Bayesian comparison of Markov models of molecular dynamics with detailed balance constraint

Sergio Bacallado,[1] John D. Chodera,[2] and Vijay Pande[1,3,a)]
[1]*Department of Structural Biology, Stanford University, Stanford, California 94305, USA*
[2]*California Institute for Quantitative Biosciences (QB3), University of California, Berkeley, California 94720, USA*
[3]*Department of Chemistry, Stanford University, Stanford, California 94305, USA*

(Received 14 April 2009; accepted 10 July 2009; published online 30 July 2009)

Discrete-space Markov models are a convenient way of describing the kinetics of biomolecules. The most common strategies used to validate these models employ statistics from simulation data, such as the eigenvalue spectrum of the inferred rate matrix, which are often associated with large uncertainties. Here, we propose a Bayesian approach, which makes it possible to differentiate between models at a fixed lag time making use of short trajectories. The hierarchical definition of the models allows one to compare instances with any number of states. We apply a conjugate prior for reversible Markov chains, which was recently introduced in the statistics literature. The method is tested in two different systems, a Monte Carlo dynamics simulation of a two-dimensional model system and molecular dynamics simulations of the terminally blocked alanine dipeptide.
 [DOI: 10.1063/1.3192309]

I. INTRODUCTION

Many molecules of interest in chemistry and medicine, such as proteins and RNA, are widely believed to have a hierarchical free energy landscape[1–3] consisting of many low-energy wells in high-dimensional space. In a molecular dynamics (MD) simulation, these regions of conformational space exhibit metastability,[4] which can hinder the sampling of uncorrelated configurations from the equilibrium distribution in a computationally viable time. Fortunately, this phenomenon enables approximations that result in discrete or continuous-time stochastic models of conformational dynamics, which are used to derive kinetic information from MD simulations shorter than the slowest timescales of the system.

Mori and Zwanzig laid the groundwork for stochastic modeling of conformational dynamics by applying the projection operator formalism to equilibrium classical mechanics.[5,6] Given a partition of conformational space into discrete states, this method results in a generalized master equation that exactly describes the evolution of the densities of each state in a canonical ensemble by the introduction of a memory function to capture the dependence on the history of previously visited states. A Markov approximation of this process, in which the memory is assumed to be infinitely short, gives rise to a continuous-time random walk with exponential waiting times from which kinetic rates may be easily obtained.[7] However, the emergence of Markovian behavior in coarse-grained dynamics has only been proven for a few microscopic models.[8,9] Chandler investigated conditions for the Markov assumption, or more generally, for the phenomenological observation that a chemical reaction between two states obeys a first-order rate law.[10] This work was later generalized to multistate systems.[12,11] It was shown that the rate constant of an infrequent transition across an energy barrier, τ_{rxn}^{-1}, is the time integral of the Mori–Zwanzig memory function. This function has a singular component at $t=0$, whose integral corresponds to the transition state theory estimate of the rate, while its nonsingular part gives rise to a transmission coefficient that corrects this estimate. The phenomenological rate law requires a separation of time scales between the dynamics along the reactive path and on the degrees of freedom orthogonal to it. This assumption leads to two expressions for τ_{rxn}^{-1} in terms of correlation functions, including the popular reactive flux formula,[10] which is independent of the placement of the dividing surface as long as a separation of time scales exists.

There is a multitude of methods to optimize reactive paths or coordinates and find rate constants.[13–18] While this task is relatively straightforward for systems with two states, as the number of metastable states increases, the problem of determining the transition rates between all pairs of states becomes prohibitively expensive. Transition path sampling techniques are being developed for computing multiple rate constants simultaneously,[19] but even these methods require well-defined metastable states. A simpler approach for large systems is to coarse grain the time domain at a fixed observation interval or *lag time*, τ_{lag}, at which the dynamics on the full, discrete state space become well approximated by a Markov chain. This would happen if the chosen τ_{lag} is greater than the time it takes to decorrelate within any given state, but much smaller that the characteristic timescale of a transition. In this regime, when a simulation starts within a state, there will be many independent attempts to leave it before a transition occurs—the system will effectively lose memory of where it started within the state, which is equivalent to the Markov condition.

Defining a Markov model is a twofold task, which involves projecting the dynamics onto a discrete partition of

a)Electronic mail: pande@stanford.edu.

configurational space and finding the shortest lag time at which a projected trajectory is well modeled by a Markov chain. Much effort has been put into algorithms that perform this task sequentially,[20–27] but their validation methods remain largely heuristic. The simplest way to evaluate a model is to compare the evolution of state populations in a simulation to that predicted by the model.[20] Another validation technique makes use of rate matrices inferred from the observed transitions at a series of lag times. The eigenvalues of the rate matrix, which are related to the characteristic time scales of concerted transitions defined by the corresponding eigenvectors, should be constant for lag times greater than the Markovian lag time. This is a consequence of the Chapman–Kolmogorov equation, which governs Markov processes[28] and is also consistent with Chandler's reactive flux theory. Swope *et al.*[29] proposed using these *implied time scales* to define the Markovian lag time as the smallest lag time that satisfies this condition. Buchete *et al.* suggested projecting the dynamics onto different eigenvectors of an inferred rate matrix and comparing the decay of the cross-correlation functions predicted by a master equation model to that inferred directly from the data.[30] Another interesting approach to model validation uses an entropy function to measure the information lost when going from a second-order Markov model to a first-order Markov model.[31]

All the cited methods rely on statistics of the simulation, which are often poor, producing large uncertainties. Here, we propose a subjective or Bayesian approach to model validation. This approach makes it possible to use all the data available to evaluate generative models quantitatively. In particular, we define a hierarchical model that can be used to select an optimal state decomposition at any fixed lag time. The method does not confirm the Markov assumption, nor does it provide an estimate of the error incurred in making it. However, when used in tandem with the cited heuristics, the method affords several advantages including the ability to statistically compare models with different numbers of states. It will also enable us to identify the most predictive Markov model among a group of poor models that cannot be validated with the usual strategies. Such a model would be inapt for drawing conclusions about the dynamics, but may nonetheless be useful in adaptive sampling methods, which would guide further refinement of the state definitions or the model parameters. The thrust of the Bayesian method is discussed further in the concluding section to clarify the meaning of a model's optimality and show how a model could be used to infer *a posteriori* distributions of kinetic properties.

