Addressing the Embeddability Problem in Transition Rate Estimation

Curtis Goolsby[1] and Mahmoud Moradi[1, a)]

Department of Chemistry and Biochemistry, University of Arkansas, Fayetteville, AR 72701, U.S.A.

(Dated: 4 August 2020)

Markov State Models (MSM) and related techniques have gained significant traction as a tool for analyzing and guiding molecular dynamics (MD) simulations due to their ability to extract structural, thermodynamic, and kinetic information on proteins using computationally feasible MD simulations. The MSM analysis often relies on spectral decomposition of empirically generated transition matrices. Here, we discuss an alternative approach for extracting the thermodynamic and kinetic information from the so-called rate/generator matrix rather than the transition matrix. Although the rate matrix is itself built from the empirical transition matrix, it provides an alternative approach for estimating both thermodynamic and kinetic quantities, particularly in diffusive processes. We particularly discuss a fundamental issue with this approach, known as the embeddability problem and offer ways to address this issue. We describe six different methods to overcome the embeddability problem. We use a one-dimensional toy model to show the workings of these methods and discuss the robustness of each method in terms of its dependence in lag time and trajectory length.

[a)]Electronic mail: moradi@uark.edu

I. INTRODUCTION

Proteins and other biological macromolecules are associated with complex conformational spaces that are virtually impossible to be fully characterized at the atomic level using current experimental and computational tools [1]. With increasing computational capabilities due to hardware improvements, all-atom molecular dynamics simulations, are used to explore the possibility of investigating the important regions of free energy landscapes of proteins and other biomolecules [2–4]. The tools provided by computational methods are limited by their computational costs. With this limitation, it is of significant interest to employ statistical techniques, which allow the inference of relevant thermodynamic and kinetic properties from shorter, less costly simulations. Markov State Models or MSMs[5–10] provide some of the most powerful tools for analyzing the ensembles of short molecular dynamics trajectories to extract information on both thermodynamics and kinetics of complex biomolecular systems and predict the behavior of such systems at much longer timescales than possible to simulate with current computing capabilities. These methods, however, are based on assumptions and simplifications that introduce limitations to the reliability and interpretability of these methods[11–13].

Building MSMs involves (1) discretization of conformational space, (2) extracting transition statistics from simulation trajectories, and (3) analyzing the transition statistics to estimate kinetic and thermodynamic properties[14]. Here, we only focus on the third component, which assumes an empirical transition matrix has been generated. Here we discuss a known but less common approach for analyzing empirical transition matrices that with certain remedies could provide an alternative approach to the more common eigendecomposition technique. The latter relies on the eigenvectors and eigenvalues of the empirical transition matrix to estimate thermodynamic and kinetic properties. The approach we will discuss here, however, is based on building a rate matrix, also known as a generator matrix, to estimate the kinetic and thermodynamic quantities using standard methods from chemical kinetics literature (Figure 1). Producing a rate matrix from the empirical transition matrix is known to be associated with the embeddability problem. The embeddability problem is found in taking the matrix logarithm of a time-dependent transition probability matrix in order to determine the generator matrix or the rate matrix, as known in chemical kinetics literature. In theory, the lag-time dependent transition probability matrix should only have positive eigenvalues; however, in practice, insufficient sampling can result in the presence of negative eigenvalues. Thus, its matrix logarithm has non-real values and is an invalid generator matrix.

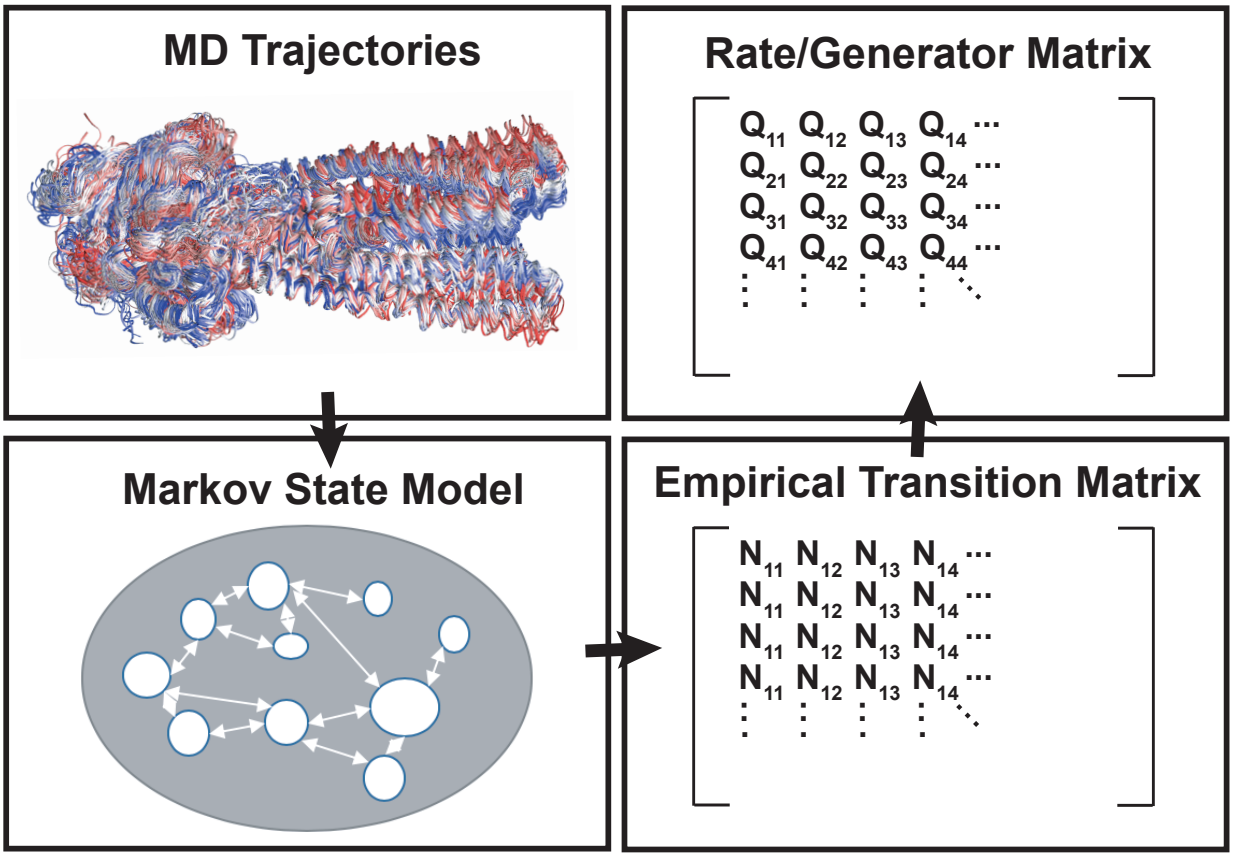

FIG. 1. The schematic representation of the proposed approach to analyze the MSM data. MD trajectories are first used within a clustering scheme to generate a Markov state model and build an empirical transition matrix using a given lag time (see MSM literature[7]). Instead of spectral decomposition of the transition matrix, however, we propose to estimate a lag-time independent rate/generator matrix from the empirical transition matrix obtained from the MD trajectories.

