A basis set for peptides for the variational approach to conformational kinetics

F. Vitalini,[1] F. Noé,[2, a)] and B.G. Keller[1, b)]

[1)] *Department of Biology, Chemistry, Pharmacy, Freie Universität Berlin, Takustraße 3, D-14195 Berlin, Germany*

[2)] *Department of Mathematics and Computer Science, Freie Universität Berlin, Arnimallee 6, D-14195 Berlin, Germany*

(Dated: 27 July 2015)

Although Markov state models have proven to be powerful tools in resolving the complex features of biomolecular kinetics, the discretization of the conformational space has been a bottleneck since the advent of the method. A recently introduced variational approach, which uses basis functions instead of crisp conformational states, opened up a route to construct kinetic models in which the discretization error can be controlled systematically. Here, we develop and test a basis set for peptides to be used in the variational approach. The basis set is constructed by combining local residue-centered kinetic modes which are obtained from kinetic models of terminally blocked amino acids. Using this basis set, we model the conformational kinetics of two hexapeptides with sequences VGLAPG and VGVAPG. Six basis functions are sufficient to represent the slow kinetic modes of these peptides. The basis set also allows for a direct interpretation of the slow kinetic modes without an additional clustering in the space of the dominant eigenvectors. Morever, changes in the conformational kinetics due to the exchange of leucine in VGLAPG to valine in VGVAPG can be directly quantified by comparing histograms of the basis set expansion coefficients.

PACS numbers: Valid PACS appear here

Keywords: Suggested keywords

[a)] Electronic mail: frank.noe@fu-berlin.de

[b)] Electronic mail: bettina.keller@fu-berlin.de

I. INTRODUCTION

Structure and function of proteins are linked by structural transitions. This is particularly evident in protein-ligand binding processes, in which induced fit, conformational selection, or allosteric regulation[1,2] directly mediate ligand recognition and biological response. But it is also true for intrinsically disordered peptides (IDPs)[3,4] which fluctuate between a variety of (partially folded) conformations. Many IDPs are involved in signaling and regulatory pathways, and adopt a specific three-dimensional structure only upon binding to their interaction partner within this pathway. Misfolded IDPs are associated to a number of diseases, such as cancer, Alzheimer's disease, and diabetes[5,6]. A description of the function and malfunction of these peptides hence requires a detailed model of their conformational kinetics.

In recent years, the estimation of transition rates across energy barriers using classical molecular dynamics (MD) simulations has become computationally tractable. Methods which intertwine the exploration of the conformational space with an estimation of transition rates, such as the milestoning approach[7–9], transition path sampling[10], transition interface sampling[11], or the string method[12,13], have been used successfully. Alternatively, molecular kinetics models can be estimated from unbiased simulations using the Markov state model (MSM) technique[14–19]. In MSMs, the conformational space is discretized into M non-overlapping states, often denoted microstates. The pairwise transition probabilities between the microstates (within a lag time τ) are estimated from the molecular-dynamics simulation data and arranged in a MSM transition matrix, which is then further analyzed.

MSMs have proven to be very useful tools in resolving the complex features of biomolecular kinetics[20,21]. Because their construction is largely independent of *a priori* assumptions as to what the actual slow conformational processes in the system might be, human bias can be minimized in these models. MSMs allow for the representation of highly complex molecular kinetics, yet the salient features of these kinetics can be converted into humanly comprehendible and visually intuitive representations, such as kinetic networks[17,22], transition path networks and density fluxes across these networks[23,24], metastable states[25,26], and conformational exchange processes[27,28]. Finally, MSMs are a very useful framework to connect data from time-resolved experiments with simulation data[29,30].

MSMs approximate the (deterministic) dynamics in the complete, high-dimensional conformational space by a stochastic dynamics in a low-dimensional subspace of the conformational space. Therefore, the discretization of the conformational space into microstates consists of two steps, which are often executed iteratively to find a suitable set of microstates. First, the dynamics of the MD simulation is projected onto a (relevant) subspace of the conformational space, which is typically chosen manually. Then this subspace is partitioned into microstates. The approximation quality of MSMs depends sensitively on how well the discretization of the conformational space represents the free-energy barriers in the system[31].

Since the features of the free-energy landscape of large molecules are not known *a priori*, the discretization of the conformational space has been a bottleneck in the construction of MSMs since the advent of the method[16–19,32,33]. In recent years, principle component analysis in conjunction with the most probable path algorithm[34,35], as well as diffusion maps[36,37] have been used to automatically identify the relevant subspace. Along the same lines, the time-lagged independent component analysis (TICA) method[27,38,39] has been proposed which, for a particular system, automatically combines a large set of user-defined order parameters into an optimal small set of order parameters for the construction of a MSM.

In highly metastable systems, i.e. in systems with few but very long-lived conformations, the dominant kinetic processes can be represented by step functions which switch between the regions of the conformational space associated to the metastable conformations. In these systems, once the relevant subspace is identified, few microstates are sufficient to obtain good approximation quality[14,25,40]. Most biomolecules however exhibit kinetic processes which vary smoothly between different regions of the conformational space[19,31]. This requires a fine discretization in these transition regions and hence a model with many microstates. Unfortunately, the model quality suffers if the number of microstates becomes too large because the statistical uncertainty of estimated transition probability increases with the numbers of microstates. To balance these two requirements, several iterative discretizations schemes have been proposed[16,41]. Also methods which abandon the crisp state definition and instead use functions which slowly vary from 0 to 1 between pairs of metatable states have been put forward[42–44].

All of these strategies have in common that they are data-driven, i.e. the projection and discretization is defined by a statistical analysis of the trajectory, rather than by considering the properties of the molecules. As a consequence, different simulation runs (even of the same system) will produce different discretizations, making the results difficult to compare; Moreover, the interpretation of the kinetic processes as conformational transitions is not straight-forward, as the microstates have no intrinsic meaning. Thus, an additional clustering step in the space of the dominant eigenvectors is applied[25] which again depends on user-defined input parameters.

Recently, we have introduced a variational approach to molecular kinetics[45] that was further developed in[27,28]. This variational approach opens up a route to construct comparable Markov state models with systematically controllable approximation quality. The mathematical properties of a propagator associated to the stochastic dynamics in the relevant subspace (self-adjointness, bound eigenvalue spectrum) allow for the formulation of a variational principle. The dominant kinetic processes can then be expanded in terms of an arbitrary basis set and the coefficients can be optimized using the method of linear variation. This is analogous to linear variation methods in quantum chemistry. The difference in the implementation is that the matrix elements are estimated as time-lagged cross-correlations from MD trajectories, rather then attempting to numerically solve the associated integral. Because the basis functions can be chosen to vary smoothly from one region of the conformational space to another, the number of basis functions needed to achieve a given approximation quality may be much less than the number of states in a corresponding conventional MSM. The better the basis functions represent the slow kinetic processes, the better the approximation quality. Hence, using prior knowledge about the system, one can customize the basis functions for a particular class of molecules. Furthermore, if the basis functions are designed such that they represent local conformational changes, the linear expansion of the slow kinetic processes in terms of this basis can be easily interpreted as a superposition of these local transitions.

In the present paper, we develop a basis set for the conformational kinetics of peptides. We model the kinetics within a single residue by up to three functions and combine these residue-centered functions into basis functions for the overall backbone dynamics. The residue-centered functions are pre-parametrized on model systems, and hence the basis set only depends on the sequence of the peptide and not on the actual simulation of the peptide. Moreover, the residue-centered functions of all (canonical) amino acids have analogous interpretations. Hence the direct comparison of conformational kinetics of the peptides with different sequence becomes possible.

II. THEORY

We present the salient points of the theory of propagators (section II A) and the variational principle for propagators (section II B). Markov state models are reviewed in appendix B. For a detailed discussion see[15,19,28,33,45,46]. The basis set for peptide dynamics is introduced in sections II C - II E.

A. The propagator

Consider an infinite ensemble of molecules in a state space X. We assume that the dynamics are Markovian, ergodic, and reversible. The time-dependent probability density of the ensemble of molecules in the state space is denoted $p_t(x)$ with $x \in X$. If $p_t(x)$ is not equal to the equilibrium distribution $\pi(x)$, it will relax gradually towards the equilibrium distribution

$$\lim_{t\to\infty} p_t(x) = \pi(x)\,. \tag{1}$$

The time-evolution of $p_t(x)$ is governed by a transition density

$$p_{t+\tau}(y) = \int_X p(x,y;\,\tau)\,p(x)\,dx \tag{2}$$

where the integral is evaluated over the entire state space X. $p(x,y;\,\tau)$ represents the conditional probability density of finding a molecule in conformation y at time t, given that it has been in $x\,dx$ at time $t-\tau$. Eq. 2 defines an operator, the so-called propagator $\mathcal{P}(\tau)$, which propagates the probability forwards in time by a fixed time interval τ

$$p_{t+\tau}(x) = \mathcal{P}(\tau)p_t(x) \tag{3}$$

$$p_{t+n\tau}(x) = \mathcal{P}^n(\tau)p_t(x)\,. \tag{4}$$

τ is a parameter of the propagator and is typically called lag time. The propagator has a bounded eigenvalue spectrum

$$\lambda_1 = 1 > |\lambda_2(\tau)| \geq |\lambda_3(\tau)| \geq .... \tag{5}$$

where $\lambda_1 = 1$ is the largest eigenvalue by absolute value. This eigenvalue always exists and is associated to an eigenvector $l_1(x)$ which is proportional to the equilibrium distribution $\pi(x)$ of the dynamic process x_t. Here we set them equal, without loss of generality:

$$l_1(x) = \pi(x) \tag{6}$$

Ergodicity and reversibility have two consequences. First, the eigenvalues $\lambda_i(\tau)$ and eigenvectors $l_i(x)$, defined by

$$\mathcal{P}(\tau) l_i(x) = \lambda_i(\tau) l_i(x)\,, \tag{7}$$

are real-valued. Second, the propagator is self-adjoint

$$\langle \mathcal{P}(\tau) g \mid f \rangle_{\pi^{-1}} = \langle g \mid \mathcal{P}(\tau) f \rangle_{\pi^{-1}} \tag{8}$$

with respect to a weighted scalar product

$$\langle g \mid f \rangle_{\pi^{-1}} = \int_X g(x) \pi^{-1}(x) f(x)\, dx\,. \tag{9}$$

Therefore, its eigenfunctions form a complete basis and the time-evolution of the probability density can be expressed as a linear combination of the eigenfunctions

$$\begin{aligned} p_t(x) &= \sum_{i=1}^{\infty} c_i \lambda_i^n(\tau) l_i(x) \\ &= \pi(x) + \sum_{i=2}^{\infty} c_i \exp\left(-\frac{t}{t_i}\right) l_i(x) \end{aligned} \tag{10}$$

where the relaxation timescales are given by $t_i = -\tau / \ln|\lambda_i(\tau)|$ (for $i > 1$), and time is in multiples of the lag time, $t = n\tau$. The coefficients c_i are determined by the probability density at time $t = 0$, and always $c_1 = 1$.

