

# Machine Learning for Molecular Dynamics on Long Timescales

Frank Noé<sup>1</sup>

*Freie Universität Berlin, Department of Mathematics and Computer Science, Arnimallee 6,  
14195 Berlin<sup>a)</sup>*

Molecular Dynamics (MD) simulation is widely used to analyze the properties of molecules and materials. Most practical applications, such as comparison with experimental measurements, designing drug molecules, or optimizing materials, rely on statistical quantities, which may be prohibitively expensive to compute from direct long-time MD simulations. Classical Machine Learning (ML) techniques have already had a profound impact on the field, especially for learning low-dimensional models of the long-time dynamics and for devising more efficient sampling schemes for computing long-time statistics. Novel ML methods have the potential to revolutionize long-timescale MD and to obtain interpretable models. ML concepts such as statistical estimator theory, end-to-end learning, representation learning and active learning are highly interesting for the MD researcher and will help to develop new solutions to hard MD problems. With the aim of better connecting the MD and ML research areas and spawning new research on this interface, we define the learning problems in long-timescale MD, present successful approaches and outline some of the unsolved ML problems in this application field.

## I. INTRODUCTION

Molecular dynamics (MD) simulation is a widely used method of computational physics and chemistry to compute properties of molecules and materials. Examples include to simulate how a drug molecule binds to and inhibits a protein, or how a battery material conducts ions. Despite its high computational cost, researchers use MD in order to get a principled understanding of how the composition and the microscopic structure of a molecular system translate into such macroscopic properties. In addition to scientific knowledge, this understanding can be used for designing molecular systems with better properties, such as drug molecules or enhanced materials.

MD has many practical problems, but at least four of them can be considered to be fundamental, in the sense that none of them is trivial for a practically relevant MD simulation, and there is extensive research on all of them. We refer to these four fundamental MD problems as SAME (Sampling, Analysis, Model, Experiment):

1. **Sampling:** To compute expectation values via MD simulations the simulation time needs to significantly exceed the slowest equilibration process in the molecular system. For most nontrivial molecules and materials, the presence of rare events and the sheer cost per MD time step make sufficient direct sampling unfeasible.
2. **Analysis:** If enough statistics can be collected, we face huge amounts of simulation data (e.g., millions of time steps, each having 100,000s of dimensions). How can we analyze such data and obtain comprehensive and comprehensible models of the most relevant states, structures and events sampled in the data?
3. **Model:** MD simulations employ an empirical model of the molecular system studied. As the simulation computes forces from an energy model, this model is often referred to a MD force field. MD energy models are build from molecular components fitted to quantum mechanical and experimental data. The accuracy of such a model is limited by the accuracy of the data used and the errors involved in transferring the training data usually obtained for small molecules to the often larger molecules simulated.
4. **Experiment:** Experiments and simulations cannot access the same observables. While in MD simulation, the positions and velocities of all particles are available at all times, experiments usually probe complex functions of the positions and velocities, such as emission or absorption

---

<sup>a)</sup>Electronic mail: frank.noe@fu-berlin.de

spectra of certain types of radiation. Computing these functions from first principles often requires the solution of a quantum-mechanical calculation with an accuracy that is unfeasible for a large molecular system. The last problem thus consists of finding good approximations to compute how an experiment would “see” a given MD state.

Machine Learning (ML) has the potential to tackle these problems, and has already had profound impact on alleviating them. Here I will focus on the analysis problem and its direct connections to the sampling problem specifically for the case of long-time MD where these problems are most difficult and interesting. I believe that the solution of these problems lies on the interface between Chemical Physics and ML, and will therefore describe these problems in a language that should be understandable to audiences from both fields.

Let me briefly link to MD problems and associated ML approaches not covered by this chapter. The present description focuses on low-dimensional models of long-time MD and these can directly be employed to attack the sampling problem. The direct effect of these models is that short MD simulations that are individually not sampling all metastable states can be integrated, and thus an effective sampling that is much longer than the individual trajectory length, and on the other of the total simulation time can be reached<sup>76</sup>. The sampling efficiency can be further improved by adaptively selecting the starting points of MD simulations based on the long-time MD model, and iterating this process<sup>19,20,34,74,75,116,127</sup>. This approach is called “adaptive sampling” in the MD community, which is an active learning approach in ML language. Using this approach, time-scales beyond seconds have been reached and protein-protein association and dissociation has recently been sampled for the first time with atomistic resolution<sup>74</sup>.

A well establish approach to speed up rare events in MD is to employ so-called enhanced sampling methods that change the thermodynamic conditions (temperature, adding bias potentials, etc.)<sup>24,28,29,46,105</sup>, and to subsequently reweight to the unbiased target ensemble<sup>4,5,22,25,55,98</sup>. Recently, ML methods have been used to adaptively learn optimal biasing functions in such approaches<sup>77,110</sup>. A conceptually different approach to sampling is the Boltzmann Generator<sup>65</sup>, a directed generative network to directly draw statistically independent samples from equilibrium distributions. While these approaches are usually limited to compute stationary properties, ML-based MD analysis models have recently been integrated with enhance sampling methods in order to also compute unbiased dynamical properties<sup>80,118,119,122</sup>. These methods can now also access all-atom protein dynamics beyond seconds timescales<sup>70</sup>.

ML methods that use MD trajectory data to obtain a low-dimensional models of the long-time dynamics are extensively discussed here. Not discussed are manifold learning methods that purely use the data distribution, such as kernel PCA<sup>89</sup>, isomap<sup>16,103</sup> or diffusion maps<sup>15,79</sup>. Likewise, there is extensive research on geometric clustering methods – both on the ML and the MD application side – which only plays a minor role in the present discussion.

Learning an accurate MD model – the so-called force-field problem – is one of the basic and most important problems of MD simulation. While this approach has traditionally been addressed by relatively *ad hoc* parametrization methods it is now becoming more and more a well-defined ML problem where universal function approximators (neural networks or kernel machines) are trained to reproduce quantum-mechanical potential energy surfaces with high accuracy<sup>6–8,82,90,91</sup>. See other chapters in this book for more details. A related approach to scale to the next-higher length-scale is the learning of coarse-grained MD models from all-atom MD data<sup>111,112,125</sup>. These approaches have demonstrated that they can reach high accuracy, but employing the kernel machine or neural network to run MD simulations is still orders of magnitude slower than simulating a highly optimized MD code with an explicitly coded model. Achieving high accuracy while approximately matching the computational performance of commonly used MD codes is an important future aim.

Much less ML work has been done on the interpretation and integration of experimental data. MD models are typically parametrized by combining the matching of energies and forces from quantum-mechanical simulations with the matching of thermodynamic quantities measured by experiments, such as solvation free energies of small molecules. As yet, there is no rigorous ML method which learns MD models following this approach. Several ML methods have been proposed to integrate simulation data on the level of a model learned from MD simulation data (e.g., a Markov state model), typically by using information-theoretic principles such as maximum entropy or maximum caliber<sup>18,35,67</sup>. Finally, there is an emerging field of ML methods that predict experimental quantities, such as spectra, from chemical or molecular structures, which is an essential task that needs to be solved to perform data integration between simulation and experiment. An important step-stone for improving our ability

to predict experimental properties are the availability of training datasets where chemical structures, geometric structures and experimental measurements under well-defined conditions are linked.

## II. LEARNING PROBLEMS FOR LONG-TIME MOLECULAR DYNAMICS

### A. What would we like to compute?

The most basic quantitative aim of MD is to compute equilibrium expectations. When  $\mathbf{x}$  is state of a molecular system, such coordinates and velocities of the atoms in a protein system in a periodic solvent box, the average value of an observable  $A$  is given by:

$$\mathbb{E}[A] = \int A(\mathbf{x}) \mu(\mathbf{x}) d\mathbf{x} \quad (1)$$

where  $\mu(\mathbf{x})$  is the equilibrium distribution, *i.e.*, the probability to find a molecule in state  $\mathbf{x}$  at equilibrium conditions. A common choice is the Boltzmann distribution in the canonical ensemble at temperature  $T$ :

$$\mu(\mathbf{x}) \propto e^{-\frac{U(\mathbf{x})}{k_B T}} \quad (2)$$

where  $U(\mathbf{x})$  is a potential energy and the input constant  $k_B T$  is the mean thermal energy per degree of freedom. The observable  $A$  can be chosen to compute, *e.g.*, the probability of a protein to be folded at a certain temperature, or the probability for a protein and a drug molecule to be bound at a certain drug concentration, which relates to how much the drug inhibits the protein's activity. Other equilibrium expectations, such as spectroscopic properties, do not directly translate to molecular function, but are useful to validate and calibrate simulation models.

Molecules are not static but change their state  $\mathbf{x}$  over time. Under equilibrium conditions, these dynamical changes are due to thermal fluctuations, leading to trajectories that are stochastic. Given configuration  $\mathbf{x}_t$  at time  $t$ , the probability of finding the molecule in configuration  $\mathbf{x}_{t+\tau}$  at a later time can be expressed by the transition density  $p_\tau$ :

$$\mathbf{x}_{t+\tau} \sim p_\tau(\mathbf{x}_{t+\tau} | \mathbf{x}_t). \quad (3)$$

Thus, a second class of relevant quantities is that of dynamical expectations:

$$\mathbb{E}[G; \tau] = \int \int \mu(\mathbf{x}_t) p_\tau(\mathbf{x}_{t+\tau} | \mathbf{x}_t) G(\mathbf{x}_t, \mathbf{x}_{t+\tau}) d\mathbf{x}_t d\mathbf{x}_{t+\tau} \quad (4)$$

As above, the observable  $G$  determines which dynamical property we are interested in. With an appropriate choice we can measure the average time a protein takes to fold or unfold, or dynamical spectroscopic expectations such as fluorescence correlations or dynamical scattering spectra.

### B. What is Molecular Dynamics?

MD simulation mimics the natural dynamics of molecules by time-propagating the state of a molecular system, such coordinates and velocities of the atoms in a protein system in a periodic solvent box. MD is a Markov process involving deterministic components such as the gradient of a model potential  $U(\mathbf{x})$  and stochastic components, *e.g.* from a thermostat. The specific choice of these components determine the transition density (3). Independent of these choices, a reasonable MD algorithm should be constructed such that it samples from  $\mu(\mathbf{x})$  in the long run:

$$\lim_{\tau \rightarrow \infty} p_\tau(\mathbf{x}_{t+\tau} | \mathbf{x}_t) = \mu(\mathbf{x}) \propto e^{-U(\mathbf{x})/k_B T}. \quad (5)$$

Thus, if a long enough MD trajectory can be generated, the expectation values (1) and (4) can be computed as direct averages. Unfortunately, this idea can only be implemented directly for very small and simple molecular systems. Most of the interesting molecular systems involve rare events, and as a result generating MD trajectories that are long enough to compute the expectation values (1) and (4) by direct averaging becomes unfeasible. For example, the currently fastest special-purpose supercomputer for MD, Anton II, can generate simulations on the order of 50  $\mu\text{s}$  per day for a protein system<sup>96</sup>. The time for two strongly binding proteins to spontaneously dissociate can take over an hour, corresponding to a simulation time of a century for single event<sup>74</sup>.

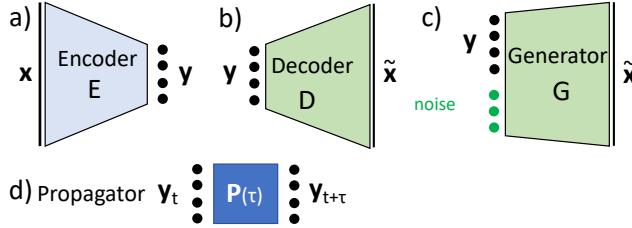


Figure 1. Overview of network structures for learning Markovian dynamical models

### C. Learning Problems for long-time MD

Repeated sampling from  $p_\tau(\mathbf{x}_{t+\tau} \mid \mathbf{x}_t)$  “simulates” the MD system in time steps of length  $\tau$  and will, due to (5), result in configurations sampled from  $\mu(\mathbf{x}_t)$ . Hence, knowing  $p_\tau(\mathbf{x}_{t+\tau} \mid \mathbf{x}_t)$  is sufficient to compute any stationary or dynamical expectation (1,4). The primary ML problem for long-time MD is thus to learn a model of the probability distribution  $p_\tau(\mathbf{x}_{t+\tau} \mid \mathbf{x}_t)$  from simulation data pairs  $(\mathbf{x}_t, \mathbf{x}_{t+\tau})$  which allows  $\mathbf{x}_{t+\tau} \sim p_\tau(\mathbf{x}_{t+\tau} \mid \mathbf{x}_t)$  to be efficiently sampled. However, this problem is almost never addressed directly, because it is unnecessarily difficult. Configurations  $\mathbf{x}$  live in a very high-dimensional space (typically  $10^3$  to  $10^6$  dimensions), the probability distributions  $p_\tau(\mathbf{x}_{t+\tau} \mid \mathbf{x}_t)$  and  $\mu(\mathbf{x})$  are multimodal and complex such that direct sampling is not tractable, and because of the exponential relationship between energies and probabilities (2), small mistakes in sampling  $\mathbf{x}$  will lead to completely unrealistic molecular structures.

Because of these difficulties, ML methods for long-time MD usually take the detour of finding a low-dimensional *representation*, often called latent space representation,  $\mathbf{y} = E(\mathbf{x})$ , using the encoder  $E$ , and learning the dynamics in that space

$$\begin{array}{ccc} \mathbf{x}_t & \xrightarrow{E} & \mathbf{y}_t \\ \text{MD } \downarrow & & \downarrow \mathbf{P} \\ \mathbf{x}_{t+\tau} & \xleftarrow{D/G} & \mathbf{y}_{t+\tau} \end{array}$$

A relatively recent but fundamental insight is that for many MD systems there exists a natural low-dimensional representation in which the stationary and dynamical properties can be represented exactly if we give up time resolution by choosing a large lag time  $\tau$ . Thus, for long-time MD the intractable problem to learn  $p_\tau(\mathbf{x}_{t+\tau} \mid \mathbf{x}_t)$  can be broken down into three learning problems (LPs) out of which two are much less difficult, and the third one does not need to be solved in order to compute stationary or dynamical expectations (1,4), and that will be treated in the remainder of the article:

1. **LP1: Learn propagator  $\mathbf{P}$  in representation  $\mathbf{y}$ .** The simplest problem is to learn a model to propagate the latent state  $\mathbf{y}_t$  in time for a given encoding  $E(\mathbf{x}_t)$ . This model is often linear using the propagator matrix  $\mathbf{P}$ , and hence shallow learning methods such as regression are used. In addition to obtaining an accurate model, it is desirable for  $\mathbf{P}$  to be compact and easily interpretable/readable for a human specialist.
2. **LP2: Learn encoding  $E$  to representation  $\mathbf{y}$ .** Learning the generally nonlinear encoding  $\mathbf{y} = E(\mathbf{x})$  is a harder problem. Both shallow methods (Regression in kernel and feature spaces, clustering and likelihood maximization) as well as deep methods (neural networks) are used. LP1 and LP2 can be coupled to an end-to-end learning problem for  $p_\tau(E(\mathbf{x}_{t+\tau}) \mid E(\mathbf{x}_t))$ . LP2 has only become a well-defined ML problem recently with the introduction of a variational approach that defines a meaningful loss function for LP2.
3. **LP3: Learn decoding  $D/G$  to configuration space.** The most difficult problem is to decode the latent representation  $\mathbf{y}$  back to configuration space. Because configuration space is much higher dimensional than latent space, this is an inverse problem. The most faithful solution is to learn a generator  $G$ , representing a conditional probability distribution,  $\mathbf{x} \sim G(\mathbf{y})$ . This problem contains the hardest parts of the full learning problem for  $p_\tau(\mathbf{x}_{t+\tau} \mid \mathbf{x}_t)$  and addressing it is still in its infancy.
  - a. These learning problems lead to different building blocks that can be implemented by neural networks or linear methods and can be combined towards different architectures (Fig. 1).

### III. LP1: LEARNING PROPAGATOR IN FEATURE SPACE

The simplest and most established learning problem is to learn a propagator,  $\mathbf{P}$ , for a given, fixed encoding  $E$ . Therefore we discuss this learning problem first before defining what a “good” encoding  $E$  is and how to find it. As will be discussed below, for most MD systems of interest, there exists an encoding  $E$  to a *spectral representation* in which the dynamics is linear and low-dimensional. Although this spectral representation can often not be found exactly, it can usually be well enough approximated such that a linear dynamic model

$$\mathbb{E}[\mathbf{y}_{t+\tau}] = \mathbf{P}^\top \mathbb{E}[\mathbf{y}_t] \quad (6)$$

is an excellent approximation as well.  $\mathbb{E}$  denotes an expectation value over time that accounts for stochasticity in the dynamics, and can be omitted for deterministic dynamical systems. For example, if  $\mathbf{y}_t$  indicates which state the system is in at time  $t$ ,  $\mathbb{E}[\mathbf{y}_t]$  corresponds to a probability distribution over states.