II. BAYESIAN MODEL COMPARISON

We begin by partitioning configurational space into small cells known as *microstates*, Z, using the sampled conformations as a guide. The configurations within any given microstate must be so similar that we can safely assume there are no substantial kinetic barriers between them. Our data will consist of a sequence of microstates $D=\{z_1,z_2,\ldots,z_n:z_i\in Z\}$, sampled from a MD trajectory at a fixed time interval, here referred to as the *lag time*. The microstates are then grouped into larger *macrostates*, Y, which are ideally approximations of the true metastable states of the system.[4] In the chosen lag time, the simulation should quickly equilibrate within the macrostate to produce an uncorrelated sample from the equilibrium distribution of microstates belonging to that macrostate. Suppose the model specifies the microstate-to-macrostate mapping or *lumping*, denoted by $L:Z\rightarrow Y$, Markovian transition probabilities between the macrostates T, and the equilibrium populations of the microstates within each macrostate, $\Theta=\{\theta_y:y\in Y\}$, where each θ_y is a vector of probabililties that sums to 1. We denote a model accordingly $M=\{L,T,\Theta\}$. We presume that the observed sequence of microstates was generated from the Markov model thusly: given the system is initially in some macrostate y_t, we choose the next macrostate from the row of the transition matrix T corresponding to macrostate y_t. Then, we choose a microstate from within macrostate y_{t+1} with probability proportional to its stationary population in y_{t+1}, given by $\theta_{y_{t+1}}$.

All the models considered will be built from the same microstate basis, and we will only be concerned with selecting the superior lumping into macrostates, L, so we let Θ and T be random variables to which we assign noninformative prior distributions. In comparing the probability of two macrostate partitionings L_1 and L_2, we marginalize over microstate probabilities Θ and macrostate transition probabilities T, so as to recover the relative probabilities of the macrostate partitionings alone. In particular, we will compute the ratio

$$\frac{P(L_1|D)}{P(L_2|D)}=\frac{P(D|L_1)}{P(D|L_2)}\frac{P(L_1)}{P(L_2)}, \tag{1}$$

where we assume the prior probabilities of the lumpings equal. By the Bayes theorem, Eq. (1) can be written as

$$\frac{P(L_1|D)}{P(L_2|D)}=\frac{\int\int dTd\Theta P(D|L_1,T,\Theta)P(T,\Theta|L_1)}{\int\int dTd\Theta P(D|L_2,T,\Theta)P(T,\Theta|L_2)}. \tag{2}$$

This ratio, known as a Bayes factor, is the main instrument of Bayesian model comparison; it expresses the probability of a model over another given some finite data, D. Unlike a simpler likelihood ratio, a Bayes factor automatically penalizes overcomplex models. It may be possible to increase the likelihood of the data $P(D|M)$ if we make the model more complex, for example, by increasing the number of macrostates and the dimensionality of T. However, when we integrate this likelihood over a prior distribution on T and Θ, the regions of parameter space of high likelihood are weighted by lower prior probabilities because they represent a smaller fraction of this high-dimensional space.[32,33]

Similar Bayesian frameworks have been used in the past to estimate the transition matrix,[21,34] as well as for model comparison.[35] The choice of prior distribution for T and Θ is critical because it expresses our *a priori* knowledge of these variables, which can affect the behavior of the method when there is little data. We will see that certain priors afford an analytical solution to probabilities of the form $P(D|L)$ in Eq. (1).

Recall that the sequence of microstates D is produced by two independent processes—a Markov chain that generates a

sequence of macrostates, which will be denoted as D_Y, and the selection of a microstate from the stationary distribution θ_y within each macrostate visited—so we can factorize the likelihood inside the integral for $P(D|L)$ to obtain

$$P(D|L) = \int\int dTd\Theta P(D_Y|T,L)P(D|D_Y,\Theta,L)P(T,\Theta)$$
$$= \int dTP(D_Y|T,L)P(T)$$
$$\times \int d\Theta P(D|D_Y,\Theta,L)P(\Theta), \quad (3)$$

where we dropped the trivial dependence of the priors on L. Both the macrostate trajectory D_Y and the selection of microstates given D_Y are sequences of independent multinomial variables parametrized by the rows of T and by Θ, respectively. This makes the Dirichlet distribution a natural prior for the parameters.[21,35] This prior distribution is conjugate to the multinomial likelihood, which means that posterior distributions obtained from the Bayes equation retain the functional form of the prior.[36] Let Z_y be the subset of microstates corresponding to macrostate y, then the Dirichlet prior on θ_y, for example, is defined by the density

$$\mathrm{Dir}(\theta_y;\alpha_y) = \frac{1}{\mathrm{B}(\alpha_y)}\prod_{z\in Z_y}\theta_y(z)^{\alpha_y(z)}, \quad (4)$$

where the normalizing constant $\mathrm{B}(\alpha_y)$ is the multinomial Beta function and α_y is a vector of the same dimension as θ_y that parametrizes the distribution.

The conjugate Dirichlet prior gives the two integrals in Eq. (3) a closed-form solution. Take, for example, the second integral,

$$P(D|D_Y,L) = \int d\Theta P(D|D_Y,L,\Theta)\prod_{y\in Y}\mathrm{Dir}(\theta_y;\alpha_y). \quad (5)$$

This can be analytically solved to yield

$$P(D|D_Y,L) = \prod_{y\in Y}\frac{\Gamma(|Z_y|)\Pi_{z\in Z_y}\Gamma(|\{i:z_i=z\}|+\alpha_y(z))}{\Gamma\left(|\{i:z_i\in Z_y\}|+\Sigma_{z\in Z_y}\alpha_y(z)\right)}, \quad (6)$$

where the parallel bars around a set denote its cardinality and $\Gamma(\cdots)$ is the gamma function. The hyperparameters α_y in the prior for each θ_y may be interpreted as a set of pseudocounts. More precisely, the distribution conveys the belief that the microstate $z\in Z_y$ has been chosen $\alpha_y(z)-1$ times *a priori*. There has been much discussion on which value of the hyperparameter produces an objective or uninformative prior.[37] Two common choices are to set all $\alpha_y(z)$ uniformly to 1 (no prior observations) and to 1/2 (the Jeffreys prior, which preserves the distribution under reparametrization[38,39]). In our examples, we have chosen the former.

We expect physical systems to satisfy detailed balance in microscopic as well as coarse-grained dynamics (see Ref. 28 for example). This property of T has been previously enforced in Bayesian computations by using an independent Dirichlet prior for each row of the matrix, as suggested above, but restricting the joint density $P(T)$ over the reversible matrices.[34] Detailed balance has also been imposed in Bayesian inference of the rate matrix.[21] These methods require Markov chain Monte Carlo (MCMC) sampling.[40] It was observed that imposing reversibility could greatly reduce the uncertainty of off-diagonal elements of the transition matrix in the posterior distribution, which has important consequences in the inference of certain kinetic observables.[34] This is due to the fact that transitions from state i to j give information about the transition rate from j to i, even if none are directly observed. We also expect this restriction to have an effect in model comparison, especially when there is little data, or when the simulations used are far from equilibrium. Here, we employ a conjugate prior for reversible Markov chains.[41] The utility of this distribution, defined strictly over matrices with detailed balance, is that it provides analytical expressions for normalization constants which we will make use of here.