Eq. 10 can be understood as superposition of dynamic modes $l_i(x)$ with time-dependent amplitudes $c_i \exp(-t/t_i)$. Since the eigenvalues are bounded by one, the amplitudes decay exponentially. t_i represents the relaxation time of this decay process. The first eigenfunction is an exception because it is associated to a constant amplitude. This is the mathematical correspondence of the physical observation that any initial distribution $p_0(x)$ will eventually decay to the equilibrium distribution $\pi(x)$.

The dominant eigenfunction-eigenvalue pairs, i.e. those with high-lying eigenvalues, contain a wealth of information on the barriers in the system, its long-lived conformational states and the dynamics between these states. An analytical solution of eq. 7 is not possible due to the high-dimensionality of the space X even for small molecules. The variational approach therefore aims at numerically approximating the dominant eigenfunction-eigenvalue pairs.

B. Variational principle and method of linear variation

The two properties of the propagator - bounded eigenvalue spectrum and self-adjointness - are sufficient to derive a variational principle for the propagator

$$\langle f \,|\, \mathcal{P}(\tau) \,|\, f \rangle_{\pi^{-1}} \leq 1\,. \tag{11}$$

where the equality holds iff $f(x) = \pi(x)$. (For more details, see Ref. 45 and to appendix A.) Based on this variational principle, we can derive variational methods to compute best approximations to the leading eigenvalues $\lambda_1(\tau), ..., \lambda_M(\tau)$ and eigenfunctions $l_1(x), ..., l_M(x)$. In particular, if one requires that ith eigenfunctions $l_i(x)$ is orthogonal (with respect to eq. 9) to the previously estimated eigenvectors $l_1(x), ...l_{i-1}(x)$, the associated estimate of the eigenvalue $\hat{\lambda}_i$ is a lower bound to the true ith eigenvalues

$$\hat{\lambda}_i(\tau) \leq \lambda_i(\tau) \qquad \forall\, i\,. \tag{12}$$

In the method of linear variation[28,45], that is analogous to the Ritz method in quantum mechanics[47], the eigenfunctions are approximated as linear combinations of a set of basis functions $\{\phi_i(x)\}$

$$l_j(x) \approx \hat{l}_j(x) = \sum_{i=1}^{M} a_{ij}\phi_i(x). \tag{13}$$

We do not require that the basis functions are orthonormal. The size of the basis set is not limited, however for any practical application a finite subset of M basis functions needs to be chosen for eq. 13. By inserting the expansion (13) into the variational principle (eq. 11) and varying the expansion coefficients $\{a_i\}_{i=1}^M$ so as to maximize $\left\langle \hat{l}_i(x) \,\middle|\, \mathcal{P}(\tau) \,\middle|\, \hat{l}_i(x) \right\rangle_\pi$, while keeping $\hat{l}_i$ orthonormal with respect to eq. 9, we obtain a generalized eigenvalue problem

$$\mathbf{C}(\tau)\mathbf{a}_i = \mathbf{S}\lambda_i(\tau)\mathbf{a}_i \tag{14}$$

where $\mathbf{C}(\tau)$ is the correlation matrix with elements

$$C_{ij}(\tau) = \langle \phi_i \mid \mathcal{P}(\tau)\phi_j \rangle_{\pi^{-1}}\,, \tag{15}$$

$\mathbf{S}$ is the overlap matrix with elements

$$S_{ij} = \langle \phi_i \mid \phi_j \rangle_{\pi^{-1}}\,, \tag{16}$$

and $\mathbf{a}$ is the vector of expansion coefficients (eq. 13). Note that only the expansion coefficients are varied, while the basis functions are kept constant. Solving this generalized eigenvalue problem yields the optimal approximation to the first M eigenfunctions in terms of the chosen basis $\{\phi_i(x)\}_{i=1}^M$ and the associated eigenvalues: $\{\hat{l}_i(x), \hat{\lambda}_i(\tau)\}_{i=1}^M$. In particular, this linear variational solution ensures that eq.12 holds for all estimated eigenvalues[45], and that several scoring functions such as the sum of eigenvalues can be used to compare different solutions (Appendix A).

Due to the high dimensionality of the conformational space X, the integral in eq. 15 cannot be evaluated directly. However, the expression $\langle \phi_i \mid \mathcal{P}(\tau)\phi_j \rangle_{\pi^{-1}}$ (eq. 15) can be interpreted as a time-lagged correlation function[28,45] which can be estimated from a time-discretized realization of the dynamical process x_t with time step Δt and length N_T

$$C_{ij}(\tau) = \lim_{N_T\to\infty} \widehat{\mathrm{cor}}(\chi_i, \chi_j, \tau) \tag{17}$$

$$= \lim_{N_T\to\infty} \frac{1}{N_T - n_\tau} \sum_{t=1}^{N_T - n_\tau} \chi_j(x_t)\chi_i(x_{t+n_\tau}) \tag{18}$$

where $n_\tau = \tau/\Delta t$. Likewise, the elements of the overlap matrix are estimated as

$$S_{ij} = \lim_{N_T\to\infty} \widehat{\mathrm{cor}}(\chi_i, \chi_j, \tau = 0) \tag{19}$$

$$= \lim_{N_T\to\infty} \frac{1}{N_T} \sum_{t=1}^{N_T} \chi_j(x_t)\chi_i(x_t)\,. \tag{20}$$

For finite N_T, $C_{ij}(\tau)$ and S_{ij} are replaced by there corresponding estimates

$$\hat{C}_{ij}(\tau) = \frac{1}{N_T - n_\tau} \sum_{t=1}^{N_T - n_\tau} \chi_j(x_t)\chi_i(x_{t+n_\tau}) \tag{21}$$

and

$$\hat{S}_{ij} = \frac{1}{N_T} \sum_{t=1}^{N_T} \chi_j(x_t)\chi_i(x_t)\,. \tag{22}$$

Note that the correlation is not defined with respect to the basis function $\{\phi_i\}$ but with respect to the co-functions $\{\chi_i\}$ which are obtained by weighting the basis functions with $\pi_i^{-1}(x)$

$$\chi_i(x) = \pi^{-1}(x)\phi_i(x) \Leftrightarrow \pi(x)\chi_i(x) = \phi_i(x)\,. \tag{23}$$

In practice, we will therefore directly work in the basis of the co-functions $\{\chi_i\}$. The theoretical background of these co-functions is discussed in section II C.

Realizations x_t of molecular dynamics can be obtained by atomistic molecular dynamics simulations, which leads to the following workflow for the variational approach to molecular dynamics

1. Generate a realization x_t of the conformational dynamics of the molecule of interest using molecular dynamics simulation.

2. Choose a (finite) basis set $\{\chi_i\}_{i=1}^M$

3. Project the x_t onto each of the basis function yielding a set of M time series $\{\chi_i(x_t)\}_{i=1}^M$.

4. Choose a lag time τ and estimate the elements of the correlation matrix using eq. 21.

5. Estimate the elements of the overlap matrix using eq. 22.

6. Solve eq. 14 to obtain the eigenvalues $\{\lambda_i\}_{i=1}^M$ and expansion coefficients $\{\mathbf{a}_i\}_{i=1}^M$ of the first M eigenfunctions.

C. The eigenfunctions of the propagator and their associated co-functions

Alternatively to the propagator formulation (eq. 1 to 10), we could choose a transfer operator formulation which is completely equivalent[14]. The transfer operator is defined in a weighted space and has eigenfunctions $r_j(x)$. For the present case of reversible dynamics, the relationship between these two sets of eigenfunctions is very simple:

$$\begin{aligned} \pi(x)\, r_j(x) &= l_j(x) \\ r_j(x) &= \pi^{-1}(x)\, l_j(x) \end{aligned} \tag{24}$$

Both sets of eigenfunctions can be interpreted as dynamic modes, which mediate the transfer of probability density between different regions of the state space X. While the functions $l_j(x)$ contain information on the probability distribution within these regions, this information is erased in $r_j(x)$ by weighting $l_j(x)$ with $\pi^{-1}(x)$. In the functions $r_i(x)$, only the specification of these regions is retained.[19].

The close connection between $l_j(x)$ and $r_j(x)$ has important consequences for interpretation of the model and for the choice of a suitable basis set. First, the variational approach yields $l_j(x)$ as a linear expansion in the basis $\{\phi_i\}_{i=1}^M$ and *simultaneously* $r_j(x)$ as a linear expansion in the basis of the co-functions $\{\chi_i\}_{i=1}^M$

$$\begin{aligned} \pi(x)\, r_j(x) &\approx \sum_{i=1}^{M} a_{ij}\phi_i(x) = \sum_{i=1}^{M} a_{ij}\pi(x)\chi_i(x) \\ &\Downarrow \\ r_j(x) &\approx \sum_{i=1}^{M} a_{ij}\chi_i(x)\,. \end{aligned} \tag{25}$$

Second, a particularly suitable basis set $\{\phi_i\}_{i=1}^M$ would be one in which the basis functions resemble the actual eigenfunctions $l_j(x)$ of the propagator. This implies that the co-functions $\{\chi_i\}_{i=1}^M$, which are used for the estimation of the matrix elements, would be similar to the eigenfunctions of the transfer operator $r_j(x)$. Therefore, when parametrizing a basis set from model systems (as suggested in the following section) one should use the eigenfunctions of the transfer operator of these model system, $r_i^{\mathrm{model}}(x)$, to construct $\{\chi_i\}_{i=1}^M$. In Markov state models (with row-normalized transition matrices), this amounts to using the right eigenvectors of the transition matrix rather than the left eigenvectors.

