# Anomalous Diffusion in Protein Dynamics

Dissertation
zur Erlangung des mathematisch-naturwissenschaftlichen Doktorgrades
"Doctor rerum naturalium"
der Georg-August-Universität Göttingen

im Promotionsprogramm PROPHYS
der Georg-August University School of Science (GAUSS)

vorgelegt von
Andreas Volkhardt
aus Eisenach
Göttingen, 2021

Betreuungsausschuss

Helmut Grubmüller, Theoretische und computergestützte Biophysik, Max-Planck-Institut für Biophysikalische Chemie

Jörg Enderlein, Drittes Physikalisches Insitut – Biophysik, Georg-August-Universität Göttingen

Mitglieder der Prüfungskommission

Referent/in: Helmut Grubmüller, Theoretische und computergestützte Biophysik, Max-Planck-Institut für Biophysikalische Chemie

Korreferent/in: Jörg Enderlein, Drittes Physikalisches Insitut – Biophysik, Georg-August-Universität Göttingen

Weitere Mitglieder der Prüfungskommission:

Marcus Müller, Institut für Theoretische Physik, Georg-August-Universität Göttingen

Matthias Krüger, Institut für Theoretische Physik, Georg-August-Universität Göttingen

Aljaz Godec, Mathematical bioPhysics, Max-Planck-Institut für Biophysikalische Chemie

Stefan Klumpp, Theoretische Biophysik, Institut für Dynamik komplexer Systeme, Georg-August-Universität Göttingen

Tag der mündlichen Prüfung: 26.10.21

# Contents

**Contents** **2**

# Chapter 1

## Introduction

Almost all processes in life are governed by proteins which are biomolecules consisting of a chain of amino acids. They perform a wide range of functions in organisms, including catalysis of metabolic reactions, transport of molecules, transcription of DNA to RNA, and ensuring the structural stability of cells.

This function is intimately linked to a proteins' three-dimensional structure and its dynamics. For example, the biological function of most proteins includes interactions with ligands or other macromolecules at specific sites of the protein. Depending on how exposed these interaction sites are to the environment in a given protein structure, interaction with ligands is more likely. Thus, changes in the structure of a protein, or conformational changes, regulate these interactions. Conversely, the binding of ligands to proteins can induce conformational changes which transmit signals, for example, in hormone-receptor binding [1]. But, also smaller structural fluctuations appear to contribute to protein function [1]. To understand the functioning of proteins, it is, therefore, necessary to investigate protein dynamics.

Protein dynamics is a complex process comprised of motions on multiple orders of magnitude both in characteristic time-scale and amplitude. Bond vibrations are the smallest and fastest motions and show an amplitude of $0.001-0.01\,\mathrm{nm}$ on a characteristic time-scale of $10^{-14}\,\mathrm{s}$ to $10^{-13}\,\mathrm{s}$ whereas the largest motions such as allosteric transitions show an amplitude of $0.1-0.5\,\mathrm{nm}$ on characteristic time scales of $10^{-5}\,\mathrm{s}$ to $1\,\mathrm{s}$ [2].

Functionally important motions are typically identified with the slowest motions of a protein as noise by the heat bath dominates bond vibrations [3]. Often those functionally relevant motions involve processes on many (long) time scales. For example, early flash-photolysis experiments of Frauenfelder et al. showed that unbinding processes of carbon monoxide from myoglobin show a stretched exponential in relaxation times [4] that can only occur if processes on multiple time-scales are involved in the unbinding. A more recent study showed that dynamics of proteins is non-equilibrium and self-similar over thirteen orders of magnitude in time [5] [6]. Finding a comprehensive model that describes such complex dynamics in detail is a very daunting task. Still, we can ask: What

is the simplest model that is complex enough to reproduce observed features of protein dynamics?

Frauenfelder [7] proposed that a hierarchical structured free-energy landscape governs the internal dynamics of proteins (see Fig. 2.2). In such hierarchical free energy landscapes, states of a protein are structured into tiers. The highest tier consists of 'taxonomic conformational states' that are typically associated with protein function. Transitions between these functional states represent the large-scale rearrangements of proteins governed by high free energy barriers, leading to long transition times. Within each of these states, smaller and faster motions between conformational substates are governed by lower free-energy barriers. Each of these conformational substates contains further substates separated by even lower free-energy barriers, where transitions describe even faster motions with smaller amplitude. Relaxation processes in this nested structure of conformational substates lead to stretched exponential decays as a whole hierarchy of decay processes with different decay times are involved.

Further, it has been shown analytically that diffusion in a simple hierarchical lattice model is anomalous, i.e., the variance of trajectories increases with a power law in time instead of linearly as expected for Brownian motion [8] [9]. This model is a 1-dimensional lattice of states where static barriers with exponentially distributed heights $p(\Delta G) = 1/\gamma exp(-\Delta G/\gamma)$ govern transitions between states. It has been shown that anomalous diffusion exponents $\alpha = 1/(1+\gamma)$ directly depend on $\gamma$, representing the 'ruggedness' of the model free-energy landscape. Higher-dimensional versions of this model that resemble protein dynamics more closely could only be treated analytically in the limit of infinite dimensions [10].

Indeed, anomalous diffusion behavior in the internal motions of proteins was observed in equilibrium molecular dynamics (MD) simulations [11] of small globular proteins [12] [13] and peptides [14]. Besides a hierarchically structured free-energy landscape [13], other causes for the observed anomalous diffusion behavior such as a fractal structure of protein configuration space [14] or a mere

projection effect arising from the analysis of collective coordinates [15] have been suggested. Assuming that the anomalous diffusion behavior arises from the hierarchical structure of protein free-energy landscapes, it should be possible to infer barrier height distributions based on anomalous diffusion exponents obtained in MD simulations.

The main aim of this work is to investigate the hierarchical structure of protein free-energy landscapes. To that end, we use a $d$-dimensional generalization of the simple hierarchical lattice model as a reference to translate anomalous diffusion exponents obtained from MD simulations into ruggedness and dimensionality estimates.

This thesis is structured in the following way.

**Chapter 2: Estimating ruggedness and dimensionality from molecular dynamics simulations** In this chapter, we present ruggedness and dimensionality estimates of 500 small globular proteins based on a $d$-dimensional hierarchical lattice model. To that end, we determined how anomalous diffusion exponents depend on ruggedness and dimensionality from random walk simulations in models $15-20\,\mathrm{kT}$ per dimension and $40-60$ dimensions. Assuming that a similar relation holds for free-energy landscapes of proteins, we determined the ruggedness and dimensionality of a set of 500 small globular proteins and obtained typical ruggedness of $15-20\,\mathrm{kT}$ per dimension and dimensionality of $40-60$ . Further, we found an interesting correlation between ruggedness and dimensionality that likely originates from proteins adapting to their particular function. This chapter is a self-contained manuskript that is currently under review.

**Chapter 3: A universal scaling relation between accessible configuration space volume and escape rates in intermediate dimensional hierarchical free-energy landscapes** Here, we asked what the cause of anomalous diffusion in the $d$-dimensional hierarchical lattice model is. Our random walk simulations showed that trajectories, especially for high ruggedness, are

trapped long times in a well-defined small region of state space. We found that ruggedness, the parameter that determines the anomalous diffusion behavior, as shown in chapter 1, affects both the escape times from these traps and their topology. A universal scaling relation describes this combined influence of the topology and escape rates. This result suggests that anomalous diffusion in hierarchical free energy landscapes is at least partly caused by the fractal structure of accessible configuration space. That means hierarchical free-energy landscapes and fractal configuration space are not competing models but rather 'two sides of the same coin'.

**Chapter 4: An efficient sampling algorithm to generate trajectories in hierarchical free-energy landscapes** To calculate random walk trajectories with sufficient sampling in higher dimensional models ruggedness, we needed to develop an enhanced sampling method, as trajectories are trapped in small regions of state space as described in chapter 2. We show that this method yields similar results to brute force sampling on average (over disorder), whereas it is orders of magnitude faster. This method also provides an interesting prospect towards an analytical solution as it is exact in the regime of high-ruggedness values and can be exploited in a renormalization group approach.

# Chapter 2

## Estimating ruggedness of free-energy landscapes of small globular proteins from principal component analysis of molecular dynamics trajectories

## 2.1 Introduction

Most processes in life are governed by proteins, macromolecules consisting of a chain of amino acids. Their biophysical function in living cells is intimately linked to their structure and, in particular, to their remarkably complex internal dynamics on time scales ranging from picoseconds to hours. These thermally activated internal motions are governed by a diffusion process on a free-energy landscape [16]. Moessbauer spectroscopy and neutron scattering experiments showed that protein free-energy landscapes with conformational coordinates as its arguments [17] are characterized by a large number of nearly isoenergetic minima. Free energy barriers between these minima are structured hierarchically, as shown by flash-photolysis experiments on myoglobin [4] (see Fig. 2.2).

Recent progress in methods and performance of computational hardware [18] [19] allows generating molecular dynamics (MD) trajectories ranging over multiple orders of magnitude in simulation length up to multiple microseconds as a standard routine. This development allows studying the hierarchical structure of protein free-energy landscapes *in silico* [20]. However, due to the large number of possible protein configurations, available MD trajectories are still not long enough to reach thermal equilibrium, which constitutes the well-known sampling problem of MD simulations. Slowest relaxation times that correspond to folding/unfolding times are on the time scale of minutes or even hours, which is still beyond what MD simulations are capable of simulating on a reasonable computing time scale.

To circumvent this problem, we use non-equilibrium methods that, rather than finding a model for a protein's equilibrium dynamics, model its dynamics as a diffusion process within its free-energy landscape. Diffusion processes in hierarchical free-energy landscapes models were explored for simple one-dimensional [8] as well as several many-dimensional models [9][21]. It has been shown that diffusion in such models is anomalous, i.e., the variance of trajectories increases with a power law in time [8][10]. In particular, it was analytically

shown how exponents depend on the barrier-height distribution [8]. Similarly, anomalous diffusion behavior would be expected for hierarchical protein free-energy landscapes and was indeed observed in MD simulations of small peptides [14] and small globular proteins [13] (see Fig. 2.1). Assuming that the observed anomalous diffusion behavior arises from the hierarchical structure of protein free-energy landscapes, it should be possible to estimate barrier height distributions from anomalous diffusion exponents obtained from MD trajectories.

In this work, we estimated barrier height distributions of 500 small globular proteins selected to cover known folds and functions from anomalous diffusion exponents observed in MD simulations. To this end, we generated for each of these proteins 1 $\mu$s molecular dynamics trajectories and carried out trajectory-length dependent principal component analysis [13] for the selected proteins. To translate the observed anomalous diffusion exponents into barrier height distributions, we used a $d$-dimensional hierarchical model. This model consists of a lattice of states: transition rates between adjacent states are governed by static free-energy barriers $\Delta G$, which are randomly distributed according to $p(\Delta G) = \frac{1}{\gamma}\, e^{-\Delta G/\gamma}$, where $\gamma$ quantifies the 'ruggedness' of the hierarchical free-energy landscape. The relation between anomalous diffusion exponents and $\gamma$ is analytically known only for 1-d models and, in a mean-field approximation, for high dimensions $d \to \infty$ [10]. However, because the essential configurational subspace of proteins is assumed to be $\sim 10 < d <\sim 100$) [22], we had to resort to a numerical approach by simulating random walks in models with $3 - 200$ dimensions. Indeed, we observed large deviations from the mean-field approximation. By cross-validation, we showed that, based on this numerically obtained relation, barrier height distributions can be estimated with an accuracy of $\sim 5\,\mathrm{kT}$.

Applying the same approach to 1 $\mu$s MD trajectories, we determined $\gamma$ for protein free-energy landscapes. We found that most ruggedness coefficients of the proteins fall within $\gamma \approx 15 - 20\,\mathrm{kT}/d$ with an estimated essential subspace dimensionality $d \approx 40 - 60$. This result provides evidence that the dynamics of

a broad range of protein folds is governed by similar barrier height distributions.

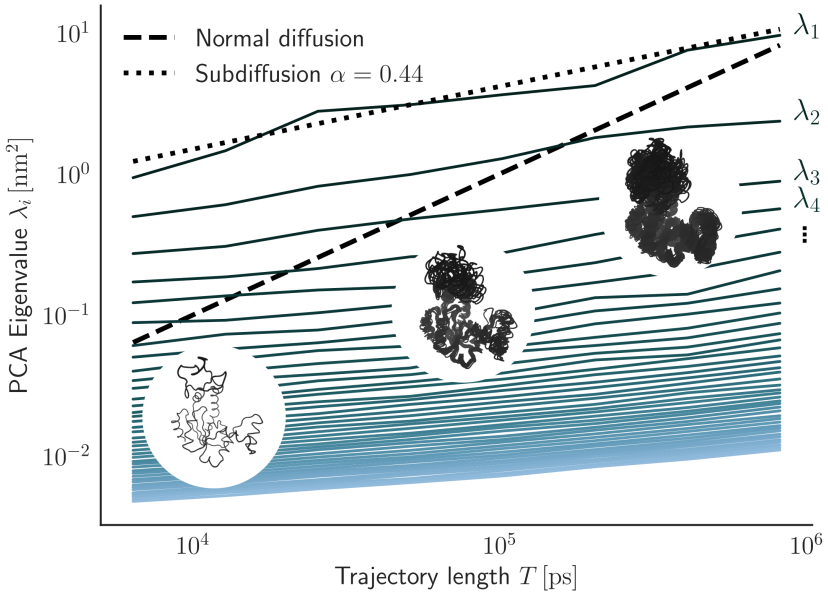


Figure 2.1: *Variance of molecular dynamics trajectories along collective coordinates shows a power law like scaling behavior in dependence of trajectory length.* The figure shows the variance of a 5 $\mu$s molecular dynamics trajectory of adenylate kinase from Escherichia coli (PDB code: 1AKE) along orthogonal collective coordinates, i.e., principal components (PC) in dependence of trajectory length $T$. These collective coordinates are (PCA) eigenvectors of the covariance matrix of the trajectory and are typically ordered according to the magnitude of their corresponding (PCA) eigenvalues $\lambda_i$, which represent the variance of the trajectory along the eigenvector. PCA eigenvalues $\lambda_i$ of MD trajectories approximately increase with a power law depending on trajectory length $T$. The scaling exponents $\alpha_i$ (slopes in a log-log plot) of these power laws show subdiffusive behavior as $\alpha_i < 1$.

