



King's Research Portal

DOI:

[10.1063/1.5027203](https://doi.org/10.1063/1.5027203)

Document Version

Peer reviewed version

[Link to publication record in King's Research Portal](#)

Citation for published version (APA):

Kells, A., Annibale, A., & Rosta, E. (2018). Limiting relaxation times from Markov state models. *Journal of Chemical Physics*, 149(7). <https://doi.org/10.1063/1.5027203>

Citing this paper

Please note that where the full-text provided on King's Research Portal is the Author Accepted Manuscript or Post-Print version this may differ from the final Published version. If citing, it is advised that you check and use the publisher's definitive version for pagination, volume/issue, and date of publication details. And where the final published version is provided on the Research Portal, if citing you are again advised to check the publisher's website for any subsequent corrections.

General rights

Copyright and moral rights for the publications made accessible in the Research Portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognize and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the Research Portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the Research Portal

Take down policy

If you believe that this document breaches copyright please contact librarypure@kcl.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.

Limiting relaxation times from Markov state models

Adam Kells¹, Alessia Annibale² and Edina Rosta^{1*}

¹King's College London, Department of Chemistry, SE1 1DB, London, UK

*e-mail address: edina.rosta@kcl.ac.uk

²King's College London, Department of Mathematics, WC2R 2LS, London, UK

Markov state models (MSMs) are more and more widely used in the analysis of molecular simulations to incorporate multiple trajectories together and obtain more accurate timescale information of the slowest processes in the system. Typically, however, multiple lagtimes are used and analyzed as input parameters, yet convergence with respect to the choice of lagtime is not always possible. Here, we present a simple method for calculating the slowest relaxation time (RT) of the system in the limit of very long lagtimes. Our approach relies on the fact that the second eigenvector's autocorrelation function of the propagator will be approximately single exponential at long lagtimes. This allows us to obtain a simple equation for the behavior of the MSM's relaxation time as a function of the lagtime with only two free parameters, one of these being the RT of the system. We demonstrate that the second parameter is a useful indicator of how Markovian a selected variable is for building the MSM. Fitting this function to data gives a limiting value for the optimal variational RT. Testing this on analytic and molecular dynamics (MD) data for Ala5 and umbrella sampling-biased ion channel simulations shows that the function accurately describes the behavior of the RT and furthermore that this RT can improve noticeably the value calculated at the longest accessible lagtime. We compare our RT limit to the hidden Markov model (HMM) approach that typically finds RTs of comparable values. However, HMMs cannot be used in conjunction with biased simulation data, require more complex algorithms to construct than MSMs, and the derived RTs are not variational, leading to ambiguity in the choice of lagtime at which to build the HMM model.

Keywords: Markov-state model (MSM), hidden Markov model (HMM), molecular dynamics (MD), relaxation time (RT), lagtime

I. INTRODUCTION

Markov state models (MSMs) have proven to be a useful tool for analyzing and understanding the results of a vast range of molecular dynamics (MD) simulations^{1,2} from folding/unfolding and conformational dynamics under applied forces^{3,4}. MSMs allow for the convenient combination of multiple MD trajectories into a single kinetic network model from which experimental observables can be more accurately computed⁵⁻¹².

The construction of MSMs involves choosing several parameters and collective variables. These include for example reaction coordinates¹³, the discretization of the state space (e.g., metastable and transition state clustering^{14,15}), and the choice of the lagtime (LT) at which transition probabilities are determined. The optimal and efficient calculation of these parameters is an active area of research and discussion¹⁶. Furthermore, there is also a number of options recently available to calculate unbiased MSMs from biased simulation data.¹⁷⁻²⁰

The slowest relaxation time (RT) of a system represents the timescale upon which the slowest process in the system takes place and can be calculated directly from an MSM of a molecular system using the eigenvalues of the propagator⁸. However, when constructing an MSM in practice, the RT will have a functional dependence, due to non-Markovian behavior, on the LT (i.e., the time at which the conditional transition probabilities are calculated) at which the MSM model is constructed. The most common condition for making a choice of a good LT value for an MSM is such that the Chapman-Kolmogorov equation is satisfied (RT of the MSM is effectively constant with respect to changes in the LT). However, in practical applications the RT will depend on the LT and the choice of what is considered effectively constant may be arbitrary and as the range of accessible LTs is limited by the length of available simulation data. Therefore, it is possible that the slowest RT of the MSM will not be observed to converge fully.

MSMs provide timescale information that follows a variational principle²¹, and the better the model the longer timescales are obtained. More accurate estimate of the slowest RTs from MD simulations is usually also sought after to compare the slowest timescales modelled in the simulations with experimental data. We are therefore interested in computing the limiting value of the RT for the largest available LTs given a set of simulation trajectories using MSMs. Alternatively, RTs are also obtained from more complex and slower algorithms, such as HMMs, which do not follow a simple variational principle with respect to the RTs, and are also more difficult to implement. Here, we

derive an approximate expression for RT behavior applicable to MSMs and, interestingly, we show that the limiting RT value can deviate and thus improve significantly from the values calculated at the longest accessible LTs. In addition to obtaining the limiting value of the lagtime, we also fit a second parameter that is related with how Markovian a selected collective variable is in constructing MSMs at reduced dimensions. We apply our method on analytical models, as well as demonstrate it's applicability on MD simulations of two systems: (i) unbiased Ala5 trajectories and (ii) Umbrella Sampling-biased pentameric GLIC ion channel simulation data.