Finding a linear model  $\mathbf{P}$  is a shallow, unsupervised learning problem that in many cases has an algebraic expression for the optimum. Having a linear propagator also has great advantages for the analysis of the dynamical system. The analyses that can be done depend on the type of the representation and the mathematical properties of  $\mathbf{P}$ . If  $E$  performs a one-hot-encoding that indicates which “state” the system is in, then the pair  $(E, \mathbf{P})$  is called Markov state model (MSM<sup>12,13,61,76,92,100</sup>), and  $\mathbf{P}$  is the transition matrix of a Markov chain whose elements  $p_{ij}$  are nonnegative and can be interpreted as the conditional probabilities to be in a state  $j$  at time  $t + \tau$  given that the system was in a state  $i$  at time  $t$  (Sec. III B and III C). For MSMs, the whole arsenal of Markov chains analysis algorithms is available, e.g. for computing limiting distributions, first passage times or the statistics of transition pathways<sup>54,63</sup>. If the transition matrix additional has a real-valued spectrum, which is associated with dynamics at thermodynamic equilibrium conditions (Sec. III C), additional analyses are applicable, such as the computation of metastable (long-lived) sets of states by spectral clustering<sup>17,61,92</sup>.

A broader class of propagators arise from encodings  $E$  that are partitions of unity, i.e. where  $y_i(\mathbf{x}) > 0$  and  $\sum_i y_i(\mathbf{x}) = 1$  for all  $\mathbf{x}$ <sup>45,51</sup>. Such encodings correspond to a “soft clustering”, where every configuration  $\mathbf{x}$  can still be assigned to a state, but the assignment is no longer unique. The resulting propagators  $\mathbf{P}$  are typically no longer transition matrices whose elements can be guaranteed to be nonnegative, but they can still be used to propagate probability densities by means of Eq. (6), and if they have a unique eigenvalue of 1, the corresponding eigenvector  $\boldsymbol{\pi} = [\pi_i]$  still corresponds to the unique equilibrium distribution:

$$\boldsymbol{\pi} = \mathbf{P}^\top \boldsymbol{\pi}. \quad (7)$$

For arbitrary functions  $E$ , we can still use  $\mathbf{P}$  to propagate state vectors according to Eq. (6), although these state vectors do no longer have a probabilistic interpretation, but are simply coefficients that model the configuration in the representation’s basis. Owing to the Markovianity of the model, we can test how well the time-propagation of the model in time coincides with an estimation of the model at longer times, by means of the Chapman-Kolmogorov equation:

$$\mathbf{P}^n(\tau) \approx \mathbf{P}(n\tau) \quad (8)$$

In order to implement this equation, one has to decide which matrix norm should be used to compare the left and right hand side. A common choice is to compare the leading eigenvalues  $\lambda_i(\tau)$ . As these decay exponentially with time in a Markov process, it is common to transform them to relaxation rates or timescales by means of:

$$t_i(\tau) = -\frac{\tau}{\log |\lambda_i(\tau)|} \quad (9)$$

A consequence of the Chapman-Kolmogorow equality is that these relaxation timescales are independent of the lag time  $\tau$  at which  $\mathbf{P}$  is estimated<sup>100</sup>. For real-valued eigenvalues,  $t_i$  corresponds to an ordinary relaxation time of the corresponding dynamical process. If  $\mathbf{P}$  has complex-valued eigenvalues,  $t_i$  is the decay time of the envelope of an oscillating process whose oscillation frequency depends on the phase of  $\lambda_i$ .

### A. Loss Function and basis statistics

Given one or many MD simulation trajectories  $\{\mathbf{x}_t\}$  and apply  $E$  in order to map them to the representation  $\{\mathbf{y}_t\}$  which define the input to LP1. The basic learning problem is the parameter estimation problem which consists of obtaining the optimal estimator  $\hat{\mathbf{P}}$  as follows:

1. Define a loss function  $\mathcal{L}(\mathbf{P}; \{\mathbf{y}_t\})$
2. Obtain the optimal estimator as  $\hat{\mathbf{P}} = \arg \min_{\mathbf{P}} \mathcal{L}(\mathbf{P}; \{\mathbf{y}_t\})$

As most texts about molecular kinetics do not use the concept of a loss function, I would like to highlight the importance of a loss (or score) function from a ML point of view. The difference between fitting a training data set  $\{\mathbf{y}_t\}$  and ML is that ML aims at finding the estimator that performs best on an independent test data set. To this end we need to not only optimize the parameters (such as the matrix elements of  $\mathbf{P}$ ), but also hyper-parameters (such as the size of  $\mathbf{P}$ ), which requires the concept of a loss function. Another important learning problem is to estimate the uncertainties of the estimator  $\hat{\mathbf{P}}$ .

To express the loss function and the optimal estimator of linear propagators  $\mathbf{P}$ , we do not actually need the full trajectory  $\{\mathbf{y}_t\}$ , but only certain sufficient statistics that are usually more compact than  $\{\mathbf{y}_t\}$  and thus may require less storage space and lead to faster algorithms. The most prominent statistics are the empirical means and covariance matrices:

$$\boldsymbol{\mu}_0 = \frac{1}{T} \sum_{t=1}^{T-\tau} \mathbf{y}_t \quad (10)$$

$$\boldsymbol{\mu}_\tau = \frac{1}{T} \sum_{t=1}^{T-\tau} \mathbf{y}_{t+\tau} \quad (11)$$

$$\mathbf{C}_{00} = \frac{1}{T} \sum_{t=1}^{T-\tau} \mathbf{y}_t \mathbf{y}_t^\top \quad (12)$$

$$\mathbf{C}_{0\tau} = \frac{1}{T} \sum_{t=1}^{T-\tau} \mathbf{y}_t \mathbf{y}_{t+\tau}^\top \quad (13)$$

$$\mathbf{C}_{\tau\tau} = \frac{1}{T} \sum_{t=1}^{T-\tau} \mathbf{y}_{t+\tau} \mathbf{y}_{t+\tau}^\top \quad (14)$$

A common modification to (12,14) is the so-called shrinkage estimator that is used in ridge or Tikhonov regularization<sup>84</sup>. Since many algorithms involve the inversion of (12,14) which might be rank-deficient, these estimators are often modified by adding a second matrix which ensures full rank, e.g.:

$$\tilde{\mathbf{C}}_{00} = \mathbf{C}_{00} + \lambda \mathbf{I} \quad (15)$$

$$\tilde{\mathbf{C}}_{\tau\tau} = \mathbf{C}_{\tau\tau} + \lambda \mathbf{I} \quad (16)$$

where the small number  $\lambda$  is a regularization hyper-parameter.

### B. Maximum Likelihood and Markov State Models

The concepts of maximum likelihood estimators and Markov State Models (MSMs) are naturally obtained by defining the following encoding:

$$y_{t,i} = \begin{cases} 1 & \mathbf{x}_t \in S_i \\ 0 & \text{else.} \end{cases} \quad (17)$$

where  $S_1, \dots, S_n$  is a partition of configuration space into  $n$  discrete states, i.e. each point  $\mathbf{x}$  is assigned to exactly one state  $S_i$ , indicated by the position of the 1 in the encoding vector. In ML, (17) is called one-hot encoding. A consequence of (17) is that the covariance matrix (13) becomes:

$$c_{0\tau,ij} = N_{ij}$$

where  $N_{ij}$  counts the total number of transitions observed from  $i$  to  $j$ . The covariance matrix (12) is a diagonal matrix with diagonal elements

$$c_{00,ii} = N_i = \sum_j N_{ij}$$

where we use  $N_i$  to count the total number of transitions starting in state  $i$ . With this encoding, a natural definition for the propagator  $\mathbf{P}$  is a transition matrix whose elements indicate the transition probability from any state  $i$  to any state  $j$  in a time step  $\tau$ :

$$p_{ij} = \mathbb{P}[y_{t+\tau,j} = 1 \mid y_{t,i} = 1]$$

A natural optimality principle is then the maximum likelihood estimator (MLE): find the transition matrix  $\hat{\mathbf{P}}$  that has the highest probability to produce the observation  $\{\mathbf{y}_t\}$ . The likelihood is given by:

$$L \propto \prod_{i,j} p_{ij}^{N_{ij}}. \quad (18)$$

Where the last term collects equal transition events along the trajectory and discards the proportionality factor. Maximizing  $L$  is equivalent to minimizing  $-L$ . However, as common in likelihood formulations we instead use  $-\log L$  as a loss, which is minimal at the same  $\hat{\mathbf{P}}$  but avoids the product:

$$\mathcal{L}_{\text{ML}}(\mathbf{P}; \{\mathbf{y}_t\}) = -\log L = -\sum_{i,j} N_{ij} \log p_{ij} \quad (19)$$

The MLE  $\hat{\mathbf{P}}$  can be easily found by minimizing (19) with the constraint  $\sum_j p_{ij} = 1$  using the method of Lagrange multipliers. The result is intuitive: the maximum likelihood transition probability equals the corresponding fraction of transitions observed out of each state:

$$p_{ij} = \frac{N_{ij}}{N_i}$$

In matrix form we can express this estimator as

$$\mathbf{P} = \mathbf{C}_{00}^{-1} \mathbf{C}_{0\tau}, \quad (20)$$

an expression that we will find also for other optimization principles. As we have a likelihood (18), we can also define priors and construct a full Bayesian estimator that not only provides the maximum likelihood result (20), but also posterior means and variances for estimating uncertainties. Efficient samplers are known that allow us to sample transition matrices directly from the distribution (18), and these samples can be used to compute uncertainties on quantities derived from  $\mathbf{P}$ <sup>33,99</sup>.

An important property of a transition matrix is its stationary distribution  $\boldsymbol{\pi}$  (which we will assume to exist and be unique here) with

$$\pi_i = \int_{\mathbf{x} \in S_i} \mu(\mathbf{x}) d\mathbf{x}.$$

$\boldsymbol{\pi}$  that can be computed by solving the eigenvalue problem (7).

### C. MSMs with Detailed Balance

In thermodynamic equilibrium, i.e., when a molecular system is evolving purely as a result of thermal energy at a given thermodynamic condition and no external force is applied, the absolute probability of paths between any two end-points is symmetric. As a consequence of this, there exists no cycle in state space which contains net flux in either direction, and no net work can be extracted from the system, consistently with the second law of thermodynamics. We call this condition *detailed balance* and write it as:

$$\mu(\mathbf{x}) p_\tau(\mathbf{y} \mid \mathbf{x}) = \mu(\mathbf{y}) p_\tau(\mathbf{x} \mid \mathbf{y}) \quad \forall \mathbf{x}, \mathbf{y}, \tau > 0 \quad (21)$$

Integrating  $\mathbf{x}$  and  $\mathbf{y}$  over the sets  $S_i$  and  $S_j$  in this equation leads to detailed balance for MSMs:

$$\pi_i p_{ij} = \pi_j p_{ji}. \quad (22)$$

When the molecular system is simulated such that equations (21) hold, we also want to ensure that the estimator  $\hat{\mathbf{P}}$  fulfills the constraint (22). Enforcing (21) in the estimator reduces the number of free parameters and thus improves the statistics. More importantly, propagators that fulfill (21) or (22) have a real-valued spectrum for which additional analyses can be made (see beginning of Sec. III).

The trivial estimator (20) does not fulfill (22), unless  $N_{ij}$  is, by chance, a symmetric matrix. Maximum likelihood estimation with (22) as a constraint can be achieved by an iterative algorithm first developed in<sup>9</sup> and reformulated as in Algorithm 1 in<sup>108</sup>. Enforcing (22) is only meaningful if there is a unique stationary distribution, which, requires the transition matrix to define a fully connected graph. For this reason, graph algorithms are commonly used to find the largest connected set of states before estimating an MSM with detailed balance<sup>9,76,86</sup>.

1. Initialize:  $\pi_i^{(0)} = \frac{\sum_{j=1}^n c_{ij}}{\sum_{i,j=1}^n c_{ij}}$
2. Iterate until convergence:  $\pi_i^{(k+1)} = \sum_{j=1}^n \frac{c_{ij} + c_{ji}}{c_i/\pi_i^{(k)} + c_j/\pi_j^{(k)}}$
3.  $p_{ij} = \frac{(c_{ij} + c_{ji})\pi_j}{c_i\pi_i + c_j\pi_i}$

**Algorithm 1: Detailed balance**  $\pi_i p_{ij} = \pi_j p_{ji}$  **with unknown  $\pi$** <sup>9,108</sup>

When the equilibrium distribution  $\pi$  is known *a priori* or obtained from another estimator as in<sup>107,118,122</sup>, the maximum likelihood estimator can be obtained by the iterative Algorithm 2 developed in<sup>108</sup>:

1. Initialize Lagrange parameters:  $\lambda_i^{(0)} = \frac{1}{2} \sum_j (c_{ij} + c_{ji})$
2. Iterate until convergence:  $\lambda_i^{(k+1)} = \sum_{j, c_{ij} + c_{ji} > 0} \frac{(c_{ij} + c_{ji})\lambda_j^{(k)}\pi_j}{\lambda_j^{(k)}\pi_i + \lambda_i^{(k)}\pi_j}$
3.  $p_{ij} = \frac{(c_{ij} + c_{ji})\pi_j}{\lambda_i\pi_i + \lambda_j\pi_i}$

**Algorithm 2: Detailed balance**  $\pi_i p_{ij} = \pi_j p_{ji}$  **with known  $\pi$** <sup>108</sup>:

As for MSMs without detailed balance, methods have been developed to perform a full Bayesian analysis of MSMs with detailed balance. No method is known to sample independent transition matrices from the likelihood (18) subject to the detailed balance constraints (22), however efficient Markov Chain Monte Carlo methods have been developed and implemented to this end<sup>3,14,53,57,86,106,108</sup>.

#### D. Minimal Regression Error

We can understand equation (6) as a regression from  $\mathbf{y}_t$  onto  $\mathbf{y}_{t+\tau}$  where  $\mathbf{P}$  contains the unknown coefficients. The regression loss is then directly minimizing the error in Eq. (6):

$$\min \mathbb{E}_t \left[ \|\mathbf{y}_{t+\tau} - \mathbf{P}^\top \mathbf{y}_t\|^2 \right]$$

and for a given dataset  $\{\mathbf{y}_t\}$  we can define matrices  $\mathbf{Y}_0 = (\mathbf{y}_0, \dots, \mathbf{y}_{T-\tau})^\top$  and  $\mathbf{Y}_\tau = (\mathbf{y}_\tau, \dots, \mathbf{y}_T)^\top$  resulting in the loss function:

$$\mathcal{L}_{\text{LSQ}}(\mathbf{P}; \{\mathbf{y}_t\}) = \|\mathbf{Y}_0 - \mathbf{Y}_\tau \mathbf{P}\|_F^2 \quad (23)$$

where  $F$  indicates the Frobenius norm, i.e. the sum over all squares. The direct solution of the least squares regression problem in (23) is identical with the trivial MSM estimator (20). Thus, the estimator (20) is more general than for MSMs – it can be applied for to any representation  $\mathbf{y}_t$ . Dynamic Mode Decomposition (DMD)<sup>81,87,88,109</sup> and Extended Dynamic Mode Decomposition (EDMD)<sup>114</sup> are also using the minimal regression error, although they usually consider low-rank approximations of  $\mathbf{P}$ .

In general, the individual dimensions of the encoding  $E$  may not be orthogonal, and if not, the matrix  $\mathbf{C}_{00}$  is not diagonal, but contains off-diagonal elements quantifying the correlation between different dimensions. When there is too much correlation between them,  $\mathbf{C}_{00}$  may have some vanishing eigenvalues, i.e. not full rank, causing it not to be invertible or only invertible with large numerical errors. A standard approach in least squares regression is to then apply the Ridge regularization (Eq. 15). Using (15) in the estimator (20) is called Ridge regression.