Solving this embeddability problem involves accurately predicting the true generator matrix from an invalid generator matrix. We aim to explore the validity of various algorithms in the literature. We show the workings of these methods using a simple one-dimensional toy model, and compare the performance of proposed algorithms empirically.

II. BACKGROUND

Enhanced sampling[15] has found success in the field of computational chemistry/biophysics as an effective remedy for the costs involved with simulating large molecular systems. Methods such as umbrella sampling (US) [16–19], metadynamics [20–23], and replica exchange [24–28] have grown increasingly popular by aiming to enhance the sampling of configuration space by manipulating the energetics of the system. Enhancing the sampling by biasing the potential or kinetic energy, of

course requires post-analysis reweighting techniques such as weighted histogram analysis method or multistate Bennett acceptance ratio method to determine the thermodynamic and kinetic quantities such as free energy and mean first passage time (MFPT). However, it is reasonable to assume that the reweighting schemes cannot fully remove the inherent bias in the MD data generated using enhanced sampling techniques. As a consequence, an alternative approach to characterize the thermodynamic and kinetic properties of biomolecular systems has been suggested that relies on Markovian analysis of transition probabilities between discrete states obtained from short but numerous unbiased MD trajectories. These methods are often known as Markov State Models or MSMs [5–10].

In this paper we explore the possibility of different algorithmic techniques for solving what is known as the "embeddability problem" [29,30] in financial literature. We present a comparison between six algorithms, used for predicting bond rating transitions, implemented in Inamura[31] and an extension by Marada[32] of Inamura's work with known and largely used methods of estimating free energies, as well as a maximum likelihood approach based on Hummer et al[33]. In order to explore the efficacy of each of these methods we have used a 1D bistable potential model.

Determination of kinetic properties is fundamental to the understanding of molecular systems. In order to use molecular dynamics as more than a virtual microscope, detailed calculations are necessary in order to search the conformational space for properties and conformations of interest. Various methods, such as those mentioned above, have been developed to improve the sampling in order to determine thermodynamic and kinetic properties of systems. These models are largely based upon defining a collective variable space[34,35] in which to reduce the dimensionality of the sampling space. The methods then address how to sample more efficiently along a predefined collective variable.

Markov State Models[36–39], MSMs, allow for the reduction of the complexities of a dynamic molecular system into a lower dimensional model. The conformational space can be reduced by clustering and then brought into a square empirical transition matrix allowing for the determination of various parameters of interest such as the relative free energies and diffusion constants among others. MSMs as with other computational methodologies still require a sufficient amount of sampling in order to obtain these parameters of interest. The present work adapts several algorithms from the field of finance in order to compare their efficacy with that of the state of the art in the chemical physics literature for addressing problems arising from insufficient sampling. In finance, Markov models are built in order to determine the probabilities of transition between rating grades

for bonds[31]. Often there is insufficient data in the bond marketplace to observe transitions from every rating to every other rating in the same fashion that the conformational space of a protein is too complex in order to sample every transition with molecular dynamics [31]. The discrete nature of the problem and the assumption of memoryless-ness cause the same models to work in both finance and chemical physics.

The algorithms, described below, are applied to a 1D bistable toy model. We test the efficacy of the various algorithms under different conditions, first in sufficient sampling to show that the algorithms do not distort the correct answer, and then in the case of increasing lag times as well as the case of insufficient data points. By doing this we are able to show the various strengths and weaknesses of each algorithm and hope to provide guidance for future researchers in their decision making about this matter.

Let us assume that we can assign any given sampled conformation from a set of MD simulations to one of the K well-defined states that are determined by clustering or other dimensionality reduction techniques. The empirical transition matrix, N, is defined as a $K \times K$ matrix, where $N_{i,j}$ is the number of observations of system being at state i at time t and being at state j at time $t+\tau$, in which τ is a given lag time. Transition matrix P can be obtained by normalizing N.

$$P_{i,j} = \frac{N_{i,j}}{\sum_k N_{i,k}}, \tag{1}$$

where $P_{i,j}$ is the probability of observing the system at state j at time $t+\tau$, given the system is observed at state i at time t.

From the transition matrix P, we can find the rate/generator matrix Q. While the P matrix is a more common quantity to work with in Markov chain based models such as MSM, the rate matrix is more well-known in chemical kinetics literature and is the focus of our work here. Q is relevant for continuous time models. There are advantages in using Q over P for such models; e.g., P is by construct dependent on the lag time, while Q is not.

Transition matrix P is a function of τ and is related to rate/generator matrix Q by $P(\tau) = \exp(Q\tau)$. Q can thus be estimated from P using:

$$Q = \frac{\log P}{\tau} \tag{2}$$

Q elements need to be real and criteria below need to be satisfied to have a valid generator matrix:

$$\begin{cases} \sum_{j=1}^{K} Q_{i,j} = 0, \ for \ 1 \leq i \leq K \\ Q_{i,j} \geq 0 \ for \ 1 \leq i,j \leq K \ with \ i \neq j. \end{cases} \tag{3}$$

The criteria above lead to:

$$Q_{i,i} \leq 0, \; for \; 1 \leq i \leq K. \tag{4}$$

Additional criteria need to be imposed in order to guarantee the validity of detailed balance relation, which is a common feature of equilibrium processes. Put simply, at equilibrium each process should be equilibrated by its reverse process. The detailed balance or reversibility feature then allows for the definition of free energies based on the equilibrium probability of states:

$$G_i = -k_B T \; \log(\pi_i), \tag{5}$$

where k_B is the Boltzmann constant and T is the temperature, G_i is the free energy of state i and π_i is the probability of observing the system at state i at equilibrium. The free energy between any two states i and j, $\Delta G_{i,j} = G(i) - G(j)$, can be calculated from Q using the detailed balance relation:

$$\Delta G_{i,j} = -k_B T \; \log(\frac{Q_{i,j}}{Q_{j,i}}). \tag{6}$$

Another interesting feature of relevance to molecular processes is the diffusivity, where no jump is allowed beyond the immediate neighbors of a state. The diffusivity condition is satisfied when the generator matrix Q is tridiagonal. For a diffusive process, the generator matrix can be used to determine the diffusion constant as[33]:

$$D_{i \leftrightarrow i+1} = \sqrt{Q_{i+1,i} Q_{i,i+1}}, \tag{7}$$

where $D_{i \leftrightarrow i+1}$ is related to the continuous diffusion constant of a diffusive process by the following approximate relationship:

$$D(\frac{x_i + x_{i+1}}{2}) \approx D_{i \leftrightarrow i+1} (x_{i+1} - x_i)^2, \tag{8}$$

where x_i represents the position along "reaction coordinate" associated with state i. If the states are equidistantly distributed along the reaction coordinate, we can estimate:

$$D_i \approx \frac{D_{i-1 \leftrightarrow i} + D_{i \leftrightarrow i+1}}{2}. \tag{9}$$

which leads to $D(x) = D_i \Delta x^2$, in which $\Delta x = x_i - x_{i-1}$ for all i.

