# Probability distributions of molecular observables computed from Markov models. II. Uncertainties in observables and their time-evolution

John D. Chodera[1,*] and Frank Noé[2,†]

[1]*Research Fellow, California Institute of Quantitative Biosciences (QB3), University of California, Berkeley, 260J Stanley Hall, Berkeley, California 94720, USA*
[2]*DFG Research Center Matheon, FU Berlin, Arnimallee 6, 14195 Berlin, Germany*

(Dated: October 23, 2009)

Discrete-state Markov (or master equation) models provide a useful simplified representation for characterizing the long-time statistical evolution of biomolecules in a manner that allows direct comparison with experiments as well as the elucidation of mechanistic pathways for an inherently stochastic process. A vital part of meaningful comparison with experiment is the characterization of the statistical uncertainty in the predicted experimental measurement, which may take the form of an equilibrium measurement of some spectroscopic signal, the time-evolution of this signal following a perturbation, or the observation of some statistic (such as the correlation function) of the equilibrium dynamics of a single molecule. Without meaningful error bars (which arise due to the finite quantity of data used to construct the model), there is no way to determine whether the deviations between model and experiment are statistically meaningful. Previous work has demonstrated that a Bayesian method that enforces microscopic reversibility can be used to characterize the correlated uncertainties in state-to-state transition probabilities (and functions thereof) for a model inferred from molecular simulation data. Here, we extend this approach to include the uncertainty in observables that are functions of molecular conformation (such as surrogate spectroscopic signals) characterizing each state, permitting the full statistical uncertainty in computed spectroscopic experiments to be assessed. We test the approach in a simple model system to demonstrate that the computed uncertainties provide a useful indictor of statistical variation, and then apply it to the computation of the fluorescence autocorrelation function measured for a dye-labeled peptide previously studied by both experiment and simulation.

Keywords: molecular dynamics, statistical error, Markov models, Bayesian error analysis

## I. INTRODUCTION

A large variety of biophysical experimental techniques are currently in use, providing information about various aspects of biomolecular structure and dynamics. Unfortunately, these methods often provide only indirect information about the microscopic origins of processes of interest. Spectroscopic methods, for example, are inherently limited by the need to collect a sufficient quantity of photons in order to distinguish relevant features of the system from statistical fluctuations of the measurement process. Assaying ensembles of molecules provides one solution to collecting sufficient photons while achieving high time resolution, at the cost of sacrificing information about heterogeneity, as only average signals are observed. Conversely, single-molecule experiments can observe statistical behavior of single molecules, but at the cost of integrating over times that may be longer than the timescales involved in processes of interest.

Molecular simulation is a powerful complementary tool for probing molecular behavior in biology. With a forcefield sufficiently representative of physical reality, it is possible to probe the stochastic behavior of individual molecules and the associated dynamical processes in atomic detail with high time resolution, filling in the gaps between experiment and intuition. In practice, however, this requires that practitioners must both validate the simulation against experimental measurements and collect sufficient data to provide insight into the mechanisms of interest. This process is frustrated by the statistical heterogeneity of dynamics (requiring many realizations of some process be observed) and the long timescales involved in processes of interest (which may greatly exceed our capacity to simulate on all but the largest computers or simplest problems).

The presence of metastable conformational states [1, 2], often the root cause of slow relaxation times and heterogeneous dynamics in many biomolecular systems, also provides a solution to this problem. A separation of timescales between fast relaxation within a metastable state and long waiting times between interstate transitions allows the statistical behavior of the system to be well-described by a discrete-state Markov model (or master equation model) over times longer than the fast mixing time within states [3]. Even if no strong separation of timescales is present, an appropriately constructed Markov model can still provide an excellent approximation to the long-time statistical dynamics of the system [4]. Because only transition rates between pairs of dynamically connected states need be estimated, these models can be efficiently constructed using molecular simulation data, where the dynamical

*Electronic Address: jchodera@berkeley.edu
†Electronic Address: frank.noe@fu-berlin.de

trajectories employed need only be longer than the lag time required for a Markov model to well approximate the original continuous dynamics (here referred to as the *Markov time* $\tau_{\mathrm{eq}}$) [5–8]. In strongly metastable systems, this Markov time may be orders of magnitude shorter than the slowest relaxation times of the system, thus making Markov models an efficient tool to model slow processes with a large but usually feasible number of short simulations, some or all of which may be generated in parallel. Methods for the construction [7–9] of such models and determination of an appropriate Markov time [4–7, 10–12] are described in detail elsewhere.

Because these models are constructed from a finite quantity of simulation data, there will be a degree of statistical uncertainty in the model parameters and in the characterization of each discrete state. This uncertainty will consequently induce uncertainties in quantities computed from the model, often in a complex way. Accurate characterization of these uncertainties is paramount to either comparison with existing experimental data (to validate the model) or the computation of new, unobserved properties (to predict unknown quantities). It is impossible to know if the model and experimental data are significantly discrepant, for example, without some estimate of both the uncertainty in the computed quantities arising from model uncertainty and the error in the experimental measurements.

Previously, a Bayesian framework was proposed for characterizing how finite quantities of simulation data produces uncertainties in the transition probabilities for a Markov model of physical systems satisfying detailed balance [13]. Here, we extend this framework to include the uncertainty in observables (such as spectroscopically-observable quantities) for each of these states, allowing the full uncertainty in equilibrium and time-resolved experimentally observable quantities computed from the model to be characterized.

This article is organized as follows. Section II reviews the statistical mechanics of common measurements and their computation from Markov models. Section III briefly describes the Bayesian inference framework for sampling transition matrices that satisfy detailed balance. Section IV introduces the proposed approach for incorporating errors in estimated observables into this framework. Section V validates the method on an analytically tractable model system, and Section VI applies it to a previously-studied fluorescent peptide for which experimental spectroscopic data is available.

## II. OBSERVABLES AND MARKOV MODELS

Biophysical spectroscopic measurements, though varied in terms of the quantity being measured, commonly fall into one of three categories:

**Equilibrium measurements** (e.g. equilibrium circular dichroism, infrared spectroscopy, and fluorescence measurements) represent ensemble averages (or *expectations*) of a conformation-sensitive spectroscopic signal $A(x)$ over some equilibrium configurational probability density $\pi(x)$, where $x$ denotes the instantaneous molecular configuration:

$$\langle A \rangle \;=\; \int dx\, \pi(x)\, A(x), \tag{1}$$

**Nonequilibrium relaxation experiments** (e.g. laser-induced temperature-jump experiments) monitor the relaxation of an ensemble-averaged spectroscopic signal following a perturbation, and can be described in terms of the initial density $\rho(x;0)$ and the time-evolution operator $\rho(x_t, t; x_0, 0)$ that describes the probability of finding a molecule in configuration $x_t$ at time $t$ given that it was initially in configuration $x_0$ at time $0$

$$\langle A(t) \rangle_{\rho_0} \;=\; \int dx_0\, \rho_0(x_0) \int dx_t\, \rho(x_t, t; x_0, 0)\, A(x_t) \tag{2}$$

**Equilibrium correlation measurements** (e.g. fluorescence correlation spectroscopy, spectral density functions) probe the auto- or cross-correlation function of some experimental observables $A$ and $B$ at equilibrium

$$\langle A(0)\, B(t) \rangle \;=\; \int dx_0\, \pi(x_0)\, A(x_0) \int dx_t\, \rho(x_t, t; x_0, 0)\, B(x_t) \tag{3}$$

A discrete-state Markov model is defined by a partition of the molecular configuration space $\mathbf{\Omega}$ into $M$ rapidly-equilibrating *states* $\mathcal{S}_i$, such that $\mathcal{S}_i \bigcap \mathcal{S}_j = 0$ if $i \neq j$ and $\bigcup_{i=1}^{M} \mathcal{S}_i = \mathbf{\Omega}$. It is convenient to define a set of *characteristic* or *membership functions* $\phi_i(x)$ which assume a value of unity within the state and zero without:

$$\phi_i(x) = \begin{cases} 1 & \text{if } x \in \mathcal{S}_i \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

Taken together, the $\phi_i(x)$ form a *partition of unity*, such that $\sum_{i=1}^{M} \phi_i(x) = 1$ for all $x$ (see, e.g. [14]).

Dynamics between the states are here characterized by a discrete-time Markov model, described by a row-stochastic transition matrix $\mathbf{T}(\tau)$ that evolves a row-vector of state occupation probabilities $\mathbf{p}(t)$ by a fixed observation interval (or *lag time*) $\tau$:

$$\mathbf{p}(\tau) \;=\; \mathbf{p}(0)\, \mathbf{T}(\tau) \tag{5}$$

Equivalently, it is common to describe the statistical dynamics in continuous time in terms of a phenomenological rate matrix $\mathbf{K}$ (see, e.g., [11]), whose off-diagonal elements contain rate constants if the states are sufficiently metastable [15–17], related to the transition matrix by

$$\mathbf{T}(\tau) \;=\; e^{\mathbf{K}\tau}, \tag{6}$$

where the exponential denotes the formal matrix exponential $e^{\mathbf{A}} \equiv \sum_{n=0}^{\infty} \mathbf{A}^n/n!$. It is critical to note that, while the statistical evolution can be described in continuous time, the model still provides no information about processes occurring on timescales shorter than the Markov time $\tau_{\mathrm{eq}}$ because they have been coarse-grained out by aggregating configuration space into discrete states. Here, we focus on the discrete-time transition matrix $\mathbf{T}$, but the method proposed in Section IV for modeling uncertainties in state observables can be combined with the inference of rate matrices $\mathbf{K}$ [18] in a straightforward manner.

