THE JOURNAL OF PHYSICAL CHEMISTRY A
A JOURNAL OF THE AMERICAN CHEMICAL SOCIETY

pubs.acs.org/JPCA

Article

Addressing the Embeddability Problem in Transition Rate Estimation

Published as part of The Journal of Physical Chemistry virtual special issue "Early-Career and Emerging Researchers in Physical Chemistry Volume 2".

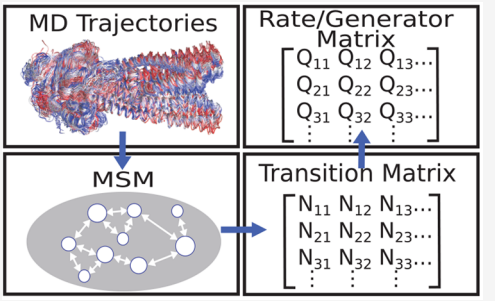
Curtis Goolsby, James Losey, Ashkan Fakharzadeh, Yuchen Xu, Marie-Christine Düker, Mila Getmansky Sherman, David S. Matteson, and Mahmoud Moradi*

Cite This: *J. Phys. Chem. A* 2023, 127, 5745−5759

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: 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. This work discusses 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 itself is built from the empirical transition matrix, it provides an alternative approach for estimating both thermodynamic and kinetic quantities, particularly in diffusive processes. A fundamental issue with this approach is known as the embeddability problem. The key contribution of this work is the introduction of a novel method to address the embeddability problem as well as the collection and utilization of existing algorithms previously used in the literature. The algorithms are tested on data from a one-dimensional toy model to show the workings of these methods and discuss the robustness of each method in dependence of lag time and trajectory length.



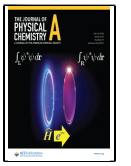
1. INTRODUCTION

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, all-atom MD simulations show a promising path toward exploring conformational free energy landscapes of proteins and other biomolecules.[2−4] However, 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.

Enhanced sampling techniques[5] have been particularly successful as an effective remedy for the costs involved with simulating large molecular systems. Methods such as umbrella sampling,[6−9] metadynamics,[10−13] replica exchange,[14−18] and their various extensions 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 energy requires postanalysis reweighting techniques 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 biased simulations. 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.[19−24] MSMs[25−28] allow for the reduction of the complexities of a dynamic molecular system into a lower dimensional model by discretizing the conformational space using a clustering technique. An empirical transition matrix can then be built to determine various parameters of interest such as the relative free energies and transition rates.

MSMs[19−24] provide some of the most powerful tools for analyzing the ensembles of short MD trajectories to extract information on both thermodynamics and kinetics of complex

Received: February 28, 2023
Revised: May 31, 2023
Published: June 28, 2023



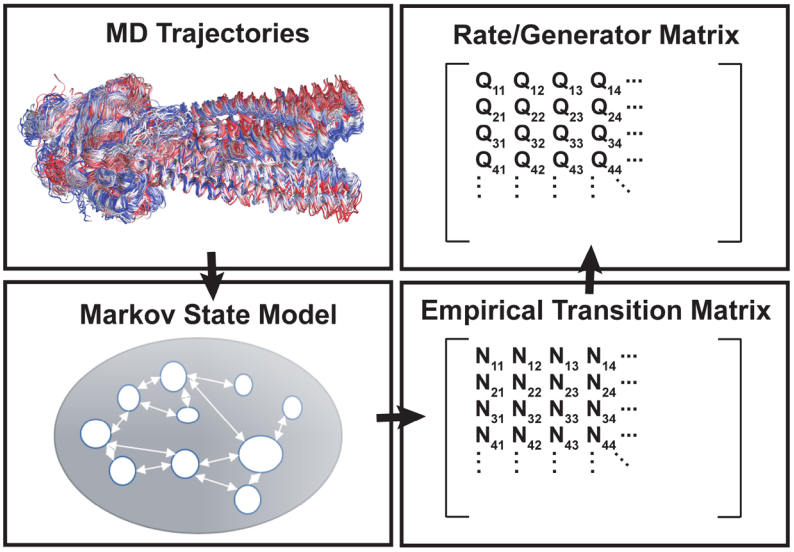


Figure 1. Schematic representation of the proposed approach to analyze the MSM data. MD trajectories are first used within a clustering scheme to generate a Markov state model and build an empirical transition matrix using a given lag time; see MSM literature.[21] Instead of spectral decomposition of 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.

biomolecular systems and predict the behavior of such systems at much longer time scales than would be 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.[27,29,30] MSMs, as with other computational methodologies, still require a sufficient amount of sampling in order to obtain the parameters of interest. A MSM is generally a continuous-time Markov model that is ubiquitous in various fields of science and engineering. In finance, for instance, such 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 MD.[31]

Building MSMs involves multiple steps: (1) discretization of conformational space, (2) extracting transition statistics from simulation trajectories, and (3) analyzing the transition statistics to estimate kinetic and thermodynamic properties.[32] Here, we only focus on the third component, which assumes an empirical transition matrix has been generated. We discuss a known[33] but less commonly used approach for analyzing empirical transition matrices that in some cases provides 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.[34] In contrast, the approach we discuss here is based on building a rate matrix to estimate the kinetic and thermodynamic quantities using standard methods from chemical kinetics literature (Figure 1). In the finance literature, the rate matrix is known as generator matrix.[35] Producing a rate matrix from the empirical transition matrix is known to be associated with the embeddability problem.[36,37] The embeddability problem is found in taking the matrix logarithm of a time-dependent transition probability matrix in order to determine the generator matrix. 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 nonreal values and is an invalid generator matrix. Solving this embeddability problem involves accurately predicting the true generator matrix from an invalid generator matrix.

Here we explore various algorithms in the chemical and financial literature that address the embeddability problem and propose a novel algorithm as well. We specifically compare eight algorithms, five of which are used for predicting bond rating transitions, implemented by Inamura[31] and Marada.[38] Other algorithms include a maximum likelihood estimator (MLE) based on the work of Hummer et al.[33] and a quadratic programming method based on the work of Commelin and Vanden-Eijnden,[39] as well as a new algorithm proposed here. The algorithms, described below, are applied to a 1D bistable toy model simulated using overdamped Langevin equation. We test the efficacy of the various algorithms with various lag times and simulation times. 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.

2. THE EMBEDDABILITY PROBLEM

The embeddability problem for Markov chains is a known problem in probability theory and was proposed in ref.[40] Ever since it has been studied by numerous authors[41−46] and there are still only partial solutions available. As stated in ref.,[40] the problem can be formulated as follows: Given a $K \times K$ stochastic matrix $\mathbf{P}$ and a parameter τ, can one find a generator matrix $\mathbf{Q}$ such that $\mathbf{P} = \exp(\mathbf{Q}\tau)$? In probability theory, this question is known as the embeddability problem for stochastic matrices or for finite Markov chains. Given a discrete-time homogeneous Markov chain with transition probability matrix $\mathbf{P}$, the embeddability problem is equivalent to asking whether we can find a continuous-time homogeneous Markov chain with transition semigroup $P = P(1)$.

https://doi.org/10.1021/acs.jpca.3c01367
J. Phys. Chem. A 2023, 127, 5745−5759

Suppose that our observations follow a Markov chain model with K states and transition probability matrix $P = (P_{i,j})_{i,j=1,\ldots,K}$. The maximum likelihood estimator $\mathbf{P}$ for the transition matrix $\mathbf{P}$ can then be written as

$$\mathbf{P} = (P_{i,j})_{i,j=1,\ldots,K}, \qquad P_{i,j} = \frac{N_{i,j}}{\sum_{k=1}^{K} N_{i,k}} \tag{1}$$

where $N_{i,j}$ denotes the number of observations in the 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. For future reference, we collect the number of transitions in the matrix $\mathbf{N} = (N_{i,j})_{i,j=1,\ldots,K}$. The question is now, whether we can obtain an estimator $\mathbf{Q}$ for the rate/generator matrix $\mathbf{Q}$ from the estimator $\mathbf{P}$ of the transition matrix $\mathbf{P}$ which satisfies $\mathbf{P}(\tau) = \exp(\mathbf{Q}\tau)$.

While the transition matrix $\mathbf{P}$ 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. The rate/generator matrix $\mathbf{Q}$ is relevant for continuous time models. There are advantages in using $\mathbf{Q}$ over $\mathbf{P}$ for such models; e.g., $\mathbf{P}$ is by construction dependent on the lag time, while $\mathbf{Q}$ is not.

Transition matrix $\mathbf{P}$ is a function of τ and is related to rate/generator matrix $\mathbf{Q}$ by $P(\tau) = \exp(Q\tau)$. The matrix $\mathbf{Q}$ can thus be estimated from an estimator for $\mathbf{P}$ within the Markovian approximation using the relation

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

However, the elements of the matrix $\mathbf{Q} = (Q_{i,j})_{i,j=1\ldots,K}$ need to be real-valued and satisfy the properties

$$\begin{cases} \sum_{j=1}^{K} Q_{i,j} = 0, & \text{for } 1 \le i \le K, \\ Q_{i,j} \ge 0, & \text{for } 1 \le i, j \le K \text{ with } i \ne j. \end{cases} \tag{3}$$

to be a valid generator matrix. The criteria in (3) lead to

$$Q_{i,i} \le 0, \qquad \text{for } 1 \le i \le K$$

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

$$G_i = -k_B T \log(\pi_i)$$

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, $\Delta G_{i,j} = G(j) - G(i)$, can be calculated from $\mathbf{Q}$ using the detailed balance relation:

$$\Delta G_{i,j} = -k_B T \log\left(\frac{Q_{i,j}}{Q_{j,i}}\right) \tag{4}$$

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 $\mathbf{Q}$ is tridiagonal. For a diffusive process, the generator matrix can be used to determine the diffusion constant as[33]

$$D_{i\leftrightarrow i+1} = \sqrt{Q_{i+1,i} Q_{i,i+1}} \tag{5}$$

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

$$D\left(\frac{x_i + x_{i+1}}{2}\right) \approx D_{i\leftrightarrow i+1}(x_{i+1} - x_i)^2$$

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

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

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, with an effective infinite free energy at x_0). Lifson and Jackson[47] and others[48] have shown the MFPT can be calculated using

$$\mathrm{MFPT}_{R\to P} = \int_{x_R}^{x_P} \frac{1}{D(x)} \exp(G(x)/k_B T) \left(\int_{x_0}^{x} \exp(-G(y)/k_B T)\, dy\right) dx \tag{6}$$

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

$$\mathrm{MFPT}_{R\to P} = \sum_{i=i_R}^{i_P} \frac{1}{D_i} \exp(G_i/k_B T) \sum_{j=1}^{i} \exp(-G_j/k_B T) \tag{7}$$

where i_P and i_R are the reactant and product states.

