The Journal of Chemical Physics

On the removal of initial state bias from simulation data

Cite as: J. Chem. Phys. **150**, 104105 (2019); https://doi.org/10.1063/1.5063556
Submitted: 28 September 2018 . Accepted: 19 February 2019 . Published Online: 13 March 2019

Marco Bacci, Amedeo Caflisch, and Andreas Vitalis

ARTICLES YOU MAY BE INTERESTED IN

Automated Markov state models for molecular dynamics simulations of aggregation and self-assembly
The Journal of Chemical Physics **150**, 115101 (2019); https://doi.org/10.1063/1.5083915

Perspective: Computational chemistry software and its advancement as illustrated through three grand challenge cases for molecular science
The Journal of Chemical Physics **149**, 180901 (2018); https://doi.org/10.1063/1.5052551

The combined force field-sampling problem in simulations of disordered amyloid-β peptides
The Journal of Chemical Physics **150**, 104108 (2019); https://doi.org/10.1063/1.5078615

AIP Publishing

On the removal of initial state bias from simulation data

Cite as: J. Chem. Phys. 150, 104105 (2019); doi: 10.1063/1.5063556
Submitted: 28 September 2018 • Accepted: 19 February 2019 •
Published Online: 13 March 2019

Marco Bacci, Amedeo Caflisch, and Andreas Vitalis[a)]

AFFILIATIONS

University of Zurich, Department of Biochemistry, Winterthurerstrasse 190, CH-8057 Zurich, Switzerland

Note: This article is part of the Special Topic "Markov Models of Molecular Kinetics" in J. Chem. Phys.

[a)]**Author to whom correspondence should be addressed:** a.vitalis@bioc.uzh.ch. **Tel./Fax:** +41 44 635 5568.

ABSTRACT

Classical atomistic simulations of biomolecules play an increasingly important role in molecular life science. The structure of current computing architectures favors methods that run multiple trajectories at once without requiring extensive communication between them. Many advanced sampling strategies in the field fit this mold. These approaches often rely on an adaptive logic and create ensembles of comparatively short trajectories whose starting points are not distributed according to the correct Boltzmann weights. This type of bias is notoriously difficult to remove, and Markov state models (MSMs) are one of the few strategies available for recovering the correct kinetics and thermodynamics from these ensembles of trajectories. In this contribution, we analyze the performance of MSMs in the thermodynamic reweighting task for a hierarchical set of systems. We show that MSMs can be rigorous tools to recover the correct equilibrium distribution for systems of sufficiently low dimensionality. This is conditional upon not tampering with local flux imbalances found in the data. For a real-world application, we find that a pure likelihood-based inference of the transition matrix produces the best results. The removal of the bias is incomplete, however, and for this system, all tested MSMs are outperformed by an alternative albeit less general approach rooted in the ideas of statistical resampling. We conclude by formulating some recommendations for how to address the reweighting issue in practice.

I. INTRODUCTION

Stochastic models have been widely used in fields such as risk prediction,[1] speech recognition,[2] bioinformatics,[3] etc. When applied to discrete systems, they generally describe the relationships, if any, between different states or entities captured in the model. In the field of atomistic simulations of molecular systems, a common strategy is to use an unsupervised data mining procedure (clustering) to discretize a complex system represented in a reduced but still high-dimensional space. The resultant discretization defines states whose interconversion can be described by stochastic network models, which are most often memoryless, i.e., they are Markov state models (MSMs).[4,5] A data-derived stochastic model allows predictions to be made regarding the evolution of the system it represents. The usual axis of propagation is time, for example, for a biomolecule evolving according to Newton's equations of motion.[6] MSMs are used to predict the different dynamic modes a system has access to (usually in a hierarchical approach prioritizing the slowest modes), to define kinetic reaction coordinates like committor probabilities[7] or mean-first passage times,[8] and to calculate relaxation/reaction rates that can be compared to observables extracted from real data.[5,9]

One of the most direct and useful predictions offered by MSMs is the implied steady state, which, given that ergodicity holds, is the unique distribution of probabilities across states that are stationary in time. It can be calculated as the eigenvector of the transition matrix (**T**) associated with eigenvalue 1.0 but also iteratively by simply applying the transition matrix to an arbitrary initial distribution until convergence. When applied to molecular systems, the steady state is associated with the thermodynamic equilibrium distribution as it is the predicted stationary distribution for an infinite amount of sampling. There is one obvious caveat in this logic, however. If both the states and the transition matrix are inferred, without adjustments, from a single simulation of finite length, the steady state is usually arbitrarily close to the sampling distribution.[10] In this limit, the MSM is merely a description of the data in a coarse-grained representation. This means that the steady state offers no information regarding convergence or sampling quality. In other words, it simply imposes the assumption of (global) equilibrium onto the sampled trajectory.

For the prediction of equilibrium distributions from MSMs to offer new insights, we thus require one of two additional criteria, which can be combined. First, the underlying data can be an ensemble of very many short trajectories instead of a single or few long trajectories.[11–13] If the distribution of their starting points is inconsistent with the steady state of a derived MSM, the predicted equilibrium will differ substantially from the raw sampling weights. Second, independent information can be used in deriving the MSM. This can be either at the level of states or at the level of transitions.[14] In the former case, a prior definition of states allows an assessment of sampling quality by consulting visitation frequencies and the recurrence of transitions to define well-sampled domains. In the latter case, prior beliefs regarding the transition matrix can be incorporated, and this will alter the predicted MSM steady state. The goal of our study is to analyze the efficacy of different ways of adding information to the transition matrix for trajectory ensembles. It is well-known that recovering the correct equilibrium distribution is a nontrivial problem in the presence of the type of initial state bias mentioned above.[15]

Conceptually, every transition matrix element can be decomposed into a kinetic and a thermodynamic component.[16] The latter is related to the weights of the target state at equilibrium and does not depend on the underlying propagator or on the chosen lag time (the effective time step of the MSM). The former does depend on both of those. For example, consider a propagator describing Hamiltonian dynamics for a system with interactions sufficient to achieve ergodicity. In this scenario, the kinetic component will be influenced by velocity variables at very short lag times, which leads to memory effects that are in conflict with the assumption of Markovianity. As the lag time increases, for an appropriate coarse-graining, the effective dynamics become diffusive.[17,18] Clearly, for the MSM to describe the true equilibrium distribution, the kinetic components must gradually cancel out as soon as the lag time approaches and eventually exceeds all relevant time scales of the system. This restates the ergodic hypothesis (supported by Liouville's theorem), and it implies that the propagator must be globally balanced. Detailed balance, i.e., requiring that the fluxes for every pair of state are balanced, implies equilibrium[19] and is a sufficient but not a necessary condition to achieve global balance. Steady states where microscopic equilibrium does not hold appear in different areas of science[20,21] and are more difficult to analyze rigorously.[22]

For real-world applications, we typically want to infer simultaneously the kinetic and the thermodynamic component of every transition matrix element from finite data. At intermediate lag times, the matrix will still be highly sparse, and this sparsity is part of the inference. Clearly, in this regime, many transitions can be safely neglected because their kinetic likelihoods are vanishingly small. However, for the ones that are marginal, i.e., those that would, for the given sampling length, give rise to single-digit transition counts, noise problems arise. There are several schools of thought to address this problem. The first one has been to rely on the theoretical fulfillment of detailed balance by the propagator to impose detailed balance onto the MSM, which removes local flux imbalances (see Sec II A). This can be done on the entire MSM or on individual strongly connected components.[10] Detailed balance can be enforced in an MSM in various ways, and, given the assumption of Markovianity, a constrained maximum likelihood (ML) approach has been developed that should keep the inferred transition matrix as consistent with the observed count matrix as possible.[4,14]

The second school of thought is the use of pseudocounts, which are a way to use a specific class of prior distribution to improve a purely likelihood-based inference (see Sec. II B). For a Markov model, each transition is in essence a throw of an N-sided die, and for this problem the most-used uninformative priors derive from the Dirichlet distribution.[23–25] It has to be pointed out that there is no universally accepted truth of what are the best uninformative or subjective priors for the problem of transition matrix inference. A third school of thought is to construct the transition matrix from scratch (see Sec. II C). This means discarding the raw kinetic information available from the trajectories.[26–29] In this framework, elements of $\mathbf{T}$ are assembled explicitly from their diffusive (using geometric similarity) and thermodynamic (from a reweighting protocol for biased simulations) elements. This strategy generally requires the use of advanced sampling techniques and implies a constant rate of diffusion across phase space. The latter does not usually hold for geometric projections.[8,16,30]

Clearly, all three approaches are in danger of introducing biases as they add information, but this type of information differs fundamentally. In this contribution, we focus on how this added information influences the prediction of the steady state of MSMs constructed from ensembles of short trajectories with biased starting positions. In this situation, deviations from the ground truth arise both in the form of random errors due to limited sampling and in the form of systematic errors due to the choice of starting positions for the individual trajectories. We perform tests on three classes of systems arranged hierarchically: (1) a discrete system where Markovianity is a given; (2) a toy model where we apply the full workflow of discretization and MSM construction on data from a propagator in continuous space; (3) a real-world application of molecular dynamics (MD) simulations of peptide folding. Where applicable, we also compare results to a conceptually unrelated strategy derived from the idea of statistical resampling (see Sec. II E).

