DIFFUSIVE AND ACTIVATED CONTRIBUTIONS IN PROTEIN DYNAMICS

by

JEREMY COPPERMAN

A DISSERTATION

Presented to the Department of Physics
and the Graduate School of the University of Oregon
in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy

June 2016

DISSERTATION APPROVAL PAGE

Student: Jeremy Copperman

Title: Diffusive and Activated Contributions in Protein Dynamics

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

Eric Corwin	Chair
Marina Guenza	Advisor
John Toner	Core Member
James Imamura	Core Member
William Cresko	Institutional Representative

and

Scott L. Pratt	Dean of the Graduate School

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

Degree awarded June 2016

DISSERTATION ABSTRACT

Jeremy Copperman

Doctor of Philosophy

Department of Physics

June 2016

Title: Diffusive and Activated Contributions in Protein Dynamics

A novel approach is developed to describe the dynamics of proteins, described as fundamentally semiflexible objects collapsed into the free energy well representing the folded state. This is a multi-scale approach, where structural correlations are used as input to an effectively linear description, which can be solved in diffusive modes. The accuracy of the LE4PD is verified by analyzing the predicted dynamics across a set of seven different proteins for which both relaxation data and NMR solution structures are available.

The biological function of proteins is encoded in their structure and expressed through the mediation of their dynamics. We present here a study of how local fluctuation relates to binding and function. This study indicates how local fluctuations are likely to initiate biologically relevant pathways as they cooperatively enhance the dynamics in specific spatial regions of the protein. The picture that emerges is a dynamically heterogenous protein where biologically active regions provide energetically-comparable conformational states that can be trapped by a reacting partner.

The long-time dynamics of proteins is controlled by an activated regime where the dynamics of the large amplitude diffusive modes becomes dominated by the presence

of energy barriers. We explicitly study the atomistic simulation-derived free energy landscape projected from the diffusive modes of the linear Langevin description of the protein, and obtain a general scaling between the fluctuation lengthscale and complexity. This hierarchical property of the free energy landscape of proteins is shown to be general across a set of six different single-domain monomeric proteins. As a consequence microscopic timescales of sub-angstrom sized fluctuations rapidly propagate out to folding timescales at the nanometer lengthscale of globular single-domain proteins.

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

CURRICULUM VITAE

NAME OF AUTHOR: Jeremy Copperman

GRADUATE AND UNDERGRADUATE SCHOOLS ATTENDED:

University of Oregon, Eugene

DEGREES AWARDED:

Doctor of Philosophy, Physics, 2016, University of Oregon

AREAS OF SPECIAL INTEREST:

Keeping Erika happy and having fun with Rylan, Maddox, and Cruise

PROFESSIONAL EXPERIENCE:

Graduate Research and Teaching Assistant

GRANTS, AWARDS AND HONORS:

STEMCORE Fellow (2014 - 2015) Fellowship in science education outreach

GK12 Fellow (2013 - 2014) Fellowship in science education outreach

PUBLICATIONS:

Copperman, J., and M. G. Guenza. "Predicting protein dynamics from structural ensembles." *The Journal of Chemical Physics* 143.24 (2015): 243131.

Copperman, J., and Marina G. Guenza. "Coarse-Grained Langevin Equation for Protein Dynamics: Global Anisotropy and a Mode Approach to Local Complexity." *The Journal of Physical Chemistry B* 119.29 (2014): 9195-9211.

McCarty, J., Clark, A. J., Copperman, J., and Guenza, M. G. "An analytical coarse-graining method which preserves the free energy, structural correlations, and thermodynamic state of polymer melts from the atomistic to the mesoscale." *The Journal of Chemical Physics*, 140.20 (2014), 204913.

Copperman, J., and Marina G. Guenza. "Mode localization in the cooperative dynamics of protein recognition." *manuscript in submission.*

Copperman, J., and Marina G. Guenza. "A hierarchical free energy landscape connects ordered and disordered protein states." *manuscript in preparation.*

ACKNOWLEDGEMENTS

I am inspired by the NSF reviewer who commented about my graduate fellowship application, "Applicant mentions a family. A person with family issues can only hope to survive in science, not excel." This ridiculous notion is completely dispelled by the support and opportunities made available during my time at the University of Oregon, and especially the physics department and Dr. Steck, Dr. Raymer, and Dr. Imamura, who took a chance to help a late in the game local get into graduate school. Thank you Dr. Corwin for the many interesting discussions and help along the way, and the rest of my committee as well for your careful consideration. There are too many faculty who have contributed greatly to my education to list, but thank you in particular to Dr. Toner, whose content and clarity of presentation is unforgettable.

I owe a huge debt to my advisor, Dr. Marina Guenza. I have enjoyed the collaborative scientific adventure from day one, and sincerely appreciate all of the opportunities made available through your hard work and scientific insight. I hope I have learned enough to maintain the commitment to the faithful description of the real physical system of interest that you have demanded in our research.

This work was supported in part by the National Science Foundation Grant CHE-1362500 and DMR-0804145, and under the Graduate STEM Fellows in K-12 Education (GK-12) program, grant DGE-0742540.

My gratitude to my family is complete, as they make up my identity. Erika, Rylan, Maddox, and Cruise, I love you forever. Mom your dinners and love keep us going, and Mike so does lunch. I dedicate this thesis to my Dad, who is the smartest person I know.

TABLE OF CONTENTS

LIST OF FIGURES

Figure Page

LIST OF TABLES

CHAPTER I

INTRODUCTION

The subject of this dissertation is a theoretical model describing the diffusive relaxation of an isolated well-folded protein (or protein cluster) in aqueous solution. The derivation of the model will be described in some detail, but primarily the work consists of the application and expansion of simple dynamical models of linear chain molecules, or polymers, to the specific case of a chain of amino acids with an equilibrium folded structure. This very finite-sized system is surprisingly rich, with diffusive and activated dynamical contributions spanning the picosecond to many milliseconds in timescale. Globular proteins are highly evolved amino acid chains, and obtaining the specific structures which these macromolecules take has been one of the great scientific feats of the previous century; this is the basis for our understanding of structural biology. This may become the century of dynamical structural biology, as the understanding grows of how biological processes are governed by the motion and cooperative rearrangement of macromolecular configuration in time. This dissertation is driven by the desire to understand the fundamental quantities governing biologically relevant fluctuations of protein structures.

Folded proteins possess intrinsic fluctuations in isolation which directly correspond to their biological function; binding pathways, catalysis, etc. Experimental efforts to determine these equilibrium motions is plagued by a lack of resolution at the microscopic lengthscale and microscopic to macroscopic timescales of interest, and the interpretation of available experimental data is often difficult because of the lack of both realistic and tractable theoretical models. This has been the goal of the research described herein, to develop a method to quantitavely describe

the motion of a protein in a realistic yet tractable model, taking the underlying static structural ensemble as input. The story that emerges over the course of the work is fascinating because of the remarkable organization of intrinsic dynamical pathways and biological function in these highly evolved macromolecules.

Chapter II describes the model in detail. This work, co-authored with Dr. Marina Guenza, was published in the Journal of Physical Chemistry B in 2014,[1] and builds upon the previously introduced formalism to include global anisotropy and a rudimentary approach to account for energy barriers in the protein free energy landscape. The method is only applied and studied for the Ubiquitin protein.

In Chapter III, the model is extended to allow the use of experimentally obtained NMR conformer ensembles of proteins, which typically contain only a handful of structures as opposed to the near-continuous ensembles obtained using atomistic molecular dynamic (MD) simulations. Additionally, the method was validated over a set of 7 different proteins using both NMR conformer ensembles and MD ensembles as input. This work, co-authored with Dr. Marina Guenza, was published in the Journal of Chemical Physics in 2015.[2]

Chapter IV addresses the biological relevance of the dynamical mode structure of the model. Additionally, the model is extended to interacting bound protein complexes, and compares and contrasts the dynamical models obtained for the free protein (apo) and bound (holo) protein complexes of the HIV Protease and Insulin

[1]Copperman, J., and Marina G. Guenza. "Coarse-Grained Langevin Equation for Protein Dynamics: Global Anisotropy and a Mode Approach to Local Complexity." *The Journal of Physical Chemistry B* 119.29 (2014): 9195-9211.

[2]Copperman, J., and M. G. Guenza. "Predicting protein dynamics from structural ensembles." *The Journal of Chemical Physics* 143.24 (2015): 243131.

Growth Factor II Receptor. This work, co-authored with Dr. Marina Guenza, is currently in submission.[3]

Surprisingly, the application of the method over many different proteins shows that general scaling relationships with similar dynamical exponents exist for all proteins so far studied. Chapter V focuses upon these scaling relationships and their consequences, as well as the possible origin and relationship to general systems dominated by disorder. This work has not yet been edited as a co-authored publication.

[3]Copperman, J., and Marina G. Guenza. "Mode localization in the cooperative dynamics of protein recognition." *manuscript in submission.*

CHAPTER II

GLOBAL ANISOTROPY AND A MODE APPROACH TO LOCAL COMPLEXITY

The dynamics of proteins, and the related biological function, are inherently determined by the complexity of their energy landscape. As proteins mostly populate the energy states at the minimum of a free energy well, the dynamics in these states mainly consist of local fluctuations around a non-trivial three-dimensional folded structure. This justifies the application of linear Langevin Equations, such as the Rouse-Zimm approach, that were originally developed to describe viscoelastic relaxation in synthetic polymer systems to describe the dynamics of the protein.[6–8] In a previous paper the structure of the Langevin equation was modified to account i) for the specific effective friction of each amino acid, which all have variable shapes, and ii) the dynamics inside the hydrophobic core of the protein, which is screened from the solvent and as such is not affected by the presence of the solvent-mediated hydrodynamic interaction.[8] In this current work the theory, which we name Langevin Equation for Protein Dynamics (LE4PD), develops further focusing in particular on the treatment of anisotropic rotational de-correlation and a mode approach to the local complexity in the folded free energy well. The theory uses short-time Molecular Dynamics (MD) simulations for the input parameters to the analytical solution of the Langevin dynamics, and as a test of the theoretically predicted dynamical properties, i.e. time correlation functions, of the protein. Theoretical predictions are then directly compared to experimental data of Nuclear Magnetic Resonance (NMR) relaxation, providing a test for the relation among the theoretical model, the MD simulation, and the experiments.

The LE4PD approach represents the protein as a collection of interacting units centered on the alpha carbon, and relies on having $N(N-1)/2$ pairwise structural bond correlations (where N is the number of residues of the protein) and N site-specific friction parameters. Bond correlation and friction parameters are extracted from relatively short simulation trajectories of $\sim 10\ ns$ or less, and with them our approach can obtain the equilibrium dynamics of processes such as global tumbling and large-amplitude internal motion, which occur in the same time regime as the simulation or longer.

The approach is based on the fundamental picture of proteins as heterogenous polymers which are collapsed into a definite tridimensional structure, which nevertheless retains some amount of flexibility. As opposed to a rigid body, where the global modes are the only degrees of freedom in the system, protein dynamics include both rotational and internal fluctuation modes. Our description includes internal dissipation due to fluctuations in the hydrophobic region by accounting for an effective protein internal viscosity and considering the relative exposure of each amino acid to the hydrophobic region (see Figure 1). We show that with the correct dissipation, the linear modes of harmonically coupled objects provide a simple but accurate description of the fluctuations of the molecule.

Because fluctuations occur in the energy well, they can be conveniently approximated by harmonic potentials, and the related intramolecular distribution of sites, e.g. α-carbons, can be well approximated by a Gaussian distribution, to at least first order. These harmonic interactions are not restricted to bonded pairs along the backbone, and construct a network of interactions which are non-local along the sequence.

In this way our approach takes advantage of the same physical principles that motivate the Gaussian Network Models (GNM) and the Normal Mode Analysis (NMA),[9–11] however it differs from those models in some important aspects. 1) In Gaussian Network Models,[9–11] the interaction between sites is taken to be a uniform harmonic interaction as long as site separation is within a distance cutoff; this cutoff is typically $\sim .7\ nm$ and the interaction strength is adjustable. In the LE4PD theory all sites are considered to have a pairwise interaction potential whose parameters are defined by the structural correlations obtained from simulations. 2) The GNM and NMA models calculate fluctuations starting from a single equilibrium structure, which is determined experimentally. One important question is how well the experimental starting structure, usually determined from a crystal phase, is representative of the ensemble of protein structures that are present in solution, in physiological conditions.[12] The LE4PD method uses configurations generated from simulations of the protein in aqueous solutions at physiological conditions, which is a more realistic representation of the protein's configurational ensemble.[13] As the LE4PD model aims at building a direct connection between protein structure and their dynamics as measured experimentally for example by NMR relaxation experiments, the representation of the protein from simulations is more realistic.

3) The LE4PD is a diffusive dynamical description with site-specific dissipation, hydrodynamic coupling, and barriers to internal fluctuations, calculated directly from the structural ensemble created by the simulation of the protein in aqueous solvent. The hydrodynamic interaction is key to this description, which is conventionally neglected in the GNM and NMA approaches which describe collective vibrational fluctuations.

In a nutshell, when comparing the LE4PD to GNM-type approaches, the advantage of the latter is their ability to represent qualitatively the type of fluctuations displayed by biomolecules within a formally simple description which requires only limited computational power, while our approach is more computationally intensive. The limitations of GNM-type approaches are related to their difficulty to obtain accurate and quantitative values of the energetics and dynamical properties from their normal modes of fluctuation, as the motion sampled by these methods is confined to the fluctuations around one equilibrium structure determined experimentally and the amplitude of the harmonic fluctuations is fictitious. Furthermore the GNM-type models have limited information of the free-energy barriers along the paths of important fluctuations,[14] and so the kinetics and the time-dependent phenomena are not accurately determined. In the LE4PD approach, the knowledge of the roughness of the free energy landscape, i.e. the sampled energy barriers, provides information on the long-time dynamics. Using the mode formalism presented in this paper, the LE4PD includes a microscopic description of the energetics and barrier crossing present in the molecular dynamic simulation.

In this paper, the Ubiquitin protein is our primary model system, though the approach has been tested on other proteins as well. Ubiquitin is a small well-folded protein whose structure and dynamics have been well characterized utilizing a number of experimental and computational techniques, including NMR backbone relaxation.[1, 15, 16] In the first section of this paper, we will discuss the general properties of the solution of the Langevin equation for biological polymers. Section 2.2 presents the methodology we used to perform Molecular Dynamic simulations of the protein in the canonical ensemble. Treatment of the rotational dynamics to

account for the anisotropic shape of the protein is presented in Section 2.3. A mode-specific dynamical renormalization of the internal modes based on the free energy surface sampled in the simulation is described in Section 2.4. A method to calculate the dynamics of any bond in the protein in terms of the LE4PD solution, necessary to obtain the $N-H$ bond dynamics probed in experimental NMR backbone relaxation, is presented in Section 2.5. A discussion of the newly presented method to calculate long-time protein dynamics from shorter-time simulations concludes the paper.

Theoretical Approach: the Langevin Equation for Protein Dynamics

In the LE4PD equation the dynamics of the protein is described as a diffusive motion across the configurational landscape,[7, 8, 17] consistent with an optimized Rouse-Zimm theory of the dynamics of macromolecules in solution.[6, 18] Proteins are anisotropic in shape and have a hydrophobic core which is only partially exposed to solvent, with this effect depending on the position of each amino acid in the protein. The LE4PD includes both rotational anisotropy and the hydrophobic core, which are features characteristic of biological macromolecules but are uncommon in synthetic polymers in solution. Local energy barriers in the interior of the protein are important to properly define its dynamics and are explicitly taken into account in the LE4PD method.

The Langevin equation formalism is derived starting from the Liouville equation for the conservation of probability density in the phase space of the full atomistic system of the protein and solvent, and using projection operators to obtain an equation of motion for the chosen sites.[19] Here the chosen coarse-grained sites are the α-carbon of each amino acid in the protein primary sequence. A covariance matrix analysis has shown that the C_α positions span the essential fluctuations of proteins.[20] To obtain a linear Langevin equation,[21] we take the coordinates

tracing the backbone configuration of the protein to be complete of the relevant slow configurational degrees of freedom, and neglect system memory. Inertial terms may be discarded as a protein in aqueous solution is safely in the overdamped limit. The intramolecular distribution around the folded state is assumed to be Gaussian, and the parameters in the distribution are directly obtained from the starting configurational ensemble.[8, 22] The coarse-grained LE4PD represents the balance of viscous dissipation with the entropic restoring force and a random Brownian force due to the random collisions of the coarse-grained protein with the fast-moving projected atoms belonging to solvent, ions, and the protein. The time evolution of the coordinate of the coarse-grained site i is well-described by the following equation

$$\overline{\zeta}\frac{\partial \vec{R}_i(t)}{\partial t} = -\frac{3k_BT}{l^2}\sum_{j,k} H_{ij}A_{jk}\vec{R}_k(t) + \vec{F}_i(t) \ , \tag{2.1}$$

where k_B is the Boltzmann constant, T is the temperature, l^2 is the squared bond distance, and $\overline{\zeta}$ is the average monomer friction coefficient, defined as $\overline{\zeta} = N^{-1}\sum_{i=1}^{N}\zeta_i$, with ζ_i the friction of the monomer i. $\vec{F}_i(t)$ is a delta-correlated random force due to projecting the system dynamics onto the coarse-grained sites, where fluctuation-dissipation requires $\langle F_{i\alpha}(t)F_{j\beta}(t')\rangle = 2k_BT\zeta_i\delta(t-t')\delta_{i,j}\delta_{\alpha,\beta}$ where α, β are cartesian indices. Eq. 2.1 is the well-known Rouse-Zimm equation for the dynamics of polymers in solution.[18, 23]

To obtain an effective linear description we assume a well-folded state where site-site correlations are Gaussian in nature. The structural force matrix $\mathbf{A}$ defines the effective mean-force potential, $V(\{\vec{R}\}) = \frac{3k_BT}{2l^2}\sum_{i,j=1}^{N} A_{ij}\vec{R}_i\cdot\vec{R}_j$, which has been successfully adopted in theories of protein folding to describe the final state of the

folding process.[24] The **A** matrix is calculated as

$$\mathbf{A} = \mathbf{M}^T \begin{pmatrix} 0 & \mathbf{0} \\ \mathbf{0} & \mathbf{U} \end{pmatrix} \mathbf{M} \ , \tag{2.2}$$

where **M** is the matrix that defines the center of gyration and the connectivity between sites, $\sum_j M_{ij}\vec{R}_j = \vec{l}_i$. In a protein the α-carbons are connected linearly, so that for $i > 1$ the matrix is defined as $M_{i,i-1} = -1$ and $M_{i,i} = 1$, with $i = 2, ..., N$, while $M_{1,i} = 1/N$ for the first row, and $M_{i,j} = 0$ otherwise. The **U** matrix is the bond correlation matrix with $(\mathbf{U}^{-1})_{ij} = \frac{\langle \vec{l}_i \cdot \vec{l}_j \rangle}{\langle |\vec{l}_i| \rangle \langle |\vec{l}_j| \rangle}$.

The matrix **H** is the hydrodynamic interaction matrix, which describes the interaction between protein sites occurring through the liquid, represented as a continuum medium. While it is standard to utilize hydrodynamical models to obtain the translational and rotational dynamics of proteins,[25] the contribution of hydrodynamical effects to protein internal motion is generally neglected. While this may be justified for very localized motion, in general the non-local hydrodynamic coupling alters the timescale and nature of the large-amplitude highly correlated internal motion and cannot be neglected.[17, 26] To maintain an effective linear description, the hydrodynamic interaction must be preaveraged. While the derivation of the hydrodynamic interaction utilizes the Oseen tensor following the general Rouse-Zimm treatment of polymer chains in dilute solution,[18] other methods such as the Rotne-Prager interaction tensor reduce to the same form upon preaveraging over the equilibrium ensemble.[27] The elements in the matrix of the hydrodynamic interaction are defined as

$$H_{ij} = \frac{\overline{\zeta}}{\zeta_i}\delta_{ij} + (1 - \delta_{ij})\overline{r}^w \langle \frac{1}{r_{ij}} \rangle \ . \tag{2.3}$$

where $\overline{r}^w = N^{-1}\sum_{i=1}^{N} r_i^w$ is the average hydrodynamic radius which is defined below. This is a perturbative hydrodynamic interaction accounting for the nature of the amino acid primary structure as a heteropolymer made up of building blocks of different chemical types, propagating through the aqueous solvent but screened in the dense hydrophobic core. The site-specific friction parameters, ζ_i, are obtained by calculating the solvent-exposed surface area, and calculating the total friction of the i_{th} site via a simple extension of Stoke's law as

$$\zeta_i = 6\pi(\eta_w r_i^w + \eta_p r_i^p) \ . \tag{2.4}$$

Here η_w and r^w denote, respectively, the viscosity of water and the radius of a spherical bead of identical surface area as the solvent-exposed surface area of the residue, the hydrodynamic radius[8], while r^p denotes the hydrodynamic radius related to the surface not exposed to the solvent. The internal viscosity is η_p, which we approximated in our previous work to be related to the water viscosity rescaled by the local energy-barrier scale $\sim k_B T$.[17, 28] The largest possible value of $\overline{r}^w$ that maintains a positive definite solution of the matrix diagonalization is adopted to avoid the well-known issue with the preaveraging of the hydrodynamic interaction in dense systems.[29] For example, in the application of the model to HIV protease, the calculated $\overline{r}^w = 2.28\mathring{A}$ is very close to the adopted value of $\overline{r}^w = 2.23\mathring{A}$, which avoids negative eigenvalues.

Because we focus only on the bond orientational dynamics and not translation, in the interest of a simpler notation we separate out the zeroth order translational mode from the internal dynamics. Following the same notation introduced for the orientational dynamics of star polymers,[30] we define $\mathbf{a}$ as the $\mathbf{M}$ matrix after suppressing the first row used to define the center of mass, and define $\mathbf{L} = \mathbf{aHa}^T$.

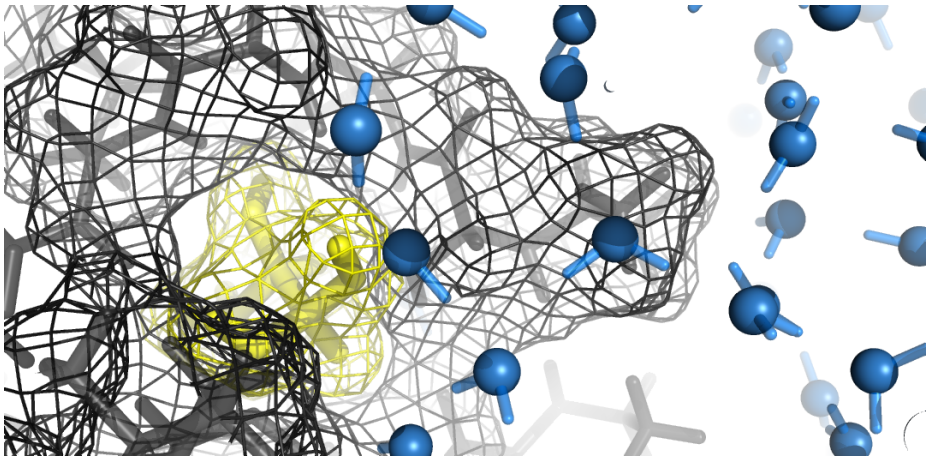

FIGURE 1. Hydrophobic and hydrophilic surfaces of a tagged amino acid. Surface of a tagged residue (Threonine7) in yellow, which is in contact with the internal protein environment (black), and the external solvent environment.

The orientational Langevin equation governing the bond dynamics is

$$\frac{\partial \vec{l}_i(t)}{\partial t} = -\sigma \sum_{j,k} L_{ij} U_{jk} \vec{l}_k(t) + \vec{v}_i(t) \ , \tag{2.5}$$

with $i, j = 1, ..., N-1$, and where $\sigma = 3k_BT/(l^2\overline{\zeta})$, and $\vec{v}_i(t)$ is the random delta-correlated bond velocity.

Eq. 2.5 represents a set of $N-1$ first-order coupled differential equations, which are solved by finding the matrix of eigenvectors $\mathbf{Q}$ which diagonalizes the product of matrices $\mathbf{LU}$. In these diffusive modes we have $N-1$ uncoupled linear equations where each mode is just a sum over the original bond vector basis $\vec{\xi}_a(t) = \sum_i Q_{ai}^{-1} \vec{l}_i(t)$. We define λ_a to be the eigenvalues of $\mathbf{LU}$ with $\sum_{i,j,k} Q_{ai}^{-1} L_{ij} U_{jk} Q_{kb} = \delta_{ab}\lambda_a$, ordered from smallest to largest λ. Like the set of bond vectors $\vec{l}_i(t)$ the set of coordinates $\vec{\xi}_a(t)$ defines the instantaneous conformation of the macromolecule. While the $\mathbf{L}$ and $\mathbf{U}$ matrices are individually symmetric, the $\mathbf{LU}$ matrix is not necessarily symmetric, making the $\mathbf{U}$ matrix only approximately diagonal in the $\mathbf{LU}$ eigenvector basis. The mean squared mode length is then $\langle \xi_a^2 \rangle \equiv \frac{l^2}{\mu_a}$ with μ_a not exactly the eigenvalues of

the bond correlation matrix alone, but defined by the sum $\sum_{i,j} Q_{ai}^{-1} U_{ij}^{-1} Q_{aj}^{-1} \equiv \frac{1}{\mu_a}$. The diffusive mode basis spans the same space as the bond vector basis with near linearity: $\langle \vec{\xi}_a \cdot \vec{\xi}_b \rangle \cong \delta_{ab} l^2 / \mu_a$.

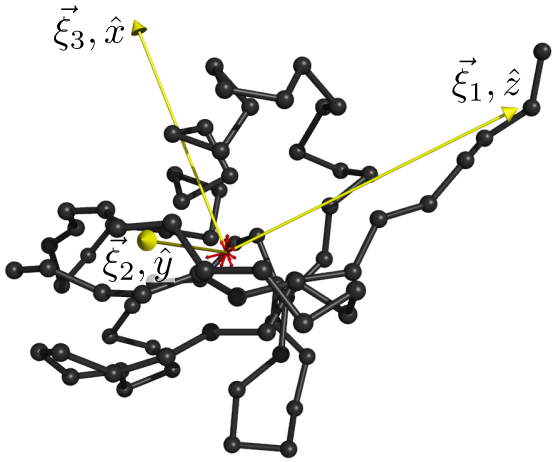

FIGURE 2. Bond basis and mode basis of ubiquitin.
A snapshot of the Ubiquitin protein, represented in the bond basis (black) and in the mode basis (global modes in yellow, internal modes in red).

Like the set of bond vectors $\vec{l}_i(t)$ the set of normal coordinates $\vec{\xi}_a(t)$ define the instantaneous conformation of the macromolecule. Figure 2 shows a snapshot of the protein Ubiquitin in the bond vector and in the mode vector representation, which is obtained by applying the linear transformation of the inverse eigenvector matrix to the coordinates of the snapshot of Ubiquitin. It shows how the first three normal modes, which are the slowest ones, are much larger in magnitude than the internal modes, indicating that the dynamics of this protein, at least in the simulation runs, largely conserves the shape of the molecule, while fluctuations do not involve large conformational transitions or slow cooperative domain motion.

Local Dynamics

The physical quantities of interest in this paper are the bond autocorrelation function and the second order Legendre polynomial of the time dependent bond orientation.

For each bond i along the backbone of the protein, the bond autocorrelation function is defined as

$$M_{1,i}(t) = \frac{\langle \vec{l_i}(t) \cdot \vec{l_i}(0) \rangle}{\langle l_i^2 \rangle} \ , \tag{2.6}$$

and in the formalism of the Langevin equation

$$\begin{aligned} M_{1,i}(t) &= \sum_{a=1}^{N-1} \frac{Q_{ia}^2}{\langle l_i^2 \rangle} \langle \vec{\xi}_a(t) \cdot \vec{\xi}_a(0) \rangle = \sum_{a=1}^{N-1} A_a^i \exp[-\sigma \lambda_a t] \ , \\ &= \sum_{a=1}^{N-1} A_{ia} \exp[-t/\tau_a] \ , \end{aligned} \tag{2.7}$$

with τ_a the correlation time for the ath mode.

Another quantity of interest is the second order Legendre polynomial of the time dependent bond orientation function $P_2(t) = \frac{3}{2}\langle \cos^2[\theta(t)] \rangle - \frac{1}{2}$ which can be related to the first order bond autocorrelation by

$$P_{2,i}(t) = 1 - 3(x^2 - \frac{2}{\pi} x^3 (1 - \frac{2}{\pi} \arctan x)) \ , \tag{2.8}$$

which is a function of $M_{1,i}(t)$ as

$$x = \frac{[1 - M_{1,i}(t)^2]^{\frac{1}{2}}}{M_{1,i}(t)} \ . \tag{2.9}$$

This expression relies on assuming a Gaussian form for the joint probabilities in normal mode coordinates, and it was derived in the paper by Perico and Guenza.[6] The theory accounts for the distribution of effective bond lengths between two sites, here alpha carbons, but does not explicitly account for anisotropy in the global modes. For dipolar relaxation, the Fourier transform of $P_{2,i}(t)$ defines the spectral density

$$J(\omega) = \frac{2}{5}\int_0^\infty P_2(t)\cos[\omega t]dt \tag{2.10}$$

from which spin-lattice (T1) and spin-spin (T2) relaxation times, and nuclear Overhauser effect (NOE) as measured in NMR can be calculated as

$$\frac{1}{T_1} = \frac{d^2}{4}[J(\omega_H - \omega_N) + 3J(\omega_N) + 6J(\omega_H + \omega_N)] + c^2 J(\omega_N)\ , \tag{2.11}$$

$$\frac{1}{T_2} = \frac{d^2}{8}[4J(0) + J(\omega_H - \omega_N) + 3J(\omega_N) + 6J(\omega_H) + 6J(\omega_H + \omega_N)] + \frac{c^2}{6}[3J(\omega_N) + 4J(0)] \tag{2.12}$$

$$NOE = 1 + \frac{d^2}{4}\frac{\gamma_H}{\gamma_N}[6J(\omega_H + \omega_N) - J(\omega_H - \omega_N)]T_1\ , \tag{2.13}$$

where $c = \frac{\omega_N \delta_N}{\sqrt{3}}$, and $d = \frac{\mu_0 h \gamma_H \gamma_N}{8\pi^2 \langle r_{NH}^3 \rangle}$. Here μ_0 is the vacuum permeability, h is Planck's constant, ω_H and ω_N are the ^1H and ^{15}N Larmor frequencies of the experimental field, γ_H and γ_N are their respective gyromagnetic ratios, δ_N is the chemical shift anisotropy of the ^{15}N nucleus, and r_{NH} is the N-H bond length.

Molecular Dynamics Simulations of Ubiquitin

Molecular Dynamic simulations have a double purpose in our studies as they provide the statistical parameters that define the conformational structure of the

protein, which enters the Langevin equation, and second because they provide a test of the time correlation functions predicted with the proposed approach in the time regime covered by the simulations. In the short-time regime sampled by the simulations the time correlation functions predicted by the theory, with input statistics from the simulations, have to be quantitatively consistent with the same functions directly sampled in the simulations.

Molecular Dynamics simulations were performed for the protein Ubiquitin in explicit solvent using the spc/e water model with the addition of the appropriate number of sodium and chloride ions to obtain a neutral system with 45 mM NaCl, identical to the NMR experiments of Lienin et al.[3] Simulations were performed in the canonical ensemble with the temperature set at 300 K. We utilized the AMBER99SB-ILDN[31] atomic force field for proteins starting from structures obtained from the RCSB protein databank (1UBQ[32] crystal structure of Ubiquitin). The GROMACS[33–36] molecular dynamics engine was utilized running multiple nodes (12-64 cores) on the local ACISS cluster at the University of Oregon. The systems were solvated and energy minimized, and then tempered for 50 ps with all bonds constrained, utilizing a 1 fs timestep. The system was then equilibrated for an additional 5 ns with a velocity rescaling thermostat. After this equilibration, a 10 ns run was performed switching to a Nose-Hoover thermostat. Following this, ten configurations separated by 1 ns were randomly chosen and these were used as initial conditions for ten production runs of 10 ns each. This was done to efficiently generate production trajectory, however these ten simulations are clearly not independent since they were not equilibrated independently. In the calculations of the free energy surfaces and the bond correlation matrix $\mathbf{U}^{-1}$, statistics from these ten production runs were used.

Ubiquitin was selected as a test system to analyze the proposed method because is a protein which has been extensively studied both experimentally[1, 15, 16] and by computational methods.[3, 37] While Ubiquitin is a well-folded protein, it also contains domains of both rigid (β-sheets, α-helices) and highly flexible (active loops, C-terminus) secondary structures and as such is an excellent test of the theoretical model.

Global Tumbling Modes and Internal Fluctuations

The dynamics of macromolecules develops in $3N$ dimensional configurational space, which for well-folded proteins largely reduces to fluctuations about an arbitrary three dimensional folded structure. The **LU** matrix, which spans the configurational space of the protein, has three eigenvectors which span the $\mathbb{R}^3$ subspace that contains the average folded structure; we will call these the global modes. For a well-folded globular protein these global eigenvectors are just the ones with the three smallest eigenvalues, λ_a, so they are the slowest modes to relax. However, this is not a strict criteria because sometimes internal modes can be as slow as the global ones. Furthermore this rule will not hold for a highly anisotropic protein for which the values of the three eigenvalues can be very different in magnitude, or for very loosely ordered proteins, for which the scaling of the modes follows more closely that of unstructured synthetic polymers.

The global rotational modes can be unambiguously identified by the fact that they are the only modes with a non-zero time average in the body fixed frame– that is $\langle\vec{\xi}_a\rangle = \vec{0}$ for all internal modes while for the global modes $\langle\vec{\xi}_a\rangle$ is a vector which defines the principal diffusion axes of the protein, as shown in Figure 2. For clarity, Figure 3 shows the average mode length $\langle|\vec{\xi}_a|\rangle$ and the length of the average mode vector $|\langle\vec{\xi}_a\rangle|$ calculated directly from the simulation coordinates of the protein Ubiquitin. The two

quantities are practically identical for the global modes, indicating the stability of the global fold, while the average internal mode vectors all tend to $\vec{0}$ by one or two orders of magnitude. The fact that the average is different from zero indicates that in Ubiquitin there is some, very small, amount of anisotropy in the internal fluctuations, mostly in the first few internal modes. Having identified them by this procedure, the protein's global orientational dynamics is spanned by these three modes of motion, [7] which describe rotations along the three main directions of the global structure.

