JCTC Journal of Chemical Theory and Computation

Article

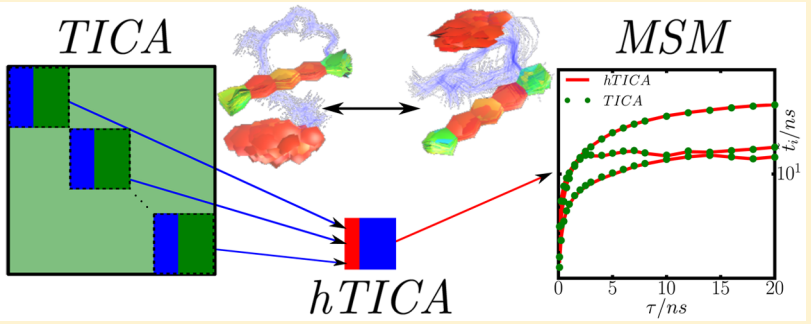
pubs.acs.org/JCTC

Hierarchical Time-Lagged Independent Component Analysis: Computing Slow Modes and Reaction Coordinates for Large Molecular Systems

Guillermo Pérez-Hernández* and Frank Noé*

Department of Mathematics and Computer Science, Freie Universitat Berlin, Arnimallee 6, Berlin, Germany 14195

Supporting Information



ABSTRACT: Analysis of molecular dynamics, for example using Markov models, often requires the identification of order parameters that are good indicators of the rare events, i.e. good reaction coordinates. Recently, it has been shown that the time-lagged independent component analysis (TICA) finds the linear combinations of input coordinates that optimally represent the slow kinetic modes and may serve in order to define reaction coordinates between the metastable states of the molecular system. A limitation of the method is that both computing time and memory requirements scale with the square of the number of input features. For large protein systems, this exacerbates the use of extensive feature sets such as the distances between all pairs of residues or even heavy atoms. Here we derive a hierarchical TICA (hTICA) method that approximates the full TICA solution by a hierarchical, divide-and-conquer calculation. By using hTICA on distances between heavy atoms we identify previously unknown relaxation processes in the bovine pancreatic trypsin inhibitor.

1. INTRODUCTION

Biological function relies on the ability of biomolecules to switch between conformational and association states with different functional roles. Examples include protein folding,[1−3] switching between active and inactive states in G-protein coupled receptors,[4] and transitions between different allosteric states in protein−ligand binding.[5,6] Often, these functional states are metastable, i.e. long-lived, and their presence has been demonstrated in various single-molecule experiments and kinetic ensemble experiments.[7−10] Recently, atomistic molecular dynamics (MD) simulations have been able to reach sufficient timescales in order to explicitly probe transitions between such metastable states and relate to experimentally measurable kinetics.[11−14]

Identification of metastable states and quantification of their equilibrium probabilities and kinetics is a prime goal in molecular simulation. A challenge is that molecular dynamics (MD) simulations are very high dimensional (3*N* Cartesian coordinates) and may comprise huge data sets, rendering a manual selection of reaction coordinates hopeless. Several procedures have been proposed to select good reaction coordinates.

Principal Component Analysis (PCA) of the Cartesian coordinates,[15,16] or internal coordinates such as dihedrals (dPCA),[17] finds linear combinations of the input parameters that contain most of the input parameters' variance. However, long-lived metastable states can sometimes be separated by small variances, or *buried* below noisy PCs,[18] and in such cases a projection onto the first principal components would loose the relevant reaction coordinates. It has been suggested to score reaction coordinates based on their ability to resolve the committor function, i.e. to provide a clear-cut definition between two metastable end-states.[19,20] This method is difficult to generalize to systems where multiple metastable states and multiple reaction coordinates are needed.

Conformation dynamics theory[21−24] has shown that the natural choice for the slow modes of the molecular system are the eigenfunctions of the Markov operator governing the molecular dynamics. These eigenfunctions define relaxation modes, each of which decay toward equilibrium with a well-defined relaxation time scales. These slow modes are moreover natural reaction coordinates for transitions between the metastable states of the system.[25]

A number of methods have used this insight in order to compute approximations to these slow modes or reaction coordinates.

Received: July 25, 2016
Published: October 28, 2016

DOI: 10.1021/acs.jctc.6b00738
J. Chem. Theory Comput. 2016, 12, 6118−6129

Markov (state) models (MSMs)[21,26−30] - see ref 31 for an overview - are a commonly used class of methods to approximate the true eigenfunctions through the eigenvectors of a transition probability matrix among discrete states. However, the process of discretizing state space itself can be very challenging. Initially, MSM studies have predominantly used clustering in relatively high-dimensional feature spaces, e.g. using the pairwise minimal root-mean-square-deviation (RMSD) as a metric.[29,32,33] While such a metric can in principle resolve all features of the conformation space when used with a fine enough cluster discretization,[30,34] it was more recently found that MSMs can tremendously benefit from a preprocessing step in which the dimensionality of the feature space is reduced while keeping the coordinates in which the slow processes can be best represented, and the metastable states are well separated.[18,35]

A rather general and direct approach is to compute the optimal linear combination of arbitrary functions of configuration space by using the Variational Approach of Conformation Dynamics (VAC).[24,36] A particularly simple choice of functions is to directly use the mean-free coordinates or some molecular order parameters such as torsion angles or distances between atoms, chemical groups, or molecules. Using the VAC with this choice then leads to an optimal approximation of the slow modes by linear combinations of these molecular order parameters.[18] The algorithm to compute this approximation has been introduced in ref 37 as a method for signal processing and compression and is known under the names blind source separation or more commonly Time-structure based or Time-lagged Independent Component Analysis (TICA). TICA has been used to identify slow collective variables in molecular dynamics and for the purpose of MSM building recently.[18,35,38,39] Although the VAC and TICA are linear approaches in the input functions, these functions can be nonlinear in the Cartesian coordinates, which means that very complex motions in configuration space can be represented. Recent developments have shown how slow modes computed from TICA or VAC can be scaled in order to define kinetically meaningful metric spaces[40,41] and how they can be estimated from short nonequilibrium data sets without imposing a strong bias.[42]

Finally, a number of intrinsically nonlinear methods for the approximation of the true reaction coordinates exist. Diffusion maps[43,44] use a linear combination of Kernel functions centered on the stored MD configurations. Diffusion coordinates are the true reaction coordinates under the assumption that the molecular configurations have emerged from a diffusion process. A combination of the diffusion map and the TICA idea in the framework of the VAC is the kernel TICA method.[35] Both methods have been practically demonstrated to provide excellent reaction coordinates,[35,44] but their computation is very time-consuming due to the large matrices that need to be parametrized and diagonalized.

In the present paper, we focus on improving the linear TICA method, although our ideas are in practice also applicable to kernel TICA or other incarnations of the VAC. It has been found that among other coordinates, distances or contacts between atoms or residues represent a suitable coordinate set.[18,35] Therefore, in order to minimize the bias of the analyst, it would be desirable to allow the use of all intramolecular distances between N_a atoms (or groups of atoms) as an input to TICA *a priori*.

Unfortunately, both the computational effort and the memory requirements of the TICA scale with the square of the number of input coordinates, and thus with N_a^4 when distances are used, quickly overwhelming any computational resource when the molecular system is large. For example, a medium-sized protein may have 300 amino acids, giving rise to nearly 45,000 nonredundant residue-distances. The memory requirements of the covariance matrices (of size 45,000 × 45,000) is about 16 GByte when stored as 8-byte floating point variables. Calculating these matrices requires more than 2×10^9 scalar products of length T, the total number of stored molecular configurations, which can be in the millions. Clearly, this approach is not scalable.

In this paper, we propose a divide-and-conquer heuristic to avoid computing and diagonalizing the above-mentioned matrices and still arrive at a *good and large enough* set of collective coordinates from which the true reaction coordinates can subsequently be approximated by e.g. the construction of a Markov model.

We present the heuristic as a two-step procedure in which a first round of distance TICA is performed N_a-times for N_a-batches of distances, the dominant subspaces of which get linearly combined in a second TICA step. In principle one can extend this heuristic and iterate the addition of basis functions (not necessarily distances) and use the variational inequality to check if the model has improved.

2. THEORY

2.1. Molecular Dynamics in the Conformation Dynamics Formulation. In the conformation dynamics formulation[21,24,34] of MD, a transition density, $p_\tau(\mathbf{y}|\mathbf{x})$ can be written

$$p_\tau(\mathbf{y}|\mathbf{x})\mathrm{d}\mathbf{y} = \mathrm{Prob}(\mathbf{x}_{t+\tau} \in \mathbf{y}\ \mathrm{d}\mathbf{y}|\mathbf{x}_t = \mathbf{x}) \tag{1}$$

for a Markovian and thermostated MD implementation. $p_\tau(\mathbf{y}|\mathbf{x})\mathrm{d}\mathbf{y}$ defines the probability density for a trajectory that has been at state $\mathbf{x}$ at time t to be in state $\mathbf{y}$ at time $t + \tau$, for any pair of phase-space vectors, $\mathbf{x},\mathbf{y} \in \Omega$. The time evolution the molecular ensemble density, $\rho_t(\mathbf{x})$, under the action of $p_\tau(\mathbf{y}|\mathbf{x})$ can be then written as

$$\rho_{t+\tau}(\mathbf{y}) = \int_\Omega p(\mathbf{x}, \mathbf{y}, \tau)\rho_t(\mathbf{x})\mathrm{d}\mathbf{x} \tag{2}$$

and further decomposed into a set of relaxation processes

$$\rho_{t+\tau}(\mathbf{x}) = \sum_i^\infty e^{-\frac{\tau}{t_i}}\langle\psi_i|\rho_t\rangle\mu(\mathbf{x})\psi_i(\mathbf{x}) \tag{3}$$

where μ is the stationary distribution (e.g., Boltzmann distribution in case of an NVT simulation) and ψ_i are the eigenvectors of the transfer operator, or Markov backward propagator, with eigenvalues $\lambda_i = e^{-\frac{\tau}{t_i}}$.

