Physical Biology

OPEN ACCESS

RECEIVED
9 June 2022

REVISED
26 July 2022

ACCEPTED FOR PUBLICATION
9 August 2022

PUBLISHED
30 August 2022

PAPER

Nonparametric Bayesian inference for meta-stable conformational dynamics

Lukas Köhs[1], Kerri Kukovetz[2], Oliver Rauh[2] and Heinz Koeppl[1,*]

[1] Centre for Synthetic Biology, Rundeturmstraße 12, Technische Universität Darmstadt, 64283 Darmstadt, Germany
[2] Department of Biology, Schnittspahnstraße 3, Technische Universität Darmstadt, 64287 Darmstadt, Germany
[*] Author to whom any correspondence should be addressed.

E-mail: heinz.koeppl@tu-darmstadt.de

Keywords: Bayesian nonparametrics, conformational switching, molecular dynamics, variational inference

Supplementary material for this article is available online

Abstract

Analyses of structural dynamics of biomolecules hold great promise to deepen the understanding of and ability to construct complex molecular systems. To this end, both experimental and computational means are available, such as fluorescence quenching experiments or molecular dynamics simulations, respectively. We argue that while seemingly disparate, both fields of study have to deal with the same type of data about the same underlying phenomenon of conformational switching. Two central challenges typically arise in both contexts: (i) the amount of obtained data is large, and (ii) it is often unknown how many distinct molecular states underlie these data. In this study, we build on the established idea of Markov state modeling and propose a generative, Bayesian nonparametric hidden Markov state model that addresses these challenges. Utilizing hierarchical Dirichlet processes, we treat different meta-stable molecule conformations as distinct Markov states, the number of which we then do not have to set *a priori*. In contrast to existing approaches to both experimental as well as simulation data that are based on the same idea, we leverage a mean-field variational inference approach, enabling scalable inference on large amounts of data. Furthermore, we specify the model also for the important case of angular data, which however proves to be computationally intractable. Addressing this issue, we propose a computationally tractable approximation to the angular model. We demonstrate the method on synthetic ground truth data and apply it to known benchmark problems as well as electrophysiological experimental data from a conformation-switching ion channel to highlight its practical utility.

1. Introduction

Computational prediction of molecular structures is a long-standing challenge in structural biology. Recently, novel frameworks for protein [1] and ribonucleic acid (RNA) structure prediction [2] have been proposed that greatly enhance the prediction accuracy and the model applicability compared to existing approaches. The focus of structure prediction however lies on *static* structures. To obtain deeper insights into molecular processes, it is necessary to complement such approaches with frameworks that also take into account the *dynamics* of molecular structure.

Both experimental and computational tools exist to study such conformational dynamics: a prime example of experimental approaches to this challenge are analyses of conformational switching of ion channels via widely adopted electrophysiological techniques, such as voltage clamping or lipid bilayer measurements [3–5]. On the other hand, ion channel switching can also be studied computationally via, e.g., molecular dynamics (MD) simulations [6]. The MD framework more generally can also be used to investigate protein and RNA folding [7, 8]. Experimentally, this can be assessed, e.g., via fluorescence quenching [9] or fluorescence resonance energy transfer measurements [10]. Since in all of these settings, both experimental and computational protocols typically yield large amounts of data, their analysis is a challenge in itself.

To understand the switching dynamics of the system under study, one needs to obtain a coarse-grained description of the *continuous* system dynamics (e.g., voltages or atom coordinates) in terms of comparatively long-lived, meta-stable *discrete* states corresponding to distinct stable structural conformations [11]. These are defined by a separation of time scales between the intra- and inter-state dynamics. In the context of MD, particularly, the framework of Markov state models (MSMs) has received much attention in recent years; besides extensive methodological research [12–15], MSMs have also been successfully applied to a wide range of use cases [16–23]. The classical MSM approach approximates the continuous simulation dynamics (mathematically expressible in terms of the *propagation operator*) directly by projecting them to a discretized space on which then a discrete-time Markov chain (DTMC) is constructed. Importantly, the binning of continuous data into discrete states introduces correlations and makes the resulting discrete process generally non-Markovian; the raw MD data, in contrast, are inherently Markovian, as they originate from the integration of stochastic differential equations (SDE).

To render the resulting discrete dynamics amenable to MSM analysis, the correlations need to be reduced via temporal thinning by some *lag time* constant [13, 24]. This results in two key parameters that need manual selection: the lag time and the state-space discretization. An explicit error bound for the reconstruction error of the MSM reconstruction and the true propagation operator can be derived in terms of these two parameters, showing that this error can be made arbitrarily small by either choosing a finer state-space discretization or decreasing the lag time [13]. This relationship between state-space discretization and lag time results in a trade-off problem: one has to balance between (i) sufficient sampling of the discrete state-space, viz, a coarse state-space discretization and (ii) a sufficiently long lag time to render the resulting process Markovian. To address this issue, tools such as the Chapman–Kolmogorov test have been introduced [13]. These tests, however, only consider the appropriateness of the lag time; the overall reconstruction error may still be off, resulting in an MSM not reproducing the long-time dynamics accurately, as detailed in [25]. Also, the lag time selection itself is acknowledged to be a major challenge in practice [26].

The identification of meta-stable conformational states of the system is then carried out after data pre-processing given some state-space discretization and lag time. Typically, this is done utilizing spectral methods such as Perron-cluster cluster analysis (PCCA) [27] or PCCA+ [28]; other approaches are however also possible, see e.g., [29, 30]. In general, the result of this procedure will depend on the chosen lag time as well as the state-space discretization.

While the framework as such has seen several refinements [31, 32] including recent deep learning extensions [33, 34], they share the two conceptual drawbacks detailed above: (i) the data needs to be thinned with a manually specified lag time, which is merely a model artifact and deteriorates the overall temporal resolution, and (ii) the number of metastable states has to be identified manually.

Both problems can be addressed utilizing nonparametric hidden Markov model (HMM) frameworks developed in statistical machine learning [36, 42, 43]: on the one hand, the introduction of a latent process to an MSM abolishes the need to temporally de-correlate the discrete projection of the data via a lag time [31], which can thus be interpreted as a generalization of classical MSMs. On the other hand, *nonparametric* probabilistic models allow one to specify distributions on unbounded spaces, such as an infinite number of topics in topic modeling [35] or an infinite number of states in a Markov chain [36].

Nonparametric Bayesian HMMs have gained attention in recent years both in experimental settings such as analyses of ion channel switching [37, 38] or single-particle tracking [39] as well as in MD studies [40, 41]. Inference of the meta-stable trajectories and the system parameters was however carried out using sampling techniques such as Markov chain Monte Carlo (MCMC); while yielding accurate results, these approaches do not scale well [42, 43]. Even for the relatively simple problem of one-dimensional ion channel voltage trajectories, they become computationally intractable for longer sequences or higher-dimensional systems.

To address both the conceptual shortcomings of MSMs and the computational tractability issues of conventional sampling approaches, we provide in this paper a scalable nonparametric Bayesian MSM inference framework for the analysis of conformational molecule dynamics. We emphasize that this framework is very widely applicable, including data generated, e.g., by voltage clamp experiments on ion channels as well as MD simulations, and can hence help bridge the gap between theory and experiment. We note also that in terms of modeling, the transition between ion channel experiments and MD simulations is gradual, as the measured ion current in the former can be interpreted as a one-dimensional reaction coordinate in the latter.

Our method does neither require manual specification of a lag time to re-establish Markovianity, nor of the number of meta-stable conformations. We model the switching dynamics between distinct structural conformations via a nonparametric HMM by defining a latent Markov process on a countable set of states (meta-stable protein or channel conformations), of which one obtains only noisy, continuous-valued observations, such as currents or atom positions. We specify noise models that are appropriate for our use cases: in the experimental and computational settings described above, observations typically are real-valued vectors, $x \in \mathbb{R}^n$, or rotation angles, $x \in [0, 2\pi)^n$. For the angular case, we furthermore propose a novel approximation enabling computational tractability.

To ensure scalability to large amounts of data, we resort to variational methods for inference: instead of drawing samples from the exact posterior distribution, we approximate the latter in a computationally efficient way by distributions of known type [44–46]. As we pursue a Bayesian approach, we treat the model parameters as random variables, specifying appropriate prior distributions and deriving their full posterior distributions conditioned on all observed data.

In the following, we will first introduce the general modeling framework and we will in particular address the issue of adequate prior distributions. Subsequently, we show how to perform scalable inference in this setting. Finally, we present our results on synthetic ground-truth data as well as real benchmark and experimental data.

2. Methods

2.1. Bayesian nonparametric Markov state models

We model the conformational molecule dynamics by utilizing the well-known HMM, consisting of two joint stochastic processes $\{Z_t, X_t : t = 1, \dots, T\}$, where t is the time index [47]. Note that we use roman upper case letters Z_t, X_t to refer to random variables and the corresponding lower case letters z_t, x_t to refer to particular realizations throughout the paper.

The distinct meta-stable conformational states are represented by the latent Markov states $Z_t \in \mathcal{Z} \subseteq \mathbb{N}$; the observed data (e.g., experimentally obtained channel voltages or simulated atom positions) are described by $X_t \in \mathbb{R}^n$. The time evolution of this joint process is given as a DTMC on the discrete state space $\mathcal{Z}$ governed by a transition probability function $\Pi : \mathcal{Z} \times \mathcal{Z} \to [0, 1]$. We represent this as a matrix $\Pi \in [0, 1]^{n \times n}$, whose kth row $\pi_k := \Pi_{k\cdot}$ specifies the probabilities for transitions to all possible states $l \in \mathcal{Z}$ from state $k \in \mathcal{Z}$,

$$\pi_{kl} = \mathsf{P}(Z_t = l | Z_{t-1} = k) =: p(l | k, \Pi). \qquad (1)$$

The observation X_t at time point t depends only on the state of the latent process at the same time. This dependency is given by the observation density

$$p(x_t | Z_t = z_t, \{\theta_1, \dots, \theta_{|\mathcal{Z}|}\}) = p(x_t | \theta_{z_t}), \qquad (2)$$

where $\{\theta_i : i = 1, \dots, |\mathcal{Z}|\}$ represents a set of generic distribution parameters for each state i. In accordance with the MSM literature, we interpret the HMM as a generalization of MSMs [31], and thus refer to this construct synonymously as *hidden* MSM.