II. METHODS

A. Maximum likelihood inference and detailed balance imposition

The simple solution for the maximum likelihood estimate of a transition matrix is dependent on the choice of a likelihood function that treats all transitions as independent (Markovian). Then,

$$\mathbf{T}_{ML} = \max[L(\mathbf{C}|\mathbf{T})] = \max\left[\prod_{i=1}^{N}\prod_{j=1}^{N} {t_{ij}}^{c_{ij}}\right] \Rightarrow t_{ij} = \frac{c_{ij}}{\sum\limits_{k=1}^{N} c_{ik}}. \quad (1)$$

In Eq. (1), $\mathbf{C}$ is the count matrix, and N is the number of states. Detailed balance can be imposed onto a transition matrix in different ways. The two popular ways of doing this naively involve deriving a symmetric count matrix $\mathbf{C}^*$ either by letting $c^*_{ij} = c^*_{ji} = 0.5(c_{ij} + c_{ji})$ or by letting $c^*_{ij} = c^*_{ji} = \max(c_{ij}, c_{ji})$ (where applicable, we used the latter in Sec. III). Bowman *et al.*[4] and later Prinz *et al.*[14] proposed numerical strategies for maximizing the (Markovian) likelihood with a symmetry constraint, i.e., to solve the implicit relation

$$\begin{aligned} \mathbf{T}_{DB,ML} &= \max\Big[L(\mathbf{C}|\mathbf{T})\Big|(\pi_i t_{ij} = \pi_j t_{ji})_{\forall i,j}\Big] \\ &= \max\left[\prod_{i=1}^{N}\prod_{j=1}^{N} t_{ij}{}^{c_{ij}}\Big|(\pi_i t_{ij} = \pi_j t_{ji})_{\forall i,j}\right]. \end{aligned} \tag{2}$$

In Eq. (2), $\boldsymbol{\pi}(\mathbf{T})$ is the vector of predicted equilibrium probabilities, which appears here in the detailed balance constraint, $(\pi_i t_{ij} = \pi_j t_{ji})_{\forall i,j}$. As with all derivations involving the likelihood L, we are restricted to the shown (Markovian) functional form. For a non-Markovian system, successive transitions become coupled, and the probabilities are no longer independent. It is important to spell out that the likelihood function and thereby all statistical procedures relying on it become incorrect for systems that do not exhibit memoryless behavior given the choices of hyperparameters (coarse-graining and lag time).

B. Posterior inference of transition matrices

Each row of a transition matrix, $\mathbf{t_i}$, describes the conditional probability of moving from state i to the putative target states j (including i). Because of the Markov property, this is equivalent to a process described by a multinomial distribution, and each row can be treated independently. Dirichlet priors, $D(\mathbf{x};\, \boldsymbol{\alpha})$, which depend on "concentration" parameters α, are conjugate to the multinomial distribution, and the resulting posterior satisfies[31]

$$p(\mathbf{t_i}|\mathbf{c_i}) \propto L(\mathbf{c_i}|\mathbf{t_i})\cdot D(\mathbf{t_i};\boldsymbol{\alpha_i}) \propto \prod_{j=1}^{N} t_{ij}{}^{c_{ij}+\alpha_{ij}-1} \propto D(\mathbf{t_i};\boldsymbol{\alpha_i}+\mathbf{c_i})\ \forall \alpha_{ij} \geq 0. \tag{3}$$

The expected value of the Dirichlet distribution on the right-hand side of Eq. (3) is

$$E(t_{ij}) = \frac{\alpha_{ij} + c_{ij}}{\sum_{k=1}^{N}(\alpha_{ik} + c_{ik})}. \tag{4}$$

Thus, if we treat the α_{ij} as pseudocounts, the expected values of the t_{ij} from the posterior distribution are straightforwardly computable and resemble an augmented maximum likelihood (ML) estimate.[25] Empirically, pseudocounts are also known under the term additive or Laplace smoothing.[32] If all the α_{ij} are 0.0, the expected value of the posterior distribution is equivalent to the ML estimate. For the uniform prior, all α_{ij} are 1.0 instead. In this case, the more natural correspondence of the ML estimate is with the maximum *a posteriori* (MAP) estimate, which has a closed-form solution if all $\alpha_{ij} \geq 1$

$$MAP(t_{ij}) = \frac{\alpha_{ij} + c_{ij} - 1}{\sum_{k=1}^{N}(\alpha_{ik} + c_{ik} - 1)}. \tag{5}$$

We can thus understand the solution in Eq. (4) equivalently as the MAP estimate for a Dirichlet prior with pseudocounts of $1 + \alpha_{ij}$. One advantage of such a prior is that the ergodicity is guaranteed. A popular choice for α appears to be $1 + 1/N$ for all i and j [equivalent to $1/N$ in the sense of Eq. (4)],[23] and all "MAP" estimates referenced in the figures below used this value. The major downside is that the matrix becomes maximally dense, which limits the number of states one can investigate in routine applications. We note that workarounds have been developed for specific tasks like the computation of mean-first passage times.[25]

C. *De novo* construction of transition matrices

As mentioned in the Introduction, every transition matrix element can be decomposed into a kinetic and a thermodynamic component.[16] Because the kinetic component must "integrate out" as long as the lag time is sufficiently large and the system undergoes ergodic dynamics, all the information about the relative weights of the states must be encoded in the thermodynamic components. Suppose now that an independent assessment of the relative weights of states is available, e.g., derived in coarse terms from an advanced sampling calculation that yielded the potential of mean force along a reaction coordinate. Converted to a set of MSM states, this becomes a vector of state probabilities at equilibrium, $\mathbf{p}^{eq}$. How do we obtain the kinetic component? Following the literature,[27–29] it is possible to derive the missing kinetic components with two additional inputs: first, a geometric threshold for excluding which pairs of states are deemed close enough to allow transitions at all; second, a base rate. The idea of the geometric threshold is similar to that of choosing a lag time in data-derived MSMs: a looser threshold corresponds to a larger lag time and vice versa. More generally, the threshold can be replaced by a continuous function H that returns something akin to an effective diffusion "kernel," i.e., a distribution of (squared) geometric distances available after an implicit time lag τ. For the literature examples cited above, H was chosen as a shifted Heaviside function, which is the functional representation of a cutoff. The base rate parameter affects all rates and is normally treated as a fitting parameter to obtain correct time scales.[27–29]

Unlike the formulation of Levy and co-workers,[29] who use the rate matrix, we construct *de novo* transition matrices as follows:

$$\begin{aligned} t^{*}_{ij} &= H(d_{ij})/p^{eq}_{i}\ \forall_{i\neq j}, \\ t^{*}_{ii} &= f\ - \sum_{j\neq i}^{N} t^{*}_{ij}. \end{aligned} \tag{6}$$

In Eq. (6), d_{ij} is the geometric distance between states i and j, H is the aforementioned kernel function, and f is the term controlling the base rate. Specifically, we choose f as a constant equivalent to the maximum row sum of $\sum_{j\neq i}^{N} t^{*}_{ij}$ across all i plus a (positive) increment, which is a free parameter. Larger values of f make the model slower overall. The final $\mathbf{T}$ is then obtained from $\mathbf{T}^{*}$ by row normalization. It is important that the symmetry of the kinetic component can be ensured easily because $d_{ij} = d_{ji}$ for a proper metric. This will lead to transition matrices that automatically imply detailed balance. Similarly to the ML estimate with detailed balance constraint in Eq. (2) above, strategies have been developed to introduce $\mathbf{p}^{eq}$ as a constraint rather than using it directly as in Eq. (6).[33]

As mentioned in the Introduction, detailed balance is a sufficient but not a necessary condition, and what ultimately matters is that the kinetic components are globally balanced. This can be illustrated using explicit Markov models where there is no notion of geometric distance. Figure 1(a) shows three alternative models

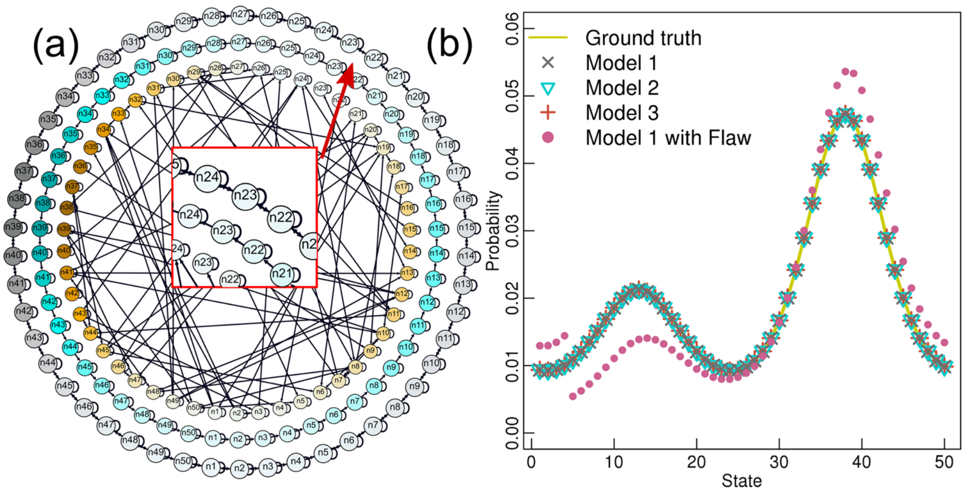

FIG. 1. Different balanced propagators integrate out. (a) The sparsity pattern of three different transition matrices is shown for the same circular graph layout. Nodes are numbered in sequence, and colors differentiate the three matrices (gray: model #1; cyan: model #2; orange: model #3). Color intensity increases with increasing equilibrium probability. The inset is shown to improve readability. Arrowheads are only drawn for connections between different states. We obtained many matrices similar to model #3 by randomly adding transitions in a symmetric fashion (both *ij* and *ji*) until ergodicity was reached. The example shown has 190 total nonzero elements. The $\mathbf{p}^{eq}$ defined in Eq. (7) enters the construction of **T** following Eq. (6) (see the text). (b) The normalized first eigenvectors of the matrices specifying models #1–#3 are plotted against the ground truth, $\mathbf{p}^{eq}$. Results for different versions of model #3 were identical to numerical precision. The prediction for a flawed version of model #1 is included to emphasize the role of balance in the propagator (see the text in Sec. III A).

constructed for the same set of states. We apply the following periodic potential, in reduced units of k_bT, to the 50 discrete states (indexed by n), which defines $\mathbf{p}^{eq}$

$$p_n^{eq} \propto \exp[-V(n)] \Leftarrow V(n) = -0.6\cos\left(\frac{\pi n}{12.5}\right) - 0.4\sin\left(\frac{\pi n}{25}\right). \quad (7)$$

Equation (7) creates two major basins of attraction with unequal weights and allows for simple visualization.

