Variational Identification of Markovian Transition States

Linda Martini,[1] Adam Kells,[1] Roberto Covino,[2] Gerhard Hummer,[2,3] Nicolae-Viorel Buchete,[4] and Edina Rosta[1,*]

[1]*Department of Chemistry, King's College London, SE1 1DB London, United Kingdom*
[2]*Department of Theoretical Biophysics, Max Planck Institute of Biophysics,*
60438 Frankfurt am Main, Germany
[3]*Institute of Biophysics, Goethe University Frankfurt, 60438 Frankfurt am Main, Germany*
[4]*School of Physics and Institute for Discovery, University College Dublin, Dublin 4, Ireland*

(Received 16 June 2016; revised manuscript received 17 March 2017; published 28 September 2017)

We present a method that enables the identification and analysis of conformational Markovian transition states from atomistic or coarse-grained molecular dynamics (MD) trajectories. Our algorithm is presented by using both analytical models and examples from MD simulations of the benchmark system helix-forming peptide Ala$_5$, and of larger, biomedically important systems: the 15-lipoxygenase-2 enzyme (15-LOX-2), the epidermal growth factor receptor (EGFR) protein, and the Mga2 fungal transcription factor. The analysis of 15-LOX-2 uses data generated exclusively from biased umbrella sampling simulations carried out at the hybrid *ab initio* density functional theory (DFT) quantum mechanics/molecular mechanics (QM/MM) level of theory. In all cases, our method automatically identifies the corresponding transition states and metastable conformations in a variationally optimal way, with the input of a set of relevant coordinates, by accurately reproducing the intrinsic slowest relaxation rate of each system. Our approach offers a general yet easy-to-implement analysis method that provides unique insight into the molecular mechanism and the rare but crucial (i.e., rate-limiting) transition states occurring along conformational transition paths in complex dynamical systems such as molecular trajectories.

DOI: 10.1103/PhysRevX.7.031060

Subject Areas: Chemical Physics,
Computational Physics

I. INTRODUCTION

Recent advances in both parallelizable computational software and the development of highly efficient supercomputers have extended the time scale accessible to atomistic molecular dynamics (MD) of biomolecules with explicit solvent representations to simulations up to the order of milliseconds [1–3]. While this enables state-of-the-art computational modeling of complex molecular processes, such as folding and binding [4], the vast amount of complex, high-dimensional data obtained from these simulations requires novel analysis methods, on the one hand, to make use of all the available information and, on the other hand, to extract comprehensible and relevant information. Several statistical analysis methods focusing on transition paths and probability distributions have been successfully applied, including transition paths sampling (TPS) [5–8], milestoning [9], Markov state models (MSMs) [10–12], and other methods using the commitment probability distribution [13]. Geometry-optimization-based algorithms such as discrete path sampling also produce kinetic transition networks, identifying key rate-determining transition states in a systematic, automated fashion [14,15].

Here, we focus on the use of MSMs, which make it possible to directly extract thermodynamic and kinetic information from MD trajectories, such as free energy profiles, equilibrium probabilities, and transition rates. MSMs have also proven to be useful tools to analyze data from both experiments and simulations [16–20]. A variety of MSM-based methods have been proposed that aggregate the conformational state space, projecting along certain reaction coordinates (RCs) of interest into macrostates, using approaches that can be implemented automatically [21–24]. However, while some of the most successful and widely used clustering algorithms such as PCCA+ [25] and its advances [26,27] or alternatives [28] are designed to identify metastable states (e.g., in protein folding terms: the folded, the unfolded, and long-lived intermediate states), none is aimed specifically at the automatic, reliable identification and characterization of the rate-determining transition states (TS), which are crucial for the thorough understanding of the underlying molecular mechanisms and which control the overall kinetics of the system. Existing methods to identify the TS in networks define optimal cuts through the network, which carry maximum flux [29–31]. In TPS, the TS is defined as a region with ½ commitment probability, using the committor as the optimal one-dimensional reaction coordinate [32–34] that defines the probability to reach the reactant state before reaching the product state. However, in complex processes, determining the committor probability for

[*] edina.rosta@kcl.ac.uk

multiple stable states is not straightforward. It is not clear how many metastable states (MSs) are important to describe the overall slow dynamics or where the key kinetic bottleneck is located. To clearly identify the TS as an ensemble, and not only as a zero-measure separating surface between MSs, we aim to define the TS ensemble as a node of a variationally optimal minimal required network consisting both of MS and the TS Markov states.

Here, we propose a method for the automatic identification and analysis of both MS and TS conformational regions, by aiming to optimally construct MSMs via the slowest relaxation time using a set of discretized RCs. The use of key RCs has long been a successful approach in many of the major enhanced sampling methods [35–38], also as a basis for a more intuitive understanding of the data [39]. Our algorithm enumerates all possible clusters that could be formed along an ordered set of states for each RC, and it selects the optimal clustering that maximizes the slowest relaxation time of the coarse network. This is both a physically and mathematically meaningful variational optimization process [40,41], as the fully discretized system's relaxation time provides an upper bound to the slowest relaxation time, and the better the coarse graining, the closer our variational objective function will be to this ideal value. Importantly, by increasing the number of coarse-grained states used to model the system, we systematically identify key MSs, until we exhaust the number of relevant MSs, and the first optimal TS is identified. This approach also enables users to find and analyze the minimal required number of MSs relevant to the kinetics of the underlying dynamic process.

II. TIME-SCALE OPTIMIZATION CLUSTERING

The slowest relaxation times have been widely used to measure the discretization error in most MSM applications [10–12,40], to validate the convergence of the model, and to extract the true Markovian time scales for the system. The slowest relaxation time is therefore a mathematically and physically meaningful optimization function that can be used to determine coarse-grained kinetic networks [25,42]. This is justified mathematically because any lumping of states into coarse-grained states decreases the measured relaxation times of the intrinsic dynamics of the model. A formal proof can be given via a variational theorem for correlation functions [41,43]. Considering our main aim of most accurately estimating the slowest relaxation process, we target our proposed clustering method towards the maximization of the slowest extracted relaxation time [25,42].

A. One-dimensional clustering

Complex, multidimensional trajectories are often analyzed using their projection on simpler 1D RCs. Here, we start by assuming that a set of RCs is available, which accounts for the slow dynamics of the system. Any continuous 1D RC can be divided into adequately small discrete bins or "microstates" (denoted here as $s_1, \ldots, s_{N_\mu}$, where N_μ is their total number), which capture the same intrinsic dynamics as the continuous RC. In practice, a 1D RC, $x(t)$, is often binned in equidistant microstate intervals over RC windows of length Δx_μ, such that, for a trajectory that samples the interval $[x_{\min}, x_{\max}]$, we have $N_\mu = (x_{\max} - x_{\min})/\Delta x_\mu$. This discretization has an intrinsic ordering inherited from the continuous RC.

Traditionally, MSMs have been constructed to derive a more accurate equilibrium distribution from multiple short trajectories and also to obtain kinetic information about the system [10–12,21,22,40,44,45]. These advances led to successful use in a number of academic and commercial drug discovery projects and studies of protein folding, etc. These traditional MSMs, however, are implemented using clustering (e.g., k-means, k-medoids, k-centers, etc. [22,46]) into metastable states only, thus lacking the resolution required for identifying TSs. We have been using discretized continuous RCs [47,48], corresponding to our MSM built from microstates termed here. These make it possible to construct MSMs with sufficiently fine resolution to be able to capture TSs and to integrate biased molecular simulation data.

Definition of cluster boundaries for ordered sets of microstates.—To reduce the dimensionality, the typically large number N_μ of microstates, we cluster them into coarse-grained "macrostates" (M-states). Throughout the paper, we refer to microstates as states, which correspond to our full-dimensional MSM, and M-states as the optimally coarse-grained network, obtained by lumping together microstates, where we aim to identify key MS and TS states.

Given the projection of a trajectory on a 1D RC, $x(t)$, for clustering it into M sorted M states over a finite domain (i.e., with no periodic boundaries), b_i cluster boundaries ($i \in \{1, 2, \ldots, M-1\}$) are needed. Discretizing $x(t)$ into M macrostates in the interval $[x_{\min}, x_{\max}]$ requires $M-1$ boundaries, such that the M macrostates (denoted by S_i) cover the following intervals: $S_1 = [x_{\min}, x_{\min} + b_1 \Delta x_\mu)$, $S_2 = [x_{\min} + b_1 \Delta x_\mu, x_{\min} + b_2 \Delta x_\mu)$, $\ldots, S_{M-1} = [x_{\min} + b_{M-2} \Delta x_\mu, x_{\min} + b_{M-1} \Delta x_\mu)$, $S_M = [x_{\min} + b_{M-1} \Delta x_\mu, x_{\max}]$. Here, we identify all possible boundary positions $(b_1, \ldots, b_{M-1}$, where $b_{i-1} < b_i < b_{i+1} \in \{1, \ldots, N_\mu - 1\})$, and thus all possible coarse-grained models with M states. For example, the first boundary can be placed at $N_\mu - 1$ positions, separating all microstates into two coarse-grained M states. For three states, we consider $(\substack{N_\mu - 1 \\ 2}) = (N_\mu - 1)(N_\mu - 2)/2$ possibilities to place the two boundaries required for identifying a three-state coarse-grained system. In general, there are $\binom{N_\mu - 1}{M-1} \approx O({N_\mu}^{M-1})$ possibilities to place the boundaries between M states, leading to a polynomial scaling algorithm in N_μ. Next, for each possible boundary position, we define a coarse-grained Markov or transition probability matrix $\mathbf{M}^{\mathrm{red}}(\tau)$, with $M_{kl}^{\mathrm{red}}(\tau)$ representing the probability that the system starting from microstate k will be in microstate l after time τ and use as our objective function the variational slowest relaxation time ($t_2 = \tau/\ln\lambda_2$, with the lag time τ) corresponding to the second largest eigenvalue (λ_2, with $\lambda_1 = 1$). Thus, we maximize t_2 to identify the optimal coarse graining ($\max\{t_2|b_1, \ldots, b_{M-1}\}$) based on the variational principle discussed above. The M-state coarse-grained system will generally not be fully Markovian; therefore, we need to make some further assumptions regarding its kinetics, as described below.

