Extended Peptide Basis Set for Variational Markov Models: Secondary Structure, Orthonormality, and Undersampled Transitions

L. Martini[1] and B.G. Keller[1, a)]

Department of Biology, Chemistry, Pharmacy, Freie Universität Berlin,
Takustraße 3, D-14195 Berlin, Germany

(Dated: 31 October 2018)

The variational approach to conformational dynamics offers a systematic way to construct kinetic models from molecular dynamics simulations, using an arbitrary set of basis functions. We have recently proposed a basis set for peptide systems that only depends on the sequence of amino acids in the system. This basis set is not data-driven and can therefore be used to compare models for different MD simulations. Here we introduce an orthonormality condition for this basis set as a requirement for the variational models to remain directly interpretable. The orthonormality condition naturally leads to a way of detecting correlations between the sampled marginal stationary probability distributions at each residue in the peptide sequence. We show how these correlations emerge from either undersampled transitions or from inherent dynamical dependencies between the residues. Our basis set relies on a tensor structure obtained from residue-centered ansatz functions. We demonstrate that this structure is sufficient to model both β-sheet and α-helix formation in peptides.

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

I. INTRODUCTION

The biological functions of biomolecules, in particular of proteins, mainly arise from their dynamical properties, that is of the time-dependent behaviour of their three-dimensional structure. For example, phenomenons that arise from protein-ligand binding like allostericity or induced fit provide evidence that the mere knowledge of a spatial shape of the molecule of interest is not sufficient for understanding its behaviour. Therefore the study of the conformational dynamics of a biomolecule is crucial for understanding the emergence of function. Classical molecular dynamics (MD) simulations are a powerful tool for research in this area, as they provide information on the structure and behaviour of the molecule at a spatial and temporal resolution currently unmatched by any experimental techniques[1]. However, due to the complexity and high-dimensionsality of the simulation data, analysis techniques are required to extract meaningful low-dimensional information from the MD trajectories.

Recently, a variational approach to conformational dynamics (VAC) has been proposed[2,3] that aims at the construction of kinetic models from molecular dynamics simulations. The underlying dynamics of such a simulation can be formalized by means of a transfer operator which governs the time evolution of functions on the state space according to the dynamical behaviour of the molecular system. The dominant eigenfunctions of such a transfer operator can be interpreted as the principal kinetic modes of the system, i.e. as the main transition processes between long-lived conformations of the molecule that can be observed in the simulation data. The associated eigenvalues can be related to the time scales at which the respective transition processes take place. Accordingly, the kinetic behaviour of a molecule can be extracted from the dominant eigenfunctions and eigenvalues of the transfer operator. The variational approach yields approximations of these objects with respect to a predefined basis set of functions on state space.

In the last two decades, several methods have been independently developed that can be formulated within the framework of the variational approach, and which can be distinguished by the choice of basis functions used for the approximation of the eigenfunctions. Most prominently, Markov state models (MSMs)[4–11] use a basis set of characteristic functions defined by a discrete partition of state space, resulting in the well-known approximation of the transfer operator as a transition matrix[12]. Further examples are core sets[8,13–15], tICA[16,17], set-free Markov state models[18], variationally optimized diffusion maps[19,20], and sparse tensor approaches[21,22]. Additionally, the variational approach has been used for the identification of transition states[23]. See also ref. 24 for an overview.

A drawback that these approaches share is the fact that the construction of basis functions is data-driven, i.e. depends on the MD simulation at hand[25]. In order to be able to compare models for different MD simulations, however, basis functions are required that only depend on the molecular system rather than individual realizations of its dynamics. Moreover, these data-driven approaches cannot be interpreted directly, but need to be subjected to further analysis.

A possible rectification of these issues could be problem-adapted basis functions that only depend on a certain class of molecules and that furthermore represent physically meaningful transition processes within the conformational space. In ref 26, we developed such a set of basis functions for the backbone dynamics of peptide systems. By making use of the heuristics that the global dynamics of such a peptide can be well-approximated by mutually independent dynamics within the individual residues, we defined these basis functions as tensors of residue-centered functions. The latter were defined as the dominant eigenfunctions that are associated with the isolated dynamics of the respective type of amino acid. In this way, the set of basis functions only depends on the sequence of the peptide and moreover each function can be straightforwardly interpreted as a combination of certain local transitions.

In ref. 26, we tested this basis set successfully on hexapeptides and were able to interpret the estimated eigenfunctions by means of their expansion coefficients in this basis. However, for larger systems such an interpretation ceased to be feasible due to the large number of basis functions that turned out to contribute to each estimated eigenfunction, contrary to our physical intuition. One of the objectives of this work is to give an explanation for this behaviour and to suggest modifications of the basis set that improve the interpretability of the results. Moreover, the from the data in ref 26 it was not clear, whether secondary structure formation can be described within a tensor ansatz for the basis set. We will therefore also address this question by discussing applications of our peptide basis set to larger peptide systems in which the formation of α-helices and β-sheet structures can be observed in the MD simulations.

We begin by reviewing the mathematical background of the transfer operator and the variational approach as well as the construction of the basis functions from ref. 26 in section II. In section III A, we show how orthonormalizing the basis functions can ameliorate the interpretability of the expansion coefficients in the estimates of the eigenfunctions. In sections III B and III C, we furthermore show how working with an orthonormal basis set provides information on the approximation quality of the stationary distribution of the system by the tensor ansatz. Finally, in section IV we discuss applications of the variational method to the identification of metastable states in two larger peptides, the β-hairpin peptide and a fragment of the human islet amyloid polypeptide (hIAPP).

II. THEORY

Consider a state space Ω containing all relevant degrees of freedom of the molecular system of interest. Let $x \in \Omega$ and $y \in \Omega$ denote particular states of the system. The dynamical behaviour of the molecular system can be represented by a Markov process $(X_t)_{t\in T}$, $T \subset \mathbb{R}_{\geq 0}$ which is realized by an MD trajectory $(x_t)_{t=0}^{T}$. On the assumption that the considered process is time-homogeneous (i.e. no time-dependent external forces are applied), a transition probability density $p : \Omega \times \Omega \times T \to [0, 1]$ can be defined via

$$p(x, y; \tau) = \mathbb{P}[X_\tau = y \mid X_0 = x] \tag{1}$$

for all $x, y \in \Omega$ and $\tau \in T$. The quantity $p(x, y; \tau)$ denotes the conditional probability density of finding the system in $y \in \Omega$ at time $t + \tau$, given that it has been in state $x \in \Omega$ at time t. Further assuming the process is ergodic, i.e. the state space cannot be decomposed into several dynamically disconnected components, the Markovian process admits a unique invariant or stationary probability measure μ. We assume that μ admits a probability density function with respect to the Lebesgue measure, i.e. a function $x \mapsto \mu(x)$ fulfilling

$$\mu(A) = \int_A \mu(x)\mathrm{d}x \tag{2}$$

for any measurable $A \subset \Omega$. For thermostatted molecular systems, $\mu(x)$ is the Boltzmann distribution. Hence $\mu(A)$ is the equilibrium probability of finding the system in $A \subset \Omega$, which is sometimes also denoted as μ_A or π_A[2,10,27].

The transition density p can be used to define the transfer operator[11,28] $\mathcal{T}_\tau : L^2(\Omega, \mu) \to L^2(\Omega, \mu)$ via

$$\mathcal{T}_\tau v(y) := \frac{1}{\mu(y)} \int_\Omega p(x, y; \tau)\mu(x)v(x)\mathrm{d}x \tag{3}$$

which is defined on the weighted Hilbert space $L^2(\Omega,\mu) = \{u : \Omega \to \mathbb{R} \mid \int_\Omega u^2(x)\mu(\mathrm{d}x) < \infty\}$ of square integrable state space functions. $L^2(\Omega,\mu)$ has the inner product

$$\langle v, w\rangle_\mu = \int_\Omega v(x)w(x)\mu(\mathrm{d}x) = \int_\Omega v(x)\mu(x)w(x)\,\mathrm{d}x. \tag{4}$$

Suppose ρ_t denotes the probability density associated with X_t at time t. Let $\nu_t \in L^2(\Omega,\mu)$ be the corresponding μ-weighted density defined by the relation $\rho_t(x) = \nu_t(x)\mu(x)$. This quantity is also referred to as the cofunction associated with ρ. By construction of the transfer operator,

$$\nu_t = \mathcal{T}_t\nu_0 \tag{5}$$

for every $t \in T$, i.e. the transfer operator transports initial (μ-weighted) probability densities forward in time according to the underlying dynamics of the Markov process. The cofunction of the stationary density is the characteristic function $\mathbf{1}_\Omega(x)$ of the entire state space[29], $\rho_\infty(x) = \mu(x) = F\mathbf{1}_\Omega(x)\mu(x)$ and is thus invariant under the action of $\mathcal{T}_\tau$

$$\mathbf{1}_\Omega(x) = \mathcal{T}_\tau\mathbf{1}_\Omega(y)\,. \tag{6}$$

Formulated differently, $\mathbf{1}_\Omega(x)$ is an eigenfunction of the transfer operator associated with the eigenvalue $\lambda_1(\tau) = 1$.

The transfer operator fulfills the Chapman-Kolmogorov equation

$$\mathcal{T}_{\tau_1+\tau_2} = \mathcal{T}_{\tau_1}\mathcal{T}_{\tau_2} \tag{7}$$

which is inherited from the equivalent property formulated for the transition density function.

For thermostatted molecular systems, one assumes that the Markov process is reversible and that the condition of detailed balance

$$\mu(x)p(x,y;\tau) = \mu(y)p(y,x;\tau) \tag{8}$$

holds for all $x, y \in \Omega$[2,3,11]. The operator $\mathcal{T}_\tau$ is self-adjoint precisely if the underlying Markov process is reversible, i.e. in this case the identity

$$\langle \mathcal{T}_\tau\nu, \omega\rangle_\mu = \langle \nu, \mathcal{T}_\tau\omega\rangle_\mu\,. \tag{9}$$

holds for all $\nu, \omega \in L^2(\Omega,\mu)$. As a consequence, the transfer operator has only real-valued eigenvalues, and its eigenfunctions form an orthonormal basis of $L^2(\Omega,\mu)$, where orthogonality is defined with respect to eq. 4.

Since the transfer operator is furthermore a bounded operator with $\|\mathcal{T}_\tau\| = 1$, its spectrum is bounded by 1 such that for all eigenvalues $\lambda_i(\tau) \in [-1, 1]$. Therefore, the characteristic function $\mathbf{1}_\Omega(x)$ is the eigenfunction associated with the largest and unique eigenvalue $\lambda_0(\tau) = 1$ with $\lambda_j(\tau) < \lambda_0(\tau)\ \forall j \neq 0$. Let $\lambda_0(\tau) = 1 > \lambda_1(\tau) > \lambda_2 > \dots$ be the eigenvalues sorted according to their absolute values and $r_0 = \mathbf{1}_\Omega, r_1, r_2, \dots$ be the corresponding eigenvectors. Any function $\nu \in L^2(\Omega, \mu)$ can then be expressed in terms of this basis, i.e.

$$\nu(x) = \sum_{j=0}^{\infty} \langle r_j, \nu \rangle_\mu \, r_j(x)\,. \tag{10}$$

Applying the transfer operator n times yields

$$\begin{aligned} \mathcal{T}_{n\tau}\nu(y) &= \sum_{j=0}^{\infty} \langle r_j, \nu \rangle_\mu \, \lambda_j^n(\tau) r_j \\ &= \mathbf{1}_\Omega + \sum_{j=1}^{\infty} \langle r_j, \nu \rangle_\mu \exp\left(-\frac{n\tau}{t_j}\right) r_j \end{aligned} \tag{11}$$

where the so-called implied timescales[5]

$$t_j = -\frac{\tau}{\log(\lambda_j(\tau))} \tag{12}$$

have been introduced. If eq. 7 holds, the t_i are independent of τ. Owing to $\lambda_j(\tau) < 1\ \forall j \neq 0$, these timescales define exponential decay rates of the expansion coefficients in eq. 11, yielding convergence of any function towards the stationary density, i.e.

$$\lim_{\tau \to \infty} \mathcal{T}_\tau \nu = \mathbf{1}_\Omega \tag{13}$$

for all $\nu \in L^2(\Omega, \mu)$. When interpreting the eigenvectors of the transfer operator as dynamical processes of the molecular system, their associated timescales can be related to physical timescales as well[30], giving rise to the notion of metastability of states and associated slow processes in between. Since the slow processes are particularly crucial to the investigation of functionality in molecular systems, kinetic models of molecular systems aim at approximating the subspace $D \subset L^2(\Omega, \mu)$ spanned by the first m eigenfunctions that dominate the sum in eq. 11

$$D = \mathrm{span}(r_0, \dots, r_m). \tag{14}$$

D is referred to as the dominant subspace. This is also the goal of the present contribution.