D. Basis set for peptide dynamics

The critical step in the workflow in section II B is the choice of the basis set. A good basis set should meet three requirements: (*i*) it should be designed such that it can distinguish all the important conformational changes of the molecule; (*ii*) the number of basis functions needed to represent the slow processes should be small; (*iii*) the basis set should be transferable, i.e. one should be able to use the same basis functions to construct dynamics models for a large range of molecules with similar chemical structure. In the following, we will demonstrate how to construct such a basis set for the conformational dynamics of peptides. The overall dynamics of peptides can usually be described to a good approximation by the ϕ- and ψ-backbone torsion angles. We therefore choose to define our basis functions in terms of these backbone-torsion angles. (requirement i). The inclusion of the side chain torsion angles (χ_1, χ_2, etc) is numerically more demanding but conceptually straight forward. Likewise, other state variables such as the distances between atoms that are far apart in the sequence, solvent and ion coordinates can be included to model larger peptides and proteins for which the space of torsions is no longer expected to be sufficient.

With this choice, the basis functions of an N-residue peptide are functions with $2N$ variables. To get these high-dimensional functions into a manageable form, we decompose them into tensor products of residue-centered two-dimensional functions

$$\chi(\phi_1, \psi_1, \phi_2, \psi_2, ...\phi_N, \psi_N) = \tag{26}$$

$$R(\phi_1, \psi_1) \otimes R(\phi_2, \psi_2) .... \otimes R(\phi_N, \psi_N) \, . \tag{27}$$

This decomposition has a biophysical underpinning: the (rigid) peptide bond by which the amino acid residues are linked in a peptide chain acts as a block to the dynamic correlations between residues. That is, typically the ϕ- and ψ-backbone torsion angle of any given residue are much higher correlated to each other than to any other backbone torsion angle in the peptide chain.

To meet requirement *ii*, the basis functions have to be close to the actual eigenfunctions of the transfer operator (see section II C). The dynamics of the ϕ-ψ-torsion angle pairs in a peptide chain is severely restricted by steric interactions of its side chain with the neighboring peptide groups. In fact, these steric interactions are so dominant that one can identify generic slow dynamic modes within the ϕ-ψ-space of each amino acid type[48]. We use these dynamic modes as the residue-centered functions $R(\phi_i, \psi_i)$ in eq. 27.

Fig. 1 shows the slow dynamic modes of of alanine (A), valine (V), leucine (L), proline (P), glycine (G), and alanine which precedes a proline ($\mathrm{A_P}$) (represented as eigenfunctions $r_i(\phi, \psi)$ of the underlying transfer operator). In the following, we denote these functions as R_k^X, where X is replaced by the one-letter code of the amino acid and $k \in \{1, 2, 3\}$ (proline: $k \in \{1, 2\}$) indicates the number of the residue-centered dynamic mode. Most amino acids, such as alanine (A), valine (V), and leucine (L), have three dynamics modes which correspond to the stationary process (R_1^A, R_1^V, R_1^L), the conformational exchange between the Lα-region and the combined α-helix and β-sheet regions (R_2^A, R_2^V, R_2^L), and the conformational exchange between the α-helix region and the β-sheet region (R_3^A, R_3^V, R_3^L). Proline (P) has only two dynamic modes because its side chain binds back to the backbone, thereby restricting the dynamics of the ϕ-torsion angle. The two modes correspond to the stationary distribution (R_1^P) and to the conformational exchange along the ψ-torsion angle (R_2^P). Glycine (G) does not have a side chain and therefore shows a different dynamics than the other amino acids. Nonetheless, its modes can be interpreted as stationary process (R_1^G), conformational exchange along the ψ-torsion angle (R_2^G), and conformational exchange along the ϕ-torsion angle (R_3^G). The dynamics of an amino acid that precedes a proline in the sequence is altered by the limited dynamics of the following residue[49] (see for example Fig. 1 in the SI). The dynamic modes of an alanine which is followed by a proline exhibits three slow dynamic modes (Fig. 1): the stationary distribution ($R_1^{A_P}$), the conformational exchange along the ψ-axis ($R_2^{A_P}$), and the conformational exchange of the minimum in the upper left corner of the graph with the rest of the ϕ-ψ-space ($R_3^{A_P}$)

The construction of example peptide basis functions as a tensor product of these residue-centered functions is illustrated in Fig. 2 for the peptide VGVAPG. The index of the basis function χ should be read as a string in which the ith element denotes which dynamic mode of the ith residue is used for this particular basis function. The index 0 denote that the corresponding residue is not included in conformational space X, and not included in the model. Excluding N-terminal residues from a kinetic model is a common approach since their dynamics tend to be decoupled from the rest of the chain. The complete basis set consists of $3^{N-N_P} \cdot 2^{N_P}$ basis functions, where N denotes the number of residues which are included in the model, and N_P denotes the number of proline residues in the peptide sequence.

E. Basis set size

The number of basis functions grows as $3^{N-P} \cdot 2^P$, where N is the number or residues in the peptide and P is the number of proline residues. This is computationally intractable for peptides beyond decamers. However, we expect that, due to the design of the peptide basis set, only very few (possibly less than N) basis functions are needed to describe the conformational subspace spanned by the slow dynamic processed of the peptide. For the molecules studied in this contribution, this expectation is confirmed. The task at hand is then to select a small number of basis functions which are likely to yield a good representation of the slow dynamic processes. The residue-centered function R_1^X can be interpreted as the dynamic ground state, and R_2^X and R_3^X as the first and second dynamically excited state of residue X. The basis function χ_{111111} approximates the dynamic ground state of a hexapeptide. Correspondingly, the basis function χ_{112111} represents a dynamic mode in which the third residues is excited, and χ_{112131} represents a dynamic mode in which two residues (residues 3 and 5) are excited. We suggest to use at least a basis set which consists of the ground state basis functions plus all singly excited basis functions, i.e. functions in which the one residue is in excited dynamic state, and all other are in the ground state. This reduces the computational complexity to $\mathcal{O}(N)$. This basis set can be systematically expanded by including doubly, triply, etc. excited basis functions.

III. METHODS

A. MD simulations

To obtain the residue-based functions R_k^X, we performed all-atom molecular-dynamics simulations of the terminally blocked amino acids Ac-A-NHMe, Ac-V-NHMe, Ac-P-NHMe, Ac-G-NHMe, and Ac-L-NHMe, and the terminally capped dipeptide Ac-AP-NHMe. We additionally simulated the terminally blocked dipeptides Ac-AV-NHMe and Ac-VA-NHMe, and the hexa-peptides VGVAPG and VGLAPG. All simulations were carried out in explicit water in the NVT ensemble, where the temperature was restrained to 300K. We used the GROMACS 4.5.5 simulation package[50] with the AMBER ff-99SB-ILDN[51] force field and the TIP3P water model[52]. The atom coordinates of the solutes were saved every picosecond. For each system, a total of c.a. 4 μs simulation time was produced. For further details on the simulation, see SI section I.

B. Markov State Models

Markov state models for all systems were constructed from the microstate trajectories using EMMA software package[53] and the recent python implementation (see pyemma.org). For each of the systems, Markov state models were constructed at a range of lag times. To construct implied timescale plots (Fig. 3, 4 and 7), at each lag time, the dominant eigenvalues of the MSM transition matrix were calculated. Additionally, the dominant left and right eigenvectors of the MSM transition matrix were extracted for the lag time τ_{Markov} at which the implied timescales of each system reached a plateau. The eigenvectors give a structural interpretation to the slow dynamic modes (Fig. 3, 5, and 6). For further details, see SI section II.

To obtain the microstate trajectories, different discretization strategies have been used for different systems. For the terminally blocked amino acids (Ac-X-NHMe) the backbone ϕ-

and ψ-torsion angles were discretized using a regular grid of 36 grid points along each angle (distance between grid points 10°), yielding a discretization of $36 \times 36 = 1\,296$ microstates. In the terminally blocked dipeptide Ac-AP-NHMe, the ϕ- and ψ-torsion angles of the alanine residue were discretized in the same fashion. The torsion angles of the proline residue were not included in the model. For the terminally blocked dipeptides Ac-AV-NHMe and Ac-VA-NHMe, a coarser discretization of the Ramachandran plane of each residue was applied. The ϕ- and ψ-torsion angle space of each residue was discretized into three bins (see SI Fig.1), which resulted in an overall discretization of $3 \times 3 = 9$ microstates. The bins were chosen such that they separate the maxima of the equilibrium distribution in the ϕ-ψ-plane of each residue (see SI Fig. 1). Similarly in the hexapeptides, the Ramachandran planes of the residues 2 to 6 were discretized into a grid of 6 (G), 3 (V, A, L), and 2 (P) states. The N-terminal residue is largely decoupled from the dynamics of the rest of the chain and was therefore excluded from discretization. Each possible configuration of bins along the peptide chain represents a microstate, resulting in $6 \times 3 \times 3 \times 2 \times 6 = 648$ microstates for the hexapeptides.

C. Variational approach

Terminally blocked amino acids serve as minimal segments which mimic the conformational dynamics of the corresponding amino acid in a peptide chain. Conventional MSMs of terminally blocked amino acids Ac-X-NHMe, where X is a wildcard for the one-letter code of as specific amino acid, and of Ac-AP-NHMe were used to obtain vector-representations of the residue-centered functions R_k^X. We use the dominant right eigenvectors of the (row-normalized) MSM transition matrix, i.e.vector-representations of the transfer operator eigenfunctions.

Given the vector-representations of the residue-centered functions, the implementation of step 3 in the workflow in section II B consists of the following substeps

3. Project the x_t onto each of the basis function yielding a set of M time series $\{\chi_i(x_t)\}_{i=1}^M$:

   (a) For each residue r, extract the $\{\phi_t, \psi_t\}_r$-torsion angle time series from the MD trajectory x_t.

(b) Project $\{\phi_t, \psi_t\}_r$ onto the grid of $36 \times 36 = 1296$ states, yielding a trajectory of (residue-centered) states $s_t^{\text{res number}}$

(c) For each basis function $\chi_i = \chi_{klmn...}$ (where the $k, l, m, n, ...$ denote the dynamic mode R_k^X of the corresponding residue), construct the time series $\chi_i(x_t)$ as

$$\chi_i(x_t) = R_k^{\text{res 1}}[s_t^{\text{res 1}}] \cdot R_l^{\text{res 2}}[s_t^{\text{res 2}}] \cdot R_m^{\text{res 3}}[s_t^{\text{res 3}}]...$$

For more details on the implementation see SI section III.