The diffusion constant and free energy of states can be used to determine the mean first passage time (MFPT). This is the time of transition from the reactant (i.e, free energy minimum at x_R) to the product (i.e., free energy minimum at x_P. Lifson and Jackson [40] and others [41] have shown the MFPT can be calculated using:

$$MFPT_{R \to P} = \int_{x_R}^{x_P} \frac{1}{D(x)} \; \exp(G(x)/k_B T) \; \big(\int_{x_0}^{x} \exp(-G(y)/k_B T) \; dy \big) \; dx, \tag{10}$$

where the free energy and diffusion constant along x are described by $G(x)$ and $D(x)$, respectively. The MFPT can be estimated from G_i and D_i values as:

$$MFPT_{R\to P} = \sum_{i=i_R}^{i_P} \frac{1}{D_i} \exp(G_i/k_BT) \sum_{j=1}^{i} \exp(-G_j/k_BT), \tag{11}$$

where i_P and i_R are the reactant and product states.

III. THE EMBEDDABILITY PROBLEM

For a given empirical transition matrix, it has been shown that with a sufficient but not necessary condition [42], an exact generator matrix exists such that $Q = \frac{\log P}{\tau}$. However, the resulting Q may not be a valid generator matrix. For instance, consider states i and j such that transitions between the two are possible given infinite time but there is no recorded transition from i to j in the empirical transition matrix. Unfortunately, this is not an uncommon case for empirical transition matrices, due to under sampling certain transitions. Such behavior in empirical transition matrices would lead to empirical transition matrices that do not not satisfy the criteria in (3). One may even calculate negative eigenvalues for the empirical transition matrix, which would result in a non-real generator matrix.

We show several methods in the following section for addressing this problem, namely either adjusting directly the generator matrix produced by the matrix logarithm of our insufficiently sampled transition matrix, performing a maximum likelihood estimation, or using a Markov Chain Monte Carlo technique.

IV. ALGORITHMS

Here we present the algorithms compared in the results section. These algorithms range from very simple methods such as the Diagonal Adjustment Algorithm to more intricate methods such as Expectation Maximization.

A. Diagonal Adjustment (DA)

The Diagonal Adjustment (DA) algorithm is a simple and sometimes effective solution to the embeddability problem. The algorithm adjusts the generator matrix in two steps:

$$Step\ 1.\ \tilde{Q}_{i,j} = \begin{cases} 0,\ if\ (i \neq j)\ and\ Q_{i,j} < 0 \\ Q_{i,j},\ otherwise \end{cases} \tag{12}$$

$$Step\ 2.\ \tilde{Q}^{DA}_{i,j} = -\sum_{j=1, j\neq i}^{K} \tilde{Q}_{i,j} \tag{13}$$

B. Weighted Adjustment (WA)

The Weighted Adjustment (WA) is another simple algorithm very similar to the DA. The algorithm adjusts the generator matrix in two steps:

$$Step\ 1.\ \ \tilde{Q}_{i,j} = \begin{cases} 0\ if\ (i \neq j)\ and\ Q_{i,j} < 0 \\ Q_{i,j}\ otherwise \end{cases} \tag{14}$$

$$Step\ 2.\ \ \tilde{Q}^{WA}_{i,j} = Q_{i,j} - |\tilde{Q}_{i,j}| \frac{\sum_{j=1}^{K} \tilde{Q}_{i,j}}{\sum_{j=1}^{K} \tilde{|}Q_{i,j}|} \tag{15}$$

C. Quasi-Optimization of the Generator (QOG)

DA and WA are very similar methodologies. Unfortunately, they are not based upon an optimization strategy. To this end, Krenin and Sidelnikova [43] have extended the above work by implementing a post-adjustment optimization method called quasi-optimization of the generator (QOG). QOG works by first noting that the generator has a restriction on each row which allows the problem to be split into K distinct minimization problems of the sum of the squared deviation between $\log P$ and $Q\tau$. Thus we can write the problem as:

$$\tilde{Q}^{QOG} = \arg\min_{Q} ||Q\tau - \log P)||. \tag{16}$$

By reducing the problem into a distinct problem for each row we can define:

$$C_i = \left\{ \mathbf{z} \in \mathbb{R}^K | \sum_{j=1}^{K} z_j = 0, z_i \leq 0, z_j \geq 0\ for\ j \neq i \right\}, \tag{17}$$

where $\mathbf{z}$ represents possible valid values for the i'th row of Q. The optimum $\mathbf{z}$ can be determined from the i'th row of $\log(P)$ (denoted by $\mathbf{a}$) as:

$$\underset{\mathbf{z}\in C_i}{\arg\min} \sum_{j=1}^{K} (a_j \tau - z_j)^2 \tag{18}$$

An easy implementation of this algorithm can be found in the Jupyter notebook contained in the supplementary information.

D. Component-Wise Optimization (CWO)

Component-Wise Optimization[32] (CWO) is a somewhat more complex version of DA or WA which bears resemblance to QOG. The general idea is to divide the problem into K-1 seperate optimization problems. First, a DA or WA is performed followed by optimizing each individual value in the generator according to:

$$\tilde{Q}_{i,j}^{CWO} = \underset{Q_{i,j}\in[0,c]}{\arg\min} ||\exp(Q(Q_{i,j})\tau) - P|| \tag{19}$$

Followed by rebalancing the row to maintain a valid generator matrix. Constant c determines the desired convergence; the smaller the c, the faster the convergence. Marada goes on to note that the CWO is not capable of distinguishing between local and global minima and as such should be used judiciously, possibly as a way to further optimize results from other algorithms [32].

E. Expectation Maximization (EM)

The procedure of the Expectation Maximization (EM)[31] is done by calculating (20) and (21) for each element, and then using (23) to construct a new generator matrix. The iteration proceeds until convergence completes the algorithm.

$$\mathbb{E}[R_i(\tau)] = \frac{1}{M}\sum_{\alpha} \frac{1}{d_\alpha}\, e_{i_\alpha}^T \int_0^\tau \exp(Q\, s)\, (e_i\, e_i^T)\, \exp(Q\, (\tau - s))\, ds\, e_{j_\alpha}, \tag{20}$$

$$\mathbb{E}[N_{i,j}(\tau)] = \frac{1}{M}\sum_{\alpha} \frac{1}{d_\alpha}\, e_{i_\alpha}^T\, Q_{i,j} \int_0^\tau \exp(Q\, s)\, (e_i\, e_i^T)\, \exp(Q\, (\tau - s))\, ds\, e_{j_\alpha}, \tag{21}$$

where α is an index pointing to a specific transition from state i_α (observed at time t_α) to state j_α (observed at time $t_\alpha + \tau$), and M is the total number of transition observations for a given lag time τ. e_i is a one-hot vector with the i'th element being unity and the rest of elements being zero. e^T is the transpose of e. d_α is defined by:

$$d_\alpha = e_{i_\alpha}^T\, \exp(-Q\, \tau)\, e_{j_\alpha}. \tag{22}$$

The above relations allow to estimate, as a function of generator matrix, the expected values of $N_{i,j}$ and R_i, i.e., the number of transitions from state i to j and the amount of time the system stays in state i, respectively. The generator matrix is then estimated as:

$$\tilde{Q}_{i,j} = \frac{\mathbb{E}[N_{i,j}(\tau)]}{\mathbb{E}[R_i(\tau)]} \tag{23}$$

The two steps above (determining the expected values from the generator matrix and determining the generator matrix from the expected values) are repeated until converged.