### E. Variational Approach for Dynamics with Detailed Balance (VAC)

Instead of using an optimality principle to estimate  $\mathbf{P}$  directly, we will now derive a variational principle for the eigenvalues and eigenvectors of  $\mathbf{P}$ , from which we can then easily assemble  $\mathbf{P}$  itself. At first, this approach seems to be a complication compared to the likelihood or least squares approach, but this approach is key in making progress on LP2 because the variational principle for  $\mathbf{P}$  has a fundamental relation to the spectral properties of the transition dynamics in configuration space (3). It also turns out that the variational approach leads to a natural representation of configurations that we can optimize in end-to-end learning frameworks. We first define the balanced propagator:

$$\tilde{\mathbf{P}} = \mathbf{C}_{00}^{-\frac{1}{2}} \mathbf{C}_{0\tau} \mathbf{C}_{\tau\tau}^{-\frac{1}{2}}. \quad (24)$$

In this section, we will assume that detailed balance holds with the a unique stationary distribution, Eq (21). In the statistical limit this means that  $\mathbf{C}_{00} = \mathbf{C}_{\tau\tau}$  holds and  $\mathbf{C}_{0\tau}$  is a symmetric matrix. Using these constraints, we find the stationary balanced propagator:

$$\tilde{\mathbf{P}} = \mathbf{C}_{00}^{-\frac{1}{2}} \mathbf{C}_{01} \mathbf{C}_{00}^{-\frac{1}{2}} = \mathbf{C}_{00}^{\frac{1}{2}} \mathbf{P} \mathbf{C}_{00}^{-\frac{1}{2}} \quad (25)$$

Where we have used Eq. (6). Due to the symmetry of  $\mathbf{C}_{0\tau}$ ,  $\tilde{\mathbf{P}}$  is also symmetric and we have the symmetric eigenvalue decomposition (EVD):

$$\tilde{\mathbf{P}} = \tilde{\mathbf{U}} \Lambda \tilde{\mathbf{U}}^\top \quad (26)$$

with eigenvector matrix  $\tilde{\mathbf{U}} = [\tilde{\mathbf{u}}_1, \dots, \tilde{\mathbf{u}}_n]$  and eigenvalue matrix  $\Lambda = \text{diag}(\lambda_1, \dots, \lambda_n)$  ordered as  $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_n$ . This EVD is related to the EVD of  $\mathbf{P}$  via a basis transformation:

$$\mathbf{P} = \mathbf{C}_{00}^{-\frac{1}{2}} \tilde{\mathbf{U}} \Lambda \left( \tilde{\mathbf{U}} \mathbf{C}_{00}^{-\frac{1}{2}} \right)^\top = \mathbf{U} \Lambda \mathbf{U}^{-1} \quad (27)$$

such that  $\mathbf{U} = \mathbf{C}_{00}^{-\frac{1}{2}} \tilde{\mathbf{U}}$  are the eigenvectors of  $\mathbf{P}$ , their inverse is given by  $\mathbf{U}^{-1} = \mathbf{C}_{00}^{-\frac{1}{2}} \tilde{\mathbf{U}}^\top$ , and both propagators share the same eigenvalues. The above construction is simply a change of viewpoint: instead of optimizing the propagator  $\mathbf{P}$ , we might as well optimize its eigenvalues and eigenvectors, and then assemble  $\mathbf{P}$  via Eq. (27).

Now we seek an optimality principle for eigenvectors and eigenvalues. For symmetric eigenvalue problems such as (26), we have the following variational principle: The dominant  $k$  eigenfunctions  $\tilde{\mathbf{r}}_1, \dots, \tilde{\mathbf{r}}_k$  are the solution of the maximization problem:

$$\begin{aligned} \sum_{i=1}^k \lambda_i &= \max_{\tilde{\mathbf{f}}_1, \dots, \tilde{\mathbf{f}}_k} \sum_{i=1}^k \frac{\tilde{\mathbf{f}}_i^\top \tilde{\mathbf{P}} \tilde{\mathbf{f}}_i}{\left( \tilde{\mathbf{f}}_i^\top \tilde{\mathbf{f}}_i \right)^{\frac{1}{2}} \left( \tilde{\mathbf{f}}_i^\top \tilde{\mathbf{f}}_i \right)^{\frac{1}{2}}} = \max_{\mathbf{f}_1, \dots, \mathbf{f}_k} \sum_{i=1}^k \frac{\mathbf{f}_i^\top \mathbf{C}_{0\tau} \mathbf{f}_i}{\left( \mathbf{f}_i^\top \mathbf{C}_{00} \mathbf{f}_i \right)^{\frac{1}{2}} \left( \mathbf{f}_i^\top \mathbf{C}_{00} \mathbf{f}_i \right)^{\frac{1}{2}}} \\ &= \sum_{i=1}^k \frac{\mathbf{u}_i^\top \mathbf{C}_{0\tau} \mathbf{u}_i}{\left( \mathbf{u}_i^\top \mathbf{C}_{00} \mathbf{u}_i \right)^{\frac{1}{2}} \left( \mathbf{u}_i^\top \mathbf{C}_{00} \mathbf{u}_i \right)^{\frac{1}{2}}} = (\mathbf{U}^\top \mathbf{C}_{00} \mathbf{U})^{-\frac{1}{2}} \mathbf{U}^\top \mathbf{C}_{0\tau} \mathbf{U} (\mathbf{U}^\top \mathbf{C}_{00} \mathbf{U})^{-\frac{1}{2}} \end{aligned} \quad (28)$$

This means: we vary a set of vectors  $\mathbf{f}_i = \mathbf{C}_{00}^{-\frac{1}{2}} \tilde{\mathbf{f}}_i$ , and when the so-called Rayleigh quotients on the right hand side are maximized, we have found the eigenvectors. In this limit, the argument of the Rayleigh quotient equals the sum of eigenvalues. As the argument above can be made for every value of  $k$  starting from  $k = 1$ , we have found each single eigenvalue and eigenvector at the end of the procedure (assuming no degeneracy). This variational principle becomes especially useful for LP2, because using the variational approach of conformation dynamics (VAC<sup>62,66</sup>), it can also be shown that

the eigenvalues of  $\mathbf{P}$  are lower bounds to the true eigenvalues of the Markov dynamics in configurations  $\mathbf{x}$  (Sec. IV B).

Now we notice that this variational principle can also be understood as a direct correlation function of the data representation. We define the *spectral representation* as:

$$\mathbf{y}_t^s = (\mathbf{y}_t^\top \mathbf{u}_1, \dots, \mathbf{y}_t^\top \mathbf{u}_n) \quad (29)$$

inserting the estimators for  $\mathbf{C}_{00}$  and  $\mathbf{C}_{0\tau}$  (Eds. 12,13) into Eq. (28), we have:

$$\sum_{i=1}^k \lambda_i = \frac{\sum_{t=1}^{T-\tau} \mathbf{y}_t^s \mathbf{y}_{t+\tau}^{s\top}}{\sum_{t=1}^{T-\tau} \mathbf{y}_t^s \mathbf{y}_t^{s\top}} = (\mathbf{C}_{00}^s)^{-\frac{1}{2}} \mathbf{C}_{0\tau}^s (\mathbf{C}_{00}^s)^{-\frac{1}{2}}$$

where the superscript  $s$  denotes the covariance matrices computed in the spectral representation.

The same calculation as above can be performed with powers of the eigenvalues, *e.g.*,  $\sum_{i=1}^k \lambda_i^2$ . We therefore get a whole family of VAC-optimization principles, but two choices are especially interesting: we define the VAC-1 loss, that is equivalent to the generalized matrix Rayleigh quotient employed in<sup>52</sup>, as:

$$\mathcal{L}_{\text{VAC-1}}(\mathbf{U}; \{\mathbf{y}_t\}) = -\text{trace} \left[ (\mathbf{U}^\top \mathbf{C}_{00} \mathbf{U})^{-\frac{1}{2}} \mathbf{U}^\top \mathbf{C}_{0\tau} \mathbf{U} (\mathbf{U}^\top \mathbf{C}_{00} \mathbf{U})^{-\frac{1}{2}} \right] \quad (30)$$

$$\mathcal{L}_{\text{VAC-1}}(\{\mathbf{y}_t^s\}) = -\text{trace} \left[ (\mathbf{C}_{00}^s)^{-\frac{1}{2}} \mathbf{C}_{0\tau}^s (\mathbf{C}_{00}^s)^{-\frac{1}{2}} \right]. \quad (31)$$

The VAC-2 loss is the Frobenius norm, i.e. the sum of squared elements of the matrix:

$$\mathcal{L}_{\text{VAC-2}}(\mathbf{U}; \{\mathbf{y}_t\}) = -\left\| (\mathbf{U}^\top \mathbf{C}_{00} \mathbf{U})^{-\frac{1}{2}} \mathbf{U}^\top \mathbf{C}_{0\tau} \mathbf{U} (\mathbf{U}^\top \mathbf{C}_{00} \mathbf{U})^{-\frac{1}{2}} \right\|_F^2 \quad (32)$$

$$\mathcal{L}_{\text{VAC-2}}(\{\mathbf{y}_t^s\}) = -\left\| (\mathbf{C}_{00}^s)^{-\frac{1}{2}} \mathbf{C}_{0\tau}^s (\mathbf{C}_{00}^s)^{-\frac{1}{2}} \right\|_F^2. \quad (33)$$

This loss induces a natural spectral embedding where the variance along each dimension equals the squared eigenvalue and geometric distances in this space are related to kinetic distances<sup>58</sup>.

## F. General Variational Approach (VAMP)

The variational approach for Markov processes (VAMP)<sup>120</sup> generalizes the above VAC approach to dynamics that do not obey detailed balance and may not even have an equilibrium distribution. We use the balanced propagator (24) that is now no longer symmetric. Without symmetry we cannot use the variational principle for eigenvalues, but there is a similar variational principle for singular values. We therefore use the singular value decomposition (SVD) of the balanced propagator:

$$\tilde{\mathbf{P}} = \tilde{\mathbf{U}} \Sigma \tilde{\mathbf{V}}^\top \quad (34)$$

Again, this SVD is related to the SVD of  $\mathbf{P}$  via a basis transformation:

$$\mathbf{P} = \mathbf{C}_{00}^{-\frac{1}{2}} \tilde{\mathbf{U}} \Sigma \left( \mathbf{C}_{\tau\tau}^{-\frac{1}{2}} \tilde{\mathbf{V}} \right)^\top = \mathbf{U} \Sigma \mathbf{V}^\top \quad (35)$$

with  $\mathbf{U} = \mathbf{C}_{00}^{-\frac{1}{2}} \tilde{\mathbf{U}}$  and  $\mathbf{V} = \mathbf{C}_{\tau\tau}^{-\frac{1}{2}} \tilde{\mathbf{V}}$ . Using two sets of search vectors  $\mathbf{f}_i = \mathbf{C}_{00}^{-\frac{1}{2}} \tilde{\mathbf{f}}_i$  and  $\mathbf{g}_i = \mathbf{C}_{\tau\tau}^{-\frac{1}{2}} \tilde{\mathbf{g}}_i$ , we can follow the same line of derivation as above and obtain:

$$\begin{aligned} \sum_{i=1}^k \sigma_i &= \max_{\tilde{\mathbf{f}}_1, \dots, \tilde{\mathbf{f}}_k, \tilde{\mathbf{g}}_1, \dots, \tilde{\mathbf{g}}_k} \sum_{i=1}^k \frac{\tilde{\mathbf{f}}_i^\top \tilde{\mathbf{P}} \tilde{\mathbf{g}}_i}{(\tilde{\mathbf{f}}_i^\top \tilde{\mathbf{f}}_i)^{\frac{1}{2}} (\tilde{\mathbf{g}}_i^\top \tilde{\mathbf{g}}_i)^{\frac{1}{2}}} \\ &= (\mathbf{U}^\top \mathbf{C}_{00} \mathbf{U})^{-\frac{1}{2}} \mathbf{U}^\top \mathbf{C}_{0\tau} \mathbf{V} (\mathbf{V}^\top \mathbf{C}_{\tau\tau} \mathbf{V})^{-\frac{1}{2}} \end{aligned}$$

Now we define again a spectral representation. If we set  $\mathbf{C}_{00} = \mathbf{C}_{\tau\tau}$  (equilibrium case) as above, we can define a single spectral representation, otherwise we need two sets of spectral coordinates:

$$\mathbf{y}_t^{s,0} = (\mathbf{y}_t^\top \mathbf{u}_1, \dots, \mathbf{y}_t^\top \mathbf{u}_n) \quad (36)$$

$$\mathbf{y}_t^{s,\tau} = (\mathbf{y}_t^\top \mathbf{v}_1, \dots, \mathbf{y}_t^\top \mathbf{v}_n) \quad (37)$$

As in the above procedure, we can define a family of VAMP scores, where the VAMP-1 and VAMP-2 scores are of special interest:

$$\mathcal{L}_{\text{VAMP-1}}(\mathbf{U}, \mathbf{V}; \{\mathbf{y}_t\}) = -\text{trace} \left[ (\mathbf{U}^\top \mathbf{C}_{00} \mathbf{U})^{-\frac{1}{2}} \mathbf{U}^\top \mathbf{C}_{0\tau} \mathbf{V} (\mathbf{V}^\top \mathbf{C}_{\tau\tau} \mathbf{V})^{-\frac{1}{2}} \right] \quad (38)$$

$$\mathcal{L}_{\text{VAMP-1}}(\{\mathbf{y}_t^{s,0}, \mathbf{y}_t^{s,\tau}\}) = -\text{trace} \left[ (\mathbf{C}_{00}^s)^{-\frac{1}{2}} \mathbf{C}_{0\tau}^s (\mathbf{C}_{\tau\tau}^s)^{-\frac{1}{2}} \right]. \quad (39)$$

The VAMP-2 score is again related to an embedding where geometric distance corresponds to kinetic distance<sup>71</sup>:

$$\mathcal{L}_{\text{VAMP-2}}(\mathbf{U}, \mathbf{V}; \{\mathbf{y}_t\}) = -\left\| (\mathbf{U}^\top \mathbf{C}_{00} \mathbf{U})^{-\frac{1}{2}} \mathbf{U}^\top \mathbf{C}_{0\tau} \mathbf{V} (\mathbf{V}^\top \mathbf{C}_{\tau\tau} \mathbf{V})^{-\frac{1}{2}} \right\|_F^2 \quad (40)$$

$$\mathcal{L}_{\text{VAMP-2}}(\{\mathbf{y}_t^{s,0}, \mathbf{y}_t^{s,\tau}\}) = -\left\| (\mathbf{C}_{00}^s)^{-\frac{1}{2}} \mathbf{C}_{0\tau}^s (\mathbf{C}_{\tau\tau}^s)^{-\frac{1}{2}} \right\|_F^2. \quad (41)$$

#### IV. SPECTRAL REPRESENTATION AND VARIATIONAL APPROACH

Before turning to LP2, we will relate the spectral decompositions in the VAC and VAMP approaches described above to spectral representations of the transition density of the underlying Markov dynamics in  $\mathbf{x}_t$ . These two representations are connected by variational principles. Exploiting this principle leads to the result that a meaningful and feasible formulation of the long-time MD learning problem is to seek a spectral representation of the dynamics. This representation may be thought of as a set of collective variables (CVs) pertaining to the long-time MD, or slow CVs<sup>59</sup>.

##### A. Spectral theory

We can express the transition density (3) as the action of the Markov propagator in continuous-space, and by its spectral decomposition<sup>83,120</sup>:

$$p(\mathbf{x}_{t+\tau}) = \int p(\mathbf{x}_{t+\tau} | \mathbf{x}_t; \tau) p(\mathbf{x}_t) d\mathbf{x}_t \quad (42)$$

$$\approx \sum_{k=1}^n \sigma_k^* \langle p(\mathbf{x}_t) | \phi(\mathbf{x}_t) \rangle \psi(\mathbf{x}_{t+\tau}) \quad (43)$$

The spectral decomposition can be read as follows: The evolution of the probability density can be approximated as the superposition of basis functions  $\psi$ . A second set of functions,  $\phi$  is required in order to compute the amplitudes of these functions.

In general, Eq. (43) is a singular value decomposition with left and right singular functions  $\phi_k, \psi_k$  and true singular values  $\sigma_k^*$ <sup>120</sup>. The approximation then is a low-rank decomposition in which the small singular values are discarded. For the special case that dynamics are in equilibrium and satisfy detailed balance (21), Eq. (43) is an eigenvalue decomposition with the choices:

$$\begin{aligned} \sigma_k^* &= \lambda_k^*(\tau) = e^{-\tau \kappa_k} \in \mathbb{R} \\ \phi_k(\mathbf{x}) &= \psi_k(\mathbf{x}) \mu(\mathbf{x}). \end{aligned}$$

Hence Eq. (43) simplifies: we only need one set of functions, the eigenfunctions  $\psi_k$ . The true eigenvalues  $\lambda_k^*$  are real-valued and decay exponentially with the time step  $\tau$  (hence Eq. 9). The characteristic decay rates  $\kappa_k$  are directly linked to experimental observables probing the processes associated with the corresponding eigenfunctions<sup>12,60</sup>. The approximation in Eq. (43) is due to truncating all terms with decay rates faster than  $\kappa_n$ . This approximation improves exponentially with increasing  $\tau$ .