The key drawback of hidden MSMs regarding the analysis of conformational switching is that the number of meta-stable molecule conformations $|\mathcal{Z}|$ needs to be specified in advance. Typically, however, this number is unknown. Quite on the contrary, it is a key quantity of interest that is to be determined from the data. This shortcoming can be addressed by utilizing a nonparametric modeling approach, which allows for countably infinite state spaces. For any finite data set, $|\mathcal{Z}|$ will however be finite and can hence be learned from the data. In more concrete terms, we specify a model for potentially infinitely many distinct molecular conformations; in any given observed trajectory from simulations or experiments, only a finite number of these conformations will be visited, the number of which can then be identified.

To set up a nonparametric HMM, one needs to construct prior distributions for transition matrices on countably infinite state spaces and for countably infinite observation parameters. This can be achieved via the hierarchical Dirichlet process (HDP) [48], giving rise to the HDP-HMM [36, 42, 43, 45]. In the following, we provide mainly the relevant definitions; we however provide an extended background section on Bayesian nonparametrics in the supporting information (https://stacks.iop.org/PB/19/056006/mmedia). An HDP-HMM is constructed hierarchically in a two-step fashion:

First, specify a Dirichlet process (DP), which is a stochastic process taking values in the space of (discrete) probability measures:

$$H_1 \sim \mathrm{DP}\,(\alpha, H_0) \qquad (3)$$

with concentration parameter $\alpha > 0$ and base probability measure H_0 over some space Θ.

A realization of H_1 is obtained by drawing independent and identically distributed (i.i.d.) samples $\theta_k \in \Theta$ from the base measure H_0,

$$\theta_k \overset{\text{i.i.d.}}{\sim} H_0 \quad \text{for } k = 1, 2, \dots, \qquad (4)$$

and assigning to each θ_k a probability mass σ_k via a *stick-breaking process*:

$$\begin{aligned} \epsilon_k &\overset{\text{i.i.d.}}{\sim} \mathrm{Beta}\,(1, \alpha) \quad \text{for } k = 1, 2, \dots, \\ \sigma_k &= \epsilon_k \prod_{j=1}^{k-1} (1 - \epsilon_j), \end{aligned} \qquad (5)$$

where $\mathrm{Beta}(r, s)$ is the beta distribution with shape parameters $r, s > 0$ [47]. Equation (5) is compactly written as $\sigma \sim \mathrm{GEM}(\alpha)$, short for *Griffiths–Engen–McCloskey* process [48]. A sample $H_1 \sim \mathrm{DP}(\alpha, H_0)$ accordingly reads

$$H_1 = \sum_{k=1}^{\infty} \sigma_k \delta_{\theta_k}, \qquad (6)$$

where δ_{θ_k} denotes the Dirac or point measure at θ_k [49], $\delta_{\theta_k}(\theta) = 1$ if $\theta = \theta_k$ and 0 otherwise. This procedure results in valid discrete probability measure, $\int \mathrm{d}H_1 = 1$, determining a prior distribution over conformational states: each k represents one distinct conformation, with σ_k its probability and θ_k its associated parameterization.

In the second step, H_1 serves as base measure of another, subordinate DP: because H_1 is discrete, all samples drawn from this subordinate DP have shared support. We consider independent draws

$$\pi_k \sim \mathrm{DP}\left(\beta + \xi, \frac{\beta H_1 + \xi \delta_{\theta_k}}{\beta + \xi}\right), \tag{7}$$

with the *stickiness* parameter $\xi \geqslant 0$, on which we will elaborate in the next paragraph. Decomposing

$$\begin{aligned} &\eta_k \sim \mathrm{GEM}\,(\beta + \xi), \\ &\varphi_{k,j} \overset{\text{i.i.d.}}{\sim} \frac{\beta\sigma + \xi\delta_k}{\beta + \xi}, \end{aligned} \tag{8}$$

with the point measure at index k yields

$$\pi_k = \sum_{j=1}^{\infty} \eta_{k,j} \delta_{\theta_{\varphi_{k,j}}}. \tag{9}$$

As the support of all π_k are the shared atoms $\{\theta_1, \theta_2, \ldots\}$ drawn in equation (4), each π_k can be understood as a realization from a probability distribution over a row of a 'countably infinite transition matrix' Π:

$$Z_t | Z_{t-1} = z_{t-1}, \{\pi_k\} \sim \pi_{z_{t-1}}.$$

Each element of the set $\theta_k \in \{\theta_1, \theta_2, \ldots\}$ corresponds to one latent state k, that is, one molecular conformation, and parameterizes the respective observation distribution,

$$p(x_t | Z_t = z_t, \{\theta_1, \theta_2, \ldots\}) = p(x_t | \theta_{z_t}).$$

In other words, the two-step HDP-HMM construction (i) defines the molecular conformations via the measure H_1, and (ii) determines their transition dynamics via all π_k.

The stickiness parameter introduces a self-transition bias, that is, it extends the sojourn times within each state. For $\xi = 0$, the classical HDP is recovered [48]. The sticky HDP-HMM has been shown to counter-balance the sensitivity of the classical HDP-HMM to within-state variability, which results in a tendency to introduce redundant states all pertaining to the same ground-truth state; see, e.g., [42]. This is exactly the setting we are interested in, as we are aiming specifically at the analysis of meta-stable states potentially exhibiting a high level of intra-state variability. Note that in comparison to classical MSMs, the stickiness parameter can be understood as a bias towards larger time scales that is to be set by the experimenter. Importantly, our approach does not discard any information, but retains all available data points; the stickiness represents only a bias, but no strict truncation of resolvable time scales. Hence, the minimum time scale this approach is able to resolve is the native time scale of the data points. As with all hyperparameters, we set ξ empirically; see the supporting information for details.

The above definition allows us to formulate the full model distribution. Denoting with $x^i_{[1,T]} := \{x^i_1, \ldots, x^i_T\}$ and $z^i_{[1,T]} := \{z^i_1, \ldots, z^i_T\}$ the ith of a total of I observed trajectories and with $x_{[1,T]} := \{x^1_{[1,T]}, \ldots, x^I_{[1,T]}\}$, $z_{[1,T]} := \{z^1_{[1,T]}, \ldots, z^I_{[1,T]}\}$, the collection of all trajectories, we can write

$$\begin{aligned} &p(x_{[1,T]}, z_{[1,T]}, \theta, \Pi, \sigma) \\ &\quad = p(\sigma) \prod_{k=1}^{|\mathcal{Z}|} p(\pi_k|\sigma) p(\theta_k) \prod_{i}^{I} p(z^i_1) p(x^i_1 | z^i_1, \theta) \\ &\quad \times \prod_{t=2}^{T} p(z^i_t | z^i_{t-1}, \Pi) p(x^i_t | z^i_t, \theta), \end{aligned} \tag{10}$$

with some initial distributions $p(z^i_1)$, constituting a fully Bayesian nonparametric HMM.

2.2. Observation models and conjugate priors

To fully specify the HDP-HMM, it remains to set up the observation distributions $p(x_t|\theta_{z_t})$ for the required spaces $x \in \mathbb{R}^n$ and $x \in [0, 2\pi)^n$ as well as the corresponding prior distributions H_0. For the purpose of inference, it is beneficial to choose priors that are *conjugate* to the respective likelihoods: a prior of a specific functional form f parameterized by γ, $p(\theta|\gamma) =: f(\theta, \gamma)$, is said to be conjugate to a given conditional probability distribution $p(x|\theta)$ if the resulting Bayesian posterior distribution $p(\theta|x) = p(x|\theta)p(\theta)/p(x)$ is of the same functional form as the prior, $p(\theta|x) = f(\theta, \gamma')$, with updated parameters γ'. This property simplifies inference, because the computation of the posterior distributions then reduces to computing the parameter updates $\gamma \to \gamma'$.

Real-valued data. A versatile model for coordinate data $x \in \mathbb{R}^n$ as often obtained through MD (e.g., 3D atom positions) as well as electrophysiological experiments is the multivariate normal (MVN) distribution,

$$p(x_t|\theta) = \mathrm{N}\,(x_t|\mu, \Sigma), \tag{11}$$

with the mean vector $\mu \in \mathbb{R}^n$ and the covariance matrix $\Sigma \in \mathbb{R}^{n \times n}$. Generally, we interpret the raw data as noisy observations of the discrete latent conformational states. The MVN is well suited for this purpose due to its unimodality as we aim to identify well-discernible, meta-stable states. For ion channel voltage data, typically $x_t \in \mathbb{R}$, which is covered by equation (11) as a special case $n = 1$. Note that normal observation models are frequently used for biophysical experiments [37, 38, 40]. We can hence cover both MD simulation data as well as experimental voltage trajectory data with the same observation model.

For the MVN, a conjugate prior exists termed the *normal inverse-Wishart* (NIW) distribution, which defines a joint distribution over means μ and covariances Σ:

$$\mathrm{NIW}\,(\mu, \Sigma|\mu_0, \lambda, \Psi, \nu) = \mathrm{N}\,\left(\mu|\mu_0, \Sigma/\lambda\right)\mathrm{IW}\,(\Sigma|\Psi, \nu), \tag{12}$$

with the inverse Wishart (IW) distribution [47]. We combine the likelihood equation (11) with the conjugate prior equation (12) to specify the HDP-HMM for real-valued data.