II. THEORY

i. Markov State Models

An MSM at an LT τ is constructed from MD simulation data as a set of conditional probabilities between an initial microstate S_i and a final microstate S_j , where these microstates are discrete regions along our chosen (possibly multidimensional) reaction coordinate x .

$$m_{ji} = \text{Prob}[x(t + \tau) \in S_j | x(t) \in S_i] \quad (1)$$

Here t is the start time between the transitions that we average over for all the available time trajectories. Typically, a sliding window¹⁰ approach is employed, such that the conditional probabilities (and consequently our system propagators) are not dependent on the choice of t . In some practical applications, one might wish to consider this dependence explicitly, and e.g., study how it influences the spectral properties of the propagators.²² Here, we assume that our probabilities are independent of the time at which the measurements are taken, and average over all available t . In the case of a continuous reaction coordinate, one must choose a certain discretization procedure.^{23,24} The corresponding matrix of conditional transition probabilities $[M(\tau)]_{ji} = m_{ji}$ can be used to propagate the current probability distribution $P(0)$ to its value at τ time later, $P(\tau)$.

$$P(\tau) = \mathbf{M}(\tau)P(0) \quad (2)$$

This is equivalent to the rate matrix formalism whereby the time evolution of the probability distribution is given by the differential equation $\frac{dP}{dt} = \mathbf{K}P$ and the rate matrix (\mathbf{K}) and Markov matrix ($\mathbf{M}(\tau)$) are hence related via the LT value by $\mathbf{M}(\tau) = e^{\mathbf{K}\tau}$.

A spectral decomposition allows the Markov matrix to be written in terms of its eigenvalues and eigenvectors which provide information about the dynamics of the system.

$$[M(\tau)]_{ji} = \sum_n \psi_n^R(j) \psi_n^L(i) e^{\lambda_n \tau} \quad (3)$$

Where ψ_n^R and ψ_n^L are the right and left eigenvectors, respectively, corresponding to eigenvalues $e^{\lambda_n \tau}$. λ_n are the eigenvalues of the associated rate matrix and they are ordered such that $0 = \lambda_1 < \lambda_2 \leq \dots \leq \lambda_N$.

The second largest eigenvalue (in magnitude) of the MSM (second smallest in magnitude in the rate matrix formalism) describes the slowest relaxation process in the system. In practice, the RT determined from an MSM will have a functional dependence on the LT at which the model is constructed. This introduces a question of what LT an MSM should be constructed at. Typically, the Chapman-Kolmogorov (CK) condition, Eq. (4) is used to assess non-Markovian effects:

$$[M(\tau)]^n P = M(n\tau)P \quad (4)$$

which states that a system is Markovian at LT τ if using the propagator $\mathbf{M}(\tau)$ n times is equivalent to using the propagator $\mathbf{M}(n\tau)$. This is equivalent to the eigenvalues (and consequently the RT) of a Markovian system being invariant under LT changes.

In practice, MSMs of biomolecular systems are not truly Markovian due to dimensionality reduction, e.g., reaction coordinate discretization or insufficient data, thus the “best” LT is typically chosen for MSMs at the largest available values, for which the change in the RT is small. Here, instead of choosing a LT where the RT values appear converged at different LTs, we propose to estimate the optimal RT by fitting the RT curve as a function of LT with a proposed analytical expression, which allows us to calculate the limiting variational RT value from this fit.

The quantity which we will use in our subsequent derivation is the normalized correlation function, $c(f,g,t,M)$ for dynamics propagated by \mathbf{M} . This can be written in terms of the eigenvalues and the eigenvectors of the Markov matrix. The equation for the correlator $c(f,g,t,M)$ between two time-dependent quantities f and g (where f and g have elements $f[i]$ and $g[i]$, respectively, and $f \cdot \psi$ represents the dot product between the two vectors f and ψ) is given by:^{8,25,26}

$$c(f, g, t, M) = \frac{\sum_{n=2}^N e^{\lambda_n t} (g \cdot \psi_n^R)(f \cdot \psi_n^R)}{\sum_{n=2}^N (g \cdot \psi_n^R)(f \cdot \psi_n^R)} \quad (5)$$

ii. Lagtime dependence of the MSM relaxation times

The founding assumption of the derivation laid out in this section is that in constructing a coarse grained MSM, these lower dimensional eigenvectors behave similarly to the eigenvectors of the exact dynamics. In many cases, a dynamic continuous Markovian system could effectively be described as an N -state system with an effective dominant eigenvalue $e^{\lambda_2^{MSM} \tau}$ and effective eigenfunctions ψ_n^{R-MSM} corresponding to the correlation function:

$$c(f, g, t, M(\tau)^{MSM}) = \frac{\sum_{n=2}^N e^{\lambda_n^{MSM} t} (g \cdot \psi_n^{R-MSM})(f \cdot \psi_n^{R-MSM})}{\sum_{n=2}^N (g \cdot \psi_n^{R-MSM})(f \cdot \psi_n^{R-MSM})} \quad (6)$$

Equivalently, we could consider the full continuous description of the system as in Eq. (7):

$$c(f, g, t, M) = \frac{\sum_{n=2}^{\infty} e^{\lambda_n t} (g \cdot \psi_n^R)(f \cdot \psi_n^R)}{\sum_{n=2}^{\infty} (g \cdot \psi_n^R)(f \cdot \psi_n^R)} \quad (7)$$

In the above equations we have kept the general expression for a correlator between two arbitrary functions f and g . Now we consider the concrete example where these are both equal to the second left eigenvector ($f = g = \psi_2^{L-MSM}$) of the MSM. From orthogonality we then find that:

$$c(\psi_2^{L-MSM}, \psi_2^{L-MSM}, t, M(\tau)^{MSM}) = \frac{\sum_{n=2}^N e^{\lambda_n^{MSM} t} (\psi_2^{L-MSM} \cdot \psi_n^{R-MSM})(\psi_2^{L-MSM} \cdot \psi_n^{R-MSM})}{\sum_{n=2}^N (\psi_2^{L-MSM} \cdot \psi_n^{R-MSM})(\psi_2^{L-MSM} \cdot \psi_n^{R-MSM})} = e^{\lambda_2^{MSM} t} \quad (8)$$

Now let us consider the same correlation function calculated with respect to the full dimensional continuous dynamics:

$$c(\hat{P}(\psi_2^{L-MSM}), \hat{P}(\psi_2^{L-MSM}), t, M) = \frac{\sum_{n=2}^{\infty} e^{\lambda_n t} (\hat{P}(\psi_2^{L-MSM}) \cdot \psi_n^R)(\hat{P}(\psi_2^{L-MSM}) \cdot \psi_n^R)}{\sum_{n=2}^{\infty} (\hat{P}(\psi_2^{L-MSM}) \cdot \psi_n^R)(\hat{P}(\psi_2^{L-MSM}) \cdot \psi_n^R)} = \sum_{n=2}^{\infty} A_n e^{\lambda_n t} \quad (9)$$

Where $\hat{P}(\psi_2^{L-MSM})$ is the projection of the second MSM eigenvector onto the full dimensional space such that the $i \in I$ -th element of the $\hat{P}(\psi_2^{L-MSM})$ vector is equal with the corresponding coarse grained element $\psi_2^{L-MSM}(I)$, and the coefficients A_i -s are the scalar product between the projected

second left eigenvector ($\hat{P}(\psi_2^{L-MSM})$) of the coarse grained MSM and the i -th eigenvector of the continuous dynamics.

