Variational approach to molecular kinetics

Feliks Nüske, Bettina G. Keller,* Guillermo Pérez-Hernández, Antonia S.J.S. Mey, and Frank Noé*

Department for Mathematics and Computer Science, FU Berlin

E-mail: bettina.keller@fu-berlin.de; frank.noe@fu-berlin.de

Abstract

The eigenvalues and eigenvectors of the molecular dynamics propagator (or transfer operator) contain the essential information about the molecular thermodynamics and kinetics. This includes the stationary distribution, the metastable states, and state-to-state transition rates. Here we present a variational approach for computing these dominant eigenvalues and eigenvectors. This approach is analogous the variational approach used for computing stationary states in quantum mechanics. A corresponding method of linear variation is formulated. It is shown that the matrices needed for the linear variation method are correlation matrices that can be estimated from simple MD simulations for a given basis set. The method proposed here is thus, to first define a basis set able to capture the relevant conformational transitions, then compute the respective correlation matrices, and then to compute their dominant eigenvalues and eigenvectors, thus obtaining the key ingredients of the slow kinetics.

1 Introduction

Biomolecules, in particular proteins, often act as small but highly complex machines. Examples range from allosteric changes[1,2] to motor proteins, such as kinesin, which literally

*To whom correspondence should be addressed

walks along microtubules,[1,3] and the ribosome, an enormous complex of RNA molecules and proteins responsible for the synthesis of proteins in the cell.[1,4] To understand how these biomolecular machines work, it does not suffice to know their structure, i.e. their three-dimensional shape. One needs to understand how the structure gives rise to the particular conformational dynamics by which the function of the molecule is achieved. Protein folding is the second field of research in which conformational dynamics plays a major role. Proteins are long polymers of amino acids which fold into particular three-dimensional structure. The astonishingly efficient search for this native conformation in the vast conformational space of the protein can be understood in terms of its conformational dynamics. Besides time-resolved experiments, molecular dynamics simulations are the main technique to investigate conformational dynamics. To date, these simulations yield information on the structure and dynamics of biomolecules at a spatial and temporal resolution which can not be paralleled by any experimental technique. However, the extraction of kinetic models from simulation data is far from trivial, since kinetic information cannot be inferred from structural similarity.[5,6] Similar structures might be separated by large kinetic barriers, and structures which are far apart in some distance measure might be kinetically close.

A natural approach towards modeling the kinetics of molecules involves the partitioning of conformation space into discrete states.[7–17] Subsequently, transition rates or probabilities between states can be calculated, either based on rate theories,[7,18,19] or based on transitions observed in MD trajectories.[6,13,15,16,20–22] The resulting models are often called transition networks, Master equation models or Markov (state) models (MSM),[23–25] where "Markovianity" means that the kinetics are modeled by a memoryless jump process between states. In Markov state models it is assumed that the molecular dynamics simulations used represent an ergodic, reversible and metastable Markov process.[25] Ergodicity means that every possible state would be visited in an infinitely long trajectory and every initial probability distribution of the system converges to a Boltzmann distribution. Reversibility reflects the assumption that the system is in thermal equilibrium. Metastability means that there are parts of the state space in which the system remains over timescales much longer than the fastest fluctuations of the molecule. In order to construct an MSM, the conformational space of the molecule

is discretized into non-overlapping microstates, and the observed transitions between pairs of microstates are counted. One obtains a square matrix with transition probabilities, the so-called transition matrix, from which a wide range of kinetic and thermodynamic properties can be calculated. The equilibrium probability distribution (in the chosen state space) is obtained as the first eigenvector of the transition matrix. Directly from the matrix elements, one can infer kinetic networks and transition paths.[26,27] The dominant eigenvectors of the transition matrix are used to identify metastable states.[28–32] Each dominant eigenvector can be interpreted as a kinetic process, and the associated eigenvalue is related to the timescale on which this process occurs.[25] All this information can be combined to reconstruct the hierarchical structure of the energy landscape.[31,33] Finally, transition matrices represent a very useful framework to connect data from time-resolved experiments with simulation data.[34,35] Over the past decade, extensive knowledge on which factors determine the quality of an MSM has been accumulated. For example, MSMs which are constructed using the internal degrees of freedom of the molecule tend to yield better results than those which were constructed using global descriptors of the structure (H-bond patterns, number of native contacts).[31] Also, degrees of freedom which are not included in the model should decorrelate on short timescales from those which are included.[36] Naturally, the sampling of the transitions limits the accuracy of an MSM, and tools to account for this error have been developed.[37–39] On the whole, the research field has matured to a point at which well-tested protocols for the construction of MSMs from MD data have been established,[25,40,41] and software to construct and validate Markov state models from MD data is freely available.[42,43] MSMs have been applied to analyze the conformational dynamics of peptides[5,31,44] and of small protein domains, such as Villin head piece,[45] pin WW,[46] FiP35 WW.[45] Recently, it has become possible to analyze the folding equilibria of full fast-folding proteins.[47–49] MSMs have also been used to investigate conformational changes, such as the self-association step in the maturation of HIV-protease,[50] ligand binding[51] or the oligomerization of peptide fragments into amyloid structures.[52]

An important aspect that has limited the routine use of MSMs is the difficulty to obtain a state space discretization that will give rise to an MSM that precisely captures the slow kinetics.

The high-dimensional molecular space is usually first discretized using clustering methods in some metric space. The form and location of these clusters, sometimes called "MSM microstates", are crucial for determining the quality of the estimated transition rates.[53–55] Various metrics and clustering methods have been attempted for different molecular systems. Small peptides can be well described by a direct discretization of their backbone dihedrals.[31] Ref.[56] has suggested to use a dihedral principal component analysis to reduce the dihedral space to a low-dimensional sub-space and subsequently cluster this space using e.g. k-means. A rather general metric is the pairwise minimal RMSD-metric in conjunction with some clustering method, such as k-centers or k-medoids.[25,30,41] Recently, the time-lagged independent component analysis (TICA) method was put forward, a dimension reduction approach in which a "slow" low-dimensional subspace is identified that has been shown to provide improved MSMs over previously employed metrics.[57,58]

In recent years, it has been established that the precision of an MSM depends on how well the discretization approximates the shape of the eigenfunction of the underlying dynamical operator (propagator or transfer operator) of the dynamics.[55] When the dynamics are metastable, these eigenfunctions will be almost constant on the metastable states, and change rapidly at the transition states.[59] Thus, methods that have sought to construct a maximally metastable discretization[30,60] have been relatively successful for metastable dynamics. However, the MSM can be improved by using a non-metastable discretization, especially when it finely discretizes the transition states, so as to trace the variation of the eigenfunction in these regions.[25,55] An alternative way of achieving a good resolution at the transition state without using a fine discretization is to use appropriately placed smooth basis functions, such as the smooth partition-of-unity basis functions suggested in.[61–63] The core-based discretization method proposed in[11] effectively employs a smooth partition-of-unity basis defined by the committor functions between sets.[64]

All of the above methods have in common that they attempt to construct an appropriate discretization based on the simulation data. This has a twofold disadvantage: (1) different simulation runs will produce different discretizations, making them hard to compare, (2) data-based clusters have no intrinsic meaning. Interpretation in terms of structural transitions

must be recovered by analyzing the molecular configurations contained in specific clusters. With all of the above methods, choosing an appropriate combination of the metric, the clustering method, the number and the location of clusters or cores, is still often a trial-and-error approach.

Following the recently introduced variational principle for metastable stochastic processes,[65] we propose a variational approach to molecular kinetics. Starting from the fact that the molecular dynamics propagator is a self-adoint operator, we can formulate a variational principle. Using the method of linear variation we derive a Roothaan-Hall-type generalized eigenvalue problem that yields an optimal representation of eigenvectors of the propagator in terms of an arbitrary basis set. Both ordinary MSMs using crisp clustering and MSMs with a smooth discretization can be understood as special cases of this variational approach. In contrast to previous MSMs using smooth discretization, our basis functions do not need to be a partition of unity, although this choice has some merits.

Besides its theoretical attractiveness, the variational approach has some advantages over MSMs. Firstly, the data-driven discretization is replaced by a user-selection of an appropriate basis set, typically of internal molecular coordinates. The chosen basis set may reflect chemical intuition - for example basis functions may be predefined to fit known transition states of backbone dihedral angles, or formation/dissociation of tertiary contacts between hydrophobically or electrostatically interacting groups. As a result, one may obtain a precise model with fewer basis functions needed than discrete MSM states. Moreover, each basis function is associated with a chemical meaning, and thus the interpretation of the estimated eigenfunctions becomes much more straightforward than for MSMs. When using the same basis set for different molecular systems of the same class, one obtains models that are directly comparable in contrast to conventional MSMs. The representation of the propagator eigenfunctions can still be systematically improved by adding more basis functions, or by varying the basis set.

Our method is analogous to the method of linear variation used in quantum chemistry.[66] The major difference is that the propagator is self-adjoint with respect to a non-Euclidean scalar product, whereas the Hamiltonian is self-adjoint with respect to the Euclidean scalar product.

The derivation of the method is detailed in section 2 and appendices A - C.

2 Theory

2.1 The dynamical propagator

Consider the conformational space X of an arbitrary molecule consisting of N atoms, i.e. the $3N-6$-dimensional space spanned by the internal degrees of freedom of the molecule. The conformational dynamics of the molecule in this space can be represented by a dynamical process $\{x_t\}$, which samples at a given time t a particular point $x_t \in X$. In this context, x_t is often called a trajectory. This process is governed by the equations of motion, and it can be simulated using standard molecular-dynamics programs. We assume that an implementation of thermostatted molecular dynamics is employed which ensures that x_t is time-homogeneous, Markovian, ergodic and reversible with respect to a unique stationary density (usually the Boltzmann distribution). We introduce a propagator formulation of these dynamics, following.[65] Readers familiar with this approach might want to skip to section 2.2.

Next, consider an infinite ensemble of molecules of the same type, distributed in the conformational space according to some initial probability density $|\rho_0(x)\rangle$. This initial probability density evolves in time in a definite manner which is determined by the aforementioned equations of motion for the individual molecules. We assume that the time evolution is Markovian

$$p(x,y;\tau)dy = \mathbb{P}(x_{t+\tau} \in y\,dy | x_t = x) \quad (1)$$

$$= \mathbb{P}(x_\tau \in y\,dy | x_0 = x) \quad (2)$$

where τ is a finite time step, and $p(x,y;\tau)$ is the so-called transition density, which is assumed to be independent of time t (time-homogeneous). Figure 1 shows an example of the time-evolution of a probability density in a one-dimensional two-well potential. Eq. 2 implies that the probability of finding a molecule in conformation $y\,dy$ at time $t+\tau$ depends only on the conformation x it has occupied one time step earlier, and not on the sequence of conformations is has visited before t. The unconditional probability density of finding a molecule

in conformation y at time $t+\tau$ is obtained by integrating over all starting conformations x

$$\rho_{t+\tau}(y) \quad = \quad \int_X p(x,y;\tau)\rho_t(x)dx\,. \tag{3}$$

This equation in fact defines an operator $\mathscr{P}(\tau)$ which propagates the probability density by a finite time step τ

$$|\rho_{t+\tau}(x)\rangle \quad = \quad \mathscr{P}(\tau)\,|\rho_t(x)\rangle \tag{4}$$

$$|\rho_{t+n\tau}(x)\rangle \quad = \quad \mathscr{P}^n(\tau)\,|\rho_t(x)\rangle\,. \tag{5}$$

$\mathscr{P}(\tau)$ is called a propagator, and the time step τ is often called the lag time of the propagator. One says the propagator is parametrized with τ. Like $p(x,y;\tau)$, the propagator $\mathscr{P}(\tau)$ in Eq. 5 is time-homogeneous, i.e. it does not depend on t. The way it acts on a density $|\rho(x,t)\rangle$ is not a function of the time t at which this density occurs, but only a function of the time step τ by which the density is propagated (Figure 1).

