JACS
JOURNAL OF THE AMERICAN CHEMICAL SOCIETY
Cite This: J. Am. Chem. Soc. 2018, 140, 2386−2396
Perspective
pubs.acs.org/JACS

Markov State Models: From an Art to a Science

Brooke E. Husic and Vijay S. Pande*

Department of Chemistry, Stanford University, Stanford, California 94305, United States

ABSTRACT: Markov state models (MSMs) are a powerful framework for analyzing dynamical systems, such as molecular dynamics (MD) simulations, that have gained widespread use over the past several decades. This perspective offers an overview of the MSM field to date, presented for a general audience as a timeline of key developments in the field. We sequentially address early studies that motivated the method, canonical papers that established the use of MSMs for MD analysis, and subsequent advances in software and analysis protocols. The derivation of a variational principle for MSMs in 2013 signified a turning point from expertise-driving MSM building to a systematic, objective protocol. The variational approach, combined with best practices for model selection and open-source software, enabled a wide range of MSM analysis for applications such as protein folding and allostery, ligand binding, and protein−protein association. To conclude, the current frontiers of methods development are highlighted, as well as exciting applications in experimental design and drug discovery.

1. INTRODUCTION

In his 1983 report on short memory approximations, Zwanzig considers the random walk of a dynamical system through a set of states. Zwanzig notes that if these states are chosen "sensibly" and the dynamics they describe are "sufficiently complex", the system will not remember how it got to its current state.[1] In this case, Zwanzig writes, "one then has a remarkably simple way to estimate the transition rates between cells."[1] This simple method to estimate transition rates has blossomed into the subfield of kinetic analysis with memoryless transition networks often referred to as Markov state models (MSMs). Paired with advances in molecular dynamics (MD) simulations, the MSM framework has grown into an advanced, statistically robust set of methods to analyze a system evolving in time. Crucially, the states no longer need to be chosen based on experience and intuition due to methods advances in the field. As a result of these advances, the full potential of MSM methods are accessible to the wider scientific community as a set of established methods implemented in several open source software packages.

In this perspective, we track the growth of the MSM field from its beginning as a specialized art driven by expertise and intuition, to a community-driven science that can be applied using objective protocols. We first summarize the MSM method, focusing on the type of information that can be obtained from transition probabilities. Ultimately, our goal is to communicate to a general scientific audience how and why the MSM framework evolved to where it is today, such that researchers across many fields can utilize this framework when analyzing dynamical systems.

In Section 2, we introduce the MSM framework and the overview information provided by the model. Sections 3 and 4 highlight critical precursors to MSM methods, and Sections 5 through 7 provide a timeline of the research that facilitated the development through the current state of the art. Finally, we highlight exciting developments at the frontiers of both methods and applications: Section 8 highlights the most recent methods advances under current development, and Section 9 describes studies that leverage the union of MSMs with experimental observables. Finally, Section 10 presents our outlook for future directions.

2. SUMMARY OF THE METHOD

Our goal is to provide a concise, contextualized review of the crucial developments in the field for a general scientific audience so that MSM methods can be used beyond the computational communities in which they were developed. We will present the concepts sometimes using algebraic terms for simplicity, but will largely omit the details of the mathematics. For excellent overviews of the theory underlying MSMs, we refer the reader to any of several books on the topic[2−4] or a recent review by Wang et al.[5] Before discussing the evolution of the field, we first cover the basics of what kind of information a MSM provides about a dynamical system.

A MSM represents a master equation framework: this means that, using just the MSM, the entire dynamics of the system can be described. The master equation formalism has been used in many scientific fields, and we refer the reader to ref 6 for commentary on its range of applications. In our case, a MSM is used to model a dynamical system, usually assumed to be in thermodynamic equilibrium. The MSM itself is an $n \times n$ square matrix, often a "transition probability matrix", where the entire configuration space spanned by the system has been divided into n states. By determining the states, we can track the dynamical progress of a system (e.g., a molecular dynamics simulation) by writing down which state it occupies at time points separated by τ, often referred to as the lag time. For the lag time τ to be Markovian, the system must be "memoryless", which means the probability that, after the next increment of τ, the system transitions to state y given it is in state x does not depend on where the system was before it entered state x.

The transition probability matrix described above is thus characterized by the n states and also by the lag time τ at which the state of the system was recorded. Each row of this matrix is a distribution that represents the probability of transitioning from the row-indexed state to the column-indexed state, where the diagonal entry represents the probability of staying in the same state. State populations and conditional pairwise transition

Received: November 17, 2017
Published: January 11, 2018

DOI: 10.1021/jacs.7b12191
J. Am. Chem. Soc. 2018, 140, 2386−2396

probabilities can be obtained from this matrix. The state populations can be easily converted to free energies using statistical mechanics, and the transition probabilities yield kinetic information as well as enumerations of possible pathways between any pair of states.

If we require that the system is in thermodynamic equilibrium (every transition happens an equal number of times forward and backward), symmetric with respect to an equilibrium distribution, ergodic (starting from any state, every other state can be reached given enough time), and aperiodic (every starting arrangement of the system will lead to the same equilibrium distribution), then the transition matrix gives us yet more information about the system through its eigendecomposition. This decomposition outputs a set of eigenvectors (column vectors) and corresponding eigenvalues (real numbers). The eigenvectors are approximations to the eigenfunctions of the transfer operator, which is the continuous integral operator that the transition probability matrix approximates.[7] Because of the properties we assumed about the transition matrix, we automatically know some things about the eigenvalues and eigenvectors:

1. The eigenvectors corresponding to each eigenvalue have n elements corresponding to each of the n states. The magnitudes and signs of these elements explain which states are contributing to the process identified by the eigenvalue.
2. The highest eigenvalue is 1, and its corresponding eigenfunction represents the equilibrium distribution.
3. All the other eigenvalues have absolute values less than 1, and represent processes that either decay to equilibrium (positive eigenvalues) or oscillate (negative eigenvalues). In practice, the latter are not physically meaningful and are not used for analysis.
4. The positive eigenvalues can be converted to physically meaningful timescales using the lag time τ at which the transition matrix was defined.

Ultimately, we want to choose the n states such that they best capture the dynamics of the system, and a lag time τ that is long enough to be Markovian but also short enough to resolve the system dynamics. If we can successfully do that, the MSM provides valuable information about the system, all from just its transition matrix. This information is summarized in Figure 1. Now that we have covered the essential theory, we are ready to see how the MSM field developed from Zwanzig's description to the present time.

3. BEFORE THE MSM: ADVANCES THAT SET THE STAGE

Although the idea of memoryless transition networks was presented by Zwanzig[1] and extensively discussed in Van Kampen,[2] the framework was not presented in full until several key papers around the year 2000.[7−9] However, the previous decade contained many important advances that led to MSMs, including some prescient studies containing ideas that would not resurface until the 2010s. Many early studies using a transition matrix or network with discrete states established foundational work upon which the MSM field was built.[10−32] We highlight three key ideas in the development of MSMs: the difficulty of choosing states, the search for appropriate collective variables, and the ability to integrate separate simulations using the MSM framework.

In their discussion of memory effects on measuring the first passage time of a system, Hünggi and Talkner[33] discussed a criterion to assess the validity of the Markovian approximation, highlighting the difficulty in determining the proper functions to approximate the dynamics as Markovian. This commentary, echoed by others,[34] foreshadows a difficulty that would characterize more than a decade of MSM methods development: namely, that the state decomposition step (choosing the n states on which to build the model) is both crucial and practically difficult, since it requires expert knowledge of the system.

Progress was also made in determining the proper collective variables along which to identify model states. In 1993, Karpen et al.[10] sought to create states (clusters) that provided useful information about the system, and found that using the similarity in dihedral angles separated groups of structures "in an energetically significant way."[10] In 2001, de Groot et al.[20] analyzed a β-heptapeptide simulation by performing PCA on the atomic

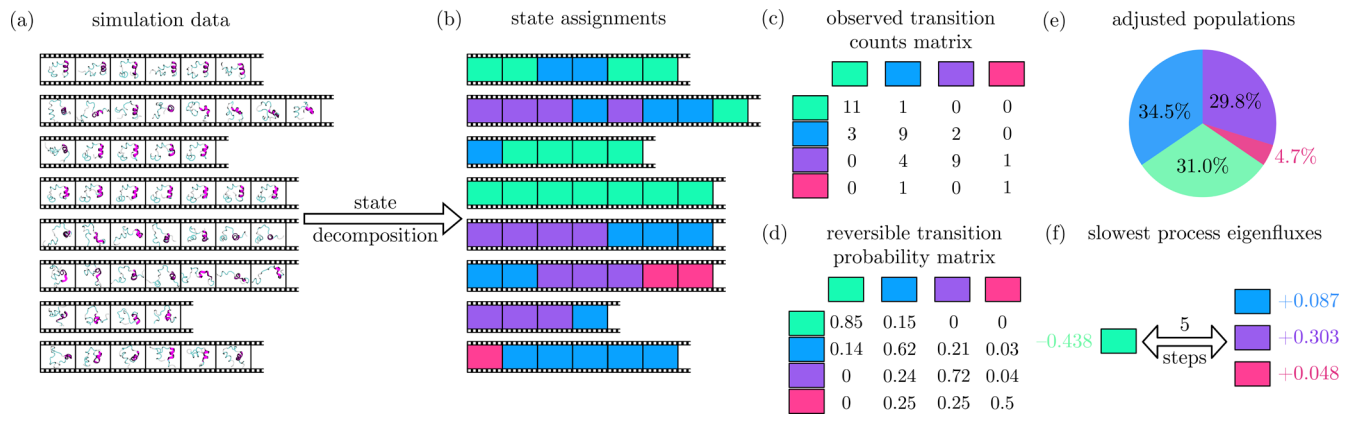

Figure 1. (a) Eight short molecular dynamics trajectories, where each frame is separated by a simulation time step. (b) A discrete state decomposition of the trajectories, in which four states (green, blue, purple, and pink) have been identified. (c) Observed transition counts matrix, which records how many times the system transitions to the column-indexed state given that it is in the row-indexed state. (d) Reversible transition probability matrix, in this case calculated by averaging the observed transition counts matrix with its transpose in order to ensure that every forward transition happens an equal number of times backward. Usually, the MLE is used instead.[46] The probabilities in each row sum to 1. (e) Pie chart of adjusted populations, which is the first eigenvector of the transition probability matrix. States with greater populations are more thermodynamically stable. (f) Schematic representing the dominant eigenprocess of the trajectory data set. The eigenflux of the green state is negative, while the eigenfluxes of all other states are positive, so this process represents movement from the green state to the other states. Whereas the populations sum to 1, the eigenfluxes for each process sum to 0. The timescale of this process is five trajectory timesteps.