F. Maximum Likelihood Estimator (MLE)

A common optimization approach is to use maximum likelihood estimation (MLE), in this case to determine a generator matrix to maximize the likelihood of observing all of the transitions, α (summarized in the empirical transition matrix for a given lag time τ)[33].

$$L = \prod_{\alpha} p(j_\alpha, t_\alpha + \tau | i_\alpha, t_\alpha) = \prod_{\alpha} \big(\exp(Q\tau)\big)_{i_\alpha, j_\alpha}, \tag{24}$$

which leads to:

$$\log L = \sum_{i,j} N_{i,j} (\exp(Q\tau)_{i,j}, \tag{25}$$

in which the off-diagonal elements of Q matrix are varied but stay positive by construct and the diagonal elements are determined by:

$$Q_{i,i} = -\sum_{j \neq i} Q_{i,j} \tag{26}$$

Other criteria can be easily imposed on Q such as reversibility (to satisfy the detailed balance) and diffusivity. To satisfy both reversibility and diffusivity, for instance, only $Q_{i,i\pm1}$ elements are varied and the other elements are determined by:

$$\tilde{Q}_{i,j} = \begin{cases} 0, \ if \ |i-j| > 1, \\ Q_{i,j}, \ if \ |i-j| = 1, \\ -Q_{i,i-1} - Q_{i,i+1} \ if \ i = j, \end{cases} \tag{27}$$

V. TOY MODEL

In order to examine the workings of different generator matrix estimators, we use a 1D bistable toy model, whose dynamics is modeled by an overdamped Langevin equation using the Euler-

Maruyama[44] method with parameters: temperature $T = 298K$, $mass = 1$, collision frequency $\gamma = 1$, and time step $\delta t = 10^{-6}$, where time and position are unitless. The potential was of the form:

$$U(r) = \frac{k}{4}(x^2 - 1)^2, \tag{28}$$

where $k = 10\frac{kcal}{mol}$. For sampling, 48 sets of simulations were performed. The starting points were equidistantly distributed between $x = -2$ and 2 with increments of $\Delta x = 4/47$. For each of the 48 starting points we ran 100 repeats of simulations, each for up to 1,000,000 steps. The simulations, however, were terminated if they visited a point outside the $[-2 : 2]$ range. After performing the simulations we created empirical transitions matrices for various lag times, which were then used as the seed information for our algorithms and subsequent detailed balance free energy calculations.

In order to test the efficacy of each algorithm we decided to vary lag time and trajectory lengths. Trajectory length was varied by cutting off our counting statistics for each copy at the appropriate point. For instance, a 50% trajectory length (denoted by $L = 50\%$) includes the first half of each trajectory. For lag time, we utilized a sliding window approach in creating our empirical transition matrix. In this approach, a transition from state i at any time t to state j at time $t + \tau$ would count towards $N_{i,j}$. The total number of transition observations based on a given trajectory would be $\tau/\delta t$ data points less than the total number of data points in the trajectory. The sliding window approach guarantees a minimal dependence on the lag time for the number of observations to avoid bias.

From the results of our algorithms we calculated the percent error from our expected values, particularly we expected a total ΔG of 0 $\frac{kcal}{mol}$ between the two energy minima at $x = \pm 1$, and a ΔG of 2.5 $\frac{kcal}{mol}$ between the energy barrier at $x = 0$ and the average of the two energy minima at $x = \pm 1$.

VI. RESULTS AND DISCUSSION

Here we will use the data generated for the toy model above to estimate the generator matrix based on various lengths of trajectories and with various lag times using six algorithms discussed above including DA, WA, QOG, CWO, EM, and MLE. We examine the robustness of these methods in terms of their performance with varying lag time and trajectory length. First, we use a fixed trajectory length (in this case, $L = 10\%$) to examine the dependence on the lag time. Figure 2

shows the estimated free energies along with the analytical potential energy for the toy model. At low τ, all algorithms work quite reasonably, at least in terms of their free energy estimates. At longer lag times, CWO is the first algorithm to break down, which can be seen visibly in the free energy profiles based on $\tau = 50\delta$ and above. Panels H and I in Fig. 2 more clearly demonstrate the τ-dependent behavior of barrier height (defined as $G(0) - \frac{G(1)+G(-1)}{2}$) and well symmetry (defined as $|G(1) - G(-1)|$) as important features of the free energy profile. In addition to CWO, whose poor performance is quite clear both in Fig. 2H and Fig. 2I, WA method is also somewhat inaccurate when compared to the other methods as it consistently overestimates the barrier height at all τ. MLE and EM that are both conceptually similar, also have a similar performance in both barrier height and well symmetry estimates. They both overestimate the barrier height as τ increases while DA and QOG seem to be more robust in barrier height estimation even at long lag times.

The poor performance of CWO and WA relative to the other four methods persists when we consider different trajectory lengths. Figure 3 illustrates the free energy profiles estimated from various methods for a fixed τ ($\tau = 100\delta t$) and varying trajectory length. We examine the performance of algorithms as we get to fewer data points (i.e., shorter trajectories). Specifically, WA and CWO over- and underestimate the barrier height even at $L = 100\%$ (see Fig. 3F). The other methods clearly have a better performance at $L = 10\%$ and beyond. When we use shorter trajectories the barrier estimates start to deviate from the correct values. From Fig. 3, it appears that the barrier simply gets flatter. This can be easily understood from observing that with shorter trajectory lengths, there is more dependence on the starting point of trajectories, which were distributed evenly over the space. As the trajectory length shortens, the free energy profiles resemble more the initial flat distribution.

From our free energy estimates, it is clear that WA and CWO have the poorest performance among the six methods implemented here. Tables I and II provide more extensive data on the dependence of these algorithms on lag time and trajectory length as they summarize various estimates based on varying both quantities. At the first row for each method, the errors associated with the barrier height and well symmetry estimates are listed in Tables I and II, respectively for full-length trajectories ($L = 100\%$) but varying lag times. In all cases, the barrier height errors are quite small at $\tau = \delta t$ but the well symmetry error is considerably higher than the rest for CWO and WA (by a factor of 25 and 5, respectively) at $L = 100\%$ and $\tau = \delta t$. The well symmetry errors are generally negligible in most cases except for CWO and WA methods. However, barrier