The infinite-time limit of the time correlation function in the *body-fixed* reference frame gives

$$\lim_{t\to\infty} M_{1,i}(t) = \frac{\langle \vec{l_i} \rangle^2}{\langle \vec{l_i^2} \rangle} \ . \tag{2.14}$$

Expressing $\vec{l_i}$ in its normal mode expansion leads to

$$\frac{\langle \vec{l_i} \rangle^2}{\langle \vec{l_i^2} \rangle} = \frac{\sum_{a,b=1}^{N-1} Q_{ia} \langle \vec{\xi_a} \rangle \cdot Q_{ib} \langle \vec{\xi_b} \rangle}{\langle \vec{l_i^2} \rangle} \ . \tag{2.15}$$

Due to the stationary nature of the global modes as a set of body-fixed axes as shown in Figure 3, $\langle \vec{\xi}_{1,2,3} \rangle^2 \approx \langle (\vec{\xi}_{1,2,3})^2 \rangle$, while due to the isotropic nature of the internal modes $\langle \vec{\xi}_{a>4} \rangle = \vec{0}$. With this property, along with mode orthogonality, in the long time limit Eq. 2.15 reduces to

$$\frac{\langle \vec{l_i} \rangle^2}{\langle \vec{l_i^2} \rangle} \approx \sum_{a=1}^{3} \frac{Q_{ia}^2}{\mu_a} = \sum_{a=1}^{3} A_{ia} \tag{2.16}$$

indicating that the sum of the global mode amplitudes is a local bond order parameter.

As another check that the LE4PD solution has separated into three global modes and $N-4$ internal modes, Figure 3 plots $1/\mu_a$ a dimensionless measure of the

amplitude of mode fluctuations. The LE4PD is solved using as input the full bond correlation matrix **U** calculated from the simulation, and that calculated from the PDB crystal structure assuming the only flexibility to be uniform backbone bond length fluctuations of typical size ($U_{ii} = 1.002$). We have ordered the modes by the size of the eigenvalues of the **LU** matrix, λ_a. Modes $a = 1, 2, 3$ are nearly identical in the two descriptions, reflecting only the small difference in structure between the solvated protein in the simulation and that in the crystal (Figure 5). Modes $a = 4, N-1$ are drastically smaller in amplitude in the LE4PD with only bond length fluctuations, indicating that these are the internal modes and the small perturbation of the bond fluctuations in the model we constructed are not nearly as collective as the larger-amplitude internal motion sampled in the simulation.

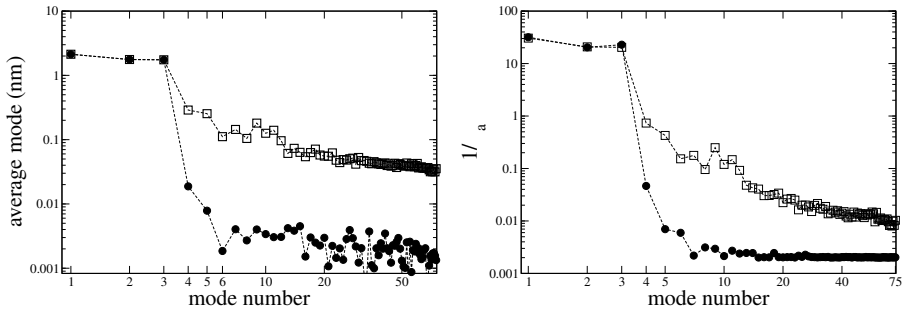

FIGURE 3. Distinguishing global and internal modes.
Left panel: The modulus of the average mode vector $|\langle\vec{\xi}_a\rangle|$ (black diamonds) and the average of the mode modulus $\langle|\vec{\xi}_a|\rangle$ (squares), calculated from the body-fixed simulation coordinates. The two quantities are nearly identical for the global modes, indicating the stability of the global structure, while $|\langle\vec{\xi}_a\rangle|$ is almost zero for all internal modes. Right panel: μ_a^{-1}, inverse eigenvalues of the bond correlation matrix **U** calculated from the PDB crystal structure assuming the only flexibility to be uniform backbone bond length fluctuations of typical size ($U_{ii} = 1.002$) in (black diamonds), and from the simulations (squares). The global modes are nearly identical while internal modes are drastically smaller in amplitude.

These three global modes define the body-fixed coordinate system. To illustrate the relation of the global modes to the structure of the molecule, we calculate the average gyration tensor of the molecule,

$$\langle S_{\alpha\beta}\rangle = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=i+1}^{N}\langle(\vec{R}_j - \vec{R}_i)\cdot\hat{\alpha}(\vec{R}_j - \vec{R}_i)\cdot\hat{\beta}\rangle \tag{2.17}$$

where α, β run over the cartesian indices x, y, z. The eigenvalues of the gyration tensor $\langle {R_g}^2_{x,y,z}\rangle$ are typically used to denote the shape of a macromolecule.[38] Noting that the bead separation vector can be written as $(\vec{R}_j - \vec{R}_i) = \sum_{k=i}^{j-1}\vec{l}_k$, and using the notation $\vec{l}_k\cdot\hat{\alpha} = l_k^\alpha$ to denote the αth cartesian component of the kth bond vector,

$$\langle S_{\alpha\beta}\rangle = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=i+1}^{N}\sum_{k,l=i}^{j-1}\langle l_k^\alpha l_l^\beta\rangle \tag{2.18}$$

and expanding into normal modes

$$\langle S_{\alpha\beta}\rangle = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=i+1}^{N}\sum_{k,l=i}^{j-1}\sum_{a,b=1}^{N-1}Q_{ka}Q_{lb}\langle\xi_a^\alpha\xi_b^\beta\rangle \tag{2.19}$$

where we can now exploit the orthogonality of the mode description.

The internal modes $a = 4, N-1$ in the harmonic theory are isotropic, reducing to $\langle\xi_a^\alpha\xi_a^\beta\rangle = (1/3)(l^2/\mu_a)\delta_{\alpha\beta}$ independent of the orientation of the x, y, z coordinate system. However, the three global modes pick out three orthogonal directions in 3D space and define a natural body-fixed coordinate system where by convention we align the z-axis with the long axis of the molecule corresponding to ξ_1, and the $x-axis$ with the shortest axis corresponding to ξ_3, and the y-axis with

the middle axis corresponding to ξ_2. With this global mode coordinate system, $\langle \xi_a^\alpha \xi_a^\beta \rangle = (l^2/\mu_a)\delta_{\alpha\beta}\delta_{a\alpha}$, and the result

$$\langle {R_g}_\alpha^2 \rangle = \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=i+1}^{N} \sum_{k,l=i}^{j-1} (Q_{k\alpha} Q_{l\alpha} \frac{l^2}{\mu_\alpha} + \frac{1}{3} \sum_{a=4}^{N-1} Q_{ka} Q_{la} \frac{l^2}{\mu_a}) \rangle \tag{2.20}$$

where α corresponds to cartesian directions x, y, z or modes $a = 3, 2, 1$. The distinction between the three global modes and the $N-4$ internal modes is clear; the global modes provide the contribution from the shape of the global folded structure, while the internal modes provide the contributions from the fluctuations. In a random walk model like the freely jointed chain where there is no global structure, the gyration radius contains only contributions from internal degrees of freedom and this expression reduces to the expected $\langle R_g^2 \rangle = Nl^2/6$. In a rigid body model the internal modes would have zero amplitude and the only contributions would be from the global modes. This calculation of the gyration radius of the protein using the LE4PD modes yields a root mean square length of $R_{g_x} = 5.509\mathring{A}, R_{g_y} = 6.126\mathring{A}, R_{g_z} = 8.575\mathring{A}$ and $< 1\%$ error from that directly calculated from the simulation trajectory.

To analyze the effect of the hydrodynamic interaction on the dynamics we report in the right panel of Figure 4 the protein with its global mode axes from the global tumbling modes without hydrodynamic contribution, which correspond to the inertial axes of the protein. The figure shows that as the hydrodynamic interaction becomes more and more relevant, the principal diffusion axes shift, as expected.[39] In addition the right panel of Figure 4 shows how the eigenvalues of $\mathbf{LU}$ change quite dramatically upon inclusion of the hydrodynamic interaction; global and low order internal modes are faster while higher order internal modes are slowed with softer scaling with mode number. This is similar to the effect of including the hydrodynamic interaction

in flexible polymer descriptions (the Rouse-Zimm theory[18]) where the scaling of relaxation rate with mode number goes from $k_a \sim a^2$ to $k_a \sim a^{3/2}$. The hydrodynamic interaction component strongly affects both the global mode relaxation rates, and the orientation of the rotational diffusion axes, and has a relevant contribution to the protein dynamics in general and cannot be discarded.

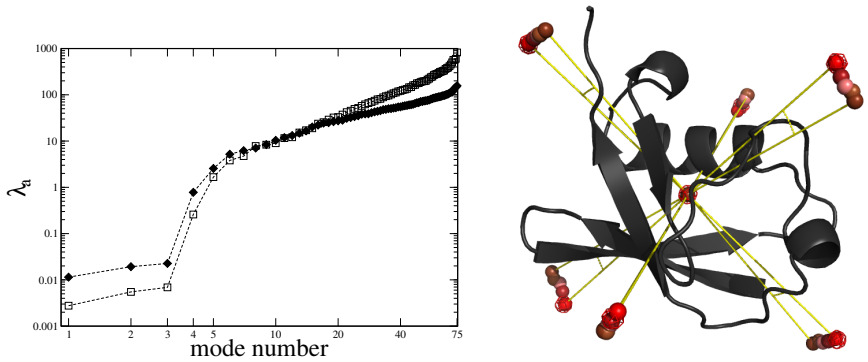

FIGURE 4. Effect of hydrodynamic interaction on eigenvalues.
Left panel: Eigenvalues of the **LU** matrix, which are proportional to the rates of the diffusive processes governing the dynamics of the Ubiquitin protein, without hydrodynamic interaction (squares) and with hydrodynamic interaction (diamonds). Right panel: The red mesh and the red ball correspond to the orientation of the principal axis of the average structure, and the orientation of the first three modes in Eq. 2.5 solution with $H_{i\neq j} = 0$. The first three eigenvectors of the **U** matrix alone correspond to the principal inertial axes of the protein. For simplicity, this model calculation is performed with identical friction coefficients for each amino acid. Using this simplification the hydrodynamic interaction is ramped up by tuning the hydrodynamic radius r^w from 0, i.e. no hydrodynamic interaction, to its maximum possible value yielding positive relaxation times. The rotational axes are modified by the presence of the hydrodynamic interaction.

Eq. 2.5 deals with the flexibility about the folded structure in an ensemble averaged fashion. For well-folded proteins, the LE4PD, with simulations as an input, is a good approximation to determine the diffusion tensor while including the contribution to the dynamics due to the fact that proteins are semiflexible objects.

Most computational approaches to determining the rates of rotational diffusion and the diffusion tensor orientation are designed around the idea of the protein as a rigid-body. While in these models the hydrodynamical treatment is often more sophisticated than in our model,[25] this sophistication comes at the cost of neglecting the contributions from internal degrees of freedom and conformational flexibility to protein dynamics.

Figure 5 shows that the average orientation of the principal diffusion axes calculated from the simulation trajectory of Ubiquitin with the LE4PD theory is very close to that calculated using other methods; the orientation of the most relevant long axis is 10.9 degrees from that found by fitting the NMR relaxation data to an axially symmetric anisotropic model, using the N-H bond orientations from the crystal structure, 6.5 degrees from that found using the HYDRO[40] program with the crystal structure, and 2.8 degrees from the orientation found using the HYDRO program and a set of snapshots taken from a Langevin simulation with the tail unhindered by crystal contacts.[1] Figure 5 shows that the difference in the orientation of the diffusion axes between the LE4PD and other methods appears to be related to the different possible configurations of the tail. Given that the tail is quite mobile the observed difference in the diffusion axes orientation could be just related to a difference in the sampling of the configurations. This suggests that Eq. 2.5 has captured the orientation of the principal diffusion axes of the protein.

We also compare the rates of rotational diffusion calculated using the LE4PD constructed from the crystal structure alone, to that calculated using the HYDRONMR program.[41] The overall rotational diffusion rate in the LE4PD $D_{av} = (1/3)TrD$ is identical to that calculated using the HYDRONMR program when the AER, the adjustable atomic element radius, is set to $a = 1.6$ $\mathring{A}$. Fit

to NMR experiment was found using the HYDRONMR program at $a = 2.2$ $\mathring{A}$;[42] typical values of the AER are much larger ~ 3.1 $\mathring{A}$. Overall anisotropy $2D_z/(D_x+D_y)$ in the HYDRONMR calculation and the LE4PD was comparable at 1.45 and 1.48 respectively. These results suggest that the simplified hydrodynamical treatment of the LE4PD with the inclusion of internal friction sources can lead to similar results for rotational diffusion as other approaches that have a more sophisticate description of hydrodynamics and exclusively consider external solvent friction sources and overlook structural fluctuations.

The differences in the rotational diffusion tensor between that found when fitting to NMR relaxation data, computational rigid-body hydrodynamical modeling, and the LE4PD, can be attributed to three main sources; namely i) the hydrodynamical treatment and its effect on the large-scale dynamics, ii) the presence of flexibility, i.e. differences between the solvated structural ensemble in solution and the crystal structure iii) the sources of dissipation considered, in our model both solvent and internal friction. The qualitative agreement between different methods observed for Ubiquitin implies that these effects in this protein are not large. Though Ubiquitin has a highly flexible C-terminal tail, there is no large-scale domain reorientation, so that the static structure of the protein is close enough to the equilibrated structure in solution and is a good representation of the equilibrium structural ensemble. In general, however, for other proteins this is not true, and we argue that starting from a static picture of a protein to calculate the diffusion tensor, can have some degree of limitation in providing an accurate representation of the rotational dynamics for the reasons just discussed.

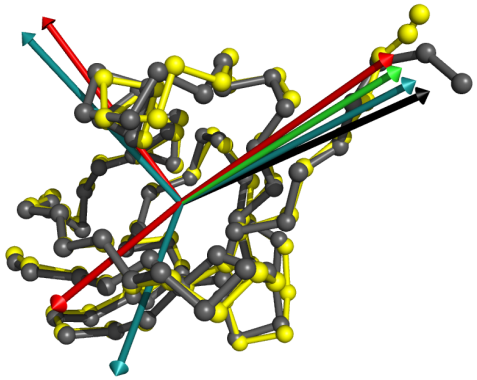

FIGURE 5. Diffusion tensor orientations of ubiquitin.
A comparison of the Ubiquitin crystal structure (gray), the ensemble averaged structure from the simulations (yellow), the orientation of the long axis of the anisotropic diffusion tensor fit to NMR experiment (black),[1] HYDRO calculation from a set of snapshots taken during a Langevin simulation allowing the tail to fluctuate (green), HYDRO calculation from the crystal structure (blue), and the average global tumbling mode orientation from the theory (red). The long-axis orientation is very close to other hydrodynamical modeling treatments when tail flexibility is taken into account, although the short axes orientation is not.

Rotational dynamics in an inertial frame

The mode solution of the LE4PD Eq. 2.5 correctly describes the dynamics of the molecule in a body-fixed coordinate system attached to the molecule. However, when calculating time correlation functions in an inertial (lab) frame of reference care must be taken to deal with the inherent anisotropy of the global tumbling modes. The internal modes, approximated here to be completely isotropic, are not affected by the transformation to an inertial coordinate system, so the first step is to isolate the global tumbling modes in Eq. 2.5. The global modes are an orthogonal set of vectors defining

a set of body-fixed coordinates, and aligning the body-fixed coordinate system

$$\begin{pmatrix} \hat{x'} \\ \hat{y'} \\ \hat{z'} \end{pmatrix} = \begin{pmatrix} \hat{\xi'}_3 \\ \hat{\xi'}_2 \\ \hat{\xi'}_1 \end{pmatrix} , \tag{2.21}$$

we obtain an orientational diffusion equation in the body-fixed system

$$\frac{\partial}{\partial t} \begin{pmatrix} \vec{\xi'}_3(t) \\ \vec{\xi'}_2(t) \\ \vec{\xi'}_1(t) \end{pmatrix} = -\sigma \begin{pmatrix} \lambda_3 & 0 & 0 \\ 0 & \lambda_2 & 0 \\ 0 & 0 & \lambda_1 \end{pmatrix} \begin{pmatrix} \vec{\xi'}_3(t) \\ \vec{\xi'}_2(t) \\ \vec{\xi'}_1(t) \end{pmatrix} + \begin{pmatrix} \vec{v'}_3(t) \\ \vec{v'}_2(t) \\ \vec{v'}_1(t) \end{pmatrix} .$$

The global modes define the orientation of the protein which in the lab-fixed frame of reference is rotating; denoting vectors in this inertial frame as unprimed $\vec{\xi}_3, \vec{\xi}_2, \vec{\xi}_1$, at $t = 0$ we align the x, y, z lab system with the $\vec{\xi'}_3, \vec{\xi'}_2, \vec{\xi'}_1$ body-axis of the protein. For $t \neq 0$, the body-fixed and lab axes are related by

$$\begin{pmatrix} \vec{\xi'}_3(t) \\ \vec{\xi'}_2(t) \\ \vec{\xi'}_1(t) \end{pmatrix} = \hat{\Re}(t) \begin{pmatrix} \vec{\xi}_3(t) \\ \vec{\xi}_2(t) \\ \vec{\xi}_1(t) \end{pmatrix} \tag{2.22}$$

where $\hat{\Re}(t)$ is the time-dependent rotation which takes the protein into its actual orientation, which is in general not aligned with xyz lab system.

In the inertial frame, the LE4PD is now

$$\frac{\partial}{\partial t} \hat{\Re}(t) \begin{pmatrix} \vec{\xi}_3(t) \\ \vec{\xi}_2(t) \\ \vec{\xi}_1(t) \end{pmatrix} = -\sigma \begin{pmatrix} \lambda_3 & 0 & 0 \\ 0 & \lambda_2 & 0 \\ 0 & 0 & \lambda_1 \end{pmatrix} \hat{\Re}(t) \begin{pmatrix} \vec{\xi}_3(t) \\ \vec{\xi}_2(t) \\ \vec{\xi}_1(t) \end{pmatrix} + \hat{\Re}(t) \begin{pmatrix} \vec{v}_3(t) \\ \vec{v}_2(t) \\ \vec{v}_1(t) \end{pmatrix} ,$$

and left multiplying by the inverse rotation which takes the lab coordinate system into the body-fixed coordinates,

$$\frac{\partial}{\partial t}\begin{pmatrix}\vec{\xi}_3(t)\\ \vec{\xi}_2(t)\\ \vec{\xi}_1(t)\end{pmatrix} = -\sigma\hat{\Re}^{-1}(t)\begin{pmatrix}\lambda_3 & 0 & 0\\ 0 & \lambda_2 & 0\\ 0 & 0 & \lambda_1\end{pmatrix}\hat{\Re}(t)\begin{pmatrix}\vec{\xi}_3(t)\\ \vec{\xi}_2(t)\\ \vec{\xi}_1(t)\end{pmatrix} + \begin{pmatrix}\vec{v}_3(t)\\ \vec{v}_2(t)\\ \vec{v}_1(t)\end{pmatrix} ,$$

where $\Re(0) = \hat{1}$ and identifying the operator

$$\hat{D}(0) = \begin{pmatrix}\lambda_3 & 0 & 0\\ 0 & \lambda_2 & 0\\ 0 & 0 & \lambda_1\end{pmatrix} , \tag{2.23}$$

and

$$\hat{D}(t) = \hat{\Re}^{-1}(t)\begin{pmatrix}\lambda_3 & 0 & 0\\ 0 & \lambda_2 & 0\\ 0 & 0 & \lambda_1\end{pmatrix}\hat{\Re}(t) . \tag{2.24}$$

The time dependent $\hat{D}(t)$ can be redefined by assuming that for short time intervals the rotation of the protein takes place in a series of small angular displacements. We note here that $(\sigma/2)\hat{D}(0)$ is the diagonalized rotational diffusion tensor. This procedure yields a time-independent operator $\hat{D}_{\Re}$ valid for rotational diffusion. While rigid-body rotational diffusion is well known,[43] we present a formal derivation for the rotational diffusion of a semiflexible, fluctuating macromolecule.

The rotation can be expressed as three consecutive rotations around the lab-fixed axes $\hat{\Re}(t) = \hat{\Re}_z(\gamma(t))\hat{\Re}_y(\beta(t))\hat{\Re}_x(\alpha(t))$. Considering small timesteps Δt we assume the angles $\alpha(t), \beta(t), \gamma(t)$ to be proportional to Δt, which gives to lowest order in Δt,

$$\hat{D}(\Delta t) = \hat{\Re}_x^{-1}(\Delta t)\hat{H}(0)\hat{\Re}_x(\Delta t) + \hat{\Re}_y^{-1}(\Delta t)\hat{H}(0)\hat{\Re}_y(\Delta t) + \hat{\Re}_z^{-1}(\Delta t)\hat{H}(0)\hat{\Re}_z(\Delta t) \quad (2.25)$$

The rotations, to linear order in Δt, are $\hat{\Re}_{x,y,z}(\Delta t) = \hat{1} + (\Delta t)\hat{L}_{x,y,z}$ where

$$\hat{L}_x = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & -1 \\ 0 & 1 & 0 \end{pmatrix}, \hat{L}_y = \begin{pmatrix} 0 & 0 & -1 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix}, \hat{L}_z = \begin{pmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} , \quad (2.26)$$

are angular momentum operators. Since we are dealing with a diffusive process, we can take the limit of infinitely small time steps

$$\lim_{\Delta t \to 0} \frac{1}{\Delta t}(\hat{\Re}_{x,y,z}(\Delta t) - \hat{\Re}_{x,y,z}(0)) = \hat{L}_{x,y,z} \ . \quad (2.27)$$

In this limit, we obtain for the time-independent operator

$$\hat{D}_{\Re} = \frac{1}{2}(\hat{L}_x^T \hat{D}(0)\hat{L}_x + \hat{L}_y^T \hat{D}(0)\hat{L}_y + \hat{L}_z^T \hat{D}(0)\hat{L}_z) \ , \quad (2.28)$$

where the factor of $\frac{1}{2}$ comes from requiring that $Tr(\hat{D}(0)) = Tr(\hat{D}_{\Re})$. Performing the matrix multiplication, the final LE4PD for the global modes in the lab-fixed inertial

coordinate system is

$$\frac{\partial}{\partial t}\begin{pmatrix}\vec{\xi}_3(t)\\ \vec{\xi}_2(t)\\ \vec{\xi}_1(t)\end{pmatrix} = -\sigma\begin{pmatrix}\frac{1}{2}(\lambda_3+\lambda_2) & 0 & 0\\ 0 & \frac{1}{2}(\lambda_3+\lambda_1) & 0\\ 0 & 0 & \frac{1}{2}(\lambda_2+\lambda_1)\end{pmatrix}\begin{pmatrix}\vec{\xi}_3(t)\\ \vec{\xi}_2(t)\\ \vec{\xi}_1(t)\end{pmatrix} + \begin{pmatrix}\vec{v}_3(t)\\ \vec{v}_2(t)\\ \vec{v}_1(t)\end{pmatrix} \quad (2.29)$$

In the LE4PD, the $a = 4, N-1$ internal modes are treated to be isotropic and are unaffected by this change from a body-fixed to in an inertial frame of reference, which is translating with the center of mass of the protein. This is an approximation, and for a more flexible protein with large-amplitude highly anisotropic fluctuations, the internal modes may need to be treated in a similar fashion.

The Rouse-Zimm model, Eq. 2.5, is typically applied to systems that are either statistically spherically symmetric (a flexible freely-jointed chain, or a semi-flexible freely-rotating chain) or possess a single long-axis (a rod). In these cases there is only one global tumbling mode, D_{LE} is just a number, not a tensor, and the rotational diffusion equation leaves the single rotational de-correlation rate unaltered. However, for proteins which have an arbitrary anisotropic folded structure, the full rotational diffusion equation of an arbitrary 3-dimensional body must be solved.

Rotational Diffusion and Solvent Viscosity in Simulations

Extracting rotational dynamics directly from a simulation is problematic due to the computational cost of long simulation runs exceeding the relaxation time of the global rotation by one or two orders of magnitude. Furthermore current water models have been seen to lead to inaccurate viscosity and surface hydration.[44] The viscosity of the spc/e water model has been estimated to be the same as real water[45] or about 27% lower.[46] After utilizing the proposed LE4PD formalism, the rates of rotational

de-correlation are found to be consistent with both the simulation and experiment. Figure 6 displays a comparison between $M_1(t)$ calculated directly from the simulation and from theory where internal barrier crossings are included as discussed in Section 2.4. The figure shows data for six selected bonds where the rotational diffusion eigenvalues so calculated improve the comparison to the simulation, which is true for 71% of the 75 bonds in Ubiquitin.

Quantitative results are obtained for the rotational dynamics in the simulation when the viscosity in the LE4PD is set to 1.33 times that of real water. For internal processes the relaxation is less sensitive to the exact viscosity used in the LE4PD because the internal timescales are dominated by the eigenvalue spectra of the **LU** matrix and by the energy barriers to relaxation. However at longer times the rotational decorrelation dominates the dynamics, making the long-time slope of the bond autocorrelation highly sensitive to the viscosity used as input to the theory. The statistical error in the in the longer-time regime of the time-correlation functions calculated from the 10 ns simulations makes the analysis of the reasons for discrepancy between spc/e water viscosity estimates and the viscosity used in the LE4PD not particularly conclusive. In particular, Wong and Case estimated the statistical error in an exponential rotational relaxation process at $t =\sim \tau$ to be about 23% when calculated from a 200 ns MD trajectory.[44] The relationship between the diffusive dynamics of proteins, solvent viscosity, internal viscosity, and hydration layer properties of the in silico protein in water model certainly warrants further study and comparison with much longer simulation runs where the simulation time exceeds the timescale of global diffusive processes by 1-2 orders of magnitude. Studies of diffusive motions of proteins in these time regimes and the tuning of water models for better viscosity have recently been undertaken,[47] but the intent of this work is to

demonstrate the utility of the LE4PD to obtain the protein dynamics from relatively short simulation runs. For comparison to experiment the correct viscosity of 10% D_2O and 90% H_2O at the experimental temperature is used, while for comparison to simulation results, the viscosity of 1.33 times that of real water at $300K$ is used in the LE4PD.

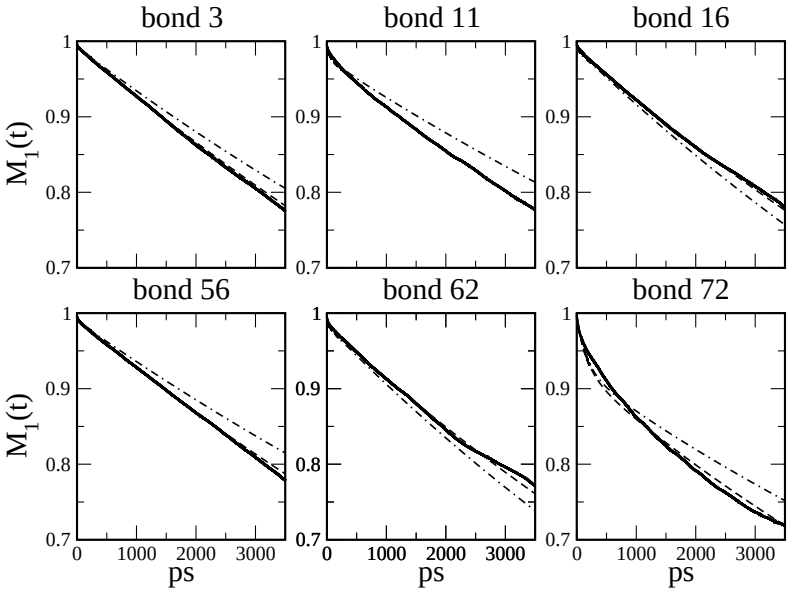

FIGURE 6. Bond autocorrelation with rotational diffusion.
$M_1(t)$ calculated from the simulation (black), LE4PD with $l = 1$ rotational diffusion eigenvalues (dashed line), and the bare LE4PD solution in the body-fixed frame (dashed-dotted line). The LE4PD approch includes the calculation of the internal mode energy barriers as described in Section 2.4.

Dynamical renormalization of the internal modes from free energy surfaces

In this section we describe how we account for local-mode barrier crossing in the Langevin equation for protein dynamics. In our previous paper this correction

was not included in the theory,[8] but we noticed how the disagreement between theoretical predictions and the simulations, from which the theoretical parameters were calculated, for the function $M_{1,i}(t)$ was largely a consequence of neglecting the local energy barriers.

Here, we explicitly account for the complexity of the free energy landscape, working with the internal mode description of the linearized Langevin equation and correcting the modes *a posteriori* by introducing the local energy barriers, as measured in simulations. We are interested in calculating the bond autocorrelation function where we separate the first three non-translational modes from the internal ones as

$$M_{1,i}(t) = \sum_{a=1}^{3} A_{ia} \exp[-\sigma\lambda_a t] + \sum_{a=4}^{N-1} A_{ia} \exp[-\sigma\lambda_a t] \ , \tag{2.30}$$

under the assumption that the first three modes represent the rotational dynamics of the molecule, while the higher modes represent the internal dynamics. This separation of local and global motion in $M_1(t)$ is due entirely to the mode structure of the solution of Eq. 2.5, and doesn't rely in any way upon the separation of the dynamics into slow and fast processes. In fact we will show that upon including internal energy barriers in the mode coordinates some internal modes can become slow enough to occur on the same timescale than the rotational decorrelation modes. It is important to note that the lack of time-scale separation between the first three modes and the other ones does not affect the validity of the treatment that is presented here. Furthermore, this separation of local and global motion occurs in the context of a solution for a first order correlation, while in higher order correlation functions, such as $P_{2,i}(t)$, global and local modes are necessarily mixed.

While the first non-zero three modes of relaxation in a largely stiff object mainly describe the rotational dynamics of the macromolecule, the modes with higher index, $a = 4, ..., N-1$, describes the internal dynamics of the protein and the collective breathing modes of an overdamped, harmonically connected object. The more cooperative the motion, the slower the characteristic timescale of the normal mode. When applied to linear synthetic polymers in solvent, which are uniform in their composition and are rotationally symmetric in their structure, the time of relaxation scales in the Rouse-Zimm theory with mode number a as $\tau_a \sim a^{-3/2}$, with $\tau_a = (\sigma\lambda_a)^{-1}$.[18] Representing the same function for a protein, for example Ubiquitin, shows a different and more interesting behavior. Figure 7 compares the Rouse-Zimm scaling law with the data from the simulations of Ubiquitin before and after accounting for the local energy barriers in the internal normal modes. Without energy barriers, the global modes are clearly slower than the internal ones. The inclusion of the local energy barriers in the local modes leads to a more complex mode-dependence of the relaxation times. The first handful of internal modes display relaxation times comparable to the rotational modes: those are the modes which span the most collective and large-amplitude fluctuations of the molecule. In general the relaxation time still decreases with mode number and, interestingly, the slowest more-cooperative internal modes also have the highest free energy barriers. This result is conceptually consistent with the behavior emerging in the first few diffusion coordinates in the LSDmap method[48] where eigenfunctions of the Fokker-Planck operator are estimated.

Each normal mode obtained from the diagonalization of Eq. 2.5 is a vector defined by the linear combination of the bond vectors weighted by the eigenvectors of the product of matrices $\mathbf{LU}$, as $\vec{\xi}_a(t) = \sum_i Q^{-1}_{ai}\vec{l}_i(t)$. In polar coordinates the

vector is represented as $\vec{\xi}_a(t) = \{|\vec{\xi}_a(t)|, \theta_a(t), \phi_a(t)\}$. In Eq. 2.5 the orientation of each Langevin internal mode diffuses isotropically about all solid angles; however, the probability distribution of the internal mode vector orientation per solid angle calculated from the simulation trajectory is unique for each mode.

As a first approximation we assume that the fluctuations in the modulus of the normal mode vector are random and independent or mode orientation, so that the relevant changes in the normal mode free energy occur as the angles, expressed in the spherical coordinates, span the configurational space. For a generic normal mode, a, the free energy surface is defined as a function of the spherical coordinate angles θ_a and ϕ_a as

$$F(\theta_a, \phi_a) = -k_B T \log \{P(\theta_a, \phi_a)\} \ , \tag{2.31}$$

with $P(\theta_a, \phi_a)$ the probability of finding the normal mode vector having the given value of the solid angle.

The internal normal modes provide information about the complexity of the configurational free energy landscape around the folded state. By calculating the mode-dependent free-energy surface from the simulation probability distribution we observe that even close to the global folded minima the free energy landscape is rough and contains multiple local minima, metastable states, and local barriers with related multiple possible pathways for the local dynamics (see for example Figure 8).

For an approximate rescaling of the dynamics, we correct the rate of reorientational diffusion of each mode by assuming thermal activation over the mode-dependent energy barrier $\langle E^{\dagger} \rangle$ where

$$\langle k_a \rangle = k_a^0 \ exp[-\langle E_a^{\dagger} \rangle / (k_B T)] \ . \tag{2.32}$$

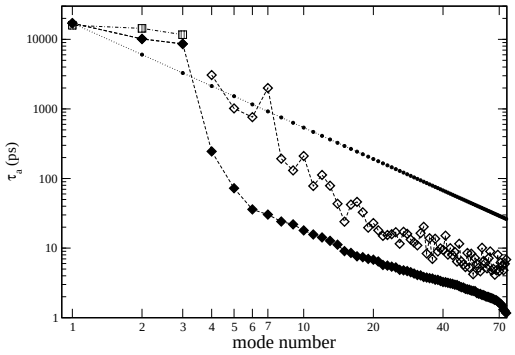

FIGURE 7. Mode-dependent relaxation times.
Mode-dependent relaxation time for the protein Ubiquitin. In the bare LE solution (filled diamonds) the rotational modes are well separated from the internal modes, while once the dynamics are modified to account for the local free-energy barriers the internal modes (open diamonds) become closer to the global modes. The three global modes after accounting for anisotropic rotational diffusion become more uniform in timescale (squares). Also reported is the standard Rouse-Zimm model scaling $a^{-3/2}$ of the normal mode dynamics (circles). Even for Ubiquitin, a well-folded protein, there is no separation in timescale between global and internal modes after dynamical renormalization.