Hence the objective of conformation dynamics, Markov state modeling, and TICA is to approximate the dominant eigenvalues λ_i and eigenfunctions ψ_i of the transfer operator, which are the slow modes of the dynamics. The eigenfunctions ψ_i are also sometimes referred to as true or intrinsic reaction coordinates, as the committor probabilities for transitions between the metastable states of the system can be represented by a linear combination of them.[25] Knowing these quantities allows for the prediction of any stationary or time-dependent property for arbitrarily long timescales (eq 3).

2.2. TICA. In a recent paper,[18] we have described a straightforward way to linearly approximate the true eigenfunctions, ψ_i, using the method of linear variation formulated in

DOI: 10.1021/acs.jctc.6b00738
J. Chem. Theory Comput. 2016, 12, 6118−6129

refs 24 and 36. Given the input coordinates, $y_i(\mathbf{x})$, we first subtract the means

$$\chi_i(\mathbf{x}) = y_i(\mathbf{x}) - \overline{y}_i(\mathbf{x}) \tag{4}$$

$\boldsymbol{\chi} = \{\chi_i\}$ is used as a basis set, i.e. we attempt to approximate the eigenfunctions ψ_i as a linear combination of Ansatz basis functions

$$\psi_i^{\ddagger} = \sum_{j=1}^{N} u_{ij}\chi_j \tag{5}$$

where the expansion coefficients u_{ij} are unknown and need to be determined. Given $\psi_i^{\ddagger}$, the corresponding eigenvalue approximation is then obtained by the normalized autocorrelation function of the eigenfunction, i.e. the Rayleigh coefficient

$$\lambda_i^{\ddagger}(\tau) = \frac{\langle \psi_i^{\ddagger}(\mathbf{x}_t)\psi_i^{\ddagger}(\mathbf{x}_{t+\tau})\rangle_t}{\langle \psi_i^{\ddagger}(\mathbf{x}_t)\psi_i^{\ddagger}(\mathbf{x}_t)\rangle_t}$$

where the optimization problem now consists of obtaining the coefficients u_{ij}, which we denote as a set of vectors $\{\mathbf{u}_i\}$, $\mathbf{u}_i \in \mathbb{R}^N$

$$\mathbf{C}^{\chi}(\tau)\mathbf{u}_i = \mathbf{C}^{\chi}(0)\mathbf{u}_i\lambda_i^{\ddagger}(\tau) \tag{6}$$

where $\mathbf{C}^{\chi}(\tau)$ is the covariance matrix of the Ansatz functions χ_i at a lag time τ. It has the elements

$$c_{ij}^{\chi}(\tau) = \langle \chi_i(\mathbf{x}), \chi_j(\mathbf{x}_{t+\tau})\rangle_t \tag{7}$$

which can be estimated from MD trajectories. This approach to approximate the slow processes in a signal was first proposed in a different context by ref 37.

In order to use TICA as a dimension reduction method, we project the mean-free input coordinates $\boldsymbol{\chi}(\mathbf{x})$ onto the first m eigenfunctions with the largest eigenvalues, where m is a parameter. We thus obtain the TIC trajectories[18]

$$z_i(t) = \mathbf{u}_i^{\mathrm{T}}\boldsymbol{\chi}(\mathbf{x}_t) = \langle \mathbf{u}_i|\boldsymbol{\chi}(\mathbf{x}_t)\rangle \tag{8}$$

where $\boldsymbol{\chi}$ is the matrix containing the chosen input coordinates as a function of the molecular conformation at a given time, $\mathbf{x}_t$.

2.3. Variational Principle. As shown in ref 18, the optimality of TICA directly follows from the fact that eq 6 is an implementation of the method of linear variation described in ref 24. A consequence, also shown in ref 24, is the following special variational principle:

1. The eigenfunction approximations $\psi_i^{\ddagger}$ computed from (6) and (5) are exact if and only if the eigenvalues are exact:

$$\lambda_i^{\ddagger} = \lambda_i \Leftrightarrow \psi_i^{\ddagger} = \psi_i$$

2. When there is a finite basis set error, i.e. the expansion (5) is not perfect, then the eigenfunctions will only be approximate. In this case, the corresponding eigenvalues will always be underestimated:

$$\lambda_i^{\ddagger} \le \lambda_i \tag{9}$$

3. As a result, the partial eigensum $M^{(m)}$ is also underestimated:

$$M^{(m)\ddagger} := \sum_{i=1}^{m} \lambda_i^{\ddagger} \le \sum_{i=1}^{m} \lambda_i =: M^{(m)}$$

consequently, $M^{(m)\ddagger}$ has recently been suggested as a score to rank kinetic models.[45] Finally, the sum of dominant relaxation rates is overestimated:

$$R^{(m)\ddagger} := \sum_{i=1}^{m} \kappa_i^{\ddagger} \le \sum_{i=1}^{m} \kappa_i =: R^{(m)}$$

The solution of the eigenvalue problem (6) finds the linear combination of basis functions that maximizes $\lambda_i^{\ddagger}$ and thereby also maximizes $M^{(m)\ddagger}$ and minimizes $R^{(m)}$. The variational method thus provides an optimal approximation to the eigenfunctions ψ_i through (5). TICA is in fact a method of linear variation with the special choice of basis functions in eq 4.

3. METHODS

3.1. Hierarchical TICA (hTICA). We propose an approximate TICA method that addresses the problem that computing the full correlation matrices (7) and solving the eigenvalue problem (6) can be computationally prohibitive for large sets of input coordinates (large N, cf. Table 1). As a consequence of the variational principle, we can directly assess the performance of such an approximate method by how large the dominant eigenvalues $\lambda_i^{\ddagger}$ or the partial eigensum $M^{(m)\ddagger}$ is or by how small $R^{(m)\ddagger}$ is. Given the basis set (4), the full TICA method provides optimal values for these scores, while any approximate TICA method will provide a further underestimate/overestimate (respectively).

Table 1. Summary of the Parameters That Characterize an hTICA Computation

symbol	meaning
N	number of input coordinates
T	number of input time steps
N_a	number of coordinate subsets
N_r	number of coordinates per subset, $N = N_aN_r$
ψ_i^k	i-th level 1 TICA eigenfunction of the k-th subset
m_a	number of stored TICs per subset, $m_a \ll N_r$
$N_{\ddagger}$	sum of stored TICs over all subsets, i.e. size of the level 2 TICA problem, $N_{\ddagger} = N_am_a$
$\tau^{\ddagger}$	lagtime of the level 1 and level 2 TICA
m	final number of TICs stored globally
n	number of discrete states finally used in the MSM

The idea of hierarchical TICA is as follows: (i) Subdivide the large set of mean-free input coordinates $\boldsymbol{\chi}(\mathbf{x}_t)$ into N_a subsets $\boldsymbol{\chi}^a$ of manageable size (N_r), conduct TICA on each of these subsets separately, and project each subset onto a small number ($m_a \ll N_r$) of time independent components (TICs). (ii) Combine all subset TICs obtained this way ($N_am_a = N_{\ddagger}$) with a second TICA and reduce again to a small number m of overall TICs. See Figure 1 for a graphical illustration.

We can write the associated approximation of eigenfunctions, $\psi_i^{\ddagger}$, as follows: the second level of hTICA approximates the eigenfunctions as a superposition of the first m_a level 1 eigenfunctions $\psi_j^{k\ddagger}$ of each k-th set [For a discussion on the parameter m_a see the SI.]:

$$\psi_i^{\ddagger} = \sum_{k=1}^{N_a}\sum_{j=1}^{m_a} b_{[i,km_a+j]}\psi_j^{k\ddagger}$$

which are obtained as TICs from each subset TICAs:

$$\psi_j^{k\ddagger} = \sum_{l=1}^{N_r} a_{jl}^k\chi_l^k$$

Combining both equations yields the hTICA expansion

$$\psi_i^{\ddagger} = \sum_{k=1}^{N_a}\sum_{j=1}^{m_a} b_{[i,km_a+j]}\sum_{l=1}^{N_r} a_{jl}^k\chi_l^k$$

DOI: 10.1021/acs.jctc.6b00738
J. Chem. Theory Comput. 2016, 12, 6118−6129

The hTICA algorithm is summarized by **hTICA**($\chi(\mathbf{x}_t)$, τ, N_a, m_a, m).

