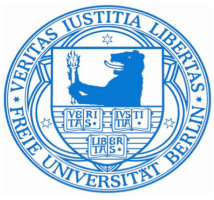

Freie Universität Berlin

# Hierarchical approaches to kinetic models of peptides

Inaugural-Dissertation

to obtain the academic degree

*Doctor rerum naturalium* (Dr. rer. nat.)

submitted to the department of Biology, Chemistry and Pharmacy

of Freie Universität Berlin


by

Vitalini Francesca


First Supervisor:

Prof. Dr. Bettina Keller

Second Supervisor:

Prof. Dr. Frank Noé

Year of submission 2015

**First Reviewer:**

Prof. Dr. Bettina Keller

Department of Biology, Chemistry and Pharmacy,

Freie Universität Berlin

**Second Reviewer:**

Prof. Dr. Frank Noé

Department of Mathematics and Informatics,

Freie Universität Berlin

Date of defence:

$8^{th}$ April 2016

## List of publications:

- **F. Vitalini**, A. S. J. S. Mey, F. Noé, and B. G. Keller, *Dynamic properties of force fields*, Journal of Chemical Physics 142, 084101, doi:10.1063/1.4909549, February 2015.
- **F. Vitalini**, F. Noé, and B. G. Keller *A Basis set for peptides for the variational approach to conformational kinetics*, Journal of Chemical Theory and Computation, 11 3992-4004. doi:10.1021/acs.jctc.5b00498, August 2015.
- **F. Vitalini** and B. G. Keller (2015) *Hierarchy in the conformational ensemble of human islet amyloid polypeptide.* Submitted to Journal of Molecular Biology.
- **F. Vitalini**, F. Noé, and B. G. Keller (2015) *Molecular Dynamics simulations of twenty encoded amino acids in different force fields.* Submitted to Data in Brief.
- F. Nüske, R. Schneider, **F. Vitalini**, and F. Noé, (2015) *Variational tensor approach for approximating the rare-event kinetics of macromolecular systems.* Submitted to Journal of Chemical Physics. (Not included in this thesis)

**Contributions to the papers:**

- *Dynamic properties of force fields*

  I provided the capped amino acids and the AVAVA peptide simulations and analysis. I was involved in the plotting of panel *a* and half of panels *b* and *c* of figure one. I produced the data for figures two and three and collaborated in the making of the figures. I made the entirety of figure four and collaborated in the making of figure five. I produced figure six panel *a* and collaborated in the production of panel *b*. I also participated in the writing of all sections of the main manuscript and of the supporting information. Additionally I made, with collaborations of the other contributors, all the figures in the supporting information.

- *A basis set for peptides for the variational approach to conformational kinetics.*

  I produced the code for implementation of the basis set and the analysis. In such context, I developed and tested the residue-centered basis functions, stored them in form of library and made them publicly available on a `git` repository. I set up and carried out the simulations of all the systems presented in the paper, as well as performing the whole analysis. Additionally I took part in the writing of the entire paper and in the generation of all the figures, with collaboration in the making of figures one, two, five and six. I wrote the supporting information and produced all the plots there included, with exception of figure three where I was co-assisted.

- *Hierarchy in the conformational ensemble of human islet amyloid polypeptide.*

  I carried out the simulations of all the peptides and constructed the Markov state model of all systems. I also performed the general analysis of all systems. I participated to the writing of the full paper and wrote the supporting information. I produced all the figures in both the main manuscript and the supporting information.

- *Molecular Dynamics simulations of twenty encoded amino acids in different force fields.*

  I carried out the simulations of all the twenty encoded amino acids and force fields combinations. I prepared all the figures and tables presented in the manuscript, as well as writing the manuscript.

- *Variational tensor approach for approximating the rare-event kinetics of macromolecular systems* (not included in this thesis)

  I simulated the Ac-VA-NHMe dimer. I performed the standard Markov model analysis of the system and wrote the corresponding part in the method section.

To the neurons heroically fallen in the process.

