Markov models and dynamical fingerprints: unraveling the complexity of molecular kinetics

Bettina G. Keller [a], Jan-Hendrik Prinz [a] and Frank Noé [a]

May 26, 2011

[a] Freie Universität Berlin, Arnimallee 6, 14195 Berlin, Germany

Corresponding Author:

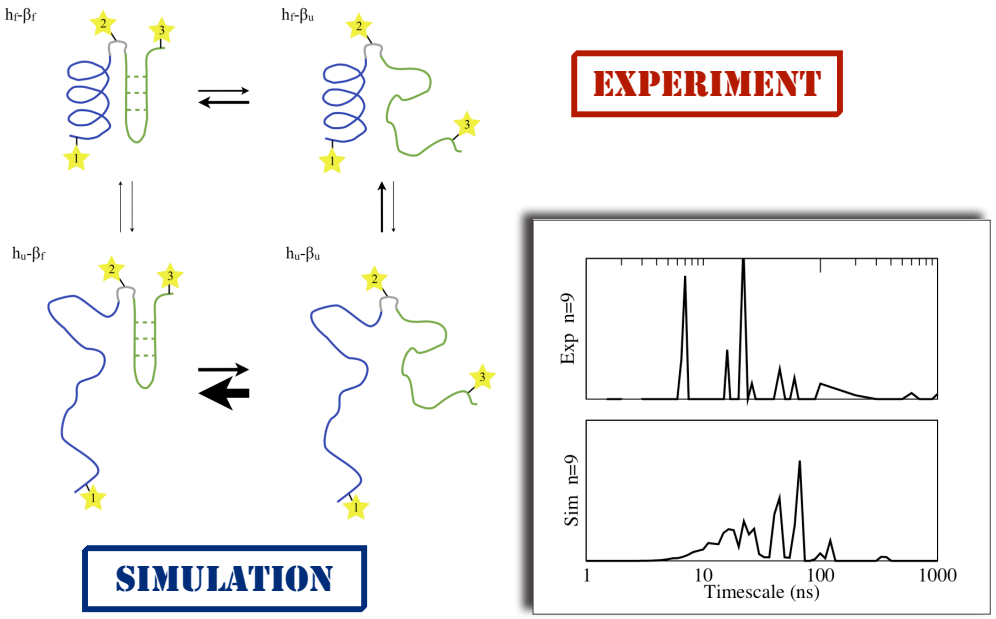
Bettina Keller	
Postal Address:	Arnimallee 6, 14195 Berlin, Gemany
Email:	bettina.keller@fu-berlin.de
Phone:	+49 (0)30 838 75366
Fax:	+49 (0)30 838 75412

Keywords: single-molecule spectroscopy, dynamical fingerprints, protein folding, Markov models, molecular dynamics, FCS, T-jump, FRET

Abstract

Dynamical fingerprints of macromolecules obtained from experiments often seem to indicate two- or three state kinetics while simulations typically reveal a more complex picture. Markov state models of molecular conformational dynamics can be used to predict these dynamical fingerprints and to reconcile experiment with simulation. This is illustrated on two model systems: a one-dimensional energy surface and a four-state model of a protein folding equilibrium. We show that (*i*) there might be no process which corresponds to our notion of folding, (*ii*) often the experiment will be insensitive to some of the processes present in the system, (*iii*) with a suitable combination the observable and initial conditions in a relaxation experiment one can selectively measure specific processes. Furthermore, our method can be used to design experiments such that specific processes appear with large amplitudes. We demonstrate that for a fluorescence quenching experiment of the MR121-G9-W peptide.

Graphical abstract



Synopsis.

Dynamical fingerprints obtained from correlation experiments can be predicted quantitatively and interpreted using Markov state models of the conformational equilibrium. This approach is of use for choosing the chromophore attachment points and designing the optimal setup of an experiment.

Highlights

- Conformational dynamics of large biomolecules represented by Markov state models.
- Quantitative prediction of the dynamical fingerprints of correlation experiments.
- The design of the experiment determines which processes can be measured.
- The slowest observed process does not always correspond to the overall folding rate.
- Optimal experimental setup and observable(s) using molecular simulation and Markov models.

1 Introduction

Complex molecular systems often possess multiple stable or metastable states which are typically associated with specific functional properties. A hallmark of this are the numerous X-ray crystallography and NMR structures in which a given macromolecule has been found to exist in multiple conformations. Famous examples are the muscle protein myosin which exists in open and closed states with different nucleotide configurations [1], DNA-enzyme complexes that have different conformations depending upon the DNA sequence [2], or the Ribosome the parts of which are found in different arrangements along the protein synthesis cycle [3]. While these systems are examples which natively have several metastable states, in natively ordered proteins, the native state is a metastable state itself. This native state is in equilibrium with unfolded states, as well as possibly aggregation-prone misfolded states that are observed in prion diseases such as Alzheimer [4]. Thus, metastability is hierarchical with metastable states containing metastable sub-states [5].

In the last years it has become increasingly clear how these metastable states are connected by dynamics. Especially single-molecule experiments such as fluorescence-based [6, 7, 8, 9] or force-probe [10, 11, 12, 13] measurements have explicitly shown that macromolecules reside in different metastable states and occasionally transit between them. Single molecule trajectories can be analyzed by advanced statistical techniques such as Hidden Markov Models or other likelihood-based methods [14, 15, 16, 17].

However, ensemble-averaged kinetic measurements remain an essential way to access molecular kinetics. Such measurements may be done by perturbation of an actual ensemble of molecules, which often can be done simpler and with a better signal- to noise ratio than manipulation of single molecules. Perturbation ensemble experiments may be done by monitoring the relaxation of an ensemble towards equilibrium after starting from a defined off-equilibrium distribution *via*, e.g., a jump in temperature [18, 19], pressure [20], a change in the chemical environment [21] or a photo flash [22, 23, 24, 25]. They may also consist of dynamical spectroscopic measurements such as X-ray or inelastic neutron scattering [26]. The resulting signal reports on the kinetic processes involved in relaxing the perturbed ensemble to the original ensemble. More precisely, the autocorrelation function of the signal can be transformed into a dynamical fingerprint of the molecule [27]. Each peak in the dynamical fingerprint corresponds to a kinetic process, and its position corresponds to the timescale of this process.

Alternatively, trajectories of single molecules or fluctuations of dilute samples may be used to accumulate correlation functions, which are then also ensemble-averaged quantities. This approach is often chosen with fluorescence methods, such as correlation spectroscopy of the fluorescence intensity [28, 29, 30, 31, 32] or fluorescence resonance energy transfer (FRET) efficiency [33, 34]. The analysis of these accumulated correlation function is analogous to the analysis of perturbation experiments, but it yields a dynamical fingerprint of the kinetic processes which are active in equilibrium.

The main limitation of kinetic experiments is that they usually probe only one or two structural coordinates simultaneously. Exceptions are NMR-based methods [24]. With these methods, however, only a low time resolution - in the order of seconds - can be achieved, and several laborious repetitions of the experiment are required to obtain a viable signal-to-noise ratio. Currently, the only technique that can access structure and dynamics simultaneously and at great detail is molecular dynamics (MD) simulations,

which are becoming increasingly accepted as a tool to investigate structural details of molecular processes and relate them to experimentally resolved features [35, 36, 37].

However, there is still a significant gap between experimental and simulation analyses: experimental analyses often allow only one or two timescales to be distinguished [29, 38], suggesting simple 2- or 3-state models are sufficient to describe their behavior. In particular, in the search for the "protein folding speed limit", a large number of fast-folding proteins have been measured - and most of them appear two be two-state systems in current measurement techniques [20, 39]. In contrast, MD simulations often reveal a considerably more complex picture with multiple metastable states and a multitude of relaxation times [40, 36, 41]. Theoretically, the macroscopically detectable changes have been proposed to arise from a stochastic walk on a rugged multidimensional energy landscape [42], possibly involving a hierarchy of barriers, resulting in a hierarchy of relaxation time scales [43], or, alternatively, a jump process on a transition network between conformational substates [44, 40, 45] for which a given structural change may involve multiple pathways [36].

Interestingly, subtle experiments with careful analysis do also indicate that there is additional complexity beyond the one or two most prevalent relaxation timescales [46, 47, 48, 49, 8, 13, 50]. In a couple of cases, enzyme kinetics has been shown to be modulated by interchanging conformational substates [51]. Some protein folding experiments have found conformational heterogeneity, hidden intermediates, and the existence of parallel pathways [52, 53, 54, 55, 56, **?**]. The identification of kinetic processes based on features of the experimental signals which are on the level of statistical or systematic measurement errors is always subject to criticism. It is, therefore, important to understand what such features mean, and how they could possibly be enhanced by an optimized experimentals setup. More specifically, we ask questions such as:

- Is the slowest rate observed always the folding process?
- Can a given experiment see all relaxation processes which are present in the dynamics of the molecule?
- In ensemble relaxation experiments such as temperature-jump, how does the initial state influence the fingerprint?
- How can specific conformational changes be assigned to the observed relaxation timescales?
- How does one design an experiment, i.e. choose the optimal attachment points for the chromophores or choose the optimal site for isotopic labelings, such that the timescale of a particular process is optimally resolved?

We attempt to assemble a systematic approach of unraveling the complex kinetics of macromolecules. This is done by building a Markov state model (MSM). Markov models approximate the molecular kinetics by decomposing the molecular state space into many small substates and specifying a transition probability between each pair of substates [57, 58, 59, 60, 61, 44, 62, 63, 64]. When the system has metastable states, the slow kinetics can often be described in terms of a reduced model that only has

transition probabilities or rates between metastable states [65, 36, 5, 66, 67, 68]. How to construct Markov models from molecular dynamics simulations and validate them has been extensively reviewed elsewhere [57]. Here, we only repeat the basic steps in Sec. 5 of the paper while the subsequent sections start by assuming that a Markov model description is given and explain how this description can be related to experimental observables to arrive at an assignment of structural rearrangement to measureable features in kinetic experiments.

In the present paper we focus on flourescence spectroscopy and FRET spectroscopy, either conducted as equilibrium measurements by correlating fluorescence fluctuation of dilute samples (FCS), or by starting an ensemble from a specific off-equilibrium distribution (e.g. as done by T-jump). However our results are generally valid and can be applied to any single-molecule experiment including experiments which are not based on spectroscopy, such as atomic force microscopy, optical or magnetic tweezer experiments.

2 Model of a protein folding equilibrium

Consider the folding equilibrium of an illustrative protein folding model shown in fig. 1. The protein consists of two domains, an α-helix and a β-sheet, linked by a short loop regions. Each of the two domains is assumed to fold and unfold independently of the other, which leads to a folding equilibrium with four possible *conformational states*: (i) both domains folded h_f-β_f, (ii) helix folded and β-sheet unfdolded h_f-β_u, (iii) helix unfolded and β-sheet folded h_u-β_f, and (iv) both domains unfolded h_u-β_u. The transition between two of these states is called a *conformational change*, and the corresponding transition probability is indicated by the thickness of the arrow between the two states.

We model the transitions between the conformations as a Markov jump process. While these models typically are a simplification of the true dynamics in the high-dimensional conformational space of the molecule, one often finds in practice that they constitute quantitatively correct and useful approximations that now have a solid theoretical foundation [69]. The *Markov state model* (MSM) directly yields the rates of the *dynamic processes* which are present in the folding equilibrium. The rates which are measured in experiments are associated to these processes. Note that, in most cases, they are not directly linked to the transition rates between individual conformational states.

The model protein has three possible attachment sites for chromophores, represented by yellow stars numbered 1 to 3 in fig. 1. In a classical single-molecule fluorescence quenching or FRET experiment, one would chose two of these attachment points and measure the conformational dynamics as a projection onto a coordinates that depends mainly on the distance between the two chromophores. By autocorrelating the measured signal one can extract the timescales of the dynamics. However, which timescales are observed not only depends on the underlying conformational equilibrium but also on the choice of chromophore attachment points. When considering a temperature jump experiment of the same system, matters become even more complex, because now the observability of the system's relaxation timescales additionally depends on the initial and final probability distribution of the system.

In section 3 we review how the measured spectrum emerges from the underlying MSM, the chosen observable and the initial distribution. In section 4.2 we apply this theory to the model protein.

3 Theory

3.1 Markov state models

Consider a *state space* Ω consisting of N discrete *microstates*

$$\Omega = \{S_1, S_2, ...S_N\}\,.$$

In the context of molecules, this state space usually is the conformational space spanned by either all or the most important conformational degrees of freedom of the molecule. A microstate is a small volume element in this high-dimensional space. The microstates cover the entire (accessible) space, but do not overlap. See [57] for a discussion how this discretization of the continuous state space affects the quality of the Markov model.