Thus, the models in Fig. 1(a) differ only in the kinetic components of their transition matrices, which were constructed using Eq. (6) but replacing the H kernel with binary functions that switch on specific transitions chosen such that global balance is preserved. For model #1, all states with neighboring indices are connected. For model #2, these connections are made one-sided. Finally, for model #3, random connections were chosen in a symmetric fashion until ergodicity was reached. Figure 1(b) shows that these three network topologies, which are differentiated exclusively by the propagator part, give exactly the same steady state. This must be so if ergodicity holds and if the global balance condition is met. For example, we could understand the randomly connected (#3) and the symmetric neighbor-only (#1) models as coming from a Monte Carlo simulation (jumps are possible) and a molecular dynamics (MD) simulation, respectively. Model #2, for which one direction was eliminated relative to model #1, illustrates that a propagator can be globally balanced without fulfilling detailed balance.

D. Computing observables reweighted by MSMs

Trajectory ensembles consist of individual snapshots. To compute expected values for an observable O from them, one typically takes just the sample mean

$$\langle O\rangle^{obs} = S^{-1}\sum_i^S O(s_i). \quad (8)$$

In Eq. (8), s_i indicates the conformation corresponding to the ith snapshot, and S is the total number of snapshots. The MSM makes a prediction for the population at equilibrium for clusters, and Eq. (8) needs to be corrected (reweighted) accordingly

$$\langle O\rangle^{MSM} = S^{-1}\sum_i^S \frac{\pi_{c_i}}{p_{c_i}^{obs}} O(s_i). \quad (9)$$

Here, c_i is the cluster that the ith snapshot belongs to, and $\boldsymbol{\pi}$ and $\mathbf{p}^{\mathbf{obs}}$ are the vectors of the predicted equilibrium and the observed cluster weights, respectively. In order to generate reweighted distributions of observables, the same logic is applied to the construction of histograms, i.e., the counts for snapshot s_i are simply weighted by the factor $\pi_{c_i}/p_{c_i}^{obs}$.

E. Casting PIGS in the framework of statistical resampling

Progress-index guided sampling (PIGS)[13] is an advanced sampling strategy that works on a trajectory ensemble of constant size in parallel. Here, we just review the method's salient features relevant to the scope of the manuscript. Initially, all replicas start from the same configuration but their evolution is stochastic, which ensures the divergence of trajectories with time. At regular intervals, the data from all trajectories are pooled and analyzed[34,35] to produce a ranking of how likely the current conformations are to explore new or undersampled areas of phase space. The ranking is used to propose stochastic reseedings favoring highly ranked replicas and disfavoring those that are likely redundant. Accepted reseedings lead to the termination of some trajectories. There is no guarantee that a different but geometrically similar trajectory will survive. There is no binning of phase space involved, and the most critical hyperparameter of PIGS is the choice of representation, which directly determines the metric for measuring conformational distance. Interested readers are referred to the original publication[13] and two recent applications for details.[36,37]

The idea of evolving a number or replicas of a system and using an informed criterion to guide the ensemble of trajectories toward more interesting states is shared by many advanced sampling methodologies.[11,12,38–42] If we consider individual trajectories as samples drawn from a well-defined distribution, the idea of weighted, statistical resampling can be applied to the ensemble.[43] For this, there are two processes to monitor, duplication and termination. The objective is to track these processes and update the statistical weights of trajectories (and thereby snapshots). Duplication of a particular sample should lead to a proportional reduction in weight, while termination should lead to a corresponding increase in the weight of a similar sample. For this manuscript, this is all done in post-processing, yet we use the acronym WE (weighted ensemble) below due to the method's origins.[41]

In detail, suppose a trajectory ensemble of constant size is evolving. Every so often, some trajectories are terminated and reseeded, which is representable as an integer map, m_i^r, where i indexes trajectories, and r time (reseeding cycles). For a terminated trajectory, $m_i^r \neq i$. The instantaneous weight, w_i^r, of a terminated trajectory is distributed to one or more similar trajectories, which is measured most often as being found in the same bin. For simplicity, let us assume that this is also representable as an integer map, $\mathbf{x}$. Then

$$w_i^r = \begin{cases} \dfrac{\sum_j \delta_{i,x_j^r} w_j^{r-1}}{\sum_j \delta_{i,m_j^r}} & \text{if } \sum_j \delta_{i,m_j^r} > 0, \\ w_{m_i^r}^r & \text{else.} \end{cases} \tag{10}$$

In Eq. (10), $\delta_{i,j}$ denotes the Kronecker delta. The top row shows the joint lumping and splitting idea of statistical weights applied at a given reseeding cycle. Trajectories that survive combine the weights of any terminated trajectories they have been deemed similar to (the map $\mathbf{x}$) and normalize it by the number of replications (manifest in the map $\mathbf{m}$). This includes the case of surviving trajectories not involved in either lumping or splitting, for which only the ith elements of $\mathbf{m}$ and $\mathbf{x}$ are equivalent to i and thus $w_i^r = w_i^{r-1}$. Terminated replicas attain the resultant, updated weight of their reseeding conformation (bottom row). This is illustrated in Table I using an (arbitrary) example. A given weight w_i^r applies to all snapshots from trajectory i collected between reseeding cycles r and $r + 1$.

When applied to a PIGS data set, where $\mathbf{m}$ is known exactly from the recorded reseeding history,[13] the main difficulty is in determining the map $\mathbf{x}$. Because PIGS does not bin phase space, the best available guess is simply to lump the weight of a terminated trajectory to the geometrically closest *and* surviving one. Without restricting ourselves to such a binary map, the best available guess would be much harder to define. In any case, the resultant guesses can be arbitrarily poor but are expected to be more robust for systems of lower complexity and larger numbers of replicas. The strategy in Eq. (10) is, unlike the MSM-based protocols above, not a general strategy, however. For example, it has no effect on an ensemble of trajectories started from incorrectly (non-Boltzmann) distributed but independent configurations. We note that the combination of PIGS and statistical resampling is a rough analog of adaptive multilevel splitting[42] when there is no reaction coordinate to define an exploitation goal. For splitting methods, many useful properties of derived estimators have been proven.[44]

TABLE I. Scheme to illustrate Eq. (10). The values of the two integer maps, $\mathbf{m}$ and $\mathbf{x}$, are shown for five reseeding cycles and four replicas. Initially (r = 1), all trajectories have equivalent weights of 1/4. The first reseeding is for r = 2, where trajectory #2 is terminated and replaced with #4, and its weight is lumped into #3. This means that the resultant weights ($\mathbf{w}^2$) will be 1/4, 1/8, 1/2, and 1/8 (for trajectories #1–#4). The next event is at r = 4 where #3 is terminated and replaced with #4. Because its weight is also lumped into #4, the updated $\mathbf{w}^4$ will be 1/4, 1/8, 5/16, and 5/16. Finally, at r = 5, both #1 and #4 are replaced by #2, and their weights are lumped into #3. Thus, $\mathbf{w}^5$ will be 1/24, 1/24, 7/8, and 1/24.

	$r = 1$	2	3	4	5
m_1^r	1	1	1	1	2
x_1^r	1	1	1	1	3
m_2^r	2	4	2	2	2
x_2^r	2	3	2	2	2
m_3^r	3	3	3	4	3
x_3^r	3	3	3	4	3
m_4^r	4	4	4	4	2
x_4^r	4	4	4	4	3

III. RESULTS AND DISCUSSION

We present the results as follows: First, we study a toy system of 50 discrete states that is explicitly Markovian (Sec. III A). We use trajectories from biased starting positions to highlight which methods of transition matrix inference are able to remove this bias. The statistical resampling (WE) approach is not applicable to this setup (see Sec. II E). Next, we advance to a similar system but in continuous space (Sec. III B). Because we employ an actual advanced sampling technique (PIGS),[13] and because we know the ground truth, the resultant trajectory ensemble can be tackled with all of the methods we consider here. Finally, we turn to a real-world application, viz., the conformational equilibrium of a 21-residue peptide prone to form α-helices (Sec. III C).[45] This is a published data set generated by atomistic MD simulations.[13]

A. An explicitly Markovian toy model

If the distribution of initial configurations for an ensemble of trajectories is not the Boltzmann one, it is expected that during the initial part there is an unbalanced probability flux toward the stationary distribution.[46] This flux imbalance is present in the network unless a suitable relaxation period is truncated. This truncation strategy is commonly used for single or few simulations of identical length, but it is not feasible for the often short trajectories produced by many advanced sampling protocols.[15] To be able to compare different methods of transition matrix construction, we remove all ambiguities regarding the coarse-graining and the desired memorylessness by turning to an explicit Markov model.

Of course, knowing the ground truth, Eq. (7), allows a straightforward (*de novo*) construction of synthetic transition matrices (Sec. II C) with arbitrarily accurate results (see Fig. 1). Clearly, the example in Fig. 1 masks the main difficulty with this approach since in practice the ground truth [$\mathbf{p}^{eq}$ in Eq. (6)] for the full phase space explored by the trajectories can be extracted only crudely from approaches like umbrella sampling[47] or replica exchange.[48] Thus, the *de novo* method's main application has been in predicting *kinetic* properties. Figure 1 includes a flawed model that differs from model #1 only in the fact that the propagator component from state 4 to state 5 was reduced by a factor of 3.0. This stresses the importance of maintaining global balance. When transition matrices are inferred fully from data, imbalances can arise sporadically and independently of $\mathbf{p}^{eq}$, and this limits the accuracy of derived predictions. In the common case that there is no independent estimate for $\mathbf{p}^{eq}$, this means that imbalance problems due to the propagator cannot be delineated from the effects of $\mathbf{p}^{eq}$. As outlined above, simple approaches to this inference problem include mandating detailed balance (Sec. II A) or using Bayesian logic (Sec. II B). The former is linked to the known symmetry property of the propagator. As shown below, these approaches can be harmful if the underlying imbalances are not sporadic but systematic.