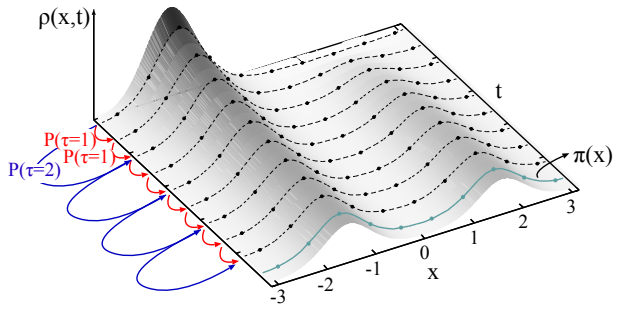

Figure 1: Illustration of two propagators acting on a probability density $|\rho_t(x)\rangle$. *Grey surface:* time evolution of $|\rho_t(x)\rangle$; *black dotted line:* snap shots of $|\rho_t(x)\rangle$; *cyan line:* equilibrium density $|\pi(x)\rangle$ to which $|\rho_t(x)\rangle$ eventually converges; *red, blue:* propagators with different lag times τ, which propagate an initial density by a time step τ in time.

The way the propagator acts on the density can be understood in terms of its eigenfunctions $\{|l_\alpha(x)\rangle\}$ and associated eigenvalues $\{\lambda_\alpha\}$, which are defined by the following eigenvalue equation

$$\mathscr{P}(\tau)\,|l_\alpha(x)\rangle = \lambda_\alpha\,|l_\alpha(x)\rangle\,. \tag{6}$$

For the class of processes which are discussed in this publication, the eigenfunctions form a complete set of $\mathbb{R}^{3N}$ (see below). Hence, any probability density (in fact any function) in this

space can be expressed as linear combination of $\{l_\alpha(x)\}$. Eq. 5 can be rewritten as

$$|\rho_{t+n\tau}(x)\rangle = \sum_\alpha c_\alpha \lambda_\alpha^n |l_\alpha(x)\rangle \quad (7)$$

$$= \sum_\alpha c_\alpha e^{-n\tau/t_\alpha} |l_\alpha(x)\rangle \,, \quad (8)$$

where n is the number of discrete time steps τ. The eigenfunctions can be interpreted as kinetic processes which transport probability density from one part of the conformational space to another, and thus modulate the shape of the overall probability density. See[25] for a detailed explanation of the interpretation of eigenfunctions. The eigenvalues are linked to the timescales t_α on which the associated kinetic processes take place by

$$t_\alpha = -\frac{\tau}{\ln(\lambda_\alpha)} \,. \quad (9)$$

These timescales are of particular interest because they may be accessible using various kinetic experiments.[35,67–69]

Given the aforementioned properties of the molecular dynamics implementation, $\mathscr{P}(\tau)$ is an operator with the following properties. A more detailed explanation can be found in appendix A.

- $\mathscr{P}(\tau)$ has a unique stationary density, i.e. there is a unique solution $|\pi(x)\rangle$ to the eigenvalue problem $\mathscr{P}(\tau)|\pi(x)\rangle = |\pi(x)\rangle$.
- Its eigenvalue spectrum is bounded from above by $\lambda_1 = 1$. Also, λ_1 is the only eigenvalue of absolute value equal to one.
- $\mathscr{P}(\tau)$ is self-adjoint w.r.t. the weighted scalar product $\langle f \,|\, g\rangle_{\pi^{-1}} = \int_\Omega f(x)g(x)\pi^{-1}(x)dx$. Consequently, its eigenfunctions $|l_\alpha(x)\rangle$ form an orthonormal basis of the Hilbert space of square-integrable functions w.r.t. this scalar product. Its eigenvalues are real and can be numbered in descending order:

$$1 = \lambda_1 > \lambda_2 \geq \lambda_3 \geq \ldots \quad (10)$$

2.2 Variational principle and the method of linear variation.

A variational principle can be derived for any operator whose eigenvalue spectrum is bound (either from above or from below) and whose eigenvectors form a complete basis set and are orthonormal with respect to a given scalar product. The variational principle for propagators was derived in.[65] The derivation is analogous to the derivation of the variational principle of the quantum-mechanical Hamilton operator.[66] For convenience, we give a compact derivation in appendix B.

The variational principle can be summarized in three steps: Firstly, for the exact eigenfunction $|l_\alpha(x)\rangle$, the following equality holds:

$$\langle l_\alpha \,|\, \mathscr{P}(\tau) \,|\, l_\alpha \rangle_{\pi^{-1}} = \lambda_\alpha(\tau) = \mathrm{e}^{-\tau/t_\alpha}. \tag{11}$$

The expression $\langle f \,|\, \mathscr{P}(\tau) \,|\, f \rangle_{\pi^{-1}}$ is the analogue of the quantum-mechanical expectation value and has the interpretation of a time-lagged auto-correlation (*c.f.* sec. 2.3). The autocorrelation of the α^{th} eigenfunction is identical to the α^{th} eigenvalue.

Secondly, for any trial function $|f\rangle$ which is normalized according to Eq. 64 the following inequality holds:

$$\begin{aligned} \langle f \,|\, \mathscr{P}(\tau) \,|\, f \rangle_{\pi^{-1}} &= \int_X f(x)\pi^{-1}(x)\mathscr{P}(\tau)f(x)\,dx & (12)\\ &\leq \lambda_1 = 1\,, & (13) \end{aligned}$$

where equality $\langle f \,|\, \mathscr{P}(\tau) \,|\, f \rangle_{\pi^{-1}} = \lambda_1$ is achieved *if and only if* $|f\rangle = |l_1\rangle$. This is at the heart of the variational principle.

Thirdly, this inequality is applicable to other eigenfunctions: When $|f\rangle$ is orthogonal to the $\alpha - 1$ first eigenfunctions, the variational principle will apply to the α^{th} eigenfunction/eigen-

value pair:

$$\langle f \,|\, \mathscr{P}(\tau) \,|\, f \rangle_{\pi^{-1}} \leq \lambda_\alpha \tag{14}$$

$$\langle f | l_\beta \rangle_{\pi^{-1}} = 0 \;\; \forall \beta = 1, ..., \alpha - 1. \tag{15}$$

This variational principle allows to formulate the method of linear variation for the propagator. Again, the derivation detailed in[65] is analogous to the derivation of the method of linear variation in quantum chemistry.[66] The trial function $|f\rangle$ is linearly expanded using a basis of n basis functions $\{|\varphi_i\rangle\}_{i=1}^n$

$$f \;\; = \;\; \sum_{i=1}^{n} a_i \,|\varphi_i\rangle \;, \tag{16}$$

where a_i are the expansion coefficients. We only choose basis sets consisting of real-valued functions because all eigenvectors of $\mathscr{P}(\tau)$ are real-valued functions. Consequently, the expansion coefficients a_i are real numbers. However, the basis set does not necessarily have to be orthonormal. In the method of linear variation, the expansion coefficients a_i are varied such that the right-hand side of Eq. 13 becomes maximal, while the basis functions are kept constant. The variation is carried out under the constraint that $|f\rangle$ remains normalized with respect to Eq. 64 using the method of Lagrange multipliers. For details, see appendix C. The derivation leads to a matrix formulation of Eq. 6

$$\mathbf{Ca} \;\; = \;\; \lambda \mathbf{Sa} \,. \tag{17}$$

$\mathbf{a}$ is the vector of expansion coefficients a_i, $\mathbf{C}$ is the (time-lagged) correlation matrix with elements

$$C_{ij} \;\; = \;\; \left\langle \varphi_i \right| \mathscr{P}(\tau) \left| \varphi_j \right\rangle_{\pi^{-1}} \;, \tag{18}$$

and $\mathbf{S}$ is the overlap matrix of the basis set, where the overlap is calculated with respect to the

weighted scalar product

$$S_{ij} \;=\; \left\langle \varphi_i \,\middle|\, \varphi_j \right\rangle_{\pi^{-1}} . \tag{19}$$

Solving the generalized eigenvalue problem in Eq. 17, one obtains the first n eigenvectors of $\mathscr{P}(\tau)$ expressed in the basis $\{|\varphi_i\rangle\}_{i=1}^{n}$ and the associated eigenvalues λ_α.

2.3 Estimating the matrix elements

To solve the generalized eigenvalue equation (Eq. 17) we need to calculate the matrix elements C_{ij}. In the quantum chemical version of the linear variation approach, the matrix elements H_{ij} for the Hamiltonian $\mathscr{H}$ (see appendix A) are calculated directly with respect to the chosen basis, either analytically or by solving the integral $H_{ij} = \left\langle \varphi_i \,\middle|\, \mathscr{H} \,\middle|\, \varphi_j \right\rangle$ numerically. Such a direct treatment is not possible for the matrix elements of the propagator. However, we can use a trajectory x_t of a single molecule, as it is generated for example by MD simulations, to sample the matrix elements and thus obtain an estimate for C_{ij}. For this, we introduce a basis set $\{\chi_i\}$ consisting of the n co-functions of the original basis set $\{\varphi_i\}$ by weighting the original functions with π^{-1}

$$\chi_i(x) = \pi^{-1}(x)\varphi_i(x) \quad \Leftrightarrow \quad \varphi_i(x) = \pi(x)\chi_i(x) . \tag{20}$$

Inserting Eq. 20 into the definition of the matrix elements C_{ij} (Eq. 18) we obtain

$$\begin{aligned} C_{ij} &= \left\langle \varphi_i \,\middle|\, \mathscr{P}(\tau) \,\middle|\, \varphi_j \right\rangle_{\pi^{-1}} \\ &= \left\langle \chi_i\pi \,\middle|\, \mathscr{P}(\tau) \,\middle|\, \pi\chi_j \right\rangle_{\pi^{-1}} \\ &= \int_X \int_X \chi_i(z)\, p(y,z,\tau)\pi(y)\, \chi_j(y)\, dy\, dz. \end{aligned} \tag{21}$$

The last line of Eq. 21 has the interpretation of a time-lagged cross-correlation between the functions χ_i and χ_j

$$\begin{aligned} \mathrm{cor}(\chi_i, \chi_j, \tau) &:= \int_X \int_X \chi_i(z) \mathbb{P}(x_{t+\tau} = z | x_t = y) \cdot & (22) \\ & \qquad \chi_j(y) \mathbb{P}(x_t = y)\, dy\, dz\,, & (23) \end{aligned}$$

which can be estimated from a time-continuous time series x_t of length T as

$$\widehat{\mathrm{cor}}_T(\chi_i, \chi_j, \tau) = \frac{1}{T-\tau} \int_0^{T-\tau} \chi_j(x_t) \chi_i(x_{t+\tau})\, dt\,, \tag{24}$$

or from a time-discretized time series x_t as

$$\widehat{\mathrm{cor}}_T(\chi_i, \chi_j, \tau) = \frac{1}{N_T - n_\tau} \sum_{t=1}^{N_T - n_\tau} \chi_j(x_t) \chi_i(x_{t+n_\tau})\,, \tag{25}$$

where $N_T = T/\Delta t$, $n_\tau = \tau/\Delta t$, and Δt is the time step of the time-discretized time series. In the limit of infinite sampling and for an ergodic process, the estimate approaches the true value

$$C_{ij} = \mathrm{cor}(\chi_i, \chi_j, \tau) = \lim_{T \to \infty} \widehat{\mathrm{cor}}_T(\chi_i, \chi_j, \tau)\,. \tag{26}$$

Note that the second line in Eq. 21 can also be read as the matrix representation of an operator which acts on the space spanned by $\{\chi_i\}$, the co-functions of $\{\varphi_i\}$ (Eq. 20). This is the so-called transfer operator $\mathscr{T}(\tau)$.