The movement of the molecule in this (conformational) state space is modeled as *time-discrete process* s with a time step τ

$$s = (s_0,\ s_\tau,\ s_{2\tau},\ s_{3\tau}...)\,. \tag{1}$$

The probability of finding the molecule in an state j at time $t = n\tau$, in principle, depends on the entire history of the process

$$\mathbb{P}(s_{n\tau} = S_j \mid s_{(n-1)\tau},\ s_{(n-2)\tau},\ s_{(n-3)\tau}, ...\ s_0).$$

Only if this long conditional probability can be truncated to

$$\mathbb{P}(s_{n\tau} = S_j \mid s_{(n-1)\tau}),$$

the process is called *Markovian [70].* The probability of finding the molecule in state j at time $t = n\tau$ then only depends on the state the molecule has been in at the previous time step (memory-free process). These probabilities do not change in the course of the process (time invariant). They only depend on the pair of microstates $\{s_{n\tau} = S_{j,}s_{(n-1)\tau} = S_i\}$ and the time step τ of the process. Arranged in a $N \times N$ matrix, they form the *transition matrix* $\mathbf{T}(\tau)$ with

$$T_{ij} = \mathbb{P}(s_{n\tau} = S_j \mid s_{(n-1)\tau} = S_i)\,,$$

the central property of a Markov state model. The matrix elements represent the probability that the molecule is found in microstate S_j provided that it has been in microstate S_i a time τ earlier. The ith row of this transition matrix represents all possibilities a molecule in state i has: it can either stay in its current microstate (T_{ii}) or move to any of the other $n-1$ microstates (T_{ij}). Consequently, the elements of each row in $\mathbf{T}(\tau)$ sum up to 1

$$\sum_{j=1}^{n} T_{ij} = 1\,,\quad \forall i$$

(row-stochastic matrix).

Representing the conformational dynamics as a Markov state model is a good approximation if

1. the degrees of freedom (d.o.f.), which are not included in the model, (marginal d.o.f. or bath d.o.f.) move on faster time-scales than the d.o.f. included in the model (relevant d.o.f.) and are not coupled strongly to the latter [64], and
2. the conformational states of the molecule are projected onto disjunct regions in the space of the relevant d.o.f., i.e. they do not overlap, and
3. the transition region is sufficiently finely discretized [69, 71, 57], and
4. the time step τ is large enough [69, 57].

While these requirements now have a solid theoretical underpinning, a practical analysis must test whether the MSM is consistent with the available simulation data within statistical errors [57]

The experiments we consider in the following are conducted under equilibrium conditions. Hence, the dynamics can be represented by a single (time-invariant) transition matrix. Given a "good" Markov state model, its transition matrix can be used to generate possible coarse-grained trajectories of a single molecule in the state space of the model. More importantly, the transition matrix contains the complete information of dynamics of an ensemble of molecules in this state space. Let $\mathbf{p}(t)$ be a probability vector with N elements, where the ith element represents the fraction of molecules in the ensemble which are found in state S_i at a time t. Consequently,

$$\sum_{i=1}^{N} p_i(t) = 1 \,.$$

The time evolution of this vector is completly determined by the transition matrix $\mathbf{T}(\tau)$

$$\mathbf{p}^\top(t+\tau) = \mathbf{p}^\top(t)\mathbf{T}(\tau)\,, \tag{2}$$

where $\mathbf{p}^\top(t)$ denotes the transpose of the vector $\mathbf{p}(t)$. Given an initial probility vector $\mathbf{p}(0)$, the probability at any discrete time $k\tau$ can be calculated by repeatedly applying $\mathbf{T}(\tau)$ to it

$$\mathbf{p}^\top(k\tau) = \mathbf{p}^\top(0)\mathbf{T}^k(\tau) = \mathbf{p}^\top(0)\mathbf{T}(k\tau). \tag{3}$$

Eq. 2 and eq. 3 are equivalent, which becomes obvious if one realizes that $\mathbf{p}^\top(2\tau) = \mathbf{p}^\top(\tau)\mathbf{T}(\tau) = \mathbf{p}^\top(0)\mathbf{T}(\tau)\mathbf{T}(\tau) = \mathbf{p}^\top(0)\mathbf{T}^2(\tau)$. They are known as the Chapman-Kolmogorov equation.

If s is ergodic, and if the potential energy surface on Ω is time invariant, then from physical intuition it is obvious that there must be a unique stationary distribution π, and that this distribution must not change under the action of $\mathbf{T}(\tau)$, i.e.

$$\pi^T \quad = \quad \pi^T\mathbf{T}(\tau) \tag{4}$$

Indeed, this stationary distribution emerges as the first left eigenvector of $\mathbf{T}(\tau)$ associated with the eigenvalue $\lambda_1 = 1$ [72]. Under equilibrium conditions, the dynamics of a molecular systems always fulfills detailed balance

$$\pi_i T_{ij} \quad = \quad \pi_j T_{ji} \tag{5}$$

with respect to this stationary distribution π. This means that the number of systems in the ensemble, which go from state i to state j, is the same as the number of systems going from state j to i. This has a number of convenient consquences on the properties of the MSM, as will be explained below.

Likewise, physical intuition tells us that any initial vector $\mathbf{p}(0)$ eventually converges to the stationary distribution π. Indeed, one can show that for any $\mathbf{p}(0)$

$$\lim_{k\to\infty} \mathbf{p}^\top(0)\mathbf{T}^k(\tau) = \pi^T \,,$$

where π is the first left eigenvector of $\mathbf{T}(\tau)$.

The transition matrix is, however, considerably more than a black box which converts the probability at some point in time t to the probability at some time $k\tau$ later. The way the probability vector changes with time and eventually converges to the stationary probability vector can be understood in terms of the eigenvectors of the transition matrix. This is illustrated in Fig. 2 (adapted from [57]) for a simple example.

The upper part in Fig. 2a shows a energy landscape along a single degree of freedom with four energy minina (A, B, C, D) and a high energy barrier between minima the two minima on the left side of the coordinate (A, B) and those on the right (C, D). The coordinate is discretized into one hundred microstates. The lower part of Fig. 2a shows the corresponding equilibrium probability vector π at a given temperature T. In Fig. 2b a transition matrix, which is given by a diffusion process on this energy landscape (see [57] for details), is presented. The matrix elements are color-coded: red represents high transition probabilities between two microstates, and white or light blue represents transition probabilities which are zero or close to zero. Reading the ith row from left to right, one finds the transition probabilities of from state x_i into states which belong to minimum A ($1 \leq j < 25$), to minimum B ($25 \leq j < 50$), to minimum C ($50 \leq j < 75$), and eventually to minimum D ($75 \leq j < 100$). The four blocks along the diagonal structure of $\mathbf{T}(\tau)$ correspond to the four minima in the energy surface. They reflect the fact that transitions within a minimum are much more likely than transitions from one minimum to the other.

These properties can be used in order to identify the metastable states of the system. The mathematical foundation for this was worked out in [60] and further developed in [66]. Metastability analysis has been subject to various studies and applications [5, 73, 36, 27] and is now a major tool to reduce the complexity of macromolecular kinetics to humanly understandable terms.

Transition matrices can, as any diagonalizable matrix, be written as a linear combination of their left eigenvectors, their eigenvalues and their right eigenvectors

$$\mathbf{T}(\tau) \quad = \quad \sum_{i=1}^{n} \lambda_i(\tau) \mathbf{r}_i \mathbf{l}_i^\top \, . \tag{6}$$

and thus, for longer timescales:

$$\mathbf{T}^k(\tau) \quad = \quad \sum_{i=1}^{n} \lambda_i^k(\tau) \mathbf{r}_i \mathbf{l}_i^\top \, . \tag{7}$$

The transition matrix $\mathbf{T}(k\tau) = \mathbf{T}^k(\tau)$ which transports an initial probability k time steps forward is again a linear combination of the eigenvectors and eigenvalues. These linear combinations (eq. 6 and 7) are known as *spectral decomposition* of the transition matrix. They are very useful for connecting the dynamics of the molecule to the measured signal, which is in section 3.2.

Eq. 7 is the key for understanding how the transition matrix transforms a probability vector. The complete process consists of n subprocesses $\mathbf{r}_i \mathbf{l}_i^\top$, each of which is weighted by the eigenvalue λ_i raised to the power k. Because the transition matrix is a row-stochastic matrix, it always has one eigenvalue which is equal to one $\lambda_1 = 1$ [72]. Raising this eigenvalue to the power k does not change the weight of the corresponding subprocess $\mathbf{r}_1 \mathbf{l}_1^\top : 1^k = 1$. $\mathbf{r}_1 \mathbf{l}_1^\top$ is the stationary process, which we postulated in eq. 4, and $\mathbf{l}_1 = \pi$.

All other eigenvalues of the transition matrix are guaranteed to be smaller than one in absolute value [72]

$$|\lambda_i| \leq 1 \;\; \forall i \, .$$

The weights of the corresponding processes, hence, decay exponentially

$$\lambda_i^k = \exp\left(k \ln \lambda_i\right) = \exp\left(\frac{t}{\tau} \ln \lambda_i\right) = \exp\left(-\frac{t}{t_i}\right) \tag{8}$$

with the implied timescale t_i of the decay process

$$t_i = -\frac{\tau}{\ln \lambda_i} \, . \tag{9}$$

The smaller the eigenvalue λ_i, the smaller the implied timescale t_i, the faster the corresponding process decays. Fig. 2d shows the 15 largest eigenvalues of the transition matrix in Fig. 2b. There is one eigenvalue, λ_1, which is equal to one, followed by three eigenvalues, λ_2 to λ_4, which are close to one. These four *dominant eigenvalues* are separated by a gap from the remaining eigenvalues. Hence, the transition matrix consists of a stationary process, three slow processes and 96 processes which decay quickly. After a few time steps, only the four dominant processes contribute to the evolution of the probability vector. How these processes alter this vector, is determined by the shape of the corresponding eigenvectors.

Fig. 2c shows the four dominant right eigenvectors. The first eigenvector corresponds to the stationary

process and is, therefore, constant. The second eigenvector corresponds to the slowest process and has positive signs in regions A and B and negative signs in regions C and D. This shape effectively moves probability density across the largest barrier in the energy surface. Since the eigenvector is approximately constant within the combined region (A,B) and (C,D) left and right of the barrier, it does not alter the relative probability distribution within these regions. The third eigenvector, analogously, moves density between A and B, the fourth moves density between C and D.

A transition matrix which fulfills detailed balance (eq. 5) has several convenient properties. First, all of its eigenvalues and eigenvectors are guaranteed to be real. Second, defining a diagonal matrix Π in which the diagonal elements are equal to the equilibrium distribution π

$$\Pi: \quad \Pi_{ij} = \begin{cases} \pi_i & \text{if } i = j \\ 0 & \text{else} \end{cases},$$

the left and right eigenvectors are interconvertable [72]

$$\begin{aligned} \mathbf{l}_i &= \Pi \mathbf{r}_i \\ \mathbf{r}_i &= \Pi^{-1} \mathbf{l}_i. \end{aligned} \tag{10}$$

Hence, its spectral decomposition (eq. 7) can be written only in terms of the left eigenvectors

$$\mathbf{T}^k(\tau) = \Pi^{-1} \sum_{i=1}^{n} \lambda_i^k(\tau) \mathbf{l}_i \mathbf{l}_i^T. \tag{11}$$

In the experiments, which we discussed in the following sections, the dynamics of the molecule is governed by equilibrium dynamics (no varying forces, temperatures etc.). We will, therefore, always assume detailed balance.

3.2 Calculating experimental expectation values from Markov models

We now consider the case that an experiment is conducted which measures observable a (and possibly additional observables b, c, ...). This observable has a scalar value for every state S_i, although vector- or function-valued observables could be treated in a similar way. It is implicitly assumed that the state space discretization used in the MSM in fine enough such that the value of a varies little within individual states S_i. The discretized observable vector $\mathbf{a}$ contains the mean values of individual states, a_i.

We consider three different types of experiments: *equilibrium experiments*, *relaxation experiments* and *correlation experiments*. In *equilibrium experiments*, the observed molecule is in equilibrium with the current conditions of the surroundings (temperature, applied forces, salt concentration etc.), and the mean value of an observable a, $\mathbb{E}_\pi[a]$, is recorded. This may be either done my measuring $\mathbb{E}_\pi[a]$ directly from an unperturbed ensemble of molecules, or by recording sufficiently many and long single molecule

traces $a(t)$ and averaging over them. The expression for the expected measured value of a is purely stationary, i.e., it does not depend on the time t:

$$\mathbb{E}_\pi[a] \quad = \quad \sum_{i=1}^{N} a_i \pi_i = \langle \mathbf{a}, \pi \rangle \ . \tag{12}$$

$\langle x, y \rangle$ denotes the Euclidean scalar product between two vectors x and y and $\mathbb{E}[...]$ the expectation value.

