or beads) located at the position of the alpha-carbon along the primary sequence of a protein.[37, 38, 48, 57, 67] The LE4PD is expressed as a function of the time-dependent C_α-C_α bond coordinates, $\vec{l}_i(t)$, which allows one to discard the center-of-mass diffusion, irrelevant in the study of the internal dynamics of a protein. The LE4PD equation of motion for the bond vector i is

$$\bar{\zeta}\frac{\partial \vec{l}_i(t)}{\partial t} = -\frac{3k_BT}{l^2}\sum_{j,k}^{N-1} L_{i,j}U_{j,k}\vec{l}_k(t) + \vec{F}_i(t), \tag{2.1}$$

where k_B is the Boltzmann constant, T the temperature in Kelvin, and l^2 the mean-square bond length, which in this model is the mean-square peptide bond length.

$\vec{F}_i(t)$ is the stochastic force acting on bond i at time t, which is governed by a white-noise fluctuation-dissipation relation

$$\langle F_i^\alpha(t)F_j^\beta(t')\rangle = 2\bar{\zeta}k_BT\delta_{\alpha,\beta}\delta(t-t')\sum_{k,m=1}^{N-1} a_{i,k}\delta_{k,m}a_{m,j} \tag{2.2}$$

with $\langle \vec{F}_i(t) \rangle = 0$ and where $\langle\rangle$ defines the statistical average of a quantity . Here α, β denote Cartesian indices; δ_{ij} and $\delta(t-t')$ are the Kronecker delta and Dirac delta function, respectively. The matrix **a** is the matrix that transforms from bead coordinates to bond coordinates, while **L** is the matrix that contains the hydrodynamic interaction and **U** is the inverse of the bond correlation matrix $(\mathbf{U}^{-1})_{i,j} = \langle \vec{l}_i(t) \cdot \vec{l}_j(t) \rangle / (\langle |\vec{l}_i(t)| \rangle \langle |\vec{l}_j(t)| \rangle)$, which provides information on the protein local flexibility along its primary sequence. For convenience we define $\sigma = 3k_BT/(\overline{\zeta} l^2)$.

The statistical averages that enter the LE4PD hydrodynamic and structural matrices, as well as the amino acid friction coefficient, are calculated from trajectories of atomistic MD simulations of the protein in aqueous solvent at physiological conditions. More details on the LE4PD and on the MD simulations of ubiquitin analyzed in this and succeeding chapters are reported in Appendix A.

The LE4PD equation of motion is solved by matrix diagonalization to recover the linearly independent mode representation of the dynamics. The eigenvalues of the diffusive matrix for mode a are defined as λ_a^{LE4PD}, while the normal modes are $\vec{\xi}_a(t) = \sum_{i=1}^{N-1} \mathbf{Q}_{a,i}^{-1} \vec{l}_i(t)$, with **Q** the matrix of the eigenvectors that diagonalize the LE4PD equation (Eq. 2.1). The modes span the same space as the bond vectors with near linearity, as $\langle \vec{\xi}_a(t) \cdot \vec{\xi}_b(t) \rangle \cong \delta_{a,b} l^2/\mu_a$ with $\mu_a = \sum_{i,j} \mathbf{Q}_{a,i}^{-1} \mathbf{U}_{i,j}^{-1} \mathbf{Q}_{j,a}$. Starting from the normal mode solution of Eq. 2.1, one can calculate any structural and dynamical property of interest. A comparison of real-space structural properties predicted by the LE4PD to those calculated directly from the simulation trajectory demonstrates the accuracy of the Langevin mode description.[38]

The LE4PD normal modes, $\vec{\xi}_a(t)$, describe the dynamics of a protein over a given length scale. For a protein possessing N CG sites, the LE4PD gives $N-1$ coupled equation of motion in the bond coordinates and $N-1$ uncoupled equations of motion

in normal modes. The modes are ordered by descending diffusive timescale, with the low index modes describing the slowest, and generally highest-amplitude motions, while the highest index modes describe the fastest, lowest-amplitude motions. For a well-folded protein, the first three modes describe the rotational diffusion tensor of the protein [38]; these are the *global* modes of motion. The remaining $N-4$ modes describe internal deformations of the protein and will be referred to as *internal* modes.

We then project the MD trajectory for each bond vector onto the LE4PD mode coordinates as $\vec{\xi}_a(t)_{[trajectory]} = \sum_{i=1}^{N-1} \mathbf{Q}^{-1}_{a,i} \vec{l}_i(t)_{[trajectory]}$, thus yielding a new trajectory in the mode coordinates. Then an energy map for each mode is built by calculating the histogram of the trajectory projected onto the mode coordinates expressed in real space.[38] The map of the free-energy is reported as a function of the two spherical angles in polar coordinates: θ (describing the inclination relative to the z-axis of the simulation box) and ϕ (describing the azimuthal angle into the xy-plane of the simulation box). Note that the contour plot of the energy as a function of the modulus of the polar vector does not show relevant features and is not reported, even if its contribution enters in the normalization of the probability and thus in the evaluation of the mode-dependent free energy. The latter is given by the logarithm of the normalized probability of each molecular configuration in mode coordinates. This procedure yields $N-1$ mode-dependent FES.

As an example, Figure 1 displays the free-energy surface of mode seven for ubiquitin. The contour map of the free-energy surface displays interesting features with localized, deep minima and distinct reaction pathways between them. By applying a version of the string method[87, 88] two possible pathways emerge between the two minima, given the circular symmetry of the angle ϕ. The transitions between two minima in the free-energy landscape represent mode dependent local fluctuations,

associated with a timescale for crossing the related energy barrier, or transition time. A more in-depth description of these local fluctuations are presented in the following section.

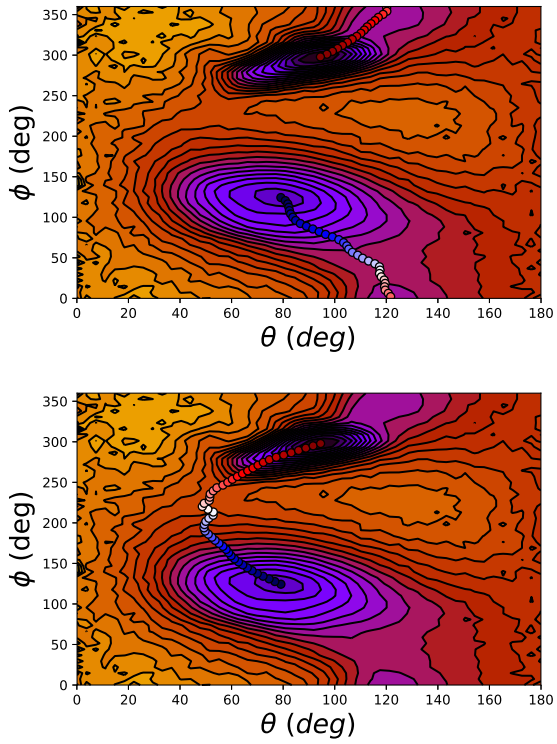

FIGURE 1. An example of the contour plot for the free-energy surface of a LE4PD mode; the example here shows data for mode seven. The red-white-blue circles show two possible low-energy pathways between the minima.

Localized Fluctuations in ubiquitin LE4PD Modes

Ubiquitin is a regulatory protein present in eukaryotic cells, whose post-translational modification is involved in multiple biological functions.[75, 89] An important function of ubiquitin is to tag misfolded proteins and to signal them for degradation via the proteasome. For this reason ubiquitin is called the "molecular kiss of death". Degradation happens through a number of steps where first ubiquitin

binds to the misfolded protein by a reaction that is catalyzed by ubiquitin ligases, and once the first ubiquitin binds, this signals the ligase for further binding of the ubiquitin proteins to form a polyubiquitin chain. The polyubiquitin chain finally binds to the proteasome, which ultimately degrades the misfolded protein: thus ubiquitin molecules form a chain that connects the misfolded protein to the proteasome.[90] The length and shape of polyubiquitin are important for the successful protein degradation.[52]

Binding of ubiquitin to a second protein by mono- or poly-ubiquitination occurs by formation of an amide bond between the carboxyl group of the last amino acid in the C-terminal tail and either the ϵ amino group in the side chain of a lysine residue or, alternatively, the amino group in the N-terminus. Ubiquitin itself has seven different Lys groups that can bind to the C-terminus of another ubiquitin: Lys6, Lys11, Lys27, Lys29, Lys33, Lys48, and Lys63 (see Figure 2). The selection of the different Lys groups for binding yields different three-dimensional structures for the resultant polyubiquitin and supports different biological processes.[52, 91] For example, the formation of polyubiquitin that leads to protein degradation is initiated by the binding of the C-terminus of the second ubiquitin to either Lys48 or Lys29 in the first ubiquitin, i.e. the one directly bound to the protein that is being degraded. Binding of ubiquitin to Lys63, instead, leads to polyubiquitin chains that are important for other functions generally related to crossing of a membrane, including for example endocytosis, membrane trafficking, and signal transduction.[64, 75]

Furthermore, it has been shown that ubiquitin may interact with other proteins through non-covalent binding, and that this non-covalent binding involves conserved regions of the protein: the hydrophobic Ile44 patch on the surface of ubiquitin binds

non-covalently to a number of ubiquitin binding domains, for example the ubiquitin interacting motif (UIM) and motif interacting with ubiquitin (MIU).[91]

Thus, given the complexity of the possible kinetic pathways involved in its binding, an accurate evaluation of the dynamics of ubiquitin is potentially enlightening in determining how this protein carries out its biological function: fluctuations may be important in characterizing the binding propensity of different regions of ubiquitin's primary sequence in agreement with the theory of conformational selection, which postulates that a protein will sample all its relevant binding conformations even in the absence of any binding partners.[11, 92] As demonstrated below, the LE4PD coupled with the MSM is effective at describing dynamics at localized sites of ubiquitin involved in both covalent and non-covalent binding.

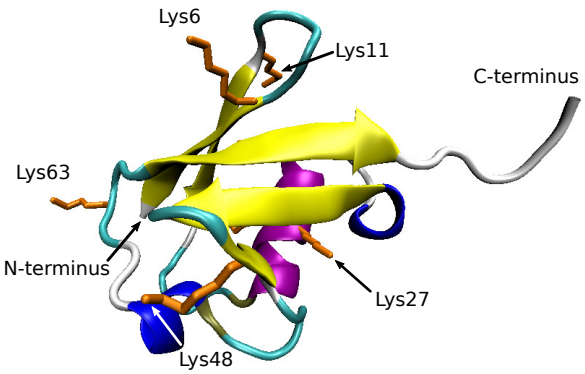

FIGURE 2. Cartoon representation of ubiquitin, with the lysine residues plus the N- and C-termini labeled. The lysines' side chains are drawn explicitly and colored orange. Two of the seven lysine residues are obscured and not labeled.

For a given LE4PD mode, the transition between two minima in the FES corresponds to well-defined local fluctuations. For each bond and a specified mode, a, the amplitude of the bond fluctuations is calculated as the Local Mode Length scale

(LML):

$$LML = \sqrt{L_{i,a}^2} = \sqrt{\frac{Q_{i,a}^2 l^2}{\mu_a}}. \tag{2.3}$$

$L_{i,a}^2$ is the mean-squared projection of mode a onto the i^{th} C_α-C_α bond.

Examples of $L_{i,a}$ are reported in Figures 3 and 4 for a LE4PD analysis applied to a 1-μs equilibrium simulation of ubiquitin. Figure 3 displays the LML for the first ten LE4PD internal modes: those are the slowest modes that contain crossing of high energy barriers and possibly rare events. For the high-amplitude, slow global modes in Figure 3 fluctuations are located mostly in the C-terminal tail (residues 71-76) of ubiquitin, which is involved in poly-ubiquitination.[75] In the slower modes smaller amplitude fluctuations are also located in the Lys11 loop (residues 6-11), while those fluctuations become dominant only for the faster and more local LE4PD modes 7, 11, and 13.

An interesting exception to the observed trend is LE4PD mode 9, which shows a large amplitude fluctuation in the stretch between residues 51 and 63; this segment is known as the 50 s loop and has been shown to be a binding site for the $A20$ zinc-finger binding motif.[59, 60, 63] The specific behavior of this LE4PD mode will be elaborated more below because it is an example of the LE4PD's ability to predict dynamics in binding regions of ubiquitin and the MSM's ability to describe accurately the slow kinetics along the LE4PD mode's FES.

Figure 4 shows the LML for a number of high index modes. The LML for these internal modes display delocalized low-amplitude fluctuations across ubiquitin's primary sequence. Dynamics on these local energy maps does not involve transition between deep wells on the free-energy landscape but rather diffusion over a rough landscape. Interestingly, this is not observed for the very last modes (modes 73-75), which describe highly-localized fluctuations along the backbone that are sensitive to

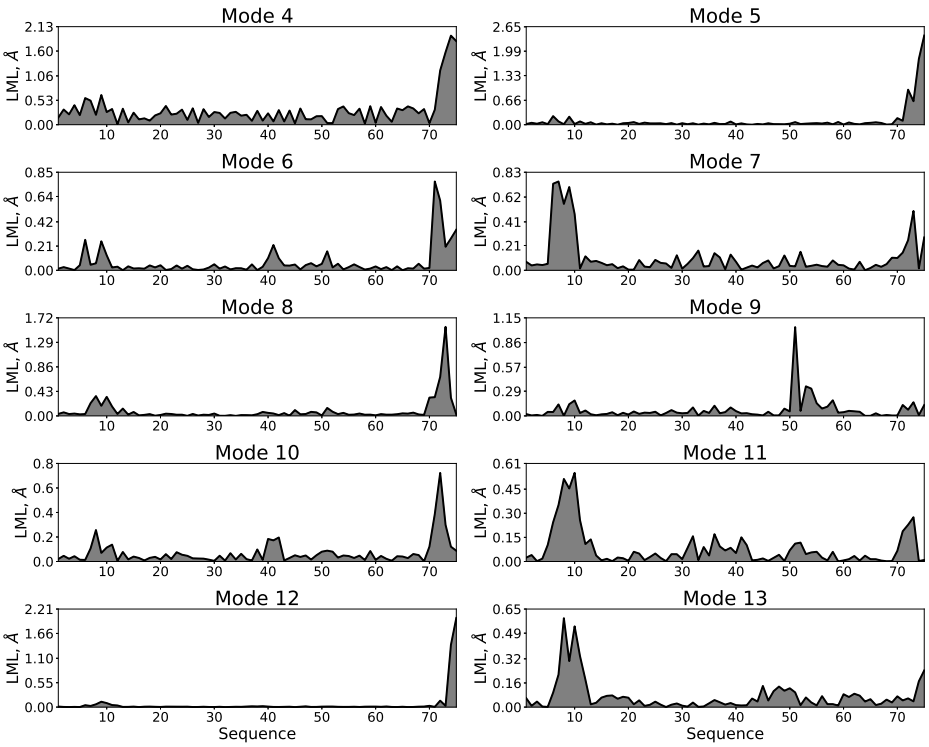

FIGURE 3. Local mode length scale (LML), a measure of mode-dependent fluctuations along the protein ubiquitin's primary alpha-carbon sequence, for the first 10 internal modes predicted by the LE4PD; these are the modes amenable to Markov state modeling.

the chemical specificity of ubiquitin's primary sequence (more details are available in the Supplementary Material of [4]).

The analysis of the LML suggests that fluctuations do not appear uniformly in all the modes, but that their localization along the primary sequence of the protein is specific of the mode number. Low index modes show fluctuations that are local in space and that occur by transition along a pathway between well-defined energy wells. Intermediate-index modes, instead, show an almost stochastic spreading of the fluctuations and delocalization along the primary sequence of the protein. The FES of high-index modes show a well-defined, highly-conserved, localization corresponding to crossing of local energy barriers of order k_BT.

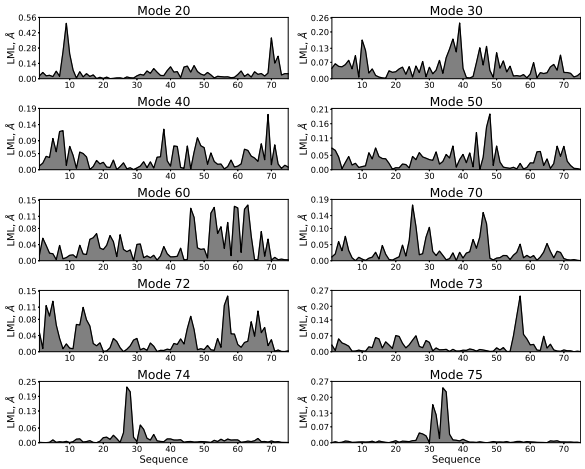

FIGURE 4. Local mode length scale (LML), a measure of mode-dependent fluctuations along the protein ubiquitin's primary alpha-carbon sequence, for a sampling of the higher-index, faster internal modes predicted by the LE4PD.

Each mode-dependent FES is associated with a timescale characterizing the fluctuations described by that mode. Thus, it appears that the different binding regions, including different lysine residues, on the surface of ubiquitin are kinetically non-equivalent: this suggests preferential binding depending on the timescale of the kinetic processes involved. To calculate the timescale associated with the crossing of the free-energy barrier for a given fluctuation we apply an MSM analysis as illustrated in the following sections.

Markovian and Non-Markovian Kinetics of the Mode-Dependent LE4PD Fluctuations

The trend observed in the mode-dependent fluctuations of Figures 3 and 4 is indicative of the structure of the related FES. Thus, one observes that low-index LE4PD modes correspond to large wavelength processes, and identify slow, cooperative dynamics of the amino acids along the primary sequence. For those slow, cooperative motions, free-energy barriers are large enough and crossing is often a

rare event. Thus, for the low-index modes kinetic transitions are uncorrelated and a Markovian statistic applies, and the mode-dependent dynamics can be analyzed by MSMs.

As one moves to consider higher-index LE4PD modes the dynamics becomes faster and the height of the free-energy barrier decreases.[57, 67] It was previously shown that the overall scaling behavior of the height of the average energy barrier scales with the mode number as $E_a^\dagger \propto (a-3)^{-0.5}$, and with the characteristic mode length, $L_a^2 = l^2/\mu_a$, as $E_a^\dagger \propto (L_{(a-3)})^{0.93}$.[67] These characteristic scaling exponents are consistent with a dynamical process where the constant fluctuations in hydrogen bonding is a source of energetic disorder in the Hamiltonian of a protein, thus supporting the mapping of protein dynamics onto the Kardar-Parisi-Zhang model.[57, 67]

Thus, the Markovian nature of the kinetics for each LE4PD mode depends on the mode that is under study. The decrease of the height of the energy barriers with increasing locality of the dynamics renders progressively more problematic the MSM analysis of the FES. For high index mode the free-energy barriers are not large enough and the trajectory cannot sample low free-energy regions for a sufficient length of time to completely lose memory of the previous transitions. For those modes the dynamics is not Markovian independent of the lag time that is selected, and the MSM approach does not apply. For these high-index modes where the roughness of the landscape dominates the dynamics, it is appropriate to adopt Kramers' diffusive renormalization of the friction coefficient, where the energy barrier is calculated through the Median Absolute Deviation (MAD) measure of the average energy roughness.[57, 93]

In the MAD measure, the free-energy barrier is calculated as

$$E_a^\dagger = \text{median}\left(E_a(\theta,\phi) - E_{\min,a}\right), \tag{2.4}$$

with $E_a^\dagger$ the free-energy barrier predicted for mode a using the MAD and $E_{\min,a}$ the global free-energy minimum of mode a.

Using the measured MAD barrier we defined an effective friction coefficient that yielded a slowing down of the dynamics calculated with Kramer's theory of diffusive barrier crossing.[38, 48, 67] By assuming that the modes diffuse along their respective FES, and that barrier crossing is a thermally-activated process, the friction coefficient can be rescaled as $\overline{\zeta} \to \overline{\zeta}\exp\left[E_a^\dagger/k_BT\right]$, leading to a slowed down decay time for each mode: $\tau_a \to l^2\overline{\zeta}\exp\left[E_a^\dagger/k_BT\right]/(3k_BT\lambda_a^{[LE4PD]}) = \tau_a^0\exp\left[E_a^\dagger/k_BT\right]$, with $\tau_a^0 = \left(\sigma\lambda_a^{[LE4PD]}\right)^{-1}$.

The use of this rescaling procedure for the mode-dependent time improves the agreement between the bond time autocorrelation function, calculated from the LE4PD theory, and the same function directly calculated from the MD simulations.[38, 48, 67] The criterion that we adopt to establish the range of LE4PD modes where the MSM applies, i.e. the Markovian nature of the kinetic transition, is a standard procedure based on indirectly measuring the Chapman-Kolmogorov condition through the implied timescales test, as illustrated in Appendix B. For the protein ubiquitin studied here, we observe that for the tenth LE4PD internal mode and higher a MSM analysis becomes impossible, and transition times are calculated by simple Kramers' rescaling of the LE4PD mode-dependent time using the MAD determination of the average energy barrier. This procedure appears to be accurate, also because the weight of each mode in the calculation of any time correlation function decreases with increasing LE4PD mode number: the possible error due to

a less accurate evaluation of the high-index modes (by MAD instead of MSM) plays a less significant rôle in the calculation of the dynamical properties of the protein in the form of time correlation functions.

A Markov State Model for the Analysis of the Mode-Dependent free-energy Surface

The MSM method models the kinetics of a molecular process as a Markov chain of uncorrelated jumps among conformational states; some additional details on MSMs are available in Appendix B. The mode-dependent trajectory is first partitioned into a finite number of discrete states, W, using the k-means++ clustering algorithm, as implemented in PyEMMA[85]. Once a lag time τ is selected, the probability of transition between different microstates is calculated and stored in the transition matrix, $\mathbf{T}(\tau)$. Thus, the evolution of the probability for the system to occupy a discrete state at a given time t, $\mathbf{p}(t)$, follows the equation[33]

$$p_j(t+\tau) = \sum_{i=1}^{W} T_{ij}(\tau) p_i(t), \tag{2.5}$$

where the matrix $T(\tau)$ is calculated from the simulation trajectory using the reversible maximum-likelihood estimate,[94]

$$T_{ij}(\tau) = \frac{(c_{ij} + c_{ji})\pi_j}{c_i \pi_j + c_j \pi_i}, \tag{2.6}$$

with $c_{ij} = c_{ij}(\tau)$ the ij^{th} element of the count matrix, which keeps track of all the transitions from state i to j in the trajectory at a lag time τ. We define $c_i = \sum_j c_{ij}$ the i^{th} row sum of the count matrix, giving the total number of observed transitions

from i, while π_i is the stationary (equilibrium) probability of state i. This definition of $T(\tau)$ satisfies detailed balance and implies reversibility of the kinetic process.

The right eigenvectors of $\mathbf{T}(\tau)$ are solutions to the eigenvalue equation

$$\mathbf{T}(\tau)\psi_i(\tau) = \lambda_i(\tau)^{[MSM]}\psi_i(\tau). \tag{2.7}$$

Since $\mathbf{T}(\tau)$ is a regular, stochastic matrix, the Perron-Frobenius theorem guarantees $\lambda_1(\tau)^{[MSM]} = 1$ is the maximum eigenvalue of $\mathbf{T}(\tau)$, and its corresponding eigenvector, ψ_1, has only positive entries.[95] The other eigenvalues obey the condition that for $i > 1$ one has $0 < \lambda_i(\tau)^{[MSM]} < 1$. Given the definition of the implied timescales, t_i,

$$t_i = -\frac{\tau}{\ln[\lambda_i(\tau)^{[MSM]}]}, \tag{2.8}$$

one observes that all the processes decay in time, excepting the one corresponding to the first eigenvalue ($t_1 = \infty$). The first left eigenvector describes the stationary distribution of the configurational states, while the eigenvectors with index higher than one describe kinetic transitions occurring at increasingly smaller timescales. The second eigenvalue, $\lambda_2(\tau)^{[MSM]}]$, gives the timescale, t_2, associated with the slowest internal motion of the protein from the given LE4PD dynamical mode. It is precisely this timescale t_2 that is of interest here because it describes the slowest kinetic process occurring on a FES. Following the procedure presented in the next section, the time t_2 is selected to correspond to the kinetic transition between the two minima in the FES of the slow LE4PD modes.

A Kinetically Informed Determination of the Mode-dependent LE4PD transition time

We calculated the mode-dependent transition time for crossing a free-energy barrier, τ, starting from the spectrum of the second right eigenvector, ψ_2, of $\mathbf{T}(\tau)$.[96] This eigenvector has been called an 'ideal reaction coordinate', because i) it corresponds to the slowest, non-stationary process of the Markov state model, where the slowest dynamics is supposed to be representative of rare events and ii) it has been shown to be a good approximation to the discrete committor function[34] (see Equation 2.9), which gives the probability of the trajectory to visit the product state before the reactant state, when initiated at any of the discrete states outside of the reactant and product regions[96, 97] (as an example of ψ_2 acting as a committor function, compare Figures 6 and 8).

To identify the relevant well-to-well transition time, we performed a series of calculations at increasing lag time, τ, and inspected the structure of the second eigenvector (ψ_2) projected onto the free-energy surface. The transition time was selected such that the discrete states with the most positive projection along ψ_2 were located in the deepest well in the free-energy surface (the 'product' state), while the discrete states with the most negative projection were located in a second, well-defined free-energy well (the 'reactant' state). This selection criterion was adopted because the goal is to construct kinetic models for each LE4PD mode describing the barrier-crossing events between the two most highly-populated regions of the free-energy surface for each LE4PD mode. An example of the resulting ψ_2 spectra are reported in Figures 5 and 6, where the free-energy maps for modes 4, 5, 7, and 9 are presented on the left panels and the discrete states for the second eigenvectors are superimposed to the FES on the right panels.

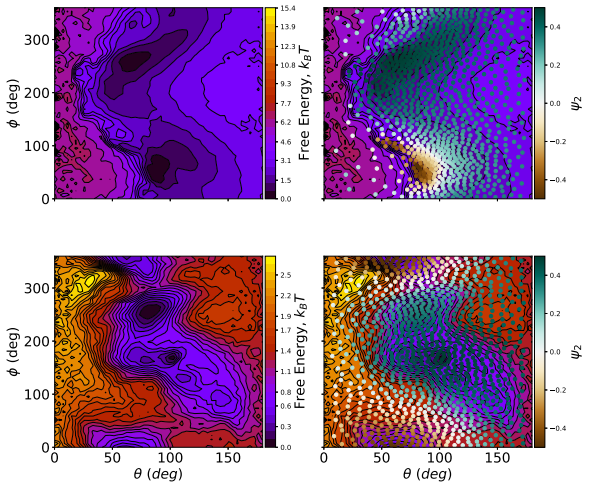

FIGURE 5. Second right eigenvector, ψ_2, of the transition probability matrix projected onto the discrete states of the kinetic models for LE4PD modes 4 (top) and 5 (bottom). For clarity of visualization and ease of comparison, a scaled version of ψ_2 is plotted: $\frac{\psi_2 - \min(\psi_2)}{\max(\psi_2) - \min(\psi_2)}$. The scaling of the contour levels of the free-energy surface is given by the colorbar on the left-hand subplot and the scaling of the eigenvector projection is given by the colorbar of the right-hand subplot.

Figure 7 reports the kinetic parameters measured using the MSM method combined with the 'kinetically informed' procedure, and a comparison of these times with the results of the Kramers rescaling procedure with the MAD determination of the energy barrier, henceforth referred to as the 'LE4PD-MAD' approach. Note that while the transition time of the slowest modes, which have pronounced and localized energy barriers, are calculated with the kinetically informed MSM procedure, for modes larger than 13 (the eleventh internal mode and higher) MSM can not be applied and the LE4PD-MAD procedure is used. Using the proposed method to construct the MSM for each LE4PD mode, the kinetics predicted by the discrete model, which corresponds to the transition time between wells on the surface, can be seen to have a similar trend in the timescales to the LE4PD-MAD approach, while the LE4PD with the kinetically informed MSM generally predicts slightly larger timescales. For

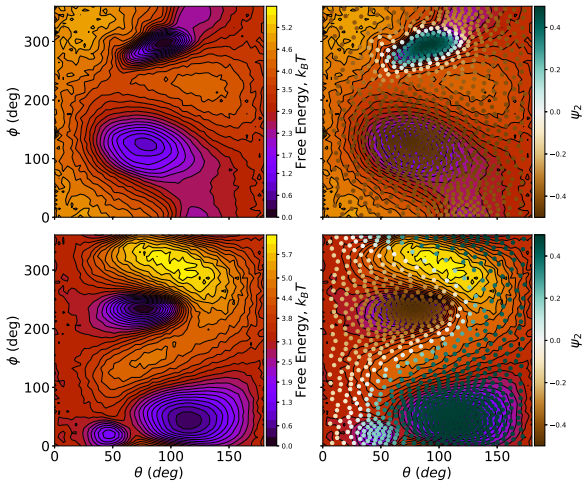

FIGURE 6. Second right eigenvector, ψ_2, of the transition probability matrix projected onto the discrete states of the kinetic models for LE4PD modes 7 (top) and 9 (bottom). For clarity of visualization and ease of comparison, a scaled version of ψ_2 is plotted: $\frac{\psi_2 - \min(\psi_2)}{\max(\psi_2) - \min(\psi_2)}$. The scaling of the eigenvector projection is given by the colorbar of the left-hand subplot and the contour levels of the free-energy surface are given by the colorbar on the right-hand subplot.

a given normal mode, the kinetically informed 'LE4PD-MSM' procedure appears to provide a more accurate determination of the implied timescale.

It is worth noticing that, if too long a lag time is selected in the kinetically-informed MSM method, the discrete states corresponding to the minimum and the maximum projections along the second eigenvector are empirically found to no longer lie within wells in the FES, as is shown for LE4PD mode 8 in Appendix B. Thus, although long-time processes become naturally uncorrelated at large lag time, and thus Markovian, the characteristic timescale of a given fluctuation has to correspond to the transition between two well-defined energetic minima. Finally, we note that the eigenvalues of the transition matrix, Equation B.2, are identically to those of the corresponding, symmetrized matrix.[96] Then, the transition time for a dynamical fluctuation is uniquely defined because its value is independent of which well is

assumed to be the initial state and which is assumed to be the final state in the free-energy barrier crossing.

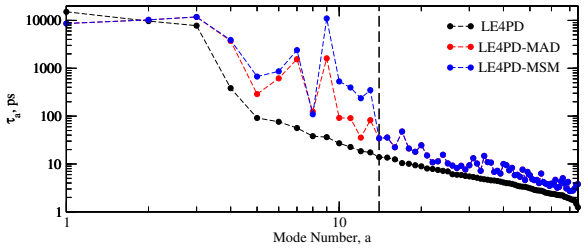

FIGURE 7. Comparison of the measured transition times as a function of mode number calculated using the kinetically informed MSM approach (blue) and compared with Kramers' diffusive energy barrier LE4PD-MAD results (red), and the predictions of LE4PD without including mode-dependent free-energy barriers (black).

Fluctuation Dynamics and Binding: the case of Mode 9

Interestingly the timescale of LE4PD mode 9 predicted from the MSM analysis has a transition time that is comparable to the rotational dynamics of the protein and the fluctuation time of the first internal mode. This indicates that important slow dynamics occurs for this internal mode. This long transition time is due to the structure of the free-energy surface for mode 9, which is shown in the left panel of Figure 8. Visible in Figure 8 is a prominent, third minimum in the FES. Using a committor analysis,[96] this minimum is shown to lie near a committor value of 0.5 and serves as a 'trap state' during transitions between the reactant and product states on the FES of LE4PD mode 9. The aforementioned committor analysis is performed using the standard approach of transition path theory.[98] The committor function for a discrete state i, q_i, is defined as [98]

$$-q_i + \sum_{k \in I} T_{ik} q_k = -\sum_{k \in P} T_{ik} \tag{2.9}$$

with I the set of intermediate states (all discrete states not belonging to the reactant or product states) and P the product set. Conceptually, the committor function describes the probability that a trajectory initiated from state i will visit (or commit) next to the product state.[34, 96] That is, $q_i = 0$ in the set of discrete states defined as the 'reactant' states and $q_i = 1$ in the set of discrete states defined as the 'product states.' Discrete states with $q_i \approx 0.5$ are transitions states, with approximately equal probability of visiting either the reactant or product state next.

The right panel of Figure 8 shows the committor function between the discrete states with the most positive and most negative projections along ψ_2; discrete states with committor values near 0 are colored dark red and discrete states with committor values near 1 are colored dark green. Since the trap state is located at an intermediate committor value ($\sim$0.5-0.6), transitions between reactant and product states get trapped there. This trapping effect extends the transition time between regions of high and low projections along ψ_2, leading to longer predicted timescales relative to the LE4PD-MAD procedure; the latter averages out these types of effects because it accounts only for a single characteristic barrier height. Thus, this mode provides a clear example of the extra information that may be afforded from adopting a more precise method of analysis of the dynamics, by using MSMs to model the kinetics of the slow LE4PD modes.

The fluctuations observed in the LML for LE4PD mode 9 are reproduced qualitatively when the transition between the reactant and product wells on the FES is modeled using a transition path found by a modified version of the zero-temperature string method.[87, 88] The minimum free-energy path is shown in the top panel of Figure 9 and the corresponding structural deformations of ubiquitin as the trajectory moves along this pathway are shown in the bottom panel; additional details regarding

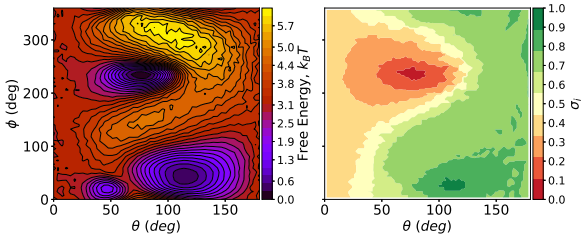

FIGURE 8. Free-energy surface (left) and the discrete committor function, q_i, superimposed on the free-energy surface for mode 9. Dark red regions denote where the committor is near 0 (reactant states), dark green regions denote where the committor is near 1 (product states), and yellow regions denote intermediate values of the committor, where the trajectory has a significant probability of visiting either the reactant or product state next. The minimum at the lower, left-hand corner of the free-energy surface is located at an intermediate committor value and serves as a trap state, which imparts the long transition time (≈10 ns) to ψ_2 for the MSM of this LE4PD mode.

the string method parameterization for the LE4PD free-energy surfaces is available in the Supplemental Material of [4]. Figure 9 demonstrates that the minimum free-energy path between wells passes through the 'trap state' mentioned above, which supports the slow timescale for this mode predicted by the MSM. The fluctuations along the primary sequence of ubiquitin, given in the bottom panel, support the localization of the fluctuations predicted by the LE4PD's LML.

Biological Interpretation

As previously observed, the conformational selection model of protein binding postulates that in the absence of its binding partner(s) a protein will still sample all energetically available states, including those states responsible for binding to the ligand(s) of interest.[11, 92] The conformational selection model implies that the C_α-C_α bond fluctuations described by the dynamics along the LE4PD FES may identify the timescales and length scales of relevant binding modes of, in the analysis shown here, ubiquitin. Thus it is useful to summarize the results that we obtained in this

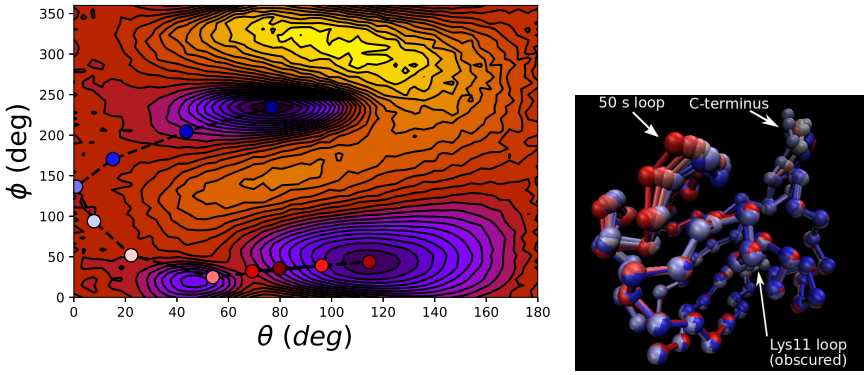

FIGURE 9. Dynamical motion of ubiquitin along the minimum free-energy pathway of LE4PD mode 9. Top: Free-energy surface for LE4PD mode 9. The colored circles discretize the string between the reactant and product states. The dark blue circle represents the reactant state; as the string moves toward the product state, the coloration changes from blue to white to red, with the dark red corresponding to the product state. Bottom: Structural deformation of ubiquitin due to movement along the minimum free-energy path. Structure color in the bottom panel corresponds to the circle color in the top panel.

study while trying to connect those data to a possible interpretation of the propensity of ubiquitin to bind at specific time scales in well-defined regions of the protein three-dimensional structure.

Re-examining the LML for LE4PD in Figure 3 we identify the slowest and more relevant fluctuations of ubiquitin. First, we observe that LE4PD predicts large-amplitude, slow dynamics in two prominent regions: the C-terminal tail (modes 4, 5, 6, 8, and 10) and the Lys11 loop, which is the flexible loop containing Lys11 (modes 7, 11, 13).[77] The C-terminal tail and the Lys11 loop regions are implicated in covalent associations with other proteins, including other ubiquitin molecules; both regions are involved in polyubiquitination events.[52, 75, 77] Due to their ability to bind covalently to numerous other proteins, it is perhaps not surprising that the LE4PD predicts motions in these two regions, which involve a relatively wide window of length- and timescales (Figures 3 and 7). We speculate that perhaps the large

variance of the size and kinetics of the fluctuations involving the C-terminal tail and the Lys11 loop is due to the need for the protein to interact with a variety of other proteins of different size and flexibility.

Different is the result for the fluctuations involving the 50 s loop.[59] In this case, fluctuations appear in a specific LE4PD mode (mode 9). X-ray crystallography, 2D-NMR experiments, and immunoblot assays have shown this region of ubiquitin is recognized by, and is bound non-covalently to, the A20 zinc-finger (ZnF) motif of Ras guanine exchange factor Rabex-5.[59, 63] Perhaps more interestingly, the X-ray structure of Rabex-5 bound to ubiquitin (PDB ID: 2C7N)[59] shows that Y25 on Rabex-5 forms a hydrogen bond with residue E51 of ubiquitin,[59] while 2D-NMR studies show that E51 on ubiquitin undergoes a large chemical shift perturbation when Rabex-5 is added to a solution of ubiquitin.[63] Interestingly, LE4PD mode 9 shows a large, local fluctuation localized on at E51 residue of ubiquitin (see Figure 3).

Conclusions

Protein dynamics is hypothesized to play a central rôle in protein function. Molecular recognition and substrate binding often occur by a conformational selection mechanism, where protein conformations that are apt to binding a specific substrate are already populated in the isolated protein.[11] Then, the binding of the substrate occurs by a simple selection of the proper protein conformation. Thus, spontaneous fluctuations in the isolated protein may provide useful information on the mechanism of protein substrate interaction and binding, where both structural and kinetic information are important to characterize the mechanisms of binding.

We have presented a detailed study of the fluctuation dynamics of the protein ubiquitin and propose a kinetically-informed method to analyze the protein motion. The protein is simulated using an MD trajectory, and is analyzed following a two-step procedure. In the first step the trajectory is projected onto diffusive normal-modes of a coarse-grained Langevin representation, the LE4PD model. This step separates the dynamics into quasi-linearly independent coordinates and allows for the identification of local fluctuations as a function of their length scale: large, cooperative fluctuations manifest themselves mostly in low-index LE4PD modes, but important exceptions naturally emerge even in this study. Normal-mode LE4PD free-energy surfaces can be constructed, which describe the energy landscape connected with the mode-dependent fluctuations. The kinetic pathways of the mode-dependent fluctuations are analyzed by Markov State Modeling, which provides the transition time for the crossing of the energy barrier for the LE4PD mode-dependent fluctuations.

To analyze the kinetics of transitions between energy wells in the free-energy surfaces of the slow, high-amplitude LE4PD modes we identify the proper timescale of transition between two energy minima using the committor function. Since we seek to analyze the slow kinetic processes predicted by the MSM, we focus mostly on the dynamic process modeled by the first non-trivial eigenmode of the MSM, which is described by the spectrum of the second right eigenvector of the transition matrix, ψ_2, and its corresponding timescale, $t_2 = -\tau/[\ln(\lambda_2^{[MSM]})]$. For low-amplitude, local modes, where the energy barriers are not high and the dynamics is not Markovian, we adopt instead a rescaling of the local friction in agreement with Kramers theory.

This method of combining the LE4PD normal modes to predict the dynamics over a specific length scale, measured using the local mode length scale (LML), and the MSM to predict the kinetics (timescales) of each LE4PD mode is applied to a 1-

μs, equilibrium simulation of the protein ubiquitin, a small, globular protein involved in the post-translational modification of many eukaryotic proteins [52, 64, 75]. The LE4PD-MSM analysis reveals slow dynamics in three flexible regions of ubiquitin: the C-terminal tail, the Lys11 loop, and the 50 s loop (residues 51-63); these three regions all play prominent rôles in ubiquitination pathways, especially in binding by ubiquitin-binding proteins.[52, 59, 63, 64] We find that a single, slow mode is dedicated to describing fluctuations in the 50 s loop while multiple slow LE4PD modes describe motions in the C-terminal tail and the Lys11 loop. This disparity could potentially be due to the small number of proteins that bind to the 50 s loop (which all contain the same ubiquitin-recognition motif, the A20 zinc finger domain[64]) while many proteins bind to the Lys11 loop and C-terminal tail, since these two regions are involved in both ubiquitin recognition and covalent binding to proteins targeted for degradation.[52, 75] Finally, the minimum free-energy pathways along the FES for the slow LE4PD modes reproduce the fluctuations predicted by the barrier-free LE4PD equation of motion, which indicates that the motion between wells of the FES of the LE4PD modes represent well the equilibrium fluctuations of ubiquitin, with the MSM analysis giving an accurate estimate of the timescales of those fluctuations.

Because MD atomistic simulations of proteins display dynamics that is coupled on multiple length scales their interpretation is not easy. The LE4PD method of characterizing mode-dependent fluctuations, starting from atomistic MD simulations, may have some advantages with respect to other approaches commonly used, such as Principal Component Analysis (PCA) or time Independent Component Analysis (tICA). The LE4PD conveniently separates the complex dynamics of a protein into linearly independent normal modes, which can be analyzed individually. With respect to PCA and tICA, LE4PD brings a straightforward physical interpretation

to the dynamics measured, and directly connects the results of the analysis to the primary and secondary structures of the protein, and to the free-energy barriers and hydrodynamics, which affect protein fluctuations.

Furthermore, we have shown here how the LE4PD can be conveniently used as a first step in a MSM analysis. Although the MSM analysis can be mathematically performed in multiple dimensions, it is useful to identify a set of coordinates allowing for a reduction of the multidimensional space into a projected, low-dimensional space where the free-energy landscape is represented as a function of two coordinates. The coordinates may be selected to be the first two principal components or time-lagged independent components, which are the two collective coordinates that maximize the variance and maximize the autocorrelation time, respectively. However, selecting the first two coordinates involves tracing the dynamics along a pair of collective coordinates that have different timescales. In the LE4PD case, one can perform an analysis of the dynamics along a single coordinate at a time, characterized by one time- and one length-scale, which is one of the main advantages of decomposing the protein's dynamics into a set of normal modes.

In conclusion, the mode-dependent LE4PD description presented here appears to be an ideal framework for the analysis of protein dynamics through MSM because it decouples the dynamics into linearly independent modes, thus representing these modes in a low dimensional space that can be easily visualized and conveniently analyzed by MSM. The decomposition in LE4PD normal modes naturally separates the dynamics in independent contributions, where fluctuations occur at a given length scale, while the MSM analysis provides a precise evaluation of the timescale and the kinetic pathway associated with the local fluctuations. Finally, this mode-dependent MSM analysis can reconstruct the dynamics of the specific protein by an inverse

transformation, $\vec{l}_i(t) = \sum_a Q_{i,a}\vec{\xi}_a(t)$, to reveal the real-space dynamics by LE4PD modes, as shown in Figure 9.

Bridge

This chapter has interfaced the LE4PD and MSM techniques to give a precise description of the timescales, lengthscales, amplitudes, and localization of the slowest LE4PD modes calculated from an MD simulation of the protein ubiquitin. Since, at least for highly metastable systems, the second right eigenfunction of the MSM transitions matrix, ψ_2, approximates the committor function,[96] inspection of ψ_2 on the free-energy surfaces of the slow LE4PD modes gives an approximate location of the relevant barrier to the dynamics described by that LE4PD mode. Furthermore, since the spectrum of ψ_2 determines what the slowest process from $\mathbf{T}(\tau)$ is,[33, 34, 99] by aligning the highest and lowest projections of ψ_2 into the energetic minima of the LE4PD free-energy surfaces, we guarantee that ψ_2 describes transitions between these minima.

Thus, the MSM is used to 1) define the slowest process on the LE4PD surface and 2) give that process' kinetics. Since the MSM estimates the eigenfunctions of $\mathbf{T}(\tau)$ using a non-linear approximation,[30, 50] it is able to account for barriers on the surface in a more precise manner than the MAD approach, which still uses the linear transformation of the eigenvectors from the LE4PD $\mathbf{LU}$ matrix, but rescales the friction coefficient to give an approximate first-order correction to the barriers that are removed by performing the analysis in a coarse-grained set of coordinates.[38, 45, 46] While the timescales predicted by the MSM for the slow LE4PD modes of ubiquitin tend to be slightly slower than those given by the MAD approach, the results are in

qualitative agreement, in that both methods agree on the relative order of the slowest modes.

The major biological result from this study is the extraction of the slow dynamics in the 50 s loop of ubiquitin, a region of the protein previously shown to unfold first in a long folding-unfolding simulation of the protein performed by the D. E. Shaw group.[16] This observation tentatively indicates that the slow LE4PD modes can identify the 'leading fluctuations' for the unfolding of proteins and that these modes could be good collective coordinates for performing advanced sampling techniques such as metadynamics [24] to explore more thoroughly the complete conformational space of proteins. This point is elucidated further in chapter V.

The eigenspectrum of the slow processes from the MSM transition matrix $\mathbf{T}(\tau)$ can also be used to coarse grain a free-energy surface into the set of metastable states where the system residues for long periods of time before transitioning to another state.[100, 101] In the next chapter, an MSM is applied to a reduced free-energy surface for the simplest single-stranded nucleic acid system, deoxyadenine dinucleotide (dApdA), to extract which conformations are important to a calculation of the system's circular dichroism (CD) spectrum. We also find that taking a single representative structure from each metastable state, the average structure within each state, the CD spectrum calculated using just the average structures is a good approximation to the CD using all the structures from the underlying simulation.

CHAPTER III

AN APPLICATION OF MARKOV STATE MODELS TO ELUCIDATE THE CONFORMATIONAL DIVERSITY OF DEOXYADENINE DINUCLEOTIDE

From Beyerle, E.R.; Dinpajooh, M.; Ji, H.; von Hippel, P.; Marcus, A.H.; and Guenza, M.G. Dinucleotides as simple models of the base stacking-unstacking component of DNA 'breathing' mechanisms. *NAR*, **49**, (2021), 1872 - 1885. Deoxyriboadenine dinucleotide monophosphate as a simple model for structural and dynamic aspects of DNA 'breathing' in duplex DNA

Introduction

Nucleic acids undergo a variety of local structural fluctuations in discharging their biological functions. These fluctuations (collectively called 'breathing') include inter-strand base-pair opening and closing, intra-strand base stacking and unstacking and conformational rearrangements of the sugar-phosphate backbone.[6, 102–106] Such thermally activated DNA 'breathing' fluctuations are thought to represent primary steps in the process by which genome regulatory proteins gain access to the double-stranded (ds) DNA interior.

Understanding thermally driven DNA fluctuations may provide a central key to structural and dynamic interpretation of the interactions between functional and regulatory proteins and their ss- and dsDNA targets during gene expression. However, many of these 'breathing' processes, if considered only in duplex DNA, are likely to represent a small fraction of the population of conformations present in duplex DNA at physiological temperatures because of 'structural cooperativity' and may thus be hard to resolve even by sensitive spectroscopic and computational techniques. One

way of reducing this problem is to focus on elementary systems, such as dinucleotides. These can be considered to represent the 'fundamental fragments' of duplex DNA, but also provide a milieu in which the only relevant breathing process is likely to be base stacking and unstacking. As a consequence, these processes can be studied in isolation in these small model systems. In addition, because of the absence of constraints imposed by neighboring and base-paired nucleotides, these stacking-unstacking fluctuations are likely to be present at higher concentrations than in larger duplex DNA molecules and thus also more amenable to study. These considerations have motivated us to reinvestigate the structure and dynamics of dApdA as a model dinucleotide fragment of duplex DNA using modern computational and molecular modeling techniques.

The relative populations of stacked and unstacked bases present in DNA molecules in solution under a variety of environmental conditions have traditionally been studied by absorbance and circular dichroism (CD) experiments. [3, 107] Initial studies of DNA stacking-unstacking fluctuations focused on dinucleotides in solution.[1, 108–113] Dinucleotides, such as dApdA, have been considered to be useful models for some of the basic interactions that control and stabilize local base conformations of dsDNA because – as indicated above – stacking interactions can be examined in these systems while avoiding the complicating features of ion condensation, cooperative stacking and inter-base hydrogen-bonding that are also present and involved in controlling the conformational behavior of long duplex DNA. In addition, homo-dinucleotides, such as dApdA, are more useful than hetero-dinucleotides as model systems for probing conformational rearrangements in these structures because the CD signals from homo-dinucleotides are strengthened by the presence of degenerate exciton coupling effects. Furthermore, dinucleotides may also

serve as partial models for deciphering the structure and energetics of some of the more complex elements of biologically important DNA structure, such as the single-stranded (ss) DNA—dsDNA forks and junctions that are essential intermediates in the pathways by which proteins that control genome expression find and interact with their target sites, but in which cooperative interactions and hydrogen bonds between strands are not significantly present.

Base and base-pair interaction free energies have typically been estimated from thermal denaturation studies of DNA oligonucleotides,[114, 115] which showed that among the contributions to the overall interaction free energies of these systems, the free energy of hydrogen bonding between complementary bases and the energetics of configurational and solvent entropy provide only small contributions to the stability of the base paired structures.14 Furthermore, base-base stacking, which is the main (enthalpic) contributor to the stability of dinucleotide conformations, appears also to be the dominant component of the overall stability of more complex DNA structures.[109, 116, 117]

Early studies of dApdA by Schellman, *et al.* [107, 109, 111–113, 116] suggested that the CD spectrum of this dinucleotide in aqueous salt solutions could be represented as the weighted sum of two conformations, one 'stacked' and the other 'unstacked', with the stacked form likely resembling (in terms of base-base overlap and helical pitch) the Watson- Crick B-form characteristic of duplex DNA. Furthermore, these workers showed that the changes induced in the CD spectrum of this dinucleotide by increasing concentrations of monovalent salt (NaCl) could be attributed to shifts in the relative populations of these same two conformations.

However, these interpretations clearly represented over-simplifications of the actual situation, since we now know that the CD spectrum of a given molecule

of this sort must comprise a sum over myriad microstate configurations that exist simultaneously in solution at equilibrium. As a consequence of this complex situation, CD spectra cannot be 'inverted' to determine the conformations that contribute uniquely to the overall spectrum. We here address this problem by means of extensive Molecular Dynamics (MD) simulations and a Markov State Model (MSM) analysis,[4, 49, 118] thus providing information on the major conformations that participate in the stacking-unstacking equilibria of dApdA, and whose excitonic transitions contribute to the overall CD spectrum.

To this end we performed a set of 2 μs MD simulations of the dApdA dinucleotide in aqueous solvent at increasing monovalent salt concentrations, using the same conditions employed for the initial spectroscopic measurements on dApdA dinucleotide.15 From our MD trajectories, each consisting of $\sim 10^7$ microstate configurations, we calculated the CD spectrum by averaging together the contributions from each MD-generated conformation using the standard method [110, 111, 119, 120] together with an extended dipole model (EDM).[120] The initial predictions generated by this method are in good general agreement with experimental spectra previously measured by others.

We next carried out an MSM analysis of our MD trajectories and identified five kinetically stable regions in the free energy landscape, which we refer to as 'macrostates.' Each macrostate contains a 'family' of conformationally-related microstates, which rapidly interconvert. Transitions between macrostates are kinetically uncoupled, because they are separated by high energy barriers and thus follow Markovian statistics.[49] The ensemble of macrostates provides a structural basis that can be used to interpret the experimental spectroscopic measurements. By combining MSM analyses with transition path theory [81, 121–124] we investigated

the kinetic pathways for base unstacking, thus revealing the roles that base 'flipping' appears to play in breathing fluctuations at the dinucleotide level. In addition, we were able to identify one average configuration for each macrostate that served, with sufficient accuracy, to represent the averaged properties of the macrostate. This simplified, five- configuration model retains the important features of the CD spectrum calculated using the full MD statistics and provides a useful minimalistic ensemble for the calculation of CD and potentially also other optical spectra obtained using more sophisticated experimental techniques.