$$\begin{aligned} C_{ij}(\tau) &= \left\langle \chi_i \pi \middle| \mathscr{P}(\tau) \middle| \pi \chi_j \right\rangle_{\pi^{-1}} & (27) \\ &= \left\langle \chi_i \middle| \mathscr{T}(\tau) \middle| \chi_j \right\rangle_{\pi}\,, & (28) \end{aligned}$$

with

$$\mathscr{T}(\tau)\, |f(z)\rangle = \frac{1}{\pi(z)} \int_X p(y, z, \tau) \pi(y) f(y)\, dy\,. \tag{29}$$

In particular, $\mathscr{T}(\tau)$ has the same eigenvalues as the propagator and its eigenfunctions are the co-functions of the propagator eigenfunctions:

$$r_\alpha(x) = \pi^{-1}(x) l_\alpha(x)\,. \tag{30}$$

We will sometimes refer to the functions r_α as right eigenfunctions. For more details on the transfer operator the reader is referred to.[59]

2.4 Crisp basis sets - conventional MSMs

Markov state models (MSMs), as they have been discussed up to now in the literature,[23–25,28,30,31,40–43,55,70] arise as a special case of the proposed method. Namely, the choice of basis sets in conventional MSMs is restricted to indicator functions, i.e. functions which have the value 1 on a particular set S_i of the conformational space X and the value 0 otherwise

$$\chi_i^{MSM}(x) \;=\; \begin{cases} 1 & \text{if } x \in S_i \\ 0 & \text{else.} \end{cases} \tag{31}$$

In effect, this is a discretization of the conformational space, for which the estimation of the matrix $\mathbf{C}$ (Eq. 25) reduces to counting the observed transitions z_{ij} between sets S_i and S_j

$$C_{ij} \;=\; \frac{1}{N_T - n_\tau} \sum_{t=1}^{N_T - n_\tau} \chi_j^{MSM}(x_t)\chi_i^{MSM}(x_{t+n_\tau}) \tag{32}$$

$$\;=\; \frac{z_{ij}}{N_T - n_\tau}\,. \tag{33}$$

It is easy to verify,[65] that the overlap matrix $\mathbf{S}$ is a diagonal matrix, with entries π_i equal to the stationary probabilities of the sets:

$$S_{ii} = \int_{S_i} \pi(x)\mathrm{dx} =: \pi_i. \tag{34}$$

Thus, the eigenvalue problem Eq. 17 becomes:

$$\mathbf{Ca} = \lambda \mathbf{\Pi a} \tag{35}$$

$$\mathbf{Ta} = \lambda \mathbf{a}, \tag{36}$$

where $\mathbf{C}$ is the correlation matrix, $\mathbf{\Pi} = \mathbf{S} = \mathrm{diag}\{\pi_1, ..., \pi_n\}$ is the diagonal matrix of stationary probabilities, and $\mathbf{T} = \mathbf{\Pi}^{-1}\mathbf{C}$ is the MSM transition matrix. Thus $\mathbf{a}$ is a right eigenvector of the MSM transition matrix. As the equations above provide the linear variation optimum, using MSMs and their eigenvectors corresponds to finding an optimal step-function approximation of the eigenfunctions. Moreover, we can use the weighted functions

$$\mathbf{b}_\alpha \quad = \quad \mathbf{\Pi a}_\alpha \tag{37}$$

and see that they are left eigenfunctions of $\mathbf{T}$:

$$\mathbf{T}\mathbf{\Pi}^{-1}\mathbf{b} = \lambda \mathbf{\Pi}^{-1}\mathbf{b} \tag{38}$$

$$\mathbf{b}^T\mathbf{\Pi}^{-1}\mathbf{C} = \lambda \mathbf{b}^T \tag{39}$$

$$\mathbf{b}^T\mathbf{T} = \lambda \mathbf{b}^T. \tag{40}$$

Note that the crisp basis functions form a partition of unity, meaning that their sum is the constant function with value one, which is the first exact eigenfunction of the transfer operator $\mathscr{T}(\tau)$. For this reason, any state space partition that is a partition of unity solves the approximation problem of the first eigenvalue/eigenvector pair exactly: the first eigenvalue is exactly $\lambda_1 = 1$, the expansion coefficients a_i^1 of the first eigenvector $|r_1\rangle$ are all equal to one. The corresponding first left eigenvector $\mathbf{b}_1 = \mathbf{\Pi a}_1$ fulfills the stationarity condition:

$$\mathbf{b}_1^T = \mathbf{b}_1^T\mathbf{T} \tag{41}$$

and is therefore, when normalized to an element sum of 1, the stationary distribution π of $\mathbf{T}$.

2.5 Stationary probability distribution in the variational approach

All previous MSM approaches – including the most common "crisp" cluster MSMs, but also the smooth basis function approaches used in[24,61,64] – have directly or indirectly used basis functions that are a partition of unity. The reason for this is that using such a partition of unity, one can recover the exact first eigenvector, and thus a meaningful stationary distribution.

In the present contribution, we give up the partition of unity condition, in order to be able to fully exploit the variational principle of the propagator with an arbitrary choice of basis sets. Therefore, we must investigate whether this approach is still meaningful and can give us "something" like the stationary distribution.

Revisiting the MSM case, the stationary probability numbers π_i can be interpreted as stationary probabilities of the sets S_i, or, in other words, they measure the contribution of these sets to the full partition function Z:

$$\begin{aligned} \pi_i &= \frac{Z_i}{Z} && (42)\\ Z_i &= \int_{S_i} \mathrm{e}^{-v(x)}\mathrm{dx} = \int_X \chi_i^{MSM}(x)\mathrm{e}^{-v(x)}\mathrm{dx} && (43)\\ \sum_i \pi_i &= \sum_i \frac{Z_i}{Z} = 1, && (44) \end{aligned}$$

where $v(x)$ is a reduced potential.

If we move on to a general basis, we can maintain a similar interpretation of the vector $\mathbf{b}_1 = \mathbf{S}\mathbf{a}_1$, as long as the first estimated eigenvalue λ_1 remains equal to one. If we use the general definition of Z_i as the local density of the basis function χ_i:

$$Z_i = \int_X \chi_i(x)\mathrm{e}^{-v(x)}\mathrm{dx}. \tag{45}$$

Then we still have

$$b_i = \frac{Z_i}{C} \tag{46}$$

for all i, where

$$C = \int_X \sum_i \chi_i(x) \mathrm{e}^{-v(x)} \mathrm{dx}. \tag{47}$$

Interestingly, this relation also becomes approximately true if the estimated eigenvalue λ_1 approaches one, as proved in appendix D. As a result, the concept of the stationary distribution is still meaningful for basis sets that do not form a partition of unity. Moreover, it is completely consistent with the variational principle, because the vector $\mathbf{b}_1$ becomes a probability distribution in the optimum $\lambda_1 = 1$.

2.6 Estimation method

We summarize by formulating a computational method to estimate the eigenvectors and eigenvalues of the associated propagator from a time series (trajectory) x_t using an arbitrary basis set.

1. Choose a basis set $\{\chi_i\}$.

2. Estimate the matrix elements of the correlation matrix $\mathbf{C}$ and of the overlap matrix $\mathbf{S}$ using Eq. 25 with lag times τ and 0, respectively.

3. Solve the generalized eigenvalue problem in Eq. 17. This yields the α^{th} eigenvalue λ_α of the propagator (and the transfer operator) and the expansion coefficients a_i^α of the associated eigenvector.

4. The eigenvectors of the transfer operator are obtained directly from the expansion coefficients a_i^α via:

$$r_\alpha \;=\; \sum_{i=1}^{n} a_i^\alpha \, |\chi_i\rangle \; . \tag{48}$$

5. If an estimate of the stationary density π is available, the eigenvectors of the propagator $\mathscr{P}(\tau)$ are obtained from

$$l_\alpha \;=\; \sum_{i=1}^{n} a_i^\alpha \, |\varphi_i\rangle = \sum_{i=1}^{n} a_i^\alpha \, |\pi\chi_i\rangle \; . \tag{49}$$

3 Methods

3.1 One-dimensional diffusion models

3.1.1 Simulations.

We first consider two examples of one-dimensional diffusion processes x_t governed by Brownian dynamics. The process is then described by the stochastic differential equation

$$dx_t = -\nabla v(x_t)dt + \sqrt{2D}\,dB_t, \tag{50}$$

where v is the reduced potential energy (measured in units of k_BT, where k_B is the Boltzmann constant and T is the temperature), D is the diffusion constant, and dB_t denotes the differential of Brownian motion. For simplicity, we set all of the above constants equal to one. The potential function is given by the harmonic potential

$$v(x) = 0.5x^2, \quad x \in \mathbb{R}, \tag{51}$$

in the first case, and by the periodic double-well potential

$$v(x) = 1 + \cos(2x), \quad x \in [-\pi, \pi), \tag{52}$$

in the second case. In order to apply our method, we first produced finite simulation trajectories for both potentials. To this end, we picked an (also artificial) time-step $\Delta t = 10^{-3}$, and then used the Euler-Maruyama method, where position x_{k+1} is computed from position x_k as

$$x_{k+1} = x_k - \Delta t \nabla v(x_k) + \sqrt{2D\Delta t}\,y_t \tag{53}$$

$$y_t \sim \mathcal{N}(0,1). \tag{54}$$

In this way, we produced simulations of $5 \cdot 10^6$ time-steps for the harmonic potential and 10^7 time-steps for the double-well potential.

3.1.2 Gaussian model.

We apply our method with Gaussian basis functions to both problems. To this end, $n = 2,3,\ldots,10$ centers are chosen at uniform distance between $x = -4$ and $x = 4$ for the harmonic potential and between $x = -\pi$ and $x = \pi$ for the double-well potential. In the latter case, the basis functions are modified to be periodic on $[-\pi,\pi)$. Subsequently, an "optimal" width of the Gaussians is picked by simply trying out several choices for the standard deviations between 0.4 and 1.0 and using the one which yields the highest second eigenvalue. From this choice, the matrices $\mathbf{C}$ and $\mathbf{S}$ are estimated and the eigenvalues, -functions and implied timescales are computed.

3.1.3 Markov models.

As a reference for our methods, we also compute Markov state models for both processes. To this end, the simulation data is clustered into $n = 2,3,\ldots,10$ disjoint clusters using the kmeans algortihm. Subsequently, the EMMA software package[43] is used to estimate the MSM transition matrices and to compute eigenvalues and timescales.

3.2 Alanine dipeptide

3.2.1 MD simulations.

We performed 20 simulations of 200 ns of all-atom explicit solvent molecular dynamics of alanine dipeptide using the AMBER ff-99SB-ILDN force field.[71] The detailed simulation setup is found in the appendix.