Spectral theory makes it clear why learning long-time MD via LP1-3 is significantly simpler than trying to model  $p(\mathbf{x}_{t+\tau} | \mathbf{x}_t; \tau)$  directly: For long time steps  $\tau$ ,  $p(\mathbf{x}_{t+\tau} | \mathbf{x}_t; \tau)$  becomes intrinsically low-dimensional, and it the problem is thus significantly simplified by learning to approximate the low-dimensional representation  $(\psi_1, \dots, \psi_n)$  for a given  $\tau$ .

### B. Variational principles

The spectral decomposition of the exact dynamics, Eq. (43), is the basis for the usefulness of the variational approaches described in Sec. III E and III F. The missing connection is filled by the following two variational principles. The VAC variational principle<sup>62</sup> is that for dynamics obeying detailed balance (21), the eigenvalues  $\lambda_k$  of a propagator matrix  $\mathbf{P}$  via any encoding  $\mathbf{y} = E(\mathbf{x})$  are, in the statistical limit, lower bounds of the true  $\lambda_k^*$ . The VAMP variational principle is more general, as it does not require detailed balance (21), and applies to the singular values:

$$\begin{aligned}\lambda_k &\leq \lambda_k^* \quad (\text{with DB}) \\ \sigma_k &\leq \sigma_k^* \quad (\text{no DB}).\end{aligned}$$

Equality is only achieved for  $E(\mathbf{x}) = \text{span}(\psi_1, \dots, \psi_n)$  when detailed balance holds, and for  $E(\mathbf{x}) = \text{span}(\psi_1, \dots, \psi_n, \phi_1, \dots, \phi_n)$  when detailed balance does not hold. Specifically, the eigenvectors or the singular vectors of the propagator then approximate the individual eigenfunctions or singular functions (assuming no degeneracy):

$$\begin{aligned}\lambda_k = \lambda_k^* &\longrightarrow \mathbf{u}_k^\top E(\mathbf{x}) = \psi(\mathbf{x}) \\ \sigma_k = \sigma_k^* &\longrightarrow \begin{cases} \mathbf{u}_k^\top E(\mathbf{x}) = \psi(\mathbf{x}) \\ \mathbf{v}_k^\top E(\mathbf{x}) = \phi(\mathbf{x}). \end{cases}\end{aligned}$$

As direct consequence of the variational principles above, the loss function associated with a given embedding  $E$  is, in the statistical limit, also an upper bound to the sum of true eigenvalues:

$$\begin{aligned}\mathcal{L}_{VAC-r} &\geq - \sum_{k=1}^n (\lambda_k^*)^r \\ \mathcal{L}_{VAMP-r} &\geq - \sum_{k=1}^n (\sigma_k^*)^r\end{aligned}$$

and for the minimum possible loss,  $E$  has identified the dominant eigenspace or singular space.

### C. Spectral representation learning

We have seen in Sec. III (LP1) that a propagator  $\mathbf{P}$  can be equivalently represented by its eigenspectrum or singular spectrum. We can thus define a spectral encoding that attempts to directly learn the encoding to the spectral representation:

$$\mathbf{y}_t^s = E^s(\mathbf{x}_t)$$

with the choices (29) or (36,37), depending on whether the dynamics obey detailed balance or not. In these representations, the dynamics are linear. After encoding to this representation, the eigenvalues or singular values can be directly estimated from:

$$\boldsymbol{\Lambda} = (\mathbf{R}^\top \mathbf{C}_{00}^s \mathbf{R})^{-1} \mathbf{R}^\top \mathbf{C}_{0\tau}^s \mathbf{R} \quad (44)$$

$$\boldsymbol{\Sigma} = (\mathbf{U}^\top \mathbf{C}_{00}^s \mathbf{U})^{-\frac{1}{2}} \mathbf{U}^\top \mathbf{C}_{0\tau}^s \mathbf{V} (\mathbf{V}^\top \mathbf{C}_{\tau\tau}^s \mathbf{V})^{-\frac{1}{2}} \quad (45)$$

Based on these results, we can formulate the learning of the spectral representation, or variants of it, as the key approach to solve LP2.

## V. LP2: LEARNING FEATURES AND REPRESENTATION

Above we have denoted the full MD system configuration  $\mathbf{x}$  and  $\mathbf{y}$  the latent-space representation in which linear propagators are used. We have seen that there is a special representation  $\mathbf{y}^s$ . In general there may be a whole pipeline of transformations, e.g.

$$\mathbf{x} \rightarrow \mathbf{x}^f \rightarrow \mathbf{y} \rightarrow \mathbf{y}^s$$

where the first step is a featurization from full configurations  $\mathbf{x}$  to features, e.g. the selection of solute coordinates or the transformation to internal coordinates such as distances or angles. On the latent space side  $\mathbf{y}$  we may have a handcrafted or a learned spectral representation. Instead of considering these transformations individually, we may construct a direct end-to-end learning framework that performs multiple transformation steps.

To simply notation, we commit to the following notation:  $\mathbf{x}$  coordinates are the input to the learning algorithm, whether these are full Cartesian coordinates of the MD system or already transformed by some featurization.  $\mathbf{y}$  are coordinates in the latent space representations that are the output of LP2,  $\mathbf{y} = E(\mathbf{x})$ . We only explicitly distinguish between different stages within configuration or latent space (e.g.  $\mathbf{y}$  vs  $\mathbf{y}^s$ ) when this distinction is explicitly needed.

### A. Suitable and unsuitable loss functions

We first ask: what is the correct formulation for LP2? More specifically: which of the loss functions introduced in LP1 above are compatible with LP2? Looking at the sequence of learning problems:

$$\mathbf{x} \xrightarrow{LP2} \mathbf{y} \xrightarrow{LP1} \mathbf{P}$$

It is tempting to concatenate them to an end-to-end learning problem and try to solve it by minimizing any of the three losses defined for learning of  $\mathbf{P}$  in Sec. III. However, if we make the encoding  $\mathbf{y} = E(\mathbf{x})$  sufficiently flexible, we find that only one of the loss functions remains as being suitable for end-to-end learning, while two others must be discarded as they have trivial and useless minima:

**Likelihood loss:** The theoretical minimum of the likelihood loss (19) is equal to 0 and is achieved if all  $p_{ij} \equiv 1$  for the transitions observed in the dataset. However, this maximum can be trivially achieved by learning a representation that assigns all microstates to a single state, e.g. the first state:

$$\arg \max_{E, P} \mathcal{L}_{\text{ML}}(\mathbf{P}; \{E(\mathbf{x}_t)\}) = \left( \begin{array}{l} E(\mathbf{x}) \equiv 1 \\ \mathbf{P} = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ n/a & \cdots & \cdots & n/a \\ \vdots & & & \vdots \end{pmatrix} \end{array} \right).$$

Maximizing the transition matrix likelihood while varying the encoding  $E$  is therefore meaningless.

**Regression loss:** A similar problem is encountered with the regression loss. The theoretical minimum of (23) is equal to 0 and is achieved when  $\mathbf{y}_{t+\tau} \equiv \mathbf{P}^\top \mathbf{y}_t$  for all  $t$ . This, can be trivially achieved by learning the uninformative representation:

$$\arg \max_{E, P} \mathcal{L}_{\text{LSQ}}(\mathbf{P}; \{E(\mathbf{x}_t)\}) = \left( \begin{array}{l} E(\mathbf{x}) \equiv 1 \\ \mathbf{P} = \mathbf{Id} \end{array} \right).$$

Minimizing the propagator least squares error while varying the encoding  $E$  is therefore meaningless. See also discussion in<sup>69</sup>.

**Variational loss:** The variational loss (VAC or VAMP) does not have trivial minima. The reason is that, according to the variational principles<sup>62,120</sup>, the variational optimum coincides with the approximation of the dynamical dynamical components. A trivial encoding such as  $E(\mathbf{x}) \equiv 1$  only identifies a single component and is therefore variationally suboptimal. The variational loss is thus the only choice amongst the losses described in LP1 that can be used to learn both  $\mathbf{y}$  and  $\mathbf{P}$  in an end-to-end fashion.

### B. Feature selection

We first address the problem of learning the featurization  $\mathbf{x}^f$ . We can view this problem as a feature selection problem, i.e. we consider a large potential set of features and ask which of them leads to an optimal model of the long-time MD. In this view, learning the featurization is a model selection problem that can be solved by minimizing the validation loss.

We can solve this problem by employing the variational losses as follows: We compute the spectral representation  $\mathbf{R}$  or  $\mathbf{U}, \mathbf{V}$  directly from the training set  $\mathbf{X}^{\text{train}} = (\mathbf{x}_0^f, \dots, \mathbf{x}_T^f)^\top$  and then recompute

the covariance matrices in the validation set  $\mathbf{X}^{\text{val}}$ . We then compute the following matrices that are diagonal in the training set but only approximately diagonal in the validation set. The VAC and VAMP validation scores can then be computed as  $\mathcal{L}_{\text{VAC}}(\mathbf{U}^{\text{train}}; \{\mathbf{y}_t^{\text{test}}\})$  (Eq. 30,32) or  $\mathcal{L}_{\text{VAMP}}(\mathbf{U}^{\text{train}}; \{\mathbf{y}_t^{\text{test}}\})$  (Eq. 38,40). In<sup>85</sup> we perform VAMP-2 validation in order to select optimal features for describing protein folding and find that a combination of torsion backbone angles and  $\exp(-d_{ij})$  with  $d_{ij}$  being the minimum distances between amino acids.

### C. Blind Source Separation and TICA

For a given featurization, a widely used linear learning method to obtain the spectral representation is an algorithm first introduced in<sup>56</sup> as a method for blind source separation that later became known as time-lagged independent component analysis (TICA) method<sup>1,73,93</sup>, sketched in Algorithm 3. In<sup>73</sup>, it was shown that the TICA algorithm directly follows from the minimization of the VAC variational loss (31,33) to best approximate the Markov operator eigenfunctions by a linear combination of input features. As a consequence, TICA approximate the eigenvalues and eigenfunctions of Markov operators that obey detailed balance (21), and therefore approximates the slowest relaxation processes of the dynamics.

Algorithm 3 performs a symmetrized estimation of covariance matrices in order to guarantee that the eigenvalue spectrum is real. In most early formulations, one usually symmetrizes only  $\mathbf{C}_{0\tau}$  while computing  $\mathbf{C}_{00}$  by (12), which is automatically symmetric. However these formulations might lead to eigenvalues larger than 1, which do not correspond to any meaningful relaxation timescale in the present context – this problem is avoided by the step 1 in Algorithm 3<sup>121</sup>. Note that symmetrization of  $\mathbf{C}_{0\tau}$  introduces an estimation bias if the data is non-stationary, e.g. because short MD trajectories are used that have not been started from the equilibrium distribution. To avoid this problem, please refer to Ref.<sup>121</sup> which introduces the Koopman reweighting procedure to estimate symmetric covariance matrices without this bias, although at the price of an increased estimator variance.

Furthermore, the covariance matrices in step 1 of Algorithm 3 are computed after removing the mean. Removing the mean has the effect of removing the eigenvalue 1 and the corresponding stationary eigenvector, hence all components return by Algorithm 3 approximate dynamical relaxation processes with finite relaxation timescales estimates according to Eq. (9).

The TICA propagator can be directly computed as  $\bar{\mathbf{P}} = \bar{\mathbf{C}}_{00}^{-1} \bar{\mathbf{C}}_{0\tau}$ , and is a least-squares result in the sense of Sec. III D. Various extensions of the TICA algorithm were developed: Kernel formulations of TICA were first presented in machine learning<sup>30</sup> and later in other fields<sup>94,115</sup>. An efficient way to solve TICA for multiple lag times simultaneously was introduced as TDSEP<sup>126</sup>. Efficient computation of TICA for very large feature sets can be performed with a hierarchical decomposition<sup>72</sup> compressed sensing approach<sup>49</sup>. TICA is closely related to the Dynamic Mode Decomposition (DMD)<sup>81,87,88,109</sup> and the Extended Dynamic Mode Decomposition (EDMD) algorithms<sup>114</sup>. DMD approximates the left eigenvectors (“modes”) instead of the Markov operator eigenfunctions described here. EDMD is algorithmically identical to VAC/TICA, but is in practice also used for dynamics that do not fulfill detailed balance (21), although this leads to complex-valued eigenfunctions.

### D. TCCA / VAMP

When the dynamics do not satisfy detailed balance (21), e.g., because they are driven by an external force or field, the TICA algorithm is not meaningful, as it will not even in the limit of infinite data approximate the true spectral representation. If detailed balance holds for the dynamical equations, but the data is non-stationary, i.e. because short simulation trajectories started from a non-equilibrium distribution are used, the symmetrized covariance estimation in Algorithm 3 introduces a potentially large bias.

These problems can be avoided by going from TICA to the time-lagged canonical correlation analysis (TCCA, Algorithm 4) which is a direct implementation of the VAMP approach<sup>120</sup>, i.e. it results from minimizing the VAMP variational loss (39,41), when approximating the Markov operator singular functions with a linear combination of features. The TCCA algorithm performs a canonical correlation analysis (CCA) applied to time series. TCCA returns two sets of features approximating the left and right singular functions of the Markov operator and that can be interpreted as the optimal spectral

1. Compute symmetrized mean free covariance matrices

$$\begin{aligned}\bar{\mathbf{C}}_{00} &= \lambda \mathbf{I} + \sum_{t=1}^{T-\tau} (\mathbf{x}_t - \boldsymbol{\mu}_0)(\mathbf{x}_t - \boldsymbol{\mu}_0)^\top + (\mathbf{x}_{t+\tau} - \boldsymbol{\mu}_\tau)(\mathbf{x}_{t+\tau} - \boldsymbol{\mu}_\tau)^\top \\ \bar{\mathbf{C}}_{0\tau} &= \sum_{t=1}^{T-\tau} (\mathbf{x}_t - \boldsymbol{\mu}_0)(\mathbf{x}_{t+\tau} - \boldsymbol{\mu}_\tau)^\top + (\mathbf{x}_{t+\tau} - \boldsymbol{\mu}_\tau)(\mathbf{x}_t - \boldsymbol{\mu}_0)^\top\end{aligned}$$

with means  $\boldsymbol{\mu}_0, \boldsymbol{\mu}_\tau$  defined analogously as in (10-11), where  $\lambda$  is an optional ridge parameter.

2. Compute the largest  $n$  Eigenvalues and Eigenvectors of:

$$\bar{\mathbf{C}}_{0\tau} \mathbf{u}_i = \lambda_i \bar{\mathbf{C}}_{00} \mathbf{u}_i$$

3. Project to spectral representation:  $\mathbf{y}_t = (\mathbf{x}_t^\top \mathbf{u}_1, \dots, \mathbf{x}_t^\top \mathbf{u}_n)$  for all  $t$

4. Return  $\{\mathbf{y}_t\}$

**Algorithm 3: TICA**( $\{\mathbf{x}_t\}, \tau, n$ ).

representation to characterize state of the system “before” and “after” the transition with time step  $\tau$ . For non-stationary dynamical systems, these representations are valid for particular points in time,  $t$  and  $t + \tau$ <sup>44</sup>.

VAMP/TCCA as a method to obtain a low-dimensional spectral representation of the long time MD is discussed in detail in<sup>71</sup>, where the algorithm is used to identify low-dimensional embeddings of driven dynamical systems, such as an ion channel in an external electrostatic potential.

1. Compute covariance matrices  $\mathbf{C}_{00}, \mathbf{C}_{0\tau}, \mathbf{C}_{\tau\tau}$  from  $\{\mathbf{x}_t\}$ , as in Eqs. (12-14) or Eqs. (15,16).

2. Perform the truncated SVD:

$$\tilde{\mathbf{P}} = \mathbf{C}_{00}^{-\frac{1}{2}} \mathbf{C}_{0\tau} \mathbf{C}_{\tau\tau}^{-\frac{1}{2}} \approx \mathbf{U}' \mathbf{S} \mathbf{V}'^\top$$

where  $\tilde{\mathbf{P}}$  is the propagator for the representations  $\mathbf{C}_{00}^{-\frac{1}{2}} \mathbf{x}_t$  and  $\mathbf{C}_{\tau\tau}^{-\frac{1}{2}} \mathbf{x}_{t+\tau}$ ,  $\mathbf{S} = \text{diag}(s_1, \dots, s_k)$  is a diagonal matrix of the first  $k$  singular values that approximate the true singular values  $\sigma_1, \dots, \sigma_k$ , and  $\mathbf{U}'$  and  $\mathbf{V}'$  consist of the  $k$  corresponding left and right singular vectors respectively.