1. Subdivide the N input coordinates χ into N_a sets χ^k, each of length $N_r = N/N_a$.
2. For each $k = 1, \ldots, N_a$:
(a) Compute level 1 correlation matrices $\mathbf{C}^k(0)$ and $\mathbf{C}^k(\tau)$

$$c_{ij}^k(\tau) = \langle \chi_i^k(\mathbf{x}_t)\chi_j^k(\mathbf{x}_{t+\tau})\rangle_t$$

(b) Compute the first m_a subset TICs by solving the level 1 TICA problem:

$$\mathbf{C}^k(\tau)\mathbf{a}_i^k = \mathbf{C}^k(0)\mathbf{a}_i^k\lambda_i^{k\ddagger}(\tau)$$

3. Compute level 2 correlation matrices $\mathbf{C}(0)$ and $\mathbf{C}(\tau)$ by

$$c_{[km_a+i,lm_a+j]}(\tau) = \langle \mathbf{a}_i^{k\mathrm{T}}\chi^k(\mathbf{x}_t)\mathbf{a}_j^{(l)\mathrm{T}}\chi^{(l)}(\mathbf{x}_{t+\tau})\rangle_t$$

4. Solve the level 2 TICA problem:

$$\mathbf{C}(\tau)\mathbf{b}_i = \mathbf{C}(0)\mathbf{b}_i\lambda_i^{\ddagger}(\tau)$$

5. Approximate the TICs by

$$\psi_i^{\ddagger} = \sum_{k=1}^{N_a}\sum_{j=1}^{m_a} b_{[i,km_a+j]}\mathbf{a}_j^{k\mathrm{T}}\chi^k \tag{10}$$

The correlation functions $\langle\cdot\rangle_t$ are in practice computed by a straightforward time-average. While $\mathbf{C}(0)$ and $\mathbf{C}^k(0)$ are automatically symmetric, we must enforce symmetry in $\mathbf{C}(\tau)$ and $\mathbf{C}^k(\tau)$ after the time-average. Eq 10 can alternatively be written as

$$\psi_i^{\ddagger}(\mathbf{x}) = \mathbf{b}_i^{\mathrm{T}}\mathbf{A}\chi(\mathbf{x})$$

where $\mathbf{b}_i^{\mathrm{T}}$ is the row vector projecting the level 1 TICs to the i-th level 2 TICs, and the coefficients from TICA level 1 are collected in the sparse matrix:

$$\mathbf{A} = \begin{pmatrix} (\mathbf{a}_1^1)^{\mathrm{T}} & & & \\ \vdots & & & \\ (\mathbf{a}_m^1)^{\mathrm{T}} & & & \\ & (\mathbf{a}_1^2)^{\mathrm{T}} & & \\ & \vdots & & \\ & (\mathbf{a}_m^2)^{\mathrm{T}} & & \\ & & \ddots & \\ & & & (\mathbf{a}_1^N)^{\mathrm{T}} \\ & & & \vdots \\ & & & (\mathbf{a}_m^N)^{\mathrm{T}} \end{pmatrix}$$

One can then (formally) rewrite the entire hTICA transformation (level 1 + level 2) as the following matrix multiplication:

$$\mathbf{\Psi}^{\ddagger}(\mathbf{x}) = \mathbf{BA}\chi(\mathbf{x}) \tag{11}$$

3.2. Partition Schemes and Computational Cost. The choice of how many level 1 TICA problems will be constructed (N_a) and the size (N_r) of these blocks is, in principle, arbitrary. So far, we have only stated that these level 1 problems need to be of manageable size, meaning that the effort of evaluating and diagonalizing them has to be significantly smaller than evaluating the full matrices (cf. Figure 1). For simplicity, we here use a balanced partition, i.e. the convention that all N_a level 1 problems (or blocks) have equal size. For a given number of blocks (N_a) and a given number of input coordinates (N), the size of each block in a balanced partition would be

$$N_r = \frac{N}{N_a}$$

In principle, other partition schemes could be chosen using information known *a priori* about the molecular topology. For example, one could choose different blocks representing different domains of the molecule, or different monomers of a heteromer, or substrate and ligand, etc. In this paper we will consider a relatively extensive set of features and consider partitions in which each block represents one residue or atom and all its contacts/distances with all other residues or atoms. There are, however, a combinatorially large number of possible partition schemes. In section 4.2 we have sampled a number of random partitions and found that the hTICA results are robust with respect to the choice of the partition.

Finally, the computational cost of (h)TICA is dominated by the calculation of the correlation matrices and by the memory requirements of storing them. In the normal TICA case, for N input features, the matrix contains N^2 elements (= memory requirements). Each element of this matrix has been computed as the sum over T pair products, hence the cost of N^2T.

In hTICA, level 1 comprises of N_a TICA problems, each of size $(N/N_a)^2$, and level 2 contains one problem of size $(N_am_a)^2$.

In a balanced partition we have that $N_r \approx N_a$ and $N = N_a^2$, leading to the computational effort of hTICA being proportional to N_a^3 rather than to N_a^4. Table 2 summarizes the order $O(\cdot)$ of hTICA's computational cost and memory cost.

3.3. Hierarchical Principal Component Analysis. Based on the above ideas we can straightforwardly formulate a hierarchical principal component analysis (hPCA) method by making the following substitutions in the hTICA algorithm

$$\mathbf{C}^k(\tau) \leftarrow \mathbf{C}^k(0)$$

$$\mathbf{C}^k(0) \leftarrow \mathbf{Id}$$

$$\mathbf{C}(\tau) \leftarrow \mathbf{C}(0)$$

$$\mathbf{C}(0) \leftarrow \mathbf{Id}$$

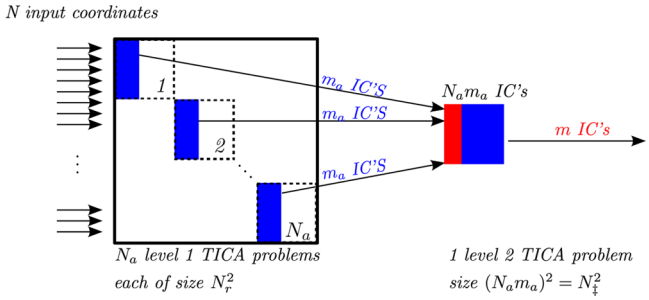


Figure 1. hTICA scheme. Instead of solving the full N^2 TICA problem (the square with solid black outline), the N input coordinates are partitioned into N_a subsets of size N_r, giving rise to N_a smaller level 1 TICA problems of sizes N_r^2 (the dotted blocks). Once these level 1 TICA eigenproblems are solved, m_a independent components are kept of each subset (the blue column vectors). The resulting N_am_a independent components are then parsed into a "level 2" TICA problem of size $(N_am_a)^2$ (the blue square matrix), the solution of which is truncated at m independent components (the red column vectors).

DOI: 10.1021/acs.jctc.6b00738
J. Chem. Theory Comput. 2016, 12, 6118−6129

Table 2. Order $O(\cdot)$ of Computational Cost and Memory Requirement[a]

	full TICA	hTICA level 1	hTICA level 2	hTICA(1 + 2)
memory requirements	N^2	$N_a\left(\frac{N}{N_a}\right)^2 = \frac{N^2}{N_a}$	$(N_a m_a)^2$	$\frac{N^2}{N_a} + (N_a m_a)^2$
calculation of $\mathbf{C}(0)$, $\mathbf{C}(\tau)$	$N^2 T$	$N_a\left(\frac{N}{N_a}\right)^2 T = \frac{N^2}{N_a} T$	$(N_a m_a)^2 T$	$\frac{N^2}{N_a} T + (N_a m_a)^2 T$

[a] N is the number of input features (distances, in our examples), and T is the number of timesteps. N_a is the number of level 1 TICA problems, and m_a is the number of level 1 TICs that are kept and used for the level 2 TICA. If the partition is balanced ($N_r \approx N_a$), we have $N = N_a^2$, and the computational effort of hTICA is proportional to N_a^3 rather than to N_a^4. See the SI for more details.

We will not use hPCA here as PCA aims at maximizing the variances rather than autocorrelations and is therefore not optimal for the calculation of slow modes and thus for the construction of a Markov model (see Table 3).

Table 3. Notation Summary of the Different Levels of Approximation and the Spaces They Are Computed in

symbol	approximation	space
λ	true	continuous
$\lambda^{(a)\ddagger}$	level 1 TICA	continuous
$\lambda^{\ddagger}$	level 2 TICA	continuous
$\hat{\lambda}$	full TICA	continuous
$\tilde{\lambda}$	MSM	discrete

3.4. Markov Model in the Dominant TICA/hTICA Subspace. Markov State Models (MSMs) have been successfully applied to biomolecular MD-data analysis on many occasions and are introduced here as the last step in the hTICA/hTICA methodology. The construction of an MSM will usually improve the reaction coordinates estimated by TICA/hTICA and increase the eigenvalues (see Figure 2). The TICs $\psi_i^{\ddagger}$ approximate the true eigenfunctions, ψ_i, only optimally in terms of linear combinations of the chosen input parameters, χ_i. This restriction can be overcome by an MSM which approximates the nonlinear eigenfunctions by step-functions that are constant on the discrete states.[30,34]

Here we employ the recent Python version of the software EMMA[46] (http://www.pyemma.org) in order to construct MSMs. We conduct a *kmeans* clustering discretization in the space of the dominant hTICs and use the resulting Voronoi discretization to estimate the MSM transition matrix $\mathbf{T}(\tau)$ using the reversible maximum likelihood estimation (MLE) algorithm.[30,47] This estimate is conducted for several values of τ in order to find an appropriate choice of the lag time, that is subsequently validated using a Chapman-Kolmogorov test.[30] The left and right eigenfunctions of $\mathbf{T}(\tau)$, $\tilde{\phi}_i$ and $\tilde{\psi}_i$, respectively, together with the eigenvalues, $\tilde{\lambda}$, and associated timescales, $\tilde{t}_i$, are the ultimate MSM results for the sake of this paper (for an overview on notation see Table 3).

4. RESULTS

4.1. hTICA Using Interatomic Distances in MR121-GSGS-W. We first use hTICA on a data set belonging to the synthetic fluorescent peptide MR121-GSGS-W,[48] a small peptide only six residues long, with a flexible glycine-serine-glycine-serine chain linking two rigid fluorescent groups: MR121 (a dye) and a tryptophan (Trp or W). A picture of the peptide is shown in Figure 3. The data set, available at http://simtk.org/home/emma, consists of two explicit-solvent simulations, each 4 μs long. The details of the simulation setup are described in ref 30.

MR121-GSGS-W has been extensively studied experimentally and theoretically.[11,18,48,49] The slowest relaxation timescales of the data has been estimated to be between 20 and 30 ns, and it has been found that the slowest processes are dominated by the interaction between MR121 and the tryptophan residue (Trp). Furthermore, the size of the molecule also allows for a direct comparison of TICA vs hTICA, as a means to validate the proposed heuristic.