Of the five macrostates, three are statistically the most populated, with the CD spectrum being largely determined as the sum of contributions from only two configurational states, consistent with early experimental observations.[109] While the original studies interpreted those spectra in terms of a single stacked and a single unstacked configuration of dApdA, our analysis shows that, of the two conformational states that contribute significantly to the features of the CD spectra, the most populated corresponds to an ensemble of hybrid dinucleotide conformations that include one base that has flipped into a syn conformation, which in dsDNA results in Hoogsteen base-pairing, [125–132] while the relatively less populated state corresponds to an ensemble in which both bases of the dinucleotide remain in the canonical anti conformation, compatible with right-handed B form (Watson-Crick) base-pairing in dsDNA. The third highly populated dApdA conformation, which is partially unstacked and contains one syn base, does not contribute significantly to the CD signal. However, these results do indicate that conformations compatible with the Hoogsten structure could well play an important role in some types of breathing fluctuations—at least at the dinucleotide level – thus confirming its possible relevance

to biologically important breathing fluctuations in larger DNA molecules as well.[125–132]

Our studies of the orientations and distributions of counterions in aqueous solutions of dApdA have revealed an abrupt structural transition in the positioning and distribution of these ions around the dinucleotide at a NaCl concentration slightly above 1 M, indicating that counterion concentrations are also involved in controlling breathing fluctuations at the dinucleotide level,[117, 133] and thus likely to play a role in the 'breathing' of larger DNA molecules as well. We show that the above abrupt salt-concentration-dependent transition is correlated with a shift in the equilibria between the three most populated macrostates of the dApdA dinucleotide, and is consistent with early thermal studies of DNA stability at increasing monovalent salt concentration.[134–136] We have shown that this transition is not seen in MD simulations of the isolated phosphate anion in ionic solutions, suggesting that this salt-dependent transition depends also on other (uncharged) components of the dinucleotide structure.

Material and Methods

Molecular dynamics (MD) simulations.

MD simulations of the dApdA dinucleotide monophosphate molecule in aqueous solution were performed at increasing salt concentrations ([NaCl]= 0.1, 0.5, 1.0, 1.05, 1.2 and 1.5 M) in the NPT ensemble using the GROMACS software program.[137] The length of the simulation box was allowed to fluctuate, so that the average distance between the box boundary and the dApdA molecule was approximately 20 $\mathring{A}$. The initial configuration for the dApdA dinucleotide was selected as the B-form conformation, for which we obtained atomic coordinates from the ambertools

software package (http://casegroup.rutgers.edu/). Simulations were performed with the Amber03 force-field [138] and the TIP3P water model [139] to model the dApdA molecule and the water component of the solvent, respectively. While these models were not specifically parameterized to achieve accurate CD calculations of the dApdA dinucleotide, they have been used successfully for nucleic acid systems in the past and represent present state-of-the-art for simulations of DNA in solution. A sufficient number of sodium and chloride ions were included to achieve the target salt concentration. The energy of the solvated structure was minimized using the Steepest Descent algorithm for 500 steps. The system was then heated to 300 K and equilibrated in the isothermal- isobaric (NPT) ensemble using a time step of 2 fs over a period between 50 - 100 ns.

Production runs at each salt concentration were performed for a total duration of 2 μs in the NPT ensemble in order to ensure sufficient sampling of the conformational landscape. These simulations used the stochastic velocity-rescaling thermostat [140] with a time constant of 0.2 ps, and the Parrinello-Rahman barostat (using an isotropic pressure coupling time constant of 1.0 ps). We implemented the Leap-Frog algorithm to integrate Newton's equations of motion using the LINCS constraints fourth order in the expansion of the constraint coupling matrix, which included one iteration to correct for rotational lengthening. [141] We set the time step to 2 fs, and truncated the Lennard-Jones interactions using a cutoff distance of 10.0 $\mathring{A}$. We additionally used a particle mesh Ewald sum to handle long-range electrostatic interactions with a real space cutoff of 10.0 $\mathring{A}$ and a grid spacing of 1.0 $\mathring{A}$. The Verlet neighbor list algorithm was applied with a frequency of 10 MD steps to enhance the computational speed. Trajectory frames were stored every 0.2 ps. In Figure 10, we show a sample

frame from one of our MD trajectories. At each salt concentration we included ~ 10 million such frames in our CD calculations for the dApdA system.

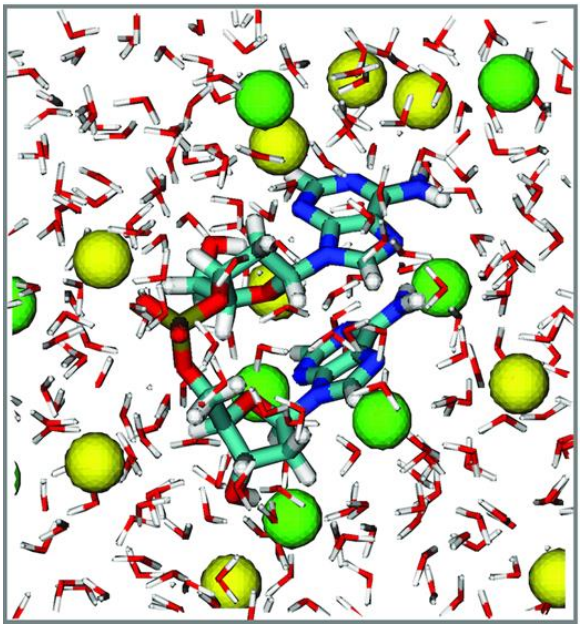

FIGURE 10. A sample configuration frame taken from an MD simulation run of dApdA dinucleotide monophosphate in TIP3P water with [NaCl] = 0.1 M. Sodium ions are shown as yellow spheres, and chloride ions as green spheres. The atoms of the dApdA and water are colored according to CPK rules, except for carbon, which is colored light blue.

Theoretical modeling of circular dichroism (CD).

The CD spectra for the dApdA dinucleotide monophosphate were calculated from the molecular coordinates of each simulation frame by summing over the contributions from each individual k electronic transition, according to $\Delta\epsilon(\nu) = \sum_{k=1}^{n_{\rm tot}} \Delta\epsilon(\nu_k)$. For the kth electronic transition, we approximate its contribution to the CD spectral line shape as a Gaussian function $\Delta\epsilon(\nu_k) = \Delta\bar{\epsilon}_k \exp\left\{-\left[(\nu_k - \overline{\nu}_k)^2 / 2\sigma_k^2\right]\right\}$, where σ_k is the Gaussian standard deviation, $\overline{\nu}_k$ $(= E_k/h)$ is the mean transition frequency and $\Delta\bar{\epsilon}_k = R_k\overline{\nu}_k/A\sqrt{\sigma_k}$ is the magnitude. A is a numerical constant; more details regarding the

theory behind the CD calculations are given in Appendix C . For a given transition k, the rotational strength, R_k, depends on the relative orientation of the monomer electric dipole transition moments, and is calculated from the diagonalization of the Hamiltonian that models the delocalized electronic states of the dApdA dinucleotide as a function of base stacking conformation. The Hamiltonian was formalized using the extended-dipole model (EDM). The parameters adopted in the EDM model are discussed in Appendix C.

Markov State Model analysis.

The MD trajectories were analyzed using the Markov state model (MSM) PyEMMA software program.[85] Briefly, we used the k-means++ algorithm[142, 143] to construct a kinetically-relevant, balanced clustering of the trajectories (using the Euclidean criterion) by partitioning the 107 conformations into 100 initial microstates. A transition rate matrix was constructed for these microstates and then diagonalized into eigenvalues and eigenvectors. From the eigen-spectra of the transition probability matrix, we constructed five macrostates by implementing a minimum error propagation version of the Perron-cluster cluster analysis (PCCA+). We justified our choice for these five macrostates by considering the related conformational landscape and the implied interconversion time scales (additional details are available in the Supplementary information of [66]). Rapidly interconverting molecular conformations were assigned to the same macrostate, while slowly interconverting conformations, which are separated by large barriers, occur between conformations that lie within different macrostates. By identifying and separating slowly interconverting conformations from rapidly interconverting ones, the MSM ensures that the slow processes obey Markovian statistics. To sample slow transitions, we adopted a lag-

time of 500 ps, and confirmed that under these conditions Markovian behavior was satisfied by checking that the Chapman-Kolmogorov condition applies[99, 144, 145] (see Fig. S5 of the SI). We then calculated the CD spectrum for the configurations of dApdA that are contained in each macrostate.

Results

Structural parameters of dinucleotides.

As pointed out above, it has long been known that under physiological salt conditions the adjacent bases of each strand of duplex DNA in aqueous solution adopt helical conformations close to the Watson-Crick B-form, with an average inter-base separation ~ 3.5 $\mathring{A}$ and a relative twist angle $\sim 36^\circ$ (see Figure 11 for parameter definitions). Spectroscopic studies of small oligonucleotides in solution have examined the various contributions to base stacking stability in duplex and ssDNA – i.e., the effects of hydrophobic bonding, backbone interactions, inter-base hydrogen bonding and cooperativity.[107, 116]

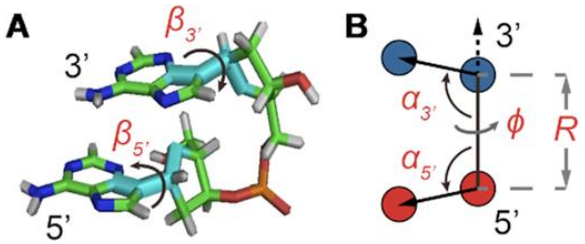

FIGURE 11. Structural coordinates for the dApdA dinucleotide monophosphate used in this chapter. (A) An atomic scale structure is shown with interbase roll angles $\beta_{3'}$ and $\beta_{5'}$. (B) Virtual atoms are shown with blue and red spheres positioned within the planes of the $5'$ and $3'$ bases, respectively, with inter-base separation R, tilt angles $\alpha_{3'}$ and $\alpha_{5'}$, and dihedral twist ϕ (see [66] for further details).

Free energy landscapes as a function of structural parameters and varying salt concentration.

Prior studies of the dApdA dinucleotide monophosphate used CD spectroscopy to investigate changes in base conformation as a function of salt concentration, in order to elucidate the roles of the solvent ions in controlling dinucleotide conformation.[107, 116][107, 116] These studies concluded that the predominant conformation for these truncated ssDNA molecules at physiological salt conditions is a stacked form close to the right-handed Watson-Crick B-form conformation, and that increasing the salt concentration appeared to destabilize this B-form conformation. As we discuss further below, the results of our analyses suggest that the dApdA system is, in fact, more accurately described as an equilibrium distribution of primarily three distinct stacked conformations.

We performed MD simulations of dApdA in aqueous solution at increasing salt concentrations with [NaCl] = 0.1, 0.5, 1.0, 1.05, 1.2 and 1.5 M, as described in the Materials and Methods (MM) section (see Figure 10 for a snapshot of a sample configuration from the 0.1 M simulation). For each 2-μs simulation run, we constructed a histogram representing the probability $P(R, \phi)$ of finding a configuration at a given value of the structural parameters R and ϕ and the related free energy values, $G(R, \phi) = -k_B T \ln(R, \phi)$ (see also Figure S4, which reports the parameters adopted to calculate the FES and perform the MSM analysis, and related discussion in the SI). An example of such a free energy contour diagram plot is shown in Figure 12A. To ensure that the FES represents the system at equilibrium, we show, in Fig. S8 of the SI, the time autocorrelation function of the fluctuations in inter-base separation. Because the function is found to decay in $\sim$ 10 ns, which compares well with the 2 μs of simulation time, the system can be considered to be in equilibrium.

Two additional sets of orientational coordinates per base – the base tilt angles, $\alpha_{3'}$ and $\alpha_{5'}$, and the roll angles, $\beta_{3'}$ and $\beta_{3'}$ – are needed to fully specify the dinucleotide conformation. However, our results indicate that the positions of the local minima in the FES depend largely on the inter-base separation R and the dihedral twist angle ϕ and are less sensitive to changes in the tilt and roll angles. While all of the above structural parameters are specified in our calculations of the CD spectra and of structural and dynamical distribution functions, the visual representation of the Free Energy Surface (FES) is conveniently reported as a function of R and ϕ. The FES $G(R, \phi)$ of the dApdA dinucleotide shown in Fig. 3A applies to [NaCl] = 0.1 M, which is close to the monovalent salt concentration under physiological conditions. Using the same procedure, we also determined $G(R, \phi)$ at increasing salt concentrations (surfaces not shown). To test the validity of the FESs shown in Fig. 3A, we used the molecular configurations obtained from our 2-μs MD trajectories to calculate the CD spectra for dApdA. The results, as a function of salt concentration and using the procedures described in the MM and Appendix C, are shown in Figure 12B.

For the lowest salt concentration, [NaCl] = 0.1 M, we compared our calculations to the experimental CD spectrum of the dApdA dinucleotide obtained under these same conditions (see Figure 12B).[1] We obtained good agreement between the experimental and calculated CD in the long wavelength region of the spectrum (240 – 300 nm). In principle one could achieve a quantitative agreement with the experimental spectrum by optimizing some parameters in the calculation of the theoretical spectra, and thus obtain a better fit of the theoretical predictions to the experimental spectra. However, given the number of possible adjustable parameters, such a procedure would not provide any new information. We prefer, instead, to independently set the parameters in our calculations and then to discuss their

predictions. We note that the agreement is less favorable in the short wavelength region (200 – 240 nm) of the spectrum, where the peak features are slightly blue-shifted and exhibit smaller amplitudes than the experiment. This latter disagreement is not surprising, given that the CD spectra at shorter wavelengths are strongly perturbed by the high density of nearly degenerate electronic states, which makes the theoretical methods we employ in our calculations less accurate in this wavelength range.

In general, we find that the positions of the local minima within the free energy surfaces do not change with salt concentration, while their relative stabilities and equilibrium distributions do depend on this variable. The FES in Figure 12A shows that the dApdA dinucleotide exists primarily as a mixture of the two chiral conformations with opposite handedness ($\phi = 40°$ and -80°) and nearly stacked inter-base separation $R = 3.8$ $\mathring{A}$, together with an achiral conformation that shows no stacking of the bases ($\phi = 0°$) and a significantly larger inter-base separation $R = 4.7$ $\mathring{A}$. Henceforth, we will designate as 'chiral' a conformation that exhibits chiral stacking of the bases, and as 'achiral' conformations with no stacking of the bases, even though some components of the molecule, like the sugar, do of course retain their 'chemical chirality.'

To study the effects of increasing salt concentration on the population of the chiral and achiral conformational states, we report in Figure 12C the local probabilities calculated as the sum of the states contained within the areas of the FES defined by the red and white squares (panel A), respectively, for the chiral state with coordinates (3.8 $\mathring{A}$, 40°) and for the achiral state with coordinates (4.7 $\mathring{A}$, 0°) as a function of the salt concentration. We note that as the salt concentration is increased to [NaCl] = 0.5 M, the local probability of the chiral state with coordinates

(3.8 $\mathring{A}$, 40°) increases, while the weight of the achiral state slightly decreases. A further increase of the salt concentration to [NaCl] $\sim$ 1 M begins to destabilize both of the stacked conformations in favor of the unstacked one, with the weight of the achiral state with coordinates (4.7 $\mathring{A}$, 0°) increasing strongly. We observe a similar dependence on the salt concentration for the CD spectrum, which depends on the distribution of stacked bases. In Figure 12D, we plot the difference CD spectrum for incremental changes of the salt concentration. For incremental increases of the salt concentration below [NaCl] = 1 M (0.1 $\rightarrow$ 0.5 M, 0.5 $\rightarrow$ 1.0 M), the difference CD spectrum shows little variation. However, for the incremental increase of 1.0 $\rightarrow$ 1.5 M, the difference CD spectrum undergoes a pronounced change. This change is also reflected by the value of the peak-to-peak amplitude of the difference CD spectrum (i.e., the difference between the positive peak value at 245 nm and the negative peak value at 270 nm), which is shown in the inset of Figure 12D. Note that these findings are in agreement with the salt-dependent changes in the CD amplitude of this dinucleotide reported in the work of Johnson and Schleich.[116]

The above findings are in qualitative agreement with experiments involving the thermal melting of duplex DNA structures in NaCl, where increases in the concentration of monovalent ions tend to first stabilize the stacked conformation, resulting in an increase in the melting temperature. Then, at higher salt concentration (around [NaCl] = 1 M) this trend reverses, and the further addition of counterions slightly decreases the stability of the dsDNA conformation.[134–136] For duplex DNA, this behavior is generally explained by assuming that an increase in salt concentration facilitates the screening of the negative charges situated on the phosphates in the DNA backbone, rendering the backbone more stable. However, at monovalent salt concentrations around 1 M, the concentration of ions in solutions becomes

equivalent to the concentration of counterions closely bound to the phosphate backbone under ion condensation conditions. As a consequence, additional increases in salt concentration cannot further stabilize the double helix and other mechanisms (presumably 'Hofmeister effects' [117, 146–149]) come into play. Mechanisms involving the stabilization of long duplex DNA molecules by screening the repulsion between backbone phosphates cannot apply to dApdA, since only one phosphate is present. However, the counterions can alter the relative stabilities of the various conformations available to the dApdA dinucleotide by effectively neutralizing the negative charge of the single phosphate.

Distributions of ions and water molecules around dinucleotides.

To examine the roles of salt concentration on the observed structural transition we used the results of our MD simulations to calculate the distributions of the ions and water molecules of the solvent environment in the immediate vicinity of dApdA. This study provides physical insights into the origins of the changes in equilibrium base stacking conformations of this dinucleotide with increasing salt concentration. [1, 6, 102–104] The radial distribution function (RDF) of species j around species i is defined:

$$g_{i--j} = \frac{\langle \rho_j(r) \rangle}{\langle \rho_j \rangle} \tag{3.1}$$

Figures 13A and 13B show the RDFs of the dApdA system at the lowest and highest salt concentrations we examined; i.e., [NaCl] = 0.1 and 1.5 M, respectively. The position-dependent oscillations of the RDFs reflect the local solvation shells of the water hydrogen atoms and of the ionic species relative to the central phosphate. At salt concentrations close to physiological conditions, ([NaCl] = 0.1 M Figure 13A), the phosphate is coordinated with concentric ion shells, with the water hydrogen atoms

forming interstitial layers between the shells. The RDF for water hydrogen atoms appears to be independent of salt concentration, with its first peak centered at r = 2.8 $\mathring{A}$ and its second peak at r = 4.2 $\mathring{A}$. The RDFs for sodium and chloride ions, on the other hand, oscillate at half the spatial frequency of that of the water hydrogen atoms. The RDF for sodium ions has its first peak at r = 3.6 $\mathring{A}$, which coincides with a trough for the water hydrogen atoms at this distance. Similarly, a trough for sodium ions occurs at r = 4.2 $\mathring{A}$, which coincides with the second hydration shell for the water hydrogen atoms. The first ion shell for chloride ions occurs at r = 5.8 $\mathring{A}$, which is the same position as the second ion shell for sodium ions. In general, the nth chloride ion shell occurs at approximately the same position as the (n+1)th sodium ion shell, indicating that these ion shells have mixed compositions. As shown in Figure 13B, the relatively well-defined boundaries between successive ion shells seen at the lowest salt concentrations become diffuse at the highest salt concentration tested ([NaCl] = 1.5 M).

Our observation of a well-ordered structure of successive ion shells at low salt concentration is largely consistent with simple models of counterion condensation, which is an important contributing factor to the stability of larger nucleic acid molecules.[150, 151] Figures 3.1C and 3.1D show, respectively, the RDFs of sodium and chloride ions, each as a function of salt concentration. For both ions, the RDFs appear to change little over salt concentrations between [NaCl] = 0.1 – 1.0 M, yet exhibit an abrupt loss of ion shell structure at salt concentrations slightly greater than 1.0 M.

To illuminate the role(s) of the adenine bases in this situation, we performed a set of 400 ns simulations of $H_2PO_4^-$ at increasing monovalent salt concentration ([NaCl] = 0.1, 0.5, 1.0 and 1.5 M), and studied the ion distributions around a singly-charged

phosphate ion, $H_2PO_4^-$, in aqueous solution (see Figures. 13E and 13F). In $H_2PO_4^-$ we observed an alternating structure of positive and negative ion shells consistent with simple models of counterion condensation. However, we found no signature of the abrupt disruption of the ion shell structure at salt concentrations greater than 1.0 M that was observed with the dApdA dinucleotide. We next turned our attention to a closer examination of the solvent orientation around dApdA. As mentioned previously, the structure of the water, which is reported as the position-dependent RDF of the water hydrogen atoms relative to P, $g_{\mathrm{P}--\mathrm{H}}(r)$, does not change significantly with salt concentration (see Figures 3.1A and 3.1B). More detailed behavior is observed in the position-dependent orientational distribution function (ODF) of the water dipole moment as a function of its separation from the central P atom. The ODF is defined as the average cosine, $\langle\cos(\theta)\rangle$, of the angle θ that subtends the permanent dipole moment of the water molecule, $\vec{\mu}_{\mathrm{H_2O}}$, and the vector connecting the P atom to the water O atom, $\vec{\mathrm{PO}}_{\mathrm{H_2O}}$, as shown in Figures 14A.

Figure 14B shows the ODFs of water relative to the central phosphate of dApdA as a function of salt concentration. It also shows the RDF of the water hydrogen atoms. The position dependence of the ODFs shown in Fig. 5B exhibits damped oscillations that vary across successive hydration layers for all salt concentrations, similar to the behavior observed for the ion shell structures shown in Fig. 4. The ODFs show a sharply pronounced feature centered at $r = 2.8\ \mathring{A}$, which is coincident with the first peak of the RDF. The shapes of the underlying distributions of the angle θ within a narrow range of distances r ensures that $\langle\cos(\theta)\rangle \approx \cos\left(\langle\theta\rangle\right)$ (orange points in Figure 14B). The distributions of the angle θ for a given hydration shell, with each distribution corresponding to one orange point in Figure 14B, are reported in Fig. S7 of the SI. Thus, the narrow feature at $r = 2.8\ \mathring{A}$ has an approximate

peak value of $\cos(\langle\theta\rangle) = -0.8$, which indicates that the water H atoms within this first hydration shell are highly oriented with dipole moment $\vec{\mu}_{H_2O}$ directed toward the central P. Furthermore, the presence of the broadened shoulder centered near the second hydration layer (at $r = 4.2$ $\mathring{A}$), with peak value approximately $\cos(\langle\theta\rangle) = -0.6$, indicates the preferential orientation of the O-H bond vectors of water molecules within the second hydration shell towards the oxygens of water molecules within the first hydration shell. We thus see that hydrogen bonding interactions between water molecules of the first and second hydration shells are stronger than the Coulomb interaction between the negatively charged phosphate and the water dipole moments of the second hydration shell. We further note that the distribution of angles θ over a given range of distances r broadens nonuniformly as the distance from the central P increases, indicating the presence of hydrogen bonding between successive hydration layers and the ensuing loss of orientational correlation between the water dipoles and the central P. At the separation $r = 5$ $\mathring{A}$, the values of the ODFs are approximately zero, indicating the absence of orientational alignment. A recurrence of partial orientational order occurs at separation $r = 6$ $\mathring{A}$, which appears to coincide approximately with the position of the first ion shell of the Cl- ions.

We note that the ODF exhibits a weak, but clear, dependence on the salt concentration. For the case of [NaCl] = 0.1 M, the sharp feature at $r = 2.8$ $\mathring{A}$ indicates a pronounced orientation, which becomes slightly less ordered for [NaCl] = 0.5 M. At the higher salt concentrations of [NaCl] = 1.0 M and 1.5 M, the orientation of the water dipole moments become slightly more ordered. The changes in the ODF of the water molecules as a function of increasing ion concentration are small. Rather, the leading factor in determining the stabilities of the conformations of the dinucleotide structure in solution appears to involve the distribution of monovalent

ions, and the modification of this distribution with increasing ion concentrations (see Figures 13A – D). Water, however, does appear to play a role through its orientation, which is both distance and weakly salt- concentration-dependent. Interestingly, this study also shows that the stabilization of dApdA stacking by increasing counterion concentrations, and the observed sharp transition of the ion structure around 1 M are dependent on the presence of the bases of the dApdA dinucleotide, and do not occur when the ionized phosphate molecule is present alone (Figures 13E – F). The consistency of the observed trend with the effects of increasing salt concentration on the experimental melting curves of DNA suggests that this ion-related base stacking mechanism of DNA stabilization is already present and operational, even at the level of the isolated dinucleotide.

Markov state model analysis of the free energy landscapes of the dApdA dinucleotide and comparison with CD spectral analysis.

The theoretical representation of the CD spectrum for a flexible molecule in solution is the summation of contributions from the myriad microscopic conformational states (i.e. microstates) that exist at equilibrium. Intuitively, we expect the dApdA dinucleotide to fluctuate between various 'open' and 'closed' base conformations, which in turn are stabilized (or destabilized) by the surrounding hydration and ion shells. We first determined the CD spectrum by summing over equally weighted contributions from the 10 million microstates that are sampled from each of our MD simulations (see Figure 12B). Although the above 'brute-force' approach is straightforward, it suffers from two significant limitations: (i) it provides little insight into the interpretation of CD in terms of specific molecular conformations; and (ii) it becomes computationally inefficient if one adopts more

sophisticated quantum chemical models to calculate the CD spectrum beyond the extended dipole model used here, because one would need to perform advanced calculations for each of the 10 million microstates. In reality, only a relatively small subset of the total number of possible conformational states is expected to contribute significantly to the measured CD spectrum. The specific states that dominate the CD are the stacked and chiral conformations of the dinucleotide, for which both the electronic coupling between monomer electric dipole transition moments (or EDTMs) and the rotational strengths resulting from these couplings are significant (see SI). Conformational states that are unstacked, in addition to those that are stacked and essentially achiral, contribute much less to the CD spectrum.

To determine the dApdA configurations that are most relevant for the interpretation of the CD spectra, we used a Markov state model (MSM) analysis[4, 49, 118] to subdivide the 10 million microstates obtained from our MD simulations into a relatively small number (five) of 'macrostates', each of which is associated with a distinctive region of the free energy landscape (see Figure 15A). Each macrostate represents a collection of conformationally related states, or 'microstates', which rapidly interconvert during the simulation, while slow transitions between macrostates require crossing large energy barriers. Starting from the MSM analysis, we calculated the CD spectrum as the sum of those, unequally weighted, macrostates.

Thus, the kinetic processes that occur in the simulations are partitioned between those that occur faster than the 'lag time' (in this study $\tau = 500$ ps) and those that occur more slowly than this time scale. Transitions between conformations within a given macrostate occur frequently and are non-Markovian, while transitions between conformations belonging to different macrostates occur less frequently and are Markovian. The 'lag time' is defined as the time that fulfills the above-stated

condition of Markovian transitions between conformations belonging to different macrostates (for details see SI and, in particular, Fig. S5, which tests the Chapman-Kolmogorov condition, thus ensuring the Markovian nature of our partitioning of the FES into five states. Table S3 shows that the MSM analysis is insensitive to the choice of the number of microstates).

To further reduce the computational requirements for the calculation of the CD spectrum, we identified one averaged structure, together with its relative weight, for each of the five key macrostates that are relevant to the CD observable. We found that the total CD spectrum can be accurately represented by the weighted sum of the contributions from these five averaged structures (see Figs. S6 A-F in the SI), which could be used for modeling of the CD spectrum using more advanced quantum chemical models.

In Fig. 15A, we show the free energy landscape for the dApdA dinucleotide monophosphate in 0.1 M salt (NaCl), and its subdivision into five macrostates (indicated by dark blue contour lines and labeled S1 – S5), which we established using our MSM analysis approach. Each of the five macrostates exhibits qualitatively different behavior in terms of the relative stabilities of the dinucleotide conformation. The S1 macrostate includes 264,608 microstate configurations (2.6% of the total 10 million) and is dominated by a relatively shallow free energy basin with a narrow range of values for the inter-base separation $R < 6$ $\mathring{A}$ and relative twist angle: $100^\circ < \phi < 180^\circ$. The S2 macrostate, on the other hand, describes a relatively broad and featureless region of the free energy landscape, which encompasses a wide range of 'open' and 'unstacked' values for the inter-base separation ($R < 10$ $\mathring{A}$) and unrestricted twist angle: $-180^\circ < \phi < 180^\circ$. Like the S1 macrostate, the S2

macrostate represents a minority of the total population, with just 276,786 microstates (2.8% of the total 10 million).

The majority of the total conformation population is contained in the combined S3, S4 and S5 macrostates, with the number of microstates in S3: 895,636 (9.0%); in S4: 3,729,206 (37.3%); and in S5: 4,833,758 (48.3%). Moreover, the S3 macrostate contains the free energy minimum with $R = 3.8$ $\mathring{A}$ and $\phi = 40°$, the S4 macrostate contains a minimum with $R = 3.8$ $\mathring{A}$ and ϕ = -80°, and the S5 macrostate contains a minimum with $R = 4.7$ $\mathring{A}$ and $\phi = 0°$. We thus identify the S3 macrostate with an ensemble of stacked right-handed chiral conformations that include the *anti* form; the S4 macrostate with an ensemble of stacked left-handed chiral conformations, including the precursor of the Hoogsteen structure; and the S5 macrostate with one slightly less stacked and more achiral conformation, which also includes a *syn* structure.[125–132] The borders between macrostates show a 'fine structure' that represents the maximum of the energy at the top of the free energy barriers, where the states are less frequently sampled by the simulation. Thus, the high energy regions in the free energy map may display roughness, which can be smoothed to avoid overfitting.[101, 152] However this step is not needed in our study because the results of our analysis depend largely on the minima of the free energy maps, which are statistically well sampled. To confirm the presence of Hoogsteen-like structures in the S4 and S5 regions, we present – in Figs. 15B and 15C – a study of the roll angles $\beta_{5'}$ and $\beta_{3'}$, for the 5′ and 3′ base, respectively. It is known, for structures containing a Hoogsteen conformation, that one of the two bases in the dApdA dinucleotide is 'flipped' relative to the 'standard' conformation characteristic of the Watson-Crick geometry. Figure 15B shows that in the S3 macrostate the most stable structures have a positive roll angle $\beta_{5'}$ for the 5′ base (green). Figure 15C shows, instead, that while the roll angle for the 3′ base is

still positive in the microstate S3, the same 3′ base is flipped in microstates S4 and S5 (purple), confirming the presence of a syn Hoogsteen-like conformation in macrostates S4 and S5, and the anti Watson-Crick-like form in macrostate S3.

In Figures 17A – E, we compare the experimental CD spectrum (dashed black curve) to our CD calculations corresponding to each of the five macrostates (blue curves), which are based on summing over the microstate configurations that lie within the partitioned boundaries of the free energy landscape shown in Fig. 6A. We also show the proportionately weighted contribution of each macrostate to the CD spectrum (red). From Figures 17A and 17B, we see that the S1 and S2 macrostates, which represent minority fractions of the total population (2.6% and 2.8%, respectively), do not contribute significantly to the CD spectrum.

Similarly, the S3 macrostate (Fig. 7C), which contains the coordinates of the B-form conformation, also represents a comparably small fraction of the total population (9.0%). On the other hand, the S4 macrostate (Figure 17D) contains a significant fraction of the total population (37.3%) and is largely composed of left-handed base-stacked conformations, which gives rise to a strong CD signal. We note that the calculated CD spectrum of the S4 macrostate has a similar 'right-handed' shape (in the long wavelength regime) to that of the S3 macrostate, in spite of their apparent opposite chiral symmetries. This is consistent with the syn structure (i.e., with roll angles $\beta_{3'} \approx 180^\circ$ and $\beta_{5'} \approx 0^\circ$). The detailed calculation of the spectra for all five macrostates is reported in Figure 16, which shows the spectral decomposition of the degenerate CD spectrum for the average structure of each macrostate. From the spectral decomposition it is straightforward to see that the flipping of one base is responsible for a CD spectrum that is consistent in the Watson-Crick structure and in the *syn* conformation of dApdA. We note that this behavior may not be observed

in dinucleotides with different base compositions, because the transition dipoles are different. Although the S5 macrostate (Fig. 7E) represents the highest fraction of the total population (48.3%), it is dominated by an achiral and slightly unstacked *syn* conformation which, because of its symmetry, results in a negligible CD contribution to the total spectrum. In Fig. 17F, we show the individual weighted contributions for the S3 (9.0%), S4 (37.3%) and S5 (48.3%) macrostates, in addition to the weighted sum of all the macrostates (gray curve). We thus see that the favorable agreement we observe between experiment and theory in the long wavelength regime is essentially the result of two significant contributions, a minor contribution from macrostate S3 and a larger contribution from macrostate S4.

Having identified the key macrostates relevant to the CD observable, we used this information to determine the smallest number of structural parameters necessary to characterize these macrostates. We thus identified five averaged conformations, one for each macrostate, which, properly weighted, were used to calculate the CD spectrum. The comparison between the contribution to the CD spectrum from all the conformational states in a macrostate and the contribution from the averaged macrostate structure are shown in Fig. S6 of the SI, with structural parameters listed in Table S6. The calculation of the CD spectrum with only five conformations is in good agreement with the complete calculation, while it greatly speeds up the computation time needed to calculate the CD spectrum. In principle, such structural models can be used for the general interpretation of any spectroscopic measurement performed on the dApdA system.

In reconsidering the previous interpretations of the CD spectrum by Lowe and Schellman, and given that the signal from the un-stacked mono- nucleotide is comparatively negligible, our study suggests that the stacked 'native' form of the

dinucleotide is primarily given by the sum of the S3 and S4 states, because the S1 state is less densely populated. Analogously, the unstacked 'denaturate' state corresponds in this study to the S5 state, which is more populated than the unstacked S2 state.

The large degree of conformational disorder that characterizes macrostate S2 contrasts with the highly ordered macrostates S3 and S4. The stabilities of macrostates S3 and S4, relative to macrostate S5, are reminiscent of the 'solvophobic' models for nucleic acid base stacking,[103, 110, 111] in which the 'stacked' macrostates S3 and S4 are favored due to enthalpic base stacking interactions,[153] which offsets the configurational entropy of the disordered S2 macrostate. Solvophobic base stacking is known to be favored by a decrease in the enthalpy ΔH and opposed by a decrease in the entropy ΔS. Solvophobic bonding, as defined here, is 'enthalpically driven' and differs significantly from hydrophobic bonding, which is generally thought to drive protein folding[154] by a positive change in the entropy of the system. Such physical models are supported by studies that examine the stabilizing and destabilizing effects on base stacking by various salts and other solvent additives.[109, 116]

Mean first passage times (MFPTs) for dApdA macrostates and pathways of macrostate interconversion.

While CD spectra provide a useful measure of the stationary (equilibrium) properties of the dApdA system, they do not provide information about the dynamic processes involved in state-to-state interconversion. In this section we apply the results of our MSM analysis of MD trajectories to the investigation of the kinetic pathways associated with the free energy landscape, and to identify pathways of interconversion between the various stacked and unstacked macrostates.

To characterize the kinetic properties of the dApdA system, we examined the mean first passage times (MFPTs) of the five macrostates, which are assigned to the regions of the free energy landscape shown in Figure 15A. The MFPT $\tau_{i\to f}$ is the average time for the system to undergo a transition to state f, provided that it was initially in state [118, 155] We determined the MFPTs for the free energy landscape of dApdA at salt concentration [NaCl] = 0.1 M (see full data set in Table S4 of the SI. Also, Table S5 shows that the MFPTs are insensitive to the number of microstates selected in the MSM analysis). Macrostate S2 represents the region of the free energy landscape with the greatest degree of conformational disorder; thus, it can be considered to serve as an end-state for base-unstacking.

Moreover, while macrostate S3 is approximately B-form in character, the relative roll angles of macrostates S4 and S5 are greater than 90°, which in each case corresponds to a base configuration that has been flipped into the Hoogsteen-like conformation. Thus, the process of 'base-flipping' may play an important role in the dynamics of the dApdA system, although in longer strands of (especially) duplex DNA, such flipping may be suppressed by the overall cooperativity that controls the order-disorder transitions for these larger macromolecular species. Nevertheless, these studies of the less cooperatively stabilized dinucleotide may provide insight into structural rearrangements that in principle could, and likely – with some frequency – do, occur in larger biologically relevant DNA macromolecules.

We use the transition path theory (TPT) method [81, 121–124] to determine the frequency of events in which an initially base-stacked macrostate (e.g., S3 or S4) undergoes successive conformational changes that permit entry into the region of the free energy landscape characterized by the 'final' unstacked macrostate S2. When the system initially occupies macrostate S3, which corresponds to the average base

stacking of the Watson-Crick B-form, we found that the dominant pathway leading to macrostate S2 (with 46% probability) was S3 → S5 → S2. Thus, base-unstacking from the right- handed B-form conformation occurs predominantly by a two-step process through the achiral S5 intermediate, in which one of the adenine bases has been flipped. The remaining, less prevalent base-unstacking pathways were S3 → S4 → S2 (with 26% probability); S3 → S5 → S4 → S2 (with 15% probability); and S3 → S2 (with 10% probability). When the system occupied initially the left-handed and base-flipped macrostate S4, which corresponds to a Hoogsteen base-stacking configuration, the two most prevalent unstacking pathways were the one-step S4 → S2 pathway (with 47% probability) and the two-step S4 → S5 → S2 pathway (with 40% probability).

We see that, in general, transitions to the most sparsely populated macrostates S1 and S2 occur relatively slowly (in ˜35 to 60 ns), while transitions to the most highly populated macrostate S5 occur relatively quickly (in ˜2 to 5 ns), suggesting that the macrostate S5 acts as a common intermediate for the pathways between the other macrostates for the stacking-unstacking transition.

Because the energy barriers in dApdA are small, the height(s) of the barrier(s) that the system has to overcome to transition between any two macrostates is close to the difference in free energy between the two states. Thus, the kinetics of the interconversion between macrostates are driven primarily by their relative stabilities. It is reasonable to expect that cooperativity in base stacking will increase the heights of the energy barriers between conformational states in both the ssDNA and the dsDNA. Such information can provide new insights into the mechanisms of base stacking-unstacking transitions in nucleic acids and the possible role of these processes in biologically important protein-nucleic acid interactions.

Discussion

Structural and dynamic characterization of 'breathing' fluctuations at the dinucleotide level.

Thermally activated breathing fluctuations, in which flanking nucleic acid bases spontaneously move away from their stacked and hydrogen-bonded conformations, are thought to be important initial steps in the pathways that lead to DNA denaturation and the specific binding of proteins to DNA.[6, 102–106] Despite their relevance, the details of the interactions and kinetics that control breathing fluctuations are still largely not understood. It is known, however, that the stacking interactions of the bases within nucleic acids are the dominant stabilizing forces of the native conformations that oppose the melting of DNA, while inter-strand base-base hydrogen bonding and cooperativity play less important stabilizing roles.[109, 115, 153] Traditionally the equilibrium between stacked and unstacked base conformations has been studied by circular dichroism (CD) experiments, which are sensitive to the conformational chirality of the base stacking.[1] Such measurements, however, are limited in the amount of information that they can provide because CD spectra cannot be directly inverted to determine the conformations that contribute to these spectroscopic signals.

CD spectroscopy is an important biophysical tool for the analysis of nucleic acid structure, in that the relationship between CD spectra and local nucleic acid base conformation can be understood in terms of quantum chemical principles. Nevertheless, for many of these systems the free energy landscape can favor the simultaneous presence of multiple conformations at equilibrium, many of which may interconvert due to thermal fluctuations. Thus, the complexity of the free energy

landscapes of nucleic acid systems is a significant obstacle for achieving a meaningful interpretation of CD spectra.

Solvophobic effects on the conformational stability of dinucleotides.

Early studies by Lowe and Schellman of the base stacking-unstacking equilibrium focused on the CD spectrum of the dApdA dinucleotide monophosphate as a function of increasing monovalent salt concentration, because the stacking interactions of the elementary dinucleotide unit could be isolated and studied independently of other stabilizing factors.[109] These studies concluded that the stacking-unstacking equilibrium of dinucleotides can be modeled as a two state transition, where the driving force for the stacking of the bases is 'solvophobic' in nature; i.e., driven by a decrease in the enthalpy of the process ($\Delta H \approx -6.6$ kcal mol^{-1} at T = 293 K), and opposed by a decrease in the entropy of the system ($\Delta S = -23$ e.u. such that -T$\Delta S \approx 6.7$ kcal mol^{-1} at T = 293 K).[1, 109, 113, 153] Thus, these workers concluded that the transition as a whole was likely driven – to a major extent – by rearrangements of the molecules of the solvent environment present (here water molecules and ionic species). However, these studies could not exclude the possibility that more than two states might contribute to the overall CD signal, and thus could not define the precise nature of the underlying conformations.[109] They did determine, however, that each of the two states of the dApdA dinucleotide that contributed to the CD signal was most likely present as a number of similar configurations, and that the state with highest disorder and entropy was likely to be more stable at high temperatures and at higher monovalent salt concentrations.

Conclusions and Overview.

In the present study we have established a methodology that can be used to relate the CD spectrum to the underlying relevant molecular conformations. We combined extensive MD simulations (μs in duration) with direct calculations of the CD spectrum. Our CD calculations were based on standard methods[120] and an extended-dipole model (EDM)[156] to estimate the exciton coupling between the electric dipole transition moments (EDTMs) of the adenine bases of dApdA. The EDM takes into account the finite length of the electronic transition charge distribution across the adenine chromophore, and it correctly describes the dependence of the electronic coupling on the inter-base twist angle ϕ and the relative tilt angle $\alpha_{5'}$ - $\alpha_{3'}$ (coordinates defined in Figure 11). By calculating the CD spectrum for each of the 10 million conformations in the MD simulations, we obtained good agreement between our CD calculations and previously published experimental spectra of the dApdA system at approximately physiological salt concentration [NaCl] = 0.1 M.[1, 109] Nevertheless, the calculation of the CD spectrum by these procedures provided little insight into the important conformational states contributing to the CD spectrum, and can become computationally too expensive if sophisticated quantum chemical calculations are adopted to calculate the exciton couplings from the detailed electronic structure of the adenine bases.

To surmount this problem, we performed a Markov State Modeling (MSM) analysis of the free energy landscape of the dApdA dinucleotide and identified five kinetically separable macrostates, each containing conformational species that can rapidly interconvert. We then calculated a single averaged conformation to represent each of the five MSM macrostates, and we found that the total CD spectrum can be

represented accurately by the weighted sum of the contributions from the averaged structures of these macrostates.

We foind that only two states exhibit both stacked and chiral conformations, which are necessary to provide significant exciton coupling between monomer EDTMs and rotational strengths, thus contributing to the CD observable. The two states are conformational ensembles with opposite chirality, which contain the anti (Watson-Crick B) form (S3) and a *syn* (Hoogsteen) flipped-base conformation (S4), respectively. A third highly populated state is an achiral *syn* state, with a slightly unstacked conformation (S5) that does not contribute significantly to the CD signal. We observed that both the S3 and the S4 states provide right-handed CD features in the long wavelength region of the dApdA spectrum. These results are qualitatively consistent with the early hypothesis that two leading states dominate the CD spectrum, but now provide more detailed information about the nature of those states.[109] We conclude that both the S3 and the S4 states contribute to the stacked conformation detected by Lowe and Schellman, while their unstacked conformation likely comprises the S5 state, which is the most populated, and to a lesser extent the fully unstacked state S2. Furthermore, our study shows that the Hoogsteen structure plays a key role in the mechanism of the stacking and unstacking pathways of the bases in dApdA, and possibly in DNA as well, as it is present in the highly populated stacked S4 and unstacked S5 conformations at all salt concentrations.

By connecting the CD spectrum of the dApdA dinucleotide to five leading conformational states as a function of salt concentration, we were able to obtain information about how the distribution of ion shell structures affects local base-stacking interactions. In agreement with early experiments by Johnson and Schleich,[116] we observed that the effect of increasing salt was to decrease the

magnitude of the CD signal over the 240 – 300 nm regime. In the CD experiments by Lowe and Schellman,[109] the decrease in CD signal was accompanied by a shift in the equilibrium population of open (unstacked) achiral conformations relative to closed chiral conformations. Our findings show an initial increase of the base stacking stability with increasing monovalent salt concentration (NaCl or KCl), followed by a decrease of stability at salt concentrations higher than 1.0 M. By analyzing radial distribution functions and orientation distribution functions of the ions and water solvent, respectively, we observed that the changes in local base stacking conformation at high salt concentrations are correlated strongly with the disruption of the ion shell boundaries, and weakly with a change in the orientations of the water dipole moments. Over the full range of salt concentrations, the orientations of water molecules within successive hydration shells are highly correlated, from layer to layer, through hydrogen bonding. Thus, the relatively large negative change in solvent entropy is attributed to the emergence of order of the ion shell structure upon base stacking, rather than the restructuring of the water dipole moments. These findings provide a more detailed picture in the context of the solvophobic bonding model, in which the enthalpic base stacking interaction is closely balanced by the decrease in entropy of the solvent environment. In contrast, this behavior is not observed for the singly-charged phosphate anion in isolation, suggesting that the presence of the bases in dApdA structure may be responsible for the disruption of the ion shell structure upon base un- stacking.

We also find that the trend in base stacking stability with increasing monovalent ion concentration for the dApdA dinucleotide is consistent with trends observed for the more complex duplex DNA.[114, 125] Although other factors, such as H-bonding and the cooperative stacking of multiple bases are known to play an important

role in determining the stability of dsDNA structures, our results suggest that the restructuring of the ion shells about the central phosphate ion with increasing salt concentration, observed in dApdA, may also play a role in regulating the stability of larger DNA macromolecules.

Bridge

In Chapter II, Markov state models (MSMs) were introduced and used to define the timecsales and dynamics of the slowest collective near-equilibrium motions of the protein ubiquitin. In this chapter, MSMs have been used to coarse-grain the (reduced) conformational space sampled by a single-stranded deoxyadenine dinucleotide (dApdA). We find that the coarse graining of the state space afforded by the PCCA+ approach is effective in separating the major conformational families contributing to the calculated CD spectrum of dApdA. That is, the CD spectrum of dApdA is mainly composed of signal from the conformations populating a right-handed minimum encapsulating the Watson-Crick B-form structure and conformations populating a left-handed minimum encapsulating Hoogsteen-type structures; the other three metastable states contribute little to the overall CD spectrum of dApdA. Furthermore, it was found that coarse-graining to the level of taking a single structure derived from each metastable state, the average intra-state structure, the overall CD spectrum could be well reproduced using just these five conformations of dApdA. These results support the results of [109] that two states contribute to the CD spectrum, but we find that the two states are not a right-handed and unstacked state, as postulated in [109], but rather a left- and right-handed pair of states.

This chapter also examined the salt dependence of the CD for dApdA. We find that adding salt continuously decreases the amplitude of the CD spectrum of dApdA because dApdA samples more frequently the achiral state near a twist angle of 0 degrees. This change in the CD spectrum coincides with a change in the salt structure surrounding the dApdA, as measuring by calculating the radial distribution function of the ions surrounding the phosphate group in the dApdA backbone; however, it should be noted that artificial crystallization [157] of the Na^+ and Cl^- ions used in the simulation was observed above salt concentrations of 1.0 M in the dApdA, although the artificial crystallization is not observed in the simulations of dihydrogen phosphate ion, likely due to the smaller box volume occupied by dihydrogen phosphate compared to dApdA. This shortcoming of the parameterization of the ions used in the simulation could contribute to the change in ion structure above 1.0 M concentration of NaCl.

In the next chapter, we return to extending the LE4PD method to describe the dynamics of proteins. The next chapter develops an anisotropic version of the LE4PD, and this new toolkit is used to re-analyze the 1-microsecond simulation of ubiquitin from Chapter II. This new LE4PD method maps directly onto a principal component analysis of the alpha-carbon trajectory of the protein when hydrodynamic effects, residue-specific friction coefficients, and free-energy barriers along each mode are neglected. As with the isotropic LE4PD, this new, anisotropic version of the LE4PD is also interfaced with Markov state models to desrcibes the kinetics and dynamics of the protein's slow collective motions.

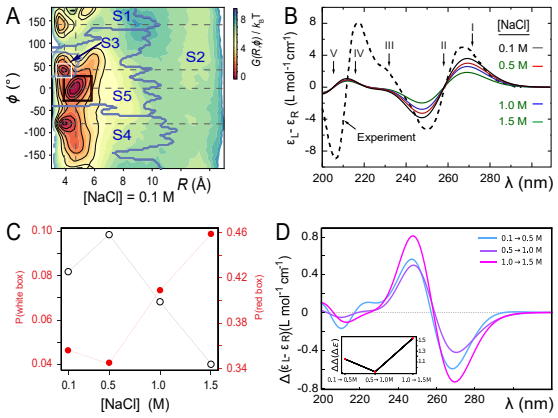

FIGURE 12. (A) Free energy landscape $G(R, \phi)$ as a function of the inter-base separation R and the dihedral twist angle ϕ, as obtained from 2-μs MD simulations of the dApdA dinucleotide at [NaCl] = 0.1 M. The coordinates corresponding to the canonical (average) B-form conformation (R = 3.6 $\mathring{A}$ and $\phi = 36°$) are included in the white (outlined) square, while the unstacked 'achiral' conformations are included in the red (outlined) square. The five macrostate regions, labeled S1-S5, were identified through the Markov State Modeling procedure described previously. (B) CD spectra of dApdA were determined from 2-μs MD simulations at salt concentrations [NaCl]=0.1 (black), 0.5 (red), 1.0 (blue) and 1.5 M (green). Differences between the calculated spectra are greater than the error bars (shown as the width of the colored lines), which were determined from the standard error of the mean from five block averages. The experimental CD spectrum (dashed black curve) was taken from ([1]). Roman numerals indicate the wavelengths of the electronic transitions of the uncoupled adenine monomers, which are used as input for our calculations. The experimental parameters used in the CD calculations are given in Appendix C (C) The local probabilities of the B-like stacked conformation and the unstacked 'achiral' conformation was calculated as the sum of states contained within the boundaries defined by the white and red squares, respectively, shown in panel (A). These probabilities are shown as a function of salt concentration [NaCl] = 0.1, 0.5, 1.0 and 1.5 M. The relative population of stacked and unstacked conformations changes abruptly around 1 M concentration. (D) Differences between the CD spectra at increasing salt concentration. The difference between calculated CD spectra are shown for incremental increases of the salt concentration. The peak-to-peak amplitude of the difference CD spectra decreases dramatically when the concentration is raised above [NaCl] = 1 M. This is reflected by the abrupt change in the peak-to-peak amplitude of the difference CD spectra shown in the inset.

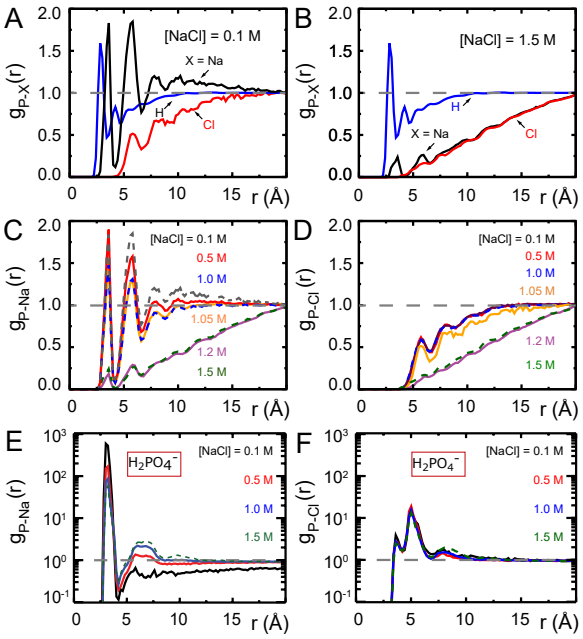

FIGURE 13. Radial distribution functions (RDFs) [Equation (3.1)] obtained from MD simulations of dApdA between Na^+, Cl^-, and the H atoms of water and the P atom of the anionic phosphate of the dApdA dinucleotide at salt concentrations ([NaCl]) of (A) 0.1 M and (B) 1.5 M. RDFs for (C) sodium ions and (D) chloride ions over the range of salt concentrations [NaCl] = 0.1, 0.5, 1.0, 1.05, 1.2 and 1.5 M. RDFs for sodium (E) and chloride (F) ions obtained from MD simulations of the phosphate anion $H_2PO_4^-$ - at the salt concentrations ([NaCl] = 0.1 (black), 0.5 (red), 1.0 (blue), 1.5 M (light blue)). Unlike the RDF plots for dApdA, the RDFs of $H_2PO_4^-$ - in aqueous solutions do not show the sharp change in the ion shell structure at [NaCl] $\approx$ 1.0 M (see also text).

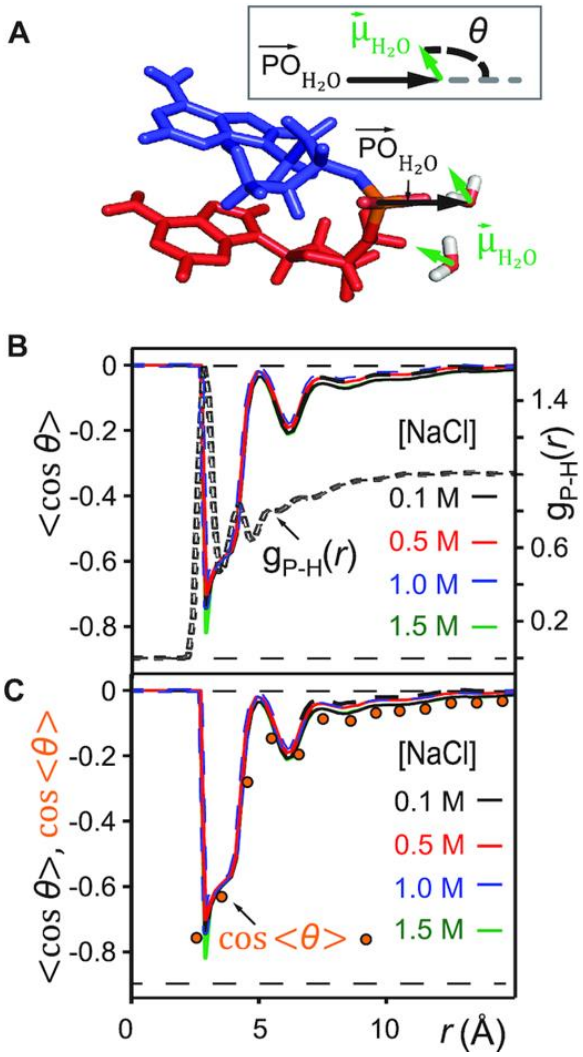

FIGURE 14. (A) Definition of the angle θ, which subtends the permanent dipole moment of the water molecule $\vec{\mu}_{H_2O}$ and the vector $\vec{PO}_{H_2O}$ connecting the phosphorous atom to the oxygen atom of the water molecule. (B) Orientation distribution functions (ODFs) for the dipole of the water molecule relative to the phosphate–oxygen (water) bond and RDF of the hydrogen of water, $g_{P--H}(r)$, of Figure 13A. (C) Superimposed on the ODFs defined as $\langle\cos(\theta)\rangle$, are the orange points indicating the cosine of the average angle, $\cos(\langle\theta\rangle)$.