This simple dynamical renormalization provides an average correction to the dynamics of the LE, which approximately accounts for the local barrier crossing, and is in agreement with free energy landscape theories suggesting an underlying dynamical glass transition at low enough temperature[49, 50]. Other more detailed methods can be used to model barrier crossing, such as Markov network models[51] and approaches which explicitly deal with the complexity of the free energy landscape[48, 52–54]. As a first approximation, the depth of the minimum free-energy well in the mode serves as the relevant barrier to transport, and by requiring thermal activation over this barrier the timescale of all mode-dependent dynamical quantities are simply renormalized by these factors. This simple model gives a realistic first order correction to the barrier-free Langevin dynamics.

The presence of internal free energy barriers is not limited to the first internal modes. Figure 11 shows that the complexity in the free energy surface is present in any normal mode, extending out to fluctuations on mode $17th$, which is an intermediate mode, to mode $75th$, which is the most local and the last of the internal modes. In this way the renormalization of the dynamics has to be included for each internal normal mode of motion.

Protein real-space configurations from the normal mode pathways

In simulations each time frame corresponds to a specific molecular configuration. Each molecular configuration corresponding to a given simulation frame at time t can be described by the set of bond vectors $\vec{l}_1(t), \vec{l}_2(t), ...\vec{l}_{N-1}(t)$ or equivalently by the set of mode vectors $\vec{\xi}_1(t), \vec{\xi}_2(t), ...\vec{\xi}_{N-1}(t)$. However, a single mode vector orientation is associated with an ensemble of molecular configurations. Given that we are interested in the explicit representation of the structure along a pathway of interest for each normal mode, all structures from the simulation ensemble which pertain to a particular θ, ϕ orientation of a mode vector of interest are extracted and averaged. By calculating this average structure along a particular pathway in the free energy surface of a particular mode, for example along the path of lowest free energy connecting two minima, the structural pathway of relaxation in that mode can be obtained. This is a very different process than projecting the protein structures along particular directions, often used for example to visualize eigenvector directions in Essential Dynamics Analysis[55]. Figure 8 displays the free energy surface and structural pathway of the first of the internal modes of Ubiquitin. Because the modes are in general ordered from the most collective to the most local, this is the most collective internal motion for Ubiquitin. Using this process of transforming the protein

coordinates from the simulation into the mode representation, and then extracting back the average structure for each specific mode orientation, is possible to reconstruct the ensemble of protein configurations along a structural pathway for each mode.

The transform shows that the first internal mode mainly involves large-amplitude fluctuations of the C-terminal tail, which are biologically important as the C-tail is the linkage point for poly-ubiquitination[2]. Figure 9 shows the free energy surface of the fourth internal mode which captures concerted fluctuations of the tail and the loop containing Lys11, a relatively abundant linkage site involved in cell-cycle regulation[2]. The most collective internal modes are those which span the important functional motion of the protein, illustrating the fundamental relationship between protein dynamics and biological function[56].

Coarse-gained representation, dynamics, and internal energy rescaling

In the independent normal modes representation, the LE4PD theory renormalizes the rough free energy surface measured in simulations and replaces it with a smooth isotropic surface upon which dynamical processes become faster than in the real system[57]. From the solution of Eq. 2.5, reorientational diffusion takes place at a rate $k_a^0 = \sigma\lambda_a$, while the real free energy surface of the normal mode, as it emerges from the molecular dynamic simulations, presents complex roughness. Dynamical processes need to include activate dynamics to overcome internal energy barriers[58]. To find this average activation barrier $\langle E_a^\dagger \rangle$, we calculate the depth of the well in which the minimum free energy state lies, for each normal mode. Figures 10 shows, as an example, the average free-energy barriers in the first and fourth internal mode of Ubiquitin. The difference in energy is taken between the deepest value of the energy

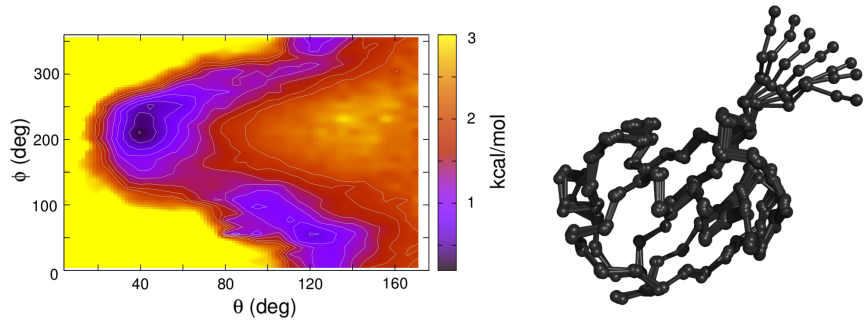

FIGURE 8. Free energy surface of the first internal mode.
Left panel: free energy surface of the first internal mode. Right panel: structural fluctuations in the first internal mode along the path of minimum energy.

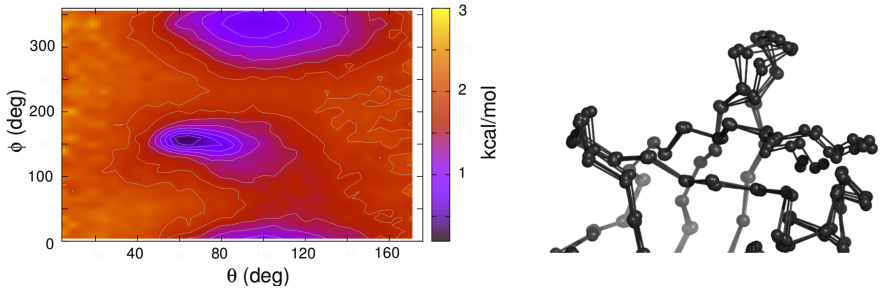

FIGURE 9. Free energy surface of the fourth internal mode.
Left panel: free energy surface of the fourth internal mode. Right panel: structural fluctuations in the fourth internal mode, involving the loop containing Lys11, a relatively abundant linkage site involved in cell-cycle regulation[2]

in the energy well and the barrier that the system needs to overcome to escape the same.

The height of the energy barrier is mode-dependent (see Figure 12). Complex energy landscape is observed for any mode and at all length scales, but large energy barriers, present in the first internal modes, converge to smaller barriers for more local modes, and finally to a minimum value given by $\langle E^{\dagger}_{int} \rangle \sim k_B T$. Assuming that the slowing down of the dynamics in the hydrophobic region is largely dominated by

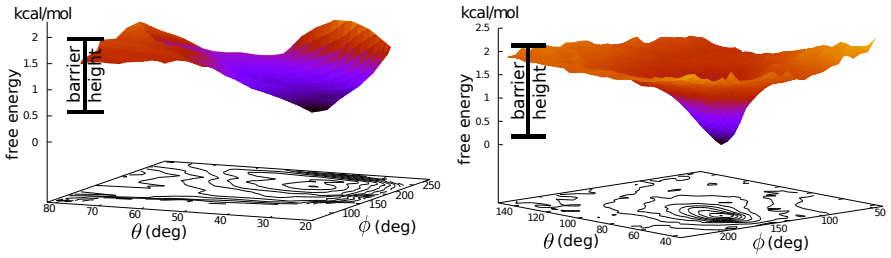

FIGURE 10. The free-energy barrier to mode fluctuations.
The depth of the free energy well is used as $E\dagger$, the free-energy barrier to mode fluctuations. Left panel: first internal mode of Ubiquitin. Right panel: fourth internal mode of Ubiquitin.

local barrier-crossing we evaluated the protein internal viscosity as

$$\eta_p = \eta_s e^{\frac{\langle E^{\dagger}_{int} \rangle}{k_B T}} \ , \tag{2.33}$$

where the external solvent viscosity is taken as the reference point in the calculation.[28] The solvent accessible surface area calculation shows that all protein residues are at least partially exposed to the solvent– which justifies the use of the solvent viscosity as the valid reference point, as all protein sites are at least partially solvated, and all dissipation must ultimately lead to dissipation into the solvent to maintain equilibrium. Because the protein's free-energy landscape is in general rough and funnel-shaped, containing barriers of all heights[49], it seems typical that at any finite temperature the most highly sampled and relevant local internal barriers should be of the order of $\sim k_B T$. This simple estimate of the protein internal viscosity in the hydrophobic core is used as an input in Eq. 2.4. At this point all the information needed to solve the equation of motion for protein dynamics is defined.

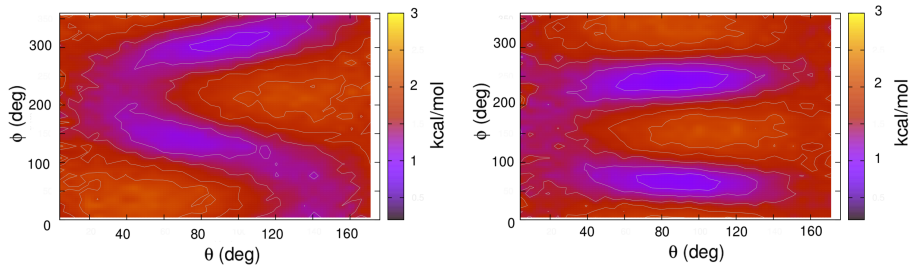

FIGURE 11. Free energy surface of local internal modes.
Free energy surface of the 17th internal mode (left panel) and of the 75th internal mode (right panel). Both at intermediate and local lengthscales there is still complexity in the internal mode FES.

Testing the free energy surface approximation against simulations

As a first test of the proposed approach we compare the decay of the autocorrelation function of a C_α-C_α bond, $M_{1,i}(t)$ defined in Eq. 2.30, with the data from simulations. The theoretical function is calculated from the solution of the Langevin Equation for Protein Dynamics, with and without taking into account the barrier crossing of the internal normal modes. Figure 13 reports as an example the bond autocorrelation for bond seven: bond seven lies in a particularly active loop with large barriers to fluctuations; the inclusion of the renormalization of the local modes dynamics, due to the internal energy barrier, drastically improves the agreement with simulations. Further comparison of the LE4PD theory predictions for $M_1(t)$, with the renormalization from the energy barriers in the internal modes, can be seen in Figures 6 and 16.

Figure 15 displays similar agreement between theory and simulations for bonds belonging to α-helices, β-sheets, other loops, and the flexible C-terminal tail, where rotational contributions are removed from both the simulation and the theoretical time correlation functions to highlight the internal dynamics.

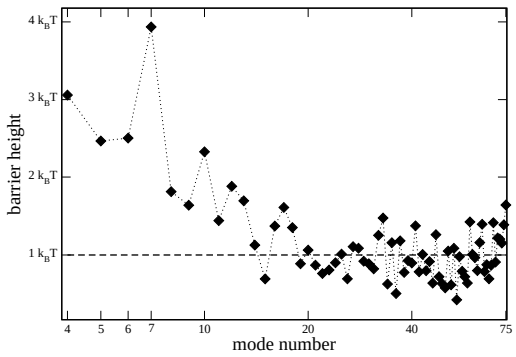

FIGURE 12. Height of the free-energy barrier as a function of mode number. Height of the local-mode free-energy barrier to conformational diffusion as a function of mode number. The first handful of internal modes display large energy barriers, while the average energy barrier for the higher order internal modes goes to $\sim k_B T$.

Dynamics of the N-H dipole vector

When comparing to experiments probing dipolar relaxation, it is important to account for the fact that the experimental probe either refers to the dipole along the N-H or to the one along the C-H bond. The orientation and dynamics of both these probes is not necessarily equivalent to the dynamics of the C_α-C_α bond vector, described by Eq. 2.5. The left panel of Figure 14 shows, as an example, the relative orientation of the N-H dipole and the C_α-C_α bond vector in one sample configuration of the protein Ubiquitin from simulations. The N-H dipole, relevant to 15N NMR backbone relaxation experiments of Ubiquitin in this paper, has a different orientation than the related C_α-C_α bond vector. The right panel of the same figure represents a model for the relative orientation of the instantaneous vectors, $\vec{l}_{NH,i}(t)$ and $\vec{l}_{C_\alpha,i}(t)$, and the averaged vectors, $\langle\vec{l}_{NH,i}\rangle$ and $\langle\vec{l}_{C_\alpha,i}\rangle$. In addition to a difference in average orientation, the C_α-C_α vector and the N-H dipole vector can have different contributions to their internal dynamics, both additional short-time

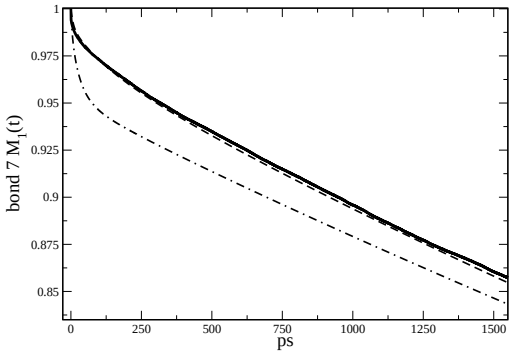

FIGURE 13. Bond auto-correlation function in a highly active loop.
Bond auto-correlation function $M_1(t)$, with both rotational and internal motion, for C_α-C_α bond 7 located in a highly active loop. Data calculated directly from the simulations (solid line), from Eq. 2.5 with dynamical renormalization of internal modes (dotted line), and without renormalization of the internal modes (dashed line).

librational processes, and additional slower processes where the N-H bond can rotate with the peptide plane without affecting the overall backbone orientation.

The orientational bond autocorrelation for N-H bond relaxation

$$M_{1,NH}(t) = \frac{\langle \vec{l}_{NH}(t) \cdot \vec{l}_{NH}(0) \rangle}{\langle (l_{NH})^2 \rangle} \ , \tag{2.34}$$

is calculated from the LE4PD theory by assuming that the modes form a complete set, which is equivalent to say that we can express the N-H bond vector as an expansion in the LE4PD normal modes as

$$\vec{l}_{NH,i}(t) = \sum_{a=1}^{N-1} Q_{NH,ia} \vec{\xi}_a(t) \ , \tag{2.35}$$

where the transformation matrix $\mathbf{Q}_{NH}$ is defined as

$$Q_{NH,ia} = \langle \vec{l}_{NH,i} \cdot \vec{\xi}_a \rangle / \langle \xi_a^2 \rangle \ . \tag{2.36}$$

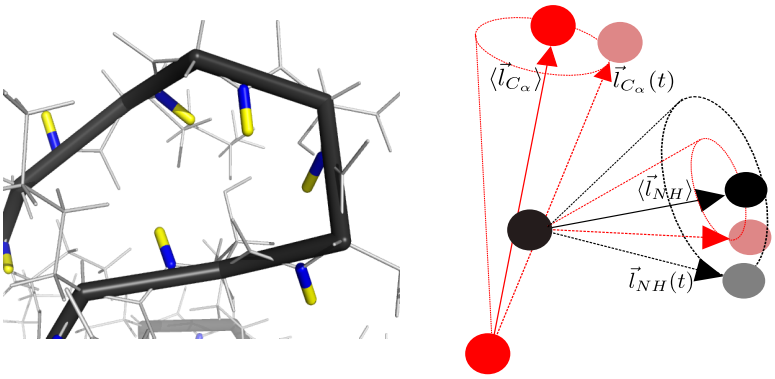

FIGURE 14. Relative position for the N-H and the C_α-C_α bonds. Left panel: the orientation of the bond basis (C_α-C_α) is in black and the N-H vector measured in 15N NMR is blue-yellow. Right panel: model representation of the relative position and orientation of the instantaneous and averaged bond vectors for the N-H and the C_α-C_α bonds.

The bond autocorrelation function is then

$$M_{1,NH,i}(t) = \sum_{a=1}^{N-1} A_{NH,ia} \exp[-\sigma \lambda_a t] \ , \tag{2.37}$$

with

$$A_{NH,ia} = \frac{Q_{NH,ia}^2 \langle \xi_a^2 \rangle}{l_{NH}^2} \ . \tag{2.38}$$

The expansion coefficient $Q_{NH,ia}$ can be calculated directly from the simulation using Eq. 2.36, in this fashion it is possible to obtain the dynamics of any protein-based vector from the LE4PD modes.

For accuracy, when calculating the normal mode expansion of the N-H bond vector from the simulation coordinates, the local bond order parameter $\sum_{a=1}^{3} A_{NH,ia}$ is enforced to be the same as $\langle \vec{l}_{NH,i} \rangle^2 / \langle \vec{l}_{NH,i}^2 \rangle$ calculated directly from the simulation. This normalization is performed prior to enforcing the overall

normalization $\sum_{a=1}^{N-1} A_{NH,ia} = 1$. This is only a small correction of a few percent for all bonds except for those located in the C-terminal tail bonds 72-75. Here, the mode expansion fails and $\sum_{a=1}^{N-1} A_{NH,ia}$ calculated from the simulation before normalization is much less than 1. For these highly flexible C-terminal tail bonds, the C_α-C_α bond dynamics are obtained well by the LE4PD theory but the predictions for the N-H bond dynamics are much too fast, showing that there are large amplitude slow processes which are uncorrelated with protein backbone fluctuations in these bonds.

Comparing the N-H bond and the C_α-C_α dynamics to simulation

As a test of the normal mode solution of the Langevin Equation for Protein Dynamics to describe the dynamics of both the $N-H$ bond vector and the C_α-C_α, we compare the decay of the first order autocorrelation function with data from computer simulations. Calculations were performed for each bond, and as an example Figure 15 displays the comparison between theoretical predictions and simulations for the autocorrelation functions of both the N-H bond and the C_α-C_α bond related to a representative bond in all secondary structure types. To emphasize the difference in the internal dynamics, in this figure overall rotation is removed from simulation before calculating the time correlation functions, and the global mode contributions are removed from the theoretical calculation. Figure 16 compares the results of the full theory with rotation included for all of the poly-ubiquitination linkage sites, showing the high variability in the local dynamics at these biologically important residues. The comparison shows that the agreement is quantitative; similar quality of agreement is observed for all bonds along the primary sequence, up to the anomalous relaxation of the N-H bond vectors in some of the tail residues, supporting the validity of the model proposed in this paper. The theoretical model just presented proves to

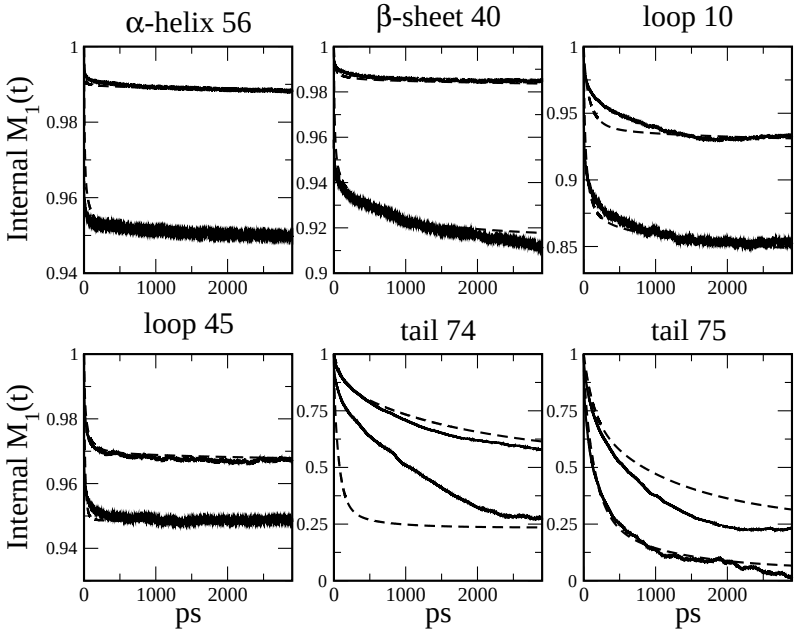

FIGURE 15. Bond autocorrelation functions due to internal processes. Temporal decay of the bond autocorrelation functions due to internal processes, where rotational relaxations have been subtracted, for the C_α-C_α vector (top) and for the N-H vector (bottom) for bonds in many secondary structure types along the protein primary sequence. The dotted lines are calculated from the LE4PD normal mode expansion, solid lines are calculated directly from the simulation.

be fully consistent with the simulations, which provide the input quantities to the theory, in the short time regime where the simulations efficiently sample the local configurational space.

Comparing the LE4PD predictions to NMR experiment

As a second step in our study we use the theoretical predictions for $P_{2,i}(t) = \frac{1}{2}(3\cos^2\theta_i(t) - 1)$, obtained from $M_{1,i}(t)$ using Eq. 2.8, to calculate T1 and T2 relaxation times, and NOE, which are measured experimentally. 15N NMR backbone relaxation experiments are very sensitive to the site-specific dynamics in the

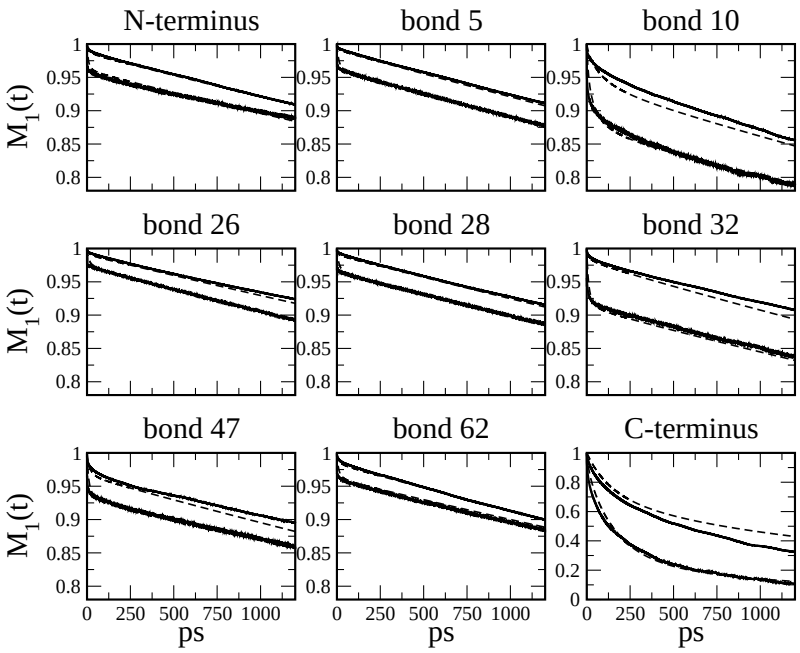

FIGURE 16. Bond autocorrelation at poly-ubiquitination linkages. Temporal decay of the bond autocorrelation functions due to internal and rotational processes, for the C_α-C_α vector (top) and for the N-H vector (bottom) for all poly-ubiquitination linkage sites. The dotted lines are calculated from the LE4PD normal mode expansion, solid lines are calculated directly from the simulation.

picosecond to the nanosecond regimes[59]. The target function is the bond relaxation for the N-H vector, which is calculated using information from the simulations. It is known that the force-fields used to calculate the N-H bond interaction in the atomistic simulation are approximate, because they are not designed for high accuracy at the single-hydrogen level. In addition, the LE4PD modes span the configurational space of the protein backbone but are not necessarily complete for the local N-H bond dynamics. To overcome these issues we present three different methods to estimate the internal N-H bond dynamics. In all three methods, the global mode contributions $A_{NH,ia}$ for $a = 1, 2, 3$ are calculated by evaluating Eq. 2.38 from the simulation, while the internal dynamics and the overall normalization between global and internal modes differ.

Each method is tested against three different sets of NMR experiments, those of Lee and Wand[15] at 298K, and those of Tjandra et al.[1] and Lienin et al.[3] at 300K. The solvent viscosity used in the theoretical expression is adjusted to account for the mixture of 90% H_2O and 10% D_2O used in all experiments[44]. The simulation conditions (temperature and salt concentration) were set to match the experiments of Lienin et al., and only the temperature and the temperature dependence of the viscosity in the theoretical expression was set to match the different experiments. The experimental data are for the most part self-consistent, and the theory agrees quite well with all experiments, with correlation coefficients between .792 and .983, depending on the method used to describe the internal dynamics of the N-H bond, as can be seen in the Table 1. Figure 17 shows the comparison between the theoretical results and the experimental data of T1, T2, and NOE from Lienin et al.[3] The different theoretical models proposed here provide comparable agreement with the experiments.

In the first method, the internal dynamics are taken to be identical to that of the C_α-C_α vector, with the addition of a trivial contribution from the fast independent librational motion of the N-H bond. This librational motion is taken to be a sub-picosecond process with identical magnitude for all bonds. From NMR experiments in oriented media, Ottiger and Bax estimated an average order parameter S for independent librational motion of the N-H bond from the backbone basis to be $S = .94$[60], corresponding to an approximate amplitude of the independent librational process in $M_{1,NH}(t)$ of $A_{librational} = .02$. This gives the expression for the relaxation of the ith N-H bond

$$M_{1,NH,i}(t) = C \sum_{a=1}^{3} A_{NH,ia} \exp[-\sigma\lambda_a t] + \sum_{a=4}^{N-1} A_{C\alpha,ia} \exp[-\sigma\lambda_a t] + A_{librational} \exp[-\frac{t}{\tau_{lib}}] \quad (2.39)$$

where $\tau_{lib} = .2\ ps$. The results are practically independent of τ_{lib} as long as it is taken to be a fast process. To enforce the correct normalization, the constant $C = (\sum_{a=1}^{3} A_{C\alpha,ia} - A_{librational}) / \sum_{a=1}^{3} A_{NH,ia}$ so that $M_{1,NH,i}(0) = 1$. As can be seen in the Table 1, the α-carbon based internal dynamics of the backbone (model 1) correlates highly with the N-H bond dynamics measured experimentally for all three different sets of data.

The second method allows for an arbitrary independent relaxational mode for each N-H bond that is directly fit to the simulation, that is

$$M_{1,NH,i}(t) = C \sum_{a=1}^{3} A_{NH,ia} \exp[-\sigma\lambda_a t] + \sum_{a=4}^{N-1} A_{C\alpha,ia} \exp[-\sigma\lambda_a t] + A_{indNH}(t)] \ , \quad (2.40)$$

where $A_{indNH}(t)$ is a three-exponential fit to simulations of the difference in the internal dynamics of the N-H bond and the C_α-C_α vector,

$$A_{indNH}(t) = \frac{\langle \vec{l}_{NH}(t) \cdot \vec{l}_{NH}(0) \rangle}{\langle (l_{NH})^2 \rangle} - \frac{\langle \vec{l}_{C\alpha}(t) \cdot \vec{l}_{C\alpha}(0) \rangle}{\langle (l_{C\alpha})^2 \rangle} \ , \tag{2.41}$$

evaluated in the body-fixed frame. Normalization is set by $C = (\sum_{a=1}^{3} A_{C\alpha,ia} - A_{indNH}(0))/\sum_{a=1}^{3} A_{NH,ia}$. Table 1 shows that accounting for the difference between the local N-H bond dynamics and the backbone leads to lower error and higher correlation to experiments than the first method. The dynamics of the N-H bonds calculated with the second method for the highly flexible C-terminal tail agrees well with simulations, however it disagrees with the experimental data for the tail-bond 75. This suggests the simulations predict dynamics that are too fast when compared with the experimental findings indicating that some slow energy barrier is not accounted for in the simulation.

In the third method the full LE4PD mode expansion is used, $M_{1,NH,i}(t) = C\sum_{a=1}^{3} A_{NH,ia} \exp[-\sigma\lambda_a t]$ with no additional independent processes added. Despite the errors in the C-terminal tail bonds, the overall agreement and correlation to experiment is nearly identical to the second method in which the difference in the internal dynamics between the α-carbon basis and N-H bond is fitted (see the Table and Figure 17). Only the second method requires any fitting to a time-dependent quantity, and, as can be seen in the Table, this improves the agreement with experiment only modestly, if at all.

It is important to point out that the good agreement between theory and experiment has been obtained without the need of fitting the theory to experimental data, and in this way our approach differs from other approaches commonly used to interpret NMR relaxation.[61] When a model is parametrized by fitting to the

TABLE 1. Correlation coefficients with experimental data of NMR relaxation.

Correlation coefficient of the theoretical models with experimental data of NMR relaxation, using the three different methods to estimate the independent N-H bond fluctuations, (1) assuming identical internal dynamics to the α-carbon bond basis up to an independent librational process, (2) fitting the independent dynamics of the N-H bond to the simulation, and (3) using the full LE4PD mode expansion of the N-H bond vector.

NMR reference and theoretical model	T1 Corr.	T2 Corr.	NOE Corr.	RMS Rel. Error
Lee[15] et al. 1	.792	.885	.937	7.7%
Lee[15] et al. 2	.875	.974	.967	5.9%
Lee[15] et al. 3	.880	.944	.940	14.8%
Tjandra[1] et al. 1	.806	.879	.948	13.7%
Tjandra[1] et al. 2	.890	.971	.978	12.1%
Tjandra[1] et al. 3	.914	.938	.955	20.6%
Lienen[3] et al. 1	.959	.936	.967	7.0%
Lienin[3] et al. 2	.960	.960	.970	6.6%
Lienin[3] et al. 3	.880	.983	.982	16.6%

experimental data it carries their uncertainty and errors, and it cannot directly relate the measured data to actual structural relaxation processes. In this respect our approach has a clear advantage. The level of agreement between theoretical prediction, simulation, and experiment suggests that the LE4PD approach models the protein backbone dynamics with accuracy, while the disagreement observed for specific bonds can be related to insufficient sampling of the free energy landscape that enters the relaxation dynamics of NMR, or possible experimental errors. The LE4PD approach has the advantage of being firmly grounded in the underlying physical processes which relax the dipole orientation.

Conclusions

We have presented a coarse-grained description for the dynamics of a folded protein in aqueous solution. The theory is closely related to the Rouse-Zimm model of the dynamics of synthetic, unfolded, polymers in dilute solutions but carefully

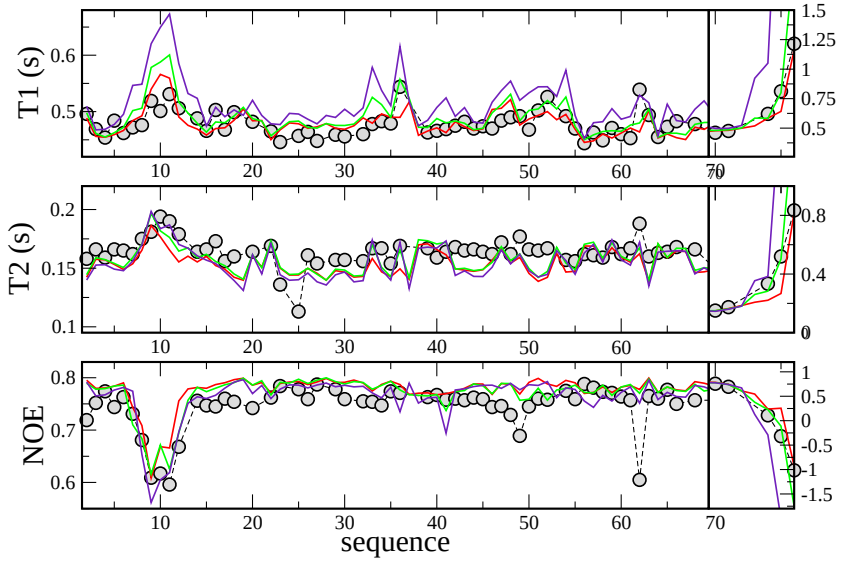

FIGURE 17. T1, T2, and NOE 15N NMR backbone relaxation.
T1, T2, and NOE 15N NMR backbone relaxation, comparison between experiment (black)[3], the first theoretical method (red) correcting for N-H orientation and only adding an identical librational process onto the $C-\alpha$ internal motion, the second theoretical method (green) accounting for differences in internal N-H fluctuations by fitting indendent N-H relaxation to simulation, and the third theoretical method with the full LE4PD expansion of the N-H bond vector (purple).

incorporates the complexity of the protein free energy lansdscape into a linear Langevin description of the protein dynamics. The theory is conveniently expressed in normal modes of motion. Because the modes are linearly independent, this facilitates the inclusion of correction terms that are specific of the dynamics spanned by each single mode.

The theory, which we call the Langevin equation for protein dynamics (LE4PD), modifies the Rouse-Zimm equation for both the global and internal modes. Since proteins have a specific tridimensional structure which is anisotropic, global modes must describe the fully anisotropic rotational diffusion of the molecule. The three global modes, properly modified, are able to capture the fully anisotropic rotational dynamics of the folded structure. The mode framework also allows for the calculation of the dynamics of bonds in the protein structure other than the α-carbon bond basis chosen for the coarse-grained description. This is especially important when comparing to experiments which probe the dynamics of specific dipoles measured in experimental techniques, here the N-H dipole whose relaxation is measured in 15N NMR backbone relaxation.

To obtain a formalism that is solvable analytically, our treatment of the hydrodynamics involves preaveraging; in principle other bead or shell hydrodynamical models may be more accurate in this regard[25]. However, while the treatment of the hydrodynamics can be more exact for a completely rigid protein, our approach captures correctly the essential physics of a collapsed polymer system with internal fluctuations. Typical synthetic polymers have an isotropic shape and all the monomers are statistically equally exposed to the solvent. This is not true for proteins; residues that are inside the protein, in the hydrophobic "core", have very little contact with the solvent, however they experience friction due to interactions with

other protein residues. The physics of the tumbling of a rigid body, no matter how detailed the parameterization and hydrodynamical treatment, incorrectly attributes all frictional sources to solvent contacts. This would imply that the sites belonging to the hydrophobic core of the protein have little or no friction, whch is unphysical. Furthermore, for the system of a protein in solvent there is no conservation law mandating that the friction associated with global diffusive processes be due only to direct solvent contacts. And it has been observed that the fit to experimental data in bead and shell hydrodynamical models for rotational diffusion always requires extra friction. We argue here that while the conventional inclusion of an adjustable bound layer of water and ions leads to quantitative comparison between rigid body hydrodynamical calculations and experiment,[25] the neglect of internal friction may be an additional factor not accounted for in the hydrodynamic modeling of biological polymers as rigid objects.

The dynamics of the internal modes predicted by the Rouse-Zimm equation for synthetic polymers is unrealistically fast due to neglecting the complex nature of the internal free energy landscape. We have presented here a new model of a Langevin Equation for Protein Dynamics, which includes a first-order correction where the timescale of relaxation in each internal mode is rescaled by the mode-specific mean free-energy barrier to orientational diffusion. After rescaling we observe that there is no longer a separation in timescale between internal and global processes even in the well-folded Ubiquitin protein. Accounting for global anisotropy and the complexity of the internal free energy landscape leads to simultaneous quantitative agreement with simulation and NMR backbone relaxation rates.