Correlation functions that define the MSM Markov matrix elements, correspond to the choice of

the coarse grained indicator function $f_i(J) = \begin{cases} 1, & \text{if } J = I \\ 0, & \text{if } J \neq I \end{cases}$, are exactly equal at $t = \tau$ to the full

dimensional correlation functions with projection vectors of the full dimensional space,

$$g_I(i) = \begin{cases} 1, & \text{if } i \in I \\ 0, & \text{if } i \notin I \end{cases} :$$

$$M^{MSM}(I, J)P(J) = \sum_{n=2}^N e^{\lambda_n^{MSM} \tau} (f_I \bullet \psi_n^{R-MSM})(f_J \bullet \psi_n^{R-MSM}) = \sum_{i \in I} \sum_{j \in J} \sum_{n=2}^{\infty} e^{\lambda_n \tau} (g_I \bullet \psi_n^R)(g_J \bullet \psi_n^R) = \sum_{i \in I} \sum_{j \in J} M^{full}(i, j)P(j) \quad (10)$$

Using Eq. (10) and writing ψ_2^{L-MSM} as a linear combination of weighted sum of f_i -s as basis vectors, we can equate the two correlation functions of Eqs. (8) and (9) exactly at $t = \tau$:

$$c(\psi_2^{L-MSM}, \psi_2^{L-MSM}, \tau, M(\tau)^{MSM}) = c(\hat{P}(\psi_2^{L-MSM}), \hat{P}(\psi_2^{L-MSM}), \tau, M) \quad (11)$$

If we also assume that the MSM is a faithful reproduction of the full dynamics then we expect that $A_2 \gg A_{i>2}$ in Eq. (9). Moreover, at long LTs, only the λ_2 eigenvalue will dominate the expression in Eq. (9):

$$\sum_{n=2}^{\infty} A_n e^{\lambda_n t} \approx A_2 e^{\lambda_2 t} \quad (12)$$

If we once again assume that the coarse grained picture is accurate and A_2 does not have a significant LT dependence, then from Eq. (11) we can obtain the coarse grained second eigenvalue as a function of the LT:

$$e^{\lambda_2^{MSM} \tau} = \sum_{n=2}^{\infty} A_n e^{\lambda_n \tau} \approx A_2 e^{\lambda_2 \tau} \quad (13)$$

$$\lambda_2^{MSM} = \lambda_2 + \frac{\varepsilon}{\tau} \quad (14)$$

Where $\varepsilon = \log(A_2)$ and the relaxation timescales of the system are the inverse of the eigenvalues ($\mu_n^{relax} = \frac{1}{\lambda_n}$). This leads to Eq. (15), which describes the RT, $\mu_2^{relax-MSM}$, as a function of the LT, τ :

$$\mu_2^{relax-MSM} = \frac{\tau \times \mu_2^{relax}}{\tau + \varepsilon \mu_2^{relax}} \quad (15)$$

We have thus obtained a functional dependence for the observed RT as a function of the LT τ at large values, with two free parameters: the true (limiting) relaxation timescale μ_2^{relax} and initial rate of change of the effective RT ε . The latter is related to how well the second eigenvector is captured

with the MSM, therefore how good the collective variable used for the MSM is in describing the slowest process. If $\varepsilon = 0$, the system is perfectly Markovian with respect to the slowest relaxation time (there is no LT dependence). Hence, by generating the RT for a range of large LTs, fitting this curve to the data will yield a value for both the intrinsic limiting RT parameter (μ_2^{relax}) and the quality of the MSM via ε .

The derivation demonstrated above is built upon the assumption that the correlation functions obtained from a full dimensional MSM and a coarse grained MSM behave similarly. As such, Eq. (15) is applicable to any set of data from which an MSM can be constructed. This includes umbrella sampling simulations, since a number of unbiasing methods exist¹⁷⁻²⁰ which can construct MSMs from biased data.

iii. Hidden Markov Models

For comparison, the approach derived above is contrasted with the results of using a hidden Markov Model (HMM) formalism.²⁷ This method is outlined in detail in a recent publication by Noe et al.²⁸ The central idea of HMMs is that there exist some set of unobserved (hidden) states on which the dynamics of the system are Markovian. Then from these underlying hidden states $\{h_i\}$, at each observation time the system will project onto one of our observed states $\{S_j\}$ with a given probability E_{ij} .

Given a set of observation data amongst the observed states one can then construct a Markov model that describes the dynamics amongst the hidden states and proceed to analyze the resulting model as one would with a regular MSM. This approach has been shown to be successful in analyzing MD simulation data and to accurately identify relaxation times from short LTs.²⁸

However, the RTs obtained by implementing the HMM method do not generally follow a variational principle, and might result in longer RTs than the true value. Moreover, the RTs do not follow the functional dependence as a function of the LTs as MSMs do, as there is no corresponding theoretical description. Therefore, our fitting procedure is not applicable for HMM data as there are situations where the HMM does not display the same functional dependence as our derived fitting equation. Such examples are also demonstrated here in the context of the pentalanine simulation data.