## Abstract

In this thesis, we address different approaches for the construction of kinetic models of peptides in the framework of Markov State Models. Using human amylin polypeptide as a test case, we construct kinetic models of peptide's fragments of increasing length and uncover an underlying hierarchy of the dynamics. The slow kinetic modes of groups of highly collaborative residues are combined together to build a model of the system.
Markov state models are, however, sensitively dependent on the discretization of the configuration space. A newly introduced variational approach to conformation dynamics permits to overcome a crisp-state discretization and systematically control the quality of the model. Here, a basis set for peptides kinetics for the variational approach is developed and tested. The basis functions are constructed by combining local residue-centered kinetic modes, obtained from pre-parametrized kinetic models of terminally blocked amino acids.
However, the quality of the approach depends on how well the basis functions capture the features of the underlying energy landscape. Thus, the effect of Molecular Dynamics force fields in capturing kinetic properties, is called into question. By comparing the dynamic properties of blocked amino acids and short peptides, a strong force field dependance is identified. Therefore a library of force field dependent residue-centered basis functions is developed and made available for further applications of the method.

## Zusammenfassung

In dieser Arbeit behandeln wir verschiedene Ansätze für den Aufbau kinetischer Modelle von Peptiden im Rahmen von Markov State Models. Für den Fall des Polypeptids Amylin, entwickeln wir kinetische Modelle für unterschiedlich lange Fragmente und finden so eine grundlegende Hierarchie innerhalb der Dynamiken. Durch Kombination der langsamen kinetischen Moden von Gruppen hoch kooperativer Aminosäuren wird ein Model für das ganze System erzeugt.
Markov State Models weisen eine hohe Empfindlichkeit bezüglich der Diskretisierung des Konfigurationsraums auf. Ein neu eingeführter variationeller Ansatz für die Konformationsdynamik erlaubt es, die diskrete Zustandsdefinition zu umgehen und die Qualität des Modells systematisch zu steuern. Hier entwickeln und testen wir einen Basissatz zur Beschreibung von Peptidkinetiken mittels des variationellen Ansatzes. Die Basisfunktionen werden dabei durch Kombination von lokalen Aminosäure-zentrierten kinetischen Moden konstruiert, die zuvor durch vor-parametrisierte kinetische Modelle von terminal blockierten Aminosäuren bestimmt wurden.
Die Qualität dieses Ansatzes hängt jedoch stark davon ab, wie gut die zugrunde liegende Energielandschaft durch die Basisfunktionen beschrieben wird. Daher gehen wir der Frage nach, in wie Fern kinetische Eigenschaften von Kraftfeldern in Molekül-dynamik Simulationen beeinflusst werden. Durch den Vergleich der dynamischen Eigenschaften von blockierten Aminosäuren und kurzen Peptiden, können wir eine starke Kraftfeldabhängigkeit aufzeigen. Deshalb wird eine Bibliothek von kraftfeld-abhäängigen Aminosäure-zentrierten Basisfunktionen angelegt und für zukünftige Anwendungen der Methode zugänglich gemacht.

# Contents

Chapter 1

# Introduction

Proteins and peptides are biomolecules, which play a central role in living organisms. They participate in virtually every process in a cell and cover a vast array of functions [1]. For example, they are involved in catalyzing metabolic reactions [2], replicating DNA [3], responding to stimuli [4] and transporting molecules [5]. The building block of proteins are the amino acids (fig 1.1). They contain an amine group (-$NH_2$), a carboxylic acid (-COOH), and differentiate for the specific side-chain (R). When connected together, two amino acids form a covalent bond, known as the peptide bond [6] (fig 1.1). Amino acids connected by a peptide bond and part of a protein/peptide are also referred to as residues.