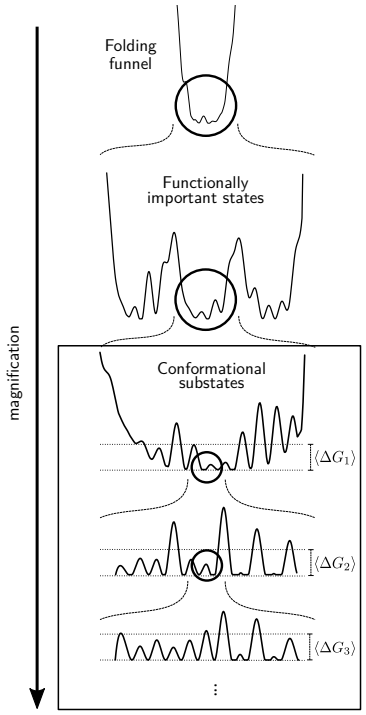


Figure 2.2: *Hierarchical structure of protein free-energy landscapes proposed by Frauenfelder [4].* (Top) On a large scale, the folding funnel dominates free-energy landscapes of globular proteins [16]. Within this folding, on a smaller scale, functionally important states are found. Each of these contains a hierarchy of conformational substates, where hierarchy tiers $i$ are characterized by mean barrier heights $\langle \Delta G_i \rangle$. Due to the multitude of different isoenergetic conformational substates, the free energy landscape of proteins on this scale is best described in statistical terms, i.e., in terms of barrier distributions. [23]

## 2.2 Theory

To explain non-exponential kinetics, e.g., ligand binding experiments [4], Frauenfelder proposed early on that, in the folded state (Fig. 2.2, top), the underlying intramolecular protein dynamics is governed by a hierarchical (free) energy landscape [4] (Fig. 2.2 bottom and magnification). Accordingly, the kinetics

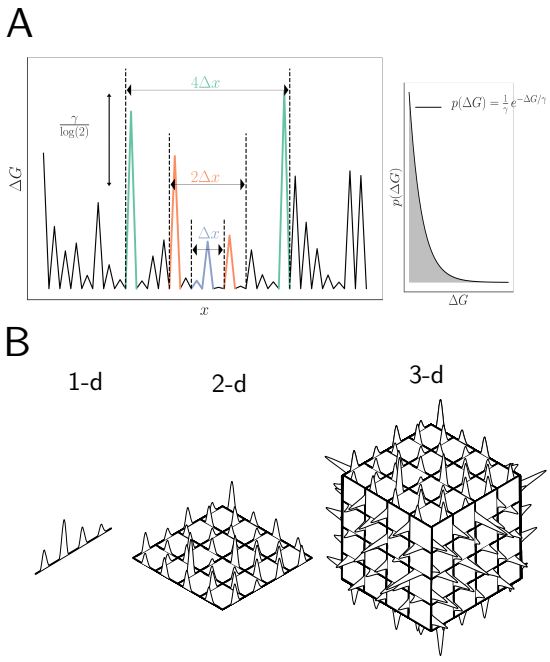


Figure 2.3: (A) Sketch of a 1-d hierarchical model free energy landscape, which is characterized by a lattice of free energy minima with equal free energy separated by barriers with random heights distributed according to an exponential distribution. With increasing length scales, highest barrier heights increase on average with a characteristic height $\gamma$. (B) Sketch of a $d$-dimensional generalization procedure of a 1-dimensional hierarchical model. Whereas the exponential barrier height distribution is kept, the 1-dimensional lattice is generalized to $d$-dimensional lattices (2-d and 3-d cases are shown).

of larger, typically functional protein motions are governed by higher free energy barriers located at correspondingly larger distances in configurational space (top sketch in the box of Fig. 2.2). These barriers separate 'taxonomic conformational states'. Within each of these functional states, increasingly smaller and faster motions between 'statistical substates' [23] are described by more frequent crossings of increasingly lower barriers separated by correspondingly smaller distances (lower sketches in the box of Fig. 2.2). Overall, the protein

free energy landscape is thus described by a hierarchy of energy barriers with characteristic barrier heights and separations at each tier. Precisely how the barrier height increases with increasing mutual distance between the barriers is described by the ruggedness $\gamma$. Hence, $\gamma$ determines the subdiffusive dynamics of the protein over many orders of magnitude [8].

### 2.2.1 Simple hierarchical model free energy landscape

To relate this subdiffusive behavior to ruggedness $\gamma$ , we used a $d$-dimensional lattice model inspired by the subdiffusive behavior of the simple 1-dimensional model [8]. The near power law type behavior of essential degrees of freedom of many proteins, as illustrated in Fig. 2.1, suggests using a hierarchy of barriers, the heights $\gamma$ of which are distributed exponentially,

$$p(\Delta G) = \frac{1}{\gamma}\, e^{-\Delta G/\gamma}. \tag{2.1}$$

In this model, shown in Fig 2.3 A, the ruggedness $\gamma$ describes how the height of the barriers $\Delta G$ increases with increasing average distance $\Delta x$ between these barriers. Specifically, for a distance increase by a factor of two, the barrier heights increase by $\gamma/\log(2)$.

Generalizations to $d$-dimensions have been suggested, which may be considered as a model for the high-dimensional protein free-energy landscape, such as in [9]. Here, we rather choose the model proposed in [21] because we expect it to give more isotropic diffusion than the other. This hierarchical lattice model is a $d$-dimensional cubic lattice where exponentially distributed barriers heights (see Fig. 2.3) govern transitions between states.

It has been shown that this model exhibits anomalous diffusion both for 1-dimensional and in the limit of high-dimensional models [10].

For the former, the subdiffusion exponent $\alpha$ is

$$\alpha = \frac{2}{1+\gamma}. \tag{2.2}$$

For high dimensions $d$, a mean-field approximation [10] yields

$$\alpha = \frac{2\,d}{\gamma}. \tag{2.3}$$

This approximation assumes that a random walk trajectory never visits any state twice, which is strictly only fulfilled for $d \to \infty$ and is equivalent [10] to continuous time random walks (CTRW) [24]. Because protein dynamics is well described by typically tens or hundreds of collective coordinates [25], we assumed that neither approximation is sufficiently accurate in this intermediate range and therefore resorted to studying this model numerically.

### 2.2.2 Determining anomalous diffusion exponents from MD trajectories

To this end, we followed common practice in protein dynamics simulations and calculated anomalous diffusion exponents from trajectory length dependent PCA [13] [14]. This approach differs from traditional analyses of subdiffusion in statistical free energy landscapes in that anomalous diffusion exponents are determined from the trajectory length dependence of the eigenvalues of the time-averaged covariance matrix,

$$C_{ij}(T) = \frac{1}{T}\left\langle \int_0^T \mathrm{d}t\,(\mathbf{x}_i(t) - \boldsymbol{\mu}_i)\cdot(\mathbf{x}_j(t) - \boldsymbol{\mu}_j)\right\rangle_{\mathrm{ens}}, \tag{2.4}$$

rather than time dependence of the ensemble-averaged covariance matrix

$$C_{ij}(t) = \langle(\mathbf{x}_i(t) - \boldsymbol{\mu}_i)\cdot(\mathbf{x}_j(t) - \boldsymbol{\mu}_j)\rangle_{\mathrm{ens}}\,. \tag{2.5}$$

In the above two equations 2.4 and 2.5, $\mathbf{x}_i$ denotes the $3N$ Cartesian coordinates of $N$ selected atoms of a protein and $\boldsymbol{\mu}_i = \int_0^T \mathrm{d}t\,\mathbf{x}_i(t)$ as well as $\boldsymbol{\mu}_i = \langle\mathbf{x}_i(t)\rangle_{\mathrm{ens}}$ their corresponding means. Note that $\mathbf{C}$ is by construction a non-negative symmetric matrix and is therefore diagonalizable. Its non-negative eigenvalues ("PCA eigenvalues") $\lambda_i$ represent the variance of a trajectory along corresponding eigenvectors $\mathbf{v}_i$ ("PCA eigenvectors"), which represent collective motions. If $\lambda_i$ follows a power law

$$\lambda_i(T) \propto T^{\alpha_i}, \tag{2.6}$$

we define anomalous diffusion exponents as $\alpha_i$. For Brownian motion in the limit of high-dimensional spaces, it has been shown that this definition is equivalent to the conventional one derived from equation 2.5 [26]. In supplement I, we show that this equivalence also holds for anomalous diffusion in high-dimensional hierarchical lattice models.

## 2.3 Methods

### 2.3.1 Trajectory length dependent principal component analysis (tPCA)

Anomalous diffusion exponents $\alpha_i$ were determined in a similar way from both MD trajectories and random walk trajectories in hierarchical model free energy landscapes. To estimate $\alpha_i$ via equations 2.4 and 2.6, each trajectory of length $T_0$ was split into 100 windows $[nT_0]$ of length $T$. On each of these windows, PCA was performed and the resulting set of PCA eigenvalues was averaged. 10 different time window lengths $T$ were chosen, distributed exponentially from 100 ps to 300 ns, such that the overlap of consecutive windows was below 10% to maximize information content. For these data, anomalous diffusion exponents $\alpha_i$ were estimated from the slopes of linear least-squares fits to the logarithm of window length $T$ and PCA eigenvalues $\lambda_i$.

### 2.3.2 Random walk generation

We generated 40,000 random walks, each in a separately generated hierarchical free energy landscape. Random walks were generated using the Gilliespie algorithm [27] with parameter ranges summarized in Table 2.1. Due to the high dimensionality of the energy landscapes, sections of these were generated dynamically on demand. To this end, for each visited state, adjacent barriers to previously visited states were recovered from memory, whereas new adjacent barriers were chosen randomly from the exponential distribution $p(\Delta G) = \frac{1}{\gamma} e^{-\Delta G/\gamma}$ and stored (equation 2.1 and Fig. 2.2). The next barrier crossing was chosen as

described in the Gillespie algorithm [27] and the time was advanced accordingly. The temperature was set to 1 such that barrier heights and ruggedness $\gamma$ is given in units of kT in the following. For intermediate dimensional models ($d > 10$) with high ruggedness ($\gamma/d > 20$ kT), an enhanced sampling algorithm (see chapter 4) was used to generate random walk trajectories that were long enough to sample a sufficiently large PCA subspace.

| | |
|---|---|
| $T_0$ | $10^{5..12}$ [trajectory length] |
| $d$ | 3..200 |
| $\gamma/d$ | 3..30 [kT] |

Table 2.1: Parameter range of random walk simulations of the hierarchical free-energy landscape model

### 2.3.3 Estimation of the dependence of anomalous diffusion exponents on ruggedness

To determine which functional form $f$ describes best how the obtained anomalous diffusion exponents $\alpha_i$ decrease with ruggedness $\gamma$, we considered the exponential $\alpha^{f=1,\beta_1,\beta_2}(\gamma/d) = \beta_1 \exp(-(\gamma/d)/\beta_2)$, power law $\alpha^{f=2,\beta_1,\beta_2}(\gamma/d) = \beta_1 /(\gamma/d)^{\beta_2}$and linear $\alpha^{f=3,\beta_1,\beta_2}(\gamma/d) = \beta_1 \gamma/d + \beta_2$ dependence. Further, the data suggested normalizing the ruggedness by dimension $d$. To calculate the posterior probability $P(f, \beta_1, \beta_2|\{\alpha_i\})$ for each of these functional forms $f$, given the obtained $\{\alpha_i\}$, a Bayesian approach was used

$$P(f, \beta_1, \beta_2|\{\alpha_i\}) \propto P(\{\alpha_i\}|f, \beta_1, \beta_2)\, P(f, \beta_1, \beta_2).$$

Here, the likelihood $P(\{\alpha_i\}|f, \beta_1, \beta_2)$ of observing a set of scaling exponents at a given ruggedness value was described by a Gaussian distribution

$$P(\{\alpha_i\}|f, \beta_1, \beta_2) \propto \exp\left(-\sum_{\gamma/d}\sum_{i} \frac{\left(\alpha_i - \alpha^{f,\beta_1,\beta_2}(\gamma/d)\right)^2}{\sigma^2(\gamma/d)}\right), \tag{2.7}$$

and $\alpha^{f,\beta_1,\beta_2}(\gamma/d)$ was chosen as a linear, exponential or power law function as described above. A constant prior $P(f, \beta_1, \beta_2)$ was assumed for each param-

eter of the likelihood function. The variance $\sigma^2(\gamma/d)$ of the Gaussian probability distributions was approximated by $\sigma^2(\gamma/d) = a \cdot \gamma/d + b$. To determine the two parameters $a$ and $b$, we generated 1,000 trajectories each for $\gamma/d \in 5\,\mathrm{kT}, 15\,\mathrm{kT}, 17\,\mathrm{kT}, 20\,\mathrm{kT}$, respectively, and calculated the respective variances. To these, the above linear function was fitted. Posterior probabilities for the function and their two parameters were determined using Gibbs sampling [28] with 100,000 steps.

### 2.3.4 Ruggedness and dimensionality estimates

To estimate ruggedness and dimensionality for given $\alpha_i$, in the absence of an analytical expression, the likelihood of observing anomalous diffusion exponents in random walk trajectories that were generated in models with given ruggedness and dimensionality was estimated. To that end, a joint kernel density was estimated from ruggedness $\gamma$ and dimensionality $d$ parameters as well as $n$ PCA eigenvalue anomalous diffusion exponents $\alpha_i$ a $(n+2)$-dimensional using a kernel density estimator [29] (see Fig. 2.4). Given a set of $n$ PCA scaling exponents $\{\alpha_i\}$, a likelihood $p(\gamma/d, d|\{\alpha_i\})$ was calculated from the joint kernel density. The number $n$ of PCA eigenvalues that are sufficiently large ($> 10^{-5}$) depends on the length of a random walk/trajectory. Due to the limited amount of sampling in random walks, $n = 6$ PCA eigenvalues were sufficiently large to determine anomalous diffusion exponents in all of the generated random walk trajectories and were used for the kernel density estimation. For a set of anomalous diffusion exponents obtained both from random walks as well as from MD trajectories, the most likely ruggedness and dimensionality values as an estimate were used.

### 2.3.5 Protein selection

The 500 proteins were selected using the protocol found in [30]. In this protocol, nonhomologous proteins were selected from the protein data bank (PDB) [31] such that a large range of small globular proteins with less than 90% sequence

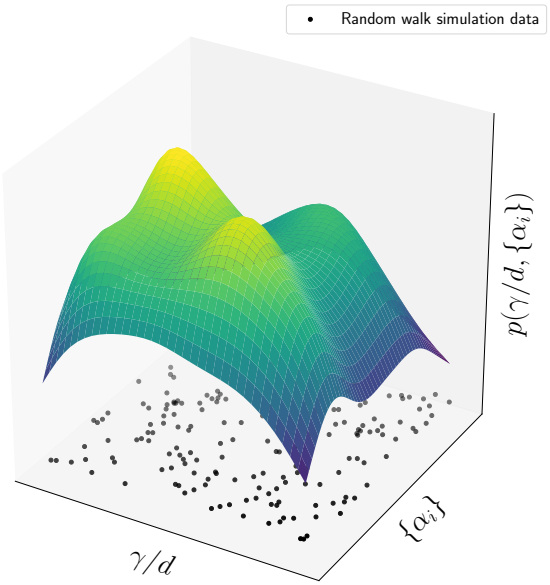


Figure 2.4: Dimension reduced sketch of the probability density of observing a set of anomalous diffusion exponents $\{\alpha_i\}$ for a given ruggedness $\gamma/d$ and dimensionality $d$. To estimate this probability density, random walk trajectories in 3 to 200 dimensions and ruggedness values of 3 to 30kT were generated and anomalous diffusion exponents $\alpha_i$ were determined. A Gaussian kernel density estimator to obtain a probability density from the simulation results.

identity was retrieved. Monomeric structures without gaps consisting of only standard residues were used. From the remaining protein structures, those containing polymeric or non-constitutive ligands were excluded. The selected pdb codes can be found in the supplement 2.6.2 and 100 ofthe selected structures in supplement 2.6.3.

Among the 500 selected proteins, 100 enzymes and non-enzymes were selected to perform three additional 1 $\mu$s MD-simulations.

### 2.3.6 Generation of MD trajectories

For each of the 500 selected proteins, MD simulations were performed using the simulation package software GROMACS 2018 [32]. Starting structures were obtained as described above in section 2.3.5. Solvent (TIP4P-Ew water model [33]) and ions ($Na^+$ and $Cl^-$) were added, establishing a salt concentration of $0.15\,\mathrm{mol\,l^{-1}}$ and neutralizing the overall system charge. A triclinic box with periodic boundary conditions was used with a 1.5 nm distance between solute and box boundary. Prior to each simulation run, energy minimization was performed using the GROMACS steepest descent algorithm until convergence was reached. This energy minimization was followed by a 1 ns (NPT) MD simulation to equilibrate the system. After energy minimization and equilibration a $1\,\mu$s MD trajectory was generated for each protein using Amber99*ildn force field [34] with a 2.5 fs time step with virtual sites [35]. All bond lengths were constrained, using the Settle algorithm [36] for the solvent and Lincs algorithm [37] for the solute, with a Lincs order of 4 during energy minimization and equilibration and 6 in the production run. Van-der-Waals forces were ignored for distances $> 1$ nm and Coulomb forces were calculated using the particle mesh Ewald method [38] with a real-space cutoff of 1 nm, PME order of four and a Fourier grid spacing of 1.2 Å.

For 200 of the selected proteins comprising 100 enzymes and 100 non-enzymes, three additional microsecond trajectories were calculated following the same protocol to estimate the statistical uncertainty of the determined ruggedness and dimensionality.