III. RESULTS

The derived equation for the slowest RT is tested on three different systems: (i) a trajectory in an unbiased analytic potential, (ii) MD trajectories of unbiased pentalanine in explicit water and (iii) an umbrella sampling simulations of an ion passing through a pentameric GLIC ion channel. The results are compared to the RTs predicted by the HMM approach implemented in PyEMMA²⁹ for the unbiased cases. A series of Markov models are constructed at different LTs and the values for the fitting parameters that minimize the error are calculated. The fitting parameters are obtained by doing least squares fitting over the range of LTs shown in the figures.

i. Analytic potential

The first system we tested is an analytic potential given by $V(x) = -2\sin[(x-\pi)/2] + \frac{x}{8\pi} + c_0$ where c_0 is a number such that the minimum of the function in the domain $-4\pi \leq x \leq 4\pi$ is 0 (Fig. 1). The system's dynamics is constrained within this domain. We identified the elements of the associated rate matrix \mathbf{K} by discretizing the x -axis into 100 bins and used an Arrhenius-like expression of $K_{ij} = Ae^{-\beta\left(\frac{V(j)-V(i)}{2}\right)}$ to calculate the transition rates (with $A = 2.5$), where our analytic potential is given by the aforementioned function $V(x)$. From this rate matrix, a series of trajectories were generated by a simple Markov chain propagation approach on a 100 state model created by discretizing the interval x into equally sized bins. We also tested the Gillespie algorithm³⁰ as an alternative, that gives largely similar results (data not shown). Each MSM was constructed using 100 independent simulations of length 40000 which is approximately twice the exact relaxation time of the system (~ 19200). This process was repeated 10 times to estimate errors bar on the calculated values. These trajectories are used to construct a series of MSMs at different LTs as shown in Fig 1(b). The advantage of using an artificial potential is that by construction we know the exact relaxation time of the system and we can evaluate whether Eq. (15) is an effective method of calculating the RT.

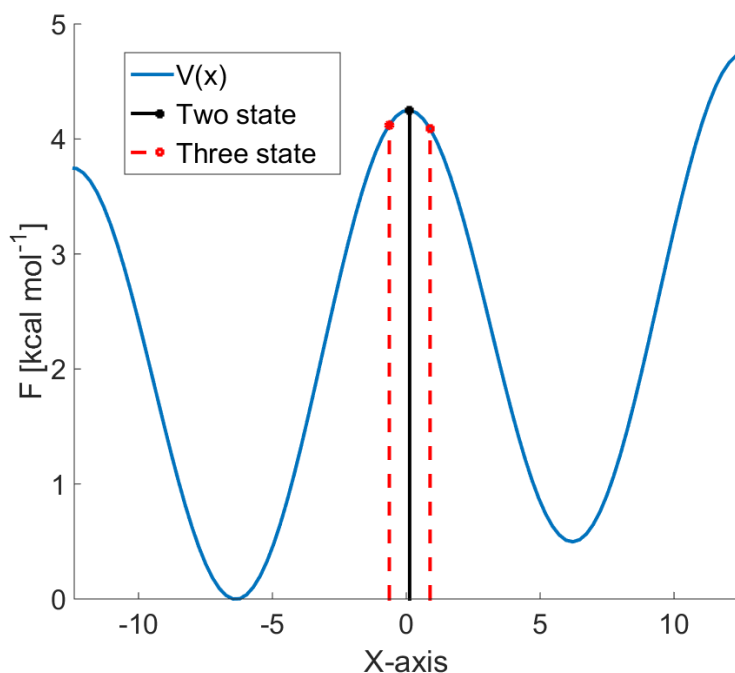


Figure 1: Analytic free energy profile used to determine the system's dynamics and corresponding relaxation times. Optimal cluster boundaries for 2-state (black vertical line) and 3-state (red dashed vertical lines) clustering are shown. The optimal boundaries are calculated using a recent variational method for obtaining cluster boundaries¹⁵.

We considered here two interesting trajectory types based on different initial conditions. In the first case, we ran a series of trajectories with randomly chosen initial state and clustered the states in the resulting trajectory into two clusters (with a boundary shown as black vertical line, Fig 1). The boundaries for the (2- and 3-state) clusterings (Fig. 1) are the variationally optimal clustering boundaries for coarse graining as described in Ref. 15. We then extracted the RTs obtained by constructing an MSM and a 2-state HMM using the PyEMMA software. Our derived method is then used to fit to the MSM relaxation times and find the limiting value of the RT. For comparison the analysis is repeated using only 25% of the data, such that each trajectory is of length 10000 and hence is approximately half of the RT of the system (Fig. 2a.). This can be contrasted with the full data in Fig 2b. As previously observed, the 2-state HMM finds RTs much closer to the true value than the MSMs and moreover it finds these values at shorter LTs. However, we found here that performing a best fit to the MSM data leads to a limiting RT that is approximately the same as the HMM value, yet using a much simpler approach. We find that, as expected, increasing the length of the trajectories results in calculated RTs closer to the exact value.

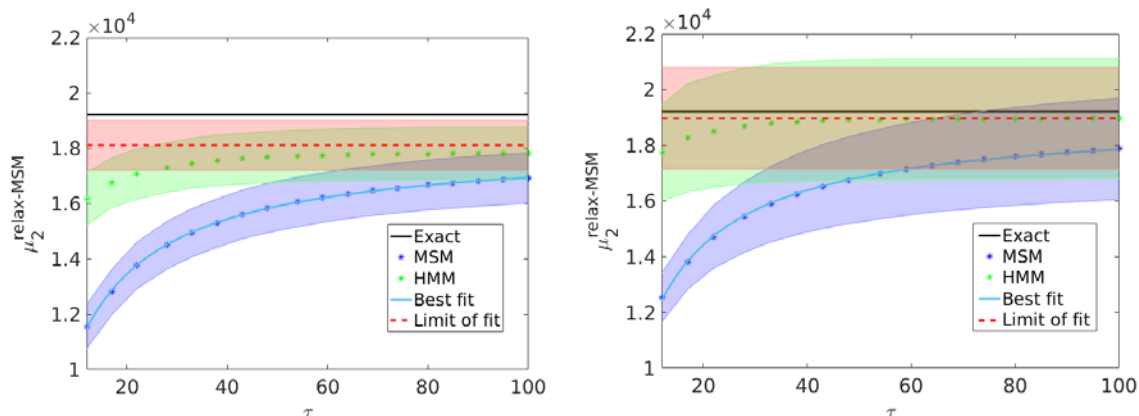


Figure 2: Slowest relaxation times constructed from MSMs (blue stars), HMMs (green stars) and from the fit using Eq. (15) (Fig. 1, black) of data obtained using an analytic potential with an exact relaxation time shown in black. The best fit is shown by a light blue line and the limit of the best fit is given by a dashed red line. The fitting is performed with a least squares approach on LTs in the range 1 to 25.