DOI: 10.1021/jacs.7b12191
J. Am. Chem. Soc. 2018, 140, 2386−2396

coordinates before determining states using the *k*-medoids clustering algorithm on the first three principal components (PCs). They assessed that the degrees of freedom not captured by the first three PCs "must be negligible in terms of both structure and dynamics."[20]

Ultimately, one of the strengths of MSM analysis in the 2000s would prove to be its ability to integrate many separate simulations, such as those produced on the distributed computing platform Folding@home.[35] However, well before the creation of Folding@home, Luty and McCammon[11] created a 4-state Markovian model for a bimolecular reaction by aggregating the results of separate simulations performed in different domains, demonstrating that Markov state models can integrate separate simulations into the same model. Since the system is memoryless, valid paths among the states do not need to be fully sampled by any single trajectory: instead, separate trajectories can be threaded together when they occupy common states. In a following paper, Luty et al.[12] suggested that this method could be used to analyze ligand binding.

4. HARDWARE PIONEERS THAT CATALYZED MSM DEVELOPMENT

In 2000, Shirts and Pande[35] introduced Folding@home, a distributed platform for performing many MD simulations in parallel motivated by the ensemble dynamics method.[36] This method leveraged the stochasticity of a swarm of trajectories by moving every trajectory across a barrier once the first trajectory crossed that barrier.[37] In other words, the time trajectories spend waiting to surmount a barrier is distributed, enabling more efficient simulations.[38] While the ensemble dynamics method would soon be supplanted by MSM methods for sets of distributed trajectories,[8] the innovation of Folding@home not only catalyzed methods development for distributed simulations, but also fostered community interest in MD as a way to gain insight into how protein dynamics relate to various diseases.

Contemporaneously, IBM introduced their BlueGene project,[39−41] a massively parallel computer architecture designed to run biomolecular dynamics simulations. The BlueGene project represented a shift from analyzing MD based on one-off observations to an extensive sampling approach, wherein many observations are analyzed in aggregate in order to draw statistically meaningful conclusions. This shift set the stage for researchers at IBM to adapt the theory of Markov processes for analyzing simulations generated using BlueGene.[9] A decade later, Buch et al.[42] introduced GPUGRID, a volunteer-based distributed computing platform leveraging GPUs for distributed molecular dynamics simulations.[43,44]

5. CRITICAL ADVANCES IN THEORY, VALIDATION, AND ERROR ANALYSIS

The oft-cited papers credited with the advent of MSM analyses are the 1999 report by Schütte et al.,[7] a pair of 2004 companion papers by Swope and colleagues,[9,45] and the work of Singhal et al.[8] In their 1999 report, Schütte et al.[7] presented a formal foundation of MSM theory using the transfer operator. Notably, they suggest that conformations of a dynamical system should be defined not geometrically, but instead according to metastability, by approximating almost invariant sets using the eigendecomposition of the transition matrix.

Swope et al.[9] present a validation method for assessing the Markovian assumption: if the dynamics are indeed Markovian at the chosen lag time τ, then the timescales of the dynamical processes identified by the decomposition of the transition matrix will be constant for all longer lag times short enough to resolve the process. This technique, often referred to as an implied timescale analysis, is frequently used for MSM validation, and is a special case of the Chapman-Kolmogorov test used for MSM validation.[46] In the companion paper, Swope et al.[45] find that different mechanisms for β-hairpin dynamics were identified when different state definitions were used. This highlighted the difficult process of choosing MSM states, suggesting that the proper variables to describe a system are nonobvious, and that an improper choice can produce to misleading results.

In the first to explicitly identify the MSM method, Singhal et al.[8] presented MSMs as a "probabilistic roadmap", advocating for the method's efficiency because it incorporates all simulation data, in contrast to a traditional transition path sampling analysis. A follow up study in 2005 focused on error in MSM analyses, identifying two main sources.[47] First, errors can result from poor state decomposition, such as grouping two conformations that transition more slowly to each other than to conformations in other states. However, even if the state decomposition is perfect, there is still a second kind of error that arises from finite sampling of transitions between states. To address the latter problem, Singhal and Pande[47] proposed an adaptive sampling method based on resampling transitions that contribute the most uncertainty. In 2007, Hinrichs (née Singhal) and Pande[48] extended this analysis to uncertainties in the eigenvalues and eigenvectors of the transition matrix.

Several other notable papers included the work of Noé et al.,[49] which provided an extensive discussion of how to approach MSM state decomposition for proteins, noting that internal degrees of freedom, such as dihedral angles, produce better results than "global descriptors" such as native hydrogen bonding patterns. In 2008, Noé[50] introduced a method to approximate the complete distributions of observables from MD simulations by sampling all MSM transition matrices statistically consistent with the observed data, which was extended by Chodera and Noé[51] to include the sampling of experimental observables. A 2008 study by Buchete and Hummer[52] also reported insight into MSM construction and state decomposition: first, that states should be chosen according to transition paths instead of structure; and second, that models with longer lag times are more accurate. The authors also suggested not using a full partition of configuration space but instead avoiding the states in the transition region, which ultimately developed into the core set MSM framework.[53]

6. BIGGER SYSTEMS AND BETTER SOFTWARE

The theoretical advances of the 2000s showcased the advantages of using MSMs to analyze MD simulations. In 2009, these methods were made available to the scientific community through a set of Python scripts that were published as the first version of the MSMBuilder software.[54] In a subsequent publication, Bowman et al.[55] recommended determining MSM microstates using the pairwise root-mean-square deviation (RMSD) to assess the similarity of all pairs of structures, and using the computationally fast *k*-centers clustering algorithm to group structures into microstates. Since the *k*-centers algorithm will choose outlier conformations as cluster centers, the authors recommended subsampling the data in order to reduce the number of outlier data points.[55] The authors anticipated models would have tens of thousands of states, and thus would be impossible to interpret in spite of their quantitative accuracy, so a coarse-grained model could

DOI: 10.1021/jacs.7b12191
J. Am. Chem. Soc. 2018, 140, 2386−2396

be built with kinetic clustering methods[56,57] to aid with model interpretability.[58]

The protocol enumerated by Bowman et al.[55] (and later extended to include the hybrid *k*-medoids method[59]) has been used to guide MSM building in many studies, including recent work.[60,61] This procedure facilitated several key investigations of protein folding using MSMs. In 2010, Voelz et al.[62] presented the first model of ab initio protein folding on a millisecond timescale using Folding@home[35] simulations of the 39-residue protein NTL9. In 2011, Bowman et al.[63] analyzed an 80-residue fragment of λ-repressor, capturing folding on a 10 ms timescale, following up previous research showing that native states have hub-like behavior.[64,65] This idea was later expanded by Dickson and Brooks[66]

With the availability of MSMBuilder and other software packages,[67−69] MSMs were increasingly used to study biomolecular dynamics. Additional notable advances included the incorporation of Bayesian statistics into various aspects of MSM construction, such as approximating the transition matrix,[48,50,51,70−73] creating the MSM from parallel tempering simulations,[74−78] comparing models,[79−81] core set MSM sampling[52,53] and parameter estimation,[82] as well as incorporating transition path theory (TPT) into MSM methods,[83−87] such as to determine ensembles of protein folding pathways[88] or in combination with clustering to identify folding "tubes" containing groups of similar paths.[89−91]

The early 2010s also represented significant progress toward the theory of MSMs. In 2010, Sarich et al.[92] rigorously showed that the approximation error due to state space discretization decreases as the state partitioning becomes finer and as the lag time of the model increases.[73,92] In 2012, Djurdjevac et al.[93] derived upper bounds for the error in timescales between the MSM and the raw trajectory data and showed that this error also decreases as lag time increases. In the same year, Sarich and Schütte[94] demonstrated that a true eigenvalue of the continuous transfer operator approximated by the MSM transition matrix can be obtained with an appropriate choice of subspace.

7. VARIATIONAL APPROACH TO CONFORMATIONAL DYNAMICS

A 2011 report by Prinz et al.[46] represented a paradigm shift in the motivation of the MSM state decomposition by framing it as an approximation to the eigenfunctions of the continuous transfer operator (recall Section 2). Two years later, Nóe and Nüske[95] published a report that initiated a crucial change in the construction of MSMs. This paper derived a variational principle for the eigenfunctions of a MSM, analogous to the variational principle for choosing wave functions in quantum mechanics.[95,96] Thus, the highest eigenvalue produced by any of the state decompositions is closest to the true value, and its corresponding eigenfunction (in practice, eigenvector) is closest to the true eigenfunction for that process.

This procedure can be performed iteratively to approximate every eigenvalue in order to produce the set of eigenfunctions, assuming that every higher eigenvalue is known exactly.[95,97] Importantly, nothing about this variational optimization requires the use of states at all; this analysis can be performed with any set of input functions, and a MSM is generated in the special case that these are indicator functions that identify disjoint, discrete states.[95,97] This method, called the variational approach to conformational dynamics or VAC, transformed MSM analysis: expertise and experience were no longer needed to perform the state decomposition; instead, this process could be automated based on an objective criterion, i.e., the magnitude of the eigenvalue being approximated.

