Theoretical restrictions on longest implicit time scales in Markov state models of biomolecular dynamics

Anton V. Sinitskiy, and Vijay S. Pande

Citation: The Journal of Chemical Physics **148**, 044111 (2018); doi: 10.1063/1.5005058
View online: https://doi.org/10.1063/1.5005058
View Table of Contents: http://aip.scitation.org/toc/jcp/148/4
Published by the American Institute of Physics

Theoretical restrictions on longest implicit time scales in Markov state models of biomolecular dynamics

Anton V. Sinitskiy[a] and Vijay S. Pande[a]
Department of Bioengineering, Stanford University, Stanford, California 94305, USA

(Received 16 September 2017; accepted 5 January 2018; published online 26 January 2018)

Markov state models (MSMs) have been widely used to analyze computer simulations of various biomolecular systems. They can capture conformational transitions much slower than an average or maximal length of a single molecular dynamics (MD) trajectory from the set of trajectories used to build the MSM. A rule of thumb claiming that the slowest implicit time scale captured by an MSM should be comparable by the order of magnitude to the aggregate duration of all MD trajectories used to build this MSM has been known in the field. However, this rule has never been formally proved. In this work, we present analytical results for the slowest time scale in several types of MSMs, supporting the above rule. We conclude that the slowest implicit time scale equals the product of the aggregate sampling and four factors that quantify: (1) how much statistics on the conformational transitions corresponding to the longest implicit time scale is available, (2) how good the sampling of the destination Markov state is, (3) the gain in statistics from using a sliding window for counting transitions between Markov states, and (4) a bias in the estimate of the implicit time scale arising from finite sampling of the conformational transitions. We demonstrate that in many practically important cases all these four factors are on the order of unity, and we analyze possible scenarios that could lead to their significant deviation from unity. Overall, we provide for the first time analytical results on the slowest time scales captured by MSMs. These results can guide further practical applications of MSMs to biomolecular dynamics and allow for higher computational efficiency of simulations. *Published by AIP Publishing.* https://doi.org/10.1063/1.5005058

I. INTRODUCTION

Markov state models (MSMs) have been one of the main tools for computational studies of biomolecules.[1–3] These models represent the dynamics of a biomolecule in the form of discrete transitions between a finite number of states. In MSMs, the system is assumed to be memoryless, that is, the probabilities of transition between Markov states depend only on the current state but not on the previous states of the system.

MSMs are particularly useful when multiple molecular dynamics (MD) trajectories are available.[4] In this case, an MSM of a simulated biomolecular system can integrate all the dynamic information from different trajectories, yielding estimates of the equilibrium and long-term dynamic properties of a biomolecular system under investigation. It has been known for a long time that this approach can capture slow dynamics of the system, with characteristic time scales much longer than lengths of single MD trajectories in the set of trajectories employed to construct an MSM.

How far in time scale can MSMs proceed in aggregating data from multiple trajectories? To the best of our knowledge, no strict answer to this question based on analytical derivations has ever been published. However, an empirical rule has been known in the community of specialists applying MSMs to study biomolecules, namely, that the slowest implicit time scale captured by an MSM can go up to the values on the same order of magnitude as the aggregate sampling achieved in the used MD dataset. In previous studies of specific biomolecular systems with MSMs based on MD simulations, reported longest implicit time scales turned out to be smaller than the aggregate sampling by at least one order of magnitude[5–17] or comparable to the aggregate sampling.[18–22]

In this paper, we analyze several MSMs with different topologies to uncover the reasons for the above empirical rule. These MSMs represent typical cases when slow dynamics is observed in biomolecular systems.

II. THEORY

A. Simplest two-state MSM of a rare conformational transition

A simplest two-state Markov model with rare transitions between two states (Fig. 1) can capture many aspects of the slowest-time scale dynamics of larger MSMs, while allowing for exact analytical solutions. State 1 in this model represents the native state of the biomolecule (corresponding to most Markov states in a multistate MSM with frequent transitions between them), while state 0 represents its metastable state [corresponding to an outlying Markov state(s)

[a]Author to whom correspondence should be addressed: sinitskiy@stanford.edu or pande@stanford.edu.

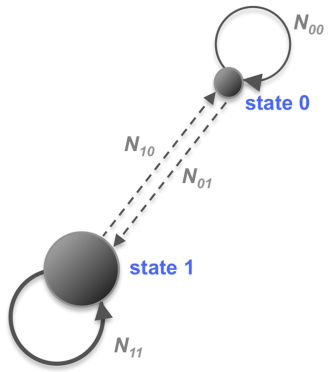

FIG. 1. Two-state MSM allows for exact analytical solutions, and can also serve as a simplified version of a multistate MSM with rare transitions between two subgraphs.

in a multistate MSM reached in only a small fraction of available MD trajectories].

For a given set of MD trajectories, the count matrix C equals

$$C = \begin{pmatrix} N_{00} & N_{01} \\ N_{10} & N - N_{00} - N_{01} - N_{10} \end{pmatrix}, \tag{1}$$

where N is the total number of observed transitions; N_{00}, N_{01}, and N_{10} are the numbers of transitions from state 0 to state 0, from state 0 to state 1, and from state 1 to state 0, respectively. To find the transition matrix T, we use here the direct symmetrization of the number of counts to enforce the reversibility of the transition matrix (the results derived without the symmetrization are considered in Appendix A). The transition matrix T assumes the following form:

$$T = \begin{pmatrix} 1 - \frac{\omega}{\omega_0 + \omega} & \frac{\omega}{\omega_0 + \omega} \\ \frac{\omega}{1 - \omega_0 - \omega} & 1 - \frac{\omega}{1 - \omega_0 - \omega} \end{pmatrix}, \tag{2}$$

where $\omega_0 = N_{00}/N$ is the relative sampling of state 0, and $\omega = (N_{01} + N_{10})/2N$ is the relative sampling of transitions between states 0 and 1. Due to the physical interpretation of the model,

$$\omega \ll \omega_0 < 1. \tag{3}$$

One of the eigenvalues of matrix T is $\lambda_0 = 1$ (corresponding to the stationary/equilibrium state), while the other is

$$\lambda_1 = 1 - \frac{\omega}{(\omega_0 + \omega)(1 - \omega_0 - \omega)}. \tag{4}$$

The corresponding implicit time scale τ_1, in the units of the lagtime $\tau_{lagtime}$, is

$$\frac{\tau_1}{\tau_{lagtime}} = -\frac{1}{\ln\left[1 - \frac{\omega}{(\omega_0+\omega)(1-\omega_0-\omega)}\right]}, \tag{5}$$

which, taking into consideration Eq. (3), simplifies to

$$\frac{\tau_1}{\tau_{lagtime}} \approx \frac{2\omega_0(1-\omega_0)N}{N_{01} + N_{10}}. \tag{6}$$

Based on Eq. (6), we formulate the following general expression for the longest implicit time scale computed from an MSM as the product

$$\tau_1 = f_{rare\ state} f_{transitions} f_{sw} T_s, \tag{7}$$

where T_s is the total sampling (in the units of time) and $f_{rare\ state}, f_{transitions}$, and f_{sw} are dimensionless factors. Specifically, $f_{rare\ state}$ is determined by the fraction of MD frames in which the system is in the rare state (state 0),

$$f_{rare\ state} = 2\omega_0(1 - \omega_0). \tag{8}$$

$f_{transitions}$ depends on how much sampling of the rare transition events is available,

$$f_{transitions} = \frac{1}{N_{01} + N_{10}}, \tag{9}$$

and f_{sw} is determined by the ratio of the lagtime to the output frequency in the used MD trajectories, and also may include a correction for the effect of counting transitions using a sliding window. The derivation above refers to the case where a sliding window is not used, and the intermediate points in the MD trajectories are ignored, and then an exact expression for f_{sw} can be obtained from Eqs. (6) and (7),

$$f_{sw} = \frac{\tau_{lagtime}}{T_s/N}. \tag{10}$$

In practice, either the lagtime is taken to be equal to the output frequency, implying $f_{sw} = 1$, or the lagtime is greater than the output frequency, and a sliding window is used. In the latter case, the transitions cannot be considered as statistically independent, and Eq. (10), corresponding to independent data, presents an overestimation. Therefore, in the case of a sliding window, we can give an estimate of

$$1 \le f_{sw} \le \frac{\tau_{lagtime}}{T_s/N}. \tag{11}$$

Currently, we do not have an analytical expression for f_{sw} in the case of a sliding window transition counts. We speculate that in this case a reasonable estimate might be

$$f_{sw} \sim \sqrt{\frac{\tau_{lagtime}}{T_s/N}}. \tag{12}$$

Detailed analysis of factors $f_{rare\ state}$, $f_{transitions}$, and f_{sw}, as well as the corresponding implications for using MSMs to study the dynamics of biomolecules, are provided below, in Sec. III.

Next, consider how an estimate of the longest time scale from a finite sample relates to the exact value of this time scale. Denote the exact probabilities of transition from state 0 to state 1 and from state 1 to state 0 per lagtime by Ω_{01} and Ω_{10}, respectively. These variables are similar to the above variables ω_{01} and ω_{10}, with the difference that Ω_{01} and Ω_{10} are the exact probabilities, while ω_{01} and ω_{10} are the corresponding estimates from a finite statistical sample (a given set of MD trajectories). We do not introduce a separate variable for the exact probability of staying in state 0 because, due to the physical interpretation of the model, there are much more events of staying in state 0 than transitioning between states 0 and 1, and therefore, the difference between the exact probability of staying in state 0 and its

estimate from a finite sample can be neglected in comparison to the differences between Ω_{01} and ω_{01} and between Ω_{10} and ω_{10}.