The second case we examined involves a series of downhill trajectories with a 3-state clustering (Fig 1, red dashed vertical lines show cluster boundaries). Similarly, we ran 100 trajectories of length 40,000 with each trajectory initiated from the top of the barrier (inside the transition state). In this case we observed that a 3-state MSM (blue symbols) or a 2-state HMM (green symbols) correctly identified the true RT while a 3-state HMM (red symbols) slightly exceeded this value.

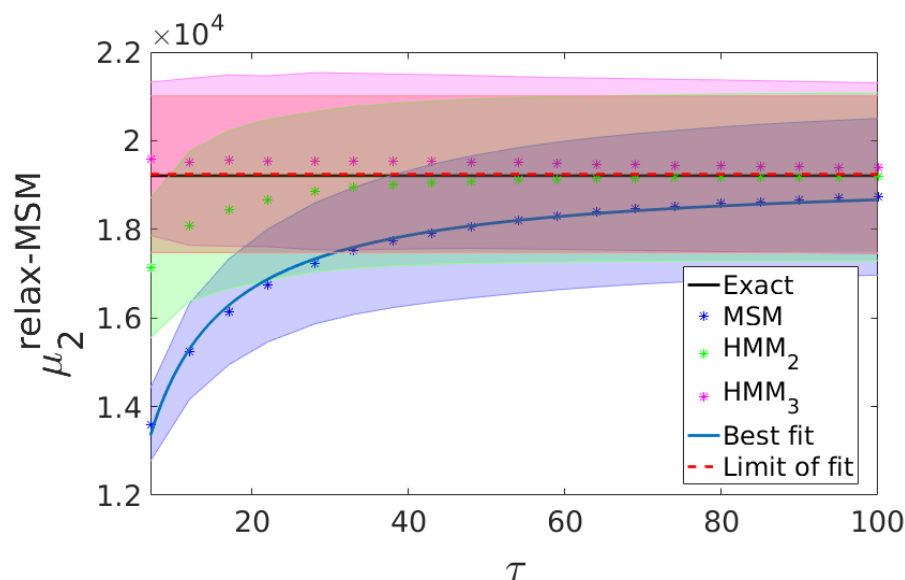


Figure 3: Relaxation timescale plots for a series of 100 downhill trajectories (simulations initialized from the top of the barrier). Timescales are extracted from 3-state MSM (blue stars), 2-state (green stars) and 3-state (magenta stars) HMM. The fitting method (solid blue line) used in conjunction with the MSM data finds a limiting value (dashed red line) of almost exactly the correct timescale (solid black line), as does the two state HMM. The 3-state HMM slightly exceeds the correct value although this is within the margin of error (shaded color).

These analytic examples demonstrate that using the fitting procedure in conjunction with MSMs obtained at different LTs results in RTs much closer to the true value than using the RT calculated at the largest LT. This limiting value is close to the value found by the HMM method and has error bars of similar magnitude.

ii. Unbiased Ala₅ simulation data

In the following section we apply our approach to an MD simulation of pentalanine (Ala₅). This system has become one of the more commonly used example systems for new MSM analysis methods,⁹ as it is one of the simplest systems allowing it to be simulated until convergence, whilst still demonstrating interesting kinetic behavior for helix-coil transition. Here, the Ala₅ MD simulations of Ref. 8 are analyzed via the 10 Ramachandran angles: five φ and five ψ backbone dihedral angles.

The data used for our study was obtained from the study of Buchete and Hummer⁸ and further details of the simulation methods and parameters can be found there. The relevant simulation details for our analysis is that the data consists of four 250 ns long independent unbiased MD simulations at different initial conditions with frames saved every 1 ps. The RT is calculated by creating a single MSM from the full dataset. The error on this number is taken to be the variance on the RT obtained from splitting the data into four sets based on simulation time and creating MSMs for each set separately. Here we demonstrate the results using one of the coordinates, ϕ_3 , a summary of the key information of the other coordinates is presented in Table 1 and in the SI.

We analyzed the LT dependence of the RTs for all angles, and found that even the MSMs constructed at the longest accessible LTs are not fully converged and do not satisfy the CK condition. Accordingly, a best fit with Eq. (15) a limiting RT is obtained that is much longer than the longest RT of the MSMs in many cases, particularly for the ϕ angles that are generally less good reaction coordinates. Intriguingly, however, the limiting RT values are very similar for all but one (φ_5) angles (Table 1), and well approximate the $\sim 6-7$ ns RT obtained⁸ by an analysis that simultaneously considers all angles and uses a transition-based state assignment. Therefore, our results demonstrate that building MSMs using almost any of our finely discretized 1D Ramachandran angle provides very similar RTs as obtained from the more complex analysis, but coarse grained to strictly metastable states. Moreover, we can also evaluate how good a reaction coordinate is by comparing

the fitted ϵ values (Table 1), these values tell us how fast the RC converges to its long time limit. Smaller values indicate a faster convergence to the long time limit and hence more Markovian behavior. In particular, for a perfect reaction coordinate, we expect ϵ to vanish and thus show no LT dependence, therefore the smaller the value the better the reaction coordinate is in describing the slowest process. In general, smaller ϵ values also correspond to larger RTs of the MSMs at the longest, but not necessarily at the smallest LTs.

Examining the data in Table 1, we see that ϕ_5 has the largest ϵ value and hence the slowest convergence to its long time value. This may explain why our estimate for the long time value is so different from those from the other coordinates. Also the Ψ coordinates, which are known to be good reaction coordinates all have much smaller values of ϵ (<0.2 , except for Ψ_5 that is also known to correspond to the flexible end of the peptide). This suggests that our fitted ϵ parameter may be useful in identifying good RCs.

For comparison, the same data is examined using the HMM formalism implemented in PyEMMA. Using the full data set, HMMs find approximately the same RTs as the limiting fit (green data points in Fig. 4). However, while it finds consistent RTs at short LTs, the HMM appears to break down at earlier LTs than the MSM for longer LTs, giving rise to larger numerical error. Similar results are also obtained for the other Ramachandran angles of Ala₅ that are included in the SI. Typically the HMM and limiting fit results in similar RT estimates with similar errors.

