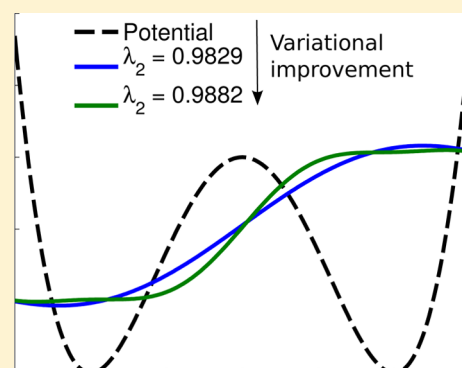


Variational Approach to Molecular Kinetics

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

Department for Mathematics and Computer Science, Fiere Universitat of Berlin, 14195 Berlin, Germany

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 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, that is, 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 that 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 cannot 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 that are far apart in some distance measure might be kinetically close.

A natural approach toward 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 time scales much longer than the fastest fluctuations of the molecule. In order to construct an MSM, the conformational space of the molecule is discretized into nonoverlapping 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 time scale 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 that are constructed using the internal degrees of freedom of the molecule tend to yield better results than those that were constructed using global descriptors of the structure (H-bond patterns, number of native contacts).³¹

Received: October 21, 2013

83 Also, degrees of freedom that are not included in the model should
84 decorrelate on short time scales from those that are included.³⁶
85 Naturally, the sampling of the transitions limits the accuracy of an
86 MSM, and tools to account for this error have been
87 developed.^{37–39} On the whole, the research field has matured to
88 a point at which well-tested protocols for the construction of
89 MSMs from MD data have been established,^{25,40,41} and software
90 to construct and validate Markov state models from MD data is
91 freely available.^{42,43} MSMs have been applied to analyze the
92 conformational dynamics of peptides^{5,31,44} and of small protein
93 domains, such as Villin head piece,⁴⁵ pin WW,⁴⁶ FiP35 WW,⁴⁵
94 Recently, it has become possible to analyze the folding
95 equilibria of full fast-folding proteins.^{47–49} MSMs have also
96 been used to investigate conformational changes, such as the
97 self-association step in the maturation of HIV-protease,⁵⁰
98 ligand binding,⁵¹ or the oligomerization of peptide fragments
99 into amyloid structures.⁵²

100 An important aspect that has limited the routine use of
101 MSMs is the difficulty to obtain a state space discretization that
102 will give rise to an MSM that precisely captures the slow
103 kinetics. The high-dimensional molecular space is usually first
104 discretized using clustering methods in some metric space. The
105 form and location of these clusters, sometimes called “MSM
106 microstates”, are crucial for determining the quality of the
107 estimated transition rates.^{53–55} Various metrics and clustering
108 methods have been attempted for different molecular systems.
109 Small peptides can be well described by a direct discretization
110 of their backbone dihedrals.³¹ It was suggested in ref 56 to use a
111 dihedral principal component analysis to reduce the dihedral
112 space to a low-dimensional subspace and subsequently cluster
113 this space using, for example, *k*-means. A rather general metric
114 is the pairwise minimal RMSD-metric in conjunction with some
115 clustering method, such as *k*-centers or *k*-medoids.^{25,30,41}
116 Recently, the time-lagged independent component analysis
117 (TICA) method was put forward, a dimension reduction
118 approach in which a “slow” low-dimensional subspace is
119 identified, which has been shown to provide improved MSMs
120 over previously employed metrics.^{57,58}

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

139 All of the above methods have in common that they attempt
140 to construct an appropriate discretization based on the
141 simulation data. This has a two-fold disadvantage: (1) different
142 simulation runs will produce different discretizations, making
143 them hard to compare; (2) data-based clusters have no intrinsic
144 meaning. Interpretation in terms of structural transitions must
145 be recovered by analyzing the molecular configurations

146 contained in specific clusters. With all of the above methods,
147 choosing an appropriate combination of the metric, the
148 clustering method, and the number and the location of clusters
149 or cores is still often a trial-and-error approach.

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

164 Besides its theoretical attractiveness, the variational approach
165 has some advantages over MSMs. First, the data-driven
166 discretization is replaced by a user-selection of an appropriate
167 basis set, typically of internal molecular coordinates. The
168 chosen basis set may reflect chemical intuition—for example,
169 basis functions may be predefined to fit known transition states
170 of backbone dihedral angles or formation/dissociation of
171 tertiary contacts between hydrophobically or electrostatically
172 interacting groups. As a result, one may obtain a precise model
173 with fewer basis functions needed than discrete MSM states.
174 Moreover, each basis function is associated with a chemical
175 meaning, and thus, the interpretation of the estimated
176 eigenfunctions becomes much more straightforward than for
177 MSMs. When using the same basis set for different molecular
178 systems of the same class, one obtains models that are directly
179 comparable in contrast to conventional MSMs. The represen-
180 tation of the propagator eigenfunctions can still be systemati-
181 cally improved by adding more basis functions or by varying the
182 basis set.

183 Our method is analogous to the method of linear variation
184 used in quantum chemistry.⁶⁶ The major difference is that the
185 propagator is self-adjoint with respect to a non-Euclidean scalar
186 product, whereas the Hamiltonian is self-adjoint with respect to
187 the Euclidean scalar product. The derivation of the method is
188 detailed in section 2 and Appendices A–C.

2. THEORY

2.1. **Dynamical Propagator.** Consider the conformational
189 space X of an arbitrary molecule consisting of N atoms, that is,
190 the $3N-6$ -dimensional space spanned by the internal degrees of
191 freedom of the molecule. The conformational dynamics of the
192 molecule in this space can be represented by a dynamical
193 process $\{x_t\}$, which samples at a given time t a particular point
194 $x_t \in X$. In this context, x_t is often called a trajectory. This
195 process is governed by the equations of motion, and it can be
196 simulated using standard molecular-dynamics programs. We
197 assume that an implementation of thermostatted molecular
198 dynamics is employed, which ensures that x_t is time-
199 homogeneous, Markovian, ergodic, and reversible with respect
200 to a unique stationary density (usually the Boltzmann
201 distribution). We introduce a propagator formulation of these
202 dynamics, following.⁶⁵ Readers familiar with this approach might
203 want to skip to section 2.2.
204

205 Next, consider an infinite ensemble of molecules of the same
206 type, distributed in the conformational space according to some

207 initial probability density $\rho_0(x)$. This initial probability density
 208 evolves in time in a definite manner that is determined by the
 209 aforementioned equations of motion for the individual
 210 molecules. We assume that the time evolution is Markovian

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

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

211 where τ is a finite time step, and $p(x, y; \tau)$ is the so-called
 212 transition density, which is assumed to be independent of time
 213 t (time-homogeneous). Figure 1 shows an example of the

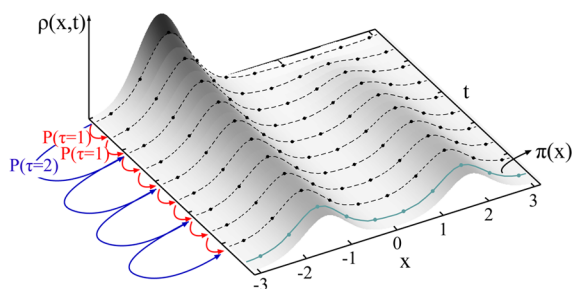


Figure 1. Illustration of two propagators acting on a probability density $\rho_i(x)$. Gray surface: time evolution of $\rho_i(x)$. Black dotted line: snapshots of $\rho_i(x)$. Cyan line: equilibrium density $\pi(x)$ to which $\rho_i(x)$ eventually converges. Red, blue: propagators with different lag times τ , which propagate an initial density by a time step τ in time.