In the second type of experiments, *relaxation experiments,* the observed molecule or ensemble is allowed to equilibrate under a given set of conditions to the distribution $\mathbf{p}(0)$. At time $t = 0$ these conditions are changed virtually instantaneously to another set of conditions which are associated with a different equilibrium distribution π. Now an observable a is traced over time whose mean value decays from the old expectation $\mathbb{E}_{\mathbf{p}(0)}[a]$ to the new expectation $\mathbb{E}_\pi[a]$. The way this relaxation, $\mathbb{E}[a(t)]$, happens in time, allows conclusions on the intrinsic dynamical processes of the molecule. This principle is used in temperature- and pressure jump experiments. It can likewise be reproduced by single molecule experiments by measuring many trajectories whose conditions are rapidly changed at certain points in time, and then averaging over this trajectory ensemble. Single-molecule relaxation measurements can be realized e.g. by cycling the Mg^{2+} concentration in single-molecule FRET experiments or by changing the reference positions in optical tweezer experiments. Computationally, the dynamics of the molecule after $t = 0$ are governed by a transition matrix $\mathbf{T}(\tau)$ which reflects the conditions after the jump. At each time $t = k\tau$, the ensemble will be distributed as $\mathbf{p}^T(k\tau) = \mathbf{p}^T(0)\mathbf{T}^k(\tau)$. The expectation value of $a(t)$ changes accordingly with time:

$$\mathbb{E}_{\boldsymbol{p}(0)}[a(k\tau)] \quad = \quad \sum_{i=1}^{N} a_i p_i(k\tau) = \langle \mathbf{a}, \mathbf{p}(k\tau) \rangle. \tag{13}$$

Using eq. 3, and eq. 11 one can expand eq. 13 to

$$\begin{aligned} \mathbb{E}[a(k\tau)] \quad &= \quad \left\langle \mathbf{a}, \left[\mathbf{p}^\top(0)\mathbf{T}^k(\tau) \right]^\top \right\rangle \\ &= \quad \left\langle \mathbf{a}, \left[\mathbf{p}^\top(0)\Pi^{-1} \sum_{i=1}^{N} \lambda_i^k(\tau)\mathbf{l}_i\mathbf{l}_i^\top \right]^\top \right\rangle \\ &= \quad \left\langle \mathbf{a}, \left[\mathbf{p'}^\top(0) \sum_{i=1}^{N} \lambda_i^k(\tau)\mathbf{l}_i\mathbf{l}_i^\top \right]^\top \right\rangle \end{aligned} \tag{14}$$

where we have replaced the probability distribution $\mathbf{p}(0)$ by the excess probability distribution

$$\mathbf{p'}(0) \quad = \quad \Pi^{-1}\mathbf{p}(0) \tag{15}$$

with $p_i^{'}(0) = {}^{p_i(0)}/_{\pi_i}$. Rearranging the sum and the scalar products, one obtains

$$\begin{aligned}
\mathbb{E}[a(k\tau)] &= \left\langle \mathbf{a},\ \left[\sum_{i=1}^{N} \lambda_i^k(\tau) \left(\mathbf{p}^{'\top}(0),\ \mathbf{l}_i\right) \mathbf{l}_i^{\top}\right]^{\top} \right\rangle && (16)\\
&= \sum_{i=1}^{N} \lambda_i^k(\tau) \left\langle \mathbf{a},\ \left[\left(\mathbf{p}^{'\top}(0),\ \mathbf{l}_i\right) \mathbf{l}_i^{\top}\right]^{\top} \right\rangle \\
&= \sum_{i=1}^{N} \lambda_i^k(\tau) \left\langle \mathbf{a}, \mathbf{l}_i \right\rangle \left\langle \mathbf{p}^{'\top}(0),\ \mathbf{l}_i \right\rangle \\
&= \left\langle \mathbf{a}, \pi \right\rangle \left\langle \mathbf{p}^{'\top}(0),\ \pi \right\rangle + \sum_{i=2}^{N} \exp\left(-\frac{k\tau}{t_i}\right) \left\langle \mathbf{a}, \mathbf{l}_i \right\rangle \left\langle \mathbf{p}^{'\top}(0),\ \mathbf{l}_i \right\rangle . && (17)
\end{aligned}$$

In this notation, it becomes obvious that the time-dependence of the expected measured signal $\mathbb{E}\left[a(k\tau)\right]$ has the fom of a multiexponential decay function

$$f(t) = \gamma_1 + \sum_{i=2} \gamma_i^{\text{relax}} \exp\left(-\frac{t}{t_i}\right) , \tag{18}$$

with $t = k\tau$. γ_i is the amplitude of the a ith decay process and is given as

$$\gamma_i^{\text{relax}} = \left\langle \mathbf{a}, \mathbf{l}_i \right\rangle \left\langle \mathbf{p}^{'\top}(0),\ \mathbf{l}_i \right\rangle . \tag{19}$$

The respective decay constant t_i is equal to the ith implied timescale of the underlying transition matrix. Note that the individual components of the signal decay until the expected measured signal of the *equilibrium experiment* under the target conditions is reached

$$\lim_{k\to\infty} \mathbb{E}[a(k\tau)] = \left\langle \mathbf{a}, \pi \right\rangle \left\langle \mathbf{p}^{'\top}(0),\ \pi \right\rangle = \left\langle \mathbf{a}, \pi \right\rangle = \mathbb{E}_{\pi}[a] .$$

The amplitudes γ_i^{relax} in Eq. (eq. 19) reflect the extent to which a given mode (eigenvector) of the dynamics influences the time-evolution of $\langle a(k\tau)\rangle_{\mathbf{p}(0)}$. This depends on two factors

1. how much probability density is transported *via* this mode during the relaxation from $\boldsymbol{p}(0)$ to π, represented by the scalar product $\left\langle \mathbf{p}'^{\top}(0),\ \mathbf{l}_i \right\rangle$
2. how sensitive $\mathbf{a}$ is to changes along this mode, represented by the scalar product, $\langle \mathbf{a}, \mathbf{l}_i \rangle$.

A third type experiments considered here are *correlation experiments* which report on the intrinsic molecular kinetics *via* time correlation functions of certain observables. One way to measure such correlation functions is by tracing the equilibrium fluctuations of a molecule subsequently correlating this signal in time. This is e.g. done in fluorescence correlation spectrosopy (FCS). In experiments, in which two signals, a and b, are measured simultaneously, also cross-corrleation functions can be extracted from the measured signal Multiparameter-FRET experiments [74, 75] or Multichromophore FRET experiments

[7] are examples of this type of experiment. A way of directly measuring time correlation functions of atomic positions are X-ray and neutron scattering experiments.

3.3 Calculating experimental correlation functions from Markov models

We now use the existing formalism to derive expressions which predict the auto-correlation function of observable a and the cross-correlation function of observable a and b. Altough, to the best of our knowledge, the auto- or crosscorrelation analysis of the measured signal has not been applied to relaxation experiments yet, we also include this possibility into our derivation for completeness. In total, we obtain four different expressions for the four possible experimental situations (equilibrium or relaxation experiment combined with either auto- or cross-correlation function). The respective expressions of the amplitudes are summarized in Tab. 8.

We start with the most complex case: cross-correlation function in a relaxation experiment. All other results are specializations of this case. The movement of the molecule is represented by the jump process on the discrete microstates S_i (eq. 1). Each state is associated with a value of each of the measured signal, represented by the signal vectors $\mathbf{a}$ and $\mathbf{b}$. The correlation of $a(t)$ and $b(t)$, given an initial probability represented by $\mathbf{p}(0)$, is defined as

$$\begin{aligned}\mathbb{E}_{\mathbf{p}(0)}[cor(a,b;\,\tau)] &= \sum_{i=1}^{N}\sum_{j=1}^{N} a_i \mathbb{P}(s_0 = S_i)\cdot b_j \mathbb{P}(s_{k\tau} = S_j \mid s_0 = S_i) \\ &= \sum_{i=1}^{N}\sum_{j=1}^{N} a_i p_i(0)\cdot b_j \mathbb{P}(s_{k\tau} = S_j \mid s_0 = S_i)\end{aligned}$$

If s_t is a Markov processes with transition matrix $\mathbf{T}(\tau)$, then the conditional probability $\mathbb{P}(s_{k\tau} = S_j \mid s_0 = S_i)$ can be replaced by the corresponding matrix element $\left[\mathbf{T}^k(\tau)\right]_{ij}$ of the transition matrix raised to the power k. Introducing a diagonal matrix $\mathbf{P}(0)$ in which the diagonal elements are equal to the initial probability vector $P_{ii}(0) = p_i(0)$, we can formulate the cross-correlation function as a vector-matrix equation

$$\mathbb{E}_{\mathbf{p}(0)}[cor(a,b;\,k\tau)] = \mathbf{a}^\top \mathbf{P}(0)\mathbf{T}^k(\tau)\mathbf{b}. \tag{20}$$

We introduce an excess initial density $\mathbf{P}'(0) = \Pi^{-1}\mathbf{P}(0)$ (analogous to eq. 15), replace the transition matrix by its spectral decomposition (eq. 11), use the definition of the implied timescale (eq. 8) and obtain an expression which has the same structure as eq. 18

$$
\begin{aligned}
\mathbb{E}_{\mathbf{p}(0)}[cor(a,b;\,k\tau)] &= \mathbf{a}^T\mathbf{P}(0)\Pi^{-1}\left[\sum_{i=1}^{N}\lambda_i^k\mathbf{l}_i\mathbf{l}_i^\top\right]\mathbf{b} \\
&= \sum_{i=1}^{N}\lambda_i^k\sum_{r,s=1}^{N}a_r\frac{p_r(0)}{\pi_r}\left\{\mathbf{l}_i\mathbf{l}_i^\top\right\}_{rs}b_s \\
&= \sum_{i=1}^{N}\lambda_i^k\sum_{r,s=1}^{N}a_r\frac{p_r(0)}{\pi_r}\left\{\boldsymbol{l}_i\right\}_s\left\{\boldsymbol{l}_i\right\}_s b_s \\
&= \sum_{i=1}^{N}\lambda_i^k\left\langle\mathbf{a},\mathbf{P}'(0)\mathbf{l}_i\right\rangle\left\langle\mathbf{b},\mathbf{l}_i\right\rangle \\
&= \left\langle\mathbf{a},\mathbf{P}'(0)\pi\right\rangle\left\langle\mathbf{b},\pi\right\rangle + \sum_{i=2}^{N}\exp\left(-\frac{k\tau}{t_i}\right)\left\langle\mathbf{a},\mathbf{P}'(0)\mathbf{l}_i\right\rangle\left\langle\mathbf{b},\mathbf{l}_i\right\rangle \qquad (21)
\end{aligned}
$$

The ith decay constant of this multiexponential decay is given as the implied timescale associated with the ith eigenvector of the transition matrix. The corresponding amplitude is given as

$$
\gamma_i^{\text{relax, cross-cor}} = \left\langle\mathbf{a},\mathbf{P}'(0)\mathbf{l}_i\right\rangle\left\langle\mathbf{b},\mathbf{l}_i\right\rangle\ . \qquad (22)
$$

The autocorrelation function of a relaxation experiment is obtained by replacing the signal vector $\mathbf{b}$ by $\mathbf{a}$ in eq. 20 and 21

$$
\mathbb{E}_{\mathbf{p}(0)}[cor(a,a;\,k\tau)] = \left\langle\mathbf{a},\mathbf{p}(0)\right\rangle\left\langle\mathbf{a},\pi\right\rangle + \sum_{i=1}^{N}\exp\left(-\frac{k\tau}{t_i}\right)\left\langle\mathbf{a},\mathbf{P}'(0)\mathbf{l}_i\right\rangle\left\langle\mathbf{a},\mathbf{l}_i\right\rangle.
$$

with the amplitudes

$$
\gamma_i^{\text{relax, auto-cor}} = \left\langle\mathbf{a},\mathbf{P}'(0)\mathbf{l}_i\right\rangle^2\ . \qquad (23)
$$

In the more common case that the correlation functions are measured under equilibrium conditions, the initial density equals the equilibrium density and consequently $\mathbf{P}'(0)$ is equal to the identity matrix. The cross- and autocorrelation are thus given as

$$
\begin{aligned}
\mathbb{E}_{\pi}[cor(a,b;k\tau)] &= \sum_{i=1}^{N}\lambda_i^k\left\langle\mathbf{a},\mathbf{l}_i\right\rangle\left\langle\mathbf{b},\mathbf{l}_i\right\rangle \\
&= \left\langle\mathbf{a},\pi\right\rangle\left\langle\mathbf{b},\pi\right\rangle + \sum_{i=2}^{N}\exp\left(-\frac{k\tau}{t_i}\right)\left\langle\mathbf{a},\mathbf{l}_i\right\rangle\left\langle\mathbf{b},\mathbf{l}_i\right\rangle
\end{aligned}
$$

with the amplitudes

$$\gamma_i^{\text{cross-cor}} = \langle \mathbf{a}, \mathbf{l}_i \rangle \langle \mathbf{b}, \mathbf{l}_i \rangle \ . \tag{24}$$

and

$$\begin{aligned} \mathbb{E}_\pi[cor(a, a;\, k\tau)] &= \sum_{i=1}^{N} \lambda_i^k \langle \mathbf{a}, \mathbf{l}_i \rangle^2 \\ &= \langle \mathbf{a}, \pi \rangle^2 + \sum_{i=2}^{N} \exp\left(-\frac{k\tau}{t_i}\right) \langle \mathbf{a}, \mathbf{l}_i \rangle^2 \ . \end{aligned} \tag{25}$$

with the amplitudes

$$\gamma_i^{\text{cross-cor}} = \langle \mathbf{a}, \mathbf{l}_i \rangle^2 \ . \tag{26}$$

4 Application to model systems

4.1 1D energy surface

Figure 3 shows the eigenvectors of our model of an one-dimensional energy surface (fig. 2), two different observables and two different initial distributions. The observables model a fluorescence quenching experiment. With $\boldsymbol{a}_1$ the chromophore fluoresces if the system is in state A or B, whereas fluorescence is quenched in state C and D. With $\boldsymbol{a}_2$ fluorescence is quenched in A, B, and D.