Angular data. Another natural and widely used way in MD of specifying the spatial arrangement of complex molecules are the dihedral angles between adjacent atom or molecule planes [50]. Hence, data are often angular and constrained to the unit circle, $x_t \in [0, 2\pi)^n$. Observation models particularly suited for these spaces are von Mises (vM) type distributions [51, 52]. Since it is customary to characterize amino acid chains such as proteins by sets of *pairs* of torsion angles (ϕ_i, ψ_i), we focus on the two-dimensional case here; the extension to multiple pairs is then straightforward. We utilize a well-known and in the context of protein modeling established parameterization of the bivariate vM distribution [53] for $x_t \equiv (\phi, \psi) \in [0, 2\pi)^2$,

$$\begin{aligned} p(\phi, \psi|\theta) &= \mathrm{BvM}\,(x_t|\zeta, \nu, \kappa_1, \kappa_2, \kappa_3) \\ &=: c^{-1}(\kappa_1, \kappa_2, \kappa_3)\exp\{\kappa_1\ \cos(\phi-\zeta) \\ &\quad + \kappa_2\ \cos(\psi-\nu) \\ &\quad - \kappa_3\ \cos(\phi-\zeta-\psi+\nu)\}, \end{aligned} \tag{13}$$

where

$$\begin{aligned} c(\kappa_1, \kappa_2, \kappa_3) = (2\pi)^2 \Bigg[& I_0(\kappa_1)I_0(\kappa_2)I_0(\kappa_3) \\ & + 2\sum_{k=1}^{\infty} I_k(\kappa_1)I_k(\kappa_2)I_k(\kappa_3)\Bigg] \end{aligned} \tag{14}$$

and I_i is the modified Bessel function of the first kind and order i. The location parameters ζ and ν control the position of the mode of the distribution, as can be seen from the trigonometric terms in equation (13). The parameters $\kappa_1, \kappa_2, \kappa_3$ specify the spatial correlations. Note that marginalizing over ϕ and setting $\kappa_1 = \kappa_3 = 0$ recovers the conventional one-dimensional vM distribution,

$$p(\psi|\theta) = \frac{\exp\{\kappa_2\ \cos(\psi-\nu)\}}{2\pi I_0(\kappa_2)}. \tag{15}$$

While analytical expressions for a conjugate prior exist also for the bivariate vM distribution [51], the infinite sum of Bessel functions in equation (13) renders the normalizer c intractable in a Bayesian setting. Computing the exact posterior vM is hence not possible, as this requires computing integrals over all κ parameters in equation (14). Additionally, this distribution is not guaranteed to be unimodal; intricate conditions exist on the relation of the concentration parameters $\kappa_1, \kappa_2, \kappa_3$ to achieve unimodality [52]. For high concentration values in specific regimes, however, it is known that the bivariate vM distribution is well approximated by a bivariate normal distribution [51]. This is unsurprising, as in general, vM-type distributions and normal distributions are tightly linked: the former can be constructed from the latter [54]. To ensure tractability and interpretability, we utilize this circumstance and propose an approximation via

$$\mathrm{BvM}\,(x_t|\zeta, \nu, \kappa_1, \kappa_2, \kappa_3) \approx \mathrm{N}\,(x_t|\mu, \Sigma). \tag{16}$$

The mode position ζ, ν roughly corresponds to the mean vector μ; the covariance depends on the κ-parameters. The precise analytical expressions for these dependencies are rather involved and not relevant to our approximation—we hence refer the interested reader to [51, 52] for an in-depth analysis. As we focus on systems exhibiting distinct, separable meta-stable states, we in fact expect peaked angular distributions, for which this approximation is valid. To gain a better intuition, see also figure 1, where we compare a one-dimensional vM with the corresponding normal distribution; as is immediately clear, for high concentration values the approximation error becomes negligible. Additionally, equation (16) allows for straightforward debugging: as long as the probability assigned to the area outside the unit circle is small, the approximation can be assumed valid; vice versa, it deteriorates if this probability becomes non-negligible. We consequentially accept this error in the observation model to arrive at a tractable expression. Note that the gained tractability may greatly aid the practical utility of the framework, as it is otherwise also customary to resort to 3D coordinates to avoid mathematical complexity, disregarding crucial structural information about the biological problem [40]. In the following, we refer to our approximation as the *approximate* vM model.

2.3. Scalable inference of meta-stable states

The goal of Bayesian inference is to compute the posterior distribution over the latent sequences $z_{[1,T]}$ and the model parameters given the observed trajectories $x_{[1,T]}$,

$$p\left(z_{[1,T]}, \theta, \Pi, \sigma|x_{[1,T]}\right) = \frac{p\left(x_{[1,T]}|z_{[1,T]}, \theta\right)p\left(z_{[1,T]}|\Pi\right)p(\theta)p\left(\Pi|\sigma\right)p(\sigma)}{p\left(x_{[1,T]}\right)}. \tag{17}$$

This distribution cannot be evaluated analytically. In principle, one can employ standard sampling techniques such as MCMC and obtain the posterior empirically [36]. In our case, however, the typically large data sets from simulations or long-duration

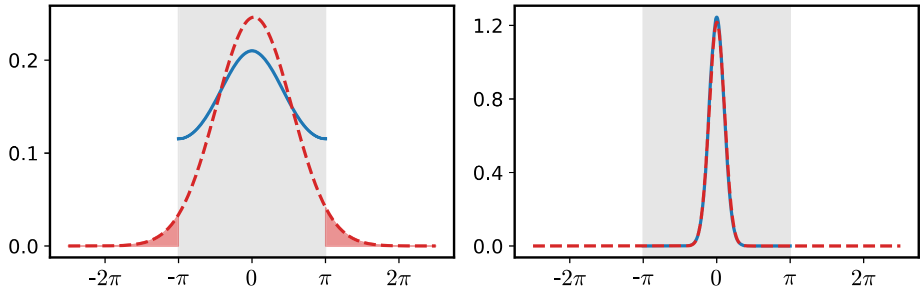


Figure 1. Comparison of a one-dimensional vM with the approximate vM model. Left: inferred approximate vM (red dashed line) for $N = 1000$ data points generated from a low-concentration vM ($\kappa = 0.3$, blue line). Gray shaded area: unit circle. Red shaded area: probability mass outside the unit circle boundaries. Right: inferred approximate vM for $N = 1000$ data points generated from a high-concentration vM $\kappa = 10$. No significant probability mass is placed outside the unit circle.

experiments (see, e.g., the discussion in [37]), render this computationally infeasible, because every draw from the posterior requires one full pass through the data.

To alleviate these computational issues, we utilize a variational inference (VI) approach. The core idea of VI methods is to cast the inference task as an optimization problem. We briefly lay out the general method, but refer the interested reader to, e.g., [44]. The objective is to find an approximate (or *variational*) distribution q^* that minimizes the Kullback–Leibler (KL) divergence to the exact posterior p, equation (17):

$$q^*\left(z_{[1,T]}, \Pi, \theta, \sigma\right) = \mathrm{argmin}_q\, \mathrm{KL}\left(q\left(z_{[1,T]}, \Pi, \theta, \sigma\right) \times \|p\left(z_{[1,T]}, \Pi, \theta, \sigma | x_{[1,T]}\right)\right). \tag{18}$$

This KL divergence is still computationally intractable because of the evidence $p(x_{[1,N]})$ in the denominator of equation (17). The evidence is in principle obtained by integrating over all unobserved model components, $p(x) = \sum_z \int p(x, z, \Pi, \theta, \sigma) \mathrm{d}\theta\, \mathrm{d}\Pi\, \mathrm{d}\sigma$, which requires a summation of $|\mathcal{Z}|^T$ terms, each of which contains the full integrals over the parameters Π, θ and σ and hence is impossible to compute for realistic state space sizes and sequence lengths.

The optimization problem equation (18) can be however transformed into an equivalent, but tractable problem. To do so, one re-writes the KL divergence as

$$\mathrm{KL}\left(q\left(z_{[1,T]}, \Pi, \theta, \sigma\right) \| p\left(z_{[1,T]}, \Pi, \theta, \sigma | x_{[1,T]}\right)\right) = -\mathcal{L} + \log(p(x_{[1,T]})), \tag{19}$$

where

$$\mathcal{L} = \mathsf{E}_q[\log p(z_{[1,T]}, \theta, \Pi, \sigma, x_{[1,T]})] - \mathsf{E}_q[\log q(z_{[1,T]}, \Pi, \theta, \sigma)] \tag{20}$$

and E_q is the expectation operator, where the expectation is to be taken with respect to q. This yields a lower bound on the log evidence, $\mathcal{L} \leqslant \log p(x_{[1,T]})$, since the KL divergence satisfies $\mathrm{KL}(q\|p) \geqslant 0$ for any two distributions q, p. Accordingly, the quantity $\mathcal{L}$ is termed the evidence lower-bound (ELBO). Since the log evidence is constant with respect to the model parameters, *minimizing* the KL divergence is equivalent to *maximizing* the ELBO. The intractable computation of the log evidence is hence not needed to evaluate $\mathcal{L}$.