3.2.2 Gaussian model.

Similar to the previous example, we use periodic Gaussian functions which only depend on one of the two significant dihedral angles of the system (see Sec. 4.2) to apply our method. For both dihedrals, we separately perform a pre-selection of the Gaussian trial functions. To this end, we first project the data onto the coordinate, then we solve the projected optimization problem for all possible choices of centers and widths, and then pick the ones yielding

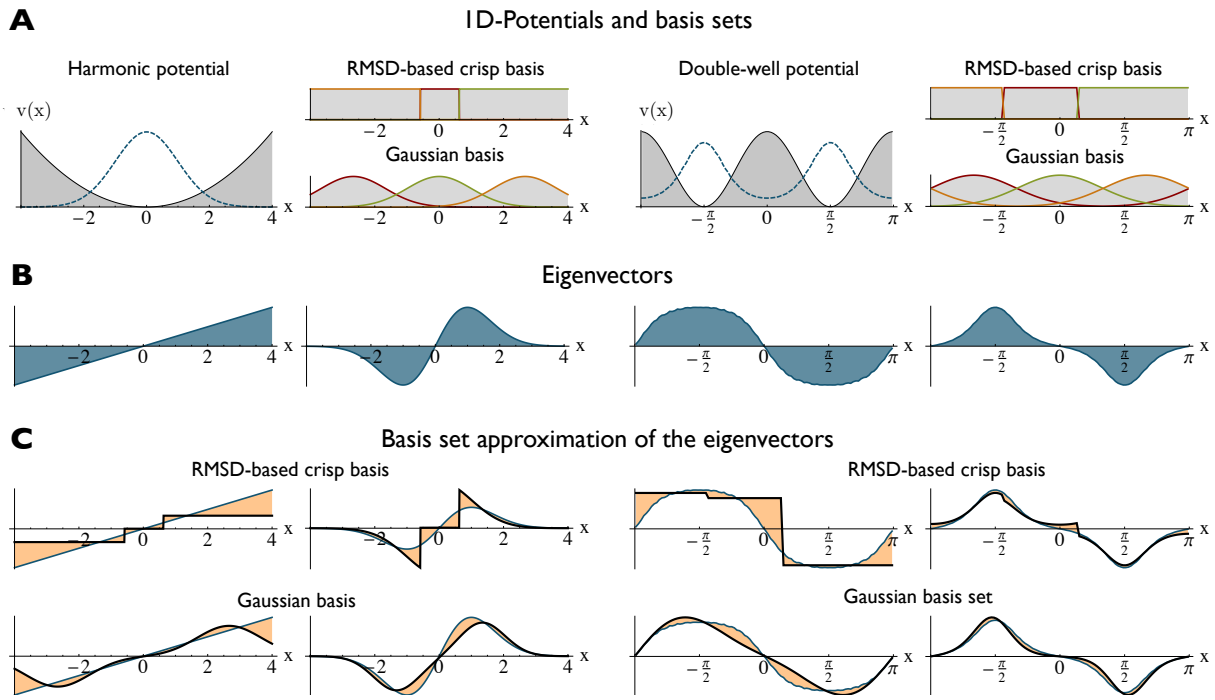

Figure 2: Illustration of the method with two one-dimensional potentials, the harmonic potential in the left half and a periodic double-well potential in the right half of the figure. Panel A shows the potential v together with its invariant distribution π (shaded) next to two possible choices of basis functions: A three-element crisp basis and a set of three Gaussian functions. Panel B shows the exact right and left second eigenfunctions, $|r_2\rangle$ and $|l_2\rangle$. In Panel C, the approximation results for these second eigenfunctions obtained from the basis sets shown above are displayed.

the highest eigenvalues. In every step of the optimization, we select three out of four equidistributed centers between $-\pi$ and π, and one of eleven standard deviations between 0.04π and 0.4π. In this way, we obtain three Gaussian trial functions per coordinate, resulting in a full basis set of six functions. Having determined the parameters for both angles, we use the resulting trial functions to apply our method as before. A bootstrapping procedure is used to estimate the statistical uncertainty of the implied timescales.

Note that the variations of basis functions described here to find a "good" basis set could be conducted once for each amino acid (or short sequences of amino acids) for a given force field, and then be reused.

3.2.3 Markov models.

This time, we cluster the data into $n = 5, 6, 10, 15, 20, 30, 50$ clusters, again using the k-means algorithm. From these clustercenters, we build Markov models and estimate the eigenvalues and eigenvectors using the EMMA software.

3.3 Deca-alanine.

3.3.1 MD simulations.

We performed six 500 ns all-atom explicit solvent molecular-dynamics simulations of deca-alanine using the Amber03 force field. See appendix for the detailed simulation setup

3.3.2 Gaussian model.

As before, we use Gaussian basis functions which depend on the backbone dihedral angles of the peptide, which means that we now have a total of 18 internal coordinates. A pre-selection of the trial functions is performed for every coordinate independently, similar to the alanine dipeptide example. In order to keep the number of basis functions acceptably small, we select two trial functions per coordinate. As before, their centers are chosen from four equidistributed centers along the coordinate, and their standard deviations are chosen from eleven different values between 0.04π and 0.4π. We also build a second Gaussian model using five functions per coordinate, with equidistributed centers and standard deviations optimized from the same values as in the first model. Having determined the trial functions, we estimate the matrices $\mathbf{C}$ and $\mathbf{S}$ and compute the eigenvalues and eigenvectors, and again use bootstrapping to estimate uncertainties.

3.3.3 Markov models.

We construct two different Markov models from the dihedral angle data. The first is built using kmeans clustering with 1000 cluster centers on the full data set, whereas for the second, we divide the $\phi - \psi$ plane of every dihedral pair along the chain into three regions corresponding to the α-helix, β-sheet and left-handed α-helix conformation, see section 4.2. Thus, we have three discretization boxes for all dihedral pairs, which yields a total of 8^3 discrete states to which the trajectory points are assigned.

4 Results

We now turn to the results obtained for the four systems presented in the previous section.

4.1 One-dimensional potentials

The two one-dimensional systems are toy examples where all important properties are either analytically known or can be computed arbitrarily well from approximations. For the harmonic potential, the stationary distribution is just a Gaussian function

$$|\pi(x)\rangle = |l_1(x)\rangle = \frac{1}{\sqrt{2\pi}} \exp(-\frac{x^2}{2}). \tag{55}$$

The exact eigenvalues $\lambda_\alpha(\tau)$ are given by

$$\lambda_\alpha(\tau) = \exp(-(\alpha-1)\tau), \tag{56}$$

and the associated right eigenfunction r_α is given by the $(\alpha-1)$-th normalized Hermite polynomial

$$|r_\alpha(x)\rangle = |H_{\alpha-1}(x)\rangle \sim (-1)^{\alpha-1} \exp(\frac{x^2}{2}) \frac{d^{\alpha-1}}{dx^{\alpha-1}} \exp(-\frac{x^2}{2}). \tag{57}$$

The left halves of Figure 2.A and Figure 2.B show the harmonic potential and its stationary distribution, as well as the second right and left eigenfunction. The sign change of $|l_2\rangle$ indicates the oscillation around the potential minimum, which is the slowest equilibration process. Note, however, that there is no energy barrier in the system, i.e. this process is not metastable. On the right hand sides of Figure 2.A and Figure 2.B, we see the same for the periodic double-well potential. The invariant density is equal to the Boltzmann distribution, where the normalization constant was computed numerically. The second eigenfunction was computed by a very fine finite-element approximation of the corresponding Fokker-Planck equation, using 1000 linear elements. The slowest transition in the system is the crossing of the barrier between the left and right minimum. This is reflected in the characteristic sign change of the second eigenfunction. Figure 2.A and Figure 2.B also show two choices of

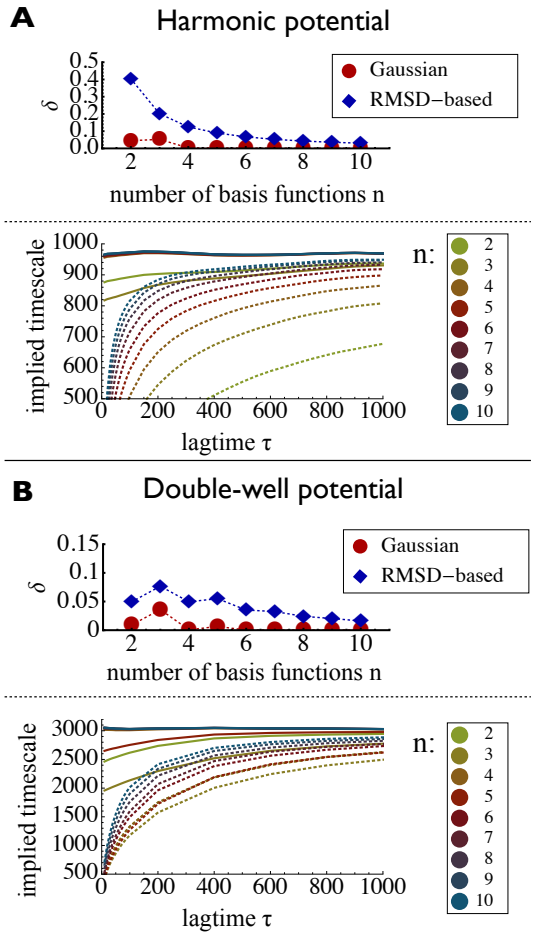

Figure 3: Analysis of the discretization error for both 1D-potentials. In the upper figure of both panels, we show the L^2-approximation error of the second eigenfunction from both crisp basis functions and Gaussian basis functions, dependent on the size of the basis set. The lower figures show the convergence of the second implied timescales $t_2(\tau)$ dependent on the lag time τ. Dotted lines represent the crips basis sets and solid lines the Gaussian basis sets. The colours indicate the size of the basis.

basis sets which can be used to approximate these eigenfunctions: A three element Gaussian basis set and a three state crisp set. The resulting estimates of the right and left eigenfunctions are displayed in Figure 2.C. Already with these small basis sets, a good approximation is achieved.

Let us analyze the approximation quality of both methods in more detail. To this end, we first compute the L^2-approximation error between the estimated second eigenfunction $\widehat{|r_2\rangle}$ and the exact solution $|r_2\rangle$, i.e. the integral

$$\delta = \int_X (|r_2\rangle(x) - \widehat{|r_2\rangle}(x))^2 \pi(x) \mathrm{d}x. \tag{58}$$

We expect this error to decay if the basis sets grow. Indeed, this is the case, as can be seen in the upper graphics of Figure 3.A and Figure 3.B, but the error produced by the Gaussian basis sets decays faster. Even for the ten state MSM, we still have a significant approximation error. Another important indicator is the implied timescale $t_\alpha(\tau)$, associated to the eigenvalue $\lambda_\alpha(\tau)$. It is the inverse rate of exponential decay of the eigenvalue, given by $t_\alpha(\tau) = -\frac{\tau}{\lambda_\alpha(\tau)}$ and corresponds to the equilibration time of the associated slow transition. The exact value of t_α is independent of the lag time τ. But if we estimate the timescale from the approximate eigenvalues, the estimate will be too small due to the variational principle. However, with increasing lag time, the error is expected to decay, as the approximation error also decays with the lag time. The faster this decay occurs, the better the approximation will be. In the lower graphics of Figure 3.A and Figure 3.B, we see the lag time dependence of the second timescale t_2 for growing crisp and Gaussian basis sets. We observe that it takes only four to five Gaussian basis functions to achieve much faster convergence compared even to a ten state Markov model. For 7 or more Gaussian basis functions, we achieve precise estimates even for very short lag times, which can not be achieved with Markov models with a reasonable number of states.

4.2 Alanine dipetide

Alanine dipeptide (Ac-Ala-NHMe, i.e. an alanine linked at either end to a protection group) is designed to mimic the dynamics of the amino acid alanine in a peptide chain. Unlike the previous examples, the eigenfunctions and eigenvalues of alanine dipeptide cannot be calculated directly from its potential energy function, but have to be estimated from simulations of its conformational dynamics. However, alanine dipeptide is a thoroughly studied system, many important properties are well-known, though their estimated values depend on the precise potential energy function (force field) used in the simulations. Most importantly, it is known that the dynamical behaviour can be essentially understood in terms of the two backbone dihedral angles ϕ and ψ: Figure 4.A shows the free energy landscape obtained from population inversion of the simulation, where white regions correspond to non-populated states. We find the three characteristic minima in the upper left, central left, and central right part of

the plane, which correspond to the β-sheet, α-helix and left-handed α-helix conformation of the amino acid. The two slowest transitions occur between the left half and the left handed α-helix, and from β-sheet to α-helix within the main well on the left, respectively.

Figure 4.B shows the weighted second and third eigenfunctions. They are obtained from applying our method with a total of six basis functions (3 for each dihedral), and from an MSM constructed from thirty clustercenters. The resulting estimates of $|r_2\rangle$ and $|r_3\rangle$ are then weighted with the population estimated from the trajectory, in order to emphasize the regions of phase space which are related to the structural transitions. Almost identical results are achieved, and the sign pattern of both approximations clearly indicates the aforementioned processes.