A. Method of linear variation

Since neither the transfer operator $\mathcal{T}_\tau$ nor the eigenvalues $\lambda_j(\tau)$ and eigenfunctions r_j of the transfer operator are analytically available for any large molecule, numerical methods are required that aim at approximating these quantities from MD simulation data. A recent and very successful method is the variational approach to conformational dynamics (VAC)[2,3,27], which is mathematically analogous to the Rayleigh-Ritz method used in quantum mechanics.

Since the transfer operator is bounded and self-adjoint, one can derive a variational principle for this operator[2,3]: Let $\hat{r}_k \in L^2(\Omega, \mu)$ be a normalized function that is orthogonal to the first $k-1$ eigenfunctions r_j of the transfer operator that are ordered with respect to the absolute values of their corresponding eigenvalues $\lambda_j(\tau)$. Then

$$\langle \hat{r}_k, \mathcal{T}_\tau \hat{r}_k \rangle_\mu = \hat{\lambda}_k(\tau) \leq \lambda_k(\tau) \tag{15}$$

and the equality holds if and only if $\hat{r}_k = r_k$. Therefore the eigenfunctions can be approximated by means of finding functions $\hat{r}_i$ that are orthonormal and that maximize the left-hand side in eq. 15. This task can be achieved by the method of linear variation.

Define a set of basis functions $\mathcal{B} = \{\chi_0, \ldots, \chi_n\} \subset L^2(\Omega, \mu)$ that span an $n+1$-dimensional ansatz space

$$B = \mathrm{span}(\chi_0, \ldots, \chi_n). \tag{16}$$

The method of linear variation aims at finding approximations of the first $n+1$ eigenfunctions $\hat{r}_j \in B$ within this ansatz space, i.e.

$$\hat{r}_j = \sum_{i=0}^{n} a_{ij} \chi_i \tag{17}$$

where the coefficients $a_{ij}, 0 \leq i, j \leq n$ are varied such that $\langle \hat{r}_j, \mathcal{T}_\tau \hat{r}_j \rangle_\mu$ becomes maximal for all $0 \leq j \leq n$ under the constraint of the set of estimated functions $\{\hat{r}_0, \ldots, \hat{r}_n\}$ remaining orthonormal with respect to μ. This results in the generalized eigenvalue problem

$$\boldsymbol{C}(\tau)\boldsymbol{A} = \hat{\boldsymbol{\Lambda}}(\tau)\boldsymbol{S}\boldsymbol{A} \tag{18}$$

where $\boldsymbol{A} = (a_{ij})_{0 \leq i,j \leq n}$ contains the expansion coefficients and $\hat{\boldsymbol{\Lambda}}(\tau) = \mathrm{diag}(\hat{\lambda}_0(\tau), \ldots, \hat{\lambda}_n(\tau))$ comprises the estimated eigenvalues. The correlation matrix $\boldsymbol{C}(\tau)$ and the overlap matrix

$\boldsymbol{S}$ are defined via

$$C_{ij}(\tau) = \langle \chi_i, \mathcal{T}_\tau \chi_j \rangle_\mu \tag{19}$$

$$S_{ij} = \langle \chi_i, \chi_j \rangle_\mu . \tag{20}$$

Hence, the optimal approximations of the first n eigenfunctions $\hat{r}_j$ within the subspace V as well as the corresponding estimates of the eigenvalues $\hat{\lambda}_j(\tau)$ are obtained by solving the generalized eigenvalue problem in eq. 18.

Given a set of basis functions $\mathcal{B} = \{\chi_0, \ldots, \chi_n\}$, the integrals in eqs. 19 and 20 have to be calculated in order to apply the variational method via solving eq 18. Since the transfer operator is usually not known analytically for non-trivial systems, an analytical solution of eqs. 19 and 20 is not available. However, these integrals have the interpretation of (time-lagged) correlations with respect to the given Markov process $(X_t)_{t\in T}$ via

$$\begin{aligned} C_{ij}(\tau) &= \mathrm{cor}(\chi_i, \chi_j; \tau) && (21)\\ &= \mathbb{E}(\chi_i(X_t)\chi_j(X_{t+\tau})) \\ &= \int_\Omega \chi_i(x)\mathcal{T}_\tau\chi_j(x)\mu(\mathrm{d}x) \\ &= \langle \chi_i, \mathcal{T}_\tau \chi_j \rangle_\mu && (22) \end{aligned}$$

where $\tau = 0$ simply yields the inner product of the basis functions, that is the elements of the overlap matrix S_{ij}. These correlations can be estimated from realizations of the process such as an MD trajectory. Let $(x_t)_{t=0}^{N_T}$ be a time-discretized time series, where Δ is the time step of the time series and the total lenght of the time series is N_T time steps. The lag time is given as $\tau = n_\tau \Delta$. Estimates of the above quantities are then given by

$$\widehat{C}_{ij}(\tau) = \widehat{\mathrm{cor}}(\chi_i, \chi_j; \tau) = \frac{1}{N_T - n_\tau} \sum_{t=0}^{N_T - n_\tau} \chi_i(x_t)\chi_j(x_{t+n_\tau}) \tag{23}$$

and

$$\widehat{S}_{ij} = \widehat{\mathrm{cor}}(\chi_i, \chi_j; \tau = 0) = \frac{1}{N_T} \sum_{t=0}^{N_T} \chi_i(x_t)\chi_j(x_t). \tag{24}$$

B. Coordinates and basis sets for peptide dynamics

The space B (eq. 16) spanned by the basis functions is not the same as the dominant subspace D (eq. 14). To obtain an accurate model, D should be contained in B. To achieve this,

1. the basis functions should be defined on coordinates that can encompass all important conformational changes within the peptide dynamics.

2. the basis functions should model the relevant dynamical processes of the system, i.e. they should be as close to the actual eigenfunctions of the propagator as possible.

One would like to compare models of different systems directly. Thus,

3. the basis functions should be designed independently of the actual system of interest such that they can be used for any kind of peptide sequence, i.e. they should not be data-driven.

Requirement 2 is trivially achieved by choosing a large basis set. However, finite sampling induces a statistical error in eq. 23 which increases with the basis set size. Thus,

4. the basis set size should be kept as small as possible.

In Ref. 27 we developed a basis set that accommodates these requirements. We briefly summarize the construction of the peptide basis set and explain how it relates to the four requirements. We denote the stationary density associated to a particular force field by $\mu(x)$ and remind the reader that different force fields give rise to different stationary probability densities[31].

Ad 1. The slow dynamical processes of peptides and proteins can be well described by the φ- and ψ-backbone torsion angles, and thus dynamic models are frequently constructed on the backbone torsion angles[26]. The associated state space is

$$\Omega^{\mathrm{bb}} = \Omega_1 \times \cdots \times \Omega_N \,, \tag{25}$$

where $\Omega_r = S^1 \times S^1$ is the Ramachandran space of the rth residue, parametrized by the ϕ- and ψ-backbone dihedral angles (here S^1 denotes the circle). N is the number of residues in the peptide. Let $p\colon \Omega \to \Omega^{\mathrm{bb}}$ be the associated projection. The underlying assumption is that the projected Markov process $p(X_t)$ is still Markovian and inherits all the properties from the original process. If this is satisfied, then the transfer operator of $p(X_t)$ will be the projection of $\mathcal{T}$ along p and we may determine all dynamical properties of the original process by investigating those of the projected process.

Thus, by abuse of notation, we consider the Markov process X_t to be Ω^{bb}-valued and ignore the intermediate projection step. Hence, $X_t = (X_t^{(1)}, \dots, X_t^{(N)})$ with $X_t^{(r)}$ being the projection of X_t onto Ω_r. We will refer to the process $X_t^{(r)}$ as the *local Markov process at residue* r.

Ad 2. The state space Ω^{bb} is still very high-dimensional. To systematically construct basis functions on this state space, we approximate the stationary probability distribution μ as a tensor product

$$\mu \approx \mu^{\mathrm{tensor}} = \mu_1 \otimes \cdots \otimes \mu_N. \tag{26}$$

where $\mu_r : \Omega_r \to [0, 1]$ is the marginal probability measure on Ω_r defined via

$$\mu_r(A) = \mu(\Omega_1 \times \cdots \times \Omega_{r-1} \times A \times \Omega_{r+1} \times \cdots \times \Omega_N). \tag{27}$$

Note that this measure is just the stationary measure of the local Markov process $X_t^{(r)}$ at residue r.

The above approximation is justified if and only if the random variables $X_t^{(r)}$ are mutually independent. This can be assumed because neighboring pairs of Ramachandran spaces Ω_r and Ω_{r+1} are separated by a rigid torsion angle, the peptide bond, which minimizes the correlation of residues r and $r \pm 1$. The approximation is frequently used in Markov state models of peptide dynamics, in which the microstates are constructed by discretizing the Ramachandran plane[26]. It implies

$$L^2(\Omega^{\mathrm{bb}}, \mu^{\mathrm{tensor}}) \cong L^2(\Omega_1, \mu_1) \otimes \cdots \otimes L^2(\Omega_N, \mu_N), \tag{28}$$

thus the dominant eigenfunctions r_k can be approximated by tensors of residue-centered functions. However, the resulting space is still infinite-dimensional as the local L^2-spaces are. We therefore approximate the local dynamics at each residue r by a subspace spanned by m_r residue-centered functions R_l^r, which represent the m_r slow dynamic modes within this residue. The associated local Hilbert space is

$$D_r \coloneqq \mathrm{span}\left(R_0^r, \ldots, R_{m_r}^r\right) \subset L^2(\Omega_r, \mu_r)\,, \tag{29}$$

with $l \in 0, \ldots m_r$, and for most residues $m_r = 2$. The local Hilbert space has a weighted scalar product, where the weight is given by the marginal probability distribution μ_r. We subsequently define the ansatz space B as

$$B \coloneqq D_1 \otimes \cdots \otimes D_N. \tag{30}$$

Thus, if the stationary measure is well-approximated by a tensor product (eq. 26), it suffices to know the residue-centered functions R^r_l of each residue r to construct a basis set which fulfills requirement 2. The full basis set then consists of all possible combinations of residue-centered functions, i.e.

$$\mathcal{B} = \left\{ R^1_{l_1} \otimes \cdots \otimes R^N_{l_N} \;\middle|\; 0 \leq l_r \leq m_r \right\}. \tag{31}$$

Ad 3. In principle, one could estimate the residue-centered functions from a simulation of the full peptide. In this case, the residue-centered functions would vary from system to system, and one would obtain a data-driven basis set. We choose a different approach and obtain the residue-centered functions from a set of reference simulations. The approach is justified, because the residue centered functions $R^r_{l_r}$ depend on the type of residue, but are known to be largely independent from the remaining peptide sequence[27,31]. Thus, it should be possible to replace the residue-centered Hilbert space D_r of a specific residue r by a generic Hilbert space D_X of the corresponding type of residue, i.e. $D_r \approx D_X$, where X is a place-holder for the single-letter amino acid code of residue r. As an example, consider the pentapeptide Ac-Ala-Val-Ala-Val-Ala-NHMe, where Ac denotes an acetyl group, and NHMe denotes a methyl amine group. We replace the space spanned by the five residue-centered Hilbert spaces $D_{r=1,\ldots,5}$ by an appropriate combination of the generic amino-acid specific Hilbert spaces D_{A} and D_V: $D_1 \otimes D_2 \otimes D_3 \otimes D_4 \otimes D_5 \approx D_{\mathrm{A}} \otimes D_{\mathrm{V}} \otimes D_{\mathrm{A}} \otimes D_{\mathrm{V}} \otimes D_{\mathrm{A}}$.

We obtain the generic amino-acid specific Hilbert space $D_X = \mathrm{span}(R^X_0, \ldots, R^X_{m_X})$ by constructing Markov state models of the corresponding capped amino acid Ac-X-NHMe. Fig. 1 shows the dominant transfer operator eigenfunctions of alanine-dipeptide (Ac-A-NHMe)[26] that span $D_{\mathrm{A}} = \mathrm{span}(R^A_0, R^A_1, R^A_2)$. R^A_0 is the first right eigenvector of the MSM transition matrix (approximates the first eigenvector of the associated transfer operator) and corresponds to the characteristic function of the accessible conformational space $\mathbf{1}_{\Omega_r}$. R^A_1 and R^A_2 represent kinetic exchange processes between regions with the negative sign (blue) and regions with positive sign (red). That is, R^A_1 represents a transition along the φ-torsion angle, and R^A_2 represents a transition along the ψ-torsion angle. The residue-centered functions of each type of amino acid are pre-calculated. For all residues except proline, the generic residue-centered Hilbert space D_X is spanned by three functions as exemplified by alanine above. In the case of proline, the ϕ-torsion angle is rigid, thus its generic residue-centered Hilbert space D_{P} is spanned by only two functions, i.e. $D_{\mathrm{P}} = \mathrm{span}(R^{\mathrm{P}}_0, R^{\mathrm{P}}_1)$ where R^{P}_0 again corresponds to the characteristic function of the total accessible space and R_1^{P} represents the ψ-torsion. The basis set for a given peptide is eventually obtained by combining the residue-centered functions ordered according to the sequence of the peptide (eq. 31). In this way we obtain a basis set that is independent of the peptide simulation.