In order to compare the different data-driven methodologies for the inference of $\mathbf{T}$, we create synthetic trajectories from random walkers on model #1 in Fig. 1(a). The initialization of a fixed number of trajectories was uniform except that they started with 10-times higher likelihood in the first 25 out of the 50 total states [left side in Fig. 1(b)]. This creates an initial state bias, which is what we hope for the steady state to correct. For shorter trajectories, there is more initial state bias, while for fewer trajectories, there is lower statistical precision overall. Figures 2(a) and 2(b) show the exhaustively sampled case (1.5×10^6 total steps), either in a scenario of vanishing bias (0.15×10^6 steps per trajectory), (a), or in a scenario of a strong and consistent initial state bias (150 steps per trajectory), (b). Among the methods tested, only the unconstrained maximum likelihood

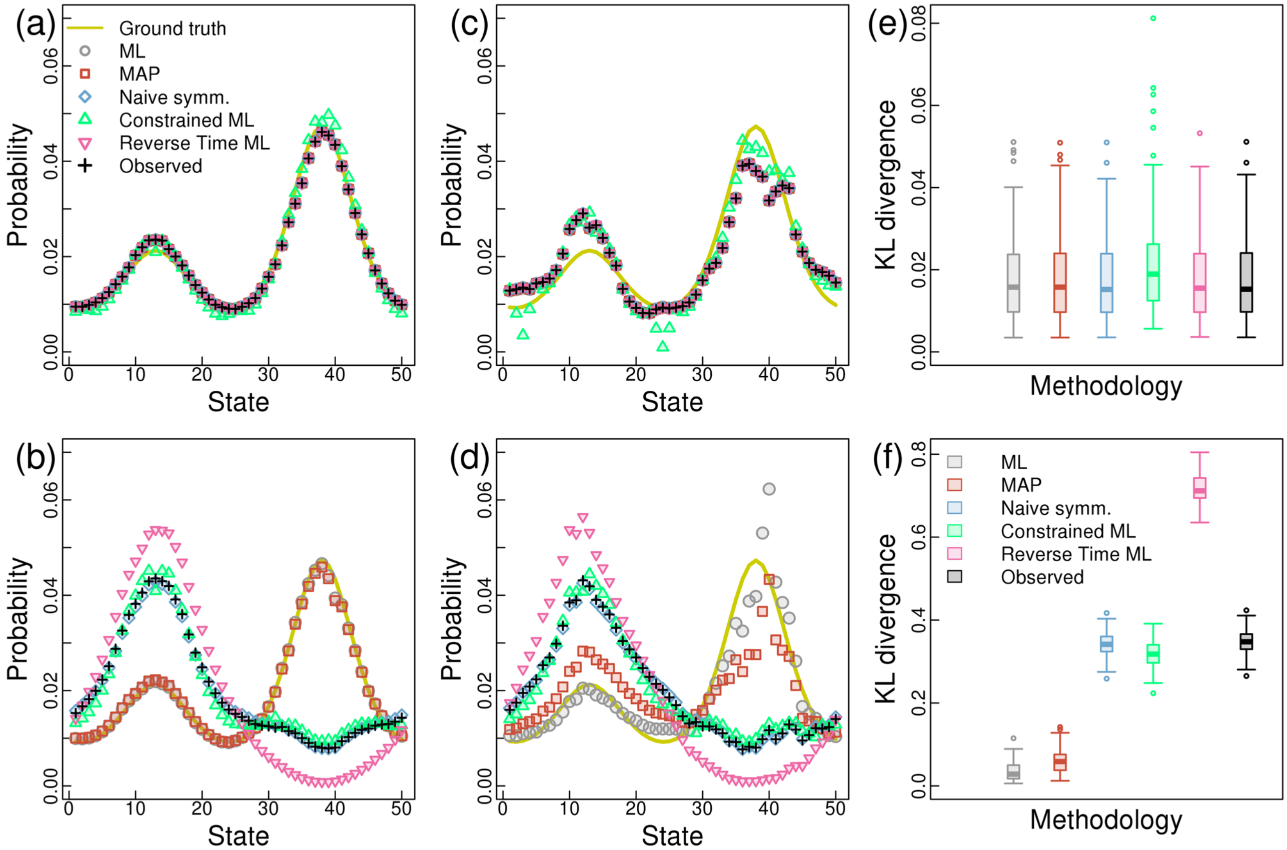

FIG. 2. Local flux imbalances must be preserved for thermodynamic reweighting. (a) Data for 10 trajectories of 150 000 steps each. (b) Data for 10 000 trajectories of 150 steps each. (c) Data for 1 trajectory of 45 000 steps. (d) Data for 300 trajectories of 150 steps each. The results in (c) and (d) are individual examples of noisy data sets. (e) Tukey-rule boxplots of the Kullback-Leibler (KL) divergences between the ground truth and the predicted steady state for 100 repetitions of the settings in (c). See (f) for the legend. (f) Tukey-rule boxplots of the KL divergences between the ground truth and the predicted steady state for 100 repetitions of the settings in (d).

(ML) estimate of the transition matrix, Eq. (1), and the maximum *a posteriori* (MAP) estimate, Eqs. (3)–(5), predict the correct steady state in both cases. It is expected that the pseudocounts from the prior distribution do not matter in this well-sampled regime. The imposition of detailed balance, regardless of methodology, appears to fix the steady state to the observed sampling weights. This is appropriate only if the observed sampling weights are approximately correct. The failure in Fig. 2(b) can be understood as the result of explicitly removing local flux imbalances, which is precisely where the initial state bias appears to be encoded. This phenomenon is also highlighted effectively by considering the reverse-time transition matrix, which, consistent with this explanation, amplifies rather than weakens the initial state bias.

Figures 2(c)–2(f) demonstrate what happens as the amount of overall sampling is decreased to 3% of that in (a)–(b). If the initial state bias is small [only a single random walker was used, Fig. 2(c)], the only estimator that deviates from the sampled distribution is the ML-estimated transition matrix under the detailed balance constraint [Eq. (2)]; see Fig. 2(e). Rather than improving the estimate, however, it worsens it slightly but significantly (Welch two-sample t-test relative to straight ML estimate indicates difference in means with ~98% confidence). The fact that the reverse-time ML estimate exhibits no significant difference confirms that this simulation is at equilibrium. In the presence of initial state bias [Fig. 2(d)], which is theoretically identical in magnitude to the scenario in Fig. 2(b), it becomes clear that the ML estimator performs best for this system [Fig. 2(f)]. Importantly, the results from Fig. 2(b) are confirmed when considering repeated but noisier measurements: detailed balance imposition performs similarly to the observed distribution, and the reverse-time ML estimate amplifies the bias. Furthermore, the ML estimate performs significantly better also than the MAP estimate (Sec. II B). Indeed, all differences in Fig. 2(f) are highly significant (p-values of 10^{-6} or lower) with a single exception: the observed distribution and those from a naïve imposition of detailed balance are indistinguishable.

A few remarks are in order regarding the differences in Fig. 2(f). First, the prior chosen for the MAP estimate is not in the category of typical objective priors. It uses for the concentration parameters in the Dirichlet distribution a fixed value of $1 + 1/N$ [see Eq. (5)]. For comparison, a Jeffreys prior, whose guiding principle is for the posterior distribution to become invariant with reparametrizations of the problem, would use a value of 0.5. It thus seems as if the primary purpose of this prior in MSMs is not necessarily to prevent bias or incorporate prior information but to guarantee ergodicity in a "safe" manner. Second, the detailed balance-constrained ML estimate predicts a steady state that is slightly but significantly better than the raw sampling distribution. Because this method introduces fractional counts, it appears to add noise [see Fig. 2(c), in particular]. Evidently, this noise is not white: it is slightly harmful at equilibrium [Fig. 2(e)] but slightly beneficial in the presence of systematic biases [Fig. 2(f)].

B. A continuous space toy model

We next turn to a one-dimensional system. Like the system in Figs. 1 and 2, it is characterized by having two main states with unequal weights. However, we add the following elements that create a much more realistic use case. First, the system is defined and propagated in continuous space. This necessitates coarse-graining as a post-processing step,[34] which can introduce errors by compromising Markovianity given a choice of lag time. It also implies that all states have explicit and well-defined (mutually consistent) geometric distances from each other. Second, we employ an actual advanced sampling method, PIGS,[13] to generate a trajectory ensemble carrying initial state bias. This brings the statistical resampling (WE) strategy for reweighting into play (see the first paragraph of Sec. II E for a brief description of PIGS). Third, there is an explicit propagator controlling the evolution of individual trajectories outside of reseeding events. This allows us to use this information to construct an alternative Dirichlet prior for deriving a MAP estimate of the transition matrix as shown below.

At equilibrium, the population of the second (right) state is very low, so we represent positional distributions in logarithmic (free energy) space. This low population and the relatively slow barrier-crossing rate mean that distributions from 16 independent trajectories with no reseedings (conventional sampling, CS) carry significant errors [Fig. 3(a)]. For the data set shown, there is only about one crossing event per replica on average. Due to the absence of reseedings, the WE post-processing is inapplicable to CS. Figure 3(b) shows that the reweighting of these CS data with MSMs is largely ineffectual. Similarly to the results in Fig. 2(e), the constrained ML approach performs slightly but significantly worse. Given that initial state bias appears to be present in these trajectories [Fig. 3(a), they all started from the left state], it may seem confusing that MSMs fail to detect and remove any flux imbalance [Fig. 3(a), top]. However, as shown, the heuristic procedure of discarding the first half of the data for every replica also fails. This means that the primary source of error is not the initial state but the poor statistics, and these are clearly not corrigible by any of the MSMs evaluated.