III. A CONJUGATE PRIOR FOR REVERSIBLE MARKOV CHAINS

There is an equivalence between a reversible Markov chain and a random walk on an edge-weighted, undirected graph $\{Y,E\}$.[42] The set of vertices corresponds to the macrostates Y and there is an edge in E for every unordered pair of macrostates. The random walk proceeds as follows: if we start at vertex y_t at time t, the next vertex y_{t+1} is chosen with probability proportional to the non-negative weight, $k_{\{y_t,y_{t+1}\}}$. The set of normalized edge weights, defined by

$$x = \left\{x_{\{i,j\}} = \frac{k_{\{i,j\}}}{\Sigma_{\{h,l\}\in E}k_{\{h,l\}}}:\{i,j\}\in E\right\}, \quad (7)$$

is sufficient to parametrize the Markov chain. The row-stochastic transition probability from macrostate i to j, T_{ij}, is simply given by

$$T_{ij} = \frac{x_{\{i,j\}}}{\Sigma_l x_{\{i,l\}}} = \frac{x_{\{i,j\}}}{x_i}, \quad (8)$$

where x_i, the sum of normalized weights of all the edges adjacent to i, is equivalent to the stationary probability of macrostate i multiplied by 2. In the following, we use x or T interchangeably as parameters of the Markov chain.

The conjugate prior on reversible chains is based on a related stochastic process on graphs known as edge-reinforced random walk (ERW), which proceeds as the above random walk, with the difference that every time we traverse an edge $\{i,j\}$ in either direction, we increase its weight $k_{\{i,j\}}$ by one. So, the initial conditions of an ERW are fully specified by the vertex of origin, y_0, and the initial set of unnormalized edge weights, which will be denoted a. Let $P_{y_0,a}(D_Y)$ be the probability that the ERW traverses a given sequence of macrostates D_Y. By de Finetti's theorem for Markov chains, it was shown that the ERW is a mixture of reversible chains,[41] which means that $P_{y_0,a}(D_Y)$ is the expectation of the probability of D_Y as a random Markov chain, whose parameters x have a well-defined distribution. We can write this as

$$P_{y_0,a}(D_Y) = \int_\Delta d\sigma(x) P(D_Y|x) \phi_{y_0,a}(x), \tag{9}$$

where $P(D_Y|x)$ is the likelihood of a reversible Markov chain parametrized by x and the integral is over the Lebesgue measure $d\sigma(x)$ on the unit simplex Δ in which x lives. The density $\phi_{y_0,a}(x)$ on the parameters of the Markov chain is a function of the initial conditions of the ERW. This density was found to be continuous on the unit simplex and has the closed form[41]

$$\phi_{y_0,a}(x) = Z^{-1}_{y_0,a} \frac{\prod_{\{i,j\}\in E} x_{\{i,j\}}^{a_{\{i,j\}}-(1/2)}}{x_{y_0}^{a_{y_0}/2} \prod_{y\neq y_0} x_y^{(a_y+1)/2}} \sqrt{\det(A(x))}. \tag{10}$$

A combinatorial proof yields a closed form for the function $A(x)$.[41] The constant $Z_{y_0,a}$ is given by

$$Z_{y_0,a} = \frac{\prod_{\{i,j\}\in E} \Gamma(a_{\{i,j\}})}{\Gamma\left(\frac{a_{y_0}}{2}\right) \prod_{y\neq y_0} \Gamma\left(\frac{a_y+1}{2}\right)} \frac{(|E|-1)!\,\pi^{n-1/2}}{2^{1-n+\Sigma_{\{i,j\}\in E} k_{\{i,j\}}}}. \tag{11}$$

It is worth noting that all these results have been generalized to graphs with *loops*, which are edges that connect every vertex to itself, to account for the self-transitions expected in physical systems.[41]

In model comparison, we are interested in computing the two integrals presented in Eq. (3). We have seen that both have closed-form solutions when we choose Dirichlet priors, but a Dirichlet prior for T is defined over all stochastic matrices, neglecting our knowledge of detailed balance. The first integral in Eq. (3), for $P(D_Y|L)$, is greatly simplified if we choose the prior $P(T)$ to have the form of $\phi_{y_0,a}(x)$. By Eq. (9), the integral becomes the probability $P_{y_0,a}(\cdots)$ of an ERW, with initial conditions y_0 and a, and the same path as D_Y. This probability has a very simple form given by the reinforcement scheme of the ERW. However it is not necessary to trace the walk and compute the probability of each transition; the measure $P_{y_0,a}(\cdots)$ is just a function of the transition count matrix, which is written in Eq. 4.13 in Ref. 41. Computing the first integral in Eq. (3) given a single sequence D_Y becomes as straightforward as computing the probability of an ERW through D_Y.

This prior has the further advantage that it is conjugate for the reversible Markov chain. If we assign a prior density of $\phi_{y_0,a}(x)$ to the parameters of a reversible chain, and in an experiment we observe a sequence D_Y starting at y_0, the posterior distribution of x would be given by $\phi_{y_f,b}(x)$, where y_f is the final vertex this sequence visits and b is the set of unnormalized edge weights resulting from an ERW with the same path as D_Y. Due to the conjugacy of the prior, the initial edge weights a can be thought of as pseudocounts, fulfilling the same role as the hyperparameters of the Dirichlet prior used before. So, to make this prior noninformative, we could set the edge weights uniformly to 1, for example. The choice of hyperparameters is discussed further in Appendix B.

These derivations have taken D_Y to be a single macrostate sequence, but we run into difficulties when the data are composed of many independent sequences $\{D_1, D_2, \ldots, D_m\}$ starting from given initial states, which is the case for most MD simulations. We outline a method to compute $P(D_Y|L)$ in this case, which is developed in detail in Appendix A. This integral may not be split into factors for each sequence, but we can factorize the likelihood inside the integral to obtain

$$P(D_Y|L) = \int_\Delta d\sigma(x) \prod_{D_i\in D_Y} P(D_i|L,x) \phi_{y_0,a}(x). \tag{12}$$

Then, we manipulate the integrand, taking advantage of the conjugate prior, to bring out factors which are just ERW probabilities for each sequence D_i. We are left with an integral of density ratios, which can be further simplified by taking out factors that do not depend on x, yielding an expression of the form,

$$P(D_Y|L) = \mathcal{G}(D_Y) \int_\Delta d\sigma(x) W(x;D_Y) \phi_{y_f,a'}(x), \tag{13}$$

where $\mathcal{G}$ is a closed-form function of the data, $W(x;D_Y)$ is a product of vertex weights dependent only on the end states of the observed sequences, and $\phi_{y_f,a'}(x)$ is a pseudoposterior density. The last term is approximated by Monte Carlo integration, which makes use of the ERW as a sampling scheme for the density $\phi_{y_f,a'}(x)$.