Lastly, in Figure 4.C, we again investigate the convergence of the slowest implied timescales. Different MSMs with a growing number of crisp basis functions (cluster centers) were used and compared to the six basis function Gaussian model. The colors indicate the number of basis functions used, the thinner lines correspond to the Markov models, whereas the thick solid line is obtained from the Gaussian model. In agreement with the previous results, we find that thirty or more crisp basis functions are needed to reproduce a similar approximation quality like a six-Gaussian basis set.

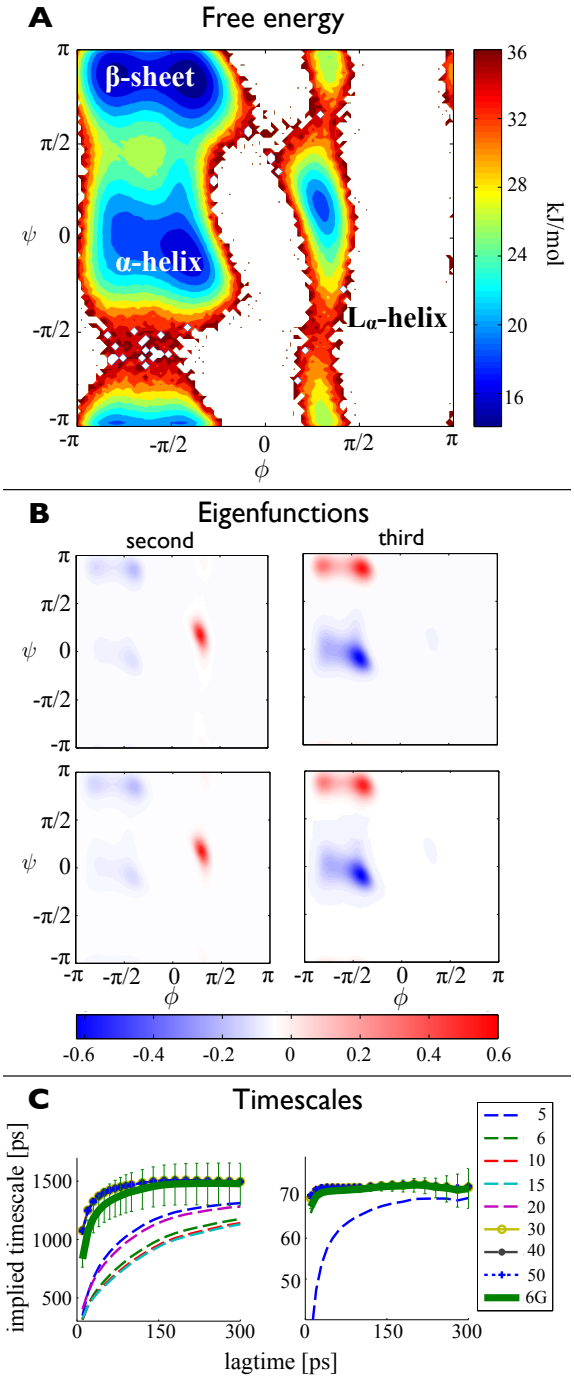

Figure 4: Illustration of the method using the 2D dihedral angle space (ϕ, ψ) of alanine dipeptide trajectory data. A) Free energy landscape obtained by direct population inversion of the trajectory data. B1 and B2) Color-coded contour plots of the second and third eigenfunctions of the propagator $(|l_2\rangle, |l_3\rangle)$, obtained by approximating the functions $|r_2\rangle$ and $|r_3\rangle$ by a Gaussian basis set with six functions, cf Eq. 48, and weighting the results with the estimated stationary distribution from A). C1 and C2) Color-coded contour plots of the second and third eigenfunctions of the propagator $(|l_2\rangle, |l_3\rangle)$, obtained by approximating the functions $|r_2\rangle$ and $|r_3\rangle$ by a Markov state model with thirty clustercenters, cf Eq. 48, and weighting the results with the estimated stationary distribution from A). D1 and D2) Convergence of implied timescales $t_\alpha(\tau)$ (in picoseconds) corresponding to the second and third eigenfunction, as obtained from Markov models using $n = 5, 6, 10, 15, 20, 30, 50$ clustercenters (thin lines), compared to the timescales obtained from the Gaussian model with a total of six basis functions (thick green line). Thin vertical bars indicate the error estimated by a bootstrapping procedure.

4.3 Deca alanine

As a third and last example, we study deca alanine, a small peptide which is about five times the size of alanine dipeptide. A sketch of the peptide is displayed in Figure 5.A.

The slow structural processes of deca alanine are less obvious compared to alanine dipeptide. The Amber03 force field used in our simulation produces a relatively fast transition between the elongated and the helical state of the system, with an associated timescale of 5 to 10 nanoseconds. As we can see in Figure 5.B, we are able to recover this slowest timescale with our method, t_2 converges to roughly 6.5 ns for both models. Comparing this to the two Markov models constructed from the same simulation data, we see that both yield slightly higher timescales: The k-means based MSM returns a value of about 8 ns and the finely discretized one ends up with 8.5 ns. Note that the underestimate of the present Gaussian basis set is systematic, and likely due to the fact that all basis functions were constructed as a function of single dihedral angles only, thereby neglecting the coupling between multiple dihedrals.

Despite this approximation, we are able to determine the correct structural transition. In order to analyse this, we evaluate the second eigenfunction $|r_2\rangle$, obtained from the smaller model, for all trajectory points, and plot a histogram of these values as displayed in Figure 5.C. We then select all frames which are within close distance of the peaks of that histogram, and produce overlays of these frames as shown underneath. Clearly, large negative values of the second eigenfunction indicate that the peptide is elongated, whereas large positive values indicate that the helical conformation is attained. This is in accord with a similar analysis of the second right Markov model eigenvector: In Figure 5.D, we show overlays of structures taken from states with the most negative and most positive values of the second eigenvector, and we find that the same transition is indicated, although the most negative values correspond to a slightly more bent arrangement of the system.

In summary, it is possible to use a comparatively small basis of 36 Gaussian functions to achieve results about the slowest structural transition which are comparable to those of MSMs constructed from about 1000 and 6500 discrete states, respectively. However, the differences in the timescales point to a weakness of the method: The fact that increasing the number of

basis functions does not alter the computed timescale indicates that coordinate correlation cannot be appropriately captured using sums of one-coordinate basis functions. In order to use the method for larger systems, we will have to study ways to overcome this problem.

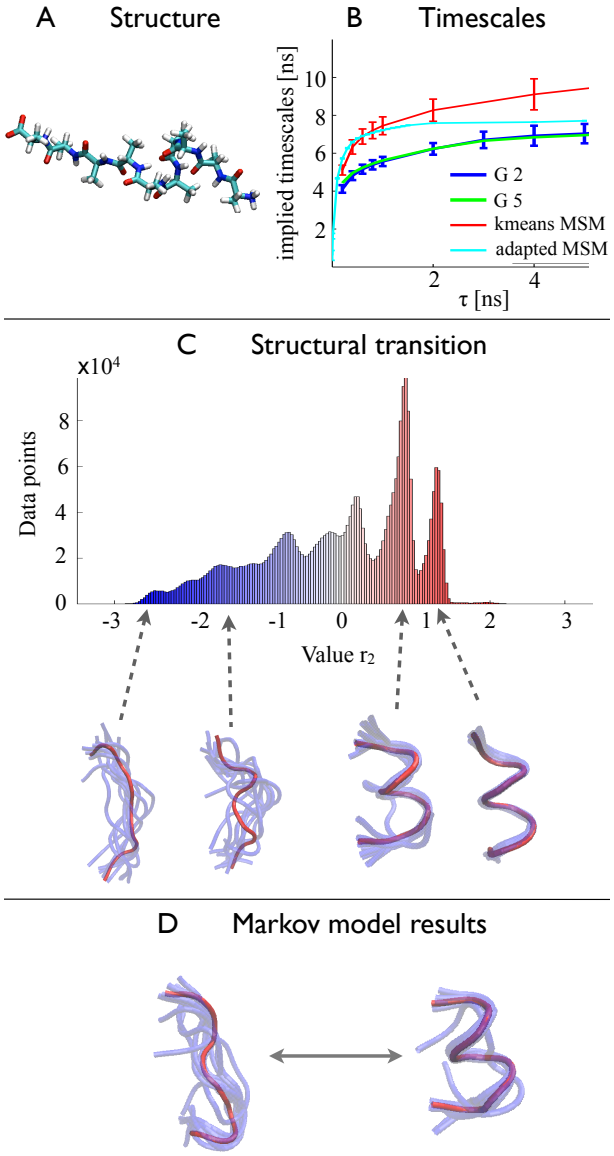

Figure 5: Illustration of the method using dihedral angle coordinates of the deca alanine molecule. A) Graphical representation of the system. B) Convergence of the estimated second implied timescale (in nanoseconds) depending on the lag time. We show the results of both Gaussian models and of both the kmeans based MSM and the adapted MSM. Thin vertical bars indicate the error estimated by a bootstrapping procedure. C) Assignment of representative structures for the second slowest process: The histogram shows how the values of the second estimated eigenfunction $|r_2\rangle$ of the smaller model are distributed over all simulation trajectories. Underneath, we show an overlay of structures taken at random from the vicinity of the peaks at -2.7, -1.6, 0.7 and 1.3. D) Overlays of structures corresponding to the most negative (left) and most positive (right) values of the second Markov model eigenvector, taken from the kmeans MSM.

5 Conclusions

We have presented a variational approach for computing the slow kinetics of biomolecules. This approach is analogous to the variational approach used for computing stationary states in quantum mechanics, but uses the molecular dynamics propagator (or transfer operator) rather than the quantum-mechanical Hamiltonian. A corresponding method of linear variation is formulated. Since the MD propagator is not analytically tractable for practically relevant cases, the matrix elements cannot be directly computed. Fortunately, these matrix elements can be shown to be correlation functions that can be estimated from simple MD simulations. The method proposed here is thus, to first define a basis set able to capture the relevant conformational dynamics, then compute the respective correlation matrices, and then to compute their dominant eigenvalues and eigenvectors, thus obtaining the key ingredients of the slow kinetics.

Markov state models (MSMs) are found to be a special case of the variational principle formulated here, namely for the case that indicator functions (also known as crisp sets or step functions) on the MSM clusters are used as a basis set.

We have applied the variational approach using Gaussian basis functions on a number of model examples, including one-dimensional diffusion systems and simulations of the alanine dipeptide and deca alanine in explicit solvent. Here we have used only one-dimensional basis sets that were constructed on single coordinates (e.g. dihedral angles), but it is clear that multidimensional basis functions could be straightforwardly used. Despite the simplicity of our bases, we could recover, and in most cases improve the results of n-state MSMs with much less than n basis functions in the applications shown here.

Note that practically all MSM approaches presented thus far use data-driven approaches to find the clusters on which these indicator functions are defined. Such a data-driven approach impairs the comparability of Markov state models of different simulations of the same system, and even more so of Markov state models of different systems. (Essentially, every Markov state model which has been published so far has been parametrized with respect to its own unique basis set). In contrast, the method proposed here allows to define basis sets which are in principle transferable between different molecular systems. This improves the compara-

bility of models made for different molecular systems. The second — and possibly decisive — advantage of the proposed method is that the basis sets can be chosen such that they reflect knowledge about the conformational dynamics or about the forcefield with which x_t has been simulated. It is thus conceivable that optimal basis sets are constructed for certain classes of small molecules or molecule fragments (e.g. amino acids or short amino acid sequences), and then combined for computing the kinetics of complex molecular systems.