The transition matrix $\mathbf{T}(\tau)$ has a well-defined meaning in terms of correlation functions [5]

$$T_{ij}(\tau) \equiv \frac{\langle \phi_i(0)\, \phi_j(t) \rangle}{\langle \phi_i \rangle}. \tag{7}$$

If dynamics is Markovian, the transition probability for an integral multiplier $n$ of the time $\tau$ is given by exponentiating the transition matrix:

$$\mathbf{T}(n\tau) = [\mathbf{T}(\tau)]^n, \quad \tau \geq \tau_{\mathrm{eq}} \tag{8}$$

where the minimal lag time $\tau$ for which this relationship holds approximately is termed the *Markov time* $\tau_{\mathrm{eq}}$ [7].

Because each state mixes quickly compared to the transition time between states, the average value of some spectroscopic signal $A(x)$ over each state can be easily estimated from short trajectories:

$$A_i \equiv \langle A \rangle_i = \int dx\, \pi_i(x)\, A(x) \tag{9}$$

where $\pi_i(x) = \pi(x)\, \phi_i(x)/\pi_i$ is the equilibrium probability restricted to the state defned by the membership function $\phi_i(x)$, and $\pi_i = \int dx\, \pi_i(x)\, \phi_i(x)$ is the equilibrium probability of occupying state $i$, which can also be determined from the equilibrium eigenvector of $\mathbf{T}(\tau)$.

This allows us to write the three types of experimental observables (Eqs. 1–3), computed from the Markov model defined by $\mathbf{T}(\tau)$ for a fixed $\tau \geq \tau_{\mathrm{eq}}$, as

$$\langle A \rangle = \sum_{i=1}^{M} \pi_i\, A_i \tag{10}$$

$$\langle A(n\tau) \rangle_0 = \sum_{i=1}^{M} \sum_{j=1}^{M} p_i(0)\, ([\mathbf{T}(\tau)]^n)_{ij}\, A_j \tag{11}$$

$$\langle A(0) B(n\tau) \rangle = \sum_{i=1}^{M} \sum_{j=1}^{M} \pi_i\, A_i\, ([\mathbf{T}(\tau)]^n)_{ij}\, B_j \tag{12}$$

Because the transition matrix $\mathbf{T}(\tau)$ is *a priori* unknown, it must be estimated or *inferred* from some data, such as a set of transition counts between discrete states observed in one or more molecular dynamics simulations. Since the quantity of data is necessarily finite, the true value of $\mathbf{T}(\tau)$, defined in terms of transfer probabilities, will be uncertain.

For a given set of observed transitions from trajectory data, what is the uncertainty of $\mathbf{T}$ and how does this affect the uncertainty of equilibrium or kinetic properties computed from the model? Adequate assessment of this uncertainty and how it influences (potentially complicated, nonlinear) observables of interest is critical in determining when a sufficient quantity of data has been collected for the hypotheses in question to be decided, and whether any discrepancy observed with experimental observation is to be deemed statistically significant. Additionally, one may further exploit knowledge of these uncertainties by planning new simulations such as to most reduce the uncertainties in the observables of interest, and thus to get converged observables with a minimal amount of simulation effort [19–22].

Due to the expense of generating the simulation data required to construct a reliable model, it is generally not practical to partition the data into independent subsets to assess the variation in properties of interested among smaller datasets. Bootstrap methods have been applied to the estimation of error in Markov models [6], but it is challenging to correctly preserve the correlation structure in the data unless the model is constructed from numerous independent trajectories sampled from equilibrium.

Much effort has therefore been devoted to Bayesian approaches to uncertainty analysis for Markov models. Bayesian methods provide a powerful and simple framework for describing the posterior uncertainty in both model parameters and observables of interest computed from the model, in addition to the potential to quantify uncertainties with information-theoretic tools. Previously, Bayesian approaches have been proposed for inferring transition matrices that do not satisfy detailed balance [19], but recent work has shown that use of prior knowledge that the dynamics must obey detailed balance [23] can significantly improve the quality of the resulting inference [24]. To this end, Bayesian approaches for inferring either reversible transition matrices [13, 24] or rate matrices [25, 26] have been proposed. Here, we consider the extension of the method for sampling reversible transition matrices proposed in [13] to also consider the uncertainty due to insufficient characterization of the averages of molecular observables within the states.

## III. UNCERTAINTY IN TRANSITION PROBABILITIES

Consider a discrete-state Markovian trajectory of duration $N\tau$ starting in a given state $s_0$, with observations of the current state made with a time resolution $\tau \geq \tau_{\mathrm{eq}}$, denoted $\mathbf{s} \equiv \{s_0, s_1, \ldots, s_N\}$. Given one or more such trajectories, the $M \times M$ transition count matrix $\mathbf{C}$, where $c_{ij}$ is the number of times a discrete trajectory $\mathbf{s}$ appears in state $i$ at some time index $t$ and state $j$ at time index $t+1$, is a sufficient statistic for capturing all information about the stochastic behavior of this system [13]. For one

trajectory, $\mathbf{C}$ can be written

$$c_{ij} \;=\; \sum_{t=1}^{T} \delta_{is_{t-1}} \delta_{js_t} \tag{13}$$

In the limit of an infinitely long trajectory, the elements of the true transition matrix are given by the trivial estimator:

$$\hat{T}_{ij}(\tau) = \frac{c_{ij}}{\sum_{k=1}^{M} c_{ik}} = \frac{c_{ij}}{c_i}, \tag{14}$$

where $c_i \equiv \sum_{k=1}^{M} c_{ik}$ gives the total number of observed transitions leaving state $i$. For a trajectory of finite length, the underlying transition matrix $\mathbf{T}(\tau)$ cannot be computed exactly from the observed transition count matrix $\mathbf{C}$. Instead, we can compute the probability that the true (unknown) transition matrix $\mathbf{T}$ generated the observed counts, assuming the process is Markovian (henceforth suppressing the argument $\tau$):

$$p(\mathbf{C}|\mathbf{T}) = \prod_{t=1}^{N} T_{s_{t-1},s_t} = p(\mathbf{C}|\mathbf{T}) = \prod_{i,j=1}^{M} T_{ij}^{c_{ij}} \tag{15}$$

By Bayes' theorem, the probability that $\mathbf{T}$ is the true transition matrix that generated the data $\mathbf{C}$ is then

$$p(\mathbf{T}|\mathbf{C}) \;\propto\; p(\mathbf{C}|\mathbf{T})\, p(\mathbf{T}) = \left[\prod_{i,j=1}^{M} T_{ij}^{c_{ij}}\right] p(\mathbf{T}), \tag{16}$$

where $p(\mathbf{T})$ is the prior probability of transition matrices, reflecting our knowledge about $\mathbf{T}$ before observing any data. For uniform prior $p(\mathbf{T})$, the maximum of the likelihood function Eq. 15 is given by the trivial estimator of Eq. 14. In the limit that the number of samples $N \to \infty$, $p(\mathbf{T}|\mathbf{C})$ becomes progressively more peaked around $\hat{\mathbf{T}}(\tau)$.

For a physical system, the condition of detailed balance $[\pi_i\, T_{ij}(\tau) \;=\; \pi_j\, T_{ji}(\tau)]$ must be satisfied for any $\tau$ [23], where

$$\pi_i \;\equiv\; \int dx\, \pi(x)\, \phi_i(x) \tag{17}$$

We presume also that the Markov chain described by $\mathbf{T}$ is *indecomposable*, so that the unit eigenvalue is nondegenerate. Such a stochastic matrix has a single, dominant eigenvector of 1, and the corresponding left eigenvector $\boldsymbol{\pi}$

$$\boldsymbol{\pi}\mathbf{T} \;=\; \boldsymbol{\pi} \tag{18}$$

which yields the stationary distribution of the transition matrix when normalized to a 1-norm of 1, such that $\sum_j \pi_j = 1$. All other eigenvalues are real, and lie on the interval $[-1, +1)$.

| | | | |
|---|---|---|---|
| C1 | Element-wise nonnegativity | $0 \le T_{ij}$ | $\forall\; i, j$ |
| C2 | Row-stochasticity | $\sum_{j=1}^{M} T_{ij} = 1$ | $\forall\; i$ |
| C3 | Detailed Balance | $\pi_i T_{ij} = \pi_j T_{ji}$ | $\forall\; i, j$ |

We wish to impose the following constraints on the transition matrix $\mathbf{T}$ through the choice of prior $p(\mathbf{T})$: Which formally corresponds to the prior

$$p(\mathbf{T}) \propto \prod_{i,j=1}^{M} T_{ij}^{(b_{ij}-1)}\, \delta(\pi_i T_{ij} - \pi_j T_{ji})\, h(T_i j) \tag{19}$$

where $h_{ij}(x)$ is the Heaviside function, and $\mathbf{B} \equiv (b_{ij})$ is an $M \times M$ matrix of prior *pseudocounts*. Typical choices of the prior are the uniform prior, $b_{ij} = 1\; \forall\, i, j$ [13], and the null prior, $b_{ij} = 0\; \forall\, i, j$ [27]. The uniform prior assigns a uniform a priori distribution to all matrix elements by adding a full pseudocount to each of them, which takes much observation data to be overriden if the system consists of many states. The null prior, on the other hand, gives most impact to the observed data, as it forces the maximum likelihood estimator and mean of the transition matrix to coincide, and all transition matrix elements in which no count has been observed to zero.