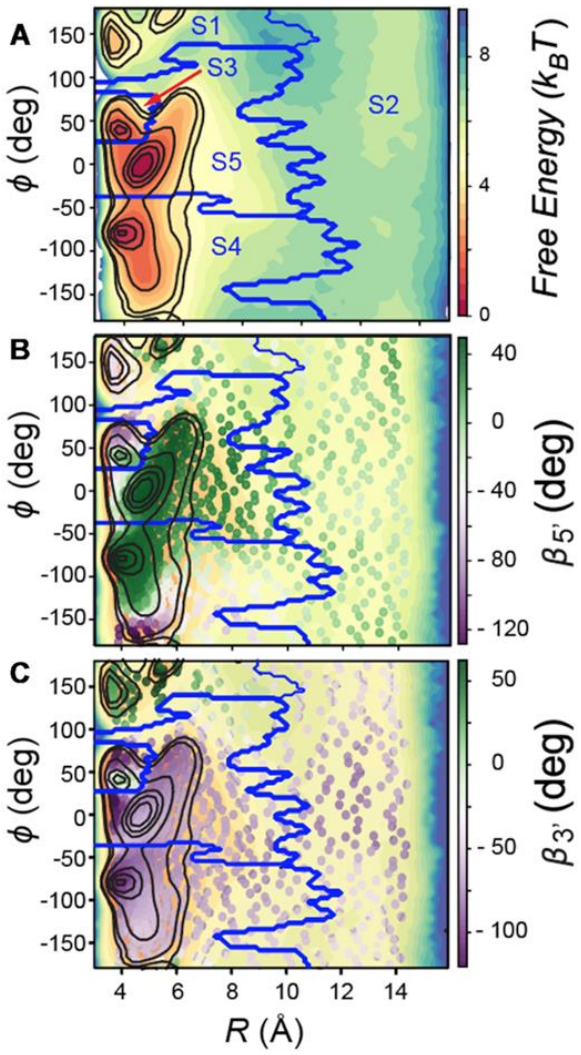

FIGURE 15. (A) The free energy landscape $G(R, \phi)$ of the dApdA dinucleotide (shown in Figure 11A) is sub-divided by dark blue boundaries into five regions (labeled S1 – S5), which are called 'macrostates.' The macrostate assignments were derived by performing a Markov state model (MSM) analysis of MD simulation data for [NaCl] = 0.1 M. The anti (Watson–Crick) form conformation is contained within the boundaries of the S3 macrostate, while the *syn* (Hoogsteen) containing form is included within the boundaries of the S4 macrostate. Superimposed on the free energy landscape $G(R, \phi)$ of the dApdA dinucleotide we show the orientation of the 5′ base (B) and of the 3′ base (C), respectively. The macrostate S3, which contains the anti form conformation, correctly displays both bases with positive orientation (green free energy minimum of S3 in both panels B and C), while macrostates S4 and S5, which contain a syn base presents the 3′ base flipped with respect to the 5′ base (green in panel B and purple in panel C).

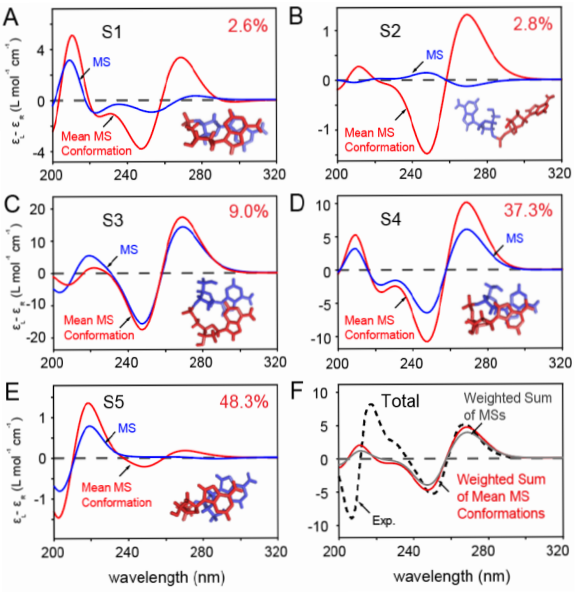

FIGURE 16. (A) – (E) Each panel shows, for each macrostate, the comparison between the contribution to the CD spectrum from all the conformational states in the macrostate (blue curve) and the contribution from the averaged macrostate structure (red curve), with structural parameters listed in Table S6 from the Supplementary Information of [66]. The molecular models representative of the averaged dApdA structures are shown as insets in each panel. The 5′ nucleotide is shown in blue, and the 3′ nucleotide and phosphate are shown in red. (F) The weighted sum of the macrostate contributions to the total CD are shown in gray, and the weighted sum from the averaged structures in red. The experimental CD spectrum13 is shown as a dashed black curve.

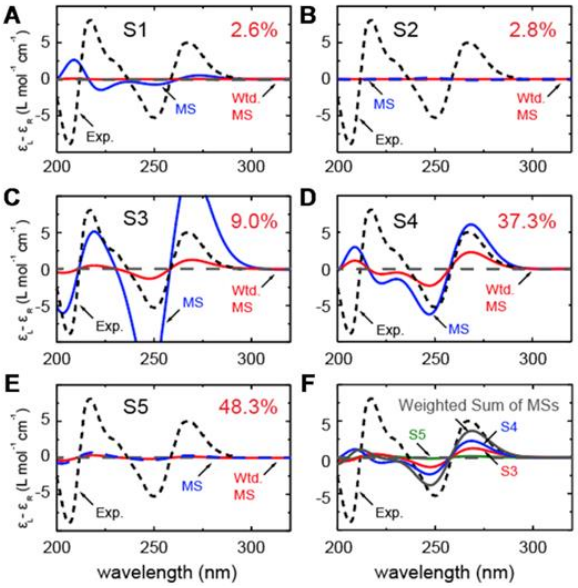

FIGURE 17. Macrostate decomposition of the CD spectrum of the dApdA dinucleotide by Markov state model (MSM) analysis of MD simulation data for [NaCl] = 0.1 M. The total CD spectrum is calculated from 10 million MD frames (or microstates), and the component spectra for macrostates (A) S1, (B) S2, (C) S3, (D) S4 and (E) S5 constitute 2.6, 2.8, 9.0, 37.3 and 48.3% of the total CD spectrum, respectively. The component CD spectra for each macrostate are shown in blue, and the number-fraction weighted contributions are shown in red. (F) The sum of number-fraction weighted macrostate contributions to the total CD is shown in gray. Also shown separately are the number-weighted contributions of macrostates S3, S4 and S5 (green, blue and red, respectively). In all panels, the experimental CD spectrum (from ([111])) is shown as dashed black curves.

CHAPTER IV

AN ANISOTROPIC LANGEVIN EQUATION FOR PROTEIN DYNAMICS: THE LE4PD-XYZ MODEL

From Beyerle, E. R. and Guenza, M.G. Comparison between slow, anisotropic LE4PD fluctuations and the Principal Component Analysis modes of Ubiquitin. *J. Chem. Phys.*, **154**, 12411 (2021).

Proteins are semi-flexible objects whose function is determined by the combined effect of their three-dimensional structure and local fluctuations [158]. 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 [159–162]. A common technique used to determine the slow, functional motions of proteins is the Principal Component Analysis (PCA) [160, 162–165]. PCA is a dimensionality reduction procedure commonly used in signal analysis to highlight important persistent features underlying noisy data. [166] 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. [167] 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 [26, 164, 168].

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) [4, 37, 38, 48, 57, 67]. 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.[161, 169] The direction and magnitude of the anisotropic fluctuations have been described using either the so-called 'porcupine plots' [169–172] or a simple linear interpolation between the extreme structures in the simulation trajectory[173]. 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.[87, 88] 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) [48]. 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.[4, 33, 34, 174] A similar analysis 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.

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 i^{th} 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 i^{th} 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

$$\frac{d\Delta\vec{R}_i^{\alpha}(t)}{dt} = -\frac{k_BT}{\overline{\zeta}}\sum_{\beta,\gamma\in\{x,y,z\}}\sum_{i=1}^{N}\sum_{j=1}^{N}H_{ij}^{\alpha\beta}A_{jk}^{\beta\gamma}\Delta\vec{R}_k^{\gamma}(t) + \vec{\Delta v}_i^{\,\alpha}(t), \tag{4.1}$$

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 of [51].

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. 4.1, the hydrodynamic interaction matrix $H_{ij}^{\alpha\beta}$ is given by

$$H_{ij}^{\alpha\beta} = \frac{\overline{\zeta}}{\zeta_i}\delta_{ij}\delta_{\alpha\beta} + (1-\delta_{ij})\delta_{\alpha\beta}\overline{r}_w \left\langle \frac{1}{r_{ij}} \right\rangle \ , \tag{4.2}$$

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 $\overline{r}_w = \frac{1}{N}\sum_i r_{w,i}$ 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

$$A_{jk}^{\beta\gamma} = \left(\left[a \otimes \widehat{I} \right]^T \mathbf{U} \left[a \otimes \widehat{I} \right] \right)_{jk}^{\beta\gamma}, \tag{4.3}$$

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.[175] A detailed derivation of $\mathbf{H}$ as well as a formal connection of the $\mathbf{U}$ matrix defined here to the $\mathbf{U}$ in the previously developed, isotropic version of the LE4PD[37, 38, 48, 57] are given in Appendix C.

It should be noted that the Langevin equation given in Eq.4.1 is identical in form to the optimized Rouse-Zimm equation for describing polymer dynamics derived by Zwanzig,[39] 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.[4, 37, 38, 48, 57, 176] 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.[45, 177] The Rouse-Zimm equation is equivalent to a Fokker-Planck-Smoluchowski for polymer dynamics.[40, 178] In this respect, the LE4PD-XYZ equation presented here is founded on well-established first-principles approaches.

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

$$C_{ij}^{\alpha\beta} = \langle \Delta \vec{R}_i^{\alpha} \Delta \vec{R}_j^{\beta} \rangle \tag{4.4}$$

where $\langle \cdots \rangle$ denotes, as above, an average over frames in the simulation trajectory. Using the definition of **A** given in Eq. 4.3 and that of **C** given in Eq. 4.4 it follows that

$$\mathbf{A} = \mathbf{C}^{-1}, \tag{4.5}$$

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

The set of coupled Langevin equations, in Eq. 4.1, is diagonalized by the eigenvector **Q**, as $\mathbf{Q}^{-1}\mathbf{HAQ} = \Lambda$, with Λ the diagonal matrix of eigenvalues of **HA**. By applying the eigenvector transformation, Eq.4.1 can be written in terms of its normal modes, $\Delta\vec{\xi}_a(t) = \sum_i (Q^{-1})_{ai} \Delta\vec{R}_i(t)$ as

$$\frac{d\Delta\vec{\xi}_a(t)}{dt} = -\sigma_a \Delta\vec{\xi}_a(t) + \Delta\vec{v}_a(t), \tag{4.6}$$

where $\sigma_a = \frac{k_B T \lambda_a}{\overline{\zeta}}$ is the characteristic diffusive rate of mode a,[41] 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, $\vec{v}_a(t) = \sum_i (Q^{-1})_{ai} \vec{v}_i(t)$ 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. 4.5, 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 **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 a^{th} 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,[4, 38, 48, 57] and is not part of the conventional PCA.

Methods

Molecular Dynamics Simulations

We generated the MD simulation of ubiquitin using GROMACS version 5.0.4[137] with the AMBER99SB-ILDN atomistic force field[179] on the Comet supercomputer at the San Diego Supercomputing Center. The starting structure was selected from the Protein Databank, PDB ID: 1UBQ.[53] 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.[54] 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.[38, 48, 57] 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[141] 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 $(10^6$ ps$)/(0.2$ ps/frame$) = 5 \times 10^6$ frames for analysis.

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.D.13, is calculated from an extended Stoke's law as [37]

$$\zeta_i = 6\pi(\eta_w r_i^w + \eta_p r_i^p) \ . \tag{4.7}$$

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,[43] $\eta_p = \eta_w \exp[< E_{int} > /k_B T]$ 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 $\mathbf{C}$ and $\mathbf{A}$, 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.[180] 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.[181] 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,[182] which implies that $\mu_a, \lambda_a \geq 0, \ \forall a$.

Validation of the LE4PD-XYZ theory

In Section 4.1 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. D.13 is formally consistent with a Fokker-Planck-Smoluchowski equation, where the dynamics are expressed as a function of the probability density function.[40, 99] In reference [178], 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 18 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 4.2 and 4.3. 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 4.3 for details).[183] The agreement is close to quantitative for the

time correlation functions (tcfs) shown in Figure 18. 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 [178] between simulations and the Fokker-Planck-Smoluchowski equation, where hydrodynamics and energy barriers are not included. However, in reference [178] 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 18 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.[40]

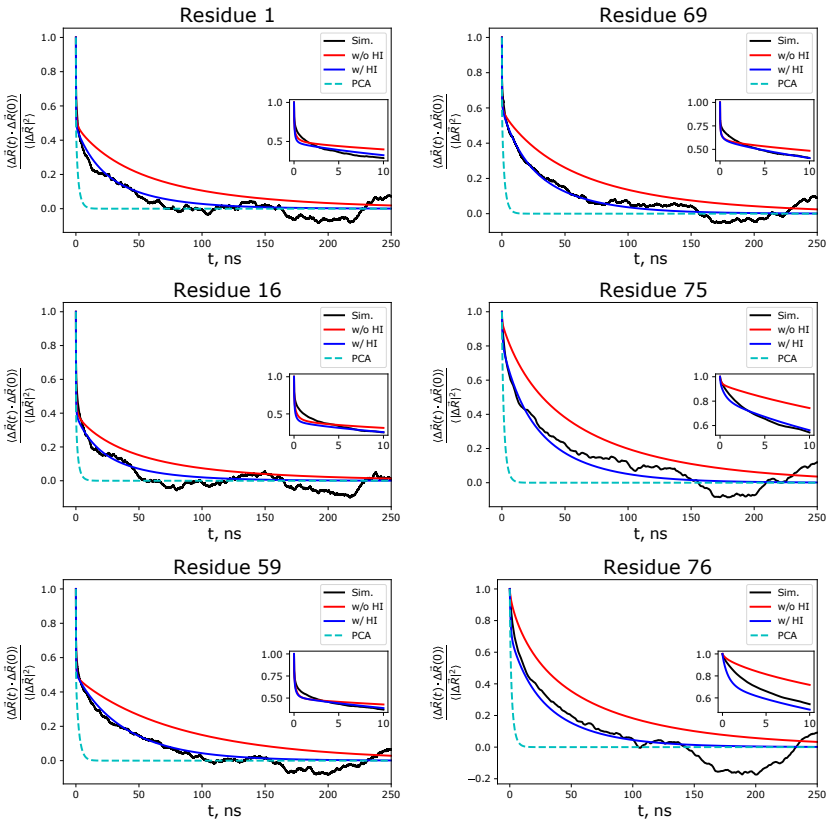

FIGURE 18. Time decay of the position fluctuation for different residues in the ubiquitin. The predictions of the LE4PD-XYZ theory with hydrodynamic interactions included (blue) are compared with the theoretical predictions without hydrodynamics (red) and with simulations (black). The predictions of 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.

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. D.13).

When the hydrodynamic interaction matrix reduces to an identity matrix because residue-specific friction coefficients 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 4.2. 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.[184, 185] 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 4.4.

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.[186] 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 4.4.

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.4.1, 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[4, 48] (see Section 4.3).

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

$$Q_a^{-1} = Q_{a,x}^{-1} \otimes \hat{x}^T + Q_{a,y}^{-1} \otimes \hat{y}^T + Q_{a,z}^{-1} \otimes \hat{z}^T, \tag{4.8}$$

with Q_a^{-1} the a^{th} 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., $\widehat{x} = (1, 0, 0)^T$. 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

$$F(\theta_a, \phi_a) = -k_B T \ln\left[P(\theta_a, \phi_a)\right] \ , \tag{4.9}$$

averaged over all the values of $|\vec{\xi_a}|$. A further discussion of this step is presented in the Supplementary Material of [51].

To account for the effects of energy barriers in the decay of the correlations of the residue fluctuations, such as those shown in Figure 18, the friction coefficient in eq. 4.6 is re-normalized using a Kramers-type approach [46, 67]: $\overline{\zeta} \rightarrow \overline{\zeta} \exp\left[F_{\mathrm{MAD},a}/K_B T\right]$, where $F_{\mathrm{MAD},a} = \mathrm{median}\left(\left|F\left(\theta_a, \phi_a\right) - \min\left(F\left(\theta_a, \phi_a\right)\right)\right|\right)$ is an average free-energy barrier calculated for LE4PD-XYZ mode a using the median absolute deviation (MAD)[47, 187] from the minimum of energy on the surface. This approach rescales the diffusive mode timescales as $\tau_a = \frac{\overline{\zeta}}{k_B T \lambda_a} \rightarrow \frac{\overline{\zeta}}{k_B T \lambda_a} \exp\left[F_{\mathrm{MAD},a}/K_B T\right]$. 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,[38] 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 19). 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.4.1 is solved while assuming $\mathbf{H} := \mathbf{I}$. In that case, the mode solutions and the free energy maps from both approaches are identical.

Comparing Pathways through the Free-Energy Surface

Figure 19 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,[4, 87] and are reported in the Figure 19a and Figure 19b, top panels. For the PCA, instead, the path is defined as the linear interpolation between the most extreme configurations sampled by the simulation[173] 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 19c 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 19c, 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.

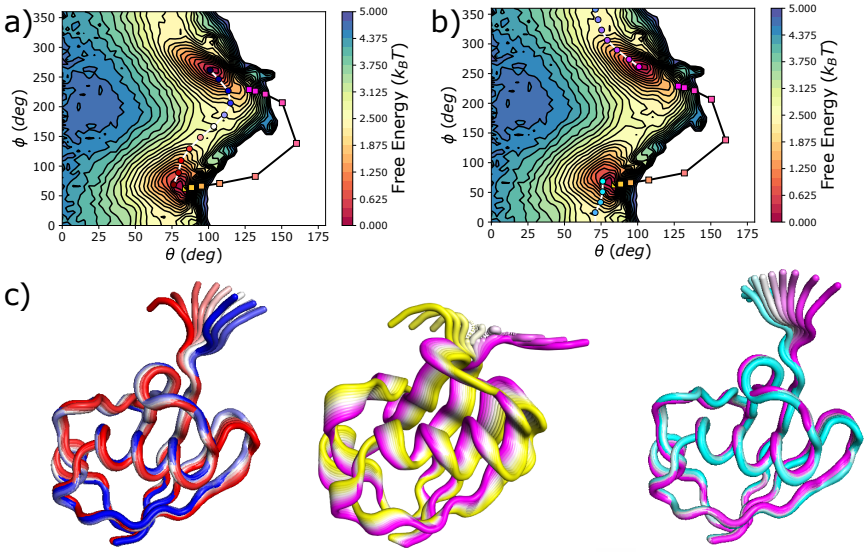

FIGURE 19. Free-energy surface for the first LE4PD-XYZ mode, solved for the case where $\mathbf{H} := \mathbf{I}$, so that the LE4PD-XYZ mode solutions are identical to the calculated PCA modes. The FES in the a) and b) panes are identical. a) The blue-white-red path is a minimum energy pathway between 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.

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 20 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 20 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.

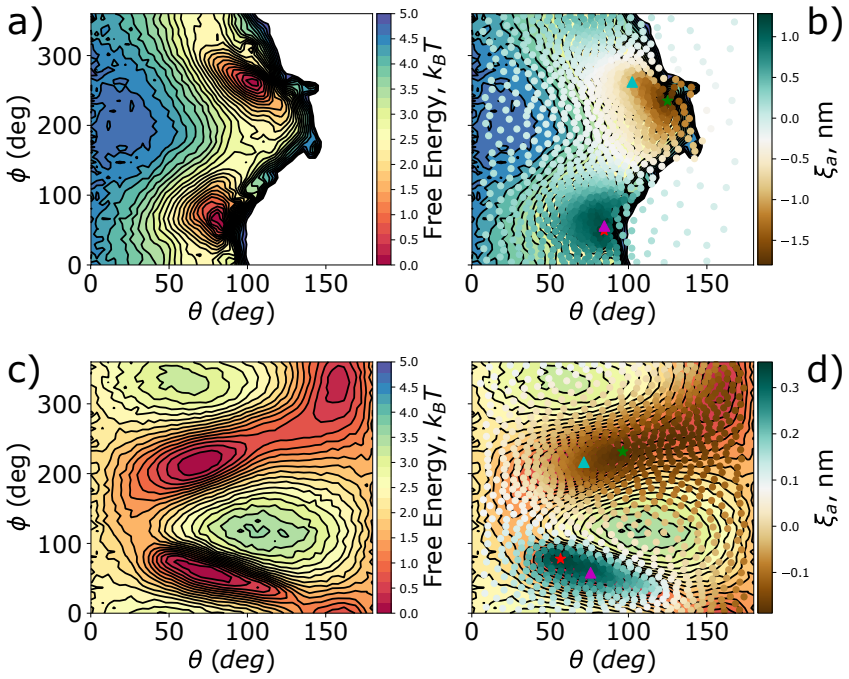

FIGURE 20. a) FES for the first internal LE4PD-XYZ mode. b) Projection of ξ_a (green-white-brown markers) onto the two-dimensional FES for a=1. c) FES for the seventh internal LE4PD-XYZ mode. d) Projection of ξ_a (green-white-brown markers) onto the two-dimensional FES for a=7. Green and red stars mark 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.

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 21 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 .[40, 57]

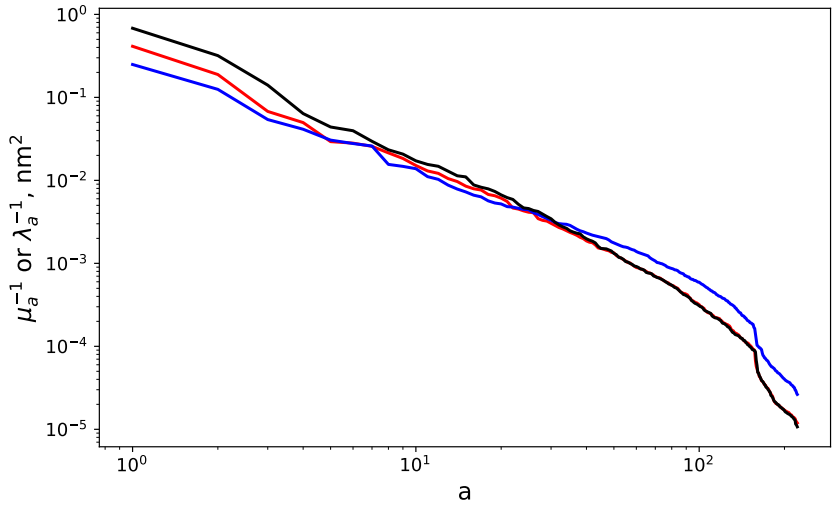

FIGURE 21. Comparison of the eigenvalue spectrum without hydrodynamic interaction (black curve), with the ones where the reside-specific friction coefficients are included, to account for the chemical specificity of each residue (red curve), and with the eigenvalues resulting from the diagonalization of the product of matrices containing the full hydrodynamic interaction matrix (blue curve).

Modifying the description of the hydrodynamic interaction is likely to affect the values of the eigenvectors of the matrix product **HA** as well. Figure 22 compares the eigenvector projections into the x-, y-, and z-coordinates, with and without 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

$$O_{ab} = Q_a \cdot Q'_b \ . \tag{4.10}$$

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 $\times$ 222 matrix. A more in-depth explanation of the significance of O is given in the Supplementary Material of [51].

Figure 23 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

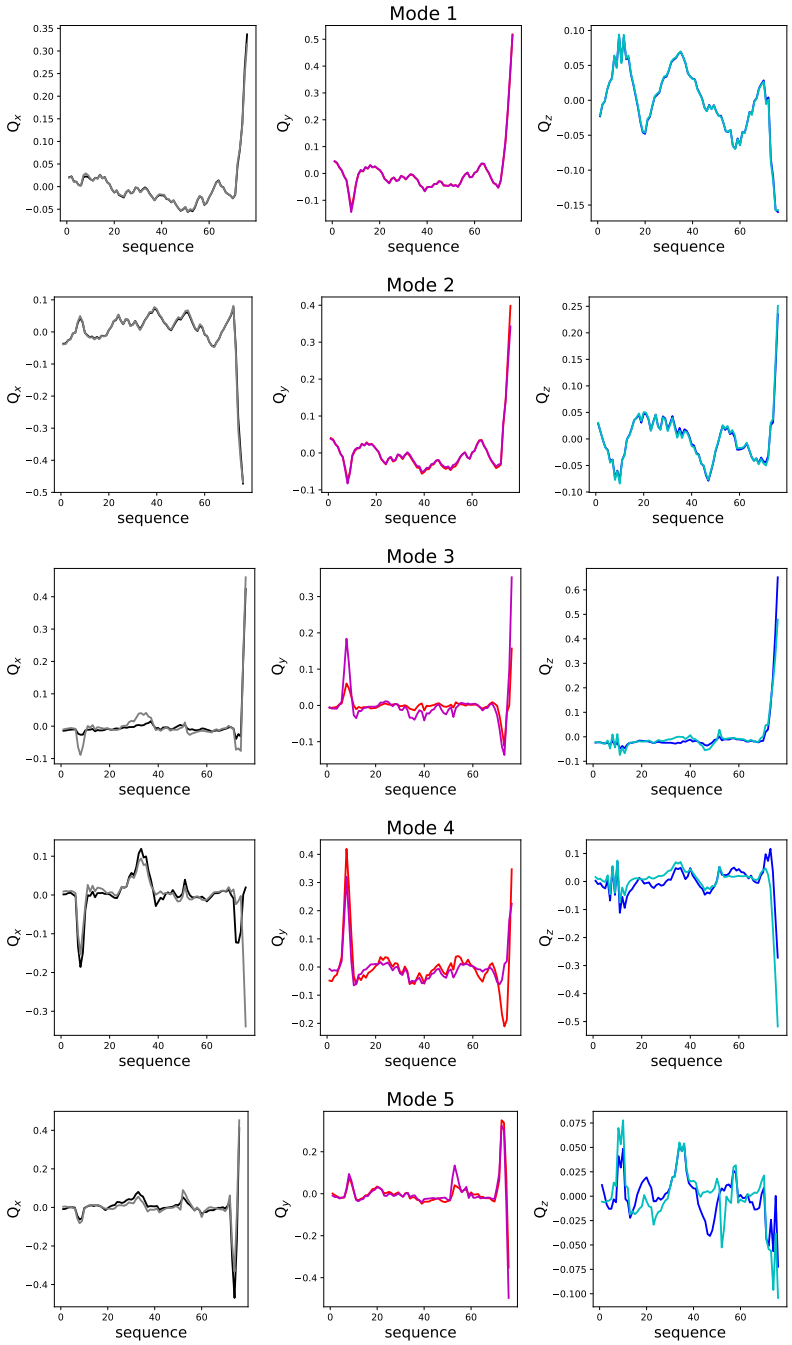

FIGURE 22. Comparison of the Q_a^x, Q_a^y, Q_a^z without (black, red, blue) and with (grey, magenta, cyan) hydrodynamic interactions for the first 5 LE4PD-XYZ modes.

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 23, 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 highest 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.

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 chapter 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

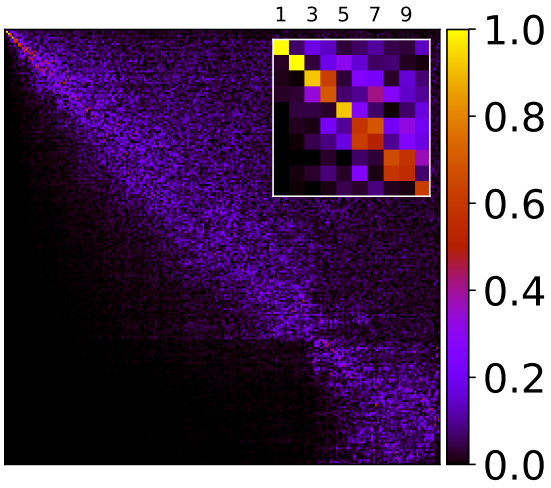

FIGURE 23. Overlap matrix O, defined in Eq. 4.10, between the right eigenvectors of the LE4PD-XYZ without hydrodynamics and the right eigenvectors of the LE4PD-XYZ with hydrodynamic interactions included. Overall, there is weak diagonal trend in this matrix, indicating similarity between the analogous 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.

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 22) while the corresponding eigenvalues are modified by the presence of hydrodynamic interactions (see Figure 21). 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 24 and 25. The ordering of the modes in Figures 24 and 25 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. 4.10. In agreement with Figure 22, 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 23. Note that mode 6 with HI (LE4PD-XYZ) has the maximum overlap with mode 7 without HI (PCA), and they will be directly compared.

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:

$$\sum_i \langle \Delta\vec{R}_i \cdot \Delta\vec{R}_i \rangle = \sum_a \mu_a^{-1} \sum_i Q_{ia}^2, \tag{4.11}$$

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. 4.11 we obtain the definition of the mean-square fluctuations at residue i due to the process described by mode a, which we will call the mean-squared local mode lengthscale, LML_{ia}^2:

$$\mathrm{LML}_{ia}^2 = Q_{ia}^2 \mu_a^{-1} \ . \tag{4.12}$$

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_{ia}^2 can be

decomposed into $x-$, $y-$, and $z-$projections:

$$\mathrm{LML}^2_{ia,x} = \left(Q^x_{ia}\right)^2 \mu_a^{-1} \tag{4.13}$$

$$\mathrm{LML}^2_{ia,y} = \left(Q^y_{ia}\right)^2 \mu_a^{-1} \tag{4.14}$$

$$\mathrm{LML}^2_{ia,z} = \left(Q^z_{ia}\right)^2 \mu_a^{-1}. \tag{4.15}$$

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 26 and 27; 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 22). 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 22), 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

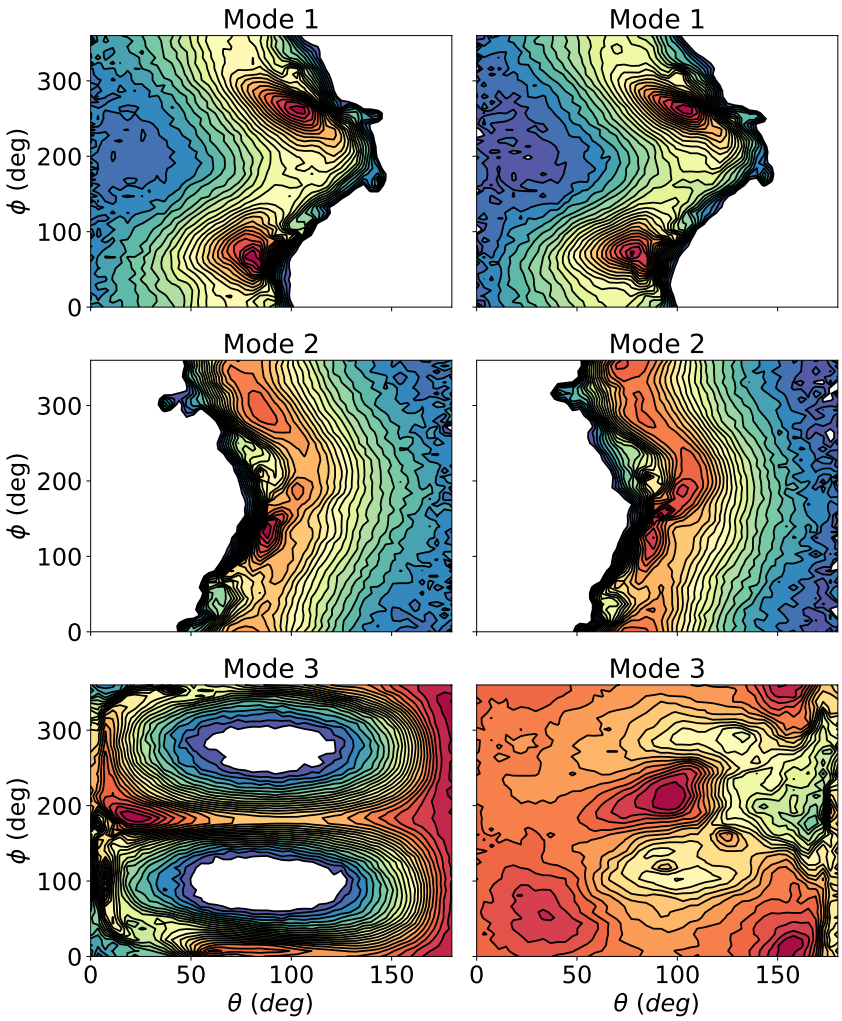

FIGURE 24. Comparing the three slowest modes without (left) and with (right) hydrodynamics in the LE4PD-XYZ analysis. Red corresponds to low energy and blue to high energy; all regions with a free-energy above 5 k_BT are 'masked' as white. The scaling of the free energy is the same as that in Figures 19 and 20.

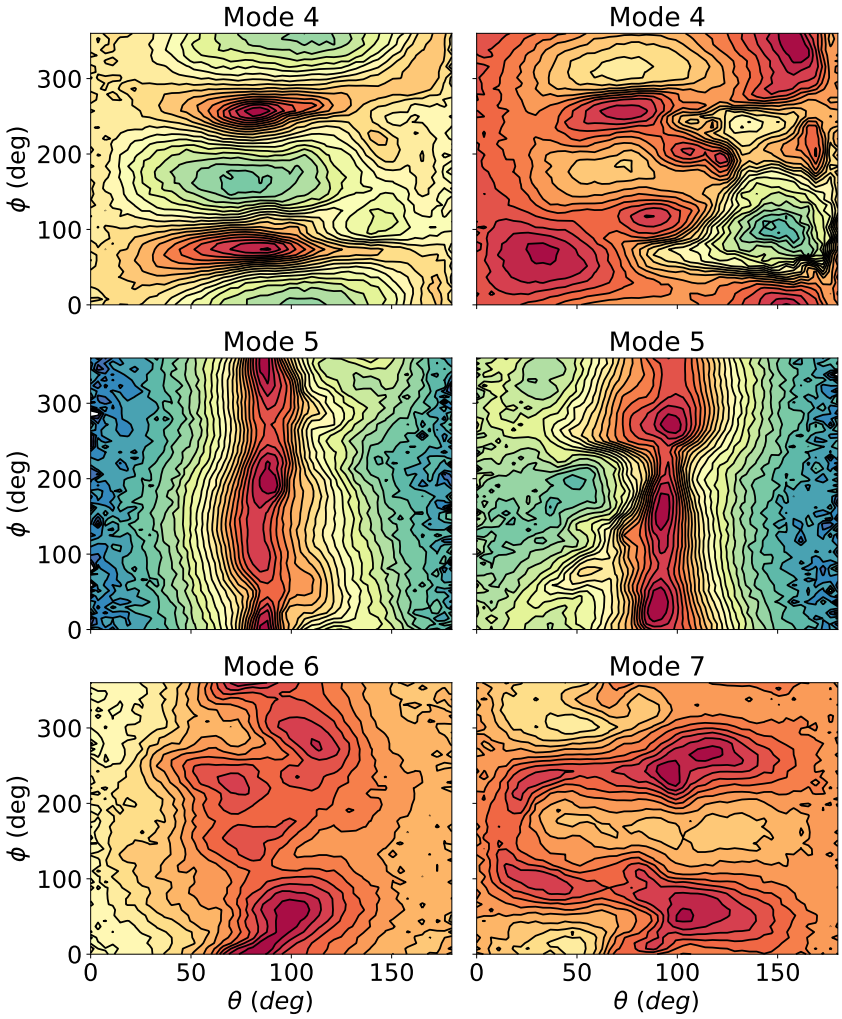

FIGURE 25. Comparing the three next slowest modes without (left) and with (right) hydrodynamics in the LE4PD-XYZ analysis. Red corresponds to low energy and blue to high energy; all regions with a free-energy above 5 k_BT are 'masked' as white. Mode 7 is swapped with mode 6 in the hydrodynamics treatment because it overlaps more strongly with mode 6 without hydrodynamics.

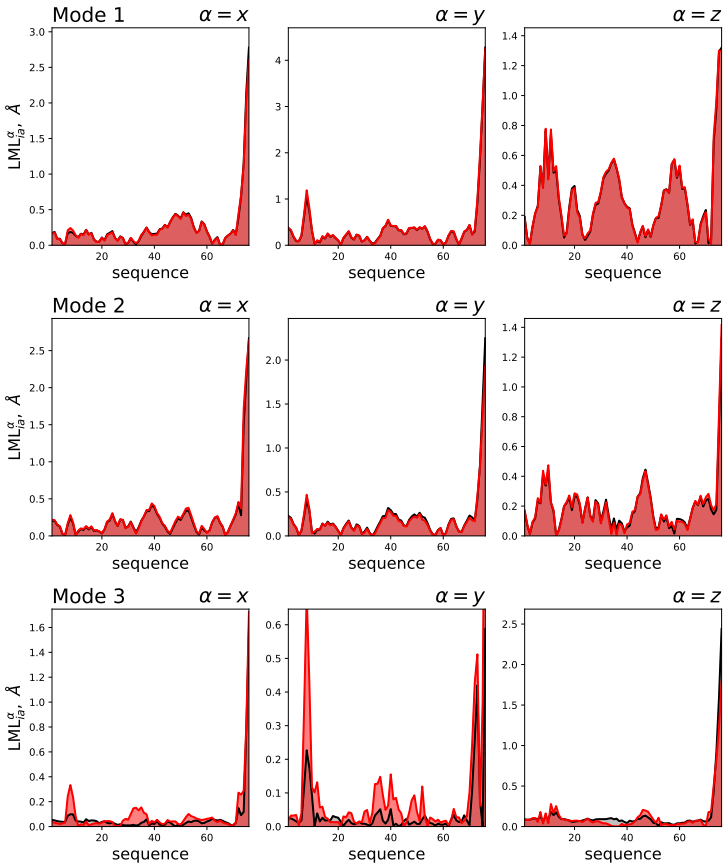

FIGURE 26. Anisotropic LML, $\mathrm{LML}_{ia}^{\alpha}$ for the first three LE4PD-XYZ modes, as ordered by the λ_a eigenvalues, for the case where hydrodynamic effects are neglected (black) and with hydrodynamic effects included (red).

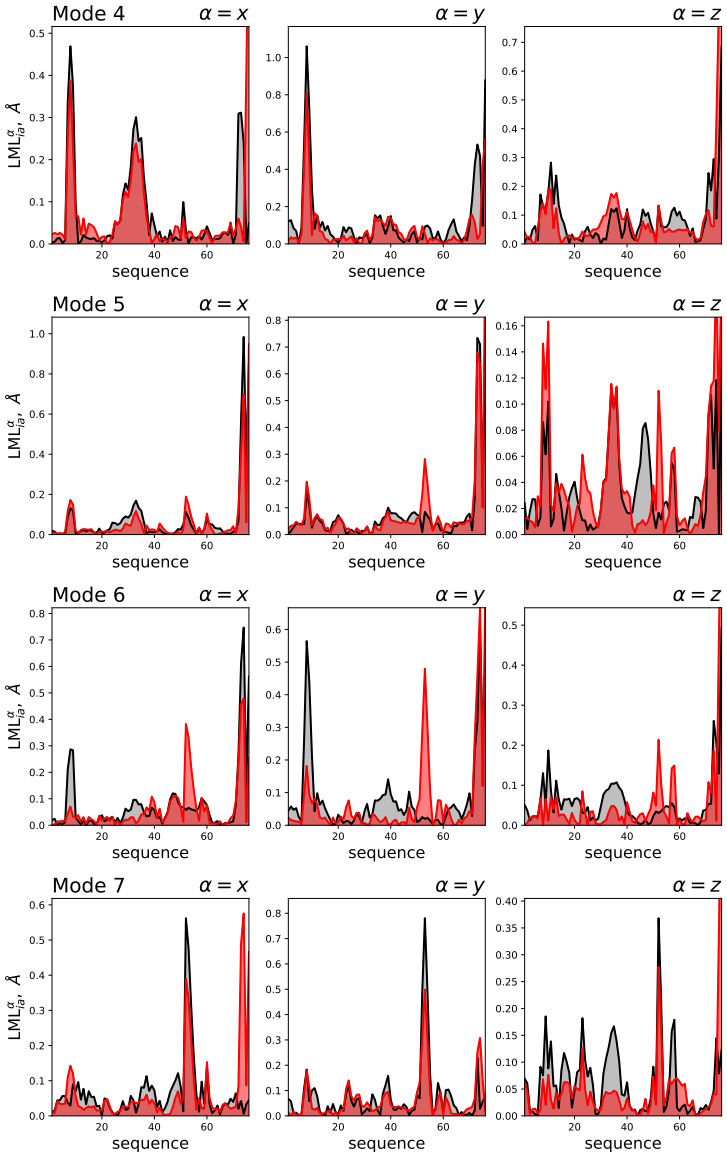

FIGURE 27. Anisotropic LML, $\mathrm{LML}_{ia}^{\alpha}$ for the next four slowest LE4PD-XYZ modes, as ordered by the λ_a eigenvalues, for the case where hydrodynamic effects are neglected (black) and with hydrodynamic effects included (red).

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 28 demonstrates how hydrodynamic effects can alter the predicted conformational changes along mode-dependent transition pathways. For the slowest LE4PD-XYZ mode, shown in Figure 28a, 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 28b 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.

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 barrier 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.[49] 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[5, 33, 49]. 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. [95] 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; $\phi_1 = \pi$.[95, 99]

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[5, 33, 49]:

$$t_i = -\frac{\tau}{\ln\left(|\lambda_i^{\mathrm{MSM}(\tau))}|\right)}. \tag{4.16}$$

This definition of the timescale of the transition relies on the Chapman-Kolmogorov theorem of Markovian statistics.[49] 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.[4] 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.[183]

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.[4] 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,[34, 174] which is the value for a transition state on the surface.[34, 96] 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[188–190] performed on the presented MSMs. $\mathbf{T}(\tau)$ is constructed using the reversible, maximum likelihood estimator given in [94]. An in-depth example of how the MSM is constructed for LE4PD-XYZ mode 7 is given in the Supplementary Material of [51].

Table 1 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 1 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 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

TABLE 1. Lagtimes, τ, and predicted timescales of the slowest process from the MSM, t_2, (both in ns) of the ten slowest LE4PD-XYZ modes 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)

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,[4] 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 27).

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,[4] 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. [4]

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,[167, 169] defined as:

$$\tau_a^{autocorr} = \int_0^\infty \frac{\langle \xi_a(t)\, \xi_a(0) \rangle}{\langle \xi_a(0)^2 \rangle} dt. \tag{4.17}$$

This approach gives the decorrelation time for any arbitrary stochastic process.[191] 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.[167, 169] 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.[192, 193] 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 2) or with (Table 3) hydrodynamic interactions included. For all the modes shown here, the τ_a calculated from the equation of motion 4.6 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.[4] 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.

TABLE 2. 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 3. 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

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 generally, for the protein's biological function.[92, 194] A popular method to select these large fluctuations is a Principal Component Analysis or PCA.[166] 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.[26, 164, 168]

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.[4, 37, 38, 48, 57, 67] 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.4.1) is diagonalized into a set of equations describing the independent, uncorrelated LE4PD-XYZ normal modes (Eq. 4.6).

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.[70, 71, 195] 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.[4]

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.[169]

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 1, 2, and 3).

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.[4] 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 [52, 75] and non-covalent .[78] 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,[10, 11] 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 22, 26, 27), eigenvalues (Figure 21), free-energy surfaces (Figure 24, 25, 28), and timescales (Table 1). 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 19, 20). 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.

Bridge

This chapter has developed an anisotropic extension of the original, isotropic LE4PD method [37, 38] by writing the Langevin equation in terms of the body-centered frame and using the deviations of each alpha-carbon in the protein as the relevant variables. This approach, termed the LE4PD-XYZ method, generates a set of collective coordinates identical to a principal component analysis (PCA) performed on the alpha-carbon degrees of freedom when 1) long-ranged hydrodynamic effects are neglected, 2) all the friction coefficients are assumed to be uniform and equal to the average friction coefficient of each amino acid, and 3) energetic barriers along the mode coordinates are ignored.

However, although the collective coordinates generated by the LE4PD-XYZ and PCA are identical in the case where hydrodynamics effects and residue-specific friction coefficients are neglected, PCA ignores energetic barriers along these collective modes, which the LE4PD-XYZ takes into account by generating a mode-dependent free-energy surface in a manner analogous to that done for the isotropic LE4PD, [38] even when hydrodynamic effects and residue-specific friction coefficients are not input to the model. These free-energy surfaces can significantly affect the interpretation of the predicted dynamics for the slow LE4PD-XYZ modes compared to the linear interpolation procedure [173] frequently used to interpret the motions described by the principal components.

Furthermore, it was shown that inclusion of the hydrodynamic interactions is necessary to model effectively the residue-residue autocorrelation functions calculated directly from the simulation even though hydrodynamic effects do not significantly alter the structure of the eignevectors of the slowest LE4PD-XYZ modes. We also couple the LE4PD-XYZ and MSM approaches to model the kinetics and dynamics of the slow LE4PD-XYZ modes in a manner analogous to that done for the isotropic LE4PD method [4] and find that the LE4PD-XYZ is also able to extract the slow motions of ubiquitin in the C-terminal tail, 50 s loop, and Lys11 loop. However, the LE4PD-XYZ analysis splits the dynamics of the 50 s loop into two modes, so there is no single, slow mode describing the conformational fluctuations of that loop.

Next, both the isotropic LE4PD and the anisotropic LE4PD-XYZ methods are compared to the results generated by a time-lagged independent component analysis (tICA) of the same 1-microsecond simulation of ubiquitin. tICA is a type of independent component analysis that selects the most slowly decorrelating set of collective motions from the input coordinates or 'features.' [30, 31] For the ubiquitin trajectory, this approach compresses the slow dynamics from all three flexible regions of ubiquitin into a single mode, but does not necessarily select any slow motions not captured by either the LE4PD or LE4PD-XYZ analyses. Re-constructing the residue-residue correlation function this chapter using the tICA modes also gives a poorer comparison to the simulation compared to the reconstruction using the LE4PD-XYZ modes.

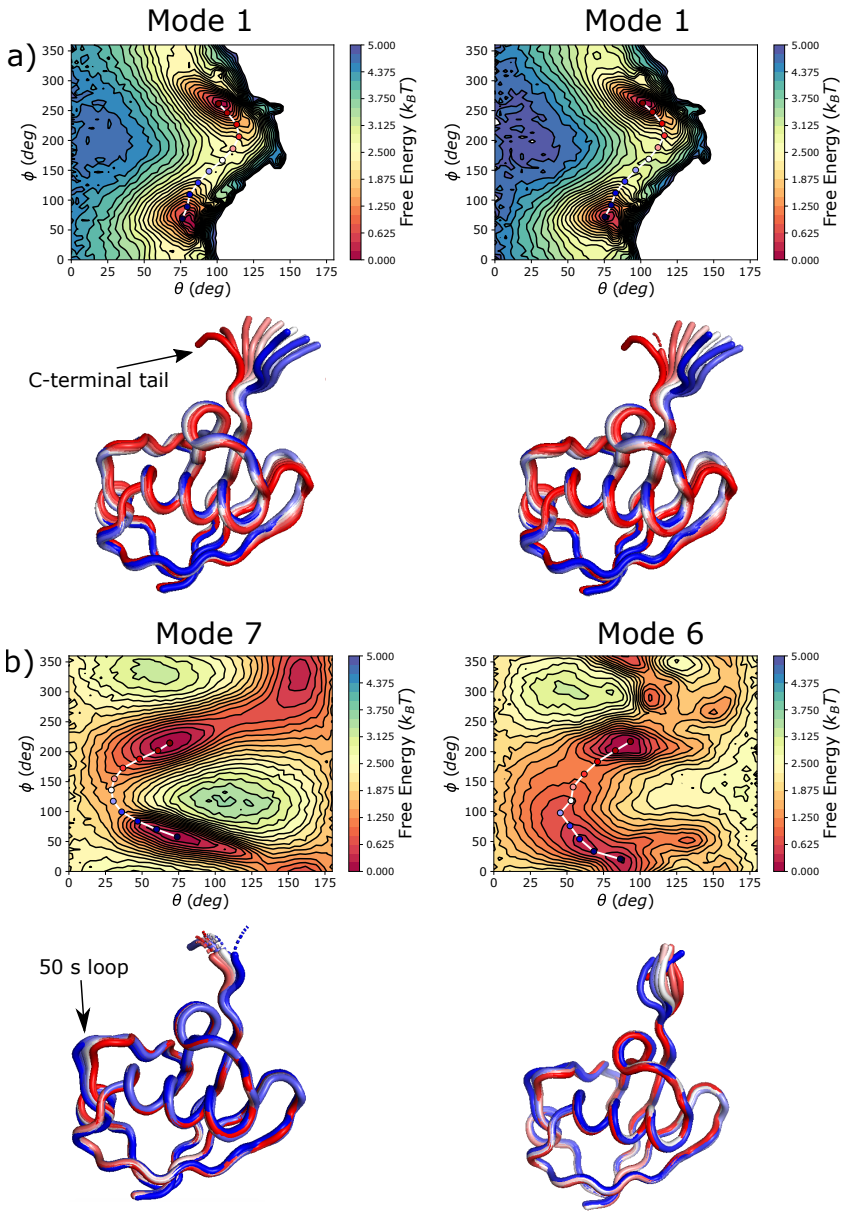

FIGURE 28. Comparing the mode-dependent fluctuations along a path in the free-energy surface for the slow LE4PD-XYZ modes when hydrodynamics are neglected (left), which is equivalent to PCA, or included (right) for a) LE4PD-XYZ mode 1 and b) LE4PD-XYZ mode 7 without HI (left) and mode 6 with HI (right). 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.

CHAPTER V

COMPARING THE SLOW DYNAMICS IN UBIQUITIN PREDICTED BY THE LE4PD, LE4PD-XYZ, AND TIME-LAGGED INDEPENDENT COMPONENT ANALYSIS METHODS

From Beyerle, E. R. and Guenza, M.G. Identifying the leading dynamics of ubiquitin: a comparison between the tICA and the LE4PD slow fluctuations in amino acids' position, submitted to *J. Chem. Phys.*

Large-scale fluctuations and global structural rearrangements play an essential role in the biological functions of biopolymers. Processes such as DNA transcription and replication involve the self-assembly of large multiprotein complexes that spontaneously form through step-by-step processes where binding of proteins is facilitated by the molecular flexibility and global scale rearrangements of the macromolecules comprising the overall structure.[6] At the single molecule level, folding of the proteins to their most probable conformation involves large-scale molecular fluctuations and slow global structural rearrangements of the protein conformation guided by cooperative dynamics.[196–199] These slow, large-scale, dynamical transitions drive the global biological processes that are important for the protein's biological function. [22, 23, 26, 27, 29, 70, 200, 201]

Molecular dynamics (MD) simulations of proteins in solvent are a powerful method to identify fluctuations and investigate the role that the chemical structure, or primary sequence, of a protein play in multiscale dynamics. However, the information contained in the simulation trajectory is difficult to analyze because dynamical processes are often coupled on multiple lengthscales. Therefore, it is crucial to devise statistical procedures that conveniently separate the multidimensional trajectory of a

simulation into a set of *independent* dynamical processes that, when added together, form the observed data. These different contributions should be as independent as possible to be able to analyze and classify their dynamical response separately. Traditionally, this issue has been addressed by adopting statistical tools from signal processing to extract from a noisy response the most critical information, which is usually a slowly fluctuating signal or a collection of slowly fluctuating signals.

A widely used analysis method for simulation trajectories is the principal component analysis or PCA method, based on the definition of a covariance matrix of the selected variables. [163] Correlation is a linear association measure, and uncorrelated processes are defined as having the cross terms in the covariance matrix equal to zero. However, independent processes and uncorrelated processes are different from the mathematical point of view. Independent processes are defined as having a joint probability distribution that can be separated into a product of individual distributions.[202] In practice, linearly uncorrelated processes identified using the covariance method are not always independent.