Definition of the coarse-grained dynamics.—To define a Markovian dynamics of the coarse-grained system, besides the basic requirements regarding equilibrium, we are also able to impose some additional constraints. We describe the original dynamics with an N_μ-dimensional rate matrix $\mathbf{K}$, where K_{kl} corresponds to the rate from k to l, or alternatively, we can also use a full-dimensional Markov matrix $\mathbf{M}^{\mathrm{full}}(\tau) = e^{\mathbf{K}^{\mathrm{full}}\tau}$ at lag time τ. In general, we want to ensure that the equilibrium populations are exactly maintained in the reduced system's dynamics. Thus, all coarse-grained states S_i, $i \in \{1, \ldots, M\}$ have to preserve the total equilibrium population of the corresponding microstates with indices in the range of $[b_{i-1}+1, \ldots, b_i]$:

$$P_i^{\mathrm{eq}} = \sum_{k=b_{i-1}+1}^{b_i} p_k^{\mathrm{eq}}, \tag{1}$$

where $b_0 = 0$ and $b_M = N_\mu$, P_i^{eq} is the equilibrium population of the coarse-grained state S_i obtained from the normalized first left eigenvector of $\mathbf{K}^{\mathrm{full}}$, or the corresponding Markov matrix $\mathbf{M}^{\mathrm{full}}(\tau)$ at lag time τ, and p_k^{eq} is the kth component of the normalized first left eigenvector corresponding to eigenvalue 0 of $\mathbf{K}^{\mathrm{full}}$ [or 1 of $\mathbf{M}^{\mathrm{full}}(\tau)$].

In addition to projecting the correct equilibrium populations, we can consider two alternative methods to build the kinetics of the coarse-grained system: (i) a local equilibrium (LE) lag-time-dependent method, or (ii) the Hummer-Szabo definition (HS) for a lag-time-independent coarse-rate matrix [42]. In both cases, the main criteria for setting the kinetics of the reduced system are defined via correlation functions that allow us to count the number of transitions between different states. The number correlation function $C_{ij}(t)$ is the probability that the system is in coarse-grained state S_i at time 0 and in coarse-grained state S_j at time t, as would be seen in a long-equilibrium run. It can be written in terms of the rate matrix exponential propagating a conditional equilibrium in S_i and collecting the probability in S_j, written using a bra-ket notation as

$$\begin{aligned} C_{ij}^{\mathrm{red,LE}}(\tau) &= \langle \mathbf{e}_j | \exp(\mathbf{K}^{\mathrm{full},T}\tau) | \boldsymbol{\pi}_i^{\mathrm{eq}} \rangle \\ &= \langle \mathbf{e}_j | \mathbf{M}^{\mathrm{full},T}(\tau) | \boldsymbol{\pi}_i^{\mathrm{eq}} \rangle, \end{aligned} \tag{2}$$

where the projection (i.e., coarse-graining) vector $\langle \mathbf{e}_i| = (0, \ldots, 0, 1, \ldots, 1, 0, \ldots, 0)$ is an N_μ-dimensional column vector, having nonzero elements only for indices in the range $[b_{i-1}+1, \ldots, b_i]$ (with $b_0 = 0$ and $b_M = N_\mu$), that corresponds to coarse-grained state S_i. Analogously, the renormalized equilibrium distribution for the components of coarse state S_i is column vector $\boldsymbol{\pi}_i^{\mathrm{eq}} = (0, \ldots, 0, p_{b_{i-1}+1}^{\mathrm{eq}}, \ldots, p_{b_i}^{\mathrm{eq}}, 0, \ldots, 0)^T / P_i^{\mathrm{eq}}$, with $\mathbf{p}^{\mathrm{eq}}$ being the first normalized left eigenvector, as before. Here, $\mathbf{K}^{\mathrm{full},T}$ and $\mathbf{M}^{\mathrm{full},T}$ are the transposed full rate and Markov matrices, respectively. Therefore, $C_{ij}^{\mathrm{red,LE}}(\tau)$ represents the probability that starting from state S_i, the system will be in state S_j after lag time τ, corresponding to Kronecker delta $\lim_{\tau\to 0} C_{ij}^{\mathrm{red,LE}}(\tau) = \delta_{ij}$ and $\lim_{\tau\to\infty} C_{ij}^{\mathrm{red,LE}}(\tau) = P_j^{\mathrm{eq}}$.

The LE approximation sets the coarse-grained transition probability matrix elements $M_{ij}^{\mathrm{red,LE}}(\tau)$ (i.e., defined as the transition probability to be at state S_j after lag time τ, starting from S_i) equal to the correlation function at the chosen lag time τ:

$$\begin{aligned} M_{ij}^{\mathrm{red,LE}}(\tau) &\equiv C_{ij}^{\mathrm{red,LE}}(\tau) \\ &= \sum_{k=b_{i-1}+1}^{b_i} \sum_{l=b_{j-1}+1}^{b_j} \pi_k^{\mathrm{eq}} C_{kl}^{\mathrm{full}}(\tau) \\ &= \frac{1}{P_i^{\mathrm{eq}}} \sum_{k=b_{i-1}+1}^{b_i} \sum_{l=b_{j-1}+1}^{b_j} p_k^{\mathrm{eq}} C_{kl}^{\mathrm{full}}(\tau). \end{aligned} \tag{3}$$

This ensures that the coarse-grained dynamics reproduces exactly the average number of transition counts from the full microstate dynamics [$P_i^{\mathrm{eq}} M_{ij}^{\mathrm{red,LE}}(\tau) = P_i^{\mathrm{eq}} C_{ij}^{\mathrm{red,LE}}(\tau) = \sum_{k=b_{i-1}+1}^{b_i} \sum_{l=b_{j-1}+1}^{b_j} p_k^{\mathrm{eq}} C_{kl}^{\mathrm{full}}(\tau)$] between the corresponding coarse-grained M states at lag time τ, with respect to the full N_μ-dimensional correlation function $C_{kl}^{\mathrm{full}}(\tau)$ for microstates k, $l \in \{1, \ldots, N_\mu\}$. This equation can be intuitively interpreted as simply calculating the reduced matrix elements by (equilibrium) averaging over the elements of the full matrix, which are being clustered. Note that, in this work, we do not discuss the zero lag-time limit within the LE case. This would correspond to a reduced rate matrix (infinitesimal generator matrix), instead of a Markov transition probability matrix, calculated via the well-known local equilibrium approximation.

Using molecular simulation data, we can calculate the transition count matrix $\mathbf{T}(\tau)$, with elements $T_{ij}(\tau)$ corresponding to the number of transitions observed from S_i to S_j [or s_k to s_l microstates, for estimating $M_{kl}^{\mathrm{full}}(\tau)$] along the trajectory between any two data points separated by time τ. Subsequently, the corresponding transition probability or Markov matrix (both the full-dimensional and the coarse-grained ones, depending on whether microstates or macrostates are used to count the transitions, respectively) can be numerically estimated with the normalized elements defined as $M_{ij}(\tau) \approx T_{ij}(\tau)/\sum_{k=1}^{M} T_{ik}(\tau)$. To ensure detailed, balance $M_{ij}(\tau)p_i = M_{ji}(\tau)p_j$, we can also use a reversible estimator via an iterative approach, or for transition counts from long equilibrium simulations, we can simply symmetrize the count matrix: $\mathbf{T}^{\mathrm{symm}}(\tau) \coloneqq (\mathbf{T}(\tau) + \mathbf{T}^T(\tau))/2$ [44,45,49,50].

To avoid the arbitrary choice in the lag time and seek approximate kinetics that considers a range of lag times, we also used the HS method to define a coarse-grained rate matrix $\mathbf{K}^{\mathrm{red,HS}}$ [42] and the corresponding correlation matrices $\mathbf{C}^{\mathrm{red,HS}}(\tau) = \mathbf{M}^{\mathrm{red,HS}}(\tau) = \exp(\mathbf{K}^{\mathrm{red,HS}}\tau)$. Rather than enforcing the correlation functions to be equal at a given lag time (as in the LE case seen previously), here we enforce that the integral over all possible lag times of the correlation functions [51–54] be equal and thus define a lag-time-independent reduced rate matrix via

$$\begin{aligned}
&\int_0^\infty d\tau (C_{ij}^{\mathrm{red,HS}}(\tau) - P_j^{\mathrm{eq}}) \\
&\quad = \int_0^\infty d\tau ([\exp(\mathbf{K}^{\mathrm{red,HS}}\tau)]_{ij} - P_j^{\mathrm{eq}}) \\
&\quad = \int_0^\infty d\tau \sum_{k=b_{i-1}+1}^{b_i} \sum_{l=b_{j-1}+1}^{b_j} (\pi_k^{\mathrm{eq}} C_{kl}^{\mathrm{full}}(\tau) - p_l^{\mathrm{eq}}). \qquad (4)
\end{aligned}$$

The rate matrix $\mathbf{K}^{\mathrm{red,HS}}$ can be further expressed via Eq. (8) in Ref. [42]:

$$\begin{aligned}
&(\mathbf{P}_{\mathrm{eq}}\mathbf{1}_M^T - (\mathbf{K}^{\mathrm{red,HS}})^T)_{ji}^{-1} P_i^{\mathrm{eq}} \\
&\quad = \sum_{k=b_{i-1}+1}^{b_i} \sum_{l=b_{j-1}+1}^{b_j} (\mathbf{p}_{\mathrm{eq}}\mathbf{1}_{N_\mu}^T - \mathbf{K}^{\mathrm{full},T})_{lk}^{-1} p_k^{\mathrm{eq}} \qquad (5)
\end{aligned}$$

which leads to the explicit working equation via Eq. (12) in Ref. [42]:

$$(\mathbf{K}^{\mathrm{red,HS}})^T = \mathbf{P}_{\mathrm{eq}}\mathbf{1}_M^T - \mathbf{D}_M(\mathbf{A}^T(\mathbf{p}_{\mathrm{eq}}\mathbf{1}_{N_\mu}^T - \mathbf{K}^{\mathrm{full},T})^{-1}\mathbf{D}_{N_\mu}\mathbf{A})^{-1}, \qquad (6)$$

where $A_{ki} = \begin{cases} 1 & \text{if } k \in S_i \\ 0 & \text{otherwise} \end{cases}$ defines the partitioning, and $\mathbf{D}$'s are diagonal matrices containing the equilibrium populations for microstates k, $l \in \{1, \ldots, N_\mu\}$: $(D_{N_\mu})_{kl} = p_k^{\mathrm{eq}}\delta_{kl}$ or M states i, $j \in \{1, \ldots, M\}$: $(D_M)_{ij} = P_i^{\mathrm{eq}}\delta_{ij}$.