IV. IMPLEMENTATION

In this section we summarize the procedure of Bayesian model comparison for MD in practice. We begin with a decomposition of configurational space into microstates, to which a number of MD trajectories are projected, as well as several different lumpings. If the lag time we are using is longer than the interval between conformations in the MD trajectories, we must first extract a data set at the correct lag time. For each trajectory, we choose an initial conformation z_t at a random time $0 \leq t \leq \tau_{\text{lag}}$ and take the sequence $D = (z_t, z_{t+\tau_{\text{lag}}}, z_{t+2\tau_{\text{lag}}}, \ldots)$.

Once a data set D has been chosen, we compute the evidence $P(D|L)$ for each of the lumpings. We use independent Dirichlet priors for the parameters $\theta_i \in \Theta$, such that the second integral in Eq. (3) for $P(D|D_Y,L)$ can be obtained analytically from Eq. (6). The first integral in Eq. (3), for $P(D_Y|L)$, is computed using either an independent Dirichlet prior for each row of the transition matrix, which yields an analytical solution, or the conjugate prior for reversible Markov chains introduced in Sec. III. Obtaining $P(D_Y|L)$ with the latter prior requires a Monte Carlo integration. The most expensive part of this computation is sampling from the asymptotic distribution of the ERW. Using the algorithm proposed in Appendix A, this integration scales as $O(N)$, where N is the number of macrostates, and it is highly parallelizable. In the examples of the following section, this computation took only a few minutes.

Note that the evidence $P(D|L)$ is a function of the data set, which in turn depends on the initial conformations used for sampling microstate sequences at the desired lag time from the MD trajectories. To control for the variation that might arise from this, we perform the computation described

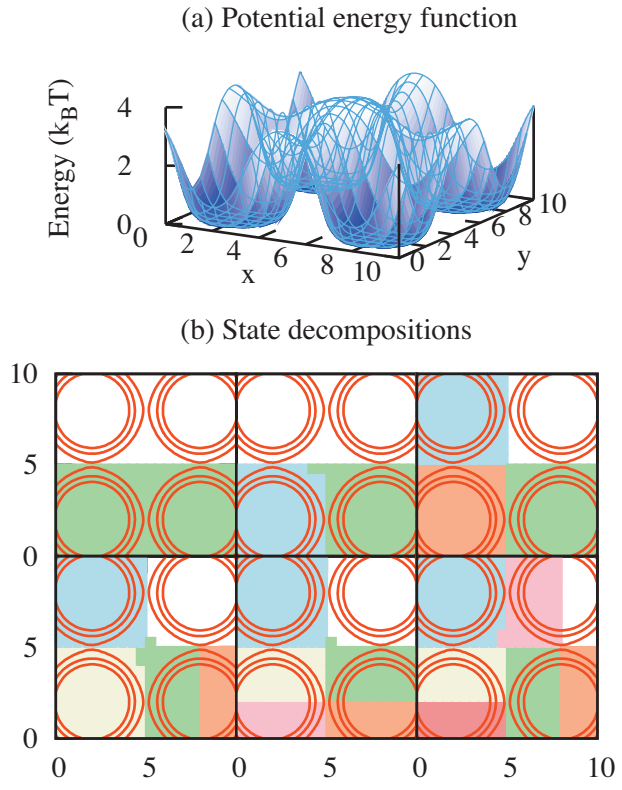

FIG. 1. (a) Artificial energy surface simulated by Metropolis–Hastings Monte Carlo. (b) Space partitions of maximal metastability for 2 through 7 macrostates; each state is highlighted in a different color on the *xy* plane over a contour plot of the energy.

above for 200 different data sets, which are extracted from the trajectories by randomly selecting the initial conformations between time 0 and τ_{lag}. All of the results in the following section are reported with a confidence interval that reflects this variation.

We can compute a Bayes factor to compare any two models by taking the ratio of their evidence. The models are defined by the lumping L and the type of prior distribution assigned to T. A Bayes factor has a meaningful scale related to betting odds; it tells one how many times more likely one model is over the other, given the data. Harold Jeffreys proposed some guidelines for its interpretation, which are summarized in Ref. 33. In short, a Bayes factor greater than 100 gives decisive evidence in favor of the model in the numerator of Eq. (1).

V. APPLICATIONS

A. Metropolis–Hastings Monte Carlo simulation of a model two-dimensional potential

The energy function, plotted in Fig. 1(a), is given by

$$H(q) = \frac{4}{1 + \sum_{o \in C} \frac{1000}{d(o,q)^8}}, \tag{14}$$

where $d(o,q)$ is the Euclidean distance between two points o and q, and C is the set of points $\{(2,2),(2,8),(8,2),(8,8)\}$. This potential is simulated by Metropolis–Hastings Monte Carlo,[43] with an asymmetric proposal kernel that is uniform on the intersection of a square $\sigma(q_0)$ of side 0.5 centered at the current position q_0 with the state space Ω extending from 0 to 10 in the x and y axes. So, a transition from q_0 to q_f is accepted with probability

$$\min\left(1, e^{(-H(q_f)+H(q_0))} \frac{1/\text{Area}(\sigma(q_f) \cap \Omega)}{1/\text{Area}(\sigma(q_0) \cap \Omega)}\right).$$

The small step size ensures that the simulation behaves like a form of stochastic dynamics on the potential. A total of 4000 trajectories of 1000 steps were generated, with an equal number starting at the minima of each of the four wells in the energy surface. We define 400 microstates by dividing the entire area with a 20 by 20 uniform grid.

To partition the set of microstates into metastable macrostates, we used a spectral clustering method known as Perron cluster cluster analysis (PCCA).[44,45] More specifically, our method is based on the lumping step of the iterative algorithm proposed by Chodera *et al.*[20] We first apply PCCA, using as input a transition matrix inferred from the trajectories with the smallest possible lag time, then, we optimize the partition in 20 rounds of simulated annealing aimed at maximizing the *metastability*, which is measured as the sum of the macrostates' self-transition probabilities. The implementation of this algorithm by Bowman *et al.*[46] was used to produce metastable partitions with different numbers of macrostates; these are shown in Fig. 1(b).

The eigenvalue spectrum $\{\lambda_i\}$ of the transition probability matrix taken at a lag time of 20 steps and its implied timescales $\tau_i = -\tau_{\text{lag}}/\log(\lambda_i)$, provide evidence that a dynamics with four states would be Markovian. At this lag time, there are three time scales greater than 20, $\{60.5 \pm 0.5, 58.8 \pm 0.8, 31.1 \pm 0.3\}$, which are stable with increasing lag time, while the rest are all smaller than 6.3. For each of the six models in Fig. 1(b), $P(D|L)$ was computed by the procedure described in Sec. IV.