The Langevin Equation for Protein Dynamics is based upon the inherent nature of proteins as polymers with both flexibility and global structure; it is a modified

Langevin approach for polymer dynamics. The approach seamlessly describes both internal and rotational fluctuations, as well as dissipation in the internal hydrophobic core of the protein and the external solvent environment. With the ease of use of current simulation packages and the availability of computational power, when starting structures are available, the LE4PD approach can help bridging the gap between the often short timescales of simulations and longer timescales probed by experiments, providing a direct formal connection between the protein's primary sequence, three dimensional structure, free energy landscape, and the dynamics.

CHAPTER III

PREDICTING PROTEIN DYNAMICS FROM PROTEIN STRUCTURAL ENSEMBLES

The evolved amino acid sequence of a native protein encodes its folded structure and inherent dynamical properties in aqueous solution.[56, 62, 63] The latter determines the dynamics of specific residues in a protein primary sequence, which are active participants in the pathways of the biological function. Biologically active segments are often mobile and adaptable to assume a proper configuration when binding to a reaction partner. The multiple configurational states that an active segment may populate are not randomly selected: configurations with minimal energy are connected by energy barriers, and as such are thermally activated, enabling emerging regions of high mobility, which can behave like "switches" along the binding pathway.[63]

Different experiments and computational models exist to probe the dynamical processes of proteins, spanning the femtosecond regime of bond and angle vibrational modes to the millisecond and longer time regimes of folding and enzymatic kinetics. Important information in the picosecond to tens of nanosecond regime can be collected through NMR relaxation experiments, such as T_1, T_2 and NOE, however their interpretation is model dependent. Atomistic Molecular Dynamic (MD) simulations can provide a realistic dynamical model, but for most proteins of interest sufficient sampling to obtain converged dynamical correlations is prohibitively costly, and a theoretical approach is needed.

The theory we present here is the Langevin Equation for Protein Dynamics (LE4PD), which provides a coarse-grained but still physically realistic representation

of biological macromolecules at the lengthscale of a single amino acid and larger. The LE4PD theory describes the amino acid dynamics quantitatively, as the theory contains information about the extent of the intramolecular energy barriers, specific amino acid friction coefficient, semiflexibility, degree of hydrophobicity, as well as hydrodynamics. The LE4PD accurately predicts the sequence-dependent dynamics starting from the ensemble of metastable structural configurations around the folded state measured by NMR, or from MD simulations.

The LE4PD model is unique in that it is a minimal dynamical model which projects the local and global diffusive dynamics of proteins from the protein structural ensemble with no adjustable parameters. This is possible because it is a coarse-grained yet microscopic model whose parameters are set directly from the microscopic physical system. This is in contrast to most methods constructed to model protein dynamics which rely upon site-specific adjustable parameters, such as the model-free formalism of Lipari and Szabo.[61] Other methods attempt to define the internal diffusion of proteins as fractional Brownian processes,[64] which is a more accurate description of the general nature of the internal motion of proteins but is not predictive in nature. Nodet, Abergel, and Bodenhausen have modeled the dynamics of proteins as a coupled network of rotators under the assumption of a single conformational minima and small displacements.[65] This approach, which attempts to predict fluctuations and dynamics from a single protein structure, is not directly comparable to the LE4PD model we present here, where we model the dynamics and take the structural ensemble from experiment or by sampling an underlying atomistic model via MD simulation. Like other elastic network models[9–11, 66, 67] the coupled rotator model is capable of capturing the local variation in flexibility along the protein chain with no site-specific adjustable parameters, but because it begins from an empirical network description it

requires a large amount of parameterization and specification of an overall rotational diffusion time τ_0, a scaling factor k_0, a cut-off distance R_c, and a characteristic internal diffusion time t_D. The model is explicitly limited to small displacements around a single conformational minima, and relaxation times centered upon a short characteristic internal diffusion time of $\sim 300ps$. In contrast, the LE4PD model is capable of simultaneously describing the global rotational diffusion, as well as local motion spanning the picosecond to many nanosecond and microsecond regimes. In particular, the long-time, highly correlated, large-amplitude dynamical motion of proteins is of great biological interest.

Input to the LE4PD is an ensemble of structural configurations, which has to be representative of the distribution of folded states of the protein. While proteins sample a very large $3N$-dimensional configurational space, with N the number of independent sites comprising the protein, at the bottom of the funnel-like energy landscape the conformational diversity is much smaller.[49, 50, 68] A common paradigm is that the important internal fluctuations of a folded protein span a limited number of specific structures,[69, 70] and these can be well sampled experimentally by NMR.[12] If that is the case, NMR conformer ensembles should provide a structural ensemble consistent with well-sampled MD simulations, and the LE4PD coupled with structural NMR should provide predictions of the protein dynamics without need of performing lengthy computer simulations. In practice, NMR solution structures encode a structural diversity that is due to a combination of thermal fluctuations and a possible lack of complete experimental information. The LE4PD method provides the ability to test the capability of an input structural ensemble to produce experimentally determined dynamical measurements such as site-specific NMR relaxation.

The diffusive mode solution of the LE4PD organizes the configurational landscape, defining fluctuations on a set of well-defined length and timescales encompassing the relative motion between neighboring α-carbons to the global rotations of the structure as a whole.[8, 17] In the diffusive mode description the LE4PD identifies the regions of local flexibility and cooperative motion of the residues inside a protein. As an example we project the MD trajectory onto the diffusive modes of the HIV protease monomer, and obtain a free energy landscape barrier height distribution which scales with mode cooperativity. Using the scaling form for this barrier height distribution, which appears to be a general feature of protein dynamics, leads to accurate dynamical timescales in the simulation-free conformer-based LE4PD model.

Mode-based descriptions are extremely useful in computational approaches to protein dynamics.[20, 55] Analysis of the free energy landscape in covariance modes have been used to describe the folding of small proteins.[71] The covariance matrix of the spatial functions of the nuclear spin interactions from MD simulation have been used to calculate NMR relaxation, as fit to the trajectory correlation times and experimental values.[72] The characteristic difference between the LE4PD approach and these other approaches is that we study the modes of an appropriate equation of motion, and as such are associated directly with the timescale and pathway of a quasi-independent structural relaxation process. Other mode-based approaches are based upon studying the abstract covariance modes of a set of variables, and as such any time-dependence in these modes comes purely from a fit to the simulation trajectory.

The dynamical predictions of the LE4PD model starting from an ensemble of structures generated from experimentally determined NMR conformers are compared with a second ensemble of structures generated in the course of an MD simulation in

the timescale of $50 - 150$ nanoseconds. To validate the accuracy of the theoretical predictions of the dynamics using the LE4PD approach we test its predictions against experimental data of NMR relaxation for seven different proteins and 1876 site-specific NMR relaxation measurements. Using either the MD generated or NMR solution structure ensembles we obtain quantitatively self-consistent predictions, with similar overall correlation of $\rho_{MD} = .95$ and $\rho_{NMR} = .93$. We find that, in general, the MD-generated ensembles provide through the LE4PD a closer agreement with experimental data than the LE4PD informed by NMR ensembles, with 17% lower relative error.

Structural ensembles of proteins

The LE4PD model predicts the dynamics of the protein using the structural ensemble as input. By generating a structural ensemble through relatively short time ($\sim 10\ ns$) MD simulations, the needed input for the LE4PD was evaluated leading to accurate predictions for the global and site-specific dynamics of Ubiquitin[17] and the signal transduction protein CheY.[8] The accuracy of the simulations, however, depends upon the accuracy of the force-field used, and sampling the full configurational space can become computationally expensive depending upon the size of the protein and the extent of configurational rearrangements.

Building statistical ensembles from metastable configurations

We take as an ansatz that the configurational space of a folded protein is spanned by limited number of conformational states, and that these conformational states are known *a priori*. As an alternative procedure to performing MD simulations we assume as starting configurational ensembles the conformers that were measured

experimentally by NMR. The extent to which NMR solution structure conformers represent important metastable states of the protein, as opposed to uncertainty due to incomplete experimental information, is controversial and varies between different NMR structures.[73] It is certainly clear that NMR structural ensembles do encode some measure of the conformational variability of the protein, as NMR structural ensembles have been shown to correlate highly with structural ensembles generated by MD simulation,[74] and have been used to gain valuable insight into protein flexibility in computational studies of ligand binding.[75] In our model we investigate the assumption that all conformers represent metastable protein configurations which contribute equally to the full ensemble, and use the resulting dynamical predictions to evaluate the ability of the input structural ensemble to span the experimentally observed dynamics.

Fluctuations around the local conformational states are imposed by applying a Gaussian Network Model (GNM).[9] While many elastic network models of varying complexity are in routine use, the differences in the predicted local flexibilities are usually small and affect only the short-time dynamics in the picosecond regime. The GNM builds a harmonic network of interactions around each residue based on a distance cutoff criteria, and solves the resulting site-site fluctuations as a linear matrix equation. GNM models have been shown to reproduce well crystalline state fluctuations measured as Debye-Waller Temperature factors (B-factors) and thus are a good representation of the short-time fluctuations while they require minimal computational effort. Once combined with the LE4PD the theory provides a realistic and computationally inexpensive prediction of the dynamics of proteins on a wide range of time scales, from the local fluctuations to the large, concerted, conformational transitions.

From the GNM we define the bond correlation matrix locally around the αth conformer $U_{\alpha,ij}$. The GNM defines the pairwise fluctuations $\langle \Delta\vec{R}_i \cdot \Delta\vec{R}_j \rangle = \frac{3k_BT}{\gamma}\Gamma^{-1}_{ij}$ where $\mathbf{\Gamma}$ is the Kirchoff adjacency matrix defined using a cutoff radius of $7.0\mathring{A}$[9, 10] and γ is the harmonic interaction strength. We found that in general a value of $\gamma = 0.06\frac{kcal}{\mathring{molA}^2}$ is needed to match the short-time $1 - 10\ ps$ orientational fluctuations of the protein from the MD simulation.

An interaction strength of $\sim 1\frac{kcal}{\mathring{molA}^2}$ is typically used with the GNM to predict crystallographic B-factors; this order of magnitude difference in interaction strength may be due to the local anharmonic softening of the orientational potential energy surface due to the aqueous solvent.[76] The boundary water layer of hydrated proteins in aqueous solution is highly mobile in the picosecond regime[77]; the constant shifting of the protein-water hydrogen bonds may lead to enhanced orientational fluctuations which are completely local in nature. This effect is absent in the crystalline state, where the hydration water is much more static.

Recognizing that in the body-fixed reference frame $\vec{l}_i(t) - \langle\vec{l}_i\rangle = [\vec{R}_{i+1}(t) - \vec{R}_i(t)] - [\langle\vec{R}_{i+1}\rangle - \langle\vec{R}_i\rangle]$ we can determine the local bond correlation matrix around each conformer as

$$(\mathbf{U})^{-1}_{\alpha,ij} = \frac{1}{\langle|\vec{l}_i|\rangle\langle|\vec{l}_j|\rangle}\left[\langle\vec{l}_i\rangle \cdot \langle\vec{l}_j\rangle + \frac{3k_BT}{\gamma} (\Gamma^{-1}_{ij} + \Gamma^{-1}_{i+1,j+1} - \Gamma^{-1}_{i,j+1} - \Gamma^{-1}_{i+1,j})\right]. \quad (3.1)$$

The total $\mathbf{U}$ matrix is then simply the average $U_{jk} = \frac{1}{N_c}\sum_{\alpha=1}^{N_c} U_{\alpha,jk}$ with N_c the number of conformers in the NMR structural ensemble. Similarly, the hydrodynamic matrix $\mathbf{H}$, the site friction coefficient ζ_i, and all other input quantities to the LE4PD, are calculated separately for each conformer and then the statistical average

is taken over all the conformers. This is an extremely simplistic picture of the structural ensemble of a protein; however the dynamics predicted by this structural ensemble generated by the set of NMR conformers is consistent in many ways with the much more detailed ensemble generated through the sophisticated process of explicit solvent MD simulation. The set of NMR conformers provides us an ensemble of important metastable structural minima in the free energy landscape; and the GNM provides fluctuations around these minima. Molecular anisotropy, rotational diffusion, hydrodynamic interactions, and local energy barriers are included through the LE4PD.

Building statistical ensembles from Molecular Dynamics simulations

Because both the determination of NMR conformers and the experimental measurements of NMR relaxation are affected by errors, we performed as a further test MD simulations of the same systems to evaluate the quality of agreement of the LE4PD starting from the NMR and the MD conformers. Simulations were performed in explicit solvent using the spc/e water model. We utilized the AMBER99SB-ILDN[31] atomic force field for proteins and the GROMACS[33–36] molecular dynamics engine was utilized on the TRESTLES supercomputer at San Diego.[78] All system conditions, e.g. temperature and salt concentration, were set to reproduce the experimental conditions. The systems were solvated and energy minimized, and then underwent a 500 *ps* tempering and equilibration routine including pressure coupling. The production simulations were performed in the canonical ensemble, using a velocity rescaling thermostat.[79]

For the PR95 protease monomer, simulations were performed starting from each of the twenty conformers in the NMR structure, resulting in a set of twenty

production ensembles used as input to the LE4PD, with averages taken over all twenty results. The same set of production trajectories for Ubiquitin were used from our previous work.[17] For the remaining five proteins, the first conformer was chosen as the starting structure, and only one simulation was performed for each protein. Each simulation had 50 ns of production. For each trajectory the root mean square deviation (RMSD) was calculated and statistics were only collected in the equilibrated sections of the trajectory. The trajectories were also required to contain only reversible transitions, as monitored by the RMSD. The simulation time effectively used in the LE4PD for each trajectory ranged from 10 to 30 ns. The simulations performed and the protein databank structures used are summarized in Table 2.

TABLE 2. Systems and Structural Ensembles

Protein	MD Sim.	Starting Struct.	Temp.	NMR + GNM
Protease	20 x 50 ns	1Q9P (1-20)	293K	1Q9P (1-20)
1GF2R	160 ns	2M6T (1)	273K	2M6T (1-20)
N-TIMP-1	50 ns	1D2B (1)	293K	1D2B (1-30)
S836	50 ns	2JUA (1)	298K	2JUA (1-20)
CPB1	50 ns	1MX7 (1)	298K	1MX7 (1-22)
KAPP	50 ns	1MZK (1)	298K	1MZK (1-30)
Ubiquitin	10 x 10ns	1UBQ (1)	300K	1XQQ (1-128)

The configurational ensembles that emerge from the NMR ensembles and from the MD simulations are reported in Figure 18. For all but the Ubiquitin protein, the starting configuration was from the NMR structure, yet the equilibrated simulation conformations do not exactly resemble this initial structure. Overall, however, the global fold is fully preserved, and the conformational differences are specific in nature. This indicates the NMR structures were not necessarily in an exact free energy

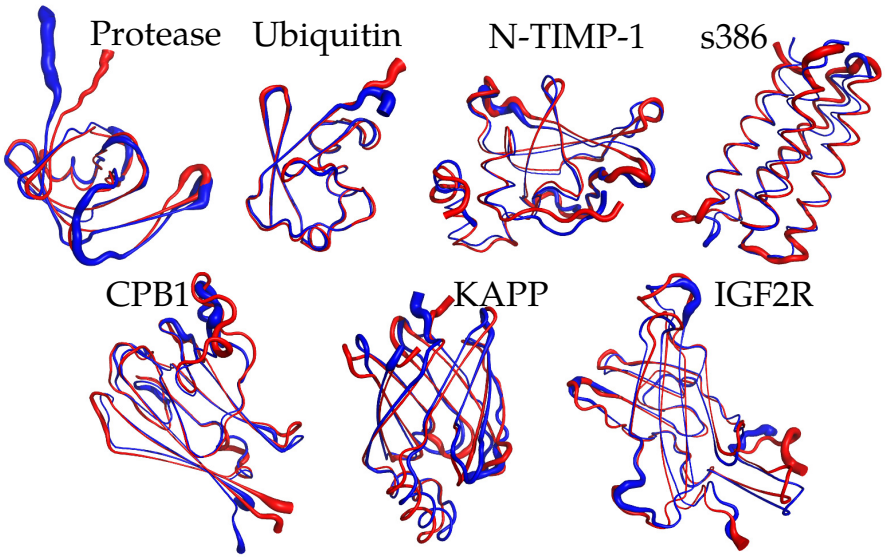

FIGURE 18. Average configuration in simulation and NMR ensembles. Average configuration from the MD simulation ensemble (red) and from the NMR structural ensemble (blue), with the thickness of the ribbon accurate to the local orientational distribution.

minimum of the AMBER force-field model, though this does not indicate whether the MD equilibrated protein structures are necessarily more accurate. For all but the s836 protein, the structural variation in the NMR ensembles is slightly larger than the MD simulation. This is primarily true in the intrinsically disordered regions of the protein, such as the C-terminal and N-terminal tails. This may be because the limited simulation times do not fully sample the configurational space. A study over a test set of 140 proteins found high correlation between the fluctuations of NMR ensembles and MD simulations, and found that the increased sampling allowed by using a coarse-grained protein model led to even higher correlation between simulations and NMR ensembles.[74] What does agree quite remarkably are the locations of enhanced flexibility and the timescales of the motion, which can be seen in Figures 22 and 23, showing the calculated NMR relaxation times from the ensembles.

To evaluate the consistency between the dynamics generated using the MD ensembles as input, and the NMR conformer ensembles, we compare the full decay of the $P_{2,i}$ correlation function of the ith C_α-C_α segment in the HIV protease protein, with the data from simulations (see Figure 19). While there are many differences between the analytical predictions from the NMR ensemble and the MD ensemble, and differences especially at short times, overall Figure 19 shows that the agreement is quite good as the LE4PD, from both ensembles, can model quite accurately the site-specific internal and rotational dynamics of the protein.

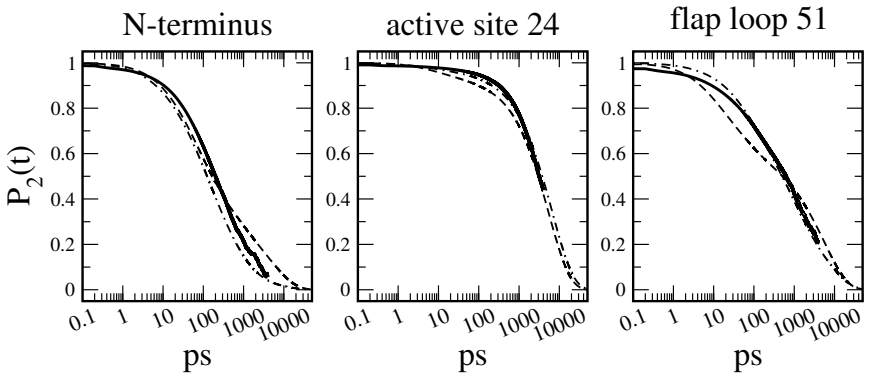

FIGURE 19. $P_{2,i}(t)$ time correlation function.
$P_{2,i}(t)$ time correlation function 3 different sites along the protein sequence of the HIV protease monomer PR95, calculated directly from the conformer simulations (solid line), from the LE4PD theory with the conformer simulations as input (dashed line), and from the LE4PD with the NMR conformer ensemble as input (dashed-dotted line).

Dynamical barriers and cooperativity

Analysis of the collective fluctuations obtained from simulations of proteins [80] has shown that the dynamics around the minima of energy is well described by small fluctuations inside metastable states at low local energy and by the crossings between them. By reverting the LE4PD equation to its mode form, the structural representations of these important metastable minima can be identified as a function

of mode number. We investigate the nature of the free energy surface of a protein around its folded ground state.

Each diffusive mode obtained from the diagonalization of Eq. 2.5 is a vector defined by the linear combination of the bond vectors weighted by the eigenvectors of the product of matrices **LU**, as $\vec{\xi}_a(t) = \sum_i Q_{ai}^{-1}\vec{l}_i(t)$. In polar coordinates the vector is represented as $\vec{\xi}_a(t) = \{|\vec{\xi}_a(t)|, \theta_a(t), \phi_a(t)\}$. The most relevant changes in the diffusive mode free energy occur as the angles, expressed in the spherical coordinates, span the configurational space. For any diffusive mode a, the free energy surface is defined as a function of the spherical coordinate angles θ_a and ϕ_a as $F(\theta_a, \phi_a) = -k_B T \log\{P(\theta_a, \phi_a)\}$,with $P(\theta_a, \phi_a)$ the probability of finding the diffusive mode vector having the given value of the solid angle. Given that we are interested in the explicit representation of the structure at the minima of interest, all structures from the simulation ensemble which pertain to a particular θ, ϕ orientation, which is a relatively deep minima in the mode free energy, are extracted and averaged.

By calculating the average structure at each minima we obtain the structural ensemble of metastable states spanning each internal mode of fluctuation for the protein. As a representative example, the free energy landscape in the LE4PD modes from the MD simulation of the HIV protease monomeric construct is presented in Figure 20. The ensemble of structural minima on the mode free energy surfaces generated from MD simulations, and the structural ensembles directly measured by NMR experiments, are compared as well. The full configurational landscape for each mode is generated from the combination of twenty well-equilibrated independent simulation trajectories. Each trajectory starts from a different experimental NMR conformer and runs to 50 ns of simulation time. Superimposed to the full configurational landscape from simulations, the twenty starting configurations

measured experimentally by NMR are reported as red stars. The combination of the trajectories creates a complex free energy landscape, which is only partially spanned by the NMR conformers. The starting NMR configurations are often close to energy minima (reported as green triangles), but they do not exactly correspond to them. Nor they are fully representative of all the minima that define the configurational landscape obtained from the simulation trajectories.

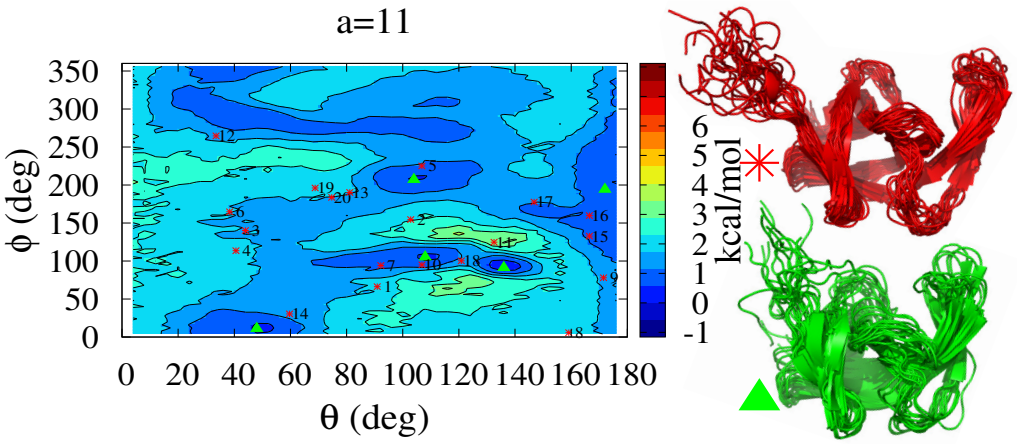

FIGURE 20. Free energy surfaces, protease.
Internal mode $a = 11$ free energy surface of the HIV protease ($a = 4$ is the first internal mode) on the left. Projections of the NMR conformer structures, labeled by conformer index, are plotted with red stars, and the simulation minima from which structures were calculated are marked on the free energy surface with green triangles. Structural minima from simulation modes $a = 4 - 20$ are on the bottom right, and the set of NMR conformers on the top right.

The fluctuations in each mode appear to be spanned by a handful of metastable minima. As the mode index increases the fluctuations progress from collective in nature to more local. The typical energy barrier in each mode, a, is evaluated from the simulation as the Median Absolute Deviation[81] from the global minimum E_{gs}, that is $E_a^\dagger = median_{(\theta,\phi)}(E_a(\theta,\phi) - E_{gs,a})$. The depth of these minima, or the barriers between them, are largest for the low mode numbers corresponding to the most collective, large-amplitude fluctuations. Figure 21 shows that the energy barriers

$E_a^\dagger$ as observed in the simulation trajectory can be well described as scaling with the mode index, a measure of the mode cooperativity, over a large range of the protein fluctuations. The observer scaling with mode number follows $E_a^\dagger \propto (a-3)^{-.5}$, where the first three rotational modes have been separated out. At a local enough length scale, where the specific chemical nature of the amino acid is most important, the energy barriers are no longer described by this expression.

The observed scaling law is consistent with the hierarchical nature of the protein free energy landscape. Each mode describes dynamics involving a number of bonds in the protein, which need to move collectively in a cooperative fashion. At short times the bonds fluctuate independently, while large-amplitude correlated fluctuations occur when all the bonds transition collectively.[82] The equilibrium probability for the Z gating bonds to independently transition away from the "correct" orientation with energy preference E is $P(Z) \propto \exp(-\frac{ZE}{k_B T})$. In a transition state perspective this can be interpreted as a free energy barrier which scales proportionally with the number of bonds cooperatively rearranging. This model is similar to the Adam-Gibbs theory of the glass transition,[83, 84] relating the complex hierarchical nature of the free energy landscape the protein in solution to a structured glassy fluid.[50, 85] The observed scaling form is included in the simulation-free LE4PD approach, which adopts the set of NMR conformers as the input structural ensemble.

Predictions of NMR relaxation are compared to experiments

Theoretical predictions for $P_{2,i}(t) = \frac{1}{2}(3\cos^2\theta_i(t) - 1)$ are obtained from $M_{1,i}(t)$ using Eq. 2.8, and are used to calculate T_1 and T_2 relaxation times, and NOE, which are measured experimentally. 15N NMR backbone relaxation experiments are very sensitive to the site-specific dynamics in the picosecond to

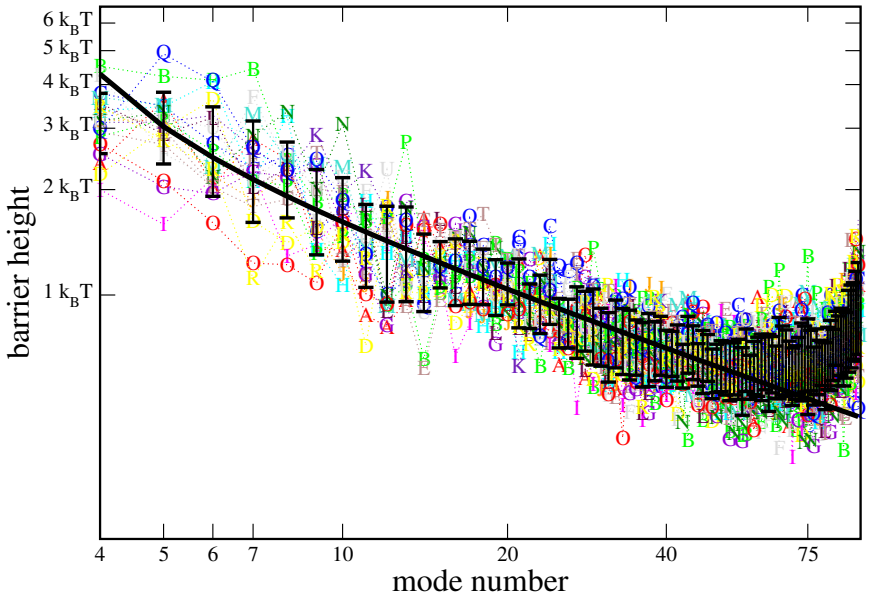

FIGURE 21. The free energy barrier to diffusion in the modes, protease. The free energy barrier to diffusion in the modes $E_a^\dagger$ calculated directly from the set of 20 simulations of the HIV protease at $T = 293K$, indexed by letter A-T, with standard deviations around the average as black bars. The black line is the fit to the scaling form $E_a^\dagger \propto (a-3)^{-.50}$.

the nanosecond regimes.[59] To test the LE4PD approach using the NMR solution structures to generate the structural ensemble, we constructed dynamical models for seven proteins for which NMR relaxation data and NMR solution structures were available. These proteins were N-TIMP-1 (1D2B)[86], a de Novo α-helix bundle protein s836 (2JUA)[87], Cellular retinol-binding protein I CPB1 (1MX7)[88], Kinase-associated protein phosphatase KAPP (1MZK)[89, 90], Insulin Growth Factor 2 Receptor IFG2R domain 11 (2M6T)[91], Ubiquitin (1UBQ)[3, 32], and HIV Protease monomer (1Q9P).[92, 93]

The input parameters to the LE4PD equation change from protein to protein: the structural parameters such as bond length, monomer friction, hydrodynamic radius, and the pairwise bond correlations are determined from the structural ensemble,

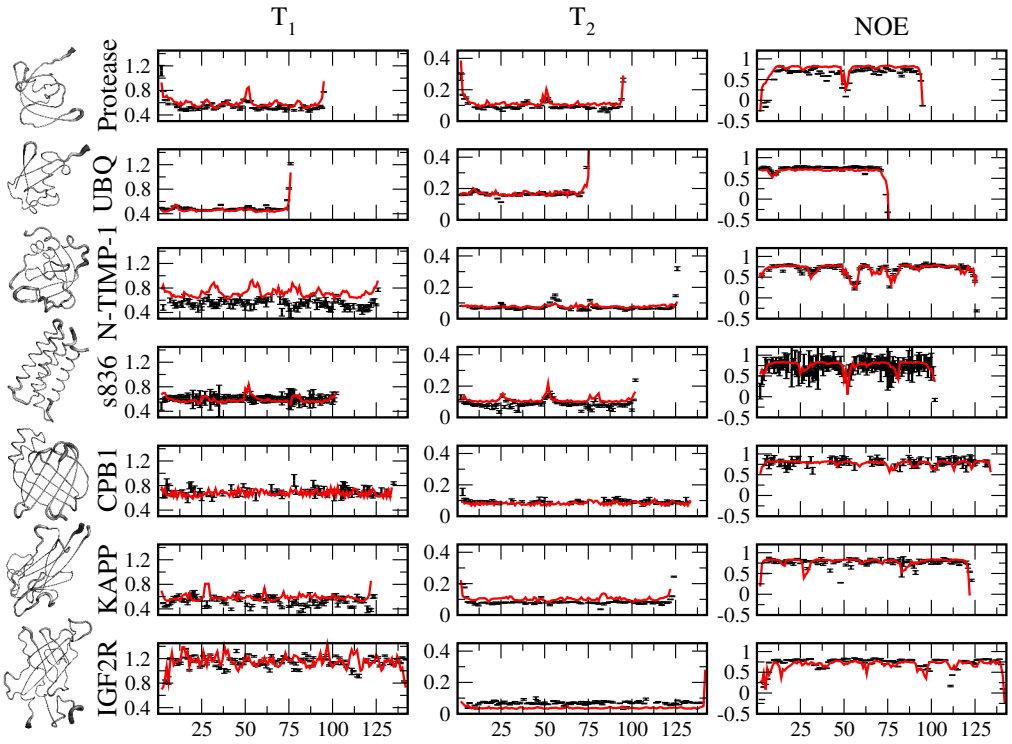

FIGURE 22. T_1, T_2, and NOE relaxation times using simulation ensembles. T_1, T_2, and NOE relaxation times (see Table 1) for seven different proteins. Comparison between experimental (black) and theoretical values from LE4PD theory from MD generated ensembles (red).

while the thermodynamic parameters such as solvent viscosity, and temperature, are defined by the experimental conditions. The viscosity was set to account for temperature dependence and content of deuterated water.[44] Parameters such as the protein internal viscosity η_p, the proportionality constant between cooperativity and energy barriers, and the characteristic parameters needed to calculate the NMR relaxation times, such as the chemical shift or $\langle 1/r_{NH}^3 \rangle$, were assumed to be identical for all proteins in this study and identical to those used in our previous work.[17]

Figure 22 and 23 displays the calculations of T_1, T_2 and NOE relaxation times as they are directly predicted by the LE4PD approach and the NMR experimental data. NMR experimental data of relaxation times are not used at all in any point to

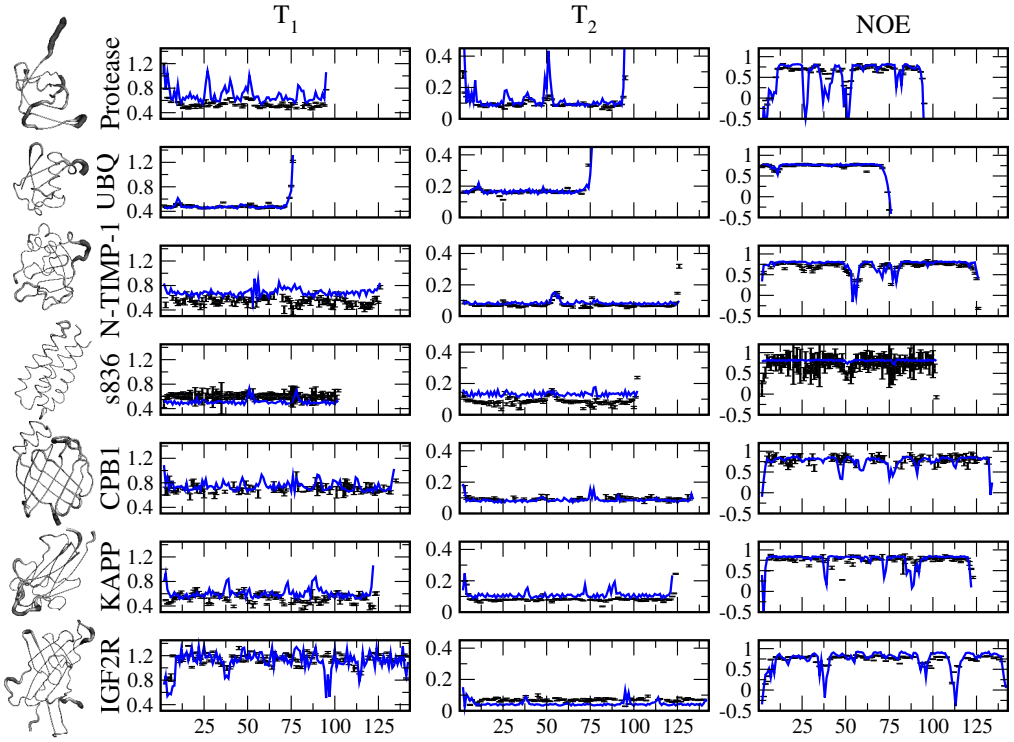

FIGURE 23. T_1, T_2, and NOE relaxation times using NMR ensembles. T_1, T_2, and NOE relaxation times (see Table 1) for seven different proteins. Comparison between experimental (black) and theoretical values from LE4PD theory from ensembles generated from NMR conformers (blue).

optimize the theoretical calculations, so these are independent theoretical predictions. The comparison between theory and experiments is performed for each amino acid in the protein and reported as a function of the protein primary sequence. Also reported are the experimental uncertainties for the NMR data of each protein.