The chosen input coordinates are all the interatomic distances between the 81 atoms of the peptide, hence $N = 81^2 - 81 = 6480$. For the level 1 TICA, the most straightforward partition is to group the input parameters in $N_a = 81$ subsets, each a-th set containing all distances that include the a-th atom ($N_r = 80$). Note that we have not excluded redundant distances as to have balanced, equally sized subsets. For these subsets, level 1 and

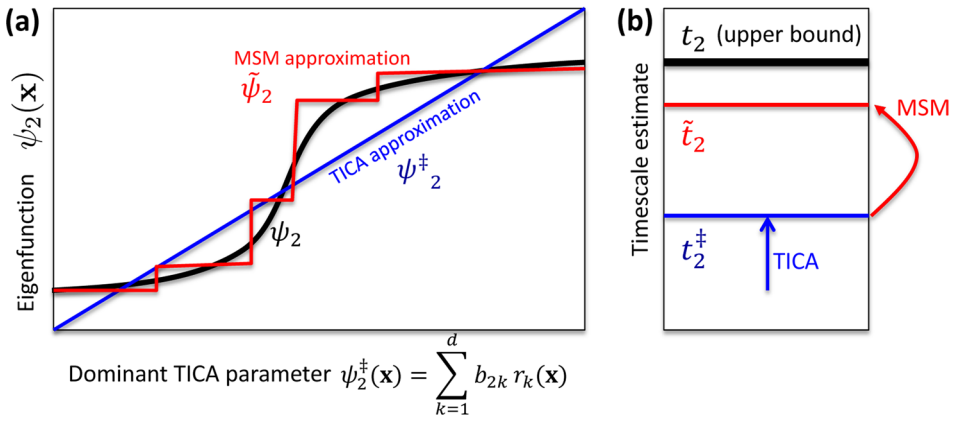


Figure 2. a) Scheme illustrating different approximations to the dominant eigenfunction of the molecular dynamics propagator and b) the associated approximations to the slowest relaxation timescales.

DOI: 10.1021/acs.jctc.6b00738
J. Chem. Theory Comput. 2016, 12, 6118−6129

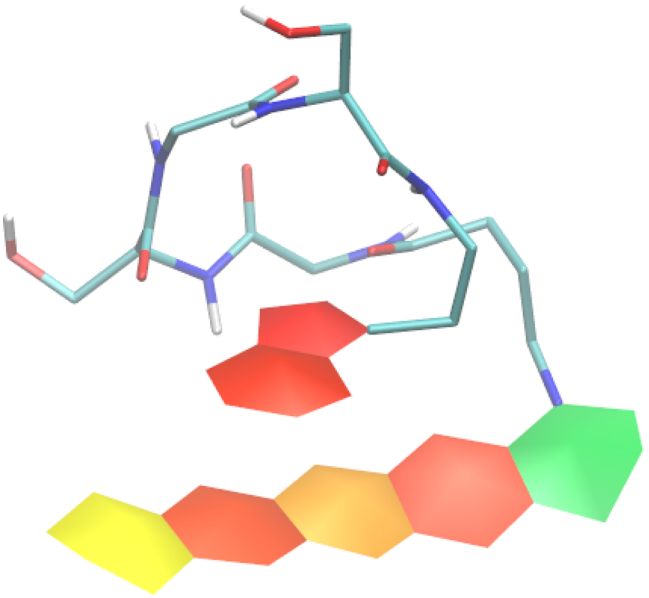


Figure 3. Small peptide MR121-GSGS-W. The flexible, GSGS linker is shown in the licorice representation, whereas the rigid moieties MR121 and TRP are shown using the so-called paperchain representation style.

level 2 TICA are performed at a lagtime of 5 ns ($\tau^{\ddagger} = 5$), using the first 10 ($m_a = 10$) TICs of each subset. All these parameters are summarized in Table 2 in the Supporting Information. A comparison of level 1 vs level 2 eigenvalues is shown in Figure 4. See the Supporting Information for a short note on how the lagtime is chosen.

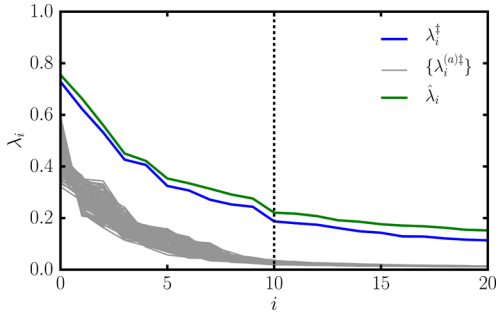


Figure 4. MR121-GSGS-W. Overlayed in gray are all N_a sets of level 1 TICA eigenvalues, $\lambda_i^{(a)\ddagger}$, in blue the set of level 2 TICA, $\lambda^{\ddagger}$. The vertical dotted line marks the chosen $m_a = 10$ value as a visual aid: the blue curve results of mixing all the level 1 eigenfunctions, $\psi^{(a)\ddagger}$ associated with the gray eigenvalues to the left of the dotted line. As a comparison, we also show the eigenvalues of the full TICA, $\hat{\lambda}_i$. See the SI for a brief discussion on the difference between $\hat{\lambda}_i$ and $\lambda^{\ddagger}$.

Figure 4 illustrates two of the concepts that we want to put forward in this paper. First, the variational principle behind the inequality of eq 9 ensures that the level 2 TICA approximation, $\psi_i^{\ddagger}$, can only get better if we keep adding new basis functions. Note that in the particular case of hTICA, the new basis functions are themselves TICs from other subsets, $\psi^{(a)\ddagger}$, but that does not need to be the case. One could resort to input coordinates, and still the level 2 eigenfunctions would necessarily be equal or better than level 1 eigenfunctions. Second, there is almost no difference between hTICA($\lambda_i^{\ddagger}$) and full TICA($\tilde{\lambda}_i$). The divide-and-conquer approach that avoids evaluating all $N^2 \approx 81^4 \approx 4 \times 10^7$ covariance matrix elements (cf. eq 7) recovers the dominant eigenvalues of the full TICA almost exactly, and it does so by evaluating only roughly $N_a N_r^2 \approx 81^3 \approx 5 \times 10^5$ elements in the level 1 TICA and $m_a N_a \approx 10 \times 81 \approx 8 \times 10^2$ elements in level 2 TICA.

Finally, we move on to the MSM construction. For this, we choose $m = 10$ and $n = 1000$; that is, we choose to discretize the space spun by the first 10 $\psi_i^{\ddagger}$ eigenfunctions into 1000 microstates, by finding 1000 *kmeans*-clustercenters.

After discretizing the trajectories, we estimate the transition matrix $\mathbf{T}(\tau)$ at different lagtimes. The associated eigenvalues $\tilde{\lambda}_i(\tau)$ are represented as timescales $\tilde{t}_i(\tau) = -\frac{\tau}{\log(\tilde{\lambda}_i)}$ in the implied timescales plot in Figure 5.

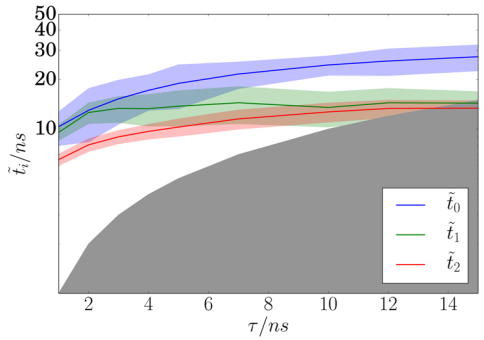


Figure 5. MR121-GSGS-W. *Semilog* plot of the implied timescales (ITS, or t_i) for a Markov State Model built on the *kmeans*-discretization (1000 centers) in the space spun by the first 10 hTICs, $\psi_i^{\ddagger}$. The color-shaded areas mark the 2σ confidence intervals of a bootstrapped sample ($N = 500$). The gray shaded area marks the $\tilde{t}_i \leq \tau$ region, for which the timescales estimation is not valid by construction.

The implied timescales plot displays the first three time scales $\tilde{t}_i(\tau)$ estimated at τ values between 1 and 15 ns. The slowest process appears with an associated timescales of $\tilde{t}_1 \approx 25$ ns, followed by two almost equally slow processes at $\tilde{t}_2 \approx \tilde{t}_3 \approx 13$ ns, in agreement with previously reported values.[11,18] As to the structural changes associated with these timescales, we look at the left eigenvectors using so-called kinetic maps. Plotting $\tilde{\phi}_{ik}$ against $\tilde{\phi}_{jk}$ for every k-th clustercenter highlights the most likely transitions as

$$\underset{k}{\text{argmin}}[\tilde{\phi}_{ik}] \leftrightarrow \underset{k}{\text{argmax}}[\tilde{\phi}_{ik}]$$

given that left eigenvectors are weighted by the stationary distribution, $\tilde{\mu}_k$. Such a kinetic map, for a lagtime of $\tau = 10$ ns, is shown in Figure 6.

In the transition along the horizontal axis of Figure 6, that is, in the slowest process of the system, the Trp-moiety changes its position relative to the MR121-moiety from above to below and *vice versa*, in accordance with previous studies. The second slowest process (the vertical axis of Figure 6) is a transition from a *folded* conformation, where the linker is folded between the Trp and MR121 moieties, to both the *unfolded* conformations (Trp above or below the MR121) and *vice versa*. Interestingly, this transition did not appear as the second slowest transition in our previous study.[18] However, the quasi degeneracy and uncertainties of $\tilde{t}_2$ and $\tilde{t}_3$ (cf. Figure 5) explain the inversion of the third and second processes in this case. As a matter of fact, inspection of the $\underset{k}{\text{argmin}}[\tilde{\phi}_{3k}] \leftrightarrow \underset{k}{\text{argmax}}[\tilde{\phi}_{3k}]$ transition reveals the known rotation of the terminal TRP-side chain.