The results of the model comparison are shown in Table I. As expected, the partition with four states had the highest probability given the data and the table contains Bayes factors that compare every other model to the four-state model. The Bayes factors obtained with both of the priors for T lead to similar conclusions in this case. The results indicate that

TABLE I. Dynamics of 2D model potential with $\tau_{\text{lag}}=20$. Logarithm of the Bayes factors comparing the four-state model, L_4, to all the other models, denoted L_N, where N is the number of macrostates. The columns reflect the choice of prior for T, which is either the conjugate prior on reversible chains suggested here, or the symmetric Dirichlet prior used previously (Refs. 35 and 21). The first number in each cell is the mean value of the logarithmic Bayes factor from 200 data sets extracted from the MD trajectories as explained in Sec. IV and the range between parentheses is its 68% confidence interval.

	$\ln(P(L_4\|D)/P(L_N\|D))$	
N	Reversible Markov prior	General Markov prior
2	5570 (5541, 5598)	5565 (5536, 5594)
3	3128 (3099, 3156)	3121 (3091, 3147)
5	13.2 (3.9, 22.5)	19.4 (11.2, 27.9)
6	18.4 (11.4, 25.5)	33.8 (27.0, 39.8)
7	42.5 (32.5, 52.6)	66.8 (57.2, 75.9)

lumping two kinetically resolved macrostates is penalized much more heavily than splitting a macrostate in two. This might be because a model with more macrostates than necessary can fit the data well, but it represents only a slight increase in complexity with respect to the four-state model.

B. MD simulation of the terminally blocked alanine peptide

The second example is an MD simulation of the peptide Ace-Ala-Nme in explicit solvent. The data consist of 975 trajectories from the 400 K replica of a parallel tempering simulation. They come from an equilibrium pool of constant-energy, constant-volume trajectories 20 ps long, and sampled at a period of 0.1 ps. The data set is thoroughly described in Ref. 47. This peptide has been used previously to test state decomposition algorithms,[20] as well as Bayesian model comparison,[35] because it is possible to identify metastable states directly from a potential of mean force in torsional space.

The microstates were defined using a k-centers algorithm that approximates average-linkage hierarchical clustering.[48] Using the heavy-atom root mean squared deviation (RMSD) as a distance metric, the algorithm produced 3900 microstates with an average diameter of 0.14 Å, where the diameter is the maximum distance between any two structures in a cluster. The microstates were then clustered automatically into six macrostates by the algorithm described in the previous example, to form L_{good}. A poor six-macrostate decomposition, L_{poor}, was generated by clustering the microstates to match the manual decomposition defined in Ref. 20.

Note that the trajectories are relatively short, so we must choose a lag time smaller than 10 ps if we want to have sequences of more than two steps and control for the initial frame noise alluded to above. However, we do not expect either of the models to be Markovian at lag times smaller than 8 ps, because the solvent degrees of freedom, neglected in our state definitions, relax in comparable time scales.[20,49] We compared the two models at seven different lag times, using both of the priors that were assigned to T in the previous example.

The results are shown in Table II. They suggest that L_{good} is better than L_{poor} at every lag time. Even though the good model is only approximately Markovian at these lag times, the Bayes factors can help choose the superior state decomposition. As in the previous example, the comparisons between L_{good} and L_{poor} are similar when we use either prior for T. However, Bayes factors can also be used to compare the models where detailed balance is imposed in the prior to those where it is not. For the good state decomposition of this system, the prior on reversible Markov chains is strongly preferred at every lag time. The natural logarithm of this Bayes factor is in every case greater than 11 within 68% confidence.

TABLE II. MD of the terminally blocked alanine peptide. Logarithm of the Bayes factor comparing L_{good} to L_{poor} at different lag times. As in Table I, the columns reflect the choice of prior for T. The first number in each cell is the mean value of the logarithmic Bayes factor from 200 data sets extracted from the MD trajectories as explained in Sec. IV and the range between parentheses is its 68% confidence interval.

	$\ln(P(L_{\text{good}}\|D)/P(L_{\text{poor}}\|D))$	
τ_{lag} (ps)	Reversible Markov prior	General Markov prior
3	1626 (1571, 1671)	1603 (1561, 1639)
4	1032 (981, 1088)	1014 (978, 1048)
5	701 (654, 752)	691 (660, 726)
6	496 (457, 540)	489 (457, 522)
7	367 (319, 424)	360 (328, 397)
8	283 (230, 336)	274 (238, 308)
9	216 (161, 269)	208 (170, 243)

VI. DISCUSSION AND CONCLUSIONS

Statistical hypothesis testing quantifies the evidence for different models of conformational dynamics allowing us to choose the optimal or most predictive one. In a Bayesian framework, models specify a distribution over the data. They must be defined with complete ignorance of the sampled data, so a hierarchical structure of model parameters is often useful. Then, a Bayes factor measures the relative evidence of two models at every value of their parameters. In our case, the hypotheses tested are the metastable state decomposition, L, as well as the detailed balance implicit in the prior of T. In a frequentist framework, one would device a statistic to test a hypothesis, which in this case need not specify the distribution of the data completely. The test would consist of determining how typical the sample statistic is under said hypothesis.

These methods, Bayesian or frequentist, do not establish bounds on the error due to assuming the optimal Markov model, which might still be overly parsimonious. So validating the Markovian assumption by heuristic or other means is necessary to assess predictions. The advantage of a Bayesian framework is that the uncertainty of model parameters and physical observables associated with them, *assuming* the model, emerge naturally from their posterior distributions. The process of deriving these distributions, known as *inference*, is perhaps more widely applied and has a similar mathematical structure as model comparison. Appendix D hints at how to apply the prior on reversible matrices to inference problems.

Bayes factors hold several advantages over previous approaches to Markov model validation. One distinction, which was made clear by the examples, is that it is easy to compare models with different numbers of macrostates provided the microstate basis is the same for both models. A perhaps more important aspect of the Bayesian approach, which sets it apart from other methods that depend on estimates of the transition matrix, is that one is able to compare models when there is little data available. Even when the estimates have not converged, a Bayes factor will indicate the relative probabilities of two models. On the other hand, if none of the available models are able to reproduce the dynamics well at the desired lag time, one would be able to identify the best model in a set of poor models.