A recent study of ours on the analysis of proteins' MD trajectories has shown that slow processes identified by PCA often follow a pathway that is different from the most probable path of barrier crossing for the dynamical process.[51] This is not surprising, given that the kinetic paths of fluctuations in proteins are largely nonlinear, and it is unlikely for the linear processes of PCA to capture them. Furthermore, PCA is primarily based on the protein's structure and does not explicitly provide information on the time-dependent phenomena unless it relates to an equation of motion.[51, 203, 204]

To overcome the shortcomings of the PCA procedure in the study of the kinetics of large-scale protein fluctuations, it was proposed to use the time-lagged independent

component analysis or tICA. [30, 31, 205] In this time-dependent version of ICA, the dynamics of proteins are separated into mode signals that are uncorrelated at both zero lag time and a specific lag time of interest, τ_{tICA}, at which time the extracted modes also possess their maximal autocorrelation.[50] These constraints on the tICA substitute for the stringent independence criteria normally required from an ICA, with the independence at τ_{tICA} substituting for independence of nonlinear zero-lag correlations, [202] while allowing efficient temporal separation of the underlying mode dynamics.[30, 31, 202, 206] When paired with Markov state modeling of the kinetic of transition between modes, tICA accurately detects dominant slow modes of motion[31, 84], the tICA modes has been used as variationally-optimal collective coordinates for enhanced sampling in metadynamics.[207, 208] Interestingly, while tICA remains a rigorous statistical analysis of the multidimensional simulation trajectory, it still doesn't provide a physical interpretation of the slow dynamics or the connection between slow motions and protein's atomistic structure and interactions. That is, the degrees of freedom or 'features' input to the tICA are chosen based on their ability to predict the slowest dynamics, but are not necessarily connected to an equation of motion for describing the time evolution of the input coordinates.

A similar approach to tICA is the Relaxation Mode Analysis (RMA) by Takano and coworkers.[32, 203, 209, 210] Both RMA and tICA maximize the time-dependent correlation matrix of the fluctuations at a given lag time, τ, and at an initial time, t_0, while dynamics faster than t_0 is averaged out.[32] The difference between the two is that RMA calculates the covariance matrix at a time $t_0 \neq 0$, while tICA is a particular case of RMA, where $t_0 = 0$.[32] The RMA has also some similarities with our LE4PD approach, described below, as both accurately model with a Langevin

equation of motion the slow dynamics of the protein, even if the details of the two dynamical equations are different.

In recent years, the Guenza group has developed a coarse-grained protein dynamics representation called the Langevin Equation for Protein Dynamics or LE4PD.[37, 38, 48, 51, 57] The LE4PD is a powerful method that identifies the slow dynamical processes in a simulation trajectory of proteins in an aqueous solvent. The LE4PD separates the dynamics sampled in a long MD simulation, or in a set of short MD simulations, into a set of diffusive normal modes that are largely independent. The LE4PD approximately accounts *a posteriori* for the nonlinearities in the dynamics through the construction of free-energy landscapes for each mode and the rescaling of the timescale of barrier crossing via Kramers' theory.[4, 38, 51] These modes directly depend on real-space information, as the dynamical picture relates to each aminoacid's local friction, the water's viscosity, the potential of mean force, and the internal energy barriers. For each mode, the LE4PD provides a free energy landscape where one can identify the pathways of local fluctuations. Relaxation dynamics predicted by LE4PD have been shown to be accurate when compared with experimental data of T_1, T_2, and NOE NMR relaxation,[38, 57] as well as to short-time Debye-Waller factors from X-ray scattering experiments.[37]

While the LE4PD formalism is based on the physical representation of the dynamics of a polymer in solution, as defined in the famous Rouse-Zimm dynamical equation,[40, 46, 211] to describe the dynamics of a protein in solution the Rouse-Zimm approach has been modified in the LE4PD, which includes physical characteristics that are specific of folded proteins: typically i) inside the hydrophobic core of a protein, where atoms are not exposed to the solvent, the hydrodynamic interaction is screened, but atoms still experience friction, and ii) molecular

rearrangements of the protein during fluctuations involve the crossing of energy barriers that play a major role in protein dynamics and folding.

Recently, the LE4PD has been extended to treat anisotropic fluctuations of the alpha-carbon sequence, in the so-called LE4PD-XYZ approach. Beyerle et al.[51] have shown that LE4PD-XYZ directly maps on the PCA model and thus it provides PCA with a related equation of motion when hydrodynamic interactions and residue-dependent friction coefficients are neglected. For each slow mode, the LE4PD-XYZ identifies a mode-dependent free-energy barrier and the pathway of the non-linear fluctuations, indicating that those barriers are important in defining the correct timescale of the PCA slow modes.

In this study we move another step forward and compare the timescale of fluctuations for the slow modes measured by tICA and the ones described by the LE4PD-XYZ. With this goal in mind, we analyze several possible procedures to identify an "optimal" tICA lag time, and directly compare the predictions of tICA's slow-modes fluctuations (i.e. location along the primary sequence, amplitude, and timescale) with the ones predicted by the LE4PD. The question we aim to address is if a Langevin-mode decomposition can be effective in isolating the leading dynamical processes from a protein trajectory. While the tICA modes are designed for this purpose, contrary to the LE4PD modes, they do not have associated a formal equation of motion, which could be used to perform simulations of the protein in a reduced ensemble. We observe that both methods identify the same slow relevant motions when analyzing an extensive, 1-μs long MD simulation of the protein ubiquitin in a solution of sodium chloride at physiological conditions, albeit the two methods partition the dynamics into different quasi-independent modes from the decomposition of the trajectory. When directly compared, the time correlation

functions (tcfs) described by the LE4PD-XYZ are in almost quantitative agreement with the simulations, while the tcfs calculated using the optimal tICA description appear to be in less quantitative agreement.

The Langevin Equation for Protein Dynamics (LE4PD)

Isotropic LE4PD

We start by briefly reviewing the LE4PD approach in its isotropic and anisotropic versions. In recent years we have developed a coarse-grained model to describe protein fluctuations in the amino acid positions, called the Langevin Equation for Protein Dynamics (LE4PD). The approach has been extensively presented in several publications, so that we only briefly discuss it here.[4, 37, 38, 48, 57, 67] LE4PD projects the MD trajectory of a protein onto the slow coordinates of the alpha-carbon of each residue represented by the vector $\vec{R}(t)$. It models the time evolution of these coordinates using an overdamped Langevin equation, where the residues interact through the potential of mean force, defined by the matrix $U_{jk} = \langle \vec{l_i} \cdot \vec{l_j} \rangle / \langle |\vec{l_i}| \rangle \langle |\vec{l_j}| \rangle$. Here $\vec{l_i} = \vec{R}_{i+1} - \vec{R}_i$ is the bond vector between residue i and residue $i+1$ along the protein's primary sequence and the bracket defines the statistical average over all the trajectory's conformations. The dynamics is guided by the intramolecular potential of mean force (matrix $\mathbf{A}$, which defines the potential of mean force in the set of $\vec{R}$ coordinates) and hydrodynamic interactions, as well as the random forces generated by the collisions with the surrounding solvent. Thus, the propagation in time of the protein's dynamics follows a Langevin equation that in the α-carbon coordinates reads:

$$\frac{d\vec{R}_i(t)}{dt} = -\frac{3k_BT}{l^2\overline{\zeta}} \sum_{j=1}^{N} \sum_{k=1}^{N} H_{ij} A_{jk} \vec{R}_k(t) + \frac{\vec{F}_i(t)}{\overline{\zeta}}, \tag{5.1}$$

where k_B is the Boltzmann constant, T is the temperature of the protein-solvent system, l^2 is the mean-square bond length between alpha-carbons, $\overline{\zeta}$ is the average amino-acid friction coefficient, and H_{ij} describes the hydrodynamic interaction between residues i and j. $\vec{F}_i(t)$ is a random force modelling the effect of solvent collisions with the protein, and obeys the following fluctuation-dissipation theorem: $\langle \vec{F}_i(t) \cdot \vec{F}_j(t) \rangle = 6\overline{\zeta} k_B T \delta_{ij}$. The transformation from bead to bond coordinates effectively removes the global center-of-mass translation.

The LE4PD takes into account hydrodynamic effects and the chemical specificity of each residue in semiflexibility and friction coefficient. Diagonalizing the LE4PD leads to a Langevin equation of motion in a set of quasi-linearly independent, diffusive normal modes. Eq. (5.1) is solved using the eigenvalue decomposition of the **HA** matrix product, $\mathbf{Q}^{-1}\mathbf{HAQ} = \Lambda$,

$$\frac{d\vec{\xi}_a(t)}{dt} = -\frac{3k_B T}{l^2 \overline{\zeta}} \lambda_a \vec{\xi}_a(t) + \frac{\vec{F}_a(t)}{\overline{\zeta}}, \tag{5.2}$$

with $\vec{\xi}_a(t) = \sum_i \left(\mathbf{Q}^{-1}\right)_{ai} \vec{R}_i(t)$ the a^{th} LE4PD mode, and $\vec{F}_a(t)$ the random force vector transformed into the normal mode coordinates. The equation of motion, Eq. 5.1, can be written as a function of the bond vector coordinates, $\vec{l}$, thus uncoupling the center-of-mass translation from the internal dynamics of proteins. The two approaches yield equivalent information; however, for all the isotropic LE4PD results presented here, our analysis starts from the bond vector basis, $\vec{l}$. In the LE4PD formalism in bond coordinates, the first three modes represents the rotational dynamics of the protein, while modes with index higher than three describe the internal dynamics of the protein.[38] Since, in this study, we are interested only in describing the internal dynamics of a protein, we ignore the three isotropic LE4PD rotational modes, and,

when referring to isotropic LE4PD mode a, we implicitly mean isotropic LE4PD *internal* mode a.

Free-energy maps in isotropic coordinates and measuring fluctuation timescales

Each mode is associated with a free energy map describing mode-dependent local fluctuations of the aminoacids at specific locations along the protein's primary sequence (see Figure 29). The maps are constructed as follows: each isotropic LE4PD mode is a linear transformation of the amino acid position vectors,

$$\vec{R}_i(t) = (R_{i,x}(t), R_{i,y}(t), R_{i,z}(t))^T ,$$

through the eigenvector matrix $\mathbf{Q}^{-1}$, giving a mode vector with $x-, y-$, and $z-$components:

$$\vec{\xi}_a(t) = (\xi_{a,x}(t), \xi_{a,y}(t), \xi_{a,z}(t))^T .$$

For each LE4PD mode one can construct a free-energy surface in spherical coordinates, using the $x-, y-$, and $z-$components of $\vec{\xi}_a(t)$ as

$$\theta_a(t) = \arccos\left(\frac{\xi_{a,z}(t)}{|\vec{\xi}_a(t)|}\right) \tag{5.3}$$

$$\phi_a(t) = \arctan\left(\frac{\xi_{a,y}(t)}{|\xi_{a,x}|}\right) \tag{5.4}$$

$$F(\theta_a, \phi_a) = -k_B T \ln\left[P(\theta_a, \phi_a)\right] \ . \tag{5.5}$$

In Eq. (5.5), the dependence on the radial coordinate $|\vec{\xi}_a(t)|$ is averaged over to obtain the joint probability used in the definition of $F(\theta_a, \phi_a)$:

$$P(\theta_a, \phi_a) = \int P\left(|\vec{\xi}_a|, \theta_a, \phi_a\right) d|\vec{\xi}_a|.$$

To each mode is associated an energy map with a complex energy landscape where fluctuations have defined pathways, with characteristic amplitudes and timescales. From the linear combination of all the modes one can reconstruct the overall dynamics of the protein and its time correlation functions.[4, 38, 48, 57, 67, 212] Among these LE4PD modes one can identify and separate the slow, important motions of the protein based on their timescale. However the information of each mode is retained during the whole process.

Anisotropic LE4PD or LE4PD-XYZ

The isotropic LE4PD model has recently been extended to the related anisotropic formalism, called the LE4PD-XYZ method or anisotropic LE4PD. When hydrodynamic interactions are neglected, the friction coefficient is assumed to be identical for all aminoacids, and the internal energy barriers are neglected, the LE4PD-XYZ directly maps onto the PCA. Under these approximations, the force matrix in the Langevin equation is the inverse of the covariance matrix for the alpha carbon coordinates (special care needs to be taken when taking the inverse of the covariance matrix, as six eigenvalues are equal to zero after translation and rotation have been removed). Thus, we have shown that the LE4PD-XYZ and the PCA have identical eigenvectors and inverse eigenvalues when the above conditions are satisfied.

The first step in developing the anisotropic LE4PD is to define as the leading variables the deviations of the position of the protein's alpha-carbons from their average values, $\Delta\vec{R}_i(t) = \vec{R}_i(t) - \langle\vec{R}_i(t)\rangle$.[51] Each component of the position vector fluctuation follows the anisotropic LE4PD equation of motion

$$\frac{d\Delta R_i^{\alpha}(t)}{dt} = -\frac{k_B T}{\overline{\zeta}} \sum_{\beta,\gamma\in\{x,y,z\}} \sum_{j=1}^{N}\sum_{k=1}^{N} H_{ij}^{'\alpha\beta} A_{jk}^{'\beta\gamma} \Delta R_k^{\gamma}(t) + \Delta v_i^{\alpha}(t), \tag{5.6}$$

where $\alpha, \beta, \gamma \in \{x, y, z\}$. In this equation, the dynamics is defined in a body-fixed system of coordinates, where both translation and rotation dynamics have been eliminated. The trajectory of the protein, analyzed to build the $\mathbf{H}'$ and $\mathbf{A}'$ matrices for example, is also in a body-fixed reference system, where translation and rotation are absent. This is important, because the transformation of a trajectory directly from lab-system to body-fixed system of coordinates can lead to coupling terms that, in principle, cannot be ignored.[181, 213]

The matrix $H_{ij}^{'\alpha\beta}$ describes the hydrodynamic interaction between the α component of residue i and the β component of residue j, while the matrix $A_{jk}^{'\beta\gamma}$ describes the covariance between the β component of residue j and the γ component of residue k. More details on the anisotropic LE4PD model, and how it is formally related to the isotropic LE4PD, are given in [51]. As with the isotropic LE4PD, Eq. (5.6) is solved using the eigenvalue decomposition of the $\mathbf{H}'\mathbf{A}'$ matrix product, $\mathbf{Q}'^{-1}\mathbf{H}'\mathbf{A}'\mathbf{Q}' = \Lambda'$, which gives the equation of motion for the evolution of the LE4PD-XYZ modes:

$$\frac{d\Delta\vec{\xi}_a'(t)}{dt} = -\frac{k_B T}{\bar{\zeta}}\lambda_a'\Delta\vec{\xi}_a'(t) + \Delta\vec{v}_a'(t). \tag{5.7}$$

Free-energy maps in anisotropic coordinates and measuring fluctuation timescales

Using the decomposition of $\mathbf{Q}'$ for the anisotropic $\mathbf{H}'\mathbf{A}'$ matrix, the mode coordinate $\xi_a'(t)$ of the anisotropic LE4PD can be separated into its $x-, y-$, and

$z-$ components as

$$\begin{aligned}
\vec{\xi}'_a(t) &= \sum_{i=1}^{3N} Q'^{-1}_{ai} \Delta\vec{R}_i(t) \\
&= \sum_{i=1}^{3N} \left[\left(Q'^{-1}_{a,x} \otimes \hat{x}^T\right)_i + \left(Q'^{-1}_{a,y} \otimes \hat{y}^T\right)_i + \left(Q'^{-1}_{a,z} \otimes \hat{z}^T\right)_i \right] \Delta\vec{R}_i(t) \\
&= \sum_{i'=1}^{N} Q'^{-1}_{ai',x} \Delta x_{i'(t)} + Q'^{-1}_{ai',y} \Delta y_{i'}(t) + Q'^{-1}_{ai',z} \Delta z_{i'}(t) \\
&= \xi'_{a,x}(t) + \xi'_{a,y}(t) + \xi'_{a,z}(t) \ , \qquad (5.8)
\end{aligned}$$

and the spherical mode coordinates and free-energy surfaces can be defined analogously to the isotropic case as

$$\theta'_a(t) = \arccos\left(\xi'_{a,z}(t)/|\xi'_a(t)|\right)$$

$$\phi'_a(t) = \arctan\left(\xi'_{a,y}(t)/\xi'_{a,x}(t)\right),$$

$$F'(\theta'_a, \phi'_a) = -k_B T \ln\left[P'(\theta'_a, \phi'_a)\right] \ , \qquad (5.9)$$

where

$$P'(\theta'_a, \phi'_a) = \int P'\left(|\vec{\xi}'_a|\theta'_a, \phi'_a\right) d|\vec{\xi}'_a|.$$

As with the isotropic LE4PD modes, the linear combination of all the anisotropic modes leads to the structural and time-dependent properties, which can be directly compared with simulations or experimental data. These anisotropic free-energy surfaces are used to calculate fluctuations in the three spatial directions. The analysis of the mode-dependent free energy landscapes identifies the location of the protein fluctuations (i.e. loops, tails etc.), as well as it provides the pathways and the energy barriers related to those fluctuations. As an example, Figure 29 shows in panel a) the

FES in the mode coordinates for the first LE4PD-XYZ mode. The FES displays two minima separated by a small energy barrier. The protein's conformations along the pathway of transition between these two minima are displayed in panel b). Panels c) and d) report data from a Markov State Model analysis (see Section 5.7) of the mode trajectory, which shows the projection of the second MSM eigenvector, ψ_2. The second eigenvector of the MSM transition matrix identifies the top of the energy barrier and the transition state between the two minima (panel c)). Panel d) shows the calculation of the transition time that corresponds to the crossing between the two minima defined by the second MSM eigenvector. Note that Figure 29 displays results for the LE4PD-XYZ theory without hydrodynamic interactions (see for a discussion Section 5.4), and that identical calculations performed for the LE4PD-XYZ theory with hydrodynamic interactions are reported in the Supplemental Material document of [65]. The two calculations give free-energy maps and MSM analyses for the first mode that are quite similar.

When comparing the data from the LE4PD-XYZ analysis and the similar analysis of the tICA modes, one needs to account for the fact that in the simulation trajectory both translation and rotation have been eliminated. Thus, the first six modes in the diagonalization have zero eigenvalues. From the free energy surfaces we calculate the average energy barrier for each mode, and, using an extension of Kramers' kinetic theory, we calculate the slowing down of the dynamics due to the presence of barrier-crossing trajectories. Once this mode-dependent slowing down is accounted for, we build from the linear combination of the rescaled modes all the dynamical quantities, such as time correlation functions, reported in this manuscript. More details on this procedure are reported in our previous publications.[4, 51]

Time-lagged independent component analysis or tICA

The time-lagged independent component analysis is a method extensively used in the field of signal processing, information theory, artificial neural networks to identify hidden factors that are shared and underlie the observed multivariate data.[206] This technique has been applied in several fields, including the analysis of protein dynamics to identify the prevalent large-scale motion inside a simulation trajectory. With Independent Component Analysis (ICA), it is possible to identify collective slow dynamical components that are as statistically independent as possible. By introducing a time lag in the sampling of the data, one effectively includes the temporal dimension in the analysis of the leading fluctuations making it possible to model kinetic processes. The time-lagged ICA is an extension of the principal component analysis (PCA) method, where one takes care of isolating the most slowly decorrelating dynamics while including the time dependence of the data as an explicit variable in the analysis. The tICA method has been reviewed in several recent publications and will be only summarized here.[84, 85, 169, 214, 215]

While tICA is a general approach that applies to any set of coordinates, here, we are interested in performing a tICA of the alpha-carbon trajectory of a protein with N residues. We define as tICA coordinates the $\Delta\mathbf{R}(t)^T = \vec{R}_1(t) - \langle\vec{R}_1(t)\rangle, \vec{R}_2(t) - \langle\vec{R}_2(t)\rangle, \ldots, \vec{R}_n(t) - \langle\vec{R}_n(t)\rangle$, where $\Delta\vec{R}_i(t) = \vec{R}_i(t) - \langle\vec{R}_i(t)\rangle$ represents the fluctuations out of the equilibrium structure of the position of the space coordinates $\vec{R}_i(t)$, with $\vec{R}_i(t) = x_i(t), y_i(t), z_i(t)$ and $i = 1, \ldots, N$ with N the number of amino acids in the protein. The time dependent covariance matrix is defined, for a lag time τ, as $\mathbf{C}^r(\tau) = \langle\Delta\mathbf{R}(t+\tau)^T\Delta\mathbf{R}(t)\rangle_\tau$, and for $\tau = 0$ the covariance matrix recovers the static, structural matrix that is used in PCA, as $\mathbf{C}^r(0) = \langle\Delta\mathbf{R}(t)^T\Delta\mathbf{R}(t)\rangle$.

The tICA modes, or tICs, are found by solving the following *generalized eigenvalue equation*[30, 202]:

$$\mathbf{C}^r(\tau)\Omega = \mathbf{C}^r(0)\Omega\Lambda_{IC}(\tau), \tag{5.10}$$

where Ω is the matrix of right eigenvectors of $\mathbf{C}^r(\tau)$, and $\Lambda_{IC}(\tau)$ is the diagonal matrix of the related eigenvalues. In addition, $\langle a(t)\rangle = \frac{1}{M}\sum_{t=1}^{M} a(t)$ denotes the usual static average calculated over a trajectory of length M frames and $\langle a(t+\tau)b(t)\rangle_\tau = \frac{1}{M-\tau}\sum_{t=1}^{M-\tau} a(t+\tau)b(t)$ denotes an average over the time-lagged trajectory.

From the solution of the generalized eigenvalue problem, one has that the eigenvector matrix, Ω, diagonalizes both $\mathbf{C}^r(\tau)$ and $\mathbf{C}^r(0)$:

$$\begin{aligned}\Omega^T\mathbf{C}^r(\tau)\Omega &= \Lambda_{IC}(\tau)\\ \Omega^T\mathbf{C}^r(0)\Omega &= \Lambda'_{IC}(0) = \mathbf{I},\end{aligned} \tag{5.11}$$

where $\mathbf{I}$ is an identity matrix of the same dimensions as $\mathbf{C}^r(\tau)$ and $\mathbf{C}^r(0)$. The tICA modes, $\mathbf{z}(t)$, are determined by transforming the input coordinates $\Delta\mathbf{R}(t)$ by $\mathbf{z}(t) = \Omega^T\Delta\mathbf{R}(t)$.

The second line of Eq.5.11 gives a method to interpret the meaning of the transformation above as follows. We start by decomposing $\mathbf{I}$ into two orthogonal matrices $\mathbf{V}$ as $\mathbf{I} = \mathbf{V}^T\mathbf{V}$ and give the eigenvalue decomposition of the zero-lag time covariance matrix as $\mathbf{C}^r(0) = \mathbf{W}\Lambda_{IC}(0)\mathbf{W}^T = \mathbf{W}\Lambda_{IC}(0)^{\frac{1}{2}}\Lambda_{IC}(0)^{\frac{1}{2}}\mathbf{W}^T$. The

eigenvector matrix Ω can be expressed as a function of the eigenvectors $\mathbf{W}$ as follows:

$$\begin{aligned} \Omega^T \mathbf{W} \Lambda^{\frac{1}{2}}(0) \Lambda^{\frac{1}{2}}(0) \mathbf{W}^T \Omega &= \mathbf{V}^T \mathbf{V} \\ \Rightarrow \left[\Lambda^{\frac{1}{2}}(0) \mathbf{W}^T \Omega \right]^T \Lambda^{\frac{1}{2}}(0) \mathbf{W}^T \Omega &= \mathbf{V}^T \mathbf{V}. \end{aligned} \tag{5.12}$$

Equating the sides of Eq. 5.12 gives

$$\begin{aligned} \mathbf{V} &= \Lambda^{\frac{1}{2}}(0) \mathbf{W}^T \Omega \\ \Rightarrow \Omega &= \mathbf{W} \Lambda^{-\frac{1}{2}}(0) \mathbf{V} \\ \Rightarrow \Omega^T &= \mathbf{V}^T \Lambda^{-\frac{1}{2}}(0) \mathbf{W}^T, \end{aligned} \tag{5.13}$$

which agrees with the result given in [30]. Since the principal component modes (PCs), $\Xi(t)$, of $\Delta\mathbf{R}$ are defined as $\Xi(t) = \mathbf{W}^T \Delta\mathbf{R}$, the tICA modes,$\mathbf{z}(t) = \Omega^T \Delta\mathbf{R}$, can be written in terms of the PCs as

$$\mathbf{z}(t) = \mathbf{V}^T \Lambda^{-\frac{1}{2}}(0) \Xi(t) = \mathbf{V}^T \widehat{\Xi}(t), \tag{5.14}$$

where $\widehat{\Xi}(t) = \Lambda^{-\frac{1}{2}}(0)\Xi(t)$ are the whitened (unit variance and zero-mean) PCs. Thus, the tICA modes have a straightforward interpretation: since $\mathbf{V}^T$ is an orthogonal matrix, it defines a rotation in $3N$-dimensional space, and the tICs are just rotations of a linear combination of the whitened PCs, $\widehat{\Xi}(t)$. For example, the i^{th} tICA mode, $z_i(t)$, can be written in terms of the whitened PC modes as

$$z_i(t) = \sum_j V_{ij}^T \widehat{\Xi}_j(t).$$

Furthermore, since Ω is just a rotation of the (scaled) eigenvectors of the zero-lag covariance matrix, its elements can be decomposed into their $x-$, $y-$, and $z-$projections, as is the case for the eigenvectors of the zero-lag covariance matrix:[51]

$$\Omega = \Omega^x \otimes \widehat{x} + \Omega^y \otimes \widehat{y} + \Omega^z \otimes \widehat{z}, \tag{5.15}$$

where $\widehat{x}$, $\widehat{y}$, and $\widehat{z}$ are the unit vectors in the x-, y-, and z-directions, and $\otimes$ denotes the Kronecker product.[175] This decomposition is useful as it allows for the creation of tIC-dependent free-energy surfaces, which can be compared directly with the LE4PD free energy surfaces (see Sections 5.1 and 5.2).

Converting to spherical coordinates creates a free-energy surface for each tICA mode

To define a Free-Energy Surface (FES) for each of the tICA mode coordinates, we start by projecting the space coordinates of the fluctuations onto tICA modes using the tICA eigenvectors. For the tICA modes, the eigenvector matrix Ω^T, which transforms the $\Delta\vec{R}(t)$ into the $\mathbf{z}(t)$ tIC coordinate system, can be decomposed into its contributions from the $x-, y-$, and $z-$components of $\Delta\vec{R}(t)$,

$$\Omega^T = \Omega^{T,x} \otimes \widehat{x}^T + \Omega^{T,y} \otimes \widehat{y}^T + \Omega^{T,z} \otimes \widehat{z}^T, \tag{5.16}$$

which allows for the decomposition of each tIC $z_a(t)$ into its contributions from the $x-, y-$, and $z-$components of the input coordinates $\Delta\vec{R}(t)$:

$$z_{a,x}(t) = \sum_{i=1}^{N} (\Omega^x)_{ai}^T \, \Delta x_i(t) \tag{5.17}$$

$$z_{a,y}(t) = \sum_{i=1}^{N} (\Omega^y)_{ai}^T \, \Delta y_i(t) \tag{5.18}$$

$$z_{a,z}(t) = \sum_{i=1}^{N} (\Omega^z)_{ai}^T \, \Delta z_i(t). \tag{5.19}$$

This decomposition can be used to describe each tIC in a new spherical coordinate system:

$$R_a(t) = z_{a,r}(t) = \sqrt{z_{a,x}(t)^2 + z_{a,y}(t)^2 + z_{a,z}(t)^2} \tag{5.20}$$

$$\theta_a(t) = z_{a,\theta}(t) = \arccos\left(\frac{z_{a,z}(t)}{|\mathbf{z}_a(t)|}\right) \tag{5.21}$$

$$\phi_a(t) = z_{a,\phi}(t) = \arctan\left(\frac{z_{a,y}(t)}{z_{a,x}(t)}\right). \tag{5.22}$$

This decomposition of the tICs into the contributions from the $x-, y-$, and $z-$components of $\Delta\vec{R}(t)$ is completely analogous to that given for the principal components and LE4PD-XYZ modes in [51]. In this way, an analysis of the slow tICs in the coordinate system defined by Eqs. (5.20), (5.21), (5.22) can be seen as an extension of the analysis preformed for the principal components in [51] to the domain where the covariance matrix contains a time-lag.

With the definitions of $\theta_a(t), \phi_a(t)$ and $R_a(t)$, two-dimensional free-energy surfaces in (θ_a, ϕ_a) can be created analogously to that done for the anisotropic LE4PD

modes by averaging over the radial coordinate $R_a(t)$:

$$F(\theta_a, \phi_a) = -k_B T \ln \left[P(\theta_a, \phi_a)\right] = -k_B T \ln \left[\int P\left(R_a, \theta_a, \phi_a\right) dR_a \right], \tag{5.23}$$

which is used to determine tIC-specific dynamics, free-energy barriers, and timescales. The main advantages of constructing the free-energy surfaces in this manner are 1) each surface is tIC-specific, so the dynamics among tICs are decoupled, and 2) energetic pathways and fluctuations along this surfaces are facile to visualize for each tIC. As with previous LE4PD analyses, a variant of the string method is utilized to find minimum free-energy pathways between energy wells on the surface.[4, 51, 87]

Selection of the tICA lag time using the free-energy surfaces

The tICA approach is general and applies to any time-dependent set of coordinates. After selecting the input coordinates to the tICA, which in this study are the coordinates of the fluctuations away from the average structure calculated over the MD trajectory, $\mathbf{\Delta R}$, there remains a single adjustable parameter: the observation lag time, τ_{tICA}. This time parameter is used to construct the time-lagged covariance matrix (see Eq. 5.10). Identifying an appropriate lag time for the system of interest is vital to obtaining relevant results from the tICA.[30, 31, 169] In general, one selects the τ_{tICA} that captures the relevant dynamical fluctuations: the tICs identify the dynamics taking place over a timescale longer than τ_{tICA}, while dynamical phenomena that are faster than the selected lag time are averaged out and cannot be detected. Thus, only selecting the proper lag time can lead to the correct sampling of the dynamical phenomena that one desires to study. Here, we select the optimum τ_{tICA} from the mode-dependent free energy surface. Figure 30 shows that by increasing the

lag time, the system samples a rising barrier with slow fluctuations that move from the C-terminal tail and Lys11 loop into the 50 s loop. The barrier height (see Figure 31) increases until $\tau_{tICA} = 2.0$ ns, when it starts decreasing, thus identifying as an optimal tICA lag time the two nanosecond time interval for sampling. Figure 31 also reports the calculated Markov State Model (MSM)time, t_2, which is related to the second MSM eigenvector (more details on the MSM method are in Section 5.7). t_2 is the time needed by the system to cross the barrier and shows a nice correlation with the barrier height for long tICA lag times.

Note that this procedure to find the optimal time lag doesn't show a sharp transition in the shape of the FES at the selected $\tau_{tICA} = 2.0$ ns. At the same time, a more evident change of behavior is visible in the correlation of the time with the barrier's height. One can find similar qualitative results for the FES by selecting lag times of one order of magnitude below or above the two nanosecond threshold. Still, both these choices (Figures 30 and 31) result in surfaces where the barriers are lower and where the tICA is less able to model the overall dynamics observed in the simulation (Section 5.5).

The non-homogeneity of ubiquitin's dynamics when changing the tICA lag time is likely associated with the well-known hierarchical energy landscape of proteins in the folded state.[23, 197] At short lag times the tICA is selecting faster dynamics, which crosses small barriers within a single well on the folded energy landscape. As the tICA lag time is increased, the analysis picks up inter-basin correlations, with a corresponding increase in predicted timescales and barrier heights. The saturation observed in t_2 between tICA lag times of 0.2 and 2.0 ns likely indicates that, over these timescales, ubiquitin is able to sample almost completely the energy wells within its native basin. The fall-off in t_2 (and in barrier heights) at longer tICA lag times is likely

due to a loss of statistics as the lag time is made large and the system makes direct 'hops' between deep minima, thus avoiding the sampling of the barriers. That is, since the t_2 from the MSM is being reported as the timescale of the slowest processes found by the tICA and at both long and short τ_{tICA} there are no large barriers sampled, the tICA coordinates, which are unit-free and do not encode lengthscales, return a similar quadratic or barrier-free surface to the MSM analysis, although in the latter case, the width of the effective energy well is larger, leading to the larger value of t_2. The inhomogeneity of the slow tICA dynamics is elaborated further in the Supplemental Material of [65], where the self-self overlap or stability of the eigenvectors of the slowest tICA modes is calculated as a function of tICA lag time.

Other procedures may be adopted to select the optimal tICA lag time, as discussed in the Supplemental Material of [65]. For example, Section 5.5 identifies the optimal lag time by optimizing the decay of the tICA time correlation functions (tcfs) of the local fluctuations in comparison with simulations. In that case as well the lag time of $\tau_{tICA} = 2.0$ ns appears to give a slightly better agreement with the decay of the tcfs measured directly from the simulation trajectory than other values of the lag time.

Free energy surface for a tICA mode at the selected lag time

Once the tICA lag time is defined, here $\tau_{tICA} = 2.0$ ns, we can compare the slow tICA modes with the slow LE4PD-XYZ modes. With the goal of determining if the LE4PD-XYZ approach provides an analysis that is comparable to the tICA, we build a procedure for the tICA modes that follows the steps of the LE4PD procedure, starting from the calculation of the tICs' free-energy surfaces. We define a Free Energy Surface (FES) for each of the tICA mode coordinates, starting from the space coordinates of

the fluctuations, projected into tICA modes using the tICA eigenvectors. Then, we compare the fluctuations and the time of barrier-crossing for each tICA mode to the results of the similar analysis performed using the LE4PD-XYZ mode coordinates (as defined in Section 5.1). Because the tICA coordinates are commonly identified as the coordinates defining the order parameters for the slow fluctuations, comparing the slow LE4PD-XYZ modes with the tICA predictions is of great interest. In the LE4PD-XYZ, slow diffusive modes emerge from the natural decomposition in normal modes of the dynamics of a protein starting from its equation of motion. Thus, the LE4PD-XYZ has the advantage over tICA of associating to each set of slow coordinates a specific equation of motion, providing a precise representation in time of the slow dynamics of interest. The question is if these slow LE4PD modes are able to reproduce with accuracy the slow dynamics dominating the protein trajectory, as identified by tICA.

As an example of the information inherent in $F(\theta_a, \phi_a)$ for the tICs, Figure 32 shows the results of the analysis in the (θ_a, ϕ_a) coordinate space for the slowest tIC extracted from the 1-μs simulation of ubiquitin. Figure 32a shows the free energy map, $F(\theta_1, \phi_1)$, for the first tICA mode, $z_1(t)$, with a pathway drawn between the two prominent minima on the surface. Figure 32b displays the fluctuations along the alpha-carbon backbone of ubiquitin when moving along the pathway given in Figure 32a; the colors of the structures in Figure 32b correspond to the colors of the images along the pathway in Figure 32a. Movement along the minimum energy pathway for $z_1(t)$ shows concerted fluctuations in the 50 s loop (blue arrow), the C-terminal tail (black arrow), and the Lys11 loop (red arrow), each of which is a known binding region of ubiquitin to other proteins.[52, 59, 75] Figure 32c shows the projection of the most slowly decaying eigenfunction, ψ_2, from the MSM transition matrix constructed

on this surface starting from the MD trajectory; the most positive projection of ψ_2 lies in the minimum in the bottom half of the surface, and the maximum projection of ψ_2 lies in the minimum in the top half of the surface. This spectrum indicates that the slowest process described by the MSM corresponds to transitions between the two minima on the surface, whose fluctuations should be described well by the extracted structures from the pathway given in Figure 32b. Finally, Figure 32d shows the implied timescale of t_2, the timescale of the process described by ψ_2, as a function of MSM lag time τ_{MSM}. The vertical dashed line marks the lag time used in the construction of the MSM shown in Figure 32c. Thus, for the ψ_2 shown in Figure 32c, the MSM transition matrix, $\mathbf{T}$, is constructed at a lag time $\tau_{MSM} = 4.0$ ns, and the predicted timescale is $t_2\,(\tau_{MSM} = 4.0 \text{ ns}) = 52.6$ ns. In summary, combining the tIC free-energy surface in (θ_a, ϕ_a) with the Markov state modeling analysis predicts that the timescale of movement between the two minima in Figure 32a is approximately 53 ns. The corresponding dynamics along the alpha-carbon backbone during this event are illustrated in Figure 32b.

Figure 33 illustrates the analogous analysis for the (θ_a, ϕ_a) surface spanned by the *second-slowest* tIC. Drawing a transition pathway between the two minima on the surface (Figure 33a) and extracting the structures along that pathway from the MD simulation shows that this tIC describes fluctuations in the Lys11 loop and C-terminal tail regions of ubiquitin (Figure 33b).[52, 75] Again, using the decomposition of ψ_2 from the MSM on this surface to choose the lag time of the MSM (Figure 33c), the process of transitioning between the minima on the surface is predicted to occur over a timescale of 6.7 ns (Figure 33d). Thus, the (θ_a, ϕ_a) surface for the slowest tIC predicts concerted motions in the C-terminal tail, 50 s loop, and Lys11 loop occurring

over a timescale of 52.6 ns while the (θ_a, ϕ_a) surface for the second-slowest tIC predicts mainly motion in the tail and Lys11 loop, occurring over a timescale of 6.7 ns.

Mapping the tICA modes onto the slow fluctuations predicted by the LE4PD-XYZ

The direct comparison of the free energy maps and fluctuation transitions for the first LE4PD-XYZ mode (Figure 29) and for the first (Figure 32) and the second (Figure 33) tICA modes show that the dynamics of these modes are different, even if the fluctuations involve comparable segments of the protein. It is, thus, interesting to understand how an identical trajectory analyzed by different methods can lead to different slow fluctuations. One wonders how the LE4PD and tICA methods differ and which method may be most useful in identifying the modes of proteins' slow dynamics.

An important detail that we have overlooked so far is the following: because each internal mode displays energy barriers that slow down the dynamics with respect to the harmonic fluctuations represented by the Langevin equation, the timescale of fluctuations of each Langevin mode is in practice slower than what is predicted by the straightforward diagonalization of the Langevin equation.[4, 38, 51] Different modes are slowed down differently, and some more internal modes may have larger energy barriers than the first mode so that it is possible that the slowest Langevin fluctuation in LE4PD is not the one predicted by the first Langevin mode but is the fluctuation belonging to some other internal mode. This is, in fact, the case for ubiquitin.

To calculate the transition times, we construct the MSM for each mode and estimate the timescales via the MSM's most slowly decaying timescale, t_2, either using the mapping of the second MSM eigenvector onto the FES,[4] or using the

markovian criterion of the transition (i.e. the Chapman-Kolmogorov [CK] condition) for the mode trajectories; the results from both approaches are reported in Tables 4 and 5, respectively. All dashed entries in the table denote surfaces where the extreme projections of ψ_2 are never located in minima on the surface and are thus not suited for Markov state modeling in the manner desired here.

From the timescales listed in Table 4, all the LE4PD methods give roughly the same timescales for the slowest motions of the system. The first tICA mode, however, displays dynamics that is one order of magnitude slower than LE4PD. The first tIC corresponds to the concerted motions in the three flexible binding regions of ubiquitin, as shown in Figure 32, and predicts this motion occurs almost ten times slower than the roughly analogous motion predicted by the isotropic LE4PD mode 6 and LE4PD-XYZ mode 7 with hydrodynamics, respectively. However, when the MSM lag time is selected using the CK condition, which does not always coincide with the lag time selected by optimizing the projection of ψ_2 from the MSM,[4] the gap between the predicted timescales of the slow LE4PD and tICA modes is reduced, as shown in Table 5.

These data indicate that the tICA procedure can group the slowest dynamics in a smaller number of modes than the LE4PD, which, instead, partitions the protein's slow dynamics into a number of leading modes with different time and length scales. When the goal is identifying the slowest fluctuations in one mode, tICA appears to be more efficient than the LE4PD in isolating the slow fluctuations. However, if the ultimate goal is the accurate analysis of the protein's slow dynamics, the LE4PD approach has a more desirable outcome. As shown in Section 5.5, the LE4PD can predict the dynamics as measured by time correlation functions with higher accuracy than the tICA modes.

TABLE 4. Comparing the slowest timescales from the isotropic LE4PD, the LE4PD-XYZ (ansiotropic LE4PD), and tICA for the 1-μs simulation of ubiquitin at the MSM lag time where the spectrum of ψ_2 on the free-energy surface is optimized.[4] The isotropic LE4PD modes are indexed by internal mode number.

	LE4PD	LE4PD-XYZ		tICA
		w/ HI	w/o HI	
Mode	$t_2(\tau)$, ns	$t_2(\tau)$, ns	$t_2(\tau)$, ns	$t_2(\tau)$, ns
1	3.9(1.05)	8.0(3.2)	6.5(2.8)	52.6 (4.0)
2	0.7(0.1)	3.7(1.1)	4.6(1.5)	6.7 (1.6)
3	0.9(0.35)	4.3(2.5)	4.3(2.0)	4.8(1.6)
4	2.4(0.5)	6.4(4.0)	1.0(0.3)	—(—)
5	0.1(0.01)	4.8(4.0)	5.5(4.9)	2.5(1.0)
6	11.0(0.9)	3.3(1.0)	3.1(2.0)	—(—)
7	0.5(0.25)	3.6(2.0)	7.4(3.0)	2.5(1.6)
8	0.4(0.11)	0.6(0.2)	—(—)	6.1(5.0)
9	0.24(0.1)	0.3(0.1)	1.3(0.5)	0.8(0.3)
10	0.35(0.3)	0.4(0.1)	0.4(0.2)	7.0(5.0)

TABLE 5. Comparing the slowest timescales from the isotropic LE4PD, the LE4PD-XYZ (ansiotropic LE4PD), and tICA for the 1-μs simulation of ubiquitin in the long-lag time regime where the dynamics best satisfy the Chapman-Kolmogorov condition.[5] The isotropic LE4PD modes are indexed by internal mode number.

	LE4PD	LE4PD-XYZ		tICA
		w/ HI	w/o HI	
Mode	$t_2(\tau)$, ns	$t_2(\tau)$, ns	$t_2(\tau)$, ns	$t_2(\tau)$, ns
1	5.3(1.8)	14.6(12.0)	16.2(12.0)	54.0 (5.0)
2	3.3(1.6)	14.4(10.0)	16.6(12.0)	12.6 (5.0)
3	1.9(1.2)	9.6(8.0)	9.2(8.0)	10.5 (5.0)
4	4.7(1.6)	7.2(6.0)	9.5(8.0)	9.1 (5.0)
5	3.6(1.6)	4.8(4.0)	7.7(6.0)	9.3(5.0)
6	33.7(25.0)	4.6(4.0)	21.5(12.0)	6.6(5.0)
7	1.2(1.0)	19.9(12.0)	12.6(10.0)	5.7(5.0)
8	3.0(1.6)	2.4(2.0)	4.6(4.0)	6.1(5.0)
9	0.5(0.4)	4.0(3.5)	1.8(1.5)	6.4(5.0)
10	0.35(0.3)	1.3(1.0)	3.7(3.0)	7.0(5.0)

One detail to note from Table 4 is that even though the tICs are sorted in descending order of decorrelation time (i.e., in the order of the tICA eigenvalues)

when the barriers or anharmonicity along the tICA coordinates are accounted through the Markov state modeling, the relative timescales of the tICs change, as they do for the LE4PD modes.[4, 38, 51] An inspection of the internal mode-dependent energy barriers and related transition times (see Table 4) identifies as the slowest LE4PD-XYZ fluctuations those of modes 6 and 7. Thus, we compare the predictions of those slow LE4PD-XYZ modes with the ones emerging from tICA.

Similarities between the tICA predictions and the predictions of the isotropic and anisotropic LE4PD

The main point of this section is the comparison between the slowest tICs and the slowest modes of the anisotropic LE4PD model without hydrodynamic interaction included. In chapter IV we show how the covariance matrix in the PCA is equivalent to the matrix of forces that lead the dynamics in the anisotropic LE4PD-XYZ model *when the effect of hydrodynamic interactions is neglected*, which is equivalent to setting the **H** matrix in Eq.5.6 equal to the identity matrix.

Other approximations needed for the two approaches to be consistent are the assumption of a uniform friction coefficient for all amino acids and the approximation of neglecting internal energy barriers. Our study showed that hydrodynamics modifies the dynamics predicted by the equation of motion. We also observed an almost quantitative agreement between the time correlation functions directly calculated from the simulation and the ones obtained by solving the anisotropic Langevin equation when hydrodynamics is included.[51] This result confirms the importance of hydrodynamics in the Langevin dynamics of proteins in solution, which is not surprising given that the Langevin is an equation of motion in the protein's reduced coordinates, where the effect of the solvent enters through friction, random forces, and

hydrodynamic interactions. Thus, the hydrodynamic forces that enter the LE4PD equations result from the projection of the forces due to the solvent and the protein's atomistic fast degrees of freedom onto the reduced coordinates of the alpha carbons.

Nevertheless, the formal connection between LE4PD-XYZ and the PCA is obtained when hydrodynamic interactions are not included.[51] Because PCA is, in effect, the zero lag time limit of the tICA formalism, in this section, we compare modes from tICA and from LE4PD-XYZ *where hydrodynamics is discarded*. We also include data from the full isotropic LE4PD-XYZ (with hydrodynamics) and compare them to tICA.

Comparing the tICA, the isotropic LE4PD (with hydrodynamics), and the LE4PD-XYZ (without hydrodynamics) free energy surfaces

Figure 34 displays in each row the comparison between the LE4PD slowest modes and the tICA slowest mode for the two LE4PD models we study, namely the isotropic and the anisotropic LE4PD theory. In the first column, Figure 34 shows the FES of the LE4PD projected trajectory, which displays energy minima for the most populated conformations of the protein. For this FES, the second column of Figure 34 presents the second eigenvector obtained from the Markov State Model (MSM) analysis of the FES. The second MSM eigenvector, ψ_2, provides information on the slowest kinetic transition occurring on the FES. The node in the MSM second eigenfunction identifies the position of the maximum of the energy barrier, while the maximum and minimum values of the second eigenvector defines the two most important energetic basins in the FES.[33, 99, 144] The superposition of the second MSM eigenvector to the LE4PD energy map indicates which transition represents the slowest fluctuation for the given LE4PD mode.

Finally, the third column in Figure 34 shows the comparison between a tICA mode and the LE4PD mode. The superposition is accomplished by projecting the first tIC onto the LE4PD free energy map and testing if the most extreme tICA conformations are the ones that correspond to the minima in the LE4PD FES. To perform this comparison, we assign each conformation in the tICA mode trajectory to the closest MSM microstate in the LE4PD-mode FES surface, using the root mean square distance from each MSM microstate as the assignment metric. Then the tICA mode trajectory populates the FES, giving a projection of the tICA mode that is completely analogous, in both meaning and interpretation, to the projection of an eigenvector ψ_i from the MSM onto the LE4PD FES (see the second column of Figure 34). The approach of projecting a tICA mode onto a free-energy surface has been previously applied by Sultan and Pande[207] to verify the interpretation for the slowest tIC from a simulation of alanine dipeptide.

When projecting the tICs, $z(t)$, onto the (θ_a, ϕ_a) surfaces, the average of $z(t)$ within each MSM LE4PD microstate i, S_i, is calculated as

$$\langle z(t) \rangle_i = \frac{1}{M_i} \sum_{k=1}^{M_i} z(k), \qquad \forall \, (\theta_a(k), \phi_a(k)) \in S_i,$$

with M_i the total number of frames the $z(t)$ trajectory resides in the S_i LE4PD microstate over the course of the simulation. This local average of $z(t)$ within each of the discrete states is what is reported in Figure 34.

Since the slowest tIC is the optimal linear approximation to the full-space Markov propagator of the system [30]. The ψ_2 from the MSM on the slowest LE4PD modes are also estimators of the slowest processes of the system; a high similarity between the projected spectra of the slow tICs and ψ_2 indicate high similarity between the predicted dynamics from the two models. That is, if the slow dynamics predicted

in each approach are consistent with each other, then the spectra of both the slow tICs and ψ_2 should predict probability flow between the deep minima on the (θ_a, ϕ_a) surfaces of the slowest LE4PD modes. The ψ_2 are already parameterized to do so, [4] but the slow tICs are, in principle, ignorant of the LE4PD (θ_a, ϕ_a) surface. We use this technique to confirm that the slow LE4PD modes can extract the slow dynamics compatible with tICA modes.[22]

Let's analyze Figure 34 by looking at one row at a time. We see that the first row shows the slowest tICA mode (mode 1) in comparison with the slowest mode of the *isotropic* LE4PD (mode 6) with hydrodynamics, as reported in Table 4. We observe that the extrema of the tICA mode correspond to the two most populated regions of the isotropic LE4PD mode. This indicates that movement along the tICA's slowest collective coordinate corresponds to movement between the minima on the LE4PD FES or that the fluctuations selected by the two modes are strongly related. Similar behavior is observed for the slowest modes in the isotropic LE4PD without hydrodynamic interactions (which is not reported here).

The second row compares the slowest tICA mode (mode 1) with the slowest mode (mode 7) of the *anisotropic* LE4PD without hydrodynamics, as reported in Table 4. Again the agreement between the two methods is compelling.

In the third row, instead, Figure 34 shows a comparison between the third tICA mode (mode 3) and the fifth mode of the *anisotropic* LE4PD without hydrodynamics. It is clear from these results that the slow dynamics detected by tICA and LE4PD-XYZ are similar.

The technique used here of projecting the tICs onto the (θ_a, ϕ_a) surfaces of the LE4PD modes is analogous to the technique used in [216–219] to model experimental observables using Markov state models. Like an experimental observable, the

separation of two minima of the (θ_a, ϕ_a) surfaces into 'high z' and 'low z' states indicates that transitions on the (θ_a, ϕ_a) surface correspond to transitions between a high z state and a low z state, similar to how fluorescence experiments on a protein search for transitions between a high fluorescence state, indicating the protein is sampling conformations where the fluorophores are far apart, and a low fluorescence state, where the protein is sampling conformations where the fluorophors are close together.[220, 221]

In conclusion, Figure 34 shows that the slowest tICA mode is representing well the slowest *isotropic* LE4PD mode, which is mode 6, and the slowest *anisotropic* LE4PD without hydrodynamics, which is mode 7.

Note that no anisotropic LE4PD mode with hydrodynamics included shows the extreme values of z_1 projected into the minima of the free-energy surface, despite the high correlation between $z_1(t)$ and the seventh anisotropic LE4PD mode with hydrodynamics, as shown in the Supplementary Material of of [65]. This absence is likely related to the anisotropic LE4PD with hydrodynamics failing to isolated into a single mode the slow deformations of the 50 s loop of ubiquitin,[51] which is the dynamics described by $z_1(t)$, as seen in Figure 32.

In conclusion, Figure 34 demonstrates that both LE4PD approaches are able to capture the same slow motion as the tICA, and, furthermore, they both funnel the dynamics into the slowest modes. The correlation between the time series of z_1 and ψ_2 from the MSM of the slowest isotropic LE4PD mode is high ($\rho = 0.92$), indicating that both z_1 and ψ_2 are predictive of the slow dynamics in ubiquitin. The correlation coefficient between the time series of z_1 and ψ_2 from the MSM of the slowest anisotropic LE4PD mode is $\rho = 0.73$, which is still acceptable.

Comparing tICA's and LE4PD's local fluctuations

Next, we compare the dynamics predicted by the slow tICs and LE4PD modes by calculating the mode-dependent fluctuation profiles as a function of amino acid sequence along the backbone of the protein. Local fluctuations are well represented by the local mode lengthscale (LML). The definition of the LML for the isotropic[4, 48] and the anisotropic[51] LE4PD models have been described previously. For the isotropic LE4PD, the local mode lengthscale (LML) is defined as a function of the mode, a, and the aminoacid position, i, as:

$$\mathrm{LML}_{ia}^2 = Q_{ia}^2 l^2 \mu_{a,\mathrm{LE4PD}}^{-1} \ , \tag{5.24}$$

where $\mu_{a,\mathrm{LE4PD}}$ determines the mean-square amplitude of LE4PD mode a (see Figure 35).[37, 38] In the anisotropic formalism of LE4PD-XYZ, where $\langle \Delta \vec{R}_i \cdot \Delta \vec{R}_i \rangle = \langle \Delta x_i^2 \rangle + \langle \Delta y_i^2 \rangle + \langle \Delta z_i^2 \rangle$, the eigenvectors are partitioned into their $x-$, $y-$, and $z-$components, thus isolating the $x-$, $y-$, and $z-$projections of LML_{ia}^2 as:

$$\mathrm{LML}_{ia,x}^2 = \left(Q_{ia}^x\right)^2 \mu_{a,\mathrm{LE4PD\text{-}XYZ}}^{-1} \tag{5.25}$$

$$\mathrm{LML}_{ia,y}^2 = \left(Q_{ia}^y\right)^2 \mu_{a,\mathrm{LE4PD\text{-}XYZ}}^{-1} \tag{5.26}$$

$$\mathrm{LML}_{ia,z}^2 = \left(Q_{ia}^z\right)^2 \mu_{a,\mathrm{LE4PD\text{-}XYZ}}^{-1}, \tag{5.27}$$

where $\mu_{a,\mathrm{LE4PD\text{-}XYZ}}$ are the eigenvalues of $\mathbf{A}'$ [51]. These three isotropic terms can be summed to generate an isotropic LML profile analogous to eq. 5.24, the square root of which is what is shown in Figure 36.