Subsequent work used Gaussian functions,[97] force-field dependent functions based on individual amino acids,[98] and sparse tensor products of one-dimensional functions[99] instead of indicator functions as initial functions to estimate the eigenfunctions. Alternatively, features can be extracted from a MD data set, such as dihedral angles or pairwise contact distances of a protein. To make a MSM, these features can be used as collective variables along which to determine the states. However, creating linear combinations of these features chosen such that their decorrelation time is maximized are themselves a set of initial functions to which the VAC can be applied: this process is called time structure-based (or time-lagged) independent component analysis, abbreviated as tICA or TICA.[100−104]

The tICA algorithm was reported in 1994 by Molgedey and Schuster[100] as a way to solve the blind source separation problem[101] and was first used to identify slow modes in proteins by Naritomi and Fuchigami.[102] In 2013, Schwantes and Pande[103] and Pérez-Hernández et al.[104] independently introduced the tICA algorithm as an intermediate step in MSM construction; in other words, the time structure-based independent components (tICs) were used to determine the states of the system. Instead of using structural similarity as a proxy for kinetic similarity,[55,59] Schwantes and Pande[103] and Pérez-Hernández et al.[104] sought to encode kinetic similarity in the states explicitly by choosing the states along the tICs, which serve as reaction coordinates for the dynamical system.[105] Previous studies had also used lower dimensionality spaces to encode protein dynamics[106−108] or a kinetically motivated state decomposition step to build MSMs.[109−113]

Using tICA to build MSMs resulted in models with hundreds of states, instead of the tens of thousands of states often created from clustering with structural metrics,[55,62,63] which contained more conformations and were thus more statistically robust. In their 2014 perspective article, Schwantes et al.[114] wrote that such models should be interpretable from the outset, instead of relying on kinetic clustering for interpretability.[114] tICA was widely adapted by the MSM community and was augmented by incorporating kinetic and commute distances.[115,116] Kernel,[117−119] hierarchical,[120] and sparse[105] versions of the tICA algorithm were also developed. Very recent applications have used tICA as a collective variable for metadynamics.[121−123]

The derivation of tICA as a special case of the VAC by Pérez-Hernández et al.[104] was the first to use the tICA model to approximate the eigenfunctions directly, and showed that tICA is the linearly optimal estimator. However, since using a linear combination of input features harshly constrains the input functions, only the dominant eigenfunction and eigenvalue can be variationally approximated with this method, since the approximation of subsequent eigenfunctions requires the previously determined eigenfunctions to be the true ones.[104]

For this reason, kernel tICA[117,126,127] was developed to alleviate the linear constraint on the tICA solutions, such that all the MSM eigenfunctions could be directly approximated. In 2017, Harrigan and Pande[118] introduced landmark kernel tICA, a kernel approximation of tICA using the RMSD to assess structural similarity and a set of conformations as landmarks. This formulation is equivalent to building a MSM with soft states, which had been suggested or accomplished by several other groups over the past decade or so.[53,93,128−133]

Harrigan[119] built upon landmark kernel tICA utilizing a differentiable form of RMSD. These variationally optimized, learnable

DOI: 10.1021/jacs.7b12191
J. Am. Chem. Soc. 2018, 140, 2386−2396

soft MSMs produced superior models to optimized nonlearnable landmark kernel tICA models, as well as to variationally optimized traditional "crisp" MSMs built using linear tICA, and required much fewer states.[119] A similar approach presented by Mardt et al.[125] uses a neural network to transform superposed Cartesian coordinates directly into soft states. The authors also report that just a few states are sufficient to encode system dynamics. However, for realistic systems, the Cartesian coordinates needed to be featurized before input into the network.[125] Additional advances in using deep neural networks to create analogs to tICA or MSMs include the use of nonlinear autoencoders.[134−137]

In their 2015 article, McGibbon and Pande[124] highlighted the problem of overfitting to finite data, showing that the approximated eigenvalues can exceed the variational bound when the model is overfit. To address this problem, the authors advocated for cross-validated model in which the MSM is trained on a subset of the data and then evaluated on the part of the data that was originally left out. To develop a scoring function for cross-validation, McGibbon and Pande[124] leveraged the results of Noé and Nüske[95] by presenting the VAC as the simultaneous optimization of eigenfunctions corresponding to a set of the highest eigenvalues.

This MSM score was termed the GMRQ, which stands for generalized matrix Rayleigh quotient, the form of the approximator (also referred to as the Rayleigh trace).[124] The GMRQ on the validation set will be poor if the model was overfit on the training set but better if the model identifies the underlying dynamics common to both sets. In 2016, Noé and Clementi[115] demonstrated that kinetic variance in a data set can be explained by summing the squared tICA eigenvalues. Since the variational principle derived in Noé and Nüske[95] holds for any strictly nonincreasing weights applied to the scored eigenvalues,[96] the kinetic variance can also be used to score models, or to determine how many tICs are needed to explain a given amount of kinetic variance in the data.

Combined with MD analysis software packages MDTraj[138] and HTMD,[44] the third version of MSMBuilder,[139] PyEMMA,[140] and the Osprey software for performing MSM hyperparameter searches,[141] model selection using the cross-validated GMRQ across many hyperparameter options became a powerful method to construct MSMs capturing slow dynamical processes in an automated fashion. This type of optimization assumes that the model is created at an appropriate lag time, which cannot be chosen by hyperparameter search and must be determined based on the dynamics of interest. Alternatively, continuous-time MSMs can be used, which do not have a lag time.[142] A summary of major improvements to the MSM pipeline is presented in Figure 2.

8. BEYOND MSMS: FRONTIERS IN METHODS DEVELOPMENT

Now that we have established the current state of the field, We now highlight a few themes that represent current areas of key methods advances that represent guided MD sampling techniques, methods to coarse-grain models to increase their interpretability, multi-ensemble approaches for aggregating data sets at different thermodynamic states, and the extension of MSM-inspired methods to nonreversible processes. Figure 3 summarizes the many extensions to MSMs and how they relate to a standard MSM.

8.1. Adaptive Sampling Guides Molecular Dynamics.

Several methodological advances have been published in the past few years. MSMs have been used to motivate adaptive sampling strategies for more than a decade;[47,153−156] however, the past few years have seen substantial advances in this area.[121−123,157−160] In 2014, Voelz et al.[158] introduced a surprisal metric for quantifying transitions between metastable states for two similar MSMs, and that the information−theoretic similarity between the two models converges upon sampling the transitions with the greatest surprisal. In 2015, Zimmerman and Bowman[159] presented the fluctuation amplification of specific traits (FAST) adaptive sampling scheme, which chooses states to sample by balancing structural metrics of interest and exploration of undersampled states. A key advantage of this algorithm is that the results maintain proper thermodynamics and

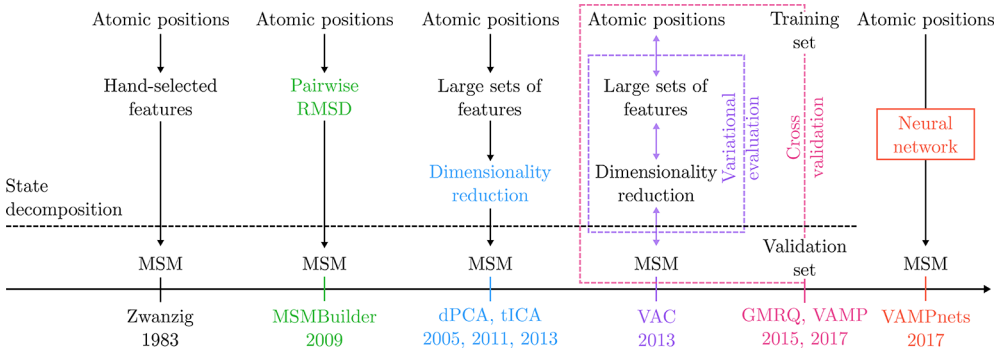

Figure 2. Chronological summary of major improvements to the MSM pipeline from 1983 to 2015. In 1983,[1] proposed a theoretical outline of the method, suggesting the transformation of atomic positions to a kinetic model via features based on physical intuition and past experience. In 2009, the first version of the MSMBuilder software[54] and a subsequent article by Bowman et al.[55] advocated for the transformation of atomic positions into states by using pairwise RMSD similarity as a proxy for kinetic similarity. Several groups independently developed dimensionality reduction techniques, including dPCA in 2005[107] and tICA in 2011[102] and 2013,[103,104] which were used to transform larger sets of atomic features into meaningful coordinates. Such dimensionality reductions weight features according to importance, so many features can be input into the model. In 2013, the derivation of the VAC[95] enabled an objective protocol for evaluating different state choices. Finally, in 2015, the incorporation of cross-validation into the variational evaluation of MSM states as proposed by McGibbon and Pande[124] signified a practical application of the VAC to account for the finite length of data sets. In this scheme, the set of atomic positions is divided into training and validation sets. The MSM is created from the training set and evaluated using the GMRQ on the validation set. Several folds of cross validation can be performed by using multiple training and validation splits from the same data. Finally, very recent reports have utilized neural network architectures to obtain MSMs directly from atomic positions without an explicit state decomposition.[125] For realistic systems, this method currently still requires selecting features.[125]