Ad 4. With three dominant eigenfunction for most residues, the basis set grows as 3^N, where N is the number of residues. This full basis set becomes computationally intractable for even short peptides. However, the basis functions can be classified according to the number of generic residue-centered functions R_l^X in the tensor product, which differ from R_1^X. The basis function $\chi_0 = R_0^1 \otimes R_0^2 \cdots \otimes R_0^N$, which only consists of residue-centered functions R_0^X, always needs to be included in the basis set, because it best approximates the stationary process r_0. In fact, $\chi_0 = R_0^1 \otimes R_0^2 \cdots \otimes R_0^N$ is the characteristic function of the accessible space of the peptide as estimated from the tensor product of the local accessible spaces. Thus, if the conformational ensemble of the peptide only contains conformations which are part of this accessible space, $\hat{r}_0 = \chi_0$, and the corresponding vector of expansion coefficients is $\mathbf{a}_0 = (1, 0, \dots 0)^\top$. The basis set can be systematically expanded by adding basis functions in which one residue differs from R_0^X (single basis set), two residue differ from R_0^X (double basis set), three residue differ from R_0^X (triple basis set) etc. The single basis set grows linearly, and the double basis set grows quadratically with the number or residues.

C. Discretization of state space

In practice, the local functions R_l^X are obtained by constructing a Markov state model from MD simulations of the respective capped amino acid Ac-X-NHMe. This procedure requires a partition of Ramachandran space Ω_r into (disjoint and non-overlapping) microstates. We use the same partition of $36 \times 36 = 1296$ microstates for each amino acid and denote the set of microstates by $S = \{s_1, \dots s_{1296}\}$. We calculate the MSM transition matrix $T(\tau)$ on this discrete state space and approximate the dominant eigenfunctions of the underlying transfer operator as its right eigenvectors

$$R_l^X = \sum_{s_i \in S} \alpha_l^X(s_i)\, \mathbf{1}_i \tag{32}$$

where $\mathbf{1}_i$ is the characteristic function of microstate s_i and $\alpha_l^X(s_i)$ is the ith component of the respective eigenvector. Thus R_l^X is a step function that is piecewise constant on each microstate $s_i \in S$. This drastically simplifies the calculation of the tensor product in eq. 31. Given a peptide conformation (MD snap shot) x_t, the (ϕ, ψ)-coordinates of each residue are projected onto the corresponding microstates, which yields a discretized trajectory $\overline{x}_t \in \mathbb{N}^N$. The basis functions as defined in eq. 31 are then given as

$$R_{l_1}^1 \otimes \cdots \otimes R_{l_N}^N(x_t) = \prod_{r=1}^{N} \alpha_{l_r}^{X_r}(\overline{x}_t) \tag{33}$$

where X_r is the residue type of residue r.

Straightforwardly, the elements of the correlation matrix and the overlap matrix can be estimated by inserting the equation above into eq. 23 which in turn allows for the approximation of the dominant subspace D via solving the generalized eigenvalue problem in eq. 18.

III. EXTENSION OF THE METHOD

We recomputed a variational model of the VGVAPG peptide using the basis set as discussed so far (Fig. 2.A). The model is analogous to the model reported in Fig. 6c and 6d in ref. 26 . The scatter plot shows trajectory snapshots projected onto the dominant subspace of the model. We manually clustered the data points into metastable states, whose structures are shown below the scatter plot. Note that since the publication of ref. 26 we changed the index convention of the transfer operator and propagator eigenfunctions. Previously, the stationary process was denoted r_1 and l_1 respectively. Now we “start counting at zero” and the stationary process is r_0 and l_0 respectively. Hence, the “first process” in Fig. 2.A corrsponds to “process 2” in Fig. 6c in ref. 26 etc. The histograms below the scatter plot show the absolute values of the expansion coefficients a_{ij} for $j = 0, 1, 2, 3$, and are consistent with Fig. 6c in ref. 26. The stationary process r_0 is represented by a single basis set $\chi_0 = R_0^{\mathrm{G}} \otimes R_0^{\mathrm{V}} \otimes R_0^{\mathrm{A}} \otimes R_0^{\mathrm{P}} \otimes R_0^{\mathrm{G}}$, which indicates that the stationary probability density can be well approximated by a tensor product (eq. 26) and by extension that the tensor approximation holds for this peptide. We additionally introduce an activity display below each of the slow processes which shows to which degree each of the torsion angles is affected by the slow process. For each residue, we show two squares representing the ϕ-

and ψ-torsion angle (left and right square). The color-code in each square was calculated such that the color intensity is proportional to the sum of all expansion coefficients that correspond to those basis functions which model the corresponding local exchange at the respective residue. If only few basis functions contribute to given process the activity display is somewhat redundant to directly looking up the interpretation of basis functions with large expansion coefficients. However, the display becomes very useful if the basis set is large, and if more than a hand full of basis functions contribute to a given process.

A. Renormalizing the basis set

A major benefit of constucting the basis set as a tensor of residue-centered functions is that the results can be interpreted in a straightforward manner. Since each basis function χ_i represents a probability exchange between well-defined local conformations, an eigenfunction r_j can be interpreted as a superposition of these exchange processes (see eq. 17). The interpretation is particularly simple if the superposition is dominated by a few large expansion coefficients a_{ij}. For example, the first slow process r_1 is dominated by the basis functions $\chi_5 = R_0^G \otimes R_0^V \otimes R_1^A \otimes R_0^P \otimes R_0^G$ and $\chi_{26} = R_0^G \otimes R_2^V \otimes R_1^A \otimes R_0^P \otimes R_0^G$, corresponding to a ϕ-torsion in A_4 coupled to a ψ-torsion in V_3. In theory, the expansion coefficient vectors in eq. 18 should hence suffice to obtain a structural interpretation of the dominant eigenfunctions. This interpretation however assumes that the basis functions are orthonormal. As a consequence, we add a fifth requirement to the list in section II B:

5. the basis set should be orthonormal.

Orthonormality is defined with respect to the inner product as defined in eq. 4. That is, it is weighted by the stationary probability distribution of the current molecular system. This means that the inner product changes with each molecular system and with each approximation to the dynamics of the system.

Let us survey the stationary probability distributions that have been introduced so far. The stationary probability distribution of the system is defined in eq. 2, and approximations are introduced in eq 26

$$\mu \approx \mu^{\text{tensor}} = \mu_1 \otimes \cdots \otimes \mu_N \,,$$

where $\mu_r : \Omega_r \to [0, 1]$ is the marginal probability distribution on the Ramachandran space of residue r defined in eq. 27. By "stationary probability distribution of the system" we mean the Boltzmann distribution assocatiated to the force-field used in the simulation. It is well understood that this probability distribution will differ from force field to force field and that each of these force-field-asscociated distributions is only an approximation to the true stationary probability distribution of the molecular system. Additionally, for finite sampling there is always a deviation between the stationary probability distribution as estimated from the MD trajectory $\hat{\mu}$ and the stationary probability distribution associated to the force field μ. However, for now we will assume that $\hat{\mu} = \mu$. Effects due to finite sampling will be discussed in section III B.

The method of linear variation guarantees that the estimated eigenvectors $\hat{r}_i$ are orthonormal with respect to $\hat{\mu} = \mu$, i.e. $\langle \hat{r}_i, \hat{r}_j \rangle_{\hat{\mu}=\mu} = \delta_{ij}$. The generic basis set $\mathcal{B} = \{\chi_0, \ldots, \chi_n\}$ discussed so far will typically not be orthonormal with respect to this scalar product, i.e. $\langle \chi_i, \chi_j \rangle_{\hat{\mu}=\mu} \neq \delta_{ij}$, because the mariginal stationary density of a residue within a peptide chain differs from the stationary density of the corresponding capped amino acid used to constuct the residue-centered functions for the basis set.

In the variational model with the generic basis set for VGVAPG (fig. 2.A and 3A (red bars)) one can see in the third process that the generic basis set is not fully orthonormal with respect to $\hat{\mu} = \mu$. In the model, the stationary process $\hat{r}_0$ is represented by a single basis function: $\hat{r}_0 = \chi_0$. Thus, if the basis set was orthonormal, χ_0 should not contribute to any other process. However in the third process, clearly $a_{03} \neq 0$. Similar to this, other spurious expansion coefficients might arise to ensure that the dominant eigenfunctions are orthonormal with respect to $\hat{\mu} = \mu$. Interpreting these additional expansion coefficients as conformational exchange processes would be incorrect. We therefore recommend to orthonormalize the basis set with respect to the marginal probability densities of the peptide residues prior to the construction of the variational model. After renormalizing the basis set, the expansion coefficient for χ_0 vanishes (fig. 3A (blue bars)). Other expansion coefficients in VGVAPG were not altered by the renormalization.

Note that the generic model is nonetheless a valid model of the peptide dynamics, and that the dominant eigenvectors can be submitted to the usual analyses, such as extraction of metastable states by a PCCA+ analysis. The generic model only lacks the additional benefit that the expansion coefficients can be interpreted directly. This is also demonstrated by the fact that both models, generic basis set and orthonormalized basis set, yield identical implied timescale plots (fig. 3B).

How is the basis set orthonormalized? We constructed the basis set as a tensor product (eq. 31). As a consequence, we can achieve exact orthonormalization with respect to μ^{tensor}, but only approximate orthonormalization with respect to $\hat{\mu} = \mu$ (eq. 26). By construction, the generic basis functions are orthonormal with respect to a generic probability measure

$$\mu^{\text{generic}} := \mu_1^{\text{generic}} \otimes \cdots \otimes \mu_N^{\text{generic}} \tag{34}$$

where μ_r^{generic} is the stationary probability distribution of the capped amino acid Ac-X-NHMe corresponding to residue r. That is, μ^{generic} is the tensor product of the equilibrium probability densities μ_r^{generic} of the individual residues as obtained from the reference simulations of the capped amino acids.

μ^{tensor} deviates from μ^{generic} if $\mu_r \neq \mu_r^{\text{generic}}$ for any residue r, i.e. if the probability density of a residues r in the peptide chain differs from the probability density in the reference simulation of the corresponding capped amino acid. It is hence straightforward to achieve orthonormality with respect to μ^{tensor} by ensuring that the residue centered basis functions $\{R_1^r \dots R_{l_r}^r\}$ of each residue r are orthonormal with respect to the corresponding marginal probability density μ_r. Because of the tensor product structure of the basis functions, the scalar product is equal to a product of residue-centered scalar products

$$\left\langle R_{k_1}^1 \otimes \cdots \otimes R_{k_N}^N \ , \ R_{l_1}^1 \otimes \cdots \otimes R_{l_N}^N \right\rangle_{\mu_1 \otimes \cdots \otimes \mu_N} = \prod_{r=1}^{N} \left\langle R_{k_r}^r \ , \ R_{l_r}^r \right\rangle_{\mu_r} \tag{35}$$

with

$$\left\langle R_{l_r}^r, R_{k_r}^r \right\rangle_{\mu_r} = \int_{\Omega_r} R_{l_r}^r(x) R_{k_r}^r(x) \mu_r(\mathrm{d}x). \tag{36}$$

Using the discretization in eq. 32, eq. 36 reads

$$\begin{aligned} \left\langle R_{l_r}^r, R_{k_r}^r \right\rangle_{\mu_r} &= \left\langle \sum_{s_r \in P_r} \alpha_{l_r}(s_r) \mathbf{1}_{s_r}, \sum_{s_r \in P_r} \alpha_{k_r}(s_r) \mathbf{1}_{s_r} \right\rangle_{\mu_r} \\ &= \sum_{s_i \in S} \alpha_{l_r}(s_i) \alpha_{k_r}(s_i) \mu_r(s_i). \end{aligned} \tag{37}$$

$\mu_r(s_i)$ is estimated from the simulation of the peptide, and based on eq. 37 the residue-centered functions are orthonormalized using the Gram-Schmidt method.