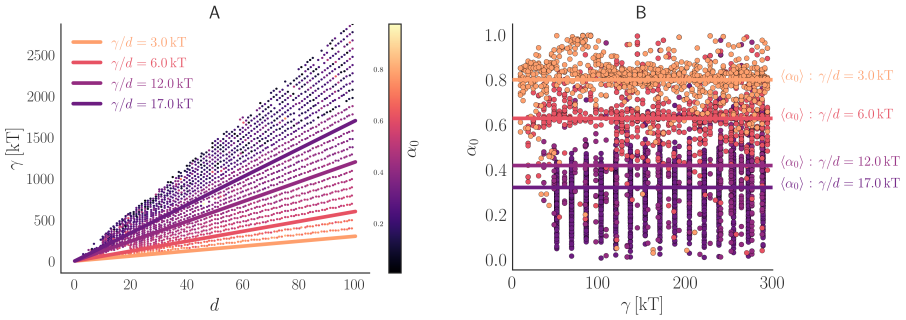


Figure 2.5: (A) Dependence of the scaling exponent of the largest PCA eigenvalue $\alpha_0$ on ruggedness $\gamma/d$ and dimensionality $d$. The color of each point represents an average scaling exponent averaged over all simulations with corresponding ruggedness and dimensionality values. (B) Dependence of anomalous diffusion exponents on ruggedness $\gamma$ for different ratios of $\gamma/d$

## 2.4 Results and Discussion

### 2.4.1 Anomalous diffusion in intermediate dimensional hierarchical models

Using random walk trajectories, generated as described in section 2.3.2, we first determined how anomalous diffusion exponents depend on the ruggedness and dimensionality of the hierarchical lattice model shown in Fig. 2.5. Almost normal diffusion ($\alpha = 1$) is seen for small ruggedness parameters $\gamma$, with increasingly strong subdiffusion ($\alpha < 0.1$) for larger $\gamma$ as shown in Fig. 2.5 A. Notably, similar $\alpha$ are seen for regions of similar $\gamma/d$ ratios as shown in Fig. 2.5 B.

For an explanation of this behavior, note that a trajectory is dominated by crossings of the lowest barriers $\Delta G_{\mathrm{min}}$. For the hierarchical lattice model, it follows from eq. 2.1 that for each visited state, the lowest of the $2d$ adjacent

barriers is distributed as

$$p(\Delta G_{\min}) \;=\; \frac{2d}{\gamma}\,\exp\left(-\frac{2d\,\Delta G_{\min}}{\gamma}\right). \tag{2.8}$$

As this distribution is a function of $\gamma/d$, similar anomalous diffusion exponents are expected for equal ratios of $\gamma/d$. This idea also explains why strong subdiffusion is only observed for unexpectedly high $\gamma$, particularly for large dimensionalities $d$. This finding motivates the use of this ratio or 'normalized' ruggedness as an argument for the functional dependence $\alpha(\gamma/d)$ further below.

Next, we compared our numerically obtained anomalous diffusion exponents $\alpha$ to the mean-field approximation and asked how accurately $\gamma/d$ can be estimated from $\alpha$. To that aim, Fig. 2.6 shows, as a violin plot for the largest four PCA eigenvalues, how much scaling exponents $\alpha$ scatter when derived from single trajectories for different landscapes as a function of $\gamma/d$. Note that the considerable width of these distributions results not only from the stochastic nature of the individual trajectories and the underlying energy landscapes but also from their different ruggedness and dimension for given $\gamma/d$. As expected, increasing subdiffusion is seen for increasing $\gamma/d$. For the smaller eigenvalues and small $\gamma/d$, some superdiffusion is seen as was already explained in terms of ballistic motion [26]. Overall, much weaker subdiffusion is seen compared to the mean field approximation (dashed lines), with decreasing discrepancy for larger dimensions, as also expected. Scaling exponents of large PCA eigenvalues decrease faster with increasing $\gamma/d$ and show a lower variance, mainly due to better sampling of these coordinates. Notably, the shown scatters generally exhibit large overlaps for adjacent $\gamma/d$ values, particularly for larger $\gamma/d$, which suggests that reconstructions of ruggedness and dimension from subdiffusion exponents $\alpha$ involve considerable uncertainties. These will be explored further below.

As our observed anomalous diffusion exponents deviate considerably from the functional relation 2.3 derived in a mean-field approximation (black dashed line in Fig. 2.6), we asked which function $\alpha^{f,\beta_1,\beta_2}(\gamma/d)$ describes the observed mean anomalous diffusion exponents best. As functional forms, we considered a

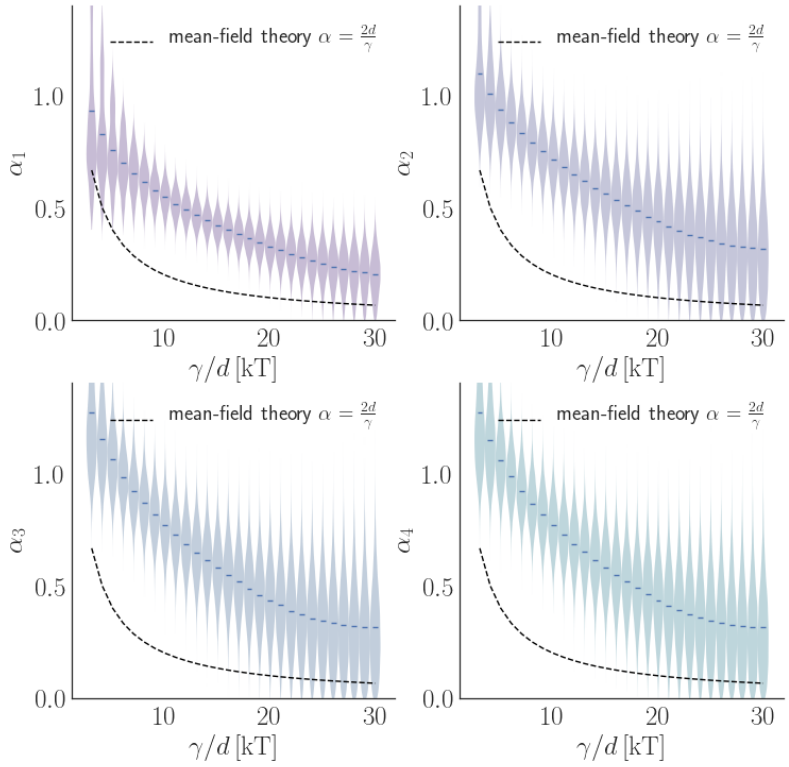


Figure 2.6: Distributions of scaling exponents of the first 4 PCA eigenvalues $\alpha_0$ (top left), $\alpha_1$ (top right), $\alpha_2$ (bottom left) and $\alpha_3$ (bottom right) in dependence of ruggedness $\gamma/d$. Probabilities of observing an anomalous diffusion exponent at a given ruggedness value are represented by violins. The dashed black line indicates the mean-field approximation.

power law (see caption of Fig. 2.7) as the generalization of the mean-field result, an exponential decay as a plausible alternative, and a linear function for comparison. For these three functional forms, posterior probabilities obtained via Gibbs sampling (see methods) are shown in Fig. 2.7 A. As an example, Fig. 2.7 B shows the posterior distributions for the respective parameters $(\beta_1, \beta_2)$ for

the exponential function. As can be seen in the figure, the highest posterior probability is obtained for the exponential function and in that sense describes our numerical results the best among the three considered functions. Remarkably, the power law, for which the mean-field theory eq. 2.3 is a special case for $\beta_1 = 2, \beta_2 = 1$, turns out to be the least probable, which suggests that a simple modification of the mean-field theory will most likely not suffice for a quantitative explanation of the anomalous diffusion exponents. We conclude that the underlying assumption that no trajectory visits any state twice most likely does not provide a good approximation for intermediate dimensional models.

However, although the exponential function fits the data best, it is only a slightly better description of the numerical anomalous diffusion exponents (see colored lines in Fig. 2.7). Therefore, we did not use any of these three functions to extract ruggedness and dimensionality from the anomalous diffusion exponents extracted from protein MD simulations further below, but rather resort to a probabilistic approach. To that end, we used a kernel density estimator to

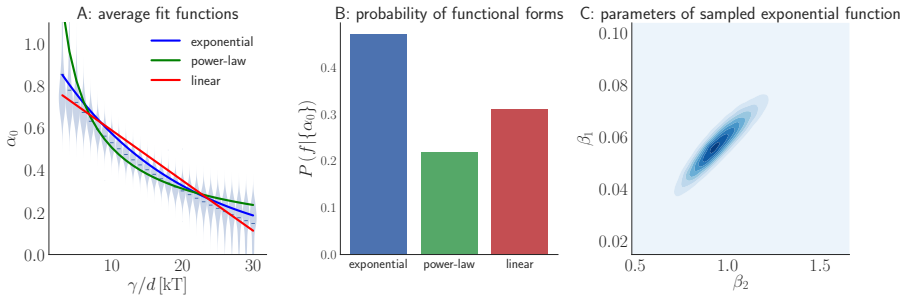


Figure 2.7: (A) Average fit function for three selected functional forms (i.e., exponential, power law and linear) of the relation between mean scaling exponent $\langle\alpha_0\rangle$ of the first PCA eigenvalue and $\gamma/d$ based on the scaling exponent distributions of generated random walks.(B) Posterior probabilities of exponential $\alpha^{f=1,\beta_1,\beta_2}(\gamma/d) = \beta_1 \exp(-(\gamma/d)/\beta_2)$, power law $\alpha^{f=2,\beta_1,\beta_2}(\gamma/d) = \beta_1 /(\gamma/d)^{\beta_2}$ and linear $\alpha^{f=3,\beta_1,\beta_2}(\gamma/d) = \beta_1 \gamma/d + \beta_2$ functions obtained from a Gibbs sampling. (C) Distribution of posterior probability of parameters for the exponential function.

model the joint probability density of anomalous diffusion exponents $\alpha_1, \ldots, \alpha_4$, $\gamma/d$, and $d$ as observed in a total of 40,000 random walks (see methods and dimension reduced sketch in Fig. 2.4). This probability density will serve to obtain distributions of $\gamma/d$ and $d$ for given $\alpha_1, \ldots, \alpha_4$ by marginalization.

Before discussing the obtained $\gamma/d$ and $d$, we used 2200 trajectories to estimate the uncertainty of these values by two independent approaches, from the variance of the marginalized distribution of $\gamma/d$ and $d$ via cross-validation. Figure 2.8 shows for each of the trajectories (blue dots) the actual error of the estimate, i.e., the deviation of the estimated normalized ruggedness (A) and dimensionality (B) from their known values that were used to build the respective underlying energy landscapes. For comparison, the average error estimate obtained from the marginalized distributions is shown as a black dashed line. In addition, the red line shows the cross-validation in terms of the average actual error for 100 trajectories with the same ruggedness (or dimensionality) values, which have not been used for the training of the kernel density. We obtained an overall mean error of 4.2 kT for the ruggedness estimate and 10 dimensions for the dimensionality estimate. The mean relative error of ruggedness estimation is moderate and increases with increasing ruggedness as expected, whereas the mean relative error of dimensionality estimates is substantially larger and independent of dimensionality. Overall, the hierarchical lattice model suggests that it should be possible to estimate the ruggedness of proteins rather reliably from anomalous diffusion exponents that were obtained via trajectory length dependent principal component analysis of atomistic simulations.

### 2.4.2 Anomalous diffusion in realistic protein free-energy landscapes

To this end, we used the above probabilistic model to explore ruggedness and dimensionality of the free-energy landscapes of 500 small globular proteins selected to cover known folds and functions as described in methods 2.3.5. We carried out a 1 $\mu$s MD-simulation for each of these 500 proteins and performed

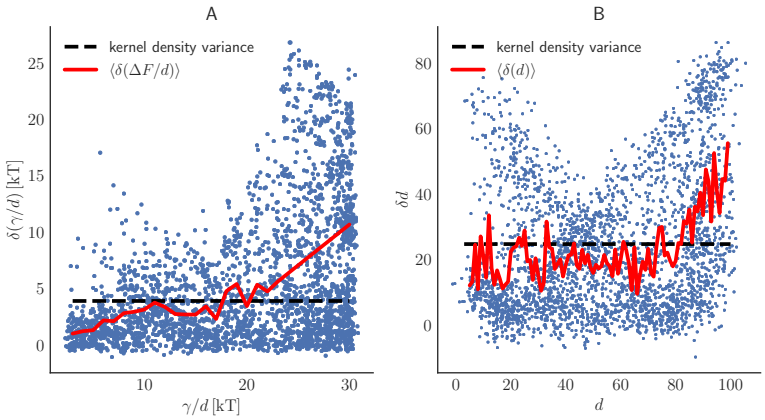


Figure 2.8: Error estimates of ruggedness and dimensionality based on cross-validation: We randomly picked 100 random walk trajectories per ruggedness value and estimated ruggedness and dimensionality of the corresponding intermediate dimensional hierarchical model using a maximum likelihood estimator. (A) Dependence of the error in ruggedness estimates on ruggedness $\gamma/d$. The mean error (red line) of ruggedness estimates increases with higher ruggedness values. (B) Dependence of the error in dimensionality estimates on dimensionality. The mean error in dimensionality estimates shows no clear dependence on dimensionality. Gaussian noise with 1% variance was added in both plots for visualization purposes.

a trajectory length dependent principal component analysis (tPCA) for each of the trajectories as described in methods 2.3.1. From least square fits to the trajectory length dependent largest eigenvalue scaling exponents $\alpha_1$ were obtained. Figure 2.9 A shows the distribution of scaling exponents jointly as a function of protein size (number of $C_\alpha$ atoms $N$) as described by the dimension $d_{\text{conf}} = 3N$ by the respective configurational space.

As can be seen, almost all of the obtained scaling exponents show subdiffusion ($\alpha_0 < 1$). In fact, ca. 90% of the scaling exponents are smaller than 0.6 and the mean anomalous diffusion exponent is 0.3. Remarkably, no significant corre-

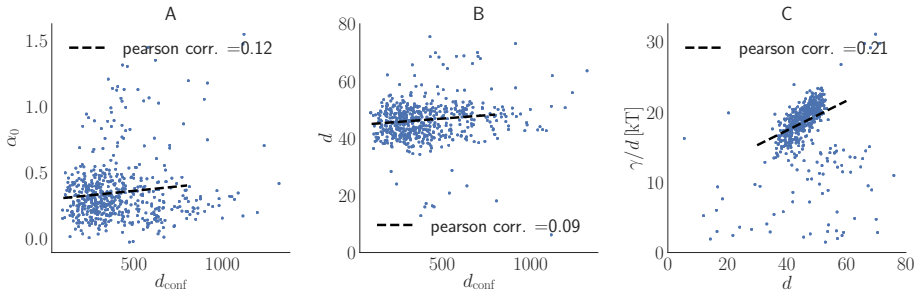


Figure 2.9: (A) Subdiffusion exponents $a_0$ as a function of configuration space dimensionality obtained from MD trajectories of 500 small globular proteins with configuration space dimensionality $d_{\mathrm{conf}} = 3N$, where $N$ is the number of $C_\alpha$ atoms.(B) Estimated effective dimensionality $d$ and configuration space dimensionality.(C) Frequency of dimensionality and ruggedness estimates of 500 small globular proteins obtained from microsecond molecular dynamics simulation.

lation between $\alpha_0$ and configuration space dimensionality $d_{\mathrm{conf}}$ is seen (Pearson correlation coefficient $c = 0.12$).

Using these anomalous diffusion exponents, we estimated the ruggedness $\gamma$ and effective dimensionality $d$ of the 500 protein free-energy landscapes via the above probabilistic model. Fig. 2.9 C shows the distribution of dimensionality and ruggedness estimates of $\gamma/d$ and $d$, where the ruggedness has been normalized by the effective dimension as suggested for the simple hierarchical grid model and shown in Fig. 2.5.

As can be seen, ruggedness values between $15 - 20\,\mathrm{kT}$ per dimension dominate, as well as effective dimensionalities $d$ between 40 and 60. This result is reproducible for four independent sets of MD simulations of a test set of 200 proteins as described in methods 2.3.5. For this test set, an average standard deviation of 1.1 kT for ruggedness and 4.8 dimensions for the dimensionality was obtained, which is lower than the expected error from our random walk simulations. We attribute these low errors to the fact that MD simulations were started from the same starting structure and therefore explore similar regions

in their free energy landscape, whereas in the case of the hierarchical lattice model, each trajectory is simulated in a different free energy landscape.