For a given empirical transition matrix, it has been shown that with a sufficient but not necessary condition,[49] an exact generator matrix exists such that $Q = \frac{\log(P)}{\tau}$ as in (2). However, the resulting $\mathbf{Q}$ 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 undersampling certain transitions. Such behavior in empirical transition matrices would lead to empirical transition matrices that do not satisfy the criteria in (3). One may even calculate negative eigenvalues for the empirical transition matrix, which would result in a nonreal generator matrix. In Section 3 below, we present several methods to address this problem.

3. ALGORITHMS

We present eight algorithms to estimate the rate matrix. These algorithms range from very simple methods such as the Diagonal Adjustment algorithm to more intricate methods such as Expectation Maximization and Maximum Likelihood Estimation. Besides using existing algorithms suggested in related literature for comparison, we also introduce a novel algorithm in order to find an estimator $\tilde{\mathbf{Q}}$ for the generator matrix $\mathbf{Q}$. The procedure is called Polynomial Adjustment

https://doi.org/10.1021/acs.jpca.3c01367
J. Phys. Chem. A 2023, 127, 5745−5759

(PA) algorithm and is based on the common eigenvector structure of $\tilde{\mathbf{Q}}$ and the estimated transition matrix $\mathbf{P}$. A comparison of these algorithms can be found in Section 5.

3.1. Diagonal Adjustment (DA). The Diagonal Adjustment (DA)[31] algorithm is a simple and sometimes effective solution to the embeddability problem. The DA algorithm adjusts the generator matrix in two steps after using Rel. (2) to estimate the $\mathbf{Q}$ matrix from the $\mathbf{P}$ matrix:

$$\text{Step 1: } \tilde{Q}_{i,j}^{DA} = \begin{cases} 0, & \text{if } i \neq j \text{ and } Q_{i,j} < 0, \\ Q_{i,j}, & \text{otherwise.} \end{cases}$$

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

where $\tilde{Q}_{i,j}^{DA}$ denotes the elements of an estimated generator matrix obtained by the DA algorithm.

3.2. Weighted Adjustment (WA). The Weighted Adjustment (WA)[31] is another simple algorithm very similar to the DA. The algorithm adjusts the generator matrix in two steps:

$$\text{Step 1: } \tilde{Q}_{i,j} = \begin{cases} 0, & \text{if } i \neq j \text{ and } Q_{i,j} < 0, \\ Q_{i,j}, & \text{otherwise.} \end{cases}$$

$$\text{Step 2: } \tilde{Q}_{i,j}^{WA} = \tilde{Q}_{i,j} - |\tilde{Q}_{i,j}| \frac{\sum_{k=1}^{K} \tilde{Q}_{i,k}}{\sum_{k=1}^{K} |\tilde{Q}_{i,k}|},$$

where $\tilde{Q}_{i,j}^{WA}$ denotes the elements of an estimated generator matrix obtained by the WA algorithm.

3.3. Quasi-optimization of the Generator (QOG). DA and WA are very similar methodologies. Unfortunately, they are not based upon an optimization strategy and become increasingly hard to trust in sparse data situations due to their unphysicality. To this end, Krenin and Sidelnikova[50] have extended the above work by implementing a postadjustment 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 $\log(\mathbf{P})$ and $\mathbf{Q}\tau$. Thus, we can write the problem as

$$\tilde{Q}^{QOG} = \arg\min_{\mathbf{Q}} \|\mathbf{Q}\tau - \log(\mathbf{P})\|$$

where $\|\cdot\|$ denotes the so-called Frobenius norm defined as $\|\mathbf{A}\|^2 = \sum_{i,j=1}^{K} A_{i,j}^2$ for a real-valued matrix $\mathbf{A} = (A_{i,j})_{i,j=1,\ldots,K}$. By reducing the problem to a distinct problem for each row, we can define

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

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

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

3.4. Component-Wise Optimization (CWO). Component-Wise Optimization (CWO)[38] is a somewhat more complex version of DA or WA which bears resemblance to QOG. The general idea is to divide the problem into $K - 1$ separate optimization problems. First, a DA or WA optimization is performed to find an initial estimator $\tilde{\mathbf{Q}}$. The first step is followed by optimizing each individual value in the generator according to

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

where $\tilde{\mathbf{Q}}$ refers to all values of the matrix $\tilde{\mathbf{Q}}$ being fixed except for the element $Q_{i,j}$. To ensure that the generator matrix $\tilde{\mathbf{Q}}$ remains valid, one must adjust the elements in the ith row of $\tilde{\mathbf{Q}}$ in the same step when changing an element $Q_{i,j}$. The constant c determines the desired convergence; the smaller the c, the faster the convergence. We defined c as 0.0001 and used DA algorithm to an initial $\tilde{Q}$ in our work. 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.[38]

3.5. Expectation Maximization (EM). The Expectation Maximization (EM)[31] algorithm is based on iterating two steps, the Expectation-step and the Maximization-step. To be more precise, recall that $N_{i,j}$ denotes the number of transitions from state i to j. We further write R_i for the number of time steps the system stays in state i. Then, write

$$\mathbb{E}[R_i(\tau)] = \frac{1}{M} \sum_{\alpha} \frac{1}{d_\alpha} e_{i_\alpha}^{\mathrm{T}} \int_0^\tau \exp(\mathbf{Q}s)(e_i e_i^{\mathrm{T}}) \exp(\mathbf{Q}(\tau - s)) \, \mathrm{d}s \, e_{j_\alpha} \tag{8}$$

$$\mathbb{E}[N_{i,j}(\tau)] = \frac{1}{M} \sum_{\alpha} \frac{1}{d_\alpha} e_{i_\alpha}^{\mathrm{T}} Q_{i,j} \int_0^\tau \exp(\mathbf{Q}s)(e_i e_i^{\mathrm{T}}) \exp(\mathbf{Q}(\tau - s)) \, \mathrm{d}s \, e_{j_\alpha} \tag{9}$$

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. Furthermore, e_i denotes a unit vector with the i'th element being one and the rest of elements being zero, we write $e^{\mathbb{T}}$ for the transpose of e. The quantity d_α is defined by

$$d_\alpha = e_{i_\alpha}^{\mathrm{T}} \exp(-\mathbf{Q}\tau) e_{j_\alpha}$$

The relations (8) and (9) allow us to estimate the expected values of $N_{i,j}$ and R_i as a function of the generator matrix $\mathbf{Q}$. The generator matrix is then estimated as

$$\tilde{Q}_{i,j}^{EM} = \frac{\mathbb{E}[N_{i,j}(\tau)]}{\mathbb{E}[R_i(\tau)]} \tag{10}$$

The procedure of the EM is done by calculating (8) and (9) for each element, and then using (10) to construct a new generator matrix. The iteration proceeds until convergence completes the algorithm.

3.6. Maximum Likelihood Estimator (MLE). A common optimization approach is to use 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] The relation,

https://doi.org/10.1021/acs.jpca.3c01367
J. Phys. Chem. A 2023, 127, 5745−5759

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

which leads to

$$\log(L) = \sum_{i,j=1}^{K} N_{i,j} \log((\exp(\mathbf{Q}\tau))_{i,j})$$

in which the off-diagonal elements of $\mathbf{Q}$ matrix are varied but stay positive by construct and the diagonal elements are determined by

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

Other criteria can be easily imposed on $\mathbf{Q}$ 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

$$\tilde{Q}_{i,j}^{MLE} = \begin{cases} 0, & \text{if } |i-j| > 1, \\ Q_{i,j}, & \text{if } |i-j| = 1, \\ -Q_{i,i-1} - Q_{i,i+1}, & \text{if } i = j. \end{cases}$$

3.7. Quadratic Programming (QP). A quadratic programming approach was proposed by D. T. Crommelin and E. Vanden-Eijnden[39] to take into account the eigen-structures of the transition probability matrix $\mathbf{P}$ and the generator matrix $\mathbf{Q}$. Indeed, if $\mathbf{P}$ has the eigendecomposition

$$\mathbf{P} = \sum_{i=1}^{K} \Lambda_i \phi_i \psi_i$$

where $\mathbf{P}\,\phi_i = \Lambda_i\,\phi_i$, $\psi_i\,\mathbf{P} = \Lambda_i\,\psi_i$, $\Lambda_i \in \mathbb{C}$, $\forall\, i = 1, ..., K$, then with $\lambda_i = \tau^{-1}\log(\Lambda_i)$,

$$\tau^{-1}\log(\mathbf{P}) = \sum_{i=1}^{K} \lambda_i \phi_i \psi_i \tag{11}$$

Note that $(\phi_i, \psi_i, \Lambda_i, \lambda_i)$, for $1 \le i \le K$, can be complex-valued.

Since the generator matrix $\mathbf{Q}$ has a similar eigen-structure as $\tau^{-1}\log(\mathbf{P})$ given in (11), an estimator $\tilde{\mathbf{Q}}$ can be obtained by solving the following optimization problem

$$\tilde{\mathbf{Q}}^{QP} = \arg\min_{\mathbf{Q}\in\mathbb{D}} \sum_{i=1}^{K} (\alpha_i \|\mathbf{Q}\phi_i - \lambda_i\phi_i\|^2 + \beta_i \|\psi_i\mathbf{Q} - \lambda_i\psi_i\|^2 + \gamma_i |\psi_i\mathbf{Q}\phi_i - \lambda_i|^2) \tag{12}$$

where $\alpha_i, \beta_i, \gamma_i$ for $i = 1, ..., K$ are positive weights chosen to stabilize the algorithm numerically, and domain $\mathbb{D}$ is a subspace of $\mathbb{R}^{K\times K}$ defined by (3). The norm-operators and the absolute values in (12) are compatible with complex cases. Note that the objective function in (12) is a quadratic function of the entries $\mathbf{Q}_{i,j}$ for all $1 \le i, j \le K$, and (3) only imposes linear constraints to the domain space $\mathbb{D}$, the problem (12) can thus be well handled by quadratic programming.[51]

3.8. Polynomial Adjustment (PA). We propose here a novel routine of Polynomial Adjustment (PA) that not only meets the generator matrix constraints defined in (3) but also strictly maintains the eigenvectors structure.