Due to the hierachical nature of the energy landscape an interpretation of the measured timescales in terms of individual conformational changes can be misleading. In a four state system, there are six possible transitions, i.e. six possible conformational changes. Yet the dynamics in this state space is described by only three relaxation processes (non-stationary eigenvectors of the corresponding transition matrix). Processes three and four indeed correspond *mostly* to transitions from one conformational state to another. However, process two represents the transition between the group $\{A, B\}$ and the group $\{C, D\}$, i.e. it can be associated to the transition across the barrier separating B and C.

The stationary process is always detected. The overlap between the observables and the initial distributions with the eigenvectors of the model, represented by the respective scalar product, are shown to the left and right of the eigenvector plots in Fig. 3. Because the stationary distribution $\mathbf{l}_1 = \pi$ has only positive entries, the scalar product with any observable or any initial distribution is greater than zero. Consequently, γ_1 in eq. 18 is always greater than zero. Although dynamical fingerprints do normally not include this stationary part [27], we here include the overlap of observable and initial distributions with the stationary process for completeness.

Not all dynamical processes can be detected. Whether a given process appears in the experimental fingerprint depends on the overlap of the observable with this process. For example, the overlap of $\mathbf{a}_1$ with the third and the fourth process is nearly zero. These processes correspond to swaps between states which have the same signal value (A $\leftrightarrow$ B and C $\leftrightarrow$ D). Hence, $\mathbf{a}_2$ is insensitve to them, and $\gamma_3 \approx 0$, and $\gamma_4 \approx 0$ in an autocorrelation experiment (eq. 25). Only the second process can be observed with $\mathbf{a}_1$. $\mathbf{a}_2$ is sensitive to the second and fourth process but not to the third. Compare the scalar products in Fig. 3 with the Fig. 4a and 4b.

By a clever choice of $\mathbf{a}$ and $\mathbf{p}(0)$ one can selectively measure a specific process. It is not possible to observer processes in a relaxation experiment which would be invisible in an equilibrium experiment (Fig. 4c and 4d), because the amplitude is proportional to the overlap of the observable with the eigenvector (Eq. 23). The amplitude is also proportional to the overlap of the eigenvector with the initial distribution. By choosing the initial distribition appropriately one can "hide" processes which are visible in the equilibrium experiment. This allows for the selective measurement of processes which might be hard to extract from the multiexponential decay in the corresponding equilibrium experiment, for example processes which decay on short timescales. This is shown in Fig. 4f. An unwise combination of observable and initial distribution, however, may lead to a spectrum in which only the stationary process can be observed (Fig. 4e).

4.2 Protein folding model

We model the folding equilibrium of the model protein (Fig. 1) as a Markov model with four states which are defined as: state 1 = $h_f\beta_f$ (both domains folded), state 2 = $h_f\beta_u$ (helix folded, β-sheet unfolded), state 3= $h_u\beta_f$ (helix unfolded, β-sheet folded), state 4 = $h_u\beta_u$ (both domains unfolded). Suppose, we have observed the protein and took note of the transitions after each time step τ. The matrix

$$C(\tau) = \begin{pmatrix} 12000 & 20 & 2 & 0 \\ 20 & 7000 & 0 & 2 \\ 2 & 0 & 6000 & 20 \\ 0 & 2 & 20 & 1000 \end{pmatrix} . \tag{27}$$

contains the total number of observed transitions. By normalizing each row one obtains the corresponding transition matrix

$$T(\tau) \approx \begin{pmatrix} 0.9982 & 0.0017 & 0.0002 & 0 \\ 0.0028 & 0.9969 & 0 & 0.0003 \\ 0.0003 & 0 & 0.9963 & 0.0033 \\ 0 & 0.0020 & 0.0196 & 0.9785 \end{pmatrix} \tag{28}$$

which represents the Markov model. Note that due to rounding errors, the rows in eq. 28 do not exactly sum up to one.

The thickness of the arrows in Fig. 1 reflect the transition probabilities between the states. There is a fast equilibrium between the folded and the unfolded conformation of the β-sheet if the helix is unfolded. The folding of the complete protein mainly occurs through a cooperative folding pathway *via* the state $h_f\beta_u$. Folding of the helix when the β-sheet is already formed is considerably less likely. The eigenvalue spectrum, as well as the left and right eigenvectors of $\mathbf{T}$ are shown in Fig. 5.

The slowest rate in the system is not *a priori* the "folding rate" of the protein. The timescales observed in single-molecule experiments are often interpreted in terms of conformational changes in the examined molecule, and the slowest process is typically associated with the overall folding and unfolding. In the present example, the folding rate could either be defined as the rate of going from state 4 to state 1, or as the rate of going from the ensemble of states 2, 3, and 4 to state 1. However, none of the eigenvectors corresponds to either of the two processes. Rather they have the following interpretation: $\mathbf{l}_2$ represents the folding and unfolding of the helix, $\mathbf{l}_3$ represents the folding equilibrium of the β-sheet when the helix is already formed, and $\mathbf{l}_4$ represents the same equilibrium when the helix is unfolded. Care should be taken to differentiate between the folding rate and the rate limiting step in a folding equilibrium which in this case is the formation of the helix.

Fig. 5b shows two initial distributions, as they could be used in jump experiments. The first one ($\mathbf{p}_1(0)$) represents an ensemble in which all systems are folded, the second one ($\mathbf{p}_2(0)$) an ensemble in which all systems are completely unfolded. Fig. 5c shows observable vectors which correspond to a FRET experiment in which the chromophores are attached at sites 1 and 2 ($\mathbf{a}_1$), sites 2 and 3 ($\mathbf{a}_2$), and

sites 1 and 3 ($\mathbf{a}_3$). For all three observables, we discuss the autocorrelation fingerprints of equilibrium experiments. We also discuss an equilibrium multichromophore experiment in which observables $\mathbf{a}_1$ and $\mathbf{a}_2$ are combined (donor at site 2, first acceptor at site 1, second acceptor at site 3). As for the relaxation experiments, we discuss the combination of the two initial distributions with observable $\mathbf{a}_3$.

The three observables are an intuitive example why some observables do not resolve all processes present in the system. From Fig. 1 it is clear that, if the chromophores are attached at site 1 and 2 (observable $\mathbf{a}_1$), the experiment will only be sensitive to processes which involve the folding or unfolding of the helix. This is reflected in the scalar products of $\mathbf{a}_1$ with the $\mathbf{l}_2$, $\mathbf{l}_3$, and $\mathbf{l}_4$ (table 6). $\mathbf{a}_1$ has a large overlap with $\mathbf{l}_2$, but only small or virtually no overlap with $\mathbf{l}_4$, and $\mathbf{l}_3$. Correspondingly, $\mathbf{a}_2$ (chromophores attached at sites 2 and 3) is sensitive to $\mathbf{l}_3$, and $\mathbf{l}_4$, which represent the folding of the β-sheet, but rather insensitive to $\mathbf{l}_2$. $\mathbf{a}_3$ (chromophores attached to sites 1 and 3) is sensitive to all three processes. The expected amplitudes of equilibrium experiments with $\mathbf{a}_1$, $\mathbf{a}_2$, or $\mathbf{a}_3$ are shown in Fig. 6a-c.

Given only the three-dimensional structure of a molecule it is often impossible to decide whether a particular observable can resolve all processes in the conformational equilibrium. However, with the help of MD simulations one can quantify the sensitivity of the observable to any process in the equilibrium. This is discussed in section 5.

The two initial probability distributions illustrate a pitfall of jump experiments. Not all processes are used when the system relaxes from a particular initial distribution to the equilibrium distribution. For example, in the relaxation from the folded state ($\mathbf{p}_1(0)$) the equilibrium between the folded and the unfolded conformation of the β-sheet is entirely achieved via $\mathbf{l}_3$, and not via $\mathbf{l}_4$ (table 2: $\langle \mathbf{p}_1(0), \mathbf{l}_3 \rangle$ =0.50, $\langle \mathbf{p}_1(0), \mathbf{l}_4 \rangle$ =0.00). When the system is relaxed from the unfolded state ($\mathbf{p}_2(0)$), however, the situation is reversed: $\mathbf{l}_4$ is active, wheras $\mathbf{l}_3$ is not (table 2: $\langle \mathbf{p}_2(0), \mathbf{l}_3 \rangle$ =0.06, $\langle \mathbf{p}_2(0), \mathbf{l}_4 \rangle$ =1.09). Therefore, even when an observable which is sensitive to all processes is chosen, like $\mathbf{a}_3$ in the present example, some processes might still be undetectable in a relaxation experiment. Fig. 6d and 6e. shows the expected amplitudes for the two relaxation experiments. For $\mathbf{p}_1(0)$ the fourth process has no amplitude, and for $\mathbf{p}_2(0)$ the third process has a very small amplitdue.

With multichromophore experiments the trade-off between selectivity and comprehensiveness is alleviated. An observable like $\mathbf{a}_3$ has the advantage of comprehensiveness. However, it can be very tedious and difficult to extract multiple timescales from a possibly noisy data set. In principle, it would possible to perform several experiment on a given system, each with a different observable, and combine the obtained results. Unless the sensitivity of the observables to the processes in the system is known, it will be hard to decide whether peaks which appear with similar timescales in two different experiments are the same conformational process slightly shifted or two different conformational processes with similar timescales. By performing a multiple-chromophore experiment one obtains the information of the two individual experiments, and additionally can use the information from the cross-correlation from the two signals to match peaks from the individual experiments (Fig. 6f) If two peaks in the individual experiments correspond to the same conformational process i, the amplitude in the cross-correlation fingerprint should be $\langle \mathbf{a}_1, \mathbf{l}_i \rangle \langle \mathbf{a}_2, \mathbf{l}_i \rangle$, where $\langle \mathbf{a}_1, \mathbf{l}_i \rangle$ and $\langle \mathbf{a}_2, \mathbf{l}_i \rangle$ are obtained as the square-root of the amplitudes in the respective auto-correlation fingerprint. If, on the other hand, the two individual experiments measure

disjunct sets of processes (as in our example), the amplitudes of in the cross-correlation fingerprint should be close to zero.

5 Experimental design using MD and MSM

5.1 Experimental dynamical fingerprints

To reconcile our Markov model analysis with measured data, it is useful to transform the experimental relaxation curve into timescales and amplitudes. In practice, this is often done by fitting a single- or multiexponential model. This approach is not objective as it requires the number of timescales to be fixed. For example, multiple exponentials with similar timescales, or a double-exponential where the larger timescale has a small amplitude will both yield visually excellent single-exponential fits with an effective timescale that may not exist in the underlying system (see [40] and SI of [27]). To prepare the experimental data for a systematic analysis, we propose to use a method that uniquely transforms the observed relaxation profile into an amplitude density of relaxation timescales (here called dynamical fingerprints). Several such methods have been developed especially maximum entropy or least squares based methods [76, 77]. In [27] we have developed a maximum-likelihood method which is available through the package SCIMEX (e.g. https://simtk.org/home/scimex) which is briefly discussed here.

Suppose a correlation or relaxation function $x_j = x(t_j)$ is given (e.g. from an experiment) at real time points $t_1, ..., t_o$. We expect from physical principles that this signal is a noisy realization of a function that is in fact a sum of multiple exponentials with initially unknown timescales and amplitudes, i.e. a function that can be represented by

$$y_\Phi(t) = \int_{t'} dt'\, \gamma(t')\, \exp\left(-\frac{t}{t'}\right),$$

i.e. the Laplace transform of the amplitude spectrum, or "fingerprint", $\gamma(t')$ is expected to consist of peaks. To computationally determine this fingerprint the timescale axis t' needs to be discretized using n spectral time points $t'_1, ..., t'_n$. With a fine timescale discretization we obtain a good approximation of the fingerprint:

$$y_\Phi(t) \approx \sum_{i=1}^{n} a_i \exp\left(-\frac{t}{t'}\right).$$

where the amplitudes a_i define a set of parameters $\Phi = \{a_i = a(t'_i)\}$ defining the fingerprint that needs to be determined. When each observation x_j comes with a Gaussian-shaped uncertainty σ_j, the log-Likelihood of a given fingerprint having generated the observed signal x is given by (up to an irrelevant additive constant):

$$\log p(x|\Phi) = \sum_{j=1}^{o} \frac{(x_j - \sum_{i=1}^{n} a_i \exp(-t_j/t'_i))^2}{2\sigma_j^2} \tag{29}$$

And the amplitudes are estimated as the maximum of this function, yielding the discretized maximum-likelihood fingerprint $[(t'_1, a_1), ..., (t'_n, a_n)]$. As an example, we consider a hypothetical measurement of a

correlation function of the form

$$y(t) = 0.9 \exp\left(\frac{t}{50}\right) + 0.1 \exp\left(\frac{t}{250}\right) \tag{30}$$

with additive Gaussian error having intensities of $\sigma = 0.5/\sqrt{t}$. Fig. 7c shows the curve of Eq. (30) along with the measured correlation function, while Fig 7a (black) shows the corresponding fingerprint. Figs 7a (red), b, and c(green) show the results of the fingerprint estimation procedure. The experimental fingerprint shown in Figs 7a is then used for the further analysis.