The HS coarse dynamics is preferred, as no lag time is required, and for our analytical examples below, we compare it to the LE coarse graining. The reduced M-dimensional rate matrix for each choice of a set of state boundaries b_i can thus be obtained using the lag-time-independent HS expressions. We chose the LE method for analyzing numerical simulation data because this corresponds to the standard MSM approaches using a lag-time-dependent estimator count matrix for the transition probability matrix. Approximate reduced Markov matrices can also be built using alternative methods; see, for example, Fačkovec *et al.* [19].

We also demonstrate that both the LE and the HS coarse graining lead to a variational maximum of t_2, and all our analytical examples show that this maximum is limited by the exact second eigenvalue of the full-dimensional system (see, e.g., Fig. S2B). Analytically, the exact relaxation times are obtained at infinitely long lag times ($\tau \to \infty$) for the LE matrices [in practice, however, the matrices become degenerate with $M_{ij}^{\mathrm{red}}(\infty) = P_j^{\mathrm{eq}}$, and numerically, the diagonalization cannot be carried out at very long lag times]. While there is no previous formal proof that the HS matrices also correspond to a variational maximum for the t_2 relaxation times as a function of boundaries between coarse-grained states (Fig. S2B), this is justified by the fact that the integral on the right-hand side of Eq. (4) can be rewritten by introducing the LE correlation matrices at different lag times using Eq. (3) to obtain the following relationship:

$$\int_0^\infty d\tau (C_{ij}^{\mathrm{red,HS}}(\tau) - P_j^{\mathrm{eq}}) = \int_0^\infty d\tau (C_{ij}^{\mathrm{red,LE}}(\tau) - P_j^{\mathrm{eq}}). \qquad (7)$$

Therefore, the HS eigenvalue is also bound, as each of the LE eigenvalues is bound by the exact t_2. This is demonstrated by the slightly lower t_2 HS relaxation time as compared with the LE eigenvalues at very long lag times (Fig. S2B), and the exact eigenvalue is obtained only for the full-dimensional matrix.

Exhaustive cluster boundary search.—The optimal boundary positions will be determined via the slowest relaxation time obtained by solving the eigenvalue problem of the Markov matrices $\mathbf{M}^{\mathrm{red,LE}}(\tau)$ using the LE method, or rate matrices $\mathbf{K}^{\mathrm{red,HS}}$ using the HS method for each possible set of boundaries (Fig. 1). The optimal M-state clustering corresponds to the best $M-1$ boundary positions, for which the slowest relaxation time t_2 is maximum, and the boundary positions have been exhaustively searched at all possible positions. Our algorithm therefore does not require higher-order eigenvectors and eigenvalues besides the second eigenvalue (unlike, e.g., PCCA+ as used in Ref. [42]). Previous information from the optimal $M-1$-state clustering is therefore not needed either to determine the optimal M-state clustering. However, for large M, $O(N_\mu^{M-1})$

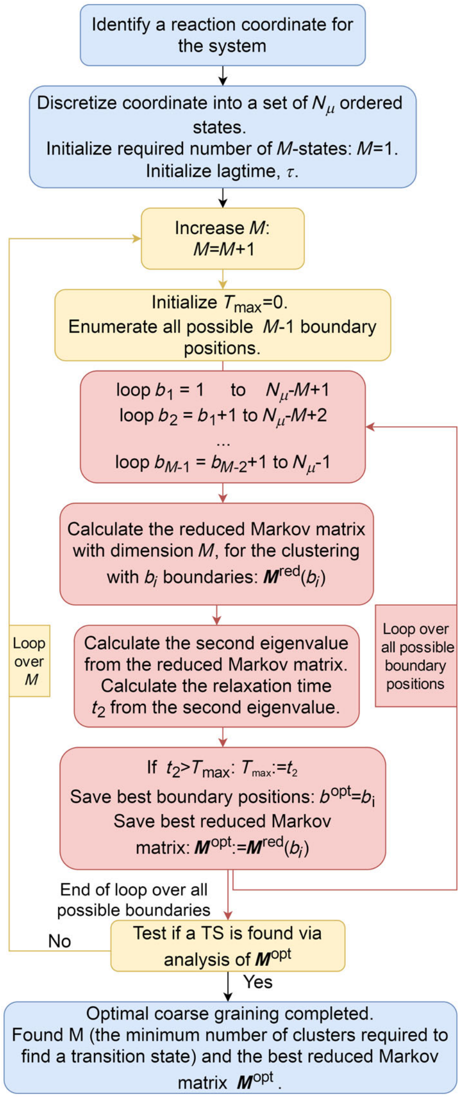
Identify a reaction coordinate for the system
Discretize coordinate into a set of N_μ ordered states.
Initialize required number of M-states: M=1.
Initialize lagtime, τ.
Increase M: M=M+1
Initialize $T_{\max}$=0.
Enumerate all possible M-1 boundary positions.
loop b_1 = 1 to N_μ-M+1
loop b_2 = b_1+1 to N_μ-M+2
...
loop b_{M-1} = b_{M-2}+1 to N_μ-1
Calculate the reduced Markov matrix with dimension M, for the clustering with b_i boundaries: $\boldsymbol{M}^{\text{red}}(b_i)$
Calculate the second eigenvalue from the reduced Markov matrix. Calculate the relaxation time t_2 from the second eigenvalue.
If t_2>$T_{\max}$: $T_{\max}$:=t_2
Save best boundary positions: b^{opt}=b_i
Save best reduced Markov matrix: $\boldsymbol{M}^{\text{opt}}$:=$\boldsymbol{M}^{\text{red}}(b_i)$
Loop over M
Loop over all possible boundary positions
End of loop over all possible boundaries
Test if a TS is found via analysis of $\boldsymbol{M}^{\text{opt}}$
No
Yes
Optimal coarse graining completed.
Found M (the minimum number of clusters required to find a transition state) and the best reduced Markov matrix $\boldsymbol{M}^{\text{opt}}$.

FIG. 1. Flowchart illustrating our 1D coarse-graining algorithm for the initial N_μ states, defined along a 1D reaction coordinate, into an optimal set of M states.

polynomial scaling of the computation time may prove problematic, and here, we also implement an iterative algorithm by taking into account the previous optimal boundary positions. In this approximation, we only change two boundaries at a time, and we iteratively update all neighboring pairs of boundaries, until there is no further improvement in the second eigenvalue. We find that this algorithm arrives at the same optimum as the full exhaustive search in the applications we discuss in more detail in Sec. III.

Importantly, by increasing the number of M states one by one, we first obtain all the key MSs, corresponding to diagonally dominated reduced transition probability matrices with elements $M_{ij}^{\text{red}}(\tau)$. Our algorithm automatically evaluates the transition probabilities to neighboring states [$M_{ij}^{\text{red}}(\tau)$ with all $j \neq i$], and these are compared with the probability to remain in the same state [$M_{ii}^{\text{red}}(\tau)$], for appropriate lag times (see Ref. [55], Sec. III).

Definition of the TS.—In this work, we define a TS as an unstable, optimally coarse-grained state S_i from which the outgoing probabilities to at least two other states (e.g., S_m and S_n) are greater than the probability to stay in the same state [i.e., $\exists m,\ n \neq i$: $M_{im}^{\text{red}}(\tau) > M_{ii}^{\text{red}}(\tau)$ & $M_{in}^{\text{red}}(\tau) > M_{ii}^{\text{red}}(\tau)$]. This is an operational definition that requires an appropriate choice of lag time. We can verify the nature of our TS by checking that the largest fraction of the reactant flux goes through this state (see Ref. [55], Sec. I, Fig. S1) rather than around it, in accordance with the maximum flux criterion of Huo and Straub [56].

First and foremost, our TS is one of the states obtained from an optimal coarse graining. This distinguishes saddles from maxima because the maxima will not be important to contribute to the overall relaxation time of the system. Importantly, the TS will be the first nonmetastable state that will appear as one of the optimal states from our coarse graining that maximizes the overall relaxation time. Our TS is therefore an ensemble of unstable states that can be optimally separated from the metastable states to best describe the variational slowest relaxation time in the coarse-grained system.

This new TS definition, together with the exhaustive variational optimization of the states, is different from previously used TS definitions. Common definitions considered TSs with low-equilibrium populations and/or with committor probability values close to 0.5 [32–34]. We note here that the 0.5 value of the commitment probability alone cannot, in general, be used to define TSs in complex systems with more than two MSs (see the example illustrated in Fig. S8). Alternatively, more elaborate definitions required the evaluation of (free) energy gradients and/or of second derivatives, as done, e.g., in discrete path sampling methods [14,15]. Our definition is also different from TSs identified in the context of MSMs, where the initial starting point is a clustering aimed at identifying MSs. At our starting point, the microstates have a resolution higher than the TS itself, which is not the case in MSM-based methods. In that context, high-energy intermediates can be identified that are closest to TSs [57]; however, they represent local minima, which do not, in general, fulfil our TS definition. This method is also different from the hierarchical coarse-graining approach presented in Ref. [42], as we do not search for additional MSs and TS by maintaining previous MS boundaries but rather by exhaustively searching for new optimal additional boundaries that may change several of the previously identified MSs.