The generator matrix $\mathbf{Q}$ can be expressed as a polynomial of the transition probability matrix $\mathbf{P}$ with order at most $K-1$, or equivalently, the vectorization of the generator matrix can be given by a linear combination of the column vectors of the matrix

$$\mathbf{S} = (\mathrm{vec}(I_K), \mathrm{vec}(\mathbf{P}), ..., \mathrm{vec}(\mathbf{P}^{K-1})) \in \mathbb{R}^{K^2\times K}$$

with $vec(\cdot)$ the operator that stacks all the columns of a matrix, and $\mathbf{I}_K$ the $K \times K$ identity matrix. A singular value decomposition (SVD) separates a matrix into an ordered sum of matrices. The SVD of $\mathbf{S}$ yields nonincreasing singular values $\lambda_1, ..., \lambda_K$ and corresponding left singular eigenvectors $u_1, ..., u_K \in \mathbb{R}^{K^2}$, while r is the index that satisfies $\lambda_r > \varepsilon \ge \lambda_{r+1}$ (set $\lambda_{K+1} := 0$) for a given threshold $\varepsilon \ge 0$ for numeric stability. It is then possible to parametrize the adjustment $\tilde{\mathbf{Q}}$ as

$$vec(\tilde{\mathbf{Q}}^{PA}) = \tilde{\mathbf{S}}x,\ x \in \mathbb{R}^r \tag{13}$$

with $\tilde{\mathbf{S}} = (\lambda_1\, u_1, ..., \lambda_r\, u_r)$ the column space basis of $\mathbf{S}$. With $\mathbf{Q}$ based on the power series expression of matrix logarithm of $\mathbf{P}$, the minimizer $\hat{x}$ in the optimization problem

$$\hat{x} = \arg\min_{x\in\mathbb{R}^r} \left\| \tilde{\mathbf{S}}x - \mathrm{vec}(\mathbf{Q}) \right\|^2 \text{ subject to } \begin{cases} (\mathbf{1}_K \otimes e_1)\tilde{\mathbf{S}}x = 0, \\ (I_K \otimes I_K - 2\,\mathrm{diag}(\mathrm{vec}(I_K)))\tilde{\mathbf{S}}x \ge 0 \end{cases} \tag{14}$$

is available from quadratic programming with all linear constraints.[51] Here $\otimes$ is the Kronecker product, $\mathbf{1}_K = (1, ..., 1) \in \mathbb{R}^K$, $e_1 = (1, 0, ..., 0) \in \mathbb{R}^K$, $\mathrm{diag}(\cdot)$ generates a diagonal matrix, and the last inequality holds entry-wise. The adjusted generator matrix $\tilde{\mathbf{Q}}^{PA}$ can be obtained by reshaping the vector $\tilde{\mathbf{S}}\,\hat{x}$ according to (13). After this optimization, some off-diagonal elements of $\tilde{\mathbf{Q}}^{PA}$ can be negative. As a final step, the negative values are set to zero to meet the criteria set in (3).

Though QP and PA both use the property of shared eigen-structures and incorporate the quadratic programming toolkit, the two methods still differ in many ways. First, the QP algorithm explicitly computes the possibly complex eigendecompositions of $\mathbf{P}$ and $\mathbf{Q}$, while PA avoids doing it by using a different idea of parametrization. Second, the QP algorithm is supposed to find $\mathbf{Q}$ with eigenvectors that are sufficiently close to those of $\mathbf{P}$, while PA forces the eigenvectors of $\mathbf{P}$ and $\mathbf{Q}$ to be strictly the same. In addition, due to the different parametrizations, PA also reduces the dimension of the search domain to $O(K)$, compared to $O(K^2)$ in QP.

4. 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[52] method with parameters: temperature $T = 298$ K, 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

$$U(x) = \frac{k}{4}(x^2 - 1)^2 \tag{15}$$

where $k = 10\ \frac{\mathrm{kcal}}{\mathrm{mol}}$. To perform the sampling, we bin the space between $x = -2$ and 2 using the bin width $\Delta x = 4/47$, resulting in a total of 48 discrete states. Consequently, we simulate an independent trajectory of length 10^9 time steps starting from each bin, generating 48 independent trajectories. 302

https://doi.org/10.1021/acs.jpca.3c01367
J. Phys. Chem. A 2023, 127, 5745−5759

Table 1. Barrier Height ($\Delta G = G(0) - \frac{G(-1) + G(1)}{2}$) Based on Different Free Energy Predictions Obtained from a Theoretical Ideal Generator Matrix and Full and Individual Data Sets[a]

τ	1	5	10	50	100	500	1000	
DA	2.46	2.46	2.45	2.43	2.41	2.34	2.31	theoretical
	2.48	2.47	2.46	2.44	2.43	2.35	2.31	full data set
	2.48 ± 0.02	2.47 ± 0.02	2.47 ± 0.02	2.44 ± 0.01	2.43 ± 0.01	2.36 ± 0.02	2.31 ± 0.01	individual data sets
WA	2.52	2.51	2.50	2.48	2.46	2.40	2.35	theoretical
	2.53	2.53	2.52	2.50	2.48	2.41	2.36	full data set
	2.54 ± 0.02	2.53 ± 0.02	2.53 ± 0.02	2.50 ± 0.02	2.48 ± 0.02	2.41 ± 0.02	2.36 ± 0.02	individual data sets
QOG	2.46	2.45	2.45	2.42	2.40	2.32	2.28	theoretical
	2.48	2.47	2.46	2.44	2.42	2.33	2.28	full data set
	2.48 ± 0.02	2.47 ± 0.02	2.47 ± 0.02	2.44 ± 0.01	2.42 ± 0.01	2.34 ± 0.01	2.28 ± 0.01	individual data sets
CWO	2.48	11.55	8.36	2.43	2.41	2.34	2.30	theoretical
	2.50	14.09	2.46	2.44	2.42	2.35	2.32	full data set
	2.50 ± 0.02	14.08 ± 0.07	2.47 ± 0.02	2.44 ± 0.01	2.43 ± 0.01	2.36 ± 0.01	2.31 ± 0.02	individual data sets
EM	2.46	2.44	2.43	2.39	2.35	2.20	2.12	theoretical
	2.48	2.48	2.48	2.48	2.49	2.49	2.48	full data set
	2.47 ± 0.12	2.47 ± 0.11	2.47 ± 0.11	2.47 ± 0.10	2.47 ± 0.09	2.48 ± 0.07	2.48 ± 0.07	individual data sets
MLE	2.48	2.47	2.47	2.46	2.46	2.46	2.48	theoretical
	2.48	2.49	2.48	2.48	2.48	2.48	2.49	full data set
	2.49 ± 0.01	2.49 ± 0.01	2.49 ± 0.01	2.49 ± 0.01	2.49 ± 0.03	2.49 ± 0.02	2.49 ± 0.03	individual data sets
QP	2.47	2.47	2.47	2.48	2.48	2.51	2.54	theoretical
	2.49	2.49	2.49	2.50	2.51	2.54	2.49	full data set
	2.49 ± 0.01	2.49 ± 0.01	2.50 ± 0.01	2.50 ± 0.02	2.51 ± 0.02	2.92 ± 0.76	2.77 ± 0.50	individual data sets
PA	2.47	2.47	2.47	2.47	2.47	2.47	2.47	theoretical
	2.48	2.48	2.48	2.49	2.48	2.49	2.48	full data set
	2.49 ± 0.02	2.49 ± 0.02	2.49 ± 0.02	2.49 ± 0.02	2.49 ± 0.02	2.49 ± 0.02	2.49 ± 0.02	indivdual data sets
pyEmma	4.17	4.17	4.17	4.17	4.17	4.17	4.18	theoretical
	4.19	4.19	4.19	4.19	4.19	4.20	4.20	full data set
	3.68 ± 0.48	3.69 ± 0.48	3.71 ± 0.49	3.77 ± 0.50	3.82 ± 0.51	4.02 ± 0.55	4.17 ± 0.58	indivdual data sets

[a]The predictions are based on various estimates of the generator matrix including: DA, WA, QOG, CWO, EM, MLE, QP, and PA, as well as the direct use of transition matrices via the pyEmma software, for a given τ.

independent sets of simulations described above are then generated. Each set can be used independently to estimate the quantities of interest. Empirical transitions matrices are recorded in each simulation repeat, which were then used as the seed information for our rate matrix algorithms. In addition to individual sets, one may combine all sets to generate a single empirical transition matrix to be the basis of the same algorithms. We denote the combined data sets generated from all 302 individual data sets as the "full data set", which represents the abundance of data.

In order to test the efficacy of each algorithm we varied the lag time and trajectory lengths for the recorded transition matrix in individual data sets. Trajectory length was varied by cutting off our counting statistics for each copy at the appropriate point. For instance, a 10% trajectory length (denoted by $L = 10\%$) includes the first tenth 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 toward $N_{i,j}$. 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.

Each algorithm is used to estimate a generator matrix for an empirical transition matrix at each combination of τ and L for individual data sets and at each τ for the full data set. From the results of our algorithms for individual data sets, we calculate the average and standard deviation of several metrics for different lag times and trajectory lengths. Some algorithms generated errors for a given repeat or did not converge quickly enough. The number of such repeats in each algorithm is shown in the Supporting Information, Table S1.

While, the algorithms presented in this work are expected to work assuming Markovianity of the model, it is important to note that deviations from Markovianity are expected in many cases, particularly when a model based on continuous space is discretized. Specifically, our toy model described above would present an almost ideal Markovian space if very small bins are used for discretization. However, this work is more concerned with the application of the methodology in a nonideal case as is typical in realistic applications of MSM and molecular dynamics in general. Therefore, rather than a perfect Markovian model, we are working with a toy model that deviates from Markovianity assumption, precisely due to its discretization. While it is easy to theoretically or empirically estimate various thermodynamic and kinetic quantities associated with the continuous space model, it is more difficult to estimate such quantities for its discretized space model, which is not fully Markovian. To overcome this issue, we use a strategy that relies on the Markovianity assumption for a discretized space model with a much smaller bin width as compared to that used above (namely, 1000 times smaller). We start by building an ideal generator matrix ($\mathbf{Q}$) of size 47,000 × 47,000. We employ relations (15), (4), and (5), by using $G(x) = U(x)$ and $D = k_B T/m\gamma = 0.59$:

https://doi.org/10.1021/acs.jpca.3c01367
J. Phys. Chem. A 2023, 127, 5745−5759

Table 2. Well Symmetry, or the Absolute Errors in $\Delta G = |G(1) - G(-1)|$, the Free Energy Difference between the Two Minima Obtained from a Theoretical Ideal Generator Matrix and Full Data Set[a]