DOI: 10.1021/jacs.7b12191
J. Am. Chem. Soc. 2018, 140, 2386−2396

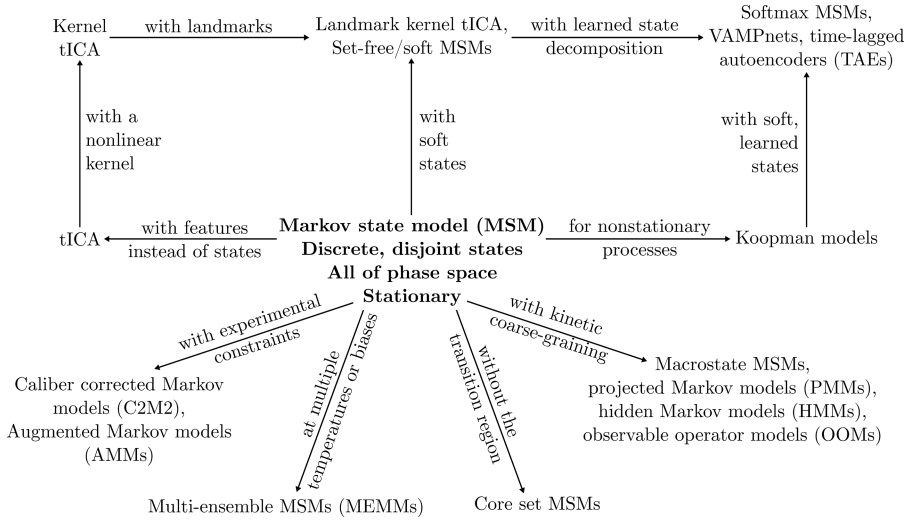

Figure 3. Many extensions to MSMs are characterized by changes to the standard protocol, which uses discrete, disjoint states spanning all of phase space (in practice, configuration space), and assumes a stationary process. When trajectories are described using features, such as side chain dihedral angles or contact pairs, instead of state labels, the model becomes a tICA model.[103,104] Using a nonlinear kernel for tICA yields kernel tICA.[117] Using landmarks to effectively regularize kernel tICA results in landmark kernel tICA,[118] which is shown to be equivalent to MSMs with soft states, also called set-free MSMs.[133] When the soft state decomposition is learned using a neural network, the resulting model is a softmax MSM.[119] Generalizing MSM methods to nonstationary processes involves the use of Koopman models.[143,144] Learnable soft states for Koopman models have been presented using VAMPnets[125] or time-lagged autoencoders (TAEs).[135] When a MSM is coarse-grained according to its transition matrix or eigendecomposotion (i.e., kinetic coarse-graining via PCCA,[56] PCCA+,[57] BACE,[80] HNEG,[145] or MVCA[146]) the model becomes a macrostate MSM. Projected Markov models (PMMs) are also macrostate MSMs, which can be approximated using hidden Markov models (HMMs)[147,148] or observable operator models (OOMs).[149] Another formulation of MSMs, core set MSMs,[52] involves partitioning a subset of full phase space by avoiding the transition region. Multi-ensemble MSMs, or MEMMs, involve combining the data from multiple MSMs from different thermodynamic states or biases using transition-based reweighting analysis (TRAM) or related methods.[150] Finally, experimental constraints can be incorporated into MSM estimation using caliber corrected Markov models (C_2M_2)[151] or augmented Markov models (AMMs).[152]

kinetics.[159] Recent publications have used tICA[121−123] and evolutionary coupling information[160] to guide sampling.

8.2. Coarse-Graining Provides Model Interpretability. Methods for the kinetic lumping of a microstate model into a coarse-grained macrostate model have been developed for at least half a century,[161] including for master equation models.[162] The first coarse-graining developed specifically for MSMs was presented in 2000 by Deu hard et al.,[56] who introduced the Perron cluster−cluster analysis (PCCA) method for automatically assembling macrostates from the original microstates by leveraging the idea that the eigenvalues of the transition matrix are expected to be well-separated. PCCA was improved by a series of investigators over several years, ultimately resulting in robust PCCA, often referred to as PCCA+, which modifies the treatment of states at boundaries between macrostates.[57,163] In 2013, the hierarchical Nyström extension graph method (HNEG)[145] and BACE[80] were shown to outperform PCCA and PCCA+.[164]

In 2013, Noé et al.[147] and others[148,165] advocated for the use of hidden Markov models (HMMs) as approximations to projected Markov models (PMMs) to coarse-grain MSMs into a few macrostates. This analysis has been used in recent studies by Olsson and Noé[166] and Plattner et al.,[167] as well as on experimental data sets.[168] In 2015, Wu et al.[149] presented observable operator models (OOMs), a less restrictive method for the approximation of PMMs. Although OOMs, like HMMs, are not Markovian, when the system is metatable, OOMs approximate the timescales exactly.[149] In 2017, Nüske et al.[169] showed that OOM theory can be used to estimate a MSM transition matrix such that it is not biased by its initial distribution, addressing a long-standing issue in MSM analysis.[46,54,88,153]

New coarse-graining approaches have been recently presented, such as a renormalization group theory approach[170,171] and a method using key reaction coordinates and hierarchical construction of MSMs to identify both metastable and transition regions.[172] In 2017, Husic et al.[146] developed a minimum variance clustering analysis (MVCA) that demonstrated advantages over several earlier methods. Notably, MVCA can be used not only to coarse-grain individual models, but to group an aggregated set of MSMs based on their dynamical similarity. The authors demonstrated this method for protein folding simulations in multiple force fields, and it could also be used for a set of systems with different mutations.[146]

8.3. Multi-ensemble MSMs Unify Simulations. There have been many recent developments in multi-ensemble MSMs spanning different temperatures or biases. A simple way to analyze such a set of MSMs is to choose states at the target temperature and directly count the instances of these states at the other temperatures.[173] Reweighting these instances with respect to the temperature of interest produces the weighted histogram analysis method (WHAM)[174,175] or, without binning states, the multi-state Bennett acceptance ratio (MBAR);[176] however, these approaches are only valid if simulations at each state equilibrate quickly, sacrificing the ability of MSMs to capture long timescale dynamics. In 2014, Mey et al.[173] presented a transition-based reweighting analysis (TRAM). Partly inspired by MSM methods, TRAM leverages conditional probabilities and thus does not require the analyzed data to be in global equilibrium.[173] Several related algorithms were developed around the same time,[173,177−183] including dTRAM,[179] a generalization of WHAM and reversible MSMs representing first statistically optimal TRAM method.

DOI: 10.1021/jacs.7b12191
J. Am. Chem. Soc. 2018, 140, 2386−2396

In 2016, Wu et al.[150] published a general TRAM estimator that combines reweighting simulations at different temperatures, conditional transition probabilities from MSMs, and a maximum likelihood approach that maintains thermodynamic and kinetic information. Wu et al.[150] showed that TRAM is a formal generalization of MBAR and dTRAM. In 2017, Paul et al.[184] extended TRAM to enable the combination of unbiased, very short replica-exchange simulations. This method, called TRAMMBAR, was the first to estimate protein-peptide association kinetics at timescales greater than one second with small uncertainties.[184] These methods are expected to augment the analysis of multi-body systems, which is discussed in a perspective by Zhu et al.[185]

8.4. Modeling Nonstationary Processes Extends the Domain of Relevant Systems. Several extensions to MSMs for nonequilibrium processes have been recently proposed,[135,143,144,186−190] In 2017, Wu et al.[143] introduced variational Koopman models, which formulate the VAC in a more general way in order to describe nonreversible processes. In the special case of a reversible process, the Koopman models are equivalent to tICA or MSMs when linear transformations of features or discrete state decompositions are used, respectively. Wu and Noé[144] also introduced variational approach for Markov processes (VAMP), which applies to both reversible and nonreversible processes, the latter of which is ubiquitous in biophysics and fluid mechanics.[144]

The recent application of deep learning to learn overlapping states of a system[125] are based on the Koopman model framework and can thus be applied to more complicated, nonstationary systems such as protein folding landscapes with kinetic sinks,[188] tightly bound complexes that do not unbind,[187] or systems subject to a periodic time-dependent potential.[186,189,190] The continued development of a more general set of methods that can accommodate nonstationary processes in addition to stationary processes will greatly expand the domain of systems for which MSM-inspired analysis can be applied. Such methods developments will be capable of lending insight into systems in which the reversibility assumption is not appropriate, such fluid mechanics applications to atmospheric and oceanic currents.

9. EXPERIMENTAL INSIGHT: FRONTIERS IN APPLICATIONS

The union of MSM methods with experimental methods and outcomes represents an exciting area for new insight into chemistry. Here, we discuss a few of the most recent advances involving MSMs and experimental data: the incorporation of experimental quantities into MSM construction in order to generate rich dynamical descriptions, the use of a unifying framework to describe relaxation processes and design experiments, and the ability of MSMs along with nontraditional docking to identify druggable sites and hit compounds.

9.1. Incorporating Experimental Quantities Enables Consistency between Models and Observables. Theoretical models are designed to provide a richer picture of a system than experiments can resolve. However, these models must be consistent with coarser experimental results in order for their finer details to be credible. For this reason, several groups have developed methods to tune MSMs so their results reflect experimental observables.[51,151,152,191] In 2016, Rudzinski et al.[191] introduced a method to incorporate experimental measurements into MSM construction using a tunable bias parameter. This bias parameter is designed to be flexible such that it does not prescribe the type of experimental measurement that is used to bias the model. In 2017, Dixit and Dill[151] presented a maximum caliber method to create a caliber corrected Markov model (C_2M_2) by updating the thermodynamic and kinetic information with constraints determined from experimental results in the form of position-dependent diffusion coefficients.

Alternatively, MSMs can be designed with the intention of aligning with coarse experimental data. In 2017, McKiernan et al.[192] codified the collective degrees of freedom of a fast-folding β-hairpin to correspond to the same structural elements probed in a temperature-jump experiment for the same system. By performing the same MSM analysis for simulations in three different force fields, force field-agnostic conclusions could be drawn about the system that aligned with experimental findings while also providing a richer description of the dynamics.[192] In the same year, Olsson et al.[152] introduced augmented Markov models (AMMs), which are designed to systematically treat simulation and experimental errors. For example, Olsson et al.[152] show that discrepancies in equilibrium stabilities between MSMs for ubiqitin in two different force fields can be reconciled by AMMs that have been augmented with NMR observables. This powerful approach opens the door for the union of computational and experimental methods to accurately model dynamical systems.