Unlike for the first toy system, here the assumed lag time matters because the propagator operates in continuous space and Markovianity is not a given. Figures 3(d)–3(f) show data from a PIGS run using the same number of replicas and overall number of steps as CS. In terms of observed counts, this data set dramatically overestimates the population of the second state (on the right). This is because PIGS penalizes sampling redundancy, i.e., trajectories are rewarded for visiting both states equally. However, Fig. 3(e) demonstrates that, irrespective of lag time, all MSM-based reweighting approaches except those imposing detailed balance provide a better prediction of the ground truth than raw or reweighted CS data. The only exception to this rule is the MAP estimator at short lag times. For all other cases, the combination of the underlying Monte Carlo sampler and the chosen discretization appear to lead to the shortest possible lag time (we recorded data with a frequency of 10 elementary steps) being most appropriate. Why does the performance of the MAP estimator deteriorate at very short lag times? Clearly, the network is very sparsely and locally connected in this regime, and the propagator (see the Introduction and Sec. II C) fundamentally limits the reachable states. The imposition of prior information assuming complete connectivity thus violates the properties of the propagator encoded in the observed counts. We hypothesize that, as Fig. 3(e) shows, this can be harmful despite the relatively low weight of the pseudocounts. An alternative to the uniform prior is to use a prior distribution that is aware of the propagator, and we show one possible strategy for this as the MAP+ estimator in

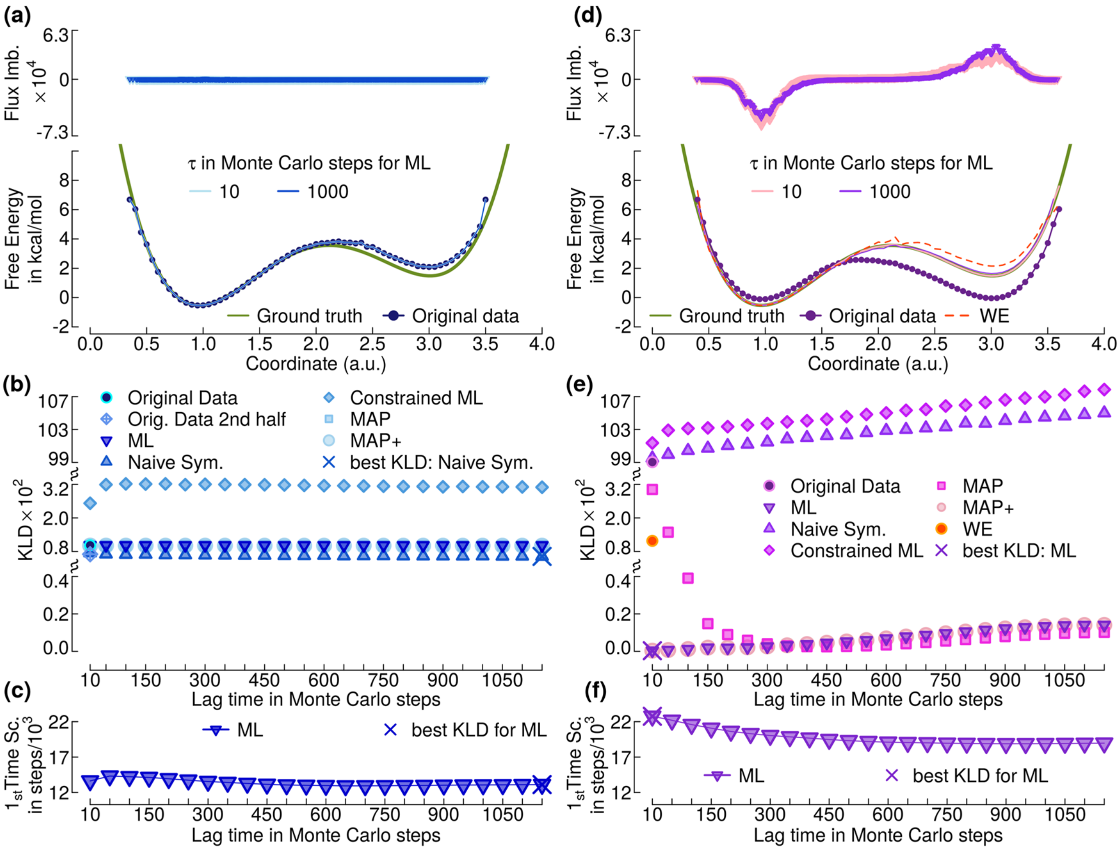

FIG. 3. **Results for a continuous space toy system.** A particle was propagated in the potential shown in olive in (a) and (d) by a Metropolis Monte Carlo sampler using a single move type at 298K: symmetric displacements sampled uniformly from an interval of 0.1a.u. All replicas started in the left basin throughout. (a)–(c) Data for 16 trajectories of 10^5 steps each with no reseedings (CS). The ground truth along with the potential inferred from the observed distribution is displayed in (a). Potentials from ML reweighting are shown as well but the lines superpose with the observed data (clustering yielded 425 states). The top of (a) (second y-axis) is a measure of the direction of the flux for two lag times (the same color code and x-axis apply). It is computed per state as the row sum of the matrix **B** where $b_{ij} = p_{raw,i}\, t_{ij} - p_{raw,j}\, t_{ji}$, and negative values indicate incoming flux, while positive values indicate outgoing flux. Panel (b) shows KL divergences from the ground truth for all attempted reweighting strategies and lag times. The point of best agreement across all attempts is highlighted. Finally, (c) shows the value of $-\tau/\ln\lambda_2$ where λ_2 is the 2nd largest eigenvalue of the ML estimate of $\mathbf{T}(\tau)$, and τ is the lag time. Here, this corresponds to the time to cross the barrier between the two states. (d)–(f) The same as (a)–(c) for a PIGS data set of identical extent (clustering yielded 445 states). In addition to the analogous data shown in (a)–(c), we add here the WE result in panels (d) (dashed line) and (e) (orange-red symbol). In panel (d), both ML results overlap with the ground truth. Note the y-axis discontinuities in (e). Figure S1 shows the same data analyzed with regular space binning instead of clustering. KLD is an abbreviation for KL divergence used throughout.

Fig. 3. It is derived by changing the α_i for the Dirichlet prior as follows:

$$\alpha_{ij} \propto p_{obs}(d_{ij}) \quad \text{and} \quad \sum_{j}^{N} \alpha_{ij} = 1. \tag{11}$$

This means that for each row, we add a total pseudocount weight of 1.0. However, rather than distributing it uniformly, it is added preferentially for geometrically reachable states, where p_{obs} is the estimated distribution of snapshot-to-snapshot distances connected in (temporal) sequence, which is conditional upon the propagator, metric, and lag time. As in Eq. (6), d_{ij} is the geometric distance of states, here represented by their cluster centroids. In practice, we estimate p_{obs} independently of the discretization by considering all time-connected pairs of snapshots (using a sliding window approach for lag times differing from unity). These data are binned finely, truncated to 99% of the cumulative distribution function, and finally rebinned to exactly 50 bins. This is to achieve consistent noise levels when analyzing the same data with different settings while avoiding the requirement to fit a function. The estimate of p_{obs} is frequently zero (when d_{ij} is large), and the distribution of pseudocounts resembles the structure of the ML-estimated transition matrix itself. However, this approach still adds prior information and thus works as a regularizer. For example, for the system in Fig. 3, for a case with 486 clusters of the particle position and a lag time of 1000 elementary steps, the raw data gave rise to ~15% of nonzero elements in **T**. While the uniform prior obviously ensures that all 100% of the elements are not zero, application of the nonuniform one still resulted in ~34% of all possible links having nonzero weights. As seen in Fig. 3(e), the MAP+ estimator performs as well as the ML one for all lag times.

The apparent absence of issues with Markovianity in Fig. 3 and Fig. S1 results from the relatively fine discretizations. We next asked what would happen for a much coarser partitioning, in particular the one that is informed by the true nature of the (free) energy landscape. Figure S2 shows results where the MSM was constructed with only two states, one to the left of the barrier and one to the right. Artifacts due to lack of Markovianity now appear, as expected, at short lag times. They arise because the states are so large that at short lag time the real system has significant memory as to where it was within a state. As seen in Fig. S2, these errors are avoidable by simply choosing larger lag times. With only two states, the pseudocounts added by the MAP methods are inconsequential. The important conclusions from Fig. S2 are (i) that the imposition of detailed balance is equally harmful as in Fig. 3; and (ii) that the reweighting with the ML and MAP MSMs still works quantitatively, albeit with the caveat that the resolution is limited [see Fig. S2(d)].

The final question we pose for this system regards the mechanism of reweighting. As shown in Figs. 3(d) and 3(e), MSMs are able to reweight a PIGS data set to high precision. The WE strategy also performs well although it carries more noise and slightly *under*estimates the weight of the right state. We were thus curious whether the sets of MSM- *vs.* WE-derived snapshot weights are comparable to each other. Figure 4 reveals that the MSMs are able to recognize the two-state nature of the system and allow the inference of weights that correct directly for the flux imbalance across the boundary. These weights are constant within a cluster. In contrast, the weights from statistical resampling can vary with each individual trajectory and are much noisier overall. They appear to differ fundamentally but, upon averaging, a similar albeit slightly more erratic curve is obtained.