By analogy with Eq. (6), the exact implicit time scale, in the units of the lagtime, is given by

$$\frac{\tau_1^{exact}}{\tau_{lagtime}} = \frac{2\omega_0\,(1-\omega_0)}{\Omega_{01}+\Omega_{10}}. \tag{13}$$

On the other hand, the expectation value of the implicit time scale estimated from a finite sample is given by

$$\frac{\langle\tau_1\rangle}{\tau_{lagtime}} = 2\omega_0\,(1-\omega_0)\,N\left\langle\frac{1}{N_{01}+N_{10}}\right\rangle, \tag{14}$$

where the values of N_{01} and N_{10} are random variables following the Poisson distribution,

$$N_{01} \sim Pois\,(\Omega_{01}N)\,, \qquad N_{10} \sim Pois\,(\Omega_{10}N)\,, \tag{15}$$

and the average on the right-hand side of Eq. (14) is taken over the zero-truncated distribution (elementary events with $N_{01} = 0$ and simultaneously $N_{10} = 0$ are omitted because the corresponding MSMs do not provide a finite estimate of the implicit time scale). Therefore,

$$\frac{\langle\tau_1\rangle}{\tau_{lagtime}} = 2\omega_0\,(1-\omega_0)\,N\frac{\sum\limits_{\substack{N_{01},N_{10}=0\\ \neg(N_{01}=0\wedge N_{10}=0)}}^{\infty}\frac{1}{N_{01}+N_{10}}\frac{(\Omega_{01}N)^{N_{01}}}{N_{01}!}\frac{(\Omega_{10}N)^{N_{10}}}{N_{10}!}e^{-(\Omega_{01}+\Omega_{10})N}}{\sum\limits_{\substack{N_{01},N_{10}=0\\ \neg(N_{01}=0\wedge N_{10}=0)}}^{\infty}\frac{(\Omega_{01}N)^{N_{01}}}{N_{01}!}\frac{(\Omega_{10}N)^{N_{10}}}{N_{10}!}e^{-(\Omega_{01}+\Omega_{10})N}}. \tag{16}$$

The sums in the right-hand side of Eq. (16) can be computed by using the following integral representation:

$$\frac{1}{N_{01}+N_{10}} = \int_0^1 \frac{d\xi}{\xi}\xi^{N_{01}+N_{10}}, \tag{17}$$

switching the order of integration and summation in Eq. (16), and recognizing the Taylor series of the exponential function. As a result, the ratio of the average estimated time scale to its true value can be simplified to

$$\frac{\langle\tau_1\rangle}{\tau_1^{exact}} = \frac{ne^{-n}\left[\mathrm{Ei}(n)-\gamma-\ln(n)\right]}{1-e^{-n}}, \tag{18}$$

where $n = (\Omega_{01}+\Omega_{10})N$ is the average expected number of transitions between states 0 and 1, Ei is the exponential integral function, and γ is the Euler–Mascheroni constant. Detailed analysis of the function on the right-hand side of Eq. (18) is provided in Sec. III.

To sum up, the simplest two-state MSM with rare transitions reveals many important aspects of the relationship between the slowest time scale estimated from finite data, the exact value of this time scale, and the aggregate sampling. But how transferable these results are to MSMs with more states? We explore this question in Secs. II B–II D.

B. Perturbative analysis of an arbitrarily large MSM with a rarely sampled state

Consider an MSM with N_B closely connected states and one more state with rare transitions to/from it (Fig. 2; the notation "B" for the set of closely connected states is used for comparability of the formulas below with the analysis in Sec. II C). Physically, this model may correspond to a biomolecule with two distinct functional states. One of these functional states is structurally and dynamically diverse and has to be represented by a large number of Markov states forming a dense subgraph, while the other functional state is homogeneous and can be reasonably approximated by a single Markov state. For example, a simple MSM of folding and unfolding of a polypeptide or a protein fragment may include a single Markov state for the folded functional state and multiple Markov states for various unfolded conformations. Besides possible practical interpretations, the model presented in this section is designed to help the reader follow a more technically complicated analysis of the third model presented later in Sec. II C. Without loss of generality, we denote the rarely sampled state as state 0, and all other states as states $1, \ldots, N_B$.

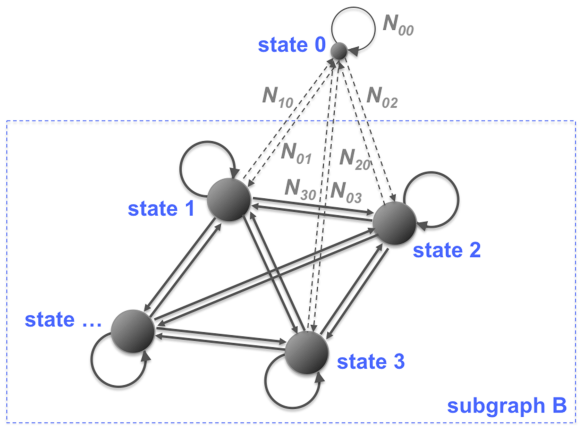

FIG. 2. Multistate MSM with rare transitions between one state (state 0) and all other states is a typical situation where long implicit time scales emerge in practice.

The count matrix can be written as

$$C = \begin{pmatrix} N_{00} & N_{01} & N_{02} & \cdots & N_{0N_B} \\ N_{10} & N_{11} & N_{12} & \cdots & N_{1N_B} \\ N_{20} & N_{21} & N_{22} & & \\ \vdots & & & \ddots & \\ N_{N_B0} & N_{N_B1} & & & N_{N_BN_B} \end{pmatrix}, \quad (19)$$

where N_{ij} is the number of observed transitions from state i to state j. Then, the transition matrix T can be written in a block-diagonal form as follows (the direct symmetrization of the number of counts is implied):

$$T = \begin{pmatrix} 1 - \dfrac{\omega}{\omega_0 + \bar{\omega}} & T^{0B} \\ T^{B0} & T^{BB} \end{pmatrix}, \quad (20)$$

where T^{B0} is a column formed by the elements

$$T_i^{B0} = \frac{(\omega_{0i} + \omega_{i0})/2}{(\omega_{0i} + \omega_{i0})/2 + \frac{1}{N}\sum\limits_{j=1}^{N_{st}} \frac{N_{ij}+N_{ji}}{2}}, \quad i = 1, ..., N_B, \quad (21)$$

T^{0B} is a row formed by the elements

$$T_i^{0B} = \frac{(\omega_{0i} + \omega_{i0})/2}{\omega_0 + \bar{\omega}}, \quad i = 1, \ldots, N_B, \quad (22)$$

T^{BB} is an $N_B \times N_B$ square matrix formed by the elements

$$T_{ij}^{BB} = \frac{\left(N_{ij} + N_{ji}\right)/2}{N\,(\omega_{0i} + \omega_{i0})/2 + \sum\limits_{k=1}^{N_{st}} (N_{ik} + N_{ki})/2}, \quad (23)$$

$$i = 1, \ldots, N_B, \quad j = 1, \ldots, N_B,$$

$\omega_0 = N_{00}/N$ is the relative sampling of state 0, $\omega_{ij} = N_{ij}/N$, and $\bar{\omega}$ is the relative sampling of transitions to or from state 0,

$$\bar{\omega} = \sum_{j=1}^{N_{st}} \frac{\omega_{0i} + \omega_{i0}}{2}. \quad (24)$$

If transitions to and from state 0 are relatively rare,

$$\omega_{0i} \ll 1, \quad \omega_{i0} \ll 1, \quad i = 1, \ldots, N_B, \quad (25)$$

then eigenvalues of the transition matrix T, and therefore the implicit time scales, can be found by perturbation expansion. First, define the zeroth-order terms. In the limit of

$$\omega_{0i} \to 0, \quad \omega_{i0} \to 0, \quad i = 1, \ldots, N_B, \quad (26)$$

the transition matrix reduces to

$$T \to \tilde{T} = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & & & \\ \vdots & & \tilde{T}^{BB} & \\ 0 & & & \end{pmatrix}, \quad (27)$$

where $\tilde{T}^B$ is an $N_B \times N_B$ square matrix formed by the elements

$$\tilde{T}_{ij}^{BB} = \frac{N_{ij} + N_{ji}}{\sum\limits_{k=1}^{N_{st}} N_{ik} + N_{ki}}, \quad i = 1, \ldots, N_B, \quad j = 1, \ldots, N_B. \quad (28)$$

Matrix $\tilde{T}$ has two eigenvalues equal to 1, one of which corresponds to the stationary state in the subsystem of states 1, ..., N_B, and the other to the absence of transitions (infinitely large characteristic time scale) between state 0 and the other states. The corresponding left and right eigenvectors are

$$\begin{aligned} v_0^\dagger &= \left(0\ \tilde{p}_1 \cdots \tilde{p}_{N_B}\right), \quad w_0 = \begin{pmatrix} 0 \\ 1 \\ \vdots \\ 1 \end{pmatrix}, \\ v_1^\dagger &= \left(1\ 0 \cdots 0\right), \quad w_1 = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \end{aligned} \quad (29)$$

where $\tilde{p}_1, \ldots, \tilde{p}_{N_B}$ are the occupations of states 1, ..., N_B in subgraph B at equilibrium in the absence of transitions between state 0 and subgraph B. The eigenvectors satisfy the conditions

$$\begin{aligned} &\tilde{T}w_0 = w_0, \quad \tilde{T}w_1 = w_1, \quad v_0^\dagger \tilde{T} = v_0^\dagger, \quad v_1^\dagger \tilde{T} = v_1^\dagger, \\ &v_0^\dagger w_0 = 1, \quad v_1^\dagger w_1 = 1, \quad v_0^\dagger w_1 = 0, \quad v_1^\dagger w_0 = 0. \end{aligned} \quad (30)$$