9.2. Assigning Simulated Processes to Experimental Observables Facilitates Experimental Design. In 2011, Noé et al.[193] introduced dynamical fingerprint theory, which enables the assignment of structural relaxation processes observed in simulation to experimentally observable processes.[194] In this report, the authors demonstrate its use for MD simulations fluorescent peptides and corresponding fluorescence correlation spectroscopy (FCS) experiments. Both Noé et al.[193] and, in a subsequent report, Keller et al.[194] highlight the ability to design FCS, FRET, or temperature-jump experiments using dynamical fingerprints by targeting specific relaxation processes from simulation. In 2013, dynamical fingerprint theory was used to reconstruct inelastic neutron scattering spectra based on MSMs from simulation data.[195] In a more recent study, Olsson and Noé[166] connect MSM properties to chemical exchange induced relaxation from NMR experiments, and show how separate processes contribute observed chemical exchange signals.

Other techniques with similar motivation have also been developed. In 2011, Zhuang et al.[196] used MSMs to predict experimental observables from two-dimensional and time-resolved infrared (IR) spectroscopy, as well as from temperature jump experiments. In 2017, a method reported by Mittal and Shukla[197] demonstrates the use of MSMs to predict an ideal set of double electron−electron resonance (DEER) spin-label positions to maximize information gained from experimental investigations of protein conformational change.

9.3. Dynamical Fluctuations Reveal Target Sites for Drug Design. MD simulation data sets contain dynamical information about a protein and thus generate many conformations that may contain druggable sites not present in published crystal structures. In 2015, Bowman et al.[198] demonstrated that analyzing simulations with MSMs can enable the identification of such sites by identifying MSM states with pockets that resemble known active sites. This method is powerful because, since it first identifies empty pockets during simulation, it does not require ligands to be simulated or proposed outright. In this report, the authors discovered several hidden allosteric sites in the TEM-1 β-lactamase protein.[198]

The following year, Hart et al.[60] analyzed the same system to investigate the specificity of these hidden conformations using docking scores weighted by equilibrium populations according

DOI: 10.1021/jacs.7b12191
J. Am. Chem. Soc. 2018, 140, 2386−2396

to the MSM, referred to as Boltzmann docking. The authors contrasted this method with traditional docking, which was unable to predict site specificity. Using their results, the authors designed variants for the purpose of controlling these sites' populations and experimentally demonstrated their success in stabilizing certain sites. A subsequent report by Hart et al.[199] identified three small molecules that inhibit or activate the TEM system using crystal structures, MSMs,[198] and Boltzmann docking.[60] These compounds were experimentally verified to bind to the MSM-predicted sites. This method represents a powerful approach to the investigation of allosteric modulation, especially for targets for which the available crystal structures do not contain druggable sites.

Finally, for the same system, a recent report by Zimmerman et al.[61] uses MSMs to understand a known mutation in the TEM system and subsequently predict new mutations, including a new stabilizing mutation. Notably, the prediction effects were experimentally tested using many complementary experimental methods, such as crystallography, NMR, and measurements performed in vivo. This work exemplifies the prospect of understanding atomistic dynamics using MSMs, using that understanding to predict experimental outcomes, and verifying the predicted experimental outcomes in real systems.

10. OUTLOOK

Nearly 35 years after Zwanzig's enumeration of a "remarkably simple" method for estimating transition probabilities among states of a dynamical system, the contributions of a great number of scientists have led to the current state of MSM analysis methods and availability of software and objective model selection protocols. In addition to the improved accessibility of these tools, the past several years have also yielded exciting advances both in methods development such as improvements in adaptive sampling, coarse-graining, multi-ensemble analysis, and the modeling of nonstationary processes. MSMs have also demonstrated their utility in enriching or informing experimental results through methods that incorporate measured observables, unify the description of relaxation processes across simulation and experiment, and facilitate drug discovery for hidden binding sites.

There are still important questions the MSM field must address. Modeling is often guided by a variational approach that maximizes the timescale of the slowest process within a data set. The underlying assumption of this approach is that the slowest dynamical processes are the most interesting or important processes. There are at least two cases in which this assumption is challenged: first, and perhaps more straightforward, protein simulations can feature dihedral flips or distance contractions described by sparse collective variables that maybe due to finite sampling or systematic force field error that occur more rarely than the event of interest. Second, a dynamical system may feature multiple interesting events, such as both conformational change and ligand binding, and the event of analytical interest may be the faster of the two.

In either case, the slowest event in the data set is not the event of interest. How to quantify and thus optimize for what is interesting, when it is not the same as what is slow, is an important question for the field to address. We believe that the answer is rooted in the determination of how to select and transform collective variables that encode the dynamics. Extensions of tICA and other applications of the VAC that can be tuned to prioritize accurate modeling of the process of interest—especially in an objective fashion—would address a notable limitation of current approaches.

An important direction for the MSM field is to develop methods that can perturb or combine systems without additional simulations. For example, the ability to adjust a MSM for one system such that it describes a mutated form of that system would enable efficient comparison of thermodynamics and kinetics without additional simulation.[137,200] Another key challenge is the construction of a MSM for a large, computationally intractable system by combining a set of MSMs from different parts of that system. Ultimately, these advances would enable MSMs to produce testable predictions for sets of related systems that may be unwieldy to simulate or experimentally probe, so that additional simulations and experiments can be directed by theoretical results.

MSM methods development has been characterized by rigorous methods combined with optimized, objective tools for analyzing dynamical systems and their interesting processes. Key challenges remain in isolating events of interest and in using MSMs to explore mutations and large systems that can be subdivided. The continued development and application of MSM-inspired methods will lead to the increased growth and widespread use of MSMs in chemistry, biophysics, materials science, fluid dynamics, and other related fields.

AUTHOR INFORMATION

Corresponding Author

*pande@stanford.edu

ORCID

Brooke E. Husic: 0000-0002-8020-3750

Notes

The authors declare the following competing financial interest(s): V.S.P. is a consultant and SAB member of Schrodinger, LLC and Globavir, sits on the Board of Directors of Apeel Inc., Freenome Inc., Omada Health, Patient Ping, and Rigetti Computing, and is a General Partner at Andreessen Horowitz.

ACKNOWLEDGMENTS

The authors are grateful to Muneeb Sultan, Fátima Pardo-Avila, Cathrine Bergh, Greg Bowman, and Frank Noé for manuscript feedback. We acknowledge the National Institutes of Health under No. NIH R01-GM62868 for funding.

REFERENCES

(1) Zwanzig, R. *J. Stat. Phys.* **1983**, *30*, 255−262.
(2) Van Kampen, N. G. *Stochastic Processes in Physics and Chemistry*; North-Holland Physics Publishing: New York, 1992.
(3) Schütte, C.; Sarich, M. *Courant Lecture Notes*; American Mathematical Society, 2013; Vol. *24*.
(4) Bowman, G. R.; Pande, V. S.; Noé, F. In *Advances in Experimental Medicine and Biology*; Bowman, G. R., Pande, V. S., Noé, F., Eds.; Springer: The Netherlands, 2014; Vol. *797*.
(5) Wang, W.; Cao, S.; Zhu, L.; Huang, X. *WIREs Comput. Mol. Sci.* **2018**, *8*, e1343.
(6) Klippenstein, S. J.; Pande, V. S.; Truhlar, D. G. *J. Am. Chem. Soc.* **2014**, *136*, 528−546.
(7) Schütte, C.; Fischer, A.; Huisinga, W.; Deuflhard, P. *J. Comput. Phys.* **1999**, *151*, 146−168.
(8) Singhal, N.; Snow, C. D.; Pande, V. S. *J. Chem. Phys.* **2004**, *121*, 415−425.
(9) Swope, W. C.; Pitera, J. W.; Suits, F. *J. Phys. Chem. B* **2004**, *108*, 6571−6581.
(10) Karpen, M. E.; Tobias, D. J.; Brooks, C. L. *Biochemistry* **1993**, *32*, 412−420.
(11) Luty, B. A.; McCammon, J. A. *Mol. Simul.* **1993**, *10*, 61−65.