These two properties make a Bayesian approach useful in adaptive sampling schemes, in which the simulation strat-

egy is constantly updated based on short trajectories given fixed state definitions (for an example, see Ref. 50). The comparison method presented here could be used to identify the most predictive Markov model at a short lag time, even if the dynamics at this time resolution are only loosely Markovian. The Markov model could nonetheless be able to identify the regions of conformational space where simulation is most needed. As suggested by Singhal,[50] the ability to extract a posterior distribution of T from the model would facilitate this task, allowing us to perturb eigenvectors, eigenvalues, and other functions of T to predict which states contribute most to their uncertainties. It would also be possible to apply more expensive information theoretic strategies to this problem.

In a more direct application, Bayes factors could be integrated into a state-partitioning algorithm. For example, the algorithm proposed by Chodera *et al.*[20] could be modified by substituting the metastability measure by the weight of evidence of a partition, $P(D|L)$. This would allow one to vary the number of macrostates, N, in the lumping step, where the microstate-to-macrostate mapping is optimized. The computation of the evidence using the prior that enforces detailed balance scales with system size as $O(N)$, its most expensive step being the Monte Carlo integration.

In conclusion, we put forward Bayesian model comparison as a method to test hypotheses on MD. We introduced a hierarchical model for metastable dynamics on an arbitrary number of states, which enforces detailed balance. The prior distribution of Diaconis and Rolles[41] is shown to be analytically advantageous. The structure of the method is quite general, in that any hypothesis expressed as a generative model for microstate sequences may be tested. This allows for the possibility of extending our current model, or formulating altogether new ones to incorporate different physical insights about simulations of one or multiple ensembles.

ACKNOWLEDGMENTS

We would like to thank Persi Diaconis, Gregory Bowman, Xuhui Huang, Wai Wai Liu, Jian Sun, and Yuan Yao for stimulating insights, and an anonymous referee for a helpful review. S.B. was supported by a Stanford Graduate Fellowship and J.D.C. acknowledges support from a QB3-Berkeley Distinguished Postdoctoral Fellowship. This work was funded by NIH Grant No. R01-GM062868, as well as NSF Award No. CNS-0619926 for computer resources.

APPENDIX A: INTEGRATING THE LIKELIHOOD OF MULTITRAJECTORY DATA

In the following, we develop a method for estimating the probability of a set D_Y of independent macrostate sequences, given the microstate-to-macrostate mapping L. We will assume the data consist of m independent sequences, $D_Y=\{D_1,D_2,\ldots,D_m\}$, with known initial states.

We will use the following notation: the first and last states of sequence D_i are, respectively, $y_{0,i}$ and $y_{f,i}$, and its edge traversal counts will be denoted by k_i. The likelihood of a sequence D_i, given a fixed set of parameters x for the macrostate Markov chain, will be written $P(D_i|x,L)$. The subscripted probability $P_{y,k}(D_i|L)$ denotes the integrated likelihood of D_i, with a prior density $\phi_{y,k}(x)$ on the parameters, x.

For notational convenience, we will let the prior density of x be $\phi_{y_{f,0},k_0}(x)$, with an initial vector of unnormalized edge weights k_0, and a vertex of origin $y_{f,0}$. The probability of D_Y under this prior is

$$P_{y_{f,0},k_0}(D_Y|L)=\int_\Delta d\sigma(x)P(D_Y|x,L)\phi_{y_{f,0},k_0}(x)$$
$$=\int_\Delta d\sigma(x)\prod_{i\geq 1}P(D_i|x)\phi_{y_{f,0},k_0}(x), \tag{A1}$$

where we used the conditional independence of the chains given x to factorize the probability inside the integral. Now, we can multiply the integrand by a factor of 1,

$$P_{y_{f,0},k_0}(D_Y|L)=\int_\Delta d\sigma(x)\prod_{i\geq 1}P(D_i|x,L)\frac{\phi_{y_{0,1},k_0}(x)}{\phi_{y_{0,1},k_0}(x)}\phi_{y_{f,0},k_0}(x). \tag{A2}$$

We can extract the factor $P(D_1|x,L)\phi_{y_{0,1},k_0}(x)$ from the integrand, which is just the joint density of D_1 and x, with D_1 fixed. Using Bayes theorem,

$$P(D|x)P(x)=P(x|D)P(D), \tag{A3}$$

we can rewrite this term as $\phi_{y_{f,1},k_0+k_1}(x)P_{y_{0,1},k_0}(D_1|L)$, where the new density is the posterior of x given D_1. We have taken advantage of the conjugacy of the prior ϕ for reversible Markov chains, as well as the closed-form expression for $P_{y_{0,1},k_0}(D_1|L)$, which is given by Eq. 4.13 in Ref. 41. We are left with

$$P_{y_{f,0},k_0}(D_Y|L)=P_{y_{0,1},k_0}(D_1|L)\int_\Delta d\sigma(x)$$
$$\times\frac{\phi_{y_{f,0},k_0}(x)}{\phi_{y_{0,1},k_0}(x)}\prod_{i>1}P(D_i|x,L)\phi_{y_{f,1},k_0+k_1}(x). \tag{A4}$$

We can repeat the last step for all other sequences. For simplicity, let us define $K_i=\Sigma_{j<i}k_j$. We obtain,

$$P_{y_{f,0},k_0}(D_Y|L)=\prod_{i\geq 1}P_{y_{0,i},K_i}(D_i|L)\times\mathcal{I}, \tag{A5}$$

$$\mathcal{I}=\int_\Delta d\sigma(x)\prod_{i\geq 1}\frac{\phi_{y_{f,i-1},K_i}(x)}{\phi_{y_{0,i},K_i}(x)}\phi_{y_{f,m},K_{m+1}}(x). \tag{A6}$$

Finally, we can simplify the product inside the integral, taking advantage of the closed form for the density [Eq. (10)], to get

$$\mathcal{I} = \prod_{i\geq 1}\left[\frac{\Gamma\left(\frac{K_i(y_{f,i-1})}{2}\right)\Gamma\left(\frac{K_i(y_{0,i})+1}{2}\right)}{\Gamma\left(\frac{K_i(y_{0,i})}{2}\right)\Gamma\left(\frac{K_i(y_{f,i-1})+1}{2}\right)}\right] \times \int_\Delta d\sigma(x)\prod_{i\geq 1} w_i(x)\phi_{y_{f,m},K_{m+1}}(x), \quad \text{(A7)}$$

where we define

$$w_i(x) = \left[\frac{x_{y_{f,i-1}}}{x_{y_{0,i}}}\right]^{1/2}.$$

Everything outside the integral has a simple analytical form and it is the result we would get if we took the data to be the concatenation of all the D_i. The quantity in brackets inside the integral takes care of the difference that results from considering each sequence separate and independent. The integral can be estimated by Monte Carlo integration.