To sample the distribution given by Eq. 16, a sampling procedure based on the Metropolis-Hasting algorithm is used. Given a current matrix $T$ and a proposed new matrix $T'$, the acceptance probability is computed by:

$$p_{accept}(\mathbf{T} \to \mathbf{T}') = \frac{p(\mathbf{T}' \to \mathbf{T})}{p(\mathbf{T} \to \mathbf{T}')} \frac{p(\mathbf{T}'|\mathbf{C})}{p(\mathbf{T}|\mathbf{C})} \tag{20}$$

where $p(\mathbf{T} \to \mathbf{T}')$ and $p(\mathbf{T}' \to \mathbf{T})$ denote the probability to propose a stochastic move to $\mathbf{T}'$ given $\mathbf{T}$, and vice versa. Two types of proposal steps (which together generate an ergodic chain) are used, and are described briefly below. To ensure the ratio of priors remains calculable, the moves are restricted to only generate proposals $\mathbf{T}'$ that conform to constraints C1, C2, and C3. This is achieved by the two move types described in the subsequent subsections, each of which is chosen with a 50% probability in each iteration.

The sampling procedure is initialized with a matrix that fulfills detailed balance:

$$T_{ij}^{\text{init}} = \frac{c_{ij} + c_{ji}}{\sum_{k=1}^{M} [c_{ik} + c_{ki}]}. \tag{21}$$

When the sampling data is not generated from a global equilibrium, this matrix can be very different from the probability maximum or even outside the region of main probability mass. Therefore, it is necessary to run the sampler for a "burn-in" phase long enough until the estimated properties of interest (e.g. uncertainties of the expectation or the correlation values) become stable. This burn-in phase is discarded and the actual estimation of the uncertainties is based on the subsequent production phase of the sampler.

Further details and formal proofs of correctness of this sampling procedure are given in Ref. [13].

### A. Reversible Element Shift

Consider a pair of states $(i, j)$, $i \neq j$. The changed elements in the proposed transition matrix, $\mathbf{T}'$, after a move parameterized by $\Delta$, are given by:

$$T'_{ij} = T_{ij} - \Delta \;;\; T'_{ji} = T_{ji} - \frac{\pi_i}{\pi_j}\Delta$$
$$T'_{ii} = T_{ii} + \Delta \;;\; T'_{jj} = T_{jj} + \frac{\pi_i}{\pi_j}\Delta$$

If $\mathbf{T}$ fulfills the detailed balance condition (C3), $\mathbf{T}'$ will also fulfill C3, and the stationary distribution remains unchanged:

$$\boldsymbol{\pi}' = \boldsymbol{\pi}$$

In order to maintain the stochasticity of $\boldsymbol{T}$ (constraint C2), the parameter $\Delta$ is restricted to:

$$\Delta \in [\max(-T_{ii}, -\frac{\pi_j}{\pi_i}T_{jj}), T_{ij}]$$

If $\Delta$ is chosen uniformly from this range, the proposal probabilites for the reversible element shift are given by

$$\frac{p(\boldsymbol{T}' \to \boldsymbol{T})}{p(\boldsymbol{T} \to \boldsymbol{T}')} = \sqrt{\frac{(T_{ij} - \Delta)^2 + (T_{ji} - \frac{\pi_i}{\pi_j}\Delta)^2}{(T_{ij})^2 + (T_{ji})^2}}. \tag{22}$$

The ratio of posterior probabilities is given by

$$\frac{p(\boldsymbol{T}'|\boldsymbol{C})}{p(\boldsymbol{T}|\boldsymbol{C})} = \left(\frac{T_{ii} + \Delta}{T_{ii}}\right)^{C_{ii}} \left(\frac{T_{ij} - \Delta}{T_{ij}}\right)^{C_{ij}} \times \left(\frac{T_{jj} + \frac{\pi_i}{\pi_j}\Delta}{T_{jj}}\right)^{C_{jj}} \left(\frac{T_{ji} - \frac{\pi_i}{\pi_j}\Delta}{T_{ij}}\right)^{C_{ji}} \tag{23}$$

### B. Row Shift

In a row shift move, a row $i$ of $\boldsymbol{T}$ is selected with uniform probability $1/M$, and probability mass is moved from the diagonal element $T_{ii}$ to all outgoing probabilities $T_{ij}$ for $j = 1, \ldots, M$:

$$T'_{ij} = \alpha T_{ij}$$
$$T'_{ii} = \alpha(T_{ii} - 1) + 1$$

To maintain stochasticity, the parameter $\alpha$ is drawn uniformly from the range:

$$\alpha \in \left[0, \frac{1}{1 - T_{ii}}\right]$$

The ratio of proposal probabilities is given by:

$$\frac{p(\boldsymbol{T}' \to \boldsymbol{T})}{p(\boldsymbol{T} \to \boldsymbol{T}')} = \alpha^{(m-2)}. \tag{24}$$

where $m$ is the number of non-zero transition probabilities in the modified row.

The ratio of posterior probabilities is given by

$$\frac{p(\boldsymbol{T}' \to \boldsymbol{T})}{p(\boldsymbol{T} \to \boldsymbol{T}')} = \alpha^{(c_i - c_{ii})} \left(\frac{1 - \alpha(1 - T_{ii})}{T_{ii}}\right)^{c_{ii}} \tag{25}$$

The row shift operation will change the stationary distribution $\pi$, but $\pi$ may be efficiently updated:

$$\pi'_i = \frac{\pi_i}{\pi_i + \alpha(1 - \pi_i)} \;;\; \pi'_j = \frac{\alpha\,\pi_j}{\pi_i + \alpha(1 - \pi_i)}.$$

Since this update scheme is incremental, it will accumulate numerical errors over time that cause the updated $\pi$ to drift away from the stationary distribution of the current transition matrix. To avoid that, $\pi$ is recomputed from the current sample of the transition matrix in regular intervals (here, every 100 sampling steps).

## IV. UNCERTAINTY IN MOLECULAR OBSERVABLES

We now consider the problem of estimating the expectation of some molecular observable $A(x)$ over each discrete state:

$$\langle A \rangle_i \;\equiv\; \int dx\, A(x) \tag{26}$$

From a single trajectory, the straightforward estimator of this quantity is simply the sample mean over those samples within state $i$:

$$\hat{A}_i = \frac{\sum_{t=1}^{T} \phi_i(x_t)\, A(x_t)}{\sum_{t=1}^{T} \phi_i(x_t)} \tag{27}$$

Temporally sequential samples $A_t \equiv A(x_t)$ collected with a temporal resolution of the Markov time $\tau_{\mathrm{eq}}$ are presumed to be uncorrelated, because by definition, this is the maximum time required for the system to decorrelate within any discrete state. We presume that the set of samples $A(x_t)$ for those configurations $x_t$ appearing in state $i$ are collected in the set $\{A_n\}_{n=1}^{N}$ in the remainder of this section, generally abbreviated as $\{A_n\}$.

Because only a finite number of samples are collected for each state, there will be a degree of uncertainty in this estimate. Unlike the problem of inferring the transition matrix elements, however, we cannot write an exact expression for the probability of observing the $A_n$ in terms of a simple parametric form, since its probability distribution may be arbitrarily complex:

$$p_i(A_n) \;=\; \int dx\, \delta(A_n - A(x))\, \pi_i(x) \tag{28}$$

Despite this, the central limit theorem states that the behavior of $\hat{A}_i$ approaches a normal distribution (generally very rapidly) as the number of samples $N$ increases. We will therefore make the assumption that $p_i(A_n)$ is normal, and demonstrate this allows us to do a very good job of inferring the distribution of the error in $\delta\hat{A}_i \equiv \hat{A}_i - \langle A \rangle_i$ for a very reasonable number of samples, and generally gives an overestimate of the error (which is arguably less dangerous than an underestimate) for smaller sample sizes.

Consider the sample mean estimator for $\langle A \rangle_i$:

$$\hat{\mu} \;=\; \frac{1}{N}\sum_{n=1}^{N} A_n \tag{29}$$

The asymptotic variance of $\hat{\mu}$, which provides a good estimate of the statistical uncertainty in $\hat{\mu}$ in the large-sample limit, is given as a simple consequence of the central limit theorem

$$\delta^2\hat{\mu} \;\equiv\; E[(\hat{\mu} - E[\hat{\mu}])^2] = \frac{\mathrm{var}\, A_n}{N} \approx \frac{\hat{\sigma}^2}{N} \tag{30}$$

where the unbiased estimator for the variance $\sigma^2 \equiv \mathrm{var}\, A_n$ is given by

$$\hat{\sigma^2} \;\equiv\; \frac{1}{N-1}\sum_{n=1}^{N}(A_n - \hat{\mu})^2 \tag{31}$$

Suppose we now *assume* the distribution of $A$ from state $i$ is normal, even though there is no reason to expect it will be:

$$A \,|\, \mu, \sigma^2 \;\sim\; N(\mu, \sigma^2) \tag{32}$$

Were this to be a reasonable model, we could model the timeseries of the observable $A_t \equiv A(x_t)$ by the hierarchical process:

$$\begin{aligned} s_t \,|\, s_{t-t}, \mathbf{T} \;&\sim\; \mathrm{BS}(T_{s_{t-1}\,1}, \ldots, T_{s_{t-1}\,N}) \\ A_t \,|\, \mu_{s_t}, \sigma^2_{s_t} \;&\sim\; \mathrm{N}(\mu_{s_t}, \sigma^2_{s_t}) \end{aligned} \tag{33}$$

Here, the notation $\mathrm{BS}(\pi_1, \ldots, \pi_N)$ denotes a Bernoulli scheme where discrete outcome $n$ has associated probability $\pi_n$, and $\mathrm{N}(\mu, \sigma^2)$ denotes the normal distribution with mean $\mu$ and variance $\sigma^2$. We will demonstrate below how this model does in fact recapitulate the expected behavior in the limit where there are sufficient samples from each state.