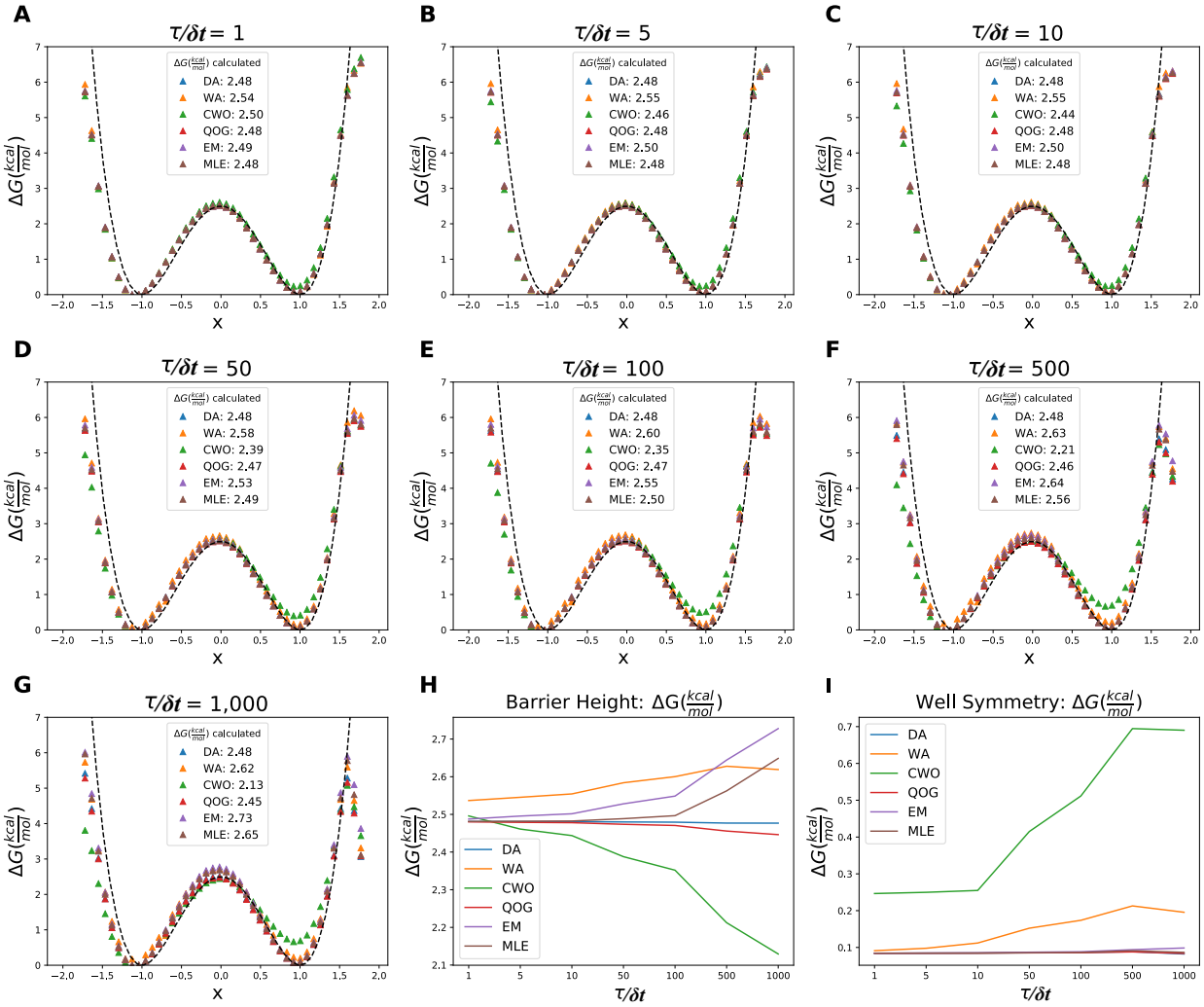

FIG. 2. (A-G) Estimated free energy along x from DA (blue), WA (orange), CWO (green), QOG (red), EM (purple) and MLE (brown) algorithms for various lag times (expressed in terms of the number of time steps) based on 10% of the data ($L = 10\%$); (A) $\tau = \delta t$, (B) $\tau = 5\delta t$, (C) $\tau = 10\delta t$, (D) $\tau = 50\delta t$, (E) $\tau = 100\delta t$, (F) $\tau = 500\delta t$, and (G) $\tau = 1000\delta t$. The black reference line is the model potential: $2.5(x^2-1)^2$. (H) The estimated energy barrier height as a function of τ, with an analytical value of 2.5. (I) The absolute difference of the energy minima as a function of τ with an analytical value of 0.

height estimates worsen in most cases as the trajectory length decreases or as lag time increases. At $L = 0.1\%$, all algorithms perform similarly in barrier height estimates.

The above analysis implies that most algorithms above (except for CWO and WA) are more robust for estimating thermodynamic quantities (here the free energy difference between the minima) than kinetic ones (here the barrier height, which is related to activation energy in chemical

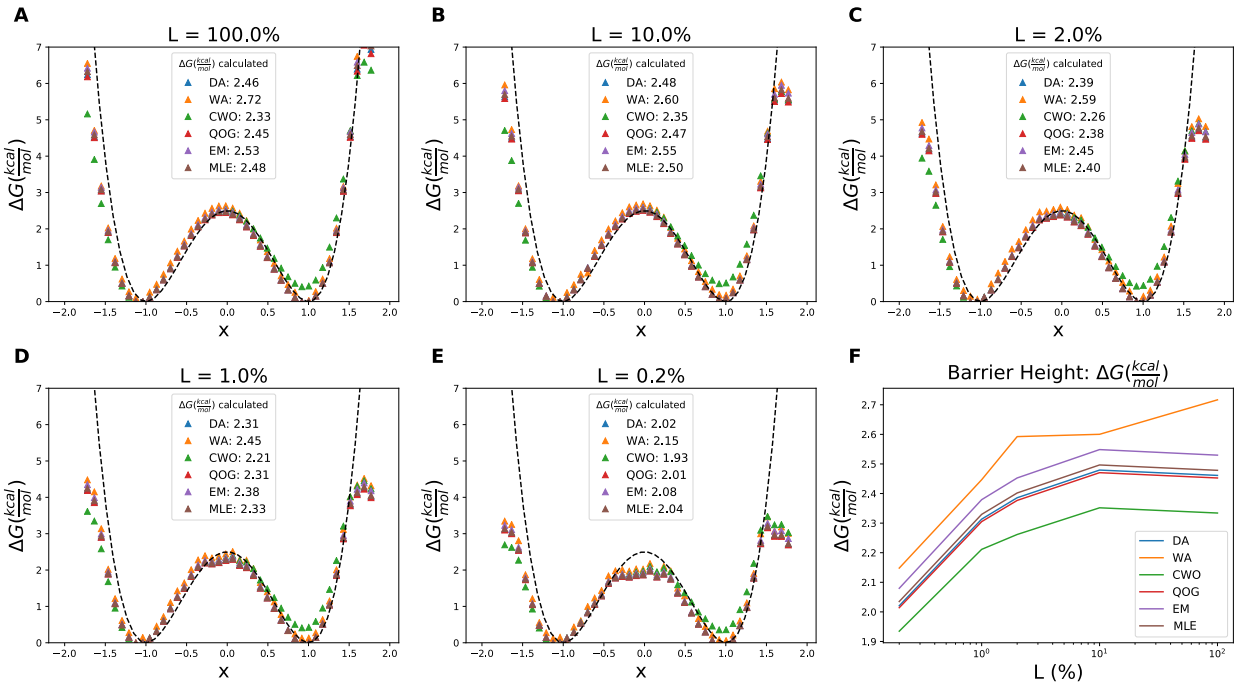

FIG. 3. (A-E) Estimated free energy for $\tau = 100\delta t$ along x from DA (blue), WA (orange), CWO (green), QOG (red), EM (purple), and MLE (brown) for various trajectory lengths; (A) $L = 100\%$, (B) $L = 10\%$, (C) $L = 2\%$, (D) $L = 1\%$, (E) $L = 0.2\%$. (F) The energy barrier height ΔG (as a function of trajectory length) with an analytical value of 2.5 $\frac{kcal}{mol}$.