For models of the terminally blocked amino acids Ac-A-NHMe and Ac-V-NHMe, the basis set consisted simply of the corresponding residue centered functions $\{R_1^A, R_2^A, R_3^A\}$ and $\{R_1^V, R_2^V, R_3^V\}$. For the models of the terminally blocked dipeptides Ac-AV-NHMe and Ac-VA-NHMe, a basis set consisting of all possible combinations of the residue-centered functions was used (Tab. IA). For the hexa-peptides VGLAPG and VGVAPG, a truncated basis set consisting of all singly and doubly excited basis functions was used. The corresponding indices are reported in Tab. IB. Analogous to the Markov state models, the correlation matrices were estimated at a range of lag times τ. The generalized eigenvalue problem (eq. 14) was solved, and the dominant eigenvalues were used to construct the implied-timescale plots (Fig. 3 and 4). A structural interpretation of the slow dynamic modes was obtained by analyzing the corresponding vector of the expansion coefficients **a** (Fig. 3, 5, 6). The software for the variational approach was implemented in python[54] in conjunction with the packages NumPy[55], SciPy[55], and Matplotlib[56]. We are planning to add this implementation as a package to the EMMA project (pyemma.org).

IV. RESULTS

A. Terminally blocked amino acids

To check consistency, we applied the variational approach to the terminally blocked amino acids Ac-A-NHMe and Ac-V-NHMe, which were used to obtain the residue-centered functions $\{R_1^A, R_2^A, R_3^A\}$ and $\{R_1^V, R_2^V, R_3^V\}$, respectively. The basis sets consisted simply of these functions. The variational results are compared to the conventional MSM that was used to parametrize the residue-centered functions. As expected, the variational approach yields the same implied timescales as the MSM (Fig. 3a and 3b). The stationary process and the two slow kinetic processes of the MSM are given by the three residue-centered functions shown in Fig. 1. The corresponding processes of the MSM are given as a superposition of the three basis functions (weighted by the corresponding equilibrium distribution). Fig. 3c and 3d show that, in each process, only one of the basis functions makes a contribution to this superposition. That is, the variational approach correctly recovers the results of the MSM.

B. Terminally blocked dipeptides

We applied the variational approach to the terminally blocked dipeptides Ac-AV-NHMe and Ac-VA-NHMe using the complete set of nine basis functions for these molecules. The basis functions have the form $\chi_{ij} = R_i^A \otimes R_j^V$ for Ac-AV-NHMe and $\chi_{ij} = R_i^V \otimes R_j^A$ for Ac-VA-NHMe. For the mapping between $\{i, j\}$ and the index of the basis function, see Tab. IA. The results are compared to direct MSM in which the ϕ-ψ-space of each amino acid was discretized into 3 states, yielding $3 \cdot 3 = 9$ microstates for the dipeptides. (The estimation of a model using the same discretization as the basis functions is numerically still feasible for two residues (see SI sec.V A). However, the model is constructed on a number of discretized states of the same order of magnitude than the available data points, therefore making the estimation of the transition probabilities subject to high statistical uncertainty). Both the variational estimate and the 9-microstate MSM yield the same implied time scales for Ac-AV-NHMe and Ac-AV-NHMe (Fig. 4.a and 4.b). The two slow kinetic processes of Ac-AV-NHMe have implied timescales of 4 ns and 3.5 ns, respectively. The slow kinetic processes of Ac-AV-NHMe have slightly larger implied timescale: 6.5 ns and 4 ns.

Fig. 5 compares the dynamical processes identified by the two models, constructed at lag time $\tau = 1$ ns. Fig. 5a and 5b show histograms of the absolute values of the expansion coefficients in eq. 13. In Ac-AV-NHMe, each of the processes is dominated by a single basis function. Process one, which is the stationary process, is represented by $\chi_{11} = R_1^A \otimes R_1^V$, i.e. the stationary process in both residues. Process two (4 ns) and three (3.5 ns) are dominated by $\chi_{12} = R_1^A \otimes R_2^V$ and $\chi_{21} = R_2^A \otimes R_1^V$, respectively. χ_{12} represents a conformational transition across the barrier ϕ=0 in the second residue (valine), and χ_{21} a transition across ϕ=0 in first residue (alanine). In Ac-VA-NHMe, the first and the third process are each represented by a single basis function: $\chi_{11} = R_1^V \otimes R_1^A$ and $\chi_{21} = R_2^V \otimes R_1^A$, respectively. Interestingly, the second dynamic mode (6.5 ns) shows the contribution of two basis functions, $\chi_{12} = R_1^V \otimes R_2^A$ and $\chi_{22} = R_2^A \otimes R_2^V$, suggesting a coupled motion of the residues.

How do the variational results compare to the results of the direct MSM? Fig. 5c and 5d show an analysis of the eigenvectors of the 9-micorstate MSM eigenvectors. As the microstates have no intrinsic meaning, a direct interpretation of the eigenvectors is not feasible. One therefore first identifies long-lived conformations, and then interprets the eigenvectors as transitions between these conformations[19,25]. The scatterplots show projections of the (visited) microstates onto the second and third right MSM eigenvector (second and third kinetic process). The size of each point is proportional to the stationary probability of the corresponding microstate. The emerging clusters of microstates can be interpreted as long-lived conformational states[19,26,40]. The structural characterization of each cluster is shown next to the scatter plots as the Ramachandran plots of both residues, where the coloring of the Ramachandran planes indicates whether the corresponding region is populated in the respective cluster.

In Ac-AV-NHMe (Fig. 5c) cluster one corresponds to backbone conformations, in which both residues are in the α-helical or the β-sheet conformation. By contrast, in cluster two, V_2 is in the L_α conformation ($\phi_2 > 0$), and in cluster three, A_1 is in the L_α conformation. Process two represents the conformational exchange between cluster one and two and hence requires a rotation around the ϕ-angle in V_2. Process three represents the conformational exchange between cluster one and three and is mediated by a rotation around the ϕ-angle in A_1. This is in agreement with the results of the variational approach.

Fig 5d shows the structural interpretation of the clusters in the 9-states MSM of Ac-VA-NHMe. Cluster one again corresponds to conformations in which both residues are in the α-helical or the β-sheet conformation. However, in cluster two both residues show some population in the L_α conformation and hence the conformational exchange between these two clusters (process two) requires rotations around both ϕ-torsion angles. This is in line with the result of the variational approach. Cluster three comprises structures in which V_1 is in the L_α conformation and A_2 is in the α-helical or the β-sheet conformation. Consequently, process three corresponds to a conformational transition in the ϕ angle of V_1 without coupling to a transition in A_2, which is in agreement with the results of the variational approach.

C. Hexapeptides VGLAPG and VGVAPG

We constructed variational models of the hexapeptides VGLAPG and VGVAPG, including singly and doubly excited states (SD basis set, 42 basis functions, Tab. I B) into the basis set. The results were compared to direct MSMs constructed on a Ramachandran-based discretization of 648 states (Fig 1 SI). Fig. 4c and 4d show that, for both peptides, the variational model (solid lines) and the direct MSM (dashed lines) yield converged results and similar estimates for the implied timescales.

Besides the stationary process, VGLAPG has two slow kinetic processes, with implied timescales of 15.6 ns and 2.2 ns (both models for a lag time $\tau = 2$ ns), respectively (Fig. 4c). The expansion coefficients of the corresponding eigenvectors in the variational model (Fig. 6a) yield a structural interpretation of these slow processes. The stationary process (process 1) is mapped to the dynamic ground state (χ_{011111}). The second process is dominated by basis function 4 (χ_{012111}), which represents a torsion around the ϕ-angle of residue L_3. The third process is a superposition of the dynamic ground state and basis function 5 (χ_{013111}), which represents a torsion around the ψ-angle of residue L_3. In essence, the slow dynamics of this hexapeptide is dominated by conformational transitions in the backbone torsion angles of L_3, with all other amino acids equilibrating on timescales shorter than 2.2 ns. This interpretation is confirmed by the analysis of the MSM eigenvectors (Fig. 6b). Analogously to Fig. 5, the microstates are projected onto the second and third right eigenvector of the MSM transition matrix and the emerging clusters are characterized using Ramachandran plots (Fig 4 SI). Of all residues in the peptide, only the Ramachandran plots of L_3 varied from cluster to cluster, confirming that the slow dynamics of VGLAPG is governed by this residue. In cluster one, L_3 is in the L_α conformation and the kinetic exchange with cluster two and three (process two) requires a torsion around its ϕ angle. Furthermore, the kinetic exchange between cluster two and three (process two) is mediated via conformational exchange along the ψ-torsion angle of L_3. Both processes are hence in agreement with the variational model.

The conformational dynamics of VGVAPG is governed by three slow kinetic processes with relaxation timescales of 8.7 ns (MSM: 8.6 ns), 8.2 (MSM: 8.4 ns), and 4.5 ns (both models for a lag time of $\tau = 3$ ns). By comparing Fig. 6c to Fig. 6a, one can directly assess the effect of substituting L_3 by V_3 on the conformational dynamics of the hexapeptide.

As in the previous examples, stationary process (process one) is mapped to the dynamic ground state (χ_{011111}). The expansion coefficients of process four are very similar to those of process three in VGLAPG. Both processes represent a torsion around the ϕ-angle in the third residue, where the process is associated to a slightly higher relaxation timescales in VGVAPG than in VGLAPG. Process three in VGVAPG represents a torsion around the ϕ-angle in the third residue and is therefore related to process two in VGLAPG. It has however an additional contribution from basis function 26 (χ_{012311}), which couples this torsion to a conformational transition in the ψ-angle of A_4. Process two in VGVAPG is not related to any slow process in VGLAPG. It represents a torsion around the ϕ-angle of A_4 coupled to a torsion around the ψ-angle in V_3. Overall, the conformational kinetics of VGVAPG is governed by correlated conformational transitions in V_3 and A_4. The coupling between V_3 and A_4 is most likely caused by the branching at the C_β-atom in the valine side chain which induces a stronger steric interaction with the backbone than the leucine side chain in VGLAPG (which is only branched at the C_γ-atom).