3. Compute  $\mathbf{U} = \mathbf{C}_{00}^{-\frac{1}{2}} \mathbf{U}'$ ,  $\mathbf{V} = \mathbf{C}_{\tau\tau}^{-\frac{1}{2}} \mathbf{V}'$

4. Project to spectral representation:  $\mathbf{y}_t^0 = (\mathbf{x}_t^\top \mathbf{u}_1, \dots, \mathbf{x}_t^\top \mathbf{u}_n)$  and  $\mathbf{y}_t^\tau = (\mathbf{x}_t^\top \mathbf{v}_1, \dots, \mathbf{x}_t^\top \mathbf{v}_n)$  for all  $t$

5. Return  $\{(\mathbf{y}_t^0, \mathbf{y}_t^\tau)\}$

**Algorithm 4: TCCA**( $\{\mathbf{y}_t\}, \tau, n$ )

## E. MSMs based on geometric clustering

For the spectral representations found by TICA and TCCA, a propagator  $\mathbf{P}(\tau)$  can be computed by means of Eq. (6), however this propagator is harder to interpret than a MSM propagator whose elements correspond to transition probabilities between states. For this reason, TICA, TCCA and other dimension reduction algorithms are frequently used as a first step towards building an MSM<sup>72,73,93</sup>. Before TICA and TCCA were introduced into the MD field, MSMs were directly built upon manually constructed features such as distances, torsions or in other metric spaces that define features only indirectly, such as the pairwise distance of aligned molecules<sup>40,104</sup> – see Ref.<sup>38</sup> for an extensive discussion.

In this approach, the trajectories in feature space,  $\{\mathbf{x}_t^f\}$ , or in the representation  $\{\mathbf{y}_t\}$ , must be further transformed into a one-hot encoding (17) before the MSM can be estimated via one of the methods described in Sec. III. In other words, the configuration space must be divided into  $n$  sets that are associated with the  $n$  MSM states. Typically, clustering methods that somehow group simulation data by means of geometric similarity. When MSMs were build on manually constructed feature spaces, research on suitable clustering methods was very active<sup>2,9,11–13,37,39,41,61,86,97,101,124</sup>. Since the introduction of TICA and TCCA that identify a spectral representation that already approximates

the leading eigenfunctions, the choice of the clustering method has become less critical, and simple methods such as  $k$ -means<sup>++</sup> lead to robust results. The final step towards an easily interpretable MSM is coarse-graining of  $\mathbf{P}$  down to a few states<sup>21,26,36,45,64,68,123</sup>.

The geometric clustering step introduces a different learning problem and objective whose relationship to the original problem of approximating long-term MD is not clear. Therefore, geometric clustering must be at the moment regarded as a pragmatic approach to construct an MSM from a given embedding, but this approach departs from the avenue of a well-defined machine learning problem.

## F. VAMPnets

VAMPnets<sup>51</sup> were introduced to replace the complicated and error-prone approach of constructing MSMs by (i) searching for optimal features  $\mathbf{x}^f$ , (ii) combining them to a representation  $\mathbf{y}$ , *e.g.*, *via* TICA, (iii) clustering it, (iv) estimating the transition matrix  $\mathbf{P}$ , and (v) coarse-graining it, by a single end-to-end learning approach in which all of these steps are replaced by a deep neural network. This is possible because with the VAC and VAMP variational principles, loss functions are available that are suitable to train the sequence of learning problems 1 and 2 simultaneously. A similar architecture is used by EDMD with dictionary learning<sup>47</sup>, which avoids the problem of the regression error to collapse to trivial encodings  $E$  (Sec. V A) by fixing some features that are not learnable.

VAMPnets contain two network lobes that transform the molecular configurations found at a time delay  $\tau$  along the simulation trajectories (Fig. 2a). VAMPnets can be minimized with any VAC or VAMP variational loss. In Ref.<sup>51</sup>, the VAMP-2 loss (41) was used, which is meaningful for both dynamics with and without detailed balance. When detailed balance (22) is enforced in the propagator obtained by (6), the loss function automatically becomes VAC-2. VAMPnets may either use two distinct network lobes to encode the spectral representation of the left and right singular functions (which is important for non-stationary dynamics<sup>43,44</sup>), whereas for MD with a stationary distribution we generally use parameter sharing and have two identical lobes. For dynamics with detailed balance, the VAMPnet output then encodes the space of the dominant Markov operator eigenfunctions (Fig. 3b).

In order to obtain a propagator that can be interpreted as an MSM,<sup>51</sup> chose to use a SoftMax layer as an output layer, thus transforming the spectral representation to a soft indicator function similar to spectral clustering methods such as PCCA+<sup>17,78</sup>. As a result, the propagator computed by Eq. (6) is *almost* a transition matrix. It is guaranteed to be a true transition matrix in the limit where the output layer performs a hard clustering, *i.e.* one-hot encoding (17). Since this is not true in general, the VAMPnet propagator may still have negative elements, but these are usually very close to zero. The propagator is still valid for transporting probability distributions in time and can therefore be interpreted as an MSM between metastable states (Fig. 4d).

The results described in<sup>51</sup> (see, *e.g.*, Fig. 3,4) were competitive with and sometimes surpassed the state-of-the-art handcrafted MSM analysis pipeline. Given the rapid improvements of training efficiency and accuracy of deep neural networks seen in a broad range of disciplines, we expect end-to-end learning approaches such as VAMPnets to dominate the field eventually.

## VI. LP3 LIGHT: LEARN REPRESENTATION AND DECODER

As discussed in Sec. V A, end-to-end learning combining LP1 and LP2 are limited in their choice of losses applied to the propagator resulting from LP2: Variational losses can be used, leading to the methods described in Sec. V, while using the likelihood and regression losses are prone to collapse to a trivial representation that does not resolve the long-time dynamical processes.

One approach to “rescue” these approaches is to add other loss functions to prevent this collapse to a trivial, uninformative representation from happening. An obvious choice is to add a decoder that is trained with some form of reconstruction loss: the representation  $\mathbf{r}$  should still contain enough information that the input ( $\mathbf{x}$  or  $\mathbf{y}$ ) can be approximately reconstructed. We discuss several approaches based on this principle. Note that if only finding the spectral embedding and learning the propagator  $\mathbf{P}$  is the objective, VAMPnets solve this problem directly and employing a reconstruction loss unnecessarily adds the difficult inverse problem of reconstructing a high-dimensional variable from a low-dimensional one. However, approximate reconstruction of inputs may be desired in some applications, and is the basis for LP3.

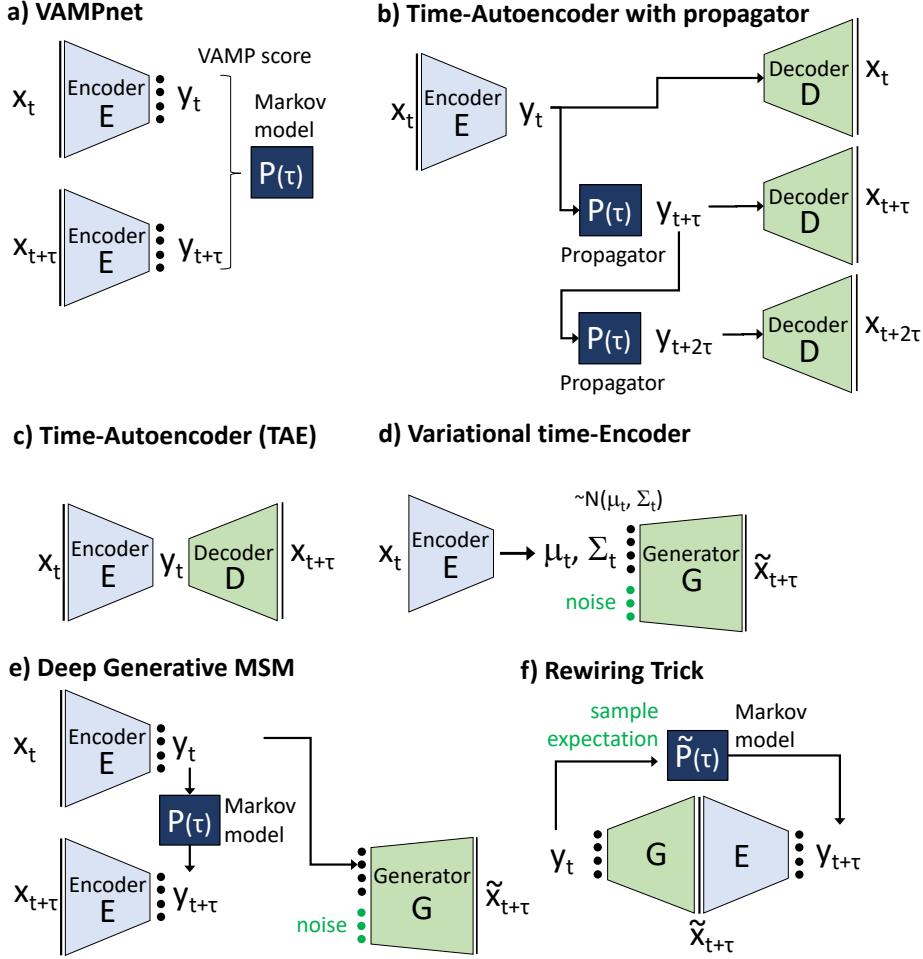


Figure 2. Overview of network structures for learning Markovian dynamical models. **a)** VAMPnets<sup>51</sup>. **b)** Time-autoencoder with propagator<sup>50,69</sup>. **c)** time-autoencoder<sup>113</sup>. **d)** variational time-Encoder<sup>32</sup>. **e)** Deep Generative Markov State Models<sup>117</sup>. **f)** The rewiring trick to compute the propagator  $\mathbf{P}$  for a deep generative MSM.

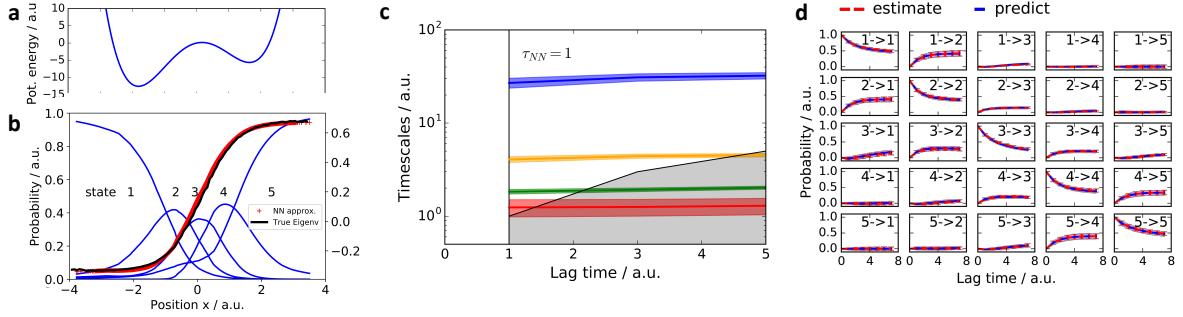


Figure 3. Figure adapted from<sup>51</sup>: Approximation of the slow transition in a bistable potential by a VAMPnet with one input node ( $x$ ) and five output nodes. **(a)** Potential energy function  $U(x) = x^4 - 6x^2 + 2x$ . **(b)** Eigenvector of the slowest process calculated by direct numerical approximation (black) and approximated by a VAMPnet with five output nodes (red). Activation of the five Softmax output nodes define the state membership probabilities (blue). **(c)** Relaxation timescales computed from the Koopman model using the VAMPnet transformation. **(d)** Chapman-Kolmogorov test comparing long-time predictions of the Koopman model estimated at  $\tau = 1$  and estimates at longer lag times. Panels (c) and (d) report 95% confidence interval error bars over 100 training runs.

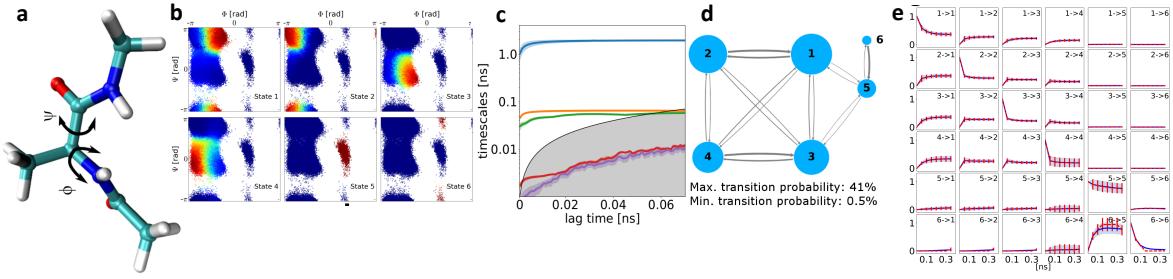


Figure 4. Figure adapted from<sup>51</sup>: Kinetic model of alanine dipeptide obtained by a VAMPnet with 30 input nodes ( $x, y, z$  Cartesian coordinates of heavy atoms) and six output nodes. (a) Structure of alanine dipeptide. The main coordinates describing the slow transitions are the backbone torsion angles  $\phi$  and  $\psi$ , however the neural network inputs are only the Cartesian coordinates of heavy atoms. (b) Assignment of all simulated molecular coordinates, plotted as a function of  $\phi$  and  $\psi$ , to the six Softmax output states. Color corresponds to activation of the respective output neuron, indicating the membership probability to the associated metastable state. (c) Relaxation timescales computed from the Koopman model using the neural network transformation. (d) Representation of the transition probabilities matrix of the Koopman model; transitions with a probability lower than 0.5% have been omitted. (e) Chapman-Kolmogorov test comparing long-time predictions of the Koopman model estimated at  $\tau = 50\text{ ps}$  and estimates at longer lag times. Panels (c) and (e) report 95% confidence interval error bars over 100 training runs excluding failed runs.

### A. Time-Autoencoder

The time-autoencoder<sup>113</sup> shortcuts LP2 and constructs a direct learning problem between  $\mathbf{x}_t$  and  $\mathbf{x}_{t+\tau}$  (Fig. 2c).

$$\mathbf{x}_t \xrightarrow{E} \mathbf{y}_? \xrightarrow{D} \mathbf{x}_{t+\tau} \quad (46)$$

The time-autoencoder is trained by reconstruction loss:

$$\mathcal{L}_{\text{TAE}}(E, D; \{\mathbf{x}_t\}) = \sum_{t=0}^{T-\tau} \|\mathbf{x}_{t+k\tau} - D(E(\mathbf{x}_t))\| \quad (47)$$

where  $\|\cdot\|$  is a suitable norm, *e.g.*, the squared 2-norm.

The TAE has an interesting interpretation: If  $E$  and  $D$  are linear transformation, i.e. encoder and decoder matrices  $\mathbf{E} \in \mathbb{R}^{N \times n}$ ,  $\mathbf{D} \in \mathbb{R}^{n \times N}$ , the minimum of (47) is found by VAMP/TCCA, and for data that is in equilibrium and obeys detailed balance by VAC/TICA<sup>113</sup>. The reverse interpretation is not true: the solution found by minimizing (47) does not lead to TICA/TCCA modes, as there is no constraint in the time-autoencoder for the components  $\mathbf{r}_t$  – they only span the same space. Within this interpretation, the time-autoencoder can be thought of a nonlinear version of TCCA/TICA in the sense of being able to find a slow but nonlinear spectral representation.

Time-autoencoders have several limitations compared to VAMPnets: (1) Adding the decoder network makes the learning problem more difficult. (2) As indicated in scheme (46), it is not clear what the time step pertaining to the spectral representation  $\mathbf{y}$  is ( $t, t+\tau$ , or something in between), as the time stepping is done throughout the entire network. (3) Since the decoding problem from any given  $\mathbf{y}$  to  $\mathbf{x}_{t+\tau}$  is underdetermined but the decoder network  $D$  is deterministic, it will only be able to decode to a “mean”  $\mathbf{x}$  for all  $\mathbf{x}$  mapping to the same  $\mathbf{y}$ . Thus, time-autoencoders cannot be used to sample the transition density (3) to generated sequences  $\mathbf{x}_t \rightarrow \mathbf{x}_{t+\tau}$ .