DOI: 10.1021/jacs.7b12191
J. Am. Chem. Soc. 2018, 140, 2386−2396

(12) Luty, B. A.; El Amrani, S.; McCammon, J. A. *J. Am. Chem. Soc.* **1993**, *115*, 11874−11877.
(13) Grubmüller, H.; Tavan, P. *J. Chem. Phys.* **1994**, *101*, 5047−5057.
(14) Huber, G.; Kim, S. *Biophys. J.* **1996**, *70*, 97−110.
(15) Becker, O. M.; Karplus, M. *J. Chem. Phys.* **1997**, *106*, 1495−1517.
(16) Orland, H.; Thirumalai, D. *J. Phys. I* **1997**, *7*, 553−560.
(17) Cieplak, M.; Henkel, M.; Karbowski, J.; Banavar, J. R. *Phys. Rev. Lett.* **1998**, *80*, 3654−3657.
(18) Ye, Y.-J.; Ripoll, D.; Scheraga, H. *Comput. Theor. Polym. Sci.* **1999**, *9*, 359−370.
(19) Bovier, A.; Eckhoff, M.; Gayrard, V.; Klein, M. *Probab. Theory Relat. Fields* **2001**, *119*, 99−161.
(20) de Groot, B. L.; Daura, X.; Mark, A. E.; Grubmüller, H. *J. Mol. Biol.* **2001**, *309*, 299−313.
(21) Wales, D. J. *Mol. Phys.* **2002**, *100*, 3285−3305.
(22) Levy, Y.; Jortner, J.; Berry, R. S. *Phys. Chem. Chem. Phys.* **2002**, *4*, 5052−5058.
(23) Schonbrun, J.; Dill, K. A. *Proc. Natl. Acad. Sci. U. S. A.* **2003**, *100*, 12678−12682.
(24) Ozkan, S. B.; Dill, K. A.; Bahar, I. *Biopolymers* **2003**, *68*, 35−46.
(25) Zhang, W.; Chen, S.-J. *J. Chem. Phys.* **2003**, *118*, 3413−3420.
(26) Hummer, G.; Kevrekidis, I. G. *J. Chem. Phys.* **2003**, *118*, 10762−10773.
(27) Lenz, P.; Zagrovic, B.; Shapiro, J.; Pande, V. S. *J. Chem. Phys.* **2004**, *120*, 6769−6778.
(28) Krivov, S. V.; Karplus, M. *Proc. Natl. Acad. Sci. U. S. A.* **2004**, *101*, 14766−14770.
(29) Rao, F.; Caflisch, A. *J. Mol. Biol.* **2004**, *342*, 299−306.
(30) Chekmarev, D. S.; Ishida, T.; Levy, R. M. *J. Phys. Chem. B* **2004**, *108*, 19487−19495.
(31) Berezhkovskii, A.; Szabo, A. *J. Chem. Phys.* **2004**, *121*, 9186−9187.
(32) Noé, F.; Krachtus, D.; Smith, J. C.; Fischer, S. *J. Chem. Theory Comput.* **2006**, *2*, 840−857.
(33) Hänggi, P.; Talkner, P. *Phys. Rev. Lett.* **1983**, *51*, 2242−2245.
(34) Shalloway, D. *J. Chem. Phys.* **1996**, *105*, 9986−10007.
(35) Shirts, M.; Pande, V. S. *Science* **2000**, *290*, 1903−1904.
(36) Voter, A. F. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1998**, *57*, R13985−R13988.
(37) Shirts, M. R.; Pande, V. S. *Phys. Rev. Lett.* **2001**, *86*, 4983−4987.
(38) Pande, V. S.; Baker, I.; Chapman, J.; Elmer, S. P.; Khaliq, S.; Larson, S. M.; Rhee, Y. M.; Shirts, M. R.; Snow, C. D.; Sorin, E. J.; Zagrovic, B. *Biopolymers* **2003**, *68*, 91−109.
(39) Allen, F.; et al. *IBM Syst. J.* **2001**, *40*, 310−327.
(40) Adiga, N. R.; et al. *Proceedings of the 2002 ACM/IEEE Conference on Supercomputing*, Baltimore, MD, Nov 16−22, 2002; IEEE: Piscataway, NJ, 2002; p 60.
(41) Fitch, B.; Germain, R.; Mendell, M.; Pitera, J.; Pitman, M.; Rayshubskiy, A.; Sham, Y.; Suits, F.; Swope, W.; Ward, T.; Zhestkov, Y.; Zhou, R. *J. Parallel Distrib. Comput.* **2003**, *63*, 759−773.
(42) Buch, I.; Harvey, M. J.; Giorgino, T.; Anderson, D. P.; De Fabritiis, G. *J. Chem. Inf. Model.* **2010**, *50*, 397−403.
(43) Harvey, M. J.; Giupponi, G.; Fabritiis, G. D. *J. Chem. Theory Comput.* **2009**, *5*, 1632−1639.
(44) Doerr, S.; Harvey, M. J.; Noé, F.; De Fabritiis, G. *J. Chem. Theory Comput.* **2016**, *12*, 1845−1852.
(45) Swope, W. C.; Pitera, J. W.; Suits, F.; Pitman, M.; Eleftheriou, M.; Fitch, B. G.; Germain, R. S.; Rayshubski, A.; Ward, T. J. C.; Zhestkov, Y.; Zhou, R. *J. Phys. Chem. B* **2004**, *108*, 6582−6594.
(46) 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.
(47) Singhal, N.; Pande, V. S. *J. Chem. Phys.* **2005**, *123*, 204909.
(48) Hinrichs, N. S.; Pande, V. S. *J. Chem. Phys.* **2007**, *126*, 244101.
(49) Noé, F.; Horenko, I.; Schütte, C.; Smith, J. C. *J. Chem. Phys.* **2007**, *126*, 155102.
(50) Noé, F. *J. Chem. Phys.* **2008**, *128*, 244103.
(51) Chodera, J. D.; Noé, F. *J. Chem. Phys.* **2010**, *133*, 105102.
(52) Buchete, N.-V.; Hummer, G. *J. Phys. Chem. B* **2008**, *112*, 6057−6069.
(53) Schütte, C.; Noé, F.; Lu, J.; Sarich, M.; Vanden-Eijnden, E. *J. Chem. Phys.* **2011**, *134*, 204105.
(54) Bowman, G. R.; Huang, X.; Pande, V. S. *Methods* **2009**, *49*, 197−201.
(55) Bowman, G. R.; Beauchamp, K. A.; Boxer, G.; Pande, V. S. *J. Chem. Phys.* **2009**, *131*, 124101.
(56) Deuflhard, P.; Huisinga, W.; Fischer, A.; Schütte, C. *Linear Algebra Appl.* **2000**, *315*, 39−59.
(57) Deuflhard, P.; Weber, M. *Linear Algebra Appl.* **2005**, *398*, 161−184.
(58) Pande, V. S.; Beauchamp, K.; Bowman, G. R. *Methods* **2010**, *52*, 99−105.
(59) Beauchamp, K. A.; Bowman, G. R.; Lane, T. J.; Maibaum, L.; Haque, I. S.; Pande, V. S. *J. Chem. Theory Comput.* **2011**, *7*, 3412−3419.
(60) Hart, K. M.; Ho, C. M. W.; Dutta, S.; Gross, M. L.; Bowman, G. R. *Nat. Commun.* **2016**, *7*, 12965.
(61) Zimmerman, M. I.; Hart, K. M.; Sibbald, C. A.; Frederick, T. E.; Jimah, J. R.; Knoverek, C. R.; Tolia, N. H.; Bowman, G. R. *ACS Cent. Sci.* **2017**, *3*, 1311−1321.
(62) Voelz, V. A.; Bowman, G. R.; Beauchamp, K.; Pande, V. S. *J. Am. Chem. Soc.* **2010**, *132*, 1526−1528.
(63) Bowman, G. R.; Voelz, V. A.; Pande, V. S. *J. Am. Chem. Soc.* **2011**, *133*, 664−667.
(64) Bowman, G. R.; Pande, V. S. *Proc. Natl. Acad. Sci. U. S. A.* **2010**, *107*, 10890−10895.
(65) Pande, V. S. *Phys. Rev. Lett.* **2010**, *105*, 198101.
(66) (a) Dickson, A.; Brooks, C. L. *J. Chem. Theory Comput.* **2012**, *8*, 3044−3052. (b) Dickson, A.; Brooks, C. L. *J. Am. Chem. Soc.* **2013**, *135*, 4729−4734.
(67) Biarnés, X.; Pietrucci, F.; Marinelli, F.; Laio, A. *Comput. Phys. Commun.* **2012**, *183*, 203−211.
(68) Senne, M.; Trendelkamp-Schroer, B.; Mey, A. S.; Schütte, C.; Noé, F. *J. Chem. Theory Comput.* **2012**, *8*, 2223−2238.
(69) Cronkite-Ratcliff, B.; Pande, V. *Bioinformatics* **2013**, *29*, 950.
(70) Sriraman, S.; Kevrekidis, I. G.; Hummer, G. *J. Phys. Chem. B* **2005**, *109*, 6479−6484.
(71) Hummer, G. *New J. Phys.* **2005**, *7*, 34.
(72) Metzner, P.; Noé, F.; Schütte, C. *Phys. Rev. E* **2009**, *80*, 021106.
(73) Trendelkamp-Schroer, B.; Noé, F. *J. Chem. Phys.* **2013**, *138*, 164113.
(74) Buchete, N.-V.; Hummer, G. *Phys. Rev. E* **2008**, *77*, 030902.
(75) Minh, D. D. L.; Chodera, J. D. *J. Chem. Phys.* **2009**, *131*, 134110.
(76) Prinz, J.-H.; Chodera, J. D.; Pande, V. S.; Swope, W. C.; Smith, J. C.; Noé, F. *J. Chem. Phys.* **2011**, *134*, 244108.
(77) Chodera, J. D.; Swope, W. C.; Noé, F.; Prinz, J.-H.; Shirts, M. R.; Pande, V. S. *J. Chem. Phys.* **2011**, *134*, 244107.
(78) Donati, L.; Hartmann, C.; Keller, B. G. *J. Chem. Phys.* **2017**, *146*, 244112.
(79) Bacallado, S.; Chodera, J. D.; Pande, V. *J. Chem. Phys.* **2009**, *131*, 045106.
(80) Bowman, G. R. *J. Chem. Phys.* **2012**, *137*, 134111.
(81) Weber, J. K.; Pande, V. S. *Biophys. J.* **2012**, *102*, 859−867.
(82) Wu, H.; Noé, F. *J. Chem. Phys.* **2015**, *142*, 084104.
(83) Weinan, E.; Vanden-Eijnden, E. *J. Stat. Phys.* **2006**, *123*, 503.
(84) Metzner, P.; Schütte, C.; Vanden-Eijnden, E. *Multiscale Model. Simul.* **2009**, *7*, 1192−1219.
(85) Berezhkovskii, A.; Hummer, G.; Szabo, A. *J. Chem. Phys.* **2009**, *130*, 205102.
(86) Venturoli, M.; Vanden-Eijnden, E.; Ciccotti, G. *J. Math. Chem.* **2009**, *45*, 188−222.
(87) E, W.; Vanden-Eijnden, E. *Annu. Rev. Phys. Chem.* **2010**, *61*, 391−420.
(88) Noé, F.; Schütte, C.; Vanden-Eijnden, E.; Reich, L.; Weikl, T. R. *Proc. Natl. Acad. Sci. U. S. A.* **2009**, *106*, 19011−19016.
(89) Zheng, W.; Gallicchio, E.; Deng, N.; Andrec, M.; Levy, R. M. *J. Phys. Chem. B* **2011**, *115*, 1512−1523.