Thus, in our analysis, as we identify new M states, we test for the TS property as defined above, and in the process, we find all the minimally required MSs before a TS is found. In our algorithm (illustrated in the flowchart of Fig. 1), we may stop the search when a TS is found, or we can continue identifying additional coarse-grained states for an even more accurate representation, but with increased complexity. Noe *et al.* [58,59] noticed that as additional Markov states are introduced in the TS region, the accuracy of the underlying Markov discretization increases. In our approach of coarse graining the initial microstates, the observed characteristic qualitative change in the transition probabilities of the last M-state system, as we add new boundaries, allows us to identify both the TS and the minimal number of key MSs that have to be accounted for in the description of the process along each 1D RC and, more generally, in higher dimensions.

B. Multidimensional clustering

To generalize our approach for the analysis of a trajectory projected onto a Z-dimensional (Z-D) state space (i.e., described by Z 1D RCs), one has to deal with an exponentially large number of microstates, $N_\mu = N_{\mu,1\mathrm{D}}{}^Z$ (assuming $N_{\mu,1\mathrm{D}}$ bins for each coordinate). However, this finer-grained representation is expected to render the actual dynamics more faithfully (see examples in Secs. III B and III C). However, in this multidimensional case, there is no obvious intrinsic ordering of the microstates, and the number of all possible different coarse grainings is exponentially large, $M^{N_\mu}/M!$, which is unfeasible to sample systematically. To overcome the additional computational costs, we use here two approximations that may be applied independently: (i) an ordering of the microstates and (ii) a hierarchical coarse graining.

To reduce the number of possible ways the microstates can be grouped into coarse-grained states, we introduce an ordering of the microstates, such that only dividing boundaries would need to be identified. This will significantly reduce the cost of the exhaustive microstate-grouping search from exponential scaling to polynomial scaling, as shown for the 1D case. The ordering could be ideally defined via the commitment probability values of the microstates, as an ideal reaction coordinate [32–34]. To estimate the commitment probability automatically for multiple states, we use the right eigenvector of the full-dimensional (with M^{N_μ} states) rate matrix or Markov transition probability matrix (with an appropriate lag time τ), corresponding to the second λ_2 eigenvalue [25,42,60], analogously to the algorithm introduced by Szabo and coworkers [42,60]. The components of the second eigenvector define the ordering of the states, and we subsequently search all possible boundary positions that separate the coarse-grained states as discussed in the 1D case.

The ordering according to the second right eigenvector is not optimal in all cases because a better coarse-grained system may be obtained, for which the corresponding t_2 relaxation time is longer, but the elements of these optimal coarse-grained states are not adjacent within the ordering along the second eigenvector. Unfortunately, the number of possibilities to test all potential clusterings into coarse-grained states is exponentially large and thus unfeasible to evaluate. Alternative orderings, based on kinetic distances, e.g., diffusion maps [61,62] or the Nystrom kinetic lumping method [28], may be useful to obtain better ordering for optimal coarse graining. However, we verify, in all examples presented here, that single-state variations do not significantly affect the obtained clustering results. We also note that by using an ordering along the second right eigenvector, the continuity of the states is no longer ensured, and therefore, kinetically disconnected states might be directly adjacent to one another and grouped into the same coarse-grained state. This problem may also be overcome by using alternative approaches to identify an ordering that also takes into account the “kinetic vicinity” of the states (for example, diffusion maps [61,62] or the Nystrom kinetic lumping method [28]).

Here, we use an algorithm where the data are first clustered independently into M macrostates for each reaction coordinate, as described for 1D in Sec. II A. We assume, for simplicity, that the optimal number of M states is the same, M, for each dimension (an assumption that is relaxed later, with no loss of generality). Taking into account Z RCs, a discretized trajectory data with M^Z states is obtained. We can thus define M^Z unique states of the form of a direct product space, e.g., $\Sigma_1 = [S_1, S_1, S_1, S_1, S_1]$, $\Sigma_2 = [S_1, S_1, S_1, S_1, S_2]$, …, $\Sigma_{M^Z} = [S_M, S_M, S_M, S_M, S_M]$ when the system along the trajectory is in the M state S_1 in all $Z = 5$ RCs. The resulting M^Z macrostates are relabeled from Σ_1 to Σ_{M^Z}, and they define our first-level coarse-grained trajectory.

To perform a second-level coarse graining, the M states $\Sigma_1, …, \Sigma_{M^Z}$ are first ordered along the second right eigenvector components for a computationally feasible algorithm, as is also done in Ref. [42]. The resulting, sorted M-state trajectory is then analyzed one more time according to the 1D clustering procedure, to find a smaller optimal clustering coarse grained into L larger “Ω states” (i.e., with a maximum corresponding slowest relaxation time t_2). The resulting second-level coarse graining into L Ω states (i.e., with $2 < L < M^Z$) is denoted here by $\Omega_1, …, \Omega_L$. We then place a single M state out of its ordered position and try to group it with each of the different Ω states to find the optimal position that maximizes the relaxation time. We perform this procedure for every individual M state and repeat it until no further improvement is possible.

This final coarse-graining procedure considers all the RCs at the same time, globally. However, importantly, for complex systems with a large number of RCs, it may be necessary to perform the final clustering by hierarchically approximating the coarse graining of the M states into the Ω states $\Omega_1, …, \Omega_L$ by clustering the M states $S_1, …, S_M$ of only a few (e.g., two) RCs at a time in a stepwise manner and including additional RCs sequentially. This avoids the clustering of an unfeasibly large number of M states $\Sigma_1, …, \Sigma_{M^Z}$ at the same time.

C. Three-state division via a transition state

Having at least three states can lead to an optimal coarse graining that presents a TS, with significantly different properties than MSs. Considering any three-state coarse graining of a 1D problem leads to a reduced 3×3 rate matrix of a linear chain:

$$K = \begin{bmatrix} -K_{12} & K_{12} & 0 \\ K_{21} & -(K_{21}+K_{23}) & K_{23} \\ 0 & K_{32} & -K_{32} \end{bmatrix}, \tag{8}$$

with elements K_{ij} (defining the rate from i to j), depending on the coarse-graining boundaries. Optimal boundaries are obtained by minimizing the magnitude of the resulting second eigenvalue $\nu_2 = -1/t_2$ of $\mathbf{K}$ (with $\nu_1 = 0$), which generally corresponds to a separation into MSs. However, for two-state-like systems, the solution of the eigenvalue optimization problem will lead to boundaries that identify a TS. Thus, when the second state has a TS-like character, this corresponds to reduced rates that approximate ν_2 (see also Ref. [55], Sec. II) as

$$-\nu_2 = \frac{K_{32}K_{21} + K_{12}K_{23}}{K_{21} + K_{23}}. \tag{9}$$

Analogously, it has also been shown previously that adding states at the transition region reduces the discretization error in constructing Markov models [58,59]. Here, we use this variational approach to identify and characterize optimal TS and MSs.

III. APPLICATIONS

We present six applications here ordered by increasing data complexity. First, we evaluate the optimal three-state coarse graining on a set of analytical potentials in one dimension. Second, we analyze MC trajectories generated on a 2D analytical model potential, where we use an ordering of the states defined on a 2D grid based on the second eigenvector. We also apply our clustering algorithm to the analysis of MD trajectories obtained for four different systems: (i) umbrella-sampling biased QM/MM simulations for the first rate-limiting step of a lipoxygenase enzyme catalytic reaction, (ii) a system previously used for benchmarking molecular kinetics [57], the helix-forming peptide Ala$_5$, and transmembrane helix dimers of (iii) the epidermal growth factor receptor (EGFR) protein, and of (iv) the Mga2 fungal transcription factor. EGFR is implicated in various cancers, and it is a validated FDA-approved drug target, for which an MD trajectory of over 120 μs is available for its transmembrane segment [63]. We discuss this application in Ref. [55]. Mga2 is a sensor for lipid packing in the ER membrane, for which over 3.6 ms of MD trajectories of its transmembrane helix dimer are available using the MARTINI force field [64].

A. Analytical 1D model potential

We first tested our clustering method using Markov state models created for a set of analytical free-energy profiles that vary continuously and monotonically between two-state-like and three-state-like dynamics. To define the kinetic model underlying the analytical free-energy function $F(x)$, we construct an N_μ-state Markov chain. The corresponding full rate matrix $\mathbf{K}$ is given by

$$K_{i,i+1} = A \exp\left[\frac{F(x_i) - F(x_{i+1})}{2k_B T}\right], \tag{10a}$$

$$K_{i,i} = -\sum_{\substack{j=1 \\ j\neq i}}^{N_\mu} K_{i,j}. \tag{10b}$$

Here, A is the Arrhenius prefactor, k_B is Boltzmann's constant, and T is the absolute temperature. Note that the optimal clustering is invariant with respect to the value of the prefactor A (which only scales the time), and it only depends on the shape of the free-energy profile in reduced units; therefore, it is also independent of the temperature or free energy alone, as long as the ratio corresponds to the same profile (see Ref. [55], Sec. IV, for more detail).

Our general aim is to obtain a clustering of the underlying Markov chain into M-state aggregates: $S_1, \ldots, S_M$, defined by optimal cluster boundaries $b_1, \ldots, b_{M-1}$ that correspond to the slowest (maximum) relaxation time. Application of our method to a multitude of free-energy profiles (Fig. 2; see also Ref. [55], Sec. IV) revealed that the slowest time-scale optimization clustering successfully identifies all metastable states in the system. Once we exhausted the number of available important MSs, the next optimal M state is qualitatively different from the metastable macrostates obtained before, and it represents a TS with large probabilities of jumping to the neighboring M states [Figs. 2(a) and 2(b)], with most of the flux going through the TS.

We also found that the HS coarse-grained rate matrices and the LE coarse-grained Markov matrices at longer lag times (close to the relaxation times) identify practically the same optimal boundary positions (Fig. 2, as well as Figs. S2b, and S3–S7 in Ref. [55]), whereas shorter lag times identify TSs preferably over MSs (Ref. [55], Sec. IV). Our numerical examples demonstrate that the optimal coarse-grained states are invariant to the details of the implementation at the appropriate limit (discretization of the profile, prefactor, coarse-graining method, etc.), and only depend on the shape of the free-energy profile in reduced units.