The projection of the microstates on the second, third, and fourth right eigenvectors of the MSM in Fig. 6d shows four clusters, whose structural characterization is presented in Fig. 5 in the SI. The clusters differ in the backbone conformations of V_3 and A_4 in line with the results of the variational model. Process two represents the kinetic exchange of cluster one with the rest of the conformational ensemble, which is mediated by transitions in the ϕ and ψ-torsion angles of A_4 possibly coupled to a conformational change in V_3. Process three, which represents the kinetic exchange of cluster one and three with the rest of the ensemble, is dominated by a transition in the ϕ-angle of V_3 coupled to a conformational change in A_4. The conformational exchange between cluster two and four (process four) involves a transition in the ψ-angle of V_3. The MSM results are in agreement with the variational model. However, Fig. 6d and Fig. 5 in the SI also highlight the difficulties in interpreting direct MSMs. First, the analysis is much more complex and time-consuming then the interpretation of the expansion coefficients in the variational approach. Second, the clustering as well as the interpretation of the conformational exchange between the clusters comprise (to a certain degree) arbitrary decisions. Third, due to the small population of the clusters (e.g. cluster one), it is not always easy to decide whether a certain conformation does not belong to the cluster or whether it is simply not sampled, which affects the interpretation of the conformational exchange processes.

D. Basis set size and choice of basis functions

The variational results for the hexapeptides suggest the slow kinetics of these peptides is dominated by only a few basis functions and that hence a very small basis set is sufficient to obtain a valid kinetic model. We thus constructed minimal variational models consisting of only the basis functions which contributed the most to the slow processes in Fig. 6. These were basis functions 1, 4, and 5 (i.e χ_{011111}, χ_{012111}, and χ_{013111}, Tab. I.B) for VGLAPG, and basis functions 1, 4, 5, 6, 26, and 27 (i.e χ_{011111}, χ_{012111}, χ_{013111}, χ_{011211}, χ_{012311}, and χ_{013211}, Tab. I.B) for VGVAPG.

For both basis sets, we obtained converged models (Fig. 7c and 7b). The implied timescales are similar to those of the variational model with singly and doubly excited basis functions (SD basis set, 42 basis functions) and to the MSM model. For VGLAPG, the amplitudes assigned to the basis functions 1, 4, and 5 in the SD variational model are compared to amplitudes in the minimal model in Fig. 7b. The amplitudes in process one and two are virtually identical. For process three, the contribution of basis function 5 (χ_{013111}) is a bit stronger in the minimal model. For VGVAPG, the amplitudes of the minimal model are compared to those of the SD variational model in Fig. 7d. Process one and four have again identical or very similar amplitudes. Process two and three are however swapped in the minimal model. The swapping of process two and process three is not very surprising since the implied timescales of these processes result so close (variational model: 8.7 ns / 8.2 ns, MSM: 8.6 ns / 8.4 ns) that the processes are in effect degenerate.

In Fig. 8 and Fig.6 in the SI, the effect of increasing the basis set is investigated. We successively added triply, quadruply and quintuply excited basis functions to the basis set. Since the N-terminal valine is not included in the model, the quintuply excited basis set corresponds to the full basis set. Increasing the basis functions had no significant effect on the estimated implied timescales (Fig.6 and Tab II in the SI). However, the amplitudes of the additional basis functions are not negligible. Especially large amplitudes are assigned to quadruply and quintuply excited basis functions (basis functions index > 98). This is shown for process two and three of VGVAPG in Fig. 8. These basis functions with multiple excitations represent kinetic processes with fast relaxation timescales, typically much faster than the lag-time of the correlation matrix. Hence, the conformational exchange associated to these processes equilibrates within the lag time τ of the model and the estimation of the time-lagged cross-correlation to slowly decaying processes is numerically instable. The large amplitudes in Fig. 8 are therefore numerical artifacts. More research is needed for an optimal strategy to pre-select the basis functions which contribute to the slow dynamics of the peptide from the full basis set and to verify the numerical accuracy of the estimated amplitudes. For now, we suggest to use the time-lagged autocorrelation of a basis functions (Fig. 8.c) as an indicator of the relaxation timescales of the cross-correlations involving this basis functions. This amounts to calculating the diagonal elements of $\mathbf{C}(\tau)$ and to truncating the basis set at excitation levels at which this autocorrelation becomes negligible.

V. CONCLUSIONS

We have proposed and tested a basis set for peptides to be used in the variational approach to molecular kinetics. The basis functions are constructed as tensor products of residue-centered functions which represent the (local) kinetic modes of the corresponding residue. That is, a given basis functions represents either a conformational transition in a specific residue (singly excited basis functions) or concerted conformational transitions in different residues (multiply excited basis functions). The slow kinetic modes of the peptides emerge as a superposition of isolated and concerted conformational transitions and can be concisely represented as histogram of the expansion coefficients. Because the basis functions have intrinsic meaning, the interpretation of the model is much simpler and considerably less tedious than the interpretation of a conventional MSM, which requires an additional clustering in the space of the dominant eigenvectors and a structural characterization of the resulting clusters.

By comparing the histograms of the expansion coefficients of different peptides one can directly quantify changes in the peptide dynamics which are induced by a modification in the peptide sequence. By comparing the hexapeptides VGLAPG and VGVAPG we demonstrated that the comparison of the histograms answers questions such as: are the relaxation timescale of a given process altered? Is a specific conformational transition suppressed? Are additional residues contributing to the slow conformational kinetics? We believe that the basis set will be particularly suited to model intrinsically disordered peptides, because their conformational kinetics often changes drastically upon the exchange of a single amino acid[5,6,57], and because they are difficult to model using conventional MSMs due to their ernormous conformational space[27].

The comparison of VGLAPG and VGVAPG also revealed that the β-branched valine introduced a coupling to A_4, which was absent in VGLAPG. This indicates that at least for β-branche side chains it might be useful to include the conformational transitions in the first side-chain torsion angle χ_1 into basis set. This can be accomplished via two routes: (*i*) the residue-centered functions are re-estimated on the space of the ϕ-,ψ- and χ_1 torsion angle: $R_k^X = R_k^X(\phi, \psi, \chi_1)$, where X is replaced by the one-letter code of the amino acid and k is the index of the dynamical mode (excitation level). (*ii*) a separate MSM is estimated for the χ_1-torsion angles on model peptides and the residue-centered functions are constructed from the dynamical modes S_l^X of the side chain (index l) and the dynamical modes of the backbone R_k^X as $R_{k,l}^X = R_k^X(\phi, \psi) \otimes S_l^X(\chi_1)$. Both routes are straight-forward to implement but numerically more demanding than the basis functions presented in this paper. Towards more general basis sets that would also be suitable for describing the kinetics of proteins, one might have to include additional terms which represent solvent exposure and long-range interactions such as salt bridges, contact formation and dissociation between different secondary structure elements.

The size of the full basis set grows rapidly with the number of residues, it therefore has the same scaling behavior as the number of crisp states in a conventional MSM. However, since the basis functions are constructed by combining the actual kinetic modes of the individual residues, only a small fraction of the full basis set is needed to accurately model the slow dynamics of the peptides. For the hexapeptides VGLAPG and VGVAPG, 3 and 6 basis functions out of 162 were sufficient. This opens up the task of devising a method which selects the important basis functions from the basis set before actually constructing the variational model. This could be accomplished by first identifying the residues which contribute to the slow kinetic modes using singly or doubly excited basis functions and then by refining the model by adding basis functions which represent multiple excitations of these residues. Alternatively, the exponential scaling behavior of the basis set can also be addressed by efficient tensor approaches, such as the tensor-train format[58]. In future work, we will combine these two strategies.

Finally, it is important to point out that the basis set is force-field dependent. Although most force fields identify the same type of conformational transitions in terminally blocked amino acids (Fig. 1) as AMBER ff-99SB-ILDN[51], which was used in the present study, the precise shape of the corresponding MSM eigenvectors differs significantly across force fields[48]. Therefore, a separated set of residue-centered functions should be parametrized when the method is applied to a simulation with a different force field.

Appendix A: Variational principle

In[45], we derived the following variational principle for the transfer operator of conformation dynamics:

1. The normalized Rayleigh coefficient of eigenfunction r_i is identical to the ith eigenvalue:
$$\frac{\langle r_i(x_t) \cdot r_i(x_{t+\tau}) \rangle_t}{\langle r_i^2(x_t) \rangle_t} = \lambda_i(\tau)$$
2. Any approximate function $\hat{r}_i$, that is orthogonal to the previous eigenfunctions $1, ..., i-1$,
$$\langle \hat{r}_i \mid r_j \rangle_\pi = 0 \quad \forall j \in \{1, ..., i-1\}$$
has a Rayleigh coefficient that underestimates the ith eigenvalue:
$$\frac{\langle \hat{r}_i(x_t) \cdot \hat{r}_i(x_{t+\tau}) \rangle_t}{\langle \hat{r}_i^2(x_t) \rangle_t} = \hat{\lambda}_i(\tau) \leq \lambda_i(\tau)$$

As the approximate eigenfunctions $\hat{r}_i(x)$ are orthonormal, one can derive a stronger variational principal for them[45]. Now the variational principle applies to each approximate eigenvalue:

$$\hat{\lambda}_i \leq \lambda_i \qquad i \in \{1, ..., M\}, \tag{A1}$$

implying that we always underestimate timescales ($\hat{t}_i \leq t_i$) and overestimate rates ($\hat{\kappa}_i = \hat{t}_i^{-1} \geq t_i^{-1} = \kappa_i$). A trivial consequence is that the sum of eigenvalues is underestimated:

$$\sum_{i=1}^{M} \hat{\lambda}_i \leq \sum_{i=1}^{M} \lambda_i \tag{A2}$$

This quantity, elsewhere named generalized matrix Rayleigh quotient[59] has been suggested to be used as a criterion to select amongst different kinetic models. Another formulation is that the sum of rates is overestimated:

$$\sum_{i=1}^{M} \hat{\kappa}_i \geq \sum_{i=1}^{M} \kappa_i \tag{A3}$$

Eq. A2 and A3 hold for any model of the conformational kinetics which can be formulated as a variational approach (see appendix B).