All other eigenvalues of matrix $\tilde{T}$ are less than 1, corresponding to finite implicit time scales of transitions in the subsystem of states 1, ..., N_B (we assumed that this subsystem is fully connected). Perturbations to the first two eigenvalues can be found as the eigenvalues of the matrix

$$\begin{pmatrix} v_0^\dagger \left(T - \tilde{T}\right) w_0 & v_0^\dagger \left(T - \tilde{T}\right) w_1 \\ v_1^\dagger \left(T - \tilde{T}\right) w_0 & v_1^\dagger \left(T - \tilde{T}\right) w_1 \end{pmatrix} = \begin{pmatrix} -\dfrac{\bar{\omega}}{\omega_0} & \dfrac{\bar{\omega}}{\omega_0} \\ \dfrac{\bar{\omega}}{1-\omega_0} & -\dfrac{\bar{\omega}}{1-\omega_0} \end{pmatrix} \quad (31)$$

[the right-hand side contains the leading terms computed with the use of Eq. (25)]. The eigenvalues of this matrix are

$$\delta\lambda_0 = 0, \quad \delta\lambda_1 = -\frac{\bar{\omega}}{\omega_0\,(1-\omega_0)}. \quad (32)$$

As expected, one of these corrections is zero since the perturbed matrix still must have one eigenvalue exactly equal to 1, corresponding to the stationary state of the whole system. The other correction is nonzero, leading to a finite implicit time scale,

$$\frac{\tau_1}{\tau_{lagtime}} = \frac{2\omega_0\,(1-\omega_0)\,N}{\sum\limits_{i=1}^{N_{st}} (N_{0i} + N_{i0})} \quad (33)$$

[again, an approximation based on Eq. (25) was used]. This result is strikingly simple and similar to Eq. (6) derived for a two-state MSM. The only difference is that instead of the total number of transitions between states 0 and 1, this expression contains the total number of transitions to or from state 0. No properties of the subsystem of states 1, ..., N_B, such as the equilibrium populations of each of these states or probabilities of transitions between them, are left in the final result, Eq. (33),

even though they were present in the intermediate formulas. Consequently, the general relationship between the slowest implicit time scale and the total sampling, Eq. (7), holds in this case, with the only change that the factor $f_{transitions}$ now assumes the following form:

$$f_{transitions} = \frac{1}{\sum_{i=1}^{N_B} (N_{0i} + N_{i0})}. \tag{34}$$

Moreover, the analysis of the relationship between the exact value of the implicit time scale and its estimates from finite data can be directly extended to this case, too. The expectation value of the time scale can be expressed as

$$\frac{\langle \tau_1 \rangle}{\tau_{lagtime}} = 2\omega_0 (1-\omega_0) N \left\langle \frac{1}{\sum_{i=1}^{N_B} (N_{0i} + N_{i0})} \right\rangle, \tag{35}$$

where the expectation value on the right-hand side of Eq. (35) is taken over zero-truncated Poisson distributions of the variables N_{0i} and N_{i0},

$$N_{0i} \sim Pois\,(\Omega_{0i} N)\,, \quad N_{i0} \sim Pois\,(\Omega_{i0} N)\,, \quad i = 1, \ldots, N_B, \tag{36}$$

where Ω_{0i} and Ω_{i0} are the exact probabilities of transitions from state 0 to state i or vice versa per lagtime. Then, Eq. (35) can be rewritten as

$$\frac{\langle \tau_1 \rangle}{\tau_{lagtime}} = 2\omega_0 (1-\omega_0) N \frac{\sum_{\substack{N_{0i}, N_{i0}=0 \\ \neg\forall i (N_{0i}0 \wedge N_{i0}=0)}}^{\infty} \frac{1}{\sum_{i=1}^{N_B} (N_{0i}+N_{i0})} \prod_{i=1}^{N_B} \frac{(\Omega_{0i}N)^{N_{0i}}}{N_{0i}!} \frac{(\Omega_{i0}N)^{N_{i0}}}{N_{i0}!} e^{-\sum_{i=1}^{N_B} (\Omega_{0i}+\Omega_{i0})N}}{\sum_{\substack{N_{0i}, N_{i0}=0 \\ \neg\forall i (N_{0i}0 \wedge N_{i0}=0)}}^{\infty} \prod_{i=1}^{N_B} \frac{(\Omega_{0i}N)^{N_{0i}}}{N_{0i}!} \frac{(\Omega_{i0}N)^{N_{i0}}}{N_{i0}!} e^{-\sum_{i=1}^{N_B} (\Omega_{0i}+\Omega_{i0})N}}. \tag{37}$$

Computing the sums leads to the result given by Eq. (18), with the following definition of n as the expectation value of the total number of transitions to or from state 0:

$$n = N \sum_{i=1}^{N_B} (\Omega_{0i} + \Omega_{i0}). \tag{38}$$

Thus, the analysis of the bias in the estimated slowest implicit time scale introduced by a finite sampling is also transferable from the simplest two-state MSM to an MSM with multiple states if transitions to and from one of these states are rarely sampled.

C. Perturbative analysis of a large MSM with rare transitions between two subgraphs

Another typical case leading to large implicit time scales is the existence of two (or more) subgraphs with rare transitions between them (Fig. 3). Physically, this case corresponds to a biomolecule switching between two functional states (active/inactive, closed/open, etc.), with each of these states being structurally and dynamically heterogeneous, and therefore represented each by a set of Markov states. Perturbative analysis of this model can be performed similar to the analysis presented in Sec. II B.

The count matrix can be written in a block-diagonal form

$$C = \begin{pmatrix} C^{AA} & C^{AB} \\ C^{BA} & C^{BB} \end{pmatrix}, \tag{39}$$

where matrices C^{AA}, C^{AB}, C^{BA}, and C^{BB} contain the number of transition between different states within subgraph A, from states in subgraph A to states in subgraph B, from states in subgraph B to states in subgraph A, and between states within subgraph B, respectively. Then, the transition matrix also assumes the block-diagonal form

$$T = \begin{pmatrix} T^{AA} & T^{AB} \\ T^{BA} & T^{BB} \end{pmatrix}. \tag{40}$$

As the unperturbed state, we choose the MSM without transitions between subgraphs A and B,

$$\tilde{T} = \begin{pmatrix} \tilde{T}^{AA} & 0 \\ 0 & \tilde{T}^{BB} \end{pmatrix}, \tag{41}$$

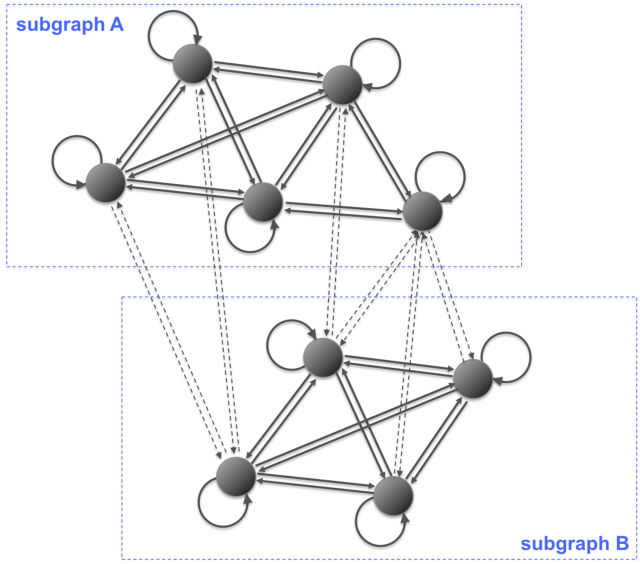

FIG. 3. Multistate MSM with rare transitions between two subgraphs is another typical situation where long implicit time scales emerge in practice.

where

$$\tilde{T}^{AA} = \lim_{C^{AB},C^{BA}\to 0} T^{AA}, \quad \tilde{T}^{BB} = \lim_{C^{AB},C^{BA}\to 0} T^{BB}. \tag{42}$$

If each of subgraphs A and B is connected, then the matrix $\tilde{T}$ has exactly two eigenvalues equal to 1, corresponding to the stationary states of isolated subgraphs A and B, while all other eigenvalues of $\tilde{T}$ are less than 1. The left and right eigenvectors corresponding to the first two eigenvalues are

$$v_0^\dagger = \left(\tilde{p}^A \ \vec{0}^\dagger\right), \quad w_0 = \begin{pmatrix}\vec{1}\\ \vec{0}\end{pmatrix}, \quad v_1^\dagger = \left(\vec{0}^\dagger \ \tilde{p}^B\right), \quad w_1 = \begin{pmatrix}\vec{0}\\ \vec{1}\end{pmatrix}, \tag{43}$$

where $\vec{0}$ and $\vec{1}$ are columns with all elements equal to 0 or 1, respectively, and $\tilde{p}^A$ and $\tilde{p}^B$ are stationary (equilibrium) distributions for subgraphs A and B in the absence of transitions between these two subgraphs,

$$\tilde{p}^A \tilde{T}_{AA} = \tilde{p}^A, \qquad \tilde{p}^B \tilde{T}_{BB} = \tilde{p}^B. \tag{44}$$