For the tICA, fluctuations are derived from the definition of the modes, $z_a(t) = \sum_i \Omega^T_{ai} \Delta R_i(t)$, and the Moore-Penrose generalized inverse[95] of Ω^T, Ω^{-1T} as

$$\Delta R_i(t) = \sum_a \Omega^{-1T}_{ia} z_a(t).$$

The mean-square fluctuations of residue i, given by $\langle \Delta x_i(t) \Delta x_i(t) \rangle + \langle \Delta y_i(t) \Delta y_i(t) \rangle + \langle \Delta z_i(t) \Delta z_i(t) \rangle$, can be written in terms of the tICs as

$$\begin{aligned}
&\langle \Delta x_i(t) \Delta x_i(t) \rangle + \langle \Delta y_i(t) \Delta y_i(t) \rangle + \langle \Delta z_i(t) \Delta z_i(t) \rangle \\
&= \sum_a \sum_b \left(\Omega^{-1T}_{ia,x} \Omega^{-1T}_{ib,x} + \Omega^{-1T}_{ia,y} \Omega^{-1T}_{ib,y} + \Omega^{-1T}_{ia,z} \Omega^{-1T}_{ib,z} \right) \langle z_a(t) z_b(t) \rangle \\
&= \sum_a \left(\Omega^{-1T}_{ia,x} \right)^2 + \left(\Omega^{-1T}_{ia,y} \right)^2 + \left(\Omega^{-1T}_{ia,z} \right)^2, \qquad (5.28)
\end{aligned}$$

where the definition $\langle z_a(t) z_b(t) \rangle = \delta_{ab}$ is used to obtain Eq. (5.28). The tICA LMLs are reported in Figure 37

Thus, Figures 35, 36, 37 show the mode-dependent fluctuations calculated from the one-microsecond ubiquitin simulation using the isotropic LE4PD, anisotropic LE4PD without hydrodynamics, and the tICA, respectively, for the first ten processes of each method. For both the isotropic with hydrodynamics and the anisotropic without hydrodynamics LE4PD approaches, most of the slow modes describe fluctuations in the C-terminal tail of the protein. For tICA, the slowest tIC describes concerted fluctuations in the Lys11 and 50 s loops of the protein, while neither of the LE4PD approaches gives a single mode describing simultaneous motion in these two regions of the protein.

The isotropic LE4PD approach shows in Figure 35 that the tail and the loop around Lys 11 have slow fluctuations that involve multiple modes, and thus multiple

length scales. From the analysis of the FES we know that the slowest fluctuations appear in the internal mode 6 where the dynamics take place mostly around residue 50. Slow dynamics in these three regions show in the modes of the anisotropic LE4PD-XYZ without hydrodynamics (see Figure 36), but the fluctuations are partitioned in multiple modes. These findings agree with the identification of the slowest fluctuations in the tICA modes (see Figure 37). We observe that there is good correspondence between the slowest tIC and the isotropic LE4PD mode 6 and anisotropic LE4PD mode 7, when hydrodynamic effects are neglected.

These conclusions are in agreement with the analysis of the energy maps in Section 5.4, where we found that the fluctuations of ubiquitin predicted from the slow LE4PD modes and the tICs agree. This finding also shows how the tICA method is most efficient in isolating the slow protein dynamics in a small number of modes. However, tICA doesn't allow one to distinguish the different timescales of the slowest fluctuations.

Testing the tICA and LE4PD predictions of time correlation functions against simulations.

While all three methods agree in identifying the regions in the protein where slow fluctuations occur, the ultimate test of the tICA's and LE4PD's ability to predict slow time dynamics is to directly compare the time correlation functions (tcfs) predicted from both approaches to the tcfs calculated from the simulation trajectory. In this case, we compare the tICA predictions with the anisotropic LE4PD model.

The normalized autocorrelation function for the fluctuations of each residue is defined as $C(t) = \frac{\langle \Delta\vec{R}(t)\cdot\Delta\vec{R}(0)\rangle}{\langle \Delta\vec{R}(0)\cdot\Delta\vec{R}(0)\rangle}$. For the LE4PD approaches, the autocorrelation function is calculated by including for each mode the slowing down of the dynamics

due to the presence of an energy barrier in the FES. This energy barrier is included by rescaling the mode-dependent timescale using Kramers' theory of reaction kinetics.[4, 51] Recently, we have shown that neglecting the hydrodynamic interaction modifies the LE4PD-XYZ curves, leading to a (moderately) worse agreement with the simulation data.[51] Figure 38 shows the fluctuation decay of the tcfs for residues sampled along the primary sequence of ubiquitin. The figure compares the LE4PD-XYZ results with hydrodynamic included to the tcfs from the simulations: the agreement is remarkable. It also shows the tcfs for the LE4PD-XYZ without hydrodynamic interactions, which are less in agreement, at least for the residues presented in the figure.

At a given lag time, $C(t)$ can be written in terms of the tICA eigenspectra by inverting the relationship $z_a(t) = \sum_i \Omega^T_{ai} \Delta R_i(t)$ as $\Delta R_i(t) = \sum_a \Omega^{-1T}_{ia} z_a(t)$, and using the (near) independence of the tICs $\langle z_a(t) z_b(0) \rangle \approx \langle z_a(0) z_b(0) \rangle \exp\left[-t/\tau_a\right] = \delta_{ab} \exp\left[-t/\tau_a\right]$ as

$$\begin{aligned} C(t) &= \frac{\langle \Delta\vec{R}(t) \cdot \Delta\vec{R}(0) \rangle}{\langle \Delta\vec{R}(0) \cdot \Delta\vec{R}(0) \rangle} \\ &= \frac{\sum_{a,b} \left[\left(\Omega^{-1T}_{ia,x} \Omega^{-1T}_{ib,x} \right) + \left(\Omega^{-1T}_{ia,y} \Omega^{-1T}_{ib,y} \right) + \left(\Omega^{-1T}_{ia,z} \Omega^{-1T}_{ib,z} \right) \right] \langle z_a(t) z_b(0) \rangle}{\sum_{a,b} \left[\left(\Omega^{-1T}_{ia,x} \Omega^{-1T}_{ib,x} \right) + \left(\Omega^{-1T}_{ia,y} \Omega^{-1T}_{ib,y} \right) + \left(\Omega^{-1T}_{ia,z} \Omega^{-1T}_{ib,z} \right) \right] \langle z_a(0) z_b(0) \rangle} \\ &= \frac{\sum_a \left[\left(\Omega^{-1T}_{ia,x} \right)^2 + \left(\Omega^{-1T}_{ia,y} \right)^2 + \left(\Omega^{-1T}_{ia,z} \right)^2 \right] e^{-t/\tau_a}}{\sum_a \left[\left(\Omega^{-1T}_{ia,x} \right)^2 + \left(\Omega^{-1T}_{ia,y} \right)^2 + \left(\Omega^{-1T}_{ia,z} \right)^2 \right]}. \end{aligned} \tag{5.29}$$

The decay timescales for each tICA mode, τ_a, are calculated empirically by the integration of the autocorrelation function $\langle z_a(t) z_b(0) \rangle / \langle z_a(0) z_b(0) \rangle$ obtained from the simulations,[169] and assuming that each mode is represented by a single exponential decay. This procedure should account for the barriers present along each tICA coordinate in, at least, a coarse manner.[51, 161] This time, τ_a, is in general different

from the inverse of the eigenvalues λ_{IC} (Eq. 5.11) because that time does not include the mode-dependent energy barrier (represented in Figure 39). If one adopted the inverse of the eigenvalues λ_{IC} as the timescale of decay, the tcfs calculated from tICA would display an even faster and more unphysical decay than the one observed when including mode-dependent energy barrier for tICA (see Fig. 38). Once a lag time is selected, we build the matrix $\mathbf{C}(\tau)$ and, by diagonalization, we derive the eigenvectors and eigenvalues that enter Eq. 5.29.

The time correlation functions calculated from the tICs (Eq. 5.29) are directly compared to the one from the simulation trajectory in Figure 38. For each residue shown, and for most residues across the primary sequence of ubiquitin, the tcfs predicted from the LE4PD-XYZ with hydrodynamics are in better agreement with the simulated tcfs than those predicted from the tICA or the LE4PD-XYZ without hydrodynamics. One may assume that this disagreement is due to the selection of the lag time and that it may be possible to improve the quality of the tICA predictions by adjusting the tICA lag time. However, Figure 39 shows that, on average, changing the lag time cannot make the tICA predictions superior with reference to the LE4PD-XYZ tcfs. This suggests that the separation of the dynamics afforded by the LE4PD-XYZ is more optimal for modeling the tcfs than the tICA mode description.

Confirming the tICA lag time using the time correlation functions

Figure 39 shows the $C(t)$ calculated both from the simulation trajectory and the tICA theory, using a range of tICA lag times from 2.0 ps to 20.0 ns, for six residues spaced along the primary sequence of ubiquitin. Even if the tICA decomposition of the dynamics offers a more efficient separation of the slow fluctuations into a few modes,[30] based on the tcf agreement, the LE4PD-XYZ with hydrodynamics

provides a more accurate representation of the general collective dynamics at all timescales sampled in the trajectory. This is observed at all the lag times and for all the residues along the entire primary sequence of ubiquitin as shown in Figure 38. In general, the decays predicted by tICA at any lag time tend to be too fast. However, at short observation times ($0 - 10$ ns; subfigure insets in Figure 39), and for the very long time scale, the tICA predictions can be quite good.

Interestingly, although the range of lag times spans five orders of magnitude, the decay is similar when adopting any lag time between 2 ps and 2 ns. To select the optimum lag time, we average the results of the tcfs at a given lag time across all residues in the primary sequence of ubiquitin. By comparing with the simulated tcfs, we identify an optimal lag time of 2.0 ns. Note that this optimal lag time is consistent with the one identified in Section 5.2 from the analysis of the free energy surfaces performed at increasing lag time.

While using a tICA lag time of 2 ns globally optimizes the agreement between the simulated tcfs and those predicted by tICA, choosing either a longer or shorter tICA lag time may give a better agreement in the tcf of specific bonds. For example, Figure 40 shows how using a shorter lag time (2 ps) when calculating the tICA time-lagged covariance matrix yields tcfs in good agreement with some residues' tcfs in the highly flexible Lys11 loop, especially at timescales less than 10 ns. Similarly, using a longer lag time (20 ns) gives tICs that agree well with the simulated tcfs of several residues in the 50s loop, where the slowest fluctuations of the protein occur (Figures 39 and 32). This analysis supports the heterogeneity of ubiquitin dynamics since one can locally optimize the residues' relaxation in different regions by varying the tICA lag time.

2D Maps with tICA slow coordinates

In what has become a fairly typical workflow for the analysis of MD simulations of biomolecules using Markov State Models, the MD trajectory is projected onto not just one mode but a number n of the slowest tICA modes. [22, 30, 31, 36, 86, 222, 223] This procedure reduces the high dimensionality of the original free energy landscape by identifying the slowest dominant modes. One then performs an MSM analysis on the reduced subspace to parse the slowest dynamics and corresponding timescales of the system.[30, 31, 50, 85, 86, 188, 224] Usually, the two slowest modes are selected, but in some cases, more than two tICA modes are considered, which can lead to even slower measured kinetic timescales. [31] This is because the transitions among the selected modes can become even less probable, while statistical insufficiencies in the necessarily finite simulation data can also play a role.

This approach allows for the improvement of the original, linear, input coordinates provided by tICA by using the eigenvectors of the transition matrix calculated on the reduced tICA-space to account for non-linearities along each tICA coordinate and correlations among the tICA coordinates. Thus, transitions of the trajectory among different slow modes represent the slowest dynamics in the MD trajectory between the intra-lobe minima on the left-hand lobe (Figure 42). The MSM predicts that this transition occurs over a timescale of $\sim$70 ns and that the transition causes motions in the Lys11 of ubiquitin. Note that the same slow fluctuations are identified by the LE4PD models.

Thus, although the timescales predicted using the space spanned by the first two tICs are slightly slower than using the (θ_a, ϕ_a) surfaces for the first two tICs individually ($\sim$40 versus 24.0 and $\sim$20 versus 10.7, respectively) the timescales are

still within a factor of 2 in both cases. The qualitative dynamics are predicted to be similar from both methods (comparing Figures 32 and 33 with Figures 41 and 42).

Here, we report the results of this type of 'traditional' tICA-MSM approach for the 1-μs ubiquitin simulation, and we compare the results to the analogous analysis performed in the (θ_a, ϕ_a) space of a tIC of index a, thus considering projections onto the single mode.

To maintain the same dimensionality as the MSM on the (θ_a, ϕ_a) surfaces, we build here the MSMs on the space spanned by the first two tICs calculated from the ubiquitin simulation at a tICA lag time of $\tau_{\text{tICA}} = 2$ ns, the same as that used in the (θ_a, ϕ_a) surfaces presented in Section5.2. Figure 41 shows the free-energy surface spanned by the first two tICs, which has two 'lobes' having two minima each. When a MSM is constructed on this surface, it predicts that the slowest motion spanned by this two-dimensional space is a transition between the two lobes, i.e. the transition between the two tICA modes, as can be seen by an examination of the spectrum of ψ_2 projected onto the free-energy surface (Figure 41, top left panel). The MSM predicts that the transition between the two lobes occurs over a timescale of approximately 200 ns. Tracing a pathway between the two deepest minima in each lobe using the same method as utilized for the (θ_a, ϕ_a) surfaces shows that inter-lobe transitions correspond to dynamics in the 50 s loop of ubiquitin. The second-slowest relaxation processes on the surface spanned by the first two tICs correspond to movement between the intra-lobe minima on the left-hand lobe (Figure 42). The MSM predicts that this transition occurs over a timescale of $\sim$70 ns and that the transition causes motions in the Lys11 of ubiquitin. Note that the same slow fluctuations are identified by the LE4PD models.

Thus, although the timescales predicted using the space spanned by the first 2 tICs are slightly slower than using the (θ_a, ϕ_a) surfaces for the first 2 tICs individually ($\sim$40 versus 24.0 and $\sim$20 versus 10.7, respectively) the timescales are still within a factor of 2 in both cases, and the qualitative dynamics are predicted to be similar from both methods (comparing Figures 32 and 33 with Figures 41 and 42).

Observation of non-homogeneous dynamics as tICA lag time is adjusted

To confirm the tICA lag time selection, we started from the 2D tICA FES, and we explored the dynamics on the tIC 1 - tIC 2 energy surface while varying the lag time and examining the trajectories between minima. This analysis is analogous to the one we completed of the FES for one tICA mode (see Figure 30). Such an analysis of the 2D FES shows how both the FESs and the associated mode-dependent dynamics change as the lag time of the tICA is adjusted. Figure 43 compares the FESs and dynamics along the pathways between minima on these surfaces when the tICA lag time is adjusted from $\tau_{tICA} = 0.2$ ps to $\tau_{tICA} = 20.0$ ns. When $\tau_{tICA} \leq$ 20.0 ps, the two slowest tICs describe mainly dynamics in the C-terminal tail of the protein (Figure 36). However, once τ_{tICA} rises above 0.2 ns, the motion shifts to a combination of fluctuations in the Lys11 and 50s loops. This shift in the foci of the dynamics coincides with an increase in the barrier between the two minima on the surface and the attainment of Markovian dynamics for the transitions between minima (Figure 43). Once the lag time rises above 2.0 ns, the barriers between minima begin to decrease, and the dynamics on the surface become again non-Markovian, as described in Figure 44. This behavior is consistent with that observed in Figure 31.

This result suggests two conclusions. First, the barrier to conformational change in the Lys11 and 50 s loops are larger than those in the C-terminal tail, because the

free-energy barrier rises when the dynamics of the two slowest tICs shifts from short lag times (where the tICs describe the dynamics in the protein's tail) to longer lag times (where the tICs describe dynamics in the Lys11 and 50 s loops). That the energy barriers in the tails are smaller than in the loop is not surprising, given the intrinsically disordered nature of the tail region. Second, at least for these surfaces, there is a strong correlation between the observed barrier height and how Markovian is the dynamics (see Figure 44). Again, this result is probably not surprising since large barriers between conformational states are required to 'erase' the intra-state memory and generate Markovian dynamics among states.[100, 225] Similar qualitative results were seen when the analysis is performed on the (θ, ϕ) surfaces (Figures 30 and 31). The conformational transitions predicted in ubiquitin's structure when moving between minima on the surface are similar between the tIC 1- tIC 2 and (θ, ϕ) at each given tICA lag time: in both approaches, at faster lag times the C-terminal tail dominates the dynamics, while at slower lag times the 50 s loop and Lys11 loop dominate.

The only significant difference is a lower correlation between the barrier height along the pathway between minima and the t_2 timescale from the MSM on the surface (see Figure 31), especially for short ($\leq$ 20.0 ps) tICA lag times. The lower correlation coefficient likely indicates that the transitions between minima on the (θ, ϕ) surface are more heterogeneous than those on the tIC 1 - tIC 2 surface, where the high correlation between the barriers and MSM timescale indicates that the pathway picks out the bottleneck in the transition pathway between minima. That is, on the (θ, ϕ) surface, there are likely more pathways than just the one over the saddle point between minima carrying significant flux between the two minima, at least for the surface made with $\tau_{tICA} \leq 20.0$ ps, where the correlation between the MSM t_2 and the barrier along

the lowest energy transition pathway is poor, while the transitions become Markovian at longer MSM time lag.

Methods: Computer simulations and Markov State Modeling

Equilibrium Molecular Dynamics Simulation of Ubiquitin

The MD simulations of ubiquitin were generated using GROMACS version 5.0.4,[137] and the AMBER99SB-ILDN atomistic force field,[179] on the Comet supercomputer at the San Diego Supercomputing Center. The starting structure was taken from the Protein Databank, PDB ID: 1UBQ.[53] 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 that used in nuclear magnetic resonance experiments of ubiquitin.[54] We subjected the protein-solvent system to two rounds of equilibration: first, a 50-ps equilibration in the NVT ensemble at 300 K, with the temperature-controlled using a Nosé-Hoover thermostat; 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 again used to maintain the temperature. We used the last frame of this burnout run as the initial configuration for the 1 μs production run, which utilized 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, the entire trajectory was deemed to fluctuate around an equilibrium value,[4] and the entire 1-μs of trajectory was used for the subsequent LE4PD and MSM analysis. We used the LINCS algorithm[141] to constrain all hydrogen-to-heavy-atom bonds in the system and adopted an integration

timestep of 2 fs during both the equilibration and MD simulation. We saved the trajectory to file every 100 integration steps (every 0.2 ps), obtaining a total of $\frac{10^6 \text{ ps}}{0.2 \text{ ps/frame}} = 5 \times 10^6$ frames for analysis.

The MD simulation protocol is the same as that given in Appendix A. However, the post-processing steps are different. Before performing the tICA, the 'raw' MD trajectory is processed to remove the rigid-body rotational and translational motions. First, the reference frame, the first frame in the MD simulation, is centered at the origin of the simulation box. Then, all subsequent frames are centered on this reference structure, and all frames where the protein is broken across the periodic boundaries are made whole. Finally, the rotational motion is removed by fitting each frame in the trajectory to the first, centered frame of the trajectory.

Markov State Modeling

This study adopts the Markov State Model approach to evaluate each normal mode's fluctuations' timescale for both the LE4PD and the tICA. Given the number of resources available describing the theory and application of Markov state models (MSMs) to the analysis of protein dynamics,[31, 33, 49, 81, 85, 226–228] we only present a brief review of the method. To construct an MSM from an MD simulation, one first identifies a subset of the degrees of freedom or important coordinates. Then one constructs MSM in the state space of these essential collective coordinates. Here, these collective coordinates are the isotropic or anisotropic LE4PD modes or the tICs. Second, a sample space of a small number of these important coordinates is discretized by assigning frames from the trajectory to an appropriate volume of the sample space. Third, the transitions among these discrete volumes of the sample

space are counted to build a transition matrix $\mathbf{T}$, with the elements T_{ij} defining the conditional probability of transitioning from discrete state i to discrete state j.

The MSM transition matrix is parameterized by a lag time τ_{MSM}, such that the eigenvalues and eigenvectors of $\mathbf{T} = \mathbf{T}(\tau_{MSM})$ are generally functions of the MSM lag time; for truly Markovian dynamics, the eigenvalue decomposition of $\mathbf{T}(\tau_{MSM})$ is independent of τ_{MSM}.[99, 229] The eigenvalues of the MSM transition matrix are ordering by descending value of its eigenvalues $\Lambda^{MSM}(\tau_{MSM})$, with the first eigenvalue $\lambda_1^{MSM}(\tau_{MSM}) = 1$ and all other eigenvalues of modulus strictly less than 1. The first eigenprocess from the transition operator describes the stationary distribution, and all other eigenprocesses describe dynamic (i.e. decaying) processes of varying timescale. The timescale for the i^{th} process, t_i, is given by

$$t_i = -\frac{\tau_{MSM}}{\ln(\lambda_i^{\mathrm{MSM}})}.$$

Furthermore, the spectrum of the corresponding right eigenfunction of $\mathbf{T}((\tau_{MSM})$, ψ_i, details the dynamics described on the sample space over the timescale given by t_i.[34, 96, 174]

For the MSMs presented here, all steps are performed using the PyEMMA package (http://emma-project.org).[230] For all free-energy surfaces, the state space was broken into 1000 discrete states using the k-means++ algorithm,[142] which we found previously to be acceptably optimal for ubiquitin.[4] The transition matrix between discrete states is estimated using the reversible estimator given in [94]. lag times for the MSMs on the (θ_a, ϕ_a) surfaces are selected using the spectrum of ψ_2, as described in the following section. Lag times for the MSMs constructed on the surface spanned by the two slowest tICs are selected using the implied timescales test,[49, 231] as shown in the Supplemental Material of [65].

For the MSMs constructed on the (θ, ϕ) surfaces, the lag time, τ_{MSM}, is identified using the same procedure as in [4, 51]. Briefly, the spectrum of the second right eigenfunction ψ_2 of the MSM transition matrix is examined, and the largest MSM lag time such that the maximum and minimum projection of ψ_2 reside in the minima on the given free-energy surface is selected for construction of the MSM to be analyzed.

Discussion and conclusions

Atomistic MD simulations of proteins have been shown to describe with accuracy relevant biological processes. However, the leading behavior that guides the properties in biological systems, including the slow cooperative dynamics involved in protein binding and function, is often hidden by the broad spectrum of phenomena and the multitude of atomistic details displayed in simulations. Many studies have begun to rely on machine learning techniques to distill the essential leading kinetic information. Two of the most straightforward analytical tools in ML are the principal component analysis (PCA) or the time-lagged independent analysis (tICA).

The difference between the two methods is that PCA identifies large uncorrelated fluctuations but maps these fluctuations into linearly-interpolated processes.[173] Because protein fluctuations are intrinsically not linear, the PCA method provides an approximated and often unrealistic picture of the protein's fluctuation pathways.[51] One can obtain a more realistic representation of the anharmonic protein fluctuations by combining the tICA, where the covariance matrix samples fluctuations at a given lag time, with the MSM analysis of the trajectory projected onto the slowest tICA modes. [30, 31] Similarly to tICA, the LE4PD and LE4PD-XYZ approaches accurately model nonlinear protein motions by including the mode-dependent free-energy surfaces obtained by analyzing the protein's MD trajectory. [4, 4, 38, 48, 57, 67]

Among deep learning methods, more sophisticated approaches than tICA have been proposed to model nonlinear effects in protein motions, such as kernel tICA [232], state-free reversible VAMPnets [233], and time-lagged autoencoders (TAEs).[36, 234] However, the use of such deep learning approaches to modeling nonlinearities in dynamics often comes with an increased computational cost, paired with a loss of physical intuition for the system under study. The tICA coordinates are considered, in general, the optimal linear approximation to the order parameters for relevant slow processes in proteins' dynamics.[84, 85, 169, 214, 215] Unfortunately, the tICA modes have no *a priori* physical basis or any associated equation of motion. Nevertheless, the tICA modes can be seen as the limiting case of the more general relaxation mode analysis technique where the initial covariance matrix is calculated at time $t_0 \to 0$.[32, 203, 209]

In a previous publication of ours,[51] we compared the slow diffusive dynamics from the LE4PD with the slowest PCA modes. That study shows that the effects of hydrodynamics, residue-dependent friction coefficients, and mode-dependent energy barriers, which are explicit components of the LE4PD equation but are not part of the PCA, are essential to describe with precision the decay of protein fluctuations over the whole range of timescales involved.

Here, we compare the predictions of the tICA method to the isotropic and anisotropic LE4PD models. To do so, we associate to each tICA mode a free energy landscape obtained by the eigenvector projection of the simulation trajectory onto the tIC modes. This representation is convenient because it allows one to analyze the tICA's predictions based on the time evolution of fluctuations onto the mode-dependent free energy landscape.

Both tICA and LE4PD consistently identify the leading slow modes in the dynamics of the proteins. However, while tICA tends to collect all the slow processes in the first or a few modes, the LE4PD provides a more detailed picture of the time- and length-dependence of the slow dynamics, which are partitioned into a larger number of modes.

In general, the tICA captures the slow fluctuations that occur at a timescale longer than the given lag time, while faster dynamics are averaged out. The LE4PD method, instead, which is based on the solution of a "bead-and-spring" model of macromolecular dynamics, provides detailed information on the dynamics at the different length scales. It follows that the LE4PD is accurate in reproducing the time decay of aminoacid fluctuations at all timescales when the dynamics is represented by the time correlation functions calculated from the simulation trajectory (Figures 38 and 39). The similar calculation performed using tICA modes is, with a few exceptions, much less accurate (Figures 39 and 40).

The tICA's lack of accuracy in the description of the time dependence of the decay of the slow fluctuations as described by the simulated tcfs is not surprising because the tICA averages out the information at times shorter than the lag time. Setting a lag time for tICA affects the modality of sampling the dynamics in the free energy landscape. For example, if the lag time is too short or too long, the tICA cannot properly sample the free energy barriers.

In the meantime, the tICA is convenient because it can isolate in a small number of modes the slowest protein motions. When one performs a Markov State Model analysis of the kinetic transitions associated with the first few tICA modes, the method provides realistic values for the timescale associated with the slow, leading processes. On the contrary, the LE4PD approach, which accurately represents the

time correlation functions, requires information from many slow modes to recover the correct transition times.

In conclusion, if a rapid identification of the leading slow dynamics is required, the tICA analysis is a practical and valuable strategy to collect that information. However, suppose the time propagation of the slow leading dynamics is of interest. In that case, the LE4PD provides a more accurate representation of the slow processes based on its superior ability to reproduce the protein's dynamics at all times.

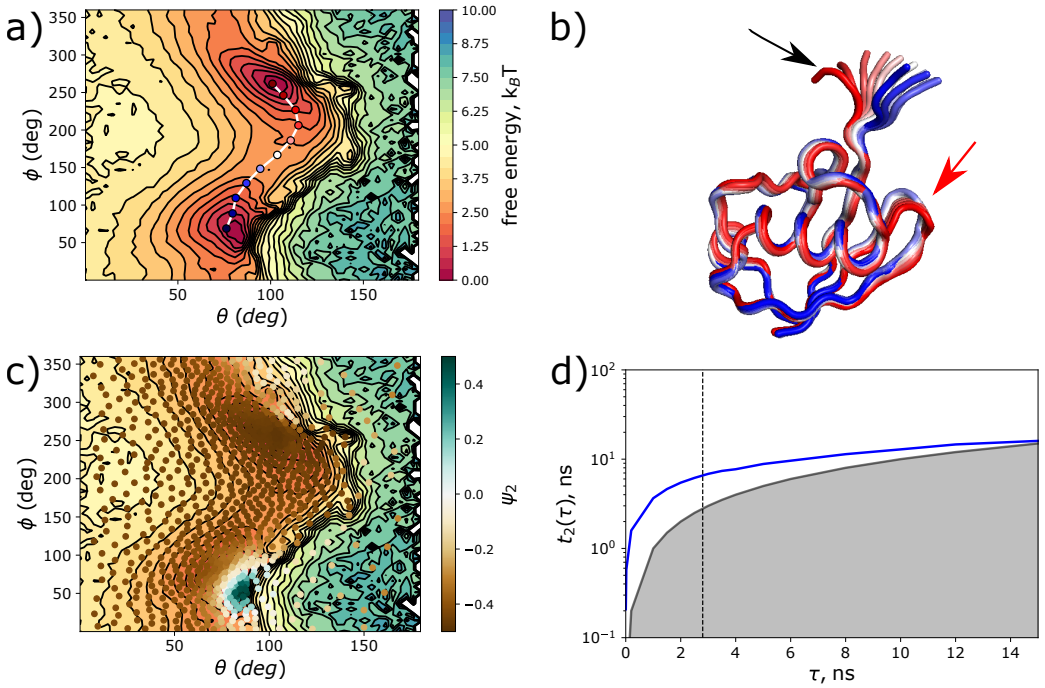

FIGURE 29. Analysis of the free energy map of the first LE4PD-XYZ mode without hydrodynamics (the map of the first LE4PD-XYZ mode with hydrodynamics is reported in the Supplemental Material of [65]. Panel a) shows the free energy landscape of the first LE4PD-XYZ mode in the two spherical coordinate reference system. The pathway of crossing the energy barrier between the two minima is identified with a rubber band, using a variant of the string method.[4] Panel b) shows ubiquitin's conformations that correspond to the pathway shown in panel a) with the red conformation referring to the energy minimum at the top of the map, and the blue conformation corresponding to the energy minimum at the bottom of the map. Arrows point to the two regions of ubiquitin showing the largest amplitude fluctuations: the C-terminal tail (black arrow), and the Lys11 loop (red arrow). The second eigenvector resulting from the diagonalization of the transition matrix defined in the Markov State Model (MSM) procedure for this mode identifies the two minima in the FES. The projection of ψ_2 onto the discrete states of the MSM has colors that correspond to the scaled-and-shifted value of ψ_2 at that discrete state, $\psi_2 = \frac{\psi_2 - \min(\psi_2)}{\max(\psi_2) - \min(\psi_2)} - 0.5$. Panel d) shows how the transition time for the second MSM eigenvector changes when we select a different lag time in the calculation of the MSM transition matrix (see Section 5.7). The black, vertical line demarcates the lag time corresponding to the second eigenvector mapping the two minima, as reported in panel c).

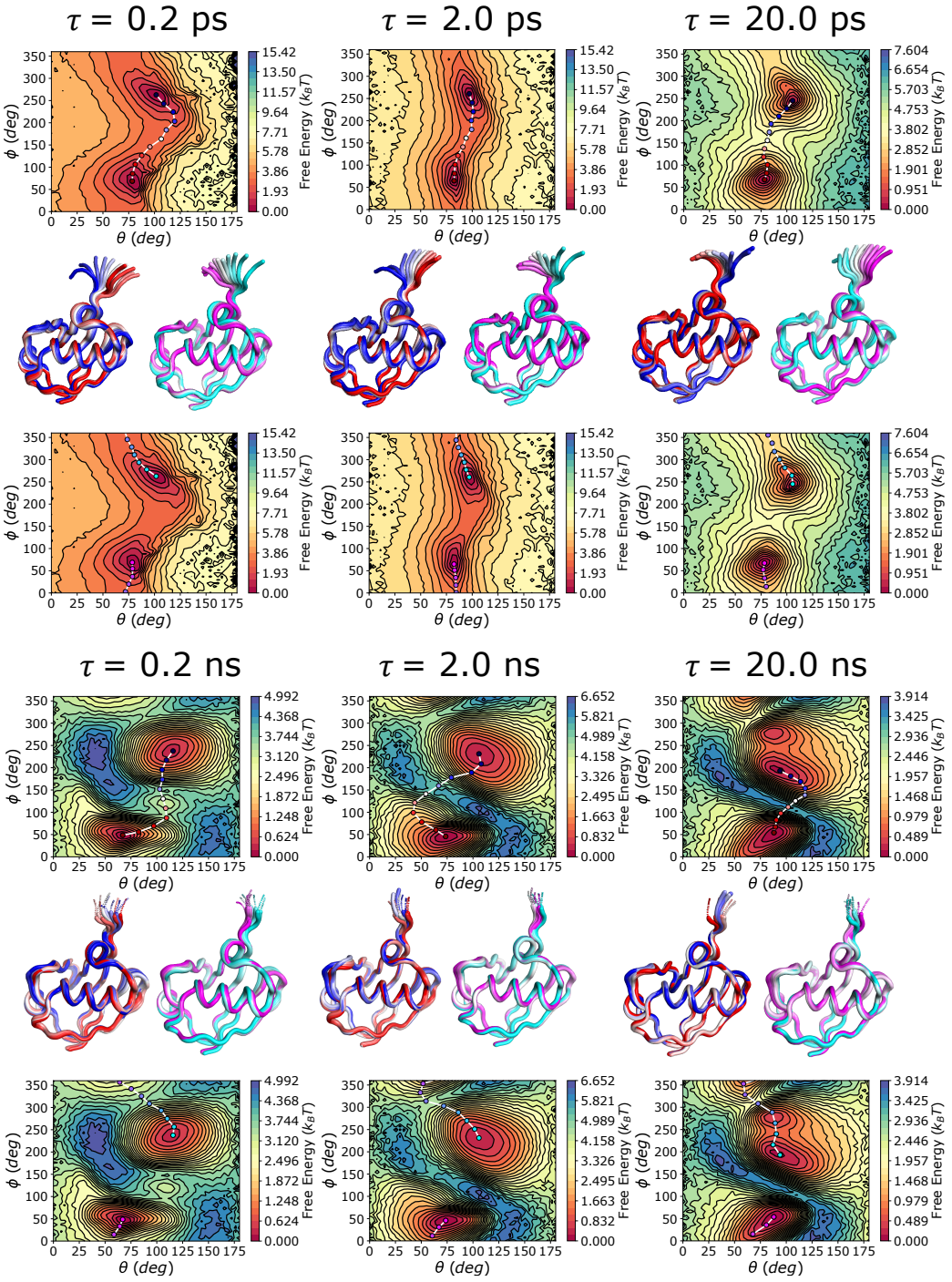

FIGURE 30. Effect of changing the tICA lag time on the first tICA mode free energy surface (FES) and the associated fluctuations. Note that each FES has two possible pathways to transition between the two energy minima, depicted in the panels above and below the protein fluctuations pictures. As one increases the lag time, the FES detects an increasing internal energy barrier. When the system crosses the barrier, it samples fluctuations in the tail and in the important loops. As the lag time increases, the predicted motion moves from the C-terminal tail and Lys11 loop into the 50 s loop. Concurrently, the barrier between the two minima on the surface rises until $\tau_{tICA} = 2.0$ ns, when the barrier starts to decrease. This decrease in the barrier height coincides with the loss of Markovian behavior at lag times above $\tau_{tICA} = 2.0$ ns, as seen in the plot of the implied timescales reported in the Supplemental Material

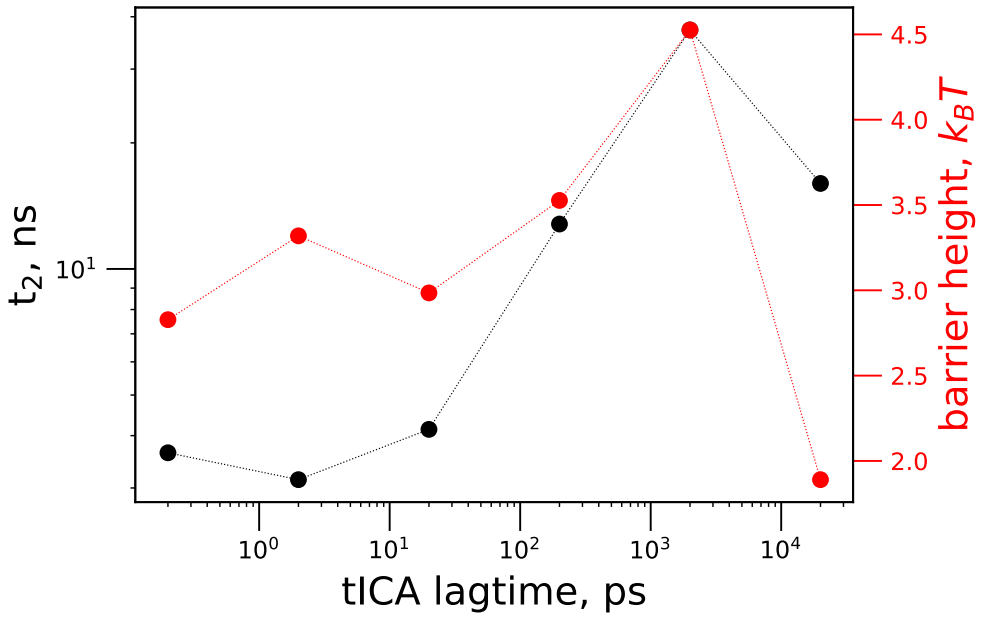

FIGURE 31. Correlation between the barrier surmounted by the red-white-blue pathway between minima in Figure 30 (red markers) and the t_2 timescale of the MSM constructed on the surface (black markers), as a function of tICA lag time. The correlation coefficient between the two sets of data, ρ, is 0.56. Dotted lines between markers are a guide to the eye.

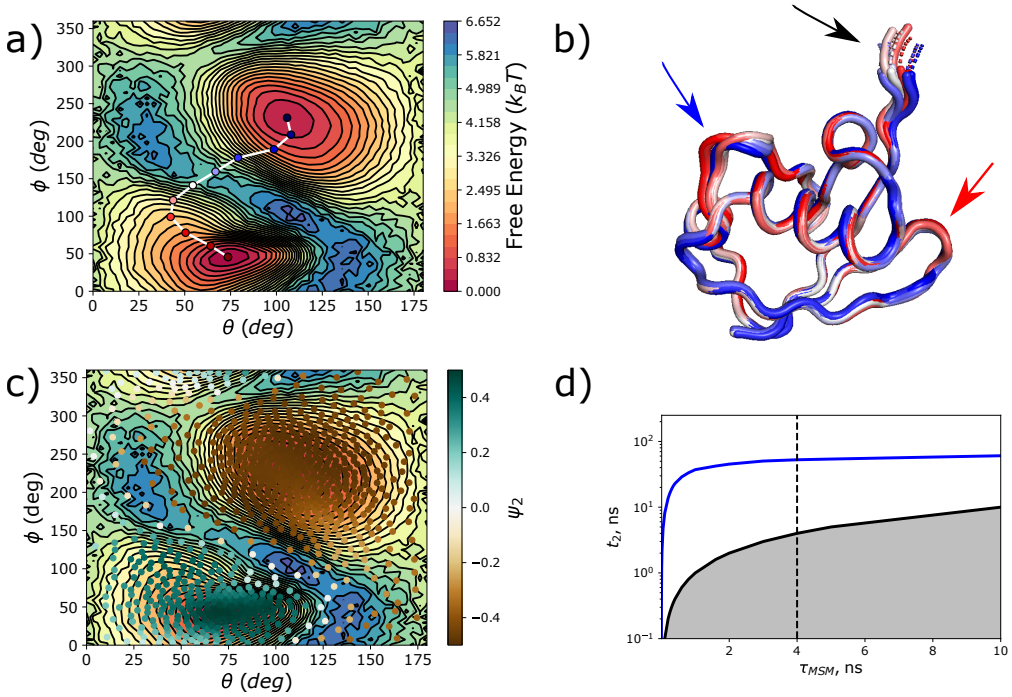

FIGURE 32. Results for the MSM in the two-dimensional (θ_a, ϕ_a) coordinate space for the slowest tIC. a): Free-energy surface along the (θ_a, ϕ_a) coordinates for the slowest tIC. b): Structures of ubiquitin from the trajectory along the free-energy surface given in a). The colors of the structures correspond to the given colored marker along the transition pathway. Arrows point to the three regions of ubiquitin showing the largest amplitude fluctuations: the C-terminal tail (black arrow), the 50 s loop (blue arrow), and the Lys11 loop (red arrow). c): projection of ψ_2 onto the discrete states of the MSM; colors correspond to the scaled-and-shifted value of ψ_2 at that discrete state, $\psi_2 = \frac{\psi_2 - \min(\psi_2)}{\max(\psi_2) - \min(\psi_2)} - 0.5$. d): implied timescales of the MSM as a function of MSM lag time. The black vertical line demarcates the lag time selected when constructing the MSM, 4 ns.

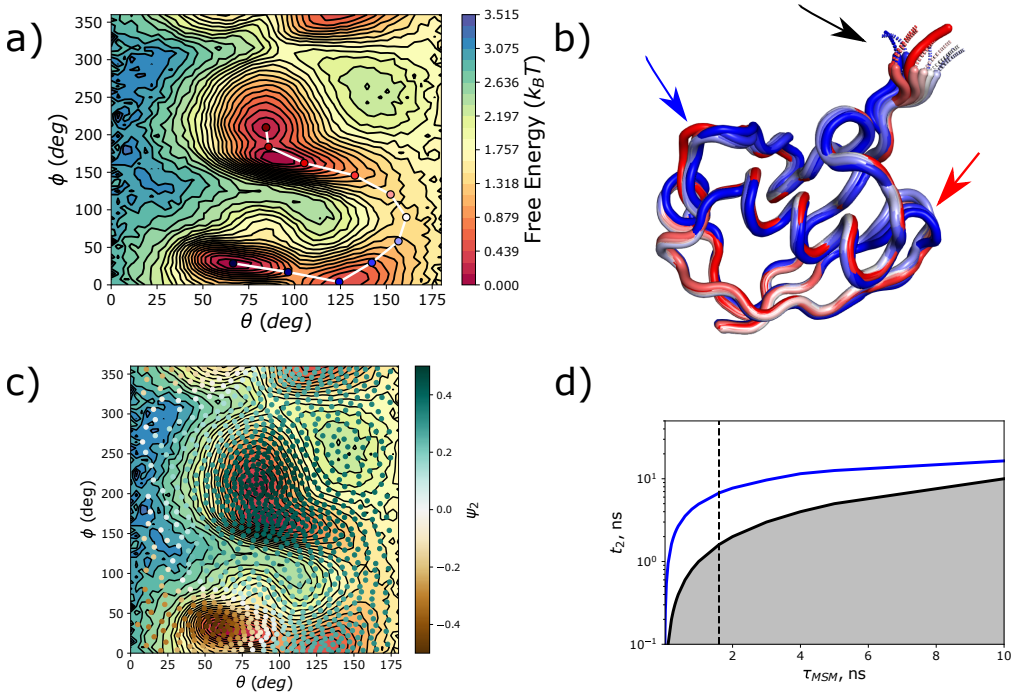

FIGURE 33. Results for the MSM in the two-dimensional (θ_a, ϕ_a) coordinate space for the second slowest tIC. a): Free-energy surface along the (θ_a, ϕ_a) coordinates for the slowest tIC. b): Structures of ubiquitin from the trajectory along the free-energy surface given in a). The colors of the structures correspond to the given colored marker along the transition pathway. Movement along the pathway in a) correspond to fluctuations mostly in the Lys11 loop (blue arrow) and C-terminal tail (black arrow) of ubiquitin, as well as smaller amplitude motions in the 50 s loop (blue arrow) c): projection of ψ_2 onto the discrete states of the MSM; colors correspond to the scaled-and-shifted value of ψ_2 at that discrete state, $\psi_2 = \frac{\psi_2 - \min(\psi_2)}{\max(\psi_2) - \min(\psi_2)} - 0.5$. d): implied timescales of the MSM as a function of MSM lag time. The black vertical line demarcates the lag time selected when constructing the MSM, 1.6 ns.

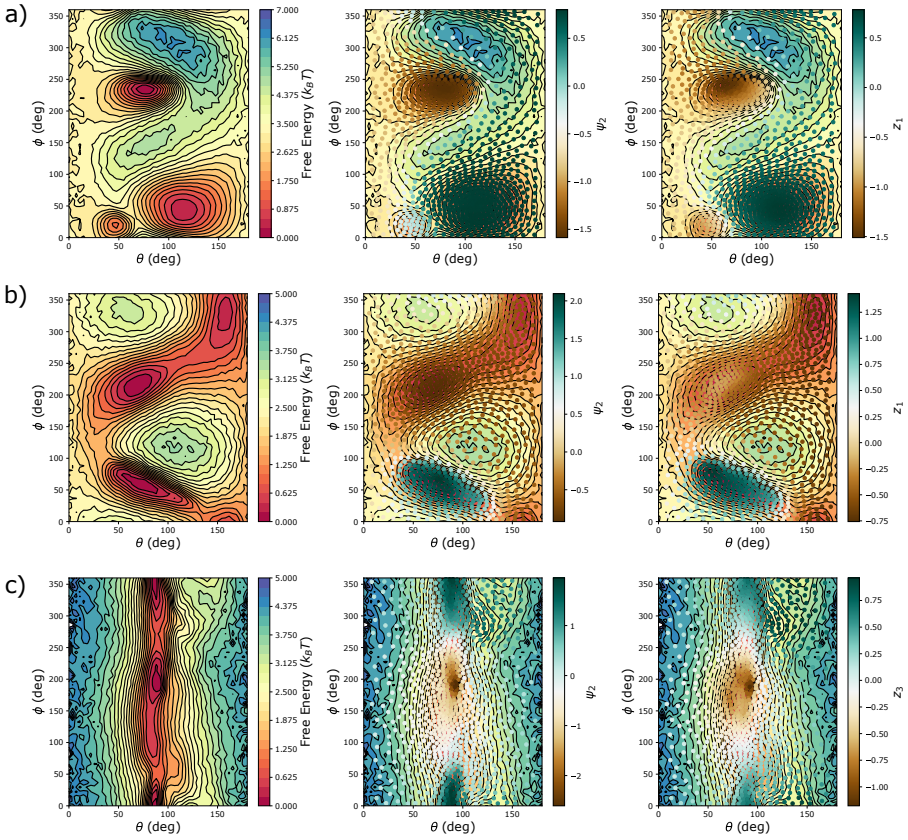

FIGURE 34. a) From left to right: Free-energy surface of isotropic LE4PD internal mode 6 with hydrodynamics from the one-microsecond ubiquitin simulation; projection of ψ_2 from the MSM of the trajectory on the (θ, ϕ) surface; and projection of the first tIC $z_1(t)$ onto the (θ, ϕ) surface. b) Same as a), but the displayed free-energy surface is for *anisotropic* LE4PD mode 7 without hydrodynamics. c) Same as a) and b), except for *anisotropic* LE4PD mode 5 without hydrodynamics, with the projection in the right-most panel being the third tIC $z_3(t)$ onto the surface.

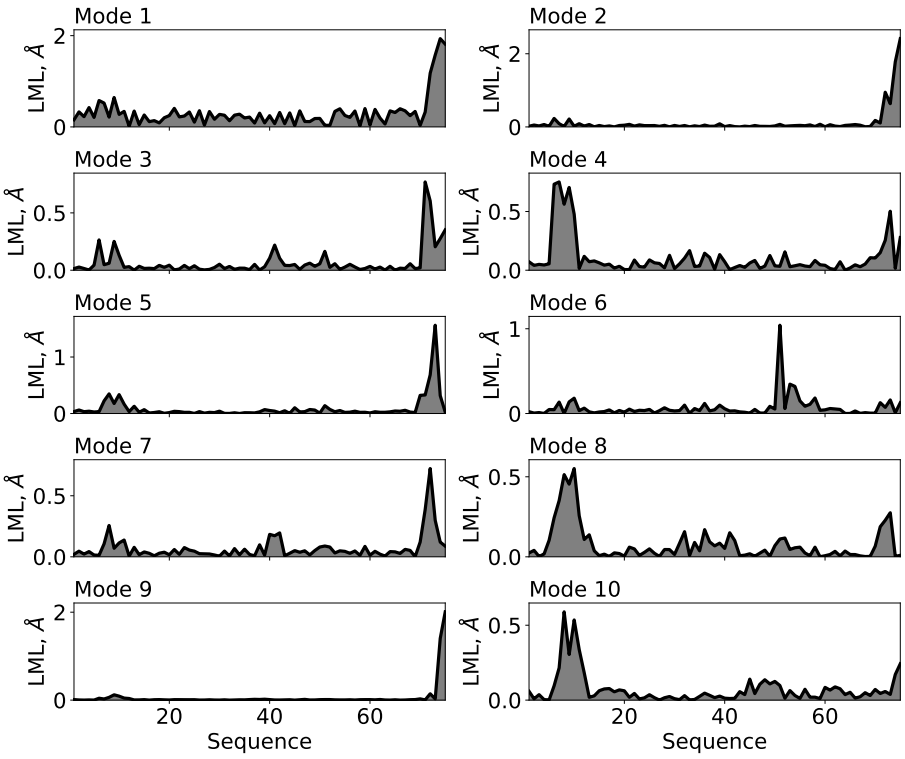

FIGURE 35. Mode-dependent fluctuations or local mode lengthscale (LML) for the ten slowest internal modes captured from the isotropic LE4PD analysis, with hydrodynamics, of the 1-μs simulation of ubiquitin.

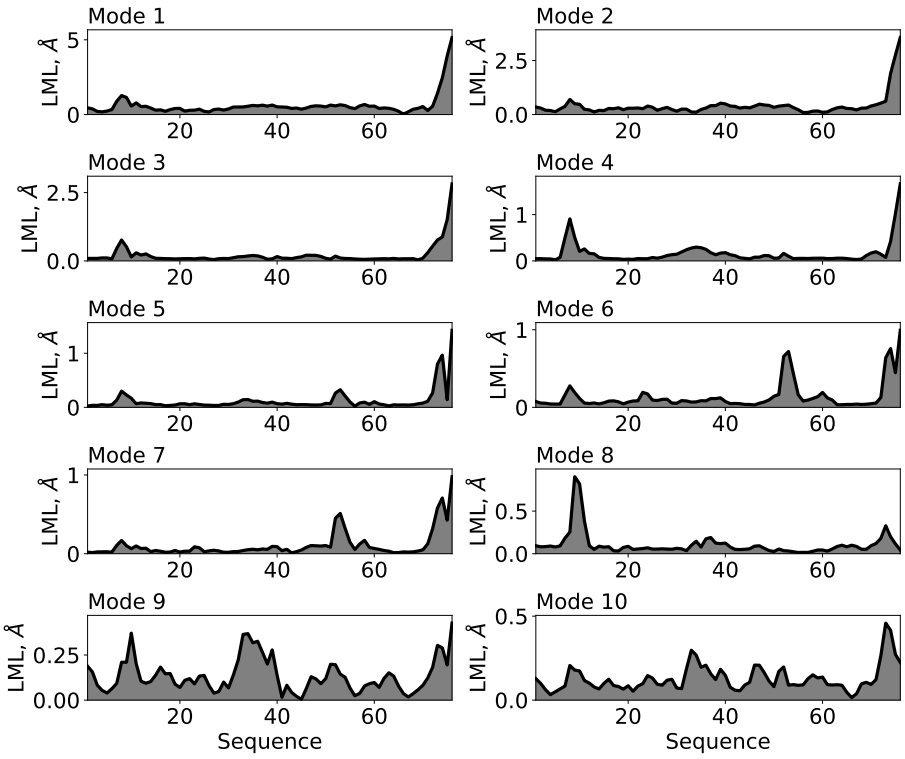

FIGURE 36. Mode-dependent fluctuations or local mode lengthscale (LML) for the ten slowest modes captured from the anisotropic LE4PD analysis, without hydrodynamics, of the 1-μs simulation of ubiquitin.

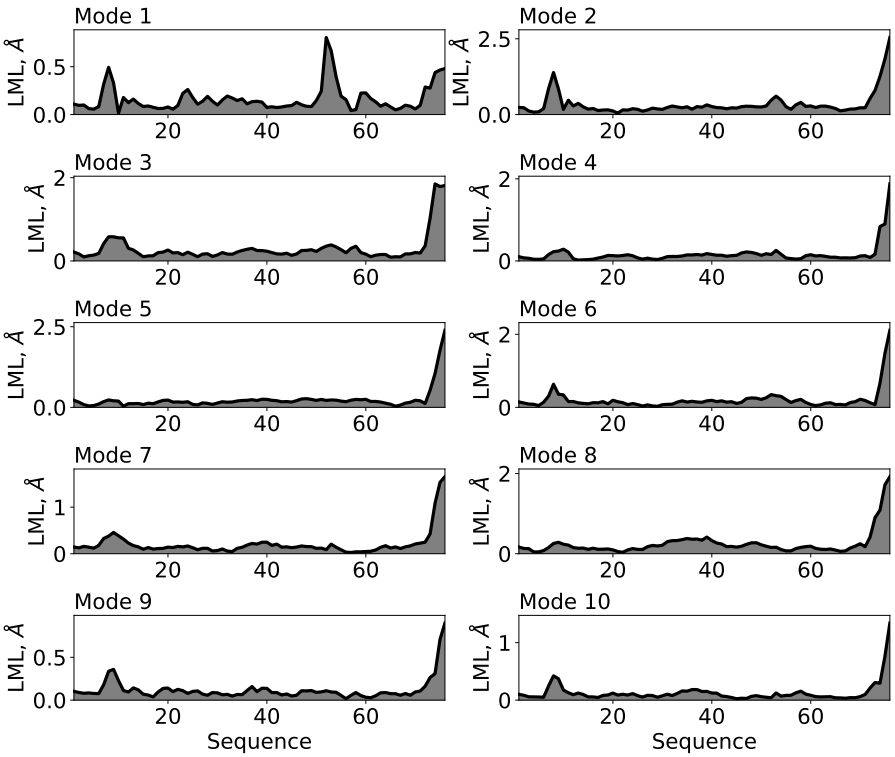

FIGURE 37. Mode-dependent fluctuations or local mode lengthscale (LML) for the ten slowest modes captured from the tICA of the 1-μs simulation of ubiquitin, with a tICA lag time of 2 ns.