B. Validity of the tensor ansatz

Dynamic processes which require the concerted movements of two residues l and s are represented within the peptide basis by either superpositioning the corresponding singly active basis functions with expansion coefficients a and b (e.g. $\hat{r}_i = a(R_0^1 \otimes \dots R_1^l \otimes R_0^s \dots \otimes R_0^N) + b(R_0^1 \otimes \dots R_0^l \otimes R_2^s \dots \otimes R_0^N) + \dots$); or by including the corresponding doubly active basis function in the superposition with expansion coefficient c (e.g. $\hat{r}_i = c(R_0^1 \otimes \dots R_1^l \otimes R_2^s \dots \otimes R_0^N) + \dots$). However, peptide dynamics in which the local stationary probability densities of two residues l and s are mutually dependent cannot be represented within the peptide basis set. This is the case if the probability of some $A \subset D_r$ depends on the conformation of residue s and vice versa. Then the tensor ansatz for the probability distribution fails $\mu \neq \mu_1 \otimes \dots \mu_l \otimes \mu_s \dots \otimes \mu_N$, and should, in principle, be replaced by the corresponding joint distribution of the two local subspaces $\mu = \mu_1 \otimes \dots \mu_{l,s} \dots \otimes \mu_N$.

After the basis set has been orthonormalized with respect to μ^{tensor}, any remaining deviation from a true orthonormal basis set with respect to $\hat{\mu} = \mu$ is due to a deviation of the tensor approximation μ^{tensor} (eq. 26) from the full probability distribution of the system $\hat{\mu} = \mu$. A telltale sign for a deviation from the tensor approximation is if the stationary process is not exclusively represented by the first basis function, i.e. if $\hat{r}_0 \neq \chi_0$ or equivalently $a_{i0} > 0$ for $i > 0$. The variational model of VGVAPG does not exhibit this warning sign (Fig. 3A, blue lines). However, the variational model of the β-hairpin shows several expansion coefficients $a_{i0} > 0$ for the stationary process after renormalization (Fig. 5.A).

The residual between μ^{tensor} and $\hat{\mu} = \mu$ represents the mutual dependence between the residue-centered processes. We can test for this by analyzing the overlap matrix $\boldsymbol{S}$. An orthonormal basis set gives rise to an overlap matrix $\boldsymbol{S}$ which is equal to the identity matrix $\boldsymbol{I}$. The elements of the overlap matrix $\boldsymbol{S}$ are calculated via eq. 24, and because the MD trajectory (x_t) samples the full probability density of the system $\hat{\mu} = \mu$ rather than its tensor approximation μ^{tensor}, the elements of the overlap matrix $\boldsymbol{S}$ represent scalar products $\langle \chi_i, \chi_j \rangle_\mu$ with respect to $\hat{\mu} = \mu$. Thus, if after orthonormalizing the basis set with respect to μ^{tensor} the overlap matrix $\boldsymbol{S}$ deviates from $\boldsymbol{I}$, there must be a residual between μ^{tensor} and $\hat{\mu} = \mu$. To trace back this residual to specific mutually dependent local Markov processes, we analyze the diagonal elements of the overlap matrix that deviate from 1 by more than an order of magnitude, i.e. $\langle \chi_i, \chi_i \rangle_\mu < 0.1$ or $\langle \chi_i, \chi_i \rangle_\mu > 10$.

Fig. 3C shows the diagonal elements of the overlap matrix of VGVAPG (double basis set, orthonormalized) sorted by size. We find only minor deviations from the target value 1. However, a single basis functions gives rise to a very small value close to zero, which is indicated by the blue-shaded area in the figure. The function in question is a double basis function encoding ϕ-exchange at the third residue (valine) and ψ-exchange at the forth position (alanine). This suggests that the marginal probability distributions at these positions are mutually dependent, i.e. $\mu_3 \otimes \mu_4 \neq \mu_{3,4}$. In the next section we provide evidence that this is indeed the case.

Figure 6A shows the diagonal elements of the overlap matrix of a β-hairpin peptide (double basis set, orthonormalized) sorted by size. A majority of basis functions are found to be normalized with respect to $\hat{\mu} = \mu$ but a small subset of functions can be identified for which the squared norm deviates from 1 by more than an order of magnitude (highlighted by the blue-shaded area in the figure). This indicates that some of the local Markov processes defined by the dynamics of the β-hairpin peptide are mutually dependent. This brings up the question: what causes these mutual dependencies?

C. Accounting for undersampled transitions

Up to now, we have assumed that the probability measure estimated from the simulation is identical to the probability measure associated to the force field, i.e. that $\hat{\mu} = \mu$. However, due to finite sampling the estimated probability distribution $\hat{\mu}$ always slightly deviates from the probability distribution of the force field μ. Since the matrix elements of the overlap matrix are estimated from the MD trajectory (eq. 24), an analysis of the overlap matrix can only test whether $\hat{\mu}$ deviates from the tensor structure. The sampled probability measure can deviated from the tensor structure, either because the actual probability measure of the force field μ deviates from the tensor structure, or because finite sampling induces an apparent mutual dependence in $\hat{\mu}$.

Figure 4A,B shows an example of a basis function from the VGVAPG peptide for which the diagonal element of the overlap matrix is smaller than 0.1. The basis function in fig. 4A,B is the doubly active function $R_1^{\mathrm{V}} \otimes R_1^{\mathrm{G}} \otimes \boldsymbol{R_2^{\mathrm{V}}} \otimes \boldsymbol{R_2^{\mathrm{A}}} \otimes R_1^{\mathrm{P}} \otimes R_1^{\mathrm{G}}$. It represents the correlated transition along the ϕ-torsion angles of residues V_3 and A_4; or equivalently: the correlated kinetic exchange between the L_α regions and the α/β regions of the two Ramachandran spaces. Figure 4B shows R_2^{V} and R_2^{A} separately. Figure 4A shows $R_2^{\mathrm{V}} \otimes R_2^{\mathrm{A}}$ on the combined state space $\Omega_3 \times \Omega_4$ of residues 3 and 4. We sorted the underlying 1296 microstates of each residue by region (A_{L_α} or $A_{\alpha\beta}$), which gives rise to the quadrant structure. The sign structure of the function shows that it represents probability exchange between the conformations in which both residues are in the L_α conformation (negative signs in the upper right quadrant $A^3_{L_\alpha} \times A^4_{L_\alpha}$) and the remaining state space (nonnegative signs in all other quadrants).

The fact that the corresponding diagonal element of $\boldsymbol{S}$ deviates from 1 implies that, within the local space $\Omega_3 \times \Omega_4$, the sampled joint measure $\hat{\mu}_{3,4}$ deviates from the tensor measure $\hat{\mu}_3 \otimes \hat{\mu}_4$. Fig. 4.C shows the relative error of these local measures

$$\Delta_{\mathrm{rel}}(A) \coloneqq \frac{\hat{\mu}_{3,4}(A) - \hat{\mu}_3 \otimes \hat{\mu}_4(A)}{\mu_{3,4}(A)} \tag{38}$$

for measurable $A \subset \Omega_3 \times \Omega_4$, where we define sets of the form $A^3_i \times A^4_j$ with $i, j \in \{\alpha, \beta, L_\alpha\}$. Negative values of $\Delta_{\mathrm{rel}}(A)$ mean that the simulation visits the set A less frequently than would be expected from the tensor measure $\hat{\mu}_3 \otimes \hat{\mu}_4$. This is the case for the conformation $A^3_{L_\alpha} \times A^4_{L_\alpha}$ and to a lesser extent for $A^3_\alpha \times A^4_{L_\alpha}$ In fact, the region $A^4_{L_\alpha}$ is visited by less than 1% of all simulation data points. If this 1% percent fraction of the MD data set is not sufficient to fully sample the marginal distribtion μ_3 in the neighboring residue, an apparent mutual dependence between Ω_3 and Ω_4 arises. This is indeed the case. The upper panel in fig. 4.D shows the marginal distribution $\hat{\mu}_3$ estimated from the full MD data set, whereas the lower panel shows the conditional marginal distribution $\hat{\mu}_3^{L_\alpha}$ estimated from the 1% percent fraction of the MD data set where residue 4 populates $A^4_{L_\alpha}$. The region $A^3_{L_\alpha}$ is almost never visited within this conditional fraction of the MD data set, and this generates a mutual dependence in $\hat{\mu}_{3,4}$. Since residues 3 and 4 are valine and alanine, there is no reason why the conformation $A^3_{L_\alpha} \times A^4_{L_\alpha}$ should be sterically prohibited. We thus conclude that the mutual dependence is caused by insufficient sampling rather than by an actual mutual dependence in the underlying force field.

In figure 6B-E, an analogous situation as above is shown for the β-hairpin peptide. The β-hairpin peptide is a 14-residue peptide with the sequence RGKITVNGKTYEGR. The basis function in fig. 6B,C is the doubly active function $R_1^{R} \otimes R_1^{\mathrm{G}} \otimes R_1^{\mathrm{K}} \otimes R_1^{\mathrm{I}} \otimes R_1^{\mathrm{T}} \otimes R_1^{\mathrm{V}} \otimes R_1^{\mathrm{N}} \otimes R_1^{\mathrm{G}} \otimes \boldsymbol{R}_{\boldsymbol{2}}^{\mathrm{K}} \otimes R_1^{\mathrm{T}} \otimes \boldsymbol{R}_{\boldsymbol{2}}^{\mathrm{Y}} \otimes R_1^{\mathrm{E}} \otimes R_1^{\mathrm{G}} \otimes R_1^{\mathrm{R}}$. As above, this function gives rise to a diagonal element of the overlap matrix that is smaller than 0.1. In figure 6D the relative error between the joint measure $\mu_{9,11}$ and the tensor measure $\mu_9 \otimes \mu_{11}$ is depicted. Similar to the situation for the VGVAPG peptide, both measures deviate significantly within both $A^9_\alpha \times A^{11}_{L_\alpha}$ and $A^9_{L_\alpha} \times A^{11}_{L_\alpha}$. In both cases, $\mu_{9,11}$ is dominated by $\mu_9 \otimes \mu_{11}$. Accordingly, the deviation between the marginal probability distribution in the upper panel of figure 6E and the conditional probability distribution $\mu_9^{L_\alpha}$ in the lower panel of figure 6E is predominantly present in the α- and L_α-region of residue 9. Again, this deviation can be explained by insufficient sampling as the conditional distribution has been generated by less than 1% of the frames.

Apparent mutual dependencies will generate misleading expansion coefficients for the corresponding basis functions. We therefore suggest to construct a reduced orthonormal basis set by removing basis functions for which the diagonal element of $\boldsymbol{S}$ differs by more than an order of magnitude from 1 from the basis set. The effect of the reduced orthonormal basis set on the histogram of the expansion coefficients is shown in fig. 2B and fig. 3 for VGVAPG, and in fig. 5A for the β-hairpin peptide. In fact, the histograms and consequently the intepretation changes slightly if undersampled basis functions are removed from the basis set. Alternatively, one could use an iterative approach in which undersampled transitions are identified by an analysis of the overlap matrix, and additional simulations are started from the undersampled regions, until apparent mutual dependencies due to limited sampling can be separated from actual mutual dependencies in the underlying force field.

IV. SECONDARY STRUCTURE FORMATION

It is not a priori clear that our basis set which is based on residue-centered local functions can model a concerted conformational rearrangements such as secondary structure formation. We test this be constructing variational models for the β-hairpin peptide RGKITVNGKTYEGR and a fifteen-residue fragment of the human islet amylin polypeptide, which from a wide variety of conformations including an α-helix. In both cases, we use an orthonormalized reduced double basis set. Both models pass the implied timescale test (figs. 5.C and 9.B), indicating that our basis set indeed yields well-converged models of secondary structure formation.

A. β-Hairpin Peptide

Fig. 7B shows the expansion coefficients of the variational model associated with the stationary process and the three dominant conformational exchange processes. Process r_1 is mostly represented by a single basis function which can be mapped to a transition in the ϕ-torsion angle of resdiue E_{12}. The basis function corresponding ot the remaining expansion coefficients affect the ϕ-torsion angle in N_7 and Y_{11}. This is summarized in an activity display below the histogram of expansion coefficients. The colour intensity is proportional to the sum of expansion coefficients of the basis functions that model a conformational transition at this position. Since two local transition processes (along the φ- and ψ-axis, respectively) are distinguished, we show two squares for each residue of which the left one represents φ- and the right one represents ψ-exchange. Process r_2 is affects the ϕ torsion angles for residues N_7 and E_{12}, and the ψ-torsion angles of T_5, V_6, and K_9. Process r_3 affects the full Ramachandran space of residues N_7, T_{10}, and Y_{11}, as well as the ψ-torsion angles of I_4, and V_6.