We choose the (improper) Jeffreys prior [28],

$$p(\mu, \sigma^2) \;\propto\; \sigma^{-2} \tag{34}$$

because it satisfies intuitively reasonable reparameterization [28] and information-theoretic [29] invariance principles. Note that this prior is uniform in $(\mu, \log\sigma)$.

The posterior is then given by

$$\begin{aligned} p(\mu, \sigma^2|\{A_n\}) &\propto \left[\prod_{n=1}^{N} p(A_n|\mu, \sigma^2)\right] p(\mu, \sigma^2) \\ &\propto\; \sigma^{-(N+2)} \exp\left[-\frac{1}{2\sigma^2}\sum_{n=1}^{N}(A_n - \mu)^2\right] \end{aligned} \tag{35}$$

Rewriting in terms of the sample statistics $\hat{\mu}$ and $\hat{\sigma}^2$, we obtain

$$\begin{aligned} &p(\mu, \sigma^2|\{A_n\}) \\ &\propto\; \sigma^{-(N+2)} \exp\left\{-\frac{1}{2\sigma^2}\left[\sum_{n=1}^{N}(A_n - \hat{\mu})^2 + N(\hat{\mu} - \mu)^2\right]\right\} \\ &\propto\; \sigma^{-(N+2)} \exp\left\{-\frac{1}{2\sigma^2}\left[(N-1)\hat{\sigma}^2 + N(\hat{\mu} - \mu)^2\right]\right\} \end{aligned} \tag{36}$$

The posterior has marginal distributions

$$\begin{aligned} \sigma^2 \,|\, \{A_n\} \;&\sim\; \mathrm{Inv}{-}\chi^2(N-1, \hat{\sigma}^2) \\ \mu \,|\, \{A_n\} \;&\sim\; \mathrm{t}_{N-1}(\hat{\mu}, \hat{\sigma}^2/N) \end{aligned} \tag{37}$$

where $\sigma^2$ is distributed according to scaled inverse chi-square distribution with $N-1$ degrees of freedom, and $\mu$ according to Student's t-distribution with $N-1$ degrees of freedom that has been shifted to be centered about $\hat{\mu}$ and whose width has been scaled by $\hat{\sigma}^2/N$.

As can be seen in Figure 1, as the number of degrees of freedom increases, the marginal posterior for $\mu$ approaches the normal distribution with the asymptotic behavior expected from standard frequentist analysis for the standard error of the mean, namely

$$\mu \to \mathrm{N}(\hat{\mu}, \hat{\sigma}^2/N) \tag{38}$$

At low sample counts, the t-distribution is lower and wider than the normal distribution, meaning that confidence intervals computed from this distribution will be somewhat larger than those of the corresponding normal estimate for small samples. In some sense, this partly compensates for $\hat{\sigma}^2$ being a poor estimate of the true variance for small sample sizes, which would naturally lead to underestimates of the statistical uncertainty. In any case, this is also far from the asymptotic limit where the normal distribution with variance $\hat{\sigma}^2/N$ is expected to model the uncertainty well.

The posterior can be decomposed as

$$p(\mu, \sigma^2|\{A_n\}) \;=\; p(\mu|\sigma^2, \{A_n\})\, p(\sigma^2|\{A_n\}) \tag{39}$$

This readily suggests a two-step sampling scheme for generating uncorrelated samples of $(\mu, \sigma^2)$, in which we first sample $\sigma^2$ from its marginal distribution, and then $\mu$ from its distribution conditional on $\sigma^2$

$$\begin{aligned} \sigma^2 \,|\, \{A_n\} \;&\sim\; \mathrm{Inv}{-}\chi^2(N-1, \hat{\sigma}^2) \\ \mu \,|\, \sigma^2,\, \{A_n\} \;&\sim\; \mathrm{N}(\hat{\mu}, \sigma^2/N) \end{aligned} \tag{40}$$

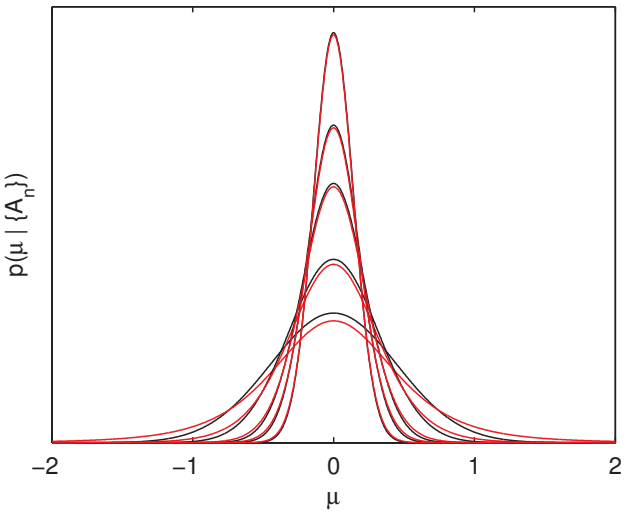


FIG. 1: **Approach to normality for marginal distribution of the mean** $p(\mu|\{A_n\})$**.** For fixed $\hat{\mu}$ and $\hat{\sigma}^2$, the marginal posterior distribution of $\mu$ (red), a scaled and shifted Student t-distribution, rapidly approaches the normal distribution (black) expected from asymptotic statistics. The PDF is shown for sample sizes of $N = 5$ (the broadest), 10, 20, and 30.

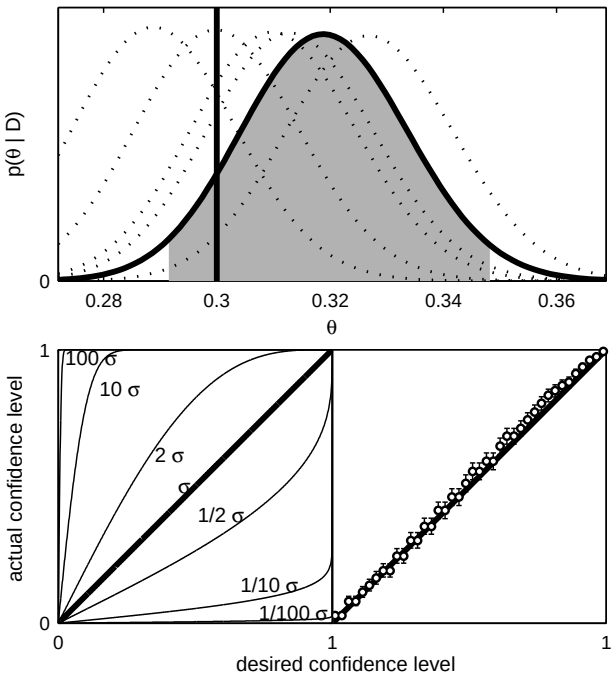


FIG. 2: **Testing the posterior for inference of a biased coin flip experiment.** *Top:* Posterior distribution for inferring the probability of heads, $\theta$, for a biased coin from an sequence of $N = 1000$ coin flips (dark line) with 95% symmetric confidence interval about the mean (shaded area). The true probability of heads is 0.3 (vertical thick line). Posteriors from five different experiments are shown as dotted lines. *Bottom left:* Desired and actual confidence levels for an idealized normal posterior distribution that either overestimates (upper left curves) or underestimates (bottom right curves) the true posterior variance by different degrees. *Bottom right:* Desired and actual confidence levels for the Binomial-Beta posterior for the coin flip problem depicted in upper panel. Error bars show 95% confidence intervals estimates from 1000 independent experimental trials. For inference, we use a likelihood function such that the observed number of heads is $N_H|\theta \sim \text{Binomial}(N_H, N, \theta)$ and conjugate Jeffreys prior [28, 29] $\theta \sim \text{Beta}(1/2, 1/2)$ which produces posterior $\theta|N_H \sim \text{Beta}(N_H + 1/2, N_T + 1/2)$ along with constraint $N_H + N_T = N$.

Alternatively, if the scaled inverse-chi-square distribution is not available, the $\chi^2$-distribution (among others) can be used:

$$(N-1)(\hat{\sigma}^2/\sigma^2) \,|\, \{A_n\} \;\sim\; \chi^2(N-1) \qquad (41)$$

where the first argument is the shape parameter and the second argument is the scale parameter.

## V. VALIDATION IN A MODEL SYSTEM

How can we test a Bayesian posterior distribution? One of the more powerful features of a Bayesian model is its ability to provide confidence intervals that correctly reflect the level of certainty that the true value will lie within it. For example, if the experiment were to be repeated many times, the true value of the parameter being estimated should fall within the confidence interval for a 95% confidence level 95% of the time. As an illustrative example, consider a biased coin where the probability of turning heads is $\theta$. From an observed sample of N coin flips, we can estimate $\theta$ using a Binomial model for the number of coin flips that turn up heads and a conjugate Beta Jeffreys prior [28, 29]. Each time we run experiment and generate a new independent collection of $N$ samples, we get a different posterior estimate for $\theta$, and a different confidence interval (Figure 2, top). If we run many trials and record what fraction of the time the true (unknown) value of $\theta$ falls within the confidence interval estimated from that trial, we can see if our model is correct. If correct, the observed confidence level should match the desired confidence level (Figure 2, bottom right). Deviation from parity means that the posterior is either two broad or too narrow, and that the statistical uncertainty is being either over- or underestimated (Figure 2, bottom left).