B. Three-state system on a 2D model potential

We constructed an analytical 2D-model free-energy potential $F(x, y)$ (in kcal/mol, x, $y \in [-3, 3]$) using the following functional form:

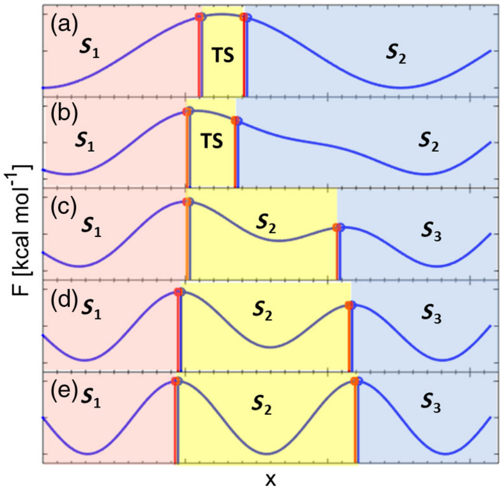

FIG. 2. Illustration of automatic optimal clustering into $M = 3$ macrostates for a set of 1D analytical potentials [panels (a) to (e); see Ref. [55], Sec. III) in which the intrinsic kinetics is tuned continuously as a function of a control parameter from being two-state-like (a) to three-state-like (e). The middle region (yellow) between the two boundaries identified by the algorithm (vertical lines) is either a TS, in (a) and (b), or it becomes the third MS in (c)—(e). Vertical lines correspond to the optimal clustering boundaries that maximize the second eigenvalue of the HS coarse-grained rate matrices (blue, Ref. [42]) and the LE transition probability matrices at $\tau = 1000$ (red).

$$F(x,y) = -0.7\ln[e^{-(x+2)^2-(y+2)^2} + e^{-(x-2)^2-(y-1)^2} + e^{-(x+3)^2-5(y-2)^2}] + c. \tag{11}$$

The constant c was chosen to give a potential of zero at the minimum on the 2D surface. We carried out Monte Carlo (MC) simulations on this free-energy surface by using a uniformly distributed 22-by-22 grid in the range of $[-2.7, 2.7]$ to initialize trajectories. Each trajectory was run for 4000 MC steps at 300 K using random displacement MC steps with a radius of $\rho = 0.3\sqrt{2\ln(1/r)}$, where r is a uniformly distributed random number between 0 and 1. The trajectory was analyzed using a lag time of 1000, and the corresponding Markov model was discretized into coarse-grained M states using a uniform grid of 15 states for each coordinate (x and y). A three-state model correctly identified the three stable minima (Fig. 3, right panel), and we obtained an optimal four-state network with the TS indicated as black stars on the reconstructed 2D profile (Fig. 3, left panel).

C. Umbrella sampling QM/MM free energy calculations of the 15-lipoxygenase-2 enzyme

Biased QM/MM calculations were performed previously [65] using an umbrella sampling protocol with harmonic biasing potential along a 1D reaction coordinate with 20 windows. We defined the reaction coordinate as the difference, $r_1 - r_2$, between the two relevant bond-breaking and forming distances. The simulation was unbiased using DHAM [48] on microstates defined via a 2D grid with 35 bins for both r_1 and r_2 (Fig. 4, symbols). The resulting free-energy surface is shown in the inset of Fig. 4, together with the TS state identified from the three-state coarse graining. We used a lag time of 10^9 fs for the coarse graining by propagating the original Markov matrix of lag time = 1 fs that was calculated with DHAM [48]. In the optimal coarse-grained system at this discretization, the probability to jump from the TS to state 1 (RS) is 0.024, to state 2 (TS) is very small (7.54e-14), and to state 3 (PS) is 0.976.

D. Conformational dynamics of Ala$_5$

We used our algorithm to study the dynamics of alanine pentapeptide (Ala$_5$)—a commonly used test system for evaluating the intrinsic conformational kinetics—using

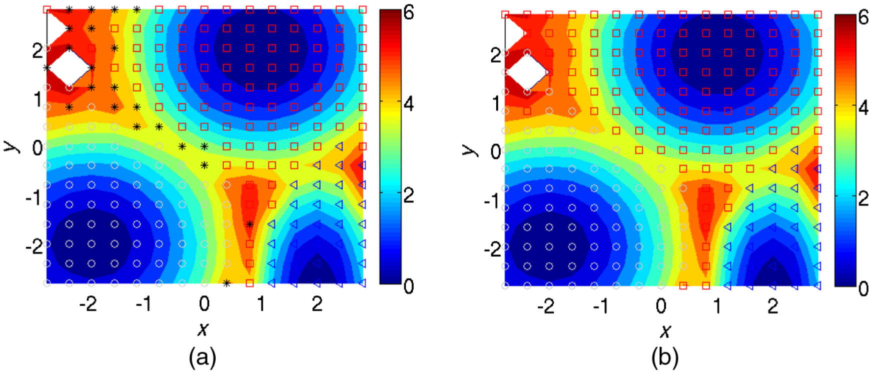

FIG. 3. The transition probability matrix and 2D free-energy surface were calculated for the direct product of 15 uniformly distributed M states at both x and y coordinates, using a lag time $\tau = 1000$, and by obtaining a relaxation time of 8532.7 (arbitrary units). The final coarse graining results in 4 M states (a), which are displayed as symbols (grey circle, black star, red square, and blue triangle), with a relaxation time of 8513.0. The TS ensemble corresponds to state 2 (black stars), with transition probabilities of 0.35, 0.64, and 0.01 to states 1, 3, and 4, respectively. A coarse graining to 3 M states results in the state assignment shown on the right panel (b), with a relaxation time of 8499.9.

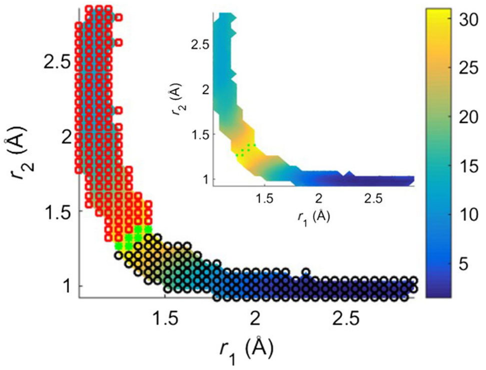

FIG. 4. 2D discretization and coarse graining of the kinetics of the hydride transfer in the 15-LOX-2 enzyme. QM/MM umbrella sampling simulations [65] were unbiased using DHAM [48] and projected on a 2D space of the bond-breaking (r_1) and forming (r_2) distances. Colored symbols indicate the three states identified from coarse graining (red for reactants, green for TS, black for product states). The free-energy surface with the TS ensemble is also shown in the inset (color bar in kcal/mol).

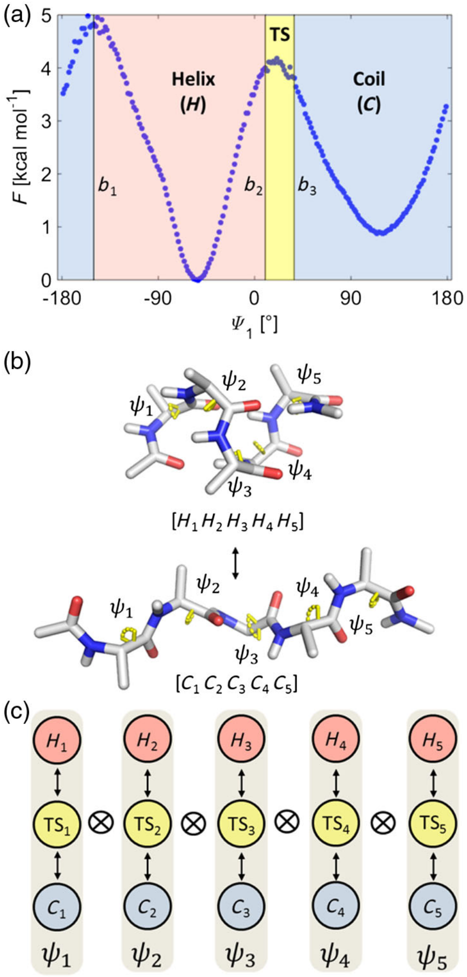

FIG. 5. (a) Free energy for the backbone angle Ψ_1 of Ala$_5$ calculated for $N_\mu = 200$ microstates at 350 K and lag time $\tau = 1$ ps. Resulting M states (vertical boundary lines: $b1$, $b2$, $b3$) are shown for three-state clustering with periodic boundaries [($b1$, $b2$) being identical for two-state clustering]. (b) All-helical, 00000, and all-coil, 22222, conformations of Ala$_5$. The five Ψ backbone dihedral angles (yellow) are used as RCs. (c) For N-D clustering ($N = 5$), microstate clustering first identifies $M = 3$ M states along each RC (H in red, C in blue, and TS in yellow), which constitute a product space (with 3^5 states) using all five RCs for the final coarse-graining to Ω states.

atomistic MD trajectories [57,66]. The Ala$_5$ system (Fig. 5) has the advantage of being sufficiently small for generating converged MD sampling, even with an explicit representation of water, using relatively modest computational resources. At the same time, it can form a helical turn, enabling the study of secondary structure formation in both theoretical [57,67–70] and experimental [71,72] studies. We analyzed MD trajectories of Ala$_5$ as described in detail in Ref. [57].

To cluster the MD trajectory of the Ala$_5$ peptide, the five Ramachandran Ψ angles have been chosen as RCs (Fig. S9) because the free-energy barriers along the Φ angles are much smaller and thus contribute less to the slowest relaxation modes [57]. The clustering result for the 1D profile for Ψ_1 is demonstrated in Fig. 5(a), with cluster boundaries along the free-energy profile calculated from $N_\mu = 200$ Markov microstates at 350 K and lag time $\tau = 1$ ps. For all 5 Ψ angles, the two-state assignment successfully identified the two metastable states corresponding to the two free-energy basins, and the slowest relaxation time is in good accordance with the full 200-state model. We also tested a range of bins (Fig. S10) and lag times (Fig. S11) to determine the Markovian limit of our full-dimensional model (see discussion in Ref. [55], Sec. IV). As demonstrated in Fig. 5(a), three-state clustering for meaningful lag times up to the magnitude of the relaxation time (τ~500 ps) leads to the detection of the small TS state. We also compared the optimal boundary positions using the HS clustering to obtain the initial M states (Fig. S12), and these are fully consistent with the ones obtained using the LE coarse-graining method. All TS states are characterized by a small equilibrium population and survival probability, and similar transition probabilities to either the H or C states.