τ	1	5	10	50	100	500	1000	
DA	0.05	0.05	0.04	0.05	0.04	0.04	0.04	theoretical
	0.04	0.04	0.04	0.04	0.04	0.03	0.04	full data set
	0.04 ± 0.03	0.04 ± 0.03	0.04 ± 0.03	0.04 ± 0.03	0.04 ± 0.03	0.04 ± 0.03	0.04 ± 0.03	individual data sets
WA	0.06	0.06	0.06	0.06	0.06	0.06	0.06	theoretical
	0.06	0.06	0.06	0.06	0.06	0.06	0.05	full data set
	0.06 ± 0.03	0.06 ± 0.03	0.06 ± 0.03	0.06 ± 0.03	0.06 ± 0.03	0.06 ± 0.03	0.06 ± 0.03	individual data sets
QOG	0.05	0.04	0.04	0.05	0.04	0.05	0.05	theoretical
	0.04	0.04	0.04	0.04	0.04	0.04	0.05	full data set
	0.04 ± 0.03	0.04 ± 0.03	0.04 ± 0.03	0.04 ± 0.03	0.04 ± 0.03	0.04 ± 0.03	0.04 ± 0.03	individual data sets
CWO	0.09	2.21	0.04	0.05	0.04	0.04	0.05	theoretical
	0.09	0.31	0.04	0.04	0.04	0.04	0.04	full data set
	0.09 ± 0.03	0.33 ± 0.14	0.04 ± 0.03	0.04 ± 0.03	0.04 ± 0.03	0.04 ± 0.03	0.04 ± 0.03	individual data sets
EM	0.05	0.05	0.05	0.05	0.04	0.04	0.04	theoretical
	0.04	0.04	0.04	0.04	0.04	0.04	0.05	full data set
	0.04 ± 0.03	0.04 ± 0.03	0.04 ± 0.03	0.04 ± 0.03	0.04 ± 0.03	0.04 ± 0.03	0.04 ± 0.03	individual data sets
MLE	0.06	0.06	0.04	0.03	0.04	0.04	0.04	theoretical
	0.04	0.05	0.05	0.04	0.03	0.04	0.05	full data set
	0.05 ± 0.03	0.05 ± 0.03	0.05 ± 0.03	0.05 ± 0.03	0.05 ± 0.06	0.06 ± 0.04	0.06 ± 0.05	individual data sets
QP	0.05	0.05	0.05	0.05	0.05	0.04	0.04	theoretical
	0.04	0.04	0.04	0.04	0.04	0.03	0.05	full data set
	0.04 ± 0.03	0.04 ± 0.03	0.05 ± 0.03	0.05 ± 0.03	0.05 ± 0.03	0.05 ± 0.03	0.08 ± 0.06	individual data sets
PA	0.05	0.05	0.05	0.05	0.05	0.05	0.04	theoretical
	0.04	0.04	0.04	0.04	0.04	0.04	0.05	full data set
	0.04 ± 0.03	0.04 ± 0.03	0.05 ± 0.03	0.04 ± 0.03	0.04 ± 0.03	0.04 ± 0.03	0.04 ± 0.03	individual data sets
pyEmma	0.0	0.0	0.0	0.0	0.0	0.0	0.0	theoretical
	0.01	0.01	0.01	0.01	0.01	0.01	0.01	full data set
	0.11 ± 0.08	0.12 ± 0.09	0.14 ± 0.10	0.20 ± 0.15	0.24 ± 0.19	0.45 ± 0.35	0.59 ± 0.46	individual data sets

[a]The predictions are based on various estimates of the generator matrix including: DA, WA, QOG, CWO, EM, MLE, QP, and PA, as well as the direct use of transition matrices via the pyEmma software for a given τ.

$$Q_{i,j} = \begin{cases} D_{i\leftrightarrow j} \exp((G_i - G_j)/2k_BT), & \text{if } i, j = i+1, \\ D_{i\leftrightarrow j} \exp(-(G_i - G_j)/2k_BT), & \text{if } i+1, j = i, \\ -\sum_{k=1}^{K} Q_{i,k}, & \text{if } i = j, \\ 0, & \text{otherwise.} \end{cases}$$

Subsequently, we compute the transition probability matrix or normalized transition matrix ($\mathbf{P}$) using $\mathbf{P}(\tau) = \exp(\mathbf{Q}\tau)$, for a given τ. Transition matrix ($\mathbf{N}$) is then constructed by denormalizing the matrix ($\mathbf{P}$) using Relation (1), while noting that $\Sigma_i N_{i,k} = \exp(-\beta U_k)$ holds, given that $\Sigma_i P_{ik} = 1$. Note that all elements of $\mathbf{N}$ can be multiplied by any given number without any effect in the subsequent steps of the process. Finally, we estimate a 47 × 47 transition matrix, representing the non-Markovian discretized model of bin width 4/47 from the $\mathbf{N}$ matrix above by summing the elements of each 1,000 × 1,000 block in $\mathbf{N}$, which represents the corresponding element in the reduced 47 × 47 transition matrix. This reduction step is justified by the fact that the number of transitions between any pair of bins that are divided into sub-bins is equal to the sum of the number of all transitions between all sub-bins ($N_{ij} = \sum_{I,J} N_{IJ}$, where I and J represent any sub-bin of i and j, respectively). The resulting 47 × 47 transition matrix is then utilized to construct a normalized 47 × 47 transition matrix, which is in turn used to estimate rate matrices using various algorithms, namely DA, WA, QOG, CWO, EM, MLE, QP, and PA. This approach provides a theoretical estimate for metric values under the assumption of Markovian behavior. Note that the 47 × 47 normalized transition matrix above truly approximates the expected normalized transition matrices that we obtain from the simulations without assuming Markovianity for the discretized space model of bin width 4/47 but only relying on the Markovianity of the discretized space model of bin width 4/47,000.

5. RESULTS

We used the data generated for the toy model to estimate the generator matrix for various lag times using the algorithms (DA, WA, QOG, CWO, EM, MLE, QP, and PA) discussed above, and compared the performance of these algorithms under two scenarios: (i) full data set, and (ii) individual data sets, for a given lag time. Our investigation involved a comparison of multiple metrics, each offering unique insights into the algorithms' performance and convergence. We examined the free energy profile, barrier height (defined as $G(0) - \frac{G(1)+G(-1)}{2}$), well symmetry (defined as $|G(1) - G(-1)|$) as well as the average diffusion constant and MFPT at different lag times. Moreover, to underscore the significance of embeddability approaches, we compared our algorithmic predictions with more conventional methods that rely on empirically generated transition matrices using the spectral decomposition techniques. To accomplish this, we employed the pyEmma software,[53] which, for instance, calculates the first

https://doi.org/10.1021/acs.jpca.3c01367
J. Phys. Chem. A 2023, 127, 5745−5759

Table 3. Average Diffusion Constant $D(x)$ Estimated from D_i Values over the Range $x = -2$ to $x = 2$ Obtained from a Theoretical Ideal Generator Matrix and Full Data Set[a]

τ	1	5	10	50	100	500	1000	
DA	36.76	16.40	11.57	5.12	3.59	1.56	1.07	theoretical
	36.52	14.57	10.90	5.03	3.55	1.55	1.06	full data set
	36.51 ± 0.03	14.57 ± 0.01	10.90 ± 0.01	5.04 ± 0.00	3.57 ± 0.00	1.63 ± 0.00	1.22 ± 0.00	individual data sets
WA	36.84	16.51	11.69	5.26	3.73	1.70	1.21	theoretical
	36.79	14.74	11.07	5.19	3.71	1.69	1.17	full data set
	36.79 ± 0.03	14.74 ± 0.01	11.07 ± 0.01	5.20 ± 0.00	3.72 ± 0.00	1.77 ± 0.00	1.33 ± 0.00	individual data sets
QOG	36.77	16.41	11.58	5.13	3.61	1.57	1.08	theoretical
	36.53	14.58	10.91	5.05	3.56	1.56	1.08	full data set
	36.53 ± 0.03	14.58 ± 0.01	10.91 ± 0.01	5.06 ± 0.00	3.58 ± 0.00	1.64 ± 0.00	1.23 ± 0.00	individual data sets
CWO	36.95	58.22	22.14	5.38	3.78	1.64	1.12	theoretical
	36.69	37.41	11.44	5.29	3.73	1.62	1.11	full data set
	36.69 ± 0.03	37.40 ± 0.03	11.44 ± 0.01	5.29 ± 0.00	3.74 ± 0.00	1.70 ± 0.00	1.27 ± 0.00	individual data sets
EM	37.04	16.68	11.85	5.41	3.89	1.86	1.39	theoretical
	36.80	14.80	11.15	5.32	3.84	1.84	1.38	full data set
	36.80 ± 0.03	14.80 ± 0.01	11.16 ± 0.01	5.32 ± 0.00	3.85 ± 0.00	1.92 ± 0.00	1.53 ± 0.00	individual data sets
MLE	37.02	16.68	11.85	5.42	3.89	1.86	1.39	theoretical
	36.79	14.79	11.15	5.32	3.84	1.85	1.38	full data set
	36.79 ± 0.03	14.79 ± 0.01	11.16 ± 0.01	5.32 ± 0.00	3.85 ± 0.00	>100.0	>100.0	individual data sets
QP	37.16	16.80	11.98	5.55	4.03	2.08	1.68	theoretical
	36.92	14.89	11.27	5.45	3.99	2.05	1.37	full data set
	36.91 ± 0.03	14.89 ± 0.01	11.27 ± 0.01	5.46 ± 0.01	4.00 ± 0.00	2.13 ± 0.01	1.76 ± 0.05	individual data sets
PA	37.04	16.67	11.83	5.38	3.86	1.87	1.41	theoretical
	36.67	14.76	11.19	5.57	4.17	2.09	1.64	full data set
	36.73 ± 0.21	15.11 ± 0.83	12.19 ± 2.18	8.82 ± 3.26	6.41 ± 1.89	1.71 ± 0.02	1.86 ± 0.05	individual data sets

[a]The predictions are based on various estimates of the generator matrix including: DA, WA, QOG, CWO, EM, MLE, QP, and PA, for a given τ. Outliers were excluded from the individual data sets when calculating the averages and standard deviations reported in the table, using the interquartile range (IQR) method.

eigenvector of the transition matrix as the stationary distribution.

5.1. Comparative Analysis of the Algorithms. We begin with an extensive evaluation of the algorithmic methods, aiming to unravel their performance when presented with the full and individual data sets, all with full length (L = 100%), across various lag times. To account for the non-Markovianity, we have also calculated theoretical values of the quantities estimated by reducing a large, ideal generator matrix, as described in the toy model section. These theoretical values, denoted as "theoretical", were included alongside the results obtained from the full and individual data sets. These theoretical values serve as our reference points, representing the ground truth for comparison with the other outcomes.

The results obtained from the full and individual data sets and theoretical values generally aligned well for all the metrics analyzed. Specifically, focusing on the barrier height (Table 1), all algorithms provided reasonable predictions, albeit with slight deviations form the analytical value (2.5). The barrier height estimates for DA, WA, QOG, and CWO decreased as τ increased for the theoretical, full and individual data sets, with QOG performing the worst when τ was largest. The CWO results at τ = 5 exhibited an unexpectedly large deviation, which was consistently observed across all analyzed metrics, and even τ = 10. The QP estimates were accurate for τ < 1000, but slightly overestimated at τ = 1000. Notably, the EM estimates showed discrepancies between the theoretical values and both the full and individual data sets, particularly for large τ values. On the other hand, MLE and PA consistently provided the most precise estimates across all τ values, with PA demonstrating the best overall performance. Interestingly, the outcomes obtained from the pyEmma software consistently yielded significantly higher values compared to our algorithmic methods.

Well symmetry, another measure of predictive accuracy, was generally low for all algorithms, as shown in Table 2, with exceptions for CWO at τ = 5. Among the algorithms, the CWO algorithm displayed the highest level of asymmetry at τ = 0.09. While all methods exhibited a noteworthy consistency between the theoretical values and both the full and individual data sets, the QP estimates at τ = 1000 showcased the greatest level of inconsistency across the individual data sets, theory, and full data sets. The DA, MLE, and QP algorithms demonstrated the greatest well symmetry, with an absolute error of 0.03 observed across various lag times. For most algorithms, the well symmetry remained relatively constant, hovering around 0.05 for DA, QOG, EM, QP, and PA and around 0.06 for WA. The theoretical and full data sets values obtained from the direct use of transition matrices via pyEmma software were close to zero, whereas the individual data sets exhibited the largest values. This indicates the high sensitivity of the spectral decomposition based methods to the amount of data.