DOI: 10.1021/acs.jctc.6b00738
J. Chem. Theory Comput. 2016, 12, 6118−6129

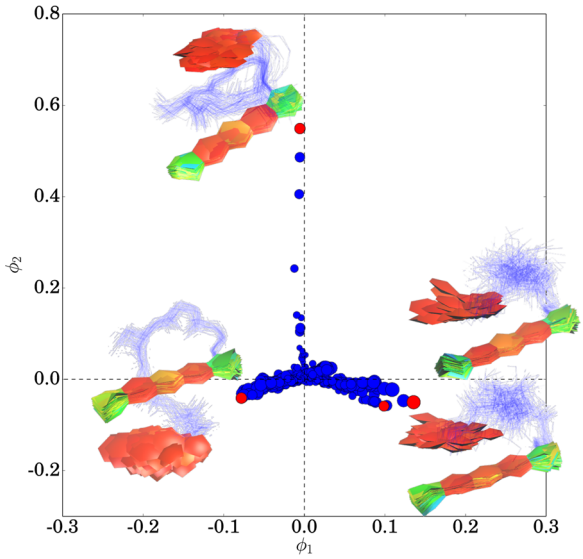


Figure 6. MR121-GSGS-W. Kinetic map of the first two left eigenvectors, $\tilde{\phi}_{1k}$ vs $\tilde{\phi}_{2k}$. Each k-th clustercenter is plotted as a dot, where the pair $(\tilde{\phi}_{1k}, \tilde{\phi}_{2k})$ is used as (x,y) coordinates. The stationary distribution, $\tilde{\mu}_k = \tilde{\phi}_{0k}$, gives the area of each dot. The most representative transitions, $\underset{k}{\text{argmin}}[\tilde{\phi}_{ik}] \leftrightarrow \underset{k}{\text{argmax}}[\tilde{\phi}_{ik}]$, of the first and second slowest processes are shown (i = 1, 2 cf. Figure 5). The shown molecular structures are an overlay of 50 members of the respective clustercenters (marked in red).

Finally, we compare hTICA and TICA as dimensionality reduction methods for MSM construction in Figure 7. Whereas panels a) and b) show small differences in the eigenvalues (and corresponding timescales), at the MSM level these differences do not play any role, as can be clearly seen in panel c).

4.2. hTICA with C_α-Distances in BPTI. As a second molecular system for validation of the hTICA method, we use a well-known data set of one millisecond of bovine pancreatic trypsin inhibitor (BPTI) in explicit water, as published (and made available) by Shaw and co-workers in 2010.[50] BPTI has been extensively studied and deemed the *workhorse* of molecular dynamics, given that its small size allows for longer simulation times, containing a number of conformational changes. Indeed, this one millisecond long BPTI trajectory allowed for the characterization of processes with lifetimes up to ∼30 μs.[50]

With 58 residues, BPTI is small as a protein but still considerably larger than the MR121-GSGS-W peptide. Hence, as input parameters for the hTICA run, we choose C_α-atom distances, instead of atom-wise distances. This choice results in $N = 58^2 - 58 = 3306$ input parameters. Except for the choice of a larger TICA lagtime, $\tau^\ddagger$ = 250 ns, the rest of the hTICA partition is similar to that of MR121-GSGS-W, which we summarize in Table 2 in the Supporting Information.

The level 1 and level 2 TICA eigenvalues are shown in Figure 8. Except for the first two indexes, level 2 eigenvalues decay much slower than the level 1 eigenvalues. In other words, the quality of the approximation for the lower eigenfunctions (ψ_1 and ψ_2) is already acceptable at level 1 TICA. For higher indexes, the approximation is considerably improved by level 2 TICA. This was not the case for MR1-GSGS-W, where level 2 TICA improved all eigenvalues comparably. Again, the full TICA solution is presented for comparison, and, again, the hTICA approximation is almost indistinguishable from the TICA solution.

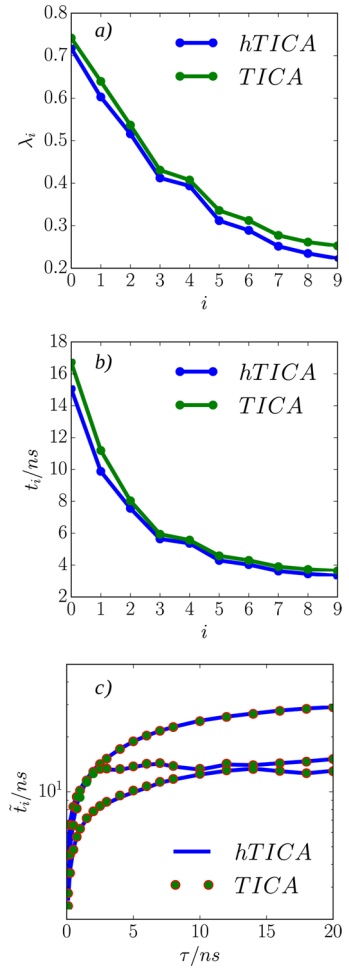


Figure 7. Comparison of TICA and hTICA through a) eigenvalues, b) implied timescales (continuous), and c) implied timescales of the resulting discrete MSM models. Although hTICA does not recover the full TICA solution exactly (a, b), this does not have any effect on the quality of hTICA as dimensionality reduction (c).

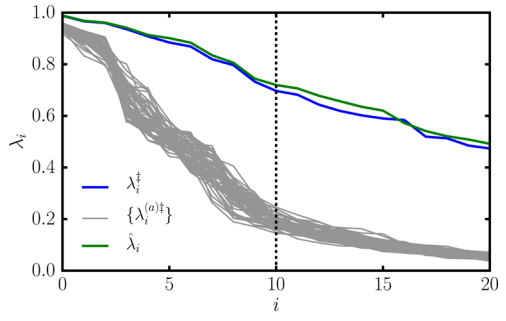


Figure 8. BPTI-C$_\alpha$. Overlayed in gray are all N_a sets of level 1 TICA eigenvalues, $\lambda_i^{(a)\ddagger}$, in blue the one set of level 2 TICA, $\lambda^\ddagger$. The vertical dotted line marks the chosen m_a = 10 value as a visual aid: the blue curve results of mixing all the level 1 eigenfunctions, $\psi^{(a)\ddagger}$ associated with the gray eigenvalues to the left of the dotted line. As a comparison, we also show the eigenvalues of the full TICA, $\hat{\lambda}_i$.

Before moving on to MSM construction, we briefly present hTICA results of randomized balanced partition schemes. While keeping the block size equal (N_r = 57, in this case), we randomize what C_α-distances are grouped together in each block. So far, we have always ensured that each block represents one atom and contains all possible contacts of that atom with

DOI: 10.1021/acs.jctc.6b00738
J. Chem. Theory Comput. 2016, 12, 6118−6129

the rest of the molecule. However, hTICA results appear very robust with respect to other partition choices where this is not necessarily ensured. The highly redundant space of all possible distances and the two-step nature of hTICA avoid missing important correlations in either one of the two steps. Figure 9

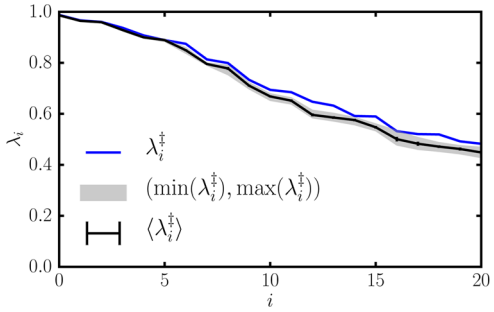


Figure 9. BPTI-C_α. Comparison of the level 2 TICA eigenvalues for different partition schemes. In black is the expected value and standard deviation of the eigenvalue, $\langle\lambda_i^{\ddagger}\rangle$, over 1000 randomized balanced partitions. The interval $(\min(\lambda_i^{\ddagger}),\max(\lambda_i^{\ddagger}))$ for each i is shown in gray. The spectrum of the atom-aware partition is shown (cf. Figure 8) in blue.

displays the robustness of the spectrum with respect to the partition scheme. The expectation of the dominant hTICA spectrum over the sampled partitions is only slightly below the spectrum of the atom-aware partition. Although small differences in the hTICA spectrum have almost no impact in the MSM construction (cf. Figure 7), we continue to use the atom-based partition, that appears to be slightly better than randomized partitions and ensures reproducibility.

Following the same procedure as above, the first 10 level 2 TICs are clustered onto 1000 *kmeans* clustercenters, and a transition matrix is estimated for different lagtimes until a convergence of the timescales is observed. The resulting implied-timescales plot is shown in Figure 10.

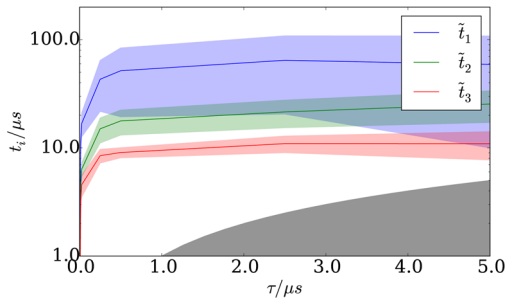


Figure 10. BPTI-C_α. *Semilog* plot of the implied timescales (ITS, or t_i) for a Markov State Model built on the *kmeans*-discretization (1000 centers) in the space spun by the first 10 hTICs, $\psi_i^{\ddagger}$. The color-shaded areas mark the 2σ confidence intervals of a bootstrapped sample (N = 500). The gray shaded area marks the $\tilde{t}_i \leq \tau$ region, for which the timescales estimation is not valid by construction.