Appendix B: Markov state models

In conventional Markov state models, the state space X is discretized into a set of M non-overlapping states $\{s_i\}_{i=1}^{M}$, and the transition probabilities $t_{ij}(\tau)$ between pairs of states are estimated from MD trajectories. The states can be represented by indicator functions[19]

$$\chi_i^{\mathrm{MSM}}(x) = \begin{cases} 1 & \text{if } x \in s_i \\ 0 & \text{otherwise}\,. \end{cases} \tag{B1}$$

The set of indicator functions $\{\chi_i^{\mathrm{MSM}}(x)\}_{i=1}^{M}$ can be regarded as a basis set and can be used in equation eq. 21 and eq. 22. The resulting expressions for the estimates of C_{ij} and S_{ij} are[28,45]

$$\hat{C}_{ij}^{\mathrm{MSM}}(\tau) = \frac{1}{N_T - n_\tau} \sum_{t=1}^{N_T - n_\tau} \chi_j^{\mathrm{MSM}}(x_t) \chi_i^{\mathrm{MSM}}(x_{t+n_\tau}) = \frac{z_{ij}}{N_T - n_\tau} \tag{B2}$$

and

$$\hat{S}_{ij}^{\mathrm{MSM}} = \frac{1}{N_T} \sum_{t=1}^{N_T} \chi_j^{\mathrm{MSM}}(x_t) \chi_i^{\mathrm{MSM}}(x_t) = \pi_i \, \delta_{ij}\,. \tag{B3}$$

where z_{ij} are the number of observed transitions from state s_i to state s_j within lag time τ, π_i is the relative equilibrium probability of state s_i, and δ_{ij} is the Kronecker delta. Finally, the familiar expressions for the estimation of a Markov state model transition matrix $\mathbf{T}(\tau)$ arise[28,45], if one considers that $\mathbf{T}(\tau) = [\mathbf{S}^{\mathrm{MSM}}]^{-1}\,\mathbf{C}^{\mathrm{MSM}}(\tau)$, and hence

$$\hat{T}_{ij}(\tau) = \frac{1}{\pi_i} \frac{z_{ij}}{N_T - n_\tau} = \frac{N_T - n_\tau}{\sum_j z_{ij}} \frac{z_{ij}}{N_T - n_\tau} = \frac{z_{ij}}{\sum_j z_{ij}}\,. \tag{B4}$$

In summary, Markov state models are a special case of the variational approach in which indication functions (eq. B1) are used as a basis set. However, since the slow dynamic modes of larger molecules tend to have a smoothly sloped rather than a step-like shape, typically many of these indicator functions are needed to achieve a good approximation of the dynamics of the system[19,31].

Similarly, other versions of Markov state model analysis can be formulated as special cases of the variational approach. For example, the time-lagged independent component analysis (TICA) method uses the mean-free molecular coordinates as basis functions[27]. In the core-based Markov state models, committor functions between cores are used as basis set[60].

ACKNOWLEDGMENTS

The authors would like to thank Feliks Nüske for the interesting discussions on the variational principle and on the tensor-train format and Johannes Dietschreit for productive discussion during his internship. F.V. acknowledges funding from the Dahlem research school. F.N. acknowledges funding from ERC starting grant "pcCell" B.G.K. acknowledges the funding of the German Research Foundation (DFG) through the Collaborative Research Center (SFB) 765 "Multivalency as chemical organization and action principle: new architectures, functions and applications". The computer facilities of the Freie Universität Berlin (ZEDAT) are acknowledged for computer time.

Supporting Information

Simulation details, further information regarding the MSM construction and the discretizations, a pseudo-algorithm to construct the basis functions, information on error estimation and additional supporting figures to the results presented in this manuscript are included in the Supporting Information. The Supporting Information is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES

[1] Jardetzky, O. *Prog. Biophys. Mol. Biol.* **1996**, *65*, 171–219.

[2] Bahar, I.; Chennubhotla, C.; Tobi, D. *Curr. Opin. Struct. Biol.* **2007**, *17*, 633–40.

[3] Dunker, A. K. et al. *J. Mol. Graph. Model.* **2001**, *19*, 26–59.

[4] Dyson, H. J.; Wright, P. E. *Nat. Rev. Mol. Cell Biol.* **2005**, *6*, 197–208.

[5] Dobson, C. M. *Nature* **2003**, *426*, 884–90.

[6] Uversky, V. N.; Oldfield, C. J.; Dunker, A. K. *Annu. Rev. Biophys.* **2008**, *37*, 215–46.

[7] Faradjian, A. K.; Elber, R. *J. Chem. Phys.* **2004**, *120*, 10880–9.

[8] Vanden-Eijnden, E.; Venturoli, M. *J. Chem. Phys.* **2009**, *130*, 194101.

[9] Bello-Rivas, J. M.; Elber, R. *J. Chem. Phys.* **2015**, *142*, 094102.

[10] Bolhuis, P. G.; Chandler, D.; Dellago, C.; Geissler, P. L. *Annu. Rev. Phys. Chem.* **2002**, *53*, 291–318.

[11] van Erp, T. S.; Moroni, D.; Bolhuis, P. G. *J. Chem. Phys.* **2003**, *118*, 7762.

[12] Weinan, E.; Ren, W.; Vanden-Eijnden, E. *Phys. Rev. B* **2002**, *66*, 052301.

[13] Maragliano, L.; Fischer, A.; Vanden-Eijnden, E.; Ciccotti, G. *J. Chem. Phys.* **2006**, *125*, 24106.

[14] Schütte, C.; Fischer, A.; Huisinga, W.; Deuflhard, P. *J. Comput. Phys.* **1999**, 146–168.

[15] Swope, W. C.; Pitera, J. W.; Suits, F. *J. Phys. Chem. B* **2004**, *108*, 6571–6581.

[16] Chodera, J. D.; Singhal, N.; Pande, V. S.; Dill, K. A.; Swope, W. C. *J. Chem. Phys.* **2007**, *126*, 155101.

[17] Noé, F.; Horenko, I.; Schütte, C.; Smith, J. C. *J. Chem. Phys.* **2007**, *126*, 155102.

[18] Buchete, N.-V.; Hummer, G. *J. Phys. Chem. B* **2008**, *112*, 6057–69.

[19] 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.

[20] Chodera, J. D.; Noé, F. *Curr. Opin. Struct. Biol.* **2014**, *25*, 135–44.

[21] Schwantes, C. R.; McGibbon, R. T.; Pande, V. S. *J. Chem. Phys.* **2014**, *141*, 090901.

[22] Bowman, G. R.; Pande, V. S. *Proc. Natl. Acad. Sci. USA* **2010**, *107*, 10890–5.

[23] Metzner, P.; Schütte, C.; Vanden-Eijnden, E. *J. Chem. Phys.* **2006**, *125*, 084110.

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

[25] Deuflhard, P.; Weber, M. *Linear Algebra Appl.* **2005**, *398*, 161–184.

[26] Keller, B.; Daura, X.; van Gunsteren, W. F. *J. Chem. Phys.* **2010**, *132*, 074110.

[27] Pérez-Hernández, G.; Paul, F.; Giorgino, T.; De Fabritiis, G.; Noé, F. *J. Chem. Phys.* **2013**, *139*, 015102.

[28] Nüske, F.; Keller, B. G.; Pérez-Hernández, G.; Mey, A. S. J. S.; Noé, F. *J. Chem. Theory Comput.* **2014**, *10*, 1739–1752.

[29] Noé, F.; Doose, S.; Daidone, I.; Löllmann, M.; Sauer, M.; Chodera, J. D.; Smith, J. C. *Proc. Natl. Acad. Sci. USA* **2011**, *108*, 4822–7.

[30] Keller, B. G.; Prinz, J.-H.; Noé, F. *J. Chem. Phys* **2012**, *396*, 92–107.

[31] Sarich, M.; Noé, F.; Schütte, C. *Multiscale Model Simul* **2010**, *8*, 1154–1177.

[32] Kube, S.; Weber, M. *J. Chem. Phys* **2007**, *126*, 024103.

[33] Keller, B.; Hünenberger, P.; van Gunsteren, W. F. *J. Chem. Theory Comput.* **2011**, *7*, 1032–1044.

[34] Jain, A.; Stock, G. *J. Chem. Theory Comput.* **2012**, *8*, 3810–3819.

[35] Jain, A.; Stock, G. *J. Chem. Phys. B* **2014**,

[36] Coifman, R. R.; Lafon, S.; Lee, A. B.; Maggioni, M.; Nadler, B.; Warner, F.; Zucker, S. W. *Proc. Natl. Acad. Sci. USA* **2005**, *102*, 7426–7431.

[37] Rohrdanz, M. A.; Zheng, W.; Maggioni, M.; Clementi, C. *J. Chem. Phys.* **2011**, *134*.

[38] Schwantes, C. R.; Pande, V. S. *J. Chem. Theory Comput.* **2013**, *9*, 2000–2009.

[39] Schwantes, C. R.; Pande, V. S. *J. Chem. Theory Comput.* **2015**, 150103163622003.

[40] Deuflhard, P.; Huisinga, W.; Fischer, A.; Schütte, C. *Linear Algebra Appl.* **2000**, *315*, 39–59.

[41] Bowman, G. R.; Ensign, D. L.; Pande, V. S. *J. Chem. Theory Comput.* **2010**, *6*, 787–794.

[42] Röblitz, S.; Weber, M. *Adv. Data Anal. Classif.* **2013**, *7*, 147–179.

[43] Sarich, M.; Schütte, C. *Comm. Math. Sci.* **2012**, *10*, 1001–1013.

[44] Sarich, M.; Banisch, R.; Hartmann, C.; Schütte, C. *Entropy* **2013**, *16*, 258–286.

[45] Noé, F.; Nüske, F. *Multiscale Model Simul* **2013**, *11*, 635–655.

[46] Bowman, G. R.; Pande, V. S.; Noé, F. In *An Introduction to Markov State Models and Their Application to Long Timescale Molecular Simulation*; Dordrecht, S. S. M., Ed.; Advances in Experimental Medicine and Biology; Springer Netherlands, 2014; Vol. Vol. 797.

[47] Ritz, W. *J. Reine Angew. Math.* **1909**, *135*, 1–61.

[48] Vitalini, F.; Mey, A. S. J. S.; Noé, F.; Keller, B. G. *J. Chem. Phys.* **2015**, *142*, 084101.

[49] Lovell, S. C.; Davis, I. W.; Arendall, W. B.; de Bakker, P. I. W.; Word, J. M.; Prisant, M. G.; Richardson, J. S.; Richardson, D. C. *Proteins* **2003**, *50*, 437–50.