214 time-evolution of a probability density in a one-dimensional
 215 two-well potential. Equation 2 implies that the probability of
 216 finding a molecule in conformation y dy at time $t + \tau$ depends
 217 only on the conformation x it has occupied one time step
 218 earlier, and not on the sequence of conformations it has visited
 219 before t . The unconditional probability density of finding a
 220 molecule in conformation y at time $t + \tau$ is obtained by
 221 integrating over all starting conformations x

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

223 This equation, in fact, defines an operator $\mathcal{P}(\tau)$ that propagates
 224 the probability density by a finite time step τ

$$\rho_{t+\tau}(x) = \mathcal{P}(\tau) \rho_t(x) \quad (4)$$

$$\rho_{t+n\tau}(x) = \mathcal{P}^n(\tau) \rho_t(x) \quad (5)$$

$\mathcal{P}(\tau)$ is called a propagator, and the time step τ is often called
 227 the lag time of the propagator. One says the propagator is
 228 parametrized with τ . Such as $p(x, y; \tau)$, the propagator $\mathcal{P}(\tau)$ in
 229 eq 5 is time-homogeneous; that is, it does not depend on t . The
 230 way it acts on a density $\rho(x, t)$ is not a function of the time t at
 231 which this density occurs but only a function of the time step τ
 232 by which the density is propagated (Figure 1).

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

$$\mathcal{P}(\tau) l_\alpha(x) = \lambda_\alpha l_\alpha(x) \quad (6)$$

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

space can be expressed as linear combination of $\{l_\alpha(x)\}$.
 Equation 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 that
 transport probability density from one part of the conforma-
 tional space to another and thus modulate the shape of the
 overall probability density. See ref 25 for a detailed explanation
 of the interpretation of eigenfunctions. The eigenvalues are
 linked to the time scales t_α on which the associated kinetic
 processes take place by

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

These time scales 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, $\mathcal{P}(\tau)$ is an operator with the
 following properties. A more detailed explanation can be
 found in Appendix A.

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

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

**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. First,
 for the exact eigenfunction $|l_\alpha(x)\rangle$, the following equality holds:

$$\langle l_\alpha | \mathcal{P}(\tau) | l_\alpha \rangle_{\pi^{-1}} = \lambda_\alpha(\tau) = e^{-\tau/t_\alpha} \quad (11)$$

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

Second, for any trial function $|f\rangle$ that is normalized according
 to eq 64, the following inequality holds:

$$\langle f | \mathcal{P}(\tau) | f \rangle_{\pi^{-1}} = \int_X f(x)\pi^{-1}(x)\mathcal{P}(\tau)f(x)dx \quad (12)$$

$$\leq \lambda_1 = 1 \quad (13)$$

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

Third, this inequality is applicable to other eigenfunctions: When $|f\rangle$ is orthogonal to the $\alpha - 1$ eigenfunctions, the variational principle will apply to the α -th eigenfunction/eigenvalue pair:

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

$$\langle f|\lambda_\beta\rangle_{\pi^{-1}} = 0 \quad \forall \beta = 1, \dots, \alpha - 1 \quad (15)$$

This variational principle allows to formulate the method of linear variation for the propagator. Again, the derivation detailed in ref 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 $\{|\varphi_i\rangle\}_{i=1}^n$

$$f = \sum_{i=1}^n a_i |\varphi_i\rangle \quad (16)$$

where a_i are the expansion coefficients. We only choose basis sets consisting of real-valued functions because all eigenvectors of $\mathcal{P}(\tau)$ 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:

$$\mathbf{C}\mathbf{a} = \lambda\mathbf{S}\mathbf{a} \quad (17)$$

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

$$C_{ij} = \langle \varphi_i|\mathcal{P}(\tau)|\varphi_j\rangle_{\pi^{-1}} \quad (18)$$

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

$$S_{ij} = \langle \varphi_i|\varphi_j\rangle_{\pi^{-1}} \quad (19)$$

Solving the generalized eigenvalue problem in eq 17, one obtains the first n eigenvectors of $\mathcal{P}(\tau)$ 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 \mathcal{H} (see Appendix A) are calculated directly with respect to the chosen basis, either analytically or by solving the integral $H_{ij} = \langle \varphi_i|\mathcal{H}|\varphi_j\rangle$ 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 cofunctions of the original basis set $\{\varphi_i\}$ by weighting the original functions with π^{-1}

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

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

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

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

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

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

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

or from a time-discretized time series x_t as

$$\widehat{\text{cor}}_T(\chi_i, \chi_j, \tau) = \frac{1}{N_T - n_\tau} \sum_{i=1}^{N_T - n_\tau} \chi_j(x_i)\chi_i(x_{i+n_\tau}) \quad (25)$$

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

$$C_{ij} = \text{cor}(\chi_i, \chi_j, \tau) = \lim_{T \rightarrow \infty} \widehat{\text{cor}}_T(\chi_i, \chi_j, \tau) \quad (26)$$

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 cofunctions of $\{\varphi_i\}$ (eq 20). This is the so-called transfer operator $\mathcal{J}(\tau)$.

$$C_{ij}(\tau) = \langle \chi_i|\pi|\mathcal{P}(\tau)|\pi\chi_j\rangle_{\pi^{-1}} \quad (27)$$

$$= \langle \chi_i|\mathcal{J}(\tau)|\chi_j\rangle_{\pi} \quad (28)$$

$$= \langle \chi_i|\mathcal{J}(\tau)|\chi_j\rangle_{\pi} \quad (28)$$

with

$$\mathcal{J}(\tau)|f(z)\rangle = \frac{1}{\pi(z)} \int_X p(y, z, \tau)\pi(y)f(y)dy \quad (29)$$

In particular, $\mathcal{J}(\tau)$ has the same eigenvalues as the propagator and its eigenfunctions are the cofunctions of the propagator eigenfunctions:

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

We will sometimes refer to the functions r_α as right eigenfunctions. For more details on the transfer operator the reader is referred to ref 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, that is, functions that have the value 1 on a particular set S_i of the conformational space X and the value 0 otherwise

$$\chi_i^{\text{MSM}}(x) = \begin{cases} 1 & \text{if } x \in S_i \\ 0 & \text{otherwise} \end{cases} \quad (31)$$

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

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

$$= \frac{z_{ij}}{N_T - n_\tau} \quad (33)$$

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:

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

Thus, the eigenvalue problem eq 17 becomes

$$\mathbf{C}\mathbf{a} = \lambda \mathbf{\Pi}\mathbf{a} \quad (35)$$

$$\mathbf{T}\mathbf{a} = \lambda \mathbf{a} \quad (36)$$

where \mathbf{C} is the correlation matrix, $\mathbf{\Pi} = \mathbf{S} = \text{diag}\{\pi_1, \dots, \pi_n\}$ 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

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

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

$$\mathbf{T}\mathbf{\Pi}^{-1}\mathbf{b} = \lambda \mathbf{\Pi}^{-1}\mathbf{b} \quad (38)$$

$$\mathbf{b}^T \mathbf{\Pi}^{-1} \mathbf{C} = \lambda \mathbf{b}^T \quad (39)$$

$$\mathbf{b}^T \mathbf{T} = \lambda \mathbf{b}^T \quad (40)$$

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 $\mathcal{J}(\tau)$. 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 $\mathbf{b}_1 = \mathbf{\Pi}\mathbf{a}_1$ fulfills the stationarity condition:

$$\mathbf{b}_1^T = \mathbf{b}_1^T \mathbf{T} \quad (41)$$

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 refs 24, 61, and 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 :

$$\pi_i = \frac{Z_i}{Z} \quad (42)$$

$$Z_i = \int_{S_i} e^{-v(x)} dx = \int_X \chi_i^{\text{MSM}}(x) e^{-v(x)} dx \quad (43)$$

$$\sum_i \pi_i = \sum_i \frac{Z_i}{Z} = 1 \quad (44)$$

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 :

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

Then, we still have

$$b_i = \frac{Z_i}{C} \quad (46)$$

for all i , where

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

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

$$r_\alpha = \sum_{i=1}^n a_i^\alpha |\chi_i\rangle \quad (48)$$

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

$$l_\alpha = \sum_{i=1}^n a_i^\alpha |\varphi_i\rangle = \sum_{i=1}^n a_i^\alpha |\pi\chi_i\rangle \quad (49)$$