We performed a similar test on a three-state model system, where the (reversible, row-stochastic) transition matrix for one Markov time is given by

$$\mathbf{T}(1) \;=\; \begin{bmatrix} 0.86207 & 0.12931 & 0.00862 \\ 0.15625 & 0.83333 & 0.01041 \\ 0.00199 & 0.00199 & 0.99602 \end{bmatrix} \qquad (42)$$

Each state is characterized by a mean value of the observable $A(x)$, fixed to $3, 2$, and $1$ for the first, second,

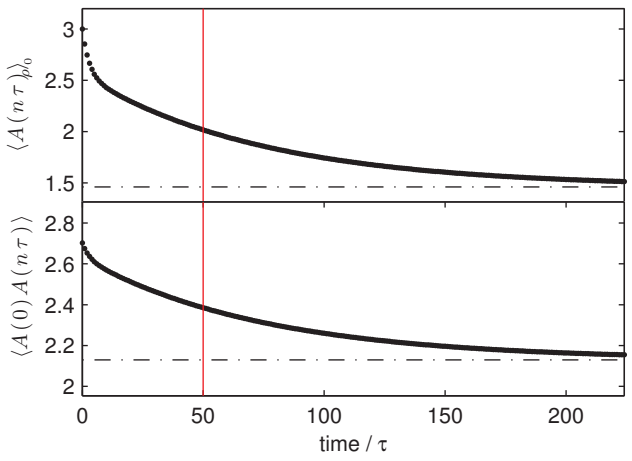


FIG. 3: **Observables for three-state model system** *Top:* Relaxation of $\langle A(t)\rangle_{\rho_0}$ (solid line) from initial distribution $\rho_0 = [100]$ to equilibrium expectation $\langle A\rangle$ (dash-dotted line). *Bottom:* Equilibrium autocorrelation function $\langle A(0)\,A(t)\rangle$ (solid line) to $\langle A\rangle^2$ (dash-dotted line). The estimates of both $\langle A(t)\rangle_{\rho_0}$ and $\langle A(0)\,A(t)\rangle$ at 50 timesteps (red vertical line) were assessed in the validation tests described here.

and third states, respectively. The equilibrium populations are $\pi \approx [0.1625\ 0.1345\ 0.7031]$. Simulation from this model involves a stochastic transition according to the transition element $T_{ij}$ followed by observation of the value of $A(x)$ sampled i.i.d. from the current state's probability distribution $p_i(A)$. The nonequilibrium relaxation $\langle A\rangle_{\rho_0}$ from the initial condition $\rho_0 = [100]$ in which all density is concentrated in state 1, as well as the autocorrelation function $\langle A(0)\,A(t)\rangle$, is shown in Figure 3.

With the means of $p_i(A)$ within each state fixed as above, we considered models for $p_i(A)$ that were either *normal* or *exponential*, using the probability density functions:

$$p_i(A) = (2\pi)^{-1/2}\sigma_i^{-1}\exp\left[-\frac{1}{2\sigma_i^2}(A-\mu_i)^2\right] \qquad \text{normal}$$

$$p_i(A) = \mu_i^{-1}\exp[-A/\mu_i]\ ,\ A \geq 0 \qquad \text{exponential}$$

While the normal output distribution for $p_i(A)$ corresponds to the hierarchical Bayesian model that forms the basis for our approach, the exponential distribution is significantly different, and represents a challenging test case.

Figure 4 depicts the resulting uncertainty estimates for both normal (top) and exponential (bottom) densities for the observable $A$. In both cases, the confidence intervals are *underestimated* for short trajectory lengths (1 000 steps) where, in many realizations, few samples are observed in one or more states, so that the variance is underestimated or the effective asymptotic limit has not yet been reached. As the simulation length is increased to 10 000 or 100 000 steps so that it is much more likely there are a sufficient number of samples in each state to reach the asymptotic limit, however, the confidence intervals predicted by the Bayesian posterior become quite good. For the exponential model for observing values of $A$ (which might be the case in, say, fluorescence lifetimes), we observe similar behavior. Except for what appears to be a slight, consistent underestimation of $\langle A(t)\rangle_{\rho_0}$ (much less than half a standard deviation) there appears to be excellent agreement between the expected and observed confidence intervals, confirming that this method is expected to be a useful approach to modeling statistical uncertainties in equilibrium and kinetic observables.

## VI. APPLICATION TO FLUORESCENCE CORRELATION IN A PEPTIDE

Time-resolved single-molecule fluorescence experiments provide a way to monitor the microscopic features of folding by probing the fluctuations of a conformation-sensitive spectroscopic probe for individual molecules. In contrast to ensemble measurements, which yield information only on average properties, single-molecule experiments provide information on distributions and time trajectories of properties that would otherwise be hidden [30]. Due to recent advances in instrumentation, single-molecule fluorescence spectroscopy has received a particular surge of interest. In contrast to traditional fluorescence experiments, nonequilibrium perturbations to synchronize an ensemble of molecules are avoided, allowing equilibrium conformational dynamics to be studied [30–32]. Fluorescence correlation spectroscopy (FCS) takes advantage of Brownian diffusion to bring one or a few fluorescent molecules into the observation volume at any given time, recording bursts of fluorescent emission. By calculating the autocorrelation function (ACF) of the fluorescence intensity fluctuations, information about the temporal statistics of dynamics can be obtained for a wide range of time scales spanning nanoseconds to seconds [32]. Because the fluorescence signal is only an indirect probe of molecular conformation, molecular dynamics simulations (typically limited to nanosecond to microsecond timescales) have proven to be of great utility in the interpretation of fluorescence autocorrelation experiments [33]. By employing Markov models to describe the long-time dynamics, it is possible to extend the utility of molecular dynamics simulation for predicting and interpreting these experiments into the microsecond regime and beyond [27].

Here, we illustrate the use of the Bayesian uncertainty analysis procedure described above in interpreting an experimental study of the loop-closure dynamics of a small fluorescently-labeled peptide using a Markov model constructed from molecular dynamics simulation data. To probe the fastest processes in protein folding, Krieger et al. and Neuweiler et al. collected single-