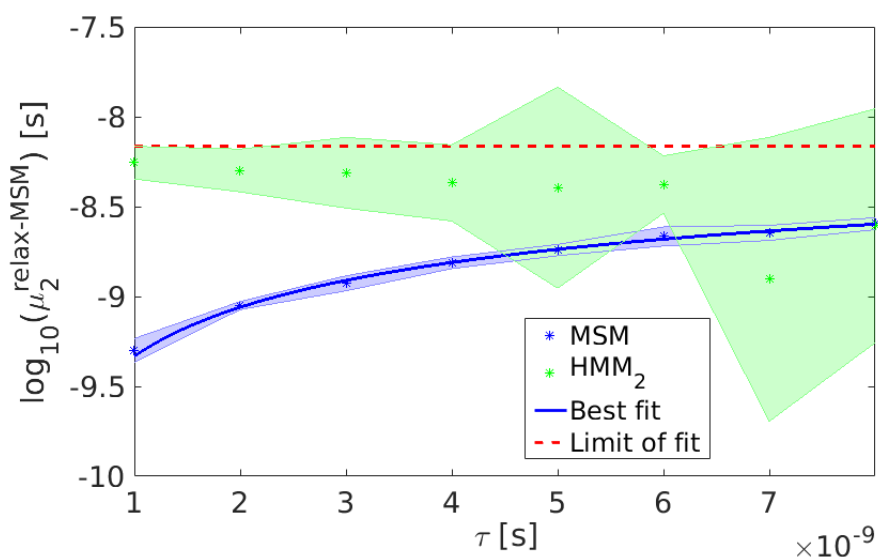


Figure 4: Relaxation timescale plot for angle ϕ_3 of the Ala₅ simulation data using a 100-state MSM model (blue symbols and color shaded error bars). The fit with Eq. (15) is in the range 1 to 8 ns LTs (red dashed line). The relaxation times on the y-axis are on a log₁₀ scale. The HMM relaxation times are also shown for 2-state (green symbols and color shaded error bars).

Table 1: Relaxation times (in ps) calculated for Markov models at short ($\tau = 1$ ps) and long ($\tau = 1000$ ps) LTs for each Ramachandran angle of Ala5. The fitted values to Eq. (15) μ_2^{relax} are also given.

COORDINATE	LT=1	LT=1000	EPSILON	LIMITING RT
1 ($\Phi1$)	6.5	516.1	1.81	6976.3
2 ($\Psi1$)	952.2	2700.7	0.23	4711.3
3 ($\Phi2$)	25.5	567.7	1.75	6042.0
4 ($\Psi2$)	687.2	3353.6	0.17	6571.1
5 ($\Phi3$)	33.9	515.8	2.01	6875.1
6 ($\Psi3$)	653.2	2813.0	0.22	5101.8
7 ($\Phi4$)	65.8	424.7	2.47	9421.1
8 ($\Psi4$)	490.0	1929.3	0.47	5325.4
9 ($\Phi5$)	27.1	302.9	3.43	11303.5
10 ($\Psi5$)	189.5	740.5	1.06	5594.0

iii. Umbrella sampling biased GLIC simulation data

Analysis on a more complex system is also presented based on umbrella sampling MD simulations carried out previously for an ion passing through a GLIC channel (Fig. 5, top).^{31,32} This system is of particular interest for our method as the data here was generated from a series of harmonically biased simulations with Hamiltonian replica exchange^{3,9} steps attempted every 200 fs (simulation timestep 1 fs, full simulation parameters are described in detail in the referenced publications), unlike the unbiased data presented in the above examples. As mentioned previously, our fitting method can be applied to biased data provided one has an unbiasing procedure with which to construct an MSM. In this application since our data was generated using umbrella sampling, an MSM can be constructed using the DHAM method of Rosta and Hummer.¹⁷

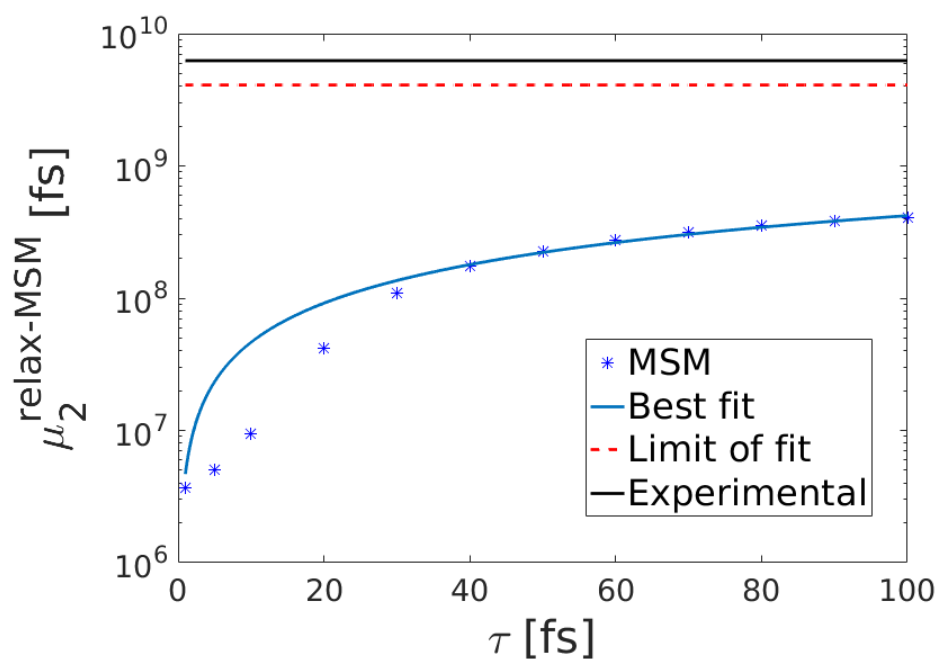
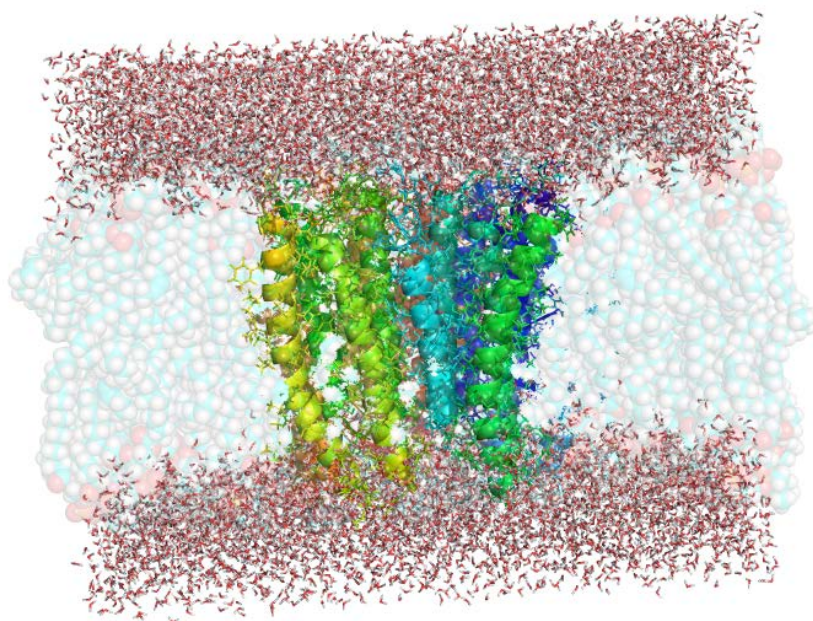