Without further assumptions, maximization of equation (20) yields the exact posterior, $q^* = p(z_{[1,T]}, \Pi, \theta | x_{[1,T]})$, but does not provide a practical way of actually performing the optimization. To enable a practical computational scheme, we employ a standard mean-field assumption on the variational distributions [44]:

$$q\left(z_{[1,T]}, \Pi, \theta, \sigma\right) = q\left(z_{[1,T]}\right) q(\sigma) \prod_{k=1}^{|\mathcal{Z}|} q(\theta_k) q(\pi_k). \tag{21}$$

This enables a computationally tractable, iterative coordinate-wise ascent optimization procedure [47]: one variational factor of equation (21) is optimized at a time while keeping all others fixed, and one pass through all variational factors constitutes a VI iteration. The generic distribution update for any quantity $\alpha \in \{z_{[1,T]}, \{\theta_k\}_k, \{\pi_k\}_k, \sigma\}$ is obtained as

$$q(\alpha) \propto \exp\left\{\mathsf{E}_{q^{\backslash\alpha}}\left[\ln p(x_{[1,T]}, z_{[1,T]}, \Pi, \theta)\right]\right\}, \tag{22}$$

where $\mathsf{E}_{q^{\backslash\alpha}}$ denotes the expectation with respect to all variational distributions except $q(\alpha)$. Note that while the ELBO is not convex with respect to all variational distributions jointly [44], it is convex with respect to any factor individually [55]. This coordinate-wise ascent algorithm hence converges to a local optimum which in general depends on the initialization of the variational factors. To alleviate this initialization-dependency, we additionally utilize a multi-start approach; we run several instances of the inference algorithm until convergence and then select the one with the maximal ELBO score as the overall optimum.

Since the HDP-HMM specifies distributions over countably infinite objects, VI in this case requires an additional variational parameter. To be able to instantiate the q-distributions, it is necessary to truncate the number of *variational* states to some maximum number K. This number could in principle be set to the number of data points; in practice, for computational reasons one chooses some number which is large compared to the expected number of HMM states [36, 42]. Note that this only affects the variational distributions; the original model equation (10) remains unchanged [56]. We choose the 'direct assignment' truncation method, setting $q(z_t) = 0$ for any $z_t > K$ [36]. The resulting update equations follow in closed form, as will be shown in the following. The variational model then can, but not necessarily does utilize all clusters up to K [57]. This also allows for straightforward debugging, as it is directly apparent whether all K states are occupied. If $q(z_t) > 0$ for all states z_t, one might incur a non-negligible truncation error, as intuitively speaking more states might be needed to explain the data, and a double-check with increased K is due. If however $q(z_t) = 0$ for some states, the variational approximation is expressive enough and will not result in a significant truncation error. Note that the direct assignment scheme can be utilized for automated search algorithms over the truncation depths [58].

We provide the detailed mathematical derivations of all updates as well as the used initializations in the supporting information and state here only the update equations.

Latent state sequence. The marginal probabilities of the sequence of meta-stable states, $q(z_t)$, can be computed by a forward-backward message-passing algorithm [59]. The forward messages α_t and the backward messages β_t are computed as

$$\begin{aligned}\alpha_t(z_t) = \exp\{\mathsf{E}\left[\ln\, p(x_t|\theta_{z_t})\right]\}\sum_{z_{t-1}}\alpha_{t-1}(z_{t-1})\\ \times \exp\{\mathsf{E}\left[\ln\, p(z_t|z_{t-1},\Pi)\right]\},\end{aligned} \tag{23}$$

$$\begin{aligned}\beta_t(z_t) = \sum_{z_{t+1}}\exp\{\mathsf{E}\left[\ln\, p(x_{t+1}|\theta_{z_{t+1}})\right]\}\beta_{t+1}(z_{t+1})\\ \times \exp\{\mathsf{E}\left[\ln\, p(z_{t+1}|z_t,\Pi)\right]\},\end{aligned} \tag{24}$$

and yield the marginals via $q(z_t) \propto \alpha_t(z_t)\beta_t(z_t)$. The expectations occurring in these expressions can be evaluated in closed form because of conjugacy between the variational distributions over Π and $\{\theta_1,\ldots,\theta_K\}$ and the corresponding likelihoods. Note that the forward and backward messages are the only part of the framework that accepts the trajectory data x to be analyzed as input. Due to the mean-field assumption, the remaining variational distribution updates do not require x.

Transition distributions. The DP can be shown to be conjugate to the (infinite) categorical distributions defined by a DP-draw, equation (9) [48]. Note that this is completely analogous to the finite case, where the Dirichlet distribution is a conjugate prior for the categorical distribution. As we constrain the variational posterior to a maximum of K states, this induces a partition on the base measure space Θ of equation (9). This allows one to write the prior as a finite Dirichlet distribution with $K+1$ dimensions corresponding to K states and the 'rest' of the state space where no transitions are observed (see the supporting information for details). Because of conjugacy, the updated variational transition probabilities for the ith row of Π then read

$$q(\pi_i|\{\eta_{i,k}\}_k) = \mathrm{Dir}\,(\eta_{i,1},\ldots,\eta_{i,K},\eta_{i,-}), \tag{25}$$

with the posterior concentration parameters

$$\begin{aligned}\eta_{i,j} &= (\beta\sigma_j + \delta_{i,j}\xi)\\ &\quad + \sum_t q(Z_t = j, Z_{t-1} = i) \quad \text{for}\, j = 1,\ldots,K,\\ \eta_{i,-} &= \eta_- = \beta\cdot\left(1 - \sum_{i=1}^{K}\sigma_i\right),\end{aligned}$$

where the Kronecker delta $\delta_{i,j} = 1$ if $i = j$ and 0 otherwise.

Observation distributions. Due to conjugacy between the MVN observation likelihoods and the corresponding variational NIW distributions, the variational posterior of the observation model parameters

$$q(\theta_k) = \mathrm{NIW}\,(\mu_k, \Sigma_k|\mu_{0,k}, \lambda_k, \Psi_k, \nu_k), \tag{26}$$

where

$$\begin{aligned}\mu_{0,k} &= \frac{\lambda\mu_0 + IT\bar{x}_k}{\lambda + Q_k}, \quad \lambda_k = \lambda + Q_k, \quad \nu_k = \nu + Q_k,\\ \Psi_k &= \frac{\lambda(Q_k - IT)}{\lambda + Q_k}\mu_0\mu_0^\top + M_k + S_k + \Psi_0.\end{aligned} \tag{27}$$

Recall that I is the number of trajectories and T is the number of time points; $\mu_0, \lambda, \Psi, \nu$ are the parameters of the prior distribution H_0 and

$$\begin{aligned}\bar{x}_k &= \frac{1}{I\cdot T}\sum_{i,t} x_t^i q(Z_t^i = k), \qquad Q_k = \sum_{i,t} q(Z_t^i = k),\\ M_k &= \frac{\lambda I\cdot T}{\lambda + Q_k}(\bar{x}_k - \mu_{0,k})(\bar{x}_k - \mu_{0,k})^\top,\\ S_k &= \sum_{i,t}\left[q(Z_t^i = k)x_t^i x_t^{i,\top} - \frac{\lambda + I\cdot T}{\lambda + Q_k}\bar{x}_k\bar{x}_k^\top\right].\end{aligned} \tag{28}$$

For the approximate vM model, we deal with the periodicity by projecting the data into an interval $[-\pi, +\pi]$ around each mean:

$$x_t^k \leftarrow x_t - 2\pi\cdot\mathrm{sgn}(x_t - \mu_{0,k}). \tag{29}$$

Note that this necessarily leads to an underestimation of the covariance, as we treat the data as if it were produced by a normal distribution, where in reality, it has been generated by a vM; data outside of $[0, 2\pi)$ do not occur. This is tolerable for two reasons: first, as detailed above, we assume the data to be peaked for

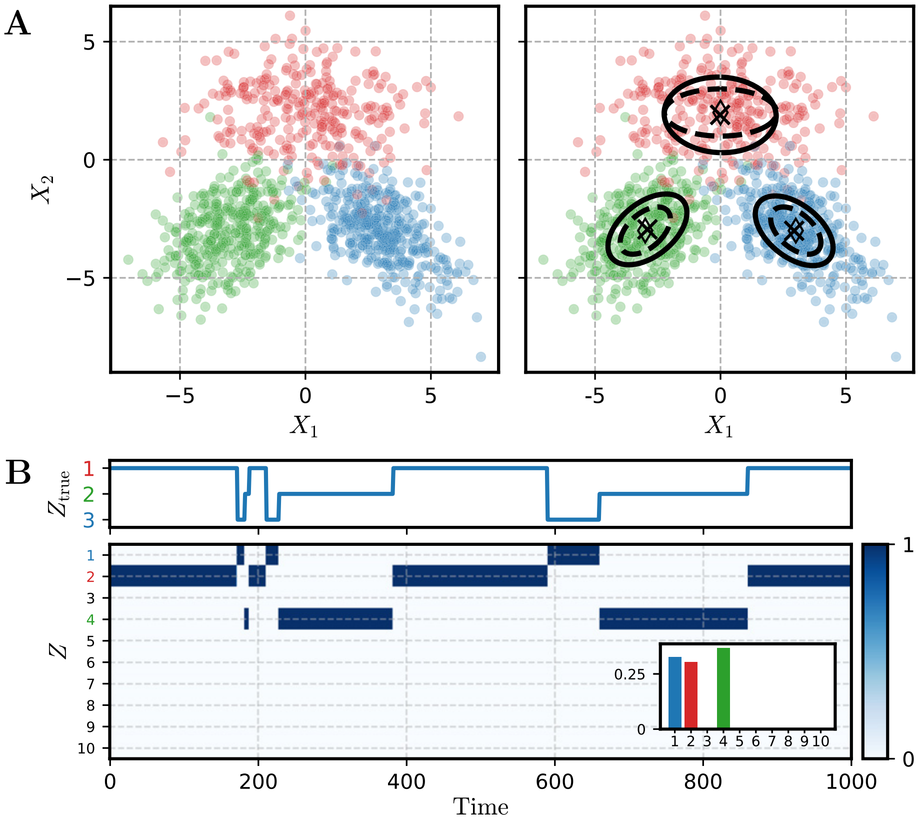


Figure 2. Inference on synthetic data generated from an HMM with 3 states and 2D normal observation noise. For visualization purposes, only one of the 10 simulated sequences is shown. (A) Left: ground truth data. Coloring indicates the corresponding latent state z_t for each data point x_t. Right: inferred state assignments. Coloring is based on the MAP estimate of the latent state. Diamonds and dashed ellipses indicate the ground truth means and covariances. The latter are represented by the $1-\sigma$ covariance ellipses. Black crosses and solid ellipses indicate the inferred variational means $\mu_{0,z}$ and expected variational covariances $\Psi_z/(\nu_z-3)$. (B) Top: ground truth latent sequence $z_{[1,T]}$. Bottom: reconstructed marginal probabilities. Every row $z\in\{1,2,\dots,10\}$ indicates the posterior probability to be in state z at time point t, $q(Z_t=z)\in[0,1]$. Inset: average state occupation $\langle q(z)\rangle=(IT)^{-1}\sum_{i,t}q(Z_t^i=z)$ for all z. Coloring indicates latent states.