In figure 7A, the projection of the β-hairpin trajectory onto the dominant subspace of the first three processes is shown. As expected from previous theoretical and empirical results, the projected trajectory resides within a 3-simplex with its vertices being interpreted as the metastable conformations[32]. We therefore identify four clusters that should represent the long-lived conformations of the system. A more thorough characterization of these clusters is given in fig. 8 in terms of hydrogen bonds and secondary structures, as determined by the DSSP algorithm[33]. We find these structures to be in excellent agreement with previous findings using the core set method[15]. In fact, the assignement of trajectory frames to clusters C_2, C_3, and C_4 is almost identical in the core-set model and variational model. These three cluster represent various β-hairpin conformations that mainly differ in the position and size of the central loop region. The variational model additionally identifies cluster C_1, a fairly loose β-hairpin that is less stabilized by hydrogen bonds compared to the other structures. C_1 has a low population of only 3 % of the trajectory frames. The core-set method discourages clusters with low population, which might explain why the cluster is only identified by the variational model.

The patterns of dynamically active residues in the estimates of the eigenfunctions can be linked to the structural transitions between the metastable conformations. As these mainly comprise positional shifts of the β-hairpin, the primarily active regions within the sequence are those that undergo transitions between loop and β-bridge conformations. Hence, we find the active residues in the eigenfunctions to be part of transitional regions of the molecule, i.e. regions that are found to undergo conformational transitions between the metastable structures. Without additional information on the conformation of the non-active residues, the patterns of dynamically active residues are however not enough to long-lived conformations and the transitions between them.

B. hIAPP Fragment

Figure 9 shows the variational Markov model for the hIAPP fragment HSSNNFGAILSSTNV. This molecule is particularly difficult to model with conventional MSMs, because it explores such a wide range of different conformations. This is also reflected in the histogram of expansion coefficients, which - in contrast to the β-hairpin peptide - show that a large number of basis functions contribute to each of the slow processes in the hIAPP fragment. Nonetheless, each process can be localized onto torsional angle transitions in only a few residues. In the three slow processes, residues S_3 and N_4 move concertedly with neighboring residues and with various residues from the central segment AILSS.

Analogous to the model for the β-hairpin peptide, the MD trajectory forms a 3-simplex within the coordinates of the dominant eigenfunctions and its vertices are interpreted as the metastable conformations of the system. As opposed to the β-hairpin system, however, the center of the simplex comprises a highly ordered α-helical structure (Cluster 0 in fig 9A) that moreover corresponds to the highest poplated conformation in the simulation data. These findings suggest a kinetic model where the central conformation (Cluster 0) occupies the lowest free energy level of the conformational space and the dominant processes represent transitions of this conformation into the surrounding clusters.

The stationary process r_0 is not represented exclusively by the first basis functions. This implies that the tensor approximation is not as cleanly fulfilled as in the β-hairpin peptide. Nonetheless, the well-converged implied timescales (fig 9A)) and the clear separation of the conformations in the dominant eigenspaces show that the variational model is an accurate model of the conformational transitions from the α-helical structure to various other folded structure.

V. COMPUTATIONAL DETAILS

A. MD simulations

a. hIAPP fragment. We performed all-atom MD simulations of the fragment HSSNNFGAILSSTNV (residues 18-32) of human islet amyloid polypeptide (hIAPP). The peptide was acetylated at the N-terminus and methylated at the C-terminus. Starting structures where obtained from an NMR structure of hIAPP in an membrane environment (PDB ID: 2L86)[34]. The simulations were performed with the AMBER ff99SB-ILDN[35] force field in explicit water (TIP3P[36] water model), using the GROMACS simulation package[37] (versions 4.4.5 and 5.0.2). The NVT ensemble was applied, where the V-Rescale thermostat[38] was used to restrain the temperature to 300 K. Cubic boxes, with a minimum distance between solute and box walls of 1 nm, were used. After an initial equilibration of 100 ps, ten structures were selected randomly from the trajectory and used as starting conformations for independent simulation runs, yielding a total simulation time of 13.5 μs. The atom positions of the solute were saved every 1 ps. We used the leap-frog integrator and applied periodic boundary conditions in all directions. The LINCS algorithm[39] was used to constrain all bonds to hydrogen atoms (lincs_iter = 1, lincs_order = 4), allowing for a integration time step of 2 fs. Lennard-Jones interactions were cut off at 1 nm. The Particle-Mesh Ewald (PME) algorithm[40] was applied to treat electrostatic interactions, with a real space cutoff of 1 nm, a grid spacing of 0.15 nm, and an interpolation order of 4.

b. Hairpin peptide The all-atom MD simulations of the β-hairpin peptide RGKITVNGKTYEGR have been reported previously[15]. Briefly, the peptide was simulated for 7.4 μs in explicit water (TIP3P water model[36]) at a temperature of 300 K and constant volume (NVT ensemble). We used charged termini, protonated the arginine and lysine residues, and deprotonated the glutamic acid residue, and added 3 chlorine anions to obtain an uncharged simulation box.

B. Variational peptide dynamics

The ϕ and ψ torsion angle trajectories of each residue were extracted using the GROMACS command `g_rama`. The variational models were constructed using an in-house developed software package which is freely available at GitHub[41]. The residue-centered functions

$R^r_l(\phi, \psi)$ (eq. 29) for the AMBER ff99SB-ILDN[35] were constructed from simulations which have been reported in Ref. 26. For this we discretized the ϕ and the ψ-torsion angles into 36 bins and constructed a Markov state model on the resulting grid of 1296 microstates[27]. The residue-centered functions are also freely available at GitHub[41].

VI. SUMMARY AND DISCUSSION

A Markov model analysis (conventional or variational) yields two important pieces of information: long-lived conformations including the timescale of the transitions between them, and conformational degrees of freedom which mediate the transitions between the long-lived conformations. In conventional MSM analysis, one first identifies long-lived conformations via clustering in the dominant subspace, then interprets the eigenfunctions as transitions between them[10], and finally one identifies the conformational degrees of freedom which dominate these transitions. That is, these two analyses depend on each other, and any error in the dominant subspace analysis (e.g. choice of clustering parameters) will carry over to the identification of relevant conformational degrees of freedom. In variational Markov models, these two steps can be decoupled if the chosen basis functions represent a meaningful physical transition. Then, one can directly interpret the histograms of the expansion coefficients in terms of these physical transition without the intermediate step of clustering in the dominant eigenspace (see activity displays in figs. 2, 7.C, and 9.C). Independent of this, one can additionally identify long-lived conformations by clustering in the dominant eigenspace.

In deriving a physically meaningful basis set for peptide dynamics in ref. 27, we have made three main assumptions: The dominant eigenfunctions (*i*) are well approximated by functions of the flexible backbone torsion angles, and (*ii*) can be represented by tensors of residue-centered functions; (*iii*) the residue-centered functions can be approximated by the dominant eigenfunctions of the corresponding capped amino acid. Since the full basis set comprising all tensors of residue-centered basis functions would be far too large even for medium-sized peptides, we have added a fourth assumption: (*iv*) The dominant eigenfunctions of the transfer operator are well approximated by those tensors of residue-centered functions that model concerted conformational exchange at not more than two residues simultaneously. In other words, the dominant eigenfunctions are close to the subspace spanned by the doubly-active basis functions. Here, we improve and expand this generic peptide basis set in two important aspects: (v) we ensure orthonormality of the basis set, and (vi) we account for undersampled transitions.

Orthonormality. The built-in constraint of the VAC forces the approximated eigenfunctions to be orthonormal with respect to a scalar product which is weighted by the sampled stationary distribution $\hat{\mu}$. This results in the emergence of physically meaningless expansion coefficients that compensate the lack of orthonormality in the generic basis set with respect to this system-specific scalar product. The basis set can be straightforwardly orthonormalized with respect to the tensor approximation of the system-specific stationary density $\hat{\mu}^{\text{tensor}} = \hat{\mu}_1 \otimes \cdots \otimes \hat{\mu}_N$ by orthormalizing the residue-centered basis functions (R_0^i, R_1^i, R_2^i) with respect to $\hat{\mu}_i$. This strategy mildly alters the expansion coefficient pattern, mainly by reducing the contribution of the constant basis function χ_0 to the estimates of the higher eigenfunctions r_i for $i > 0$. Since this observation concurs with our physical intuition, these results may be considered a slight improvement in terms of interpretability of the expansion coefficient pattern. Notably, both the implied timescales and the identified metastable conformations are not affected by the orthonormalization, therefore we conclude that the estimates of the eigenfunctions within the ansatz space remain virtually unaffected by this transformation of the basis set, and clustering in the dominant eigenspace is possible both for the generic and the orthonormalized basis set.

Accounting for undersampled transitions. We can only ensure orthonormality with respect to $\hat{\mu}^{\text{tensor}}$. Any remaining deviation from orthonormality directly depends on the approximation quality of $\hat{\mu}$ by the tensor measure. By analyzing the overlap matrix, we found that, in all considered systems, a small share of the basis functions exhibit a sizeable deviation from orthonormality with respect to $\hat{\mu}$. These functions are doubly active basis functions that model concerted probability exchange at two residues in the peptide sequence. Therefore these basis functions signal that the underlying joint distribution of the affected residues must deviate from the tensor product of the individual marginal distributions. Such a deviation can be explained either by an actual mutual dependence of the two residues, or by insufficient sampling of the joint distribution by the MD simulation data. If two residues are mutually dependent, the tensor measure that correspond to the affected pair of residues should be replaced by the joint measure of both residues, and tensors of the two residue-centered basis functions should be replaced by joint functions. However, we found that in our examples the observed deviations from the tensor product structure are most likely explained by insufficient sampling, as the relevant conditional distributions have been generated by only a very small share of the frames in the simulation data. We chose to remove the affected basis functions from the basis set. This approach is analogous to removing disconnected microstates from the set of microstates in MSM analysis. We find that by reducing the basis set in this manner, the implied timescales and the identified metastable conformations are not affected. However, the histogram of the expansion coefficients changes.

We tested whether secondary structure formation can be resolved with our basis set, by applying the reduced orthonormal basis sets on the β-hairpin peptide as well as on a fragment of the human islet amyloid polypeptide (hIAPP), which forms a variety of structures including an α-helix. Regarding the β-hairpin peptide, we were able to reproduce the large and medium-sized metastable conformations that have previously been identified using the core set method[15] with remarkable accuracy. The variational model identifies one additional metastable state with low population that could not be detected by the core set method. Since the conformational dynamics of the β-hairpin peptide primarily comprises several slightly different β-sheet structures, we conclude that the VAC in combination with the peptide basis set is well suited for the detection of these secondary structures. An obvious question is whether the idea of inferring structural information from the expansion coefficients of the variational models can be realized in the above examples. In the case of the β-hairpin peptide, this is partially possible, in the sense that the patterns of dynamically active residues in the estimates of the eigenfunctions can be linked to the loci of the structural transitions between the corresponding metastable conformations. However, this connection is too unspecific to manually predict structural information solely from the expansion coefficients.

For the hIAPP fragment, we also obtained a very detailed dynamic model using reduced orthonormal basis sets basis set. The dynamical network of metastable conformations consists of a central dominant α-helical structure which transitions into several conformations with lower population. These results show that our basis set can be used to detect and model the formation of α-helices. However, the hIAPP is much more flexible than the β-hairpin peptide and the residues seem to move in a more concerted fashion. Thus, interpreting the histogram of expansion coefficients was not possible for hIAPP.

In summary, the reduced orthonormal peptide basis set is well suited to model the notoriously complex dynamics of unfolded peptides. We believe that it can be used to model the dynamics of long intrisically disordered peptides and in particular the emergence of secondary structure elements in amyloid formation or during folding upon binding of the intrinsically disordered peptides to protein. Additionally, two possible ways to extend our method arise from the current study. The analysis of the overlap matrix identifies undersampled transitions. Our current strategy is to remove the corresponding basis functions from the data set. Alternatively, one could use this as starting point for an adaptive sampling strategy in which the thus identified transitions are resampled. The second extension concerns the identification of collective variables. A number of methods have been proposed recently in which collective variables are proposed based on an analysis of the dominant eigenspace of the system[42–44]. Our activity displays identify those torsion angles which are affected by a given dominant eigenfunction and can thus be regarded as a set of collective variables. We are curious to see whether the torsional collective variables identified by our basis set can be used to enhance the sampling and whether they can be combined with path reweighting methods[45,46].

ACKNOWLEDGMENTS

This research has been funded by Deutsche Forschungs-gemeinschaft (DFG) through Grant No. CRC 1114 *Scaling Cascades in Complex Systems*, B05 *Origin of the scaling cascades in protein dynamics*. The computer facilities of the Freie Universitt Berlin (ZEDAT) are acknowledged for computer time.

REFERENCES

[1] van Gunsteren, W. F. et al. *Angew. Chem., Int. Ed.* **2006**, *45*, 4064–4092.