3. METHODS

464 **3.1. One-Dimensional Diffusion Models.** 3.1.1. *Simu-*
 465 *lations.* We first consider two examples of one-dimensional
 466 diffusion processes x_t governed by Brownian dynamics. The
 467 process is then described by the stochastic differential equation

$$468 \quad dx_t = -\nabla v(x_t)dt + \sqrt{2D} dB_t \quad (50)$$

469 where v is the reduced potential energy (measured in units of
 470 $k_B T$, where k_B is the Boltzmann constant and T is the
 471 temperature), D is the diffusion constant, and dB_t denotes the
 472 differential of Brownian motion. For simplicity, we set all of the
 473 above constants equal to one. The potential function is given by
 474 the harmonic potential

$$475 \quad v(x) = 0.5x^2, \quad x \in \mathbb{R} \quad (51)$$

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

$$477 \quad v(x) = 1 + \cos(2x), \quad x \in [-\pi, \pi] \quad (52)$$

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

$$483 \quad x_{k+1} = x_k - \Delta t \nabla v(x_k) + \sqrt{2D\Delta t} y_t \quad (53)$$

$$484 \quad y_t \sim \mathcal{N}(0, 1) \quad (54)$$

485 In this way, we produced simulations of 5×10^6 time-steps
 486 for the harmonic potential and 10^7 time-steps for the double-
 487 well potential.

488 **3.1.2. Gaussian Model.** We apply our method with Gaussian
 489 basis functions to both problems. To this end, $n = 2, 3, \dots, 10$
 490 centers are chosen at uniform distance between $x = -4$ and $x =$
 491 4 for the harmonic potential and between $x = -\pi$ and $x = \pi$ for
 492 the double-well potential. In the latter case, the basis functions
 493 are modified to be periodic on $[-\pi, \pi]$. Subsequently, an
 494 “optimal” width of the Gaussians is picked by simply trying out
 495 several choices for the standard deviations between 0.4 and 1.0
 496 and using the one which yields the highest second eigenvalue.
 497 From this choice, the matrices \mathbf{C} and \mathbf{S} are estimated and the
 498 eigenvalues, functions, and implied time scales are computed.

499 **3.1.3. Markov Models.** As a reference for our methods, we
 500 also compute Markov state models for both processes. To this
 501 end, the simulation data is clustered into $n = 2, 3, \dots, 10$ disjoint
 502 clusters using the k -means algorithm. Subsequently, the EMMA
 503 software package⁴³ is used to estimate the MSM transition
 504 matrices and to compute eigenvalues and time scales.

505 **3.2. Alanine Dipeptide.** 3.2.1. *MD Simulations.* We
 506 performed 20 simulations of 200 ns of all-atom explicit solvent
 507 molecular dynamics of alanine dipeptide using the AMBER
 508 ff-99SB-ILDN force field.⁷¹ The detailed simulation setup is
 509 found in Appendix E.

510 **3.2.2. Gaussian Model.** Similar to the previous example, we
 511 use periodic Gaussian functions that only depend on one of the
 512 two significant dihedral angles of the system (see section 4.2)
 513 to apply our method. For both dihedrals, we separately perform
 514 a preselection of the Gaussian trial functions. To this end, we
 515 first project the data onto the coordinate, then we solve the
 516 projected optimization problem for all possible choices of
 517 centers and widths, and then pick the ones yielding the highest
 518 eigenvalues. In every step of the optimization, we select three
 519 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 time scales.

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 cluster-centers, 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 E for the detailed simulation setup.

3.3.2. Gaussian Model. As before, we use Gaussian basis
 functions that depend on the backbone dihedral angles of the
 peptide, which means that we now have a total of 18 internal
 coordinates. A preselection 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
 k -means clustering with 1000 cluster centers on the full data set,
 whereas for the second, we divide the ϕ – ψ 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

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

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

$$\lambda_\alpha(\tau) = \exp(-(\alpha - 1)\tau) \quad (56)$$

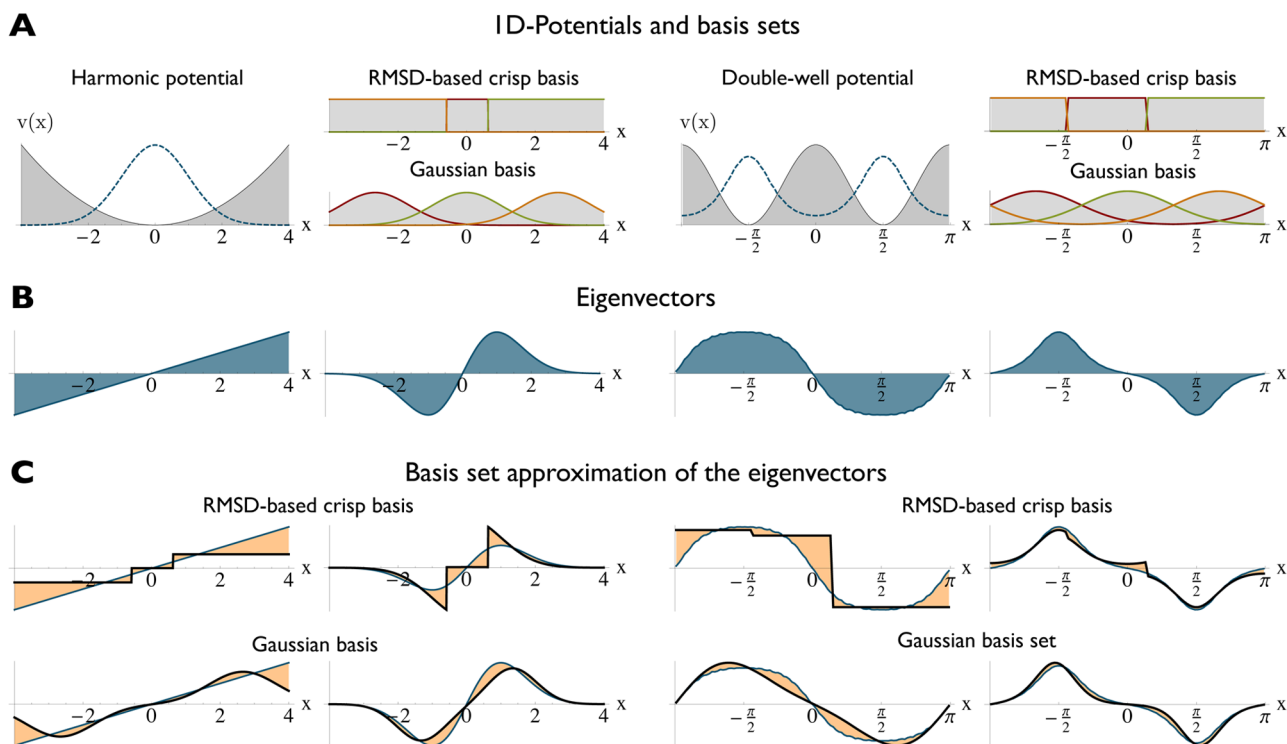


Figure 2. Illustration of the method with two one-dimensional potentials, the harmonic potential in the left half and a periodic double-well potential in the right half of the figure. (A) Potential v together with its invariant distribution π (shaded) next to two possible choices of basis functions: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

576 and the associated right eigenfunction r_α is given by the
577 $(\alpha - 1)$ -th normalized Hermite polynomial

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

578

579 The left halves of panels A and B in Figure 2 show the
580 harmonic potential and its stationary distribution, as well as the
581 second right and left eigenfunction. The sign change of $|l_2\rangle$
582 indicates the oscillation around the potential minimum, which
583 is the slowest equilibration process. Note, however, that there is
584 no energy barrier in the system; that is, this process is not
585 metastable. On the right-hand sides of parts A and B in Figure 2,
586 we see the same for the periodic double-well potential. The
587 invariant density is equal to the Boltzmann distribution, where the
588 normalization constant was computed numerically. The second
589 eigenfunction was computed by a very fine finite-element
590 approximation of the corresponding Fokker–Planck equation,
591 using 1000 linear elements. The slowest transition in the system is
592 the crossing of the barrier between the left and right minimum.
593 This is reflected in the characteristic sign change of the second
594 eigenfunction. Parts A and B of Figure 2 also show two choices of
595 basis sets that can be used to approximate these eigenfunctions: A
596 three element Gaussian basis set and a three state crisp set. The
597 resulting estimates of the right and left eigenfunctions are
598 displayed in Figure 2C. Already with these small basis sets, a
599 good approximation is achieved.