There is a simple algorithm to obtain approximate, independent samples from the density $\phi_{y_{f,m},K_{m+1}}(x)$. In an ERW starting at $y_{f,m}$ with initial edge weights K_{m+1}, let $\{k_{\{i,j\}}(n)\}$ be the edge-traversal counts after n steps. Consider the statistic $\kappa(n)=\{k_{\{i,j\}}(n)/\Sigma_{\{h,l\}\in E}k_{\{h,l\}}(n)\}$. It was also shown in Ref. 41 that as n goes to infinity, $\kappa(n)$ converges in probability to a random vector κ with density $\phi_{y_{f,m},K_{m+1}}(\kappa)$. So, if we simulate this ERW for long enough, we would expect the normalized edge weights to converge to an independent sample of this density. We can draw a set of samples X in this way to estimate the integral in question by a Monte Carlo sample average:

$$\int_\Delta d\sigma(x)\prod_{i\geq 1} w_i(x)\phi_{y_{f,m},K_{m+1}}(x) \approx \frac{1}{|X|}\sum_{x\in X}\prod_{i\geq 1} w_i(x). \quad \text{(A8)}$$

In the examples of the text, we estimated every integral from 300 samples generated in this fashion. One advantage of this scheme, compared to MCMC, is that every sample obtained is truly independent. However, it is difficult to know how long we must simulate the ERW to obtain sufficient precision in the elements of each sample to ensure the dominant error in our estimate of this integral is from the statistical uncertainty in the above expression. In our computations, for a system with N macrostates, we ran the ERW for $1000\times N$ steps. We expect that after this, the normalized vertex weights $\{\kappa_i(n)\}$ needed to compute the sum will not deviate greatly from the realization of the random vector obtained as n goes to infinity. The effect in precision of having approximate samples can be determined if we assume that any error resulting from this will not be systematic. If we make this assumption, we can test the accuracy of the method by looking at the convergence of the Monte Carlo expectation. This convergence was checked by bootstrapping.[51] The error of the numerical method was significantly smaller than the variation resulting from the choice of an initial reading frame for the trajectories, which is why it was not given in the results.

APPENDIX B: CHOICE OF HYPERPARAMETERS FOR THE PRIOR

The prior distribution has most influence when performing inference with little data, while it should not matter when there is enough data. We would like to choose an objective or uninformative prior such that conclusions are based solely on information from the simulation. We suggested that one make all the initial, unnormalized edge weights defined in the prior, k_0, equal to 1. This could be interpreted as giving no preference to any transition (including self-transitions) over another. We made this choice in our examples. There may be more objective priors, such as Jeffreys prior, which is proportional to the square root of the Fisher information. However, their derivation was not pursued in this study.

We must also choose a vertex of origin for the prior, which was denoted by $y_{f,0}$. When we choose any single vertex, we lose the notion of symmetry in the prior. We could also use a uniform mixture of priors of the form $P(x)=\Sigma_{y\in Y}N^{-1}\phi_{y,k_0}(x)$. Note that the distribution $\phi_{y_{f,m},K_{m+1}}(x)$ we must sample from in order to estimate the integral in $\mathcal{I}$ does not depend on the vertex $y_{f,0}$ defined in the prior. Thus, from a large sample of this distribution, we can estimate the integral regardless of the vertex of origin we chose for the prior. This makes it easy to employ the uniform mixture of priors shown above and this is what was done in the examples.

APPENDIX C: SPEEDING UP THE ERW

We found a way to speed up the simulation of an ERW that is worth explaining here. In a physical dynamical system with metastability, we expect macrostates to have large self-transition probabilities. So, an ERW with initial edge weights defined by K_{m+1} (the added edge traversals observed in all the sequences D_i) will likely spend a lot of time traversing loops. We can avoid simulating this by showing that the self-transition probabilities have a Beta distribution, because of an equivalence between the edge-reinforced walk and the well-known Polya-urn process.

Consider any vertex, y. We will denote the weight of the adjacent loop, or the edge for self-transitions, $k_l(n)$, and the summed weights of the edges that connect y to every other vertex $k_r(n)$, both indexed by the number of steps the ERW has taken. We want to study the asymptotic behavior of the self-transition probability $p_n=k_l(n)/(k_l(n)+k_r(n))$. Suppose we begin an ERW at y, we will choose to traverse the adjacent loop with probability p_0, in which case we increase k_l by 2; alternatively, we will go to a different vertex and return to y in a finite number of steps almost surely,[41] in which case we increase k_r by 2. This is equivalent to the Polya-urn scheme, in which we have an urn with some number of red and blue balls, analogous to k_l and k_r, and at each step we take a ball at random from the urn and put it back in along with two balls of the same color.

It is known that in a Polya-urn process, p_n converges almost surely to a random variable p distributed as

$$p \sim \text{Beta}\left(\frac{k_l(0)}{2}, \frac{k_r(0)}{2}\right).$$

If we wanted to make the simulation of the ERW faster, we could first sample an ERW on the graph without loops. Then, for each vertex y in the graph take a sample of p from the corresponding Beta distribution and from this value of p and the final value of k_r for the vertex, generate a sample of the final loop weight k_l. In addition, the Beta distribution of p gives us a graphical picture of the marginal prior we assign to the self-transition probabilities, based on which we could adjust its hyperparameters. This procedure was not followed here.

APPENDIX D: BAYESIAN INFERENCE PROBLEMS

The numerical techniques developed here for model comparison may also be applied to Bayesian inference problems. Suppose we want to know the posterior average of some function of the transition matrix $A(T)$, given a set of MD trajectories sampled at a fixed lag time. We would like to compute the integral

$$\mathbb{E}[A(T)|D,L] = \int\int dTd\Theta A(T)P(T,\Theta|D,L), \tag{D1}$$

where we are keeping the macrostate definition L fixed. Applying Bayes equation to the probability inside the integral, we get

$$\mathbb{E}[A(T)|D,L] = \frac{\int\int dTd\Theta A(T)P(D|T,\Theta,L)P(T,\Theta)}{\int\int dTd\Theta P(D|T,\Theta,L)P(T,\Theta)}. \tag{D2}$$

We can factorize the integrals in the numerator and denominator as we did in Eq. (3) in the main text. We obtain the same integral over $d\Theta$ above and below, which cancels out leaving,

$$\mathbb{E}[A(T)|D,L] = \frac{\int dTA(T)P(D_Y|T,L)P(T)}{\int dTP(D_Y|T,L)P(T)}. \tag{D3}$$

It is possible to marginalize out T in the denominator, as we show in Eq. (A1), and the integral in the numerator can be manipulated in the same way. We can go through the steps followed in that derivation to obtain a series of prefactors with closed-form expressions, which will be the same in the numerator and denominator. Using the notation of Eq, (A1), we are left with the ratio of integrals

$$\mathbb{E}[A(T)|D,L] = \frac{\int_\Delta d\sigma(x)A(x)\Pi_{i\geq 1}w_i(x)\phi_{y_{f,m},K_{m+1}}(x)}{\int_\Delta d\sigma(x)\Pi_{i\geq 1}w_i(x)\phi_{y_{f,m},K_{m+1}}(x)}. \tag{D4}$$

Each integral in this ratio may be approximated by a Monte Carlo integration procedure, where we draw many samples,

$$x' \sim \phi_{y_{f,m},K_{m+1}}(x),$$

and take the sample mean of $A(x')\Pi_{i\geq 1}w_i(x')$ and $\Pi_{i\geq 1}w_i(x')$, respectively. The rate of convergence of this algorithm will depend on the variance of the integrands under this measure.