Then, perturbations to the first two eigenvalues can be found as the eigenvalues of the matrix

$$\begin{pmatrix} v_0^\dagger \left(T-\tilde{T}\right) w_0 & v_0^\dagger \left(T-\tilde{T}\right) w_1 \\ v_1^\dagger \left(T-\tilde{T}\right) w_0 & v_1^\dagger \left(T-\tilde{T}\right) w_1 \end{pmatrix} = \begin{pmatrix} -\frac{N_{AB}+N_{BA}}{2N_{AA}} & \frac{N_{AB}+N_{BA}}{2N_{AA}} \\ \frac{N_{AB}+N_{BA}}{2N_{BB}} & -\frac{N_{AB}+N_{BA}}{2N_{BB}} \end{pmatrix}, \tag{45}$$

where N_{AA}, N_{AB}, N_{BA}, and N_{BB} are the total number of transitions between different states within subgraph A, from states in subgraph A to states in subgraph B, from states in subgraph B to states in subgraph A, and between states within subgraph B, respectively,

$$N_{AA} = \sum_{i_A=1}^{N_{st,A}} \sum_{j_A=1}^{N_{st,A}} C^{AA}_{i_A j_A}, \quad N_{AB} = \sum_{i_A=1}^{N_{st,A}} \sum_{j_B=1}^{N_{st,B}} C^{AB}_{i_A j_B}, N_{BA} = \sum_{i_B=1}^{N_{st,B}} \sum_{j_A=1}^{N_{st,A}} C^{BA}_{i_B j_A}, \quad N_{BB} = \sum_{i_B=1}^{N_{st,B}} \sum_{j_B=1}^{N_{st,B}} C^{BB}_{i_B j_B}. \tag{46}$$

For example,

$$v_0^\dagger \left(T-\tilde{T}\right) w_0 = \sum_{i_A=1}^{N_{st,A}} \sum_{j_A=1}^{N_{st,A}} \tilde{p}^A_{i_A} \left(T^{AA}_{i_A j_A} - \tilde{T}^{AA}_{i_A j_A}\right) = \sum_{i_A=1}^{N_{st,A}} \sum_{j_A=1}^{N_{st,A}} \tilde{p}^A_{i_A} \left(\frac{\frac{C^{AA}_{i_A j_A}+C^{AA}_{j_A i_A}}{2}}{\sum_{k_A=1}^{N_{st,A}} \frac{C^{AA}_{i_A k_A}+C^{AA}_{k_A i_A}}{2} + \sum_{k_B=1}^{N_{st,B}} \frac{C^{AB}_{i_A k_B}+C^{BA}_{k_B i_A}}{2}} - \frac{\frac{C^{AA}_{i_A j_A}+C^{AA}_{j_A i_A}}{2}}{\sum_{k_A=1}^{N_{st,A}} \frac{C^{AA}_{i_A k_A}+C^{AA}_{k_A i_A}}{2}} \right)$$

$$= -\sum_{i_A=1}^{N_{st,A}} \tilde{p}^A_{i_A} \frac{\sum_{k_B=1}^{N_{st,B}} \frac{C^{AB}_{i_A k_B}+C^{BA}_{k_B i_A}}{2}}{\sum_{k_A=1}^{N_{st,A}} \frac{C^{AA}_{i_A k_A}+C^{AA}_{k_A i_A}}{2} + \sum_{k_B=1}^{N_{st,B}} \frac{C^{AB}_{i_A k_B}+C^{BA}_{k_B i_A}}{2}}. \tag{47}$$

If transitions between the two subgraphs are rare in comparison to transitions within each subgraph, the leading term in Eq. (47) is

$$v_0^\dagger \left(T-\tilde{T}\right) w_0 \approx -\sum_{i_A=1}^{N_{st,A}} \tilde{p}^A_{i_A} \frac{\sum_{k_B=1}^{N_{st,B}} \frac{C^{AB}_{i_A k_B}+C^{BA}_{k_B i_A}}{2}}{\sum_{k_A=1}^{N_{st,A}} \frac{C^{AA}_{i_A k_A}+C^{AA}_{k_A i_A}}{2}} = -\sum_{i_A=1}^{N_{st,A}} \tilde{p}^A_{i_A} \frac{\sum_{k_B=1}^{N_{st,B}} \frac{C^{AB}_{i_A k_B}+C^{BA}_{k_B i_A}}{2}}{\tilde{p}^A_{i_A} \sum_{k_A=1}^{N_{st,A}} \sum_{l_A=1}^{N_{st,A}} \frac{C^{AA}_{l_A k_A}+C^{AA}_{k_A l_A}}{2}} = -\frac{\sum_{i_A=1}^{N_{st,A}} \sum_{k_B=1}^{N_{st,B}} \frac{C^{AB}_{i_A k_B}+C^{BA}_{k_B i_A}}{2}}{\sum_{k_A=1}^{N_{st,A}} \sum_{l_A=1}^{N_{st,A}} \frac{C^{AA}_{l_A k_A}+C^{AA}_{k_A l_A}}{2}} = -\frac{N_{AB}+N_{BA}}{2N_{AA}}. \tag{48}$$

The other elements of the matrix in Eq. (45) can be computed in a similar way. Then, the eigenvalues of this matrix are

$$\delta\lambda_0 = 0, \qquad \delta\lambda_1 = -\frac{(N_{AB}+N_{BA})(N_{AA}+N_{BB})}{2N_{AA}N_{BB}} \approx -\frac{(N_{AB}+N_{BA})N}{2N_{AA}N_{BB}}, \tag{49}$$

where N is the total number of observed transitions,

$$N = N_{AA} + N_{BB} + N_{AB} + N_{BA} \approx N_{AA} + N_{BB}. \tag{50}$$

As expected, one of the eigenvalues stays unchanged and equals to 1, corresponding to the stationary state of the whole system. The correction to the other eigenvalue is nonzero, leading to a finite implicit time scale,

$$\frac{\tau_1}{\tau_{lagtime}} = \frac{2\omega_A\omega_B N}{N_{AB}+N_{BA}}, \tag{51}$$

where $\omega_A = N_{AA}/N$ and $\omega_B = N_{BB}/N$. This result is also similar to the expression for the implicit time scale derived from the simplest two-state MSM, Eq. (6). The subsequent analysis,

including the relationship between the slowest implicit time scale and the total sampling, Eq. (7), holds in this case, too, with the factor $f_{transitions}$ now assuming the following form:

$$f_{transitions} = \frac{1}{N_{AB} + N_{BA}}. \tag{52}$$

The analysis of the relationship between the exact value of the implicit time scale and its estimates from finite datasets can also be directly extended to this case, leading to the result given in Eq. (18), with the following definition of n as the expectation value of the total number of transitions between subgraphs A and B:

$$n = N\left(\sum_{i_A=1}^{N_{st,A}}\sum_{j_B=1}^{N_{st,B}}\Omega_{i_A j_B}^{AB} + \sum_{i_B=1}^{N_{st,B}}\sum_{j_A=1}^{N_{st,A}}\Omega_{i_B j_A}^{BA}\right), \tag{53}$$

where Ω values are the exact probabilities of transitions between specific states in subgraphs A and B.

Among four models analyzed in this paper, the model described in this Sec. II C is arguably the most typical for practical applications. A numerical illustration of this model for the case of the terminally blocked alanine peptide is given below in Appendix B.

D. Three-state MSM with rare transitions to two states

The analysis in Secs. II A–II C referred to MSMs with one slow implicit time scale. In this section, we analyze cooperative effects between two different sets of rare transitions, leading to the appearance of two slow implicit time scales. Taking into consideration the results of Secs. II B and II C, where we demonstrated that the relationships for the slow implicit time scale are transferrable from the simplest two-state MSM to more general models, in this section, we limit ourselves to the analysis of a three-state MSM, assuming that the results will be qualitatively applicable to more general models with two slow implicit time scales and an arbitrary number of Markov states.

This model (Fig. 4) includes three states: state A, corresponding to the native geometry of a biomolecule, and states B and C, which correspond to two metastable states of the biomolecule. Transitions between A and B and between A and C are rare, and there are no direct transitions between B and C. In practice, this model can be used to describe a biomolecule with three functional states, each of which is structurally and dynamically homogeneous. The count matrix can be written as

$$C = \begin{pmatrix} N_{AA} & N_{AB} & N_{AC} \\ N_{BA} & N_{BB} & 0 \\ N_{CA} & 0 & N_{CC} \end{pmatrix}, \tag{54}$$

where N_{ij} is the number of observed transitions from state i to state j. The transition matrix T (after a direct symmetrization of the number of counts) assumes the following form:

$$T = \begin{pmatrix} 1-\alpha_{AB}-\alpha_{AC} & \alpha_{AB} & \alpha_{AC} \\ \alpha_{BA} & 1-\alpha_{BA} & 0 \\ \alpha_{CA} & 0 & 1-\alpha_{CA} \end{pmatrix}, \tag{55}$$

where

$$\begin{aligned} \alpha_{AB} &= \frac{N_{AB}+N_{BA}}{2N_{AA}+N_{AB}+N_{BA}+N_{AC}+N_{CA}}, \quad \alpha_{AC} = \frac{N_{AC}+N_{CA}}{2N_{AA}+N_{AB}+N_{BA}+N_{AC}+N_{CA}}, \\ \alpha_{AB} &= \frac{N_{AB}+N_{BA}}{2N_{BB}+N_{AB}+N_{BA}}, \quad \alpha_{AC} = \frac{N_{AC}+N_{CA}}{2N_{CC}+N_{AC}+N_{CA}}. \end{aligned} \tag{56}$$

The transition matrix has the following eigenvalues:

$$\lambda_0 = 1, \quad \lambda_{1,2} = 1 - \frac{(\alpha_{AB}+\alpha_{BA}+\alpha_{AC}+\alpha_{CA}) \pm \sqrt{(\alpha_{AB}+\alpha_{BA}-\alpha_{AC}-\alpha_{CA})^2 + 4\alpha_{AB}\alpha_{AC}}}{2}, \tag{57}$$

and hence, the exact expressions for the two implicit time scales are

$$\frac{\tau_{1,2}}{\tau_{lagtime}} = -\frac{1}{\ln\left[1 - \frac{(\alpha_{AB}+\alpha_{BA}+\alpha_{AC}+\alpha_{CA}) \pm \sqrt{(\alpha_{AB}+\alpha_{BA}-\alpha_{AC}-\alpha_{CA})^2 + 4\alpha_{AB}\alpha_{AC}}}{2}\right]}. \tag{58}$$