our approximation to hold; if this assumption is valid, the probability mass outside the interval $[-\pi,\pi]$ is negligible, cf figure 1. Second, due to the well-known mode-seeking property of VI methods, we anyway incur an underestimation of uncertainty, to which equation (29) should add little [44].

Top-level stick-breaking measure. Setting up the transition distributions $p(\pi_k|\sigma)$ as above (cf 'transition distributions') as $K+1$-Dirichlet distributions results in the relation between $p(\pi_k|\sigma)$ and the stick-breaking measure $\sigma\sim\mathrm{GEM}(\alpha)$ being non-conjugate [36]. Hence, a closed-form update for σ is not available. It is customary to instead utilize a point estimate $q(\sigma)=\delta_{\sigma^*}(\sigma)$, rendering the expectation in equation (22) tractable [36]. The optimum still has no closed-form solution, however; thus, we utilize a gradient optimization scheme and update $\sigma^*\leftarrow\sigma^*+\omega\nabla_{\sigma^*}\mathcal{L}$. To set the step size ω, we utilize a back-tracking line search algorithm [64].

3. Results

We apply the laid-out framework to a range of different data sets. First, we demonstrate the framework on ground truth 2D HMM data to provide an intuition about its functionality. We then apply the model to synthetic continuous-valued SDE data generated from a standard three-well benchmark potential often utilized in the MSM literature [13, 15, 31] and demonstrate its ability to learn a readily interpretable discrete structure from continuous dynamics. Subsequently, we show that our vM approximation works well on synthetic 2D vM data and then employ this approximation on a standard MD benchmark dataset from the protein alanine dipeptide [15, 33, 60, 61]. Lastly, we show the model's utility on a large dataset from voltage clamp experiments on the viral potassium channel Kcv$_{\text{PBCV}-1}$ [62]. We provide the parameters used to generate all synthetic data in the supporting information.

3.1. Synthetic HMM data

To demonstrate the method, we set up a cyclic three-state HMM with transition probabilities

$$\Pi=\begin{pmatrix}0.99 & 0.01 & 0\\ 0 & 0.99 & 0.01\\ 0.01 & 0 & 0.99\end{pmatrix}. \tag{30}$$

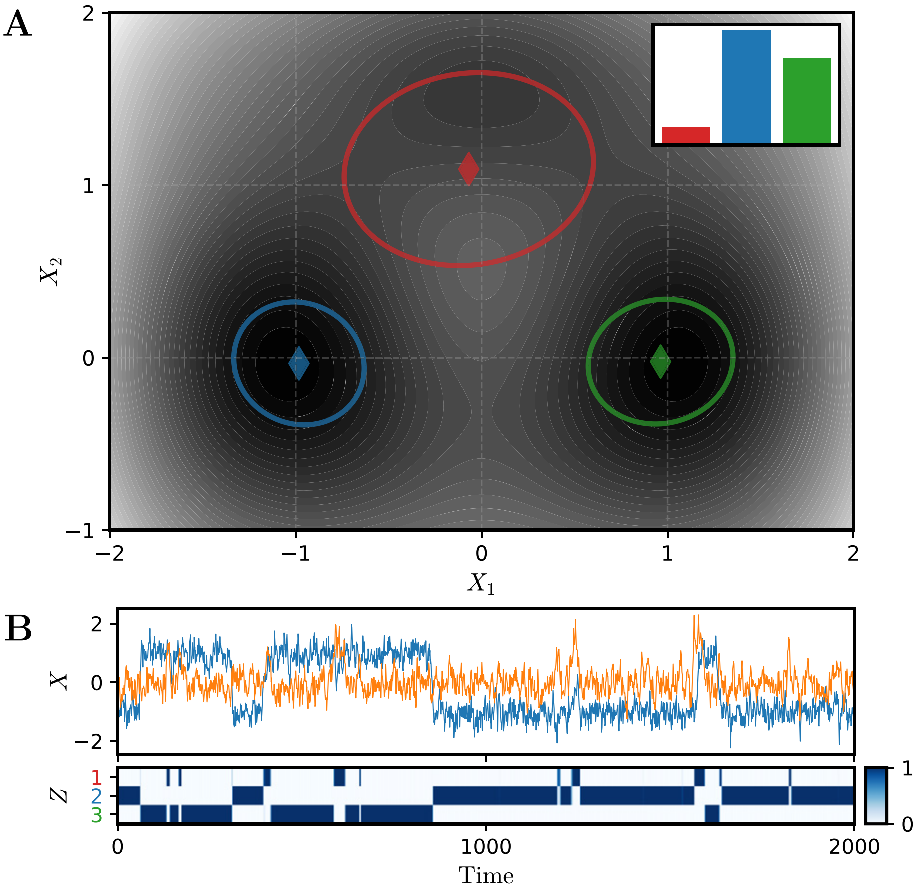


Figure 3. Inference of meta-stable states of 2D SDE dynamics. (A) The heatmap shows the potential landscape used to simulate the continuous dynamics (brighter colors indicate higher values). Colored diamonds and ellipses indicate the inferred variational means and $1-\sigma$ ellipses of the expected variational covariances $\Psi_z/(\nu_z-3)$, cf equation (32). Inset: average state occupation $\langle q(z)\rangle = (IT)^{-1}\sum_{i,t} q(Z_t^i = z)$ for all 3 identified meta-stable states. (B) Top: part of one simulated trajectory (X_1-component in black, X_2-component in orange). Bottom: corresponding latent sequence reconstruction; label colors indicate the variational modes.

From this HMM, we generate 10 independent latent sequences $z^i_{[1,T]}$ consisting of 1000 time points each. Each observation $x^i_t \in \mathbb{R}^2$ is drawn from a normal distribution with anisotropic covariances, $x^i_t \sim \mathrm{N}\,(x^i_t|\mu_{z^i_t}, \Sigma_{z^i_t})$. We provide the values of all $\mu_{z^i_t}, \Sigma_{z^i_t}$ in the supporting information. As discussed above, we utilize a multi-start scheme to alleviate the problem of local optima; we run 10 randomly initialized instances until convergence and pick the one with the highest ELBO value as the optimum. The method accurately recovers three latent states, as shown in the inset of figure 2(B). We show the inferred marginals $q(z^i_t)$ of one latent state sequence in figure 2, demonstrating also the accurate recovery of the ground truth sequence. In particular, the maximum *a posteriori* state assignment z_t of each data point x_t defined via

$$z_t^{\mathrm{MAP}} = \mathrm{argmax}_{z_t}\, q(z_t). \tag{31}$$

precisely matches the corresponding ground truth. Accordingly, also the inferred posterior means $\mu_{0,z}$ and expected covariances (black crosses and circles in figure 2)

$$\mathsf{E}\,[\Sigma_z] = \frac{\Psi_z}{\nu_z - n - 1}, \tag{32}$$

with $n = 2$ the dimensionality of the system, faithfully resemble their true counterparts (diamonds and dashed ellipses in figure 2). Note that the labels of the inferred states of course do not need to correspond to the ground-truth labels; this is an interpretation to be done by the experimenter after convergence of the model. We hence show in figure 2 explicitly the trajectories of all K states, of which only 3 are significantly occupied. In all following figures, we will omit all unoccupied states.