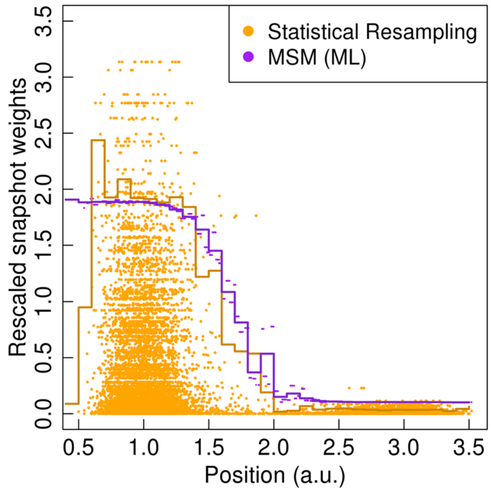

FIG. 4. **Comparison of weights for the continuous-space toy system of Fig. 3.** For an example PIGS run, per-snapshot weights are compared for the WE, Eq. (10), and MSM strategies [the latter are given as $\pi_{c_i}/p^{obs}_{c_i}$, compare Eq. (9)]. Individual weights are plotted as a function of position as dots, and averages across bins in position space (width of 0.1 a.u.) are shown as steps.

C. A real-world application: Molecular dynamics simulations of the FS-peptide

The final system we consider in this manuscript is a 21-residue peptide with sequence Acetyl-$A_5(AAARA)_3A$-N'-methylamide simulated using a Cartesian Langevin dynamics integrator in implicit solvent. At low enough temperatures, this peptide, known as FS-peptide, undergoes reversible folding transitions between states rich in α-helix, coil-like states, and collapsed globules in the chosen computational model. We showed previously that the application of PIGS to this system uncovered a complex network of states with many low-likelihood but significantly metastable states. At the time, we did not attempt to reweight the observed distribution to a predicted equilibrium one, and this is the objective pursued here. For this, we reanalyzed the published data sets obtained at 250 K,[13] which are described further below.

Compared to the one-dimensional system considered above, the FS-peptide poses the same tasks to address. The complexity is massively increased, however. Even when considering a simplistic model of two states per residue, the resultant state space is of size 2^{21}. Thus, we are forced to rely on a data-driven discretization (here, tree-based clustering),[34] and the choice of representation is nontrivial. The metrics used for generating the PIGS data were high-dimensional and based on either 76 dihedral angles (DPIGS below) or 145 interatomic distances (RPIGS below). For the analyses presented here, we focused instead on a single representation composed of the ϕ/ψ-angles of the 17 central residues of the FS-peptide. In such high-dimensional spaces, phenomena summarized under the umbrella term "curse of dimensionality" come into play. In particular, the spectrum of conformational distances becomes highly compressed, which means that neighbor relations are difficult to establish in purely geometric terms.

Before presenting the new analyses, it is important to recapitulate relevant findings from our earlier work.[13] First, we use raw results from long MD trajectories as our gold standard (GS). This is justified because we showed that, for two completely different starting conformations, the distributions of simple order parameters like size and helicity converged to the same distributions. This was also true for PIGS runs for two different metrics. Importantly, the resultant PIGS distributions differed systematically from the GS ones, and the observed bias was not only independent of the starting conformations but also nearly independent of the metric. Thus, at least in low-dimensional projections, we face a statistically robust bias that we wish to remove. We also noted that the GS simulations did not visit all of the states explored by PIGS. For a high-dimensional system, it is inevitable that the limits of the sampling domains differ between data sets, which means that it is unclear how reliable the GS actually is. This is why it is important that the aforementioned two starting conformations were either the dominant state (straight α-helix) or one of the nonhelical states discovered only by PIGS. Due to the clear convergence between those results and additional results from replica exchange simulations,[13] we are prone to trust the GS quantitatively.

Figure 5(a) shows a comparative analysis of distributions of the radius of gyration. For the FS-peptide, this quantity has a main peak close to 10 Å that is due to the straight α-helix.[49] In contrast, the peak (or peaks) at smaller values results from a mix of compact states including both helix bundles and nonhelical states. In the raw

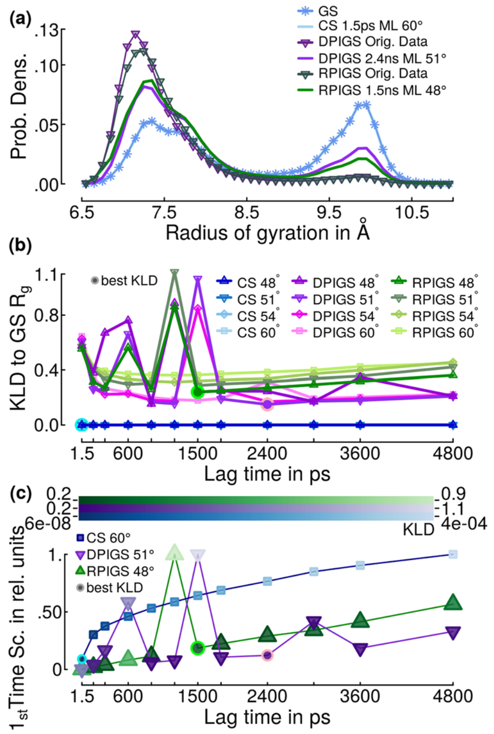

FIG. 5. **Reweighting of the radius of gyration from three data sets using ML inference**. RPIGS and DPIGS are PIGS runs whose only difference was in the metric (see the text). CS stands for conventional sampling, and the cumulative simulation time was identical to the PIGS runs. The gold standard (GS) is the raw distribution from CS. (a) Only results from the best-performing MSMs for the three data sets are shown (all are based on ML inference with lag time and clustering resolution given in the legend). (b) KL divergences from the GS. The measure has to ignore empty bins in the GS. Data are shown for networks differing in clustering resolutions and lag times. MSMs constructed for CS data have a negligible impact on the distribution. The point of maximal agreement is highlighted for every data set. (c) The slowest time scale is shown as a function of lag time and in relative units (to facilitate visualization on the same graph). The value corresponds to $-\tau/\ln\lambda_2$ where λ_2 is the 2nd largest eigenvalue of the ML estimate of $\mathbf{T}(\tau)$, and τ is the lag time. The normalization is by the corresponding maximum value within each data set (CS, DPIGS, or RPIGS). Only trends derived from a single clustering are shown for each data set, and the best-performing MSMs are highlighted as in (b). The color of the symbols indicates the KL divergence (color legend on top). Note that the CS-ML result overlaps with the GS in (a), and that all CS results overlap in (b).

PIGS data, the population of the straight helix is much lower than that in the GS. The limited view offered by a projection onto a relevant geometric variable like the radius of gyration is enough to make the following point. Even fine discretizations of very large data sets (6.656×10^6 snapshots, up to 268 342 states) coupled to an exhaustive scan of lag times do not allow us to find an MSM that is able to quantitatively recover the GS distribution. This holds for either PIGS data set. Clearly, the MSMs do detect a flux imbalance and the reweighted distributions are closer to the GS than the unweighted ones, but the performance is not comparable to that in Fig. 2 or Fig. 3. Thus, in practice, if there is no estimate of the GS, it will be difficult to deduce much more than a direction for the required correction.

From Fig. 5(b), it becomes clear why a regularization of **T** is often desired. The lag time and resolution dependencies of the deviations [Fig. 5(b)] are generally systematic and suggest that both too short and too long lag times as well as too coarse representations are inappropriate (unless the data are already at equilibrium as for CS), which is in accord with observations and recommendations by others.[14,50] However, these systematic trends are interrupted by individual outliers, which arise preferably for smaller lag times and finer resolutions. For these particular MSMs, the prediction of equilibrium deteriorates dramatically, and this feature is mirrored in kinetic analyses. The slowest relaxation time scales shown in Fig. 5(c) follow a similar trend for the two discretizations of PIGS data shown: clear outliers in kinetics appear to be predictive of the failure to reweight. This can be explained, for example, by a very poorly balanced transition into a normally insignificant state. Conversely, the general trend of increasing relaxation time with increasing lag time is not predictive of MSM performance. There is no consistent plateau for the three different data sets in Fig. 5(c), yet the performance metric in Fig. 5(b) has already stepped through a minimum region.

We were of course curious to see if the various regularization schemes could improve the reweighting. As expected, a naïve imposition of detailed balance leads to a very well-behaved transition matrix that performs no appreciable reweighting (Fig. S3). While such a matrix may be useful to study specific kinetic processes, its utility as a quantitative prediction tool is limited. This has been noted in recent applications in the literature.[37,51] The constrained ML scheme [see Eq. (2)] unfortunately seems to combine the downsides of both worlds: it neither makes **T** well-behaved nor is the equilibrium distribution reweighted appreciably (Fig. S4). With the simple Bayesian regularizer (MAP, flat prior), we are constrained here by the fact that it creates a maximally dense matrix. As a result, numbers of states exceeding ~10^4 are difficult to deal with routinely because of the numerical complexity. Figure S5 shows that, possibly due to the limits on clustering resolution, the best-case reweighted efforts are worse than those for the ML case seen in Fig. 5. The MAP+ prior has the advantage that it retains the sparsity of **T** to a significant degree. In the analysis here, this is the alternative to the ML estimate that gets closest in peak performance (see Table II below). However, there is no real gain as the results are both worse and more erratic than those for the ML estimate (compare Fig. S6 to Fig. 5). Figure 6 shows equivalent data for another observable, viz., α-helical content. Here, the raw PIGS data underestimate the sampling weight of helix-rich states and (correspondingly) overestimate that of nonhelical states. This is not fully corrigible by any of the evaluated MSMs (see also Figs. S7–S10).

The remaining errors in Figs. 5 and 6 can originate from at least two sources: poor statistics as in Fig. 3(a) or a failure to find a suitable combination of discretization and lag time to preserve Markovianity. In the above, we have restricted ourselves to MSMs built on the idea of a unique and exhaustive mapping from conformation to states. A rich body of literature exists on using objective functions rooted in kinetic properties (metastability) to derive, improve, or optimize such models.[53–56] An important alternative is to drop the requirement of a unique mapping. For example, transition state theory and transition path sampling are concerned with sets of states that leave the transition regions between them unassigned.[57,58] In this logic, the state vector changes only when the

TABLE II. Comparison of KL divergences from the GS for the best cases for the reweighting of PIGS data. The numbers provided are the minima across the two PIGS data sets and, for MSMs, across all lag times and resolutions we scanned. Results are provided separately for the tested inference methods. The MSM representation was always the same, viz., the ϕ- and ψ-angles of the central 17 residues of FS-peptide. The WE approach has no parameters except the metric with its underlying representation (here, R_g and α-content).