### B. Time-Autoencoder with Propagator

Both<sup>50,69</sup> have introduced time-autoencoders that additionally learn the propagator in the spectral representation, and thus fix problem (2) of time-autoencoders, while problems (1) and (3) still remain. Instead of scheme (46), time-autoencoders with propagator introduce a time propagation step that makes the time step explicit for every step:

$$\mathbf{x}_t \xrightarrow{E} \mathbf{y}_t \xrightarrow{\mathbf{P}} \mathbf{y}_{t+\tau} \xrightarrow{D} \mathbf{x}_{t+\tau} \quad (48)$$

where  $\mathbf{P}$  is the matrix defined by a  $n \times n$  linear layer. Training this network exclusively with the standard autoencoder loss would not impose the correct internal structure – in particular, it would not be possible to control that  $E$  learns only the representation and  $\mathbf{P}$  performs the time step.<sup>50,69</sup> enforce the dynamical consistency by training several lag times simultaneously with variants of the following type of loss:

$$\mathcal{L}_{\text{TAE-P}} = \sum_{t=0}^{T-k\tau} \left( \sum_{k=0}^K \alpha_k \|\mathbf{x}_{t+k\tau} - D(\mathbf{P}^k E(\mathbf{x}_t))\| + \sum_{k=1}^K \beta_k \|E(\mathbf{x}_{t+k\tau}) - \mathbf{P}^k E(\mathbf{x}_t)\| \right) \quad (49)$$

where  $\alpha_k, \beta_k$  are coefficients, the first term correspond to a autoencoder reconstruction loss and the second term trains the correct time-propagation of  $\mathbf{P}$  in latent space. The number of lag times,  $K$ , to be considered is a user-defined choice. Note that it is not a typical hyper-parameter as matching the dynamics at more lag times makes the learning problem harder, and thus the cross-validation score of (49) cannot be used to select  $K$ . Unrolling the network for  $K = 2$  results in Fig. 2b. This approach works excellently in deterministic (but highly nonlinear) dynamical systems with short time steps<sup>50,69</sup>. In stochastic systems such as MD, it appears more difficult to learn  $\mathbf{r}_t$  and  $\mathbf{P}$  such that they span the spectral components of the underlying propagator and recover its largest eigenvalues. While this observation needs more study, potential explanations are that in long-time MD we need large time steps  $\tau$ , in order to make the spectral representation learning problem low-dimension (see Sec. IV A), and that the stochastic fluctuations are large which makes learning a decoder  $D$  difficult.

### C. Variational (time-)Autoencoders

Several recent approaches employ variational autoencoders (VAEs) for the long-time MD or related learning problems. Variational autoencoders<sup>42</sup> learn to sample a probability distribution that approximates the distribution underlying observation data. To this end, VAEs employ variational Bayesian inference<sup>23</sup> in order to approximately minimize the KL divergence between the generated and the observed distribution. VAEs have a similar structure as usual autoencoders, with an inference network mapping from a high-dimensional variable  $\mathbf{x}$  to a typically lower-dimensional latent variable  $\mathbf{r}$ , and attempting to reconstruct  $\mathbf{x}$  in a decoder network. The main difference is that every latent point  $\mathbf{r}$  encodes the moments of a distribution which are used to sample  $\mathbf{x}$  such that the distributions become similar.

VAEs have been used in RAVE<sup>77</sup> for enhancing the sampling by identifying a space of “reaction coordinates” in which MD sampling can be efficiently driven, and in Autograin<sup>112</sup> to find a way to coarse-grain a molecule into effective beads. Both methods use VAEs without an inference network that employs a time step  $\tau$ , and therefore they address learning problems that are conceptually different from long-time MD learning problem as treated here.

A much more closely related work are variational time-encoders<sup>32</sup> (Fig. 2d), which employ a VAE between time steps  $\mathbf{x}_t$  at the input and  $\mathbf{x}_{t+\tau}$  at the output:

$$\begin{array}{c} \mathbf{x}_t \xrightarrow{E} \mu(\mathbf{x}_t) \rightarrow \mathbf{y}_t \rightarrow \oplus \rightarrow \mathbf{y}_{t+\tau} \xrightarrow{D} \mathbf{x}_{t+\tau} \\ \uparrow \\ \mathcal{N}(0, 1) \end{array}$$

As<sup>32</sup> note, this approach does not achieve the sampling of the  $\mathbf{x}_{t+\tau}$  distribution (the variational theory underlying VAEs requires that the same type of variable is used at input and output) and hence does not act as a propagator  $\mathbf{x}_t \rightarrow \mathbf{x}_{t+\tau}$ , but succeeds in learning a spectral representation of the system. For this reason, the variational time-encoder is listed in this section rather than in LP3.

## VII. LP3 HEAVY: LEARN GENERATIVE MODELS

The full solution of LP3 involves learning to generate samples  $\mathbf{x}_{t+\tau}$  from the lower-dimensional feature embedding or spectral representation. This is a very important goal as its solution would yield an

ability to sample the MD propagator  $\mathbf{x}_t \rightarrow \mathbf{x}_{t+\tau}$  at long time-steps  $\tau$ , which would yield a very efficient simulator. However, because of the high dimensionality of configuration space and the complexity of distributions there, this aim is extremely difficult and still in its infancy.

Clearly standard tools for learning directed generative networks, such as Variational Autoencoders<sup>42</sup> and generative adversarial nets<sup>27</sup> are “usual suspects” for the solution of this problem. However, existing applications of VAEs and GANs on the long-time MD problem have focused on learning a latent representation that is suitable to encode the long-time processes or a coarse-graining, and the decoder has been mostly used to regularize the problem (Sec. VI C). The first approach to actually reconstruct molecular structures in configuration space, so as to achieve long-time-step sampling, was made in<sup>117</sup>, which will be analyzed in some detail below.

### A. Deep Generative MSMs

The deep generative MSMs described<sup>117</sup> (Fig. 2e), we propose to address LP1-3 in the following manner. We first formulate a machine learning problem to learn the following two functions:

- An probabilistic encoding of the input configuration to a low-dimensional latent space,  $\mathbf{x}_t \rightarrow E(\mathbf{x}_t)$ . Similar to VAMPnets with a probabilistic output (Sec. VF),  $\chi$  has  $n$  elements, and each element represents the probability of configuration  $\mathbf{x}$  to be in a metastable (long-lived) state  $i$ :

$$E_i(\mathbf{x}) = \mathbb{P}(\mathbf{x}_t \in \text{state } i \mid \mathbf{x}_t = \mathbf{x}).$$

Consequently, these functions are nonnegative ( $E_i(x) \geq 0 \forall x$ ) and sum up to one ( $\sum_i E_i(x) = 1 \forall x$ ). The functions  $E(x)$  can, e.g., be represented by a neural network mapping from  $\mathbb{R}^d$  to  $\mathbb{R}^m$  with a SoftMax output layer.

- An  $n$ -element probability distribution  $\mathbf{q}(\mathbf{x}; \tau) = (q_1(\mathbf{x}; \tau), \dots, q_n(\mathbf{x}; \tau))$ , which assigns to each configuration  $\mathbf{x}$  a probability density that a configuration that was in metastable state  $i$  at time  $t$ , will “land” in  $\mathbf{x}$  at time  $t + \tau$ :

$$q_i(\mathbf{x}; \tau) = \mathbb{P}(\mathbf{x}_{t+\tau} = \mathbf{x} \mid \mathbf{x}_t \in \text{state } i).$$

We thus briefly call these densities “landing densities”.

Schematically, Deep generative MSMs treat LP1-3 in the way:

$$\begin{array}{ccc} \mathbf{x}_t & \xrightarrow{E} & \mathbf{y}_t \\ & \swarrow q & \\ & \mathbf{x}_{t+\tau} & \end{array}$$

Deep generative MSMs represent the transition density (3) in the following form (Fig. 2e):

$$p_\tau(\mathbf{x}_{t+\tau} \mid \mathbf{x}_t) = E(\mathbf{x}_{t+\tau})^\top \mathbf{q}(\mathbf{y}; \tau) = \sum_{i=1}^m E_i(\mathbf{x}_t) q_i(\mathbf{x}_{t+\tau}; \tau). \quad (50)$$

To work with this approach we finally need a Generator  $G$ , which is a structure that samples from the density  $\mathbf{q}$ :

$$G(i, \epsilon; \tau) = \mathbf{y} \sim q_i(\mathbf{y}; \tau) \quad (51)$$

It appears that Deep generative MSMs do not learn the propagator explicitly. However, the propagator can be obtained from  $E$  and  $\mathbf{q}$  by means the “rewiring” trick (Fig. 2f): By exchanging the order in which  $E$  and  $G$  are applied and then computing the propagator  $\mathbf{P}$  as a sample average over  $\mathbf{q}$ , obtained from repeatedly applying the generator:

$$p_{ij}(\tau) = \mathbb{E}_G [E_j(G(i, \epsilon; \tau))]. \quad (52)$$

In contrast to VAMPnets (Sec. VF), it is guaranteed that the propagator (52) is a true transition matrix with nonnegative elements.

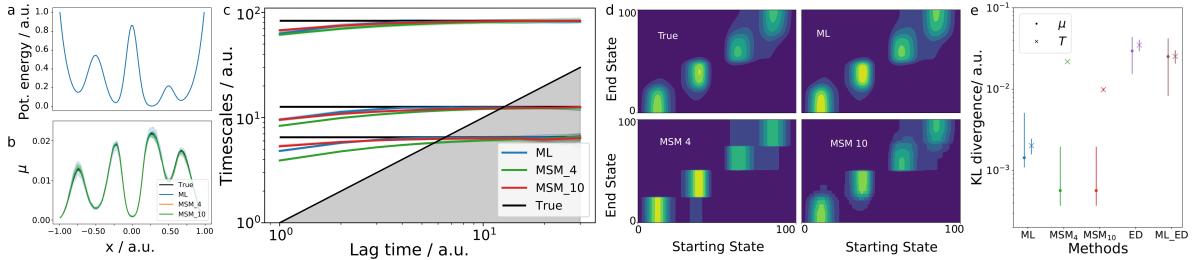


Figure 5. Reproduced from<sup>117</sup>: Performance of deep versus standard MSMs for diffusion in the Prinz Potential. (a) Potential energy as a function of position  $x$ . (b) Stationary distribution estimates of all methods with the exact distribution (black). (c) Implied timescales of the Prinz potential compared to the real ones (black line). (d) True transition density and approximations using maximum likelihood (ML) DeepResampleMSM, four and ten state MSMs. (e) KL-divergence of the stationary and transition distributions with respect to the true ones for all presented methods (also DeepResampleMSM).

## B. Deep Resampling MSMs

We first describe a very simple generator that generates no new (unseen) configurations, but only learns a function  $\mathbf{q}$  that can be used to resample known configurations<sup>117</sup>. While this approach is clearly limited, it has two advantages: it will not generate any illegal configuration, and it can be trained with maximum likelihood. For this approach, we model the landing densities by

$$q_i(\mathbf{x}_{t+\tau}) = \frac{w(\mathbf{x}_{t+\tau})\gamma_i(\mathbf{x}_{t+\tau})}{\sum_{s=0}^{T-\tau} w(\mathbf{x}_{s+\tau})\gamma_i(\mathbf{x}_{s+\tau})}. \quad (53)$$

Where  $\gamma_i(\mathbf{x}_{t+\tau})$  is a trainable, unnormalized density function and  $w$  is an additional weight function which may be employed to change the weights of configurations, but is usually identical to 1. In<sup>117</sup>,  $\gamma_i(\mathbf{y})$  is a deep neural network that receives  $\mathbf{y}$  as an input as well as the condition  $i$  by means of a one-hot-encoding with  $n$  input units, and has a single output node encoding the probability weight. The normalized density  $\mathbf{q}$  is computed by evaluating the  $\gamma$ -network for all configurations at time points  $\tau, \dots, T$  and then normalizing over all time points.

Deep resampling MSMs can be trained by maximizing the likelihood based on expression (50), resulting in the following loss function:

$$\begin{aligned} \mathcal{L}_{\text{DeepResampleMSM}} &= \sum_{t=1}^{T-\tau} p_\tau(\mathbf{x}_{t+\tau} | \mathbf{x}_t) \\ &= \sum_{t=1}^{T-\tau} \sum_{i=1}^m E_i(\mathbf{x}_t) q_i(\mathbf{x}_{t+\tau}; \tau) \end{aligned}$$

where  $q_i$  is evaluated by (53). Alternatively, we can optimize  $\chi_i$  and  $\gamma_i$  using the Variational Approach for Markov Processes (VAMP)<sup>120</sup>. However, we found the ML approach to perform significantly better in<sup>117</sup>.

In Deep Resample MSMs, the propagator according to (52) becomes simply:

$$\mathbf{P} = \frac{1}{T-\tau} \sum_{t=1}^{T-\tau} \mathbf{q}(\mathbf{x}_{t+\tau}) E(\mathbf{x}_{t+\tau})^\top. \quad (54)$$

Deep Resample MSMs were found to accurately reproduce the eigenfunctions and dominant relaxation timescales of benchmark examples<sup>117</sup>, and learn to represent the transition density in configuration space (Fig. 5).

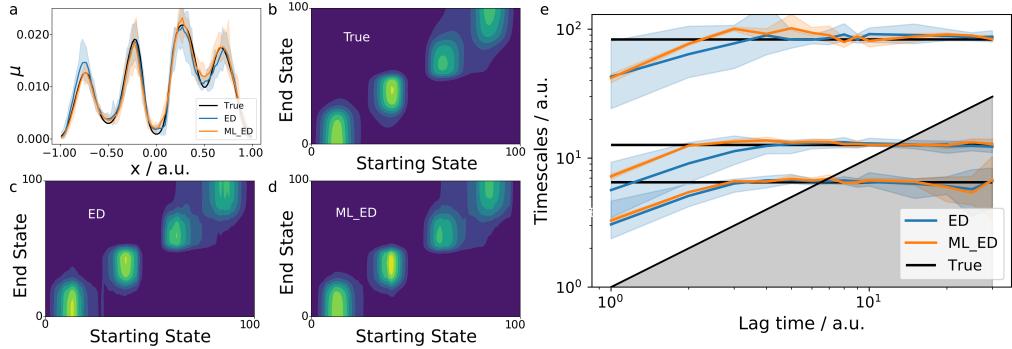


Figure 6. Reproduced from<sup>117</sup>: Performance of deep generative MSMs for diffusion in the Prinz Potential. Comparison between exact reference (black), deep generative MSMs estimated using only energy distance (ED) or combined ML-ED training. (a) Stationary distribution. (b-d) Transition densities. (e) Relaxation timescales.

### C. Deep Generative MSMs with Energy Distance Loss

In contrast to resampling MSMs, we now want to generate MSMs, which can produce genuinely new configurations. This makes the method promising for performing active learning in MD<sup>10,74</sup>, and to predict the future evolution of the system in other contexts. To this end, we train a directed generative network to represent (51). Such a generator can be trained with various principles, e.g. by means of a variational autoencoder or with adversarial training<sup>27,42</sup>. In<sup>117</sup>, we found that a third principle works well: training the generator  $G$  by minimizing the conditional Energy Distance (ED). The standard ED, introduced in<sup>102</sup>, is a metric between the distributions of random vectors, defined as

$$D_E(\mathbb{P}(\mathbf{x}), \mathbb{P}(\mathbf{y})) = \mathbb{E}[2\|\mathbf{x} - \mathbf{y}\| - \|\mathbf{x} - \mathbf{x}'\| - \|\mathbf{y} - \mathbf{y}'\|] \quad (55)$$

for two real-valued random vectors  $\mathbf{x}$  and  $\mathbf{y}$ .  $\mathbf{x}', \mathbf{y}'$  are independently distributed according to the distributions of  $\mathbf{x}, \mathbf{y}$ . Based on this metric, we introduce the conditional energy distance between the transition density of the system and that of the generative model:

$$\begin{aligned} D &\triangleq \mathbb{E}[D_E(\mathbb{P}(\mathbf{x}_{t+\tau} | \mathbf{x}_t), \mathbb{P}(\hat{\mathbf{x}}_{t+\tau} | \mathbf{x}_t)) | \mathbf{x}_t] \\ &= \mathbb{E}[2\|\hat{\mathbf{x}}_{t+\tau} - \mathbf{x}_{t+\tau}\| - \|\hat{\mathbf{x}}_{t+\tau} - \hat{\mathbf{x}}'_{t+\tau}\| - \|\mathbf{x}_{t+\tau} - \hat{\mathbf{x}}'_{t+\tau}\|] \end{aligned} \quad (56)$$

Here  $\mathbf{x}_{t+\tau}$  and  $\mathbf{x}'_{t+\tau}$  are distributed according to the transition density for given  $\mathbf{x}_t$  and  $\hat{\mathbf{x}}_{t+\tau}$ ,  $\hat{\mathbf{x}}'_{t+\tau}$  are independent outputs of the generative model. Implementing the expectation value with an empirical average results in an estimate for  $D$  that is unbiased, up to an additive constant. We train  $G$  to minimize  $D$ , and subsequently estimate  $\mathbf{P}$  by using the rewiring trick and sampling (52).