We 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, assuming a continuous space, is 34.3. This value is estimated from relation (6), using $G(x) = U(x)$ (from relation (15)) and $D = k_B T/m\gamma$ = 0.59. The integrals in (6) were evaluated numerically, with the first integral taken over the range of −1 to 1, and the second integral taken from −5 (a point further

https://doi.org/10.1021/acs.jpca.3c01367
J. Phys. Chem. A 2023, 127, 5745−5759

Table 4. MFPT Predicted Using Various Estimates of the Generator Matrix, Including DA, WA, QOG, CWO, EM, MLE, QP, and PA, as well as the Direct Use of Transition Matrices via the pyEmma Software[a]

τ	1	5	10	50	100	500	1000	
DA	0.55	1.21	1.71	3.75	5.25	11.15	15.66	theoretical
	0.54	1.35	1.78	3.77	5.25	11.14	15.38	full data set
	0.54 ± 0.01	1.36 ± 0.04	1.80 ± 0.05	3.79 ± 0.10	5.25 ± 0.15	10.64 ± 0.27	13.45 ± 0.36	individual data sets
WA	0.57	1.27	1.78	3.84	5.30	10.78	14.17	theoretical
	0.61	1.51	1.99	4.13	5.66	11.33	15.55	full data set
	0.62 ± 0.02	1.52 ± 0.04	2.01 ± 0.05	4.15 ± 0.11	5.67 ± 0.15	10.87 ± 0.27	13.57 ± 0.35	individual data sets
QOG	0.55	1.21	1.70	3.72	5.16	10.69	14.80	theoretical
	0.54	1.34	1.78	3.72	5.16	10.77	14.56	full data set
	0.54 ± 0.01	1.35 ± 0.03	1.79 ± 0.05	3.74 ± 0.10	5.17 ± 0.14	10.26 ± 0.26	12.80 ± 0.33	individual data sets
CWO	0.62	>100.0	1.71	3.57	4.98	10.59	14.63	theoretical
	0.61	>100.0	1.16	3.57	5.00	10.57	14.84	full data set
	0.62 ± 0.02	>100.0	1.18 ± 0.04	3.53 ± 0.11	5.00 ± 0.14	10.15 ± 0.25	12.89 ± 0.34	individual data sets
EM	0.54	1.17	1.63	3.37	4.52	7.93	9.73	theoretical
	0.54	1.34	1.78	3.73	5.18	10.76	14.27	full data set
	0.55 ± 0.02	1.36 ± 0.04	1.80 ± 0.05	3.76 ± 0.10	5.20 ± 0.15	10.36 ± 0.27	12.91 ± 0.36	individual data sets
MLE	0.55	1.20	1.71	3.76	5.15	10.84	14.74	theoretical
	0.54	1.34	1.78	3.73	5.15	10.67	14.11	full data set
	0.55 ± 0.02	1.36 ± 0.04	1.81 ± 0.06	3.77 ± 0.12	5.06 ± 0.27	1.48 ± 0.22	0.96 ± 0.24	individual data sets
QP	0.55	1.21	1.70	3.98	4.48	10.86	14.47	theoretical
	0.54	1.35	1.79	3.74	5.17	10.56	14.29	full data set
	0.54 ± 0.01	1.35 ± 0.03	1.79 ± 0.05	3.76 ± 0.11	5.20 ± 0.16	10.97 ± 2.96	15.79 ± 8.17	individual data sets
PA	0.55	1.21	1.71	3.76	5.29	10.93	14.71	theoretical
	0.54	1.35	1.78	3.59	4.80	9.58	12.10	full data set
	0.54 ± 0.01	1.33 ± 0.07	1.69 ± 0.22	2.64 ± 0.96	3.49 ± 1.19	11.58 ± 0.36	10.71 ± 0.39	individual data sets
pyEmma	0.54	1.20	1.70	3.80	5.37	12.01	16.79	theoretical
	0.55	1.37	1.83	3.93	5.53	12.33	17.26	full data set
	0.34 ± 0.08	0.85 ± 0.20	1.14 ± 0.28	2.57 ± 0.68	3.74 ± 1.04	9.83 ± 3.72	15.5 ± 6.91	individual data sets

[a]The predictions were obtained from a theoretical ideal generator matrix, full and individual data sets for a given τ. Outliers were excluded from the individual data sets when calculating the averages and standard deviations reported in the table, using the interquartile range (IQR) method.

than −1) to 1. An empirical estimate of the MFPT was also obtained as 33.2 by averaging the results of more than a million simulations, each initiated at $x = -1$ and ran until reaching $x = +1$. A histogram of the reaching times can be found in Supporting Information. To estimate the MFPT from the theoretical and the full data set, we employ relation (7), which requires an estimate for the free energies (G_i) as well as D_i. The average estimate for $D(x)$ for the algorithmically estimated generator matrices are given in Table 3. The estimates were drastically influenced by τ. Unlike some of the free energy estimates that tend to be more accurate for lower τ values, the diffusion constant estimates were more accurate for longer lag times. The DA and QOG algorithms generated the closest estimates to the analytical diffusion (∼0.59), using long lag times, for both the theoretical and full data sets, though they still overestimate by approximately 45% which indicates deviation from Markovinity. The MLE, QP, and PA algorithms were the least accurate estimators at the long lag times. With the exception of the MLE algorithm beyond $\tau > 100$, all algorithms gave results within the same order of magnitude.

MFPT estimates, based on D_i and G_i estimates, for the theoretical and full and individual data sets are shown in Table 4 for varying lag times. In comparison to the analytical value of 34.3 and the empirical value of 33.20, all estimates were lower than expected, which is the direct result of the overestimation of $D(x)$, which in turn is the result of deviation from Markovian behavior. Similar to $D(x)$, the estimates are more accurate at longer lag times. Once again, all the methods, except MLE for very large τ's, demonstrated agreement between the theoretical predictions (taking into account the non-Markovianity) and the full and individual data sets. Notably, when comparing the values obtained from the full and individual data sets with those acquired through the theoretical approach, the EM method generally yielded higher values, while the PA method yielded lower values. When considering the longest lag times, the MFPT estimates that were somewhat closer to the analytical/empirical value (34.3/33.2) were obtained from DA for the theoretical data, WA for the full data, and QP for the individual predictions. In contrast, MLE exhibited the least accurate predictions for the individual data sets, particularly at $\tau = 1000$. The values obtained from the pyEmma software were closer to the analytical and empirical values, off by a factor of 2, and accompanied by a large error. On the other hand, our generator matrix based algorithms generally produced low errors, with the exception of QP for large lag times.

5.2. Convergence Behavior of the Algorithms. In this section, we focus on convergence behavior of discussed algorithmic methods when confronted with individual data sets, with various trajectory lengths. Utilizing the data obtained from our toy model, we applied these algorithms to estimate the generator matrix at various lag times and trajectory lengths. Through a comparison of metrics such as the free energy profile and MFPT, we gained valuable insights into how these algorithms improved their estimates and approached convergence.

https://doi.org/10.1021/acs.jpca.3c01367
J. Phys. Chem. A 2023, 127, 5745−5759

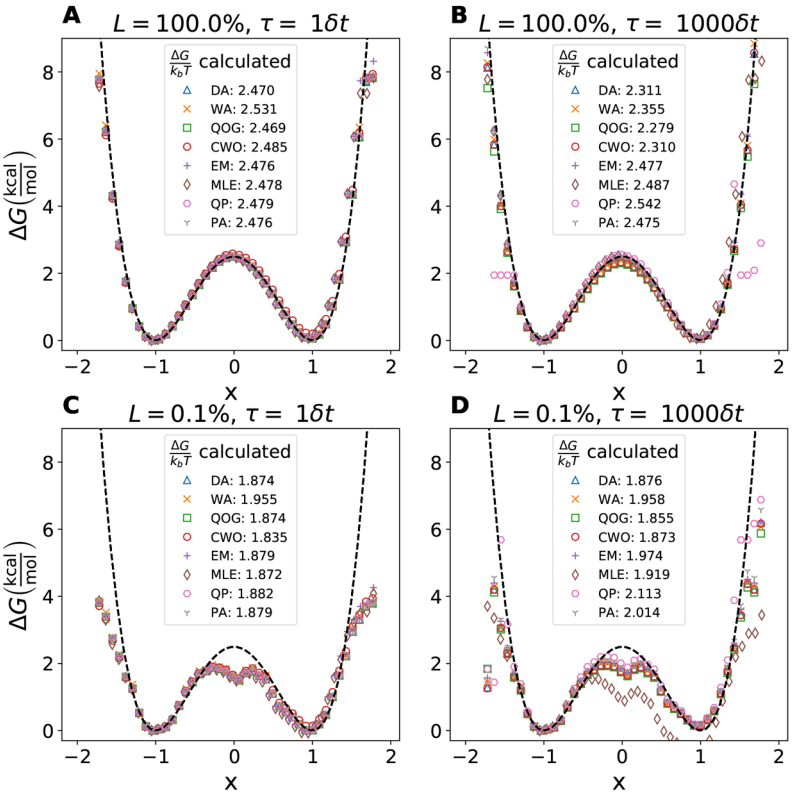


Figure 2. (A–D) Examples of the predicted free energy with a particular random seed as a function of x is shown for each algorithm using the extremes of trajectory length and τ. (A) L = 100% and $\tau = 1\delta t$, (B) L = 100% and $\tau = 1\delta t$, (C) L = 0.1% and $\tau = 1000\delta t$, and (D) L = 0.1% and $\tau = 1000\delta t$. Algorithms shown are DA (blue triangle), WA (orange ×), QOG (green square), CWO (red circle), EM (purple +), MLE (brown diamond), QP (pink hexagon), and PA (gray Y).

The estimated free energy for all algorithms has varying degrees of accuracy, with Figure 2 showing a single replication for the extremes of τ and L.

In the case of the complete trajectory and a short lag time of $\tau = 1\delta t$ (Figure 2A), all algorithms produced reasonable predictions, particularly in terms of their estimations of free energy. However, noticeable discrepancies from the model potential function were observed primarily when $|x| > 1.5$. At longer lag times in (Figure 2B), some algorithms began to break down where the free energy profiles have a $\tau = 1000\delta t$. For this iteration, the QP algorithm deviates at $|x| > 1.5$. For short trajectory lengths (Figure 2C,D), L = 0.1%, the algorithms behaved similarly with $\tau = 1\delta t$ and $\tau = 1000\delta t$, with reasonable well symmetry (with the exception of the MLE algorithm at $\tau = 1000\delta t$), but markedly less accurate estimates for the barrier and when $|x| > 1.5$.

To better summarize the barrier height data, the average values for the extremes of L and τ are shown in Figure 3. This more clearly demonstrates the observed τ-dependent behavior of barrier height of the free energy profile.