The second-level clustering was performed by considering the three M states of all five Ψ RCs together (3^5 states in total, with M-state boundaries given in Table S1). For convenience, we have labeled the S_i M states with 0 if they are in a helical (H) region of Ψ, 1 if they are in a coil (C) region of Ψ, and 2 if they are in a TS region of Ψ. Thus, the all-helical macrostate of Ala$_5$ is denoted as [00000], while, for example, [02221] will denote an M state with the N-terminal residue helical, the C-terminal residue coil, and the middle residues in their respective TS regions of their Ψ angles.

We obtained 8 Ω macrostates at 350 K using lag times $\tau = 10$ ps as depicted in Fig. 6 and presented in Table S2 (including the corresponding Ω-state populations and the

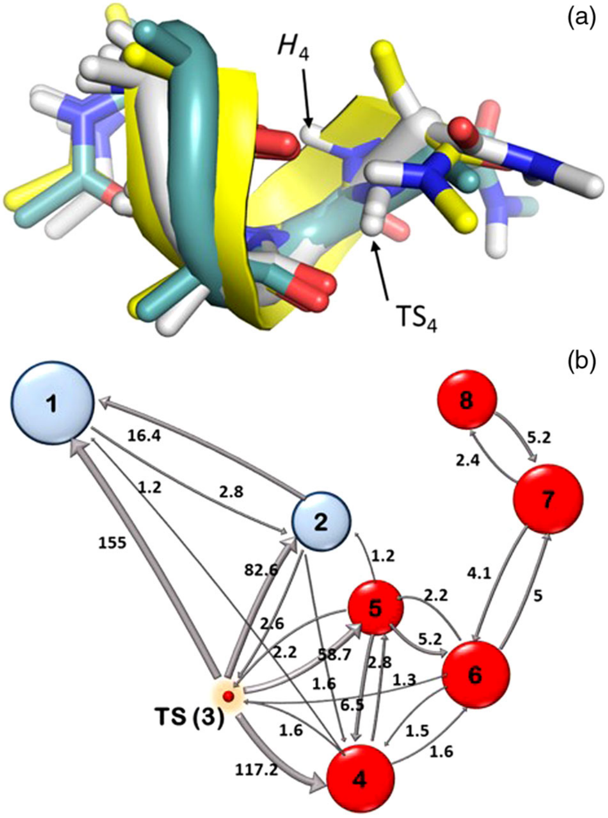

FIG. 6. (a) Transitions in the TS region are dominated by conformational dynamics at the fourth alanine residue of Ala_5. Representative conformations are shown for helical [00000] (yellow) and TS states [00020] (blue), and [00021] (grey). (b) Kinetic network for the optimal coarse-grained Ω states for the Ala_5 (350 K, Tables S1–S3, Fig. S13). Macrostates Ω_1 and Ω_2 (blue) form the folded ensemble, and five Ω-states (labeled 4–8, red) form the unfolded ensemble. The numbers near each arrow indicate the corresponding transition rates (ns^{-1}).

equilibrium probabilities of their most-populated M states). During the second-level clustering, for up to 3 Ω states, the boundaries were all varied at the same time exhaustively. To identify the optimal Ω states for four or more states, we first placed a new boundary while keeping the previous ones fixed, and we subsequently optimized every two adjacent boundaries exhaustively, iterating over all pairs while keeping all other boundaries fixed. Iterations were repeated until there were no more changes for any boundary pairs, and the corresponding solution was verified to be the optimal one by also exhaustively searching all the boundary positions at the same time for up to 7 Ω states. This iterative boundary sampling algorithm was also used in the subsequent examples below for four or more states.

The macrostates Ω_1 and Ω_2 have the largest population and form the "folded" ensemble characterized by the [00000] and [00001] configurations, respectively. As illustrated in Fig. 6(b), the "unfolded ensemble," states Ω_4 to Ω_8, presents, at 350 K, (i) a specific connectivity that depends on how many residues can change their state cooperatively and (ii) transition rates that are lower than in the folded ensemble.

The range of transitions near the TS is largely dominated by conformational dynamics at the first, second, and fourth residues of Ala_5. We estimated that the TS carries about 35%–45% of the overall flux when calculated using $k_D(k_A + k_B)/(k_D k_A + k_D k_B + k_{A'} k_B)$ (see Ref. [55], Sec. I). These observations, enabled by our method, agree well with previous studies of Ala_5 [57] while offering a more detailed, automatic analysis, also identifying TS states.

E. Dynamics of the EGF receptor

We also applied our algorithm to analyze the dynamics of the EGFR transmembrane helices. The MD data were obtained using the Anton supercomputer [1] and were first presented in Ref. [63] by the Shaw group. We analyze in detail a 124.51 μs-long trajectory of the N-terminal transmembrane dimer and used the time-lagged independent component analysis (tICA) to generate RCs [Figs. 7(a) and S14 of Ref. [55]] [45,50]. Our analysis (discussed in detail in Ref. [55], Sec. V) has identified six coarse-grained states [Fig. 7(c)]. The TS (state Ω_5) contributes less than 1% to the total equilibrium population, yet it carries about 90% of the overall flux, with transition probabilities to Ω_4 and Ω_6 nearly identical, 49.22% [Fig. 7(b)]. Our analysis can thus identify an overall coarse-grained kinetic network based of the current simulation data [Fig. 7(c)] and suggests key starting points for additional simulations from the TS configurations that could be aimed at efficient sampling of the dynamics.

F. Dynamic clustering of the Mga2 transmembrane helix conformational space

Mga2 is a fungal transcription factor forming homodimers in the ER and sensing the state of lipid membranes via rotational conformational changes of the transmembrane helices [64]. We have coarse grained the conformational dynamics of the Mga2 transmembrane helix dimer, focusing on the rotational orientations of the helices, using as a proxy, the relative position of W1042, a key residue involved in the sensing mechanism [64]. The coarse graining identified five MSs and two TSs (three MSs and one TS, considering the symmetry of the dimers). The two TS separate MS3 from MS1 and MS2, and from MS4 and MS5 (respectively), using the ordering of states given in Fig. 8. We used the following two coordinates: $\theta = \phi + \pi/2$ and $\chi = \Psi - \phi + \pi/2$, where ϕ and Ψ are angles that correspond to the relative orientation of the two helices in the dimer (Fig. 8, right). Note that θ and χ bring all the relevant clusters in the range $[0, 2\pi]$. Since the two helices have identical sequence and adopt similar structures, the representation by θ and χ exhibits mirror

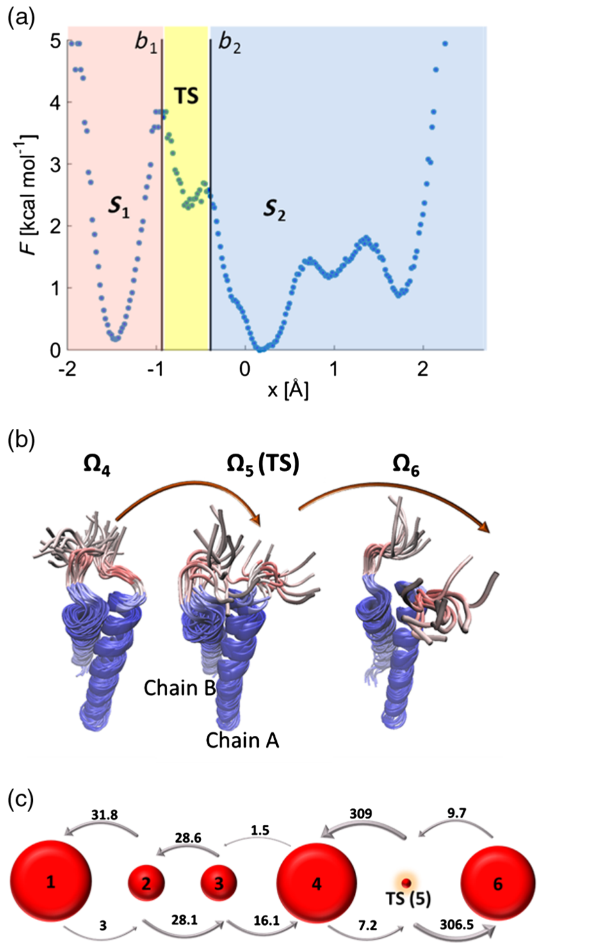

FIG. 7. (a) Free energy along the second tICA component of the EGFR transmembrane dimer calculated using an initial 200-state MSM. Results are shown for the $M = 3$-state system (boundaries $b1$ and $b2$) clustering for a lag time of $\tau = 1$ μs. The TS region (yellow) separates the S_1 macrostate from S_2. (b) Configurations of the transmembrane module of the EGF receptor that are representative of the Ω_4, Ω_5 (TS), and Ω_6 states at the N-terminal end of the dimer. The structures are colored according to their rms correlation with the second tICA component (blue to red: low to high). (c) Global kinetic network (Ω_1 to Ω_6) illustrated using four tICA components at lag time $= 1$ μs. Kinetic rates are shown in $1/100\ \mu\mathrm{s}^{-1}$.

symmetry about the line $\theta + \chi = 2\pi$. Accordingly, we symmetrized the data, counting each transition for also its symmetric counterpart $(\theta', \chi') = (2\pi - \chi, 2\pi - \theta)$. Our full-dimensional system consisted of 17 bins for each reaction coordinate θ and χ, which are in the Markovian limit for the full-dimensional MSM (Fig. S16). Our choice of $\tau = 100$ ns lag time is also in the Markovian limit and allows a meaningful coarse graining (Fig. S17). We obtained relaxation rates of 644.4 ns for the full 288-microstate system, and 609.2 ns for the seven-state system with the two TSs. The population of each TS state is about 4.32%, with a probability of about 0.1 to stay in the TS after the $\tau = 100$ ns lag time.