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

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

[4] Schütte, C.; Fischer, A.; Huisinga, W.; Deuflhard, P. *J. Comput. Phys.* **1999**, *151*, 146–168.

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

[6] Singhal, N.; Snow, C. D.; Pande, V. S. *J. Chem. Phys.* **2004**, *121*, 415.

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

[8] Buchete, N.-V.; Hummer, G. *J. Phys. Chem. B* **2008**, *112*, 6057–6069.

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

[10] 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.

[11] Schütte, C.; Sarich, M. *Metastability and Markov state models in molecular dynamics: modeling, analysis, algorithmic approaches*; American Mathematical Soc., 2013; Vol. 24.

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

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

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

[15] Lemke, O.; Keller, B. G. *J. Chem. Phys.* **2016**, *145*, 164104.

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

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

[18] Weber, M.; Fackeldey, K.; Schütte, C. *J. Chem. Phys.* **2017**, *146*, 124133.

[19] Nedialkova, L. V.; Amat, M. A.; Kevrekidis, I. G.; Hummer, G. *J. Chem. Phys.* **2014**, *141*, 114102.

[20] Boninsegna, L.; Gobbo, G.; Noé, F.; Clementi, C. *J. Chem. Theory Comput.* **2015**, *11*, 5947–5960.

[21] Orioli, S.; Faccioli, P. *J. Chem. Phys.* **2016**, *145*, 124120.

[22] Nüske, F.; Schneider, R.; Vitalini, F.; Noé, F. *J. Chem. Phys.* **2016**, *144*, 054105.

[23] Martini, L.; Kells, A.; Covino, R.; Hummer, G.; Buchete, N.-V.; Rosta, E. *Phys. Rev. X* **2017**, *7*, 031060.

[24] Husic, B. E.; Pande, V. S. *J. Am. Chem. Soc.* **2018**, *140*, 2386–2396.

[25] Klus, S.; Nüske, F.; Koltai, P.; Wu, H.; Kevrekidis, I.; Schütte, C.; Noé, F. *J. Nonlinear Sci.* **2018**, *28*, 985–1010.

[26] Vitalini, F.; Noé, F.; Keller, B. G. *Data in Brief* **2016**, *7*, 582–590.

[27] Vitalini, F.; Noé, F.; Keller, B. G. *J. Chem. Theory Comput.* **2015**, *11*, 3992–4004.

[28] Schütte, C.; Huisinga, W.; Deuflhard, P. *Ergodic theory, analysis, and efficient simulation of dynamical systems*; Springer, 2001; pp 191–223.

[29] Schütte, C.; Jahnke, T. *ESAIM: Math. Modell. Numer. Anal.* **2009**, *43*, 721–742.

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

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

[32] Weber, M. *Classification and Data Mining*; Springer, 2013; pp 147–154.

[33] Kabsch, W.; Sander, C. *Biopolymers* **1983**, *22*, 2577–2637.

[34] Nanga, R. P. R.; Brender, J. R.; Vivekanandan, S.; Ramamoorthy, A. *Biochim. Biophys. Acta* **2011**, *1808*, 2337–2342.

[35] Lindorff-Larsen, K.; Piana, S.; Palmo, K.; Maragakis, P.; Klepeis, J. L.; Dror, R. O.; Shaw, D. E. *Proteins: Struct., Funct., Bioinf.* **2010**, *78*, 1950–1958.

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

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

[38] Bussi, G.; Donadio, D.; Parrinello, M. *J. Chem. Phys.* **2007**, *126*, 014101.

[39] Hess, B.; Bekker, H.; Berendsen, H. J. C.; Fraaije, J. G. E. M. *J. Comput. Chem.* **1997**, *18*, 1463–1472.

[40] Darden, T.; York, D.; Pedersen, L. *J. Chem. Phys.* **1993**, *98*, 10089–10092.

[41] Variational peptide dynamics software. `https://github.com/BDGSoftware/VariationalProteinDynamics`, Accessed: 2017-04-20.

[42] McGibbon, R. T.; Husic, B. E.; Pande, V. S. *J. Chem. Phys.* **2017**, *146*, 044109.

[43] Yang, Y. I.; Parrinello, M. *J. Chem. Theory Comput.* **2018**, *14*, 2889–2894.

[44] Bittracher, A.; Koltai, P.; Klus, S.; Banisch, R.; Dellnitz, M.; Schütte, C. *J. Nonlinear Sci.* **2018**, *28*, 471–512.

[45] Donati, L.; Hartmann, C.; Keller, B. G. *J. Chem. Phys.* **2017**, *146*, 244112.

[46] Donati, L.; Keller, B. G. *J. Chem. Phys.* **2018**, *149*, 072335.

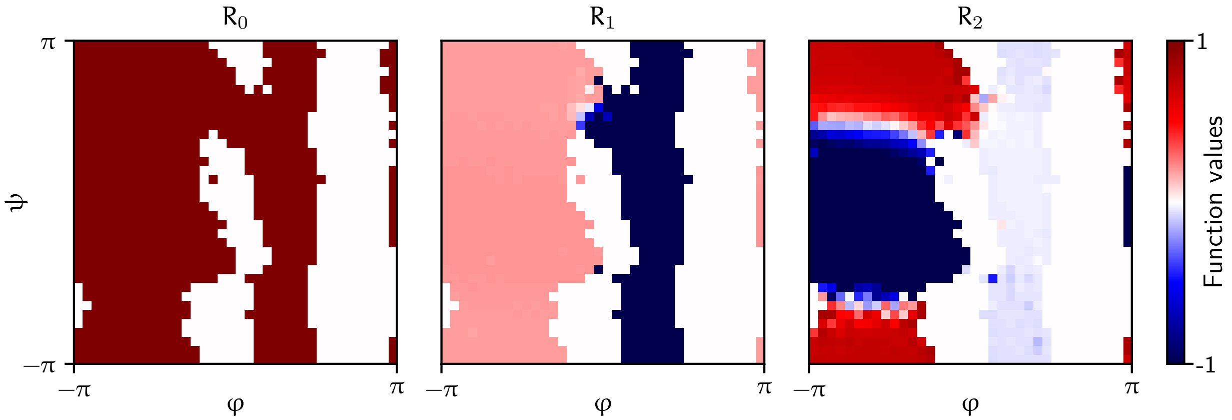

FIG. 1. Residue-centered functions of alanine, obtained from a Markov state model of the terminally capped amino acid based on ad discretization of the Ramachandran space by a regular 36 × 36-grid.

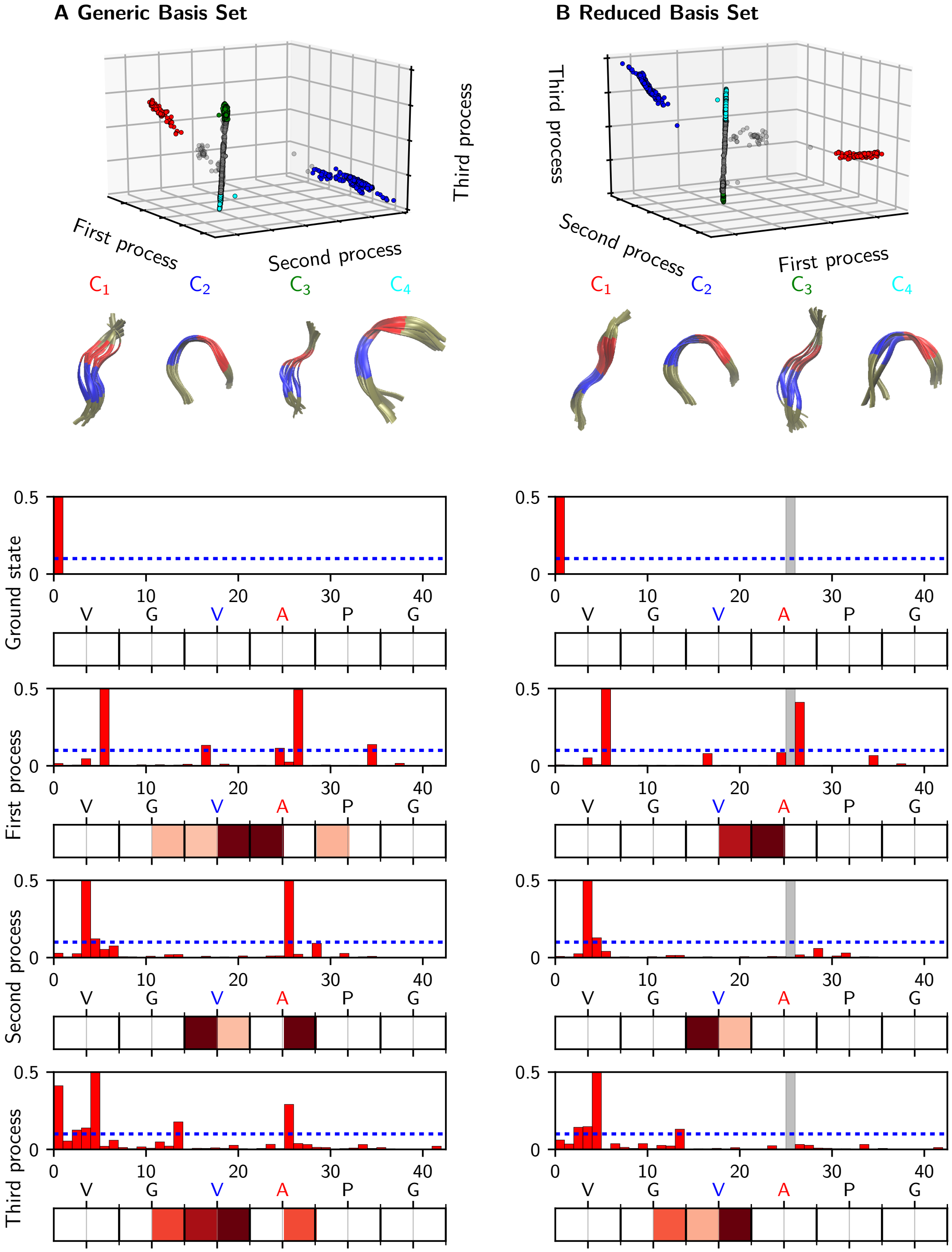

FIG. 2. Variational model of VGVAPG with generic and reduced peptide basis set. Top: Projection of the trajectory in the dominant subspace and representative molecular structures of each of the identified clusters. Bottom: Expansion coefficients of the four dominant eigenfunctions. Below each histogram the residues are highlighted at which the probability exchange associated with the large expansion coefficients is located.