Deep Generative MSMs trained with the energy distance were also found to accurately reproduce the eigenfunctions and dominant relaxation timescales of benchmark examples<sup>117</sup>, and learn to represent the transition density in configuration space (Fig. 6). In contrast to Resampling MSMs described in the previous section, they can also be used to generalize to sampling new, previously unseen, configurations, and are therefore a first approach to sample the long-time propagator  $\mathbf{x}_t \rightarrow \mathbf{x}_{t+\tau}$  in configuration space (Fig. 7).

## VIII. DATA AND SOFTWARE

Many of the algorithms described above are implemented in the PyEMMA<sup>86,95</sup> software – [www.pyemma.org](http://www.pyemma.org) and in MSMbuilder<sup>31</sup>. Some of the deep learning algorithms can be found at <https://github.com/markovmodel/deeptime>. The field is still lacking good resources with public datasets, partially because long-time MD data of nontrivial systems is typically extremely large (giga- to terabytes), and due to the unsupervised nature of the learning problems, the role of a benchmarking dataset is less straightforward as in supervised learning. Commonly used datasets for the evaluation of long-time MD models are the fast folding protein trajectories produced by D. E. Shaw research on the Anton supercomputer<sup>48</sup>, which can be obtained from them on request. We provide datasets for small peptides via the Python package mdshare <https://markovmodel.github.io/mdshare/>.

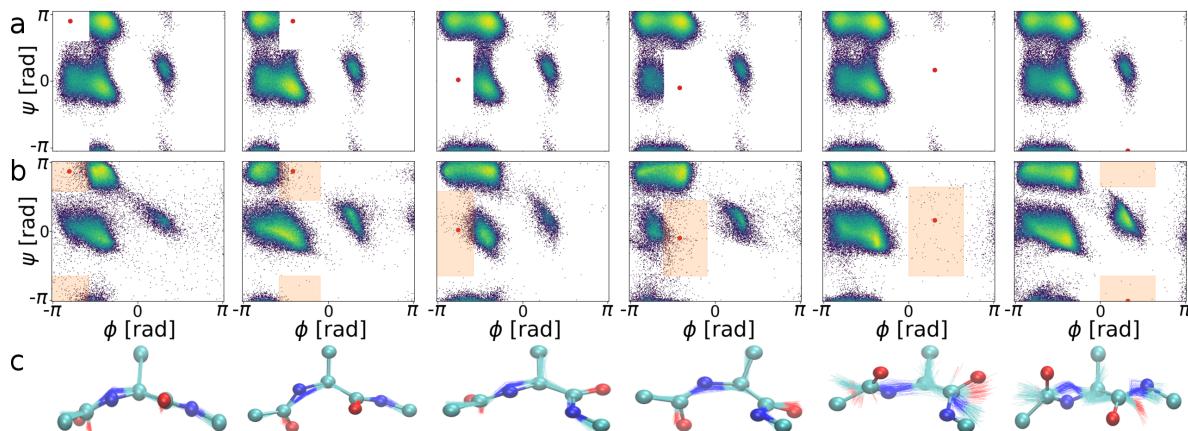


Figure 7. Reproduced from<sup>117</sup>: DeepGenMSMs can generate physically realistic structures in areas that were not included in the training data. (a) Distribution of training data. (b) Generated stationary distribution. (c) Representative “real” molecular configuration (from MD simulation) in each of the metastable states (sticks and balls), and the 100 closest configurations generated by the deep generative MSM (lines).

## REFERENCES

- <sup>1</sup>Erkki Oja Aapo Hyvärinen, Juha Karhunen. *Independent Component Analysis*. John Wiley & Sons, 2001.
- <sup>2</sup>A. Altis, P. H. Nguyen, R. Hegger, and G. Stock. Dihedral angle principal component analysis of molecular dynamics simulations. *J. Chem. Phys.*, 126:244111, 2007.
- <sup>3</sup>S. Bacallado, J. D. Chodera, and V. S. Pande. Bayesian comparison of Markov models of molecular dynamics with detailed balance constraint. *J. Chem. Phys.*, 131:045106, 2009.
- <sup>4</sup>C. Bartels. Analyzing biased monte carlo and molecular dynamics simulations. *Chem. Phys Lett.*, 331:446–454, 2000.
- <sup>5</sup>Christian Bartels and Martin Karplus. Multidimensional adaptive umbrella sampling: Application to main chain and side chain peptide conformations. *J. Comp. Chem.*, 18:1450–1462, 1997.
- <sup>6</sup>A. P. Bartók, R. Kondor, and G. Csányi. On representing chemical environments. *Phys. Rev. B*, 87:184115, 2013.
- <sup>7</sup>J. Behler and M. Parrinello. Generalized neural-network representation of high-dimensional potential-energy surfaces. *Phys. Rev. Lett.*, 98:146401, 2007.
- <sup>8</sup>T. Bereau, R. A. DiStasio Jr, A. Tkatchenko, and O. A. Von Lilienfeld. Non-covalent interactions across organic and biological subsets of chemical space: Physics-based potentials parametrized from machine learning. *J. Chem. Phys.*, 148:241706, 2018.
- <sup>9</sup>G. R. Bowman, K. A. Beauchamp, G. Boxer, and V. S. Pande. Progress and challenges in the automated construction of Markov state models for full protein systems. *J. Chem. Phys.*, 131:124101, 2009.
- <sup>10</sup>G. R. Bowman, D. L. Ensign, and V. S. Pande. Enhanced Modeling via Network Theory: Adaptive Sampling of Markov State Models. *J. Chem. Theory Comput.*, 6(3):787–794, 2010.
- <sup>11</sup>G. R. Bowman, V. S. Pande, and F. Noé, editors. *An Introduction to Markov State Models and Their Application to Long Timescale Molecular Simulation*., volume 797 of *Advances in Experimental Medicine and Biology*. Springer Heidelberg, 2014.
- <sup>12</sup>N. V. Buchete and G. Hummer. Coarse Master Equations for Peptide Folding Dynamics. *J. Phys. Chem. B*, 112:6057–6069, 2008.
- <sup>13</sup>J. D. Chodera, K. A. Dill, N. Singhal, V. S. Pande, W. C. Swope, and J. W. Pitera. Automatic discovery of metastable states for the construction of Markov models of macromolecular conformational dynamics. *J. Chem. Phys.*, 126:155101, 2007.
- <sup>14</sup>J. D. Chodera and F. Noé. Probability distributions of molecular observables computed from markov models. ii: Uncertainties in observables and their time-evolution. *J. Chem. Phys.*, 133:105102, 2010.
- <sup>15</sup>R. R. Coifman, S. Lafon, A. B. Lee, M. Maggioni, B. Nadler, F. Warner, and S. W. Zucker. Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps. *Proc. Natl. Acad. Sci. USA*, 102:7426–7431, 2005.
- <sup>16</sup>P. Das, M. Moll, H. Stamati, L. E. Kavraki, and C. Clementi. Low-dimensional, free-energy landscapes of protein-folding reactions by nonlinear dimensionality reduction. *Proc. Natl. Acad. Sci. USA*, 103:9885–9890, 2008.
- <sup>17</sup>P. Deuflhard and M. Weber. Robust perron cluster analysis in conformation dynamics. In M. Dellnitz, S. Kirkland, M. Neumann, and C. Schütte, editors, *Linear Algebra Appl.*, volume 398C, pages 161–184. Elsevier, New York, 2005.
- <sup>18</sup>P. D. Dixit and K. A. Dill. Caliber corrected markov modeling (c2m2): Correcting equilibrium markov models pd dixit, ka dill. *J. Chem. Theor. Comput.*, 14:1111–1119, 2018.
- <sup>19</sup>S. Doerr and G. De Fabritiis. On-the-fly learning and sampling of ligand binding by high-throughput molecular simulations. *J. Chem. Theory Comput.*, 10:2064–2069, 2014.
- <sup>20</sup>S. Doerr, M. J. Harvey, F. Noé, and G. De Fabritiis. HTMD: High-Throughput Molecular Dynamics for Molecular Discovery. *J. Chem. Theory Comput.*, 12:1845–1852, 2016.
- <sup>21</sup>K. Fackeldey and M. Weber. Genpcca – markov state models for non-equilibrium steady states. *WIAS Report*, 29:70–80, 2017.
- <sup>22</sup>A. M. Ferrenberg and R. H. Swendsen. Optimized monte carlo data analysis. *Phys. Rev. Lett.*, 63:1195–1198, 1989.

- <sup>23</sup>Charles W. Fox and Stephen J. Roberts. A tutorial on variational bayesian inference. *Artificial Intelligence Review*, 38:85–95, 2012.
- <sup>24</sup>Hiroaki Fukunishi, Osamu Watanabe, and Shoji Takada. On the hamiltonian replica exchange method for efficient sampling of biomolecular systems: Application to protein structure prediction. *J. Chem. Phys.*, 116:9058, 2002.
- <sup>25</sup>Emilio Gallicchio, Michael Andrec, Anthony K. Felts, and Ronald M. Levy. Temperature weighted histogram analysis method, replica exchange, and transition paths. *J. Phys. Chem. B*, 109:6722–6731, 2005.
- <sup>26</sup>S. Gerber and I. Horenko. Toward a direct and scalable identification of reduced models for categorical processes. *Proc. Natl. Acad. Sci. USA*, 114:4863–4868, 2017.
- <sup>27</sup>I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and J. Bengio. Generative adversarial networks. In *NIPS'14 Proceedings of the 27th International Conference on Neural Information Processing Systems*, volume 2, pages 2672–2680, MA, USA, 2014. MIT Press Cambridge.
- <sup>28</sup>H. Grubmüller. Predicting slow structural transitions in macromolecular systems: conformational flooding. *Phys. Rev. E*, 52:2893, 1995.
- <sup>29</sup>Ulrich H.E. Hansmann. Parallel tempering algorithm for conformational studies of biological molecules. *Chem. Phys. Lett.*, 281(1-3):140 – 150, 1997.
- <sup>30</sup>S. Harmeling, A. Ziehe, M. Kawanabe, and K.-R. Müller. Kernel-based nonlinear blind source separation. *Neur. Comp.*, 15:1089–1124, 2003.
- <sup>31</sup>M. P. Harrigan, M. M. Sultan, C. X. Hernández, B. E. Husic, P. Eastman, C. R. Schwantes, K. A. Beauchamp, R. T. McGibbon, and V. S. Pande. Msmbuilder: Statistical models for biomolecular dynamics. *Biophys J.*, 112:10–15, 2017.
- <sup>32</sup>Carlos X. Hernández, Hannah K. Wayment-Steele, Mohammad M. Sultan, Brooke E. Husic, and Vijay S. Pande. Variational encoding of complex dynamics. *Phys. Rev. E*, 97:062412, 2018.
- <sup>33</sup>N. S. Hinrichs and V. S. Pande. Calculation of the distribution of eigenvalues and eigenvectors in Markovian state models for molecular dynamics. *J. Chem. Phys.*, 126:244101, 2007.
- <sup>34</sup>Nina S. Hinrichs and Vijay S. Pande. Calculation of the distribution of eigenvalues and eigenvectors in markovian state models for molecular dynamics. *J. Chem. Phys.*, 126:244101, 2007.
- <sup>35</sup>G. Hummer and J. Köfinger. Bayesian ensemble refinement by replica simulations and reweighting. *J. Chem. Phys.*, 143:243150, 2015.
- <sup>36</sup>G. Hummer and A. Szabo. Optimal dimensionality reduction of multistate kinetic and markov-state models. *J. Phys. Chem. B*, 119:9029–9037, 2015.
- <sup>37</sup>B. E. Husic and V. S. Pande. Ward clustering improves cross-validated markov state models of protein folding. *J. Chem. Theo. Comp.*, 13:963–967, 2017.
- <sup>38</sup>B. E. Husic and V. S. Pande. Markov state models: From an art to a science. *J. Am. Chem. Soc.*, 140:2386–2396, 2018.
- <sup>39</sup>A. Jain and G. Stock. Identifying metastable states of folding proteins. *J. Chem. Theory Comput.*, 8:3810–3819.
- <sup>40</sup>Wolfgang Kabsch. A solution for the best rotation to relate two sets of vectors. *Acta Cryst.*, A32:922–923, 1976.
- <sup>41</sup>B. G. Keller, X. Daura, and W. F. van Gunsteren. Comparing geometric and kinetic cluster algorithms for molecular simulation data. *J. Chem. Phys.*, 132:074110, 2010.
- <sup>42</sup>D. P. Kingma and M. Welling. Auto-encoding variational bayes. In *Proceedings of the 2nd International Conference on Learning Representations (ICLR)*, arXiv:1312.6114, 2014.
- <sup>43</sup>P. Kolta, G. Ciccotti, and Ch. Schütte. On metastability and markov state models for non-stationary molecular dynamics. *J. Chem. Phys.*, 145:174103, 2016.
- <sup>44</sup>P. Kolta, H. Wu, F. Noé, and C. Schütte. Optimal data-driven estimation of generalized markov state models for non-equilibrium dynamics. *Computation*, 6:22, 2018.
- <sup>45</sup>S. Kube and M. Weber. A coarse graining method for the identification of transition rates between molecular conformations. *J. Chem. Phys.*, 126:024103, 2007.
- <sup>46</sup>A. Laio and M. Parrinello. Escaping free energy minima. *Proc. Natl. Acad. Sci. USA*, 99:12562–12566, 2002.
- <sup>47</sup>Q. Li, F. Dietrich, E. M. Boltt, and I. G. Kevrekidis. Extended dynamic mode decomposition with dictionary learning: A data-driven adaptive spectral decomposition of the koopman operator. *Chaos*, 27:103111, 2017.
- <sup>48</sup>K. Lindorff-Larsen, S. Piana, R. O. Dror, and D. E. Shaw. How fast-folding proteins fold. *Science*, 334:517–520, 2011.
- <sup>49</sup>F. Litzinger, L. Boninsegna, H. Wu, F. Nüske, R. Patel, R. Baraniuk, F. Noé, and C. Clementi. Rapid calculation of molecular kinetics using compressed sensing. *J. Chem. Theory Comput.*, 24:2771–2783, 2018.
- <sup>50</sup>B. Lusch and S. L. Brunton J . N. Kutz. Deep learning for universal linear embeddings of nonlinear dynamics. arXiv:1712.09707, 2017.
- <sup>51</sup>A. Mardt, L. Pasquali, H. Wu, and F. Noé. Vampnets: Deep learning of molecular kinetics. *Nat. Commun.*, 9:5, 2018.
- <sup>52</sup>R. T. McGibbon and V. S. Pande. Variational cross-validation of slow dynamical modes in molecular kinetics. *J. Chem. Phys.*, 142:124105, 2015.
- <sup>53</sup>P. Metzner, F. Noé, and C. Schütte. Estimation of transition matrix distributions by monte carlo sampling. *Phys. Rev. E*, 80:021106, 2009.
- <sup>54</sup>P. Metzner, C. Schütte, and E. Vanden-Eijnden. Transition Path Theory for Markov Jump Processes. *Multiscale Model. Simul.*, 7:1192–1219, 2009.
- <sup>55</sup>A. S. J. S. Mey, H. Wu, and F. Noé. xTRAM: Estimating equilibrium expectations from time-correlated simulation data at multiple thermodynamic states. *Phys. Rev. X*, 4:041018, 2014.
- <sup>56</sup>L. Molgedey and H. G. Schuster. Separation of a mixture of independent signals using time delayed correlations. *Phys. Rev. Lett.*, 72:3634–3637, 1994.
- <sup>57</sup>F. Noé. Probability Distributions of Molecular Observables computed from Markov Models. *J. Chem. Phys.*, 128:244103, 2008.
- <sup>58</sup>F. Noé and C. Clementi. Kinetic distance and kinetic maps from molecular dynamics simulation. *J. Chem. Theory Comput.*, 11:5002–5011, 2015.
- <sup>59</sup>F. Noé and C. Clementi. Collective variables for the study of long-time kinetics from molecular trajectories: theory