600 Let us analyze the approximation quality of both methods
601 in more detail. To this end, we first compute the
602 L^2 -approximation error between the estimated second
603 eigenfunction $|\widehat{r}_2\rangle$ and the exact solution $|r_2\rangle$, that is, the
604 integral

$$\delta = \int_X (|r_2\rangle(x) - |\widehat{r}_2\rangle(x))^2 \pi(x) dx \quad (58)$$

We expect this error to decay if the basis sets grow. Indeed, 605
this is the case, as can be seen in the upper graphics of Figure 606
3A and B, but the error produced by the Gaussian basis sets 607
decays faster. Even for the 10-state MSM, we still have a 608
significant approximation error. Another important indicator is 609
the implied time scale $t_\alpha(\tau)$, associated to the eigenvalue $\lambda_\alpha(\tau)$. 610
It is the inverse rate of exponential decay of the eigenvalue, 611
given by $t_\alpha(\tau) = -\tau/\lambda_\alpha(\tau)$ and corresponds to the equilibration 612
time of the associated slow transition. The exact value of t_α is 613
independent of the lag time τ . However, if we estimate the 614
time scale from the approximate eigenvalues, the estimate 615
will be too small due to the variational principle. However, 616
with increasing lag time, the error is expected to decay, as 617
the approximation error also decays with the lag time. The 618
faster this decay occurs, the better the approximation will 619
be. In the lower graphics of Figure 3A and B, we see the lag 620
time dependence of the second time scale t_2 for growing 621
crisp and Gaussian basis sets. We observe that it takes only 622
four to five Gaussian basis functions to achieve much faster 623
convergence compared even to a 10-state Markov model. 624
For seven or more Gaussian basis functions, we achieve 625
precise estimates even for very short lag times, which cannot 626
be achieved with Markov models with a reasonable number 627
of states. 628

4.2. Alanine Dipetide. Alanine dipetide (Ac-Ala-NHMe, 629
i.e. an alanine linked at either end to a protection group) is 630
designed to mimic the dynamics of the amino acid alanine in a 631
peptide chain. Unlike the previous examples, the eigenfunctions 632
and eigenvalues of alanine dipetide cannot be calculated 633
directly from its potential energy function but have to be 634

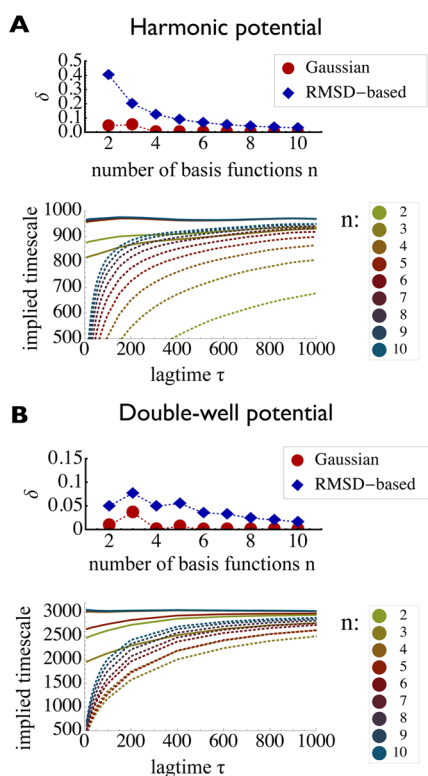


Figure 3. Analysis of the discretization error for both 1D-potentials. In the upper figure of both panels, we show the L^2 -approximation error of the second eigenfunction from both crisp basis functions and Gaussian basis functions, dependent on the size of the basis set. The lower figures show the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

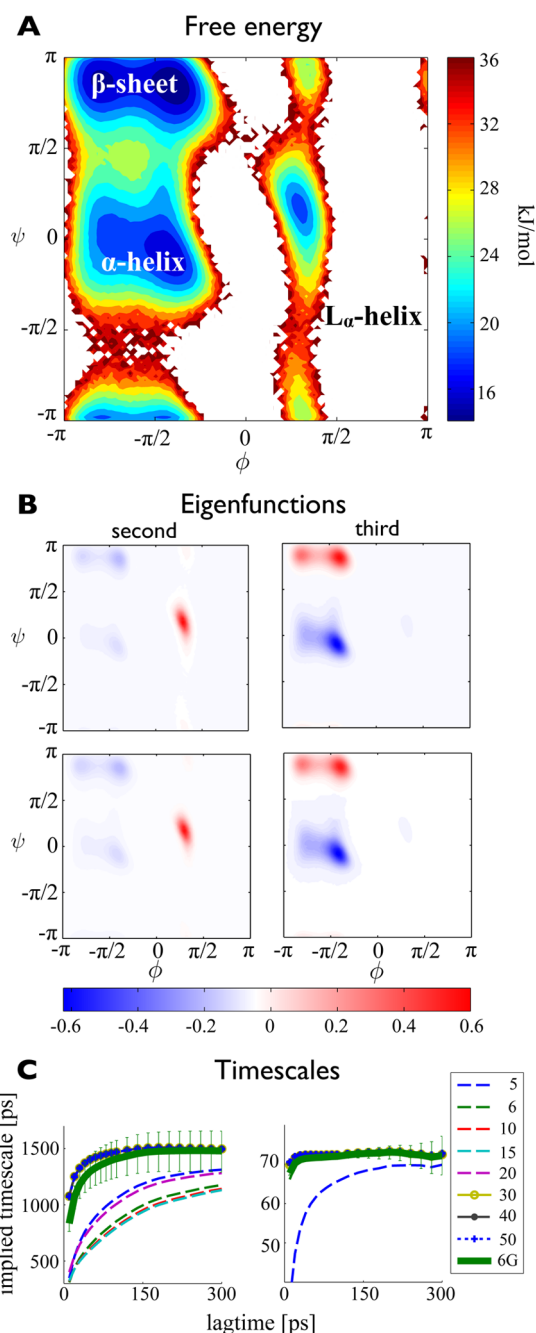


Figure 4. Illustration of the method using the 2D dihedral angle space (ϕ, ψ) of alanine dipeptide trajectory data. (A) Free energy landscape obtained by direct population inversion of the trajectory data. (B1 and B2) Color-coded contour plots of the second and third eigenfunctions of the propagator $(|l_2\rangle, |l_3\rangle)$, 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 part 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 30 cluster-centers, c.f. eq 48, and weighting the results with the estimated stationary distribution from part A. (D1 and D2) Convergence of implied time scales $t_a(\tau)$ (in picoseconds) corresponding to the second and third eigenfunction, as obtained from Markov models using $n = 5, 6, 10, 15, 20, 30, 50$ cluster-centers (thin lines), compared to the time scales 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.

635 estimated from simulations of its conformational dynamics.
 636 However, alanine dipeptide is a thoroughly studied system;
 637 many important properties are well-known, though their
 638 estimated values depend on the precise potential energy
 639 function (force field) used in the simulations. Most
 640 importantly, it is known that the dynamical behavior can
 641 essentially be understood in terms of the two backbone dihedral
 642 angles ϕ and ψ : Figure 4A shows the free energy landscape
 643 obtained from population inversion of the simulation, where
 644 white regions correspond to nonpopulated states. We find the
 645 three characteristic minima in the upper left, central left, and
 646 central right part of the plane, which correspond to the β -sheet,
 647 α -helix, and left-handed α -helix conformation of the amino
 648 acid. The two slowest transitions occur between the left half
 649 and the left handed α -helix, and from β -sheet to α -helix within
 650 the main well on the left, respectively.