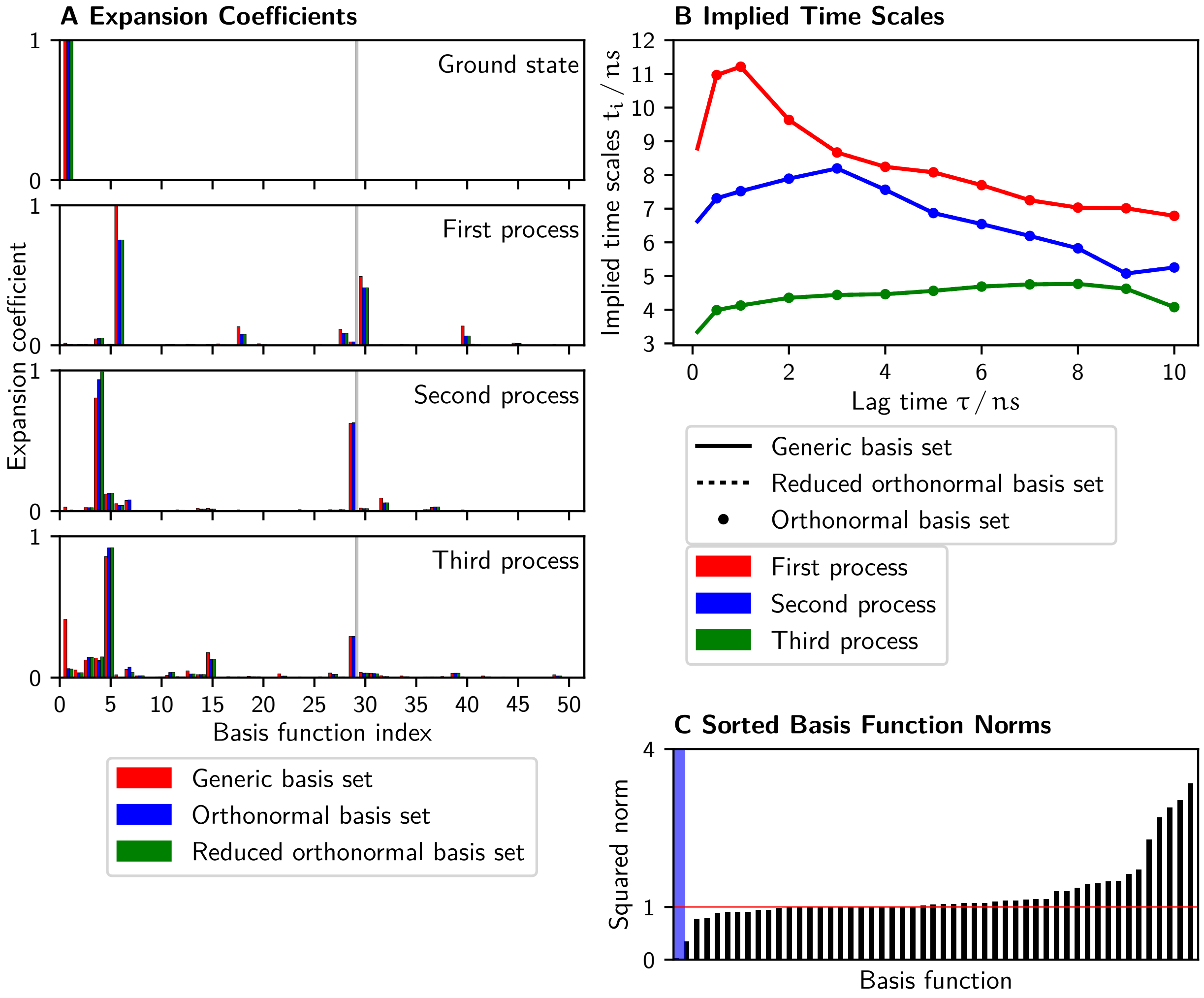

FIG. 3. **A**: Absolute values of the expansion coefficients of the first four eigenvectors estimated for the VGVAPG peptide at 2 ns lag time using the generic basis set (red), the full orthonormal basis set (blue) and the reduced orthonormal basis set (green), all of which comprise doubly active basis functions. The gray bar highlights the position of a basis function that is excluded from the reduced basis set. **B**: Implied time scales associated with the first three processes estimated for the VGVAPG peptide. **C**: Sorted squared norm values of the orthonormal basis set for the VGVAPG peptide. The blue shaded area indicates the basis function with squared norm value less than 0.1.

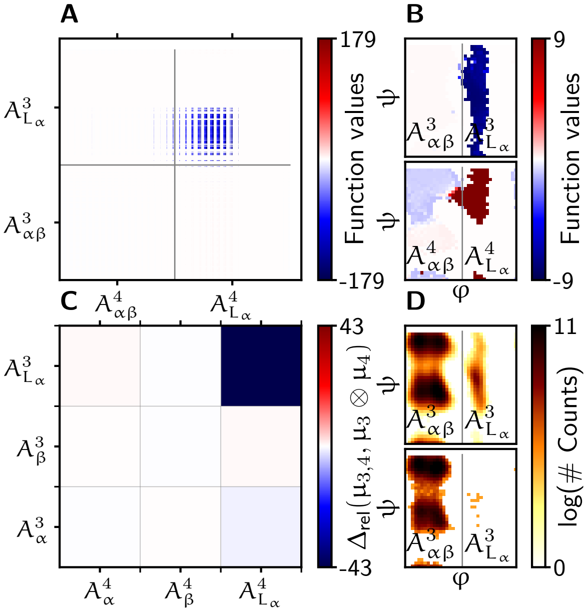

FIG. 4. **A**, **B**: Orthonormalized residue-centered functions mapping probability exchange along the φ-axis of residues 3 (valine) and 4 (alanine) of VGVAPG and the doubly active basis function that comprises these local functions projected onto their joint state space $\Omega_3 \times \Omega_4$ (**B**). $A^r_{L_\alpha} \subset \Omega_r$ and $A^r_{\alpha\beta} \subset \Omega_r$ refer to the areas of Ramachandran space of residue r containing the L_α minimum and the α- and β-minima, respectively. **C**: Relative difference $\Delta_{\mathrm{rel}}(\mu_{3,4}, \mu_3 \otimes \mu_4)$ of probabilities measured by the joint distribution $\mu_{3,4}$ and the tensor distribution $\mu_3 \otimes \mu_4$. **D**: Illustrations of the stationary marginal probability measure μ_3 on Ω_3 (top) as well as the stationary conditional probability measure $\mu_3^{L_\alpha}$ that quantifies the probabilities of torsion angles on Ω_3 given that residue 4 populates the L_α minimum (bottom).

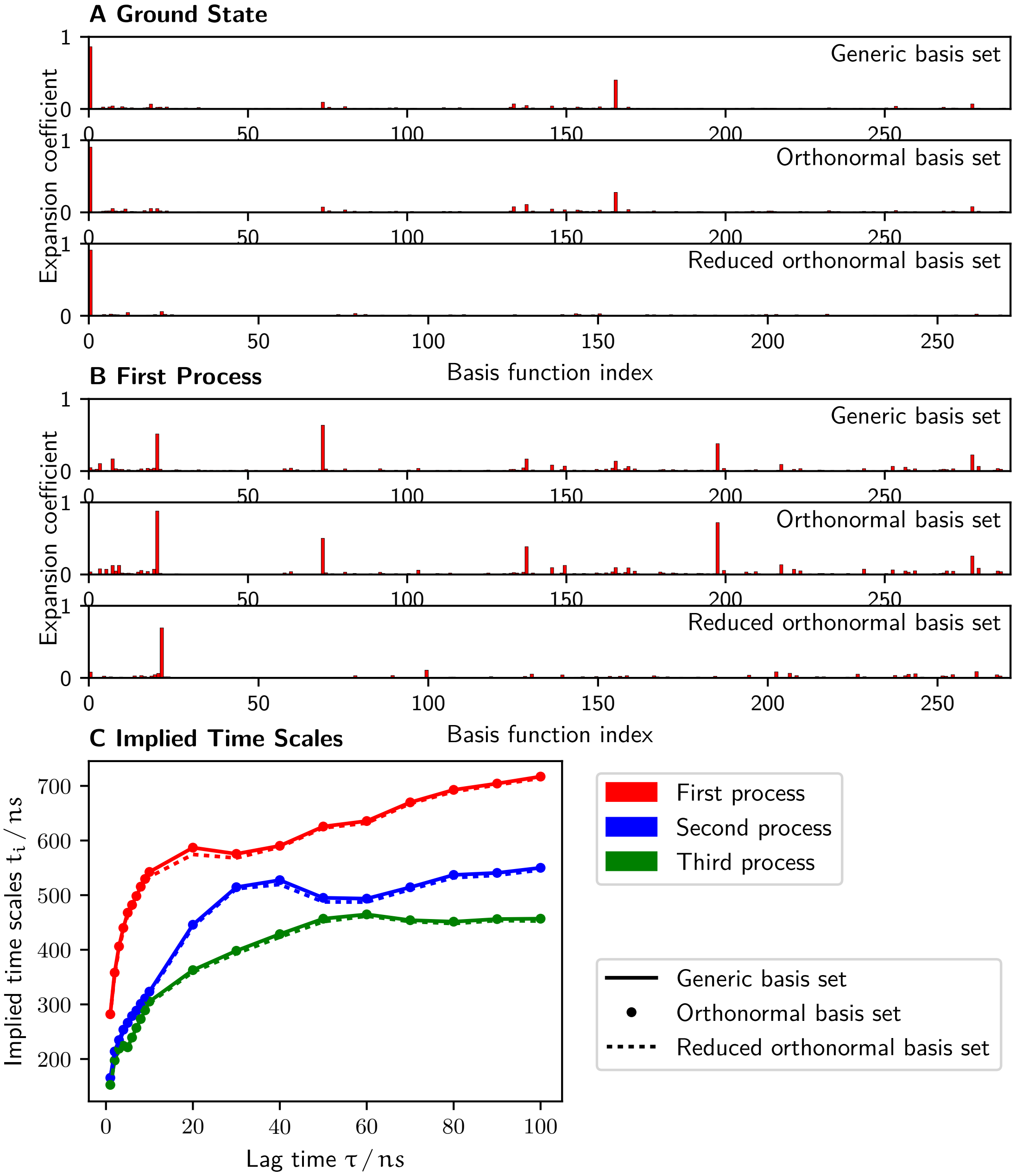

FIG. 5. **A,B**: Absolute values of the expansion coefficients corresponding to each basis function in the estimate of the first (**A**) and second (**B**) eigenfunctions of the underlying transfer operator that associates with the dynamics of the β-hairpin peptide, estimated at a lag time of 30 ns. **C**: Implied time scale plots of of the second, third and fourth eigenfunctions computed for the three different basis sets.

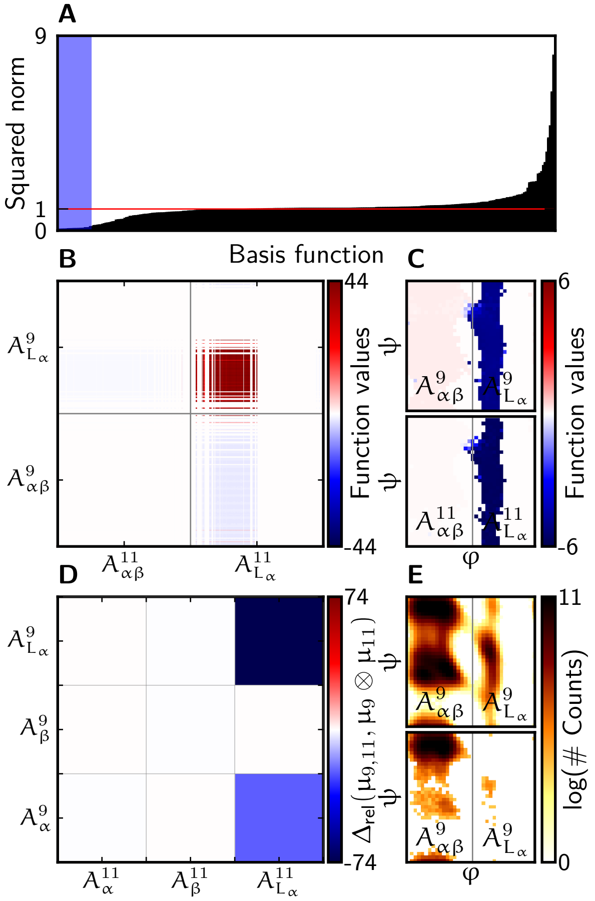

FIG. 6. **A**: Sorted squared norm values of the orthonormalized doubly active basis set generated for simulations of the β-hairpin peptide. The blue shaded area indicates basis functions with squared norm values less than 0.1. **B**, **C**: Orthonormalized local functions mapping probability exchange along the φ-axis of residues 9 (lysin) and 11 (tyrosine) of the β-hairpin peptide (**C**) and the doubly active basis function that comprises these local functions projected onto their joint state space $\Omega_9 \times \Omega_{11}$ (**B**). $A^r_{L_\alpha} \subset \Omega_r$ and $A^r_{\alpha\beta} \subset \Omega_r$ refer to the areas of Ramachandran space of residue r containing the L_α minimum and the α- and β-minima, respectively. **D**: Relative difference $\Delta_{\mathrm{rel}}(\mu_{9,11}, \mu_9 \otimes \mu_{11})$ of probabilities measured by the joint distribution $\mu_{3,4}$ and the tensor distribution $\mu_9 \otimes \mu_{11}$, respectively. **E**: Illustrations of the stationary marginal probability measure μ_9 on Ω_9 (top) as well as the stationary conditional probability measure $\mu_9^{L_\alpha}$ that quantifies the probabilities of torsion angles on Ω_9 given that residue 11 populates the L_α minimum (bottom).