In Figure 3A, all models gave reasonable barrier height estimates at low τ with slight under-estimations, WA being the highest. CWO drastically overestimated the barrier height for $\tau = 5\delta t$, while QP overestimated at $\tau = 1000\delta t$. The QP estimates are reasonable for $\tau \leq 100\delta t$ but increased to unreasonably high estimates for the highest τ values. DA, WA, CWO and QOG estimates all decreased with increasing τ, where QOG had the worst barrier height estimate when τ was largest. The EM, MLE, and PA models were the most accurate and remained consistent with a slight underestimation of the barrier height for all τ values. For the short trajectory (Figure 3B), the general trend was for underestimation of the barrier height when τ was small, but the estimate increased toward the true barrier height value as τ increased. EM, MLE, and PA made the closest barrier height predictions when $\tau = 1000\delta t$. Again, CWO shows poor estimates for $\tau = 5\delta t$ and $\tau = 10\delta t$. For short lag time (Figure 3C), all the algorithms have qualitative similarities in that they underestimate the barrier height when $\tau = 1\delta t$ and for short trajectories and they improve as L increases with only small gains made after L > 10%. For longer lag times (Figure 3D), the average barrier height predictions are consistent across all trajectories, with EM, MLE, and PA being the most accurate. DA, WA, QOG, and CWO all underestimated the barrier for long lag times, while QP overestimated.

Average well symmetry results in Figure 4, were generally low for all algorithms with exceptions for CWO. When the full trajectory was used (Figure 4A), the average well symmetry was flat for most algorithms along the $\tau/\delta t$ axis, with the exceptions of CWO at $\tau = 5\delta t$ and $\tau = 10\delta t$ and to a lesser extent MLE and QP for longer lag times. $\tau > 100\delta t$. Also, the well symmetry from WA predictions were consistently higher than most other algorithms for all τ. For the shortest trajectory (Figure 4B), the general trend is less accurate well symmetry as lag times increased. The poor predictions in CWO at $\tau = 5\delta t$ and $\tau = 10\delta t$ were even worse for the shortest trajectory, while MLE is a poor predictor for $\tau = 1000\delta t$. With a short lag time of $\tau = 1\delta t$ (Figure 4C), the well symmetry estimates generally decreased with trajectory length, but all the algorithms show a peak at L = 50%. For this measure, the short τ is more sensitive

https://doi.org/10.1021/acs.jpca.3c01367
J. Phys. Chem. A 2023, 127, 5745–5759

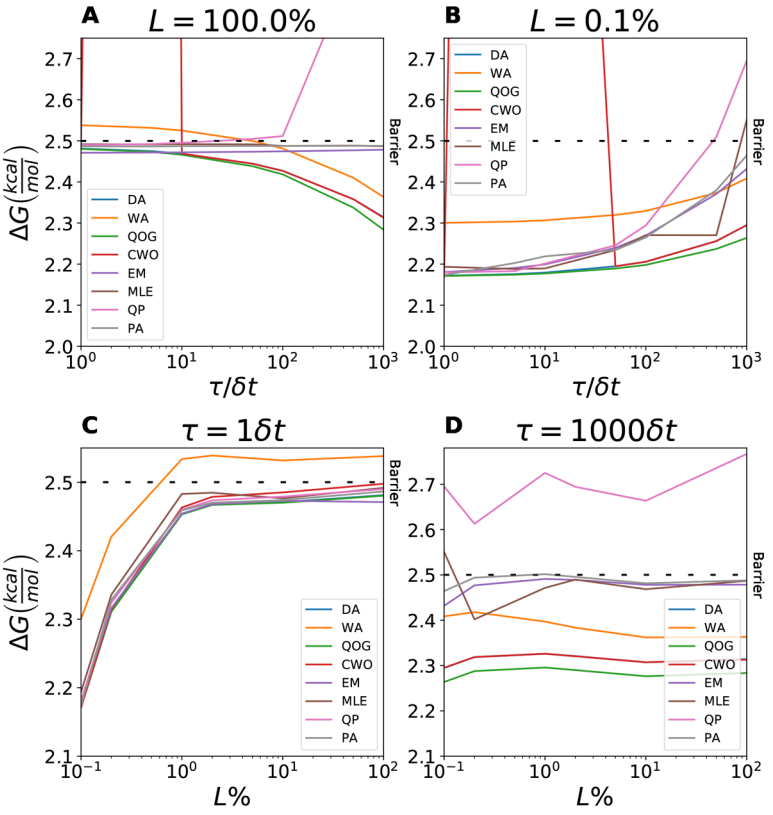


Figure 3. Average predicted barrier height as a function of τ for each algorithm, where (A) L = 100% and (B) L = 0.1%. Also, average predicted barrier height as a function of trajectory length for constant τ is shown for each algorithm, where (C) $\tau = 1\delta t$ and (D) $\tau = 1000\delta t$. The dashed line represents the analytic value for the well barrier of 2.5 (kcal/mol). The graph with error bars is given in Figure S2.

to any structure within trajectory data. The notable differences in algorithms are higher errors in CWO compared with the other algorithms for $L > 0.1\%$. MLE shows the greatest error for the shortest trajectory, $L = 0.1\%$. For long lag times of $\tau = 1000\delta t$ (Figure 4D), the errors were generally higher for all trajectory lengths. Most of the algorithms showed similar performance with decreasing error as the trajectory length increased. MLE has the highest well symmetry error at the shortest trajectory lengths, and QP performs the worst at $L = 100\%$. The increased lag times reduced sensitivity to structure in the data captured with generally monotonically decreasing error as L increases, except for the QP algorithm that has a small increase at $\tau = 50\delta t$.

Barrier height and well symmetry measures provide information on the accuracy of the free energy profile at specific points. The Kullback–Leibler divergence[54] (D_{KL}) quantifies the "distance" from the analytical free energy profile for all points. D_{KL} is a measure of difference between probability distributions, so it will include deviations from the free energy profile at the edges of the trajectory, near ±2. Figures of the evolution of D_{KL} with τ and L are shown in Figures S8–S10 along with a description of the metric. In general, most of the algorithms showed low D_{KL} values, and high values reflected the already observed errors from barrier height and well symmetry. CWO has high divergence at $\tau = 5\delta t$ and $\tau = 10\delta t$, while QP divergence increases at $\tau > 100\delta t$. Also, MLE divergence increases when trajectory length is short and at the longest lag time, $\tau = 1000\delta t$. It is also apparent that QOG showed small but notable improvements in divergences for long lag times, $\tau = 1000\delta t$, and short trajectory lengths, $L \leq 0.2\%$.

Figures with intermediary values for τ and L are in the Supporting Information (Figures S4–S9) and give a fuller picture of the evolution of the metrics mentioned above for the different algorithms. The average values in these figures as well as standard deviation is also presented in tabular form (Tables S2–S4) to further show the dependence of these algorithms on lag time and trajectory length as they summarize various estimates based on varying both quantities.

The average estimate for $D(x)$ (averaged over x for each replication) for the algorithmically estimated generator matrices are in the Supporting Information (Table S5). Overall, the estimates demonstrated greater sensitivity to τ rather than to L. Despite the scarcity of data, the estimates were relatively stable with the closest ones, specifically those produced by DA and QOG for the longest lag time, being approximately 50% larger than the analytical diffusion. It is worth noting that this difference remained consistent at around 45% when the full data set was used. On the other hand, the MLE algorithm exhibited the least accurate estimations, particularly for longer lag times.

Average MFPT estimates, based on D_i and G_i estimates, for the all trajectory repeats and standard deviation presented as error are shown in Table 5 for varying lag times and trajectory lengths.

Across various percentages of trajectory length and lag time, all the estimates displayed a tendency toward convergence, characterized by low error bars. Remarkably, the DA, WA, and PA algorithms exhibited highly favorable behavior, yielding estimates that converged closely to the expected values with minimal deviation. In contrast, the CWO and QP algorithms displayed relatively poorer performance in terms of con-

https://doi.org/10.1021/acs.jpca.3c01367
J. Phys. Chem. A 2023, 127, 5745−5759

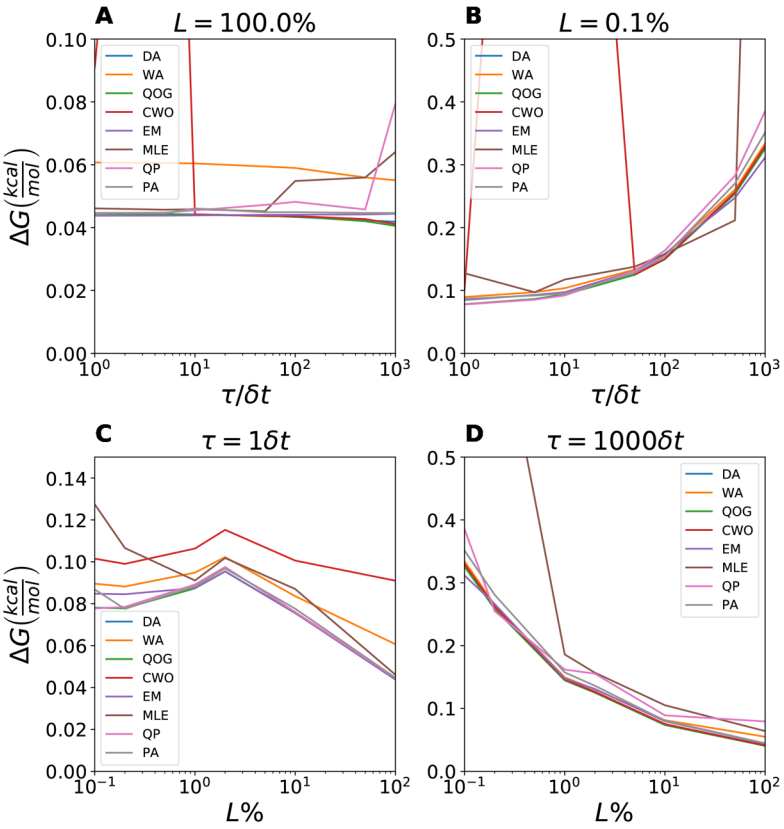


Figure 4. Average well symmetry as a function of τ for trajectory length is shown for each algorithm, where (A) L = 100% and (B) L = 0.1%. Well symmetry as a function of trajectory length for constant τ is shown for each algorithm, where (C) $\tau = 1\delta t$ and (D) $\tau = 1000\delta t$. The graph with error bars is given in Figure S3.

vergence. The MLE algorithm stood out by consistently generating unexpectedly low values specifically for τ = 1000, regardless of the trajectory length. Despite these variations, the overall convergence and relatively small error bars indicate the robustness of the estimation process across different algorithms and scenarios.