Unexpectedly, despite the fact that there is no correlation between the ruggedness coefficient and the protein size, Fig. 2.9 C shows a strong correlation between normalized ruggedness $\gamma/d$ and *effective* dimensionality $d$ (Pearson correlation coefficient $c = 0.21$). We asked if this correlation is due to a possible correlation between protein size and effective dimensionality. However, no such correlation is seen in the respective scatter plot (Fig. 2.9 B, Pearson correlation coefficient $c = 0.09$). Taken together, these results suggest that both the effective dimensionality and normalized ruggedness of a protein do not depend on its size and rather are adapted to the particular function of each single protein. Furthermore, it is remarkable that the ranges of both normalized ruggedness and effective dimensionality (by a factor of about 1.5) are much smaller than the scatter of protein sizes (by a factor of ca. 5) among the selected 500 proteins. This finding suggests that, quite generally, these narrow ranges are optimal for the function of essentially every protein.

## 2.5 Conclusion

In this work, we developed a method to connect simple hierarchical free-energy lattice models to atomistic simulations of biological macromolecules. To this end, we have characterized the high-dimensional free-energy landscape of 500 small globular proteins in terms of effective dimensionality and distribution of free energy barrier heights. These quantities have been obtained from anomalous diffusion exponents observed in microsecond molecular dynamics trajectories of these proteins.

For the hierarchical free-energy lattice model, we assumed an exponential distribution $p(\Delta G) \propto \exp(-\Delta G/\gamma)$ of static barrier heights, where $\gamma$ denotes the 'ruggedness' of the energy landscape, similar to a disorder temperature. While analytic expressions have been derived for 1-dimensional [9] and high-dimensional lattices [10], we are not aware of any result for the intermediate

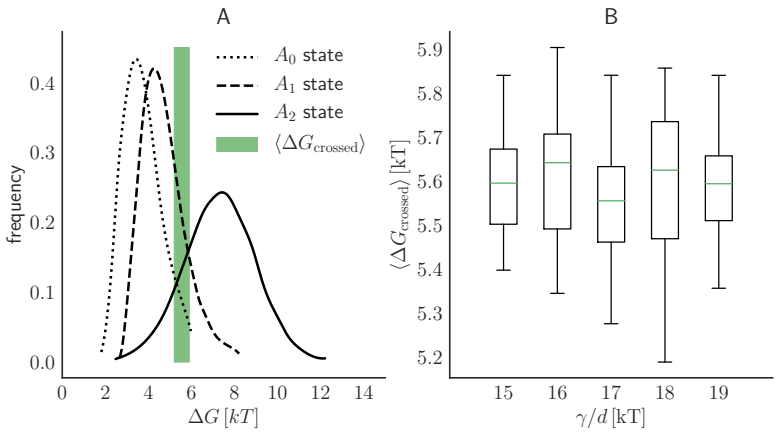


Figure 2.10: (A) Distributions of barrier heights of 3 states of myoglobin $(A_0, A_1, A_2)$ estimated from flash photolysis experiments [23](B). Boxplots of distributions of barrier heights $\delta G_{\text{crossed}}$ that were crossed in random walks in intermediate dimensional hierarchical lattice models with typical ruggedness and dimensionality as estimated for the selected proteins $(40-60$ and $15-20\,\text{kT}$ per dimension).

effective dimensions relevant for biomolecules; we therefore resorted to a numerical approach.

To this end, we carried out random walk simulations and found indeed anomalous diffusion exponents that deviate from both the 1-dimensional and high-dimensional limiting cases. A Bayesian analysis showed that, overall, anomalous diffusion exponents decrease less steeply with increasing ruggedness and most likely not by the inverse-law of the two limiting cases. These significant deviations suggest different mechanisms from which anomalous diffusion behavior arises.

In the limit of infinite dimensions, random walks are equivalent to high dimensional continuous time random walks (CTRW), for which no state is visited more than once and therefore the mean-field description is accurate. In particular, the probability of returning to a previously visited state approaches zero.

As a result, anomalous diffusion in CTRWs is caused by waiting time distributions that are heavy-tailed (i.e., diverging averages) [39] and dominated by a few extreme waiting times in states enclosed by high barriers. In contrast, for the intermediate dimensional models discussed here, recrossings were observed with consequences, the analysis of which is beyond the scope of this work. Based on our observations (data not shown), we speculate that the anomalous diffusion in intermediate-dimensional models most likely originates – similar to the 1-d case – from high free energy barriers confining random walk trajectories to a region in conformational space. However, anomalous diffusion in intermediate-dimensional models differs from that of 1-d models in that, due to the higher dimensionality of these regions, high free-energy barries are circumvented. This effect results in a fractal-like topology of the subregion actually accessed by trajectories [14]. This scenario is also supported by our observation that the height distribution of actually crossed barriers is much lower than the overall barrier distribution of the free-energy landscape.

In that sense, this study reveals a connection of the two main conceptual frameworks explaining anomalous diffusion in protein dynamics, diffusion on fractal geometries and hierarchical free-energy landscapes. Specifically, we have shown that, for intermediate dimensionality, fractal-like topologies of accessible configurational space arise necessarily from dynamics in hierarchical energy landscapes with very high ruggedness.

Using our numerical results, we asked how accurately ruggedness and dimensionality can be estimated based on anomalous diffusion exponents obtained from non-equilibrium trajectories. For the hierarchical lattice model, we showed via cross-validation that a maximum likelihood estimate yields an accuracy of $4.2\,\mathrm{kT}$ for ruggedness and 10 dimensions for the effective dimensionality.

This result enabled us to use our method to estimate ruggedness and dimensionality based on anomalous diffusion exponents we observed in MD trajectories of 500 small globular proteins. We obtained typical ruggedness estimates in the range of $15 - 20\,\mathrm{kT}$ per dimension and effective dimensionality of $40 - 60$ . The robustness of the ruggedness and dimensionality estimates for three inde-

pendent MD simulations shows that the intermediate-dimensional hierarchical model indeed captures features of protein free-energy landscapes that govern the anomalous diffusion behavior in protein dynamics.

It is remarkable that neither the effective dimensionality nor the ruggedness correlates with protein size, whereas there is a significant correlation between effective dimension and ruggedness. Further, the ranges of both normalized ruggedness and effective dimensionality are much smaller than the scatter of protein sizes (by factors of about 1.5 and 5, respectively) among the selected 500 proteins. Taken together, we conclude that these two properties of the free-energy landscape of a protein are rather adapted to the particular function of each single protein, and that, quite generally, these narrow ranges are optimal for the function of essentially every small globular protein.

## 2.6 Supplements

### 2.6.1 Scaling of PCA eigenvalues for high dimensional hierarchical models

The dimensionality in hierarchical lattice models determines the amount of possible transitions $2d$ of states to their neighboring states. The amount of possible transitions is inversely proportional to the probability of returning to the state which was previously visited as the number of possible transitions increase. In the limit of high-dimensionality, this probability vanishes as recrossings get less likely with an increasing amount of possible transitions. Therefore, all barriers of visited states can be neglected in this limit, such that it can be assumed that in every step a new set of free energy barriers is encountered. These free energy barriers both determine the probability of the direction of the next step and the waiting time $t$ in the current state until a barrier crossing event occurs [27]. Because barriers are distributed isotropically within the grid, a random walk trajectory resembles a free diffusion process in this approximation, if only transitions between states are considered and waiting times are neglected. It has been shown [26] that PCA eigenvalues of free diffusion processes scale linearly with the number of steps $n$

$$\lambda(T) \propto n(T). \tag{2.9}$$

The total time $T = \sum_{i=0}^{n} t_i$ is given by the sum of all waiting times in the individual states, which are random variables. The waiting time distribution within $p(t)$ a state depends on ruggedness $\gamma$ and dimensionality $d$. Because in the high dimensional limit no state is revisited, waiting time distributions $p(t)$ are obtained directly by a probability transformation of the barrier height distribution and is given by

$$p(t) \propto t^{1-\frac{\gamma}{2d}}. \tag{2.10}$$

For $\gamma/d < 1$, waiting times $t_i$ have a well-defined expectation value, such that for $n \to \inf$

$$T \approx n \langle t_i \rangle \tag{2.11}$$

$$\lambda(T) \propto \frac{T}{\langle t_i \rangle} \tag{2.12}$$

which leads to normal diffusion behavior. Anomalous diffusion behavior emerges for $\gamma/2d > 1$, because the expectation value for the waiting times $\langle t \rangle$ diverges for $n \to \infty$. However, for a finite number of transitions $n$, this expectation value is also finite and depends asymptotically on $T$ as

$$\langle t_i \rangle \propto T^{1-\frac{2d}{\gamma}}. \tag{2.13}$$

Inserting this result into 2.11 yields the scaling behavior of PCA eigenvalues in the approximation of high dimensional random walks

$$\lambda(T) \propto T^{\frac{2d}{\gamma}}. \tag{2.14}$$

For $\gamma/2d > 1$ the expectation value for the waiting times diverges.

### 2.6.2 List of PDB codes of the selected proteins

'1WIT', '3CHY', '1ENH', '1SHF', '1UBQ', '1MJC', '1A6N', '1ARB', '1CUN', '1WAS', '1EP0', '2PTH', '1QAU', '1EBD', '4WBC', '1DYN', '1JAM', '2GO0', '2GIW', '1IFC', '1BP5', '1HH8', '1BS2', '1IXA', '1IER', '1JD1', '1YPR', '1BFD', '1EQK', '11AS', '1WE8', '1Z9D', '1Y8B', '1YEL', '1XX3', '1XHK', '1XQO', '1XKR', '1WHN', '1WIN', '1YB3', '1WK1', '1WFY', '1XSZ', '1WJW', '1WHB', '1WJG', '1WJJ', '1WFT', '1WB7', '1IAD', '1FKB', '1FZW', '2HNP', '1EV4', '1ESJ', '1J1Y', '3GRS', '1EHE', '1G5B', '1FZT', '1E09', '2IFE', '1IYU', '2GLT', '1E6Y', '5HPG', '1G6L', '1FHQ', '1I11', '1IMF', '1J22', '1G61', '1HPL', '1I6A', '1FQN', '1G5M', '1IHC', '1GEF', '1JPU', '1C3P', '1GC7', '1FVL', '1G03', '1G24', '1JH3', '1EJF', '1GQY', '1ITV', '1FID', '1IGP', '1FUO', '1EGL', '1GD5', '1FSZ', '1EO9', '1H8H', '1JR2', '1H41', '1HJZ', '3GAR', '1JAW', '1ILE', '1EY1', '1EW4', '2HGF', '1FVA', '1H5P', '1I39', '1EPU', '2FU3', '1E9T', '1IAZ', '1IJA', '1IVH', '1FX3', '1JR3', '1HT2', '1JYH', '1EUV', '1JHF', '1IFG', '1HUS', '3GCC', '1HD8', '1I4J', '2GYK', '1JW3', '1IMU', '1F7T', '1EQ1', '2G03', '1IJY', '1ETH', '1G41', '1GM7', '1IUH', '1BM8', '1EQ6', '1J33', '2EZN', '1GZT', '1G1E', '1JW2', '1J2M', '1GH9', '1IS1', '1GHH', '1J26', '1FM7', '2IBS', '1EMW', '1IU3', '1GXL', '1IQO', '1JI8', '1HH2', '1HUF', '1G2R', '1GD8', '1F7W', '1GGG', '1GHT', '1EM8', '1HF2', '1JOB', '1EJ5', '1EZA', '1GGW', '1IUR', '2END', '1J0T', '1JR5', '1G9L', '1DVO', '1FOA', '1FJR', '1I4W', '1LBD', '1L8L', '2TRC', '1U9A', '2TGI', '1TFB', '1T0G', '1SWB', '1LMH', '1TUL', '1V4E', '1SRV', '4MAT', '1SSO', '1TYA', '1M6B', '1KK9', '1TPG', '1S04', '1MOP', '1KNB', '1L1D', '1KSV', '1KEO', '1L5I', '1LTU', '1UNN', '1KO7', '1LBV', '1KVN', '1K0F', '1K3C', '2LIS', '2MOB', '1MT6', '1KRA', '1L8R', '1K6K', '1MI1', '1M3I', '1TE2', '1S12', '1VDH', '1M1L', '1L9V', '1LR0', '1MWP', '1T3B', '1SQR', '1K0S', '1K3W', '1LNS', '1UG2', '1VMG', '1MP1', '1LRE', '1K19', '1LML', '1U02', '1QLP', '1OTP', '1RV9', '1RYU', '1P9Y', '1OBL', '1QK9', '2NMU', '2NEF', '1OP4', '1RKE', '1QGI', '1PP1', '1RLH', '1Q57', '1RKI', '1Q60', '1QKF', '1PU1', '1N6Z', '1PV5', '1OQV', '1NIJ', '1POZ', '1OBG', '1ODG', '1QZM', '1Q92', '1O99', '1P35', '1NI5', '1R5E', '1RYK', '1RP4', '1R0D', '1RXQ', '1QZ4', '1Q5Z', '1PVE', '1NY9', '1R9K',

’1NKT’, ’1Q1V’, ’1QC7’, ’1PUJ’, ’1NG6’, ’1NK6’, ’1AB2’, ’1QAZ’, ’1B6B’, ’1NTN’, ’1EZG’, ’1R4V’, ’1BSG’, ’1D6T’, ’1CVZ’, ’1CEQ’, ’1D0N’, ’1HCC’, ’1BF0’, ’1UXC’, ’1PHP’, ’1DDG’, ’1D8V’, ’1DP0’, ’1AH6’, ’1DXK’, ’1D0B’, ’1AH2’, ’3PMG’, ’1PNO’, ’1O6D’, ’1B10’, ’1FAD’, ’1R3B’, ’1EVL’, ’1UNK’, ’1DSX’, ’1C25’, ’2PGI’, ’3PRO’, ’1WWR’, ’1QCV’, ’1BW3’, ’1DVG’, ’1F17’, ’1AI9’, ’1CD5’, ’1DCO’, ’1B2P’, ’1DTW’, ’1LDL’, ’1CDZ’, ’1BKP’, ’1BG2’, ’3RUB’, ’1BYR’, ’1BEA’, ’1QPM’, ’1CRN’, ’1OAG’, ’1DPT’, ’1C2A’, ’1AKO’, ’1CVR’, ’1AKZ’, ’1E0L’, ’1H9F’, ’1BOL’, ’1CMZ’, ’1BM0’, ’1K2Y’, ’1BY1’, ’1FD3’, ’1B78’, ’1B8W’, ’1A6S’, ’1WDE’, ’1WU2’, ’1OPS’, ’1NGN’, ’6RHN’, ’1AT0’, ’1F68’, ’1BPX’, ’1N2J’, ’2FUS’, ’1DHN’, ’1RWC’, ’1CIP’, ’1YCQ’, ’1YS9’, ’1A0G’, ’1NOG’, ’1CZ4’, ’1PKY’, ’1ERD’, ’1DPB’, ’1MM0’, ’1EAI’, ’1AA3’, ’1RL6’, ’1C3G’, ’1TSF’, ’1R2Z’, ’1UK3’, ’1CIY’, ’1BUH’, ’1AUZ’, ’1IHN’, ’1HIC’, ’1BX8’, ’1B6Z’, ’1A3A’, ’1RHX’, ’1DU5’, ’1I2T’, ’1DZO’, ’1EO0’, ’1CMI’, ’1WGR’, ’1C44’, ’2BES’, ’1A1X’, ’1D1L’, ’1Y6X’, ’1RLO’, ’1B04’, ’1RQL’, ’1XJH’, ’1AP8’, ’1A5M’, ’1XD3’, ’1XHS’, ’1NYN’, ’1Z52’, ’1CFE’, ’1QTS’, ’1NYO’, ’1W4H’, ’1UJ8’, ’1QQV’, ’1PDO’, ’1QQH’, ’1RQS’, ’1MGT’, ’1D1R’, ’1MMS’, ’1XS8’, ’2AHC’, ’1BLE’, ’1RLK’, ’1OH1’, ’1WOO’, ’1GMU’, ’1JJU’, ’1XWM’, ’1KJS’, ’1VG5’, ’1BSH’, ’1BC9’, ’1UTG’, ’1DWU’, ’1LFP’, ’1BGW’, ’1YGE’, ’1D5T’, ’1PVS’, ’1IUQ’, ’1YGY’, ’1DZF’, ’1C8Z’, ’1JRM’, ’1CBY’, ’1B75’, ’1NWB’, ’1J1V’, ’1H3L’, ’1BEG’, ’1EWS’, ’1WJ2’, ’1E8P’, ’1ADN’, ’1CO4’, ’1DCQ’, ’1D0Q’, ’1P7A’, ’1ZFD’, ’1UOY’, ’1AFP’, ’1AD2’, ’1DT9’, ’1SYX’, ’1KPT’, ’1WHR’, ’1WIH’, ’1XHJ’, ’1WHZ’, ’1WHQ’, ’1OUO’, ’1B2V’, ’1WN9’, ’1YDL’, ’1WPS’, ’2FFM’, ’1QW2’, ’1Q5F’, ’1O8R’, ’1XRS’, ’1MK0’, ’1WOT’, ’1YLQ’, ’1JBI’, ’1KAF’, ’1NNV’, ’1WJ6’, ’1NYR’, ’1DUJ’, ’1VCC’, ’1WFR’, ’1QHK’, ’2HBB’, ’1C97’, ’1WIK’, ’1S3A’, ’1OGD’, ’1RZW’, ’1RO7’, ’1AUA’, ’1AQT’, ’1WFJ’, ’1WFM’, ’1X7F’, ’1CL3’, ’1HOE’, ’1YEZ’, ’1FGP’, ’1WHM’, ’1WFW’, ’1NKG’, ’1CQ3’, ’1B7Y’, ’2B97’, ’1NPR’, ’1BGF’, ’1ABV’, ’1AF7’, ’1DP3’, ’1WHC’, ’1WIX’, ’1RQ6’, ’1OYW’, ’1S7E’, ’1TDP’, ’1YOZ’, ’1WGW’, ’1U84’, ’1PUZ’, ’1UG0’, ’1PV0’, ’1NO1’, ’2HP8’, ’1Z5B’, ’1HNR’, ’1WIJ’, ’1C1K’, ’1X9B’, ’1S2O’, ’1B0A’, ’1DOT’, ’1XS5’, ’1XVI’, ’1CBF’, ’1DJ0’, ’1RK6’, ’1Y51’, ’1W6X’, ’1IMT’, ’1AA7’, ’1ST7’, ’1G71’, ’1WPB’, ’1AMM’, ’1ALC’, ’2AYH’, ’2OVO’,