Unlike previous cases, it does not seem possible to analytically simplify Eq. (58) for small N_{AB}/N, N_{BA}/N, N_{AC}/N, N_{CA}/N, and finite $\omega_B = N_{BB}/N$ and $\omega_C = N_{CC}/N$. However, in the practically important case of

$$\frac{N_{AB}}{N}, \frac{N_{BA}}{N}, \frac{N_{AC}}{N}, \frac{N_{CA}}{N} \ll \omega_B, \omega_C \ll 1, \tag{59}$$

Eq. (58) simplifies to

$$\begin{aligned} \frac{\tau_1}{\tau_{lagtime}} &= \frac{2\omega_B(1-\omega_B)N}{N_{AB}+N_{BA}}\left(1+O\left(\omega^2\right)\right), \\ \frac{\tau_2}{\tau_{lagtime}} &= \frac{2\omega_C(1-\omega_C)N}{N_{AC}+N_{CA}}\left(1+O\left(\omega^2\right)\right), \end{aligned} \tag{60}$$

where the small parameter $\omega = \max(\omega_B, \omega_C)$. These expressions for the implicit time scales are similar to the result for the two-state MSM, Eq. (6), and demonstrate decoupling of rare transitions from each other, at least in the case of rare sampling of the corresponding metastable states.

The analysis of the bias in the estimates of implicit time scales can be performed similar to the cases analyzed in

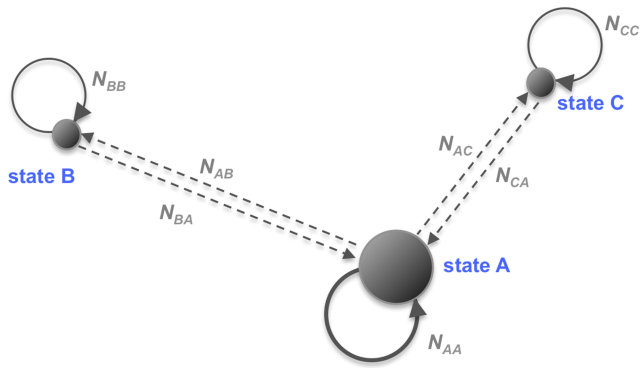

FIG. 4. Three-state MSM with rare transitions between state A and state B, and state A and state C can serve as a simplest example allowing for an exact analysis of coupling of two processes with slow time scales.

Secs. II A–II C, leading to the results similar to Eq. (18), namely,

$$\frac{\langle \tau_1 \rangle}{\tau_1^{exact}} = \frac{n_1 e^{-n_1} \left[\mathrm{Ei}(n_1) - \gamma - \ln(n_1)\right]}{1 - e^{-n_1}}, \; n_1 = N\left(\Omega_{AB} + \Omega_{BA}\right),$$
$$\frac{\langle \tau_2 \rangle}{\tau_2^{exact}} = \frac{n_2 e^{-n_2} \left[\mathrm{Ei}(n_2) - \gamma - \ln(n_2)\right]}{1 - e^{-n_2}}, \; n_2 = N\left(\Omega_{AC} + \Omega_{CA}\right), \tag{61}$$

where Ω values are the exact probabilities of transitions between the corresponding states.

III. DISCUSSION

In all the four models analyzed above, the slowest implicit time scale(s) of rarely sampled conformational transitions has been expressed as a product of the aggregate sampling T_s and three dimensionless factors,

$$\tau = f_{rare\ state} f_{transitions} f_{sw} T_s. \tag{62}$$

Moreover, it has been demonstrated that this estimate from a finite MD dataset is, in general, biased, while the unbiased estimate is given by

$$\tau^{exact} = f_{rare\ state} f_{transitions} f_{sw} f_{bias} T_s, \tag{63}$$

where f_{bias} is defined as the ratio of the exact slowest implicit time scale τ^{exact} of a rarely sampled conformational transition to its value τ estimated from an MSM based on finite sampling.

Consider now what might be typical values of these four dimensionless factors and their product in most cases, and what scenario may lead to significantly greater or smaller values of this product.

The first factor $f_{rare\ state}$ is determined by the shares of MD frames in which the system is in one or the other subgraph. Its maximum value is 1/2, which is achieved when half of all sampled frames refer to one subgraph, and the other half to the other subgraph. If the metastable state is less sampled, which is a typical situation, then $f_{rare\ state}$ is approximately equal to the fraction of frames in this metastable state (ω_0 in Secs. II A and II B, ω_A or ω_B in Sec. II C, or ω_B or ω_C in Sec. II D). In such cases, the longest time scale is limited by the time the system spent in the rare state, rather than the aggregate sampling for all possible states.

The maximum value of $f_{transitions}$ is 1, and it is achieved when only one transition between two subgraphs occurred in the simulations. This factor falls to 1/2 if two transitions are observed in either direction, or one transition in each direction. With an increase in the sampling, $f_{transitions}$ decreases, but not so drastically, and remains on the order of unity as long as the total number of observed rare transitions stays a single-digit number.

The third factor f_{sw} is greater than 1 if a sliding window is used for counting transitions and equals 1 otherwise. In practice, in most situations, the lagtime is ~10-100 times greater than the output frequency, a sliding window is used, and therefore, based on the assumption of Eq. (12), $f_{sw} \sim$ 3-10.

Finally, the bias f_{bias} in the time scale estimate, according to Eqs. (18) and (61), can be written as

$$f_{bias} = \frac{1 - e^{-n}}{n e^{-n} \left[\mathrm{Ei}(n) - \gamma - \ln(n)\right]}. \tag{64}$$

The right-hand side of Eq. (64) monotonically decreases from $+\infty$ to 0.7574 when n goes from 0 to 3.75, and after that monotonically increases, with the asymptotic value of 1 reached in the limit of $n \to \infty$ (Fig. 5, Table I).

The asymptotic behavior of f_{bias} in the limit of small n is given by

$$f_{bias} \underset{n \to 0}{\sim} \frac{1}{n} + \frac{1}{4} + O(n), \tag{65}$$

and therefore, in the regime when transitions are rare compared to simulated time scales, the estimates of the implicit time scales from MSMs are significantly underestimated (e.g., by a factor of ~10 when the probability of observing a transition is 0.1). This happens because of the truncation of MSMs with no transitions between the states corresponding to this implicit time scale. The average value of $\langle 1/N_{01} + N_{10} \rangle$ over the non-truncated models reduces to 1, as evident from Eq. (16), while the exact value of this average in the chosen notations is $1/n$.

The large n limit of f_{bias} is

$$f_{bias} \underset{n \to \infty}{\sim} 1 - \frac{1}{n} + O\left(\frac{1}{n^2}\right). \tag{66}$$

Though the bias vanishes in the limit of infinitely many transitions, the speed with which it happens is not so high. Even at $n = 10$, which is a good sampling of a rare transition from

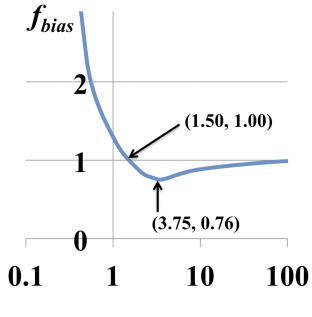

FIG. 5. The dependence of the bias f_{bias} on the expected number of rare transitions n has a nontrivial character, as follows from Eq. (64).

TABLE I. Correction f_{bias} to a slow implicit time scale of a Markov state model as a function of the expectation value of the numbers of transitions n to and from the corresponding rarely sampled state or subgraph. f_{bias} is defined as the ratio of the exact value of the implicit time scale to the expectation value of its estimate from MSMs based on finite sampling. The exact expression for f_{bias} is given by Eq. (64). Values at $n < 0.1$ and $n > 10$ can be estimated with the use of asymptotic expansions, Eqs. (65) and (66).

n	f_{bias}
0.1	10.255
0.5	2.276
1	1.304
2	0.867
3	0.770
4	0.758
5	0.776
6	0.802
7	0.828
8	0.851
9	0.870
10	0.885

the practical viewpoint, the bias is still noticeable (~10%). In Appendix C, we demonstrate that another approach to determining the slowest characteristic time scale, namely, fitting an exponential, has the same $1/n$ term in the large n series for the bias. Therefore, the approach to calculating the time scale used in this work, based on the diagonalization of the transition matrix, is not interior relative to other approaches.

Interestingly, the estimate of the implicit time scale is unbiased not only in the limit of infinite sampling, due to the law of large numbers, but also at $n = 1.503$, due to a cancellation of the effects of zero-truncation in the Poisson distributions on the numerator and denominator in Eqs. (16) and (37) and similar equations for MSMs with more states. In the practically important range of $n \sim$ 1–10, the bias may be positive (at $n < 1.503$) or negative ($n > 1.503$), reaching +30% at $n = 1$ and −24% at $n = 3.75$. This analysis of the systematic bias is based on the approximation of rare transitions between subgraphs in various types of MSMs [Eq. (3), (25), or (59)], which allowed us to proceed with analytical derivations beyond the results previously published in the literature[23–27] and obtain the analytical result given by Eq. (64).