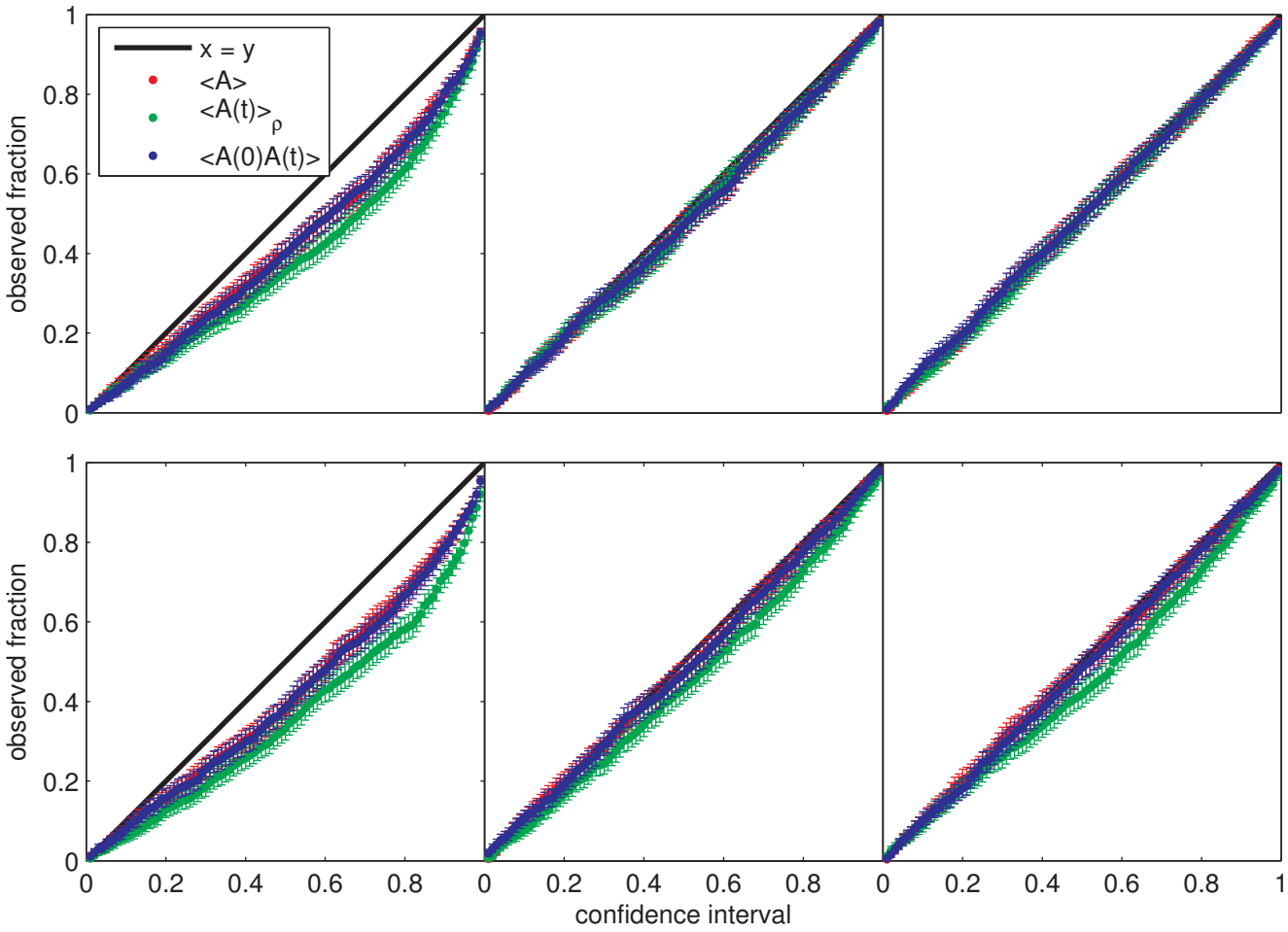


FIG. 4: **Confidence interval tests for model system** *Top:* Expected and observed confidence intervals for three-state system with normal distribution for observable $A$ with unit variance for simulations of length 1 000 (left), 10 000 (middle), and 100 000 (right) steps. Confidence intervals were estimated from generating 10 000 samples from the Bayesian posterior. Estimates of the fraction of observed times the true value was within the confidence interval estimated from the Bayesian posterior were computed from generating 1 000 independent experimental realizations. The resulting curves are shown for the equilibrium estimate $\langle A \rangle$ (red), nonequilibrium relaxation $\langle A \rangle_{\rho_0}$ (green), and the equilibrium correlation function $\langle A(0)A(t) \rangle$ (blue). *Bottom:* Same as top, except an exponential distribution with the same mean was used for the probability of observing a particular value of $A$ within each state.

molecule fluorescence data on a series of small peptides incorporating fluorophores and quenchers [34, 35]. In the experiment considered here, end-to-end contact formation is studied by monitoring selective fluorescence quenching of an N-terminal fluorescent oxazine derivative MR121 by a C-terminal tryptophan residue, an efficient natural amino acid quencher, with an intervening Gly-Ser-Gly-Ser repeat (hereafter called MR121-GSGS-W, see Fig. 5). To first order, when the tryptophan is sufficiently close to the MR121 dye, fluorescence is quenched (resulting in an "off" state); conversely, when the tryptophan is far from the dye and the peptide is in a more extended conformation, the dye is fluorescent (resulting in an "on" state). The quenching process has been shown to be diffusion limited [35, 36], enabling the underlying contact-formation kinetics to be probed by fluorescence correlation spectroscopy (FCS) with nanosecond time-resolution [34, 35].

### A. Simulation details and Markov model construction

Here, a 3 $\mu$s molecular dynamics simulation of MR121-GSGS-W is analyzed. The simulations were performed in explicit water at 293 K using the GROMOS96 force field 43a1 [37, 38] and the GROMACS program version 3.2.1 [39, 40]. Partial atomic charges for the dye MR121 were taken from Vaiana et al. [36]. One peptide molecule in an extended conformation was solvated with SPC water [41] and placed in a periodic rhombic

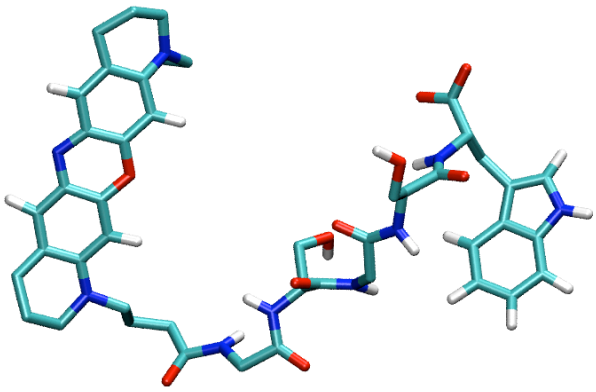

FIG. 5: **Fluorescent peptide MR121-GSGS-W.** The large fluorescent dye molecule, MR121, is the fused multiring structure visible at the N-terminus on the left of the figure.

dodecahedron box large enough to contain the peptide molecule and $\approx 1.0$ nm of solvent on all sides at a liquid density of 55.32 mol/l ($\approx$1 g/cm$^3$), resulting in 1155 water molecules in the simulation box. The simulation volume was held fixed, and thermal control was enforced [1] using the Berendsen weak-coupling algorithm [45] with a coupling time of 0.1 ps. All bond lengths were fixed using the LINCS algorithm of order 4 with a tolerance of $10^{-4}$ nm [46]. and a time step of 2 fs for numerical integration was used. Periodic boundary conditions were applied to the simulation box and the long-range electrostatic interactions were treated with the particle mesh Ewald method [47] using a grid spacing of 0.12 nm combined with a fourth-order B-spline interpolation to compute the potential and forces in between grid points and an Ewald tolerance of $10^{-5}$. The real space cut-off distance was set to 0.9 nm. The C-terminal end of the peptide was modeled as $COO^-$ to reproduce a pH of about 7 as in the experimental conditions [35]. No counterions were added since the simulation box was already neutral (one positive charge on MR121 and one negative charge on the terminal $COO^-$). The coordinates were saved every $\Delta t = 0.2$ ps.

As a crude model of fluorescence quenching, all configurations in which the heavy atoms of the rings systems of MR121 and the tryptophan had a nearest-neighbor distance of greater than 0.45 nm were defined to be fluorescent, and the remainder as dark (quenched). A cutoff of 0.45 nm was selected since quenching occurs upon van der Waals contact [35]. Thus, a fluorescence observable can be defined by the fluorescence observable $f(x)$ :

$$f(x) = \begin{cases} 0 & x \text{ dark} \\ 1 & x \text{ fluorescent} \end{cases}, \tag{43}$$

while the fluorescence of state $i$ is given by

$$f_i = \int dx\, \pi_i(x)\, f(x), \tag{44}$$

which, of course, must be estimated from the simulation data collected within state $i$[2].

From the simulation trajectory, 1 000 configurations equally spaced in time were used to define the *generators* of an initial partitioning of configuration space into Voronoi polytopes (as in the microstates of [48]), where each structure in the trajectory is assigned to the generator it is closest to using the least-squares-aligned RMSD [49, 50] as a metric [51] using the efficient RMSD calculation procedure of Theobald [52]. A lag time of $\tau$ = 1 ns was then used, as examination of the implied timescales [48] showed they become approximately independent of lag time for larger values of $\tau$. In order to obtain reasonable estimates of the per-state fluorescence $f_i$ and transition probabilities $T_{ij}$, the state space was further coarse-grained by computing the right eigenvectors of the 1000×1000 row-stochastic transition matrix and lumping states that had a Euclidean distance of less than 0.05 in the space of the 20 dominant eigenvectors which were normalized to have unit 2-norm. This generated a reduced state space of 163 states, and it was verified that the implied timescales did not significantly deteriorate.

The unnormalized fluorescence autocorrelation function is given in terms of $f_i$ and $T_{ij}$ as

$$\langle f(0) f(n\tau) \rangle = \sum_{i=1}^{M} \sum_{j=1}^{M} \pi_i\, f_i\, ([T(\tau)^n])_{ij}\, f_j. \tag{45}$$

which corresponds to the part of the fluorescence autocorrelation curve that is measured by FCS experiments. Note that the measured FCS curves will also contain additional contributions due to effects such as triplet states and diffusion which are not considered in our analysis.

The normalized form of the fluorescence autocorrelation function, which decays from an initial value of

[1] Note that the Berendsen weak-coupling algorithm does not generate a true NVT ensemble, and only produces statistics equivalent to an NVT ensemble in the thermodynamic limit [42–44].

[2] Note that, for this particular case, the observable $f(x)$ could be modeled *exactly* by a Bernoulli trial where the state fluorescence $f_i$ are the (unknown) parameters. However, due to the broad applicability of the normal model to many types of observables, we do not examine the Bernoulli model here. In any case, any significant difference between the two models should vanish as the number of samples reaches the asymptotic limit.

unity to zero at large $t$, is given by:

$$C(t) \equiv \frac{\langle f(0)f(n\tau)\rangle - \langle f\rangle^2}{\langle f^2\rangle - \langle f\rangle^2} = \frac{\sum_{i=1}^M \sum_{j=1}^M \pi_i\, f_i\, ([T(\tau)^n])_{ij}\, f_j \; - \sum_{i=1}^M (\pi_i\, f_i)^2}{\sum_{i=1}^M \pi_i\, f_i^2 (1-\pi_i)} \quad (46)$$

### B. Uncertainties in model-derived observables

To determine how the uncertainties in the computed unnormalized and normalized fluorescence correlation functions depend on the quantity of simulation data used, the initial portions of the clustered MD trajectory 0.1, 0.25, 0.5, 1, 2 and 3 $\mu$s were used to construct the count matrix $\mathbf{C}$ and number of fluorescent samples per state $f_i$. The sampling procedure described above was used to sample from the model posterior, with a burn-in phase of $10^6$ samples that were discarded. Due to the local nature of the Monte Carlo transition matrix moves detailed in Section III, subsequent transition matrix samples are generally strongly correlated. Since the procedure ergodically samples from the full distribution, it is asymptotically correct to use all transition matrix samples for estimating uncertainties in the observables, however this is inefficient for large systems. Therefore, here $10^5$ sampling steps were conducted for each transition matrix used to estimate the fluorescence correlation function, and 1 000 fluorescence correlation functions were sampled in this way (corresponding to $10^8$ sampling steps in total). This procedure required 2 – 3 minutes on a standard 2.4 GHz Intel CPU core in total.