'1NKR', '5CPA', '1F7U', '4AKE', '1AST', '1BPI', '1CNV', '1BD8', '1GAD', '1A3H', '1A4V', '1BQG', '1WEK', '1LB6', '1CEM', '2BAA', '1AKE', '1AAJ', '1GSO', '1DDE', '2F21', '2HBA'

### 2.6.3 Structures of the 500 selected proteins

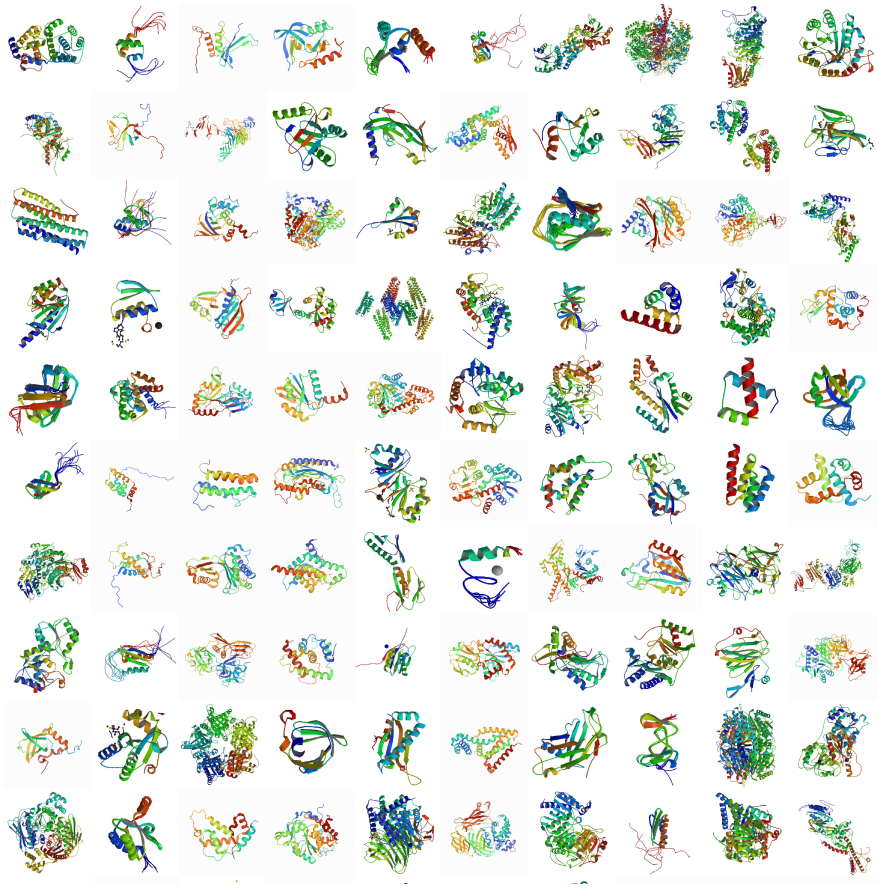

Figure 2.11: Structures of 100 of the selected 500 proteins used as starting structures for the MD simulations

# Chapter 3

## A universal scaling relation between accessible configuration space volume and escape rates in intermediate dimensional hierarchical free-energy landscapes

## 3.1 Introduction

The internal dynamics of proteins, the essential building blocks of life, govern their function. Due to the complex structure of proteins and their many-faceted interactions among their constituents and surroundings, their internal motions occur on a broad range of time scales from picoseconds to seconds. Early flash-photolysis experiments on the unbinding kinetics of carbon monoxide from myoglobin by Frauenfelder et al. suggested that this broad range arises from a hierarchical structure of protein free-energy landscapes [4].

The free-energy landscapes of complex biomolecules such as proteins consist of a vast number of isoenergetic configurations [40]. Early flash-photolysis experiments suggested that the observed non-exponential kinetics originates from hierarchically structured free-energy barriers between different protein configurations [4]. These free-energy barriers restrict the accessible configurations to a smaller subset of accessible configurations for a given time scale of the dynamics. The hierarchical structure of these free energy barriers ensures sufficiently high free-energy barriers to restrict the accessible configuration space to increasing but finite volumes on all time scales. Due to this relation between time scales and corresponding accessible configuration space, diffusion processes in such a hierarchically structured free-energy landscape are anomalous [9].

Such anomalous diffusion behavior was observed in collective motions in molecular dynamics trajectories of proteins [13] and small peptides [14]. However, different sources of this anomalous diffusion behavior have been suggested, besides a hierarchically structured protein free-energy landscape: a fractal structure of accessible configuration space [14] and a projection effect arising from the analysis of collective coordinates [15]. Here, we assume that the observed anomalous diffusion behavior is a feature of protein dynamics and not a projection artifact. Our results from chapter 2 suggested that the two remaining sources, i.e. a hierarchically structured free-energy landscape and a fractal accessible configuration space, are compatible. We observed frequent revisiting of

states in a well-defined region of accessible configuration space in higher dimensional hierarchical free energy landscapes, which lead us to the conjecture that these regions exhibit a fractal structure.

Such a hierarchical structure implies that a diffusion process like the internal dynamics of proteins as they explore their configuration space show anomalous diffusion. Indeed, anomalous diffusion was observed in current molecular dynamics simulations of peptides and small globular proteins [13] [14]. The source of this anomalous diffusion behavior was suggested to be either the hierarchical structure of protein free-energy landscapes [13] or the fractal structure of accessible protein configuration space [14]. Our results of chapter 2 suggested that both models are compatible because accessible configuration space in hierarchical free-energy landscapes has a fractal structure.

To test this conjecture and to elucidate the structure of accessible configuration space in hierarchical free-energy landscapes, we investigated the accessible configuration space in a $d$-dimensional hierarchical model free energy landscape that was used to model anomalous diffusion observed in proteins. These models are $d$-dimensional cubic lattices where static exponentially distributed barriers determine transitions between states. Here, the distribution of barriers $p(\Delta G) = 1/\gamma \exp -\Delta G/\gamma$ is governed by a parameter $\gamma$ that represents the 'ruggedness' of a model free-energy landscape. We have shown in the previous chapter that $\gamma/d$ determines the anomalous diffusion behavior of random walks in the model. Depending on how much $\gamma/d$ affects accessible configuration space volume and escape rates, it should be possible to determine to what extent the two contribute to the anomalous diffusion behavior. We found that $\gamma/d$ influences both escape rates and accessible configuration space volume from which we conclude that both quantities influence the subdiffusion behavior in hierarchical free-energy landscape. To our surprise, we found a universal scaling relation between the accessible configuration space volume and corresponding escape rate that quantifies the collective influence of both on the anomalous diffusion behavior.

Following this conjecture, we investigated in this work the accessible config-

uration space in random walks in 3 to 10 dimensional hierarchical models.

In particular, we investigated how accessible configuration space volumes and escape rates from these confined regions depend on the barrier height distribution which determines the anomalous diffusion behavior as we showed in chapter 2. How much accessible configuration space volumes and escape rates depend on the barrier height distribution shows how much both contribute to the anomalous diffusion behavior.

## 3.2 Theory

Our random walk simulations in the intermediate dimensional hierarchical lattice model presented in chapter 2 showed that trajectories only visited a limited set of states instead of exploring the full state space. Such a set of states is what we call accessible configuration space in the following. Additionally, our results suggested that the topology of the accessible configuration space is a source of anomalous diffusion, which led us to investigate the properties regions further.

To this end, we first specify in the following how accessible configuration space and corresponding escape rates govern diffusion processes in intermediate dimensional hierarchical models. In this model, states in configuration space are represented by nodes in a $d$-dimensional cubic lattice. Edges represent transitions between states such that each state in a $d$-dimensional cubic lattice has $2d$ neighbouring states. Transition rates $r_{ij}$ between states are determined by free energy barriers $\Delta G_i$

$$r_i = e^{-\Delta G_i} \tag{3.1}$$

where $\Delta G_i$ are uncorrelated random variables which are distributed according to

$$p(\Delta G_i) = \frac{1}{\gamma} e^{\Delta G_i/\gamma} \tag{3.2}$$

where $\gamma$ represents the 'ruggedness' of a model free energy landscape. We showed that random walks are dominated by the highest transition rates (lowest barriers). Both ruggedness and dimensionality affect the maximal transition rate

(minimal barrier) to leave a state: With increasing ruggedness it is less likely to encounter a low barrier within a random walk while with increasing dimensionality it becomes more likely because of the increasing number of possible transitions. We have shown that this effect approximately is compensated by increasing ruggedness such that $\gamma/d \approx$ const. Therefore, we introduced a normalized ruggedness $\gamma/d$ and found that this quantity determines anomalous diffusion in intermediate dimensional hiearchical free-energy landscapes.

Configuration space exploration in this model is described by diffusion process described by the propagator equation

$$\mathbf{p}(t + \Delta t) \; = \; \mathbf{M}(\Delta t) \cdot \mathbf{p}(t), \tag{3.3}$$

where $\vec{p} = (p_i \dots p_N)$ represents probabilities of occupying states in the cubic lattice with $N$ states. $M$ is a symmetric transition rate matrix where rates are normalized and determined by barriers such that

$$M_{ij} = \begin{cases} \frac{1}{\sum_{k=0}^{2d} e^{-\Delta G_{ik}}} e^{-\Delta G_{ij}} & \text{if transition in grid} \\ 0 & \text{else} \end{cases} \tag{3.4}$$

where $d$ is the grid dimension and $\Delta G_{ij} = \Delta G_{ji}$ are exponentially distributed random barriers as described in the previous section. Random walks can be generated using eq. 3.3 iteratively with delta distributed probability densities $p(t) = \delta(t)_{ij}$ which represent the occupied state in a random walk at time $t$. In each step of the next occupied state is drawn at random with probability $\mathbf{p}(t + \Delta t)$.

Probabilities of occupying a specific state $i$ at time $T$ are composed of the probability $P_i(n+1)$ of having arrived at this state after $n$ prior transitions and with occupation times $\{\Delta t_j\}_{j=0}^{n+1}$ such that $T = \sum_j^{n+1} \Delta t_j$ such that

$$p_i(T) = \tilde{p}_i(n+1)\, p(\{\Delta t_j\}_{j=0}^{n+1}) \tag{3.5}$$

where $p(\{\Delta t_j\}_{j=0}^{n+1})$ is the probability of observing a sequence of occupation times. Here, the first term $\tilde{p}_i(n+1)$ describes the effective configuration space

exploration while the latter describes the time spent in individual states. For a large number of (effective) steps $n \to \infty$ with frequent revisiting of states, the time spent in individual states is subject to the central limit theorem, such that $T \approx n \langle \Delta t \rangle$ and $p(\{\Delta t_j\}_{j=0}^{n+1}) \approx \delta(T - n \langle \Delta t \rangle)$. In this limit, the term is replaced with its expectation value and does not contribute to the anomalous diffusion behavior.

The remaining effective accessible configuration space is also given by a propergator equation

$$\tilde{\mathbf{p}}(n+1) = \tilde{\mathbf{M}} \cdot \tilde{\mathbf{p}}(n), \tag{3.6}$$

where $\tilde{\mathbf{p}}(n) = (p_1 \dots p_n)$ are the probabilities of arriving at a state after $n$ transitions between states. The entries of the transition matrix $\tilde{\mathbf{M}}$ is given by

$$\tilde{M}_{ij} = \begin{cases} 0 & i = j \\ \frac{1}{r_i} M_{ij} & i \neq j \end{cases} \tag{3.7}$$

and is non-symmetric and row stochastic such that an eigenvalue decomposition exists with real eigenvalues $1 \geq \lambda_i > -1$. Due to the row stochasticity there is one eigenvalue $\lambda_0 = 1$ with a corresponding right eigenvector $v_0 = \vec{\mu}$ which represents the probability $\mu_i$ of visiting a state $i$ in phase space. All remaining eigenvalues are degenerate due to disorder introduced through the random barriers and corresponding left and right eigenvectors have a finite support where they are non zero. The number of states $V$ which are included in the support of an (left or right) eigenvector $v_i$ represent the accessible configuration space volume on a time scale given by $|\lambda_i|$.

To leave the accessible configuration space region with volume $V$ a state at its boundary $\Omega$ (see Fig.3.1) has to be occupied and the next occupied state must be outside. Therefore, the escape rate $r$ is given by:

$$r = \sum_{i \in \Omega} \mu_i \left( \sum_{j \notin V} \tilde{M}_{ij} \right). \tag{3.8}$$

The inverse of the escape rate $r$ corresponds to the amount of steps $n$ that are on average needed in order to escape from the accessible configuration space

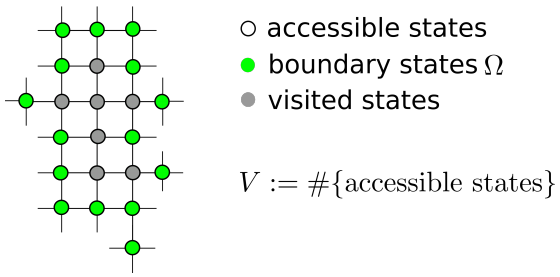


Figure 3.1: *Sketch of accessible configuration space volume in case of a 2-dimensional lattice.* Random walk trajectories in the intermediate dimensional hierarchical lattice model frequently revisit a limited of accessible states shown as circles in the figure. We define the number of these states as the accessible configuration space volume. In order to escape from

volume. To investigate how the accessible configuration space volume $V$ and escape rates $r$ depend on $\gamma/d$ we simulated random walks in intermediate dimensional hierarchical free energy landscapes as analytical results are not at hand for these expressions.

Due to the high dimensionality of the considered landscapes analytical expressions for the eigenvectors of $\tilde{\mathbf{M}}$ are not available. Also, numerical eigenvector decomposition is not feasible for a pre-computed grid as the number of states in a cubic lattice grows exponentially with the number of dimensions. Because of this circumstance, we resorted to a numerical approach where regions of accessible configuration space are created on-demand as described in chapter 2.3.2.