Figure 5: **Top.** Representation of the simulation system with water (sticks), lipids (spheres) and the protein (sticks and cartoon). **Bottom.** Relaxation timescales (blue symbols) of MSMs for ion channel simulation data where the reaction coordinate is the distance from the center of the membrane. The best fit to the data points (blue line) has a limiting value which is greatly in excess of the longest accessible LT (red dashed line). The best fit gets much closer to the experimental value (black line) than the relaxation time at the longest accessible lag time. The fitting was performed with a least squares approach on LTs in the range of 30 to 100 fs.

To ensure adequate sampling of transitions using DHAM and taking into account that Hamiltonian replica exchange was also used in the original simulations with 200 fs exchange frequency, we constructed our MSMs at up to 100 fs lagtimes. By constructing an MSM at LT 100 fs we obtained a relaxation time of 4.08×10^8 fs which is more than an order of magnitude smaller than the estimate of 6.25×10^9 fs (corresponding to a rate of $1.6 \times 10^5 \text{ s}^{-1}$) obtained from experimental data by Zhu and Hummer.^{31,32}

The HMM formalism cannot be used here since it is only valid for unbiased simulation data. However, the best fit of Eq. (15) to the data (Fig. 5, bottom) is quite good although there are some deviations at short LTs, which is likely due to contributions from the other relaxation timescales. In general, as the RT in this case is several orders of magnitude larger than the range of LTs used in the fit, we can only expect a good fit if A_2 dominates also at short times according to Eq. (12): $A_2 \gg A_{i>2}$. The best fit corresponds to a limiting value of 4.09×10^9 fs that is considerably larger than the MSM RT value at the longest accessible LT, and agrees much more closely to the value obtained from experiment.

IV. CONCLUSIONS

In the examples provided, the simple expression for the RT behavior gives values that are appreciably larger than the value calculated at the longest accessible LTs. In the analytic potential example, the hidden Markov model formalism provides similar values to the limit of the MSM fit. However, the timescale plots generated via HMMs have a different functional dependence on the LT, are not variational, and therefore the ambiguity of choosing the LT and corresponding RT remains when using this formalism.

In the example of the Ala5 MD data, all 10 Ramachandran angles were analyzed independently. We found that the limiting RTs agree closely with the values obtained from the HMMs, even though the largest accessible RT used for the fitting is less than half for all the ϕ angles. These limiting RTs also agree well to the analysis carried out using all angles simultaneously by Buchete and Hummer. Our approach using finely discretized reaction coordinates and a limiting fit therefore offers an easier to implement alternative to multidimensional analysis using coarse grained metastable MSMs. The HMM results suffer from having a greater numerical sensitivity at longer LTs due to insufficient sampling of the system, resulting in larger error bars than do the MSM derived RTs.

We can also determine limiting RT values in biased Umbrella Sampling simulations, where the HMM formalism cannot be applied. This is of particular use in the case of replica exchange data where the accessible lag times are limited by the exchange attempt frequency, regardless of how much simulation data is generated. Here, we need an additional assumption that the MSM eigenvectors are very close to the exact eigenvectors. In the ion channel example presented here, the derived expression fits well to the RT curve and finds a significantly larger RT value that agrees well with experimental ion crossing rates.

Our approach may also be useful in estimating the quality of definition for the reaction coordinate used in a set of simulation data. If the limiting RT is significantly larger than the value calculated at the longest accessible LT, then it will be necessary to generate more data to adequately satisfy the CK condition, and it can also suggest that the reaction coordinate is not complete. Our fitted ϵ parameters thus can also be used to provide a quantitative comparison between different reaction coordinates in terms of how well they are able to capture the slowest process and how close they are to the exact second eigenvector.

Our approach is straightforward to implement and does not require additional analysis besides determining MSMs at different LTs that are typically already done. One ambiguity lies in the precise choice of LTs at which to perform the fitting on the RTs. We typically chose the regions such that the RT is numerically stable with good statistics (i.e. not too large values) and long enough to allow the second eigenvalue to dominate and observe the functional dependence described by our fitting equation (i.e. not too small values). The optimal choice of data and range of LTs to fit our equations on will be subject of future work.

Our examples presented here suggest that our method has a similar accuracy in estimating the RTs than the HMM method for analytical examples, yet it is easier to implement and readily available once MSMs are constructed. Our method can also provide better RT estimates in cases when the HMM method is not applicable, such as biased simulations as shown in the ion-channel example and systems where the relaxation time is too long for sufficient sampling. Future work will involve developing HMM methods for biased simulation data.

V. SUPPLEMENTARY INFORMATION

See Supporting Information for relaxation timescale plots of the pentalanine Ramachandran angles.