From Figure 10, we see that the first 3 slowest processes have implied timescales of ~60, 25, and 11 μs and that these time scales appear converged for lagtimes of $\tau \sim 1-5$ μs. Using a lagtime of τ = 5 μs, we show the conformational transitions associated with these timescales, together with their kinetic map in Figure 11. In this case, we have decided to display together both the argmin and the argmax of each $\tilde{\phi}_i$, so that the transition can be seen clearly. We also highlight as dotted lines in Figure 11c) the input coordinates, χ_j, that most correlate

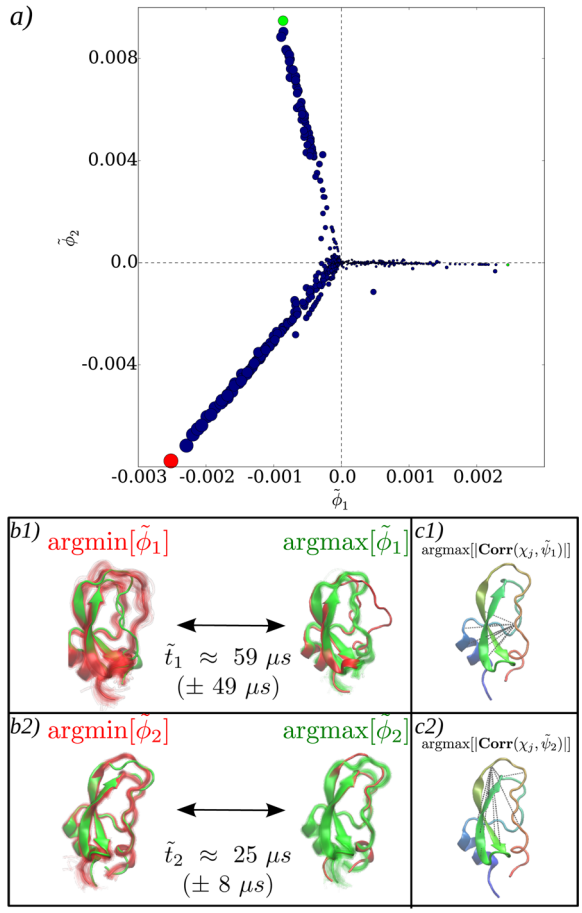


Figure 11. BPTI-C_α. a) Kinetic map (cf. Figure 5) of the first two left eigenvectors of $\mathbf{T}(\tau = 5\ \mu s,)$, $\tilde{\phi}_1$ vs $\tilde{\phi}_2$ for BPTI. The centers corresponding to the argmin and argmax of each vector appear as red and green dots, respectively. b1) and b2) show the molecular structures corresponding to the argmin and argmax of $\tilde{\phi}_1$ vs $\tilde{\phi}_2$ (same colors as in a). The argmin and the argmax are overlaid to highlight the conformational transitions. These are less obvious for BPTI than in the MR121-GSGS-W peptide. c1) and c2). To further help visualize the structural changes associated with the two slowest processes, the input parameters that are most correlated with the right eigenvectors, $\tilde{\psi}_i$, are shown. This way, one maps the MSM-information (which is occurring in discrete state space) to the continuous space of the input parameters.

with the right eigenvectors, $\tilde{\psi}_i$, of the transition matrix to further help in representing the process.

From Figure 11, it can be seen that the slowest process ($\tilde{t}_1 \approx 60$ μs) is a large amplitude motion, involving C_αs 7−11 (approximately), moving from a rather external/open conformation to an internal one. The C_α of Pro9 (which is highlighted in 11c1) is displaced by about 5 Å between the extrema of $\tilde{\phi}_1$. The second, faster process ($\tilde{t}_2 \approx 20$ μs) is also a rearrangement of the unstructured region of the protein; however it is smaller in amplitude and number of residues involved: the C_αs 13−18 (approximately) simply *flip* their orientation. The highlighted atom (C_α of Ala16, cf. 11c2)) is displaced by just ~2 Å.

These two modes of backbone motion coincide roughly with those identified in the original publication, as seen in Figure 4 B of that paper,[50] and a previous Hidden-Markov Model based analysis.[51] The colored cartoon representation of Figure 4B and our hTICA MSM highlight the same areas as most important contributors to the slowest dynamics, namely those involving the backbone.

DOI: 10.1021/acs.jctc.6b00738
J. Chem. Theory Comput. 2016, 12, 6118−6129

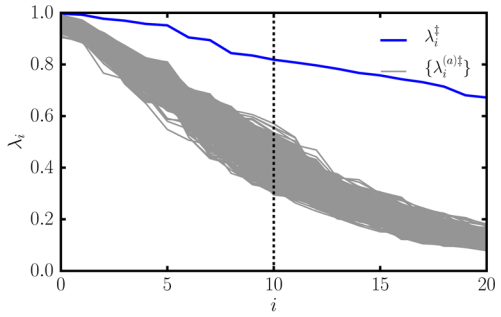


Figure 12. BPTI-heavy-atom. Overlayed in gray are all N_a sets of level 1 TICA eigenvalues, $\lambda_i^{(a)\ddagger}$, in blue the set of level 2 TICA, $\lambda^{\ddagger}$. The vertical dotted line marks the chosen $m_a = 10$ value as a visual aid: the blue curve results of mixing all the level 1 eigenfunctions, $\psi^{(a)\ddagger}$, associated with the gray eigenvalues to the left of the dotted line. In this case, there is no full TICA ($\hat{\lambda}_i$) to compare with, given the huge size of the input set (cf. Table 1).

4.3. hTICA Using Heavy-Atom Distances in BPTI. As a final validation, we use hTICA to analyze the BPTI trajectory using all heavy-atom pairwise distances. This brute-force choice of the basis set has been made deliberately to demonstrate hTICA's ability to deal with high numbers of redundant input parameters with little user input and still arrive at a converged Markov model. With 454 heavy atoms, the redundant list of pairwise distances contains $N = 454^2 - 454 = 205662$ elements, potentially resulting in matrices of size $\sim 2 \times 10^5 \times 2 \times 10^5$. The hTICA parameters are summarized in Table 2 in the Supporting Information.

The 454 sets of level 1 TICA eigenvalues, $\lambda_i^{(a)\ddagger}$, are shown in Figure 12, together with the level 2 TICA eigenvalues. As is characteristic, the level 2 TICA eigenvalues represent an improvement with respect to level 1 values.

4.3.1. TICA and Disconnectivity Issues. The localization of the slowest processes is inherent to the TICA (and hTICA) method. The slowest possible processes, captured as linear combinations of input parameters (and linear combinations thereof), *float* to the top of the list of TICs, shaping the dominant TICA subspace.

In the case of all heavy-atom pairwise distances for BPTI, this leads to a kinetically disconnected set: The first TIC captures a nonreversible drift, as seen in Figure 13. This drift accounts for two separate events that happen only once, sequentially, during the simulation: the rotation of the aromatic side chains of TYR21 and TYR23. In the original publication, this finding

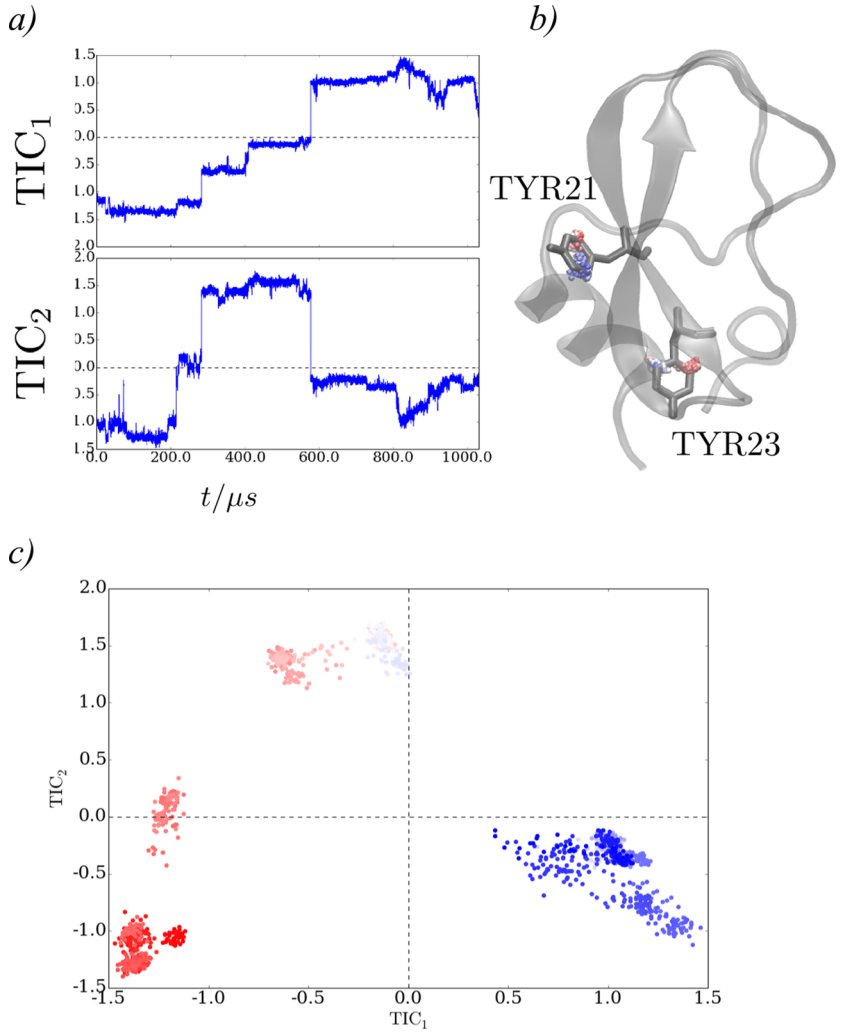


Figure 13. BPTI-heavy-atom. a) Time evolution of the first and second hTICs. The first and slowest TIC display a drift that is incompatible with reversible kinetic connectivity. The reversible MSM can only be constructed in the second half (ca. $t > 600$ μs) of the trajectory. b) BPTI structure where the transitions associated with TIC_1 have been highlighted. They correspond to χ_2 angles of the aromatic residues TYR21 and TYR23, which

DOI: 10.1021/acs.jctc.6b00738
J. Chem. Theory Comput. 2016, 12, 6118−6129

appears correctly noted as the slowest aromatic rotation in the supplementary material (Table S3 in ref 50). There, the associated timescales for TYR21 and TYR23 are listed as ca. 1 ms, which, naturally, correspond to each flip happening only one time in the 1 ms trajectory. It is worth noting that by virtue of hTICA, there is no need for the user to separately look for these ring rotations, since they readily surface as the slowest processes in the all-heavy-atom description. The equivalent *brute-force* approach needed to detect this level of atomistic detail (not present in the C_α-description) without guiding the algorithm is simply unfeasible.