## 3.3 Methods

We generated random walks in intermediate dimensional lattices with $40 - 60$ dimensions and $15 - 20\,\mathrm{kT}$ per dimension . Simulations of random walks were performed using Gillespie's algorithm [27]. In these simulations, portions of the free-energy landscape lattice were generated dynamically on demand as

During the simulations the number of visits to each state within the lattice was monitored. Once each state is revisited at least 100 times, the total number of visited states is used as the accessible configuration space volume $V$. Next, simulations are continued until a new state that was not previously occupied was found. The inverse number of transitions between states serves as an estimate for the escape rate $r$. This simulation protocol was repeated 100 times for each considered dimensionality and ruggedness value.

## 3.4 Results and Discussion

Using the simulation protocol described in section 2.3 we first asked how accessible configuration space volume $V$ and escape rates depend on normalized ruggedness and dimensionality of our model.

Figure 3.2 A shows the dependence of average accessible configuration space volume on normalized ruggedness and dimensionality. As can be seen, increased

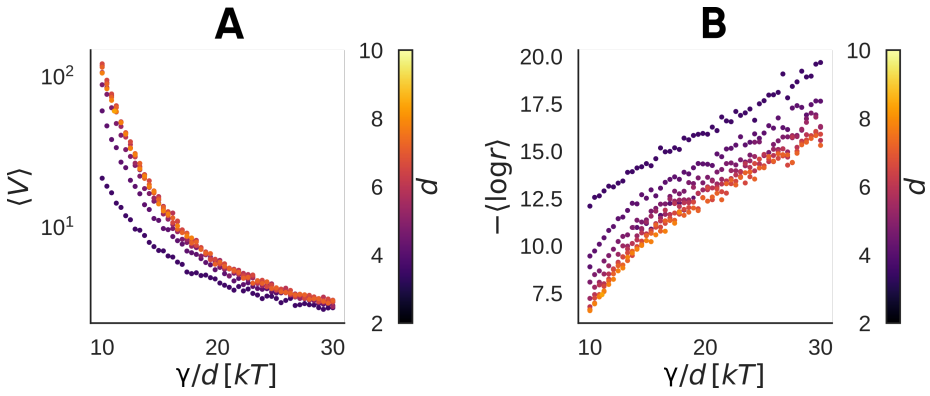


Figure 3.2: Dependence of accessible phase space volume and escape rates on ruggedness and dimensionality. (A) Accessible configuration space volume and (B) escape rates in dependence of ruggedness with color coded dimensionality dependence.

ruggedness restricts the accessible configuration space to a smaller volume. Overall, the accessible configuration space is smaller for lower dimensionality.

Note, that this effect does not arise directly from the dimensionality of configuration space as volumes increase with dimensionality. Rather, the topology of accessible configuration space resulting from the free-energy landscape depends on its dimensionality. The difference in accessible volume between configuration spaces with different dimensionalities decreases with increasing dimensionality. Only configuration spaces with dimensions smaller than four significantly deviate from the higher dimensional cases.

In the second panel B of Fig. 3.2 the obtained escape rates are shown in dependence of ruggedness. Escape rates decrease as well with increasing ruggedness. Also, in this case we observe a trend with respect to increasing dimensionality as escape rates are overall lower for lower dimensional models. This effect is explained by increasing amount of states at the boundary as the number of neighbouring states $N = 2d$ increases with increasing dimensionality. For higher dimensions this effect also seems to become independent of dimensionality, but much weaker as for the accessible configuration space volume.

Taken together, these results show that both accessible configuration space volume, as well as escape rates, are dependent on the ruggedness in the hierarchical lattice model.

To find a suitable description of this combined influence of ruggedness on accessible configuration space volumes and escape rate, we asked next how accessible configuration space volumes and escape rates are related. Fig. 3.3 shows in A the accessible configuration space volume $V$ as a function of the corresponding escape rate $r$ with color coded dimensionality and in B with color coded ruggedness. Unexpectedly, we find a scaling relation between $V$ and $r$ shown as a line in the log-log plot. The inset in Fig.3.3 A shows a linear fit to the double logarithmic data which yields the scaling relation

$$\langle V \rangle \propto - \langle \log r \rangle^{-4.24 \pm 0.24} . \tag{3.9}$$

For low $V$, deviations from this scaling behavior are observed, because here the smallest possible accessible configuration space volume of two states is approached (see dashed line in Fig.3.3). Also here, we observe the dependence of $V$

and $r$ on dimensionality explained above. While the scaling exponent is slightly lower for lower dimensionality, the overall scaling behavior is very similar.

Taken together, we conclude that both the structure of accessible configuration space and escape rates are responsible for the anomalous diffusion behavior observed in diffusion processes in the hierarchical lattice model. While the present analysis does not show the fractal nature of accessible configuration space the obtained scaling relation gives evidence for it. Such fractal structures might occur as low barriers form a network of 'bonds' between accessible states where the normalized ruggedness determines the probability of forming such a bond and thereby the volume of accessible configuration space. The normalized ruggedness also determines the statistics of the lowest barrier height over which the accessible configuration space is escaped. This lowest barrier governs the escape rate out regions of accessible configuration space. As this escape rate decreases with the accessible configuration space volume in intermediate dimensional hierarchical free-energy landscapes, it is also a source of anomalous diffusion, similar to the simple the 1-dimensional case [9].

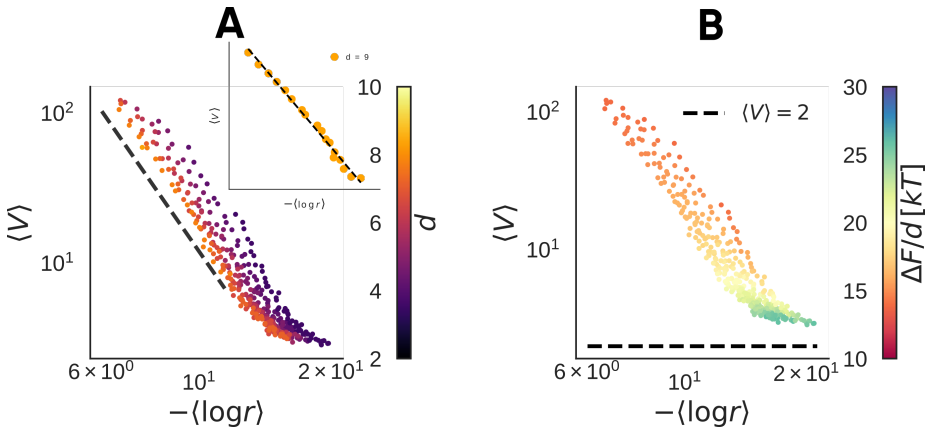


Figure 3.3: Dependence of accessible configuration space volume on escape rates with color coded dimensionality (A) and ruggedness (B). Due to the discrete grid the lowest possible phase space volume is 2 states which is indicated by the dashed line.

## 3.5 Conclusion

We have shown that random walks in simple hierarchical lattice models, used to model configuration space exploration of proteins, are goverened by a confined accessible configuration space volume and corresponding escape rates. Both quantities have been identified as sources of anomalous diffusion [24] [14], and we asked which of the two causes subdiffusion in hierarchical free-energy landscapes. In an earlier work we identified the normalized ruggedness of hierarchical free-energy landscapes as the parameter which directly drives anomalous diffusion behavior. If either of the two possible sources is affected by changes in ruggedness it most likely contributes to the anomalous diffusion behavior.

We found that the normalized ruggedness both affects accessible configuration space volume and escape rates similarly. Based on this finding, we conclude that a combination of the two parameters is the source of anomalous diffusion behavior in hierarchical free-energy landscapes. This combination is best described by a scaling relation between accessible configuration space and corresponding escape rates. This scaling relation provides evidence that accessible configuration space of diffusion processes in hierarchical free energy landscapes possesses a fractal structure formed out of a network of low barriers. However, this fractal structure is most likely not the singular source of anomalous diffusion in hierarchical free-energy landscapes and escape rates do contribute to the anomalous diffsion behavior.

The scaling relation we found also makes an exciting prediction for the internal dynamics of proteins as there should be a similar relation if protein free-energy landscapes indeed exhibit a hierarchical structure. If this prediction is correct it would be shown that anomalous diffusion in protein dynamics arises from a hierarchical structure of protein free energy landscapes. Moreover, assuming such a relationship holds, one could substantially improve enhanced sampling methods in molecular dynamics simulations, as the rate of configuration space exploration could be estimated based on the volume of accessible

configuration space. In that way, the longstanding sampling problem of molecular dynamics simulations could be addressed more systematically.

# Chapter 4

## An efficient sampling algorithm to generate trajectories in hierarchical free-energy landscapes

## 4.1 Introduction

Hierarchical free-energy landscapes occur in many different complex systems, including protein dynamics, semiconductors, and glasses. Often diffusion processes in these landscapes are particularly interesting as they govern e.g., the configuration space exploration proteins [13] or the conductance in semiconductors [41]. Whereas for lower dimensional hierarchical model free-energy landscapes, analytical solutions for diffusion in a hierarchical free-energy landscape are available, higher dimensional hierarchical models often can only be investigated via simulations of trajectories.

However, sampling trajectories in higher dimensional hierarchical free energy landscapes with high ruggedness values is a computationally challenging task in itself. Due to high free energy barriers ranging over many time scales, trajectories are often trapped in a small subset of states. This circumstance makes it computationally very costly to obtain the relevant time or ensemble-averaged quantities. Yet, because free energy landscapes are known in such models, it should be possible to find a suitable enhanced sampling method that exploits this information.

Here, we present an enhanced sampling method that facilitates orders of magnitude faster sampling in situations where trajectories are trapped in hierarchical free-energy landscapes. The idea of this method is based on Gillespie's algorithm, an established method to speed up sampling in random walk simulations. In this algorithm, instead of sampling repeated unsuccessful attempts to leave a state, the number of these until a transition to a new state occurs is estimated based on the known transition rates, and time is advanced accordingly. In that way, only successful transitions are simulated in each step, and it has been shown that this procedure still yields exact statistics of occupation times and transitions.

Applying the same idea for the sampling in traps of hierarchical free energy landscapes, instead of sampling transitions within a trap, we estimate the num-

ber of transitions before a trap is escaped, based on the known transition rates. We show that this method yields accurate results if the intra-trap dynamics is Markovian, i.e., a local equilibrium is reached before leaving the trap. However, these are precisely the cases where conventional Gillespie sampling runs into problems.

Based on this method, we propose an algorithm that drastically increases sampling speed and show that it yields approximately correct escape times and probabilities in the case of intermediate dimensional hierarchical lattice models.

## 4.2 Theory

Hierarchical free-energy landscape models have been used to describe complex systems, including protein dynamics, semiconductors, and glasses. Whereas for low dimensional models, analytical solutions exist, for high dimensional only numerical approaches are available. These numerical approaches can be very computationally costly, especially for higher-dimensional models with high free energy barriers, such as those used to model the free energy landscapes of the internal dynamics of proteins. In fact, for high-dimensional hierarchical free-energy landscape models, a very similar sampling problem appears similar to the well-known sampling problem of molecular dynamics simulations of proteins.

This sampling problem is characterized by trajectories sampling within free-energy traps enclosed by high free-energy barriers instead of exploring the entire state space. In molecular dynamics simulations of proteins, this problem was addressed by various enhanced sampling techniques. This work presents a similar approach for improving sampling in trajectories of intermediate-dimensional hierarchical lattice models. Whereas this method is in principle applicable to improving sampling of any discrete model described by a master equation, we show its applicability in the case of an intermediate-dimensional lattice model that we used to model anomalous diffusion observed in the internal dynamics of proteins. This model consists of a d-dimensional cubic lattice of states separated by exponentially distributed barriers as described in chapter 1. We showed in

chapter 2 that the normalized ruggedness parameter $\gamma/d$ determines anomalous diffusion in this model. We found that generating sampling that explores most dimensions is for high normalized ruggedness ($> 20\,\mathrm{kT}$) is computationally very costly, as random walk trajectories get trapped in confined regions of the state space.

To mitigate this sampling problem, we employed an enhanced sampling algorithm that is inspired by Gillespie's algorithm [27]. In Gillespie's algorithm, only those transitions between different states are sampled, and the number of steps the system stays within the same state is determined from a geometric distribution. Due to the Markov assumption for the dynamics within a state, this approximation is exact. We utilized the same idea for circumventing sampling within confined regions: Instead of sampling transitions within a confined region, we estimate the probability of the number of transitions $n$ until an escape event occurs and samples from a to be determined distribution as numbers of transitions. This approach is not exact because dynamics within a confinement region is not memoryless but yields a better approximation for regions with longer escape times.

### 4.2.1 Theoretical background of the method

The probability of escaping $p_{esc}$ a confinement region is given by the probability of occupying a boundary state $\mu_i$ and the probability $p_{ij}$ of transitioning from state $i$ at the boundary to a state $j$ which is outside the boundary of the confinement region as

$$p_{esc} = \sum_{i \in \Omega} \mu_i \sum_{j \notin V} p_{ij}, \tag{4.1}$$

where $\Omega$ is the set of states at the boundary of a confinement region and $V$ is the set of all states within the confinement region. If the confinement region is stable enough that a there are a lot of transitions between states, the number of transitions $n$ before leaving is memoryless,

$$Pr(n > s + t | n > s) = Pr(n > t).$$

Therefore, escape attempts are statistically independent events and the number of transitions $T$ within a confinement region until an escape event occurs are exponentially distributed as

$$p(n) \propto (1 - p_{esc})^{n-1} p_{esc} \approx e^{\log(1-p_{esc})\, n}. \tag{4.2}$$

Based on the assumption that a sufficient amount of transitions within a confinement region occurs before leaving, we estimate $\mu_i$ as the fraction of visits to state $i$ and the total amount of transitions. The number of visits to each state is sufficient to calculate time averages, as these do not depend on the particular order each state is visited. The next state $j$ outside of a confinement region is determined by the probability

$$p_j = \frac{\sum_{k\in\Omega} p_{kj}}{\sum_{l\in\bar{\Omega}} \sum_{k\in\Omega} p_{kl}} \tag{4.3}$$

where $\bar{\Omega}$ is the set of states outside of the confinement region with a connection to its boundary.

### 4.2.2 The implemented Algorithm

Taken together, equations 4.1 and 4.3 suggest the following enhanced sampling algorithm. Sampling is performed using the simple Gillespie algorithm until a trajectory is trapped. We chose a heuristic approach to determine when trapping events happen by counting the revisitings of states. If this number exceeds 100 revisitings, we assume that the dynamics within the visited states is memoryless and bootstrap the number of steps within this trap from an exponential distribution according to eq. 4.2. We determine the weight $w_i$ by which each individual state $i$ contributes to the desired time-averaged quantity is determined by $w_i = n \langle \Delta t \rangle$ where $\langle \Delta t \rangle$ is the average time spent within a state We approximate $\langle \Delta t \rangle$ with the sample-mean of the observed waiting times of the initial Gillespie sampling. The first state visited outside a trap is chosen randomly from a probability distribution according to eq. 4.3. After this state is determined, desired time averages are updated and sampling is continued with Gillespie's algorithm until the next trapping event.

## 4.3 Methods

To estimate the error introduced by this enhanced sampling algorithm, we generated 100random walks in confinement regions with normalized ruggedness values ranging from 21 kT to 29 kT and dimensionality values of 3 to 10 dimensions with brute-force Gillespie sampling until an escape event occurs. In this range of ruggedness parameters typically, sampling problems occurred, in particular, in lower-dimensional models.

The most relevant quantity for random walk dynamics that cannot be estimated from the intra-trap dynamics alone is the number of steps until an escape event occurs. We, therefore, compared escape times from brute-force Gillespie sampling with adaptive free-energy landscape generation as described in 2.3.2 with randomly sampled escape times from eq. 4.2.

To verify the improvement in sampling speed, we generated random walk trajectories between $10^7$ and $10^9$ steps with the same set of parameters with our enhanced sampling scheme and compared the resulting compute (wall-) times.

All random walk calculations were performed on a Intel Xeon CPU W3550.