651 Figure 4B shows the weighted second and third eigenfunc-
 652 tions. They are obtained from applying our method with a total
 653 of six basis functions (three for each dihedral), and from an
 654 MSM constructed from 30 cluster-centers. The resulting
 655 estimates of $|r_2\rangle$ and $|r_3\rangle$ are then weighted with the population
 656 estimated from the trajectory in order to emphasize the regions
 657 of phase space which are related to the structural transitions.
 658 Almost identical results are achieved, and the sign pattern of
 659 both approximations clearly indicates the aforementioned
 660 processes.

661 Lastly, in Figure 4C, we again investigate the convergence of
 662 the slowest implied time scales. Different MSMs with a growing

663 number of crisp basis functions (cluster-centers) were used and
 664 compared to the six basis function Gaussian model. The colors
 665 indicate the number of basis functions used; the thinner lines
 666 correspond to the Markov models, whereas the thick solid line
 667 is obtained from the Gaussian model. In agreement with the
 668 previous results, we find that 30 or more crisp basis functions
 669 are needed to reproduce an approximation quality similar to
 670 that of a six-Gaussian basis set.

671 **4.3. Deca-alanine.** As a third and last example, we study
 672 deca-alanine, a small peptide that is about five times the size
 673 of alanine dipeptide. A sketch of the peptide is displayed in
 674 Figure 5A.

675 The slow structural processes of deca-alanine are less obvious
 676 compared to alanine dipeptide. The Amber03 force field used
 677 in our simulation produces a relatively fast transition between

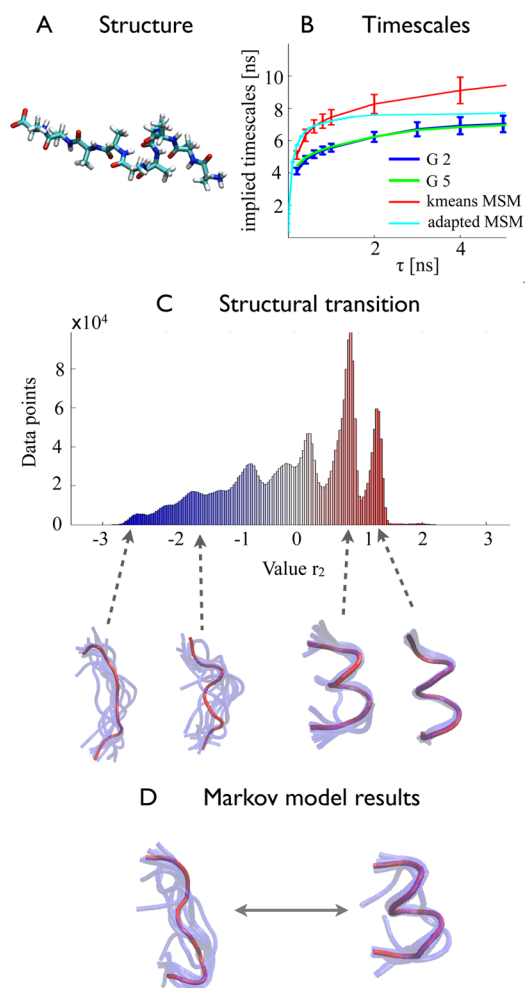


Figure 5. Illustration of the method using dihedral angle coordinates of the deca alanine molecule. (A) Graphical representation of the system. (B) Convergence of the estimated second implied time scale (in nanoseconds) depending on the lag time. We show the results of both Gaussian models and of both 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 *k*-means MSM.

the elongated and the helical state of the system, with an
 associated time scale of 5–10 ns. As we can see in Figure 5B,
 we are able to recover this slowest time scale 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 time
 scales: 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,
 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 analyze 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 5C. We then select all frames
 that 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 5D, we show overlays of structures taken
 from states with the most negative and most positive values of
 the second eigenfunction, 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 time scales point
 to a weakness of the method: The fact that increasing the
 number of basis functions does not alter the computed time
 scale indicates that coordinate correlation cannot be appropri-
 ately 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.

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 it 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 that 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 that are, in principle, transferable between different molecular systems. This improves the comparability 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.

APPENDIX A

Propagators of Reversible Processes

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

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: $\mathcal{P}(\tau)|\pi(x)\rangle = |\pi(x)\rangle$. Another way of looking at this equation is to say that $|\pi(x)\rangle$ is an eigenfunction of $\mathcal{P}(\tau)$ 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.

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

$$\int_X p(x, y, \tau) dy = 1 \quad (59)$$

That is, the probability of going from state $x_t = x$ to anywhere in the state space (including x) during time τ has to be 1.^{73,74}

Ergodicity. If the dynamics of the molecule are ergodic, then λ_1 is nondegenerate. As a consequence, there is only one unique stationary density $\pi(x)$ associated to $\mathcal{P}(\tau)$.

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

$$\pi(x)p(x, y; \tau) = \pi(y)p(y, x; \tau) \quad \forall x, y \quad (60)$$

Equation 60 implies that if the ensemble is in equilibrium, that is, 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

$$|\lambda_1| = 1 > |\lambda_2| \geq |\lambda_3| \dots \quad (61)$$

Self-adjoint Operator. Another consequence of reversibility is self-adjointness of the propagator, that is,

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

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

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

and the norm

$$|f| = \sqrt{\langle f | f \rangle_{\pi^{-1}}} \quad (64)$$

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

$$\langle \mathcal{P}(\tau) f | g \rangle_{\pi^{-1}} = \int_X \left[\int_X p(x, y, \tau) f(x) dx \right] \pi^{-1}(y) g(y) dy \quad (65)$$

$$= \int_X \left[\int_X p(y, x, \tau) \frac{\pi(y)}{\pi(x)} f(x) dx \right] \pi^{-1}(y) g(y) dy \quad (66)$$

$$= \int_X \int_X p(y, x, \tau) f(x) \pi^{-1}(x) g(y) dy dx \quad (67)$$

$$= \int_X f(x) \pi^{-1}(x) \left[\int_X p(y, x, \tau) g(y) dy \right] dx \quad (68)$$

$$= f | \mathcal{P}(\tau) g \rangle_{\pi^{-1}} \quad (69)$$

In the second line, we have used reversibility (eq 60) to replace $p(x, y, \tau)$ by $p(y, x, \tau)\pi(y)/\pi(x)$. Note that we could omit the complex conjugate in eq 63 because $f, \mathcal{P}(\tau)$, and g are real-valued functions. Self-adjointness of $\mathcal{P}(\tau)$ 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 $\langle \cdot | \cdot \rangle_{\pi^{-1}}$

$$\langle |_\alpha | |_\beta \rangle_{\pi^{-1}} = \delta_{\alpha\beta} \quad (70)$$

Comparison to the QM Hamilton Operator. With these properties of the propagator, eq 6 can be compared to the

stationary Schrödinger equation $\mathcal{H}|\chi\rangle = E|\chi\rangle$. Both equations are eigenvalue equations of self-adjoint operators with a bound eigenvalue spectrum. The equations differ in some mathematical aspects: $\mathcal{P}(\tau)$ is an integral operator, whereas \mathcal{H} is a differential operator; $\mathcal{P}(\tau)$ is self-adjoint with respect to a weighted scalar product, whereas \mathcal{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, $\mathcal{P}(\tau)$ 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.