Fig. 6 shows a sample of 20 unnormalized (left) and normalized (right) fluorescence correlation functions computed from the Bayesian model posterior. The most obvious source of uncertainty in the unnormalized correlation functions is the uncertainty in the stationary distribution $\boldsymbol{\pi}$, which results in a visually apparent dispersion in the plateau values. As expected, the variance in the plateau values becomes smaller for longer simulation times, especially when the simulation times exceeds 2 $\mu$s. The uncertainty in the rate of decay, however, appears to be rather similar despite this. This is more easily seen in the normalized fluorescence correlation functions (Fig. 6, right, shown in log scale) which has the effect of removing the effects of differences in asymptotic value. The normalized autocorrelation is plotted on a log-scale, such that a single-exponential relaxation would appear as a straight line. All curves show an asymptotic single-exponential relaxation with a timescale and rate determined by the slowest process implied by the corresponding transition matrix. At shorter times of about 10–20 ns, however, all curves appear to be multi-exponential. While multi-exponential behavior was not seen in Ref. [35], but this timescale is close to the resolution limit of this experiment. For all simulation times, the rate of decay of the slow components appears to be similar, although the amplitude (i.e. the relative amplitude of slow and fast components) varies. Note that the visually apparent increase in variance at larger at long simulation times ($\geq$1 $\mu$s) is simply a consequence of the logarithmic scale used for plotting the normalized autocorrelation function. However, it seems that the short simulation times, the slowest processes are estimated to be slower than at long simulation times.

The autocorrelation function $\langle A(0)A(n\tau)\rangle$ (Eq. 3) can be written in terms of the left eigenvectors $\mathbf{l}_i$, right eigenvectors $\mathbf{r}_i$, and eigenvalues $\lambda_i$ of $\mathbf{T}(\tau)$:

$$\begin{aligned}\langle A(0)A(n\tau)\rangle &= \sum_{i,j=1}^M A_i([\mathbf{T}(\tau)]^n)_{ij}A_j \\ &= \sum_{i,j} A_i \left(\sum_k r_{ki}\,\lambda_k^n\, l_{kj}\right) A_j \\ &= \sum_k \left[\left(\sum_i A_i\, r_{ki}\right)\left(\sum_j A_j\, l_{kj}\right)\right]\lambda_k^n \\ &\equiv \sum_k a_k\,\lambda_k^n = \sum_k a_k\, e^{-n\tau/t_k} \end{aligned} \quad (47)$$

The timescales $t_i$ associated with the two slowest processes were computed by

$$t_i = -\frac{\tau}{\log \lambda_i}, \quad (48)$$

with $i = 2, 3$, where $\lambda_2$ and $\lambda_3$ are the two largest non-unit eigenvalues of $\mathbf{T}(\tau)$. The relative amplitudes of the two processes in the fluorescence correlation function

$$a_i^{'} = \frac{a_i}{a_2 + a_3} \quad (49)$$

with $i = 2, 3$, were computed as well. The amplitudes $a_2'$, $a_3'$ are shown in Fig. 7. At simulation times of 1 $\mu$s or less, the relative amplitudes of the slowest process is somewhat smaller than that of the second-slowest process. This changes at 2 or 3 $\mu$s, where the slowest process is revealed as the one with the largest amplitude. As seen in the previous plot, the timescales, in particular the slowest timescale, are overestimated at short simulation times, while their uncertainty, being a factor of 2–3 at 100 ns, rapidly decreases. At 2 $\mu$s and 3 $\mu$s, the timescales appear well-determined, and are around 15 and 20 ns. Note that although these timescales are above the resolution limit of the experiment in Ref. [35], in the presence of noise, a signal consisting of two exponentials with similar timescales can generally not be distinguished from a single exponential.

## VII. DISCUSSION

We have demonstrated how a normal model for the probability of generating observations from each state

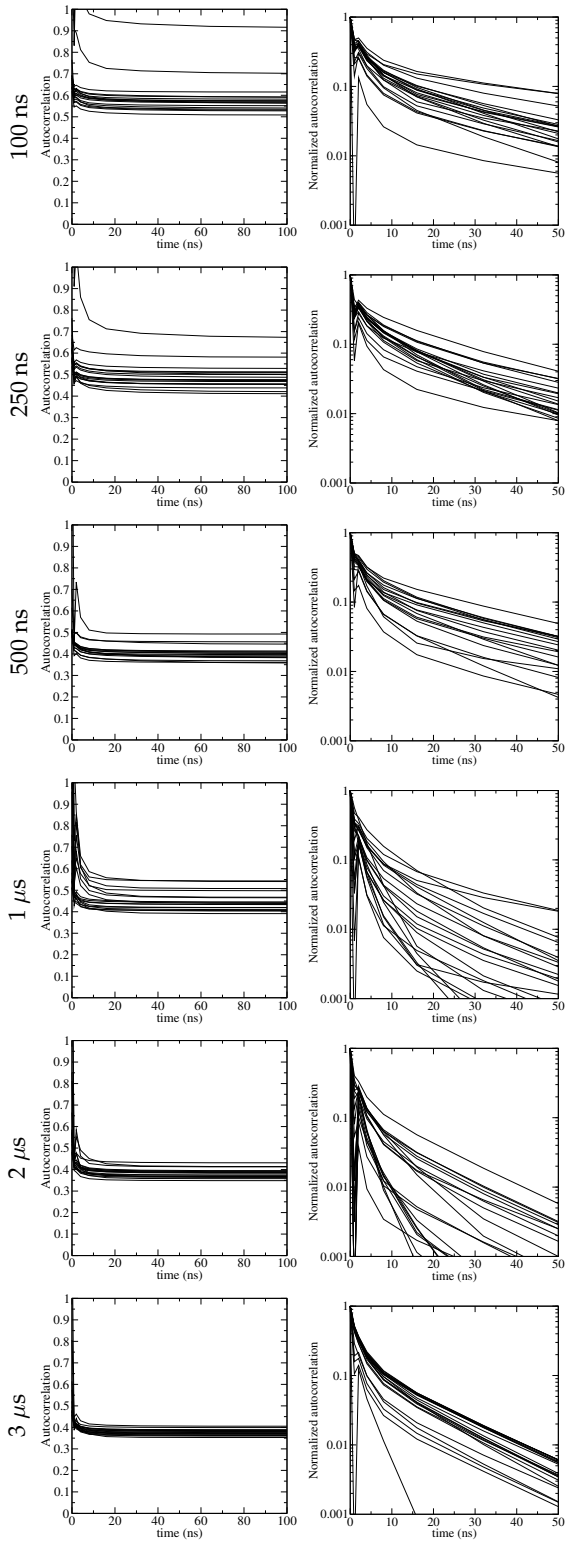


FIG. 6: **Sampled fluorescence autocorrelation functions for fluorescent peptide for different lag times.** 20 posterior samples of the computed fluorescence autocorrelation function are shown to indicate the uncertainty in the autocorrelation. Left: Unnormalized autocorrelation function on a linear scale. Right: Normalized autocorrelation function on a logarithmic scale, to better illustrate decay components.

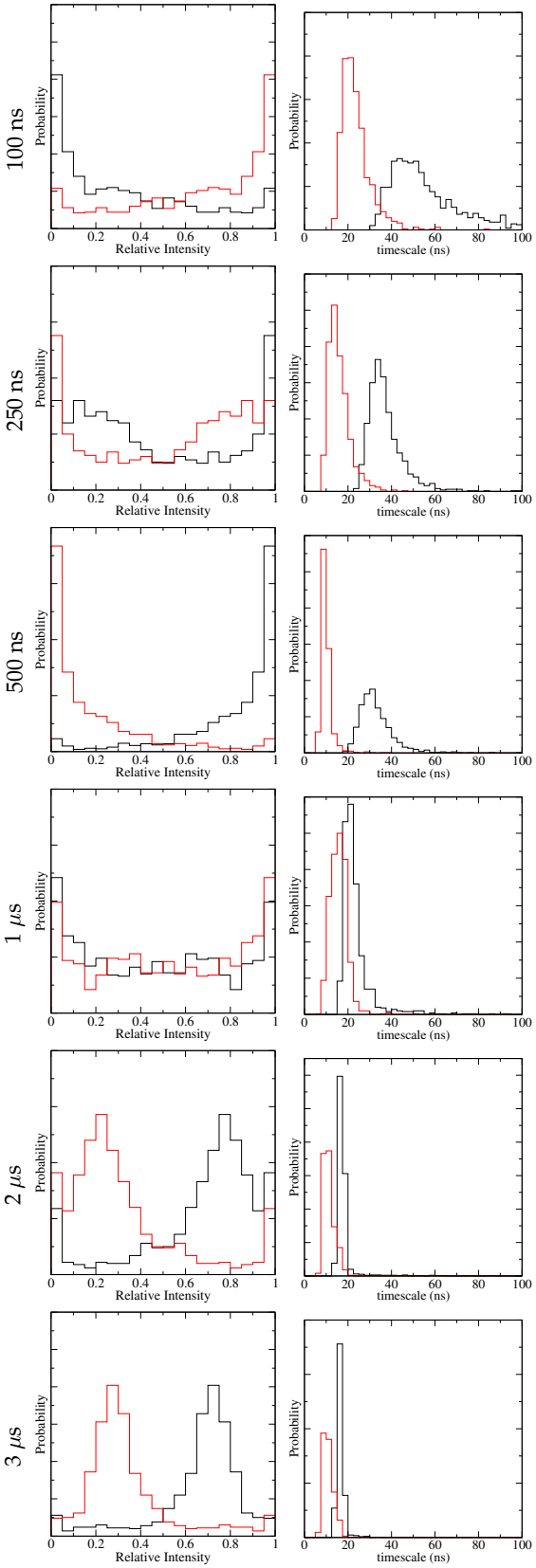


FIG. 7: FCS relative intensity and timescale of the slowest (black) and the second-slowest (red) process.

in a Markov model, even though approximate, can give very useful estimates for the uncertainties of the expectation restricted to this state, and provide a convenient means for propagating these uncertainties into the uncertainties in more complicated quantities (such as correlation functions and nonequilibrium relaxation spectra) predicted from the model. As the number of samples observed within each state approaches the asymptotic limit, the uncertainties computed from the normal model approach those expected from asymptotic theory. While in particular situations, a different (non-normal) model may be applicable to modeling the distribution of observables from each state given unknown parameters, the generality and simplicity of the normal model presented here is expected to be of broadest utility.