## 4.4 Results and Discussion

### 4.4.1 Accuracy of the enhanced sampling algorithm

The distributions of calculated and estimated escape times for different ruggedness values are shown in Fig. 4.1 (see Fig. 4.4). All generated trajectories with the same normalized ruggedness $\gamma$ contribute to each kernel density plot, in particular also those, with different dimensionality values. As can be seen, both distributions of escape times by Gillespie sampling (solid line) and random sampling (dashed line) cover the same range and are rather independent of ruggedness values. However, fluctuations in observed escape times are rather high. The mean relative error of escape times is approximately 0.4% while the

average variance of the relative error is approximately 20%. This error is not systematic as simulated and randomly sampled escape times are not correlated (see Fig. 4.5). This behavior is to be expected as models with different dimensionalities were compared among each other. We have shown in chapter 1 that anomalous diffusion behavior of models with equal normalized ruggedness but with different dimensionalities show similar anomalous diffusion behavior on average but with large fluctuations.

### 4.4.2 Inprovement in performance

Next, we asked how much our enhanced sampling algorithm improves sampling speed. Figure 4.2 shows the time it takes (wall-time) $T_{\text{wall}}$ to compute a random walk trajectory of length $T$. Overall, our implementation of the enhanced sampling algorithm (green dots) is, on average, at least an order of magnitude faster than our implementation of the standard Gillespie's algorithm (blue dots). Whereas Gillespie's algorithm, as expected, shows on average a linear increase in computing time for increasing trajectory lengths, our enhanced sampling algorithm shows sub-linear scaling with an approximate exponent of $0.78 \pm 0.07$.

Longer trajectory lengths lead to a drastically larger spread of computing times for the enhanced sampling method. For $10^7$ time steps, the spread of computing time is around one order of magnitude, it increases up to 4 orders of magnitude for $10^9$ steps. This tendency is more pronounced for the enhanced sampling method compared to Gillespie's algorithm. It is to be expected that sampling of trajectories that are not often trapped does not benefit from the enhanced sampling method. Therefore, Gillespie's algorithm and the enhanced sampling algorithm should show the same performance in these cases. This effect causes the increasing spread in computing times of the enhanced sampling method because longer trajectories are more likely to escape their 'initial' trap and explore state space more freely. For these trajectory lengths, a mixture of the performance of the Gillespie algorithm and the enhanced sampling algorithm is observed.

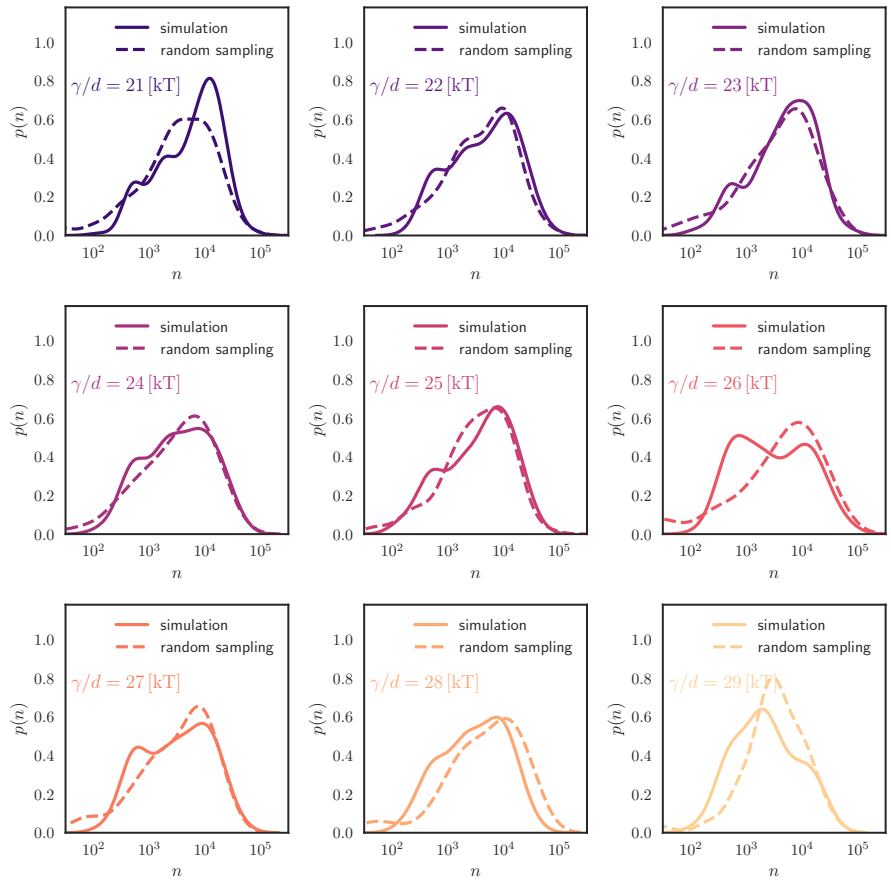


Figure 4.1: Distribution of escape times from a confinement region in random walk simulations and bootstrapped amount of steps based on the escape rate determined in eq. 4.1. The color represents different normalized ruggedness values. For a non-logarithmic plot see Fig. 4.4.

Also, Fig. 4.3 indicates that it is mainly the lower normalized ruggedness values that show the low performance. The former hypothesis also explains this effect, as lower normalized ruggedness leads to shorter escape times from traps, as shown in chapter 3.

In summary, we have shown that the computing performance of sampling random walks generally benefits from our enhanced sampling method. However,

it is only effective in high-ruggedness regimes, where trajectories are trapped for long times.

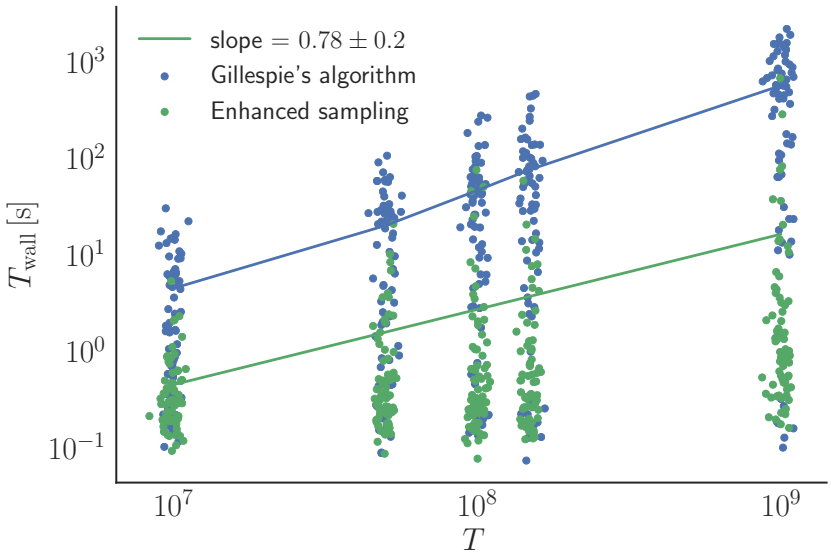


Figure 4.2: Performance of our enhanced sampling algorithm compared to Gillespie's algorithm: The figure shows how much computing time (y-axis) is needed on a single thread on a Intel Xeon CPU W3550 to generate a random walk trajectory of a certain length(x-axis). Shown as blue dots are simulations with the standard Gillespie's algorithm and shown as green dots our enhanced Gillespie's algorithm. A 0.1% Gaussian noise was added to the observed values for visualization purposes.

### 4.4.3 Conclusion

This work presented an enhanced sampling method that improves the sampling of random walks in a hierarchical free energy landscape. The principle behind this method is not restricted to random walks in hierarchical free-energy landscapes. It can be applied, quite generally, to random walks in models where a

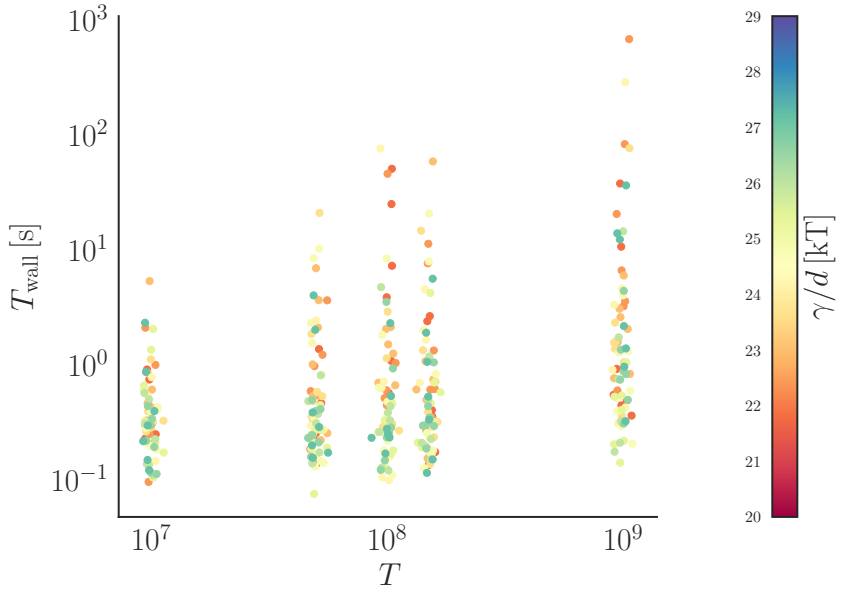


Figure 4.3: Ruggedness dependence of the performance of our enhanced sampling algorithm. The colored dots show how much computing time (y-axis) is needed for generating a random walk trajectory (x-axis). The collor shows the corresponing normalized ruggedness value.

master equation governs transitions between states. However, our results show that only models where trajectories are trapped in a confined region of state space would benefit from this method.

Strictly, it is also only exact in such models because it assumes Markovianity for the intra-trap dynamics. Here, we showed that this assumption holds in the case of $d$-dimensional hierarchical lattice models with high ruggedness to a good approximation because the method reproduced escape times of brute-force Gillespie simulations. However, we think that the accuracy of this method could be improved by explicitly taking into account the intra-trap dynamics instead of assuming Markovianity. The intra-trap dynamics is given by a master equation that can be numerically solved. In this way, the memory introduced by the system's initial conditions entering the trap at a specific state is explicitly

treated, such that the assumption of Markovianity can be dropped. This should also improve computing time in hierarchical free-energy landscapes with lower ruggedness.

An interesting outlook that the basic idea of this method offers is that it could yield an analytical solution, also for the intermediate dimensional case. The 1-dimensional case could be solved analytically using a renormalization group approach. In this approach, states that are available to a random walk on a given time scale are eliminated and replaced by a single 'macro' state. Afterwards, free-energy barriers between the remaining states are renormalized, and the reduced model is mapped onto the original lattice. Similarly, states within traps in intermediate dimensional hierarchical models could be eliminated and free-energy barriers connecting traps are renormalized accordingly. However, because the topology of the interconnections between trap 'macro' states is not a cubic grid as for the 1-dimensional case, it cannot be mapped to the original grid. This problem suggests using randomly distributed points as a state-space instead of a cubic grid as it would be possible to map the stat space back to the original in that case.

## 4.5 Supplement

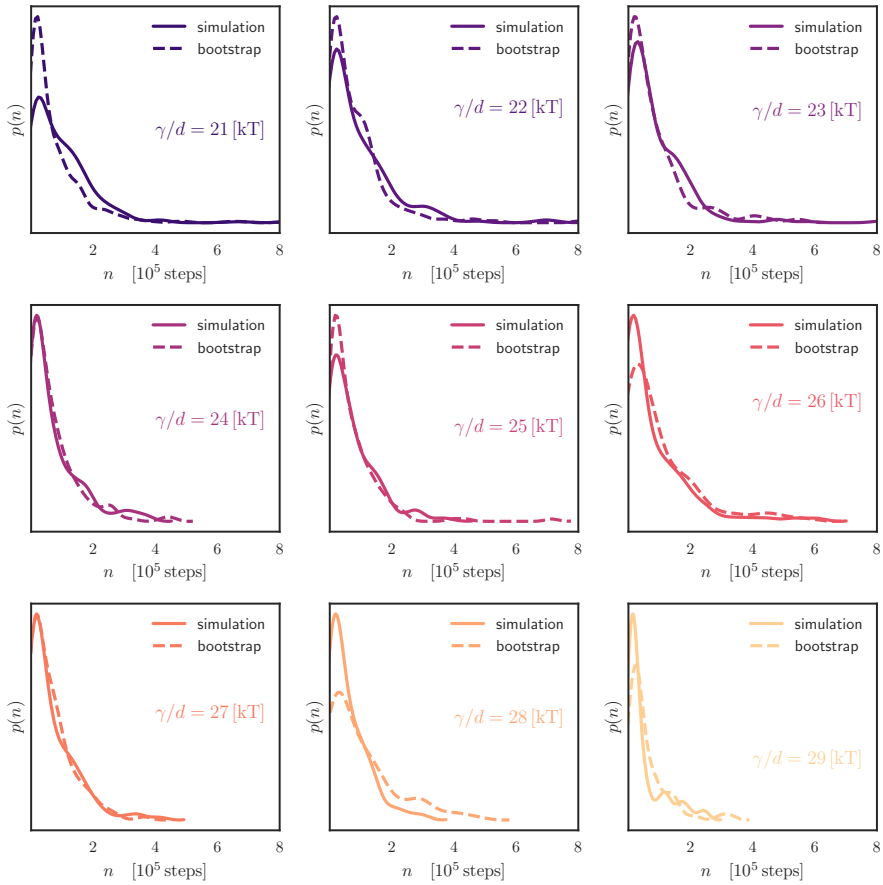


Figure 4.4: Steps needed to escape a confinement region in random walk simulations and bootstrapped amount of steps based on the escape rate determined in eq. 4.1 for random walks in different ruggedness values indicated by the color.

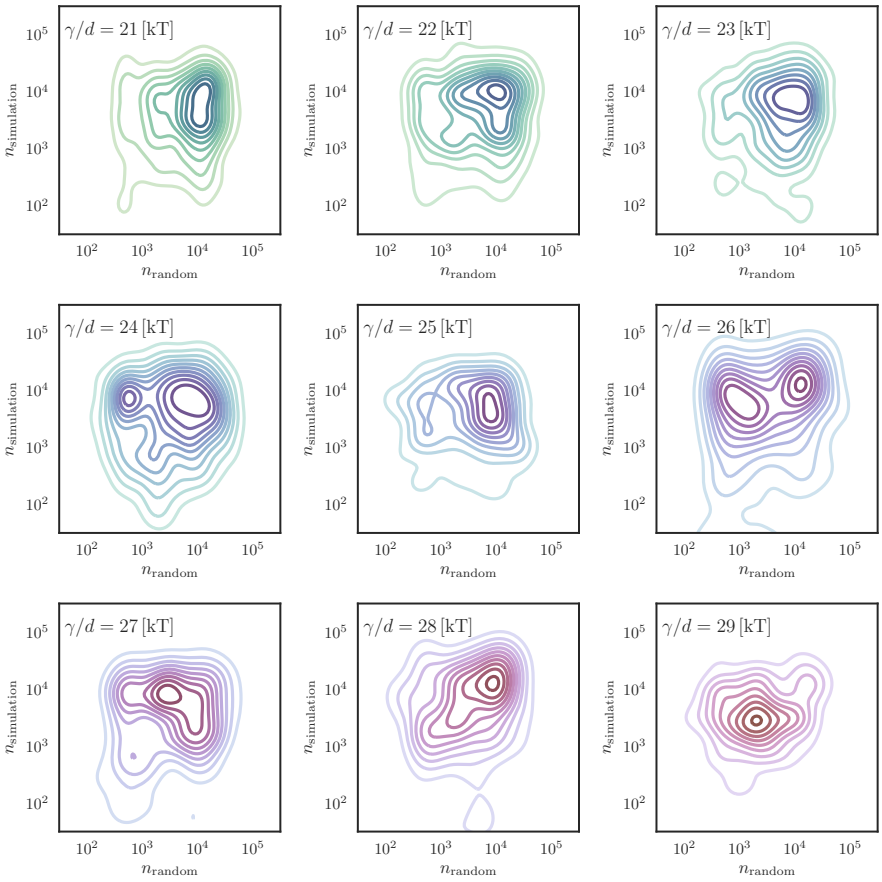


Figure 4.5: Steps needed to escape a confinement region in random walk simulations plottet against a bootstrapped amount of steps based on the escape rate determined in eq. 4.1 on a non-logarithmic scale for random walks in different ruggedness values indicated by the color.