As mentioned earlier, future work will have to focus on a systematic basis set selection and on an efficient use of multidimensional trial functions. Related to this is the question of model validation and error estimation. Due to the use of finite simulation data, use of a very fine basis set can lead to a growing statistical uncertainty of the estimated eigenvalues and eigenfunctions. In order to improve the basis set while balancing the model error and the statistical noise, a procedure to estimate this uncertainty is needed. While the special case of a Markov model allows for a solid error-theory based on the probabilistic interpretation of the model,[72] this is an open topic here and will have to be treated in the future.

Acknowledgements

The authors would like to thank Francesca Vitalini for providing the molecular dynamics simulation of alanine dipeptide.

References

(1) Alberts, B.; Johnson, A.; Lewis, J.; Raff, M.; Roberts, K.; Walter, P. *Molecular Biology of the Cell*, 5th ed.; Galand Science: New Yorck, 2008.

(2) Elber, R. *Simulations of allosteric transitions.*, 2011. `http://www.ncbi.nlm.nih.gov/pubmed/21333527`, accessed 9 January 2014.

(3) Verhey, K. J.; Kaul, N.; Soppina, V. *Annu. Rev. Biophys.* **2011**, *40*, 267–288.

(4) Dunkle, J. a.; Cate, J. H. D. *Annu. Rev. Biophys.* **2010**, *39*, 227–244.

(5) Keller, B.; Daura, X.; Van Gunsteren, W. F. *J. Chem. Phys.* **2010**, *132*, 074110.

(6) Krivov, S. V.; Karplus, M. *Proc. Nat. Acad. Sci. USA* **2004**, *101*, 14766–14770.

(7) Wales, D. J. *Energy Landscapes*, 1st ed.; Cambridge University Press, Cambridge, 2003.

(8) Noé, F.; Fischer, S. *Curr. Opin. Struc. Biol.* **2008**, *18*, 154–162.

(9) Karpen, M. E.; Tobias, D. J.; Brooks, C. L. *Biochemistry* **1993**, *32*, 412–420.

(10) Hubner, I. A.; Deeds, E. J.; Shakhnovich, E. I. *Proc. Natl. Acad. Sci. USA* **2006**, *103*, 17747–17752.

(11) Buchete, N.; Hummer, G. *J. Phys. Chem. B* **2008**, *112*, 6057–6069.

(12) Rao, F.; Caflisch, A. *J. Mol. Bio.* **2004**, *342*, 299–306.

(13) Muff, S.; Caflisch, A. *Proteins* **2007**, *70*, 1185–1195.

(14) de Groot, B.; Daura, X.; Mark, A.; Grubmüller, H. *J. Mol. Bio.* **2001**, *301*, 299–313.

(15) Schultheis, V.; Hirschberger, T.; Carstens, H.; Tavan, P. *J. Chem. Theory Comp.* **2005**, *1*, 515–526.

(16) Pan, A. C.; Roux, B. *J. Chem. Phys.* **2008**, *129*, 064107.

(17) Weber, M. *Improved Perron Cluster Analysis*; Technical Report 03-04, 2003.

(18) Noé, F.; Krachtus, D.; Smith, J. C.; Fischer, S. *J. Chem. Theory Comput.* **2006**, *2*, 840–857.

(19) Noé, F.; Oswald, M.; Reinelt, G.; Fischer, S.; Smith, J. C. *Multiscale Model. Simul.* **2006**, *5*, 393–419.

(20) Noé, F.; Horenko, I.; Schütte, C.; Smith, J. C. *J. Chem. Phys.* **2007**, *126*, 155102.

(21) Chodera, J. D.; Dill, K. A.; Singhal, N.; Pande, V. S.; Swope, W. C.; Pitera, J. W. *J. Chem. Phys.* **2007**, *126*, 155101.

(22) Swope, W. C.; Pitera, J. W.; Suits, F. *J. Phys. Chem. B* **2004**, *108*, 6571–6581.

(23) Swope, W. C.; Pitera, J. W.; Suits, F. *J. Phys. Chem. B* **2004**, *108*, 6571–6581.

(24) Buchete, N.-V.; Hummer, G. *J. Phys. Chem. B* **2008**, *112*, 6057–6069.

(25) Prinz, J.-H.; Wu, H.; Sarich, M.; Keller, B.; Senne, M.; Held, M.; Chodera, J. D.; Schütte, C.; Noé, F. *J. Chem. Phys.* **2011**, *134*, 174105.

(26) E, W.; Vanden-Eijnden, E. *J. Stat. Phys.* **2006**, *123*, 503–523.

(27) Noé, F.; Schütte, C.; Vanden-Eijnden, E.; Reich, L.; Weikl, T. R. *Proc. Natl. Acad. Sci. USA* **2009**, *106*, 19011–6.

(28) Deuflhard, P.; Weber, M. *Linear Algebra and its Applications* **2005**, *398*, 161–184.

(29) Kube, S.; Weber, M. *J. Chem. Phys.* **2007**, *126*, 024103.

(30) Chodera, J. D.; Singhal, N.; Pande, V. S.; Dill, K.; Swope, W. C. *J. Chem. Phys.* **2007**, *126*, 155101.

(31) Noé, F.; Horenko, I.; Schütte, C.; Smith, J. C. *J. Chem. Phys.* **2007**, *126*, 155102.

(32) Ruzhytska, S.; Jacobi, M. N.; Jensen, C. H.; Nerukh, D. *J. Chem. Phys.* **2010**, *133*, 164102.

(33) Bowman, G. R.; Pande, V. S. *Proc. Natl. Acad. Sci. USA* **2010**, *107*, 10890–10895.

(34) Noé, F.; Doose, S.; Daidone, I.; Löllmann, M.; Sauer, M.; Chodera, J. D.; Smith, J. C. *Proc. Natl. Acad. Sci. USA* **2011**, *108*, 4822–4827.

(35) Keller, B. G.; Prinz, J.-H.; Noé, F. *Chem. Phys.* **2012**, *396*, 92–107.

(36) Keller, B.; Hünenberger, P.; van Gunsteren, W. F. *J. Chem. Theory Comput.* **2011**, *7*, 1032–1044.

(37) Singhal, N.; Pande, V. S. *J. Chem. Phys.* **2005**, *123*, 204909.

(38) Noé, F. *J. Chem. Phys.* **2008**, *128*, 244103.

(39) Chodera, J. D.; Noé, F. *J. Chem. Phys.* **2010**, *133*, 105102.

(40) Bowman, G. R.; Beauchamp, K. A.; Boxer, G.; Pande, V. S. *J. Chem. Phys.* **2009**, *131*, 124101.

(41) Pande, V. S.; Beauchamp, K.; Bowman, G. R. *Methods* **2010**, *52*, 99–105.

(42) Beauchamp, K. A.; Bowman, G. R.; Lane, T. J.; Maibaum, L.; Haque, I. S.; Pande, V. S. *J. Chem. Theory Comput.* **2011**, *7*, 3412–3419.

(43) Senne, M.; Trendelkamp-Schroer, B.; Mey, A. S. J. S.; Schütte, C.; Noé, F. *J. Chem. Theory Comput.* **2012**, *8*, 2223–2238.

(44) Muff, S.; Caflisch, A. *Proteins: Struct. Funct. Bioinf.* **2007**, *70*, 1185–1195.

(45) Lane, T. J.; Bowman, G. R.; Beauchamp, K.; Voelz, V. a.; Pande, V. S. *J. Am. Chem. Soc.* **2011**, *133*, 18413–18419.

(46) Noé, F.; Schütte, C.; Vanden-Eijnden, E.; Reich, L.; Weikl, T. *Proc. Natl. Acad. Sci. USA* **2009**, *106*, 19011–19016.

(47) Beauchamp, K. A.; McGibbon, R.; Lin, Y.-S.; Pande, V. S. *Proc. Natl. Acad. Sci. USA* **2012**, *109*, 17807–17813.

(48) Shaw, D. E.; Maragakis, P.; Lindorff-Larsen, K.; Piana, S.; Dror, R. O.; Eastwood, M. P.; Bank, J. A.; Jumper, J. M.; Salmon, J. K.; Shan, Y.; Wriggers, W. *Science* **2010**, *330*, 341–346.

(49) Lindorff-Larsen, K.; Piana, S.; Dror, R. O.; Shaw, D. E. *Science* **2011**, *334*, 517–520.

(50) Sadiq, S. K.; Noé, F.; De Fabritiis, G. *Proc. Natl. Acad. Sci. USA* **2012**, *109*, 20449–20454.

(51) Buch, I.; Giorgino, T.; De Fabritiis, G. *Proc. Natl. Acad. Sci. USA* **2011**, *108*, 10184–10189.

(52) Kelley, N. W.; Vishal, V.; Krafft, G. A.; Pande, V. S. *J. Chem. Phys.* **2008**, *129*, 214707.

(53) Nerukh, D.; Jensen, C. H.; Glen, R. C. *J. Chem. Phys.* **2010**, *132*, 084104.

(54) Jensen, C. H.; Nerukh, D.; Glen, R. C. *J. Chem. Phys.* **2008**, *128*, 115107.

(55) Sarich, M.; Noé, F.; Schütte, C. *Multiscale Model. Simul.* **2010**, *8*, 1154–1177.

(56) Altis, A.; Nguyen, P. H.; Hegger, R.; Stock, G. *J. Chem. Phys.* **2007**, *126*, 244111.

(57) Schwantes, C.; Pande, V. *J. Chem. Theory Comput.* **2013**, *9*, 2000–2009.

(58) Pérez-Hernández, G.; Paul, F.; Giorgino, T.; De Fabritiis, G.; Noé, F. *J. Chem. Phys.* **2013**, *139*, 015102.

(59) Schütte, C.; Fischer, A.; Huisinga, W.; Deuflhard, P. *J. Comput. Phys.* **1999**, *151*, 146–168.

(60) Rains, E. K.; Andersen, H. C. *J. Chem. Phys.* **2010**, *133*, 144113.

(61) Weber, M. Ph.D. thesis, Freie Universitaet Berlin, 2006.

(62) Röblitz, S. Ph.D. thesis, Freie Universitaet Berlin, 2009.

(63) Haack, F.; Röblitz, S.; Scharkoi, O.; Schmidt, B. *AIP Conference Proceedings* **2010**, *1281*, 1585 – 1588.

(64) Schütte, C.; Noé, F.; Lu, J.; Sarich, M.; Vanden-Eijnden, E. *J. Chem. Phys.* **2011**, *134*, 204105.

(65) Noé, F.; Nüske, F. *SIAM Multiscale Model. Simul.* **2013**, *11*, 635–655.

(66) Szabo, A.; Ostlund, N. S. *Modern quantum chemistry*, 1st ed.; Dover Publications, Mineola, NY, 1996; pp 31 – 38.

(67) Noé, F.; Doose, S.; Daidone, I.; Löllmann, M.; Chodera, J.; Sauer, M.; Smith, J. *Proc. Natl. Acad. Sci. USA* **2011**, *108*, 4822–4827.

(68) Lindner, B.; Yi, Z.; Prinz, J.-H.; Smith, J.; Noé, F. *J. Chem. Phys* **2013**, *139*, 175101.

(69) Zheng, Y.; Lindner, B.; Prinz, J.-H.; Noé, F.; Smith, J. *J. Chem. Phys* **2013**, *139*, 175102.

(70) Vanden-Eijnden, E.; Venturoli, M. *J. Chem. Phys.* **2009**, *130*, 194101.

(71) Lindorff-Larsen, K.; Piana, S.; Palmo, K.; Maragakis, P.; Klepeis, J. L.; Dror, R. O.; Shaw, D. E. *Proteins* **2010**, *78*, 1950–1958.

(72) Prinz, J.-H.; Wu, H.; Sarich, M.; Keller, B.; Senne, M.; Held, M.; Chodera, J.; Schütte, C.; Noé, F. *J. Chem. Phys.* **2011**, *134*, 174105.