In conjunction with transition matrix sampling to incorporate the statistical uncertainties of the transition probabilities between Markov model states, the posterior distribution of virtually any observable accessible with experiment can be computed, provided two conditions are met: (1) The Markov time $\tau$ for which the model accurately reproduces dynamics is shorter than the timescales of the process of interest, and (2) it is possible to compute a *surrogate* for the spectroscopic signal as a function of conformation to an accuracy greater than the uncertainty of the experimental measurement. Even if the accuracy of the surrogate for the spectroscopic signal is not high, it is possible that the *timescales* present in the experimental signal may still be compared.

Another advantage conferred by having a fully *generative* model — in which complete realizations of the experiment can be produced artificially — is the possibility for predicting how additional data collection schemes will most rapidly reduce the uncertainty in the quantities desired from the model. For example, if a particular nonequilibrium relaxation spectrum is of interest, Bayesian experimental design techniques [53] could be employed in order to determine how new simulations can be initiated in a way expected to most rapidly reduce the uncertainty in this quantity.

## VIII. ACKNOWLEDGMENTS

The authors thank Vijay S. Pande and Sergio Bacallado (Stanford) for helpful feedback on the manuscript. JDC acknowledges support from a distinguished postdoctoral fellowship from the California Institute for Quantitative Biosciences (QB3) at the University of California, Berkeley. FN acknowledges support from DFG Research Center Matheon.

[1] C. Schütte, A. Fischer, W. Huisinga, and P. Deuflhard, J, Comput. Phys. **151**, 146 (1999).
[2] C. Schütte and W. Huisinga, in *Handbook of Numerical Analysis - special volume on computational chemistry*, edited by P. G. Ciaret and J.-L. Lions (Elsevier, ADDRESS, 2002), Vol. X.
[3] F. Noé and S. Fischer, Current Opinion in Structural Biology **18**, 154 (2008).
[4] M. Sarich, F. Noé, and Ch. Schütte. On the approximation quality of Markov state models. Submitted to Multiscale Model. Sim., preprint available from http://proteomics-berlin.de/771/.
[5] W. C. Swope, J. W. Pitera, and F. Suits, J. Phys. Chem. B **108**, 6571 (2004).
[6] J. D. Chodera, W. C. Swope, J. W. Pitera, and K. A. Dill, Multiscale Modeling & Simulation **5**, 1214 (2006).
[7] J. D. Chodera, N. Singhal, V. S. Pande, K. A. Dill, and W. C. Swope, J Chem Phys **126**, (2007).
[8] F. Noé, I. Horenko, C. Schütte, and J. C. Smith, J Chem Phys **126**, (2007).
[9] V. Schultheis, T. Hirschberger, H. Carstens, and P. Tavan, J. Chem. Theor. Comput. **1**, 515 (2005).
[10] S. Park and V. S. Pande, The Journal of Chemical Physics **124**, (2006).
[11] N.-V. Buchete and G. Hummer, J. Phys. Chem. B **112**, 6057 (2008).
[12] S. Bacallado, J. D. Chodera, and V. Pande, The Journal of Chemical Physics **131**, 045106+ (2009).
[13] F. Noé, J. Chem. Phys. **128**, 244103 (2008).
[14] M. Weber, PhD in Mathematics, Freie Universität Berlin — Fachbereich Mathematik und Informatik, 2006, iSBN 3-89963-307-5.
[15] D. Chandler, J. Chem. Phys. **68**, 2959 (1978).
[16] J. E. Adams and J. D. Doll, Surface Science **111**, 492 (1981).
[17] A. F. Voter and J. D. Doll, J. Chem. Phys. **82**, 80 (1985).
[18] G. Hummer, New Journal of Physics **7**, 34 (2005).
[19] N. Singhal and V. S. Pande, J. Chem. Phys. **123**, 204909 (2005).
[20] N. S. Hinrichs and V. S. Pande, J. Chem. Phys. **126**, 244101 (2007).
[21] F. Noé, M. Oswald, G. Reinelt, S. Fischer, and J. C. Smith, Multiscale Modeling and Simulation **5**, 393 (2006).
[22] F. Noé, M. Oswald, and G. Reinelt, in *Operations Research Proceedings*, edited by J. Kalcsics and S. Nickel (Springer, ADDRESS, 2007), pp. 435–440.
[23] N. G. van Kampen, *Stochastic processes in physics and chemistry*, 2nd ed. (Elsevier, ADDRESS, 1997).
[24] P. Metzner, F. Noé, and C. Schütte, Phys. Rev. E **80**, 021106 (2009).
[25] G. Hummer, New Journal of Physics **7**, 34 (2005).
[26] S. Sriraman, I. G. Kevrekidis, and G. Hummer, J. Phys. Chem. B **109**, 6479 (2005).
[27] F. Noé, C. Schütte, E. Vanden-Eijnden, L. Reich, and T. R. Weikl, accepted for Proc. Natl. Acad. Sci. USA on September 16 2009 (2009).
[28] H. Jeffreys, Proc. Royal Soc. A **186**, 453 (1946).
[29] P. Goyal, in *Bayesian Inference and Maximum Entropy Methods in Science and Engineering*, edited by K. H. Knuth, A. E. Abbas, R. D. Morriss, and J. P. Castle (American Institute of Physics, ADDRESS, 2005), pp. 366–373.

[30] X. Michalet, S. Weiss, and M. Jäger, Chem. Rev. **106**, 1785 (2006).
[31] B. Schuler, ChemPhysChem **6**, 1206 (2005).
[32] P. Tinnefeld and M. Sauer, Angewandte Chemie Intl. Ed. **44**, 2642 (2005).
[33] I.-C. Yeh and G. Hummer, Journal of the American Chemical Society **124**, 6563 (2002).
[34] F. Krieger, B. Fierz, O. Bieri, M. Drewello, and T. Kiefhaber, J. Mol. Biol. **332**, 265 (2003).
[35] H. Neuweiler, M. Löllmann, S. Doose, and M. Sauer, J. Mol. Biol. **365**, 856 (2007).
[36] A. C. Vaiana, H. Neuweiler, A. Schulz, J. Wolfrum, M. Sauer, and J. C. Smith, J. Am. Chem. Soc. **125**, 14564 (2003).
[37] W. F. van Gunsteren, S. R. Billeter, A. A. Eising, P. H. Hünenberger, P. Krueger, A. E. Mark, W. R. P. Scott, and I. G. Tironi, *Biomolecular Simulation: The GROMOS96 Manual and User Guide*.
[38] W. R. P. Scott, P. H. Hünenberger, I. G. Tironi, A. E. Mark, S. R. Billeter, J. Fennen, A. E. Torda, T. Huber, P. Kruger, and W. F. van Gunsteren, J. Phys. Chem. A **103**, 3596 (1999).
[39] H. J. C. Berendsen, D. van der Spoel, and R. van Druren, Comp. Phys. Comm. **91**, 43 (1995).
[40] E. Lindahl, B. Hess, and D. van der Spoel, J. Mol. Mod. **7**, 306 (2001).
[41] H. J. C. Berendsen, J. R. Grigera, and T. P. Straatsma, J. Phys. Chem. **91**, 6269 (1987).
[42] T. Morishita, J. Chem. Phys. **113**, 2976 (2000).
[43] M. D'Alessandro, A. Tenenbaum, and A. Amadei, J. Phys. Chem. B **106**, 5050 (2002).
[44] A. Mudi and C. Chakravarty, Mol. Phys. **102**, 681 (2004).
[45] H. J. C. Berendsen, J. P. M. Postma, W. F. van Gunsteren, A. DiNola, and J. R. Haak, J. Chem. Phys. **81**, 3684 (1984).
[46] B. Hess, H. Bekker, H. J. C. Berendsen, and J. G. E. M. Fraaije, Journal of Computational Chemistry **18**, 1463 (1997).
[47] T. Darden, D. York, and L. Pedersen, Journal of Chemical Physics **98**, 10089 (1993).
[48] J. D. Chodera, K. A. Dill, N. Singhal, V. S. Pande, W. C. Swope, and J. W. Pitera, Journal of Chemical Physics **126**, 155101 (2007).
[49] W. Kabsch, Acta Cryst. **A32**, 922 (1976).
[50] W. Kabsch, Acta Cryst. **A34**, 827 (1978).
[51] B. Steipe, Acta Cryst. **A58**, 506 (2002).
[52] D. L. Theobald, Acta Cryst. **A61**, 478 (2005).
[53] K. J. Ryan, Journal of Computational and Graphical Studies **12**, 585 (2003).