Overall, the product of the factors $f_{rare\ state}$, $f_{transitions}$, f_{sw} and, if the unbiased estimate of the time scale is required, f_{bias}, in many practical situations may be ~1, plus/minus one order of magnitude, and hence, the longest implicit time scale estimated from an MSM is indeed comparable by the order of magnitude to the total sampling used to build this MSM. This theoretical scenario seems to have been realized in numerous publications on specific biomolecular systems, including Refs. 18–22. Another scenario mentioned above is the case of a rarely sampled metastable state, implying $f_{rare\ state} f_{transitions} f_{sw} f_{bias} \sim f_{rare\ state}$, and therefore, the longest implicit time scale comparable to the time the system spent in this rare state (which is much less than the total sampling in all used MD trajectories). This scenario seems to have been realized in other publications, such as Refs. 5–17.

Practical recommendations for sampling longer time scales in MSMs follow from our theoretical analysis. To achieve statistically robust and unbiased estimates of longer time scales captured by an MSM, given the total sampling (e.g., set by an available allocation on a supercomputer), do the following:

1. *Sample rarely visited state(s) or subgraph(s) more*. This can be achieved in two mutually non-exclusive ways:
   (a) Use a larger set of diverse geometries for the initialization of MD simulations. These geometries can be based on experimentally determined (e.g., X-ray, NMR, or cryoEM) structures, or modeled by homology[28–30] or *ab initio* structure prediction from the primary amino acid sequence.[31–33]
   (b) Reseed MD simulations (preferably in many cycles), starting additional MD trajectories from the metastable state(s) rarely sampled in previously generated MD trajectories.[34,35]
2. *Do not oversample rare transitions for which the implicit time scale is sought for*. Though a certain level of sampling of such transitions is required to ensure the statistical robustness of the estimated time scale, unrestricted increase in the sampling would undermine the practical efficiency of the model. Based on the formulas for $f_{transitions}$ and f_{bias} derived in this paper, Eqs. (9), (34), (52), and (64), we estimate the optimal number of transitions to be on the order of $n \sim$ 1-10, corresponding to $f_{transitions} \sim$ 0.1-1 and $f_{bias} \sim 1$. This recommendation, which may seem at first counterintuitive, is illustrated by the case of the terminally blocked alanine peptide in Appendix B.
3. *Use a sliding window for counting transitions*. Though we do not present exact analytical results on f_{sw} in this paper, our estimate given by Eq. (12) assumes that the output frequency in MD trajectories should be chosen ~10-100 times smaller than the MSM lagtime (more precisely, the time scale on which the studied system can be treated as Markovian) and, respectively, every cycle in the sliding window should correspond to ~10-100 time steps in the trajectories. Smaller time steps improve the value of the factor f_{sw} only sublinearly but linearly increase the memory required to store the generated MD trajectories.
4. Correct for the systematic bias in the implicit time scales that appears because the time scales are estimated from finite sampling. This correction is provided by Eq. (64); for numerical values, see Table I.

Note that reseeding additional MD simulations from rarely visited state(s) has a two-fold effect on the achievable time scales. On the one hand, it increases the fraction of frames referring to the rarely sampled state(s), thereby increasing $f_{rare\ state}$. On the other hand, it increases the number of transitions to and from rarely sampled state(s), thereby decreasing $f_{transitions}$. The net effect in most cases should be an increase in the achievable time scales because such a reseeding proportionally increases the fraction of the frames in the rare states and the number of transitions from the rare states but does not affect that much the number of transitions from the native state to the rare states.

In principle, the longest time scale can exceed the aggregate sampling, with $\omega_0 \sim 1/2$ (comparable level of sampling of all subgraphs in the MSM, as opposed to poor sampling of isolated metastable states) and $f_{sw} > 1$ (the use of a sliding window) as the necessary conditions required for this result. To the best of our knowledge, no deliberate efforts have been undertaken in this direction so far. We hope that our theoretical analysis will increase the efficiency of future MSM-based modeling of biomolecular systems and, in particular, boost the power of the method in achieving longer time scales.

Previously, the sources of uncertainty in the eigenvalues of the transition matrix in MSMs were analyzed,[23] with the results easily extendable to the implicit time scales of MSMs. Based on that analysis, a strategy of adaptive sampling was suggested which reduces the variance of a specific eigenvalue (and hence, the variance of the corresponding implicit time scale) in the most computationally efficient way. In the present paper, we explore a different problem, namely, the relationship between the longest implicit time scales and the aggregate sampling, and suggest practical recipes for adaptive sampling aiming at reaching as long implicit time scales as possible given the aggregate sampling. Consequently, the strategies of adaptive sampling suggested in the cited paper and in this work are different. In practice, a researcher may pursue either of these goals or a combination of them and, correspondingly, use either of these strategies or their combination.

We do not consider changes in the number of states in an MSM as an independent parameter that could be used to achieve longer time scales. At a first glance, increasing the number of states may decrease the connectivity of the network and the number of transitions between different subgraphs, technically resulting in longer implicit time scales achieved by an MSM. However, an excessive increase in the number of states deteriorates the statistical robustness of a model due to overfitting the raw data. As demonstrated earlier, the optimal number of states in an MSM, given a set of MD trajectories, can be found by cross-validation.[36] Undertaking the actions listed above may lead to changes in the optimal number of Markov states due to changes in the raw dataset used for cross-validation, but this effect cannot be considered as an independent means of control of longest implicit time scales.

Finally, we would like to point out that this paper focuses on the analysis of the internal machinery of MSMs and does not intend to compare MSMs with other approaches to computational modeling of (bio)molecular systems. Such alternative methods, for example, milestoning or boxed molecular dynamics, have been shown to be practically efficient in speeding up simulations of rare conformational transitions in a way resembling to some degree the spirit of adaptive sampling in MSMs.[37–39] However, the mathematical apparatuses of those methods are essentially different from that of MSMs, and relationships between the longest mean passage time (or other kinetic parameters) and the aggregate sampling in such methods require a separate analysis. The practical recommendations provided in this section refer to the problem of a statistically reliable estimate of longest time scales of an MSM at the lowest computational cost. An investigator may use computational modeling to pursue other goals, e.g., aim at a detailed structural characterization of a specific conformational change. Practical recipes for achieving such goals in the most efficient way, as well as using MSMs in a combination with other enhanced sampling methods like metadynamics or milestoning, go beyond the scope of this paper.

IV. CONCLUSIONS

In this work, we have analytically investigated several representative MSMs with topologies favoring the appearance of slow implicit time scales. In all cases, unexpectedly simple relationships between the slowest implicit time scale and the aggregate sampling were derived. Our analytical results confirm the empirical rule that the slowest implicit time scale can reach the same order of magnitude as the aggregate sampling, and reveal the quantitative effects of sampling rare states and transitions to or from them on the achievable time scales. Strictly speaking, the results presented in this paper refer to four specific classes of MSMs, and it is yet to be shown (if analytically possible) that any general MSM obeys the same relationships. However, the presented four classes of MSMs cover all scenarios of the appearance of slow implicit time scales known from previous applications to specific biomolecular systems, and we expect that our results thereby provide an upper limit estimate of slowest implicit time scales achievable in all types of MSMs of biomolecules. This work can serve as a basis for a rational design of MD simulation campaigns, allowing for modeling longer-time scale conformational changes in biomolecular systems at lower cost, that is, with less computer sampling.

ACKNOWLEDGMENTS

The authors were funded by NIH R01-GM062868.

APPENDIX A: TWO-STATE MSM OF A RARE CONFORMATIONAL TRANSITION WITHOUT THE DIRECT SYMMETRIZATION OF COUNTS

In this Appendix, we repeat the analysis of the two-state MSM introduced in Sec. II A but without the direct symmetrization of counts. The need for this consideration comes from the fact that the symmetrization may create practical problems, though the final results of the models should have the same large count limits. The analytical derivations for this case are more complicated, which is the reason why the main text focuses on the case of direct symmetrization.

For the count matrix given by Eq. (1), the transition matrix T assumes the form

$$T = \begin{pmatrix} 1 - \frac{N_{01}}{N_{00}+N_{01}} & \frac{N_{01}}{N_{00}+N_{01}} \\ \frac{N_{10}}{N_{10}+N_{11}} & 1 - \frac{N_{10}}{N_{10}+N_{11}} \end{pmatrix}. \tag{A1}$$

One of the eigenvalues of T is exactly 1, while the other is

$$\lambda_1 = 1 - \frac{N_{01}}{N_{00}+N_{01}} - \frac{N_{10}}{N_{10}+N_{11}}. \tag{A2}$$

If the number of transitions between states 0 and 1 is much less than the number of transitions within state 0 or 1,

$$N_{01}, N_{10} \ll N_{00}, N_{11}, \tag{A3}$$

then the expression for the implicit time scale τ_1 assumes the following form:

$$\frac{\tau_1}{\tau_{lagtime}} \approx \frac{\omega_0 (1-\omega_0) N}{N_{01}(1-\omega_0) + N_{10}\omega_0}. \tag{A4}$$

This result is still compatible with Eq. (7), with the same definitions of $f_{rare\ state}$ and f_{sw} as before, and the following definition of $f_{transitions}$:

$$f_{transitions} = \frac{1}{2} \frac{1}{N_{01}(1-\omega_0) + N_{10}\omega_0}. \tag{A5}$$

This result coincides with the result in Sec. II A, Eq. (9), if $N_{01} = N_{10}$, or if $\omega_0 = 0$. Then, the expectation value of the implicit time scale estimated from a finite sample is given by

$$\frac{\langle \tau_1 \rangle}{\tau_{lagtime}} = \omega_0 (1-\omega_0) N \frac{\sum\limits_{\substack{N_{01},N_{10}=0 \\ \neg(N_{01}=0 \wedge N_{10}=0)}}^{\infty} \frac{1}{N_{01}(1-\omega_0)+N_{10}\omega_0} \frac{(\Omega_{01}N)^{N_{01}}}{N_{01}!} \frac{(\Omega_{10}N)^{N_{10}}}{N_{10}!} e^{-(\Omega_{01}+\Omega_{10})N}}{\sum\limits_{\substack{N_{01},N_{10}=0 \\ \neg(N_{01}=0 \wedge N_{10}=0)}}^{\infty} \frac{(\Omega_{01}N)^{N_{01}}}{N_{01}!} \frac{(\Omega_{10}N)^{N_{10}}}{N_{10}!} e^{-(\Omega_{01}+\Omega_{10})N}}. \tag{A6}$$