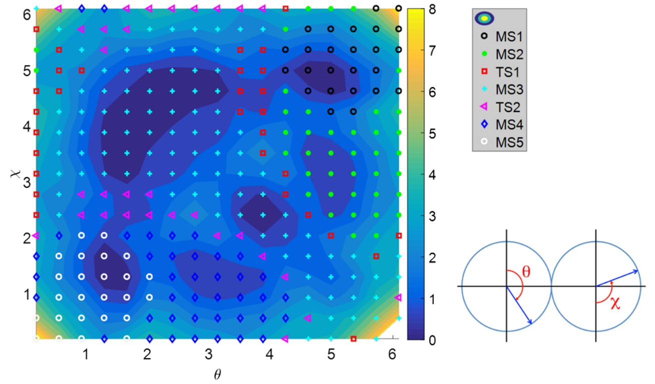

FIG. 8. 2D discretization and coarse graining of the conformational dynamics of a Mga2 transmembrane helix (TMH) dimer, using θ and χ angles (right panel) as reaction coordinates. Circles indicate the two helices in projection (bottom right of the figure). The horizontal line connects the two centers of the helices. The blue arrows point from the helix centers to the W1042 residues in each TMH. The free-energy surface is also shown with a color bar on the right in kcal/mol.

IV. CONCLUSIONS

We present a new approach to automatically identify relevant metastable and transition states along the available reaction coordinates describing a molecular process. An analytical model for three-state clustering shows two families of solutions to the optimization problem. One type leads to the identification of three metastable states, while the second type of solution yields two metastable states and one short-lived state. The short-lived state is an optimally constructed transition state, revealed automatically by our algorithm applied to analytical Markov models for arbitrary free-energy functions.

The TS identification algorithm is presented and studied by using both 1D and 2D analytical model potentials and several examples for the analysis of classical atomistic MD and QM/MM trajectories. We demonstrated the algorithm on coarse-graining QM/MM-biased umbrella sampling calculations for a catalytic reaction of the 15-LOX-2 enzyme. We also analyzed the classical-benchmark-system helix-forming peptide Ala_5 and two larger systems consisting of transmembrane helix dimers (the EGFR protein and the Mga2 lipid sensing transcription factor). In all cases, our method automatically identifies the transition states and the metastable conformations in an optimal way, with minimal input, by accurately capturing the intrinsic slowest relaxation times. We show that our new approach to identify and define transition states provides a general and easy-to-implement analysis method that provides useful insight into the underlying molecular mechanism and enables a quantitative and systematic characterization of the rare but crucial rate-limiting conformational pathways occurring in complex dynamical systems such as molecular trajectories.

In contrast to most other available clustering methods and their applications that capture metastable states or high-energy intermediates, our approach provides an automatic identification of all metastable states as well as of the significantly less populated (and thus experimentally elusive) transition states that control the slowest relaxation process in the system. This systematic method of identification of transition states and metastable states allows us to determine the underlying molecular mechanism with its key conformational ensembles. It could also lead to entirely new approaches to more efficiently simulate and analyze molecular processes, which is a central current challenge in a broad variety of biomolecular research and drug design problems. Our approach is fully general (i.e., does not rely on any system-specific molecular properties, provided that satisfactory reaction coordinates exist) in the analysis of time-dependent trajectories; therefore, it can also be applicable to time series generated for a broad range of complex systems, beyond multiscale molecular modeling studies, to identify rate-limiting events.

ACKNOWLEDGMENTS

We thank Dr. Attila Szabo (National Institutes of Health) and Dr. Alessia Annibale (King's College London) for numerous helpful discussions, and Albert Pan (D. E. Shaw Research) for his kind assistance with the EGFR simulation data. E. R. acknowledges EPSRC Grant No. EP/N020669/1. R. C. and. G. H. were supported by the Max Planck Society. N. V. B. gratefully acknowledges financial support from the Irish Research Council (IRC) and computational facilities provided by the Irish Centre for High-End Computing (ICHEC). A. K. is supported by the EPSRC Centre for Doctoral Training in Cross-Disciplinary Approaches to Non-Equilibrium Systems (CANES, EP/L015854/1). Further information can also be found at [73].

L. M. and A. K. contributed equally to this work.

[1] D. E. Shaw, R. O. Dror, J. K. Salmon, J. P. Grossman, K. M. Mackenzie, J. A. Bank, C. Young, M. M. Deneroff, B. Batson, and K. J. Bowers *et al.*, *Millisecond-scale molecular dynamics simulations on Anton*, *Proceedings of the conference on high performance computing networking, storage and analysis* (IEEE, New York, 2009), pp. 1–11.

[2] T. J. Lane, D. Shukla, K. A. Beauchamp, and V. S. Pande, *To Milliseconds and Beyond: Challenges in the Simulation of Protein Folding*, Curr. Opin. Struct. Biol. **23**, 58 (2013).

[3] P. L. Freddolino, C. B. Harrison, Y. Liu, and K. Schulten, *Challenges in Protein-Folding Simulations*, Nat. Phys. **6**, 751 (2010).

[4] K. Lindorff-Larsen, S. Piana, R. O. Dror, and D. E. Shaw, *How Fast-Folding Proteins Fold*, Science **334**, 517 (2011).

[5] P. G. Bolhuis, D. Chandler, C. Dellago, and P. L. Geissler, *Transition Path Sampling: Throwing Ropes Over Rough Mountain Passes, in the Dark*, Annu. Rev. Phys. Chem. **53**, 291 (2002).

[6] C. Dellago, P. G. Bolhuis, and P. L. Geissler, *Transition Path Sampling*, in *Advances in Chemical Physics* (John Wiley & Sons, New York, 2003), pp. 1–78.

[7] T. S. van Erp and P. G. Bolhuis, *Elaborating Transition Interface Sampling Methods*, J. Comput. Phys. **205**, 157 (2005).

[8] R. J. Allen, D. Frenkel, and P. R. ten Wolde, *Forward Flux Sampling-Type Schemes for Simulating Rare Events: Efficiency Analysis*, J. Chem. Phys. **124**, 194111 (2006).

[9] A. K. Faradjian and R. Elber, *Computing Time Scales from Reaction Coordinates by Milestoning*, J. Chem. Phys. **120**, 10880 (2004).

[10] Ch. Schütte, A. Fischer, W. Huisinga, and P. Deuflhard, *A direct approach to conformational dynamics based on hybrid Monte Carlo*, J. Comput. Phys. **151**, 146 (1999).

[11] J.-H. Prinz, H. Wu, M. Sarich, B. Keller, M. Senne, M. Held, J. D. Chodera, C. Schütte, and F. Noé, *Markov Models of Molecular Kinetics: Generation and Validation*, J. Chem. Phys. **134**, 174105 (2011).

[12] F. Noé and S. Fischer, *Transition Networks for Modeling the Kinetics of Conformational Change in Macromolecules*, Curr. Opin. Struct. Biol. **18**, 154 (2008).

[13] B. Peters and B. L. Trout, *Obtaining Reaction Coordinates by Likelihood Maximization*, J. Chem. Phys. **125**, 054108 (2006).

[14] D. J. Wales, *Perspective: Insight into Reaction Coordinates and Dynamics from the Potential Energy Landscape*, J. Chem. Phys. **142**, 130901 (2015).

[15] J. M. Carr and D. J. Wales, *Folding Pathways and Rates for the Three-Stranded β-Sheet Peptide Beta3s Using Discrete Path Sampling*, J. Phys. Chem. B **112**, 8760 (2008).

[16] P. Cossio, G. Hummer, and A. Szabo, *On Artifacts in Single-Molecule Force Spectroscopy*, Proc. Natl. Acad. Sci. U.S.A. **112**, 14248 (2015).

[17] A. Sirur, D. De Sancho, and R. B. Best, *Markov State Models of Protein Misfolding*, J. Chem. Phys. **144**, 075101 (2016).

[18] G. R. Bowman, E. R. Bolin, K. M. Hart, B. C. Maguire, and S. Marqusee, *Discovery of Multiple Hidden Allosteric Sites by Combining Markov State Models and Experiments*, Proc. Natl. Acad. Sci. U.S.A. **112**, 2734 (2015).

[19] B. Fačkovec, E. Vanden-Eijnden, and D. J. Wales, *Markov State Modeling and Dynamical Coarse-Graining via Discrete Relaxation Path Sampling*, J. Chem. Phys. **143**, 044119 (2015).

[20] H. S. Chung, S. Piana-Agostinetti, D. E. Shaw, and W. A. Eaton, *Structural Origin of Slow Diffusion in Protein Folding*, Science **349**, 1504 (2015).

[21] J. D. Chodera, N. Singhal, V. S. Pande, K. A. Dill, and W. C. Swope, *Automatic Discovery of Metastable States for the Construction of Markov Models of Macromolecular Conformational Dynamics*, J. Chem. Phys. **126**, 155101 (2007).

[22] V. S. Pande, K. Beauchamp, and G. R. Bowman, *Everything You Wanted to Know About Markov State Models but Were Afraid to Ask*, Methods **52**, 99 (2010).

[23] F. Noé, I. Horenko, C. Schütte, and J. C. Smith, *Hierarchical Analysis of Conformational Dynamics in Biomolecules: Transition Networks of Metastable States*, J. Chem. Phys. **126**, 155102 (2007).

[24] D. Prada-Gracia, J. Gómez-Gardeñes, P. Echenique, and F. Falo, *Exploring the Free Energy Landscape: From Dynamics to Networks and Back*, PLoS Comput. Biol. **5**, e1000415 (2009).

[25] P. Deuflhard, W. Huisinga, A. Fischer, and C. Schütte, *Identification of Almost Invariant Aggregates in Reversible Nearly Uncoupled Markov Chains*, Linear Algebra Applications **315**, 39 (2000).

[26] P. Deuflhard and M. Weber, *Robust Perron Cluster Analysis in Conformation Dynamics*, Linear Algebra Applications **398**, 161 (2005).