6. DISCUSSION

The MSM analysis often relies on the direct use of empirically generated transition matrices through spectral decomposition techniques. However, when it comes to continuous time models and biomolecular systems, an alternative method for extracting thermodynamic and kinetic information involves the use of generator matrices. In our study, we employed eight different estimators for generator matrices, each with its own advantages and disadvantages in terms of recovering the thermodynamic and kinetic information from the toy model. The estimations derived from both the full and individual data sets demonstrated a remarkable consistency with the theoretical predictions obtained by considering deviations from the Markovian behavior. However, there were deviations from the analytical values obtained from the Markovian continuous space model, in terms of kinetic parameters such as the diffusion constant and MFPT. These variations in behavior can be primarily attributed to the fact that our model is not completely Markovian, which is a common characteristic in molecular dynamics simulations, and the correlated nature of the data. The non-Markovianity and correlations inherent in the system introduced deviations both in the theoretical predictions and the data obtained from simulations. Nevertheless, it is worth noting the overall consistency observed across all algorithms, which suggests that the algorithms can capture important aspects of the dynamics, even when non-Markovian effects are present.

Regarding performance and convergence, the CWO method was a relatively poor free energy estimator as compared to the other methods used here, with sensitivity at specific low τ values. However, CWO performed similarly to other methods for estimating diffusion constant and MFPT. In contrast, the EM, MLE, and PA algorithms demonstrated robustness in estimating both thermodynamic and kinetic properties. They consistently provided accurate predictions of barrier heights for all lag times when utilizing the full trajectory. Furthermore, the diffusion constant and MFPT estimates obtained from these algorithms generally aligned with the estimates derived from the other methods, with the exception of MLE at long τ values. Notably, the PA algorithm, which builds upon the conceptual foundations of the QP, improved upon the long lag time errors of QP in free energy and MFPT estimations. It is important to remember that the PA algorithm was able to achieve these results while requiring significantly less computing time. The PA algorithm represents a viable approach, particularly for extracting embedded information from large transition matrices when computational resources are limited. Additionally, simpler algorithms such as DA and WA yielded marginally closer estimates, suggesting their potential usefulness in certain scenarios. Finally, our algorithms outperformed in thermodynamics metrics compared to directly using transition matrices and eigendecomposition techniques. This improvement can be

https://doi.org/10.1021/acs.jpca.3c01367
J. Phys. Chem. A 2023, 127, 5745−5759

Table 5. Average MFPT Predicted Using Various Estimates of the Generator Matrix Including DA, WA, QOG, CWO, EM, MLE, QP, and PA[a]

τ	1	10	100	1000	L (%)
DA	0.54 ± 0.01	1.80 ± 0.05	5.25 ± 0.15	13.45 ± 0.36	100.0
	0.54 ± 0.04	1.78 ± 0.14	5.21 ± 0.40	13.35 ± 0.98	10.0
	0.52 ± 0.08	1.74 ± 0.26	5.11 ± 0.79	13.59 ± 2.33	1.0
	0.33 ± 0.07	1.14 ± 0.26	3.55 ± 0.89	12.11 ± 4.58	0.1
WA	0.62 ± 0.02	2.01 ± 0.05	5.67 ± 0.15	13.57 ± 0.35	100.0
	0.61 ± 0.05	1.99 ± 0.15	5.62 ± 0.43	13.49 ± 1.01	10.0
	0.58 ± 0.09	1.92 ± 0.29	5.49 ± 0.86	13.84 ± 2.37	1.0
	0.37 ± 0.08	1.25 ± 0.30	3.82 ± 0.98	12.32 ± 4.76	0.1
QOG	0.54 ± 0.01	1.79 ± 0.05	5.17 ± 0.14	12.80 ± 0.33	100.0
	0.54 ± 0.04	1.78 ± 0.14	5.13 ± 0.39	12.72 ± 0.93	10.0
	0.52 ± 0.08	1.73 ± 0.26	5.03 ± 0.77	12.96 ± 2.16	1.0
	0.33 ± 0.07	1.13 ± 0.25	3.52 ± 0.89	11.44 ± 4.19	0.1
CWO	0.62 ± 0.02	1.18 ± 0.04	5.00 ± 0.14	12.89 ± 0.34	100.0
	0.61 ± 0.05	1.17 ± 0.11	4.96 ± 0.38	12.82 ± 0.94	10.0
	0.57 ± 0.09	35.70 ± 56.32	4.86 ± 0.75	13.08 ± 2.24	1.0
	0.37 ± 0.08	>100.0	3.40 ± 0.86	11.49 ± 4.26	0.1
EM	0.55 ± 0.02	1.80 ± 0.05	5.20 ± 0.15	12.91 ± 0.36	100.0
	0.54 ± 0.04	1.78 ± 0.14	5.18 ± 0.43	12.87 ± 1.05	10.0
	0.52 ± 0.08	1.73 ± 0.26	5.05 ± 0.84	13.02 ± 2.36	1.0
	0.34 ± 0.07	1.14 ± 0.26	3.51 ± 0.90	10.93 ± 4.15	0.1
MLE	0.55 ± 0.02	1.81 ± 0.06	5.06 ± 0.27	0.96 ± 0.24	100.0
	0.54 ± 0.04	1.76 ± 0.14	5.03 ± 0.45	0.94 ± 0.26	10.0
	0.53 ± 0.08	1.76 ± 0.29	5.02 ± 0.91	0.85 ± 0.30	1.0
	0.33 ± 0.07	1.07 ± 0.17	3.56 ± 0.80	0.09 ± 0.06	0.1
QP	0.54 ± 0.01	1.79 ± 0.05	5.20 ± 0.16	15.79 ± 8.17	100.0
	0.54 ± 0.04	1.77 ± 0.13	5.14 ± 0.40	13.68 ± 2.69	10.0
	0.52 ± 0.07	1.73 ± 0.26	5.07 ± 0.80	13.78 ± 4.07	1.0
	0.33 ± 0.07	1.14 ± 0.29	3.53 ± 0.98	13.64 ± 7.16	0.1
PA	0.54 ± 0.01	1.69 ± 0.22	3.49 ± 1.19	10.71 ± 0.39	100.0
	0.54 ± 0.04	1.59 ± 0.34	3.44 ± 1.19	10.62 ± 0.89	10.0
	0.52 ± 0.08	1.56 ± 0.40	3.42 ± 1.33	10.75 ± 2.01	1.0
	0.35 ± 0.09	1.06 ± 0.34	2.32 ± 0.94	9.53 ± 4.00	0.1

[a]Standard deviation from different iterations are shown as error. Increasing values of τ are seen along the columns while decreasing trajectory length (L) are represented along the rows. To increase readability, large values, greater than 100.0, are truncated. The full table is located in Section S6. The analytical MFPT is 34.3. Outliers were excluded from the individual data sets when calculating the averages and standard deviations reported in the table, using the interquartile range (IQR) method.

ascribed the generator matrix fulfilling the detailed balance relation, in contrast to the stationary distributions derived from the first eigenvector of the transition matrices. On the other hand, in terms of kinetic properties, the performance of our algorithms and direct use of transition matrices was generally comparable.

An important finding from these results is the influence of data quantity on estimation accuracy. It was observed that, in general, more data led to better estimates. However, it is worth noting that different algorithms exhibited different levels of sensitivity to the amount of data used; therefore, a careful consideration in the choice of estimation method is needed based on the available data and the specific properties of interest to achieve optimal results.

Another interesting finding is the relationship between lag time and the accuracy of the estimates. Generally, shorter lag times yielded more accurate free energy estimates, while longer lag times improved the accuracy of diffusion constant estimates. Shorter lag times provide a larger number of data points, and since thermodynamics is not inherently dependent on lag time, more data points contribute to better free energy estimates. Conversely, longer lag times reduce correlations, which are crucial for kinetic estimates. However, it is important to note that longer lag times may result in empirical transition matrices with more nonzero off-diagonal elements, deviating from the expected tridiagonal structure of the generator matrix. Therefore, to ensure the reliability of the results, it is necessary to examine the lag time dependence by repeating the analysis with various lag times.

7. SUMMARY AND CONCLUSION

We implemented eight algorithms for overcoming the "embeddability problem" in MSMs. The efficiency of these algorithms were tested on a 1D bistable toy model. The relative performance and convergence behavior of each algorithm were compared with regards to their ability to handle differing lag times, trajectory lengths, and ideal generator matrices. Our findings highlight the importance of Markovian assumption underlying all the algorithms and the algorithm-dependent behavior when it comes to thermodynamic calculations and kinetic characterization.

For thermodynamic characterization of a process, EM, MLE, and PA showed robust results for different lag times. PA was able to achieve comparable results in kinetic information

https://doi.org/10.1021/acs.jpca.3c01367
J. Phys. Chem. A 2023, 127, 5745−5759

estimation to EM and MLE despite having a significantly lower computational cost. CWO performed poorly at specific lag times, though it may perform better when lag times are long.

In terms of kinetic estimations such as diffusion constant, most algorithms performed similarly with the exception of MLE at long τ. MFPT estimates were similar, though large overstimation in several algorithms is observed. It is worth noting the approximate relative computational costs of the algorithms used here. The DA and WA algorithms are very low cost, with QOG, CWO, QP, and PA only slightly more expensive computationally. However, MLE and to a lesser extent EM are considerably more expensive than the other algorithms.

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 shed some light on the approaches as well as the information which can be garnered from direct analysis of 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.

■ ASSOCIATED CONTENT

Data Availability Statement

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

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jpca.3c01367.

Figures and tables showing average values and standard deviations of the metrics for barrier height, well symmetry, KL divergence, average diffusion, and MFPT for all algorithms using all the combinations of τ and L that were simulated (PDF)

■ AUTHOR INFORMATION

Corresponding Author

Mahmoud Moradi − *Department of Chemistry and Biochemistry, University of Arkansas, Fayetteville, Arkansas 72701, United States;* orcid.org/0000-0002-0601-402X; Email: moradi@uark.edu

Authors

Curtis Goolsby − *Department of Chemistry and Biochemistry, University of Arkansas, Fayetteville, Arkansas 72701, United States*

James Losey − *Department of Chemistry and Biochemistry, University of Arkansas, Fayetteville, Arkansas 72701, United States*

Ashkan Fakharzadeh − *Department of Physics, North Carolina State University, Raleigh, North Carolina 27607, United States*

Yuchen Xu − *Department of Statistics and Data Science, Cornell University, Ithaca, New York 14850, United States*

Marie-Christine Düker − *Department of Statistics and Data Science, Cornell University, Ithaca, New York 14850, United States*

Mila Getmansky Sherman − *Department of Finance, Isenberg School of Management, University of Massachusetts at Amherst, Amherst, Massachusetts 01003, United States*

David S. Matteson − *Department of Statistics and Data Science, Cornell University, Ithaca, New York 14850, United States*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jpca.3c01367

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This research is supported by the National Science Foundation under Awards 1940188, 1945465, 1934985, 1940124, 1940276, and 1940223. 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) Berendsen, H.; Hayward, S. Collective protein dynamics in relation to function. *Curr. Opin. Struct. Biol.* **2000**, *10*, 165−169.