3.2. Stochastic dynamics in a 2D potential

After validating the method, we apply it to a standard benchmark problem of Markov state modeling, which consists of stochastic particle dynamics in a 2D potential landscape with three distinct wells [13, 31, 32]. The dynamics are given by the Itô SDE [63]

$$\mathrm{d}X_t = -\nabla U(X_t)\mathrm{d}t + Q\,\mathrm{d}W_t, \tag{33}$$

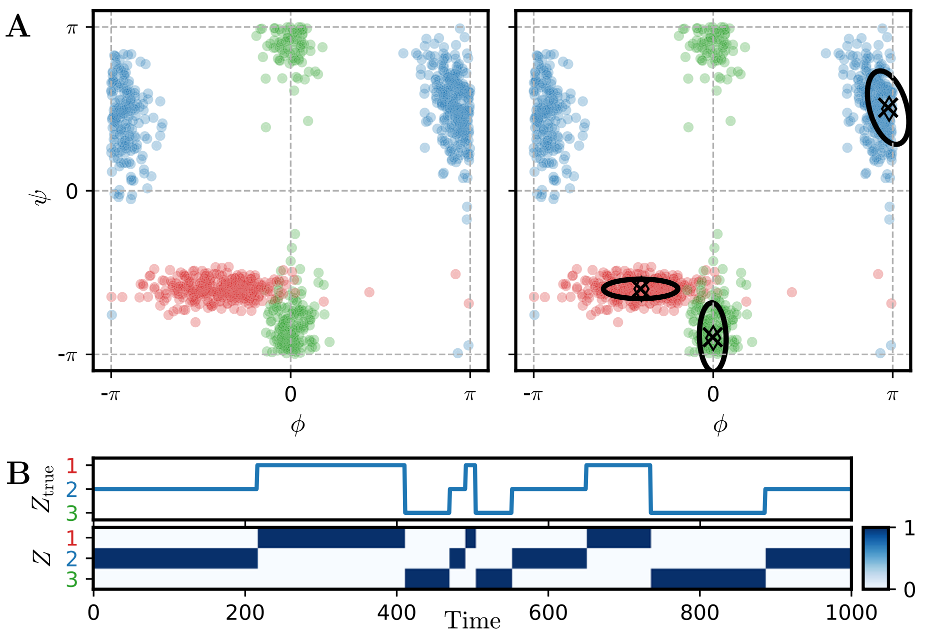


Figure 4. Inference on synthetic vM data utilizing the angular vM approximation. For visualization purposes, only one of the 10 simulated sequences is shown. (A) Left: ground truth data. Coloring indicates the corresponding latent state z_t for each data point x_t. Right: inferred state assignments. Coloring according to the MAP estimate of the latent state. Diamonds indicate the ground truth means. Black crosses and ellipses indicate the inferred variational means $\mu_{0,z}$ and expected variational covariances $\Psi_z/(\nu_z - 3)$. (B) Top: ground truth latent sequence $z_{[1,T]}$. Bottom: reconstructed marginal probabilities. Every row $z \in \{1, 2, \ldots, 10\}$ indicates the posterior probability to be in state z at time point t, $q(Z_t = z) \in [0, 1]$. Coloring indicates latent states.

with the potential function $U: \mathbb{R}^n \to \mathbb{R}$, some constant dispersion $Q \in \mathbb{R}^{n\times n}$ and the standard Brownian motion W. We provide the functional form of U in the supporting information. Using an Euler–Maruyama scheme to simulate these dynamics, we generate 10 trajectories of length $T = 10\,000$ time points each. The potential landscape together with the inferred meta-stable state means $\mu_{0,z}$ and expected covariances $\mathsf{E}\,[\Psi_z]$ is shown in figure 3.

The reconstruction captures the essential features: the locations of the potential wells are accurately identified, where the two deeper wells are fit with higher precision than the shallow minimum at the top,

$$\mu_{0,t} = [0, 1.5]^\top$$
$$\mu_{0,l} = [-1, 0]^\top$$
$$\mu_{0,r} = [1, 0]^\top$$
$$\mu^q_{0,t} = [-0.07, 1.09]^\top$$
$$\mu^q_{0,l} = [-0.98, -0.03]^\top$$
$$\mu^q_{0,r} = [0.96, -0.02]^\top$$

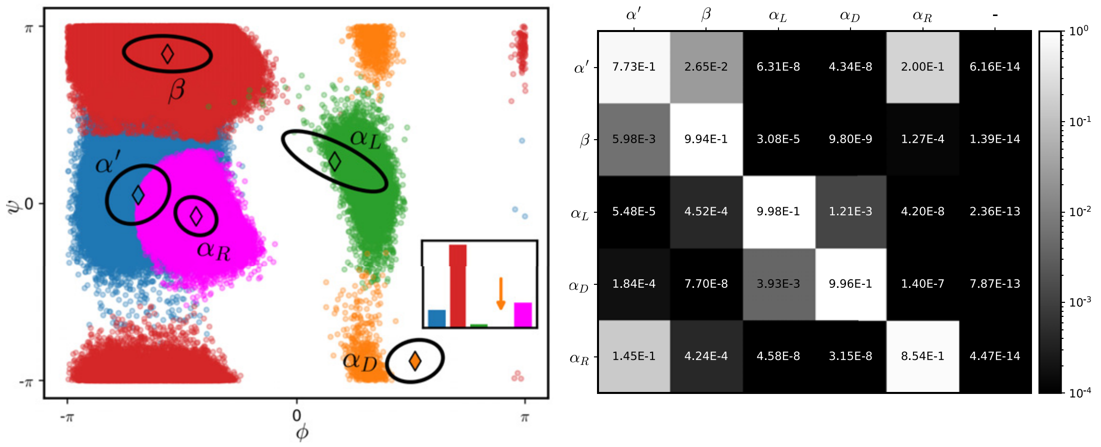
where the superscript q indicates the variational parameters and subscripts t, l, r denote the top (red), left (blue) and right (green) well in figure 3, respectively.

The total state sojourn times correspond to the well depths. The inferred sequence of meta-stable states accordingly yields highly plausible results, as can be checked by comparing the components of the true, continuous process to the inferred discrete switching process.

3.3. Synthetic HMM data with angular observations

To demonstrate the method on angular data, we generate data from the same HMM as before, but employ a vM observation model. We define two latent states to generate independent 1D vM observations along each dimension, cf equation (15); the third state includes angular correlations and is generated from the bivariate vM distribution equation (13) following the sampling scheme of [53]. As in the non-angular case, we create overlap between the individual distributions. Additionally, we include observations that wrap around the period boundary $2\pi \to 0$.

As shown in figure 4, our vM approximation recovers the ground truth means with high fidelity. Since the ground truth data were generated by true vM distributions, no ground truth covariance matrices exist and hence, the inferred covariances cannot be directly compared to them. We can however assess by comparison with the plotted data that the vM approximation produces accurate estimates. In particular, we note that our projection method equation (29) enables sensible and accurate periodic continuations across the period boundaries $2\pi \leftrightarrow 0$.



	α'	β	α_L	α_D	α_R	-
α'	7.73E-1	2.65E-2	6.31E-8	4.34E-8	2.00E-1	6.16E-14
β	5.98E-3	9.94E-1	3.08E-5	9.80E-9	1.27E-4	1.39E-14
α_L	5.48E-5	4.52E-4	9.98E-1	1.21E-3	4.20E-8	2.36E-13
α_D	1.84E-4	7.70E-8	3.93E-3	9.96E-1	1.40E-7	7.87E-13
α_R	1.45E-1	4.24E-4	4.58E-8	3.15E-8	8.54E-1	4.47E-14

Figure 5. Meta-stable states of alanine dipeptide dynamics. Left: coloring indicates the MAP state assignment. Inset shows the relative proportions of all five states occurring in the data. State 4 (orange) only occurs $<1\%$; we added the orange arrow to clarify that this also is displayed in the graph. Colored diamonds represent the variational means of the meta-stable states. Ellipses indicate the expected variational covariances. Annotations refer to known α-helix and β-sheet conformations, cf [65, 69]. Right: expected transition probabilities. Each row shows the transitions probabilities $\mathsf{E}\,[\pi]$ from one of the five found states to all others, including the 'rest' of the state space (cf equation (25)) indicated by '$-$'. Note that the transitions $\alpha' \rightarrow \alpha_R$ and $\alpha_R \rightarrow \alpha'$ are approximately symmetric.

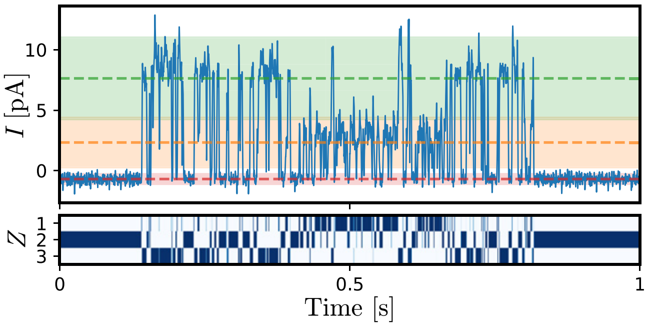


Figure 6. Conformational states of the viral potassium channel $\mathrm{Kcv}_{\mathrm{PBCV-1}}$. For illustrative purposes, we constrain the figure to an interval of $T = 1$ s, rather than showing the full trajectory. Top: measured current over time. Sampling rate is 5 kHz. Colored dashed lines represent the inferred posterior means. Shaded regions indicate the expected variational standard deviation. Green: 'open' state. Red: 'closed' state. Orange: intermediate state. Bottom: inferred latent sequence.

3.4. MD simulation data: alanine dipeptide

We apply the approximate vM model to MD simulation data from alanine dipeptide provided with the pyemma package [24].[3] Alanine dipeptide is a widely used model system in computational biology [65–67]. The data are taken from the cited, publicly available repository, and to the best of our knowledge, are simulated in explicit water using the TIP3P water model [60]. Notably, plain inspection of the raw data reveals that the simulated molecule exhibits meta-stable dynamics, underlining the relevance of our meta-stability assumption at the outset, see supporting information figure 2. This two-dimensional dataset describes the molecule dynamics in terms of the backbone torsion angles (ϕ, ψ) and consists of three independent simulation runs of length $T = 250\,000$ ps with a time step of 1 ps each.

The energy landscape in terms of ϕ and ψ exhibits an intricate fine structure of several local minima. Due to its wide adoption in the field, several computational frameworks have been applied to this dataset, yielding partitionings between three and six different states [15, 33, 60, 61, 68]. As shown in the Ramachandran plot in figure 5, our framework identifies five different states that are in line with the aforementioned literature. By comparison to in-depth MD studies of this particular molecule [65, 69], we can match the found states to known α-helix and β-sheet conformations of alanine dipeptide. Note that the transition probabilities between the two states α' and α_R are found to be very similar. It would hence be a valid interpretation of our model results that these two states could also be lumped together, which would similarly be in accordance with the literature [65, 69]. Notably, the runtime was only around 120 s for one complete optimization run until convergence on a 2.5 GHz Intel i7 processor.