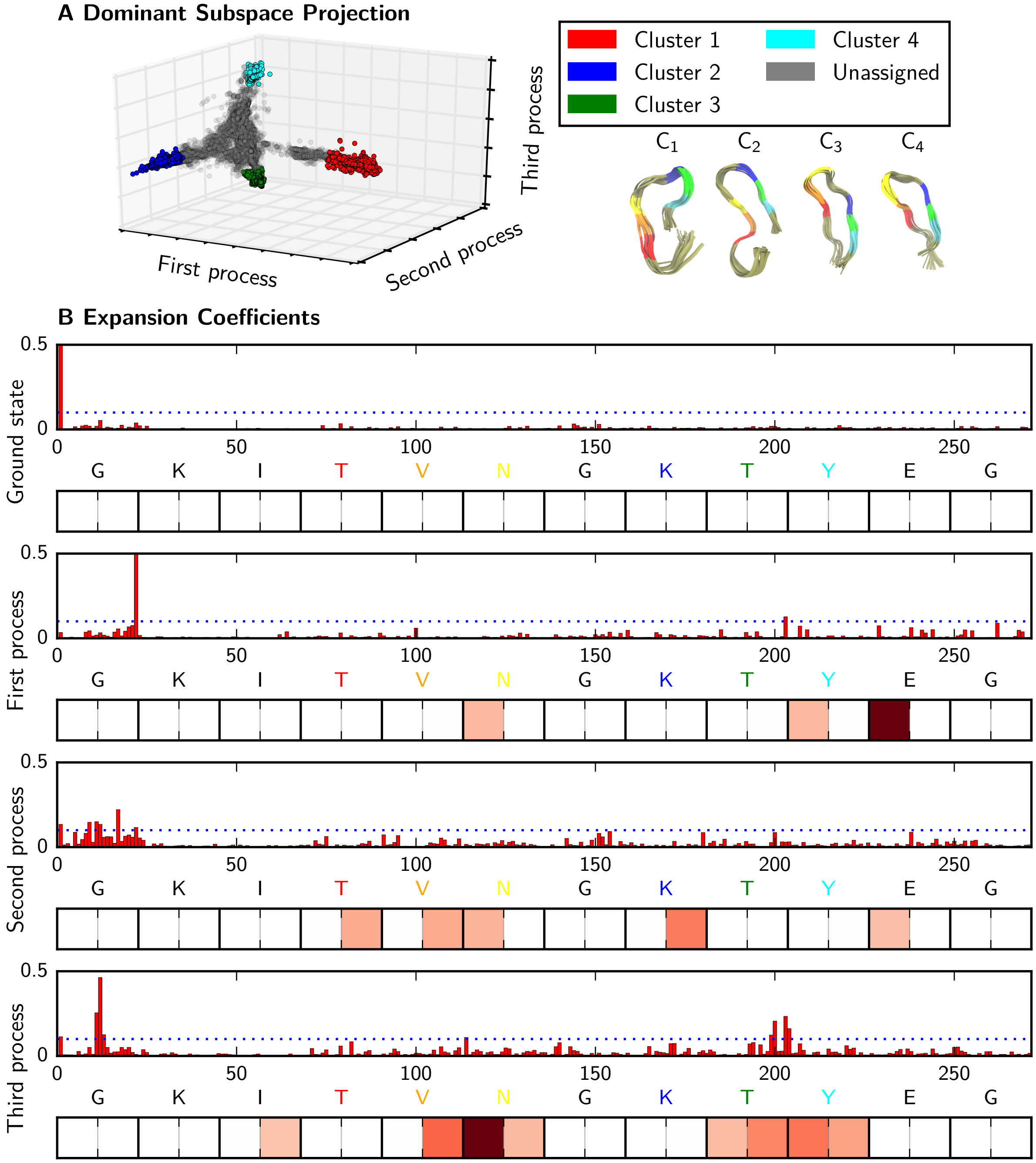

FIG. 7. **A**: Projection of the β-hairpin trajectory in the dominant subspace and representative molecular structures of each of the identified clusters. **B**: Expansion coefficients of the estimates of the three processes as well as of the ground state. Below the residues are highlighted at which the probability exchange associated with the large expansion coefficients is located.

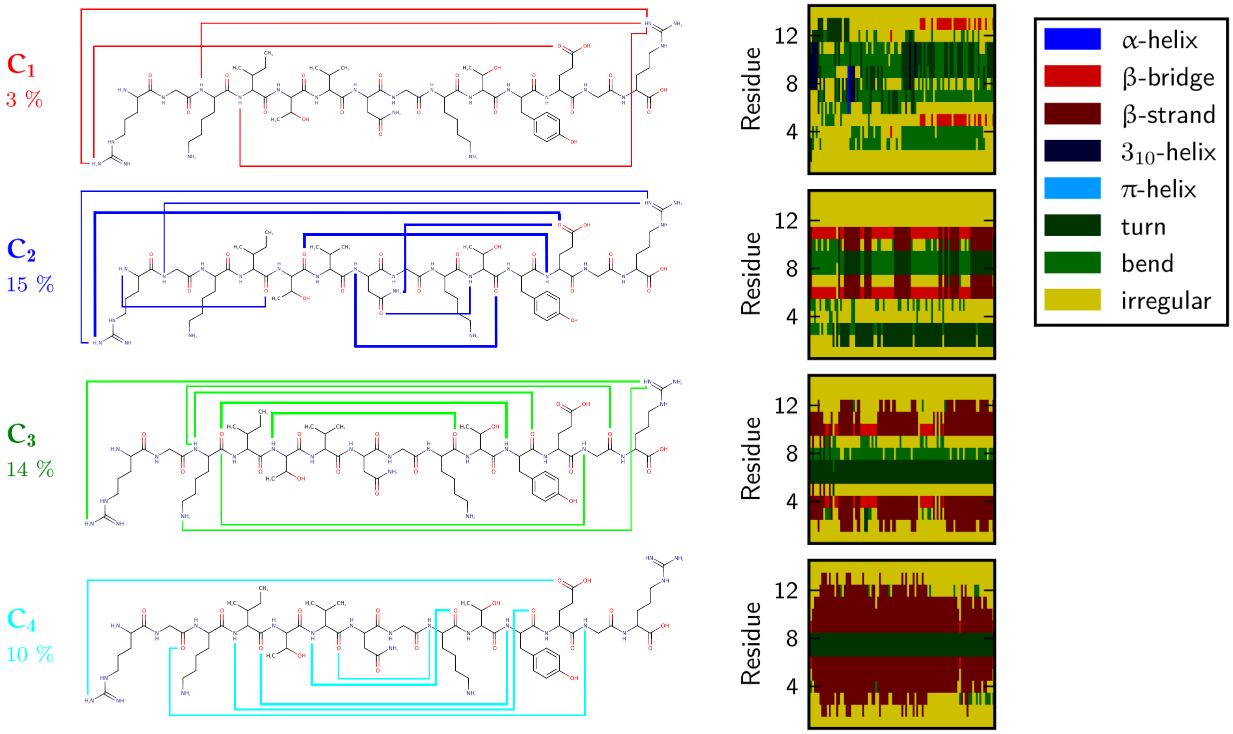

FIG. 8. Structural characterization of the identified clusters of the β-hairpin peptide in the dominant subspace. The left hand side shows the location of the most stable hydrogen bonds within each cluster. On the right hand side a characterizataion of the secondary structures via DSSP assignment is shown.

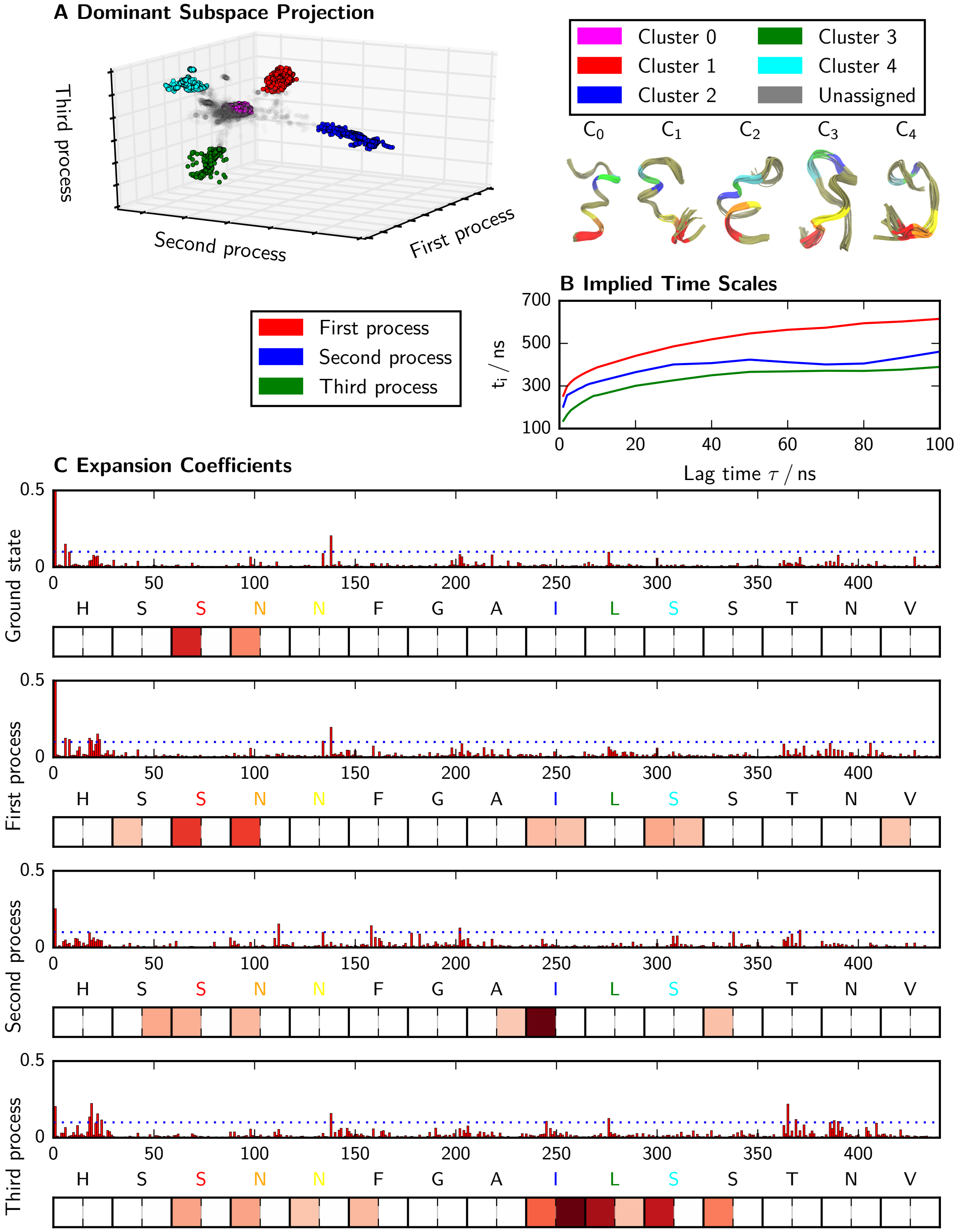

FIG. 9. **A**: Projection of the hIAPP fragment trajectory in the dominant subspace and representative molecular structures of each of the identified clusters. **B**: Implied time scales associated to the dominant three processes. **C**: Expansion coefficients of the estimates of the three processes as well as of the ground state. Below the residues are highlighted at which the probability exchange associated with the large expansion coefficients is located.

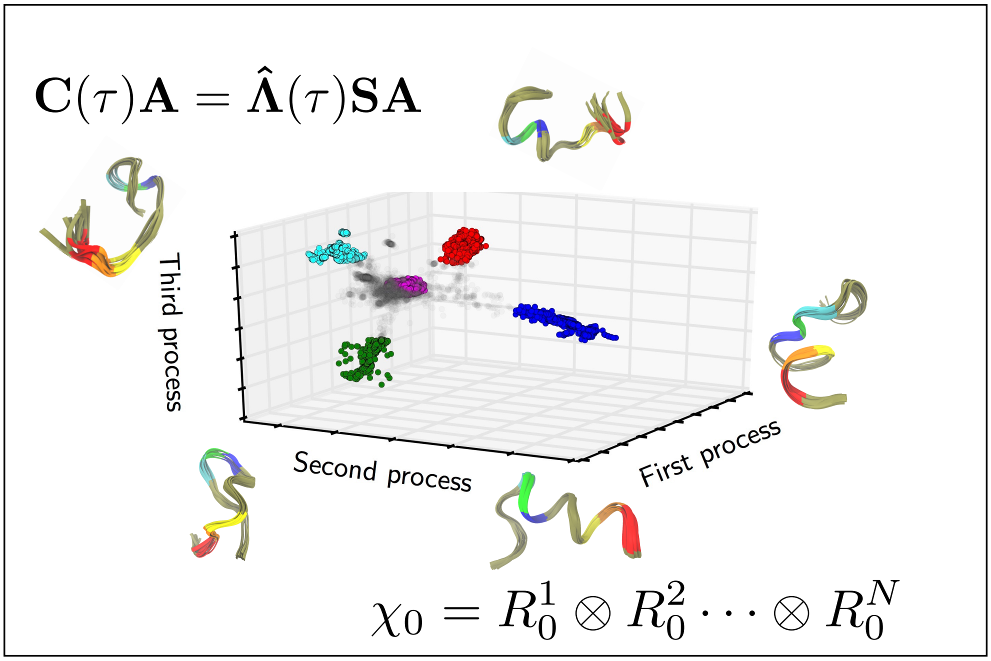
$\mathbf{C}(\tau)\mathbf{A} = \hat{\mathbf{\Lambda}}(\tau)\mathbf{S}\mathbf{A}$
Third process
Second process
First process
$\chi_0 = R_0^1 \otimes R_0^2 \cdots \otimes R_0^N$

FIG. 10. Table of content figure