	Markov state models based on TORS rep. (34 ϕ/ψ angles)					Statistical resampling	
Observable	ML	Cons. ML	Naïve sym.	MAP	MAP+	34 ϕ/ψ	2D rep.
R_g	0.15	0.61	0.59	0.27	0.23	0.07	0.01
α-helicity	0.13	0.91	0.92	0.25	0.25	0.10	0.03

domain of a new state is entered, which means that the same conformation can be assigned different states based on where that particular trajectory originated from. This logic has been taken further in methods based on milestoning[59] where MSMs are built based on (few) metastable states.[38,60,61] These are attractive methods if the dynamical properties of the system allow for their applicability. An alternative approach for allowing the mapping to become inexact is to take a probabilistic view of state memberships.[62] It is clearly possible that the reweighting quality could be improved further by some of these methods. We emphasize again, however, that it is also possible that the statistics are simply too poor, i.e., that the remaining errors are sporadic.

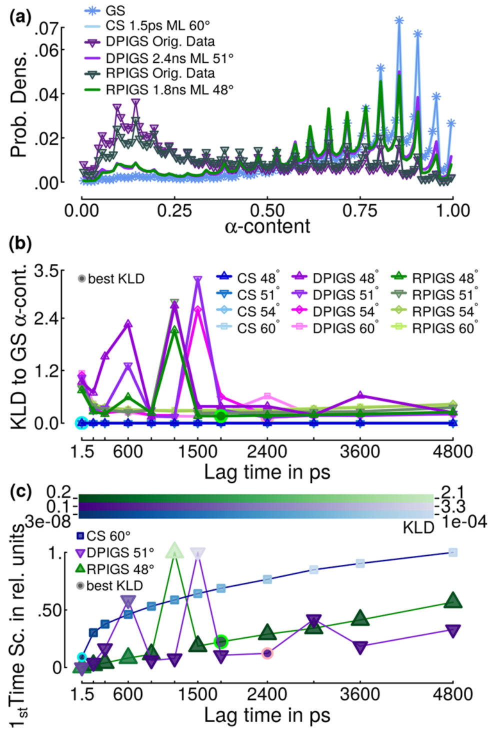

FIG. 6. Reweighting of α-helical content from three data sets using ML inference. This figure is completely analogous to Fig. 5 only that data for the peptide's α-helical content are shown. This quantity was computed using a functional form as published.[52] The α-region was a circle of 35° radius centered at −50°/−60° (ϕ/ψ), and the decay parameter was 0.002 deg^{-2}. (a) Comparison of distributions. (b) KLDs from the GS. (c) Slowest time scales. The spikes in (a) occur because the measure is a smoothed version of the (integer) number of residues in the α-basin. As in Fig. 5, the CS-ML result overlaps with the GS in (a), and all CS results overlap in (b).

Of course, we also applied the WE strategy to our PIGS data. Since the underlying data are now high-dimensional, the choice of metric becomes a parameter. As explained in Sec. II E, PIGS does not guarantee that, for a given reseeding event, a surviving trajectory is available that is also geometrically nearby. Intuitively, metrics of lower dimensionality should thus perform better as long as they are coupled to the observable of interest because they offer a higher chance of proximity/overlap. Incidentally, this is the same reason why low-dimensional projections onto geometric variables are difficult to use without optimization in transition-based analyses like MSMs.[8,58,63–66] Figure 7 presents the reweighted distributions for both observables we obtained by using different metrics in the WE formalism. These results are compared to the GS and to the best-case scenario for MSMs, and the respective KL divergences are summarized in Table II.

From Fig. 7, the WE approach relying on a low-dimensional metric composed from observables of interest emerges as a viable strategy for the thermodynamic reweighting of simulation data on complex systems carrying initial state bias. While the agreement is not perfect, these results correctly identify that the statistical weight of very compact, nonhelical states is very low despite their demonstrated metastability.[13] It is also an important observation that the equilibrium estimate obtained with the same metric as that used to construct the MSMs is still improved. This is despite the fact that there are no parameters to optimize as we did for the MSMs. Taken together, Fig. 7 and Table II suggest that the WE strategy is superior for the purpose of the removal of initial state bias from suitable trajectory ensembles, at least for low-dimensional observables and when the underlying data are high in dimensionality. This contrasts with the results for the one-dimensional system of Fig. 3 where we found MSMs to be quantitatively superior.

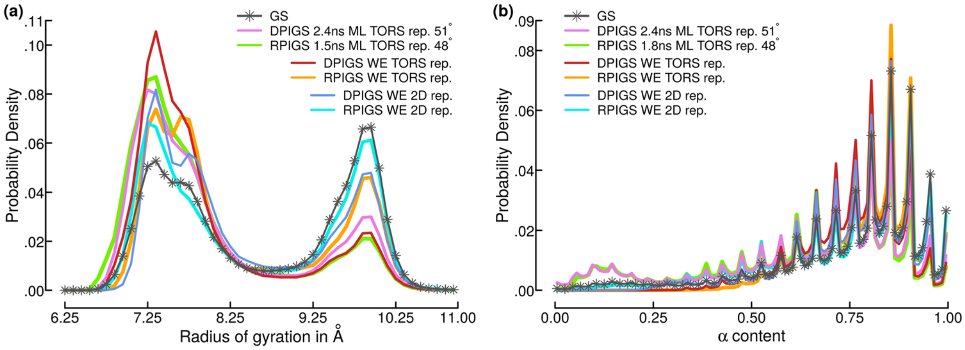

FIG. 7. **Comparison of the quality of MSM-based reweighting with the WE approach.** The acronym "TORS rep." stands for the representation of the data by 34 dihedral angles (ϕ/ψ-angles of the 17 central residues). "2D rep." instead stands for a two-dimensional representation composed of the two observables we analyze here, viz., R_g and α-content. (a) Data for the radius of gyration. Only the best MSMs obtained with "TORS rep." are shown for the two PIGS data sets, and resolution and lag time are given in the legend. (b) The same as (a) for α-content. Some of the symbols for the GS are omitted in the left half of the plot to improve readability.

IV. CONCLUSIONS

In summary, our results give rise to the following conclusions. We formulate these conclusions with a specific but common task in mind: to recover correctly weighted equilibrium distributions from data sets marred by initial state bias. Examples for those data sets are trajectory ensembles generated by advanced sampling methodologies that explore phase space in an adaptive manner. However, we emphasize that the vast majority of biological applications of molecular simulations suffer from this bias regardless of sampling strategy as they typically rely on an experimentally derived starting structure.

- Markov state models are theoretically sound tools to reweight simulation data carrying initial state bias (Figs. 2 and 3). The use of a purely data-driven methodology for discretization (clustering) is not problematic *per se.* The bias has to be consistent for it to be removable, i.e., it must not derive from (randomly) poor statistics for undersampled transitions [Fig. 3(b)].
- Consistent flux imbalances found in the raw data must be preserved in the Markov model of the data for the reweighting to be successful. This rigorously excludes all approaches imposing detailed balance onto the transition matrix from being useful for the reweighting task, in line with prior observations.[10] We showed this to hold even for a toy system that is, by construction, Markovian and free of discretization artifacts (Fig. 2).
- Markov state models may fail in recovering the correct equilibrium distribution if the underlying data are of high dimensionality (Fig. 7). Finer discretizations can work better than coarse ones but are also more erratic, and this may be difficult to diagnose.
- A plateau in relaxation time with lag time is not a useful criterion to find optimal models for thermodynamic reweighting. In the cases studied here, the suggested lag times based on this heuristic would almost all have been too large (Figs. 3, 5, and 6).
- The use of prior information generally fails to improve the predictions made based on the maximum likelihood inference of the transition matrix (Figs. 2, 3, 5–7). This holds for all the approaches we evaluated.
- For data sets that can be cast in the logic of a statistical resampling procedure, the WE approach is a viable alternative. If the raw data are of high dimensionality, a particular advantage is that it allows the metric to be optimized for observables of interest (Fig. 7), which is unfeasible for MSMs.
- The use of an advanced sampling technique along with a successful reweighting approach allows phase space to be explored more rapidly and, potentially, also more correctly (Fig. 3) than by investing the same computing time into long, independent simulations.

From these conclusions, we formulate the following strategies for high-dimensional data derived from complex systems such as proteins. When applicable, the raw data should be reweighted with the WE methodology using both general representations (like the 34 dihedral angles in our case) and observable-specific representations. If there is an approximate consensus among these, this will provide strong evidence that the reweighting is meaningful. It is unfortunately not possible to trust the WE results unconditionally, which has two main reasons. First, in our data sets, we were restricted to use the closest available trajectory for absorbing the weight of a terminated one. This is, by definition, inexact. Second, the repeated splitting and merging of weights will sometimes reduce the effective sample size dramatically and in a nonrandom manner, i.e., it can introduce noise that is difficult to interpret. For example, reweighting the DPIGS data set with the original DPIGS representation fails completely, and both the aforementioned problems likely contribute to this. In a second step, the reweighted distributions could then be used to construct a *de novo* kinetic network model following Eq. (6) where the similarity kernel, H, can be inferred from the data as we did for the MAP+ prior [Eq. (11)].