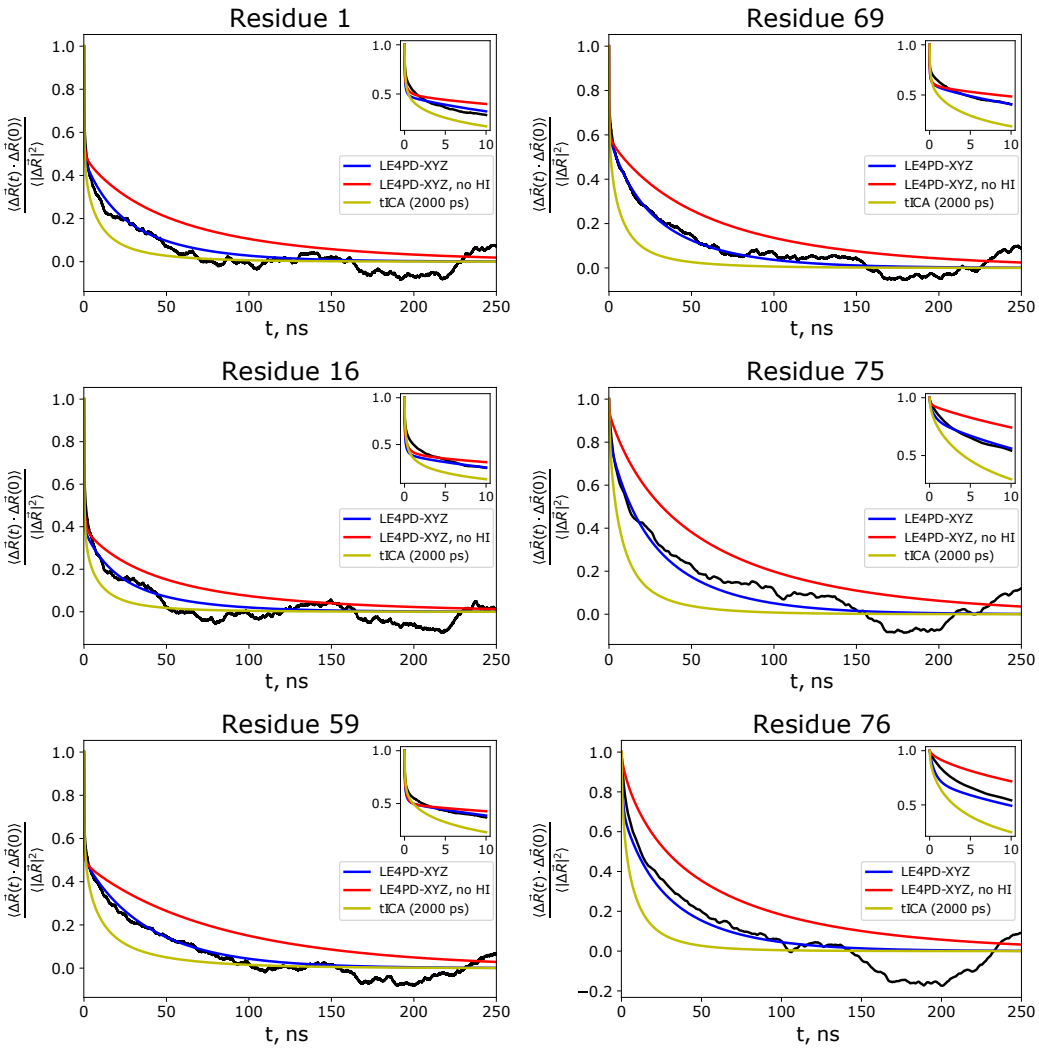

FIGURE 38. Comparison of the residue-residue time correlation functions (tcfs) for a sampling of residues along the primary sequence of ubiquitin. The black curves in each subplot show the tcf calculated from the simulation trajectory; the blue curves show the tcfs predicted from the LE4PD-XYZ theory with HI, the red curves the tcfs predicted from the LE4PD-XYZ without HI, and the yellow curves show the tcfs predicted from the tICA with a lag time of 2000 ps.

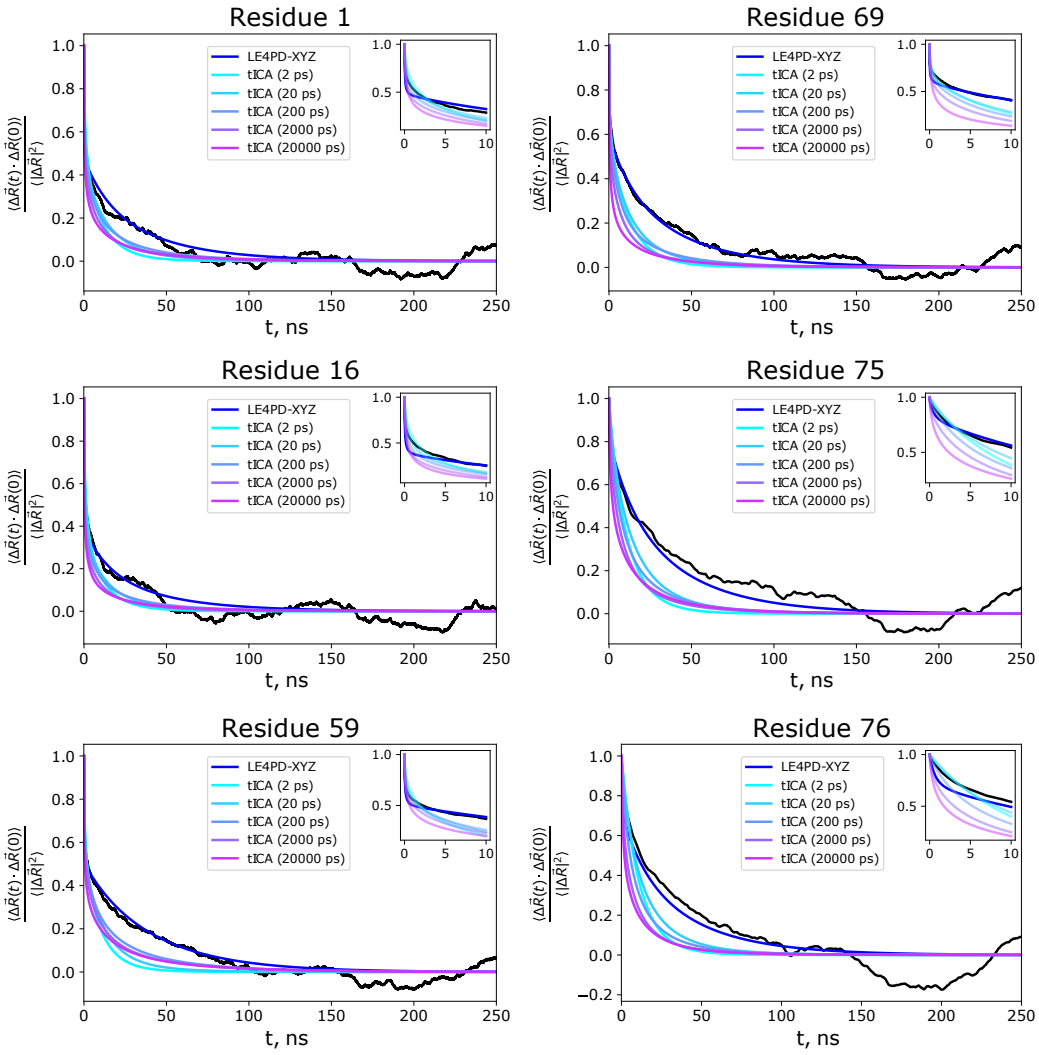

FIGURE 39. Comparison of the time correlation function (tcf) $C(t)$ calculated directly from the simulation trajectory (black) and calculated from the tICA using Eq. (5.29) for tICA lag times ranging from 20.0 to 20000.0 ps for six residues spaced along the primary sequence of ubiquitin. Using a tICA lag time of 2.0 ns gives the best agreement between the simulation and the theory for calculating the tcfs. Also reported are the time correlation functions calculated using the LE4PD-XYZ approach with hydrodynamics. The agreement between LE4PD-XYZ tcfs and the simulations is remarkable.

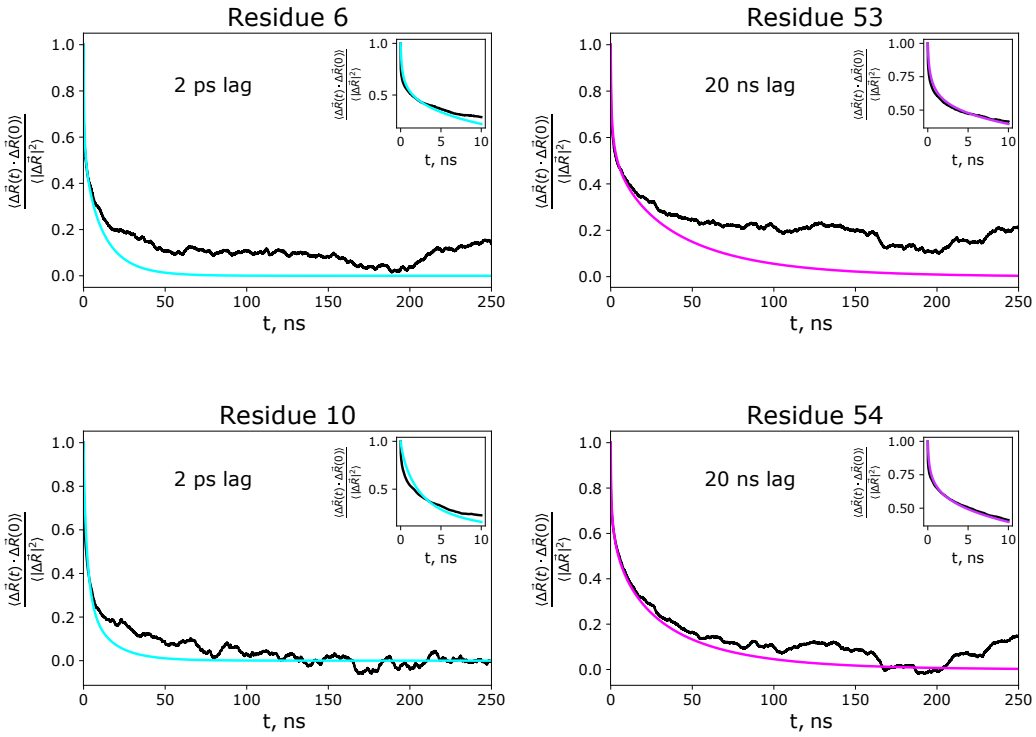

FIGURE 40. Left column: two residues in the Lys11 loop of ubiquitin whose tcfs from the simulation (black) are well approximated at timescales less than 10 ns by the tICs predicted using a lag time of 2 ps (cyan). Right column: two residues in the 50 s loop of ubiquitin whose tcfs from the simulation (black) are well approximated at timescales less than 10 ns by the tICs predicted using a lag time of 20000 ps (magenta).

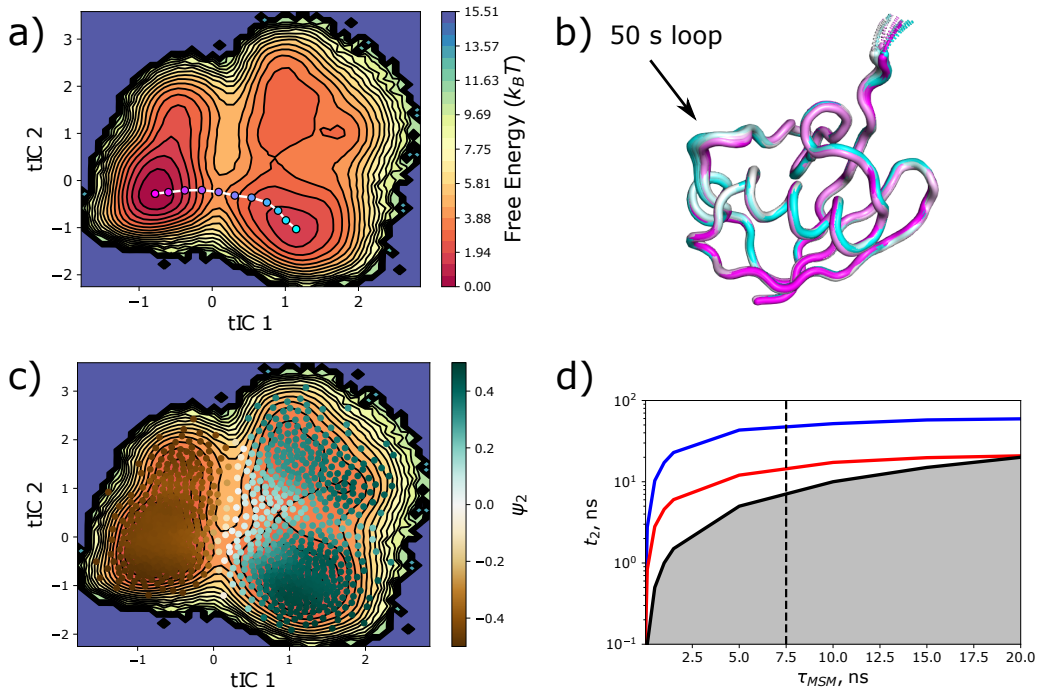

FIGURE 41. Results for the MSM of the two slowest tICs. a) Free-energy surface for the first two tICs. b) Structures of ubiquitin from the trajectory along the free-energy surface given in a). The colors of the structures correspond to the given colored marker along the transition pathway. c) projection of ψ_2 onto the discrete states of the MSM; colors correspond to the scaled-and-shifted value of ψ_2 at that discrete state, $\psi_2 = \frac{\psi_2 - \min(\psi_2)}{\max(\psi_2) - \min(\psi_2)} - 0.5$. d): the two slowest implied timescales, t_2 (blue curve) and t_3 (red curve), of the MSM as a function of MSM lag time, which is completely unrelated to the lag time used in the prior tICA step. The black vertical line demarcates the lag time selected when constructing the MSM, 7.5 ns.

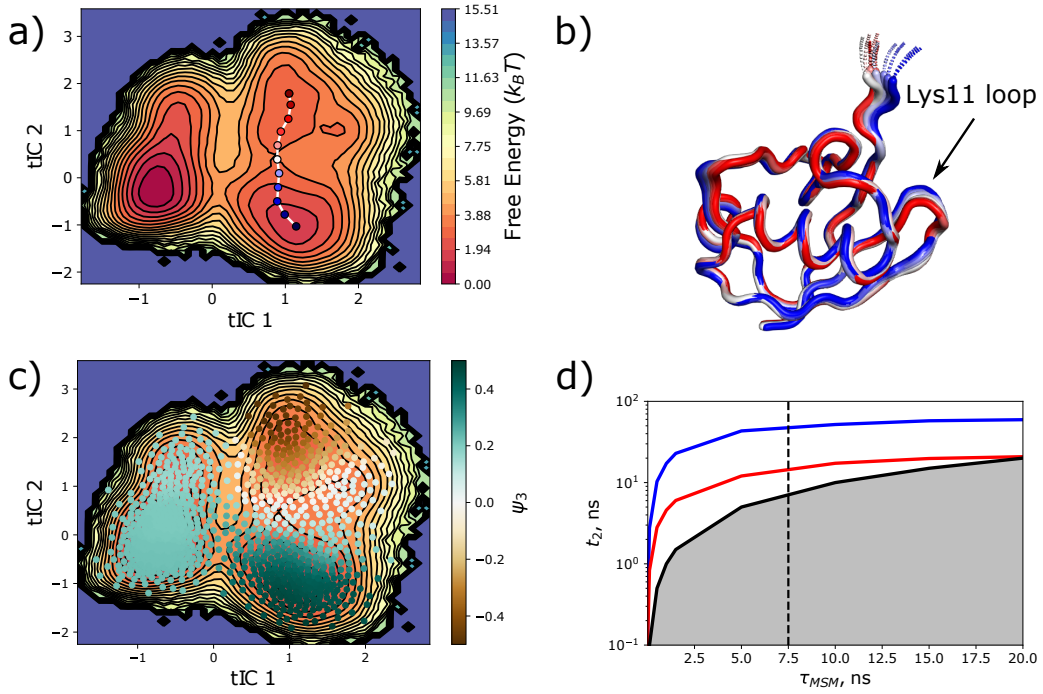

FIGURE 42. Same as Figure 41, except examining the second slowest process of the MSM, which is described by ψ_3 in c), where ψ_3 is scaled and shifted in the same manner as ψ_2 is in Figure 41.

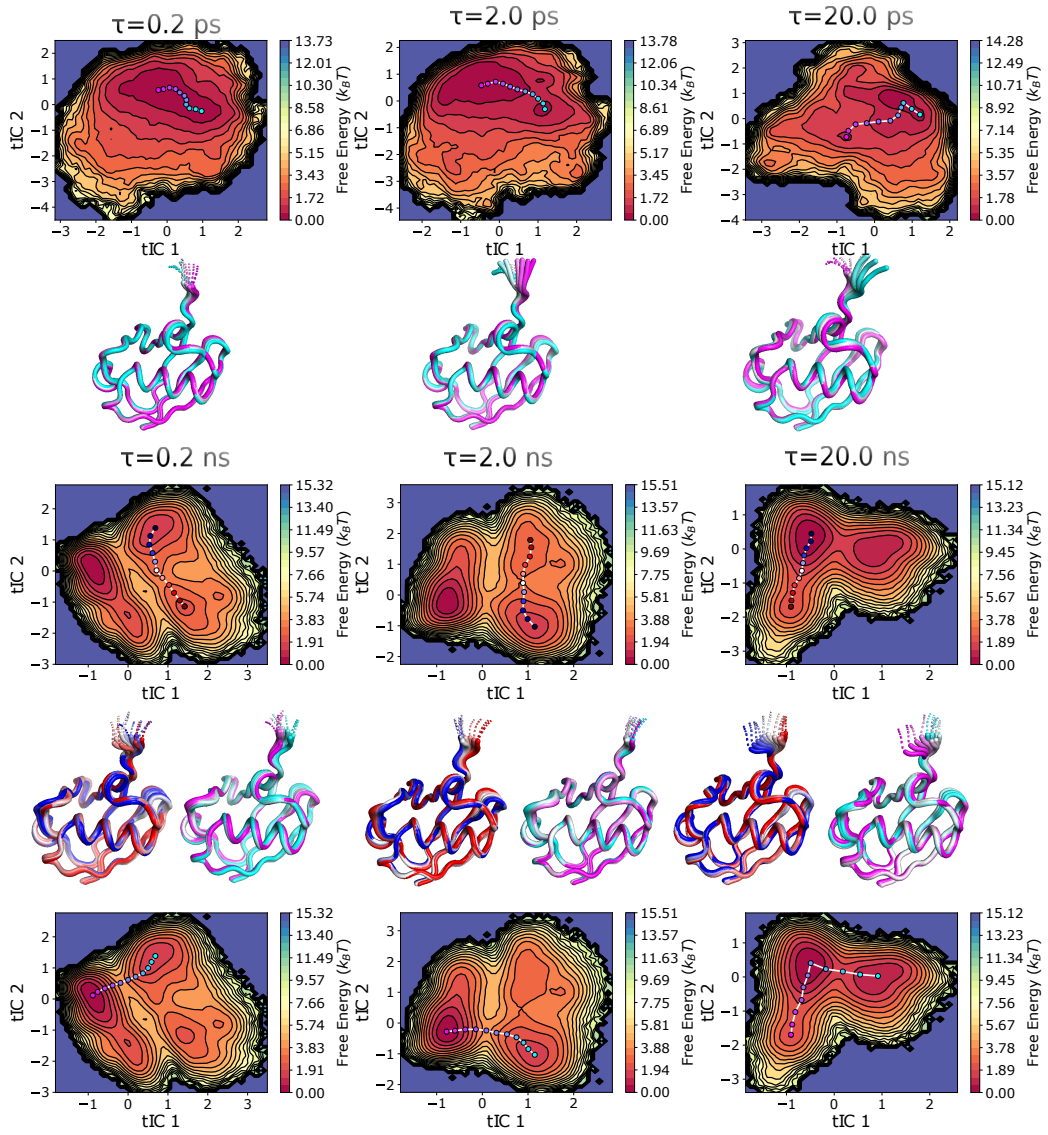

FIGURE 43. Effect of changing the tICA lag time on the resulting tIC 1 - tIC 2 FESs and associated dynamics. As the lag time is increased, the predicted motion of the slowest tIC moves from the C-terminal tail and Lys11 loop into the 50 s loop. Concurrently, the barrier between the two minima on the surface rises until $\tau = 2.0$ ns, when the barrier between minima starts to decrease. This decrease in the barrier between minima coincides with the loss of Markovian behavior at lag times above 2.0 ns seen in Figure S8 in the Supplementary Material of of [65]. Only a single pathway for $\tau \leq 20.0$ ps is drawn because there is no second minimum on the surface.

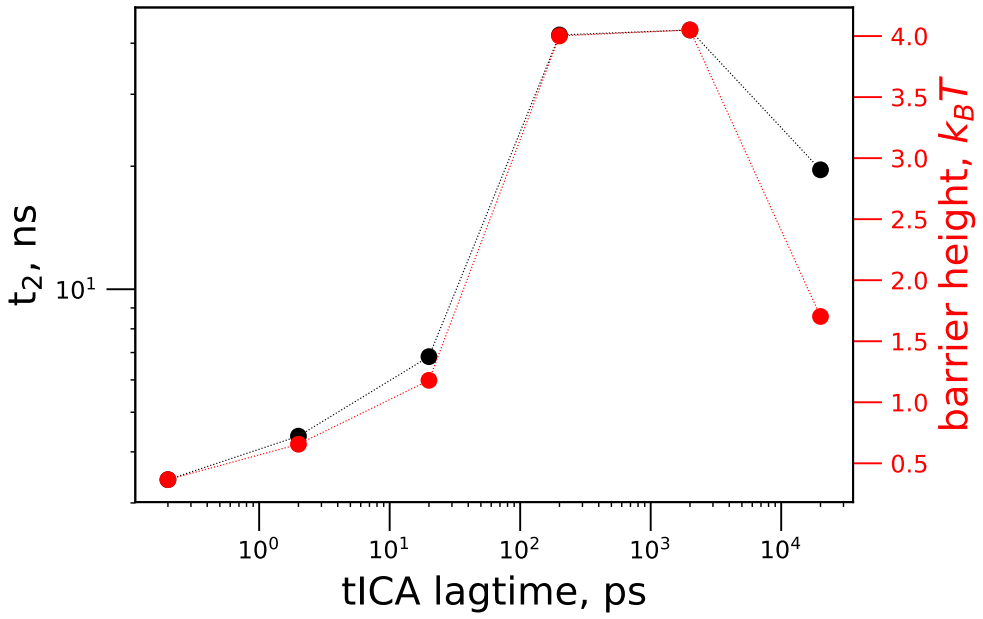

FIGURE 44. Correlation between the barrier surmounted by the red-white-blue pathway between minima in Figure 43 (red markers) and the t_2 timescale of the MSM constructed on the surface (black markers), as a function of tICA lag time. The correlation coefficient ρ is 0.99. Dotted lines between markers are a guide to the eye.

CHAPTER VI

DISCUSSION

Accurately identifying the slow, collective dynamics of proteins is important because they encode the functional dynamics of the folded state. [22, 23, 26] However, due to the intrinsic high dimensionality of biomolecular systems, efficient and effective methods for describing dynamics in a low-dimensional representation is necessary for understanding the functional behavior of biomolecules, but is challenging to perform in practice. Recently, techniques such as principal component analysis (PCA), time-lagged independent component analysis (tICA), and methods borrowed from the field of machine learning have become popular for discovering low-dimensional representations of biomolecular systems.[22] The main goal of these approaches is to find a set of *collective coordinates* that efficiently filter the most important motions of a biomolecular system into a few dimensions, where the data can be interpreted more easily. When coupled with a statistical ensemble from a molecular dynamics (MD) trajectory of an appropriate length, dimensionality reduction methods such as PCA or tICA can 'filter' the relevant dynamics from the fast motions largely irrelevant to the slow, functional dynamics in the underlying trajectory.

Another approach for discovering the slow, collective motions encoding the functional conformational dynamics of folded proteins is the Langevin equation for protein dynamics (LE4PD) [37, 38], which models the dynamics seen in the underlying MD simulation using a Langevin equation that projects the slow dynamics onto the residues of the protein, with the alpha-carbon chosen as the coarse-grained site. Unlike a PCA or tICA, the LE4PD, and its associated equation of motion, accounts for the chemical specificity of each residue, inter-residue hydrodynamic interactions,

and free-energy barriers along each of the collective coordinates or modes generated by solving the LE4PD equation of motion via diagonalization.

This approach breaks the coupled, intrinsically high-dimensional motions of a protein into a set of quasi-linearly independent normal modes, each with its own characteristic time- and lengthscale of motion. The set of slow LE4PD modes can be used to isolate and quantify the important collective motions of a protein. Thus, like the other dimensionality reduction methods mentioned previously, the LE4PD efficiently extracts and separates the 'essential' [26] motions of the protein into its slow modes. However, the LE4PD also has the advantage of an associated equation of motion, allowing for a direct determination of the decay times of each mode and a theoretical estimation of autocorrelation functions directly from the physical model provided by the theory. [38, 51] The LE4PD has previously been shown to give quantitative agreement between its predictions of B-factors [37] and T1, T2, and NOE values from NMR relaxation [38, 48, 57] and the experimental values of those quantities; also, its prediction of autocorrelation functions [38] and those calculated from the MD simulation have been shown to be in good agreement for a wide range of alpha-carbon bond vectors along the primary sequence of ubiquitin. [38]

The main theme presented in this dissertation is the extraction and analysis of slow coordinates from a long MD simulation of ubiquitin performed in its folded state using the LE4PD and LE4PD-XYZ methods. We postulate that the slowest LE4PD modes describe sampling of binding conformations, in-line with the conformational selection model of protein dynamics, [10, 11, 13] which states that free, un-bound proteins sample their binding conformations, even in the absence of ligand. When the LE4PD and LE4PD-XYZ modes are combined with Markov state models (MSMs), a detailed model of the kinetics and barrier-crossing dynamics of these slow LE4PD

modes is obtained, allowing for a precise description of the apo binding fluctuations of ubiquitin. Thus, combining MD simulations with the LE4PD and kinetic modeling tools such as MSMs, we can give significant insight regarding the functional motions of folded proteins such as ubiquitin.

In Chapter II, the slow LE4PD modes were coupled to the discrete master equation approach known as a Markov state model in order to precisely determine the timescales, amplitude, and localization of the dynamics of the regulatory protein ubiquitin. We found that the LE4PD-MSM approach was able to extract the slow dynamics in the binding regions of ubiquitin, in line with the conformational selection hypothesis that postulates an isolated protein will fluctuate about its folded structure to sample potential binding states.

In Chapter IV, the original, isotropic LE4PD was extended to account for the anistropic fluctuations of proteins, and the limit where this model maps onto the analogous PCA was described. We showed that this anisotropic model for protein dynamics, the LE4PD-XYZ model, is able to nearly quantitatively model the decay of the residue-residue autocorrelation function for a wide variety of residues across the primary sequence of ubiquitin because it takes into account hydrodynamic effects and free-energy barriers along each of the normal mode coordinates, which are both ignored by PCA.

In Chapter V, we performed a tICA on the same set of input coordinates (the configurational degrees of freedom of the alpha-carbons of each residue in ubiquitin) used in the LE4PD and LE4PD-XYZ analyses for the 1-microsecond simulation of ubiquitin studied extensively in this dissertation. We found that, while the tICA is superior at compressing the slow, high-amplitude fluctuations of the protein into a smaller number of modes, all three methods predict similar slow fluctuations in the

important binding regions of ubiquitin, even though the specific decomposition of the dynamics given by each method is different. Thus, the tICA, which has become a useful standard measure for the 'goodness' of a set of collective coordinates derived for biomolecular systems,[22] was used to verify that both the LE4PD and LE4PD-XYZ methods are able to select the directions of the 'leading fluctuations' or the slowest collective motions in the protein that are important for its function.

In addition to proteins, Chapter III used a reduced description for modeling the 'breathing' fluctuations of the simplest DNA system, the single-stranded deoxyadenine dinucleotide (dApdA). Using MSMs constructed on the two-dimensional, reduced space of dApdA, we were able to decompose the circular dichroism (CD) spectrum of dApdA into the contributions from the metastable states occupying the free-energy minima in the reduced space. The calculated CD was found to be in good agreement with the experimental spectrum in the low-wavelength region, without fitting by adjusting any input parameters, This result indicates that the coarse-graining plus MSM approach is adequate for modeling a spectroscopic observable of this simple nucleic acid system.

The studies presented here are necessarily limited in scope and are themselves capable of extension. First, the LE4PD-XYZ approach developed here has only been applied to unbound ubiquitin. By performing simulations and LE4PD-XYZ analysis of ubiquitin bound to one or more of its biological binding partners, such as another ubiquitin molecule or a ubiquitin ligase, we can observe directly how the postulated binding fluctuations are altered by binding to another protein or ligand, directly testing the conformational selection hypothesis. With the comparison of the LE4PD and LE4PD-XYZ coordinates to the tICA coordinates extracted for the same trajectory, we have postulated that the slow LE4PD and LE4PD-XYZ modes

select the leading dynamics of the protein; again, this hypothesis can be verified by running coarse-grained Langevin simulations along these coordinates or using enhanced sampling techniques, such as metadynamics,[24, 208] to extend the time- and lengthscales of the dynamics explored along these slow coordinates.

Finally, for the dApdA studies, we have postulated that this system is a model of the breathing dynamics found in more complex DNA systems. This hypothesis can be tested by gradually building up longer and longer single-stranded DNA constructs to see whether the observed breathing dynamics in dApdA are actually representative of larger DNA systems and how the addition of 'flanking' nucleotides on either end of the dApdA construct perturbs the observed conformational dynamics and thermodynamics.

APPENDIX A

MD SIMULATION OF UBIQUITIN

Simulation Details and Analysis

In chapters II, IV, and V, we study the fluctuation dynamics of ubiquitin. The first step is to perform MD simulations of the protein. We performed MD simulations using GROMACS version 5.0.4,[137] and the AMBER99SB-ILDN atomistic force field,[179] on the Comet supercomputer at the San Diego Supercomputing Center. The starting structure was taken from the Protein Databank, PDB ID: 1UBQ,[53]. We solvated the protein with spc/e water, and minimized the energy using a steepest descent algorithm. We added Na^+ and Cl^- ions until the ion concentration was 45 mM, with the concentration of ions selected to match that used in nuclear magnetic resonance experiments of ubiquitin.[54] We subjected the protein-solvent system to two rounds of equilibration: first, a 50-ps equilibration in the NVT ensemble at 300 K, with the temperature controlled using a Nosé-Hoover thermostat; 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 again used to maintain the temperature. We used the last frame of this burnout run as the initial configuration for the 1 μs production run, which utilized 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, given in the Supplementary Material, the entire trajectory was deemed to fluctuate around an equilibrium value,

and the entire 1-μs of trajectory was used for the subsequent LE4PD and MSM analysis. We used the LINCS algorithm[141] to constrain all hydrogen-to-heavy-atom bonds in the system, and adopted an integration timestep of 2 fs during both the equilibration and MD simulation. We saved the trajectory to file every 100 integration steps (every 0.2 ps), obtaining a total of $\frac{10^6 \text{ ps}}{0.2 \text{ ps/frame}} = 5\text{x}10^6$ frames for analysis.

Before performing the LE4PD and MSM analysis, we processed the 'raw' MD trajectory to remove jumps across periodic boundaries in the simulation box, and we removed global translational and rotational motions of the protein using the least-squares fitting procedure available in GROMACS.

Figure A.1 reports the root-mean-squared deviation (RMSD) of the alpha-carbons of ubiquitin, from the first frame of the trajectory, over the course of the MD trajectory. While there are many local extrema in the instantaneous trace, the running average shows that the RMSD is relatively constant throughout the simulation, allowing us to use the statistics from the entire simulation to build the structural matrices in the LE4PD and the transition matrix in the MSM for the slowest LE4PD internal modes.

Figure A.2 shows the convergence of the first LE4PD internal mode, calculated individually for time windows inside the 1-μs equilibrium MD simulation trajectory of ubiquitin. Here, we use the shape of the free energy surface (FES) of the first LE4PD internal mode to define the simulation convergence: the FES remains qualitatively the same as the length of the simulation is extended. This implies that increasing the length of the simulation increases the sampling of the configurational landscape without exploring new free energy regions. We confirmed the convergence of the first internal mode, and higher internal modes, by calculating the inner product of the

left ($\mathbf{Q}^{-1}$) eigenvector for ten variable length time slices of the simulation with the right eigenvector ($\mathbf{Q}$) from the entire 1-μs simulation. The overlap matrix calculated by taking these inner products is shown for three time slices and compared with the 1-μs self-overlap matrix in Figure A.3. It can be seen that the overlap between the left and right eigenvectors becomes becomes more strongly diagonal as the amount of sampling is increased, but by 700 ns the overlap with the eigenvectors calculated using the full simulation trajectory is nearly the identity matrix, indicating that from 700 ns on the dynamics appears well-converged.

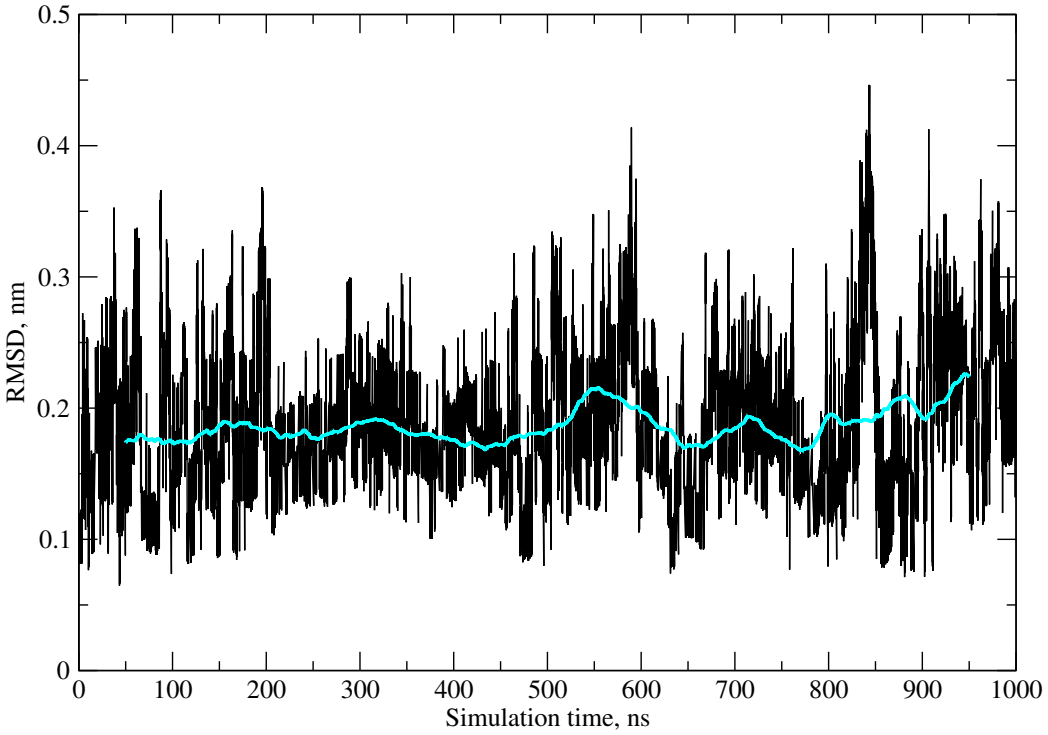

FIGURE A.1. Root-mean-squared deviation (RMSD) of the alpha-carbons of ubiquitin from the first frame of the MD trajectory over the course of the 1 μs MD trajectory analyzed in this study. The black trace gives the instantaneous RMSD at each frame of the simulation while the cyan trace gives the running average, calculated using 500000 frames at a time.

A final check of convergence of the LE4PD modes is calculating the predicted 'bare' timescales of each mode, $\tau_a^0 = (\sigma\lambda_a)^{-1}$, which are the LE4PD times without

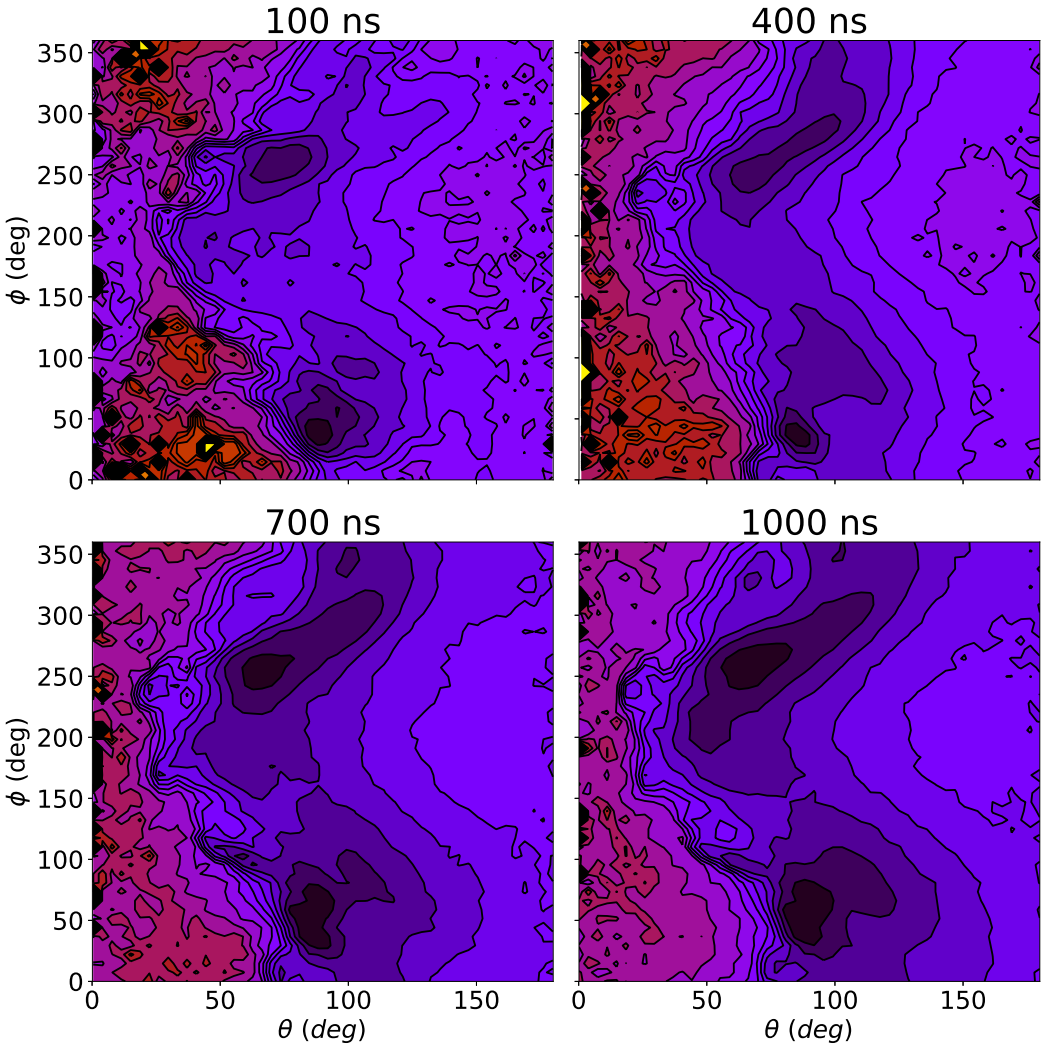

FIGURE A.2. Convergence of the free-energy landscape for the first LE4PD internal mode for a 1-μs equilibrium MD simulation of ubiquitin.

including mode-dependent free energy barriers, as the amount of simulation time used in the calculation is increased. Since $\tau_a^0 \propto \lambda_a^{-1}$ and $\sigma \approx$ const., the top panel of Figure A.4 shows the convergence of the eigenvalues of **LU**, much as Figure A.3 shows the convergence of the eigenvectors, as the simulation time is extended. The bottom panel of Figure A.4 elucidates how the scaled timescales of each LE4PD mode, $\tau_a = \tau_a^0 \exp\left[E_a^\dagger / k_B T\right]$ change as the amount of simulation time used to construct the **LU** matrix is increased. Excepting mode 9 and the 100 ns slice of simulation time,

even the τ_a across simulation slices is approximately the same, again indicating good convergence of the simulation and the LE4PD modes predicted from the statistics collected from these simulation slices.

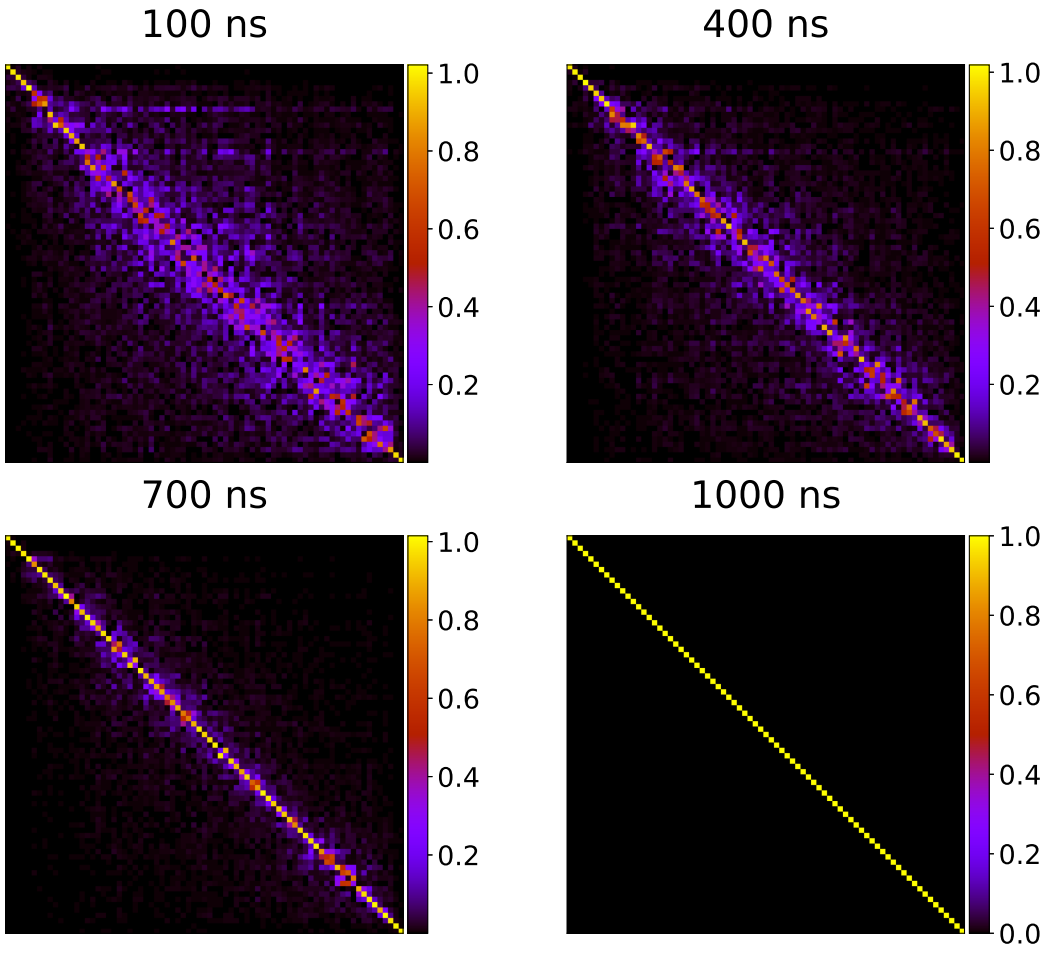

FIGURE A.3. Convergence of the eigenvectors of the **LU** matrix as the simulation time is extended. The overlap is calculated between the left eigenvectors of the **LU** matrix using the statistics harvested up to the time slice given above the plot with the right eigenvectors of the **LU** matrix calculated using the full statistics of the 1 μs simulation.

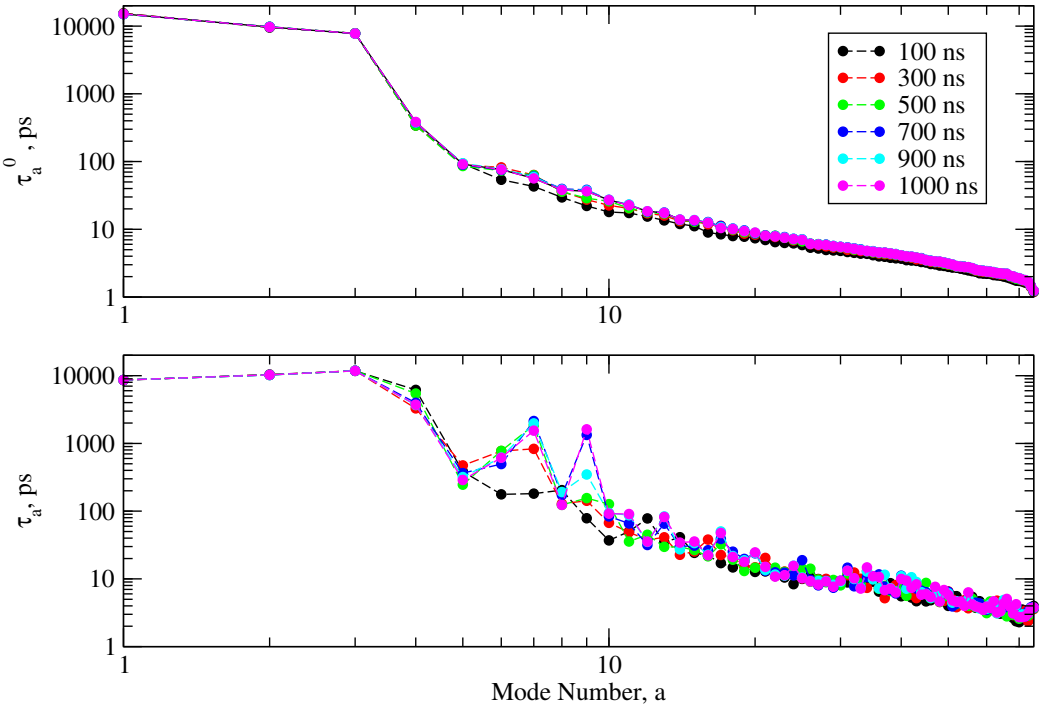

FIGURE A.4. Convergence of the timescales of the LE4PD modes as the amount of simulation time used in the analysis is extended. Top: bare timescales predicted directly from the LE4PD as $\tau_a^0 = (\sigma\lambda_a)^{-1}$ using 100 (black), 300 (red), 500 (green), 700 (blue), 900 (cyan), or 1000 (magenta) ns of simulation time. Bottom: same as the top subplot, except the timescales given for each LE4PD mode have been rescaled by the characteristic free-energy barrier, $E_a^\dagger$, predicted by the median absolute deviation (MAD) of the mode's free-energy surface: $\tau_a = (\sigma\lambda_a)^{-1} \exp\left[E_a^\dagger/k_BT\right] = \tau_a^0 \exp\left[E_a^\dagger/k_BT\right]$.

APPENDIX B

MARKOV STATE MODELS FOR THE LE4PD MODES

Markov State Model Details

Succinctly, a Markov state model or MSM is a discrete master equation approach [99] to modelling the flow of probability density in a defined state space. That is, it is a numerical approach to discretize and model the Fokker-Planck equation for the transfer of probability density on a surface. [191] In this appendix the basics of Markov state modelling and its application to the two-dimensional surfaces of the slow LE4PD modes will be described briefly; *mutatis mutandis*, the same formalism is applied to model the transition of probability density on the surfaces of the LE4PD-XYZ modes, as described in chapter IV. For more detailed views regarding the theory and applications of MSMs, the reader is advised to consult the literature; several good resources, with applications to biomolecular systems, are [30, 33, 34, 49, 50, 85, 96, 100, 144, 216], among others.

When building a MSM, first a set of input coordinates or *features* are selected; this steps amounts to a coarse-graining of the available, high-dimensional state space sampled by the system. Generally a few (less than 10, roughly) 'important' coordinates are kept from the original trajectory and used as the state space of interest; in this dissertation, all MSMs are constructed on two-dimensional state spaces of the LE4PD (θ, ϕ) (chapter II), (R, ϕ) (chapter III), LE4PD-XYZ (θ, ϕ) (chapter IV), tICA (θ, ϕ), or the two slowest tICA modes (chapter V). Next, using the trajectory in this reduced state space, $x(t)$, is discretized[144] into a set of *microstates*, which are small volumes of state space among which transition are approximately

Markovian, meaning that the conditional probability of transitioning from microstate i to microstate j over a lagtime τ depends only on the indices i and j and not where the trajectory was previously, which results in the conditional transition probability, $p\left(x(t+\tau) \in j | x(0), x(\tau), x(2\tau), \ldots, x(t) \in i\right)$ having only a 'one-step memory':

$$p\left(x(t+\tau) \in j | x(0), x(\tau), x(2\tau), \ldots, x(t) \in i\right) = p\left(x(t+\tau) \in j | x(t) \in i\right) := T_{ij}(\tau);$$

that is, the conditional probability of transitioning between states i and j over a lagtime τ, which is stored in the matrix element $T_{ij}(\tau)$ of the *transition matrix* $\mathbf{T}(\tau)$, depends only on the state occupied at time t and the time between observations of the trajectory, which is given by the lagtime τ.

The elements of $\mathbf{T}(\tau)$ are estimated empirically using the statistics from the input trajectory; some details on this process are given in [94]. Once $\mathbf{T}(\tau)$ is calculated, it is diagonalized to obtain a set of eigenmodes describing the collective kinetic processes occurring over the state space. The spectral radius of $\mathbf{T}(\tau)$ is bounded from above by 1, which is a consequence of the Perron-Frobenius theorem for stochastic matrices [95], The first (left) eigenvector of $\mathbf{T}(\tau)$, ϕ_1, gives the stationary distribution of the system over the state space, and the eigenvalues of $\mathbf{T}(\tau)$, $\lambda^{\mathrm{MSM}}(\tau)$, are related to the timescales of the eigenmodes as described in chapter II. Thus, the second (right) eigenvector, ψ_2, describes the slowest process on the state space, and the timescale of this process is given by a combination of the lagtime of the MSM and the second eigenvalue, λ_2^{MSM}, of $\mathbf{T}(\tau)$, as given in chapter II. For the LE4PD, LE4PD-XYZ, and tICA modes, this slowest process is the one of interest, since it should describe the kinetics of transitioning between wells on the mode-dependent free-energy surface of these models.

The $(\theta_a(t), \phi_a(t))$ coordinates are used to generate a two-dimensional MSM for LE4PD mode a. To make a MSM of the trajectory, the state space must split into a finite number of discrete states; then the probability of transitions between the states is calculated. We split the trajectory of the $(\theta_a(t), \phi_a(t))$ coordinates into W discrete states using the k-means++ clustering algorithm [142], as implemented in PyEMMA [85]. Given the discrete trajectory, we find a lag time τ using the criterion defined in chapter II and construct a MSM by determining the transition matrix of the system, $\mathbf{T}(\tau)$, which models the evolution of the probability vector, $\mathbf{p}(t)$, of occupying the discrete states as a Markov chain [33]. $\mathbf{T}(\tau)$ causes the system's probability to evolve as follows:

$$\begin{aligned} \mathbf{p}^T(t+\tau) &= \mathbf{p}^T(t)\mathbf{T}(\tau) \\ p_j(t+\tau) &= \sum_{i=1}^{W} T_{ij}(\tau) p_i(t), \end{aligned} \tag{B.1}$$

i.e., the probability of occupying state j at time $t+\tau$ is completely determined by summing the probability of transition from all other states i to state j at time t, given by T_{ij}, weighted by their probability of occupation. There are several ways to construct $T(\tau)$ from the simulation trajectory; here, we use the following reversible maximum-likelihood estimate of the transition matrix,[94] as described in the chapter II:

$$T_{ij}(\tau) = \frac{(c_{ij} + c_{ji})\pi_j}{c_i \pi_j + c_j \pi_i}, \tag{B.2}$$

with $c_{ij} = c_{ij}(\tau)$ the ij^{th} element of the count matrix, which counts all the transitions from states i to j in the trajectory at a lag time τ; $c_i = \sum_j c_{ij}$ the i^{th} row sum of the count matrix, giving the total number of transitions from i; and π_i the stationary (equilibrium) probability of state i.

For each LE4PD mode analyzed with a MSM, we partitioned the free-energy surface into $W = 1000$ discrete states (microstates). We based this choice on a saturation of the t_2 as the discretization was made finer; we found that after using $\sim$500 microstates, the t_2 predicted as a function of lag time became approximately constant, independent of the number of microstates in the model (data not shown). Since the coordinates used in the construction of the free-energy surface, (θ, ϕ), represent the polar and azimuthal angle, respectively, of the given LE4PD mode in real space, we assigned each frame in the trajectory to a microstate based on that frame's great-circle distance from the nearest microstate center.[235]

Figure B.1 shows the implied timescale of the most slowly decaying MSM mode, t_2, for the MSMs of the first five LE4PD internal modes as a function of lag time for the MSM. Figure B.2 shows the analogous plot for LE4PD internal modes six through ten. The vertical, dashed lines show the lag time selected for the MSM of the corresponding LE4PD mode, with the color key given in each figure's legend. These lag times are the ones used for the MSMs described and analyzed in the main text. We selected those lag times based on the spectrum of ψ_2, the second right eigenvector of the transition matrix of the MSM, for each LE4PD mode (see the following section for further details).