[50] Van Der Spoel, D.; Lindahl, E.; Hess, B.; Groenhof, G.; Mark, A. E.; Berendsen, H. J. C. *J. Comput. Chem.* **2005**, *26*, 1701–18.

[51] Lindorff-Larsen, K.; Piana, S.; Palmo, K.; Maragakis, P.; Klepeis, J. L.; Dror, R. O.; Shaw, D. E. *Proteins* **2010**, *78*, 1950–8.

[52] Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. *J. Chem. Phys.* **1983**, *79*, 926.

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

[54] Perez, F.; Granger, B. E. *Comput. Sci. Eng.* **2007**, *9*, 21–29.

[55] Oliphant, T. E. *Comput. Sci. Eng.* **2007**, *9*, 10–20.

[56] Hunter, J. D. *Comput. Sci. Eng.* **2007**, *9*, 90–95.

[57] Lin, Y. S.; Pande, V. S. *Biophys. J.* **2012**, *103*, 7–9.

[58] Nüske, F.; Schneider, R.; Noé, F. *J. Chem. Phys.* **2015**, *submitted manuscript*.

[59] McGibbon, R. T.; Pande, V. S. *http://arxiv.org/abs/1407.8083* **2015**,

[60] Schütte, C.; Noé, F.; Lu, J.; Sarich, M.; Vanden-Eijnden, E. *J. Chem. Phys.* **2011**, *134*, 204105.

A: Basis set: Ac-AV-NHMe

#	A	V	#	A	V	#	A	V
1	1	1	4	2	1	7	3	1
2	1	2	5	2	2	8	3	2
3	1	3	6	2	3	9	3	3

B: Basis set: VGVAPG

#	V	G	V	A	P	G	#	V	G	V	A	P	G
1	0	1	1	1	1	1	22	0	2	1	1	1	3
2	0	2	1	1	1	1	23	0	3	1	1	1	2
3	0	3	1	1	1	1	24	0	3	1	1	1	3
4	0	1	2	1	1	1	25	0	1	2	2	1	1
5	0	1	3	1	1	1	26	0	1	2	3	1	1
6	0	1	1	2	1	1	27	0	1	3	2	1	1
7	0	1	1	3	1	1	28	0	1	3	3	1	1
8	0	1	1	1	2	1	29	0	1	2	1	2	1
9	0	1	1	1	1	2	30	0	1	3	1	2	1
10	0	1	1	1	1	3	31	0	1	2	1	1	2
11	0	2	2	1	1	1	32	0	1	2	1	1	3
12	0	2	3	1	1	1	33	0	1	3	1	1	2
13	0	3	2	1	1	1	34	0	1	3	1	1	3
14	0	3	3	1	1	1	35	0	1	1	2	2	1
15	0	2	1	2	1	1	36	0	1	1	3	2	1
16	0	2	1	3	1	1	37	0	1	1	2	1	2
17	0	3	1	2	1	1	38	0	1	1	2	1	3
18	0	3	1	3	1	1	39	0	1	1	3	1	2
19	0	2	1	1	2	1	40	0	1	1	3	1	3
20	0	3	1	1	2	1	41	0	1	1	1	2	2
21	0	2	1	1	1	2	42	0	1	1	1	2	3

TABLE I. Map between basis function index (#) and basis function definition (notation according to Fig. 2). **A:** Ac-AV-NHMe complete basis set (analogous for Ac-VA-NHMe); **B:** VGVAPG ground state (# 1), singly (# 2-10) and doubly (# 11-42) excited states (analogous for VGLAPG);

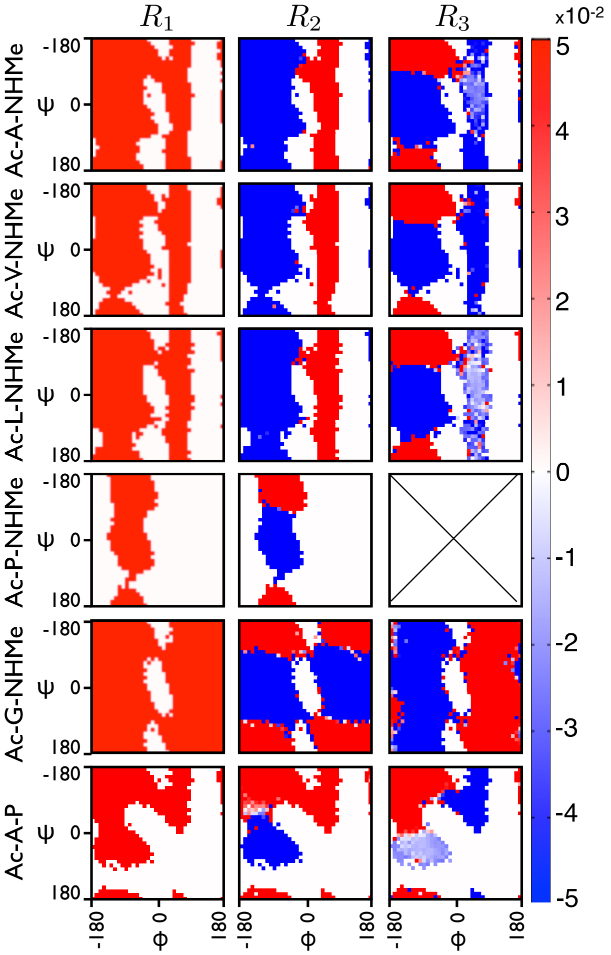

FIG. 1. Slow kinetic modes of the terminally blocked amino acids alanine (A), valine (V), leucine (L), proline (P), glycine (G), and alanine (A) followed by proline, which are used as residue-centered functions (eq. 27). The modes are obtained as the first three right eigenvectors of a MSM transition matrix estimated at lag time τ= 50 ps using a discretization of the $\phi-\psi$-space by a regular 36×36 grid.

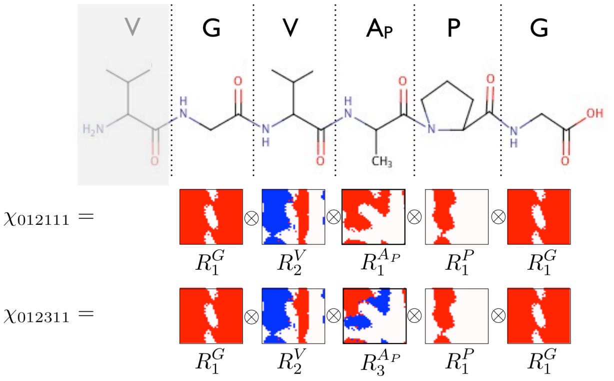

FIG. 2. Construction of basis functions for the hexapeptide VGVAPG from residue-centered functions R_k^X.

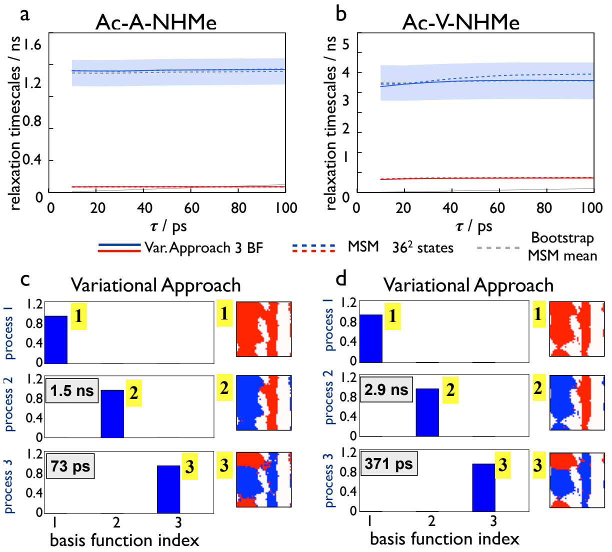

FIG. 3. Variational model for the terminally blocked amino acids: alanine (A) and valine (B). **a** and **b**: relaxation timescales of process two (blue lines) and process three (red lines). Dashed gray lines: MSM bootstrap means, shaded area: 95% confidence interval of the MSM bootstrap sample. **c** and **d**: absolute values of the expansion coefficients (eq. 13)

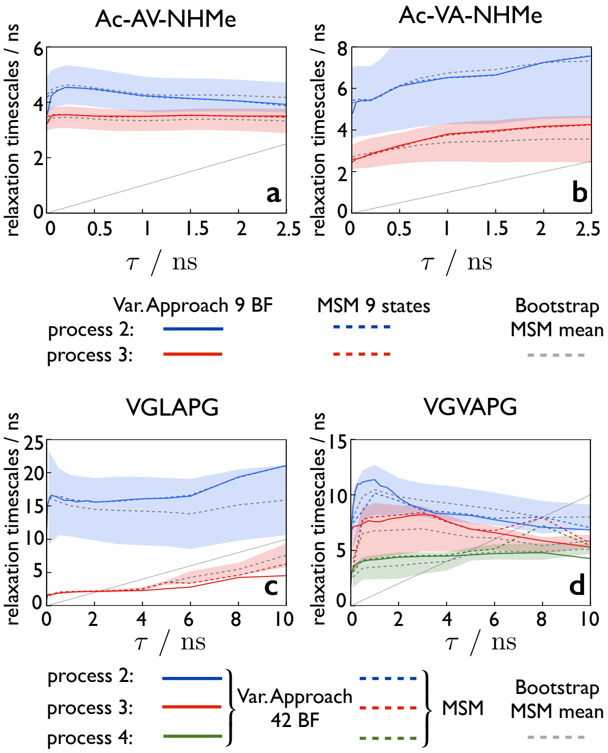

FIG. 4. Relaxation timescales of the terminally blocked dimers Ac-AV-NHMe (a), Ac-VA-NHMe (b), and hexapeptides VGLAPG (c) and VGVAPG (d) estimated using the variational approach with the SD basis set (42 basis functions, solid line) or conventional MSMs (dashed lines). Dashed gray lines: MSM bootstrap means, shaded area: 95% confidence interval of the MSM bootstrap sample.