TABLE 3. Correlation with Experimental Data of NMR Relaxation

Protein	ρ_{total}	ρ_{T1}	ρ_{T2}	ρ_{NOE}	Rel. Err.
Combined (MD)	.95	.88	.73	.70	20.8%
Combined (NMR)	.93	.83	.58	.69	24.9%
HIV Protease (MD)	.92	.73	.91	.91	32.9%
HIV Protease (NMR)	.83	.65	.90	.77	42.6%
Ubiquitin (MD)	.98	.96	.94	.97	7.0%
Ubiquitin (NMR)	.97	.96	.94	.99	7.2%
N-TIMP-1 (MD)	.92	-.10	.50	.82	20.1%
N-TIMP-1 (NMR)	.96	-.18	.57	.62	22.2%
s836 (MD)	.97	.03	.48	.57	20.6%
s836 (NMR)	.93	.18	.13	.33	34.2%
KAPP (MD)	.96	.02	.60	.60	20.1%
KAPP (NMR)	.91	-.12	.59	.46	24.6%
CPB1 (MD)	.98	.03	.06	.16	9.5%
CPB1 (NMR)	.96	.03	.14	.28	10.0%
IGF2R (MD)	.97	-.06	.84	.80	24.6%
IGF2R (NMR)	.95	.34	.15	.67	25.7%

The correlation and errors of the model, using both the MD and NMR solution structures as ensembles, are shown in Table 2. Over this set of 1876 measurements the overall correlation to the experimental values was similar for both the dynamical models constructed from NMR conformer ensembles or the MD ensembles, but with 17% lower relative error for the MD derived ensembles than the NMR conformer ensembles. Figures 22 and 23, and Table 2, in general show that MD simulations have the most detailed agreement along the primary sequence. The correlation to

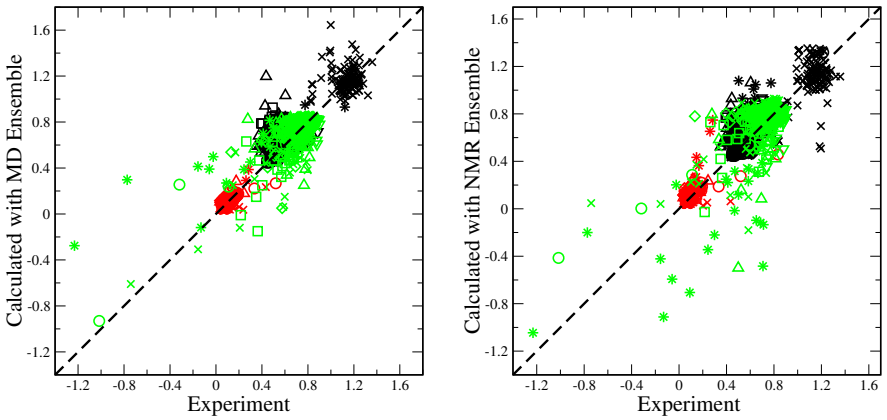

FIGURE 24. Correlation plot between experimental and calculated values. Correlation between experimental and calculated values from MD ensembles and NMR ensembles for a set of 7 different proteins. T_1 measurements in black, T_2 measurements in red, and NOE measurements in green, for Ubiquitin (circles), N-TIMP-1 (squares), s386 (diamonds), CPB1 (downward triangles), KAPP (upward triangles), IGFR2 (x), and HIV protease monomer (star).

NOE and T_1 are similar, but higher T_2 correlation for the MD ensembles. Over the seven proteins, the quality of the experimental measurements varies greatly; for example, in the measured relaxation for the s836 protein the experimental values themselves come with $\sim 30\%$ error, so that the low correlation of the theory with the experimental data is expected. For the CPB1 protein the experimental measurements in most loop and termini regions were unavailable; this is where the largest variability in the dynamics occurs and where it is possible to develop strong correlation. In general, the NOE measurements display the largest site-specific variability along the protein sequence and the highest correlation between theory and experiment for each individual protein. A scatter plot of the calculated and experimental data is shown in Figure 24. The agreement between theoretical predictions and measured NMR is supporting the quality of the predictions of the LE4PD approach.

Because the accuracy of a given NMR solution structure ensemble to represent the conformational diversity of the protein is unknown, the dynamical model built

using the LE4PD approach may be useful to evaluate the quality of an available structural ensemble. We apply the method to 9 different NMR structural ensembles of the Ubiquitin protein, PDB codes 1XQQ,[94] 2KOX,[95] 2LJ5,[96] 2NR2,[97] 1D3Z,[98] 1G6J,[99] 2KLG,[100] 2MJB,[101] and 2K39.[12] The comparison to the calculated NMR backbone relaxation in table 4 shows that all the ensembles capture the primary T_1, T_2, and NOE baselines and the enhanced flexibility of the tail region. Ensembles 1XQQ and 2NR2 have the highest correlation and lowest relative error; when the unstructured C-tail is not considered in the calculation of the correlation coefficients, it can be seen that the 1XQQ, 2NR2, 2KOX, 2LJ5 and 2K39 ensemble separate as capturing the structural variability of both the C-tail and the more structured portion of the protein, see column 2 of table 4. The primary contribution to this correlation comes from the structural variability at the loop containing lysine 6 and 11, important poly-ubiquitination linkage sites involved in cell-cycle control and DNA repair.[2] The structural ensemble generated by molecular dynamics simulation starting from the 1UBQ[32] crystal structure, with results reported in our previous paper,[17] is perhaps slightly more accurate overall, but only by a very small amount due to considering the correlation without contributions from the C-tail.

In generating the 1XQQ ensemble the NMR-derived S_2 order parameters from the model-free analysis of Lipari and Szabo[61] were used as an additional set of restraints in the generation of the ensemble. It is not surprising then that this leads to an accurate dynamical model. We have shown previously that the site-specific variability in model-free derived S_2 order parameters correlates strongly with our results,[8] despite differences in the nature of the predicted internal dynamics. This illustrates the complementary utility of the LE4PD approach, which provides a highly

TABLE 4. Correlation with Experimental Data for Ubiquitin

NMR Conformer	ρ_{NOE} (1-71)	ρ_{NOE} (all res.)	ρ_{T2}	ρ_{T1}	Rel. Error
MD Sim.	.71	.96	.94	.97	7.0%
1XQQ	.66	.99	.94	.96	7.2%
2NR2	.52	.98	.94	.95	7.3%
2LJ5	.56	.93	-.33	.97	7.7%
2K0X	.61	.94	.80	.88	8.2%
1D3Z	.02	.88	.80	.94	8.2%
2K39	.70	.88	-.63	.93	11.0%
2MJB	-.01	.96	.96	.92	11.2%
1G6J	.02	.92	.94	.92	11.5%
2KLG	-.05	.86	.92	.73	14.9%

detailed model and additional insight beyond that available when performing only a model-free analysis of NMR backbone relaxation.

The Ubiquitin ensemble 2K39 was constructed to represent the protein fluctuations in only the long-time regime beyond the global correlation time. As such this ensemble is not as accurate overall, and the dynamical model leads to high error and in particular a poor representation of the C-tail dynamics. However, we do see a separation in the mode timescales, with a slow internal process emerging on the order of $\sim 400ns$ and with $\rho_{NOE,(res1-71)} = .71$, suggesting that this ensemble has captured fluctuations in the difficult to access time regime between the global correlation time and the millisecond time regime of conformational exchange. The authors showed that this ensemble spanned the set of known bound Ubiquitin conformations, suggesting that there are configurational fluctuations of the Ubiquitin protein in the many nanosecond regime beyond the global correlation time which are relevant for the recognition of binding partners.[12]

Conclusions

The LE4PD approach was tested across a set of seven different proteins with overall consistent results for both the MD generated ensembles and the NMR conformer ensembles, with an overall correlation to the 1876 relaxation measurements of $\rho > .93$. Calculations using 9 different available NMR structural ensembles for the ubiquitin protein show that results are strongly dependent upon the quality of the input structural ensemble and experimental data, and suggest that this approach may be used as a tool to evaluate the quality of a structural ensemble to represent the important protein fluctuations around the ground folded state.

The consistent results between the MD generated ensembles and the NMR ensembles suggest that protein configurational space around the folded state can be defined by a small set of important metastable minima. However, when determining the dynamics of transitions between these minima, the hierarchical nature of the protein free energy landscape needs to be taken into account. The mode approach of the LE4PD allows one to conveniently separate contributions to the dynamics depending on the timescales involved. The LE4PD prediction of the existence of a barrier height distribution for the dynamics of folded proteins is consistent with the physics of glass-forming systems.

Building a dynamical model from NMR conformer structures using the LE4PD requires only a few seconds to a few minutes on a single processor with a standard desktop computer, with the computational time depending on the size of the protein and on the number of conformers in the NMR solution structure. While explicit solvent atomistic classical MD simulations are well-developed and can be quite accurate, achieving MD simulations with converged dynamics on the same timescale would require on the order of $10,000 - 100,000$ hours of processor time or more. The

LE4PD is not a replacement for MD simulation as a computational method to predict the fluctuations and dynamics of proteins, but it is a useful tool to quickly provide a prediction of the dynamics given an input structural ensemble.

Even though the simulation-free LE4PD requires minimal computation it is site-specific, informed of intramolecular energy barriers, hydrodynamics, and long-range correlated motion. It is a sophisticated model of protein dynamics and because of its accuracy in predicting the dynamics, with no input from the dynamical data, LE4PD is a valuable and computationally convenient model to investigate barrier-crossing processes on the suite of timescales defining the fluctuations of proteins.

CHAPTER IV

MODE LOCALIZATION IN THE COOPERATIVE DYNAMICS OF PROTEIN RECOGNITION

The biological function of proteins is uniquely defined by the protein primary sequence, which determines its three-dimensional structure and dynamics.[56, 102] The protein local motion develops along a pathway of transitions between metastable states inside a hierarchy of structures that, on the local scale, are separated by small energy barriers of the order of k_BT.[80, 103, 104] It is the unique structure of this configurational landscape that determines the specific thermodynamic and kinetic properties of a protein and defines its ability of regulating its function.[62]

Experimental studies have suggested that protein dynamics plays an important role in protein recognition. Residue-specific, localized dynamics have been shown to be relevant in the energetic pathways of binding.[105, 106] Pre-existing pathways in the free-energy landscape have been found to guide the transmission of the allosteric signals.[107] It has also been shown that point mutations of residues, and also small ligand binding, can lead to an identical folded configuration while they modify the protein biological activity. This indicates that, in those cases, not the structure but the dynamics has the dominant effect on the protein function.[108] Protein dynamics can facilitate docking of a ligand.[109] Often the bound conformation of the protein is different from its apo structure. In those cases the bound (holo) conformation populates a preexisting configuration that is just less probable than the unbound (apo) structure, and structural elements of the holo form should be already visible in a dynamic study of the apo protein.[110]

To investigate the relation between local fluctuations and protein binding, a theoretical model that relates protein structure to dynamics can play an important role.[111] The Langevin Equation for Protein Dynamics (LE4PD) is a potential candidate to fulfill this need because of its capability of predicting with accuracy protein dynamics on a wide range of timescales.[17, 112] In this paper we extend the theory to treat two interacting proteins and present the first LE4PD analysis of sequence-dependent cooperative local fluctuations, which relates these fluctuations to the protein-specific binding regions.

The LE4PD equation is a hydrodynamic diffusive approach, analytically solved, which starts from the description of a protein in its equilibrium state and predicts the dynamics in timescales from the global rotation to the single bond fluctuation.[17, 112] The theoretical formalism has been tested against NMR relaxation data (T_1, T_2, and NOE) and x-Ray Debye-Waller factors for seven different proteins for a total of 1864 point comparison.[17, 112] We used as static input structural configurational ensembles either from Molecular Dynamics simulations or from NMR experiments. In both cases the observed quality of agreement between LE4PD predicted values of the dynamical quantities and the experiments appears to support the validity of this approach as a method to predict local dynamics of proteins in their equilibrium state across a wide range of timescales, starting from the structural information.

In this paper we calculate with the LE4PD the local dynamics of two test proteins in their apo and holo states. We correlate the predicted dynamics with experiments, and with the information about conserved residues in the family of proteins performing the same biological function. We study both the isolated and the dimerized form of the HIV protease,[92, 93] and the Insulin Growth Factor II Receptor (IGF2R) domain 11 and the IGF2R:IGF2 complex.[91] Besides their

biological relevance, those proteins have been selected because experimental data of NMR configurational ensembles and relaxation times are available in the literature, and provide a further test of the quality of the LE4PD theory and its predictions.

In a recent paper[112] the predictions of the LE4PD were compared when the input ensemble of structural configurations was taken to be the one generated by the NMR or the one emerging from atomistic molecular dynamics simulations. In this paper we continue the comparison of the two procedures and present results of the LE4PD when NMR conformers are adopted as input configurations and when the sampling of the configurational space is performed by extended MD simulations of the protein in aqueous solutions. In both cases the LE4PD lead to fairly consistent results for the mode-dependent dynamics.

Some of the questions investigated in this paper are born from general considerations about protein dynamics and function. It is known, for example, that for a folded protein a number of configurational states are available in spatially well-defined regions along the primary sequence.[113] At a given temperature a part of the protein, for instance a loop or a tail, may be intrinsically disordered and populating a number of thermally activated conformational states that are metastable, energetically similar, and so equally probable.[114] Enhanced fluctuations in the spatial positions of key residues or short fragments can make possible the trapping of favorable configurations by a reactant or a substrate, following the well-established conformational selection model of binding pioneered by Monod.[69] When the correct local configuration for binding is available in the configurational landscape of the isolated protein, the extent of free energy needed for binding is reduced. As the trapping of a favorable state does not require overcoming an energy barrier, other processes, for example inter-diffusion of the reaction partners, become the

slow relevant processes that determine the rate of the reaction. This conformational selection process for the recognition event, can be followed by a relaxation or induced fit to the bound conformation.[115]

The gain in energy as the protein transitions over the energy barrier and reaches the bound state has to be small to allow the possible breaking of the reaction product when new conditions arise that are destabilizing this state: the process needs to be flexible enough to permit the progress towards the following reaction step without dramatic gains or loss of free energy. In this delicate balance of energy, modulated within an energy window of a few k_BT, the primary sequence of a protein plays a decisive role.

The presence of barriers and their height is important in the binding reaction. Transitions need to be energetically activated to render the biological process forbidden if the temperature lowers below physiological conditions. However, the barriers need to be small to make their crossing possible at physiological temperature, as the dynamics are "fueled" by the thermal fluctuations of the surrounding liquid.[63, 116]

Experimentally it has been observed that upon binding the loss in entropy of the protein, which would oppose the reaction, is often paired to the emergence of disorder in remote regions of the protein, apparently uncorrelated to the binding site.[115] New flexible regions often arise in the relaxed bound state, or exposed hydrophobic residues are found to transition to the hydrophobic region, becoming protected and increasing the entropy of the solvent in the well-known mechanism of enthalpy-entropy compensation.[117, 118] In this way, regions with multiple states available in the configurational space, which are accessed by local thermal fluctuations, play

an important role in the entropic and enthalpic balancing during recogniton and binding.

In its diffusive mode description, the LE4PD allows for the identification of a variety of dynamical processes that emerge at increasing timescales. The diffusive mode solution of the LE4PD organizes the configurational landscape in a linearly independent set of variables. Fluctuations are defined on a range of length and timescales, with dynamics encompassing the relative motion between neighboring α-carbons to the global rotations of the structure as a whole.[8, 17] In this study internal modes of motion that present energetically activated local dynamics are identified together with the characteristic length and timescales of their dynamics. A range of equilibrium dynamical processes emerge on different timescales following a hierarchical scheme, suggesting a possible sequential mechanism in the non-equilibrium reaction pathway.

The LE4PD predicts the emergence of specific regions in the protein three-dimensional structure that are dynamically active at a given timescale. The diffusive mode rendition precisely identifies the position inside the primary sequence of these energetically-guided local fluctuations, and provides information about the extent of localization of these activated dynamics. This indicates if the motion involves a single residue or a number of cooperatively moving specific residues. By identifying and analyzing the regions of local flexibility and cooperative motion of the residues inside a protein, we argue that it is possible to learn which parts of the protein will lead the kinetics of the biologically relevant processes. In this way, the LE4PD model provides a straightforward and visually intuitive representation of the locations of enhanced reactivity, or "binding regions," and the emergent length and timescales of motion. For all the proteins in this study, the LE4PD method indicates regions of high

mobility and slow, large-amplitude dynamics, which are trapped and not detected in the bound forms of the protein. We found that these regions directly correspond to regions with highly conserved residues in the family of proteins with related biological function, and are involved in the binding interactions.

Finally, the LE4PD analysis shows that upon binding, dynamics in the diffusive normal mode coordinates can be enhanced or suppressed in other remote regions of the protein. The extent of the fluctuations indicates if a new dynamically active region emerges that is likely involved in the following step in a reaction pathway (allosteric mechanism), or if entropically-relevant multiple states emerge that are distributed along the protein for an entropy-enthalpy compensation mechanism.

Fluctuation Driven Dynamics of Binding

In this paper the formalism is extended to treat the dynamics of a pair of interacting proteins. We define the protein complex as one system but we assume that there is no chemical bond between the α-carbon belonging to the C-terminus of the first protein with α-carbon belonging to the N-terminus of the second. Formally, the index labeling the protein bonds is extended to $i = 1, ...N_{complex}$ where, for example in a two-protein complex, $N_{complex} = N_1 + N_2$, and N_1 is the number of bonds in the first protein while N_2 is the number of bonds in the second protein. The N_1+1 bond is now related to the first residue of the second protein in the complex, where the ordering of choosing protein one or two is arbitrary. As just discussed, the only difference between this description and the single-protein one lies in defining the connectivity of the backbone.

In its present form, the theory describes fluctuations of the complex, but not yet possible association and disassociation processes, which exceed the timescale of the present simulations: for a proper description of the association reaction a realistic

reaction pathway should be explicitly included. However, the approach is general and applies to any kind of multi-molecular complex.

The statistical averaged structural parameters that enter the LE4PD can be calculated either from the configurational ensembles measured experimentally by NMR or from the ensemble generated from MD simulations of the same protein in solution.[112] In this paper we present results for both procedures. Details of the MD simulations are reported in a previous publication.[112] For the PR95 protease monomer, simulations were performed starting from each of the twenty conformers in the NMR structure, resulting in a set of twenty production trajectories of 50 ns. For the IGF2R protein and IGF2R:IGF2 complex, the first conformer was chosen as the starting structure, and only one simulation with $150ns$ was performed. The NMR configurational ensemble is sampling configurational states that are only partially consistent with the most stable states sampled in the MD simulation. In practice, NMR solution structures encode a structural diversity that is due to a combination of thermal fluctuations and a possible lack of complete experimental information. This is important because it indicates that both statistical ensembles we are starting with are not covering the full configurational landscape: the MD simulations are precise in sampling the states that are relevant in the time regime covered by the simulations, but other states could be possible at longer times; the NMR ensemble includes states that are away from the principal minima but their sampling is not complete. The observation that the predictions of the LE4PD theory are largely consistent in the two routes, indicates that the most relevant dynamical processes are accurately accounted for by both methods.

Because of its mathematical foundation, the LE4PD theory can predict any time dependent property of interest in the coarse-grained coordinates. For example,

the LE4PD diffusive mode description provides useful information about enhanced fluctuations and the crossing of energy barriers in localized regions of the protein, as well as the lengthscale and timescale of these fluctuations. The question we want to address is if local fluctuations that are present in the dynamics of the protein in its apo form can be correlated with the spatial regions in the protein that are directly involved with its binding to another protein or substrate. Although this question has been investigated in the past, the LE4PD approach can bring a new and useful understanding because of the diffusive mode description that allows for the convenient separation of the dynamics into quasi-normal modes with well-defined spatial regions of activity, and a defined characteristic timescale.

We define as the physical quantity that represents the sequence dependent fluctuation, the mean-squared Local Mode Lengthscale (LML)

$$L_{ia}^2 = Q_{ia}^2 \xi_a^2 \ , \tag{4.1}$$

which depends on the residue in the primary sequence of the protein and on the timescale and lengthscale of the diffusive mode. In the following analysis we report for each mode the site -dependent mode length for each protein in its apo and holo states and compare the dynamics in the two states. A sample of the type of information collected is represented, in Figure 26 and in the following figures. Through the analysis of the dynamics in diffusive modes, this study visualizes possible reaction pathways in mode coordinates for the pre-binding dynamics of the protein, and the time and space modulation of this dynamics after binding.

We describe Figure 26 as an example. In the right panel of Figure 26 we report the LML as a function of the primary sequence of the protein, for modes $a = 4$, 5, 6, 7, 8, and 9. The peaks in the LML define enhanced fluctuations in the local structure

of the protein. In the left panel of the figure we report a graphic representation of the fluctuations in the protein structure for each mode. This representation helps visualize the localization of the fluctuations. The mode amplitudes are projected along the protein sequence, by making the radius and color of the tube representing the protein to correspond directly to the extent of the local mode fluctuation. Also reported for each mode in the left panel is the correlation time of the mode. To help identify special regions in the protein sequence that are relevant for the protein biological function we included in the right panel vertical dashed lines.

The diffusive modes have quite different behavior depending on the mode number. The first three modes ($a = 1, 2, 3$) represents, for well-folded proteins as the ones in this study, the three rotational modes. These manifest themselves as very well-localized minima of energy in mode-space. However, if the simulation contains slow configurational transitions globally affecting the overall protein structure (i.e. protein folding), then the first three modes would present a more complex energy landscape including slow crossing of global energy barriers. For the proteins in this study the crossing of energy barriers does not involve global conformational transitions, but only cooperative fluctuations of the folded protein.

The internal modes present different behavior depending on if they are slow (low index modes) or fast modes (high index modes). For a protein represented by N coarse-grained units, here amino acids, there are $N - 4$ internal diffusive modes, and it would be impossible to include figures for each one of them. However their behavior is in general different depending on the mode index. The high index modes correspond to short-time processes ($\tau \sim ps$), many of which involve just small variation in the LML along the sequence. These delocalized fluctuations are largely of entropic origin, and we argue that they mainly serve to balance energetically the chemical processes in

which the protein is involved, e.g. binding reactions. Some of the high-index modes, especially the very fastest, are highly localized to specific protein regions, namely the most tightly packed amino acids, usually integral to the stability of the protein core and highly conserved. While important, the properties of protein dynamics at short times have been extensively studied and are well-represented by very simple spring network models.[119]

Here we are primarily interested in the slow internal modes (low index modes, $a \geq 4$ and $\tau > ns$), which show localized fluctuations that are very different depending on the mode. Because one timescale corresponds to one mode, we can look at pathways in which fluctuations are engaged during a window in time, as the localized fluctuations move along the primary sequence following the mode index. In this study we report only the most interesting diffusive pathways of mode fluctuations, but the complete description of the dynamics is obtained from the analysis of the modes, which represent the equation of motion of the protein with internal energy barriers and hydrodynamic interactions accounted for. Interestingly if hydrodynamics is not included or if the local barriers are not included, the dynamical response is quite different, indicating that long-time dynamics is strongly affected by long-range interactions due to hydrodynamics and by the cooperative rearrangements of the chains that overcome internal energy barriers.[112] Thus the internal friction of the protein plays an important role.[28, 120]

In Figure 25, an example of the mode-dependent free energy landscape is presented for the apo form of the HIV protease and IGF2R protein domain 11, calculated from MD simulation. In practice, to build the mode-dependent free-energy plot, the ensemble of bond vectors, $\vec{l}_i(t)$, is collected at each time step in the trajectory and transformed into mode coordinates. From the probability the energy is calculated

and represented graphically for each mode as a function of the polar coordinates θ_a and ϕ_a of the mode vector. As the trajectory is analyzed to recover the ensemble of bond coordinates, the configurations of the protein are collected and averaged together depending on their value of the polar parameters θ_a and ϕ_a. With this procedure an average configuration is associated to each point in the mode-dependent energy plot.

The NMR conformer ensemble of the holo form of each protein is depicted in magenta on the right, and projected onto the free energy landscape as magenta stars on the left. Interestingly the bound state of the protein, depicted as purple crosses, populates only a well defined energetic region among all the possible ones. This seems to indicate a mechanism of trapping of a favorable configuration during protein binding. However, the picture we observe is more complex than usually assumed. The holo structure is not simply a single conformation that was already present in the ensemble of apo structures as a less-populated member of the ensemble, and that is trapped during binding. We observe instead that the holo protein presents a restricted ensemble of configurations in the active regions of the apo protein.

Once the dynamics of the apo structure of a protein is resolved we perform the same calculation for the holo structure of the same protein and compare localized fluctuations. Through the analysis of the dynamics in diffusive modes, this study visualizes possible reaction pathways in mode coordinates for the pre-binding dynamics of the protein, and the time and space modulation of this dynamics after protein binding. After binding, different active regions emerge with large amplitude uncorrelated motion for entropic compensation and more-localized enhanced fluctations indicating possible subsequent steps in the reaction mechanism.

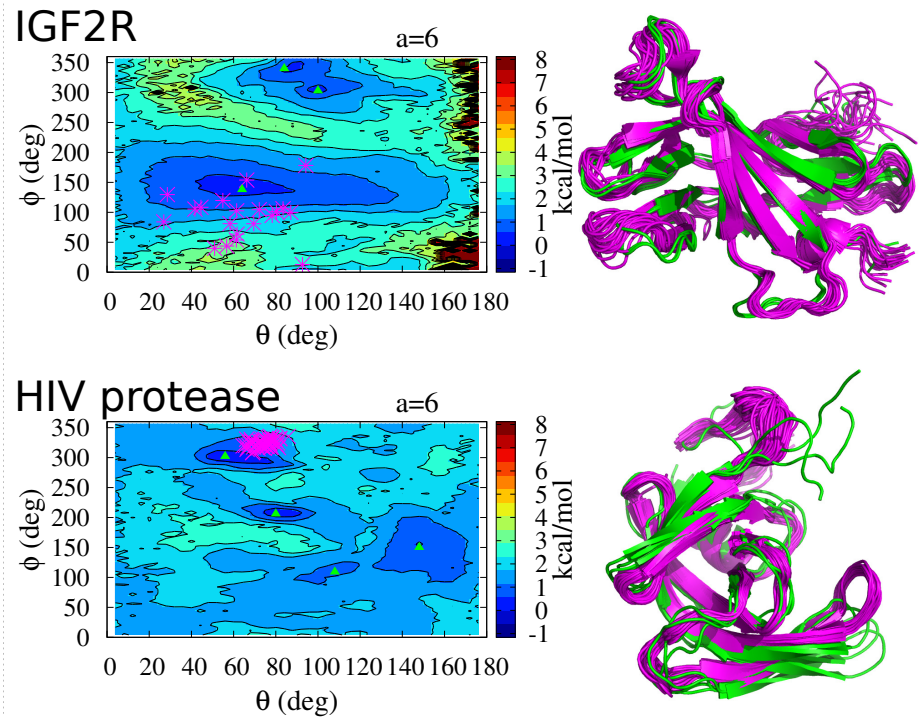

FIGURE 25. Free energy surface with apo structure projection.
Internal mode free energy surface of the IGF2R protein domain 11 (top left) and of the HIV protease monomer (bottom left) from MD simulation. NMR structural ensemble of a dimerized and inhibited HIV protease and the IGF2R:IGF2 complex are projected on the energy landscape as as magenta stars, with the structures on the right in magenta. Average structures in minima from the MD simulations (green) and marked with green triangles.

Correlation between Local Fluctuations and Binding

In this section we present the study of the correlation between the local fluctuations in diffusive mode dynamics and the binding of two test proteins: HIV protease and IGF2R protein domain 11. These proteins were selected because they have been studied structurally by NMR spectroscopy, so they provide the opportunity of comparing LE4PD predictions when starting from structural ensembles either obtained with MD simulations or from NMR experiments.

For the apo form of the two proteins, HIV protease and IGF2R protein domain 11, data of T1, T2, and NOE relaxation from NMR experiments are also available. In this way they provide a further test of the quality of the predictions by the LE4PD theory. In a previous paper we reported a detailed comparison of the theoretical LE4PD data with NMR experiments.[112] In this paper the theory is extended to treat the dynamics of the complex, and the theoretical predictions for the NMR relaxation of the holo protein for the HIV-protease dimer in complex with DMP-323 inhibitor are directly compared with the experimental data.

HIV protease

As the first example, we consider the dynamical model of the HIV protease *monomer* protein. We study LE4PD predictions starting from the NMR conformer ensemble 1Q9P.[93] In its active dimeric form aspartyl protease catalyzes the cleaving of the peptide chain for the separation of the protein products of the HIV genome.[121] The PR95 monomer construct studied here is modified to prevent the protease from self-cleavage, with five single-residue substitutions along the sequence and the deletion of the terminal residues 95 - 99. These C-terminal residues stabilize the dimer by forming antiparallel β-strands with the N-terminal residues 1 - 5.[122] Without

the C-terminal residues, both the NMR relaxation rates and the ensemble of NMR conformers[92, 93] are consistent with an N-terminal tail that is highly flexible.

The dynamic localization is represented visually in the left panel of Figure 26 using the structure of the protein. The analysis of the position inside the primary sequence of the large-amplitude internal modes shows that in this protein the fluctuations in the slow modes are mostly localized along the flap loop and the terminal region. This is consistent with the role of the flap loop in HIV protease, mediating access to the active site. The LE4PD model indicates these structural fluctuations of the flap loop take place in a large time interval of 1 ns to 500 ns for the unbound protein monomer. The motions are strongly correlated with smaller-scale conformational transitions in the active site, i.e. the sequence between amino acids 21 and 33, and the region around amino acid 40.

Important conserved regions in the family of HIV protease proteins are the residues in the active site, aminoacids 22 - 34, the flap loop residues 47 - 52 which control access to the active site, and a region containing an α-helix and part of the hydrophobic core residues 74 - 87.[123] These are consistent with the regions where enhanced mode-dependent fluctuations emerge.[119] Interestingly, the diffusive mode representation shows how the dynamics is partitioned into a number of time regimes, and indicates how chemical mutations in specific sequences of the protein could be affecting differently the kinetic of the process.

The mode-dependent dynamics of the dimerized and inhibited HIV protease is presented in Figure 27. In the bottom panel of the figure, the LE4PD theory, which has been shown to predict with accuracy the dynamics of apo proteins, is compared with the experimental data of NMR relaxation for the protein complex. When calculating data for the protein complex, we first note that while the protein

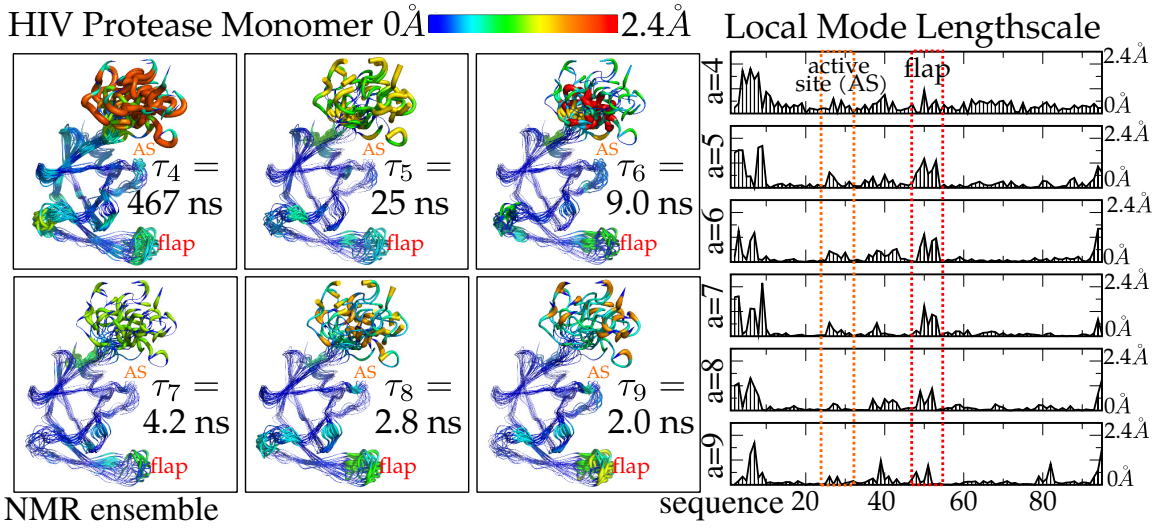

FIGURE 26. Fluctuations of the free HIV protease monomer.
Fluctuations of the free HIV protease monomer. Local mode lengthscale L_{ia} along the primary sequence (Right panel), and projected as a rainbow gradient onto the ensemble of conformer structures 1Q9P (Left panel).

complex and inhibitor posess a C2 symmetry axis, the NMR conformer ensemble breaks the equivalence between the two sides of the homodimer. We then choose to enforce the C2 symmetry by including permutations of the labeling of protein order in the conformer structures, which results in dynamical modes which are symmetric between the two proteins of the dimer. While the lack of C2 symmetry should be expected for any single NMR structure in the ensemble, since there is no difference in the two sides of the homodimer, any asymmetric structural minima of the complex should come in pairs. That this doesn't hold completely true in the NMR structural ensemble is probably an artifact of the NMR sampling. In either case, enforcing the C2 symmetry in the ensemble, or not, leads to identical agreement when comparing to site-specific NMR relaxation experiments[4] with only 12% root mean-squared relative error, as shown in the bottom panel of Figure 27. The figure shows that the LE4PD theory predicts with very good accuracy the dynamics also for the dimerized and inhibited HIV protease.

For the dimerized and inhibited HIV protease[124], the diffusive mode representation shows that the enhanced mobility of the flap loop and the terminal regions observed in the HIV monomer are not present in the dynamics of the complex, suggesting that the binding fluctuations of the monomer are suppressed upon formation of the complex (see Figure27). We notice that there is no information in this analysis about how this process occurs because we are examining the diffusive dynamics of the initial and final states of the binding reaction, and not its kinetics.