kinetics literature). We have also examined the performance of these methods in terms of estimating MFPT associated with the transition from the energy minimum at $x = -1$ to the energy minimum at $x = 1$ (i.e., from the left to the right well). The analytical MFPT is 34.3. This value is estimated from Relation (10), using $G(x) = U(x)$ (from Relations (28)) and $D = k_B T = 0.59$. To estimate the MFPT from the simulations, we employ Relations (11), which requires an estimate for the free energies (G_i) as well as D_i. Table III provides an estimate for $D(x)$ (averaged over x) from generator matrices that are estimated by DA, WA, CWO, QOG, EM, and MLE. Generally, the estimates are influenced more by τ than the trajectory lengths. Unlike the free energy estimates that tend to be more accurate for lower τ values, the diffusion constant estimates are considerably more accurate for longer lag times. The MLE method is the only exception, whose estimate for diffusion constant is equally well for any lag time (and any trajectory length). Other methods have their best estimates (which is comparable to MLE) only at the longest lag time ($\tau = 1000\delta t$) and their estimate for diffusion constant increases as the lag time decreases.

MFPT estimates based on D_i and G_i estimates are shown in Table *IV* for varying lag times

and trajectory lengths. As compared to the analytical value (34.4), all estimates are somewhat lower than expected, which is the direct result of the overestimation of $D(x)$. Similar to $D(x)$ and contrary to $G(x)$, the estimates are more accurate at longer lag times. The exception is the MLE method, which is quite robust and give similar MFPT estimates for different lag times. The MFPT estimates are also affected by the trajectory length. For $L = 100\%$ and $L = 10\%$, the difference is not significant between the MFPT estimates; howevere, at $L = 1\%$ most methods start to show more deviation from the actual MFPT values as compared to $L = 10\%$ and $L = 100\%$ estimates. The exception is again the MLE algorithm, which is clearly the most robust method in MFPT estimation. At $L = 1\%$, all methods including MLE show more deviations from the correct answer; although MLE still performs better than the other methods.

The generator matrix contains both thermodynamic and kinetic information as the above analysis clearly illustrates. The six different generator matrix estimators that we used here have their own pros and cons. There are certain features that are universal. For instance, the more data is always the more accurate the estimate is; although different algorithms have different sensitivity levels to the amount of data used. For instance the MLE method is a more robust estimator, particularly for MFEP estimation. WA and CWO methods are relatively poor free energy estimators as compared to the other methods used here; however, they perform similar to other methods (with the exception of MLE) for estimating diffusion constant and MFEP. The MLE algorithm is generally robust for estimating both thermodynamic and kinetic properties. However, for free energy estimation, it behaves quite similar to the EM algorithm and is outperformed by the simple DA method as well as the QOG method at longer lag times. An interesting observation is that free energy estimates are generally more accurate for shorter lag times but the diffusion constant estimates are more accurate for longer lag times (with the exception of MLE). The overestimation of diffusion constant (and subsequent underestimation of the MFPT) can be attributed to the correlatedness of the data generated by the overdamped langevin equation. A typical molecular dynamics simulations, we also expect to observe such a behavior, which would result in similar correlatedness in the data. The longer lag times result in more decorrelation in the empirical transition matrix. On the other hand, longer lag times result in empirical transition matrices that have more non-zero off-diagonal elements and are farther away from the generator matrix, which in this case is expected to be tridiagonal.

TABLE I. Percent errors in barrier height ($\Delta G = G(0) - \frac{G(-1)+G(1)}{2}$) based on different free energy predictions. The predictions are based on various estimates of the generator matrix including: DA, WA,QOG,CWO, EM, and MLE. Increasing values of τ are seen along the columns while decreasing trajectory length (L) are represented along the rows.

$\tau/\delta t$	1	10	100	1000	L (%)
DA	1.54	1.54	1.55	1.60	100
	0.76	0.77	0.83	0.93	10
	6.92	7.05	7.46	9.06	1
	21.64	22.52	25.35	30.14	0.1
WA	1.90	4.55	8.67	7.91	100
	1.47	2.17	4.01	4.76	10
	3.31	2.78	2.13	3.78	1
	15.30	15.48	16.09	20.91	0.1
QOG	1.57	1.65	1.90	2.73	100
	0.79	0.88	1.18	2.15	10
	6.95	7.15	7.80	10.01	1
	21.67	22.59	25.50	30.29	0.1
CWO	0.85	3.07	6.64	15.41	100
	0.17	2.25	5.93	14.82	10
	6.23	8.32	11.55	20.65	1
	20.83	23.35	27.63	36.76	0.1
EM	1.28	0.71	1.19	8.36	100
	0.50	0.06	1.94	9.11	10
	6.67	6.26	4.83	0.75	1
	21.43	21.84	23.03	23.01	0.1
MLE	1.53	1.47	0.86	5.21	100.0
	0.75	0.70	0.14	5.94	10
	6.91	6.99	6.82	2.76	1
	21.64	22.47	24.76	25.24	0.1

TABLE II. Absolute errors in $\Delta G = |G(1) - G(-1)|$, the free energy difference between the two minima, with an analytical value of 0. The predictions are based on various estimates of the generator matrix including: DA, WA,QOG,CWO, EM, and MLE. Increasing values of τ are seen along the columns while decreasing trajectory length (L) are represented along the rows.

$\tau/\delta t$	1	10	100	1000	L (%)
DA	0.01	0.01	0.01	0.01	100
	0.08	0.08	0.09	0.08	10
	0.00	0.00	0.01	0.02	1
	0.05	0.04	0.00	0.11	0.1
WA	0.05	0.11	0.16	0.17	100
	0.09	0.11	0.17	0.20	10
	0.01	0.05	0.12	0.17	1
	0.07	0.00	0.20	0.21	0.1
QOG	0.01	0.01	0.01	0.01	100
	0.08	0.08	0.09	0.08	10
	0.00	0.00	0.01	0.03	1
	0.05	0.04	0.00	0.11	0.1
CWO	0.15	0.16	0.42	0.61	100
	0.25	0.25	0.51	0.69	10
	0.17	0.18	0.44	0.64	1
	0.10	0.15	0.45	0.67	0.1
EM	0.01	0.01	0.01	0.01	100
	0.08	0.09	0.09	0.10	10
	0.00	0.00	0.01	0.02	1
	0.06	0.04	0.01	0.09	0.1
MLE	0.01	0.01	0.01	0.01	100.0
	0.08	0.08	0.09	0.09	10
	0.00	0.00	0.01	0.03	1
	0.05	0.04	0.01	0.12	0.1

TABLE III. Diffusion constant $D(x)$ averaged over $x = -2$ to $x = 2$ and estimated from D_i values. Increasing values of τ are seen along the columns while decreasing trajectory length (L) are represented along the rows. The predictions are based on various estimates of the generator matrix including: DA, WA,QOG,CWO, EM, and MLE. The underlying model has an analytical value of $D(x) = 0.59$ for all x.