5.2 Simulation, Markov model, and simulated dynamical fingerprints

Molecular simulation methods are useful to generate structures that can be assigned to experimentally measurable dynamical processes. A popular choice are atomistic molecular dynamics models, but in some cases higher-order models (such as ab initio or QM/MM) or coarser methods (coarse-grained models or Go-type models) may be useful. Furthermore, a simulation setup should be chosen which is able to generate dynamical trajectories from some well-defined ensemble. At least, one expects a constant temperature and a unique stationary density (see [57, 78] for a discussion on ensembles and thermostats that have desirable statistical properties). Based on such a setup, dynamical trajectories can be generated. At this point, we assume that the setup and the computational environment has been chosen such that a "statistically sufficient" amount of trajectories can be generated. In situations where this is not possible, see [36, 79, 80, 81, 82] for a discussion of methods that can be used to enhance the sampling.

Given the simulation data, the molecular state space is discretized by clustering. Various combinations of distance metrics and clustering methods have been proposed. Frequently used metrics include Euclidean distance after having fitted the molecule to a reference structure [36, 5], root mean square distance (RMSD) [45, 59], and various clustering methods may be used [36, 45, 59, 73, 83]. Interestingly, very simple methods such as choosing generator structures by picking simulation frames at regular time intervals or even randomly and then clustering the data by assigning all simulation frames to the nearest generator structures perform quite well [57]. Importantly, the clustering must be fine enough such that the discretization is still allows the metastable states to be distinguished in order to be useful to build a quantitative Markov model.

After having discretized the simulation data to discrete trajectories, the transition matrix $\mathbf{T}(\tau)$ is estimated. The simplest method to do this is to generate a count matrix $\mathbf{C}(\tau)$ whose entries c_{ij} contain the number of times a simulation was found in state i in time t and in j at time $t + \tau$, and then calculating $T_{ij} = c_{ij}/\sum_k c_{ik}$. However, this matrix does not necessarily fulfill detailed balance, and thus the decomposition Eq. (7) does not have a simple interpretation. It is therefore desirable to estimate a matrix $\mathbf{T}(\tau)$ that fulfills detailed balance. Reversible counting [40] can be used if one has simulation trajectories that are much longer than the slowest relaxation time, otherwise one must use an estimation method [57] which allow a reversible $\mathbf{T}(\tau)$ to be estimated based on the unbiased count matrix $\mathbf{C}(\tau)$.

In order to analyze $\mathbf{T}(\tau)$, we perform an eigenvalue decomposition, generating eigenvectors $\mathbf{l}_i$ and eigenvalues λ_i. The eigenvectors can be used to identify metastable sets [60, 66, 5] that help to understand

the essential kinetics. The eigenvectors $\mathbf{l}_i$ can be investigated in order to obtain insight between which states the relaxation process with timescale $t_i = -\tau / \ln \lambda_i$ switches.

The fingerprint is calculated by calculating the amplitudes depending on the specific type of experiment considered (see Sec. 3.2 and 3.3) and combining them with the timescales t_i. Note that this fingerprint has statistical uncertainty based on the fact that only a finite number of dynamical trajectories has been used for the estimation of $\mathbf{T}(\tau)$. This uncertainty can be characterized based on Monte Carlo methods described in [61, 84, 27].

The assignment of structural processes to experimentally-detected dynamical features can be made if peaks can be matched between

Programs to calculate Markov models from simulation data are available in the simulation package package EMMA (e.g. https://simtk.org/home/emma)

5.3 Validation and experimental design

We have discussed and shown in Sec. 4 that for each given experimental setup (i.e. combination of measurement technique and observable chosen by the label placement), the amplitude of some processes may be large, and the amplitude of many others may be small. The small-amplitude processes can often not be detected with high reliability since they might affect the signal only to a degree that is similar to statistical or systematic error present in the measurement. It is thus desirable to *design* the experiment such that specific processes appear with large amplitudes. We sketch the following systematic approach of experimental design which has been proposed in [27]:

1. Conduct MD simulations of the molecular system under investigation and estimate a Markov model to model its essential kinetics

2. For each possible experimental setup (e.g. for each placement of the labels), estimate the values of the corresponding observables, $\mathbf{a}$, $\mathbf{b}$ and calculate the expected experimental fingerprints as described in Sec. 3.2 and 3.3.

3. For each of the m slowest relaxation processes, select the experimental setup for which the amplitude of this relaxation process is largest (or largest compared to the amplitudes of the processes with similar timescales if the timescale spectrum is dense)

4. Conduct these m experiments.

This approach attempts to optimally probe each process with a single experiment, thus also keeping the number of potentially expensive experiments small. Besides yielding a useful set of complementary experiments, this approach is useful to validate the simulated results much more solidly than with a single comparison.

This approach is ideally suited for experiments with site-specific labels that do not significantly affect the kinetics. This is especially true for techniques that permit isotope labeling such as NMR, IR spectroscopy or neutron scattering. In fluorescence-based techniques this can be achieved with intrinsic dyes

(e.g. the modulation of Tryptophan fluorescence by the environment [38] or Tryptophan triplet quenching by Cysteine [85]) or with extrinsic dyes that have little effect on the conformational dynamics.

In [27], the method has been demonstrated on the MR121-GS_9-W peptide with a simple heuristic to predict fluorescence signals for each of 190 possible positions of the MR121 and W dyes along the chain. Based on this, the amplitudes of the five slowest fingerprint peaks were calculated and are shown in Fig. 9. It is apparent that for most experiments only one or two amplitudes are strong while the remaining amplitudes are weak. If this result is also true for other molecules, it is evident why so many molecules appear to have two- or three-state kinetics while they are much more complex in molecular simulations.

Based on such a comparison of predicted fingerprints, experiments can be suggested. The colored boxes in Fig. 9 highlight five experiments that are predicted to maximally probe each of the slowest relaxation processes relative to the total amplitude of the five slowest processes.

6 Conclusions

The combination of Markov models and the concept of dynamical fingerprints provides a theoretically solid and computationally feasible approach to connect molecular simulation data or molecular kinetic models to experiments that probe the kinetics of the molecular system in reality. The main advantage of this approach over traditional MD analyses is that the processes that occur at given timescales are unambiguously given by the theory. In the Markov model, this assignment is present by the one-to-one association of transition matrix eigenvalues (that correspond to measurable relaxation timescales) and eigenvectors (that describe structural changes). When the experimentally-measured relaxation data is further subjected to a spectral analysis, experiment and simulation can be reconciled on the basis of dynamical fingerprints, i.e. by matching peaks of the timescale density.

A comment is in order on the fact that in all cases, the slow relaxations in kinetic measurements are found to have the form of a sum of single exponential term, each term corresponding to an eigenvalue / eigenvector pair in our analysis. This is a general result which can also be obtained by performing the analysis in full continuous state space (as opposed to our discrete-state treatment here). The only assumptions that are made to arrive at this result are the following:

1. The dynamics of the system is Markovian in full state space (i.e. the continuous space of all positions and momenta of the molecular systems studied and the solvent molecules). This is a very weak assumption that is made in all classical simulation models. The Markovian assumption could also be applied to quantum mechanical models when the electronic degrees of freedom are included. It is thus also a reasonable assumption for real molecular systems. The only systems for which such an assumption would be unpractical are systems which have correlations over arbitrarily long lengthscales, such that no finite-size simulation setup can be made that captures all relevant processes. This can happen for glassy or crystalline systems.

2. The state space is ergodic, i.e. all states of the system can interchange. This assumption may also be untrue for glassy or crystalline systems. It is in practice also hard to fulfill for other systems if the kinetics are slow and are not measured in an ensemble but by averaging multiple single-molecule trajectories. In this case it may be difficult to collect sufficiently many trajectories that this trajectory set is effectively ergodic, and deviations from multiexponentiality may be a statistical artefact.

3. The relaxations are measured at equilibrium conditions. This does include the possibility that the system relaxes from an off-equilibrium distribution (e.g. as in temperature jump experiments), but it does so under equilibrium dynamics which fulfill detailed balance. This assumption requires that the experiment does not put energy into the system or remove energy from it. It is unclear whether laser or scattering experiments obey this condition sufficiently well.

Even in situations where these points can be assumed to be fulfilled, apparent nonexponentiality has been found over significantly long timescales, such as stretched exponentials [86, 87] or power laws [48]. Note that this is no contradiction because such apparent nonexponentialities can be easily explained by sums

of a few single exponential relaxations with particular spacings of timescales and amplitudes [88, 89, 27] - and thus also correspond to dynamical fingerprints with multiple peaks (see [27], Supplementary Fig. 1 and 2). In practice, however, care must be taken that such effects are not actually due to the measurement technique itself. Especially conditions 2 and 3 may sometimes be violated by the experimental setup itself.

It is likely that the apparent two- or three-state kinetics observed in experiments of macromolecules does not reflect the entire complexity of their conformational dynamics. In particular, the slowest measured rate is not necessarily the folding rate because (*i*) there might be no process which corresponds to our notion of folding, (*ii*) the experiment might be insensitive to this particular process. The comparison to a Markov model allows for a unambiguous interpretation of the measured fingerprints.

7 Acknowledgments

Funding from the German Science Foundation (DFG) through grant number NO 825/2 and through research center MATHEON is gratefully acknowledged.

References

[1] S. Fischer, B. Windshuegel, D. Horak, K. C. Holmes and J. C. Smith. *Proc. Natl. Acad. Sci. USA* **102** (2005), pp. 6873.

[2] P. Imhof, S. Fischer and J. C. Smith. *Biochemistry* **48** (2009), pp. 9061.

[3] A. H. Ratje, J. Loerke, A. Mikolajka, M. Brunner, P. W. Hildebrand, A. L. Starosta, A. Donhofer, S. R. Connell, P. Fucini, T. Mielke, P. C. Whitford, J. N. Onuchic, Y. Yu, K. Y. Sanbonmatsu, R. K. Hartmann, P. A. Penczek, D. N. Wilson and C. M. T. Spahn. *Nature* **468** (2010), pp. 713.

[4] J. Nguyen, M. A. Baldwin, F. E. Cohen and S. B. Prusiner. *Biochemistry* **34** (1995), pp. 4186.

[5] F. Noé, I. Horenko, C. Schütte and J. C. Smith. *J. Chem. Phys.* **126** (2007), p. 155102.

[6] B. Schuler, E. A. Lipman and W. A. Eaton. *Nature* **419** (2002), pp. 743.

[7] J. Ross, P. Buschkamp, D. Fetting, A. Donnermeyer, C. M. Roth and P. Tinnefeld. *J. Phys. Chem. B* **111** (2007), pp. 321.

[8] Y. Santoso, C. M. Joyce, O. Potapova, L. Le Reste, J. Hohlbein, J. P. Torella, N. D. F. Grindley and A. N. Kapanidis. *Proc. Natl. Acad. Sci. USA* **107** (2010), pp. 715.

[9] A. Y. Kobitski, A. Nierth, M. Helm, A. Jäschke and G. U. Nienhaus. *Nucleic Acids Res.* **35** (2007), pp. 2047.

[10] W. J. Greenleaf, M. T. Woodside and S. M. Block. *Annual review of biophysics and biomolecular structure* **36** (2007), pp. 171.

[11] J. Cellitti, R. Bernstein and S. Marqusee. *Protein Science* **16** (2007), pp. 852.

[12] M. Rief, M. Gautel, F. Oesterhelt, J. M. Fernandez and H. E. Gaub. *Science* **276** (1997), pp. 1109.

[13] Gebhardt, T. Bornschlögl and M. Rief. *Proc. Natl. Acad. Sci. USA* **107** (2010), pp. 2013.

[14] H. Wu and F. Noé. *Phys. Rev. E, published online* (2011).

[15] I. V. Gopich and A. Szabo. *J. Phys. Chem. B* **113** (2009), pp. 10965.

[16] H. Wu and F. Noé. *Multiscale Model. Simul.* **8** (2010), p. 1838.

[17] I. V. Gopich, D. Nettels, B. Schuler and A. Szabo. *J. Chem. Phys.* **131** (2009), p. 095102.

[18] M. Jäger, Y. Zhang, J. Bieschke, H. Nguyen, M. Dendle, M. E. Bowman, J. P. Noel, M. Gruebele and J. W. Kelly. *Proc. Natl. Acad. Sci. USA* **103** (2006), pp. 10648.

[19] M. Sadqi, L. J. Lapidus and V. Munoz. *Proc. Natl. Acad. Sci. USA* **100** (2003), pp. 12117.

[20] C. Dumont, T. Emilsson and M. Gruebele. *Nature Meth.* **6** (2009), pp. 515.

[21] C.-K. Chan, Y. Hu, S. Takahashi, D. L. Rousseau, W. A. Eaton and J. Hofrichter. *Proc. Natl. Acad. Sci. USA* **94** (1997), pp. 1779.

[22] A. Volkmer. *Biophys. J.* **78** (2000), pp. 1589.

[23] I. Schlichting, S. C. Almo, G. Rapp, K. Wilson, K. Petratos, A. Lentfer, A. Wittinghofer, W. Kabsch, E. F. Pai, G. A. Petsko and R. S. Goody. *Nature* **345** (1990), pp. 309.