We proceed to further construct the MSMs for the first 10 hTICs; however, we restrict ourselves to the kinetically connected data set, which corresponds roughly to the second half of the trajectory (ca. $t > 600$ μs).

Analogous to the previous cases, we cluster the trajectory in the first 10 dominant hTICA coordinates using $n = 1000$ *kmeans* clustercenters. Subsequently, we estimate the transition matrix for different lagtimes and produce the implied time scales (ITS) plot shown in Figure 14.

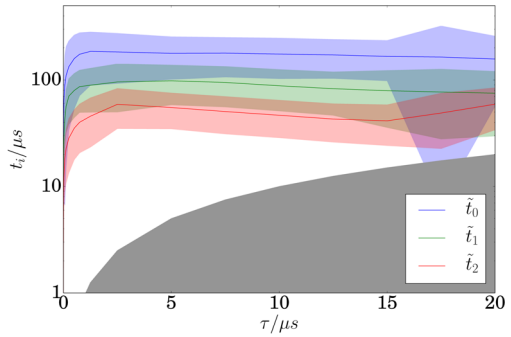


Figure 14. BPTI-heavy-atom. *Semilog* plot of the first three implied timescales (ITS, or t_i) for a Markov State Model built on the *kmeans*-discretization (1000 centers) in the space spanned by the first 10 hTICs, $\psi_i^{\dagger}$. The color-shaded areas mark the 2σ confidence intervals of a bootstrapped sample ($N = 500$). The gray shaded area marks the $\tilde{t}_i \leq \tau$ region, for which the timescales estimation is not valid by construction.

The ITSs shown in Figure 14 appear converged for lagtimes $\tau > 2.5$ μs, with the first three processes happening at timescales of ~178, 98, and 55 μs. Considering this, we choose the transition matrix at a lag $\tau = 5$ μs to investigate the conformational transitions associated with these ITSs. We show the map of the left eigenvectors, $\tilde{\phi}_1$ vs $\tilde{\phi}_2$, in Figure 15.

The two slowest processes are again aromatic ring rotations with the particularity of having timescales comparable to the backbone motions described above in the C_α-basis set (although these values carry considerable error bars). From Figure 14 and Figure 15 we assign ca. 178 μs to the TYR35 rotation and 98 μs for PHE22. Expressed as rates, these timescales correspond to ca. 6×10^3 and 1×10^3 s^{-1}, respectively (cf. Table S3 in ref 50, both these rotations appear as 1×10^3 s^{-1}).

The remaining, faster processes contain both aromatic side chain rotations and the known backbone motions. For the aromatic ring flips, we find that TYR10 presents a timescales of ca. 21 μs vs ca. 1 μs reported in ref 50 (see Figure 16).

5. CONCLUSIONS

Many existing dimensionality reduction methods, such as PCA,[15,16] TICA,[18,35] and the variational approach of conformation dynamics,[24,36] are guaranteed to find variationally optimal combinations of the input coordinates but become

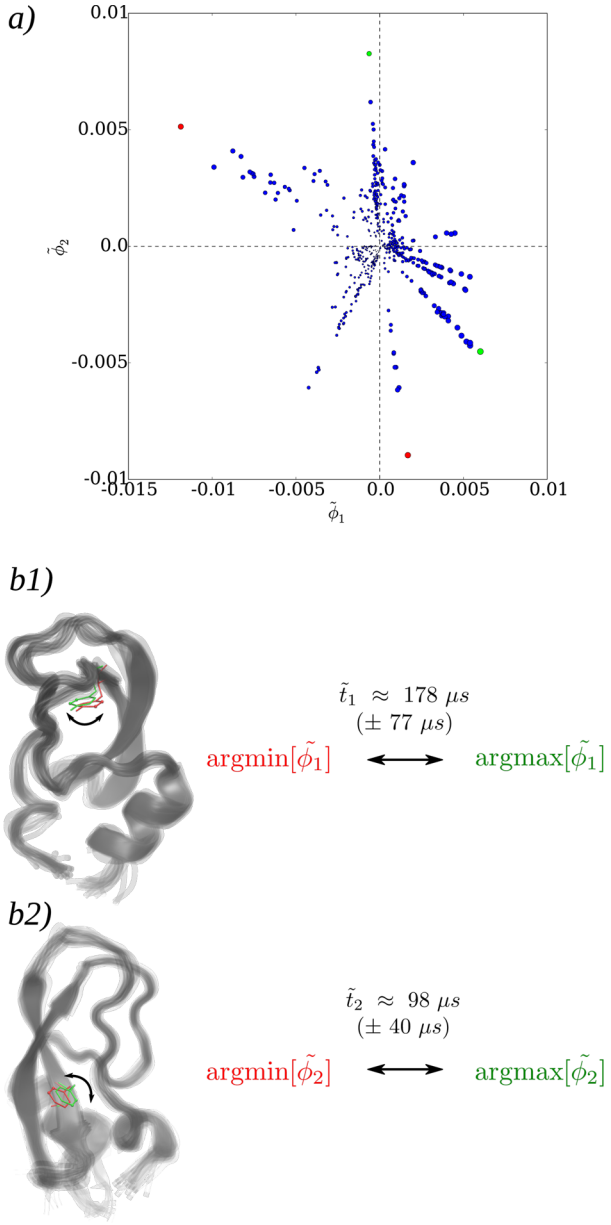


Figure 15. BPTI-heavy-atom. a) Kinetic map of the first two left eigenvectors of T($\tau = 5$ μs,), $\tilde{\phi}_1$ vs $\tilde{\phi}_2$. Highlighted in red and green are the argmin and argmax of each eigenvector. These centers provide the molecular structures associated with each processes. An overlay of theses structures is shown in b1) and b2). We have chosen to highlight only the residues in which the transition happens: TYR35 and PHE22.

computationally intractable when the set of input coordinates is too large, such as all pairs of atom−atom distances or residue−residue distances in large macromolecular systems.

We have presented a divide-and-conquer approach, hierarchical TICA (hTICA), in which the variational nature of the TICA algorithm is exploited sequentially in two steps. After a distributed first round of TICA, separately computed TICs are mixed (i.e., their time-lagged covariances computed) in a second round of TICA. The "new" TICs emerging from the second round are guaranteed to be a better projection of the slow dynamics, as is shown by larger eigenvalues of the normalized time lagged covariance matrix. Due to this hierarchical selection procedure, hTICA does not consider all pair correlations between input coordinates and can thus not be guaranteed to

DOI: 10.1021/acs.jctc.6b00738
J. Chem. Theory Comput. 2016, 12, 6118−6129

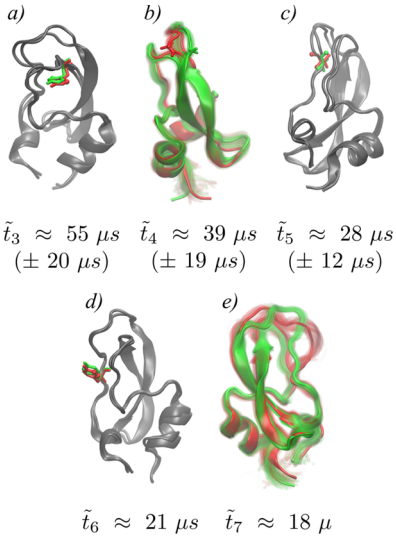


Figure 16. BPTI-heavy-atom. Molecular structures of the argmin and argmax of $\tilde{\mathbf{\Phi}}_i$, for $i = 3\cdots7$. The third and sixth slowest process correspond to χ_2-flips of the aromatic TYR moieties: a) is again TYR35 and d) is TYR10. The fourth and seventh processes (b and e) are the backbone motions of the processes already described for the pairwise C_α metric (cf. Figure 11). They appear inverted in order and at different timescales (39 μs instead of 25 μs and 18 μs instead of 59 μs, respectively); however, they fall well within the confidence intervals. Ultimately, the fifth process (c) has resolved a backbone flip in the Ramachandran angles of residue THR10.

provide an optimal linear combination of the full coordinate set. However, it is practically found to provide an excellent approximation that is well suited for dimension reduction of the input coordinate space for the construction of MSMs, at a much reduced cost when compared to the full TICA problem.

In the largest molecule investigated here (BPTI simulation data from D.E. Shaw Research[50]), hTICA correctly identified the slow modes and pointed out a previously undetected irreversible drift in in the aromatic side chain flips of residues TYR21 and TYR23.

When used in conjunction with pairwise distances between N_a atoms, residues, or chemical groups, the computational effort of hTICA scales as N_a^3 in contrast to the full TICA scaling of N_a^4. As N_a^3 might still be prohibitive for large proteins and assemblies, additional heuristics can be used to reduce the computational cost. For example, the number of coordinates may be reduced by considering only distances between atoms (groups) that ever form or break contacts or by considering only distances that experience significant contact lifetimes in the simulation.

As high-throughput MD data becomes increasingly easy to generate,[52] hTICA offers an efficient approach to approximate the relevant reaction coordinates and construct high-quality kinetic models. As it offers an affordable way to describe the slow dynamic modes of protein motion while using a high-dimensional and detailed set of structural features as an input, hTICA may help to characterize allosteric effects, e.g. understand the coupling of global motions to atomic-detail events.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jctc.6b00738.