[3] Published under the GNU Lesser General Public License v3.0, https://gnu.org/licenses/lgpl-3.0.en.html.

3.5. Electrophysiological single-molecule ion channel data

Lastly, we use our method on time course data of single channel measurements of the viral potassium channel Kcv$_{\mathrm{PBCV-1}}$. It is known that the wild-type channel switches between an 'open' and a 'closed' state; mutation of the last amino acid to histidine, however, leads to the appearance of sublevels between 'open' and 'closed' [70]. We utilize our method to quantify these sublevels. The data are obtained using the planar lipid bilayer technique as detailed in, e.g., [5]. The applied voltage is 160 mV at pH $= 6$ and data are sampled at 5 kHz over a time span of $T = 60$ s, half of which we discard due to apparent drift. The complete trajectory is shown in supporting information figure 3. Despite the high noise level in the measurements, the inferred latent sequence shows a highly plausible switching behavior, see figure 6: we find three different states: a 'closed' state and an 'open' state as well as one intermediate, subconductive state. The histidine mutation consequentially gives rise to *one* novel channel conformation that cannot be attained by the wild-type. Importantly, one full optimization run only took $\sim$25 s for a sequence of 1.5×10^5, which is orders of magnitude faster than the sampling algorithm proposed in [37] for analysis of such trajectories. Also, conventional methods of trajectory segmentation [74] require both the prespecification of the number of conformational states as well as their conductivity values, which our method does not.

4. Discussion

The nonparametric Bayesian Markov state model framework presented in this work offers a generative modeling approach for inference of global, metastable states from MD and experimental data. In particular, this allows the user to leave the number of conformational states unspecified *a priori* and rather learn it from data. This is beneficial as the number of states in typical computational and experimental settings is not known in advance. In contrast to the MSM approach, we (i) neither need to pre-process the data via discretization and temporal thinning to re-establish Markovianity, (ii) nor manually select the number of meta-stable states. Our method importantly does not deteriorate the temporal resolution of the data.

As we have demonstrated, the model is able to reliably identify the relevant meta-stable states of the system: their number has been sensibly established in all experiments. The application to the triple-well potential highlights the utility of this model on purely continuous data as generated, e.g., by MD. It hence achieves the central goal of modeling the complex dynamics via a finite set of readily interpretable discrete conformational states; in other words, one obtains a spatio-temporal clustering of the data. Furthermore, we presented a computationally tractable approximation to the classical vM distribution that yields accurate results. Application of this approximation to the canonical alanine dipeptide benchmark yielded results consistent with the literature. This is of special relevance to MD due to the frequent use of dihedral angles as system coordinates. We stress that this benchmark problem, albeit consisting of relatively short trajectories compared to MD standards, requires the use of VI methods, since MCMC-type sampling schemes would be computationally infeasible for this task. We thereby also provide a scalable alternative for inference on experimental voltage clamp data, where existing methods all resort to sampling schemes and hence require runtimes on much longer time scales than our framework. This is a significant advance: while nonparametric methods have been around for quite some time (see, e.g., [37–40]), the combination with VI is not established in the field. Furthermore, the adaptation of HDP-HMM methods to typical problems is potentially challenging: utilizing a straightforward categorical observation model $p(x_t|\theta_{z_t})$ on discretized data (see, e.g., [31, 41, 71]). We hence deem it the merit of our study to adapt the existing HDP-HMM framework to the settings commonly found in biophysical problems and to demonstrate its potential for biophysics. Note that from a technical perspective, the proposed vM approximation is novel to the best of our knowledge.

The framework lends itself to further extensions of practical relevance. One interesting direction is provided by the fact that in many MD analysis protocols, some dimensionality reduction is employed, potentially changing the geometry of the data used for analysis [72]. We believe that a natural extension of the presented model is to include observation likelihoods parameterized by neural networks. Akin to the classical variational auto-encoder this could achieve an efficient encoding to lower dimensions [73]. We note that in the field of Markov state modeling, first approaches to this challenge have been proposed recently [33, 34]. None of these proposals however build on nonparametric formulations; the number of states hence remains to be set and tuned by the user. In addition, since not only the observation distributions, but also the transition distributions are parameterized via neural networks, also the necessity to specify an artificial lag or thinning time scale is retained. Another approach from machine learning combines classical probabilistic models with complex likelihood functions in a modular way, however compromising the convexity of the ELBO [75].

Another promising extension are semi-Markovian models, in which the transition between different states is still Markovian, but the sojourn times within each state may be non-exponentially distributed. Similar analyses have already been done for ion channel data and might hence help to get a detailed understanding of more complex switching dynamics [76]. The challenge here is to obtain computationally tractable inference schemes. Notice that similar ideas are also already exploited for lumping in conventional MSM settings [29] and alleviating the lag time issue of MSMs [26].

We believe that variational nonparametric models as the one presented in this paper are a natural match for the requirements of computational and experimental data analysis in the context of structural molecular biology and hence see an untapped potential for applications to biophysical problems.

Acknowledgments

This work has been funded by the European Research Council (ERC) within the CONSYN project, Grant Agreement Number 773196, as well as within the noMAGIC project, Grant Agreement Number 695078.

Data availability statement

The data that support the findings of this study are openly available at the following URL/DOI: https://git.rwth-aachen.de/bcs/projects/lk/bnp-conf-dyn.git

ORCID iDs

Lukas Köhs https://orcid.org/0000-0001-9797-3025

References

[1] Jumper J *et al* 2021 Highly accurate protein structure prediction with AlphaFold *Nature* **596** 583–9

[2] Townshend R J L, Eismann S, Watkins A M, Rangan R, Karelina M, Das R and Dror R O 2021 Geometric deep learning of RNA structure *Science* **373** 1047–51

[3] Sakmann B and Neher E 1984 Patch clamp techniques for studying ionic channels in excitable membranes *Annu. Rev. Physiol.* **46** 455–72

[4] Chen C-C, Cang C, Fenske S, Butz E, Chao Y-K, Biel M, Ren D, Wahl-Schott C and Grimm C 2017 Patch-clamp technique to characterize ion channels in enlarged individual endolysosomes *Nat. Protoc.* **12** 1639–58

[5] Winterstein L-M, Kukovetz K, Rauh O, Turman D L, Braun C, Moroni A, Schroeder I and Thiel G 2018 Reconstitution and functional characterization of ion channels from nanodiscs in lipid bilayers *J. Gen. Physiol.* **150** 637–46

[6] Carnevale V, Delemotte L and Howard R J 2021 Molecular dynamics simulations of ion channels *Trends Biochem. Sci.* **46** 621–2

[7] Dill K A and MacCallum J L 2012 The protein-folding problem, 50 years on *Science* **338** 1042–6

[8] Sponer J *et al* 2018 RNA structural dynamics as captured by molecular simulations: a comprehensive overview *Chem. Rev.* **118** 4177–338

[9] Zhuang X, Ha T, Kim H D, Centner T, Labeit S and Chu S 2000 Fluorescence quenching: a tool for single-molecule protein-folding study *Proc. Natl Acad. Sci. USA* **97** 14241–4

[10] Kajihara D, Abe R, Iijima I, Komiyama C, Sisido M and Hohsaka T 2006 FRET analysis of protein conformational change through position-specific incorporation of fluorescent amino acids *Nat. Methods* **3** 923–9

[11] Huisinga W, Meyn S and Schütte C 2004 Phase transitions and metastability in Markovian and molecular systems *Ann. Appl. Probab.* **14** 419–58

[12] Schütte C, Huisinga W and Deuflhard P 2001 Transfer operator approach to conformational dynamics in biomolecular systems *Ergodic Theory, Analysis, and Efficient Simulation of Dynamical Systems* (Berlin: Springer) pp 191–223

[13] Prinz J-H, Wu H, Sarich M, Keller B, Senne M, Held M, Chodera J D, Schütte C and Noé F 2011 Markov models of molecular kinetics: generation and validation *J. Chem. Phys.* **134** 174105

[14] Shukla D, Hernández C X, Weber J K and Pande V S 2015 Markov state models provide insights into dynamic modulation of protein function *Acc. Chem. Res.* **48** 414–22

[15] Nüske F, Wu H, Prinz J-H, Wehmeyer C, Clementi C and Noé F 2017 Markov state models from short non-equilibrium simulations—analysis and correction of estimation bias *J. Chem. Phys.* **146** 094104

[16] Husic B E and Pande V S 2018 Markov state models: from an art to a science *J. Am. Chem. Soc.* **140** 2386–96

[17] Schwantes C R, McGibbon R T and Pande V S 2015 Perspective: Markov models for long-timescale biomolecular dynamics *J. Chem. Phys.* **141** 090901

[18] Zimmerman M I, Hart K M, Sibbald C A, Frederick T E, Jimah J R, Knoverek C R, Tolia N H and Bowman G R 2017 Prediction of new stabilizing mutations based on mechanistic insights from Markov state models *ACS Cent. Sci.* **3** 1311–21

[19] McKiernan K A, Husic B E and Pande V S 2017 Modeling the mechanism of CLN025 beta-hairpin formation *J. Chem. Phys.* **147** 104107

[20] Mittal S and Shukla D 2017 Predicting optimal deer label positions to study protein conformational heterogeneity *J. Phys. Chem.* B **121** 9761–70

[21] Hart K M, Moeder K E, Ho C M W, Zimmerman M I, Frederick T E and Bowman G R 2017 Designing small molecules to target cryptic pockets yields both positive and negative allosteric modulators *PLoS One* **12** e0178678