For traditional MSMs aiming for a complete and unique mapping from conformation to states, we recommend a pure likelihood-based inference of $\mathbf{T}$ for a fine-grained set of states. In our experience, scanning a number of resolutions for discretization and a number of lag times is necessary to distinguish robust trends from erratic results (Figs. 5 and 6 and Figs. S3–S10). This is labor-intensive, both in computational and in human terms. Unfortunately, the prediction from an individual MSM can be arbitrarily wrong, in particular, in the most promising regime. Our coarse-resolution results for the MAP estimate (Figs. S5 and S9) show that there is some value in restricting oneself to a "safe" regime, but this comes at the cost of reducing the quantitative correctness of the reweighted result further. Relaxation times can be used to diagnose outliers but should not be used to guide the choice of lag time and resolution based on the notion of plateauing. Two additional directions for future research may be to explore general ways to cast arbitrary trajectory ensembles as statistical resampling problems and to develop the MAP+ prior further.

SUPPLEMENTARY MATERIAL

Figures S1–S2 (related to Fig. 3), S3–S6 (related to Fig. 5), and S7–S10 (related to Fig. 6) are included in a single file as supplementary material.

ACKNOWLEDGMENTS

This work was supported by a grant from the Swiss National Science Foundation (SNF Grant No. 169007) to A.C. Early work on the topic was inspired by and benefitted from a grant from the Platform for Advanced Scientific Computing (PASC "Scalable Advanced Sampling in Molecular Dynamics") to A.V. and A.C., and by a CPU time grant from CSCS ("s636") to M.B., A.C., and A.V. We are grateful to Cassiano Langini, Pietro Faccioli, and Gregory R. Bowman for interesting discussions. Min Xu was a valuable contributor and collaborator in earlier attempts to address the reweighting problem.

REFERENCES

[1] A. J. G. Cairns, D. Blake, and K. Dowd, J. Risk Insur. **73**, 687 (2006).
[2] L. R. Rabiner, Proc. IEEE **77**, 257 (1989).
[3] S. R. Eddy, Curr. Opin. Struct. Biol. **6**, 361 (1996).
[4] G. R. Bowman, K. A. Beauchamp, G. Boxer, and V. S. Pande, J. Chem. Phys. **131**, 124101 (2009).
[5] J. D. Chodera and F. Noé, Curr. Opin. Struct. Biol. **25**, 135 (2014).
[6] M. Karplus and J. A. McCammon, Nat. Struct. Biol. **9**, 646 (2002).
[7] A. Berezhkovskii, G. Hummer, and A. Szabo, J. Chem. Phys. **130**, 205102 (2009).
[8] S. V. Krivov, S. Muff, A. Caflisch, and M. Karplus, J. Phys. Chem. B **112**, 8701 (2008).
[9] M. Dibak, M. J. del Razo, D. De Sancho, C. Schütte, and F. Noé, J. Chem. Phys. **48**, 214107 (2018).
[10] R. Scalco and A. Caflisch, J. Phys. Chem. B **115**, 6358 (2011).
[11] G. R. Bowman, D. L. Ensign, and V. S. Pande, J. Chem. Theory Comput. **6**, 787 (2010).
[12] A. C. Pan, D. Sezer, and B. Roux, J. Phys. Chem. B **112**, 3432 (2008).
[13] M. Bacci, A. Vitalis, and A. Caflisch, Biochim. Biophys. Acta **1850**, 889 (2015).
[14] J.-H. Prinz, H. Wu, M. Sarich, B. Keller, M. Senne, M. Held, J. D. Chodera, C. Schütte, and F. Noé, J. Chem. Phys. **134**, 174105 (2011).
[15] F. Nüske, H. Wu, J.-H. Prinz, C. Wehmeyer, C. Clementi, and F. Noé, J. Chem. Phys. **146**, 094104 (2017).
[16] G. Hummer, New J. Phys. **7**, 34 (2005).
[17] S. V. Krivov and M. Karplus, Proc. Natl. Acad. Sci. U. S. A. **105**, 13841 (2008).
[18] R. B. Best and G. Hummer, Phys. Rev. Lett. **96**, 228104 (2006).
[19] M. J. Klein, Phys. Rev. **97**, 1446 (1955).
[20] V. I. Manousiouthakis and M. W. Deem, J. Chem. Phys. **110**, 2753 (1999).
[21] M. E. Cates, Rep. Prog. Phys. **75**, 042601 (2012).
[22] R. M. L. Evans, J. Phys. A: Math. Gen. **38**, 293 (2005).
[23] G. R. Bowman, J. Chem. Phys. **137**, 134111 (2012).
[24] S. Bacallado, J. D. Chodera, and V. S. Pande, J. Chem. Phys. **131**, 045106 (2009).
[25] N. Singhal and V. S. Pande, J. Chem. Phys. **123**, 204909 (2005).
[26] N.-j. Deng, W. Zheng, E. Gallicchio, and R. M. Levy, J. Am. Chem. Soc. **133**, 9387 (2011).
[27] W. Zheng, E. Gallicchio, N. Deng, M. Andrec, and R. M. Levy, J. Phys. Chem. B **115**, 1512 (2011).
[28] W. Han and K. Schulten, J. Am. Chem. Soc. **136**, 12450 (2014).
[29] W. Zheng, M. Andrec, E. Gallicchio, and R. M. Levy, J. Phys. Chem. B **113**, 11702 (2009).
[30] A. Berezhkovskii and A. Szabo, J. Chem. Phys. **135**, 074108 (2011).
[31] J. S. Liu, A. F. Neuwald, and C. E. Lawrence, J. Am. Stat. Assoc. **90**, 1156 (1995).
[32] S. F. Chen and J. Goodman, Comput. Speech Lang. **13**, 359 (1999).
[33] B. Trendelkamp-Schroer and F. Noé, J. Chem. Phys. **138**, 164113 (2013).
[34] A. Vitalis and A. Caflisch, J. Chem. Theory Comput. **8**, 1108 (2012).
[35] N. Blöchliger, A. Vitalis, and A. Caflisch, Comput. Phys. Commun. **184**, 2446 (2013).
[36] M. Bacci, C. Langini, J. Vymětal, A. Caflisch, and A. Vitalis, J. Chem. Phys. **147**, 195102 (2017).
[37] M. Bacci, J. Vymětal, M. Mihajlovic, A. Caflisch, and A. Vitalis, J. Chem. Theory Comput. **13**, 5117 (2017).
[38] E. Vanden-Eijnden and M. Venturoli, J. Chem. Phys. **130**, 194101 (2009).
[39] M. I. Zimmerman and G. R. Bowman, J. Chem. Theory Comput. **11**, 5747 (2015).
[40] A. Dickson and C. L. Brooks III, J. Phys. Chem. B **118**, 3532 (2014).
[41] G. A. Huber and S. Kim, Biophys. J. **70**, 97 (1996).
[42] F. Cérou and A. Guyader, Stoch. Anal. Appl. **25**, 417 (2007).
[43] B. W. Zhang, D. Jasnow, and D. M. Zuckerman, J. Chem. Phys. **132**, 054107 (2010).
[44] C.-E. Brehier, M. Gazeau, L. Goudenege, T. Lelievre, and M. Rousset, Ann. Appl. Probab. **26**, 3559 (2016).
[45] D. J. Lockhart and P. S. Kim, Science **257**, 947 (1992).
[46] L. J. Smith, X. Daura, and W. F. van Gunsteren, Proteins Struct. Funct. Bioinf. **48**, 487 (2002).
[47] G. M. Torrie and J. P. Valleau, J. Comput. Phys. **23**, 187 (1977).
[48] R. H. Swendsen and J. S. Wang, Phys. Rev. Lett. **57**, 2607 (1986).
[49] A. Vitalis and R. V. Pappu, J. Chem. Phys. **141**, 034105 (2014).
[50] M. Sarich, F. Noé, and C. Schütte, Multiscale Model. Simul. **8**, 1154 (2010).
[51] G. Zhou, G. A. Pantelopulos, S. Mukherjee, and V. A. Voelz, Biophys. J. **113**, 785 (2017).
[52] A. Vitalis, N. Lyle, and R. V. Pappu, Biophys. J. **97**, 303 (2009).
[53] J. D. Chodera, N. Singhal, V. S. Pande, K. A. Dill, and W. L. Swope, J. Chem. Phys. **126**, 155101 (2009).
[54] Y. Yao, R. Z. Cui, G. R. Bowman, D.-A. Silva, J. Sun, and X. Huang, J. Chem. Phys. **138**, 174106 (2013).
[55] B. Fačkovec, E. Vanden-Eijnden, and D. J. Wales, J. Chem. Phys. **143**, 044119 (2015).
[56] R. T. McGibbon and V. S. Pande, J. Chem. Phys. **142**, 124105 (2015).
[57] P. G. Bolhuis, C. Dellago, D. Chandler, and P. L. Geissler, Annu. Rev. Phys. Chem. **53**, 291 (2002).
[58] G. Hummer, J. Chem. Phys. **120**, 516 (2004).
[59] A. K. Faradjian and R. Elber, J. Chem. Phys. **120**, 10880 (2004).
[60] C. Schütte, F. Noé, J. Lu, M. Sarich, and E. Vanden-Eijnden, J. Chem. Phys. **134**, 204105 (2011).

[61] E. Guarnera and E. Vanden-Eijnden, J. Chem. Phys. **145**, 024102 (2016).
[62] S. Röblitz and M. Weber, Adv. Data Anal. Classif. **7**, 147 (2013).
[63] P. V. Banushkina and S. V. Krivov, Wiley Interdiscip. Rev.: Comput. Mol. Sci. **6**, 748 (2016).
[64] F. Noé, C. Schütte, E. Vanden-Eijnden, L. Reich, and T. R. Weikl, Proc. Natl. Acad. Sci. U. S. A. **106**, 19011 (2009).
[65] B. Peters and B. L. Trout, J. Chem. Phys. **125**, 054108 (2006).
[66] Y. M. Rhee and V. S. Pande, J. Phys. Chem. B **109**, 6780 (2005).