Effect of Changing Lag Time on the Spectrum of ψ_2

As stated in the main text, the lag time of the MSM for each of the slow LE4PD modes is selected based on the spectrum of ψ_2; that is, the lag time is selected such that ψ_2 is a minimum in one well of the FES and a maximum in another well of the FES and approximately null-valued at saddle points or transition states on the FES. If the lag time is made too long, the discrete states with the maximum and minimum

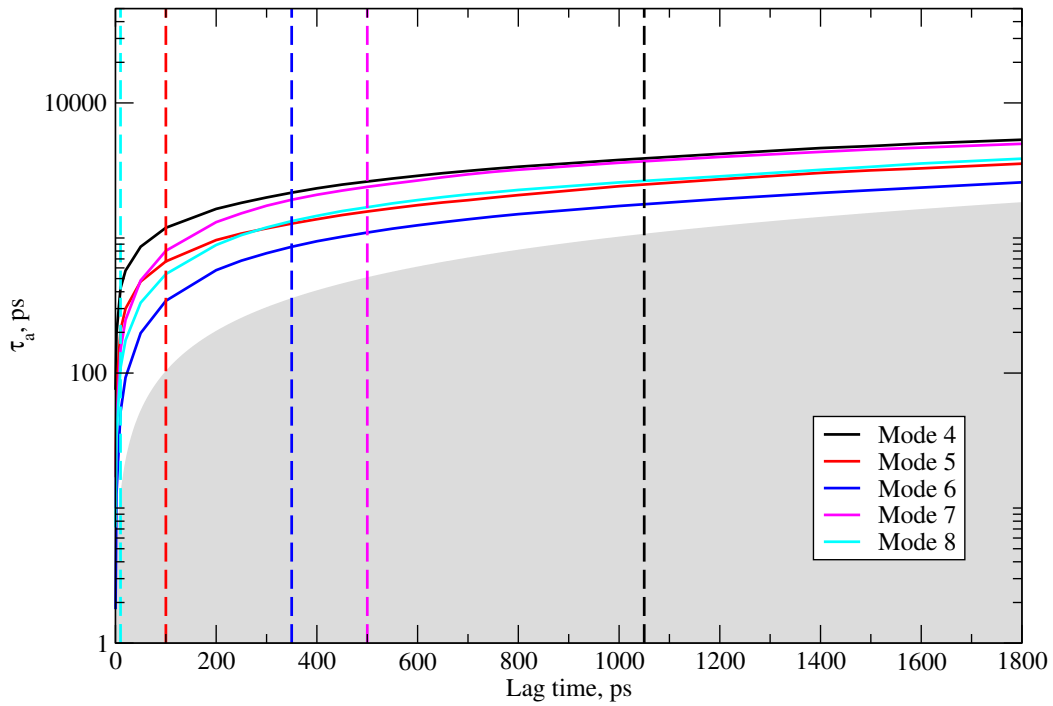

FIGURE B.1. Implied timescales versus lag time plot for the first five LE4PD internal modes. Vertical, dashed lines demarcate the lag time selected to parameterize the MSM for the corresponding LE4PD mode; e.g. the black, dashed, vertical line indicates the lag time at which the MSM for the fourth LE4PD mode was parameterized.

projection along ψ_2 will 'drift' from minima into higher free-energy regions on the FES as the lag time is increased. Figure B.3 shows an example of this phenomenon. The colored stars indicate the ten discrete states with the maximum (yellow) and minimum (cyan) projections along ψ_2 at a certain lag time, which is reported above the respective plot. The plot on the top right of Figure B.3 refers to the lag time for this mode reported in the chapter II. We observe that as the lag time is made longer, the states with the minimum and maximum projection along ψ_2 begin to drift out of their respective free-energy wells, meaning that the slowest process described by the MSM is no longer a transition between minima, but rather a transition from a well to a low-populated, high free-energy region or a transition between two high free-energy

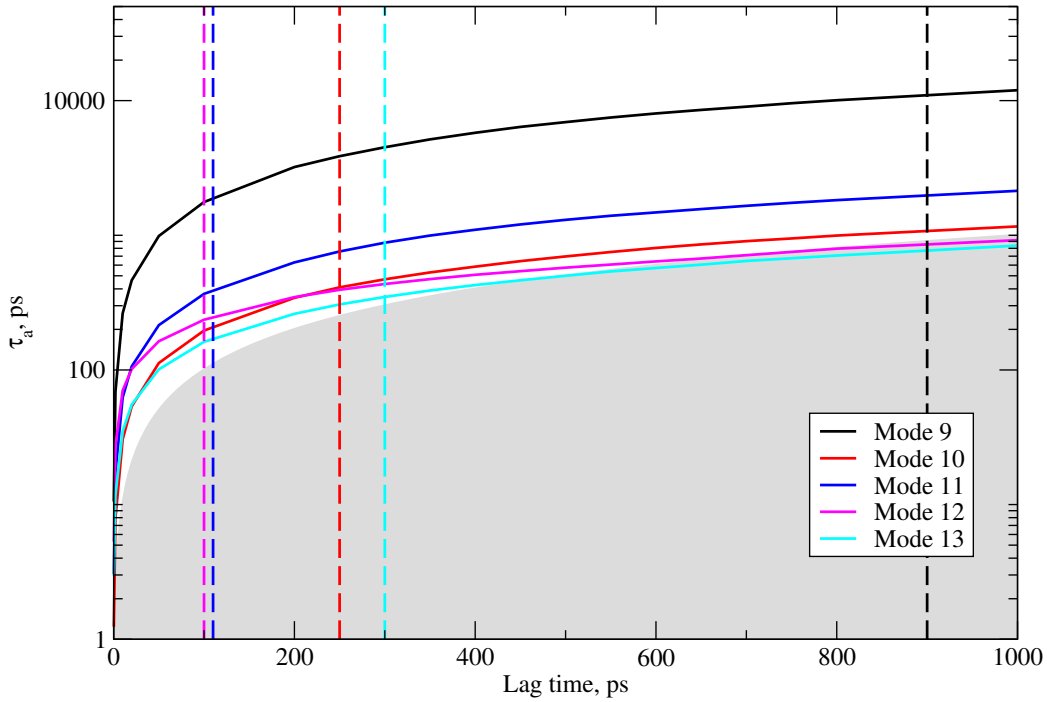

FIGURE B.2. Implied timescales versus lag time plot for the LE4PD internal modes six through ten. Vertical, dashed lines demarcate the lag time selected to parameterize the MSM for the corresponding LE4PD mode; e.g. the black, dashed, vertical line indicates the lag time at which the MSM for the ninth LE4PD mode was parameterized.

regions. The sign of ψ_2 also flips between 10 ps and 25 ps lag time, but this effect is irrelevant to the calculation of kinetic properties and the interpretation of the MSM.

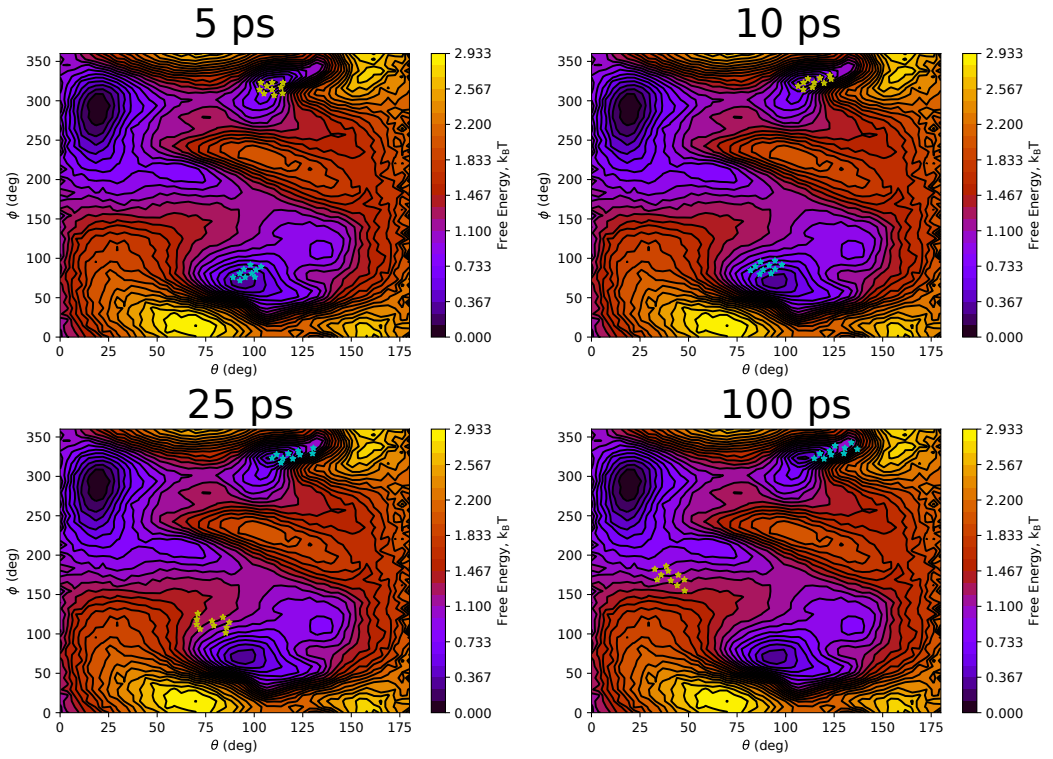

FIGURE B.3. FES of LE4PD mode 8 with the ten discrete states with the maximum (yellow) and minimum (cyan) projections along ψ_2 at the lag time specified above the plot.

APPENDIX C

CD CALCULATION DETAILS

Calculations of Circular Dichroism (CD) Spectra from Molecular Confirgurations

We applied the standard methods developed by Schellman and others to model the delocalized electronic states of the dApdA dinucleotide as a function of base stacking conformations.[110, 119, 120] In the formalism that follows, we consider only the contribution to the CD spectrum that emerges from the exciton interactions between the component adenine bases of the dApdA dinucleotide, and we neglect the minor contribution to the CD from the non-interacting adenine monomer, which provides a relatively weak signal for the fully unstacked conformation. When light of frequency ν interacts with a solution of optically active molecular chromophores, the left and right circularly polarized components are absorbed to different extents. The frequency (or wavelength) dependence of the differential extinction between left and right circular polarizations, $\Delta\epsilon(\nu) = \epsilon_L(\nu) - \epsilon_R(\nu)$, is called the CD spectrum. The CD spectrum can be understood in terms of the rotational strength R_{if} of an electronic transition from an initial state $|\Psi_i\rangle$ to a final state $|\Psi_f\rangle$, which is defined by the Rosenfeld equation

$$R_{if} = \mathrm{Im}\left[\langle\Psi_i|\widehat{\mu}|\rangle \cdot \langle\Psi_f|\widehat{\mathbf{m}}|\Psi_i\rangle\right] \tag{C.1}$$

Here $\widehat{\mathbf{u}}$ and $\widehat{\mathbf{m}}$ are the electric and magnetic dipole transition moment operators, respectively. The states $|\Psi_i\rangle$ and $|\Psi_f\rangle$ are electronic eigenstates resulting from a chiral arrangement of coupled electric dipole transition moments (or EDTMs), which are

each localized to a nucleic acid base residue. Equation C.1 shows that the rotational strength depends on the chirality of the coupled EDTMs, and its sign indicates the handedness (left versus right) of the chiral arrangement.

The Hamiltonian of the coupled system is given by

$$\widehat{H} = \widehat{H}_1 + \widehat{H}_2 + \widehat{V}_{12} \tag{C.2}$$

where $\widehat{H}_1$ and $\widehat{H}_2$ are the Hamiltonian operators of monomers 1 and 2, respectively and $\widehat{V}_{12}$ is the coupling between electronic transitions localized to each monomer as defined in the chapter III. The matrix element $V_{a1b2} = \langle\psi_{a1}|\widehat{V}_{12}|\psi_{b2}\rangle$ defines the coupling between monomer excited electronic states $\langle\psi_{a1}|$ (labeled a on monomer 1) and $\langle\psi_{b2}|$ (labeled b on monomer 2). The electronic coupling is calculated using the extended-dipole model (EDM),[236] which has been applied previously to cyanine dyes in self- assembled tubular J-aggregates,[237] to cyanine dimers in DNA,[238, 239] and to canonical nucleic acid bases in short segments of DNA.[156] In our current studies, the EDM accounts for the physical length of the adenine base by including for each monomer electronic transition a one-dimensional displacement vector, $\mathbf{l}$, that is oriented parallel to the EDTM direction. Each transition dipole moment is represented as two-point charges of equal magnitude and opposite sign ($\pm q$) separated by distance l. The coupling matrix element is given by

$$V_{a1b2} = \frac{|\mu_{a1}||\mu_{b2}|}{4\pi\epsilon\epsilon_0 l_{a1} l_{b2}} \left[\frac{1}{r_{12}^{a+b+}} - \frac{1}{r_{12}^{a-b+}} - \frac{1}{r_{12}^{a+b-}} + \frac{1}{r_{12}^{a-b-}}\right] \tag{C.3}$$

In Eq. C.3, $\mu_{a1} = q_{a1}\mathbf{l}_{a1}$ and $\mu_{b2} = q_{b2}\mathbf{l}_{b2}$ are the EDTMs of the transitions a and b on monomers 1 and 2, respectively, and the four distances $r_{12}^{a\pm b\pm}$ are those between the positive and negative point charges on monomers 1 and 2. The vacuum

permittivity of free space is given by ϵ_0, and ϵ is the local dielectric constant. For all of our calculations we used the value of the dielectric constant, $\epsilon = 2$, in accordance with prior conventions.[240]

In principle, further improvements to the accuracy of our calculations could be achieved by using more detailed, quantum chemical calculations of the electronic transition charge densities. Nevertheless, the favorable comparison between our calculations and experimental data presented below suggests that the EDM provides a reliable estimate of the electronic couplings between adjacent bases for present purposes.

We write the Hamiltonian on a monomer-site basis, such that singly-excited state wave functions are given by tensor products according to

$$|\Phi_{a1}\rangle = |\phi_{a1}\rangle|\phi_{g2}\rangle \text{ and } |\Phi_{a2}\rangle = |\phi_{a2}\rangle|\phi_{g1}\rangle \tag{C.4}$$

In Eq. C.4, $|\phi_{a1}\rangle$ and $|\phi_{a2}\rangle$ denote the a^{th} electronic excited states of monomers 1 and 2, respectively, and $|\phi_{g1}\rangle$ and $|\phi_{g2}\rangle$ are the electronic ground states. The number of distinct electronic transitions local to monomer 1 (2) is given by $n_{1(2)}$, such that the total number of site- localized transitions is $n_{\text{tot}} = n_1 + n_2$. The Hamiltonian of Eq. C.2 may thus be written on this site basis as a $n_{\text{tot}} \times n_{\text{tot}}$ matrix with diagonal elements representing the single site excitations (with energies E_{a1} and E_{b2}) and off-diagonal elements representing the couplings V_{a1b2} between monomer sites. Note that our formalism neglects the contribution from the isolated adenine monomer, which provides a signal for the fully unstacked conformation. In our calculations, however, this contribution is much smaller than the contribution due to the degenerate coupling of the adenine transitions.

Diagonalization of the Hamiltonian provides the eigen-states $|\Psi_k\rangle$ and eigen-energies E_k of the electronically coupled dinucleotide. In the so- called 'exciton' basis, the k^{th} singly-excited state $|\Psi_k\rangle$ may be written

$$|\Psi_k\rangle = \sum_{m=1}^{2} \sum_{a} C_{ma}^{k} |\Phi_{ma}\rangle, \tag{C.5}$$

where C_{ma}^{k} is the expansion coefficient corresponding to transition a local to monomer m. In the exciton basis, the ground state of the dinucleotide is given by $|\Psi_g\rangle = |\phi_{g1}\rangle|\phi_{g2}\rangle$. Using Eq. C.1, we may calculate the rotational strength $R_{gk}(= R_k)$ for the k^{th} electronic transition, where we assign the initial and final states to $|\Psi_g\rangle$ and $|\Psi_k\rangle$, respectively, and the total electric and magnetic dipole transition moment operators are given by vector sums $\widehat{\mu} = \widehat{\mu}_{a1} + \widehat{\mu}_{b2}$ and $\widehat{\mathbf{m}} = \widehat{\mathbf{m}}_{a1} + \widehat{\mathbf{m}}_{b2}$. For a given transition k, the rotational strength depends on the relative orientation of the monomer EDTMs. For the case of coupled degenerate transitions (i.e. $E_{a1} = E_{b2}$ and $E_k = E_{a1} + V_{a1b2}$), the rotational strength is given by

$$R_k = \frac{E_k}{4\hbar} \left[\mathbf{r}_{12} \cdot (\mu_{b2} \times \mu_{a1})\right] \tag{C.6}$$

For the case of non-degenerate coupled transitions (i.e.$E_{a1} \neq E_{b2}$ and $E_k \approx E_{a1}$), the rotational strength is given by

$$R_k = \frac{E_{a1} E_{b2}}{\hbar \left(E_{b2}^2 - E_{a1}^2\right)} \left[\mathbf{r}_{12} \cdot (\mu_{b2} \times \mu_{a1})\right] \tag{C.7}$$

We note that Eq. C.7 is written such that $E_{a1} > E_{b2}$.

To calculate the CD spectrum, we consider the relationship between the rotational strength and the integrated area of the CD spectrum within a finite spectral

range $\nu' \rightarrow \nu''$:

$$R = A \int_{\nu'}^{\nu''} d\nu \frac{\Delta\epsilon(\nu)}{\nu} \tag{C.8}$$

where $A = 7.659 \times 10^{-54}$ C^2 m^3 s^{-1}. The CD spectral line shape is obtained by summing over all contributions from individual transitions according to

$$\Delta\epsilon(\nu) = \sum_{k=1}^{n_{\text{tot}}} \Delta\epsilon(\nu_k) \tag{C.9}$$

For each of the k electronic transitions, we approximate the CD spectral line shape as a Gaussian function $\Delta\epsilon(\nu_k) = \Delta\bar{\epsilon}_k \exp\left\{-\left[(\nu_k - \overline{\nu}_k)^2 / 2\sigma_k^2\right]\right\}$, where σ_k is the Gaussian standard deviation, $\overline{\nu}_k$ $(= E_k/h)$ is the mean transition frequency, and $\Delta\bar{\epsilon}_k$ is the magnitude. Upon substitution of the above Gaussian function, Eq.C.9 is approximated by considering the frequency in the denominator to be constant over the width of the k spectral line, $\nu \approx \overline{\nu}_k$, and by extending the limits of the integral, $\nu' \approx \overline{\nu}_k - \Delta\nu_k$ and $\nu'' \approx \overline{\nu}_k + \Delta\nu_k$ with $\Delta\nu_k \gg 0$. This is a standard approximation used in the calculations of the CD spectra.[241] Solving the Gaussian integral, it follows that we may write the magnitude in this approximation as $\Delta\bar{\epsilon}_k = R_k\overline{\nu}_k / A\sqrt{2\pi}\sigma_k$. The whole spectrum is then calculated from Eq. C.9 by including the extended dipole modeling of the rotation strength for each k spectral line.

Selecting the Parameters for the Calculation of the CD Spectrum

For the majority of our CD calculations, we used as input parameters to Eqs. C.6 and C.7 the EDTM data for 9-methyladenine obtained by Holmén et al. (Table C.1) [2] and the dielectric constant $\epsilon = 2$. In Table C.1 we list for each transition the values we have used for the EDTM magnitude $|\mu|$, orientation δ, transition frequency ν, and extended transition dipole charge q and displacement l (see Fig. C.1). Furthermore,

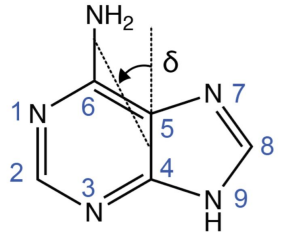

FIGURE C.1. The angle δ defines the direction of the electric dipole transition moment (EDTM) used in the CD calculations for the adenine bases of the dApdA dinucleotide monophosphate.

all transitions are in-plane $\pi \rightarrow \pi^*$, and are listed in order of increasing transition frequency. The angle δ specifies the counter-clockwise rotation of the EDTM vector within the plane of the adenine base relative to the C4-C5 bond axis (see Fig. C.1). The partial charges for the extended dipole model were derived using the relation $|\mu| = q|\mathbf{l}|$, and by representing the adenine base as an ellipse with major diameter (a) 4.6 Å and minor diameter (b) 2.6 Å such that $l = 2ab/\left[a^2\cos^2(\delta) + b^2\sin^2(\delta)\right]^{\frac{1}{2}}$.

In addition, to model the spectral line width of all monomer electronic transitions we assumed the Gaussian standard deviation $\sigma_k = 0.2$ eV. Our selection of these parameters was based on comparisons between the experimental CD spectrum of dApdA at room temperature in buffer at pH 7.2 containing 0.01 M $NaPO_3$ and 0.1 M $NaClO_4$, and CD calculations for which we assumed initially that the dApdA dinucleotide adopts only the B-form.

For comparison, we present in Table C.2 the empirical parameters from Williams et al.[3]

For all of the parameters that we tested (see Tables C.1 and C.2), we obtained moderately favorable agreement between experiment and theory. We note that the sensitivity of the calculated CD to the choice of input parameters was greatest at the

TABLE C.1. Experimental values for the magnitudes and molecular frame orientations of the electric dipole transition moments (EDTMs) for 9-methyladenine obtained by Holmén et al,[2] and which we have used to model adenine mononucleotide in chapter III.

Transition	ν (cm^{-1})	λ (nm)	$\|\mu\|$ (D)	$\delta(^\circ)$	l(Å)	q(e)
I	36 710	272.4	1.65	+66 ±7	3.96	0.09
II	38 820	257.6	3.63	+19 ±7	2.70	0.28
III	43 370	230.6	1.15	-15 ±6	2.66	0.09
IV	46 840	213.5	2.52	-21 ±7	2.72	0.19
V	48 320	207.0	2.30	-64 ±10	3.87	0.12

TABLE C.2. Empirical spectroscopic parameters from [3] for the adenine monomer.

Transition	ν (cm^{-1})	$\|\mu\|$ (D)	$\delta(^\circ)$
I	37 037	1.1	-87
II	38 022	4.0	-3
III	42 553	1.0	-87
IV	46 296	3.7	-87
V	51 282	3.7	-3
VI	53 476	4.2	-87

shortest wavelengths (200 – 250 nm) and least at the longer wavelengths (250 – 300 nm).

To demonstrate the sensitivity of the CD theoretical predictions to the choice of the empirical parameters selected in the CD modeling, we report first, in Fig. C.2A, a study of the CD spectrum for the Watson-Crick B-form of dApdA calculated using two different models: (i) the simple Point Dipole Approximation (PDA); and (ii) the Extended Dipole Model (EDM). The spectrum of the B-form, predicted by the theory is similar in both approximations, and shows a good agreement with experiments in the low energy part of the spectrum. Figure C.2B shows, instead, a study of the sensitivity of the calculations to the choice of the parameters. It reports results for the Point Dipole Approximation (PDA) calculation of the CD spectrum for the

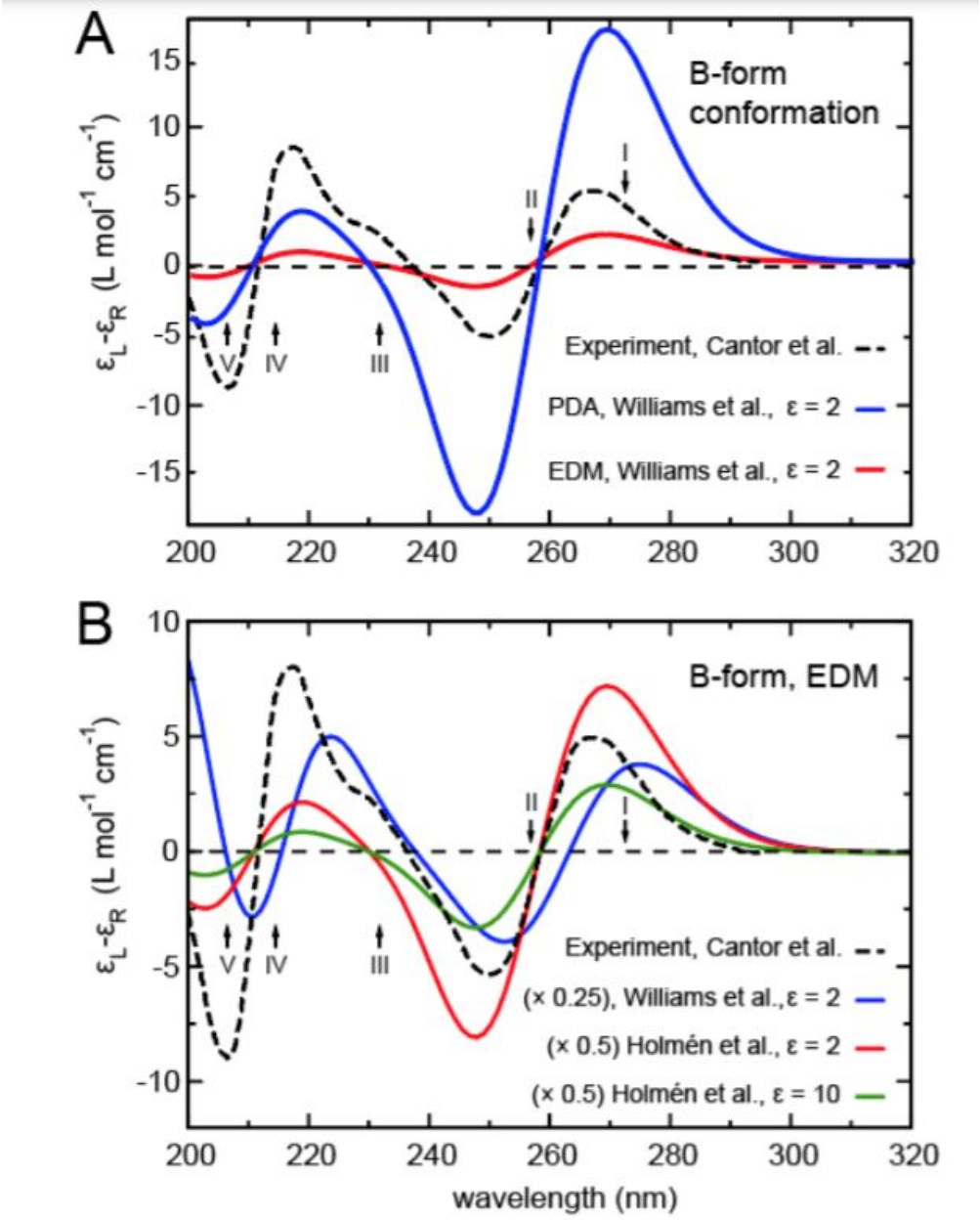

FIGURE C.2. (A) Comparison of the CD spectrum theoretically predicted for the Watson-Crick B-form of dApdA and the experimental data by Cantor at al.[1] We show both the spectra calculated using the Point Dipole Approximation (PDA) and the Extended Dipole Model (EDM). (B) Comparison of the CD spectrum theoretically predicted for the Watson-Crick B-form of dApdA and the experimental data, using either the empirical parameters from Holmén et al.[2] or from Williams et al.[3]. The effect of varying the dielectric constant (from $\epsilon = 2$ to $\epsilon = 10$) is also shown. In both panels, vertical arrows indicate the positions of the uncoupled transitions of the adenine monomer listed in Table C.1.

Watson-Crick B-form, while adopting either the empirical spectroscopic parameters from Holmén et al.[2] or those from Williams et al.[3] As can be seen, the positions and heights of the peaks change significantly at low wavelengths (high excitation energies), while they agree reasonably well at high wavelengths (low excitation energies). Thus, the results are qualitatively consistent with a slightly better agreement with the experimental values when using parameters from Table C.1. Finally, the figure shows a study of the variation of the empirical dielectric constant, which is found to produce just a change in the intensity of the spectrum, without affecting the positions of the peaks.

APPENDIX D

DERIVATION OF THE ANISOTROPIC HYDRODYNAMIC INTERACTION MATRIX AND RELATING THE LE4PD AND LE4PD-XYZ MODELS

Anisotropic Hydrodynamics

In this section, we propose an extension of the traditional derivation of the preaveraged hydrodynamic interaction while including the anisotropic formalism. The derivation is straightforward, but it may be useful, as we are not aware that it has been reported previously.

The system we are modeling is a coarse-grained description of a protein in solution, where the protein is a chain of beads, or friction points, in an effective solvent. Each bead represents one amino acid along the protein's primary sequence. The equation of motion of the fluctuations for a protein consisting of N beads, while neglecting hydrodynamic effects, is

$$\overline{\zeta}\Delta\dot{R}_i^{\alpha}(t) = -k_B T \sum_{\beta\in\{x,y,z\}} \sum_{j=1}^{N} A_{ij}^{\alpha\beta}\Delta R_j^{\beta}(t) + F_i^{\alpha}(t), \tag{D.1}$$

where $\overline{\zeta}$ is the average bead friction coefficient, $\overline{\zeta} = \frac{1}{N}\sum_{i=1}^{N}\zeta_i$, $\Delta R(t)$ is the 3N x 1 column vector of Cartesian displacements of each bead from its equilibrium position:

$$\Delta R(t) = \left[x_1(t) - \langle x_1\rangle, y_1(t) - \langle y_1\rangle, z_1(t) - \langle z_1\rangle, x_2(t) - \langle x_2\rangle, \ldots, z_N(t) - \langle z_N\rangle\right]^T, \tag{D.2}$$

with $\Delta\dot{R}_i^{\alpha}(t)$ denoting the fluctuation of the α-component of alpha-carbon i away from its equilibrium position; k_B is Boltzmann's constant, T is the temperature, A is a 3N x 3N structural matrix defining the coupling between the fluctuations of each

component of each alpha-carbon from its average position, and $F_i^{\alpha}(t)$ is a stochastic force along the α component that represents the fast collisions between the solvent and the i^{th} bead that obeys a delta-correlated, white-noise fluctuation-dissipation theorem:

$$\begin{aligned} \langle F_i^{\alpha}(t) \rangle &= 0 \\ \langle F_i{}^{\alpha}(t) F_j{}^{\beta}(t') \rangle &= 2k_B T \overline{\zeta} \delta(t-t') \delta_{ij} \delta_{\alpha\beta}, \end{aligned} \tag{D.3}$$

with $\delta_{ij}, \delta_{\alpha\beta}$ Kronecker delta symbols and α, β labeling the Cartesian indices of the force, $\alpha, \beta \in \{x, y, z\}$.

Since the protein is surrounded by a solvent, accounting for hydrodynamic effects is crucial for an accurate prediction of the protein's dynamic and kinetic behavior. The presence of the protein in the solvent will perturb the solvent's velocity; treating each bead in the protein as a solid sphere of radius s_i, the velocity of the α component of the fluid at the location of bead i is given by

$$v_i^{\alpha} = v_i^{\alpha,0} + v_i^{'\alpha}, \tag{D.4}$$

where $v_i^{\alpha,0}$ is the unperturbed α component of the velocity of the solvent (the velocity it would possess in the absence of the protein) and the perturbation in the velocity due to the presence of the other $j \neq i$ beads in the protein:

$$v_i^{'\alpha} = \sum_{\beta \in \{x,y,z\}} \sum_j T_{ij}^{\alpha\beta} F_j{}^{\beta,\zeta}. \tag{D.5}$$

As with $\Delta R(t)$, v' is an 3N x 1 column vector, with $v_i^{'\alpha}$ describing the perturbed velocity at alpha-carbon i in along the α component. $T_{ij}^{\alpha\beta}$ is an element of a $3N \times 3N$

tensor $\mathbf{T}$ that describes the coupling between the force exerted by bead j along the α component of the solvent and the resulting velocity perturbation experienced by bead i along the β component, and ${F_j}^{\alpha,\zeta}$ is the total force exerted on the solvent along component α by bead j. In equation D.5, the exact solution for $T_{ij}^{\alpha\beta}$ is (see, e.g. [186]),

$$\begin{aligned} T_{ij}^{\alpha\beta} &= \frac{1}{8\pi\eta_w r_{ij}}\left(\delta_{\alpha\beta} + (\widehat{rr})_{\alpha\beta} + \frac{s_i^2+s_j^2}{r_{ij}^2}\left[\frac{1}{3}\delta_{\alpha\beta} - (\widehat{rr})_{\alpha\beta}\right]\right),\ i \neq j \\ T_{ij}^{\alpha\beta} &= 0,\ i = j. \end{aligned} \tag{D.6}$$

In equation D.6, r_{ij} is the distance between beads i and j, s_i is the radius of bead i, and $\widehat{r}$ is a unit vector in the direction of r_{ij}, $\widehat{r} = \frac{\mathbf{r}}{r_{ij}} = \left(\frac{x_{ij}}{r_{ij}}, \frac{y_{ij}}{r_{ij}}, \frac{z_{ij}}{r_{ij}}\right)^T$. Accounting for the perturbation in the velocity at the location of bead i and the individual friction coefficient of each bead gives a modified Langevin equation:

$$\zeta_i\left(\Delta\dot{R}_i^{\alpha}(t) - {v_i}^{0,\alpha} - v_i'^{\alpha}\right) = -k_BT\sum_{\beta}\sum_{k} A_{ik}^{\alpha\beta}\Delta R_k^{\beta}(t) + F_i^{\alpha}(t)$$

$$\zeta_i\left(\Delta\dot{R}_i^{\alpha}(t) - \sum_{\beta}\sum_{j} T_{ij}^{\alpha\beta}F_j^{\beta,\zeta}\right) = -k_BT\sum_{\beta}\sum_{k} A_{ik}^{\alpha\beta}\Delta R_k^{\beta}(t) + F_i^{\alpha}(t). \tag{D.7}$$

In the second line, it has been assumed that no external force has been applied to the solvent (e.g. there are no plates at the top and bottom of the box shearing the fluid) so that $v_i^0 = 0$ and, equation D.5 has been used to write v_i' in terms of the forces exerted on the solvent by the beads. In the bead-and-spring-based model presented here, the beads can exert only two forces on the solvent: a spring force defined by A, $F_{\text{Spring},ij}^{\beta}(t) = -k_BT\sum_{\gamma}\sum_{j} A_{ij}^{\beta\gamma}\Delta R_j^{\gamma}(t)$, and a stochastic force due to random,

thermal fluctuations of the solvent moving the beads, $F_j(t)$. Thus,

$$\begin{aligned} F_i^{\beta,\zeta} &= \sum_j F^{\beta}_{\text{Spring},ij} + F_i^{\beta}(t) \\ &= -k_B T \sum_{\gamma} \sum_j A_{ij}^{\beta\gamma} \Delta R_j^{\gamma}(t) + F_i^{\beta}(t). \end{aligned} \tag{D.8}$$

Substituting equation D.8 into equation D.7 gives the explicit (anisotropic) equation of motion for bead i:

$$\zeta_i \left(\Delta \dot{R}_i^{\alpha}(t) - \sum_{\beta} \sum_j T_{ij}^{\alpha\beta} \left[\sum_{\gamma} \sum_k -k_B T A_{jk}^{\beta\gamma} \Delta R_k^{\gamma}(t) + F_j^{\beta} \right] \right) = -k_B T \sum_{\beta} \sum_k A_{ik}^{\alpha\beta} \Delta R_k^{\beta}(t) + F_i^{\alpha}(t). \tag{D.9}$$

To formulate a hydrodynamic interaction matrix, H, that describes the effect that hydrodynamics has on the motions of the other beads in the protein, there are two cases of equation D.9 to treat:

1. $i = j$: In this situation, $T_{ij}^{\alpha\beta} = 0$, so equation D.9 reduces to

$$\zeta_i \Delta \dot{R}_i^{\alpha}(t) = -k_B T \sum_{\beta} A_{ik}^{\alpha\beta} \Delta R_k^{\beta}(t) + F_i^{\alpha}(t)$$

$$\Rightarrow \Delta \dot{R}_i^{\alpha}(t) = \frac{\overline{\zeta}}{\zeta_i} \left(\frac{-k_B T}{\overline{\zeta}} \sum_{\beta} \sum_k A_{ik}^{\alpha\beta} \Delta R_k^{\beta}(t) \right) + \frac{F_i^{\alpha}(t)}{\zeta_i}. \tag{D.10}$$

   Equation D.10 implies that the form of the original Langevin equation, equation D.1, is recovered by defining $H_{ii}^{\alpha\beta}$ as $H_{ii}^{\alpha\beta} = \frac{\overline{\zeta}}{\zeta_i} \delta_{\alpha\beta}$. In equation D.10, ζ_i is the site-specific friction coefficient for residue i, which is the sum of two contributions. The first contribution is due to the partial exposure of the amino acid to the solvent, and the second is the friction due to its partial exposure to

the hydrophobic core of the protein:

$$\zeta_i = 6\pi \left(\eta_w r_{w,i} + \eta_p r_{p,i} \right), \tag{D.11}$$

where η_w is the bulk viscosity of the solvent, $r_{w,i}$ is the effective radius of residue i exposed to the solvent, η_p is the viscosity of the protein (the 'internal' viscosity), and $r_{p,i}$ is the effective radius of residue i exposed to the hydrophobic core of the protein. Both $r_{w,i}$ and $r_{p,i}$ are caclulated from the simulation and η_p is assumed to be 'slaved' to the solvent viscosity, so that it is proportional to the solvent viscosity. Specifically, it is assumed that the internal viscosity is the solvent viscosity, rescaled by a local-barrier energy scale of $k_B T$: $\eta_p = \exp[k_B T / k_B T] \eta_w \approx 2.71828 \eta_w$.

2. $i \neq j$: In this situation, the force terms on the right-hand side (RHS) of equation D.9 have already been treated in the $i = j$ case, so that, when $i \neq j$, the equation of motion reduces to

$$\zeta_i \Delta \dot{R}_i^\alpha(t) - \zeta_i \sum_\beta \sum_j T_{ij}^{\alpha\beta} \left[\sum_\gamma \sum_k -k_B T A_{jk}^{\beta\gamma} \Delta R_k^\gamma(t) + F_j^\beta \right] = 0$$

$$\Rightarrow \Delta \dot{R}_i^\alpha(t) = \overline{\zeta} \sum_\beta \sum_j T_{ij}^{\alpha\beta} \left[\sum_\gamma \sum_k -\frac{k_B T}{\overline{\zeta}} A_{jk}^{\beta\gamma} \Delta R_k^\gamma(t) + \frac{F_j^\beta}{\overline{\zeta}} \right]. \tag{D.12}$$

Again, by inspection, the form of equation D.1 is recovered by setting $H_{ij}^{\alpha\beta} = \overline{\zeta} T_{ij}^{\alpha\beta}$.

Combing these two cases, the complete hydrodynamic interaction matrix is defined piecewise as

$$H_{ij}^{\alpha\beta} = \begin{cases} \frac{\overline{\zeta}}{\zeta_i}\delta_{\alpha\beta}, & i = j \\ \\ \overline{\zeta} T_{ij}^{\alpha\beta}, & i \neq j \end{cases} \tag{D.13}$$

The off-diagonal elements of H can be simplified if we assume that only the portion of each bead exposed to solvent contributes to the hydrodynamic effect; that is, for each $H_{ij}^{\alpha\beta},\ i \neq j$, we assume $H_{ij}^{\alpha\beta} = \overline{\zeta}_w T_{ij}^{\alpha\beta}$, so that

$$\begin{aligned} H_{ij}^{\alpha\beta} = \overline{\zeta}_w T_{ij}^{\alpha\beta} &= 6\pi\eta_w \overline{r}_w T_{ij}^{\alpha\beta} \\ &= \frac{6\pi\eta_w \overline{r}_w}{8\pi\eta_w r_{ij}} \left(\delta_{\alpha\beta} + (\widehat{rr})_{\alpha\beta} + \frac{s_i^2 + s_j^2}{r_{ij}^2} \left[\frac{1}{3}\delta_{\alpha\beta} - (\widehat{rr})_{\alpha\beta} \right] \right) \\ &= \frac{3\overline{r}_w}{4r_{ij}} \left(\delta_{\alpha\beta} + (\widehat{rr})_{\alpha\beta} + \frac{s_i^2 + s_j^2}{r_{ij}^2} \left[\frac{1}{3}\delta_{\alpha\beta} - (\widehat{rr})_{\alpha\beta} \right] \right) \end{aligned} \tag{D.14}$$

When kept in the form given in equation D.14, the hydrodynamic interaction matrix is referred to as the *Rotne-Prager tensor*. This form gives the exact solution for the Stokes flow around a solid sphere moving through a viscous, incompressible fluid, and is useful because it accounts for the finite size of the beads. The hydrodynamic interaction matrix can be simplified in the limit of large inter-bead distances, i.e. in the limit of $r_{ij} \to \infty$. In this case, the term within the brackets in equation D.14 goes as $\frac{1}{r_{ij}^3}$ and is negligible compared to the first term in equation D.14. That is,

$$\begin{aligned} \lim_{r_{ij}\to\infty} H_{ij}^{\alpha\beta} &= \lim_{r_{ij}\to\infty} \frac{3\overline{r}_w}{4r_{ij}} \left(\delta_{\alpha\beta} + (\widehat{rr})_{\alpha\beta} + \frac{s_i^2 + s_j^2}{r_{ij}^2} \left[\frac{1}{3}\delta_{\alpha\beta} - (\widehat{rr})_{\alpha\beta} \right] \right) \\ &= \frac{3\overline{r}_w}{4r_{ij}} \left(\delta_{\alpha\beta} + (\widehat{rr})_{\alpha\beta} \right). \end{aligned} \tag{D.15}$$

The hydrodynamic interaction matrix defined in equation D.15 is known as the *Oseen tensor* and is simpler than what is given in equation D.14. This approximation treats the beads as point particles, and is accurate for large bead separations. However, in the case that the beads approach each other to within a bead radius, i.e. when $r_{ij} \leq \max(s_i, s_j)$, the Oseen tensor can give negative, unphysical eigenvalues because, in that instance, the point particle approximation is not really valid. Regardless if the off-diagonal elements of H are defined using equation D.14 or D.15, the equation of motion with the inclusion of hydrodynamic effects is

$$\overline{\zeta}\Delta\dot{R}_i^{\alpha}(t) = -k_B T \sum_{\beta,\gamma}\sum_{j,k} H_{ij}^{\alpha\beta} A_{jk}^{\beta\gamma} \Delta R_k^{\gamma}(t) + F_i^{\alpha}(t), \tag{D.16}$$

It should be noted that, up to this point, the pre-averaging approximation of Kirkwood and Riseman has not been invoked, so that H is still time-dependent.

Averaging the Hydrodynamic Interaction

A difficulty with the hydrodynamic interaction matrix given in equation D.13 is that it is non-linear in ΔR. Zimm's original solution to this problem was the use of the pre-averaging approximation developed by Kirkwood and Riseman in their treatment of the translational diffusion of polymers. Kirkwood and Riseman's approximation assumes that the inter-bead distribution is Gaussian and that

$$\langle H_{ij}\rangle = \frac{3\overline{r}_w}{4}\left\langle \frac{1}{r_{ij}}\left(\widehat{I} + \widehat{r}_{ij}\widehat{r}_{ij}\right)\right\rangle_{|r_{ij}|,\theta,\phi} = \frac{3\overline{r}_w}{4}\left\langle \frac{1}{r_{ij}}\right\rangle_{|r_{ij}|}\left(\langle\widehat{I}\rangle_{\theta,\phi} + \langle\widehat{r}_{ij}\widehat{r}_{ij}\rangle_{\theta,\phi}\right),\ i \neq j; \tag{D.17}$$

the second equality follows because the distribution of $\widehat{r}_{ij}$ is independent of its magnitude, r_{ij}; the average is taken over the equilibrium distribution of $\vec{r}_{ij}$, $\Psi(\vec{r}_{ij}) =$

$\left(\frac{3}{2\pi|i-j|l^2}\right)^{\frac{3}{2}} \exp\left[-\frac{3\vec{r}_{ij}^2}{2|i-j|l^2}\right]$, with l the average bond length between beads:

$$\begin{aligned}\langle\cdots\rangle_{|r_{ij}|,\theta,\phi} &= \int_0^{2\pi} d\phi \int_0^{\pi} d\theta \sin\theta \int_0^{\infty} dr_{ij} \cdots \Psi(\vec{r}_{ij}) \\ &= \int_0^{2\pi} d\phi \int_0^{\pi} d\theta \sin\theta \int_0^{\infty} dr_{ij} \cdots \Psi(r_{ij}),\end{aligned}$$

where $|r_{ij}| = r_{ij}$. Taking the angular average of $\widehat{r}_{ij}\widehat{r}_{ij}$ yields $\langle\widehat{r}_{ij}\widehat{r}_{ij}\rangle_{\theta,\phi} = \frac{4\pi}{3}\widehat{I}$ due to the tensor identity $\langle\widehat{r}_{ij}^{\alpha}\widehat{r}_{ij}^{\beta}\rangle_{\theta,\phi} = \frac{4\pi}{3}\delta_{\alpha\beta}$, where, as before, α and β index the Cartesian components of $\widehat{r}_{ij}$. So,

$$\left(\langle\widehat{I}\rangle_{\theta,\phi} + \langle\widehat{r}_{ij}\widehat{r}_{ij}\rangle_{\theta,\phi}\right) = 4\pi\left(\widehat{I} + \frac{1}{3}\widehat{I}\right) = \frac{16\pi}{3}\widehat{I},$$

and equation D.17 simplifies to

$$\begin{aligned}\langle H_{ij}\rangle &= 4\pi\overline{r}_w \left\langle\frac{1}{r_{ij}}\right\rangle_{|r_{ij}|}\widehat{I} \\ &= \overline{r}_w \left\langle\frac{1}{r_{ij}}\right\rangle\widehat{I}\end{aligned} \tag{D.18}$$

or

$$\langle H_{ij}^{\alpha\beta}\rangle = \overline{r}_w \left\langle\frac{1}{r_{ij}}\right\rangle\delta_{\alpha\beta} \tag{D.19}$$

since $\Psi(r_{ij})$ is independent of the angular integral. Because the 3 x 3 blocks of H on the diagonal are already time-independent (depending only on the individual friction coefficients of each bead), the complete pre-averaged, hydrodynamic interaction

matrix between beads i and j is given by

$$\langle H_{ij} \rangle = \frac{\overline{\zeta}}{\zeta_i} \delta_{ij} \widehat{I} + (1 - \delta_{ij}) \left\langle \frac{1}{r_{ij}} \right\rangle \widehat{I}. \tag{D.20}$$

or

$$\langle H_{ij}^{\alpha\beta} \rangle = \frac{\overline{\zeta}}{\zeta_i} \delta_{ij} \delta_{\alpha\beta} + (1 - \delta_{ij}) \left\langle \frac{1}{r_{ij}} \right\rangle \delta_{\alpha\beta}. \tag{D.21}$$

What equation D.20 means is that pre-averaging the hydrodynamic interaction, while removing the time-dependence, also removes the anisotropic effects inherent in the hydrodynamic interactions between beads. Although equation D.20 was derived using the Oseen tensor in H, since $\left(\langle \widehat{r}_{ij} \widehat{r}_{ij} \rangle\right)_{\alpha\beta} = \frac{1}{3}\delta_{\alpha\beta}$, a pre-averaging of the Rotne-Prager tensor gives the same result as using the Oseen tensor (because the second term on the RHS of equation D.14 contains $\frac{1}{3}\delta_{\alpha\beta} - \left(\widehat{r}_{ij}\widehat{r}_{ij}\right)_{\alpha\beta}$).

Relationship between the Isotropic and Anisotropic LE4PD

In the isotropic case, the relevant structural matrices are given by using the following definitions of the **U** and **A** matrices:[37, 38]

$$U_{N,ij} = \frac{\left\langle \vec{l}_i \cdot \vec{l}_j \right\rangle}{\left\langle |\vec{l}_i| \right\rangle \left\langle |\vec{l}_j| \right\rangle}$$

$$\mathbf{A}_N = \mathbf{a}^T \mathbf{U}_N^{-1} \mathbf{a},$$

with $\vec{l}_i = (r_{x,i}, r_{y,i}, r_{z,i})^T$ the bond vector between beads i and $i+1$. Taking the ij^{th} element of $\mathbf{A}^{-1}$ gives

$$\begin{aligned}
\mathbf{A}_N^{-1} &= \left(\mathbf{M}^T \begin{pmatrix} 0 & 0 \\ 0 & \mathbf{U}_N^{-1} \end{pmatrix} \mathbf{M} \right)^{-1} \\
&= \left(a^T \mathbf{U}_N^{-1} a) \right)^{-1} \\
&= a^{-1} \mathbf{U}_N \left(a^T \right)^{-1} \\
&\Rightarrow {A_N^{-1}}_{ij} \approx l^{-2} \mathbf{a}^{-1} \left\langle \vec{l}_i \cdot \vec{l}_j \right\rangle \mathbf{a}^{T^{-1}} = l^{-2} \left\langle \vec{R}_i \cdot \vec{R}_j \right\rangle \\
&\Rightarrow \mathrm{tr}\left(\mathbf{A}_N^{-1} \right) = l^{-2} \sum_i \left\langle \vec{R}_i \cdot \vec{R}_i \right\rangle = \frac{N}{l^2} \left\langle R_g^2 \right\rangle ,
\end{aligned} \tag{D.22}$$

with $\vec{R}_i$ the distance vector of bead i from the center-of-mass of the polymer, $\langle R_g^2 \rangle$ the average squared radius of gyration, and $\mathrm{tr}\,(\mathbf{B})$ the trace of a matrix $\mathbf{B}$. The approximation in the fourth line comes from taking all the average square bond lengths between beads to be equal, which is approximately true due to the stiffness of the peptide bond between the alpha-carbons in a protein. The trace of $\mathbf{A}_N^{-1}$ can be related to the trace of the 3N x 3N $\mathbf{A}$ by noting that $\mathbf{A}$ can be written as the difference of two matrices, since $\mathbf{A}^{-1} = \mathbf{C}$:

$$\begin{aligned}
\mathbf{A}^{-1} = \mathbf{C} &= \left\langle \Delta R \Delta R^T \right\rangle \\
&= \left\langle \left(R - \langle R \rangle \right) \left(R - \langle R \rangle \right)^T \right\rangle \\
&= \left\langle R R^T \right\rangle - \langle R \rangle \langle R \rangle^T \\
&= \mathbf{A}_1^{-1} - \mathbf{A}_2^{-1}.
\end{aligned} \tag{D.23}$$

Explicitly,

$$\mathbf{A}_1^{-1} = \begin{pmatrix} x_1x_1 & x_1y_1 & x_1z_1 & x_1x_2 & \cdots \\ y_1x_1 & y_1y_1 & y_1z_1 & y_1x_2 & \cdots \\ z_1x_1 & z_1y_1 & z_1z_1 & z_1x_2 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}$$

$$\Rightarrow \mathrm{tr}\left(\mathbf{A}_1^{-1}\right) = x_1x_1 + y_1y_1 + z_1z_1 + x_2x_2 + \ldots$$

$$= \vec{R}_1 \cdot \vec{R}_1 + \vec{R}_2 \cdot \vec{R}_2 + \ldots$$

$$= \sum_i \vec{R}_i \cdot \vec{R}_i = N\langle R_g^2 \rangle \Rightarrow \mathrm{tr}\left(\mathbf{A}_1^{-1}\right) \approx l^2 \mathrm{tr}\left(\mathbf{A}_N^{-1}\right). \tag{D.24}$$

Or, the trace of $\mathbf{A}_1^{-1}$ is approximately equal to the trace of $\mathbf{A}_N^{-1}$, with the approximation the equality of the bond lengths between each bead. The error in this approximation is about 0.5 %, while the error between $l^2\mathrm{tr}\left(\mathbf{A}_N^{-1}\right)$ and $\mathrm{tr}\left(\mathbf{A}^{-1}\right)$ is of the order 10^{-3}%; therefore, the equality holds to within the error of the approximation of equal bond lengths.

The accuracy of the normal modes generated from $\mathbf{A}$ can be established by examining their ability to reproduce the structural properties of the protein chain. For example, the mean-squared radius of gyration can be written in terms of the

eigenvectors and eigenvalues of $\mathbf{A_1}$:

$$\begin{aligned}\left\langle R_g^2 \right\rangle &= \frac{1}{N}\sum_{i=1}^{N}\left\langle \vec{R}_i \cdot \vec{R}_i \right\rangle \\ &= \frac{1}{N}\sum_{i=1}^{N}\left(\langle R_{i,x}R_{i,x}\rangle + \langle R_{i,y}R_{i,y}\rangle + \langle R_{i,z}R_{i,z}\rangle\right) \\ &= \frac{1}{N}\sum_{i=1}^{N}\left(\langle R_{i,x}R_{i,x}\rangle + \langle R_{i,y}R_{i,y}\rangle + \langle R_{i,z}R_{i,z}\rangle\right) \\ &= \mathrm{tr}\mathbf{A_1^{-1}} = \sum_{a=1}^{3N}\lambda_a^{-1}, \end{aligned} \tag{D.25}$$

where λ_a is the ath eigenvalue of $\mathbf{A_1}$. Similarly, $\langle R_{ete}^2 \rangle$ and the individual bond vectors can be found using the eigenvectors, $\mathbf{Q}$, from $\mathbf{A}^{-1}$:

$$\left\langle R_{ete}^2 \right\rangle = \sum_{a=1}^{3N}\left[\left(Q_{N,a}^x - Q_{1,a}^x\right)^2 + \left(Q_{N,a}^y - Q_{1,a}^y\right)^2 + \left(Q_{N,a}^z - Q_{1,a}^z\right)^2\right]\lambda_a^{-1}$$

$$\left\langle l_i^2 \right\rangle = \sum_{a=1}^{3N}\left[\left(Q_{i+1,a}^x - Q_{i,a}^x\right)^2 + \left(Q_{i+1,a}^y - Q_{i,a}^y\right)^2 + \left(Q_{i+1,a}^z - Q_{i,a}^z\right)^2\right]\lambda_a^{-1},$$

where Q_{ia}^x, Q_{ia}^y, and Q_{ia}^z have the same meaning as in the main text, but for the eigenvectors of $\mathbf{A}_1$ instead of $\mathbf{A}$. The agreement between the structural properties calculated directly from the simulation and using the modes agree to within the numerical precision of the simulation data. Analogous expressions can be found for the modes generated with the inclusion of the hydrodynamic interaction using the eigenvectors of $\mathbf{HA}_1$, but the eigenvalues of $\mathbf{A_1}$.