Using the following integral representation

$$\frac{1}{N_{01}(1-\omega_0) + N_{10}\omega_0} = \frac{1}{\omega_0 (1-\omega_0)} \int_0^1 \frac{d\xi}{\xi} \xi^{\left(\frac{N_{01}}{\omega_0} + \frac{N_{10}}{1-\omega_0}\right)}, \tag{A7}$$

the expression for the implicit time scale can be transformed to

$$\frac{\langle \tau_1 \rangle}{\tau_{lagtime}} = \frac{N e^{-n}}{1-e^{-n}} \int_0^1 \frac{d\xi}{\xi} \left[e^{N(\Omega_{01}\xi^{1/\omega_0} + \Omega_{10}\xi^{1/(1-\omega_0)})} - 1 \right]. \tag{A8}$$

We have not been able to find a general expression for the integral on the right-hand side of Eq. (A8) in terms of elementary or special functions. In the special cases of $\Omega_{01} = 0$ and arbitrary Ω_{10} or $\Omega_{10} = 0$ and arbitrary Ω_{01}, Eq. (A8) leads to the same results as with the direct symmetrization, Eqs. (18) and (64).

In the limit of $\Omega_{01} \ll 1, \quad \Omega_{10} \ll 1$, the leading term in the expression for the bias f_{bias} is

$$f_{bias} \approx \frac{C}{n},$$
$$C = \frac{\omega_0 (1-\omega_0) (\Omega_{01} + \Omega_{10})^2}{(\Omega_{01}(1-\omega_0) + \Omega_{10}\omega_0)(\Omega_{10}(1-\omega_0) + \Omega_{01}\omega_0)}. \tag{A9}$$

This asymptotic expansion is similar to Eq. (65) in the hyperbolic divergence at small n. The coefficient C has a complicated form. However, it reduces to $C = 1$ in three various cases: (1) $\Omega_{01} = 0$ and arbitrary Ω_{10} and ω_0, (2) $\Omega_{10} = 0$ and arbitrary Ω_{01} and ω_0, and (3) $\omega_0 = 1/2$ and arbitrary Ω_{01} and Ω_{10}.

Hence, the results obtained without the direct symmetrization of counts are qualitatively the same as in the case with the direct symmetrization analyzed in Sec. II A, though the corresponding formulas are technically more involved.

APPENDIX B: NUMERICAL ILLUSTRATION: AN MSM OF THE TERMINALLY BLOCKED ALANINE PEPTIDE

To illustrate the derived relationships between the longest implicit time scale and the aggregate sampling, we consider an MSM of the terminally blocked alanine peptide.[40] The choice of this molecular system is motivated by its use in an earlier work on the sources of uncertainty in the estimated eigenvectors of the transition matrix.[23] For a detailed description of this molecular system, the employed simulation settings, and the definitions of the Markov states, the reader is referred to the original paper.[40] The count matrix for the six-state MSM introduced in the cited work is

$$C = \begin{pmatrix} 4380 & 153 & 15 & 2 & 0 & 0 \\ 211 & 4788 & 1 & 0 & 0 & 0 \\ 169 & 1 & 4604 & 226 & 0 & 0 \\ 3 & 13 & 158 & 4823 & 3 & 0 \\ 0 & 0 & 0 & 4 & 4978 & 18 \\ 7 & 5 & 0 & 0 & 62 & 4926 \end{pmatrix}. \tag{B1}$$

Evidently, states 1-4 and states 5-6 form two distinct subgraphs in the transition network with rare transitions between them (19 events, or 0.06% of all transitions). Physically, this slowest motion corresponds to the rotation about the N–C bond in the alanine backbone, captured by the backbone torsion angle φ.[40] In this case, the equations derived in Sec. II C are applicable. The first two dimensionless factors can be computed based on Eqs. (51) and (52) as follows:

$$f_{rare\ state} = 2\omega_A\omega_B = 2 \cdot \frac{19\,547}{29\,550} \cdot \frac{9984}{29\,550} = 0.447, \tag{B2}$$

$$f_{transitions} = \frac{1}{N_{AB} + N_{BA}} = \frac{1}{19} = 0.0526. \tag{B3}$$

The third factor $f_{sw} = 1$ because the lagtime in this model was taken to be equal to the duration of each trajectory (0.1 ps), and a sliding window was not used. The total sampling in this set of simulations was 3.0 ns, and therefore the slowest implicit time scale predicted with Eq. (62) is

$$\tau = 0.447 \cdot 0.0526 \cdot 1 \cdot 3.0 \text{ ns} = 71 \text{ ps}. \quad \text{(B4)}$$

For comparison, the exact value of the slowest implicit time scale computed from the corresponding eigenvalue of the transition matrix is $\tau = 70$ ps.

In this example, the slowest implicit time scale is smaller than the aggregate sampling by more than one order of magnitude. The analysis presented in this paper allows us to identify the main source of inefficiency in achieving longer implicit time scales: a small value of the second factor $f_{transitions}$. According to the practical recommendations given in Sec. III (item 2), the optimal number of transitions for capturing longer time scales should be on the order of $n \sim$ 1-10, while here $n = 19$.

With ten times less sampling, the most probable value of the count matrix is

$$C' = \begin{pmatrix} 438 & 15 & 2 & 0 & 0 & 0 \\ 21 & 479 & 0 & 0 & 0 & 0 \\ 17 & 0 & 460 & 23 & 0 & 0 \\ 0 & 1 & 16 & 482 & 0 & 0 \\ 0 & 0 & 0 & 0 & 498 & 2 \\ 1 & 1 & 0 & 0 & 6 & 493 \end{pmatrix}, \quad \text{(B5)}$$

which was computed by dividing each element in matrix C, Eq. (B1), by 10 and rounding the results to the nearest integers. For this count matrix, the exact value of the longest implicit time scale computed from the eigenvalue of the transition matrix changes to $\tau' = 73$ ps, while the value predicted with Eq. (62) changes to

$$\tau' = 0.447 \cdot 0.5 \cdot 1 \cdot 0.3 \text{ ns} = 66 \text{ ps}. \quad \text{(B6)}$$

In this case, the longest implicit time scale differs from the aggregate sampling by less than one order of magnitude. Further increase in their ratio could be achieved primarily by the use of a sliding window to count the transitions, which, in its turn, would be possible after the corresponding changes in the lagtime, output frequency, and/or duration of the MD trajectories.

Evidently, this reduction in sampling leads to a dramatic increase in the variance of the estimates of implicit time scales. As demonstrated in Ref. 23, an even (in terms of the starting geometries) reduction in sampling by one order of magnitude increases the variance of the first nontrivial eigenvalue by a factor of ~40 (Fig. 6 in the cited paper, top panel), which corresponds to an approximately six-fold increase in the standard deviation of the longest implicit time scale. Depending on specific goals of a researcher, such an increase in the uncertainty may or may not be acceptable. As stated in the main text, this article focuses on the analysis of the relationship between the longest implicit time scale and the aggregate sampling in MSMs of biomolecular systems and puts forward practical recommendations for improving their ratio. If the researcher takes into account other criteria of efficiency (e.g., low standard deviations of the estimated time scales or a detailed structural characterization of a specific conformational change) or uses MSMs in a combination with other methods (e.g., metadynamics and milestoning), then the optimal strategy may be different.

APPENDIX C: DETERMINATION OF THE IMPLICIT TIME SCALE BY FITTING TO AN EXPONENTIAL

Here we consider a toy model representing another approach to calculating the characteristic time scale of a conformational transition, namely, fitting an exponential. This toy model is a two-state model, with the states labeled as state 0 and state 1, for comparability to the model in Sec. II A. Assume that we have N_{trj} parallel MD trajectories, each N/N_{trj} frames long, hence the total number of frames is N. We also assume for simplicity that the transitions occur only from state 1 to state 0, and in all the trajectories, the system was initially in state 1. Then, the expectation value of the number of transitions from state 1 to state 0 having occurred by time t equals

$$n(t) = N_{trj}\left(1 - e^{-kt}\right), \quad \text{(C1)}$$

where k is the rate constant. The curve for $n(t)$ can be obtained either from a set of MD trajectories or from experimental measurements [in the latter case, it may be a curve for an observable linearly depending on $n(t)$]. In both cases, the curve is fitted by an exponential function. From this fitting, the rate constant k can be found. Then, the characteristic time scale equals the inverse of the rate constant,

$$\tau_1^{exact} = k^{-1}. \quad \text{(C2)}$$

As follows from Eqs. (C1) and (C2), for arbitrary t,

$$\tau_1^{exact} = t\left[-\ln\left(1 - \frac{n(t)}{N_{trj}}\right)\right]^{-1}. \quad \text{(C3)}$$

Given N_{trj} and t, various samples of MD trajectories will result in various numbers of transitions from state 1 to state 0 by time t, which we further denote $N_{10}(t)$. By definition,

$$n(t) = \langle N_{10}(t)\rangle. \quad \text{(C4)}$$

Then, the expectation value of the time scale estimated from a finite sample is given by

$$\langle\tau_1\rangle = t\left\langle\left[-\ln\left(1 - \frac{N_{10}(t)}{N_{trj}}\right)\right]^{-1}\right\rangle, \quad \text{(C5)}$$

and therefore, the bias f_{bias} is

$$f_{bias} = \frac{\left[-\ln\left(1 - \frac{\langle N_{10}(t)\rangle}{N_{trj}}\right)\right]^{-1}}{\left\langle\left[-\ln\left(1 - \frac{N_{10}(t)}{N_{trj}}\right)\right]^{-1}\right\rangle}. \quad \text{(C6)}$$