VI. ACKNOWLEDGEMENTS

The authors are indebted to Drs. Attila Szabo (National Institutes of Health), Nicolae-Viorel Buchete (University College Dublin) and Karl Heinz Hoffmann (Technical University Chemnitz) for many stimulating discussions and their help with this project. A.K. is supported by the EPSRC Centre for Doctoral Training in Cross-Disciplinary Approaches to Non-Equilibrium Systems (CANES, EP/L015854/1). E.R. gratefully acknowledges financial support from the EPSRC (grant number EP/N020669/1).

VII. REFERENCES

1. Sirur, A., De Sancho, D. & Best, R. B. Markov state models of protein misfolding. *J. Chem. Phys.* **144**, 75101 (2016).
2. Cossio, P., Hummer, G. & Szabo, A. On artifacts in single-molecule force spectroscopy. *Proc. Natl. Acad. Sci.* **112**, 14248–14253 (2015).
3. Leahy, C. T., Kells, A., Hummer, G., Buchete, N.-V. & Rosta, E. Peptide dimerization-dissociation rates from replica exchange molecular dynamics. *J. Chem. Phys.* **147**, (2017).
4. Leahy, C. T., Murphy, R. D., Hummer, G., Rosta, E. & Buchete, N. V. Coarse Master Equations for Binding Kinetics of Amyloid Peptide Dimers. *J. Phys. Chem. Lett.* **7**, 2676–2682 (2016).
5. 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.* **126**, 1–17 (2007).
6. Schütte, C., Nielsen, A. & Weber, M. Markov state models and molecular alchemy. *Mol. Phys.* **5**, 1–10 (2014).
7. Wu, H. *et al.* Variational approximation of molecular kinetics from short off-equilibrium simulations. (2016).
8. Buchete, N.-V. & Hummer, G. Coarse master equations for peptide folding dynamics. *J. Phys. Chem. B* **112**, 6057–6069 (2008).
9. Buchete, N.-V. & Hummer, G. Peptide folding kinetics from replica exchange molecular dynamics. *Phys. Rev. E - Stat. Nonlinear, Soft Matter Phys.* **77**, 1–4 (2008).

10. 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.* **131**, (2009).
11. Sarich, M., Noé, F. & Schütte, C. On the Approximation Quality of Markov State Models. *Multiscale Model. Simul.* **8**, 1154–1177 (2010).
12. Noé, F. & Fischer, S. Transition networks for modeling the kinetics of conformational change in macromolecules. *Curr. Opin. Struct. Biol.* **18**, 154–162 (2008).
13. McGibbon, R. T., Husic, B. E. & Pande, V. S. Identification of simple reaction coordinates from complex dynamics. *J. Chem. Phys.* **146**, (2017).
14. Hummer, G. & Szabo, A. Optimal Dimensionality Reduction of Multistate Kinetic and Markov- State Models. (2014).
15. Martini, L. *et al.* Variational identification of markovian transition states. *Phys. Rev. X* **7**, (2017).
16. Bowman, G. R. *An Introduction to Markov State Models and Their Application to Long Timescale Molecular Simulation.*
17. Rosta, E. & Hummer, G. Free energies from dynamic weighted histogram analysis using unbiased Markov state model.
18. Stelzl, L. S., Kells, A., Rosta, E. & Hummer, G. Dynamic Histogram Analysis To Determine Free Energies and Rates from Biased Simulations. *J. Chem. Theory Comput.* **13**, (2017).
19. Donati, L., Hartmann, C. & Keller, B. G. Girsanov reweighting for path ensembles and Markov state models. *J. Chem. Phys.* **146**, (2017).
20. Noé, F., Schütte, C., Vanden-Eijnden, E., Reich, L. & Weikl, T. R. Constructing the equilibrium ensemble of folding pathways from short off-equilibrium simulations. *Proc. Natl. Acad. Sci. U. S. A.* **106**, 19011–19016 (2009).
21. Casey, F. P., Waterfall, J. J., Gutenkunst, R. N., Myers, C. R. & Sethna, J. P. Variational method for estimating the rate of convergence of Markov-chain Monte Carlo algorithms. *Phys. Rev. E* **78**, 46704 (2008).
22. Kühn, R. & Sollich, P. Spectra of empirical auto-covariance matrices. *Epl* **99**, 1–6 (2012).
23. Hartigan J. & Wong M. A K-Means Clustering Algorithm. *Source J. R. Stat. Soc. Ser. C (Applied Stat.* **28**, 100–108 (1979).
24. Sculley, D. Web-scale k-means clustering. *Proc. 19th Int. Conf. World wide web - WWW '10* 1177 (2010). doi:10.1145/1772690.1772862

25. Bicout, D. J. & Szabo, A. Electron transfer reaction dynamics in non-Debye solvents. *J Chem Phys* **109**, 2325–2338 (1998).
26. Buchner, G. S., Murphy, R. D., Buchete, N.-V. & Kubelka, J. Dynamics of protein folding: Probing the kinetic network of folding-unfolding transitions with experiment and theory. *Biochim. Biophys. Acta - Proteins Proteomics* **1814**, 1001–1020 (2011).
27. Baum, L. & Petrie, T. Statistical Inference for Probabilistic Functions of Finite State Markov Chains. **37**, 1554–1563. (1966).
28. Noé, F., Wu, H., Prinz, J. H. & Plattner, N. Projected and hidden Markov models for calculating kinetics and metastable states of complex molecules. *J. Chem. Phys.* **139**, (2013).
29. Scherer, M. K. *et al.* PyEMMA 2: A Software Package for Estimation, Validation, and Analysis of Markov Models. *J. Chem. Theory Comput.* **11**, 5525–5542 (2015).
30. Gillespie, D. T. A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *J. Comput. Phys.* **22**, 403–434 (1976).
31. Zhu, F. & Hummer, G. Pore opening and closing of a pentameric ligand-gated ion channel. *Proc. Natl. Acad. Sci.* **107**, 19814–19819 (2010).
32. Zhu, F. & Hummer, G. Theory and simulation of ion conduction in the pentameric GLIC channel. *J. Chem. Theory Comput.* **8**, 3759–3768 (2012).