[22] Plattner N, Doerr S, De Fabritiis G and Noé F 2017 Complete protein–protein association kinetics in atomic detail revealed by molecular dynamics simulations and Markov modelling *Nat. Chem.* **9** 1005–11

[23] Chu B K, Margaret J T and Sato R R 2017 Read EL Markov state models of gene regulatory networks *BMC Syst. Biol.* **11** 14

[24] Scherer M K, Trendelkamp-Schroer B, Paul F, Pérez-Hernández G, Hoffmann M, Plattner N, Wehmeyer C, Prinz J-H and Noé F 2015 PyEMMA 2: a software package for estimation, validation, and analysis of Markov models *J. Chem. Theory Comput.* **11** 5525–42

[25] Schütte C and Sarich M 2015 A critical appraisal of Markov state models *Eur. Phys. J. Spec. Top.* **224** 2445–62

[26] Cao S, Montoya-Castillo A, Wang W, Markland T E and Huang X 2020 On the advantages of exploiting memory in Markov state models for biomolecular dynamics *J. Chem. Phys.* **153** 014105

[27] Deuflhard P, Huisinga W, Fischer A and Schütte C 2000 Identification of almost invariant aggregates in reversible nearly uncoupled Markov chains *Linear Algebra Appl.* **315** 39–59

[28] Röblitz S and Weber M 2013 Fuzzy spectral clustering by PCCA+: application to Markov state models and data classification *Adv. Data Anal. Classif.* **7** 147–79

[29] Wang W, Liang T, Sheong F K, Fan X and Huang X 2018 An efficient Bayesian kinetic lumping algorithm to identify metastable conformational states via Gibbs sampling *J. Chem. Phys.* **149** 072337

[30] Bowman G R 2012 Improved coarse-graining of Markov state models via explicit consideration of statistical uncertainty *J. Chem. Phys.* **137** 134111

[31] Noé F, Wu H, Prinz J-H and Plattner N 2013 Projected and hidden Markov models for calculating kinetics and metastable states of complex molecules *J. Chem. Phys.* **139** 184114

[32] Wu H, Nüske F, Paul F, Klus S, Koltai P and Noé F 2017 Variational Koopman models: slow collective variables and molecular kinetics from short off-equilibrium simulations *J. Chem. Phys.* **146** 154104

[33] Wu H, Mardt A, Pasquali L and Noe F 2018 Deep generative Markov state models *Advances in Neural Information Processing Systems* vol 31 pp 3975–84

[34] Wehmeyer C and Noé F 2018 Time-lagged autoencoders: deep learning of slow collective variables for molecular kinetics *J. Chem. Phys.* **148** 241703

[35] Kim D and Sudderth E 2011 The doubly correlated nonparametric topic model *Advances in Neural Information Processing Systems* vol 24 pp 1980–8

[36] Johnson M J 2014 *Bayesian Time Series Models and Scalable Inference* (Cambridge: Massachusetts Institute of Technology)

[37] Hines K E, Bankston J R and Aldrich R W 2015 Analyzing single-molecule time series via nonparametric Bayesian inference *Biophys. J.* **108** 540–56

[38] Sgouralis I, Whitmore M, Lapidus L, Comstock M J and Pressé S 2018 Single molecule force spectroscopy at high data acquisition: a Bayesian nonparametric analysis *J. Chem. Phys.* **148** 123320

[39] Calderon C P and Bloom K 2015 Inferring latent states and refining force estimates via hierarchical Dirichlet process modeling in single particle tracking experiments *PLoS One* **10** e0137633

[40] Coscia B J, Calderon C P and Shirts M R 2020 Statistical inference of transport mechanisms and long time scale behavior from time series of solute trajectories in nanostructured membranes *J. Phys. Chem.* B **124** 8110–23

[41] Wu H 2014 A Bayesian nonparametric model for spectral estimation of metastable systems *Proc. 13th Conf. Uncertainty in Artificial Intelligence* pp 878–87

[42] Fox E B, Sudderth E B, Jordan M I and Willsky A S 2008 An HDP-HMM for systems with state persistence *Proc. 25th Int. Conf. Machine Learning* pp 312–9

[43] Van Gael J, Saatci Y, Teh Y W and Ghahramani Z 2008 Beam sampling for the infinite hidden Markov model *Proc. 25th Int. Conf. Machine Learning* pp 1088–95

[44] Blei D M, Kucukelbir A and McAuliffe J D 2017 Variational inference: a review for statisticians *J. Am. Stat. Assoc.* **112** 859–77

[45] Johnson M J and Willsky A S 2013 Bayesian nonparametric hidden semi-Markov models *J. Mach. Learn. Res.* **14** 673–701

[46] Zhang A, Gultekin S and Paisley J 2016 Stochastic variational inference for the HDP-HMM *Proc. 19th Int. Conf. Artificial Intelligence and Statistics* (Proceedings of Machine Learning Research vol 51) pp 800–8

[47] Bishop C 2006 *Pattern Recognition and Machine Learning* (Berlin: Springer)

[48] Teh Y W, Jordan M I, Beal M J and Blei D M 2006 Hierarchical Dirichlet processes *J. Am. Stat. Assoc.* **101** 1566–81

[49] Ghosal S and van der Vaart A 2017 *Fundamentals of Nonparametric Bayesian Inference* (Cambridge: Cambridge University Press)

[50] Leimkuhler B and Matthews C 2016 *Molecular Dynamics* (Berlin: Springer)

[51] Mardia K V 2010 Bayesian analysis for bivariate von Mises distributions *J. Appl. Stat.* **37** 515–28

[52] Mardia K V and Voss J 2014 Some fundamental properties of a multivariate von Mises distribution *Commun. Stat. Theory Methods* **43** 1132–44

[53] Boomsma W, Mardia K V, Taylor C C, Ferkinghoff-Borg J, Krogh A and Hamelryck T 2008 A generative, probabilistic model of local protein structure *Proc. Natl Acad. Sci. USA* **105** 8932–7

[54] Navarro A K W, Frellsen J and Turner R E 2017 The multivariate generalised von Mises distribution: inference and applications *31st AAAI Conf. Artificial Intelligence*

[55] Cover T M 1999 *Elements of Information Theory* (New York: Wiley)

[56] Liang P, Petrov S, Jordan M I and Klein D 2007 The infinite PCFG using hierarchical Dirichlet processes *Empirical Methods in Natural Language Processing* pp 688–97

[57] Hoffman M D, Blei D M, Wang C and Paisley J 2013 Stochastic variational inference *J. Mach. Learn. Res.* **14** 1303–47

[58] Bryant M and Sudderth E 2012 Truly nonparametric online variational inference for hierarchical Dirichlet processes *Advances in Neural Information Processing Systems* vol 25

[59] Beal M J 2003 *Variational Algorithms for Approximate Bayesian Inference* (London: University of London)

[60] Nüske F, Keller B G, Pérez-Hernández G, Mey A S J S and Noé F 2014 Variational approach to molecular kinetics *J. Chem. Theory Comput.* **10** 1739–52

[61] Schwantes C R and Pande V S 2015 Modeling molecular kinetics with tICA and the Kernel trick *J. Chem. Theory Comput.* **11** 600–8

[62] Plugge B *et al* 2000 A potassium channel protein encoded by chlorella virus PBCV-1 *Science* **287** 1641–4

[63] Särkkä S and Solin A 2019 *Applied Stochastic Differential Equations* vol 10 (Cambridge: Cambridge University Press)

[64] Boyd S and Vandenberghe L 2004 *Convex Optimization* (Cambridge: Cambridge University Press)

[65] Mironov V, Alexeev Y, Mulligan V K and Fedorov D G 2019 A systematic study of minima in alanine dipeptide *J. Comput. Chem.* **40** 297–309

[66] Grdadolnik J, Mohacek-Grosev V, Baldwin R L and Avbelj F 2011 Populations of the three major backbone conformations in 19 amino acid dipeptides *Proc. Natl Acad. Sci. USA* **108** 1794–8

[67] Ramachandran G N, Ramakrishnan C and Sasisekharan V 1963 Stereochemistry of polypeptide chain configurations *J. Mol. Biol.* **7** 95–9

[68] Sultan M M and Pande V S 2018 Transfer learning from Markov models leads to efficient sampling of related systems *J. Phys. Chem.* B **122** 5291–9

[69] Feig M 2008 Is alanine dipeptide a good model for representing the torsional preferences of protein backbones? *J. Chem. Theory Comput.* **4** 1555–64

[70] Kukovetz K 2020 *Systematic Analyses of Structure/function Variability of Viral K+ Channels for the Development of Synthetic Channels* (Darmstadt: Technische Universität Darmstadt)

[71] Fox E B 2009 *Bayesian Nonparametric Learning of Complex Dynamical Phenomena* (Cambridge: Massachusetts Institute of Technology)

[72] McGibbon R T, Husic B E and Pande V S 2017 Identification of simple reaction coordinates from complex dynamics *J. Chem. Phys.* **146** 044109

[73] Kingma D P and Welling M 2013 Auto-encoding variational Bayes *Proc. 2nd Int. Conf. Learning Representations* (ICLR)

[74] Schultze R and Draber S 1993 A nonlinear filter algorithm for the detection of jumps in patch-clamp data *J. Membr. Biol.* **132** 41–52

[75] Johnson M J, Duvenaud D K, Wiltschko A, Adams R P and Datta S R 2016 Composing graphical models with neural networks for structured representations and fast inference *Advances in Neural Information Processing Systems* vol 29 pp 2946–54

[76] Ball F, Milne R K and Yeo G F 2002 Multivariate semi-Markov analysis of burst properties of multiconductance single ion channels *J. Appl. Probab.* **39** 179–96