(73) Deuflhard, P.; Huisinga, W.; Fischer, A.; Schütte, C. *Linear Algebra and its Applications* **2000**, *315*, 39–59.

(74) MacCluer, C. R. *SIAM Review* **2000**, *42*, 487–498.

(75) Van Der Spoel, D.; Lindahl, E.; Hess, B.; Groenhof, G.; Mark, A. E.; Berendsen, H. J. C. *J. Comput. Chem.* **2005**, *26*, 1701–1718.

(76) Kritzer, J. A.; Tirado-Rives, J.; Hart, S. A.; Lear, J. D.; Jorgensen, W. L.; Schepartz, A. *J. Am. Chem. Soc.* **2005**, *127*, 167–178.

(77) Hess, B.; Bekker, H.; Berendsen, H. J. C.; Fraaije, J. G. E. M. *J. Comput. Chem.* **1997**, *18*, 1463–1472.

(78) Bussi, G.; Donadio, D.; Parrinello, M. *J. Chem. Phys* **2007**, *126*, 014101.

(79) Darden, T.; York, D.; Pedersen, L. *J. Chem. Phys* **1993**, *98*, 10089–10092.

Appendix

A Propagators of reversible processes

In the following, we explain in more detail the properties of the dynamical propagator $\mathscr{P}(\tau)$, as introduced in section 2.

A.1 Stationary density.

For any time-homogeneous propagator, there exists at least one stationary density $|\pi(x)\rangle$, which does not change under the action of the operator: $\mathscr{P}(\tau)\,|\pi(x)\rangle = |\pi(x)\rangle$. Another way of looking at this equation is to say that $|\pi(x)\rangle$ is an eigenfunction of $\mathscr{P}(\tau)$ with eigenvalue $\lambda_1 = 1$. It is guaranteed that $\pi(x) \geq 0$ everywhere as the transfer density is normalized. We additionally assume that $\pi(x) > 0$. In molecular systems, $\pi(x)$ is a Boltzmann density and $\pi(x) > 0$ is obtained when the temperature is nonzero and the energy is finite for all molecular configurations.

A.2 Bound eigenvalue spectrum.

The eigenvalue $\lambda_1 = 1$ always exists for any propagator. It is also the eigenvalue with the largest absolute value: $|\lambda_i| \leq 1$, i.e., the eigenvalue spectrum of $\mathscr{P}(\tau)$ is bound from above by the value 1. This is due to the fact that the transfer density is normalized

$$\int_X p(x,y;\tau)dy \quad = \quad 1\,, \tag{59}$$

i.e. the probability of going from state $x_t = x$ to anywhere in the state space (including x) during time τ has to be one.[73,74]

A.3 Ergodicity.

If the dynamics of the molecule are ergodic, then λ_1 is non-degenerate. As a consequence there is only one unique stationary density $|\pi(x)\rangle$ associated to $\mathscr{P}(\tau)$.

A.4 Reversibility.

If the dynamics of the individual molecules in the ensemble occur under equilibrium conditions, they fulfill reversibility (also sometimes called “detailed balance" or “micro-reversibility”) with respect to the stationary distribution π

$$\pi(x)p(x,y;\tau) \quad = \quad \pi(y)p(y,x;\tau) \;\; \forall x,y\,. \tag{60}$$

Eq. 60 implies that if the ensemble is in equilibrium, *i.e.* its systems are distributed over the state space according to $|\pi(x)\rangle$, the number of systems going from state x to state y during time τ is the same as the number of systems going from y to x. Or: the density flux from x to y is the same as in the opposite direction, and this is true for all state pairs $\{x,y\}$. For reversible processes, the stationary density becomes an equilibrium density and is equal to the Boltzmann distribution. In the following, we will only consider operators of reversible processes.

A consequence of reversibility is that λ_1 is the only eigenvalue with absolute value 1. Together with the previous properties, the eigenvalues can be sorted by their absolute value

$$|\lambda_1| = 1 > |\lambda_2| \geq |\lambda_3| ... \tag{61}$$

A.5 Self-adjoint operator.

Another consequence of reversibility is self-adjointness of the propagator, i.e.

$$\langle f\,|\,\mathscr{P}(\tau)\,|\,g\rangle_{\pi^{-1}} \quad = \quad \langle g\,|\,\mathscr{P}(\tau)\,|\,f\rangle_{\pi^{-1}}\,, \tag{62}$$

with respect to the weighted scalar product $\langle \cdot \,|\, \cdot \rangle_{\pi^{-1}}$

$$\langle f \,|\, g \rangle_{\pi^{-1}} \quad = \quad \int_X \overline{g(x)}\, \pi^{-1}(x)\, f(x) dx \,, \tag{63}$$

and the norm

$$|f| = \sqrt{\langle f \,|\, f \rangle_{\pi^{-1}}} \,, \tag{64}$$

where $\pi^{-1}(x) = 1/\pi(x)$ is the reciprocal function of $\pi(x)$ and the bar denotes complex conjugation. This is verified directly:

$$\begin{aligned}
\langle \mathscr{P}(\tau) f \,|\, g \rangle_{\pi^{-1}} &= \int_X \left[\int_X p(x,y,\tau) f(x)\, dx \right] \\
&\quad \times\ \pi^{-1}(y) g(y)\, dy && (65) \\
&= \int_X \left[\int_X p(y,x,\tau) \frac{\pi(y)}{\pi(x)} f(x)\, dx \right] \\
&\quad \times\ \pi^{-1}(y) g(y)\, dy && (66) \\
&= \int_X \int_X p(y,x,\tau) f(x) \\
&\quad \times\ \pi^{-1}(x) g(y)\, dy\, dx && (67) \\
&= \int_X f(x) \pi^{-1}(x) \\
&\quad \left[\int_X p(y,x,\tau) g(y)\, dy \right] dx && (68) \\
&= \langle f \,|\, \mathscr{P}(\tau)\, g \rangle_{\pi^{-1}} \,. && (69)
\end{aligned}$$

In the second line, we have used reversibility (Eq. 60) to replace $p(x,y,\tau)$ by $p(y,x,\tau)\frac{\pi(y)}{\pi(x)}$. Note that we could omit the complex conjugate in Eq. 63 because f, $\mathscr{P}(\tau)$, and g are real-valued functions. Self-adjointness of $\mathscr{P}(\tau)$ implies that its eigenvalues are real-valued, and its eigenfunctions form a complete basis of $\mathbb{R}^{3N}$, which is orthonormal with respect to the weighted scalar product $\langle \cdot \,|\, \cdot \rangle_{\pi^{-1}}$

$$\left\langle l_\alpha \,\middle|\, l_\beta \right\rangle_{\pi^{-1}} \quad = \quad \delta_{\alpha\beta} \,. \tag{70}$$

A.6 Comparison to the QM Hamilton operator.

With these properties of the propagator, Eq. 6 can be compared to the stationary Schrödinger equation $\mathscr{H}|\chi\rangle = E|\chi\rangle$. Both equations are eigenvalue equations of self-adjoint operators with a bound eigenvalue spectrum. The equations differ in some mathematical aspects: $\mathscr{P}(\tau)$ is an integral operator, whereas $\mathscr{H}$ is a differential operator; $\mathscr{P}(\tau)$ is self-adjoint with respect to a weighted scalar product, whereas $\mathscr{H}$ is self-adjoint with respect to the Euclidean scalar product. Also, they are not analogous in their physical interpretation. In contrast to the quantum-mechanical Hamilton operator, which acts on complex-valued wave functions, $\mathscr{P}(\tau)$ propagates real-valued probability densities. Moreover, the eigenfunctions of the propagator do not represent quantum states, such as the ground and excited states, they represent the stationary distribution and the perturbations to the stationary distribution from kinetic processes. Nonetheless, the mathematical structures of Eq. 6 and the stationary Schrödinger equation are similar enough that some methods which are applied in quantum chemistry can be reformulated for the propagator.

B Variational principle

The variational principle for propagators is derived and discussed in detail in.[65] We expand a trial function in terms of the eigenfunctions of $\mathscr{P}(\tau)$

$$|f\rangle \;=\; \sum_{\alpha} c_{\alpha}\,|l_{\alpha}\rangle\,, \tag{71}$$

where the αth expansion coefficients is given as

$$c_{\alpha} \;=\; \langle l_{\alpha}\,|\,f\rangle_{\pi^{-1}}\,. \tag{72}$$

The norm (Eq. 64) of the trial function $|f\rangle$ is then given as

$$\langle f\,|\,f\rangle_{\pi^{-1}} \;=\; \sum_{\alpha}\sum_{\beta} c_\alpha c_\beta \left\langle l_\alpha \,\middle|\, l_\beta \right\rangle_{\pi^{-1}} = \sum_{\alpha} c_\alpha^2 \,. \tag{73}$$

We therefore require that $|f\rangle$ is normalized

$$\langle f\,|\,f\rangle_{\pi^{-1}} \;=\; 1\,. \tag{74}$$

With this, an upper bound for the following expression can be found

$$\begin{aligned}
\langle f\,|\,\mathscr{P}(\tau)\,|\,f\rangle_{\pi^{-1}} &= \sum_{\alpha}\sum_{\beta} c_\alpha c_\beta \left\langle l_\alpha \,\middle|\, \mathscr{P}(\tau) \,\middle|\, l_\beta \right\rangle_{\pi^{-1}} && (75)\\
&= \sum_{\alpha}\sum_{\beta} c_\alpha c_\beta \lambda_\beta \left\langle l_\alpha \,\middle|\, l_\beta \right\rangle_{\pi^{-1}} && (76)\\
&= \sum_{\alpha} c_\alpha^2 \lambda_\alpha && (77)\\
&\leq \sum_{\alpha} c_\alpha^2 \lambda_1 = \langle f\,|\,f\rangle_{\pi^{-1}} \lambda_1 = 1, && (78)
\end{aligned}$$

and hence

$$\lambda_1 = 1 \;\geq\; \langle f\,|\,\mathscr{P}(\tau)\,|\,f\rangle_{\pi^{-1}} \,. \tag{79}$$

The above functional of any trial function is smaller than or equal to one, where the equality only holds if and only if $|f\rangle = |l_1\rangle$.