DOI: 10.1021/jacs.7b12191
J. Am. Chem. Soc. 2018, 140, 2386−2396

(90) Deng, N.-j.; Zheng, W.; Gallicchio, E.; Levy, R. M. *J. Am. Chem. Soc.* **2011**, *133*, 9387−9394.
(91) Meng, L.; Sheong, F. K.; Zeng, X.; Zhu, L.; Huang, X. *J. Chem. Phys.* **2017**, *147*, 044112.
(92) Sarich, M.; Noé, F.; Schütte, C. *Multiscale Model. Simul.* **2010**, *8*, 1154−1177.
(93) Djurdjevac, N.; Sarich, M.; Schütte, C. *Multiscale Model. Simul.* **2012**, *10*, 61−81.
(94) Sarich, M.; Schütte, C. *Comm. Math. Sci.* **2012**, *10*, 1001−1013.
(95) Noé, F.; Nüske, F. *Multiscale Model. Simul.* **2013**, *11*, 635−655.
(96) Gross, E. K. U.; Oliveira, L. N.; Kohn, W. *Phys. Rev. A: At., Mol., Opt. Phys.* **1988**, *37*, 2805−2808.
(97) Nüske, F.; Keller, B. G.; Pérez-Hernández, G.; Mey, A. S. J. S.; Noé, F. *J. Chem. Theory Comput.* **2014**, *10*, 1739−1752.
(98) Vitalini, F.; Noé, F.; Keller, B. G. *J. Chem. Theory Comput.* **2015**, *11*, 3992−4004.
(99) Nüske, F.; Schneider, R.; Vitalini, F.; Noé, F. *J. Chem. Phys.* **2016**, *144*, 054105.
(100) Molgedey, L.; Schuster, H. G. *Phys. Rev. Lett.* **1994**, *72*, 3634−3637.
(101) Ziehe, A.; Müller, K.-R. In *ICANN 98: Proceedings of the 8th International Conference on Artificial Neural Networks*, Skövde, Sweden, Sept 2−4, 1998; Niklasson, L., Bodén, M., Ziemke, T., Eds.; Springer: London, 1998; pp 675−680.
(102) Naritomi, Y.; Fuchigami, S. *J. Chem. Phys.* **2011**, *134*, 065101.
(103) Schwantes, C. R.; Pande, V. S. *J. Chem. Theory Comput.* **2013**, *9*, 2000−2009.
(104) Pérez-Hernández, G.; Paul, F.; Giorgino, T.; De Fabritiis, G.; Noé, F. *J. Chem. Phys.* **2013**, *139*, 015102.
(105) McGibbon, R. T.; Husic, B. E.; Pande, V. S. *J. Chem. Phys.* **2017**, *146*, 044109.
(106) Schütte, C.; Huisinga, W. *Handb. Numer. Anal.* **2003**, *10*, 699−744 (Special Volume, Computational Chemistry).
(107) Mu, Y.; Nguyen, P. H.; Stock, G. *Proteins: Struct., Funct., Genet.* **2005**, *58*, 45−52.
(108) Altis, A.; Nguyen, P. H.; Hegger, R.; Stock, G. *J. Chem. Phys.* **2007**, *126*, 244111.
(109) Cossio, P.; Laio, A.; Pietrucci, F. *Phys. Chem. Chem. Phys.* **2011**, *13*, 10421−10425.
(110) Kellogg, E. H.; Lange, O. F.; Baker, D. *J. Phys. Chem. B* **2012**, *116*, 11405−11413.
(111) Zhou, T.; Caflisch, A. *J. Chem. Theory Comput.* **2012**, *8*, 2930−2937.
(112) McGibbon, R. T.; Pande, V. S. *J. Chem. Theory Comput.* **2013**, *9*, 2900−2906.
(113) Sheong, F. K.; Silva, D.-A.; Meng, L.; Zhao, Y.; Huang, X. *J. Chem. Theory Comput.* **2015**, *11*, 17−27.
(114) Schwantes, C. R.; McGibbon, R. T.; Pande, V. S. *J. Chem. Phys.* **2014**, *141*, 090901.
(115) Noé, F.; Clementi, C. *J. Chem. Theory Comput.* **2015**, *11*, 5002−5011.
(116) Noé, F.; Banisch, R.; Clementi, C. *J. Chem. Theory Comput.* **2016**, *12*, 5620−5630.
(117) Schwantes, C. R.; Pande, V. S. *J. Chem. Theory Comput.* **2015**, *11*, 600−608.
(118) Harrigan, M. P.; Pande, V. S. Landmark Kernel tICA For Conformational Dynamics. *bioRxiv.org e-Print archive*, 2017; DOI: 10.1101/123752 (accessed June 28, 2017).
(119) Harrigan, M. P. Towards Robust Dynamical Models of Biomolecules. Ph.D. thesis, Stanford University, Stanford, CA, 2017.
(120) Pérez-Hérnandez, G.; Noé, F. *J. Chem. Theory Comput.* **2016**, *12*, 6118−6129.
(121) Sultan, M. M.; Pande, V. S. *J. Chem. Theory Comput.* **2017**, *13*, 2440−2447.
(122) Sultan, M. M.; Pande, V. S. *J. Phys. Chem. B* **2017**, DOI: 10.1021/acs.jpcb.7b06896.
(123) Moffett, A. S.; Shukla, D. On the transferability of time-lagged independent components between similar molecular dynamics systems. arXiv.org e-Print archive, arXiv:1710.00443, 2017; http://arxiv.org/abs/1710.00443 (accessed Oct 3, 2017).
(124) McGibbon, R. T.; Pande, V. S. *J. Chem. Phys.* **2015**, *142*, 124105.
(125) Mardt, A.; Pasquali, L.; Wu, H.; Noé, F. *Nat. Commun.* **2018**, *9*, 5.
(126) Harmeling, S.; Ziehe, A.; Kawanabe, M.; Müller, K.-R. *Neural Comput.* **2003**, *15*, 1089−1124.
(127) Williams, M. O.; Rowley, C. W.; Kevrekidis, I. G. *J. Comput. Dynam.* **2015**, *2*, 247−265.
(128) Weber, M. Meshless Methods in Conformation Dynamics. Ph.D. thesis, Freie Universiät, Berlin, 2006.
(129) Weber, M.; Kube, S. *AIP Conf. Proc.* **2008**, *1048*, 593−596.
(130) Bujotzek, A.; Weber, M. *J. Bioinf. Comput. Biol.* **2009**, *07*, 811−831.
(131) Fackeldey, K.; Röblitz, S.; Scharkoi, O.; Weber, M. In *Particle Methods II, Fundamentals and Applications*, Barcelona, Spain, Oct 26−28, 2011; Onate, E., Owen, D. R. J., Eds.; International Center for Numerical Methods in Engineering: Barcelona, Spain, 2011; pp 899−909.
(132) Fackeldey, K.; Bujotzek, A.; Weber, M. In *Meshfree Methods for Partial Differential Equations VI*; Griebel, V. I., Schweitzer, M., Eds.; Springer: Berlin/Heidelberg, 2013; pp 141−154.
(133) Weber, M.; Fackeldey, K.; Schütte, C. *J. Chem. Phys.* **2017**, *146*, 124133.
(134) Doerr, S.; Ariz, I.; Harvey, M. J.; De Fabritiis, G. Dimensionality reduction methods for molecular simulations. arXiv.org e-Print archive, arXiv:1710.10629, 2017; http://arxiv.org/abs/1710.10629 (accessed Nov 1, 2017).
(135) Wehmeyer, C.; Noé, F. Time-lagged autoencoders: Deep learning of slow collective variables for molecular kinetics. arXiv.org e-Print archive, arXiv:1710.11239, 2017; http://arxiv.org/abs/1710.11239 (accessed Nov 1, 2017).
(136) Hernández, C. X.; Wayment-Steele, H. K.; Sultan, M. M.; Husic, B. E.; Pande, V. S. Variational Encoding of Complex Dynamics. arXiv.org e-Print archive, arXiv preprint arXiv:1711.08576, 2017; http://arxiv.org/abs/1710.08576 (accessed Dec 5, 2017).
(137) Sultan, M. M.; Wayment-Steele, H. K.; Pande, V. S. Transferable neural networks for enhanced sampling of protein dynamics. arXiv.org e-Print archive, arXiv:1801.00636, 2018; http://arxiv.org/abs/1801.00636 (accessed Jan 3, 2018).
(138) McGibbon, R. T.; Beauchamp, K. A.; Harrigan, M. P.; Klein, C.; Swails, J. M.; Hernández, C. X.; Schwantes, C. R.; Wang, L.-P.; Lane, T. J.; Pande, V. S. *Biophys. J.* **2015**, *109*, 1528−1532.
(139) Harrigan, M. P.; Sultan, M. M.; Hernández, C. X.; Husic, B. E.; Eastman, P.; Schwantes, C. R.; Beauchamp, K. A.; McGibbon, R. T.; Pande, V. S. *Biophys. J.* **2017**, *112*, 10−15.
(140) Scherer, M. K.; Trendelkamp-Schroer, B.; Paul, F.; Pérez-Hernández, G.; Hoffmann, M.; Plattner, N.; Wehmeyer, C.; Prinz, J.-H.; Noé, F. *J. Chem. Theory Comput.* **2015**, *11*, 5525−5542.
(141) McGibbon, R. T.; Hernández, C. X.; Harrigan, M. P.; Kearnes, S.; Sultan, M. M.; Jastrzebski, S.; Husic, B. E.; Pande, V. S. *J. Open Source Softw.* **2016**, *1*, 34.
(142) McGibbon, R. T.; Pande, V. S. *J. Chem. Phys.* **2015**, *143*, 034109.
(143) Wu, H.; Nüske, F.; Paul, F.; Klus, S.; Koltai, P.; Noé, F. *J. Chem. Phys.* **2017**, *146*, 154104.
(144) Wu, H.; Noé, F. Variational approach for learning Markov processes from time series data. arXiv.org e-Print archive, arXiv:1707.04659, 2017; http://arxiv.org/abs/1707.04659 (accessed July 18, 2017).
(145) Yao, Y.; Cui, R. Z.; Bowman, G. R.; Silva, D.-A.; Sun, J.; Huang, X. *J. Chem. Phys.* **2013**, *138*, 174106.
(146) Husic, B. E.; McKiernan, K. A.; Wayment-Steele, H. K.; Sultan, M. M.; Pande, V. S. *J. Chem. Theory Comput.* **2017**, DOI: 10.1021/acs.jctc.7b01004.
(147) Noé, F.; Wu, H.; Prinz, J.-H.; Plattner, N. *J. Chem. Phys.* **2013**, *139*, 184114.