[24] J. Buck, B. Fürtig, J. Noeske, J. Wöhnert and H. Schwalbe. *Proc. Natl. Acad. Sci. USA* **104** (2007), pp. 15699.

[25] T. Kiefhaber. *Proc. Nat. Acad. Sci. USA* **92** (1995), pp. 9029.

[26] W. Doster, S. Cusack and W. Petry. *Nature* **337** (1989), pp. 754.

[27] F. Noé, S. Doose, I. Daidone, M. Löllmann, J. Chodera, M. Sauer and J. Smith. *Proc. Natl. Acad. Sci. USA, in press* (2011).

[28] L. J. Lapidus, W. A. Eaton and J. Hofrichter. *Proc. Natl. Acad. Sci. USA* **97** (2000), pp. 7220.

[29] H. Neuweiler, M. Löllmann, S. Doose and M. Sauer. *J. Mol. Biol.* **365** (2007), pp. 856.

[30] X. Michalet, S. Weiss and M. Jäger. *Chem. Rev.* **106** (2006), pp. 1785.

[31] P. Tinnefeld and M. Sauer. *Angew. Chem. Intl. Ed.* **44** (2005), pp. 2642.

[32] R. R. Hudgins, F. Huang, G. Gramlich and W. M. Nau. *J. Am. Chem. Soc.* **124** (2002), pp. 556.

[33] H. D. Kim, G. U. Nienhaus, T. Ha, J. W. Orr, J. R. Williamson and S. Chu. *Procl. Natl. Acad. Sci. USA* **99** (2002), pp. 4284.

[34] D. Nettels, A. Hoffmann and B. Schuler. *J. Phys. Chem. B* **112** (2008), pp. 6137.

[35] D. D. Schaeffer, A. Fersht and V. Daggett. *Curr. Opin. Struct. Biol.* **18** (2008), pp. 4.

[36] F. Noé, C. Schütte, E. Vanden-Eijnden, L. Reich and T. R. Weikl. *Proc. Natl. Acad. Sci. USA* **106** (2009), pp. 19011.

[37] W. van Gunsteren, J. Dolenc and A. Mark. *Curr. Opin. Struct. Biol.* **18** (2008), pp. 149.

[38] M. Jäger, H. Nguyen, J. C. Crane, J. W. Kelly and M. Gruebele. *J. Mol. Biol.* **311** (2001), pp. 373.

[39] O. Bieri, J. Wirz, B. Hellrung, M. Schutkowski, M. Drewello and T. Kiefhaber. *Proc. Natl. Acad. Sci. USA* **96** (1999), pp. 9597.

[40] S. Muff and A. Caflisch. *Proteins* **70** (2007), pp. 1185.

[41] D. L. Ensign, P. M. Kasson and V. S. Pande. *J. Mol. Biol.* **374** (2007), pp. 806.

[42] J. N. Onuchic and P. G. Wolynes. *Curr. Opin. Struc. Biol.* **14** (2004), pp. 70.

[43] H. Frauenfelder, G. Chen, J. Berendzen, P. W. Fenimore, H. Jansson, B. H. McMahon, I. R. Stroe, J. Swenson and R. D. Young. *Proc Natl. Acad. Sci. USA* **106** (2009), pp. 5129.

[44] F. Noé and S. Fischer. *Curr. Opin. Struc. Biol.* **18** (2008), pp. 154.

[45] G. R. Bowman, K. A. Beauchamp, G. Boxer and V. S. Pande. *J. Chem. Phys.* **131** (2009), p. 124101.

[46] A. Gansen, A. Valeri, F. Hauger, S. Felekyan, S. Kalinin, K. Tóth, J. Langowski and C. A. M. Seidel. *Proc. Natl. Acad. Sci. USA* **106** (2009), pp. 15308.

[47] H. Neubauer, N. Gaiko, S. Berger, J. Schaffer, C. Eggeling, J. Tuma, L. Verdier, C. A. Seidel, C. Griesinger and A. Volkmer. *J. Am. Chem. Soc.* **129** (2007), pp. 12746.

[48] W. Min, G. Luo, B. J. Cherayil, S. C. Kou and X. S. Xie. *Phys. Rev. Lett.* **94** (2005), p. 198302.

[49] E. Z. Eisenmesser, O. Millet, W. Labeikovsky, D. M. Korzhnev, M. Wolf-Watz, D. A. Bosco, J. J. Skalicky, L. E. Kay and D. Kern. *Nature* **438** (2005), pp. 117.

[50] B. G. Wensley, S. Batey, F. A. C. Bone, Z. M. Chan, N. R. Tumelty, A. Steward, L. G. Kwa, A. Borgia and J. Clarke. *Nature* **463** (2010), pp. 685.

[51] B. P. English, W. Min, A. M. van Oijen, K. T. Lee, G. B. Luo, H. Y. Sun, B. J. Cherayil, S. C. Kou and X. S. Xie. *Nature Chemical Biology* **2** (2006), pp. 87.

[52] M. O. Lindberg and M. Oliveberg. *Curr. Opin. Struct. Biol.* **17** (2007), pp. 21.

[53] K. Sridevi. *J. Mol. Biol.* **302** (2000), pp. 479.

[54] R. A. Goldbeck, Y. G. Thomas, E. Chen, R. M. Esquerra and D. S. Kliger. *Proc. Natl. Acad. Sci. USA* **96** (1999), pp. 2782.

[55] A. Matagne, S. E. Radford and C. M. Dobson. *J. Mol. Biol.* **267** (1997), pp. 1068.

[56] C. C. Mello and D. Barrick. *Proc. Natl. Acad. Sci. USA* **101** (2004), pp. 14102.

[57] J.-H. Prinz, H. Wu, M. Sarich, B. Keller, M. Fischbach, M. Held, J. Chodera, C. Schütte and F. Noé. *J. Chem. Phys.* (2011).

[58] W. C. Swope, J. W. Pitera, F. Suits, M. Pitman and M. Eleftheriou. *J. Phys. Chem. B* **108** (2004), pp. 6582.

[59] J. D. Chodera, K. A. Dill, N. Singhal, V. S. Pande, W. C. Swope and J. W. Pitera. *J. Chem. Phys.* **126** (2007), p. 155101.

[60] C. Schütte, A. Fischer, W. Huisinga and P. Deuflhard. *J. Comput. Phys.* **151** (1999), pp. 146.

[61] F. Noé. *J. Chem. Phys.* **128** (2008), p. 244103.

[62] V. S. Pande, K. Beauchamp and G. R. Bowman. *Methods* **52** (2010), pp. 99.

[63] N. V. Buchete and G. Hummer. *J. Phys. Chem. B* **112** (2008), pp. 6057.

[64] B. Keller, P. Hünenberger and W. van Gunsteren. *J. Chem. Theo. Comput., published online* (2011).

[65] S. Kube and M. Weber. *J. Chem. Phys.* **126** (2007), p. 024103.

[66] M. Weber. *ZIB Report* **03-04** (2003).

[67] C. Schütte, F. Noé, E. Meerbach, P. Metzner and C. Hartmann. In R. Jeltsch and G. W. (Eds), eds., *Proceedings of the International Congress on Industrial and Applied Mathematics (ICIAM)*. EMS publishing house, pp. 297–336.

[68] C. Schütte and W. Huisinga. In P. G. Ciaret and J. L. Lions, eds., *Handbook of Numerical Analysis*, volume X: Computational Chemistry. North-Holland (2003) pp. 699–744.

[69] M. Sarich, F. Noé and C. Schütte. *SIAM Multiscale Model. Simul.* **8** (2010), pp. 1154.

[70] N. G. van Kampen. *Stochastic Processes in Physics and Chemistry.* Elsevier Science Publishers B.V., 2 edition (1992).

[71] D. Nerukh, C. H. Jensen and R. C. Glen. *J. Chem. Phys.* **132** (2010), p. 084104.

[72] P. Deuflhard and H. Andreas. *Numerical Analysis in Modern Scientific Computing*, volume 43 of *Texts in Applied Mathematics.* Springer Berlin Heidelberg, 2 edition (2003).

[73] B. Keller, X. Daura and W. F. van Gunsteren. *J. Chem. Phys.* **132** (2010), p. 074110.

[74] D. Klostermeier, P. Sears, C. H. Wong, D. P. Millar and J. R. Williamson. *Nucleic Acids Research* **32** (2004), pp. 2707.

[75] E. Sisamakis, A. Valeri, S. Kalinin, P. J. Rothwell and C. Seidel. *Methods in Enzymology* **475** (2010), pp. 455.

[76] S. W. Provencher. *Comput. Phys. Commun.* **27** (1982), pp. 229.

[77] P. Steinbach. *Biophys. J.* **82** (2002), pp. 2244.

[78] J. D. Chodera, W. C. Swope, F. Noé, J.-H. Prinz and V. S. Pande. *J. Phys. Chem., submitted* (2010).

[79] N. Singhal and V. S. Pande. *J. Chem. Phys.* **123** (2005), p. 204909.

[80] C. Micheletti, G. Bussi and A. Laio. *J. Chem. Phys.* **129** (2008), p. 074105.

[81] A. Laio and M. Parrinello. *Proc Natl. Acad. Sci. USA* **99** (2002), p. 12562.

[82] G. R. Bowman, D. L. Ensign and V. S. Pande. *J. Chem. Theory Comput.* **6** (2010), pp. 787.

[83] Y. Yao, J. Sun, X. Huang, G. R. Bowman, G. Singh, M. Lesnick, L. J. Guibas, V. S. Pande and G. Carlsson. *J. Chem. Phys.* **130** (2009), p. 144115.

[84] J. D. Chodera and F. Noé. *J. Chem. Phys.* **133** (2010), p. 105102.

[85] L. J. Lapidus, W. A. Eaton and J. Hofrichter. *Phys. Rev. Lett.* **87** (2001), p. 258101.

[86] J. Klafter and M. F. Shlesinger. *Proc. Natl. Acad. Sci. USA* **83** (1986), pp. 848.

[87] R. Metzler, J. Klafter, J. Jortner and M. Volk. *Chem. Phys. Lett.* **293** (1998), pp. 477.

[88] S. J. Hagen and W. A. Eaton. *J. Chem. Phys.* **104** (1996), pp. 3395.

[89] J. B. Witkoskie and J. Cao. *J. Chem. Phys.* **121** (2004), pp. 6361.

8 Tables

	equlibrium experiment	relaxation experiment
relaxation experiment	-	$\gamma_i^{\text{relax}} = \langle \mathbf{a}, \mathbf{l}_i \rangle \left\langle \mathbf{p'}^{\top}(0),\, \mathbf{l}_i \right\rangle$
autocorrelation	$\gamma_i^{\text{eq, auto-cor}} = \langle \mathbf{a}, \mathbf{l}_i \rangle^2$	$\gamma_i^{\text{jump, auto-cor}} = \langle \mathbf{a}, \mathbf{P'}(0)\mathbf{l}_i \rangle \langle \mathbf{a}, \mathbf{l}_i \rangle$
cross-correlation	$\gamma_i^{\text{eq, cross-cor}} = \langle \mathbf{a}, \mathbf{l}_i \rangle \langle \mathbf{b}, \mathbf{l}_i \rangle$	$\gamma_i^{\text{jump, cross-cor}} = \langle \mathbf{a}, \mathbf{P'}(0)\mathbf{l}_i \rangle \langle \mathbf{b}, \mathbf{l}_i \rangle$

Table 1: Overview of the expressions for the amplitudes in correlation experiments.

$\langle \mathbf{a}_k, \mathbf{l}_i \rangle$	$\mathbf{a}$	$\mathbf{a}_2$	$\mathbf{a}$	$\langle \mathbf{p}(0), \mathbf{l}_i \rangle$	$\mathbf{p}_1(0)$	$\mathbf{p}_2(0)$
$\mathbf{l}_1$	1.45	1.39	1.08	$\mathbf{l}_1$	0.54	0.54
$\mathbf{l}_2$	0.78	0.15	0.47	$\mathbf{l}_2$	0.29	0.70
$\mathbf{l}_3$	0.01	0.59	0.57	$\mathbf{l}_3$	0.50	0.06
$\mathbf{l}_4$	0.12	0.66	0.26	$\mathbf{l}_4$	0.00	1.09

Table 2: Protein folding model: scalar products of the observable vectors and the initial distribution with the left eigenvectors.