$\tau/\delta t$	1	10	100	1000	L (%)
DA	36.64	11.52	3.47	1.01	100
	37.12	11.66	3.50	1.02	10
	37.65	11.81	3.53	1.03	1
	38.51	12.09	3.62	1.06	0.1
WA	36.08	11.43	3.51	1.11	100
	36.78	11.62	3.56	1.12	10
	37.36	11.79	3.59	1.14	1
	38.20	12.06	3.67	1.17	0.1
QOG	36.67	11.54	3.49	1.04	100
	37.14	11.68	3.52	1.05	10
	37.67	11.83	3.55	1.06	1
	38.53	12.11	3.63	1.09	0.1
CWO	36.92	11.87	3.79	1.23	100
	37.39	11.97	3.78	1.22	10
	37.90	12.08	3.77	1.22	1
	38.74	12.34	3.84	1.23	0.1
EM	36.96	11.84	3.78	1.36	100
	37.44	11.98	3.82	1.37	10
	37.97	12.13	3.84	1.39	1
	38.84	12.42	3.93	1.43	0.1
MLE	1.03	1.06	1.10	1.21	100.0
	0.96	0.99	1.03	1.16	10
	0.88	0.93	0.98	1.12	1
	0.76	0.87	0.94	1.10	0.1

VII. CONCLUSIONS

We have implemented six algorithms for overcoming the "embeddability problem" in MSMs. The efficiency of these algorithms were tested on a 1D bistable toy model. The relative strengths of each algorithm were compared with regards to their ability to handle differing lag times and trajectory lengths. We found that depending on whether one is interested in free energy calculations or kinetic characterization of a process, different algorithms may behave differently. For thermodynamic characterization of a process, all methods except for WA and CWO are recommended. For barrier height estimation, the QOG method and the simple DA method work well while EM and MLE also behave reasonably; although they work better at lower lag times. In terms of kinetic estimations such as MFPT and rate constant, the MLE method is clearly the most robust method. In general, in the abundance of data, all algorithms provide reasonable results; however, the lag time dependence should be checked by repeating the analysis for various lag times to ensure the reliability of the results.

Overall, this exploration provides the interested reader a foundation for dealing with kinetic and thermodynamic analysis of equilibrium MD trajectories, particularly in the context of MSM. Through our analysis we hope to have shed some light on the approaches as well as the information which can be garnered from direct analysis of the MD data. More data is almost always preferable but in situations where this is not possible, overcoming the embeddability problem is a possibility for estimating the thermodynamic and kinetic quantities reliably.

Supporting Information

A Jupyter notebook containing the algorithms discussed in this manuscript is provided.

Data Availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Acknowledgement

This research is supported by the National Science Foundation under CHE 1945465 and OAC 1940188. This research is also supported by the Arkansas High Performance Computing Center which is funded through multiple National Science Foundation grants and the Arkansas Economic Development Commission.

REFERENCES

[1]Herman JC Berendsen and Steven Hayward. Collective protein dynamics in relation to function. *Current opinion in structural biology*, 10(2):165–169, 2000.

[2]Peter Eastman, Jason Swails, John D Chodera, Robert T McGibbon, Yutong Zhao, Kyle A Beauchamp, Lee-Ping Wang, Andrew C Simmonett, Matthew P Harrigan, Chaya D Stern, et al. Openmm 7: Rapid development of high performance algorithms for molecular dynamics. *PLoS computational biology*, 13(7):e1005659, 2017.

[3]Andreas W Götz, Mark J Williamson, Dong Xu, Duncan Poole, Scott Le Grand, and Ross C Walker. Routine microsecond molecular dynamics simulations with amber on gpus. 1. generalized born. *Journal of chemical theory and computation*, 8(5):1542–1555, 2012.

[4]Brian Towles, JP Grossman, Brian Greskamp, and David E Shaw. Unifying on-chip and internode switching within the anton 2 network. *ACM SIGARCH Computer Architecture News*, 42(3):1–12, 2014.

[5]Christophe Andrieu, Arnaud Doucet, and Roman Holenstein. Particle Markov Chain Monte Carlo Methods. *Journal Of The Royal Statistical Society Series B-Statistical Methodology*, 72(3):269–342, 2010.

[6]Y Ephraim and N Merhav. Hidden Markov Processes. *IEEE Transactions On Information Theory*, 48(6):1518–1569, JUN 2002.

[7]Vijay S. Pande, Kyle Beauchamp, and Gregory R. Bowman. Everything You Wanted To Know About Markov State Models But Were Afraid To Ask. *Methods*, 52(1):99–105, SEP 2010.

[8]Gregory R. Bowman, Vincent A. Voelz, and Vijay S. Pande. Taming The Complexity Of Protein Folding. *Current Opinion In Structural Biology*, 21(1):4–11, FEB 2011.

[9]Sa Sisson. Transdimensional Markov Chains: A Decade Of Progress And Future Perspectives. *Journal Of The American Statistical Association*, 100(471):1077–1089, SEP 2005.

[10]Diwakar Shukla, Carlos X. Hernandez, Jeffrey K. Weber, and Vijay S. Pande. Markov State Models Provide Insights Into Dynamic Modulation Of Protein Function. *Accounts Of Chemical Research*, 48(2):414–422, FEB 2015.

[11]Gregory R Bowman, Kyle A Beauchamp, George Boxer, and Vijay S Pande. Progress and challenges in the automated construction of markov state models for full protein systems. *The Journal of chemical physics*, 131(12):124101, 2009.

[12]Natasa Djurdjevac, Marco Sarich, and Christof Schütte. Estimating the eigenvalue error of markov state models. *Multiscale Modeling & Simulation*, 10(1):61–81, 2012.

[13]Jan-Hendrik Prinz, John D Chodera, and Frank Noé. Spectral rate theory for two-state kinetics. *Physical Review X*, 4(1):011020, 2014.

[14]P.A. Gagniuc. *Markov Chains: From Theory To Implementation And Experimentation*. Wiley, 2017.

[15]Rafael C. Bernardi, Marcelo C. R. Melo, and Klaus Schulten. Enhanced Sampling Techniques In Molecular Dynamics Simulations Of Biological Systems. *Biochimica Et Biophysica Acta-General Subjects*, 1850(5, Si):872–877, May 2015.

[16]S Kumar, D Bouzida, Rh Swendsen, Pa Kollman, and Jm Rosenberg. The Weighted Histogram Analysis Method For Free-Energy Calculations On Biomolecules .1. The Method. *Journal Of Computational Chemistry*, 13(8):1011–1021, Oct 1992.

[17]Gm Torrie and Jp Valleau. Non-Physical Sampling Distributions In Monte-Carlo Free-Energy Estimation - Umbrella Sampling. *Journal Of Computational Physics*, 23(2):187–199, 1977.

[18]B Roux. The Calculation Of The Potential Of Mean Force Using Computer-Simulations. *Computer Physics Communications*, 91(1-3):275–282, Sep 1995.

[19]S Kumar, Jm Rosenberg, D Bouzida, Rh Swendsen, and Pa Kollman. Multidimensional Free-Energy Calculations Using The Weighted Histogram Analysis Method. *Journal Of Computational Chemistry*, 16(11):1339–1350, Nov 1995.