REFERENCES CITED

[1] Charles R. Cantor, Myron M. Warshaw, and Herman Shapiro. Oligonucleotide interactions. iii. circular dichroism studies of the conformation of deoxyoligonucleolides. *Biopolymers*, 9(9):1059–1077, 1970.

[2] Anders Holmén, Bengt Nordén, and Bo Albinsson. Electronic transition moments of 2-aminopurine. *Journal of the American Chemical Society*, 119(13):3114–3121, 1997.

[3] Arthur L Williams Jr, Chaejoon Cheong, Ignacio Tinico Jr, and Leigh B Clark. Vacuum ultraviolet circular dichroism as an indicator of helical handedness in nucleic acids. *Nucleic acids research*, 14(16):6649–6659, 1986.

[4] 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(16):164119, 2019.

[5] William C. Swope, Jed W. Pitera, and Frank Suits. Describing Protein Folding Kinetics by Molecular Dynamics Simulations. 1. Theory. *The Journal of Physical Chemistry B*, 108(21):6571–6581, 2004.

[6] B. Alberts, A. Johnson, J.H. Wilson, J. Lewis, T. Hunt, K. Roberts, M. Raff, and P. Walter. *Molecular Biology of the Cell*. Garland Science, 2008.

[7] J. C. Kendrew, G. Bodo, H. M. Dintzis, R. G. Parrish, H. Wyckoff, and D. C. Phillips. A three-dimensional model of the myoglobin molecule obtained by x-ray analysis. *Nature*, 181(4610):662–666, Mar 1958.

[8] R Kaptein, R Boelens, RM Scheek, and WF Van Gunsteren. Protein structures from nmr. *Biochemistry*, 27(15):5389–5395, 1988.

[9] D. E. Koshland. Application of a theory of enzyme specificity to protein synthesis. *Proceedings of the National Academy of Sciences of the United States of America*, 44(2):98–104, Feb 1958. 16590179[pmid].

[10] Jacque Monod, Jeffries Wyman, and Jean-Pierre Changeux. On the nature of allosteric transitions: A plausible model. *Journal of Molecular Biology*, 12(1):88 – 118, 1965.

[11] David D Boehr, Ruth Nussinov, and Peter E Wright. The role of dynamic conformatinal ensembles in biomolecular recognition. *Nat Chem Biol.*, 5(11):789–796, 2009.

[12] Buyong Ma, Sandeep Kumar, Chung-Jung Tsai, and Ruth Nussinov. Folding funnels and binding mechanisms. *Protein Engineering, Design and Selection*, 12(9):713–720, 09 1999.

[13] Peter Csermely, Robin Palotai, and Ruth Nussinov. Induced fit, conformational selection and independent dynamic segments: an extended view of binding events. *Trends in Biochemical Sciences*, 35(10):539–546, 2010.

[14] Christopher J Cramer. *Essentials of computational chemistry: theories and models*. John Wiley & Sons, 2013.

[15] J Andrew McCammon, Bruce R Gelin, and Martin Karplus. Dynamics of folded proteins. *Nature*, 267(5612):585–590, 1977.

[16] S. Piana, K. Lindorff-Larsen, and D. E. Shaw. Atomic-level description of ubiquitin folding. *Proceedings of the National Academy of Sciences*, 110(15):5915–5920, 2013.

[17] Kresten Lindorff-Larsen, Paul Maragakis, Stefano Piana, and David E. Shaw. Picosecond to Millisecond Structural Dynamics in Human Ubiquitin. *Journal of Physical Chemistry B*, 120(33):8313–8320, 2016.

[18] Maxwell I. Zimmerman, Justin R. Porter, Michael D. Ward, Sukrit Singh, Neha Vithani, Artur Meller, Upasana L. Mallimadugula, Catherine E. Kuhn, Jonathan H. Borowsky, Rafal P. Wiewiora, Matthew F. D. Hurley, Aoife M. Harbison, Carl A. Fogarty, Joseph E. Coffland, Elisa Fadda, Vincent A. Voelz, John D. Chodera, and Gregory R. Bowman. Sars-cov-2 simulations go exascale to predict dramatic spike opening and cryptic pockets across the proteome. *Nature Chemistry*, 13(7):651–659, Jul 2021.

[19] David E. Shaw, J. P. Grossman, Joseph A. Bank, Brannon Batson, J. Adam Butts, Jack C. Chao, Martin M. Deneroff, Ron O. Dror, Amos Even, Christopher H. Fenton, Anthony Forte, Joseph Gagliardo, Gennette Gill, Brian Greskamp, C. Richard Ho, Douglas J. Ierardi, Lev Iserovich, Jeffrey S. Kuskin, Richard H. Larson, Timothy Layman, Li-Siang Lee, Adam K. Lerer, Chester Li, Daniel Killebrew, Kenneth M. Mackenzie, Shark Yeuk-Hai Mok, Mark A. Moraes, Rolf Mueller, Lawrence J. Nociolo, Jon L. Peticolas, Terry Quan, Daniel Ramot, John K. Salmon, Daniele P. Scarpazza, U. Ben Schafer, Naseer Siddique, Christopher W. Snyder, Jochen Spengler, Ping Tak Peter Tang, Michael Theobald, Horia Toma, Brian Towles, Benjamin Vitale, Stanley C. Wang, and Cliff Young. Anton 2: Raising the bar for performance and programmability in a special-purpose molecular dynamics supercomputer. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, SC '14, page 41–53. IEEE Press, 2014.

[20] Schrödinger, LLC. The PyMOL molecular graphics system, version 1.8. November 2015.

[21] William Humphrey, Andrew Dalke, and Klaus Schulten. VMD – Visual Molecular Dynamics. *Journal of Molecular Graphics*, 14:33–38, 1996.

[22] F. Sittel and G. Stock. Perspective: Identification of collective variables and metastable states of protein dynamics. *The Journal of Chemical Physics*, 149(15):150901, 2018.

[23] P. Zhuravlev and G. Papoian. Protein functional landscapes, dynamics, allostery: a tortuous path towards a universal theoretical framework. *Quarterly Reviews of Biophysics*, 43(3):295–332, 2010.

[24] Alessandro Laio and Michele Parrinello. Escaping free-energy minima. *Proceedings of the National Academy of Sciences*, 99(20):12562–12566, 2002.

[25] M. Tuckerman. *Statistical Mechanics: Theory and Molecular Simulation*. Oxford Graduate Texts. OUP Oxford, 2010.

[26] A. Amadei, A. Linssen, and H. Berendsen. Essential dynamics of proteins. *Proteins: Structure, Function, and Bioinformatics*, 17(4):412–425, 1993.

[27] A. R. Atilgan, S. R. Durell, R. L. Jernigan, M. C. Demirel, O. Keskin, and I. Bahar. Anisotropy of fluctuation dynamics of proteins with an elastic network model. *Biophysical Journal*, 80(1):505–515, 2001.

[28] Ivet Bahar, Ali Rana Atilgan, Melik C. Demirel, and Burak Erman. Vibrational dynamics of folded proteins: Significance of slow and fast motions in relation to function and stability. *Physical Review Letters*, 80(12):2733–2736, 1998.

[29] Monique M. Tirion. Large amplitude elastic motions in proteins from a single-parameter, atomic analysis. *Physical Review Letters*, 77(9):1905–1908, 1996.

[30] Guillermo Pérez-Hernández, Fabian Paul, Toni Giorgino, Gianni De Fabritiis, and Frank Noé. Identification of slow molecular order parameters for Markov model construction. *Journal of Chemical Physics*, 139(1), 2013.

[31] Christian R. Schwantes and Vijay S. Pande. Improvements in markov state model construction reveal many non-native interactions in the folding of ntl9. *Journal of Chemical Theory and Computation*, 9(4):2000–2009, 2013. PMID: 23750122.

[32] Ayori Mitsutake and Hiroshi Takano. Relaxation mode analysis for molecular dynamics simulations of proteins. *Biophysical Reviews*, 10(2):375–389, Apr 2018.

[33] Frank Noé, Illia Horenko, Christof Schütte, and Jeremy C. Smith. Hierarchical analysis of conformational dynamics in biomolecules: Transition networks of metastable states. *Journal of Chemical Physics*, 126(15):155102, 2007.

[34] Nicolae Viorel Buchete and Gerhard Hummer. Coarse master equations for peptide folding dynamics. *Journal of Physical Chemistry B*, 112(19):6057–6069, 2008.

[35] Andreas Mardt, Luca Pasquali, Hao Wu, and Frank Noé. Vampnets for deep learning of molecular kinetics. *Nature Communications*, 9(1):5, Jan 2018.

[36] Christoph Wehmeyer and Frank Noé. Time-lagged autoencoders: Deep learning of slow collective variables for molecular kinetics. *The Journal of Chemical Physics*, 148(24):241703, 2018.

[37] Esther Caballero-Manrique, Jenelle K. Bray, William A. Deutschman, Frederick W. Dahlquist, and Marina G. Guenza. A theory of protein dynamics to predict NMR relaxation. *Biophysical Journal*, 93(12):4128–4140, 2007.

[38] 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(29):9195–9211, 2015.

[39] Robert Zwanzig. Theoretical basis for the rouse-zimm model in polymer solution dynamics. *The Journal of Chemical Physics*, 60(7):2717–2720, 1974.

[40] M. Doi and S.F. Edwards. *The Theory of Polymer Dynamics*. Clarendon Press: Oxford, 1986.

[41] P.G. de Gennes. *Scaling Concepts in Polymer Physics*. Cornell University Press, 1979.

[42] Stephen J Hagen. Solvent viscosity and friction in protein folding dynamics. *Current Protein & Peptide Science*, 11(5):385–395, 2010.

[43] Diane E. Sagnella, John 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(17):7702–7711, 2000.

[44] H. Frauenfelder, P. W. Fenimore, G. Chen, and B. H. McMahon. Protein folding is slaved to solvent motions. *Proceedings of the National Academy of Sciences*, 103(42):15469–15472, 2006.

[45] 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(3):16–18, 2011.

[46] R. Zwanzig. Diffusion in a rough potential. *Proceedings of the National Academy of Sciences*, 85(7):2029–2030, 1988.

[47] B. Iglewicz and D.C. Hoaglin. *How to Detect and Handle Outliers*. ASQC basic references in quality control. ASQC Quality Press, 1993.

[48] J. Copperman and M. G. Guenza. Mode localization in the cooperative dynamics of protein recognition. *Journal of Chemical Physics*, 145(1):015101, 2016.

[49] G.R. Bowman, V.S. 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.

[50] Stefan Klus, Feliks Nüske, Péter Koltai, Hao Wu, Ioannis Kevrekidis, Christof Schütte, and Frank Noé. Data-driven model reduction and transfer operator approximation. *Journal of Nonlinear Science*, 28(3):985–1010, Jun 2018.

[51] E. R. Beyerle and M. G. Guenza. Comparison between slow, anisotropic le4pd fluctuations and the principal component analysis modes of ubiquitin. *bioRxiv*, 2021.

[52] David Komander and Michael Rape. The ubiquitin code. *Annual Review of Biochemistry*, 81(1):203–229, 2012.

[53] Senadhi Vijay-Kumar, Charles E. Bugg, and William J. Cook. Structure of ubiquitin refined at 1.8 Å resolution. *Journal of Molecular Biology*, 194(3):531–544, 1987.

[54] Nico Tjandra, Scott E Feller, Richard W Pastor, and Ad Bax. Rotational Diffusion Anisotropy of Human Ubiquitin from 1 5 N NMR Relaxation. *J. Am. Chem. Soc.*, 117(12):12562–12566, 1995.

[55] Oliver F. Lange, Nils Alexander Lakomek, Christophe Farès, Gunnar F. Schröder, Korvin F.A. Walter, Stefan Becker, Jens Meiler, Helmut Grubmüller, Christian Griesinger, and Bert L. De Groot. Recognition dynamics up to microseconds revealed from an RDC-derived ubiquitin ensemble in solution. *Science*, 320(5882):1471–1475, 2008.

[56] S. F. Lienin, T. Bremi, B. Brutscher, R. Brüschweiler, and R. R. Ernst. Anisotropic intramolecular backbone dynamics of ubiquitin characterized by NMR relaxation and MD computer simulation. *Journal of the American Chemical Society*, 120(38):9870–9879, 1998.

[57] J. Copperman and M. G. Guenza. Predicting protein dynamics from structural ensembles. *The Journal of Chemical Physics*, 143(24):243131, 2015.

[58] George I. Makhatadze, Marin M. Lopez, John M. Richardson III, and Susan T. Thmos. Anion binding to the ubiquitin molecule. *Protein Science*, 7(3):689–697, 1998.

[59] Lorenza Penengo, Marina Mapelli, Andrea G. Murachelli, Stefano Confalonieri, Laura Magri, Andrea Musacchio, Pier Paolo Di Fiore, Simona Polo, and Thomas R. Schneider. Crystal structure of the ubiquitin binding domains of rabex-5 reveals two modes of interaction with ubiquitin. *Cell*, 124(6):1183 – 1195, 2006.

[60] Ivan Bosanac, Ingrid E. Wertz, Borlan Pan, Christine Yu, Saritha Kusam, Cynthia Lam, Lilian Phu, Qui Phung, Brigitte Maurer, David Arnott, Donald S. Kirkpatrick, Vishva M. Dixit, and Sarah G. Hymowitz. Ubiquitin binding to a20 znf4 is required for modulation of nf-κb signaling. *Molecular Cell*, 40(4):548 – 557, 2010.

[61] James H. Hurley, Sangho Lee, and Gali Prag. Ubiquitin-binding domains. *Biochemical Journal*, 399(3):361–372, 2006.

[62] Thomas P. Garner, Joanna Strachan, Elizabeth C. Shedden, Jed E. Long, James R. Cavey, Barry Shaw, Robert Layfield, and Mark S. Searle. Independent interactions of ubiquitin-binding domains in a ubiquitin-mediated ternary complex. *Biochemistry*, 50(42):9076–9087, 2011. PMID: 21923101.

[63] Sangho Lee, Yien Che Tsai, Rafael Mattera, William J. Smith, Michael S. Kostelansky, Allan M. Weissman, Juan S. Bonifacino, and James H. Hurley. Structural basis for ubiquitin recognition and autoubiquitination by rabex-5. *Nature Structural & Molecular Biology*, 13(3):264–271, 2006.

[64] Koraljka Husnjak and Ivan Dikic. Ubiquitin-binding proteins: Decoders of ubiquitin-mediated cellular functions. *Annual Review of Biochemistry*, 81(1):291–322, 2012.

[65] E. R. Beyerle and M. G. Guenza. "Identifying the leading dynamics of ubiquitin: a comparison between the tICA and the LE4PD slow fluctuations in amino acids' position," submitted to *The Journal of Chemical Physics*.

[66] Eric R Beyerle, Mohammadhasan Dinpajooh, Huiying Ji, Peter H von Hippel, Andrew H Marcus, and Marina G Guenza. Dinucleotides as simple models of the base stacking-unstacking component of DNA 'breathing' mechanisms. *Nucleic Acids Research*, 49(4):1872–1885, 01 2021.

[67] J. Copperman, M. Dinpajooh, E. R. Beyerle, and M. G. Guenza. Universality and Specificity in Protein Fluctuation Dynamics. *Physical Review Letters*, 119(15):158101, 2017.

[68] Tomasz Wlodarski and Bojan Zagrovic. Conformational selection and induced fit mechanism underlie specificity in noncovalent interactions with ubiquitin. *Proceedings of the National Academy of Sciences*, 106(46):19346–19351, 2009.

[69] K. Tai, T. Shen, U. Börjesson, M. Philippopoulos, and J. A. McCammon. Analysis of a 10-ns molecular dynamics simulation of mouse acetylcholinesterase. *Biophysical Journal*, 81(2):715–724, Aug 2001.

[70] Ivet Bahar and Robert L. Jernigan. Cooperative fluctuations and subunit communication in tryptophan synthase. *Biochemistry*, 38(12):3478–3490, 1999.

[71] Ivet Bahar, Timothy R. Lezon, Lee-Wei Yang, and Eran Eyal. Global Dynamics of Proteins: Bridging Between Structure and Function. *Annual Review of Biophysics*, 39(1):23–42, 2010.

[72] Nikita Chopra, Thomas E. Wales, Raji E. Joseph, Scott E. Boyken, John R. Engen, Robert L. Jernigan, and Amy H. Andreotti. Dynamic Allostery Mediated by a Conserved Tryptophan in the Tec Family Kinases. *PLOS Computational Biology*, 12(3):1–19, 2016.

[73] Jan H. Peters and Bert L. de Groot. Ubiquitin dynamics in complexes reveal molecular recognition mechanisms beyond induced fit and conformational selection. *PLOS Computational Biology*, 8(10):1–10, 10 2012.

[74] Katherine Henzler-Wildman and Dorothee Kern. Dynamic personalities of proteins. *Nature*, 450(7172):964–972, 2007.

[75] David Komander. The emerging complexity of protein ubiquitination. *Biochemical Society Transactions*, 37(5):937–953, 2009.

[76] Manuel Barreto Miranda and Alexander Sorkin. Regulation of receptors and transporters by ubiquitination: new insights into surprisingly similar mechanisms. *Molecular Interventions*, 7 3:157–67, 2007.

[77] Lingyan Jin, Adam Williamson, Sudeep Banerjee, Isabelle Philipp, and Michael Rape. Mechanism of ubiquitin-chain formation by the human anaphase-promoting complex. *Cell*, 133(4):653 – 665, 2008.

[78] Zongyang Lv, Katelyn M. Williams, Lingmin Yuan, James H. Atkison, and Shaun K. Olsen. Crystal structure of a human ubiquitin e1–ubiquitin complex reveals conserved functional elements essential for activity. *Journal of Biological Chemistry*, 293(47):18337–18352, 2018.

[79] Diwakar Shukla, Yilin Meng, Benoît Roux, and Vijay S. Pande. Activation pathway of src kinase reveals intermediate states as targets for drug design. *Nature Communications*, 5:3397, Mar 2014.

[80] Robert D. Malmstrom, Christopher T. Lee, Adam T. Van Wart, and Rommie E. Amaro. Application of molecular-dynamics based markov state models to functional proteins. *Journal of Chemical Theory and Computation*, 10(7):2648–2657, Jul 2014.

[81] Yilin Meng, Diwakar Shukla, Vijay S. Pande, and Benoît Roux. Transition path theory analysis of c-src kinase activation. *Proceedings of the National Academy of Sciences*, 113(33):9193–9198, 2016.

[82] Wei Wang, Siqin Cao, Lizhe Zhu, and Xuhui Huang. Constructing markov state models to elucidate the functional conformational changes of complex biomolecules. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 8(1):e1343, 2018.

[83] Raymond Jin and Lutz Maibaum. Mechanisms of dna hybridization: Transition path analysis of a simulation-informed markov model. *The Journal of Chemical Physics*, 150(10):105103, 2019.

[84] Guillermo Pérez-Hernández, Fabian Paul, Toni Giorgino, Gianni De Fabritiis, and Frank Noé. Identification of slow molecular order parameters for Markov model construction. *Journal of Chemical Physics*, 139(1), 2013.

[85] Martin K. Scherer, Benjamin Trendelkamp-Schroer, Fabian Paul, Guillermo Pérez-Hernández, Moritz Hoffmann, Nuria Plattner, Christoph Wehmeyer, Jan Hendrik Prinz, and Frank Noé. PyEMMA 2: A Software Package for Estimation, Validation, and Analysis of Markov Models. *Journal of Chemical Theory and Computation*, 11(11):5525–5542, 2015.

[86] Zhen Gang Wang. 50th Anniversary Perspective: Polymer Conformation - A Pedagogical Review. *Macromolecules*, 50(23):9073–9114, 2017.

[87] Weinan E, Weiqing Ren, and Eric Vanden-Eijnden. String method for the study of rare events. *Phys. Rev. B*, 66:052301, Aug 2002.

[88] Weinan E and Eric Vanden-Eijnden. Transition-Path Theory and Path-Finding Algorithms for the Study of Rare Events. *Annual Review of Physical Chemistry*, 61(1):391–420, 2010.

[89] Kirby N. Swatek and David Komander. Ubiquitin modifications. *Cell Research*, 26:399–42, Mar 2016.

[90] David L Nelson and Michael M Cox. *Lehninger Principles of Biochemistry, Fifth Edition*. Freeman, 2008.

[91] Ivan Dikic, Soichi Wakatsuki, and Kylie J. Walters. Ubiquitin-binding domains – from structures to functions. *Nature Reviews Molecular Cell Biology*, 10:659–671, Oct 2009.

[92] Hans Frauenfelder, Stephen G. Sligar, and Peter G. Wolynes. The energy landscapes and motions of proteins. *Science*, 254(5038):1598–1603, 1991.

[93] George Casella, Stephen Fienberg, and Ingram Olkin. *Springer Texts in Statistics*, volume 102. Springer, 2006.

[94] Benjamin Trendelkamp-Schroer, Hao Wu, Fabian Paul, and Frank Noé. Estimation and uncertainty of reversible Markov models. *Journal of Chemical Physics*, 143(17):174101, 2015.

[95] Roger A. Horn and Charles R. Johnson. *Matrix Analysis*. Cambridge University Press, New York, NY, USA, 1986.

[96] Baron Peters. *Reaction Rate Theory and Rare Events Simulations*. Elsevier, Amsterdam, 2017.

[97] Christof Schütte, Frank Noé, Jianfeng Lu, Marco Sarich, and Eric Vanden-Eijnden. Markov state models based on milestoning. *The Journal of Chemical Physics*, 134(20):204105, 2011.

[98] F. Noe, C. Schutte, E. Vanden-Eijnden, L. Reich, and T. R. Weikl. Constructing the equilibrium ensemble of folding pathways from short off-equilibrium simulations. *Proceedings of the National Academy of Sciences*, 106(45):19011–19016, 2009.

[99] L.E. Reichl. *A Modern Course in Statistical Physics*. Wiley, 1998.

[100] Peter Deuflhard and Marcus Weber. Robust Perron cluster analysis in conformation dynamics. *Linear Algebra and Its Applications*, 398(1-3):161–184, 2005.

[101] Susanna Röblitz and Marcus Weber. Fuzzy spectral clustering by PCCA+: application to Markov state models and data classification. *Advances in Data Analysis and Classification*, 7(2):147–179, jun 2013.

[102] Peter H. von Hippel, Neil P. Johnson, and Andrew H. Marcus. Fifty years of dna "breathing": Reflections on old and new approaches. *Biopolymers*, 99(12):923–954, Dec 2013. 23840028[pmid].

[103] Michel Peyrard, Santiago Cuesta-Lopez, and Guillaume James. Nonlinear analysis of the dynamics of dna breathing. *Journal of biological physics*, 35(1):73–89, 2009.

[104] Niklas Bosaeus, Anna Reymer, Tamás Beke-Somfai, Tom Brown, Masayuki Takahashi, Pernilla Wittung-Stafshede, Sandra Rocha, and Bengt Nordén. A stretched conformation of dna with a biological role? *Quarterly Reviews of Biophysics*, 50, 2017.

[105] Bobo Feng, Robert P Sosa, Anna KF Mårtensson, Kai Jiang, Alex Tong, Kevin D Dorfman, Masayuki Takahashi, Per Lincoln, Carlos J Bustamante, Fredrik Westerlund, et al. Hydrophobic catalysis and a potential biological role of dna unstacking induced by environment effects. *Proceedings of the National Academy of Sciences*, 116(35):17169–17174, 2019.

[106] Maxim Frank-Kamenetskii. How the double helix breathes. *Nature*, 328(6125):17–18, 1987.

[107] Carol L Cech, Werner Hug, and Ignacio Tinoco Jr. Polynucleotide circular dichroism calculations: use of an all-order classical coupled oscillator polarizability theory. *Biopolymers: Original Research on Biomolecules*, 15(1):131–152, 1976.

[108] Peter M. Bayley, Eigil B. Nielsen, and John A. Schellman. Rotatory properties of molecules containing two peptide groups: theory. *The Journal of Physical Chemistry*, 73(1):228–243, 1969.

[109] Margaret Jean Lowe and John A. Schellman. Solvent effects on dinucleotide conformation. *Journal of Molecular Biology*, 65(1):91 – 109, 1972.

[110] Vincenzo Rizzo and John A Schellman. Matrix-method calculation of linear and circular dichroism spectra of nucleic acids and polynucleotides. *Biopolymers: Original Research on Biomolecules*, 23(3):435–470, 1984.

[111] C Allen Bush and Ignacio Tinoco Jr. Calculation of the optical rotatory dispersion of dinucleoside phosphates. *Journal of molecular biology*, 23(3):601–614, 1967.

[112] WC Johnson Jr and I Tinoco Jr. Circular dichroism of polynucleotides: A general method applied to dimers. *Biopolymers: Original Research on Biomolecules*, 8(6):715–731, 1969.

[113] Howard DeVoe and Ignacio Tinoco Jr. The stability of helical polynucleotides: base contributions. *Journal of molecular biology*, 4(6):500–517, 1962.

[114] John SantaLucia. A unified view of polymer, dumbbell, and oligonucleotide dna nearest-neighbor thermodynamics. *Proceedings of the National Academy of Sciences*, 95(4):1460–1465, 1998.

[115] John SantaLucia Jr and Donald Hicks. The thermodynamics of dna structural motifs. *Annu. Rev. Biophys. Biomol. Struct.*, 33:415–440, 2004.

[116] Neil P Johnson and Thomas Schleich. Circular dichroism studies of the conformational stability of dinucleoside phosphates and related compounds in aqueous neutral salt solutions. *Biochemistry*, 13(5):981–987, 1974.

[117] Peter H Von Hippel and Kwok-Ying Wong. Neutral salts: the generality of their effects on the stability of macromolecular conformations. *Science*, 145(3632):577–580, 1964.

[118] John D. Chodera, Nina Singhal, Vijay S. Pande, Ken A. Dill, and William C. Swope. Automatic discovery of metastable states for the construction of markov models of macromolecular conformational dynamics. *The Journal of Chemical Physics*, 126(15):155101, 2007.

[119] A. Rodger and B. Nordén. *Circular Dichroism and Linear Dichroism*. Oxford Classical Monographs. Oxford University Press, 1997.

[120] Huiying Ji, Neil P Johnson, Peter H von Hippel, and Andrew H Marcus. Local dna base conformations and ligand intercalation in dna constructs containing optical probes. *Biophysical journal*, 117(6):1101–1115, 2019.

[121] Philipp Metzner, Christof Schütte, and Eric Vanden-Eijnden. Illustration of transition path theory on a collection of simple examples. *The Journal of Chemical Physics*, 125(8):084110, 2006.

[122] Philipp Metzner, Christof Schütte, and Eric Vanden-Eijnden. Transition path theory for markov jump processes. *Multiscale Modeling & Simulation*, 7(3):1192–1219, 2009.

[123] Alexander T Hawk and Dmitrii E Makarov. Milestoning with transition memory. *The Journal of chemical physics*, 135(22):224109, 2011.

[124] Weinan E and Eric Vanden-Eijnden. Towards a theory of transition paths. *Journal of Statistical Physics*, 123(3):503–523, 2006.

[125] Malka Kitayner, Haim Rozenberg, Remo Rohs, Oded Suad, Dov Rabinovich, Barry Honig, and Zippora Shakked. Diversity in dna recognition by p53 revealed by crystal structures with hoogsteen base pairs. *Nature structural & molecular biology*, 17(4):423–429, 2010.

[126] Barry Honig and Remo Rohs. Flipping watson and crick. *Nature*, 470(7335):472–473, 2011.

[127] Evgenia N. Nikolova, Eunae Kim, Abigail A. Wise, Patrick J. O'Brien, Ioan Andricioaei, and Hashim M. Al-Hashimi. Transient Hoogsteen base pairs in canonical duplex DNA. *Nature*, 470(7335):498–504, feb 2011.

[128] Heidi S Alvey, Federico L Gottardo, Evgenia N Nikolova, and Hashim M Al-Hashimi. Widespread transient hoogsteen base pairs in canonical duplex dna with variable energetics. *Nature communications*, 5(1):1–8, 2014.

[129] Huiqing Zhou, Bradley J Hintze, Isaac J Kimsey, Bharathwaj Sathyamoorthy, Shan Yang, Jane S Richardson, and Hashim M Al-Hashimi. New insights into hoogsteen base pairs in dna duplexes from a structure-based survey. *Nucleic acids research*, 43(7):3420–3433, 2015.

[130] Jaroslav Kypr, Iva Kejnovská, Daniel Renčiuk, and Michaela Vorlíčková. Circular dichroism and conformational polymorphism of dna. *Nucleic acids research*, 37(6):1713–1725, 2009.

[131] Remo Rohs, Sean M West, Alona Sosinsky, Peng Liu, Richard S Mann, and Barry Honig. The role of dna shape in protein–dna recognition. *Nature*, 461(7268):1248–1253, 2009.

[132] Daniel Coman and Irina M Russu. A nuclear magnetic resonance investigation of the energetics of basepair opening pathways in dna. *Biophysical journal*, 89(5):3285–3292, 2005.

[133] Josep M Huguet, Cristiano V Bizarro, Núria Forns, Steven B Smith, Carlos Bustamante, and Felix Ritort. Single-molecule derivation of salt dependent base-pair free energies in dna. *Proceedings of the National Academy of Sciences*, 107(35):15431–15436, 2010.

[134] Carl Schildkraut and Shneior Lifson. Dependence of the melting temperature of dna on salt concentration. *Biopolymers: Original Research on Biomolecules*, 3(2):195–208, 1965.

[135] Richard Owczarzy, Yong You, Bernardo G Moreira, Jeffrey A Manthey, Lingyan Huang, Mark A Behlke, and Joseph A Walder. Effects of sodium ions on dna duplex oligomers: improved predictions of melting temperatures. *Biochemistry*, 43(12):3537–3554, 2004.

[136] C.R. Cantor, P.R.S. Charles R. Cantor, R.C. Cantor, P.R. Schimmel, W. H. Freeman, and Company. *Biophysical Chemistry: Part I: The Conformation of Biological Macromolecules*. Biophysical Chemistry. W. H. Freeman, 1980.

[137] Mark James Abraham, Teemu Murtola, Roland Schulz, Szilárd Páll, Jeremy C. Smith, Berk Hess, and Erik Lindahl. Gromacs: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX*, 1-2:19–25, 2015.

[138] Yong Duan, Chun Wu, Shibasish Chowdhury, Mathew C Lee, Guoming Xiong, Wei Zhang, Rong Yang, Piotr Cieplak, Ray Luo, Taisung Lee, et al. A point-charge force field for molecular mechanics simulations of proteins based on condensed-phase quantum mechanical calculations. *Journal of computational chemistry*, 24(16):1999–2012, 2003.

[139] William L. Jorgensen, Jayaraman Chandrasekhar, Jeffry D. Madura, Roger W. Impey, and Michael L. Klein. Comparison of simple potential functions for simulating liquid water. *The Journal of Chemical Physics*, 79(2):926–935, 1983.

[140] Giovanni Bussi, Davide Donadio, and Michele Parrinello. Canonical sampling through velocity rescaling. *The Journal of Chemical Physics*, 126(1):014101, 2007.

[141] Berk Hess, Henk Bekker, Herman J.C. Berendsen, and Johannes G.E.M. Fraaije. LINCS: A Linear Constraint Solver for molecular simulations. *Journal of Computational Chemistry*, 18(12):1463–1472, 1997.

[142] David Arthur and Sergei Vassilvitskii. K-Means++: the Advantages of Careful Seeding. *Proceedings of the eighteenth annual ACM-SIAM symposium on discrete algorithms*, 8:1027–1025, 2007.

[143] M Emre Celebi, Hassan A Kingravi, and Patricio A Vela. A comparative study of efficient initialization methods for the k-means clustering algorithm. *Expert systems with applications*, 40(1):200–210, 2013.

[144] Jan Hendrik Prinz, Hao Wu, Marco Sarich, Bettina Keller, Martin Senne, Martin Held, John D. Chodera, Christof Schtte, and Frank Noé. Markov models of molecular kinetics: Generation and validation. *Journal of Chemical Physics*, 134(17), 2011.

[145] N.G. Van Kampen. *Stochastic Processes in Physics and Chemistry*. North-Holland Personal Library. Elsevier Science, 2011.

[146] Kozo Hamaguchi and E Peter Geiduschek. The effect of electrolytes on the stability of the deoxyribonucleate helix. *Journal of the American Chemical Society*, 84(8):1329–1338, 1962.

[147] Gerald S Manning. Limiting laws and counterion condensation in polyelectrolyte solutions i. colligative properties. *The journal of chemical Physics*, 51(3):924–933, 1969.

[148] Gerald S. Manning. Limiting laws and counterion condensation in polyelectrolyte solutions ii. self-diffusion of the small ions. *The Journal of Chemical Physics*, 51(3):934–938, 1969.

[149] Pierandrea Lo Nostro and Barry W. Ninham. Hofmeister phenomena: An update on ion specificity in biology. *Chemical Reviews*, 112(4):2286–2322, Apr 2012.

[150] Yu Bai, Max Greenfeld, Kevin J. Travers, Vincent B. Chu, Jan Lipfert, Sebastian Doniach, and Daniel Herschlag. Quantitative and comprehensive decomposition of the ion atmosphere around nucleic acids. *Journal of the American Chemical Society*, 129(48):14981–14988, Dec 2007.

[151] Jan Lipfert, Sebastian Doniach, Rhiju Das, and Daniel Herschlag. Understanding nucleic acid–ion interactions. *Annual Review of Biochemistry*, 83(1):813–841, 2014. PMID: 24606136.

[152] P. G. Romano and M. G. Guenza. GRadient Adaptive Decomposition (GRAD) Method: Optimized Refinement Along Macrostate Borders in Markov State Models. *Journal of Chemical Information and Modeling*, 57(11):2729–2740, 2017.

[153] P. L. Privalov. Physical basis of the dna double helix. *Journal of Biophysics and Structural Biology*, 8(1):1–7, 2020.

[154] Robert L. Baldwin and George D. Rose. How the hydrophobic factor drives protein folding. *Proceedings of the National Academy of Sciences*, 113(44):12462–12466, 2016.

[155] P.G. Hoel, S.C. Port, and C.J. Stone. *Introduction to Stochastic Processes*. Waveland Press, 1986.

[156] B. Bouvier, T. Gustavsson, D. Markovitsi, and P. Millié. Dipolar coupling between electronic transitions of the dna bases and its relevance to exciton states in double helices. *Chemical Physics*, 275(1):75–92, 2002. Photoprocesses in Multichromophoric Molecular Assemblies.

[157] In Suk Joung and Thomas E. Cheatham. Determination of alkali and halide monovalent ion parameters for use in explicitly solvated biomolecular simulations. *The Journal of Physical Chemistry B*, 112(30):9020–9041, Jul 2008.

[158] Steven Hayward, Akio Kitao, and Herman 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(3):425–437, 1997.

[159] Scott H. Northrup, Michael R. Pear, John D. Morgan, J.Andrew McCammon, and Martin Karplus. Molecular dynamics of ferrocytochrome c: Magnitude and anisotropy of atomic displacements. *Journal of Molecular Biology*, 153(4):1087 – 1109, 1981.

[160] Alexander L. Tournier and Jeremy C. Smith. Principal components of the protein dynamical transition. *Phys. Rev. Lett.*, 91:208106, Nov 2003.

[161] Justin Chan, Kazuhiro Takemura, Hong-Rui Lin, Kai-Chun Chang, Yuan-Yu Chang, Yasumasa Joti, Akio Kitao, and Lee-Wei Yang. An efficient timer and sizer of biomacromolecular motions. *Structure*, 28(2):259 – 269.e8, 2020.

[162] Angel E. García. Large-amplitude nonlinear motions in proteins. *Physical Review Letters*, 68(17):2696–2699, 1992.

[163] I.T. Jolliffe. *Principal Component Analysis*. Springer Series in Statistics. Springer, 2002.

[164] Hongfeng Lou and Robert I. Cukier. Molecular dynamics of apo-adenylate kinase: A principal component analysis. *The Journal of Physical Chemistry B*, 110(25):12796–12808, 2006.

[165] Herman JC Berendsen and Steven Hayward. Collective protein dynamics in relation to function. *Current Opinion in Structural Biology*, 10(2):165 – 169, 2000.

[166] Jakob Spiegelberg and Ján Rusz. Can we use pca to detect small signals in noisy data? *Ultramicroscopy*, 172:40 – 46, 2017.

[167] M. A. Balsera, W. Wriggers, Y. Oono, and K. Schulten. Principal component analysis and long time protein dynamics. *The Journal of Physical Chemistry*, 100(7):2567–2572, 1996.

[168] Kannan Sankar, Jie Liu, Yuan Wang, and Robert L. Jernigan. Distributions of experimental protein structures on coarse-grained free energy landscapes. *Journal of Chemical Physics*, 143(24), 2015.

[169] Yusuke Naritomi and Sotaro Fuchigami. Slow dynamics in protein fluctuations revealed by time-structure based independent component analysis: The case of domain motions. *Journal of Chemical Physics*, 134(6), 2011.

[170] Lee Wei Yang, Eran Eyal, Ivet Bahar, and Akio Kitao. Principal component analysis of native ensembles of biomolecular structures (PCA_NEST): Insights into functional dynamics. *Bioinformatics*, 25(5):606–614, 2009.

[171] Lars Skjaerven, Aurora Martinez, and Nathalie Reuter. Principal component and normal mode analysis of proteins; a quantitative comparison using the groel subunit. *Proteins: Structure, Function, and Bioinformatics*, 79(1):232–243, 2011.

[172] Tao Wu. *A MOLECULAR DYNAMICS SIMULATION BASED PRINCIPAL COMPONENT ANALYSIS FRAMEWORK FOR COMPUTATION OF MULTI-SCALE MODELING OF PROTEIN AND ITS INTERACTION WITH SOLVENT*. PhD thesis, New Jersey Institute of Technology, 2011.

[173] Barry J. Grant, Ana P. C. Rodrigues, Karim M. ElSawy, J. Andrew McCammon, and Leo S. D. Caves. Bio3d: an R package for the comparative analysis of protein structures. *Bioinformatics*, 22(21):2695–2696, 08 2006.

[174] Alexander Berezhkovskii and Attila Szabo. Ensemble of transition states for two-state protein folding from the eigenvectors of rate matrices. *The Journal of Chemical Physics*, 121(18):9186–9187, 2004.

[175] R.A. Horn and C.R. Johnson. *Topics in Matrix Analysis*. Cambridge University Press, 1994.

[176] Angelo Perico and Marina Guenza. Viscoelastic relaxation of segment orientation in dilute polymer solutions. *The Journal of Chemical Physics*, 83(6):3103–3109, 1985.

[177] Marina Guenza. Many chain correlated dynamics in polymer fluids. *The Journal of Chemical Physics*, 110(15):7574–7588, 1999.

[178] Konrad Hinsen, Andrei-Jose Petrescu, Serge Dellerue, Marie-Claire Bellissent-Funel, and Gerald R Kneller. Harmonicity in slow protein dynamics. *Chemical Physics*, 261(1-2):25–37, 2000.

[179] Kresten Lindorff-Larsen, Stefano Piana, Kim Palmo, Paul Maragakis, John L. Klepeis, Ron O. Dror, and David E. Shaw. Improved side-chain torsion potentials for the amber ff99sb protein force field. *Proteins: Structure, Function, and Bioinformatics*, 78(8):1950–1958, 2010.

[180] E.B. Wilson, J.C. Decius, and P.C. Cross. *Molecular Vibrations: The Theory of Infrared and Raman Vibrational Spectra*. Dover Books on Chemistry Series. Dover Publications, 1980.

[181] Guillaume Chevrot, Paolo Calligari, Konrad Hinsen, and Gerald R. Kneller. Least constraint approach to the extraction of internal motions from molecular dynamics trajectories of flexible macromolecules. *Journal of Chemical Physics*, 135(8), 2011.

[182] A.R. Meenakshi and C. Rajian. On a product of positive semidefinite matrices. *Linear Algebra and its Applications*, 295(1):3 – 6, 1999.

[183] H.A. Kramers. Brownian motion in a field of force and the diffusion model of chemical reactions. *Physica*, 7(4):284 – 304, 1940.

[184] George D Rose, Ari R Geselowitz, Glenn J Lesser, Richard H Lee, and Micheal H Zehfus. Hydrophobicity of amino acid residues in globular proteins. *Science*, 229(4716):834–838, 1985.

[185] Susan Miller, Joël Janin, Arthur M. Lesk, and Cyrus Chothia. Interior and surface of monomeric proteins. *Journal of Molecular Biology*, 196(3):641 – 656, 1987.

[186] L.D. Landau and E.M. Lifshitz. *Fluid Mechanics: Volume 6*. Elsevier Science, 2013.

[187] D. Ruppert and D.S. Matteson. *Statistics and Data Analysis for Financial Engineering: with R examples*. Springer Texts in Statistics. Springer New York, 2015.

[188] Brooke E. Husic, Robert T. McGibbon, Mohammad M. Sultan, and Vijay S. Pande. Optimized parameter selection reveals trends in markov state models for protein folding. *The Journal of Chemical Physics*, 145(19):194103, 2016.

[189] Hao Wu and Frank Noé. Variational approach for learning markov processes from time series data. *arXiv preprint arXiv:1707.04659*, 2017.

[190] Fabian Paul, Hao Wu, Maximilian Vossel, Bert L de Groot, and Frank Noé. Identification of kinetic order parameters for non-equilibrium dynamics. *The Journal of chemical physics*, 150(16):164120, 2019.

[191] C.W. Gardiner. *Handbook of stochastic methods for physics, chemistry, and the natural sciences*. Springer series in synergetics. Springer, 1994.

[192] Canan Baysal and Ali Rana Atilgan. Relaxation kinetics and the glassiness of native proteins: Coupling of timescales. *Biophysical Journal*, 88(3):1570 – 1576, 2005.

[193] Osman Burak Okan, Ali Rana Atilgan, and Canan Atilgan. Nanosecond motions in proteins impose bounds on the timescale distributions of local dynamics. *Biophysical Journal*, 97(7):2080 – 2088, 2009.

[194] John J. Portman, Shoji Takada, and Peter G. Wolynes. Microscopic theory of protein folding rates. II. Local reaction coordinates and chain dynamics. *Journal of Chemical Physics*, 114(11):5082–5096, 2001.

[195] 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.

[196] N. D. Socci, J. N. Onuchic, and P. G. Wolynes. Diffusive dynamics of the reaction coordinate for protein folding funnels. *The Journal of Chemical Physics*, 104(15):5860–5868, 1996.

[197] S. Wu, P. Zhuravlev, and G. Papoian. High resolution approach to the native state ensemble kinetics and thermodynamics. *Biophysical Journal*, 95(12):5524 – 5532, 2008.

[198] G Maisuradze, A Liwo, and H Scheraga. Principal component analysis for protein folding dynamics. *Journal of molecular biology*, 385(1):312–329, 2009.

[199] R. Hegger, A. Altis, P. Nguyen, and Gerhard Stock. How complex is the dynamics of peptide folding? *Physical Review Letters*, 98(2):10–13, 2007.

[200] A. Kitao and N. Go. Investigating protein dynamics in collective coordinate space. *Current Opinion in Structural Biology*, 9(2):164–169, 1999.

[201] Melik C. Demirel, Ali Rana Atilgan, Ivet Bahar, Robert L. Jernigan, and Burak Erman. Identification of kinetically hot residues in proteins. *Protein Science*, 7(12):2522–2532, 1998.

[202] A. Hyvärinen, J. Karhunen, and E. Oja. *Independent Component Analysis*. Adaptive and Cognitive Dynamic Systems: Signal Processing, Learning, Communications and Control. Wiley, 2001.

[203] Ayori Mitsutake, Hiromitsu Iijima, and Hiroshi Takano. Relaxation mode analysis of a peptide system: Comparison with principal component analysis. *The Journal of Chemical Physics*, 135(16):164102, 2011.

[204] Berk Hess. Convergence of sampling in protein simulations. *Physical Review E - Statistical Physics, Plasmas, Fluids, and Related Interdisciplinary Topics*, 65(3):1–10, 2002.

[205] Pratyush Tiwary and B. J. Berne. Spectral gap optimization of order parameters for sampling complex molecular systems. *Proceedings of the National Academy of Sciences*, 113(11):2839–2844, 2016.

[206] L. Molgedey and H. G. Schuster. Separation of a mixture of independent signals using time delayed correlations. *Physical Review Letters*, 72(23):3634–3637, 1994.

[207] Mohammad M. Sultan and Vijay S. Pande. tica-metadynamics: Accelerating metadynamics by using kinetically selected collective variables. *Journal of Chemical Theory and Computation*, 13(6):2440–2447, 2017. PMID: 28383914.

[208] James McCarty and Michele Parrinello. A variational conformational dynamics approach to the selection of collective variables in metadynamics. *Journal of Chemical Physics*, 147(20), 2017.

[209] Hiroshi Takano and Seiji Miyashita. Relaxation modes in random spin systems. *Journal of the Physical Society of Japan*, 64(10):3688–3698, 1995.

[210] Ayori Mitsutake and Hiroshi Takano. Relaxation mode analysis and markov state relaxation mode analysis for chignolin in aqueous solution near a transition temperature. *The Journal of Chemical Physics*, 143(12):124111, 2015.

[211] R.B. Bird, C.F. Curtiss, R.C. Armstrong, and O. Hassager. *Dynamics of Polymeric Liquids, Volume 2: Kinetic Theory*. Wiley, 1987.

[212] J. Copperman, M. Dinpajooh, E. R. Beyerle, and M. G. Guenza. Universality and Specificity in Protein Fluctuation Dynamics. *Physical Review Letters*, 119(15):1–11, 2017.

[213] Jesse Hall. private communication.

[214] Mohammad M. Sultan, Hannah K. Wayment-Steele, and Vijay S. Pande. Transferable Neural Networks for Enhanced Sampling of Protein Dynamics. *Journal of Chemical Theory and Computation*, 14(4):1887–1894, 2018.

[215] Frank Noé and Cecilia Clementi. Kinetic Distance and Kinetic Maps from Molecular Dynamics Simulation. *Journal of Chemical Theory and Computation*, 11(10):5002–5011, 2015.

[216] Simon Olsson, Hao Wu, Fabian Paul, Cecilia Clementi, and Frank Noé. Combining experimental and simulation data of molecular processes via augmented markov models. *Proceedings of the National Academy of Sciences*, 114(31):8265–8270, 2017.

[217] Bettina G. Keller, Jan-Hendrik Prinz, and Frank Noé. Markov models and dynamical fingerprints: Unraveling the complexity of molecular kinetics. *Chemical Physics*, 396:92 – 107, 2012.

[218] F. Noe, S. Doose, I. Daidone, M. Lollmann, M. Sauer, J. D. Chodera, and J. C. Smith. Dynamical fingerprints for probing individual relaxation processes in biomolecular dynamics with simulations and kinetic experiments. *Proceedings of the National Academy of Sciences*, 108(12):4822–4827, 2011.

[219] John D Chodera and Frank Noé. Probability distributions of molecular observables computed from markov models. ii. uncertainties in observables and their time-evolution. *The Journal of chemical physics*, 133(10):09B606, 2010.

[220] Matthew J. Comstock, Kevin D. Whitley, Haifeng Jia, Joshua Sokoloski, Timothy M. Lohman, Taekjip Ha, and Yann R. Chemla. Direct observation of structure-function relationship in a nucleic acid–processing enzyme. *Science*, 348(6232):352–354, 2015.

[221] Hisham Mazal and Gilad Haran. Single-molecule fret methods to study the dynamics of proteins at work. *Current Opinion in Biomedical Engineering*, 12:8–17, 2019. Molecular & CellularEngineering: single molecule technology Neural Engineering: High Resolution Cell Imaging.

[222] Frank Noé, Ralf Banisch, and Cecilia Clementi. Commute maps: Separating slowly mixing molecular configurations for kinetic modeling. *Journal of Chemical Theory and Computation*, 12(11):5620–5630, Nov 2016.

[223] Asghar M. Razavi and Vincent A. Voelz. Kinetic network models of tryptophan mutations in b-hairpins reveal the importance of non-native interactions. *Journal of Chemical Theory and Computation*, 11(6):2801–2812, Jun 2015.

[224] Guillermo Pérez-Hernández and Frank Noé. Hierarchical time-lagged independent component analysis: Computing slow modes and reaction coordinates for large molecular systems. *Journal of Chemical Theory and Computation*, 12(12):6118–6129, Dec 2016.

[225] P. Deuflhard, W. Huisinga, A. Fischer, and Ch Schütte. Identification of almost invariant aggregates in reversible nearly uncoupled Markov chains. *Linear Algebra and Its Applications*, 315(1-3):39–59, 2000.

[226] Vijay S Pande, Kyle Beauchamp, and Gregory R Bowman. Everything you wanted to know about markov state models but were afraid to ask. *Methods*, 52(1):99–105, 2010.

[227] Kyle A. Beauchamp, Robert McGibbon, Yu-Shan Lin, and Vijay S. Pande. Simple few-state models reveal hidden complexity in protein folding. *Proceedings of the National Academy of Sciences*, 109(44):17807–17813, 2012.

[228] Brooke E. Husic and Vijay S. Pande. Markov state models: From an art to a science. *Journal of the American Chemical Society*, 140(7):2386–2396, Feb 2018.

[229] N.G. van Kampen. *Chapter IV - MARKOV PROCESSES*. North-Holland Personal Library. Elsevier, Amsterdam, third edition edition, 2007.

[230] Martin K. Scherer, Brooke E. Husic, Moritz Hoffmann, Fabian Paul, Hao Wu, and Frank Noé. Variational selection of features for molecular kinetics. *The Journal of Chemical Physics*, 150(19):194108, 2019.

[231] William C. Swope, Jed W. Pitera, Frank Suits, Mike Pitman, Maria Eleftheriou, Blake G. Fitch, Robert S. Germain, Aleksandr Rayshubski, T. J. C. Ward, Yuriy Zhestkov, and Ruhong Zhou. Describing protein folding kinetics by molecular dynamics simulations. 2. example applications to alanine dipeptide and a β-hairpin peptide. *The Journal of Physical Chemistry B*, 108(21):6582–6594, 2004.

[232] Christian R. Schwantes and Vijay S. Pande. Modeling molecular kinetics with tica and the kernel trick. *Journal of Chemical Theory and Computation*, 11(2):600–608, Feb 2015.

[233] Wei Chen, Hythem Sidky, and Andrew L. Ferguson. Nonlinear discovery of slow molecular modes using state-free reversible vampnets. *The Journal of Chemical Physics*, 150(21):214114, 2019.

[234] Wei Chen, Hythem Sidky, and Andrew L. Ferguson. Capabilities and limitations of time-lagged autoencoders for slow mode discovery in dynamical systems. *The Journal of Chemical Physics*, 151(6):064123, 2019.

[235] W.C. Brenke. *Plane and Spherical Trigonometry*. The Dryden Press, 1943.

[236] V. Czikklely, H.D. Forsterling, and H. Kuhn. Extended dipole model for aggregates of dye molecules. *Chemical Physics Letters*, 6(3):207–210, 1970.

[237] Catalin Didraga, Audrius Pugžlys, P Ralph Hania, Hans von Berlepsch, Koos Duppen, and Jasper Knoester. Structure, spectroscopy, and microscopic model of tubular carbocyanine dye aggregates. *The Journal of Physical Chemistry B*, 108(39):14976–14985, 2004.

[238] Dylan Heussman, Justin Kittell, Loni Kringle, Amr Tamimi, Peter H. von Hippel, and Andrew H. Marcus. Measuring local conformations and conformational disorder of (cy3)2 dimer labeled dna fork junctions using absorbance, circular dichroism and two-dimensional fluorescence spectroscopy. *Faraday Discuss.*, 216:211–235, 2019.

[239] Brittany L. Cannon, Donald L. Kellis, Lance K. Patten, Paul H. Davis, Jeunghoon Lee, Elton Graugnard, Bernard Yurke, and William B. Knowlton. Coherent exciton delocalization in a two-state dna-templated dye aggregate system. *The Journal of Physical Chemistry A*, 121(37):6905–6916, 2017. PMID: 28813152.

[240] Mohammadhasan Dinpajooh, Daniel R. Martin, and Dmitry V. Matyushov. Polarizability of the active site of cytochrome c reduces the activation barrier for electron transfer. *Scientific Reports*, 6:1–7, 2016.

[241] John A Schellman. Circular dichroism and optical rotation. *Chemical Reviews*, 75(3):323–331, 1975.