9 Figures

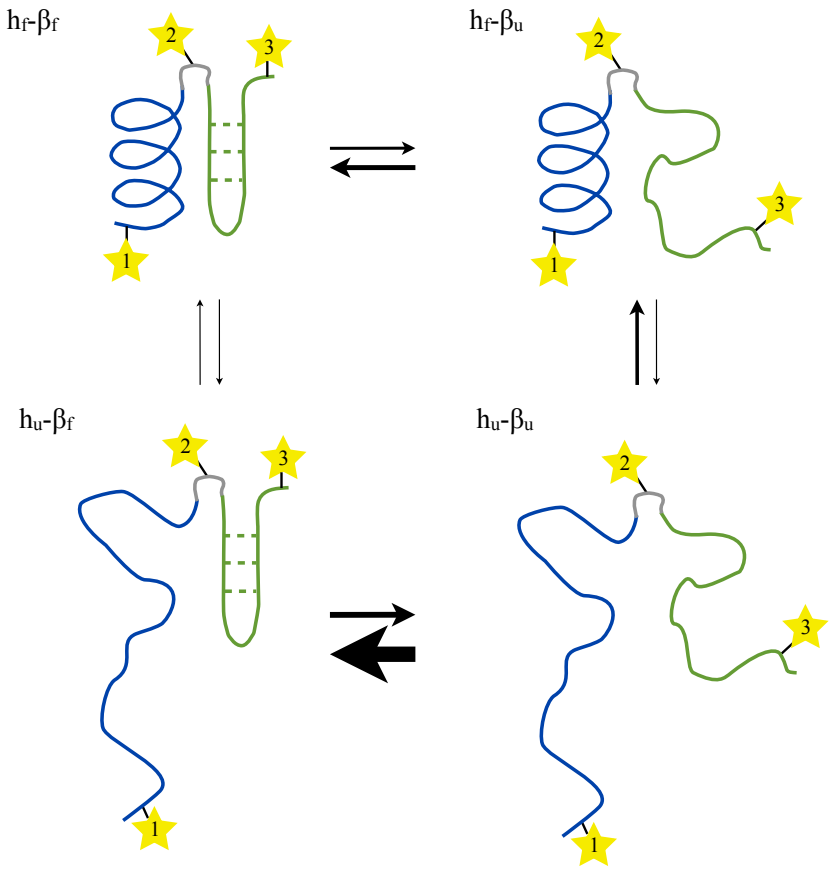


Figure 1: **Sketch of a protein folding equilibrium.** The arrows represent possible transitions between confromational states. Their thickness corrersponds to the transition probability. The yellow stars represent possible chromophore attachment points.

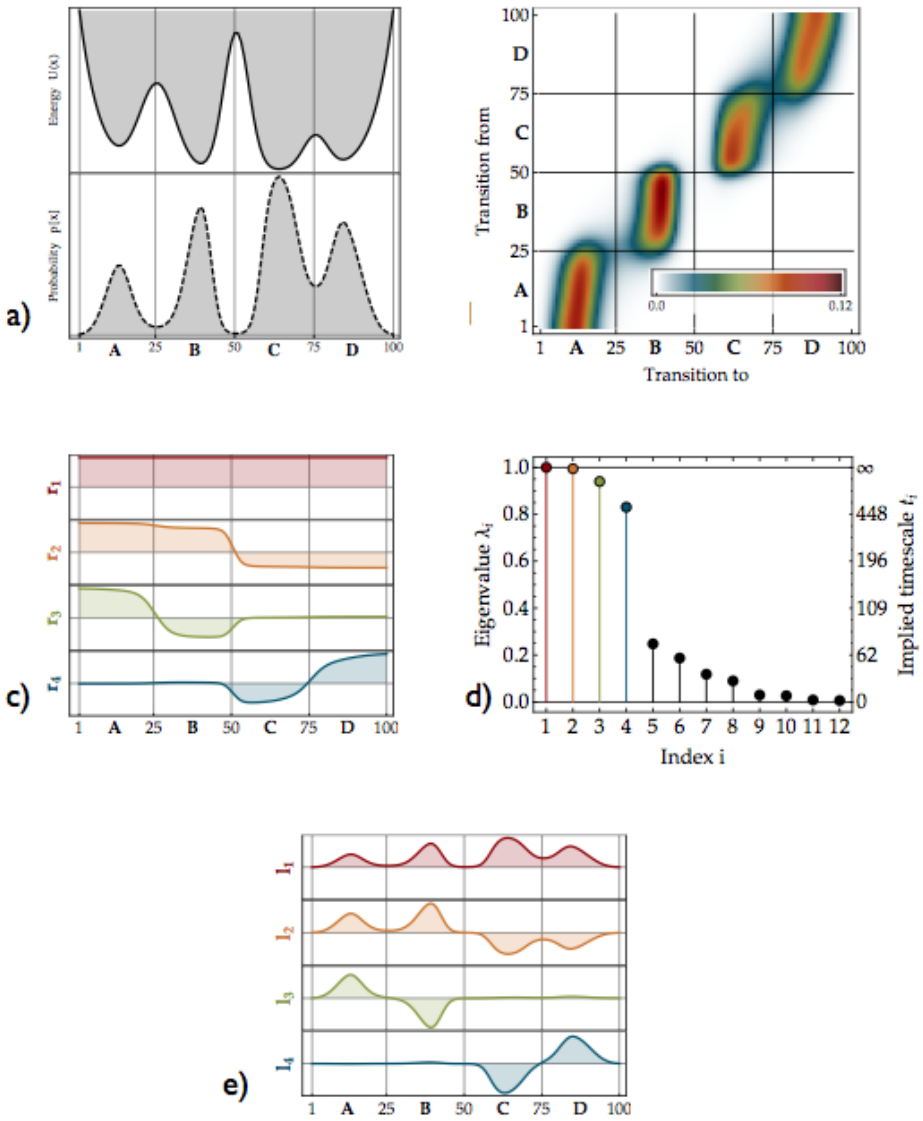


Figure 2: **Markov model of a dynamics in a 1-D energy surface.** (a) Potential energy function with four metastable states and corresponding equilibrium distribution π. (b) Plot of the transition matrix $\mathbf{T}(\tau)$ for a diffusive dynamics in this potential. $\mathbf{T}(\tau)$ is defined on a states space Ω of 100 equisized bins along the reaction coordinate. Black and orange indicate high transition probability, white zero transition probability. (c) The four dominant right eigenvectors $\mathbf{r}_i$. (d) Eigenvalue spectrum of $\mathbf{T}(\tau)$. (e) The four dominant left eigenvectors $\mathbf{l}_i$.

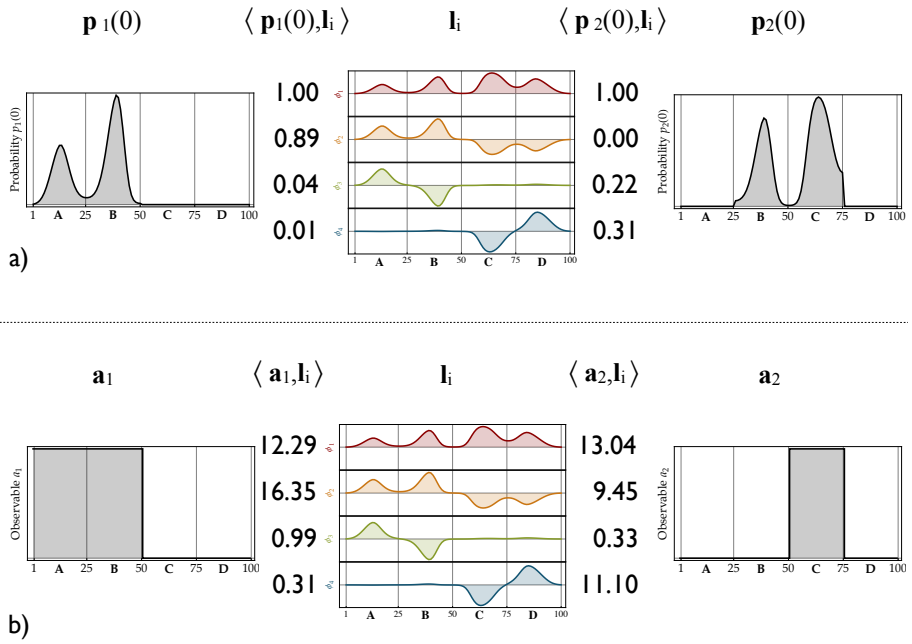


Figure 3: **Experimental setups for the 1-D energy surface model.** The middle columns shows the left eigenvectors of the model. Panel a additionnally shows two possible initial distributions, and panel c shows two possible observables. The values of the respective scalar productions are shown to the left and right of the eigenvector plots.

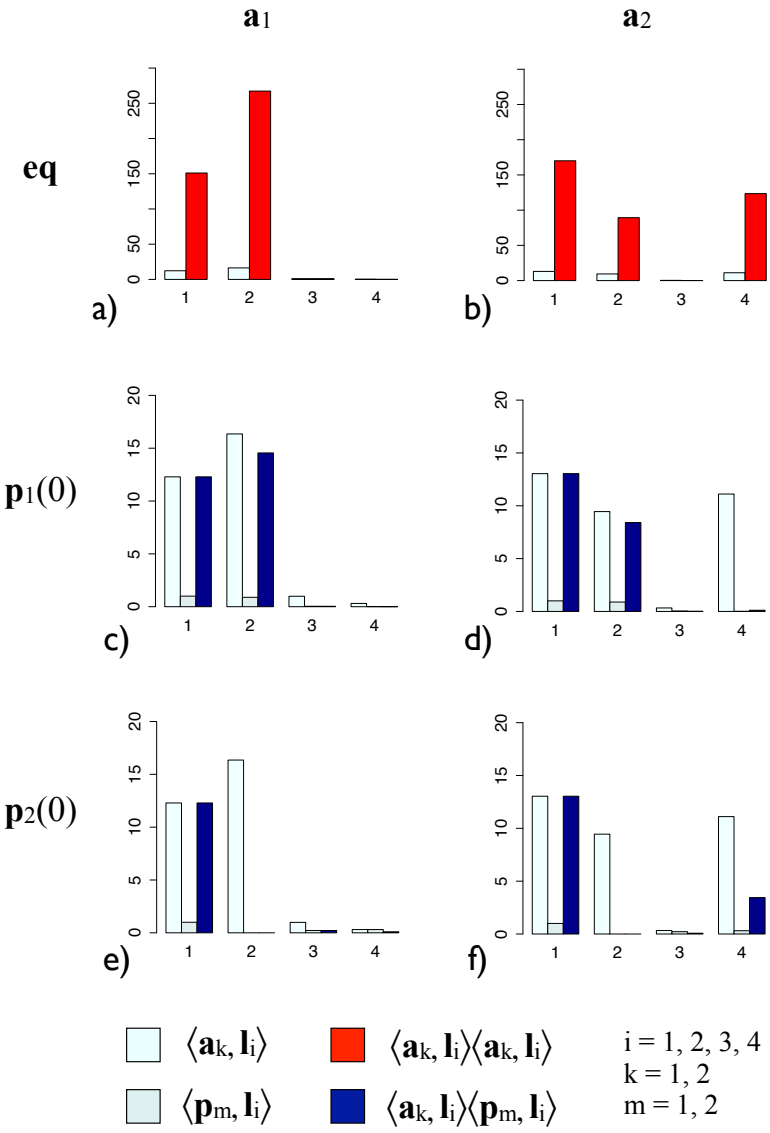


Figure 4: **Amplitdudes for the 1-D energy surface model.** Equlibrium experiments: (a) observable $\mathbf{a}_1$, (b)observable $\mathbf{a}_2$. Relaxation experiments: (c) $\mathbf{p}_1(0)$ and $\mathbf{a}_1$, combined (d)$\mathbf{p}_1(0)$ and $\mathbf{a}_2$ combined, (e) $\mathbf{p}_2(0)$ and $\mathbf{a}_1$ combined, (f) $\mathbf{p}_2(0)$ and $\mathbf{a}_2$ combined.