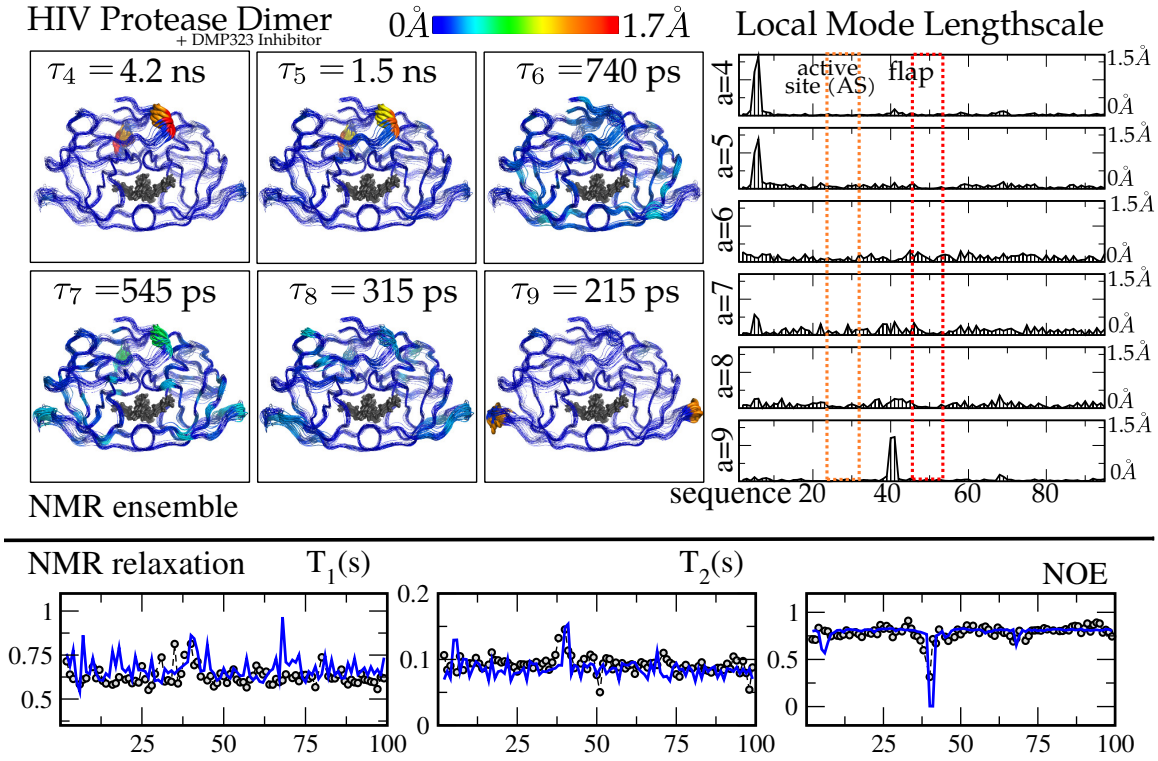

FIGURE 27. Fluctuations of the dimerized and inhibitor-bound HIV protease. Top panels: Primary fluctuations of the dimerized and inhibitor-bound HIV protease. Local mode lengthscale L_{ia} along the primary sequence (Top right panel), and projected as a rainbow gradient onto the ensemble of conformer structures 1BVE (Top left panel). Because the modes are symmetric over the two members of the homodimer, only residues 1-99 are shown. Bottom panels: Experimental[4] (circles) and predicted (blue lines) NMR relaxation of the dimerized and inhibited HIV protease.

As shown in Figure 27 the large-amplitude fluctuations observed in the slow modes of the monomer, are now trapped in the dimerized and inibitor-bound protein. The C-terminal and N-terminal tails which were previously disordered are now locked into a β-sheet structure, while motions of the flap loop and active site are inhibited by the binding of the DMP323 inhibitor. This suggests that the binding of the second protein and of the inhibitor traps specific configurational minima of active regions of the monomeric protein. A more direct observation of this phenomenon can be made in the analysis of Figure 25, where the dimerized and inhibitor bound protein structures are projected onto the mode free energy landscape calculated from MD simulations of the free monomer. The figure shows that the bound form is localized to specific regions and minima on the monomer energy landscape.

Figure 27 shows that in the bound protein new fluctuations emerge that are localized to almost single-residue regions, such as the large-amplitude fluctuations of L5:W6 and P39:G40. While in the apo protein, the fluctuations in these regions were correlated with terminal and flap motions, in the complex, the fluctuations have become highly localized. The well defined peaks that emerge, characterizing localized dynamics in the region of amino acids 5 and 40, are suggestive of a more specific step in a following reaction pathway that would involve these amino acids.

This finding is consistent with NMR relaxation experiments of the HIV protease/DMP323 inhibitor complex which found unusually large and fast motion at residue 40: this residue was speculated to be involved in the release of the reaction product after protease activity.[4] Furthermore, mutagenesis studies of the HIV protease found that non-conservative mutation of residues 5 and 40 resulted in loss of protease activity, even though the residues are separated in space and sequence from other highly conserved regions.[125] Our study shows that while these regions

already present some degree of flexibility in the apo protein,[92] their dynamics are further and specifically enhanced upon binding. The LE4PD theory indicates the presence of dynamical allostery in the dimerized complex upon binding.

IGF2R protein domain 11

As a second example, we study the Insulin Growth Factor II Receptor (IGF2R) protein domain 11 which is responsible for binding and regulating levels of insuline-like growth factor 2 (IGF2) at the cell surface. Williams et al.[91] resolved the structure of the IGF2R:IGF2 complex by x-ray, and the structure of an AB loop mutant of the protein in solution by NMR. The binding sites are composed by defined loops called AB, CD, FG, and HI. The same nomenclature is reported in our figures. Starting from this information they developed a binding model of IGF2R with the IGF2 protein where they suggested a dynamical role for two primary loops flanking a hydrophobic binding pocket.

Using as an input the ensemble of NMR configurations for the apo protein we calculated the LE4PD dynamics, which includes fluctuations around the NMR configurations, local cooperative barrier crossing, and solvent-mediated hydrodynamic interaction. The diffusive modes show for the residue-dependent Local Mode Lengthscale, L_{ia}, an interesting mode-dependent behavior. The first internal mode primarily involves independent fluctuations of the unstructured C-tail. The second and third internal mode are localized on the AB and FG loops in agreement with the model proposed by Williams et al. As can be seen in Figure 28 these modes suggest largely independent, uncorrelated motion of these loops in the tens of nanosecond regime. The dynamics move from the AB loop to the FG loop on the $3\ ns\ -\ 30\ ns$ timescale indicating a possible sequence of steps in the binding mechanism of the

protein, with the fast local dynamics ($\sim 3\ ns\ -\ 6\ ns$) occurring in the FG loop and the slow dynamics ($\sim 30\ ns$) localized in the AB loop.

An analysis of IGF2R sequences across species has shown that a gain of IGF2 binding affinity and function is related to specific, localized point mutations.[126] Most of the studied mutations were located in the IGF2R:IGF2 binding region, particularly residues 1544-1545 on the AB binding loop and residue 1600 on the FG loop.

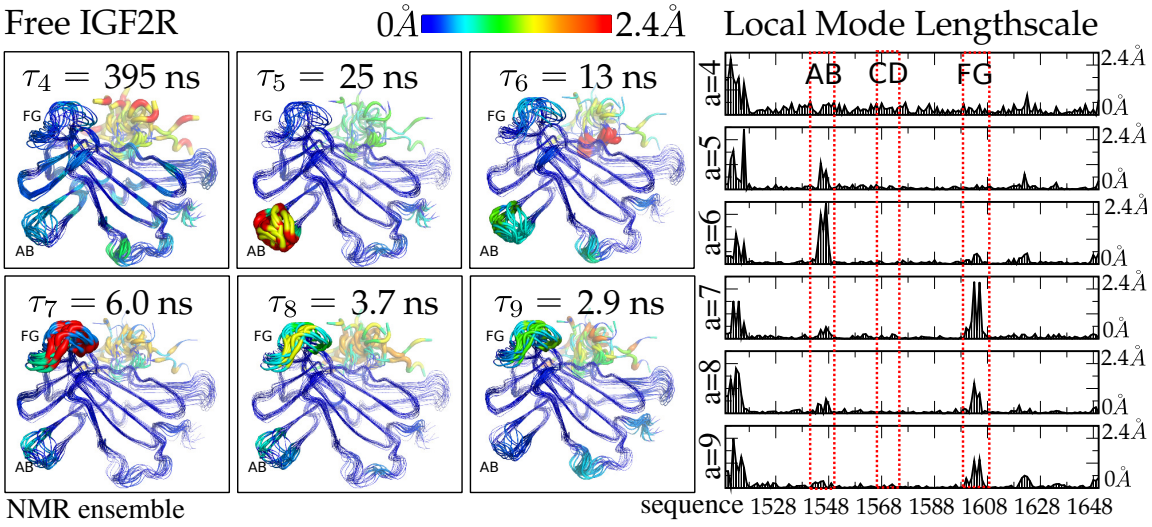

FIGURE 28. Fluctuations of the free IGF2R domain 11 protein from the NMR ensemble.
Fluctuations of the free IGF2R domain 11 protein from the NMR conformer ensemble. Local mode lengthscale L_{ia} along the primary sequence (Right panel), and projected as a rainbow gradient onto the ensemble of conformer structures 2M6T (Left panel).

The dynamics presented so far for the IGF2R protein domain in the unbound and bound configurations is calculated using the LE4PD theory with input from the experimental structures measured in NMR. However, further calculations were performed with the LE4PD theory starting from MD simulations. The simulation collected a 100 ns long trajectory with stable Root Mean Square Deviation from the initially selected structure, which is the first configuration NMR structure. Results

for the mode dependent dynamics are reported in Figure 29. The generated MD ensemble does not span the full range of dynamics observed in the NMR structural ensemble, which is not surprising since the variability in the NMR structural ensemble suggested mode dynamics in the $\sim$ 400 ns timescale, while the simulations are more limited in the timescale that they cover. However the sampling of the MD simulation in the sub 100 ns regime is more complete than in NMR experiments, as it can be observed by the results reported in the figures.

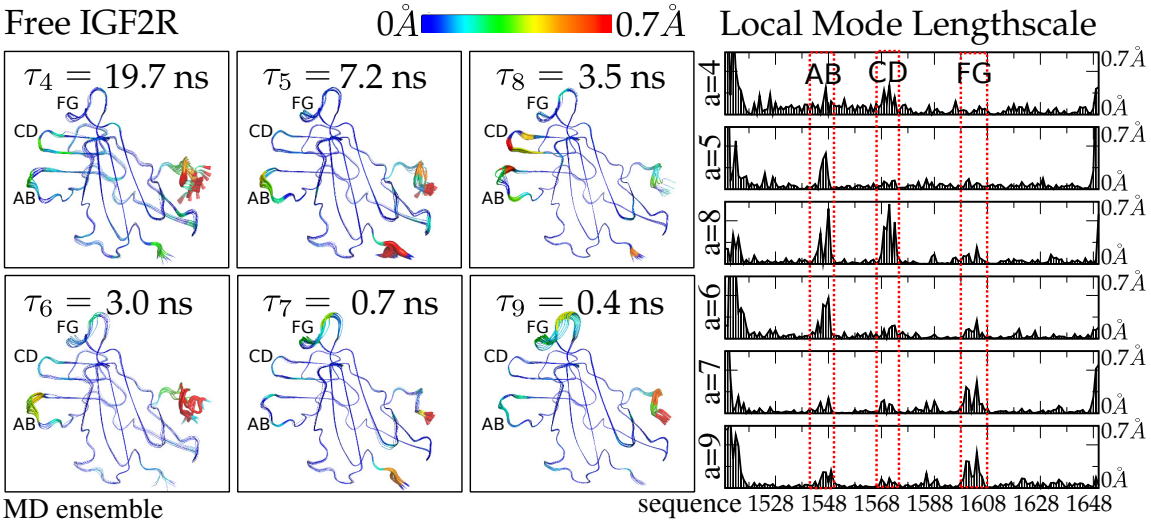

FIGURE 29. Fluctuations of the free IGF2R domain 11 protein from the MD ensemble.
Fluctuations of the free IGF2R domain 11 protein from the MD simulation ensemble. Local mode lengthscale L_{ia} along the primary sequence (Right panel), and projected as a rainbow gradient onto the ensemble of structural minima in the mode (Left panel).

The simulation shows a smaller lengthscale process propagating from the AB loop to the FG loop in the 1 ns $-$ 20 ns timescale as can be seen in Figure 29 , and in agreement with predictions from the NMR conformer ensemble. Furthermore, a new dynamical mode emerges in this intermediate timescale, which consists of correlated motion between the AB and the CD loops, as shown in right panel of Figure 30. This motion is not present in the NMR structural ensemble. When considering the

different timescales of the two calculations, the information collected from the two ensembles is consistent and complementary. Interestingly, the evolutionary analysis by Williams et al. indicates that the structure and function of the IGF2 binding site on IGF2R, i.e the CD loop, is conserved in the mammal common ancestor.

From the MD simulation we extracted the mode-dependent free energy surfaces. These surfaces display quite clearly protein fluctuations between different energetic minima. We analyze the eighth mode, represented in Figure 30, which displays an interesting correlation between the binding region CD and the AB loop. The free energy surface shows conformational transitions during the crossing between the two metastable end points. Those involve the correlated motion between the AB and CD loop conformations. The pathway is represented as a series of triangles with the red one corresponding to the top of the barrier. The transition path is not isotropic in nature. The conformational selection upon binding is observed by projecting the NMR structures of the holo form of the protein onto the landscape. Analysis of the transition states between metastable minima is important as they would mediate the dynamics of both the recognition process and the relaxation to the bound state.

Upon binding, the dynamics of the loops of IGF2R domain becomes quite different. We solve the LE4PD dynamics starting from the structure of the complex IGF2R:IGF2 from the PDB databank, reported as 2L29.[127] In the complex the dynamics of the three binding loops, AB, CD, and FG, become quenched to a large extent. As can be seen in Figure 31, in the bound state these loops maintain a small amount of flexibility but lack the cooperative long-time processes observed in the apo protein. The tails of the protein are still mobile, but these fluctuations do not cooperatively propagate to the binding loop regions of the protein. This behavior suggests that configurational fluctuations observed in the apo protein are trapped in

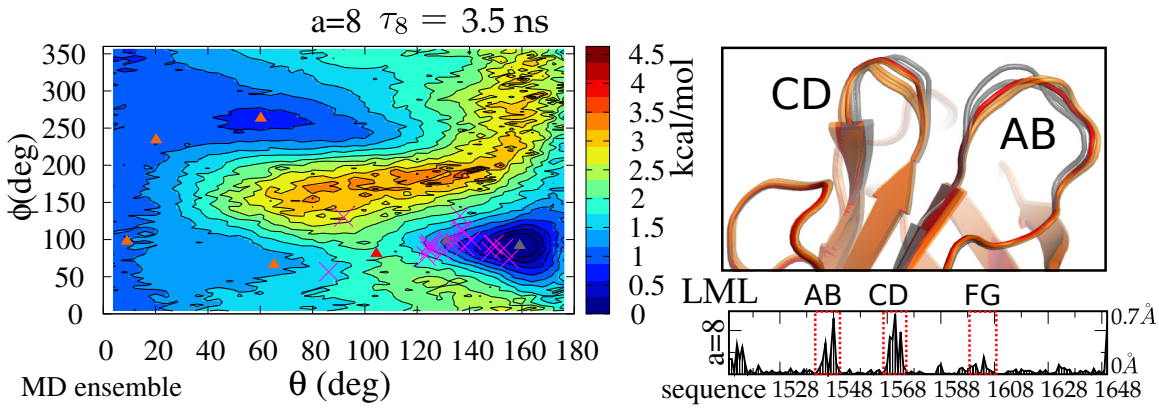

FIGURE 30. Free energy surface in the $a = 8$ internal mode of the IGF2R protein. Orientational free energy surface in the $a = 8$ internal mode of the IGF2R protein from a 100ns MD simulation conducted starting from the first conformer structure in the 2M6T ensemble. Structural conformations shown on the right are from the locations marked by triangles on the free energy landscape. Transition state region is depicted in red. The conformers of the holo protein are projected on the landscape as magenta crosses.

the bound complex. The largest fluctuations now take place in the IGF2 substrate protein. These findings are largely reproduced in the model constructed from the MD ensemble, as shown in Figure 32.

Conclusions

Residue specific localized dynamics involves the sampling of multiple conformations and of the transitions between them. Ensembles of conformational states are localized in mobile parts of the protein that are mostly loops and terminal regions. Local flexibility of the protein in selected parts of its three dimensional structure is needed to allow for the "trapping" of the conformational state useful for substrate binding, following the conformational-selection model of ligand binding.[69] The study shows that rather than ensuring the presence of specific metastable states that faithfully are present in the bound conformation, the conformational diversity in the unbound protein allows the presence of configurational states that are involved

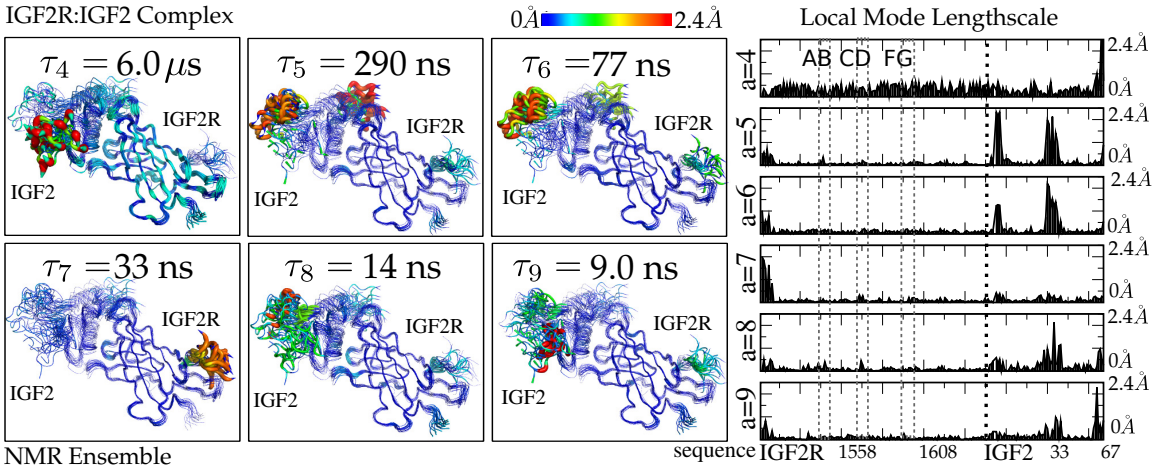

FIGURE 31. Fluctuations of the bound IGF2R domain 11 protein, NMR ensemble. Fluctuations of the bound IGF2R domain 11 protein, as calculated from the NMR conformer ensemble. Local mode lengthscale L_{ia} along the primary sequence (bottom), and projected as a rainbow gradient onto the ensemble of conformer structures 2L29 (top).

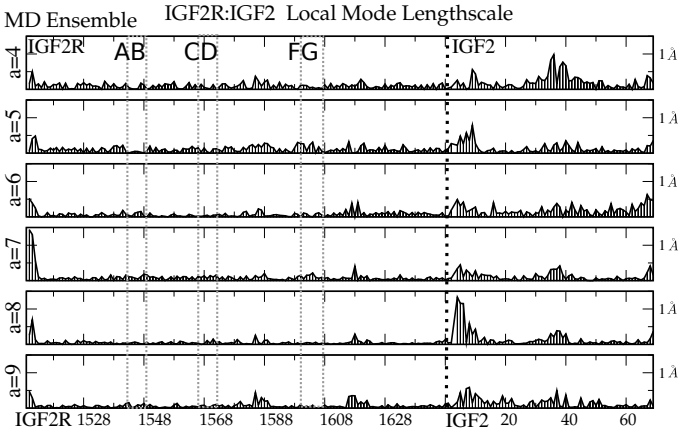

FIGURE 32. Fluctuations of the bound IGF2R domain 11 protein, MD ensemble. Fluctuations of the bound IGF2R domain 11 protein. Local mode lengthscale L_{ia} along the primary sequence as calculated from the MD ensemble.

in binding, and which ultimately relax upon binding to the desired configurational states. In this way the observed mechanism suggests firstly a conformational-selection mechanism that traps configurations, which subsequently relax to optimal structures for binding, following an induced-fit type of mechanism. Conformational diversity in the apo sturucture ensures the presence of energetically efficient binding pathways, and the dynamics of motion between these metastable states.

Local dynamics are investigated with the LE4PD approach, a coarse-grained description that accounts for local semi-flexibility, energy barriers, hydrodynamics, and the chemical structure of the protein.[112] It is important to notice that the LE4PD has as its input information the ensemble of configurational states that are detected either in an all-atom Molecular Dynamics simulations, which contains solvent molecules and counterions, or directly from NMR experiments. For this reason the properties at the coarse-grained level are informed by and reflect the properties at the atomistic-level, including hydrogen bonding, sulfur bridges, and more in general all short and long-ranged interactions.

In its diffusive mode picture, the LE4PD defines cooperative motion in localized regions of the protein. The relationship between cooperativity and energy barriers implies a sequence of relevant steps occurring to support the biological processes. In this way, the LE4PD theory has the ability to predict the emergence of localized fluctuations where mobility is enhanced at a given length and timescale, starting from the protein structural ensemble.

The most relevant information for the kinetics of binding is provided by the low index modes of motion as these large-amplitude, slow modes of motion identify cooperative fluctuations which are involved in the dynamics of recognition and binding of proteins to substrates. These slow dynamics are deemed relevant to the kinetic

selection of protein binding partners as it detects regions where amino acids move cooperatively. The dynamics for these modes are characterized by relatively small barriers that are easily crossed in the nanosecond to several microsecond timescale that is relevant to protein recognition dynamics. As a consequence, the barriers can act as a switch that allows or forbids the transition between states through the precise modulation of the delicate balance of entropy and enthalpy.

Regions of the activated dynamics are in general directly related to sequences that are conserved inside the family of proteins that perform a specific biological function. In previous work we showed how the predicted barrier-activated regions along the protein sequence corresponded to locations of signal transduction mechanisms of the CheY protein[8] and ubiqitination linkage sites.[17] Here, we show that an even more precise temporal and spatial description of the local dynamic activation can be obtained from the analysis of barrier crossing and mode dependent activated dynamics.

The solvent is a relevant player in these mechanisms of macromolecular binding, as they modulate the kinetics of the process by tuning the ordering of the solvent molecules, i.e. entropic contributions, and by breaking or forming hydrogen bonds, i.e. enthalpic contributions.[76, 77] In the LE4PD the effect of the solvent is implicitly included through friction, hydrodynamic interaction, and effective energy barriers. Because the input configurations are either calculated in an atomistic simulation with explicit representation of the solvent or from the experimental NMR conformers, a realistic description of the effect of the solvent is included in the projected dynamics of the LE4PD.

In general we observe that when the protein binds to a substrate, the original regions of energy activated motion are involved in the binding reaction mechanism

and their dynamics becomes quenched. In the protein-substrate complex we observe the quencing of the motion in the single protein enhanced dynamics regions, and the emergence of new regions of higher flexibility in parts of the protein that are remote with respect to the binding interface, following an allosteric mechanism. This emergence of new entropic states balances the reduction of entropy due to binding. It is from the sophisticated balance of all these energetics, which to a large extent tend to compensate each other, that the physiological reaction pathways emerge in the biological mechanisms that regulate the function of proteins. These pathways must be evolutionarily tuned to avoid kinetic traps and ensure that binding partners can find their bound conformation.

This study of the dynamics provides insight into protein recognition, which involves cooperative motion localized to active regions of the protein with fluctuations occurring over a hierarchy of length and timescales. Slow, correlated, spatially localized, fluctuations display a dynamical pathway which is relevant to the biological mechanism and function, while fast, uncorrelated fluctuations indicate simple entropic compensation after protein binding.

CHAPTER V

A HIERARCHICAL FREE ENERGY LANDSCAPE CONNECTS ORDERED AND DISORDERED PROTEIN STATES

Proteins are molecular machines whose structure and dynamics have been evolutionarily designed to perform functional roles. While random sequences of amino-acids exhibit strong-disorder and frustration, native sequences possess funnel-like free energy landscapes, and at physiological temperature can reversibly fold to unique global configurations.[49, 50, 68] However, the ordered, folded state must possess dynamical pathways allowing the motion required for biological function. We have recently developed a Langevin equation for protein dynamics (LE4PD), which can accurately predict the dynamics of folded proteins from the input structural ensemble. Results have been quantitatively compared to measurements of NMR relaxation,[8, 17, 112] and biological activity.[128] In the short-time regime, proteins fluctuate around a single structural minima characterized by the topology of protein connectivity.[9] In the long-time regime, proteins transition between metastable states which are of nearly equal free energy; these are the biologically relevant modes which come into play in processes such as protein-protein recognition and enzymatic activity.[110, 128]

To quantitatively capture the timescales of motion, the LE4PD accurately describes the dissipation and the cooperativity of protein fluctuations. The model is an effective linear description built by finding the effective intramolecular harmonic interactions which stabilize the folded state, obtained from the structural ensemble derived from an underlying model. We have shown how this underlying model can be taken from experimental structural ensembles, such as NMR conformer ensembles, or

from explicit solvent molecular dynamic (MD) simulations. MD simulation ensembles were found to generate the most accurate models when compared to experiment,[112] and these models for six different single-domain globular proteins are studied here.

The effective linear description allows a solution in a set of diffusive modes which specifically relate a set of cooperatively rearranging amino acids in the protein. The scaling properties of the mode solution leads to a subdiffusive scaling regime in the protein mean-squared displacement. When the MD trajectories are projected onto these modes of motion, we find an additional scaling between the cooperativity, or mode lengthscale, and the typical mode-dependent energy barrier. With the knowledge of the relevant energy barrier to diffusion, the diffusive modes can be rescaled to account for complexity on the protein free energy landscape.

To develop and analyze the coarse-grained LE4PD model, we performed explicit solvent molecular dynamics simulation of six different globular single-domain proteins, to obtain $\sim 100ns$ of equilibrium trajectory as monitored by a steady root-mean-squared displacement. Starting structures were taken from NMR or crystal structures of Ubiquitin (1UBQ),[32] HIV Protease monomer (1Q9P),[93] Kinase-associated protein phosphatase KAPP (1MZK),[89] N-TIMP-1 (1D2B),[86] Cellular retinol-binding protein I CPB1 (1MX7),[88] and Insulin Growth Factor 2 Receptor IFG2R domain 11 (2M6T).[91] These trajectories were used as structural ensembles to generate LE4PD models of the protein dynamics. These LE4PD models were shown to quantitavely predict NMR relaxation measurements, with correlation coefficient $\rho = .95$ over the 1864 site-specific measurements.[112]

We find the free energy landscape around the folded state to be like an onion with many layers; the complexity is hierarchical in nature. In this work we analyze the observed energy barriers on the protein free energy landscape in the LE4PD modes,

and the general consequences for dynamical properties. Though the sampling in our MD simulations is limited to the time regime $t < 100ns$, extrapolation to longer time and larger scale processes is shown to directly connect folded and unfolded disordered states of the protein. We also make qualitative comparison to protein folding times, and quantitative comparison to long-time folding and re-folding MD simulations of ubiquitin above its folding temperature.[37]

Subdiffusive Motion

The first three global modes of the LE4PD describe the rotations of the folded structure as the rotational diffusion tensor. The internal fluctuations of the protein is spanned by the internal modes labeled $p = 1, N_p$ with $N_p = N - 3$ because of the 3 rotational modes of the global structure (translational modes have been removed by going into bond coordinates). While the LE4PD model is site-specific and protein-specific, the general scaling between the mode number p and mode lengths $L_p \propto p^{-\beta}$, and diffusive mode relaxation times $\tau_{p,0} \propto L_p^{\alpha}$, predict the general properties of the chain dynamics.

The standard Rouse treatment of the freely jointed chain leads to $\beta = 1$ and $\alpha = 2$; and the Zimm treatment of the hydrodynamic coupling leads to $\alpha = \frac{3}{2}$.[18] In the application to proteins, both exponents are altered by the structural characteristics of the folded protein state in contrast to an ideal chain. Globular proteins have an increased connectivity beyond the linear connectivity of the protein backbone, and are characterized by a fractal dimension in $1 < d < 3$ dimensions. Specific considerations of the folded state of proteins as fractal objects has led to the conjecture that folded proteins are poised on the edge of metastability where the space-filling dimension is $d_f > 2$ and the the spectral dimension is $d_s < 2$,[129] while analysis of the vibrational spectrum of globular proteins predicts a general $d \sim 2$ dimensional

folded state.[130] Our results also show that the internal fluctuations of a globular protein scales algebraically with mode number, and with greater stability than the fully disordered freely jointed chain. Plotting mode lengths for all six proteins in the right panel of Figure 33 shows that in the LE4PD, the mode length scales with internal mode number p with the global fit $L_p = L_0 p^- \beta$ with $\beta = .42$. At large mode numbers (processes faster than tens of picoseconds), this scaling changes, but is quite consistent for all six proteins.

Hydrodynamic coupling alters the α exponent. As in the study of Granek,[26] we find that the including hydrodynamic coupling alters the scaling between the diffusive relaxation times and mode number. Figure 39 shows that the diffusive mode timescale $\tau_{0,p} = C_\tau L_p^\alpha$ (red line) with $\alpha = 2.38$ and $C_\tau = 3651 \frac{ps}{nm^\alpha}$ at 298K. Without hydrodynamic coupling $\tau_{0,p} = C_\tau L_p^\alpha$ with $\alpha = 2.65$ and $C_\tau = 8890 \frac{ps}{nm^\alpha}$ at 298K. In a similar fashion as for the Rouse model, long-range correlations induced by the hydrodynamic coupling enhances long-wavelength transport in proteins.

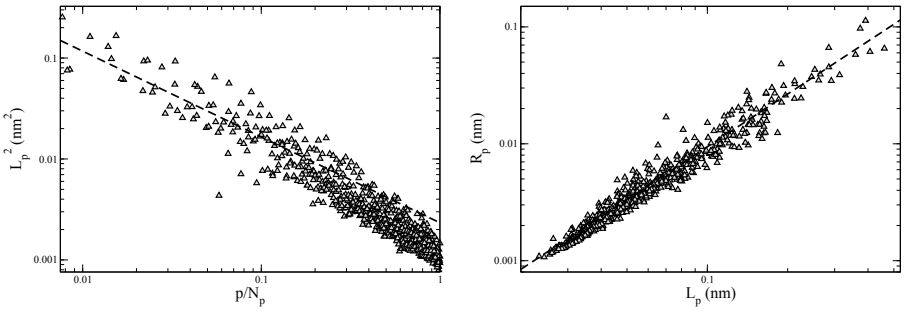

FIGURE 33. Mode length scaling.
Left: Mode length scales with internal mode number p as $L_p = L_0 p^{-\beta}$ with $\beta = .42$. This global fit is set by the scaling for the more collective modes; at large p the scaling changes. Right: Bead position lengthscale R_p scales with bond orientational mode lengthscale L_p as $R_p^2 = C_s^2 L_p^3$ with $C_s = .30 nm^{-\frac{1}{2}}$.

The α and β exponents determine general properties of the dynamical system, such as the subdiffusive exponent of the mean-squared displacement due to

configurational diffusion. The protein mean-squared displacement in the center of mass frame can be written as

$$MSD = \frac{1}{N}\sum_{i=1}^{N}\langle[\vec{R}_i(t) - \vec{R}_i(0)]^2\rangle = \frac{2}{N}\sum_{i=1}^{N}\left(\langle\vec{R}_i^2\rangle - \langle\vec{R}_i(t)\cdot\vec{R}_i(0)\rangle\right) \tag{5.1}$$

In this protein centered frame, the site-site correlation function can be defined in terms of relative distances

$$\frac{1}{N}\sum_{i=1}^{N}\langle\vec{R}_i(t)\cdot\vec{R}_i(0)\rangle = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=i+1}^{N}\langle(\vec{R}_j(t) - \vec{R}_i(t))\cdot(\vec{R}_j(0) - \vec{R}_i(0))\rangle \tag{5.2}$$

Noting that the bead separation vector can be written as $(\vec{R}_j - \vec{R}_i) = \sum_{k=i}^{j-1}\vec{l}_k$, and expanding into modes,

$$\frac{1}{N}\langle\vec{R}_i(t)\cdot\vec{R}_i(0)\rangle = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=i+1}^{N}\sum_{k,l=i}^{j-1}\sum_{a=1}^{N-1}Q_{ka}Q_{la}\langle\vec{\xi}_a(t)\cdot\vec{\xi}_a(0)\rangle \tag{5.3}$$

the bead position mode lengthscale $R_a^2 = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=i+1}^{N}\sum_{k,l=i}^{j-1}Q_{ka}Q_{la}L_a^2$ is obtained by effectively integrating the orientational modes along the chain; the bead position internal mode lengths R_p then scale with bond orientational mode lengthscale L_p as $R_p^2 = C_s^2 L_p^3$ with $C_s = .30nm^{-\frac{1}{2}}$, as can be seen in the left panel of figure 33. In the protein reference frame (fixing rotations and translations) the global modes cancel out of the MSD calculation and the $MSD = \sum_{p=1}^{N-1}R_p^2(1-\exp[-\frac{t}{\tau_p}])$. Approximating R_p and τ_p with their scaling forms, and the discrete sum as an integral, we obtain

$$MSD = 2C_s^2L_1^3\int_{p=1}^{N_p}p^{-3\beta}\left(1 - \exp\big[-\frac{tp^{\beta\alpha}}{C_\tau L_1^\alpha}\exp(-\frac{\epsilon L_1 p^{-\beta}}{k_BT})\big]\right) \tag{5.4}$$

For small fluctuations where the largest energy barrier is $\epsilon L_1 \sim k_B T$ or less, the barrier-crossing rescaling can be approximately neglected and the integral in equation 5.4 has the short time expansion $MSD \propto t^{\nu}$ with $\nu = \frac{3\beta-1}{\beta\alpha}$. For the Rouse model, this leads to the standard diffusive exponent with $\nu = 1$. The inclusion of hydrodynamic coupling (Rouse-Zimm) leads to super-diffusive transport with $\nu = \frac{4}{3}$. For proteins, using the obtained values of $\beta = .42$, and $\alpha = 2.38$, we obtain the strongly subdiffusive short-time growth exponent $\nu = 0.27$ in direct agreement with that observed in the MD simulations shown in figure 34. This also agrees with the value of $\nu \sim 0.3$ needed to model thiyl radical recombination experiments of Milanesi et al., this subdiffusive exponent was shown to be a property of the polypeptide backbone independent of sequence.[131]. Without hydrodynamic coupling, the exponent is only slightly reduced with $\nu = 0.24$. The difference in exponent with and without hydrodynamic coupling is small, and is of the scale of the variability in observed exponent between the six different proteins. However, we have shown in previous work how the hydrodynamic coupling strongly perturbs the structure and absolute timescale of the large-amplitude modes, and leads to strongly decreased correlation to experiment when neglected.[112] In general, hydrodynamic effects are large even in globular protein systems and cannot be neglected.