In general, the distribution of N_{10} is complicated and technically difficult to compute. However, some approximations are possible by analogy with the analysis in Sec. II A. If transitions from state 1 to 0 are rare, then we can assume that the number of such transitions follows the Poisson distribution,

$$N_{10}(t) \sim Pois\left(\langle N_{10}(t)\rangle\right), \quad \text{(C7)}$$

and therefore, the denominator on the right-hand side of Eq. (C6) can be explicitly computed as

$$f_{bias} = \frac{\left[-\ln\left(1 - \frac{n(t)}{N_{trj}}\right)\right]^{-1}}{\sum_{N_{10}=1}^{\infty}\left[-\ln\left(1 - \frac{N_{10}}{N_{trj}}\right)\right]^{-1}\frac{n(t)^{N_{10}}}{N_{10}!}e^{-n(t)}}. \quad \text{(C8)}$$

As before, we omit elementary events with $N_{10}(t) = 0$ because the transitions from state 1 to state 0 are not visible in such cases, and hence no information on the time scale is available. It does not seem possible to get a finite analytical expression for the sum in the denominator in Eq. (C8). However, in practice, the number of observed transitions corresponding to the slowest implicit time scales ($N_{10} \sim 1$–10) is typically much smaller than the number of MD trajectories ($N_{trj} \sim 10^2$–10^3), and the expressions in the square parentheses in Eq. (C8) can be approximated by the first term in the Taylor series expansions, leading to

$$f_{bias} = \frac{\frac{1}{n(t)} \sum\limits_{N_{10}=1}^{\infty} \frac{n(t)^{N_{10}}}{N_{10}!} e^{-n(t)}}{\sum\limits_{N_{10}=1}^{\infty} \frac{1}{N_{10}} \frac{n(t)^{N_{10}}}{N_{10}!} e^{-n(t)}}. \tag{C9}$$

Then, the sums on the right-hand side of Eq. (C9) can be computed analytically, and the final result for the bias factor coincides with the result derived in the main text, Eq. (64), but with $n(t)$ instead of n,

$$f_{bias} = \frac{1 - e^{-n(t)}}{n(t)e^{-n(t)} \left[\mathrm{Ei}(n(t)) - \gamma - \ln(n(t))\right]}. \tag{C10}$$

Respectively, the "large" $n(t)$ expansion of Eq. (C10) assumes the following form:

$$f_{bias} \underset{n(t)\to\infty}{\sim} 1 - \frac{1}{n(t)} + O\left(\frac{1}{n(t)^2}\right). \tag{C11}$$

Here, "large" $n(t)$ means that $n(t)$ is significantly greater than 1, but still much less than the number of MD trajectories N_{trj} (which is typically $\sim 10^2$–10^3), for example, $n(t) \sim 10$. This result is analogous to the result given by Eq. (66) in the main text.

The analysis in this appendix differs from the analysis in the main text [in particular, Eqs. (64) and (66)] in that the time scale here is computed by fitting to an exponential, rather than diagonalizing the transition matrix. However, the bias factor f_{bias} here also contains a leading term inversely proportional to the expectation value of the number of observed transitions. Hence, the approach of the transition matrix diagonalization to computing implicit time scales of conformational transitions is not inferior in comparison to fitting to an exponential in terms of the bias introduced into estimates of time scales due to finite sampling.

[1] V. S. Pande, K. Beauchamp, and G. R. Bowman, Methods **52**, 99 (2010).
[2] G. R. Bowman, V. S. Pande, and F. Noe, *An Introduction to Markov State Models and Their Application to Long Timescale Molecular Simulation*, 1st ed. (Springer Science & Business Media, 2013).
[3] J. D. Chodera and F. Noe, Curr. Opin. Struct. Biol. **25**, 135 (2014).
[4] M. Shirts and V. S. Pande, Science **290**, 1903 (2000).
[5] C. D. Snow, H. Nguyen, V. S. Pande, and M. Gruebele, Nature **420**, 102 (2002).
[6] D. Sezer, J. H. Freed, and B. Roux, J. Phys. Chem. B **112**, 11014 (2008).
[7] G. R. Bowman, K. A. Beauchamp, G. Boxer, and V. S. Pande, J. Chem. Phys. **131**, 124101 (2009).
[8] X. Huang, Y. Yao, G. R. Bowman, J. Sun, L. J. Guibas, G. Carlsson, and V. S. Pande, Pac. Symp. Biocomput. **2010**, 228 available at https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4423759/.
[9] I. Buch, T. Giorgino, and G. De Fabritiis, Proc. Natl. Acad. Sci. U. S. A. **108**, 10184 (2011).
[10] T. J. Lane, G. R. Bowman, K. Beauchamp, V. A. Voelz, and V. S. Pande, J. Am. Chem. Soc. **133**, 18413 (2011).
[11] F. Noe, S. Doose, I. Daidone, M. Lollmann, M. Sauer, J. D. Chodera, and J. C. Smith, Proc. Natl. Acad. Sci. U. S. A. **108**, 4822 (2011).
[12] W. Zhuang, R. Z. Cui, D. A. Silva, and X. Huang, J. Phys. Chem. B **115**, 5415 (2011).
[13] K. J. Kohlhoff, D. Shukla, M. Lawrenz, G. R. Bowman, D. E. Konerding, D. Belov, R. B. Altman, and V. S. Pande, Nat. Chem. **6**, 15 (2014).
[14] D. Shukla, Y. Meng, B. Roux, and V. S. Pande, Nat. Commun. **5**, 3397 (2014).
[15] N. Ferruz, G. Tresadern, A. Pineda-Lucena, and G. De Fabritiis, Sci. Rep. **6**, 30275 (2016).
[16] G. Pinamonti, J. Zhao, D. E. Condon, F. Paul, F. Noe, D. H. Turner, and G. Bussi, J. Chem. Theory Comput. **13**, 926 (2017).
[17] A. V. Sinitskiy, N. H. Stanley, D. H. Hackos, J. E. Hanson, B. D. Sellers, and V. S. Pande, Sci. Rep. **7**, 44578 (2017).
[18] F. Noe, C. Schutte, E. Vanden-Eijnden, L. Reich, and T. R. Weikl, Proc. Natl. Acad. Sci. U. S. A. **106**, 19011 (2009).
[19] V. A. Voelz, G. R. Bowman, K. Beauchamp, and V. S. Pande, J. Am. Chem. Soc. **132**, 1526 (2010).
[20] S. Gu, D. A. Silva, L. Meng, A. Yue, and X. Huang, PLoS Comput. Biol. **10**, e1003767 (2014).
[21] N. Plattner and F. Noe, Nat. Commun. **6**, 7653 (2015).
[22] A. M. Abramyan, S. Stolzenberg, Z. Li, C. J. Loland, F. Noe, and L. Shi, ACS Chem. Neurosci. **8**, 1735 (2017).
[23] N. S. Hinrichs and V. S. Pande, J. Chem. Phys. **126**, 244101 (2007).
[24] F. Noe, J. Chem. Phys. **128**, 244103 (2008).
[25] P. Metzner, F. Noe, and C. Schutte, Phys. Rev. E **80**, 021106 (2009).
[26] J. H. Prinz, H. Wu, M. Sarich, B. Keller, M. Senne, M. Held, J. D. Chodera, C. Schutte, and F. Noe, J. Chem. Phys. **134**, 174105 (2011).
[27] B. Trendelkamp-Schroer, H. Wu, F. Paul, and F. Noe, J. Chem. Phys. **143**, 174101 (2015).
[28] T. Schmidt, A. Bergner, and T. Schwede, Drug Discovery Today **19**, 890 (2014).
[29] A. Szilagyi and Y. Zhang, Curr. Opin. Struct. Biol. **24**, 10 (2014).
[30] A. Fiser, *From Protein Structure to Function with Bioinformatics*, edited by D. J. Rigden (Springer Netherlands, 2017), p. 91.
[31] P. D. Adams, D. Baker, A. T. Brunger, R. Das, F. DiMaio, R. J. Read, D. C. Richardson, J. S. Richardson, and T. C. Terwilliger, Annu. Rev. Biophys. **42**, 265 (2013).
[32] J. Moult, K. Fidelis, A. Kryshtafovych, T. Schwede, and A. Tramontano, Proteins: Struct., Funct., Bioinf. **84**, 4 (2016).
[33] J. Lee, P. L. Freddolino, and Y. Zhang, *From Protein Structure to Function with Bioinformatics*, edited by D. J. Rigden (Springer Netherlands, 2017), p. 3.
[34] N. Singhal and V. S. Pande, J. Chem. Phys. **123**, 204909 (2005).
[35] M. Bacci, A. Vitalis, and A. Caflisch, Biochim. Biophys. Acta **1850**, 889 (2015).
[36] R. T. McGibbon and V. S. Pande, J. Chem. Phys. **142**, 124105 (2015).
[37] D. R. Glowacki, E. Paci, and D. V. Shalashilin, J. Phys. Chem. B **113**, 16603 (2009).
[38] K. Kuczera, G. S. Jas, and R. Elber, J. Phys. Chem. A **113**, 7461 (2009).
[39] P. Majek and R. Elber, J. Chem. Theory Comput. **6**, 1805 (2010).
[40] J. D. Chodera, W. C. Swope, J. W. Pitera, and K. A. Dill, Multiscale Model. Simul. **5**, 1214 (2006).