APPENDIX B

Variational Principle

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

$$|f\rangle = \sum_{\alpha} c_{\alpha} |l_{\alpha}\rangle \quad (71)$$

where the α th expansion coefficients is given as

$$c_{\alpha} = \langle l_{\alpha} | f \rangle_{\pi^{-1}} \quad (72)$$

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

$$\langle f | f \rangle_{\pi^{-1}} = \sum_{\alpha} \sum_{\beta} c_{\alpha} c_{\beta} \langle l_{\alpha} | l_{\beta} \rangle_{\pi^{-1}} = \sum_{\alpha} c_{\alpha}^2 \quad (73)$$

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

$$\langle f | f \rangle_{\pi^{-1}} = 1 \quad (74)$$

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

$$\langle f | \mathcal{P}(\tau) | f \rangle_{\pi^{-1}} = \sum_{\alpha} \sum_{\beta} c_{\alpha} c_{\beta} \langle l_{\alpha} | \mathcal{P}(\tau) | l_{\beta} \rangle_{\pi^{-1}} \quad (75)$$

$$= \sum_{\alpha} \sum_{\beta} c_{\alpha} c_{\beta} \lambda_{\beta} \langle l_{\alpha} | l_{\beta} \rangle_{\pi^{-1}} \quad (76)$$

$$= \sum_{\alpha} c_{\alpha}^2 \lambda_{\alpha} \quad (77)$$

$$\leq \sum_{\alpha} c_{\alpha}^2 \lambda_1 = \langle f | f \rangle_{\pi^{-1}} \lambda_1 = 1 \quad (78)$$

and hence

$$\lambda_1 = 1 \geq \langle f | \mathcal{P}(\tau) | f \rangle_{\pi^{-1}} \quad (79)$$

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 $|l_1\rangle, \dots, |l_{i-1}\rangle$:

$$\langle f_i | l_j \rangle_{\pi^{-1}} = 0 \quad \forall j = 1, \dots, i-1 \quad (80)$$

the variational principle results in

$$\langle f | \mathcal{P}(\tau) | f \rangle_{\pi^{-1}} \leq \lambda_i \quad (81)$$

APPENDIX 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 $\{|\varphi_i\rangle\}_{i=1}^n$

$$f = \sum_{i=1}^n a_i |\varphi_i\rangle \quad (82)$$

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 $\mathcal{P}(\tau)$ expressed in the basis $\{|\varphi_i\rangle\}_{i=1}^n$ and the associated eigenvalues. Inserting eq 16 into eq 79 obtains

$$1 \geq \left\langle \sum_{i=1}^n a_i \varphi_i | \mathcal{P} | \sum_{j=1}^n a_j \varphi_j \right\rangle_{\pi^{-1}} \quad (83)$$

$$= \sum_{i,j=1}^n a_i a_j \langle \varphi_i | \mathcal{P} | \varphi_j \rangle_{\pi^{-1}} \quad (84)$$

$$= \sum_{i,j=1}^n a_i a_j \langle \varphi_i | \mathcal{P} | \varphi_j \rangle_{\pi^{-1}} \quad (85)$$

where we have introduced the matrix element of the correlation matrix **C**

$$C_{ij} = \langle \varphi_i | \mathcal{P} | \varphi_j \rangle_{\pi^{-1}} \quad (86)$$

The maximum of the expression of right-hand side in eq 79 is found by varying the coefficients a_i , that is,

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

$$= 0 \quad \forall k = 1, 2, \dots, n \quad (88)$$

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

$$\langle f | f \rangle_{\pi^{-1}} = \sum_{ij=1}^n a_i a_j \langle \varphi_i | \varphi_j \rangle_{\pi^{-1}} = \sum_{ij=1}^n a_i a_j S_{ij} \quad (89)$$

$$= 1 \quad (90)$$

S_{ij} is the matrix element of the overlap matrix **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

$$\mathcal{L} = \sum_{ij=1}^n a_i a_j \langle \varphi_i | \mathcal{P} | \varphi_j \rangle_{\pi^{-1}} \quad (92)$$

$$- \lambda \left[\sum_{ij=1}^n a_i a_j \langle \varphi_i | \varphi_j \rangle_{\pi^{-1}} - 1 \right] \quad (93)$$

$$= \sum_{ij=1}^n a_i a_j C_{ij} - \lambda \left[\sum_{ij=1}^n a_i a_j S_{ij} - 1 \right] \quad (94)$$

910 The variational problem then is

$$\frac{1}{2} \frac{\partial}{\partial a_k} \mathcal{L} = \frac{1}{2} \sum_{j=1}^n a_j C_{ij} + \frac{1}{2} \sum_{i=1}^n a_i C_{ij} \quad (95)$$

$$- \frac{1}{2} \lambda \left[\sum_{j=1}^n a_j S_{ij} + \sum_{i=1}^n a_i S_{ij} \right] \quad (96)$$

$$= \sum_{i=1}^n a_i C_{ij} - \lambda \sum_{i=1}^n a_i S_{ij} \quad (97)$$

$$= 0 \quad (98)$$

$$\forall k = 1, 2, \dots, n \quad (99)$$

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

$$914 \quad \mathbf{Ca} = \lambda \mathbf{Sa} \quad (100)$$

915 which is a generalized eigenvalue problem, and identical to

$$916 \quad \mathbf{S}^{-1} \mathbf{Ca} = \lambda \mathbf{a} \quad (101)$$

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

$$920 \quad \langle \mathbf{a}^f | \mathbf{S} | \mathbf{a}^g \rangle = \delta_{fg} \quad (102)$$

921 where δ_{fg} is the Kronecker delta. Then, any two functions $f =$
922 $\sum_i a_i^f | \varphi_i \rangle$ and $g = \sum_i a_i^g | \varphi_i \rangle$ are orthonormal with respect to the
923 π^{-1} -weighted inner product, as it is expected for the
924 eigenfunctions of the transfer operator

$$\langle f | g \rangle_{\pi^{-1}} = \left\langle \sum_i a_i^f \varphi_i \left| \sum_j a_j^g \varphi_j \right. \right\rangle_{\pi^{-1}} \quad (103)$$

$$= \langle \mathbf{a}^f | \mathbf{S} | \mathbf{a}^g \rangle \quad (104)$$

$$= \delta_{fg} \quad (105)$$

925 ■ APPENDIX D

926 Left Eigenvectors and Stationary Properties

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

930 Let us assume we have a sequence of basis sets $\{\chi_i\}_p$ such
931 that the corresponding first eigenvalue λ_{1j} converges to 1. Let us
932 denote the local densities of basis set j by Z_j^i , the total density
933 from eq 47 by C^j , and the entries of the normalized first left
934 eigenvector of basis set j by b_i^j . We show

$$935 \quad b_i^j - \frac{Z_i^j}{C^j} \rightarrow 0 \quad (106)$$

as $j \rightarrow \infty$, or in other words, 936

$$b_i^j C^j - Z_i^j \rightarrow 0 \quad (107) \quad 937$$

To do so, we multiply by the inverse partition function $1/Z$ 938
and rewrite this expression as 939

$$\frac{1}{Z} (b_i^j C^j - Z_i^j) = \frac{1}{Z} \frac{\sum_k a_k^j s_{ik}^j}{(\sum_{l,k} a_k^j s_{lk}^j)} \int \sum_l \chi_{lj} e^{-v(x)} - \frac{1}{Z} \int \chi_{ij} e^{-v(x)} \quad (108)$$

$$= \frac{\sum_k a_k^j \chi_{kj} | \chi_{kj} \rangle_{\pi}}{\sum_{l,k} a_k^j \chi_{lj} | \chi_{lj} \rangle_{\pi}} \langle \sum_l \chi_{lj} | 1 \rangle_{\pi} - \langle \chi_{ij} | 1 \rangle_{\pi} \quad (109)$$

We can use eq 48 to pull the summation over k into the second 940
argument of the brackets: 941

$$\frac{1}{Z} (b_i^j C^j - Z_i^j) = \frac{\langle \chi_{ij} | r_{1j} \rangle_{\pi}}{\langle \sum_l \chi_{lj} | 1 \rangle_{\pi}} \langle \sum_l \chi_{lj} | 1 \rangle_{\pi} - \langle \chi_{ij} | 1 \rangle_{\pi} \quad (110) \quad 942$$