- and methods. *Curr. Opin. Struct. Biol.*, 43:141–147, 2017.
- <sup>60</sup>F. Noé, S. Doose, I. Daidone, M. Löllmann, J. D. Chodera, M. Sauer, and J. C. Smith. Dynamical fingerprints for probing individual relaxation processes in biomolecular dynamics with simulations and kinetic experiments. *Proc. Natl. Acad. Sci. USA*, 108:4822–4827, 2011.
- <sup>61</sup>F. Noé, I. Horenko, C. Schütte, and J. C. Smith. Hierarchical Analysis of Conformational Dynamics in Biomolecules: Transition Networks of Metastable States. *J. Chem. Phys.*, 126:155102, 2007.
- <sup>62</sup>F. Noé and F. Nüske. A variational approach to modeling slow processes in stochastic dynamical systems. *Multiscale Model. Simul.*, 11:635–655, 2013.
- <sup>63</sup>F. Noé, C. Schütte, E. Vanden-Eijnden, L. Reich, and T. R. Weikl. Constructing the full ensemble of folding pathways from short off-equilibrium simulations. *Proc. Natl. Acad. Sci. USA*, 106:19011–19016, 2009.
- <sup>64</sup>F. Noé, H. Wu, J.-H. Prinz, and N. Plattner. Projected and hidden markov models for calculating kinetics and metastable states of complex molecules. *J. Chem. Phys.*, 139:184114, 2013.
- <sup>65</sup>Frank Noé and Hao Wu. Boltzmann generators - sampling equilibrium states of many-body systems with deep learning. *arXiv:1812.01729*, 2018.
- <sup>66</sup>F. Nüske, B. G. Keller, G. Pérez-Hernández, A. S. J. S. Mey, and F. Noé. Variational approach to molecular kinetics. *J. Chem. Theory Comput.*, 10:1739–1752, 2014.
- <sup>67</sup>S. Olsson, H. Wu, F. Paul, C. Clementi, and F. Noé. Combining experimental and simulation data of molecular processes via augmented markov models. *Proc. Natl. Acad. Sci. USA*, 114:8265–8270, 2017.
- <sup>68</sup>S. Orioli and P. Faccioli. Dimensional reduction of markov state models from renormalization group theory. *J. Chem. Phys.*, 145:124120, 2016.
- <sup>69</sup>S. E. Otto and C. W. Rowley. Linearly-recurrent autoencoder networks for learning dynamics. *arXiv:1712.01378*, 2017.
- <sup>70</sup>F. Paul, C. Wehmeyer, E. T. Abualrous, H. Wu, M. D. Crabtree, J. Schöneberg, J. Clarke, C. Freund, T. R. Weikl, and F. Noé. Protein-ligand kinetics on the seconds timescale from atomistic simulations. *Nat. Commun.*, 8:1095, 2017.
- <sup>71</sup>F. Paul, H. Wu, M. Vossel, B. L. de Groot, and F. Noé. Identification of kinetic order parameters for non-equilibrium dynamics. *arXiv:1811.12551*, 2018.
- <sup>72</sup>G. Perez-Hernandez and F. Noé. Hierarchical time-lagged independent component analysis: computing slow modes and reaction coordinates for large molecular systems. *J. Chem. Theory Comput.*, 12:6118–6129, 2016.
- <sup>73</sup>G. Perez-Hernandez, F. Paul, T. Giorgino, G. D Fabritiis, and Frank Noé. Identification of slow molecular order parameters for markov model construction. *J. Chem. Phys.*, 139:015102, 2013.
- <sup>74</sup>N. Plattner, S. Doerr, G. De Fabritiis, and F. Noé. Protein-protein association and binding mechanism resolved in atomic detail. *Nat. Chem.*, 9:1005–1011, 2017.
- <sup>75</sup>J. Preto and C. Clementi. Fast recovery of free energy landscapes via diffusion-map-directed molecular dynamics. *Phys. Chem. Chem. Phys.*, 16:19181–19191, 2014.
- <sup>76</sup>J.-H. Prinz, H. Wu, M. Sarich, B. G. Keller, M. Senne, M. Held, J. D. Chodera, C. Schütte, and F. Noé. Markov models of molecular kinetics: Generation and validation. *J. Chem. Phys.*, 134:174105, 2011.
- <sup>77</sup>João Marcelo Lamim Ribeiro, Pablo Bravo, Yihang Wang, and Pratyush Tiwary. Reweighted autoencoded variational bayes for enhanced sampling (rave). *J. Chem. Phys.*, 149:072301, 2018.
- <sup>78</sup>S. Röblitz and M. Weber. Fuzzy spectral clustering by PCCA+: application to Markov state models and data classification. *Adv. Data Anal. Classif.*, 7:147–179, 2013.
- <sup>79</sup>M. A. Rohrdanz, W. Zheng, M. Maggioni, and C. Clementi. Determination of reaction coordinates via locally scaled diffusion map. *J. Chem. Phys.*, 134:124116, 2011.
- <sup>80</sup>E. Rosta and G. Hummer. Free energies from dynamic weighted histogram analysis using unbiased markov state model. *J. Chem. Theory Comput.*, 11:276–285, 2015.
- <sup>81</sup>C. W. Rowley, I. Mezić, S. Bagheri, P. Schlatter, and D. S. Henningson. Spectral analysis of nonlinear flows. *J. Fluid Mech.*, 641:115, nov 2009.
- <sup>82</sup>M. Rupp, A. Tkatchenko, K.-R. Müller, and O. A. Von Lilienfeld. Fast and accurate modeling of molecular atomization energies with machine learning. *Phys. Rev. Lett.*, 108:058301, 2012.
- <sup>83</sup>M. Sarich, F. Noé, and C. Schütte. On the approximation quality of markov state models. *Multiscale Model. Simul.*, 8:1154–1177, 2010.
- <sup>84</sup>J. Schäfer and K. Strimmer. *Statistical Applications in Genetics and Molecular Biology*, volume 4, chapter A Shrinkage Approach to Large-Scale Covariance Matrix Estimation and Implications for Functional Genomics, pages 2194–6302. Walter de Gruyter GmbH & Co. KG, Berlin/Boston, 2005.
- <sup>85</sup>M. K. Scherer, B. E. Husic, M. Hoffmann, F. Paul, H. Wu, and F. Noé. Variational selection of features for molecular kinetics. *arXiv:1811.11714*, 2018.
- <sup>86</sup>M. K. Scherer, B. Trendelkamp-Schroer, F. Paul, G. Perez-Hernandez, M. Hoffmann, N. Plattner, C. Wehmeyer, J.-H. Prinz, and F. Noé. PyEMMA 2: A software package for estimation, validation and analysis of Markov models. *J. Chem. Theory Comput.*, 11:5525–5542, 2015.
- <sup>87</sup>P. J. Schmid. Dynamic mode decomposition of numerical and experimental data. *J. Fluid Mech.*, 656:5–28, jul 2010.
- <sup>88</sup>P. J. Schmid and J. Sesterhenn. Dynamic mode decomposition of numerical and experimental data. In *61st Annual Meeting of the APS Division of Fluid Dynamics. American Physical Society*, 2008.
- <sup>89</sup>Bernhard Schölkopf, Alexander Smola, and Klaus-Robert Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10:1299–1319, 1998.
- <sup>90</sup>K. T. Schütt, Farhad Arbabzadah, Stefan Chmiela, Klaus R. Müller, and Alexandre Tkatchenko. Quantum-chemical insights from deep tensor neural networks. *Nat. Commun.*, 8:13890, 2017.
- <sup>91</sup>K. T. Schütt, P.-J. Kindermans, H. E. Sauceda, S. Chmiela, A. Tkatchenko, and K.-R. Müller. Schnet: A continuous-filter convolutional neural network for modeling quantum interactions. *arXiv:1706.08566*, 2017.
- <sup>92</sup>C. Schütte, A. Fischer, W. Huisenga, and P. Deufhard. A Direct Approach to Conformational Dynamics based on Hybrid Monte Carlo. *J. Comput. Phys.*, 151:146–168, 1999.
- <sup>93</sup>C. R. Schwantes and V. S. Pande. Improvements in markov state model construction reveal many non-native inter-

- actions in the folding of ntl9. *J. Chem. Theory Comput.*, 9:2000–2009, 2013.
- <sup>94</sup>C. R. Schwantes and V. S. Pande. Modeling molecular kinetics with tica and the kernel trick. *J. Chem. Theory Comput.*, 11:600–608, 2015.
- <sup>95</sup>M. Senne, B. Trendelkamp-Schroer, A. S. J. S. Mey, C. Schütte, and F. Noé. EMMA - A software package for Markov model building and analysis. *J. Chem. Theory Comput.*, 8:2223–2238, 2012.
- <sup>96</sup>David E. Shaw, J.P. Grossman, Joseph A. Bank, Brannon Batson, J. Adam Butts, Jack C. Chao, Martin M. Deneroff, Ron O. Dror, Amos Even, Christopher H. Fenton, Anthony Forte, Joseph Gagliardo, Gennette Gill, Brian Greskamp, C. Richard Ho, Douglas J. Ierardi, Lev Iserovich, Jeffrey S. Kuskin, Richard H. Larson, Timothy Layman, Li-Siang Lee, Adam K. Lerer, Chester Li, Daniel Killebrew, Kenneth M. Mackenzie, Shark Yeuk-Hai Mok, Mark A. Moraes, Rolf Mueller, Lawrence J. Nociolo, and Jon L. Petricolas. Anton 2: Raising the bar for performance and programmability in a special-purpose molecular dynamics supercomputer. In *SC '14: Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*. IEEE, 2014.
- <sup>97</sup>F. K. Sheong, D.-A. Silva, L. Meng, Y. Zhao, and X. Huang. Automatic State Partitioning for Multibody Systems (APM): An Efficient Algorithm for Constructing Markov State Models To Elucidate Conformational Dynamics of Multibody Systems. *J. Chem. Theory Comput.*, 11:17–27, 2015.
- <sup>98</sup>Michael R. Shirts and John D. Chodera. Statistically optimal analysis of samples from multiple equilibrium states. *J. Chem. Phys.*, 129:124105, 2008.
- <sup>99</sup>N. Singhal and V. S. Pande. Error analysis and efficient sampling in Markovian state models for molecular dynamics. *J. Chem. Phys.*, 123:204909, 2005.
- <sup>100</sup>W. C. Swope, J. W. Pitera, and F. Suits. Describing protein folding kinetics by molecular dynamics simulations: 1. Theory. *J. Phys. Chem. B*, 108:6571–6581, 2004.
- <sup>101</sup>W. C. Swope, J. W. Pitera, F. Suits, M. Pitman, and M. Eleftheriou. Describing protein folding kinetics by molecular dynamics simulations: 2. Example applications to alanine dipeptide and beta-hairpin peptide. *Journal of Physical Chemistry B*, 108:6582–6594, 2004.
- <sup>102</sup>G. Székely and M. Rizzo. Testing for equal distributions in high dimension. *InterStat*, 5, 2004.
- <sup>103</sup>J. B. Tenenbaum, V. de Silva, and J. C. Langford. A Global Geometric Framework for Nonlinear Dimensionality Reduction. *Science*, 290:2319–2323, 2000.
- <sup>104</sup>Douglas L. Theobald. Rapid calculation of RMSDs using a quaternion-based characteristic polynomial. *Acta Cryst.*, A61:478–480, 2005.
- <sup>105</sup>G. M. Torrie and J. P. Valleau. Nonphysical Sampling Distributions in Monte Carlo Free-Energy Estimation: Umbrella Sampling. *J. Comp. Phys.*, 23:187–199, 1977.
- <sup>106</sup>B. Trendelkamp-Schroer and F. Noé. Efficient bayesian estimation of markov model transition matrices with given stationary distribution. *J. Phys. Chem.*, 138:164113., 2013.
- <sup>107</sup>B. Trendelkamp-Schroer and F. Noé. Efficient estimation of rare-event kinetics. *Phys. Rev. X (in press)*, preprint at [arXiv:1409.6439](https://arxiv.org/abs/1409.6439), 2015.
- <sup>108</sup>B. Trendelkamp-Schroer, H. Wu, F. Paul, and F. Noé. Estimation and uncertainty of reversible markov models. *J. Chem. Phys.*, 143:174101, 2015.
- <sup>109</sup>J. H. Tu, C. W. Rowley, D. M. Luchtenburg, S. L. Brunton, and J. N. Kutz. On dynamic mode decomposition: Theory and applications. *J. Comput. Dyn.*, 1(2):391–421, dec 2014.
- <sup>110</sup>O. Valsson and M. Parrinello. Variational approach to enhanced sampling and free energy calculations. *Phys. Rev. Lett.*, 113:090601, 2014.
- <sup>111</sup>Jiang Wang, Christoph Wehmeyer, Frank Noé, and Cecilia Clementi. Machine learning of coarse-grained molecular dynamics force fields. [arXiv:1812.01736](https://arxiv.org/abs/1812.01736), 2018.
- <sup>112</sup>Wujie Wang and Rafael Gómez-Bombarelli. Variational coarse-graining for molecular dynamics. [arXiv:1812.02706](https://arxiv.org/abs/1812.02706), 2018.
- <sup>113</sup>C. Wehmeyer and F. Noé. Time-lagged autoencoders: Deep learning of slow collective variables for molecular kinetics. *J. Chem. Phys.*, 148:241703, 2018.
- <sup>114</sup>M. O. Williams, I. G. Kevrekidis, and C. W. Rowley. A data-driven approximation of the koopman operator: Extending dynamic mode decomposition. *J. Nonlinear Sci.*, 25:1307–1346, 2015.
- <sup>115</sup>M. O. Williams, C. W. Rowley, and I. G. Kevrekidis. A kernel-based approach to data-driven koopman spectral analysis. [arXiv:1411.2260](https://arxiv.org/abs/1411.2260), 2014.
- <sup>116</sup>Wojciech Wojtas-Niziurski, Yilin Meng, Benoit Roux, and Simon Bernèche. Self-learning adaptive umbrella sampling method for the determination of free energy landscapes in multiple dimensions. *J. Chem. Theory Comput.*, 9:1885–1895, 2013.
- <sup>117</sup>H. Wu, A. Mardt, L. Pasquali, and F. Noé. Deep generative markov state models. *NIPS (in press)*. Preprint: [arXiv:1805.07601](https://arxiv.org/abs/1805.07601), 2018.
- <sup>118</sup>H. Wu, A. S. J. S. Mey, E. Rosta, and F. Noé. Statistically optimal analysis of state-discretized trajectory data from multiple thermodynamic states. *J. Chem. Phys.*, 141:214106, 2014.
- <sup>119</sup>H. Wu and F. Noé. Optimal estimation of free energies and stationary densities from multiple biased simulations. *Multiscale Model. Simul.*, 12:25–54, 2014.
- <sup>120</sup>H. Wu and F. Noé. Variational approach for learning markov processes from time series data. [arXiv:1707.04659](https://arxiv.org/abs/1707.04659), 2017.
- <sup>121</sup>H. Wu, F. Nüske, F. Paul, S. Klus, P. Kolta, and F. Noé. Variational koopman models: slow collective variables and molecular kinetics from short off-equilibrium simulations. *J. Chem. Phys.*, 146:154104, 2017.
- <sup>122</sup>H. Wu, F. Paul, C. Wehmeyer, and F. Noé. Multiensemble markov models of molecular thermodynamics and kinetics. *Proc. Natl. Acad. Sci. USA*, 113:E3221–E3230, 2016.
- <sup>123</sup>Y. Yao, R. Z. Cui, G. R. Bowman, D.-A. Silva, J. Sun, and X. Huang. Hierarchical nyström methods for constructing markov state models for conformational dynamics. *J. Chem. Phys.*, 138:174106, 2013.
- <sup>124</sup>Yuan Yao, Jian Sun, Xuhui Huang, Gregory R. Bowman, Gurjeet Singh, Michael Lesnick, Leonidas J. Guibas, Vijay S. Pande, and Gunnar Carlsson. Topological methods for exploring low-density states in biomolecular folding pathways. *J. Chem. Phys.*, 130:144115, 2009.

- <sup>125</sup>Linfeng Zhang, Jiequn Han, Han Wang, Roberto Car, and Weinan E. Deepcg: constructing coarse-grained models via deep neural networks. *J. Chem. Phys.*, 149:034101, 2018.
- <sup>126</sup>A. Ziehe and K.-R. Müller. TDSEP — an efficient algorithm for blind separation using time structure. In *ICANN 98*, pages 675–680. Springer Science and Business Media, 1998.
- <sup>127</sup>Maxwell I. Zimmerman and Gregory R. Bowman. Fast conformational searches by balancing exploration/exploitation trade-offs. *J. Chem. Theory Comput.*, 11:5747–5757, 2015.