DOI: 10.1021/jacs.7b12191
J. Am. Chem. Soc. 2018, 140, 2386−2396

(148) McGibbon, R. T.; Ramsundar, B.; Sultan, M. M.; Kiss, G.; Pande, V. S. *Proceedings of the 31st International Conference on Machine Learning*; Beijing, China, June 21−26, 2014; Xing, E. P., Jebara, T., Eds.; JMLR.org, 2014; pp 1197−1205.
(149) Wu, H.; Prinz, J.-H.; Noé, F. *J. Chem. Phys.* **2015**, *143*, 144101.
(150) Wu, H.; Paul, F.; Wehmeyer, C.; Noé, F. *Proc. Natl. Acad. Sci. U. S. A.* **2016**, *113*, E3221−E3230.
(151) Dixit, P. D.; Dill, K. *J. Chem. Theory Comput.* 2018, DOI: 10.1021/acs.jctc.7b01126.
(152) Olsson, S.; Wu, H.; Paul, F.; Clementi, C.; Noé, F. *Proc. Natl. Acad. Sci. U. S. A.* **2017**, *114*, 8265−8270.
(153) Huang, X.; Bowman, G. R.; Bacallado, S.; Pande, V. S. *Proc. Natl. Acad. Sci. U. S. A.* **2009**, *106*, 19765−19769.
(154) Bowman, G. R.; Ensign, D. L.; Pande, V. S. *J. Chem. Theory Comput.* **2010**, *6*, 787−794.
(155) Nerukh, D.; Jensen, C. H.; Glen, R. C. *J. Chem. Phys.* **2010**, *132*, 084104.
(156) Weber, J. K.; Pande, V. S. *J. Chem. Theory Comput.* **2011**, *7*, 3405−3411.
(157) Doerr, S.; De Fabritiis, G. *J. Chem. Theory Comput.* **2014**, *10*, 2064−2069.
(158) Voelz, V. A.; Elman, B.; Razavi, A. M.; Zhou, G. *J. Chem. Theory Comput.* **2014**, *10*, 5716−5728.
(159) Zimmerman, M. I.; Bowman, G. R. *J. Chem. Theory Comput.* **2015**, *11*, 5747−5757.
(160) Shamsi, Z.; Moffett, A. S.; Shukla, D. *Sci. Rep.* **2017**, *7*, 12700.
(161) (a) Wei, J.; Kuo, J. C. W. *Ind. Eng. Chem. Fundam.* **1969**, *8*, 114−123. (b) Kuo, J. C. W.; Wei, J. *Ind. Eng. Chem. Fundam.* **1969**, *8*, 124−133.
(162) Haug, K.; Truhlar, D. G.; Blais, N. C. *J. Chem. Phys.* **1987**, *86*, 2697−2716.
(163) Kube, S.; Weber, M. *J. Chem. Phys.* **2007**, *126*, 024103.
(164) Bowman, G. R.; Meng, L.; Huang, X. *J. Chem. Phys.* **2013**, *139*, 121905.
(165) Shukla, S.; Shamsi, Z.; Moffett, A. S.; Selvam, B.; Shukla, D. In *Hidden Markov Models: Methods and Protocols*; Westhead, D. R., Vijayabaskar, M. S., Eds.; Springer: New York, 2017; pp 29−41.
(166) Olsson, S.; Noé, F. *J. Am. Chem. Soc.* **2017**, *139*, 200−210.
(167) Plattner, N.; Doerr, S.; De Fabritiis, G.; Noé, F. *Nat. Chem.* **2017**, *9*, 1005−1011.
(168) Keller, B. G.; Kobitski, A.; Jäschke, A.; Nienhaus, G. U.; Noé, F. *J. Am. Chem. Soc.* **2014**, *136*, 4534−4543.
(169) Nüske, F.; Wu, H.; Prinz, J.-H.; Wehmeyer, C.; Clementi, C.; Noé, F. *J. Chem. Phys.* **2017**, *146*, 094104.
(170) Hummer, G.; Szabo, A. *J. Phys. Chem. B* **2015**, *119*, 9029−9037.
(171) Orioli, S.; Faccioli, P. *J. Chem. Phys.* **2016**, *145*, 124120.
(172) Martini, L.; Kells, A.; Hummer, G.; Buchete, N.-V.; Rosta, E. Identification and Analysis of Transition and Metastable Markov States. arXiv.org e-Print archive, arXiv:1605.04328, 2016 (accessed July 30, 2017).
(173) Mey, A. S. J. S.; Wu, H.; Noé, F. *Phys. Rev. X* **2014**, *4*, 041018.
(174) Ferrenberg, A. M.; Swendsen, R. H. *Phys. Rev. Lett.* **1989**, *63*, 1195−1198.
(175) Kumar, S.; Rosenberg, J. M.; Bouzida, D.; Swendsen, R. H.; Kollman, P. A. *J. Comput. Chem.* **1992**, *13*, 1011−1021.
(176) Shirts, M. R.; Chodera, J. D. *J. Chem. Phys.* **2008**, *129*, 124105.
(177) Sakuraba, S.; Kitao, A. *J. Comput. Chem.* **2009**, *30*, 1850−1858.
(178) Wu, H.; Noé, F. *Multiscale Model. Simul.* **2014**, *12*, 25−54.
(179) Wu, H.; Mey, A. S. J. S.; Rosta, E.; Noé, F. *J. Chem. Phys.* **2014**, *141*, 214106.
(180) Rosta, E.; Hummer, G. *J. Chem. Theory Comput.* **2015**, *11*, 276−285.
(181) Trendelkamp-Schroer, B.; Noé, F. *Phys. Rev. X* **2016**, *6*, 011009.
(182) Wan, H.; Zhou, G.; Voelz, V. A. *J. Chem. Theory Comput.* **2016**, *12*, 5768−5776.
(183) Stelzl, L. S.; Kells, A.; Rosta, E.; Hummer, G. *J. Chem. Theory Comput.* **2017**, *13*, 6328−6342.
(184) Paul, F.; Wehmeyer, C.; Abualrous, E. T.; Wu, H.; Crabtree, M. D.; Schöneberg, J.; Clarke, J.; Freund, C.; Weikl, T. R.; Noé, F. *Nat. Commun.* **2017**, *8*, 1095.
(185) Zhu, L.; Sheong, F. K.; Zeng, X.; Huang, X. *Phys. Chem. Chem. Phys.* **2016**, *18*, 30228−30235.
(186) Wang, H.; Schütte, C. *J. Chem. Theory Comput.* **2015**, *11*, 1819−1831.
(187) Weber, J. K.; Shukla, D.; Pande, V. S. *Proc. Natl. Acad. Sci. U. S. A.* **2015**, *112*, 10377−10382.
(188) Sirur, A.; De Sancho, D.; Best, R. B. *J. Chem. Phys.* **2016**, *144*, 075101.
(189) Koltai, P.; Ciccotti, G.; Schütte, C. *J. Chem. Phys.* **2016**, *145*, 174103.
(190) Koltai, P.; Schütte, C. A multi scale perturbation expansion approach for Markov state modeling of non-stationary molecular dynamics. Technical Report ZIB 17-49; Zuse Institute Berlin, Germany, 2017.
(191) Rudzinski, J. F.; Kremer, K.; Bereau, T. *J. Chem. Phys.* **2016**, *144*, 051102.
(192) McKiernan, K. A.; Husic, B. E.; Pande, V. S. *J. Chem. Phys.* **2017**, *147*, 104107.
(193) Noé, F.; Doose, S.; Daidone, I.; Löllmann, M.; Sauer, M.; Chodera, J. D.; Smith, J. C. *Proc. Natl. Acad. Sci. U. S. A.* **2011**, *108*, 4822−4827.
(194) Keller, B. G.; Prinz, J.-H.; Noé, F. *Chem. Phys.* **2012**, *396*, 92−107.
(195) (a) Lindner, B.; Yi, Z.; Prinz, J.-H.; Smith, J. C.; Noé, F. *J. Chem. Phys.* **2013**, *139*, 175101. (b) Yi, Z.; Lindner, B.; Prinz, J.-H.; Noé, F.; Smith, J. C. *J. Chem. Phys.* **2013**, *139*, 175102.
(196) Zhuang, W.; Cui, R. Z.; Silva, D.-A.; Huang, X. *J. Phys. Chem. B* **2011**, *115*, 5415−5424.
(197) Mittal, S.; Shukla, D. *J. Phys. Chem. B* **2017**, *121*, 9761−9770.
(198) 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.
(199) Hart, K. M.; Moeder, K. E.; Ho, C. M. W.; Zimmerman, M. I.; Frederick, T. E.; Bowman, G. R. *PLoS One* **2017**, *12*, e0178678.
(200) Schütte, C.; Nielsen, A.; Weber, M. *Mol. Phys.* **2015**, *113*, 69−78.

DOI: 10.1021/jacs.7b12191
J. Am. Chem. Soc. 2018, 140, 2386−2396