From the convergence of the eigenvalue λ_{1j} toward 1, it 943
follows that the approximate first eigenfunction $| r_{1j} \rangle_{\pi}$ converges 944
to the true first eigenfunction, the constant function with value 945
one, in the space L^2_{π} . This can be shown using an orthonormal 946
basis expansion. Consequently, we can use the Cauchy– 947
Schwarz inequality to estimate the expression 948

$$|\langle \chi_{ij} | r_{1j} \rangle_{\pi} - \langle \chi_{ij} | 1 \rangle_{\pi}| = |\langle \chi_{ij} | r_{1j} - 1 \rangle_{\pi}| \quad (111) \quad 949$$

$$\leq \| \chi_{ij} \| \| r_{1j} - 1 \| \quad (112) \quad 950$$

As the second term tends to zero by the L^2 -convergence, the 951
complete expression likewise decays to zero, provided that the 952
 L^2 -norms of the basis functions remain bounded, which is 953
reasonable to assume. By a similar argument, we can show that 954
the remaining fraction 955

$$\frac{\langle \sum_l \chi_{lj} | 1 \rangle_{\pi}}{\langle \sum_l \chi_{lj} | r_{1j} \rangle_{\pi}} \quad (113) \quad 956$$

converges to 1, provided that the L^2 -norm of the sum of all 957
basis functions also remains bounded. Combining these two 958
observations, we can conclude that eq 110 tends to 0, which 959
was to be shown. 960

961 ■ APPENDIX E

Simulation Setups 962

Alanine dipeptide. We performed all-atom molecular 963
dynamics simulations of acetyl-alanine-methylamide (Ac-Ala- 964
NHMe), referred to as alanine dipeptide in the text, in explicit 965
water using the GROMACS 4.5.5⁷⁵ simulation package, the 966
AMBER ff-99SB-ILDN force field,⁷¹ and the TIP3P water 967
model.⁷⁶ The simulations were performed in the canonical 968
ensemble at a temperature of 300 K. The energy-minimized 969
starting structure of Ac-Ala-NHMe was solvated into a cubic 970
box with a minimum distance between solvent and box wall of 971
1 nm, corresponding to a box volume of 2.72 nm³ and 651 water 972
molecules. After an initial equilibration of 100 ps, 20 production 973
runs of 200 ns each were performed, yielding a total simulation 974
time of 4 μ s. Covalent bonds to hydrogen atoms were constrained 975
using the LINCS algorithm⁷⁷ (lincs_iter = 1, lincs_order = 4), 976
allowing for an integration time step of 2 fs. The leapfrog 977
integrator was used. The temperature was maintained by the 978
velocity-rescale thermostat⁷⁸ with a time constant of 0.01 ps. 979
Lennard-Jones interactions were cut off at 1 nm. Electrostatic 980
interactions were treated by the Particle–Mesh Ewald (PME) 981

algorithm⁷⁹ 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 $V = 232.6 \text{ nm}^3$, 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 $T = 300 \text{ K}$. The equilibration run was followed by a 500 ns production run, again at $T = 300 \text{ K}$. 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 time step. A leapfrog integrator was used. Data was saved every 1 ps, resulting in 5×10^5 data frames. Six independent simulations from the same equilibrated configuration were carried out resulting in 3 μs total data.

AUTHOR INFORMATION

Corresponding Authors

*E-mail: bettina.keller@fu-berlin.de.

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

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

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

REFERENCES

- Alberts, B.; Johnson, A.; Lewis, J.; Raff, M.; Roberts, K.; Walter, P. *Mol. Biol. Cell*, 5th ed.; Garland Science: New York, 2008.
- Elber, R. *Simulations of allosteric transitions*. 2011; <http://www.ncbi.nlm.nih.gov/pubmed/21333527> (accessed Jan. 9, 2014).
- Verhey, K. J.; Kaul, N.; Soppina, V. *Annu. Rev. Biophys.* **2011**, *40*, 267–288.
- Dunkle, J. a.; Cate, J. H. D. *Annu. Rev. Biophys.* **2010**, *39*, 227–244.
- Keller, B.; Daura, X.; Van Gunsteren, W. F. *J. Chem. Phys.* **2010**, *132*, 074110.
- Krivov, S. V.; Karplus, M. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 14766–14770.
- Wales, D. J. *Energy Landscapes*, 1st ed.; Cambridge University Press: Cambridge, 2003.
- Noé, F.; Fischer, S. *Curr. Opin. Struc. Biol.* **2008**, *18*, 154–162.
- Karpen, M. E.; Tobias, D. J.; Brooks, C. L. *Biochemistry* **1993**, *32*, 412–420.
- Hubner, I. A.; Deeds, E. J.; Shakhnovich, E. I. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 17747–17752.
- Buchete, N.; Hummer, G. *J. Phys. Chem. B* **2008**, *112*, 6057–6069.
- Rao, F.; Cafisch, A. *J. Mol. Biol.* **2004**, *342*, 299–306.
- Muff, S.; Cafisch, A. *Proteins* **2007**, *70*, 1185–1195.

- de Groot, B.; Daura, X.; Mark, A.; Grubmüller, H. *J. Mol. Biol.* **2001**, *301*, 299–313.
- Schultheis, V.; Hirschberger, T.; Carstens, H.; Tavan, P. *J. Chem. Theory Comp.* **2005**, *1*, 515–526.
- Pan, A. C.; Roux, B. *J. Chem. Phys.* **2008**, *129*, 064107.
- Weber, M. *Improved Perron Cluster Analysis*, Technical Report 03-04; Konrad-Zuse-Zentrum für Informationstechnik Berlin: Berlin-Dahlem, Germany, 2003.
- Noé, F.; Krachtus, D.; Smith, J. C.; Fischer, S. *J. Chem. Theory Comput.* **2006**, *2*, 840–857.
- Noé, F.; Oswald, M.; Reinelt, G.; Fischer, S.; Smith, J. C. *Multiscale Model. Simul.* **2006**, *5*, 393–419.
- Noé, F.; Horenko, I.; Schütte, C.; Smith, J. C. *J. Chem. Phys.* **2007**, *126*, 155102.
- Chodera, J. D.; Dill, K. A.; Singhal, N.; Pande, V. S.; Swope, W. C.; Pitera, J. W. *J. Chem. Phys.* **2007**, *126*, 155101.
- Swope, W. C.; Pitera, J. W.; Suits, F. *J. Phys. Chem. B* **2004**, *108*, 6571–6581.
- Swope, W. C.; Pitera, J. W.; Suits, F. *J. Phys. Chem. B* **2004**, *108*, 6571–6581.
- Buchete, N.-V.; Hummer, G. *J. Phys. Chem. B* **2008**, *112*, 6057–6069.
- 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.
- E, W.; Vanden-Eijnden, E. *J. Stat. Phys.* **2006**, *123*, 503–523.
- Noé, F.; Schütte, C.; Vanden-Eijnden, E.; Reich, L.; Weikl, T. R. *Proc. Natl. Acad. Sci. U.S.A.* **2009**, *106*, 19011–6.
- Deuffhard, P.; Weber, M. *Linear Algebra and Its Applications* **2005**, *398*, 161–184.
- Kube, S.; Weber, M. *J. Chem. Phys.* **2007**, *126*, 024103.
- Chodera, J. D.; Singhal, N.; Pande, V. S.; Dill, K.; Swope, W. C. *J. Chem. Phys.* **2007**, *126*, 155101.
- Noé, F.; Horenko, I.; Schütte, C.; Smith, J. C. *J. Chem. Phys.* **2007**, *126*, 155102.
- Ruzhytska, S.; Jacobi, M. N.; Jensen, C. H.; Nerukh, D. *J. Chem. Phys.* **2010**, *133*, 164102.
- Bowman, G. R.; Pande, V. S. *Proc. Natl. Acad. Sci. U.S.A.* **2010**, *107*, 10890–10895.
- Noé, F.; Doose, S.; Daidone, I.; Löllmann, M.; Sauer, M.; Chodera, J. D.; Smith, J. C. *Proc. Natl. Acad. Sci. U.S.A.* **2011**, *108*, 4822–4827.
- Keller, B. G.; Prinz, J.-H.; Noé, F. *Chem. Phys.* **2012**, *396*, 92–107.
- Keller, B.; Hünenberger, P.; van Gunsteren, W. F. *J. Chem. Theory Comput.* **2011**, *7*, 1032–1044.
- Singhal, N.; Pande, V. S. *J. Chem. Phys.* **2005**, *123*, 204909.
- Noé, F. *J. Chem. Phys.* **2008**, *128*, 244103.
- Chodera, J. D.; Noé, F. *J. Chem. Phys.* **2010**, *133*, 105102.
- Bowman, G. R.; Beauchamp, K. A.; Boxer, G.; Pande, V. S. *J. Chem. Phys.* **2009**, *131*, 124101.
- Pande, V. S.; Beauchamp, K.; Bowman, G. R. *Methods* **2010**, *52*, 99–105.
- 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.
- Senne, M.; Trendelkamp-Schroer, B.; Mey, A. S. J. S.; Schütte, C.; Noé, F. *J. Chem. Theory Comput.* **2012**, *8*, 2223–2238.
- Muff, S.; Cafisch, A. *Proteins: Struct. Funct. Bioinf.* **2007**, *70*, 1185–1195.
- Lane, T. J.; Bowman, G. R.; Beauchamp, K.; Voelz, V. a.; Pande, V. S. *J. Am. Chem. Soc.* **2011**, *133*, 18413–18419.
- Noé, F.; Schütte, C.; Vanden-Eijnden, E.; Reich, L.; Weikl, T. *Proc. Natl. Acad. Sci. U.S.A.* **2009**, *106*, 19011–19016.
- Beauchamp, K. A.; McGibbon, R.; Lin, Y.-S.; Pande, V. S. *Proc. Natl. Acad. Sci. U.S.A.* **2012**, *109*, 17807–17813.
- 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.; Wrighers, W. *Science* **2010**, *330*, 341–346.