[20]Alessandro Barducci, Giovanni Bussi, and Michele Parrinello. Well-Tempered Metadynamics: A Smoothly Converging And Tunable Free-Energy Method. *Physical Review Letters*, 100(2), Jan 18 2008.

[21]Alessandro Laio and Francesco L. Gervasio. Metadynamics: A Method To Simulate Rare Events And Reconstruct The Free Energy In Biophysics, Chemistry And Material Science. *Reports On Progress In Physics*, 71(12), Dec 2008.

[22]Alessandro Barducci, Massimiliano Bonomi, and Michele Parrinello. Metadynamics. *Wiley Interdisciplinary Reviews-Computational Molecular Science*, 1(5):826–843, Sep-Oct 2011.

[23]A Laio, A Rodriguez-Fortea, Fl Gervasio, M Ceccarelli, and M Parrinello. Assessing The Accuracy Of Metadynamics. *Journal Of Physical Chemistry B*, 109(14):6714–6721, Apr 14 2005.

[24]Y Sugita and Y Okamoto. Replica-Exchange Molecular Dynamics Method For Protein Folding. *Chemical Physics Letters*, 314(1-2):141–151, Nov 26 1999.

[25]K Hukushima and K Nemoto. Exchange Monte Carlo Method And Application To Spin Glass Simulations. *Journal Of The Physical Society Of Japan*, 65(6):1604–1608, Jun 1996.

[26]Dj Earl and Mw Deem. Parallel Tempering: Theory, Applications, And New Perspectives. *Physical Chemistry Chemical Physics*, 7(23):3910–3916, 2005.

[27]Y Sugita, A Kitao, and Y Okamoto. Multidimensional Replica-Exchange Method For Free-Energy Calculations. *Journal Of Chemical Physics*, 113(15):6042–6051, Oct 15 2000.

[28]H Fukunishi, O Watanabe, and S Takada. On The Hamiltonian Replica Exchange Method For Efficient Sampling Of Biomolecular Systems: Application To Protein Structure Prediction. *Journal Of Chemical Physics*, 116(20):9058–9067, May 22 2002.

[29]M-A Guerry. Some Results On The Embeddable Problem For Discrete-Time Markov Models In Manpower Planning. *Communications In Statistics-Theory And Methods*, 43(7, Si):1575–1584, Apr 3 2014.

[30]M-A Guerry. On The Embedding Problem For Discrete-Time Markov Chains. *Journal Of Applied Probability*, 50(4):918–930, Dec 2013.

[31]Yasunari Inamura and Others. Estimating continuous time transition matrices from discretely observed data. *Bank Of Japan*, pages 06–07, 2006.

[32]TomÁS Marada. Quantitative credit risk modeling: Ratings under stochastic time.

[33]G Hummer. Position-Dependent Diffusion Coefficients And Free Energies From Bayesian Analysis Of Equilibrium And Replica Molecular Dynamics Simulations. *New Journal Of Physics*, 7, Jan 31 2005.

[34]Ii Kurylyak and Ir Yukhnovskii. The Method Of Collective Variables In The Equilibrium Statistical-Theory Of Bounded Systems Of Charged-Particles .1. Continuum Model Of An Electrolyte Solution Occupying A Half-Space. *Theoretical And Mathematical Physics*, 52(1):691–699, 1982.

[35]Ilaria Gimondi, Gareth A. Tribello, and Matteo Salvalaglio. Building Maps In Collective Variable Space. *Journal Of Chemical Physics*, 149(10), Sep 14 2018.

[36]John D. Chodera, Nina Singhal, Vijay S. Pande, Ken A. Dill, and William C. Swope. Automatic Discovery Of Metastable States For The Construction Of Markov Models Of Macromolecular Conformational Dynamics. *Journal Of Chemical Physics*, 126(15), Apr 21 2007.

[37]Jan-Hendrik Prinz, Hao Wu, Marco Sarich, Bettina Keller, Martin Senne, Martin Held, John D. Chodera, Christof Schuette, and Frank Noe. Markov Models Of Molecular Kinetics: Generation And Validation. *Journal Of Chemical Physics*, 134(17), May 7 2011.

[38]Gregory R. Bowman, Kyle A. Beauchamp, George Boxer, and Vijay S. Pande. Progress And Challenges In The Automated Construction Of Markov State Models For Full Protein Systems. *Journal Of Chemical Physics*, 131(12), Sep 28 2009.

[39]Kyle A. Beauchamp, Gregory R. Bowman, Thomas J. Lane, Lutz Maibaum, Imran S. Haque, and Vijay S. Pande. Msmbuilder2: Modeling Conformational Dynamics On The Picosecond To Millisecond Scale. *Journal Of Chemical Theory And Computation*, 7(10):3412–3419, Oct 2011.

[40]Shneior Lifson and Julius L Jackson. On the self-diffusion of ions in a polyelectrolyte solution. *The Journal of Chemical Physics*, 36(9):2410–2414, 1962.

[41]Anjum Ansari. Mean first passage time solution of the smoluchowski equation: Application to relaxation dynamics in myoglobin. *The Journal of Chemical Physics*, 112(5):2516–2522, 2000.

[42]Rb Israel, Js Rosenthal, and Jz Wei. Finding Generators For Markov Chains Via Empirical Transition Matricesamura , With Applications To Credit Ratings. *Mathematical Finance*, 11(2):245–265, Apr 2001.

[43]Alexander Kreinin and Marina Sidelnikova. Regularization algorithms for transition matrices. *Algo Research Quarterly*, 4(1/2):23–40, 2001.

[44]Gisiro Maruyama. Continuous markov processes and stochastic equations. *Rendiconti del Circolo Matematico di Palermo*, 4(1):48, 1955.

TABLE IV. MFPT predicted using various estimates of the generator matrix including: DA, WA,QOG,CWO, EM, and MLE. Increasing values of τ are seen along the columns while decreasing trajectory length (L) are represented along the rows. The analytical MFPT is 34.3.

$\tau/\delta t$	1	10	100	1000	L (%)
DA	0.56	1.78	5.71	19.47	100
	0.60	1.92	6.19	21.13	10
	0.49	1.54	4.92	16.30	1
	0.28	0.87	2.72	9.13	0.1
WA	0.64	2.04	6.71	21.04	100
	0.68	2.27	7.91	25.86	10
	0.54	1.82	6.34	20.37	1
	0.30	1.02	3.45	10.85	0.1
QOG	0.56	1.77	5.62	18.42	100
	0.60	1.91	6.09	19.94	10
	0.48	1.53	4.84	15.45	1
	0.28	0.87	2.70	8.77	0.1
CWO	0.64	1.87	6.28	17.58	100
	0.69	2.04	6.85	19.09	10
	0.55	1.65	5.53	15.68	1
	0.31	0.96	3.33	10.28	0.1
EM	0.56	1.77	5.69	19.78	100
	0.60	1.92	6.19	21.84	10
	0.49	1.53	4.89	16.37	1
	0.28	0.86	2.65	7.95	0.1
MLE	20.89	20.49	20.82	25.03	100.0
	23.43	22.99	23.07	27.56	10
	19.83	18.83	18.36	20.88	1
	13.06	11.43	10.60	11.08	0.1