Hierarchical Complexity Due to Cooperativity

Using the diffusive modes, we investigate the nature of the free energy surface of a protein around its folded ground state. The diffusive modes organize the fluctuations of the protein in order of cooperativity. Each diffusive mode obtained from the diagonalization of Eq. 2.5 is a vector defined by the linear combination of the bond vectors weighted by the eigenvectors of the product of matrices **LU**, as $\vec{\xi}_p(t) = \sum_i Q_{pi}^{-1} \vec{l}_i(t)$. In polar coordinates the vector is represented as

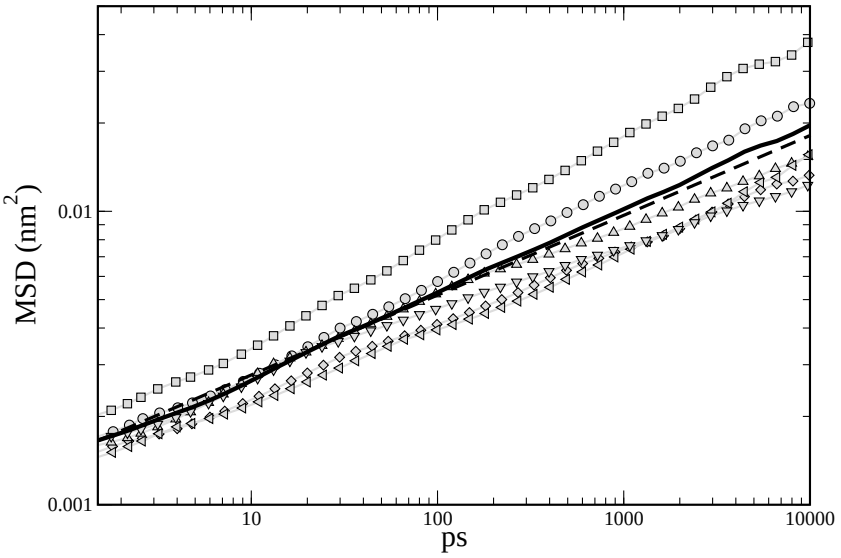

FIGURE 34. Subdiffusive regime of the mean-squared displacement. In the regime where the diffusive motion dominates, i.e. $\epsilon L_1 \sim k_B T$ or less, equation 5.4 predicts a subdiffusive regime with $\langle R^2(t) \rangle \propto t^{\nu}$ with $\nu = 0.27$ (black-dotted line). The average MSD of all six-proteins (solid-line) and the MSD of ubiquitin (circles), protease (quares), KAPP (diamonds), N-TIMP-1 (triangle up), KAPP (triangle left), and IGF2R (triangle down).

$\vec{\xi}_p(t) = \{|\vec{\xi}_p(t)|, \theta_p(t), \phi_p(t)\}$. While the mode length is variable, at any single mode orientation the mode length distribution is single-peaked about a typical value and the most relevant changes in the diffusive mode free energy occur as the angles, expressed in the spherical coordinates, span the configurational space. For any diffusive mode p, the free energy surface is defined as a function of the spherical coordinate angles θ_p and ϕ_p as $F(\theta_p, \phi_p) = -k_B T \log \{P(\theta_p, \phi_p)\}$,with $P(\theta_p, \phi_p)$ the probability of finding the diffusive mode vector having the given value of the solid angle. In general as can be seen in figure 35, the mode landscape contains multiple minima and energy barriers.

To obtain an explicit representation of the structure at a minima of interest, all structures from the simulation ensemble which pertain to a particular θ, ϕ orientation which is a strong minima in the mode free energy is extracted and averaged. By calculating the average structure at each minima we obtain the structural ensemble of metastable states spanning each internal mode of fluctuation for the protein. In figure 35 we show the mode free energy surfaces and the structures corresponding to the metatstable minima on these surfaces. The fluctuations in each mode seem to be spanned by a handful of metastable minima, and the depth of these minima, or the barriers between them, are largest for the low mode numbers corresponding to the most collective, large-amplitude fluctuations. As we progress from the lowest to highest mode number, we see that the fluctuations go from collective in nature to local in nature; this is the hierarchical property of the free energy landscape, as observed in the linear modes of the LE4PD.

The typical energy barrier in each mode, p, is evaluated from the simulation as the Median Absolute Deviation[81] from the global minimum E_{gs}, that is $E_p^\dagger = median_{(\theta,\phi)}(E_p(\theta,\phi) - E_{gs,p})$. The depth of these minima, or the barriers between them, are largest for the low mode numbers corresponding to the most collective, large-amplitude fluctuations. Figure 36 shows that the energy barriers $E_p^\dagger$ as observed in the simulation trajectory can be well described as scaling with the mode length, a measure of the mode cooperativity, over a large range of the protein fluctuations. Figure 36 shows that the observed scaling with mode number follows $E_p^\dagger = \epsilon L_p$ with $\epsilon = 6.42 \frac{kcal}{mol \cdot nm}$. At a local enough length scale, where the specific chemical nature of the amino acid is most important, the energy barriers are no longer described by this expression.

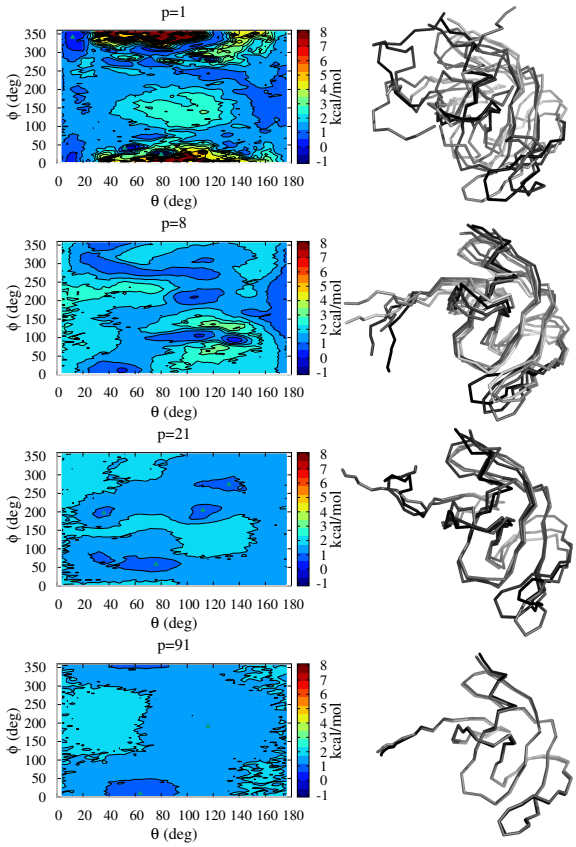

FIGURE 35. Free energy surfaces and structural minima, protease. Internal mode free energy surfaces for modes $p = 1, 8, 21, 91$ on the left and their associated structural minima on the right. Minima from which structures were calculated are marked on the free energy surface with grey triangles. The barriers between minima and the size of the fluctuations spanned by the structural minima decrease as mode number is increased.

The observed scaling law is consistent with the hierarchical nature of the protein free energy landscape. Each mode describes dynamics involving a number of bonds in the protein, which need to move collectively in a cooperative fashion. At short times the bonds fluctuate independently, while large-amplitude correlated fluctuations occur when all the bonds transition collectively.[82] The equilibrium probability for the Z gating bonds to independently transition away from the "correct" orientation with energy preference E is $P(Z) \propto \exp(-\frac{ZE}{k_B T})$. In a transition state perspective this can be interpreted as a free energy barrier which scales proportionally with the number of

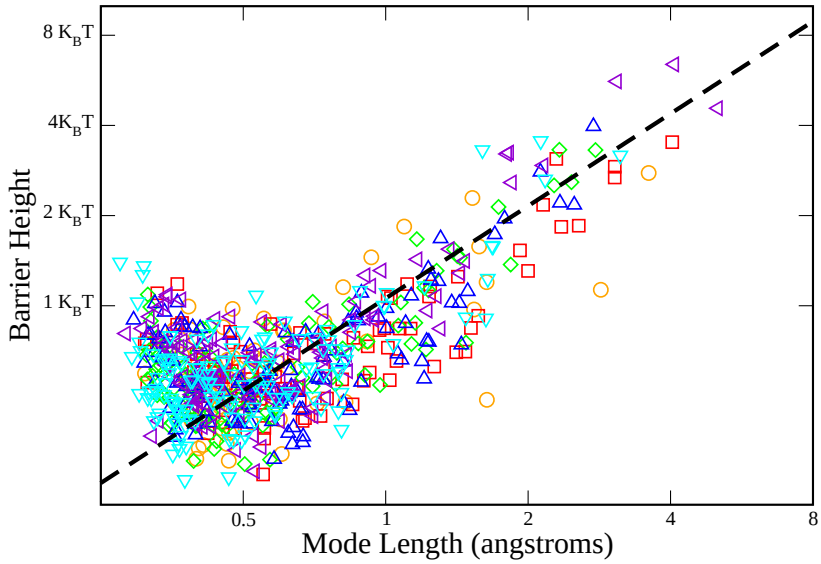

FIGURE 36. Energy barriers and mode lengthscales.
Typical energy barrier $E_p^\dagger$ taken as the median average deviation from the global minimum in the LE4PD mode free energy landscape, plotted against the mode lengthscale L_p for ubiquitin (orange circles), protease (red squares), KAPP (green diamonds), N-TIMP-1 (blue triangle up), KAPP (violet triangle left), and IGF2R (cyan triangle down). Obtained scaling relation $E_p^\dagger = \epsilon L_p$ the black dotted line with $\epsilon = 6.42 \frac{kcal}{mol \cdot nm}$.

bonds cooperatively rearranging. This model is similar to the Adam-Gibbs theory of the glass transition,[83, 84] relating the complex hierarchical nature of the free energy landscape the protein in solution to a structured glassy fluid.[50, 85]

In systems which are dominated by strong disorder, energy barriers have been observed to scale with the size of the system as $E^\dagger = (\epsilon L)^\psi$ in general.[132, 133] We go further, and claim that each mode can be treated as an independent system with energy barriers scaling with the mode size. In our notation, $E_p^\dagger = \epsilon L_p$ but L_p is the mode length, not linear length. The chemical length of the amino acid chain, $L_c = Nl$ with l the nearly constant distance between α-carbons on the protein

backbone, describes the linear 1-dimensional size of the system. This linear length is related to the mode number by $L_c \propto p^{-1}$ since each mode involves roughly $\frac{N_p}{p}$ sites.[18] Thus, in our model, the barrier-exponent ψ is just β, the exponent connecting mode length to mode number discussed in section 5.1.

To account for the affect of the local energy barriers on the internal dynamics, the friction becomes mode dependent by assuming thermal activation over the mode-dependent energy barrier $\langle E_p^\dagger \rangle$

$$\overline{\zeta} \rightarrow \overline{\zeta}\ exp[\langle E_p^\dagger \rangle / (k_B T)]\ , \tag{5.5}$$

leading to the slowing of the mode timescale $\tau_p = \frac{l^2 \overline{\zeta}}{3 k_B T \lambda_p}$ by $\tau_p = \tau_{p,0}\ exp[\langle E_p^\dagger \rangle / (k_B T)]$. This simple dynamical renormalization provides an average correction to the dynamics of the Langevin Equation, which approximately accounts for the local barrier crossing, and is in agreement with free energy landscape theories suggesting activated dynamics.[49, 50] As a first approximation, the depth of the minimum free-energy well in the mode serves as the relevant barrier to transport.

Microscopic Sources of Complexity

What is the microscopic origin of this complexity in the protein free energy landscape? There are many direct physical interactions of importance among a chain of amino acids; short-ranged Van der Waals forces of great importance in packing a stable hydrophobic core, salt-bridges, long-range electrostatic interactions between charged residues, and stacking of aromatic rings to name several. Beyond covalent bonds, the most prolific favorable direct interaction with a strength larger than $k_B T$ at biologically important temperatures is the hydrogen bond.[134, 135] In water, a hydrogen bond network forming fluid,[136, 137] this interaction become

even more ubiquitous. The ϕ, ψ amino acid dihedral angles which characterize the protein backbone, and have significant energy barriers to dihedral rotation, are often considered to be the relevant parameters describing protein structure as they characterize the α-helix and β-sheet secondary structures defining protein folds.[138] However, in aqueous solution the energy landscape of dihedral rotations is in many ways set by the favorability of intramolecular and protein/solvent hydrogen bonds.[134] We isolate and focus on the contribution of the hydrogen bonds in protein systems illustratively, but not in an exclusive sense. All of the myriad of microscopic interactions in proteins can be of significant importance in varying systems and condiitions.

The hydrogen bond is a direct microscopic interaction which reduces configurational freedom, as it requires a specific distance and angle between donor and acceptor. There is a critical temperature regime where the enthalpic gain of the interaction is exactly compensated by the entropic loss.[117, 118] For biological macromolecules, this temperature range not surprisingly is that of liquid water. Thus the hydrogen bonding network is poised at critical stability with the constant formation and breaking of the protein/water and crucial members of the protein/protein hydrogen bonding network. We conjecture that the energy scale and length scale to make and break hydrogen bonds sets the ϵ energy per unit length observed in the energy barrier scaling.

In Figure 37, we show the potential of mean force between hydrogen bond donors and acceptors, $k_B T \log\left(g_{donor-acceptor}(r)\right)$, for both protein-water hydrogen bonds and protein-protein hydrogen bonds. In both cases, the height of the energy barrier ΔF and the distance ΔL required to break a hydrogen bond matches the energy per unit length of the ϵ parameter obtained from the fit to the energy barrier distribution

in figure 36. The energy and length scales considered agree with general studies of hydrogen bond structure and energetics in liquid water.[137]

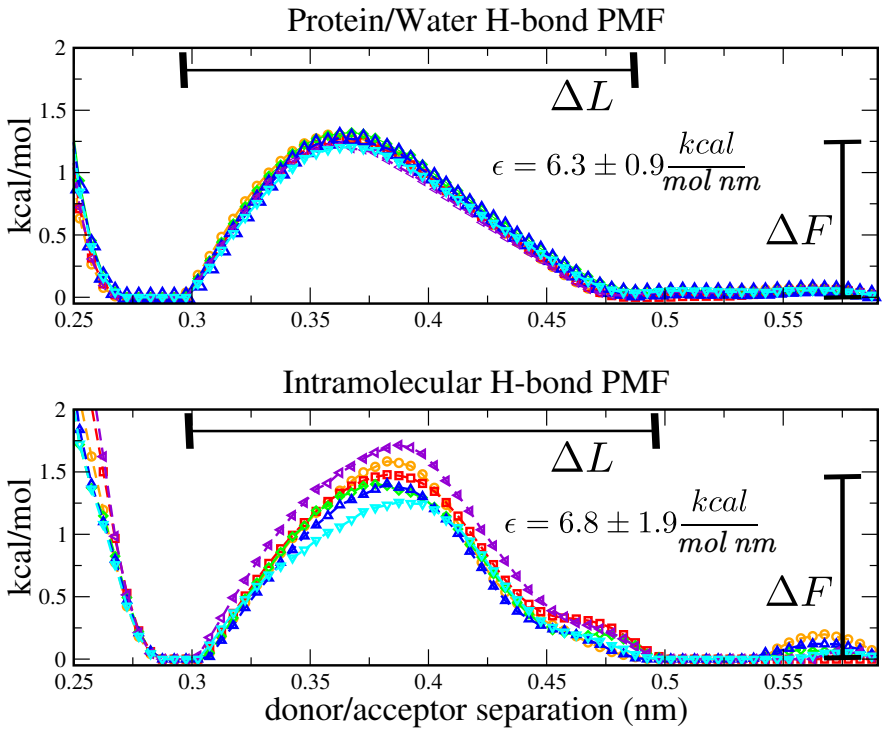

FIGURE 37. Potential of mean force between hydrogen bond donor and acceptors. Potential of mean force between hydrogen bond donor and acceptors, $k_B T \log\left(g_{donor-acceptor}(r)\right)$, normalized with the lowest energy distance set to zero free energy. Solvent-protein PMF (top panel) and Protein-protein PMF (bottom panel). The energy per unit length, $\epsilon = 6.42 \frac{kcal}{mol \cdot nm}$, setting the relation between energy barriers and fluctuation length, $E_p^{\dagger} = \epsilon L_p$, corresponds to the typical energy required to make and break hydrogen bonds.

The entropy-enthalpy compensation of individual hydrogen bonds results in the metastability of protein configurations. Proteins, even near their folded state, have many configurational minima where competing entropy and enthalpy result in configurations with similar free energy. These metastable states are coupled to large changes in the hydrogen bonding connectivity, in particular large rearrangements of the protein-water hydrogen bond network.[76, 77, 139] In figure 38, we show the

free-energy landscape in the mode orientation for the $p = 4$ mode of the IGF2R protein. This protein is responsible for binding and regulating levels of IGF2 at the cell surface,[91] and the apo and holo forms of IGF2R require concerted rearrangements of the AB, CD, and FG binding loops.[127] Figure 38 shows that the $p = 4$ mode is characterized by two metastable configurations of the AB and CD binding loops. Configurational fluctuations are accompanied by large changes in the protein-solvent hydrogen bond network, and subtle changes in the protein-protein hydrogen bond network.

Long Crossover to Activated Regime

For motions very close to the folded state, protein motion is diffusion dominated. However, as the size of the fluctuations increases the barrier distribution dominates the protein behavior. In figure 39 we plot the diffusive timescales and barrier rescaled mode timescales from the LE4PD models of all six proteins. For $\epsilon L_1 < k_B T$ the diffusive and barrier rescaled timescales roughly coincide, but as the fluctuations grow in size the energy barrier correction causes the mode relaxation time to rapidly propagate out to folding timescales at the nanometer lengthscale. This happens to be the typical size of single-domain proteins. The dataset of 2-state folding times for 52 proteins, plotted against the protein radius of gyration, sits right around the predicted line for τ_p with energy barriers included.[5] R_g is not particularly predictive of individual folding times, but using the protein size as the lengthscale in the mode timescale $\tau_p = C_\tau L_p^\alpha \exp[\frac{\epsilon L_p}{k_B T}]$ does capture the ballpark folding times of these proteins. It should be emphasized that the mode timescales were obtained without adjustable parameters and the LE4PD description was developed to be quantitative in the picosecond-nanosecond time regime. The fact that these timescales can be extended to the length and timescales of protein folding times is surprising, illustrating

that the hierarchical roughness of the free energy landscape connects folded and disordered protein configurations.

To investigate the relaxation of proteins in the long-time regime, we utilize the published results of Lindorff-Larsen et al.[37] who performed multiple millisecond regime simulations of the Ubiquitin protein above its melting temperature, and were able to reversibly sample unfolding and refolding events. We conducted a brief $100ns$ simulation using the same conditions to validate the short-time predictions of the LE4PD theory. The theory is not optimized to the simulation, but directly takes the thermodynamic parameters and solvent viscosity at the simulation temperature of 390K as input. In figure 40 we present the RMSD autocorrelation function

$$RMSDACF = \left(\frac{1}{N} \sum_{i=1}^{N} \frac{\langle \vec{R}_i(t) \cdot \vec{R}_i(0) \rangle}{\langle \vec{R}_i^2 \rangle} \right)^{\frac{1}{2}} \tag{5.6}$$

from our short-time simulations, and prediction from equation 5.4 using the obtained L_1 from the simulation. We also present the published result from the long-time simulations of Lindorff-Larsen et al., compared to the prediction from equation 5.4 using $L_1 = 1.5nm$ which is the radius of gyration of the the structural ensemble obtained in their simulations. At both long and short times the theoretical prediction of the LE4PD from equation 5.4 and the direct calculation from MD simulation quantitatively agree, using only the lengthscale of the largest fluctuation as input. In general this long time regime is set by a very slow logarthmic growth of the protein MSD; at long times it can be fit by the general coarsening law for disordered systems, $L(t) \propto \log[t]^{\frac{1}{\beta}}$.

Surprisingly the hierarchichal nature of the protein free energy landscape observed in the lengthscales accessed at short times $t < 100ns$ seems to propagate out to the lengthscale of the entire protein, at the timescales of protein folding. This

scaling of energy barriers with fluctuation lengths sets an upper limit on the domain size in proteins. The size distribution of protein domains found in biology is peaked at around 100 amino acids and near zero by 300 amino acids.[140, 141] This corresponds to proteins of maximum size $R_g \sim 2.0nm$, and by the expression obtained in this work, a longest relaxation time of $\sim$ 1 minute. Protein domains larger than the typical size found in nature we predict to have relaxation times exceeding biologically relevant timescales.

Connecting to Disordered Systems

We have shown how the microscopic CG model of the LE4PD predicts protein diffusive properties, which when coupled to the barrier distribution propagates microscopic relaxation times out to the protein-folding regime. But what is the origin of this roughness in the free energy landscape of proteins? The missing component in the diffusive description is disorder; proteins are intrinsically frustrated objects. However, well-folded globular proteins obey the principle of minimal frustration;[50] so why are the ground states of proteins disordered as well? The answer may be that biology is poised at the critical stability of the direct chemical interaction of the hydrogen bond, and the source of the disorder is the constantly shifting hydrogen-bonding network.

Explicit inclusion of time-dependent disorder qualitatively changes the resulting dynamics of linear diffusive descriptions. In particular, the description becomes non-equilibrium; the probability distribution resulting from the equation of motion is no longer $P(\mathcal{V}) \propto \exp(-\frac{1}{k_B T}\mathcal{V})$ but is governed by a now time-dependent dynamical partition function. The protein (at-least foldable proteins of typical size) are still equilibrium systems, it is just that when attempting to coarse-grain over the atomistic

degrees-of-freedom that spatial and temporal disorder affect the locally harmonic energy landscape. Explicit evaluation of the affect of disorder in the LE4PD equation 2.5 is beyond the scope of this work. However, it is known that the isotropic continuum example of a general elastic manifold embedded in a field of random pinning potentials results qualitatively in hierarchical free-energy landscapes of the type observed here for proteins.[142]

To describe directly the simplest model which captures the affect of disorder in the dynamics of proteins, we assume that every single state of the hydrogen-bond connectivity has a unique configuration of the protein backbone which is the structural minima associated with the connectivity. On the timescale of picoseconds, that hydrogen bond connectivity is being constantly perturbed; and the protein structure is constantly in an energetic battle between relaxing to the new structural minima and satisfying the random energy perturbations. The picosecond timescale is also the timescale implied by local elastic fluctuations to a single structural minima, such as those observed in the crystalline state.[8] Since we are interested in describing the affect of the disorder on only site-averaged properties, like the MSD, we describe the configurational state by the CG coordinate of the site-averaged perturbation from the average protein structure $\vec{r}(t) = \frac{1}{N}\sum_{i=1}^{N} \vec{R}_i(t) - \langle \vec{R}_i \rangle$. The motion of this vector $\vec{r}(t)$ is governed by an energy functional reflecting the balance between a transverse elastic energy cost, and the random energy perturbations induced by the shifting HB connectivity. In the spirit of writing the simplest equation that captures the properties of motion at times long compared to the timescale of the random energy perturbations and local elastic relaxation, or $t > ns$, we could describe the dynamical

process as obeying the simple energy functional

$$\mathcal{V}(t) = \frac{1}{4\Gamma}\left(\frac{d\vec{r}}{dt}\right)^2 - \eta(\vec{r},t) \tag{5.7}$$

where the random energy perturbations are white noise in space and time; that is $\langle\eta\rangle = 0$ and $\langle\eta(\vec{r},t)\eta(\vec{r}',t')\rangle = \eta^2\delta_{\vec{r}-\vec{r}'}\delta(t-t')$. Equation 5.7 is the directed polymer in random media (DPRM).

This is the simplest model of the properties of finite-temperature motion on a d-dimensional random energy landscape. For proteins we have discussed in section 33 the effective dimensionality is $d \sim 2$. This corresponds to the number of d transverse directions in the $d+1$ DRPM (the extra dimension is time). We conjecture that the DPRM, which describes the motion of a point particle in a local elastic potential with the addition of uncorrelated random perturbations, effectively describes the scaling behavior of the average motion of a tagged particle embedded in the folded protein, mapping protein dynamics to the DPRM in roughly (2+1) dimensions.

The DPRM has the dynamical partition function

$$\mathcal{Z}(\vec{x},t) = \int_{0,0}^{\vec{x},t} \mathcal{D}\vec{x}\exp\left[-\int_0^t ds\mathcal{V}(s)\right] \tag{5.8}$$

and differential form

$$\frac{d\mathcal{Z}}{dt} = \Gamma\nabla^2\mathcal{Z} + \eta\mathcal{Z} \tag{5.9}$$

There is one more illustrative mapping; applying the transformation

$$\mathcal{Z}(\vec{x},t) = \exp\left[\frac{\lambda}{2\Gamma}h(\vec{x},t)\right] \tag{5.10}$$

to equation 5.9 results in

$$\frac{dh(\vec{x},t)}{dt} = \Gamma\nabla^2 h(\vec{x},t) + \frac{\lambda}{2}\left|\vec{\nabla}h(\vec{x},t)\right|^2 + \eta'(\vec{x},t) \tag{5.11}$$

where $\eta' = \frac{2\Gamma}{\lambda}\eta$, and equation 5.11 is the celebrated Kardar-Parisi-Zhang (KPZ) equation describing the surface height of a growing solid.

The transformation equation 5.10 says that $\log[Z(\vec{x},t)] \propto h(\vec{x},t)$, or the effective free energy of the DPRM is directly proportional to the surface height of a growing solid.[143, 144] We conjecture that the origin of the universal energy barrier distribution in the free-energy landscape observed in this work for six different proteins is the rough surface-height distribution of the KPZ universality class. We note that the roughening exponent $\chi = .39$ of the $d = 2$ KPZ-equation is within numerical precision of the scaling of the exponent β observed for the scaling of free-energy roughness with chemical length $E_p^\dagger = (\epsilon L_c)^\beta$. The early-time growth exponent $L(t) \propto t^\nu$ of the (2+1) DPRM is numerically known to be $\nu = .24$.[145, 146] In the simulations, the early-time growth exponent of the RMSD was found to be $\Delta R(t) \propto t^\nu$ with $\nu = .27$, in agreement with the LE4PD with hydrodynamic coupling. However, in the LE4PD, the hydrodynamic coupling can be turned off, resulting in $\nu = .24$. The fact that the roughening and early-time growth exponents correspond to the KPZ universality class could be merely coincidental. However, examining the mapping between proteins and generalized systems with strong disorder is an attractive avenue to explain the features of protein dynamics observed in this work.

Discussion

The LE4PD was developed and shown to quantitatively predict protein dynamics local to the folded state of the protein. It is necessary to include the energy barriers in the modes of this linear description to quantitatively reproduce the experimental data of NMR relaxation and time correlation functions calculated directly from MD simulation. Here we show that the free-energy roughness distribution in the modes of the LE4PD can be extended to connect ordered with disordered protein states, and has relevance out to the scale of protein folding times.

Analyzing the scaling behavior of the mode lengths, timescale, and energy barriers in the LE4PD description, we predict a subdiffusive regime of protein motion at short times (in agreement with MD simulation and experiment) and a long crossover to disorder dominated relaxation at long times (also in agreement with millisecond scale simulations performed on the Anton supercomputer). The resulting paradigm is a hierarchical free-energy landscape connecting ordered and disordered protein states.

The source of this energy barrier scaling may be the presence of time-dependent disorder. For folded proteins which have evolved to be minimally frustrated with funnel-like energy landscapes, we conjecture that the source of this disorder is the constant breaking and formation of hydrogen bonds in the critical temperature regime of biological phenomena where enthalpy and entropy balance. This leads to multiple configurations local to the folded state which differ greatly in the protein-solvent hydrogen-bonding network, with subtle rearrangement of the protein-protein hydrogen bond network.

The predicted roughening and short-time growth exponents are very close to those of the $d = 2$ KPZ universality class. We conjecture that this is because the

constant perturbation of the hydrogen bond network results effectively in noise in the CG Hamiltonian, which maps the problem to that of a directed polymer in random media whose free energy is mapped to the KPZ equation for a growing surface.

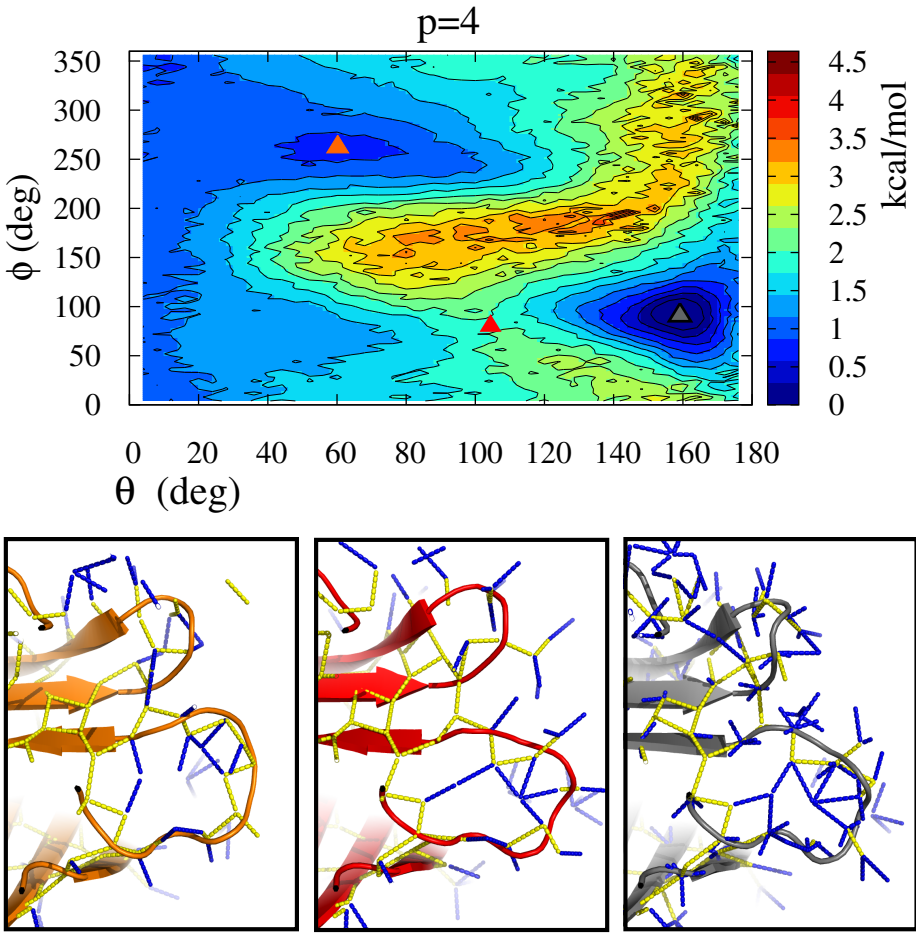

FIGURE 38. Structural minima and hydrogen-bond network.
Top: $p = 4$ orientational mode free energy landscape of the IGF2R protein, involving concerted fluctuations of the the AB and CD binding loops. Structural minima and transition state labeled with colored triangles. Bottom: Corresponding configurational snapshot from MD simulation in each mode orientation labeled in the top panel. Configurational fluctuations are accompanied by large changes in the protein-solvent hydrogen bond network, and subtle changes in the protein-protein hydrogen bond network.

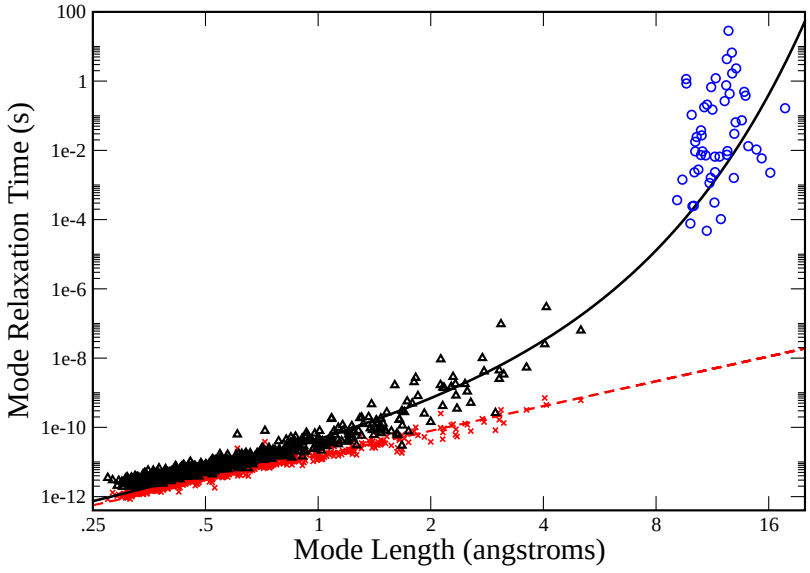

FIGURE 39. Mode timescales and lengthscales propagate to protein folding scales. LE4PD formalism built from the structural ensembles of the MD simulations predicts a scaling of the mode timescales for all six proteins (red crosses) before energy barrier correction $\tau_{0,p} = C_\tau L_p^\alpha$ (red line) with $\alpha = 2.38$ and $C_\tau = 3651 \frac{ps}{nm^\alpha}$ at 298K. Energy barrier corrected mode timescales of all six proteins (black triangles) and the scaling $\tau_p = C_\tau L_p^\alpha \exp[\frac{\epsilon L_p}{k_B T}]$ (black line). Microscopic timescales of sub-angstrom fluctuations rapidly propagate out to folding timescales at the lengthscale of single-domain proteins. Folding timescales for 52 2-state globular proteins is plotted against the protein radius of gyration (blue circles).[5]

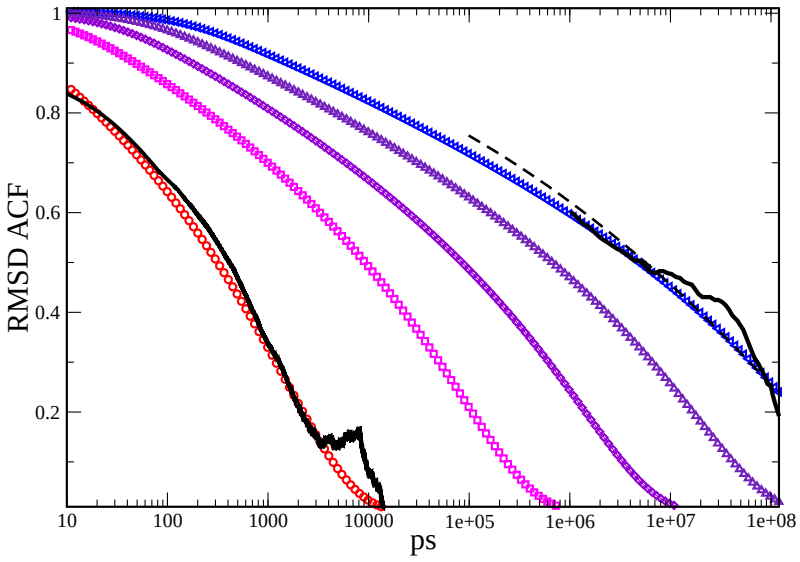

FIGURE 40. RMSD autocorrelation function of the ubiqutin protein at 390K. RMSD autocorrelation function of the ubiqutin protein at 390K, above the folding temperature. Multiple millisecond scale simulations of Shaw et al. showed reversible unfolding and folding starting from the folded configuration. RMSD autocorrelation function from the Shaw simulations, the black solid line at $t > 10^6 ps$. RMSD autocorrelation function from $100ns$ simulation performed in identical conditions, black solid line at $t < 10^4 ps$. RMSD autocorrelation function from integral 5.4 at the $100ns$ simulation derived $\langle \vec{\xi}_1^2 \rangle^{\frac{1}{2}} = 4\mathring{A}$ (red circles). Folding timescales are reached when the largest internal mode length reaches the lengthscale of the protein as derived from the Shaw simulations $\langle \vec{R}^2 \rangle^{\frac{1}{2}} = 15\mathring{A}$ (blue triangle left). Intermediate scale $\langle \vec{\xi}_1^2 \rangle^{\frac{1}{2}} = 7.5, 10, 12.5\mathring{A}$ (magenta squares, violet diamonds, and indigo triangles up). The long-time behavior can be fit by the general coarsening law $1 - A \log[t]^{\frac{1}{\beta}}$ with $A = .00077$ (black dashed line for $t > 10^5 ps$).