Furthermore, from the equations above it directly follows that for a function f_i that is orthogonal to eigenfunctions $|l_1\rangle, ..., |l_{i-1}\rangle$:

$$\left\langle f_i \,\middle|\, l_j \right\rangle_{\pi^{-1}} = 0 \;\; \forall j = 1, ..., i-1 \tag{80}$$

the variational principle results in

$$\langle f\,|\,\mathscr{P}(\tau)\,|\,f\rangle_{\pi^{-1}} \leq \lambda_i. \tag{81}$$

C Method of linear variation

Given the variational principle for the transfer operator (Eq. 79), the function $|f\rangle$ can be linearly expanded using a basis of n basis functions $\{|\varphi_i\rangle\}_{i=1}^{n}$

$$f = \sum_{i=1}^{n} a_i |\varphi_i\rangle , \tag{82}$$

where a_i are the expansion coefficients. All basis functions are real functions, but the basis set is not necessarily orthonormal. Hence, the expansion coefficients are real numbers. In the method of linear variation, the expansion coefficients a_i are varied such that the right-hand side of Eq. 79 becomes maximal, while the basis functions are kept constant. The derivation leads to matrix formulation of Eq. 6. Solving the corresponding matrix diagonalization problem, one obtains the first n eigenvectors of $\mathscr{P}(\tau)$ expressed in the basis $\{|\varphi_i\rangle\}_{i=1}^{n}$ and the associated eigenvalues. Inserting Eq. 16 into Eq. 79 one obtains

$$1 \geq \left\langle \sum_{i=1}^{n} a_i \varphi_i \middle| \mathscr{P} \middle| \sum_{j=1}^{n} a_j \varphi_j \right\rangle_{\pi^{-1}} \tag{83}$$

$$= \sum_{i,j=1}^{n} a_i a_j \left\langle \varphi_i \middle| \mathscr{P} \middle| \varphi_j \right\rangle_{\pi^{-1}} \tag{84}$$

$$= \sum_{i,j=1}^{n} a_i a_j C_{ij} , \tag{85}$$

where we have introduced the matrix element of the correlation matrix $\mathbf{C}$

$$C_{ij} = \left\langle \varphi_i \middle| \mathscr{P} \middle| \varphi_j \right\rangle_{\pi^{-1}} . \tag{86}$$

The maximum of the expression of right-hand side in Eq. 79 is found by varying the coefficients a_i, i.e.

$$\frac{\partial}{\partial a_k} \langle f | \mathscr{P} | f \rangle_{\pi^{-1}} = \frac{\partial}{\partial a_k} \sum_{ij=1}^{n} a_i a_j C_{ij} \tag{87}$$

$$= 0 \quad \forall\, k = 1, 2, \ldots n , \tag{88}$$

under the constraint that $|f\rangle$ is normalized

$$\begin{aligned} \langle f|f\rangle_{\pi^{-1}} &= \sum_{ij=1}^{n} a_i a_j \langle \varphi_i | \varphi_j \rangle_{\pi^{-1}} = \sum_{ij=1}^{n} a_i a_j S_{ij} && (89) \\ &= 1. && (90) \end{aligned}$$

S_{ij} is the matrix element of the overlap matrix $\mathbf{S}$ defined as

$$S_{ij} = \langle \varphi_i | \varphi_j \rangle_{\pi^{-1}} = \langle \varphi_j | \varphi_i \rangle_{\pi^{-1}} . \quad (91)$$

To incorporate the constraint in the optimization problem, we make use of the method of Lagrange multipliers

$$\begin{aligned} \mathscr{L} &= \sum_{ij=1}^{n} a_i a_j \langle \varphi_i | \mathscr{P} | \varphi_j \rangle_{\pi^{-1}} && (92) \\ &\quad -\lambda \left[\sum_{ij=1}^{n} a_i a_j \langle \varphi_i | \varphi_j \rangle_{\pi^{-1}} - 1 \right] && (93) \\ &= \sum_{ij=1}^{n} a_i a_j C_{ij} - \lambda \left[\sum_{ij=1}^{n} a_i a_j S_{ij} - 1 \right] . && (94) \end{aligned}$$

The variational problem then is

$$\begin{aligned} \frac{1}{2}\frac{\partial}{\partial a_k}\mathscr{L} &= \frac{1}{2}\sum_{j=1}^{n} a_j C_{ij} + \frac{1}{2}\sum_{i=1}^{n} a_i C_{ij} && (95) \\ &\quad -\frac{1}{2}\lambda \left[\sum_{j=1}^{n} a_j S_{ij} + \sum_{i=1}^{n} a_i S_{ij} \right] && (96) \\ &= \sum_{i=1}^{n} a_i C_{ij} - \lambda \sum_{i=1}^{n} a_i S_{ij} && (97) \\ &= 0 && (98) \\ &\quad \forall\, k = 1,2,...n, && (99) \end{aligned}$$

where, in the third line, we have used that $C_{ij} = C_{ji}$ and $S_{ij} = S_{ji}$ (Eqs. 62 and 91). Eq. 95 can be rewritten as a matrix equation

$$\mathbf{Ca} = \lambda \mathbf{Sa}, \quad (100)$$

which is a generalized eigenvalue problem, and identical to

$$\mathbf{S}^{-1}\mathbf{C}\mathbf{a} \;=\; \lambda\mathbf{a}\,, \tag{101}$$

where $\mathbf{a}$ is a vector which contains the coefficients a_i. The solutions of Eq. 101 are orthonormal with respect to an inner product which is weighted by the overlap matrix $\mathbf{S}$:

$$\left\langle \mathbf{a}^f \,\middle|\, S \,\middle|\, \mathbf{a}^g \right\rangle = \delta_{fg}\,, \tag{102}$$

where δ_{fg} is the Kronecker delta. Then, any two functions $f = \sum_i a_i^f \,|\varphi_i\rangle$ and $g = \sum_i a_i^g \,|\varphi_i\rangle$ are orthonormal with respect to the π^{-1}-weighted inner product, as it is expected for the eigenfunctions of the transfer operator

$$\begin{aligned}
\langle f \,|\, g\rangle_{\pi^{-1}} &= \left\langle \sum_i a_i^f \varphi_i \,\middle|\, \sum_j a_j^g \varphi_j \right\rangle_{\pi^{-1}} && (103)\\
&= \left\langle \mathbf{a}^f \,\middle|\, S \,\middle|\, \mathbf{a}^g \right\rangle && (104)\\
&= \delta_{fg}. && (105)
\end{aligned}$$

D Left eigenvectors and stationary properties

We want to show that the first "left" eigenvector $\mathbf{b}_1 = \mathbf{S}\mathbf{a}_1$ approximates the stationary distribution even for basis sets that do not form a partition of unity.

Let us assume we have a sequence of basis sets $\{\chi_i\}_j$, such that the corresponding first eigenvalue $\lambda_{1\,j}$ converges to one. Let us denote the local densities of basis set j by Z_i^j, the total density from Eq. 47 by C^j, and the entries of the normalized first left eigenvector of basis set j by b_i^j. We would like to show

$$b_i^j - \frac{Z_i^j}{C^j} \;\;\to\;\; 0 \tag{106}$$

as $j \to \infty$, or in other words

$$b_i^j C^j - Z_i^j \to 0. \tag{107}$$

To do so, we multiply by the inverse partition function $\frac{1}{Z}$ and rewrite this expression as:

$$\begin{aligned}\frac{1}{Z}(b_i^j C^j - Z_i^j) =& \frac{1}{Z}\frac{\sum_k a_k^{1j} s_{ik}^j}{\left(\sum_{l,k} a_k^{1j} s_{lk}^j\right)} \cdot \int \sum_l \chi_{lj} e^{-v(x)} \\ & - \frac{1}{Z}\int \chi_{ij} e^{-v(x)} \qquad (108) \\ =& \frac{\sum_k a_k^{1j} \langle \chi_{ij} \,|\, \chi_{kj} \rangle_\pi}{\sum_{l,k} a_k^{1j} \langle \chi_{lj} \,|\, \chi_{kj} \rangle_\pi} \cdot \left\langle \sum_l \chi_{lj} \,\middle|\, 1 \right\rangle_\pi \\ & - \langle \chi_{ij} \,|\, 1 \rangle_\pi . \qquad (109)\end{aligned}$$

We can use Eq. 48 to pull the summation over k into the second argument of the brackets:

$$\frac{1}{Z}(b_i^j C^j - Z_i^j) = \frac{\langle \chi_{ij} \,|\, r_{1j} \rangle_\pi}{\langle \sum_l \chi_{lj} \,|\, r_{1j} \rangle_\pi} \cdot \left\langle \sum_l \chi_{lj} \,\middle|\, 1 \right\rangle_\pi - \langle \chi_{ij} \,|\, 1 \rangle_\pi . \qquad (110)$$

From the convergence of the eigenvalue λ_{1j} towards one, it follows that the approximate first eigenfunction $|r_{1j}\rangle$ converges to the true first eigenfunction, the constant function with value one, in the scace L^2_π. This can be shown using an orthonormal basis expansion. Consequently, we can use the Cauchy-Schwarz inequality to estimate the expression

$$\begin{aligned}\left|\langle \chi_{ij} \,|\, r_{1j} \rangle_\pi - \langle \chi_{ij} \,|\, 1 \rangle_\pi\right| &= \left|\langle \chi_{ij} \,|\, r_{1j} - 1 \rangle_\pi\right| \qquad (111) \\ &\leq \|\chi_{ij}\| \|r_{1j} - 1\|. \qquad (112)\end{aligned}$$

As the second term tends to zero by the L^2-convergence, the complete expression likewise decays to zero, provided that the L^2-norms of the basis functions remain bounded, which is reasonable to assume. By a similar argument, we can show that the remaining fraction

$$\frac{\langle \sum_l \chi_{lj} \,|\, 1 \rangle_\pi}{\langle \sum_l \chi_{lj} \,|\, r_{1j} \rangle_\pi} \qquad (113)$$

converges to one, provided that the L^2-norm of the sum of all basis functions also remains bounded. Combining these two observations, we can conclude that Eq. 110 tends to zero, which was to be shown.

E Simulation setups

Alanine dipeptide. We performed all-atom molecular dynamics simulations of acetyl-alanine-methylamide (Ac-Ala-NHMe), referred to as alanine dipeptide in the text, in explicit water using the GROMACS 4.5.5[75] simulation package, the AMBER ff-99SB-ILDN force field,[71] and the TIP3P water model.[76] The simulations were performed in the canonical ensemble at a temperature of 300 K. The energy-minimized starting structure of Ac-Ala-NHMe was solvated into a cubic box with a minimum distance between solvent and box wall of 1 nm, corresponding to a box volume of 2.72 nm^3 and 651 water molecules. After an initial equilibration of 100 ps, 20 production runs of 200 ns each were performed, yielding a total simulation time of 4 μs. Covalent bonds to hydrogen atoms were constrained using the LINCS algorithm[77] (lincs_iter = 1, lincs_order = 4), allowing for an integration time step of 2 fs. The leap-frog integrator was used. The temperature was maintained by the velocity-rescale thermostat[78] with a time constant of 0.01 ps. Lennard-Jones interactions were cut off at 1 nm. Electrostatic interactions were treated by the Particle-Mesh Ewald (PME) algorithm[79] with a real space cutoff of 1 nm, a grid spacing of 0.15 nm, and an interpolation order of 4. Periodic boundary conditions were applied in the x, y, and z-direction. The trajectory data was stored every 1 ps.

Deca-alanine. We performed all-atom molecular dynamics simulations of deca alanine, which is protonated at the amino terminus and deprotonated at the carboxy terminus, using the GROMACS 4.5.5 simulation package,[75] the Amber03 force field and the TIP3P water model. A completely elongated conformation was chosen as an initial structure.

The structure was solvated in a cubic box of volume $V = 232.6\,\mathrm{nm}^3$, with 7647 pre-equilibrated TIP3P water molecules. First, an equilibration run of 500 ps in the NVT ensemble with full position restraints, using the velocity-rescale thermostat, was carried out. This was followed by a 500 ps NPT equilibration run. The temperature was set to $T = 300\,\mathrm{K}$. The equilibration run was followed by a 500 ns production run, again at $T = 300\,\mathrm{K}$. Two temperature coupling groups were used with a velocity-rescale thermostat and a time constant of 0.01 ps.[78] Periodic boundary conditions were applied in the x, y and z direction. For the long range electrostatic

interaction PME was used with a pme-order of 4 and a Fourier grid spacing of 0.15 nm. Covalent bonds to hydrogen bonds were constrained using the LINCS algorithm,[77] allowing for a 2 fs timestep. A leap frog integrator was used. Data was saved every 1 ps, resulting in $5 \cdot 10^5$ data frames. Six independent simulations from the same equilibrated configuration were carried out resulting in $3\,\mu$s total data.

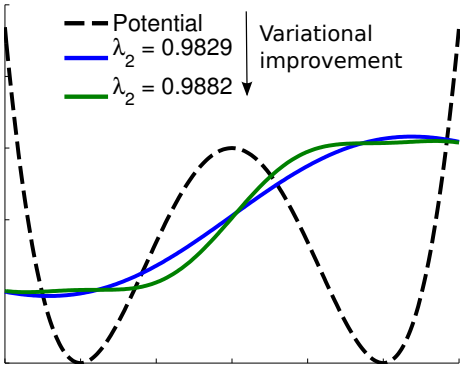

Figure 6: For Table of Contents Only.