# Chapter 5

## Conclusion

The central aim of this work was to investigate the hierarchical structure of the protein free-energy landscape. To that end, we generated molecular dynamics trajectories of 500 small globular proteins, investigated the anomalous diffusion behavior arising from the hierarchical structure of their free-energy landscapes, and obtained anomalous diffusion exponents. Using $d$-dimensional hierarchical lattice model, we estimated barrier-height distributions and effective dimensionalities of free-energy landscapes of each protein from the obtained anomalous diffusion exponents. As a result, we gained insights into how a hierarchical structured free-energy landscape governs dynamics in general and protein dynamics in particular.

We found that the internal dynamics of proteins, in general, shows anomalous diffusion. This anomalous diffusion was previously observed for small peptides and small globular proteins and was attributed to different sources. In particular, it has been discussed that it arises from a projection effect and does not reflect the structure of protein free-energy landscapes but rather the choice of the (linear) collective coordinates. For high-dimensional free diffusion, it has been shown that using the analysis method used to investigate anomalous diffusion in proteins (trajectory-length dependent principal component analysis) does not yield anomalous diffusion as a projection effect. We have also shown that anomalous diffusion behavior arises from the ruggedness of the free-energy landscapes in the case of high-dimensional hierarchical models. An explanation for the absence of projection effects is that in trajectory-length dependent principal component analysis, the trajectory is projected on different collective coordinates on different time scales. In the cases where anomalous diffusion was shown to be a projection effect, a fixed set of coordinates was used. However, we cannot rule out that the observed anomalous diffusion behavior is a projection effect for intermediate dimensional hierarchical models and proteins. This aspect is a matter of current debate and should be further explored in future work. In particular, it is assumed that anomalous diffusion arises from memory introduced by the use of linear collective coordinates, which suggests the use of non-linear collective coordinates such as Markov models. Assuming that these

non-linear collective coordinates are a good representation of the eigenfunctions of the propagator of protein dynamics, projections are free of memory. If such projections still show similar anomalous diffusion, it could be concluded that the observed anomalous diffusion is not a projection effect. Attempts in that direction have already been made in the case of peptides [42].

We think it is rather likely that the more significant part of the observed anomalous diffusion in protein dynamics does not arise from a projection effect. Based on that, essentially, two explanations of the anomalous diffusion behavior remain: a fractal structure of the accessible configuration space of proteins and a hierarchical structure of the free-energy landscape of proteins. In an attempt to decide which of the two governs anomalous diffusion in protein dynamics we learned that both arise from a common cause and, in this sense, are "two sides of the same coin". This finding rests on the evidence that the topology of accessible configuration space of trajectories in higher dimensional hierarchical free-energy landscapes is the cause for the anomalous diffusion behavior. In particular, we found that both accessible configuration space volume and escape rates are responsible for the anomalous diffusion behavior are responsible for anomalous diffusion. Unexpectedly, we found a scaling relation between these two quantities that describe their combined influence on the anomalous diffusion behavior. If protein free-energy landscapes indeed exhibit a hierarchical structure, we would expect a similar relation. This prediction could be tested in future work by analyzing the accessible configuration space volumes and escape rates in Markov state models generated from MD trajectories.

Further, we have shown that the ruggedness and dimensionality of hierarchical free energy landscapes both for proteins and hierarchical lattice models can be estimated from anomalous diffusion exponents with a reasonably small error of $\sim 5\,\mathrm{kT}$for ruggedness estimates and 10 dimensions for dimensionality estimates. Neither in the literature analytical solutions for the intermediate dimensional hierarchical model can be found, nor we have been able to derive it such that we had to resort to a numerical approach. It was possible to generate enough sampling in the relevant ruggedness and dimensionality regimes,

where trajectories typically are trapped in a small subset of states because we developed a novel enhanced sampling technique. The idea of this enhanced sampling technique is that, instead of sampling within a region where trajectories are trapped, to estimate the number of steps until the system escapes from the trap. It should be possible to expand this idea towards a renormalization group approach that could yield an analytical solution in future work. A similar renormalization group approach was employed to obtain the analytical solution of the 1-dimensional case [8] [9].

Finally, we estimated the ruggedness and dimensionality of protein free-energy landscapes based on our numerical results. We obtained typical normalized ruggedness estimates of $15 - 20\,\mathrm{kT}$ per dimension and estimates for the effective dimensionality of $40 - 60$ which are reproducible in independent MD trajectories with a standard deviation of $1.1\,\mathrm{kT}$ and 4.8 dimensions respectively. Remarkably, neither the effective dimensionality nor the ruggedness of proteins correlates with protein size, although there is a significant correlation between the two. Also, the ranges of both normalized ruggedness and effective dimensionality are much smaller than the range of protein sizes we considered. From this we conclude that evolution adapts both effective dimensionality and ruggedness of protein free energy landscapes to its respective function. This also explains why ruggedness serves as a good predictor for protein function as found in an earlier work [13]. It opens up many new questions concerning the influence of evolution on these two features of the protein free-energy landscape that could be addressed in future work.

Taken together, we conclude that the ruggedness and effective dimensionality of protein free-energy landscapes play an important role for protein function and are not mere byproducts of the complexity in protein dynamics. This opens up a new perspective on protein dynamics, where it is usually assumed that only motions on a specific time-scale, typically slow motions, contribute to the specific function of a protein. Our work suggests that protein function is rather governed by the combination of motions on different time and length scales.

# Bibliography

[1] Charles L Brooks, Martin Karplus, and B Montgomery Pettitt. *Advances in chemical physics, volume 71: Proteins: A theoretical perspective of dynamics, structure, and thermodynamics*. Wiley-Blackwell, 2006.

[2] JA McCammon. Protein dynamics. *Reports on Progress in Physics*, 47(1):1, 1984.

[3] Guillermo Pérez-Hernández and Frank Noé. Hierarchical time-lagged independent component analysis: computing slow modes and reaction coordinates for large molecular systems. *Journal of chemical theory and computation*, 12(12):6118–6129, 2016.

[4] A Ansari, J Berendzen, S F Bowne, H Frauenfelder, I E Iben, T B Sauke, E Shyamsunder, and R D Young. Protein states and proteinquakes. *Proceedings of the National Academy of Sciences*, 82(15):5000–5004, 1985.

[5] Xiaohu Hu, Liang Hong, Micholas Dean Smith, Thomas Neusius, Xiaolin Cheng, and Jeremy C Smith. The dynamics of single protein molecules is non-equilibrium and self-similar over thirteen decades in time. *Nature Physics*, 12(2):171–174, 2016.

[6] Ralf Metzler. Forever ageing. *Nature Physics*, 12(2):113–114, 2016.

[7] Fritz Parak and Hans Frauenfelder. Protein dynamics. *Physica A: Statistical Mechanics and its Applications*, 201(1-3):332–345, 1993.

[8] A. Maritan and A. L. Stella. Exact renormalization group for dynamical phase transitions in hierarchical structures. *Phys. Rev. Lett.*, 56:1754–1754, Apr 1986.

[9] S. Teitel and Eytan Domany. Dynamical phase transitions in hierarchical structures. *Phys. Rev. Lett.*, 55:2176–2179, Nov 1985.

[10] Shlomo Havlin, James E. Kiefer, and George H. Weiss. Anomalous diffusion on a random comblike structure. *Phys. Rev. A*, 36:1403–1408, Aug 1987.

[11] Tomas Hansson, Chris Oostenbrink, and WilfredF van Gunsteren. Molecular dynamics simulations. *Current Opinion in Structural Biology*, 12(2):190–196, 2002.

[12] GR Kneller and Konrad Hinsen. Fractional brownian dynamics in proteins. *The Journal of chemical physics*, 121(20):10278–10283, 2004.

[13] Ulf Hensen, Tim Meyer, Jürgen Haas, René Rex, Gert Vriend, and Helmut Grubmüller. Exploring protein dynamics space: The dynasome as the missing link between protein structure and function. *PLOS ONE*, 7(5):1–16, 05 2012.

[14] Thomas Neusius, Isabella Daidone, Igor M Sokolov, and Jeremy C Smith. Subdiffusion in peptides originates from the fractal-like structure of configuration space. *Physical review letters*, 100(18):188103, 2008.

[15] Alessio Lapolla and Aljaž Godec. Manifestations of projection-induced memory: General theory and the tilted single file. *Frontiers in Physics*, 7:182, 2019.

[16] José Nelson Onuchic, Zaida Luthey-Schulten, and Peter G Wolynes. Theory of protein folding: the energy landscape perspective. *Annual review of physical chemistry*, 48(1):545–600, 1997.

[17] H Frauenfelder, F Parak, and R D Young. Conformational substates in proteins. *Annual Review of Biophysics and Biophysical Chemistry*, 17(1):451–479, 1988. PMID: 3293595.

[18] David E Shaw, Ron O Dror, John K Salmon, JP Grossman, Kenneth M Mackenzie, Joseph A Bank, Cliff Young, Martin M Deneroff, Brannon Batson, Kevin J Bowers, et al. Millisecond-scale molecular dynamics simulations on anton. In *Proceedings of the conference on high performance computing networking, storage and analysis*, pages 1–11, 2009.

[19] Carsten Kutzner, Szilárd Páll, Martin Fechner, Ansgar Esztermann, Bert L de Groot, and Helmut Grubmüller. More bang for your buck: Improved

use of gpu nodes for gromacs 2018. *Journal of computational chemistry*, 40(27):2418–2431, 2019.

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

[21] Kalyan Kundu and Philip Phillips. Hopping transport on site-disordered d-dimensional lattices. *Physical Review A*, 35(2):857, 1987.

[22] Charles C David and Donald J Jacobs. Principal component analysis: a method for determining the essential dynamics of proteins. In *Protein dynamics*, pages 193–226. Springer, 2014.

[23] G. Ulrich Nienhaus, Joachim D. Müller, Ben H. McMahon, and Hans Frauenfelder. Exploring the conformational energy landscape of proteins. *Physica D: Nonlinear Phenomena*, 107(2):297 – 311, 1997. 16th Annual International Conference of the Center for Nonlinear Studies.

[24] Elliott W. Montroll and George H. Weiss. Random walks on lattices. ii. *Journal of Mathematical Physics*, 6(2):167–181, 1965.

[25] Isabella Daidone and Andrea Amadei. Essential dynamics: foundation and applications. *WIREs Computational Molecular Science*, 2(5):762–770, 2012.

[26] Berk Hess. Similarities between principal components of protein dynamics and random diffusion. *Phys. Rev. E*, 62:8438–8448, Dec 2000.

[27] Daniel T. Gillespie. Exact stochastic simulation of coupled chemical reactions. *The Journal of Physical Chemistry*, 81(25):2340–2361, 1977.

[28] Stuart Geman and Donald Geman. Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-6(6):721–741, 1984.

[29] David W Scott. *Multivariate density estimation: theory, practice, and visualization*. John Wiley & Sons, 2015.

[30] Tim Meyer, Marco D'Abramo, Adam Hospital, Manuel Rueda, Carles Ferrer-Costa, Alberto Pérez, Oliver Carrillo, Jordi Camps, Carles Fenollosa, Dmitry Repchevsky, Josep Lluis Gelpí, and Modesto Orozco. Model (molecular dynamics extended library): A database of atomistic molecular dynamics trajectories. *Structure*, 18(11):1399–1409, 2010.

[31] Helen M. Berman, John Westbrook, Zukang Feng, Gary Gilliland, T. N. Bhat, Helge Weissig, Ilya N. Shindyalov, and Philip E. Bourne. The Protein Data Bank. *Nucleic Acids Research*, 28(1):235–242, 01 2000.

[32] Mark James Abraham, Teemu Murtola, Roland Schulz, Szilárd Páll, Jeremy C. Smith, Berk Hess, and Erik Lindahl. Gromacs: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX*, 1-2:19–25, 2015.

[33] Hans W. Horn, William C. Swope, Jed W. Pitera, Jeffry D. Madura, Thomas J. Dick, Greg L. Hura, and Teresa Head-Gordon. Development of an improved four-site water model for biomolecular simulations: TIP4P-Ew. *The Journal of chemical physics*, 120(20):9665–9678, 2004.

[34] 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*, 78(8):1950–1958, 2010.

[35] HJC Berendsen and WF Van Gunsteren. Molecular dynamics simulations: Techniques and approaches. In *Molecular Liquids*, pages 475–500. Springer, 1984.

[36] Shuichi Miyamoto and Peter A. Kollman. Settle: An analytical version of the shake and rattle algorithm for rigid water models. *Journal of Computational Chemistry*, 13(8):952–962, 1992.

[37] Berk Hess, Henk Bekker, Herman J. C. Berendsen, and Johannes G. E. M. Fraaije. LINCS: A linear constraint solver for molecular simulations. *Journal of Computational Chemistry*, 18(12):1463–1472, 1997.

[38] T. E. III Cheatham, J. L. Miller, T. Fox, T. A. Darden, and P. A. Kollman. Molecular Dynamics Simulations on Solvated Biomolecular Systems: The Particle Mesh Ewald Method Leads to Stable Trajectories of DNA, RNA, and Proteins. *Journal of the American Chemical Society*, 117(14):4193–4194, 1995.

[39] Mark M. Meerschaert and Hans-Peter Scheffler. Limit theorems for continuous-time random walks with infinite mean waiting times. *Journal of Applied Probability*, 41(3):623–638, 2004.

[40] H Frauenfelder, SG Sligar, and PG Wolynes. The energy landscapes and motions of proteins. *Science*, 254(5038):1598–1603, 1991.

[41] Thomas B. Schrøder and Jeppe C. Dyre. Computer simulations of the random barrier model. *Phys. Chem. Chem. Phys.*, 4:3173–3178, 2002.

[42] Thomas Neusius, Isabella Daidone, Igor M Sokolov, and Jeremy C Smith. Configurational subdiffusion of peptides: A network study. *Physical Review E*, 83(2):021902, 2011.

# Danksagung

Diese Arbeit konnte nur mit der Unterstützung vieler Menschen verfasst werden. Mein Dank gilt im Besonderen meinem Doktorvater Prof. Dr. Helmut Grubmüller, unter dessen Betreuung diese Arbeit entstanden ist. Ihre Fertigstellung konnte nur aufgrund seiner Expertise und dem fortwährenden kreativen Austausch mit ihm gelingen. Sowohl seine kritischen und gerade deshalb hilfreichen Anmerkungen als auch seine persönliche Begleitung haben maßgeblich zum Gelingen dieser Arbeit beigetragen. Der große Freiraum, den er mir in der Ausarbeitung des Themas eingeräumt hat, und die Möglichkeit, meine Ideen in den Arbeitsgruppen vorzustellen, haben für ein angenehmes Arbeitsumfeld gesorgt. Ich möchte mich außerdem bei Prof. Dr. Jörg Enderlein für seine wertvollen und hilfreichen Anregungen bedanken, der mit seinem Fachwissen die Ausarbeitung dieser Arbeit bereichert hat. Den Kollegen in der Abteilung für Theoretische und Computergestützte Biophysik am Max-Planck-Institut für Biophysikalische Chemie bin ich ebenfalls für die fruchtbare Zusammenarbeit und angenehme Arbeitsatmosphäre zu Dank verpflichtet. Meinen Eltern danke ich für den Rückhalt, den ich in den letzten Jahren von ihnen erfahren durfte. Besonderen Dank möchte ich an meine Mutter richten, die diese Arbeit über all die Jahre wohlwollend begleitet hat. Der wichtigste Dank geht an meine Verlobte Eytan Celik für ihre unermüdliche Unterstützung.

**Lebenslauf**

Andreas Volkhardt

| | |
|---|---|
| Name: | Andreas Volkhardt |
| Geburtsdatum: | 13.06.1988 |
| Geburtsort: | Eisenach |
| Staatsangehörigkeit: | Deutsch |
| Aktuelle Anschrift: | Bertheaustr. 31A, 37075 Göttingen |
| Familienstand: | ledig |
| Eltern: | Udo Volkhardt<br>Cordula Volkhardt, geb. Braun |
| 1998 – 2006 | Herzog-Georg-Gymnasium, Bad Liebenstein |
| 2006 – 2007 | Zivildienst |
| 2007 – 2013 | Studium der Physik an der Georg-August-Universität<br>Bachelor of Science 2011<br>Master of Science 2013 |
| seit 2015 | Promotion am Max-Planck-Institut für biophysikalische Chemie in Göttingen. Titel der Dissertation: »Anomalous Diffusion in Protein Dynamics« |