- 1111 (49) Lindorff-Larsen, K.; Piana, S.; Dror, R. O.; Shaw, D. E. *Science*
1112 **2011**, 334, 517–520.
- 1113 (50) Sadiq, S. K.; Noé, F.; De Fabritiis, G. *Proc. Natl. Acad. Sci. U.S.A.*
1114 **2012**, 109, 20449–20454.
- 1115 (51) Buch, L.; Giorgino, T.; De Fabritiis, G. *Proc. Natl. Acad. Sci.*
1116 *U.S.A.* **2011**, 108, 10184–10189.
- 1117 (52) Kelley, N. W.; Vishal, V.; Krafft, G. A.; Pande, V. S. *J. Chem.*
1118 *Phys.* **2008**, 129, 214707.
- 1119 (53) Nerukh, D.; Jensen, C. H.; Glen, R. C. *J. Chem. Phys.* **2010**, 132,
1120 084104.
- 1121 (54) Jensen, C. H.; Nerukh, D.; Glen, R. C. *J. Chem. Phys.* **2008**, 128,
1122 115107.
- 1123 (55) Sarich, M.; Noé, F.; Schütte, C. *Multiscale Model. Simul.* **2010**, 8,
1124 1154–1177.
- 1125 (56) Altis, A.; Nguyen, P. H.; Hegger, R.; Stock, G. *J. Chem. Phys.*
1126 **2007**, 126, 244111.
- 1127 (57) Schwantes, C.; Pande, V. *J. Chem. Theory Comput.* **2013**, 9,
1128 2000–2009.
- 1129 (58) Pérez-Hernández, G.; Paul, F.; Giorgino, T.; De Fabritiis, G.;
1130 Noé, F. *J. Chem. Phys.* **2013**, 139, 015102.
- 1131 (59) Schütte, C.; Fischer, A.; Huisinga, W.; Deuffhard, P. *J. Comput.*
1132 *Phys.* **1999**, 151, 146–168.
- 1133 (60) Rains, E. K.; Andersen, H. C. *J. Chem. Phys.* **2010**, 133, 144113.
- 1134 (61) Weber, M. Ph.D. thesis, Freie Universitaet Berlin, Berlin, 2006.
- 1135 (62) Röblitz, S. Ph.D. thesis, Freie Universitaet Berlin, Berlin, 2009.
- 1136 (63) Haack, F.; Röblitz, S.; Scharkoi, O.; Schmidt, B. *AIP Conf. Proc.*
1137 **2010**, 1281, 1585–1588.
- 1138 (64) Schütte, C.; Noé, F.; Lu, J.; Sarich, M.; Vanden-Eijnden, E. *J.*
1139 *Chem. Phys.* **2011**, 134, 204105.
- 1140 (65) Noé, F.; Nüske, F. *SIAM Multiscale Model. Simul.* **2013**, 11,
1141 635–655.
- 1142 (66) Szabo, A.; Ostlund, N. S. *Modern Quantum Chemistry*, 1st ed.;
1143 Dover Publications: Mineola, NY, 1996; pp 31–38.
- 1144 (67) Noé, F.; Doose, S.; Daidone, I.; Löllmann, M.; Chodera, J.;
1145 Sauer, M.; Smith, J. *Proc. Natl. Acad. Sci. U.S.A.* **2011**, 108, 4822–4827.
- 1146 (68) Lindner, B.; Yi, Z.; Prinz, J.-H.; Smith, J.; Noé, F. *J. Chem. Phys.*
1147 **2013**, 139, 175101.
- 1148 (69) Zheng, Y.; Lindner, B.; Prinz, J.-H.; Noé, F.; Smith, J. *J. Chem.*
1149 *Phys.* **2013**, 139, 175102.
- 1150 (70) Vanden-Eijnden, E.; Venturoli, M. *J. Chem. Phys.* **2009**, 130,
1151 194101.
- 1152 (71) Lindorff-Larsen, K.; Piana, S.; Palmo, K.; Maragakis, P.; Klepeis,
1153 J. L.; Dror, R. O.; Shaw, D. E. *Proteins* **2010**, 78, 1950–1958.
- 1154 (72) Prinz, J.-H.; Wu, H.; Sarich, M.; Keller, B.; Senne, M.; Held, M.;
1155 Chodera, J.; Schütte, C.; Noé, F. *J. Chem. Phys.* **2011**, 134, 174105.
- 1156 (73) Deuffhard, P.; Huisinga, W.; Fischer, A.; Schütte, C. *Linear*
1157 *Algebra and Its Applications* **2000**, 315, 39–59.
- 1158 (74) MacCluer, C. R. *SIAM Rev.* **2000**, 42, 487–498.
- 1159 (75) Van Der Spoel, D.; Lindahl, E.; Hess, B.; Groenhof, G.; Mark, A.
1160 E.; Berendsen, H. J. C. *J. Comput. Chem.* **2005**, 26, 1701–1718.
- 1161 (76) Kritzer, J. A.; Tirado-Rives, J.; Hart, S. A.; Lear, J. D.; Jorgensen,
1162 W. L.; Schepartz, A. *J. Am. Chem. Soc.* **2005**, 127, 167–178.
- 1163 (77) Hess, B.; Bekker, H.; Berendsen, H. J. C.; Fraaije, J. G. E. M. *J.*
1164 *Comput. Chem.* **1997**, 18, 1463–1472.
- 1165 (78) Bussi, G.; Donadio, D.; Parrinello, M. *J. Chem. Phys.* **2007**, 126,
1166 014101.
- 1167 (79) Darden, T.; York, D.; Pedersen, L. *J. Chem. Phys.* **1993**, 98,
1168 10089–10092.