[1] O. Becker and M. Karplus, J. Chem. Phys. **106**, 1495 (1997).
[2] Y. Levy, J. Jortner, and O. Becker, J. Chem. Phys. **115**, 10533 (2001).
[3] Y. Levy, J. Jortner, and R. Berry, Phys. Chem. Chem. Phys. **4**, 5052 (2002).
[4] C. Schuette and W. Huisinga, *Biomolecular Conformations as Metastable Sets of Markov Chains*, in Proceedings of the 38th Annual Allerton Conference on Communication, Control, and Computing, Monticello, Illinois, 2000, Vol. 2, p. 1106.
[5] R. Zwanzig, *Lectures in Theoretical Physics (Boulder)* (Wiley, New York, 1961), Vol. 3.
[6] H. Mori, Prog. Theor. Phys. **33**, 423 (1965).
[7] R. Zwanzig, J. Stat. Phys. **30**, 255 (1983).
[8] G. Ben Arous, A. Bovier, and V. Gayrard, Commun. Math. Phys. **235**, 379 (2003).
[9] G. Ben Arous, A. Bovier, and J. Cerný, J. Stat. Mech. **4**, L04003 (2008).
[10] D. Chandler, J. Chem. Phys. **68**, 2959 (1978).
[11] A. F. Voter and J. D. Doll, J. Chem. Phys. **82**, 80 (1985).
[12] J. Adams and J. Doll, Surf. Sci. **111**, 492 (1981).
[13] A. Faradjian and R. Elber, J. Chem. Phys. **120**, 10880 (2004).
[14] D. Moroni, T. van Erp, and P. Bolhuis, Physica A **340**, 395 (2004).
[15] A. Berezhkovskii and A. Szabo, J. Chem. Phys. **122**, 014503 (2005).
[16] Y. Rhee and V. Pande, J. Phys. Chem. B **109**, 6780 (2005).
[17] S. Northrup, M. R. Pear, C. Y. Lee, J. A. McCammon, and M. Karplus, Proc. Natl. Acad. Sci. U.S.A. **79**, 4035 (1982).
[18] C. Dellago, P. Bolhuis, and P. Geissler, Adv. Chem. Phys. **123**, 1 (2002).
[19] J. Rogal and P. G. Bolhuis, J. Chem. Phys. **129**, 224107 (2008).
[20] J. Chodera, N. Singhal, V. Pande, K. Dill, and W. Swope, J. Chem. Phys. **126**, 155101 (2007).
[21] S. Sriraman, I. G. Kevrekidis, and G. Hummer, J. Phys. Chem. B **109**, 6479 (2005).
[22] B. de Groot, X. Daura, A. Mark, and H. Grubmüller, J. Mol. Biol. **309**, 299 (2001).
[23] V. Schultheis, T. Hirschberger, H. Carstens, and P. Tavan, J. Chem. Theory Comput. **1**, 515 (2005).
[24] S. Elmer, S. Park, and V. Pande, J. Chem. Phys. **123**, 114902 (2005).
[25] H. Grubmüller and P. Tavan, J. Chem. Phys. **101**, 5047 (1994).
[26] W. Zheng, M. Andrec, E. Gallicchio, and R. M. Levy, Proc. Natl. Acad. Sci. U.S.A. **104**, 15340 (2007).
[27] N. Singhal, C. Snow, and V. Pande, J. Chem. Phys. **121**, 415 (2004).
[28] N. v. Kampen, *Stochastic Processes in Physics and Chemistry* (North-Holland, Amsterdam, 1992), Chap. 5.
[29] W. Swope, J. Pitera, and F. Suits, J. Phys. Chem. B **108**, 6571 (2004).
[30] N. Buchete and G. Hummer, J. Phys. Chem. B **112**, 6057 (2008).
[31] S. Park and V. S. Pande, J. Chem. Phys. **124**, 054118 (2006).
[32] D. MacKay, *Information Theory, Inference, and Learning Algorithms* (Cambridge University Press, Cambridge, 2003), Chap. 28.
[33] R. Kass and A. Raftery, J. Am. Stat. Assoc. **90**, 773 (1995).
[34] F. Noé, J. Chem. Phys. **128**, 244103 (2008).
[35] N. Singhal, Ph.D. thesis, Stanford University, 2007.
[36] A. Gelman, J. Carlin, H. Stern, and D. Rubin, *Bayesian Data Analysis* (Chapman and Hall, London, 1995).
[37] I. J. Good, *The Estimation of Probabilities: An Essay on Modern Bayesian Methods* (MIT Press, Cambridge, 1965).
[38] H. Jeffreys, Proc. R. Soc. London **186**, 453 (1946).
[39] P. Goyal, AIP Conf. Proc. **803**, 366 (2005).
[40] J. Liu, *Monte Carlo Strategies in Scientific Computing* (Springer-Verlag, New York, 2001).
[41] P. Diaconis and S. Rolles, Ann. Stat. **34**, 1270 (2006).
[42] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. (Wiley-Interscience, Hoboken, 2006), Chap. 4.3.
[43] W. K. Hastings, Biometrika **57**, 97 (1970).
[44] P. Deuflhard, W. Huisinga, A. Fischer, and C. Schuette, Numer. Linear Algebra Appl. **315**, 39 (2000).
[45] M. Weber, Konrad-Zuse-Zentrum fur Informationstechnik Berlin, Takustrasse 7, Technical Report No. D-14195, 2004.

[46] G. Bowman, X. Huang, and V. Pande, "Using generalized ensemble simulations and Markov state models to identify conformational states," Methods (in press).
[47] J. D. Chodera, W. C. Swope, J. W. Pitera, and K. A. Dill, Multiscale Model. Simul. **5**, 1214 (2006).
[48] Y. Yao, J. Sun, and X. Huang, "A fast geometric clustering method in the conformational space of biomolecules" (unpublished).
[49] A. Ma, N. Ambarish, and A. Dinner, J. Chem. Phys. **124**, 144911 (2006).
[50] N. Singhal and V. Pande, J. Chem. Phys. **123**, 204909 (2005).
[51] B. Efron, Ann. Stat. **7**, 1 (1979).