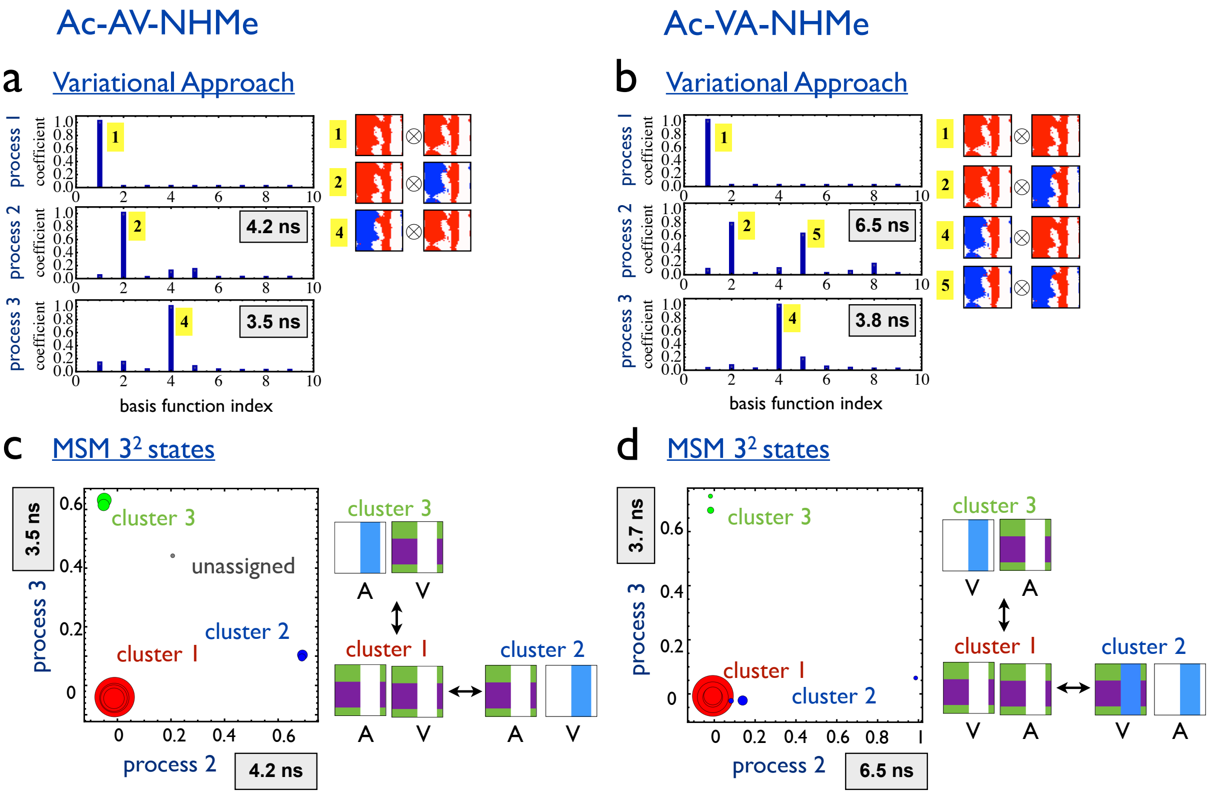

FIG. 5. Slow kinetic processes of the terminally blocked dipeptides Ac-AV-NHMe and Ac-VA-NHMe. **a** and **b**: absolute value of the expansion coefficients in the variational model (SD basis set, 42 basis functions) and representations of the most relevant basis functions; **c** and **d**: projection of the microstates on the dominant right eigenvectors of the direct MSMs and cluster-specific Ramachandran plots.

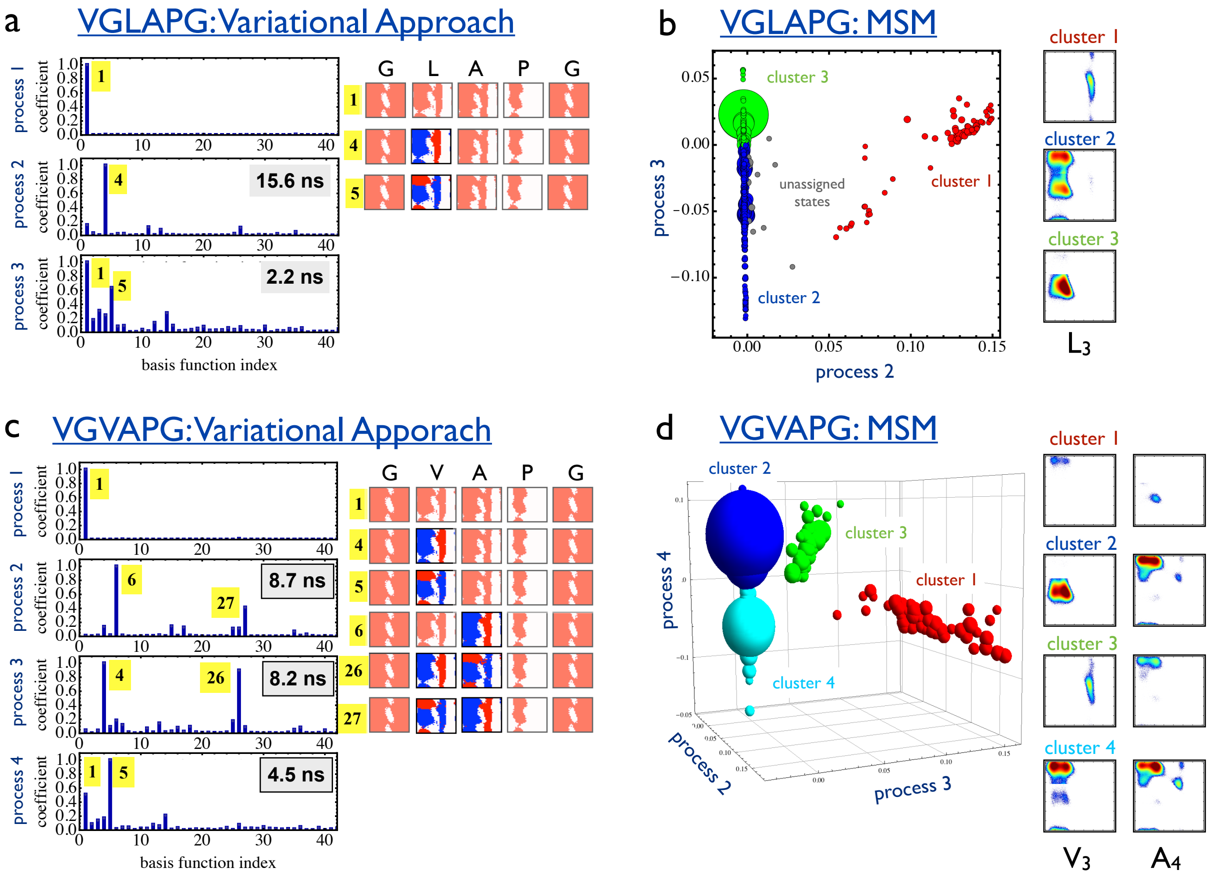

FIG. 6. Slow kinetic processes of the hexapeptides VGLAPG and VGVAPG. **a** and **c**: absolute value of the expansion coefficients in the variational model (SD basis set, 42 basis functions) and representations of the most relevant basis functions; **b** and **d**: projection of the microstates on the dominant right eigenvectors of the direct MSMs and cluster-specific Ramachandran plots.

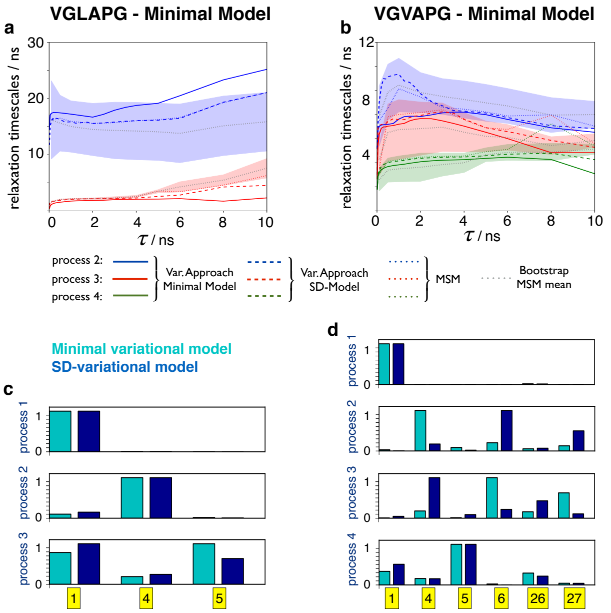

FIG. 7. Minimal variational model for VGLAPG and VGVAPG. **a** and **b**: Relaxation timescales estimated using the variational approach (Minimal Model solid line; SD basis set 42 basis functions, dashed line) or the conventional MSMs (dotted lines); Dashed gray lines: MSM bootstrap means, shaded area: 95% confidence interval of the MSM bootstrap sample. **c** and **d**: absolute value of the expansion coefficients in the minimal variational model (cyan) compared to the corresponding expansion coefficients in the variational model constructed with the SD basis set.

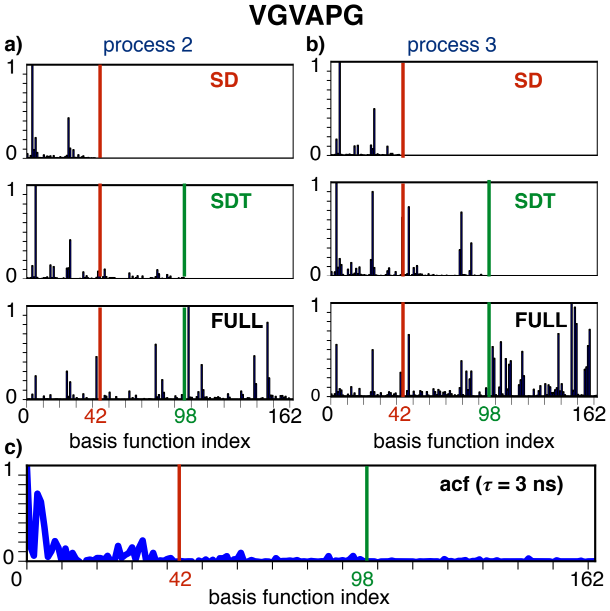

FIG. 8. Effects of basis set size. **a** and **b**: absolute value of the expansion coefficients of processes two and three of VGVAPG at lag time $\tau = 3ns$ for different basis set sizes. SD: ground state, singly and doubly excited states (42 basis functions); SDT: ground state, singly, doubly, and triply excited states (98 basis functions); FULL indicates the full set (162 basis functions). **c**: Time-lagged autocorrelations of the MD trajectory projected onto the basis functions at lag time τ= 3 ns.