Information regarding the choices for the TICA-lagtime, the parameter m_a, a summary of all used parameters, and the Chapman-Kolmogorov tests (PDF)

■ AUTHOR INFORMATION

Corresponding Authors

*E-mail: guille.perez@fu-berlin.de.
*E-mail: frank.noe@fu-berlin.de.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

We are grateful to D.E. Shaw Research for providing the BPTI simulation data from ref 50. We acknowledge funding from the Deutsche Forschungsgemeinschaft, grants NO825-3/1 (G.P.-H.), the European Commission, ERC starting grant "pcCell" (G.P.-H. and F.N.), and the federal state of Berlin (G.P.-H. and F.N.).

■ REFERENCES

(1) Goldbeck, R. A.; Thomas, Y. G.; Chen, E.; Esquerra, R. M.; Kliger, D. S. *Proc. Natl. Acad. Sci. U. S. A.* **1999**, *96*, 2782−2787.
(2) Baldwin, R. L. *J. Mol. Biol.* **2007**, *371*, 283−301.
(3) Sborgi, L.; Verma, A.; Piana, S.; Lindorff-Larsen, K.; Cerminara, M.; Santiveri, C. M.; Shaw, D. E.; de Alba, E.; Muñoz, V. *J. Am. Chem. Soc.* **2015**, *137*, 6506−6516.
(4) Katritch, V.; Cherezov, V.; Stevens, R. C. *Annu. Rev. Pharmacol. Toxicol.* **2013**, *53*, 531−56.
(5) Katz, B. A.; et al. *J. Mol. Biol.* **2001**, *307*, 1451−1486.
(6) Talhout, R.; Engberts, J. B. F. N. *Eur. J. Biochem.* **2001**, *268*, 1554−1560.
(7) Min, W.; English, B. P.; Luo, G.; Cherayil, B. J.; Kou, S. C.; Xie, X. S. *Acc. Chem. Res.* **2005**, *38*, 923−931.
(8) Kim, E.; Lee, S.; Jeon, A.; Choi, J. M.; Lee, H.-S.; Hohng, S.; Kim, H.-S. *Nat. Chem. Biol.* **2013**, *9*, 313−318.
(9) Deniz, A. A.; Mukhopadhyay, S.; Lemke, E. A. *J. R. Soc., Interface* **2008**, *5*, 15−45.
(10) Keller, B. G.; Kobitski, A.; Jäschke, A.; Nienhaus, G. U.; Noé, F. *J. Am. Chem. Soc.* **2014**, *136*, 4534−4543.
(11) Noé, F.; Doose, S.; Daidone, I.; Löllmann, M.; Chodera, J. D.; Sauer, M.; Smith, J. C. *Proc. Natl. Acad. Sci. U. S. A.* **2011**, *108*, 4822−4827.
(12) Zhuang, W.; Cui, R. Z.; Silva, D.-A.; Huang, X. *J. Phys. Chem. B* **2011**, *115*, 5415−5424.
(13) Keller, B.; Prinz, J.-H.; Noé, F. *Chem. Phys.* **2012**, *396*, 92−107.
(14) Lindner, B.; Yi, Z.; Prinz, J.-H.; Smith, J. C.; Noé, F. *J. Chem. Phys.* **2013**, *139*.
(15) Amadei, A.; Linssen, A. B.; Berendsen, H. J. C. *Proteins: Struct., Funct., Genet.* **1993**, *17*, 412−225.
(16) Hotelling, H. *J. Edu. Psych.* **1933**, *24*, 417−441.
(17) Altis, A.; Nguyen, P. H.; Hegger, R.; Stock, G. *J. Chem. Phys.* **2007**, *126*, 244111.
(18) Pérez-Hernández, G.; Paul, F.; Giorgino, T.; de Fabritiis, G.; Noé, F. *J. Chem. Phys.* **2013**, *139*, 015102.
(19) Peters, B.; Trout, B. L. *J. Chem. Phys.* **2006**, *125*, 054108.
(20) Peters, B. *J. Chem. Phys.* **2006**, *125*, 241101.
(21) Schütte, C.; Fischer, A.; Huisinga, W.; Deuflhard, P. *J. Comput. Phys.* **1999**, *151*, 146−168.
(22) Deuflhard, P.; Weber, M. In *Linear Algebra Appl.*; Dellnitz, M., Kirkland, S., Neumann, M., Schütte, C., Eds.; Elsevier: New York, 2005; Vol. *398C*, pp 161−184.
(23) Prinz, J.-H.; Chodera, J. D.; Noé, F. *Phys. Rev. X* **2014**, *4*, 011020.
(24) Noé, F.; Nüske, F. *Multiscale Model. Simul.* **2013**, *11*, 635−655.
(25) Schütte, C.; Noé, F.; Lu, J.; Sarich, M.; Vanden-Eijnden, E. *J. Chem. Phys.* **2011**, *134*, 204105.
(26) Swope, W. C.; Pitera, J. W.; Suits, F. *J. Phys. Chem. B* **2004**, *108*, 6571−6581.
(27) Buchete, N. V.; Hummer, G. *J. Phys. Chem. B* **2008**, *112*, 6057−6069.
(28) Noé, F.; Horenko, I.; Schütte, C.; Smith, J. C. *J. Chem. Phys.* **2007**, *126*, 155102.

DOI: 10.1021/acs.jctc.6b00738
J. Chem. Theory Comput. 2016, 12, 6118−6129

(29) Chodera, J. D.; Dill, K. A.; Singhal, N.; Pande, V. S.; Swope, W. C.; Pitera, J. W. *J. Chem. Phys.* **2007**, *126*, 155101.
(30) Prinz, J.-H.; Wu, H.; Sarich, M.; Keller, B.; Senne, M.; Held, M.; Chodera, J. D.; Schütte, C.; Noé, F. *J. Chem. Phys.* **2011**, *134*, 174105.
(31) *An Introduction to Markov State Models and Their Application to Long Timescale Molecular Simulation. Advances in Experimental Medicine and Biology*; Bowman, G. R., Pande, V. S., Noé, F., Eds.; Springer: Heidelberg, 2014; Vol. *797*.
(32) Voelz, V. A.; Bowman, G. R.; Beauchamp, K.; Pande, V. S. *J. Am. Chem. Soc.* **2010**, *132*, 1526−1528.
(33) Bowman, G. R.; Bolin, E. R.; Hart, K. M.; Maguire, B. C.; Marqusee, S. *Proc. Natl. Acad. Sci. U. S. A.* **2015**, *112*, 2734−2739.
(34) Sarich, M.; Noé, F.; Schütte, C. *Multiscale Model. Simul.* **2010**, *8*, 1154−1177.
(35) Schwantes, C. R.; Pande, V. S. *J. Chem. Theory Comput.* **2013**, *9*, 2000−2009.
(36) Nüske, F.; Keller, B.; Pérez-Hernández, G.; Mey, A. S. J. S.; Noé, F. *J. Chem. Theory Comput.* **2014**, *10*, 1739−1752.
(37) Molgedey, L.; Schuster, H. G. *Phys. Rev. Lett.* **1994**, *72*, 3634−3637.
(38) Naritomi, Y.; Fuchigami, S. *J. Chem. Phys.* **2011**, *134*, 065101.
(39) Naritomi, Y.; Fuchigami, S. *J. Chem. Phys.* **2013**, *139*, 215102.
(40) Noe, F.; Clementi, C. *J. Chem. Theory Comput.* **2015**, *11*, 5002−5011.
(41) Noé, F.; Banisch, R.; Clementi, C. *J. Chem. Theory Comput.* **2016**, *12* (11), 5620−5630.
(42) Wu, H.; Nüske, F.; Paul, F.; Klus, S.; Koltai, P.; Noé, F. 2016, submitted for publication.
(43) Coifman, R. R.; Lafon, S.; Lee, A. B.; Maggioni, M.; Nadler, B.; Warner, F.; Zucker, S. W. *Proc. Natl. Acad. Sci. U. S. A.* **2005**, *102*, 7426−7431.
(44) Rohrdanz, M. A.; Zheng, W.; Maggioni, M.; Clementi, C. *J. Chem. Phys.* **2011**, *134*, 124116.
(45) McGibbon, R. T.; Pande, V. S. *J. Chem. Phys.* **2015**, *142*, 124105.
(46) Senne, M.; Trendelkamp-Schroer, B.; Mey, A. S. J. S.; Schütte, C.; Noé, F. *J. Chem. Theory Comput.* **2012**, *8*, 2223−2238.
(47) Bowman, G. R.; Beauchamp, K. A.; Boxer, G.; Pande, V. S. *J. Chem. Phys.* **2009**, *131*, 124101.
(48) Neuweiler, H.; Löllmann, M.; Doose, S.; Sauer, M. *J. Mol. Biol.* **2007**, *365*, 856−869.
(49) Daidone, I.; Neuweiler, H.; Doose, S.; Sauer, M.; Smith, J. C. *PLoS Comput. Biol.* **2010**, *6*, e1000645.
(50) Shaw, D. E.; Maragakis, P.; Lindorff-Larsen, K.; Piana, S.; Dror, R.; Eastwood, M.; Bank, J.; Jumper, J.; Salmon, J.; Shan, Y.; Wriggers, W. *Science* **2010**, *330*, 341−346.
(51) Noé, F.; Wu, H.; Prinz, J.-H.; Plattner, N. *J. Chem. Phys.* **2013**, *139*, 184114.
(52) Doerr, S.; Harvey, M. J.; Noé, F.; De Fabritiis, G. *J. Chem. Theory Comput.* **2016**, *12*, 1845−1852.

DOI: 10.1021/acs.jctc.6b00738
J. Chem. Theory Comput. 2016, 12, 6118−6129