[27] S. Röblitz and M. Weber, *Fuzzy Spectral Clustering by* PCCA+*: Application to Markov State Models and Data Classification*, Adv. Data Anal. Classif. **7**, 147 (2013).

[28] G. R. Bowman, L. Meng, and X. Huang, *Quantitative Comparison of Alternative Methods for Coarse-Graining Biological Networks*, J. Chem. Phys. **139**, 121905 (2013).

[29] S. V. Krivov and M. Karplus, *Hidden Complexity of Free Energy Surfaces for Peptide (Protein) Folding*, Proc. Natl. Acad. Sci. U.S.A. **101**, 14766 (2004).

[30] S. V. Krivov and M. Karplus, *One-Dimensional Free-Energy Profiles of Complex Systems: Progress Variables that Preserve the Barriers*, J. Phys. Chem. **B110**, 12689 (2006).

[31] Y. Nagahata, S. Maeda, H. Teramoto, T. Horiyama, T. Taketsugu, and T. Komatsuzaki, *Deciphering Time Scale Hierarchy in Reaction Networks*, J. Phys. Chem. **B120**, 1961 (2016).

[32] A. Ma and A. R. Dinner, *Automatic Method for Identifying Reaction Coordinates in Complex Systems*, J. Phys. Chem. **B109**, 6769 (2005).

[33] W. E. W. Ren and E. Vanden-Eijnden, *Transition Pathways in Complex Systems: Reaction Coordinates, Isocommittor Surfaces, and Transition Tubes*, Chem. Phys. Lett. **413**, 242 (2005).

[34] R. B. Best and G. Hummer, *Reaction Coordinates and Rates from Transition Paths*, Proc. Natl. Acad. Sci. U.S.A. **102**, 6732 (2005).

[35] S. Kumar, J. M. Rosenberg, D. Bouzida, R. H. Swendsen, and P. A. Kollman, *The Weighted Histogram Analysis Method for Free-Energy Calculations on Biomolecules. I. The Method*, J. Comput. Chem. **13**, 1011 (1992).

[36] G. M. Torrie and J. P. Valleau, *Nonphysical Sampling Distributions in Monte Carlo Free-Energy Estimation: Umbrella Sampling*, J. Comput. Phys. **23**, 187 (1977).

[37] A. Laio and M. Parrinello, *Escaping Free-Energy Minima*, Proc. Natl. Acad. Sci. U.S.A. **99**, 12562 (2002).

[38] P. Tiwary and B. J. Berne, *Spectral Gap Optimization of Order Parameters for Sampling Complex Molecular Systems*, Proc. Natl. Acad. Sci. U.S.A. **113**, 2839 (2016).

[39] R. T. McGibbon and V. S. Pande, *Identification of Simple Reaction Coordinates from Complex Dynamics*, arXiv:1602.08776.

[40] J.-H. Prinz, J. D. Chodera, and F. Noé, *Estimation and Validation of Markov Models*, in *An Introduction to Markov State Models and Their Application to Long Timescale Molecular Simulation* (Springer, New York, 2014), pp. 45–60.

[41] F. Noé and F. Nüske, *A Variational Approach to Modeling Slow Processes in Stochastic Dynamical Systems*, Multiscale Model. Simul. **11**, 635 (2013).

[42] G. Hummer and A. Szabo, *Optimal Dimensionality Reduction of Multistate Kinetic and Markov-State Models*, J. Phys. Chem. **B119**, 9029 (2015).

[43] F. Nüske, B. G. Keller, G. Pérez-Hernández, A. S. J. S. Mey, and F. Noé, *Variational Approach to Molecular Kinetics*, J. Chem. Theory Comput. **10**, 1739 (2014).

[44] G. R. Bowman, K. A. Beauchamp, G. Boxer, and V. S. Pande, *Progress and Challenges in the Automated Construction of Markov State Models for Full Protein Systems*, J. Chem. Phys. **131**, 124101 (2009).

[45] C. R. Schwantes and V. S. Pande, *Improvements in Markov State Model Construction Reveal Many Non-Native Interactions in the Folding of NTL9*, J. Chem. Theory Comput. **9**, 2000 (2013).

[46] B. E. Husic and V. S. Pande, *Ward Clustering Improves Cross-Validated Markov State Models of Protein Folding*, J. Chem. Theory Comput. **13**, 963 (2017).

[47] H. Wu and F. Noé, *Optimal Estimation of Free Energies, and Stationary Densities from Multiple Biased Simulations*, Multiscale Model. Simul. **12**, 25 (2014).

[48] E. Rosta and G. Hummer, *Free Energies from Dynamic Weighted Histogram Analysis Using Unbiased Markov State Model*, J. Chem. Theory Comput. **11**, 276 (2015).

[49] L. Molgedey and H. G. Schuster, *Separation of a Mixture of Independent Signals Using Time Delayed Correlations*, Phys. Rev. Lett. **72**, 3634 (1994).

[50] G. Pérez-Hernández, F. Paul, T. Giorgino, G. De Fabritiis, and F. Noé, *Identification of Slow Molecular Order Parameters for Markov Model Construction*, J. Chem. Phys. **139**, 015102 (2013).

[51] T. Yamamoto, *Quantum Statistical Mechanical Theory of the Rate Exchange Chemical Reactions in the Gas Phase*, J. Chem. Phys. **33**, 281 (1960).

[52] H. Aroeste, *Toward an Analytic Theory of Chemical Reactions*, Adv. Chem. Phys. **6**, 1 (1964).

[53] R. Zwanzig, *Time-Correlation Functions and Transport Coefficiets in Statistical Mechanics*, Annu. Rev. Phys. Chem. **16**, 67 (1965).

[54] D. Chandler, *Statistical Mechanics of Isomerization Dynamics in Liquids and the Transition State Approximation*, J. Chem. Phys. **68**, 2959 (1978).

[55] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevX.7.031060 for further information on the definition of flux through the TS, three-state coarse graining, 1D analytical potentials, and additional analysis on the Ala5 and EGFR simulation data.

[56] S. Huo and J. E. Straub, *The MaxFlux Algorithm for Calculating Variationally Optimized Reaction Paths for Conformational Transitions in Many Body Systems at Finite Temperature*, J. Chem. Phys. **107**, 5000 (1997).

[57] N.-V. Buchete and G. Hummer, *Coarse Master Equations for Peptide Folding Dynamics*, J. Phys. Chem. **B112**, 6057 (2008).

[58] M. Sarich, F. Noé, and C. Schütte, *On the Approximation Quality of Markov State Models*, Multiscale Model. Simul. **8**, 1154 (2010).

[59] J.-H. Prinz, H. Wu, M. Sarich, B. Keller, M. Senne, M. Held, J. D. Chodera, C. Schütte, and F. Noé, *Markov Models of*

Molecular Kinetics: Generation and Validation, J. Chem. Phys. **134**, 174105 (2011).

[60] A. Berezhkovskii and A. Szabo, *Ensemble of Transition States for Two-State Protein Folding from the Eigenvectors of Rate Matrices*, J. Chem. Phys. **121**, 9186 (2004).

[61] L. V. Nedialkova, M. A. Amat, I. G. Kevrekidis, and G. Hummer, *Diffusion Maps, Clustering and Fuzzy Markov Modeling in Peptide Folding Transitions*, J. Chem. Phys. **141**, 114102 (2014).

[62] F. Noé and C. Clementi, *Kinetic Distance and Kinetic Maps from Molecular Dynamics Simulation*, J. Chem. Theory Comput. **11**, 5002 (2015).

[63] Y. Shan, M. P. Eastwood, X. Zhang, E. T. Kim, A. Arkhipov, R. O. Dror, J. Jumper, J. Kuriyan, and D. E. Shaw, *Oncogenic Mutations Counteract Intrinsic Disorder in the EGFR Kinase and Promote Receptor Dimerization*, Cell **149**, 860 (2012).

[64] R. Covino, S. Ballweg, C. Stordeur, Jonas B. Michaelis, K. Puth, F. Wernig, A. Bahrami, Andreas M. Ernst, G. Hummer, and R. Ernst, *A Eukaryotic Sensor for Membrane Lipid Saturation*, Mol. Cell **63**, 49 (2016).

[65] R. Suardíaz, P. G. Jambrina, L. Masgrau, À. González-Lafont, E. Rosta, and J. M. Lluch, *Understanding the Mechanism of the Hydrogen Abstraction from Arachidonic Acid Catalyzed by the Human Enzyme 15-Lipoxygenase-2. A Quantum Mechanics/Molecular Mechanics Free Energy Simulation*, J. Chem. Theory Comput. **12**, 2079 (2016).

[66] N.-V. Buchete and G. Hummer, *Peptide Folding Kinetics from Replica Exchange Molecular Dynamics*, Phys. Rev. E **77**, 030902 (2008).

[67] R. B. Best, N. V. Buchete, and G. Hummer, *Are Current Molecular Dynamics Force Fields Too Helical?*, Biophys. J. **95**, L07 (2008).

[68] A. E. García and K. Y. Sanbonmatsu, *α-Helical Stabilization by Side Chain Shielding of Backbone Hydrogen Bonds*, Proc. Natl. Acad. Sci. U.S.A. **99**, 2782 (2002).

[69] A. M. Berezhkovskii, F. Tofoleanu, and N.-V. Buchete, *Are Peptides Good Two-State Folders?*, J. Chem. Theory Comput. **7**, 2370 (2011).

[70] G. Hummer, A. E. García, and S. Garde, *Helix Nucleation Kinetics from Molecular Simulations in Explicit Solvent*, Proteins **42**, 77 (2001).

[71] C.-Y. Huang, Z. Getahun, Y. Zhu, J. W. Klemke, W. F. DeGrado, and F. Gai, *Helix Formation via Conformation Diffusion Search*, Proc. Natl. Acad. Sci. U.S.A. **99**, 2788 (2002).

[72] J. Graf, P. H. Nguyen, G. Stock, and H. Schwalbe, *Structure and Dynamics of the Homologous Series of Alanine Peptides: A Joint Molecular Dynamics/NMR Study*, J. Am. Chem. Soc. **129**, 1179 (2007).

[73] https://librarysearch.kcl.ac.uk.