CHAPTER VI

DISCUSSION

The biophysical community is shifting from a static view of macromolecular structure towards one dominated by the concept of the structural ensemble. The series of four papers presented in this dissertation have developed and validated an effective linear description of protein motion. The utility of the model presented lies in that it is perhaps the simplest model capable of quantitatively capturing system-specific and site-specific protein motion, as derived directly from the structural ensemble. The dynamical model quantitatively agrees with available experiments, and the large-amplitude slowest motion predicted in the dynamical modeling directly correlates to biological function.

The method explicitly separates local and global processes, and diffusive and activated contributions to protein dynamics. A general relationship between cooperativity and complexity allows the approximate inclusion of energy barrier rescaling to the diffusive timescales, resulting in a subdiffusive regime in the configurational diffusion at short times crossing over to a barrier-activated regime at long-times. However, the treatment of large-scale configurational rearrangements such as those involved in a specific allosteric transition or protein folding pathway must be made more rigorous. Using the diffusive modes of this model as reaction coordinates for path-based and Markov-chain models would be a start to more explicitly derive biologically relevant large-scale motions within the formalism.

Evolution has taken advantage of the unique properties of a specific amino acid chain in water to fine-tune intrinsic fluctuations related to biological function. Inherent to each individual protein $1-3$ nanometers in size there are specific functional

dynamical processes spanning the microscopic to macroscopic in timescale. Yet on average these proteins are members of a dynamical, if not universality, then generality class, bridging a diverse array of sequences and forms. This coexistence of specificity and universality is part of what makes these molecules fascinating.

Zooming out from the length-scale of a single atom, we find that the free energy landscape of a protein grows in complexity and the cooperativity of the allowed motions. Universality generally arises from the simplicity of the physical description as one coarse-grains from the microscopic to macroscopic, and dynamically a separation in microscopic and macroscopic timescales results in a hydrodynamic continuum limit. But even in the simplest building block of biology, the equilibrium system of a single protein, the trend seems to be going in the opposite direction.

Is this a general feature of biological molecular systems, where as the level of coarse-graining increases more and more complex and interconnected is the behavior? Zooming out beyond the single-protein lengthscale in the dynamical system of living matter, there is no scale invariance, no fixed points in the dynamical system, with spatial and temporal organisation at the molecular systems level, cell level, organism, genetic, and ecosystem level. Yet hidden in all of the specificity, are there concepts of universality applicable? Are there distinct combinations of universality and specific function yet to be discovered when coarse-graining from the microscopic to macroscopic in biological systems?

REFERENCES CITED

[1] Nico Tjandra, Scott E Feller, Richard W Pastor, and Ad Bax. Rotational diffusion anisotropy of human ubiquitin from 15n nmr relaxation. *J. A. Chem. Soc.*, 117(50):12562–12566, 1995.

[2] David Komander. The emerging complexity of protein ubiquitination. *Biochem. Soc. Trans.*, 37(Pt 5):937–953, 2009.

[3] SF Lienin, T Bremi, B Brutscher, R Brüschweiler, and RR Ernst. Anisotropic intramolecular backbone dynamics of ubiquitin characterized by nmr relaxation and md computer simulation. *J. A. Chem. Soc.*, 120(38):9870–9879, 1998.

[4] Linda K Nicholson, Toshimasa Yamazaki, Dennis A Torchia, Stephan Grzesiek, Ad Bax, Stephen J Stahl, Joshua D Kaufman, Paul T Wingfield, Patrick YS Lam, Prabhakar K Jadhav, et al. Flexibility and function in hiv-1 protease. *Nature Struct. Biol.*, 2(4):274–280, 1995.

[5] David De Sancho, Urmi Doshi, and Victor Muñoz. Protein folding rates and stability: how much is there beyond size? *J. Am. Chem. Soc.*, 131(6):2074–2075, 2009.

[6] Angelo Perico and Marina Guenza. Viscoelastic relaxation of segment orientation in dilute polymer solutions. *J. Chem. Phys.*, 83:3103, 1985.

[7] Angelo Perico, Marina Guenza, Michele Mormino, and Roberto Fioravanti. Protein dynamics: Rotational diffusion of rigid and fluctuating three dimensional structures. *Biopolym.*, 35(1):47–54, 1995.

[8] 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. *Biophys. J.*, 93(12):4128–4140, 2007.

[9] Ivet Bahar, Ali Rana Atilgan, and Burak Erman. Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential. *Folding Design*, 2(3):173–181, 1997.

[10] AR Atilgan, SR Durell, RL Jernigan, MC Demirel, O Keskin, and I Bahar. Anisotropy of fluctuation dynamics of proteins with an elastic network model. *Biophys. J.*, 80(1):505–515, 2001.

[11] Nobuhiro Go, Tosiyuki Noguti, and Testuo Nishikawa. Dynamics of a small globular protein in terms of low-frequency vibrational modes. *Proc. Natl. Acad. Sci.*, 80(12):3696–3700, 1983.

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

[13] Dong Long and Rafael Brüschweiler. In silico elucidation of the recognition dynamics of ubiquitin. *PLoS Comput. Biol.*, 7(4):e1002035, 2011.

[14] Jhih-Wei Chu and Gregory A Voth. Coarse-grained free energy functions for studying protein conformational changes: a double-well network model. *Biophys. J.*, 93(11):3860–3871, 2007.

[15] Andrew L Lee and A Joshua Wand. Assessing potential bias in the determination of rotational correlation times of proteins by nmr relaxation. *J. Biomol. NMR*, 13(2):101–112, 1999.

[16] Nils-Alexander Lakomek, Oliver Lange, KF Walter, Christopher Fares, Dalia Egger, Peter Lunkenheimer, Jens Meiler, Helmut Grubmuller, Stefan Becker, Bert L de Groot, et al. Residual dipolar couplings as a tool to study molecular recongnition of ubiquitin. *Biochem. Soc. Trans.*, 36(Pt 6):1433–1437, 2008.

[17] Jeremy Copperman and Marina G Guenza. A coarse-grained langevin equation for protein dynamics: Global anisotropy and a mode approach to local complexity. *J. Phys. Chem. B*, 2014.

[18] M Doi and SF Edwards. The theory of polymer dynamics, 1986.

[19] Robert Zwanzig. *Nonequilibrium Statistical Mechanics*. Oxford University Press, 2001.

[20] Andrea Amadei, Antonius Linssen, and Herman JC Berendsen. Essential dynamics of proteins. *Proteins: Struct., Funct., Bioinf.*, 17(4):412–425, 1993.

[21] Gene Lamm and Attila Szabo. Langevin modes of macromolecules. *J. Chem. Phys.*, 85(12):7334–7348, 1986.

[22] Marina G Guenza. Theoretical models for bridging timescales in polymer dynamics. *J. Phys.: Cond. Matt.*, 20(3):033101, 2008.

[23] Mordechai Bixon and Robert Zwanzig. Optimized rouse–zimm theory for stiff polymers. *J. Chem. Phys.*, 68(4):1896–1902, 1978.

[24] John J Portman, Shoji Takada, and Peter G Wolynes. Variational theory for site resolved protein folding free energy surfaces. *Phys. Rev. Lett.*, 81(23):5237, 1998.

[25] José García de la Torre, María L Huertas, and Beatriz Carrasco. Calculation of hydrodynamic properties of globular proteins from their atomic-level structure. *Biophys. J.*, 78(2):719–730, 2000.

[26] Rony Granek. Proteins as fractals: role of the hydrodynamic interaction. *Phys. Rev. E*, 83(2):020902, 2011.

[27] Jens Rotne and Stephen Prager. Variational treatment of hydrodynamic interaction in polymers. *J. Chem. Phys.*, 50(11):4831–4837, 1969.

[28] Diane E Sagnella, John E Straub, and D Thirumalai. Time scales and pathways for kinetic energy relaxation in solvated proteins: Application to carbonmonoxy myoglobin. *J. Chem. Phys.*, 113:7702, 2000.

[29] Robert Zwanzig, James Kiefer, and George H Weiss. On the validity of the kirkwood-riseman theory. *Proc. Natl. Acad. Sci.*, 60(2):381, 1968.

[30] Marina Guenza and Angelo Perico. A reduced description of the local dynamics of star polymers. *Macromol.*, 25(22):5942–5949, 1992.

[31] 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: Struct., Funct., Bioinf.*, 78(8):1950–1958, 2010.

[32] Senadhi Vijay-Kumar, Charles E Bugg, and William J Cook. Structure of ubiquitin refined at 1.8 åresolution. *J. Mol. Biol.*, 194(3):531–544, 1987.

[33] Herman JC Berendsen, David van der Spoel, and Rudi van Drunen. Gromacs: A message-passing parallel molecular dynamics implementation. *Comp. Phys. Comm.*, 91(1):43–56, 1995.

[34] Erik Lindahl, Berk Hess, and David Van Der Spoel. Gromacs 3.0: a package for molecular simulation and trajectory analysis. *J. Mol. Model.*, 7(8):306–317, 2001.

[35] David Van Der Spoel, Erik Lindahl, Berk Hess, Gerrit Groenhof, Alan E Mark, and Herman JC Berendsen. Gromacs: fast, flexible, and free. *J. Comput. Chem.*, 26(16):1701–1718, 2005.

[36] Berk Hess, Carsten Kutzner, David Van Der Spoel, and Erik Lindahl. Gromacs 4: Algorithms for highly efficient, load-balanced, and scalable molecular simulation. *J. Chem. Theor. Comput.*, 4(3):435–447, 2008.

[37] Stefano Piana, Kresten Lindorff-Larsen, and David E Shaw. Atomic-level description of ubiquitin folding. *Proc. Natl. Acad. Sci.*, 110(15):5915–5920, 2013.

[38] Nidhi Rawat and Parbati Biswas. Size, shape, and flexibility of proteins and dna. *J. Chem. Phys.*, 131(16):165104, 2009.

[39] Steven Harvey and Jose Garcia de La Torre. Coordinate systems for modeling the hydrodynamic resistance and diffusion coefficients of irregularly shaped rigid macromolecules. *Macromol.*, 13(4):960–964, 1980.

[40] J Garcia De La Torre, S Navarro, MC Lopez Martinez, FG Diaz, and JJ Lopez Cascales. Hydro: a computer program for the prediction of hydrodynamic properties of macromolecules. *Biophys. J.*, 67(2):530–531, 1994.

[41] J Garcıa de la Torre, ML Huertas, and B Carrasco. Hydronmr: prediction of nmr relaxation of globular proteins from atomic-level structures and hydrodynamic calculations. *J. Mag. Res.*, 147(1):138–146, 2000.

[42] Pau Bernadó, José García de la Torre, and Miquel Pons. Interpretation of 15n nmr relaxation data of globular proteins using hydrodynamic calculations with hydronmr. *J. Biomol. NMR*, 23(2):139–150, 2002.

[43] L Dale Favro. Theory of the rotational brownian motion of a free rigid body. *Phys. Rev.*, 119(1):53, 1960.

[44] Vance Wong and David A Case. Evaluating rotational diffusion from protein md simulations. *J. Phys. Chem. B*, 112(19):6013–6024, 2008.

[45] Paul E Smith and Wilfred F van Gunsteren. The viscosity of spc and spc/e water at 277 and 300 k. *Chem. Phys. Lett.*, 215(4):315–318, 1993.

[46] Guang-Jun Guo and Yi-Gang Zhang. Equilibrium molecular dynamics calculation of the bulk viscosity of liquid water. *Mol. Phys.*, 99(4):283–289, 2001.

[47] Kazuhiro Takemura and Akio Kitao. Water model tuning for improved reproduction of rotational diffusion and nmr spectral density. *J. Phys. Chem. B*, 116(22):6279–6287, 2012.

[48] Mary A Rohrdanz, Wenwei Zheng, Mauro Maggioni, and Cecilia Clementi. Determination of reaction coordinates via locally scaled diffusion map. *J. Chem. Phys.*, 134(12):124116, 2011.

[49] Joseph D Bryngelson, Jose Nelson Onuchic, Nicholas D Socci, and Peter G Wolynes. Funnels, pathways, and the energy landscape of protein folding: a synthesis. *Proteins: Struct., Funct., Bioinf.*, 21(3):167–195, 1995.

[50] José Nelson Onuchic, Zaida Luthey-Schulten, and Peter G Wolynes. Theory of protein folding: the energy landscape perspective. *Ann. Rev. Phys. Chem.*, 48(1):545–600, 1997.

[51] Frank Noé and Stefan Fischer. Transition networks for modeling the kinetics of conformational change in macromolecles. *Curr. Op. Struct. Biol.*, 18(2):154–162, 2008.

[52] E Weinan, Weiqing Ren, and Eric Vanden-Eijnden. String method for the study of rare events. *Phys. Rev. B*, 66(5):052301, 2002.

[53] Robert B Best and Gerhard Hummer. Reaction coordinates and rates from transition paths. *Proc. Natl. Acad. Sci.*, 102(19):6732–6737, 2005.

[54] Sichun Yang, José N Onuchic, and Herbert Levine. Effective stochastic dynamics on a protein folding energy landscape. *J. Chem. Phys.*, 125(5):054910, 2006.

[55] Herman JC Berendsen and Steven Hayward. Collective protein dynamics in relation to function. *Curr. Op. Struct. Biol.*, 10(2):165–169, 2000.

[56] B Montgomery Pettitt. *Advances in Chemical Physics, Proteins: A Theoretical Perspective of Dynamics, Structure, and Thermodynamics*. John Wiley & Sons, 1990.

[57] Ivan Lyubimov and MG Guenza. First-principle approach to rescale the dynamics of simulated coarse-grained macromolecular liquids. *Phys. Rev. E*, 84(3):031801, 2011.

[58] Robert Zwanzig. Rate processes with dynamical disorder. *Acc. Chem. Res.*, 23(5):148–152, 1990.

[59] Arthur G Palmer III. Nmr probes of molecular dynamics: overview and comparison with other techniques. *Ann. Rev. Biophys. Biomol. Struct.*, 30(1):129–155, 2001.

[60] Marcel Ottiger and Ad Bax. Determination of relative n-hn, n-c', cα-c', and cα-hα effective bond lengths in a protein by nmr in a dilute liquid crystalline phase. *J. A. Chem. Soc.*, 120(47):12334–12341, 1998.

[61] Giovanni Lipari and Attila Szabo. Model-free approach to the interpretation of nuclear magnetic resonance relaxation in macromolecules. 1. theory and range of validity. *J. A. Chem. Soc.*, 104(17):4546–4559, 1982.

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

[63] Józef R Lewandowski, Meghan E Halse, Martin Blackledge, and Lyndon Emsley. Direct observation of hierarchical protein dynamics. *Sci.*, 348(6234):578–581, 2015.

[64] GR Kneller and Konrad Hinsen. Fractional brownian dynamics in proteins. *J. Chem. Phys.*, 121(20):10278–10283, 2004.

[65] Daniel Abergel and Geoffrey Bodenhausen. Predicting internal protein dynamics from structures using coupled networks of hindered rotators. *J. Chem. Phys.*, 123(20):204901–204901, 2005.

[66] Lei Yang, Guang Song, and Robert L Jernigan. How well can we understand large-scale protein motions using normal modes of elastic network models? *Biophys. J.*, 93(3):920–929, 2007.

[67] Lei Yang, Guang Song, and Robert L Jernigan. Protein elastic network models and the ranges of cooperativity. *Proc. Natl. Acad. Sci.*, 106(30):12347–12352, 2009.

[68] Ken A Dill and Hue Sun Chan. From levinthal to pathways to funnels. *Nat. Struct. Biol.*, 4(1):10–19, 1997.

[69] Jacque Monod, Jeffries Wyman, and Jean-Pierre Changeux. On the nature of allosteric transitions: a plausible model. *J. Mol. Biol.*, 12(1):88–118, 1965.

[70] Cyrus Levinthal. Are there pathways for protein folding? *J. Chim. phys*, 65(1):44–45, 1968.

[71] Gia G Maisuradze, Adam Liwo, and Harold A Scheraga. Principal component analysis for protein folding dynamics. *J. Mol. Biol.*, 385(1):312–329, 2009.

[72] Jeanine J Prompers and Rafael Brüschweiler. Reorientational eigenmode dynamics: a combined md/nmr relaxation analysis method for flexible parts in globular proteins. *J. Am. Chem. Soc.*, 123(30):7305–7313, 2001.

[73] Chris AEM Spronk, Sander B Nabuurs, Alexandre MJJ Bonvin, Elmar Krieger, Geerten W Vuister, and Gert Vriend. The precision of nmr structure ensembles revisited. *J. Biomol. NMR*, 25(3):225–234, 2003.

[74] Michal Jamroz, Andrzej Kolinski, and Sebastian Kmiecik. Cabs-flex predictions of protein flexibility compared with nmr ensembles. *Bioinformatics*, page btu184, 2014.

[75] Kelly L Damm and Heather A Carlson. Exploring experimental sources of multiple protein conformations in structure-based drug design. *J. Am. Chem. Soc.*, 129(26):8225–8235, 2007.

[76] Yaakov Levy and José N Onuchic. Water mediation in protein folding and molecular recognition. *Annu. Rev. Biophys. Biomol. Struct.*, 35:389–415, 2006.

[77] Samir Kumar Pal, Jorge Peon, and Ahmed H Zewail. Biological water at the protein surface: dynamical solvation probed directly with femtosecond resolution. *Proc. Natl. Acad. Sci.*, 99(4):1763–1768, 2002.

[78] John Towns, Timothy Cockerill, Maytal Dahan, Ian Foster, Kelly Gaither, Andrew Grimshaw, Victor Hazlewood, Scott Lathrop, Dave Lifka, Gregory D Peterson, et al. Xsede: accelerating scientific discovery. *Comput. Sci. & Eng.*, 16(5):62–74, 2014.

[79] Giovanni Bussi, Davide Donadio, and Michele Parrinello. Canonical sampling through velocity rescaling. *J. Chem. Phys.*, 126(1):014101, 2007.

[80] A Amadei, BL De Groot, M-A Ceruso, M Paci, A Di Nola, and HJC Berendsen. A kinetic model for the internal motions of proteins: diffusion between multiple harmonic wells. *Proteins: Struct., Funct., Bioinf.*, 35(3):283–292, 1999.

[81] David Ruppert. *Statistics and data analysis for financial engineering*. Springer, 2010.

[82] Meyer B Jackson. On the time scale and time course of protein conformational changes. *J. Chem. Phys.*, 99(9):7253–7259, 1993.

[83] Gerold Adam and Julian H Gibbs. On the temperature dependence of cooperative relaxation properties in glass-forming liquids. *J. Chem. Phys.*, 43(1):139–146, 1965.

[84] Francis W Starr, Jack F Douglas, and Srikanth Sastry. The relationship of dynamical heterogeneity to the adam-gibbs and random first-order transition theories of glass formation. *J. Chem. Phys.*, 138(12):12A541–1–12A541–18, 2013.

[85] DJ Wales. Perspective: Insight into reaction coordinates and dynamics from the potential energy landscape. *J. Chem. Phys.*, 142(13):130901, 2015.

[86] Guanghua Gao, Valentyna Semenchenko, S Arumugam, and Steven R Van Doren. Tissue inhibitor of metalloproteinases-1 undergoes microsecond to millisecond motions at sites of matrix metalloproteinase-induced fit. *J. Mol. Biol.*, 301(2):537–552, 2000.

[87] Abigail Go, Seho Kim, Jean Baum, and Michael H Hecht. Structure and dynamics of de novo proteins from a designed superfamily of 4-helix bundles. *Protein Sci.*, 17(5):821–832, 2008.

[88] Jianyun Lu, David P Cistola, and Ellen Li. Two homologous rat cellular retinol-binding proteins differ in local conformational flexibility. *J. Mol. Biol.*, 330(4):799–812, 2003.

[89] Gui-in Lee, Zhaofeng Ding, John C Walker, and Steven R Van Doren. Nmr structure of the forkhead-associated domain from the arabidopsis receptor kinase-associated protein phosphatase. *Proc. Natl. Acad. Sci.*, 100(20):11261–11266, 2003.

[90] Gui-in Lee, Jia Li, John C Walker, and Steven R Van Doren. 1h, 13c and 15n resonance assignments of the kinase-interacting fha domain of arabidopsis thaliana kinase-associated protein phosphatase. *J. Biomol. NMR*, 25(3):253–254, 2003.

[91] Christopher Williams, Dellel Rezgui, Stuart N Prince, Oliver J Zaccheo, Emily J Foulstone, Briony E Forbes, Raymond S Norton, John Crosby, A Bassim Hassan, and Matthew P Crump. Structural insights into the interaction of insulin-like growth factor 2 with igf2r domain 11. *Struct.*, 15(9):1065–1078, 2007.

[92] Rieko Ishima, Rodolfo Ghirlando, József Tözsér, Angela M Gronenborn, Dennis A Torchia, and John M Louis. Folded monomer of hiv-1 protease. *J. Biol. Chem.*, 276(52):49110–49116, 2001.

[93] Rieko Ishima, Dennis A Torchia, Shannon M Lynch, Angela M Gronenborn, and John M Louis. Solution structure of the mature hiv-1 protease monomer: Insight into the tertiary fold and stability of a precursor. *J. Biol. Chem.*, 278(44):43311–43319, 2003.

[94] Kresten Lindorff-Larsen, Robert B Best, Mark A DePristo, Christopher M Dobson, and Michele Vendruscolo. Simultaneous determination of protein structure and dynamics. *Nature*, 433(7022):128–132, 2005.

[95] R Bryn Fenwick, Santi Esteban-Martín, Barbara Richter, Donghan Lee, Korvin FA Walter, Dragomir Milovanovic, Stefan Becker, Nils A Lakomek, Christian Griesinger, and Xavier Salvatella. Weak long-range correlated motions in a surface patch of ubiquitin involved in molecular recognition. *J. Am. Chem. Soc.*, 133(27):10336–10339, 2011.

[96] Rinaldo W Montalvao, Alfonso De Simone, and Michele Vendruscolo. Determination of structural fluctuations of proteins from structure-based calculations of residual dipolar couplings. *J. Biomol. NMR*, 53(4):281–292, 2012.

[97] Barbara Richter, Joerg Gsponer, Péter Várnai, Xavier Salvatella, and Michele Vendruscolo. The mumo (minimal under-restraining minimal over-restraining) method for the determination of native state ensembles of proteins. *J. Biomol. NMR*, 37(2):117–135, 2007.

[98] Gabriel Cornilescu, John L Marquardt, Marcel Ottiger, and Ad Bax. Validation of protein structure from anisotropic carbonyl chemical shifts in a dilute liquid crystalline phase. *J. Am. Chem. Soc.*, 120(27):6836–6837, 1998.

[99] Charles R Babu, Peter F Flynn, and A Joshua Wand. Validation of protein structure from preparations of encapsulated proteins dissolved in low viscosity fluids. *J. Am. Chem. Soc.*, 123(11):2691–2692, 2001.

[100] Tobias Madl, Wolfgang Bermel, and Klaus Zangger. Use of relaxation enhancements in a paramagnetic environment for the structure determination of proteins using nmr spectroscopy. *Angewandte Chemie Int. Ed.*, 48(44):8259–8262, 2009.

[101] Alexander S Maltsev, Alexander Grishaev, Julien Roche, Michael Zasloff, and Ad Bax. Improved cross validation of a static ubiquitin structure derived from high precision residual dipolar couplings measured in a drug-based liquid crystalline phase. *J. Am. Chem. Soc.*, 136(10):3752–3755, 2014.

[102] Liisa Holm and Chris Sander. Mapping the protein universe. *Sci.*, 273(5275):595–602, 1996.

[103] R Elber and M Karplus. Multiple conformational states of proteins: a molecular dynamics analysis of myoglobin. *Sci.*, 235(4786):318–321, 1987.

[104] Frank Noé, Illia Horenko, Christof Schütte, and Jeremy C Smith. Hierarchical analysis of conformational dynamics in biomolecules: transition networks of metastable states. *J. Chem. Phys.*, 126(15):155102, 2007.

[105] Yasuyuki Matoba and Masanori Sugiyama. Atomic resolution structure of prokaryotic phospholipase a2: analysis of internal motion and implication for a catalytic mechanism. *Proteins: Struct., Funct., Bioinf.*, 51(3):453–469, 2003.

[106] Guy G Dodson, David P Lane, and Chandra S Verma. Molecular simulations of protein dynamics: new windows on mechanisms in biology. *EMBO Rep.*, 9(2):144–150, 2008.

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

[108] Nataliya Popovych, Shangjin Sun, Richard H Ebright, and Charalampos G Kalodimos. Dynamically driven protein allostery. *Nat. Struct. Mol. Biol.*, 13(9):831–838, 2006.

[109] David L Mobley and Ken A Dill. Binding of small-molecule ligands to proteins:"what you see" is not always "what you get". *Struct.*, 17(4):489–498, 2009.

[110] Kei-ichi Okazaki and Shoji Takada. Dynamic energy landscape view of coupled binding and protein conformational change: induced-fit versus population-shift mechanisms. *Proc. Natl. Acad. Sci.*, 105(32):11182–11187, 2008.

[111] Shiou-Ru Tzeng and Charalampos G Kalodimos. Protein activity regulation by conformational entropy. *Nature*, 488(7410):236–240, 2012.

[112] J Copperman and MG Guenza. Predicting protein dynamics from structural ensembles. *J. Chem. Phys.*, 143(24):243131, 2015.

[113] JD Honeycutt and D Thirumalai. The nature of folded states of globular proteins. *Biopolym.*, 32(6):695–709, 1992.

[114] Clay Bracken, Lilia M Iakoucheva, Pedro R Romero, and A Keith Dunker. Combining prediction, computation and experiment for the characterization of protein disorder. *Curr. Op. Struct. Biol.*, 14(5):570–576, 2004.

[115] Raik Grünberg, Michael Nilges, and Johan Leckner. Flexibility and conformational entropy in protein-protein binding. *Struct.*, 14(4):683–693, 2006.

[116] DE Koshland Jr. Application of a theory of enzyme specificity to protein synthesis. *Proc. Natl. Acad. Sci.*, 44(2):98, 1958.

[117] Rufus Lumry and Shyamala Rajender. Enthalpy–entropy compensation phenomena in water solutions of proteins and small molecules: a ubiquitous properly of water. *Biopolym.*, 9(10):1125–1227, 1970.

[118] Jack D Dunitz. Win some, lose some: enthalpy-entropy compensation in weak intermolecular interactions. *Chem. & Biol.*, 2(11):709–712, 1995.

[119] 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. *Phys. Rev. Lett.*, 80(12):2733, 1998.

[120] Ignacia Echeverria, Dmitrii E Makarov, and Garegin A Papoian. Concerted dihedral rotations give rise to internal friction in unfolded proteins. *J. Am. Chem. Soc.*, 136(24):8708–8713, 2014.

[121] S Oroszlan and RB Luftig. Retroviral proteinases. In *Retroviruses*, pages 153–185. Springer, 1990.

[122] Alexander Wlodawer and John W Erickson. Structure-based inhibitors of hiv-1 protease. *Ann. Rev. Biochem.*, 62(1):543–585, 1993.

[123] Adi Doron-Faigenboim, Adi Stern, Itay Mayrose, Eran Bacharach, and Tal Pupko. Selecton: a server for detecting evolutionary forces at a single amino-acid site. *Bioinformatics*, 21(9):2101–2103, 2005.

[124] Toshimasa Yamazaki, Andrew P Hinck, Yun-Xing Wang, Linda K Nicholson, Dennis A Torchia, Paul Wingfield, Stephen J Stahl, Joshua D Kaufman, Chong-Hwan Chang, Peter J Domaille, et al. Three-dimensional solution structure of the hiv-1 protease complexed with dmp323, a novel cyclic urea-type inhibitor, determined by nuclear magnetic resonance spectroscopy. *Protein Sci.*, 5(3):495–506, 1996.

[125] Daniel D Loeb, Ronald Swanstrom, Lorraine Everitt, Marianne Manchester, Susan E Stamper, and Clyde A Hutchison. Complete mutagenesis of the hiv-1 protease. *Nature*, 340(6232):397–400, 1989.

[126] J Keith Killian, James C Byrd, James V Jirtle, Barry L Munday, Michael K Stoskopf, Richard G MacDonald, and Randy L Jirtle. M6p/igf2r imprinting evolution in mammals. *Mol. Cell*, 5(4):707–716, 2000.

[127] Christopher Williams, Hans-Jürgen Hoppe, Dellel Rezgui, Madeleine Strickland, Briony E Forbes, Frank Grutzner, Susana Frago, Rosamund Z Ellis, Pakorn Wattana-Amorn, Stuart N Prince, et al. An exon splice enhancer primes igf2: Igf2r binding site structure and function evolution. *Sci.*, 338(6111):1209–1213, 2012.

[128] J Copperman and MG Guenza. Mode localization in the cooperative dynamics of protein recognition. *manuscript in submission*, 2016.

[129] Shlomi Reuveni, Rony Granek, and Joseph Klafter. Proteins: coexistence of stability and flexibility. *Phys. Rev. Lett.*, 100(20):208101, 2008.

[130] Daniel Ben-Avraham. Vibrational normal-mode spectrum of globular proteins. *Phys. Rev. B*, 47(21):14559, 1993.

[131] Lilia Milanesi, Jonathan P Waltho, Christopher A Hunter, Daniel J Shaw, Godfrey S Beddard, Gavin D Reid, Sagarika Dev, and Martin Volk. Measurement of energy landscape roughness of folded and unfolded proteins. *Proc. Natl. Acad. Sci.*, 109(48):19563–19568, 2012.

[132] David A Huse and Christopher L Henley. Pinning and roughening of domain walls in ising systems due to random impurities. *Phys. Rev. Lett.*, 54(25):2708, 1985.

[133] Sushanta Dattagupta and Sanjay Puri. *Dissipative phenomena in condensed matter: some applications*, volume 71. Springer Science & Business Media, 2013.

[134] Douglas F Sticke, Leonard G Presta, Ken A Dill, and George D Rose. Hydrogen bonding in globular proteins. *J. Mol. Biol.*, 226(4):1143–1159, 1992.

[135] Sheh-Yi Sheu, Dah-Yen Yang, HL Selzle, and EW Schlag. Energetics of hydrogen bonds in peptides. *Proc. Natl. Acad. Sci.*, 100(22):12683–12687, 2003.

[136] Travis B Peery and Glenn T Evans. Association in a four-coordinated, water-like fluid. *J. Chem. Phys.*, 118(5):2286–2300, 2003.

[137] Jared D Smith, Christopher D Cappa, Kevin R Wilson, Benjamin M Messer, Ronald C Cohen, and Richard J Saykally. Energetics of hydrogen bond network rearrangements in liquid water. *Science*, 306(5697):851–853, 2004.

[138] Joel Janin, Shoshanna Wodak, Michael Levitt, and Bernard Maigret. Conformation of amino acid side-chains in proteins. *J. Mol. Biol.*, 125(3):357–386, 1978.

[139] M Tarek and DJ Tobias. Role of protein-water hydrogen bond dynamics in the protein dynamical transition. *Phys. Rev. Lett.*, 88(13):138101, 2002.

[140] Dong Xu and Ruth Nussinov. Favorable domain size in proteins. *Folding and Des.*, 3(1):11–17, 1998.

[141] Andreas Heger and Liisa Holm. Exhaustive enumeration of protein domain families. *J. Mol. Biol.*, 328(3):749–767, 2003.

[142] Leon Balents, Jean-Philippe Bouchaud, and Marc Mézard. The large scale energy landscape of randomly pinned objects. *J. Phys. I*, 6(8):1007–1020, 1996.

[143] Timothy Halpin-Healy and Yi-Cheng Zhang. Kinetic roughening phenomena, stochastic growth, directed polymers and all that. aspects of multidisciplinary statistical mechanics. *Phys. Rep.*, 254(4):215–414, 1995.

[144] S Roux, A Hansen, and Einar L Hinrichsen. A direct mapping between eden growth model and directed polymers in random media. *J. Phys. A: Math. Gen.*, 24(6):L295, 1991.

[145] Bruce M Forrest and Lei-Han Tang. Surface roughening in a hypercube-stacking model. *Phys. Rev. Lett.*, 64(12):1405, 1990.

[146] Lei-Han Tang, Bruce M Forrest, and Dietrich E Wolf. Kinetic surface roughening. ii. hypercube-stacking models. *Phys. Rev. A*, 45(10):7162, 1992.