(2) Eastman, P.; Swails, J.; Chodera, J.; McGibbon, R.; Zhao, Y.; Beauchamp, K.; Wang, L.-P.; Simmonett, A.; Harrigan, M.; Stern, C.; Wiewiora, R.; et al. OpenMM 7: Rapid development of high performance algorithms for molecular dynamics. *PLoS Comput. Biol.* **2017**, *13*, No. e1005659.

(3) Götz, A.; Williamson, M.; Xu, D.; Poole, D.; Le Grand, S.; Walker, R. Routine microsecond molecular dynamics simulations with AMBER on GPUs. 1. Generalized born. *J. Chem. Theory Comput.* **2012**, *8*, 1542−1555.

(4) Towles, B.; Grossman, J.; Greskamp, B.; Shaw, D. Unifying on-chip and inter-node switching within the Anton 2 network. *ACM SIGARCH Comput. Archit. News* **2014**, *42*, 1−12.

(5) Bernardi, R. C.; Melo, M. C. R.; Schulten, K. Enhanced Sampling Techniques In Molecular Dynamics Simulations Of Biological Systems. *Biochim. Biophys. Acta. Gen. Subj.* **2015**, *1850*, 872−877.

(6) Kumar, S.; Bouzida, D.; Swendsen, R.; Kollman, P.; Rosenberg, J. The Weighted Histogram Analysis Method For Free-Energy Calculations On Biomolecules 0.1. The Method. *J. Comput. Chem.* **1992**, *13*, 1011−1021.

(7) Torrie, G.; Valleau, J. Non-Physical Sampling Distributions In Monte-Carlo Free-Energy Estimation - Umbrella Sampling. *J. Comput. Phys.* **1977**, *23*, 187−199.

(8) Roux, B. The Calculation Of The Potential Of Mean Force Using Computer-Simulations. *Comput. Phys. Commun.* **1995**, *91*, 275−282.

(9) Kumar, S.; Rosenberg, J.; Bouzida, D.; Swendsen, R.; Kollman, P. Multidimensional Free-Energy Calculations Using The Weighted Histogram Analysis Method. *J. Comput. Chem.* **1995**, *16*, 1339−1350.

(10) Barducci, A.; Bussi, G.; Parrinello, M. Tempered Metadynamics: A Smoothly Converging And Tunable Free-Energy Method. *Phys. Rev. Lett.* **2008**, *100*, 020603.

(11) Laio, A.; Gervasio, F. L. Metadynamics: A Method To Simulate Rare Events And Reconstruct The Free Energy In Biophysics, Chemistry And Material Science. *Rep. Prog. Phys.* **2008**, *71*, 126601.

(12) Barducci, A.; Bonomi, M.; Parrinello, M. Metadynamics. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2011**, *1*, 826−843.

(13) Laio, A.; Rodriguez-Fortea, A.; Gervasio, F.; Ceccarelli, M.; Parrinello, M. Assessing The Accuracy Of Metadynamics. *J. Phys. Chem. B* **2005**, *109*, 6714−6721.

(14) Sugita, Y.; Okamoto, Y. Replica-Exchange Molecular Dynamics Method For Protein Folding. *Chem. Phys. Lett.* **1999**, *314*, 141−151.

(15) Hukushima, K.; Nemoto, K. Exchange Monte Carlo Method And Application To Spin Glass Simulations. *J. Phys. Soc. Jpn.* **1996**, *65*, 1604−1608.

(16) Earl, D.; Deem, M. Parallel Tempering: Theory, Applications, And New Perspectives. *Phys. Chem. Chem. Phys.* **2005**, *7*, 3910−3916.

(17) Sugita, Y.; Kitao, A.; Okamoto, Y. Multidimensional Replica-Exchange Method For Free-Energy Calculations. *J. Chem. Phys.* **2000**, *113*, 6042−6051.
(18) Fukunishi, H.; Watanabe, O.; Takada, S. On The Hamiltonian Replica Exchange Method For Efficient Sampling Of Biomolecular Systems: Application To Protein Structure Prediction. *J. Chem. Phys.* **2002**, *116*, 9058−9067.
(19) Andrieu, C.; Doucet, A.; Holenstein, R. Particle Markov chain Monte Carlo methods. *J. R. Stat. Soc. Series B Stat. Methodol.* **2010**, *72*, 269−342.
(20) Ephraim, Y.; Merhav, N. Hidden Markov processes. *IEEE Trans. Inf. Theory* **2002**, *48*, 1518−1569.
(21) Pande, V.; Beauchamp, K.; Bowman, G. Everything You Wanted To Know About Markov State Models But Were Afraid To Ask. *Methods* **2010**, *52*, 99−105.
(22) Bowman, G.; Voelz, V.; Pande, V. Taming The Complexity Of Protein Folding. *Curr. Opin. Struct. Biol.* **2011**, *21*, 4−11.
(23) Sisson, S. Transdimensional Markov Chains: A Decade Of Progress And Future Perspectives. *J. Am. Stat. Assoc.* **2005**, *100*, 1077−1089.
(24) Shukla, D.; Hernandez, C.; Weber, J.; Pande, V. Markov State Models Provide Insights Into Dynamic Modulation Of Protein Function. *Acc. Chem. Res.* **2015**, *48*, 414−422.
(25) Chodera, J. D.; Singhal, N.; Pande, V. S.; Dill, K. A.; Swope, W. C. Automatic Discovery Of Metastable States For The Construction Of Markov Models Of Macromolecular Conformational Dynamics. *J. Chem. Phys.* **2007**, *126*, 155101.
(26) Prinz, J.-H.; Wu, H.; Sarich, M.; Keller, B.; Senne, M.; Held, M.; Chodera, J. D.; Schuette, C.; Noe, F. Markov Models Of Molecular Kinetics: Generation And Validation. *J. Chem. Phys.* **2011**, *134*, 174105.
(27) Bowman, G. R.; Beauchamp, K. A.; Boxer, G.; Pande, V. S. Progress And Challenges In The Automated Construction Of Markov State Models For Full Protein Systems. *J. Chem. Phys.* **2009**, *131*, 124101.
(28) Beauchamp, K. A.; Bowman, G. R.; Lane, T. J.; Maibaum, L.; Haque, I. S.; Pande, V. S. Msmbuilder2: Modeling Conformational Dynamics On The Picosecond To Millisecond Scale. *J. Chem. Theory Comput.* **2011**, *7*, 3412−3419.
(29) Djurdjevac, N.; Sarich, M.; Schütte, C. Estimating the eigenvalue error of Markov state models. *Multiscale Model. Simul.* **2012**, *10*, 61−81.
(30) Prinz, J.-H.; Chodera, J.; Noé, F. Spectral rate theory for two-state kinetics. *Phys. Rev. X* **2014**, *4*, No. 011020.
(31) Inamura, Y. *Estimating Continuous Time Transition Matrices From Discretely Observed Data*; Bank Of Japan:2006; pp 06−07.
(32) Gagniuc, P. *Markov Chains: From Theory To Implementation And Experimentation*; Wiley: 2017.
(33) Hummer, G. Position-dependent diffusion coefficients and free energies from Bayesian analysis of equilibrium and replica molecular dynamics simulations. *New J. Phys.* **2005**, *7*, 34.
(34) Noé, F.; Horenko, I.; Schütte, C.; Smith, J. Hierarchical analysis of conformational dynamics in biomolecules: Transition networks of metastable states. *J. Chem. Phys.* **2007**, *126*, 155102.
(35) Lando, D.; Skødeberg, T. M. Analyzing rating transitions and rating drift with continuous observations. *J. Bank Financ.* **2002**, *26*, 423−444.
(36) Guerry, M.-A. Some Results On The Embeddable Problem For Discrete-Time Markov Models In Manpower Planning. *Communications in Statistics-Theory and Methods* **2014**, *43*, 1575−1584.
(37) Guerry, M.-A. On The Embedding Problem For Discrete-Time Markov Chains. *J. Appl. Probab.* **2013**, *50*, 918−930.
(38) Marada, T. Quantitative Credit Risk Modeling: Ratings Under Stochastic Time. M.Sc. thesis, VU University, Amsterdam, 2008.
(39) Crommelin, D.; Vanden-Eijnden, E. Fitting timeseries by continuous-time Markov chains: A quadratic programming approach. *J. Comput. Phys.* **2006**, *217*, 782−805.
(40) Elfving, G. Zur theorie der Markoffschen ketten. *Acta Soc. scient. Fennicae* **1937**, *A2* (9), 2−17.
(41) Kingman, J. F. C. The imbedding problem for finite Markov chains. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete* **1962**, *1*, 14−24.
(42) Runnenburg, J. T. On Elfving's problem of imbedding a time-discrete Markov chain in a time-continuous one for finitely many states. *Indagationes Mathematicae (Proceedings)* **1962**, *65*, 536−541.
(43) Johansen, S. Some results on the imbedding problem for finite Markov chains. *J. London Math. Soc.* **1974**, *s2-8*, 345−351.
(44) Davies, E. Embeddable markov matrices. *Electron. J. Probab.* **2010**, *15*, 1474−1486.
(45) Guerry, M.-A. On the embedding problem for discrete-time Markov chains. *J. Appl. Probab.* **2013**, *50*, 918−930.
(46) Jia, C. A solution to the reversible embedding problem for finite Markov chains. *Stat. Probab. Lett.* **2016**, *116*, 122−130.
(47) Lifson, S.; Jackson, J. L. On the self-diffusion of ions in a polyelectrolyte solution. *J. Chem. Phys.s* **1962**, *36*, 2410−2414.
(48) Ansari, A. Mean first passage time solution of the Smoluchowski equation: Application to relaxation dynamics in myoglobin. *J. Chem. Phys.* **2000**, *112*, 2516−2522.
(49) Israel, R. B.; Rosenthal, J. S.; Wei, J. Z. Finding generators for Markov chains via empirical transition matrices, with applications to credit ratings. *Math. Financ.* **2001**, *11*, 245−265.
(50) Kreinin, A.; Sidelnikova, M. Regularization algorithms for transition matrices. *Algo. Res. Q.* **2001**, *4*, 23−40.
(51) Frank, M.; Wolfe, P. An algorithm for quadratic programming. *Nav. Res. Logist. Q.* **1956**, *3*, 95−110.
(52) Maruyama, G. Continuous Markov processes and stochastic equations. *Rendiconti del Circolo Matematico di Palermo* **1955**, *4*, 48.
(53) Scherer, M. K.; Trendelkamp-Schroer, B.; Paul, F.; Peréz-Hernández, G.; Hoffmann, M.; Plattner, N.; Wehmeyer, C.; Prinz, J. H.; Noé, F. PyEMMA 2: A Software Package for Estimation, Validation, and Analysis of Markov Models. *J. Chem. Theory Comput.* **2015**, *11*, 5525−5542.
(54) Kullback, S.; Leibler, R. A. On Information and Sufficiency. *Ann. Math. Stat.* **1951**, *22*, 79−86.

https://doi.org/10.1021/acs.jpca.3c01367
J. Phys. Chem. A 2023, 127, 5745−5759