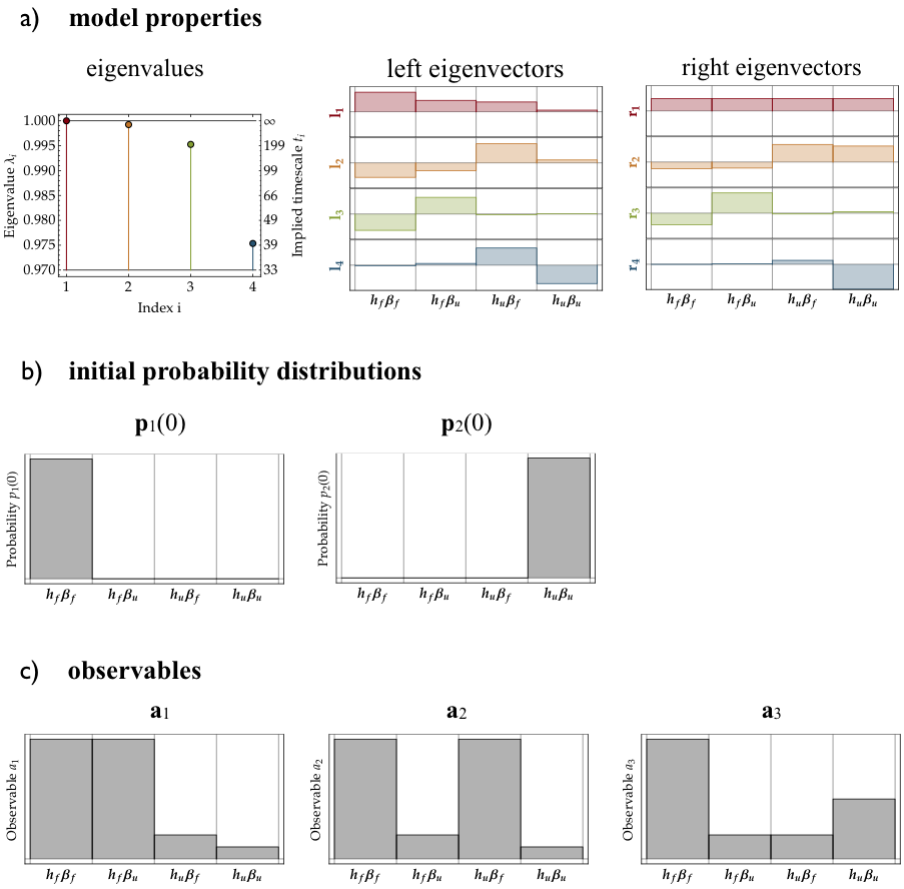


Figure 5: **Markov model and experimental setup for the protein folding model.**

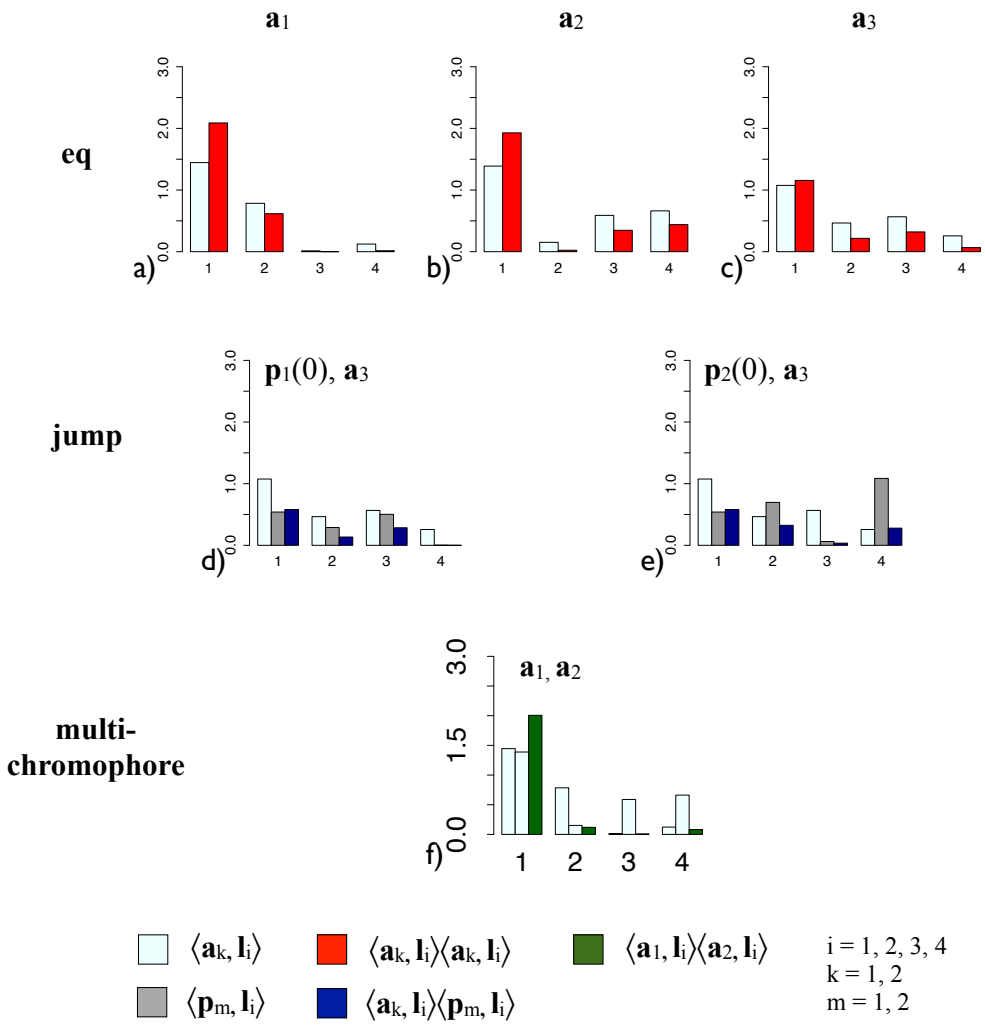


Figure 6: **Amplitudes for the protein folding model.** Equilibrium experiments: (a) observable $\mathbf{a}_1$, (b) observable $\mathbf{a}_1$, observable $\mathbf{a}_1$,

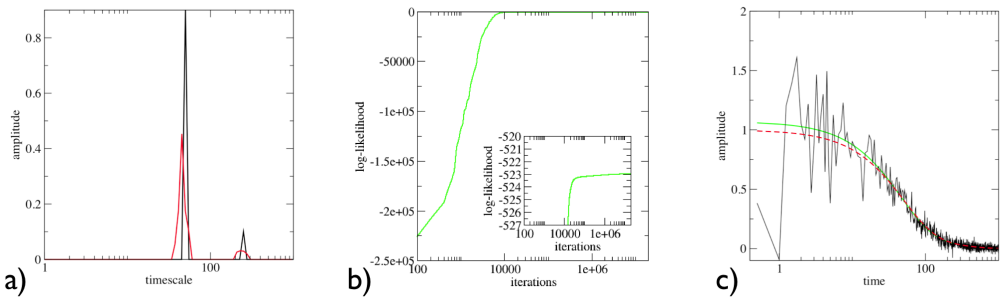


Figure 7: **Dynamical fingerprint of a model correlation function.** (a) True (black) and estimated fingerprint (red). Note that the apparent disagreement in amplitude is a result of the broadening in the estimated fingerprint which is a consequence of the noise in the data. The areas under the peaks should be the same for a correctly estimated fingerprint.
(b) Likelihood which is printed every 100 iterations to the log-file. You should inspect this likelihood and make sure that it is converged. A good rule of thumb is that it should not increase more than 1 likelihood unit within the last half of the optimization. The inset shows that this is the case here
(c) Comparison of the input (black) with the predicted relaxation curve (green). The predicted curve is a good fit to the data. The deviation at short times from the true, noiseless, signal (red) are due to statistical noise in the data.

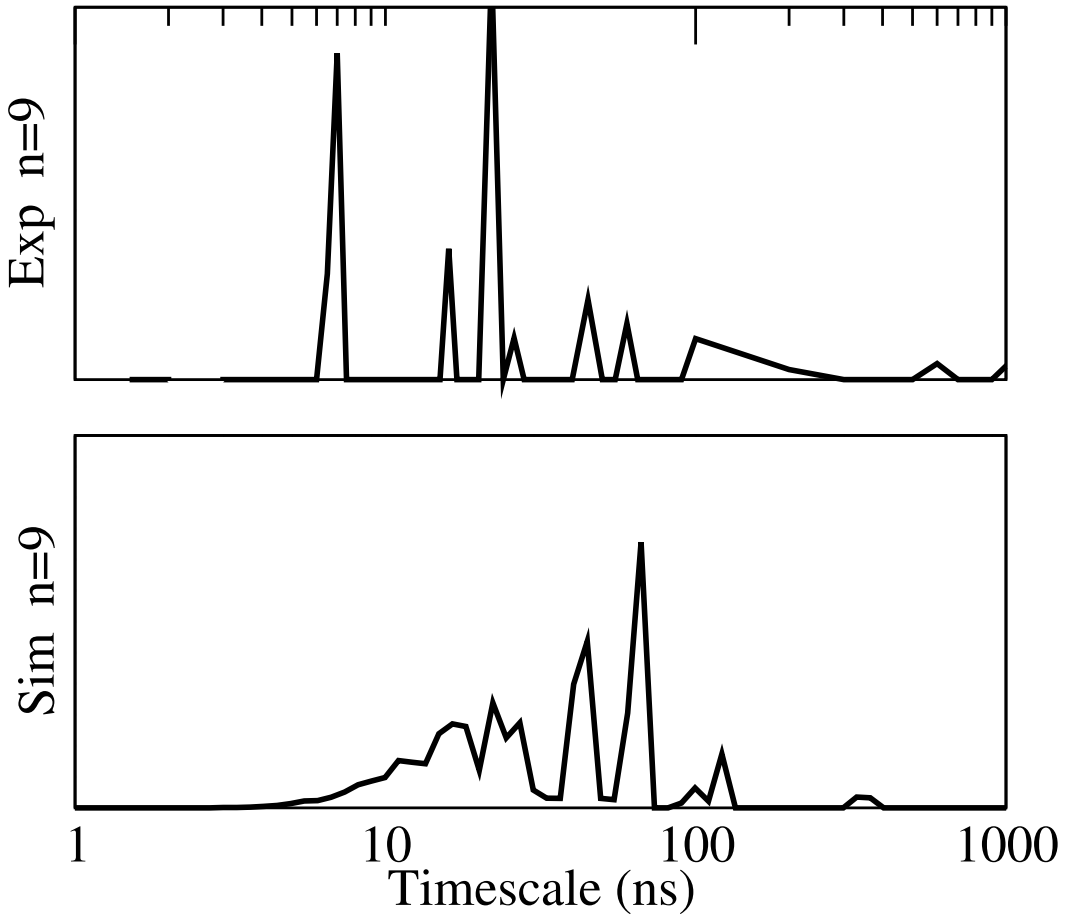


Figure 8: **Dynamical fingerprint of the MR121-GS$_9$-W peptide.** Upper panel: from experiment. Lower panel: from simulation. $n = 9$ is the chain length of the peptide.

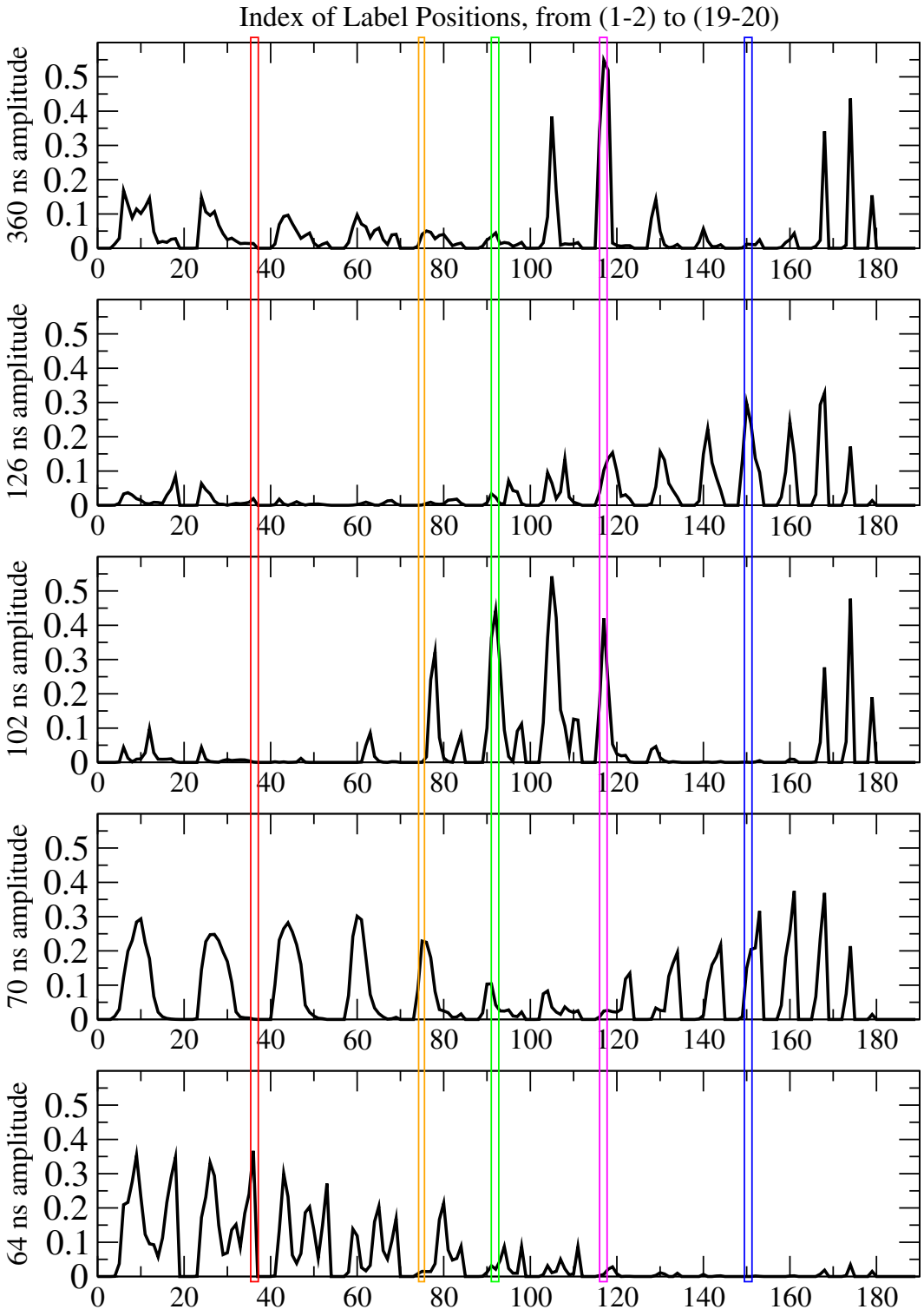


Figure 9: **Experimental design.** Prediction of the amplitudes of fingerprint peaks of the 5 slowest processes in MR121-GS_9-W when placing the fluorescence labels at any of the 190 different possible residue positions from 1-2 to 19-20. The x-Axis enumerates these 190 labeling positions. The magenta, blue, green, orange, red lines mark the proposed experimental setups to optimally probe the slowest to the fifth-slowest processes.