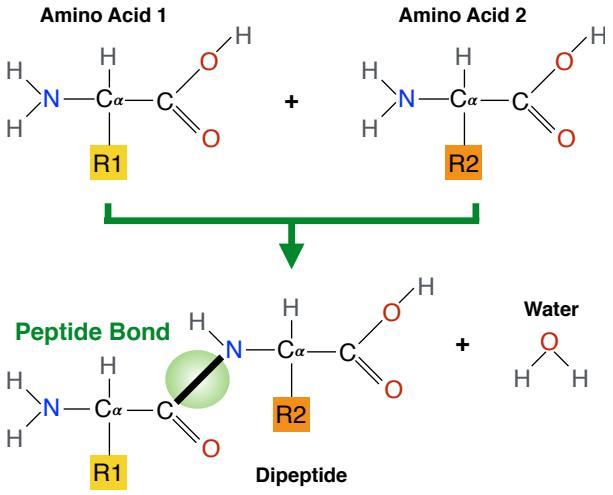


**Figure 1.1**: Schematic representation of the peptide bond between two generic amino acids.

The peptide bond is a covalent partial double-bond, due to the resonance contribution of the nitrogen (N) and the oxygen (O). Therefore the rotation around the bond is restricted. The planar character of the peptide bond imposes some conformational restrictions to the residues. Independently of the side-chain, the possible conformations of a bounded amino acid are well captured by the torsion angles $\phi$ (rotation around the N-$C_\alpha$ bond) and $\psi$ (rotation around the $C_\alpha$-C bond) (fig. 1.2.a), known as backbone dihedral angles. The different combinations of $\{\phi, \psi\}$, which are determined by steric restrictions, can be represented as a two-dimensional plot, known in the literature as the Ramachandran plane [7] (fig. 1.2.b).

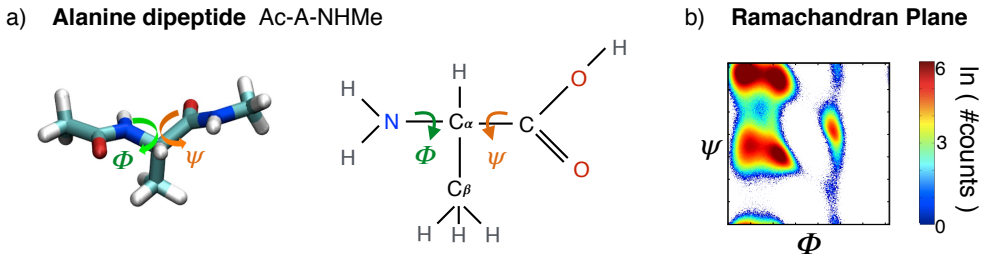


**Figure 1.2**: a: Definition of the torsion angles of an amino acid (example Ac-A-NHMe). b: Ramachandran plot of Ac-A-NHMe
obtained by a 4 $\mu$s MD simulation.

The most populated regions of the Ramachandran plane correspond to those regular motifs that constitute the second level of organization of a sequence of amino acids. These motifs, known in the literature as *secondary structures*, are repetitive arrangements of segments of the sequence, stabilized by hydrogen bonds (electrostatic interaction) between the amino (-NH) and carbonyl groups (C=O) of the main chain. There are two main motifs (fig. 1.3):

- $\alpha$-*helices*, which are right-handed helices consisting of 3.6 amino acids per turn, formed by hydrogen bonds between the carboxy group of the $i^{th}$ amino acid and the amino group of the $i^{th} - 4$ amino acid.
- $\beta$-*sheets*, which correspond to two (or more) segments linked by hydrogen bonds between complementary groups of the two chains.

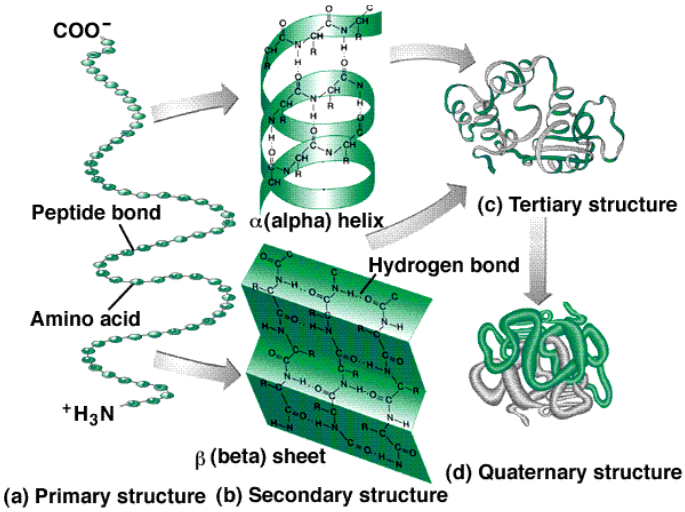


**Figure 1.3**: Representation of the structural levels of a protein: residues chain (*primary structure*), $\alpha$-helix and $\beta$-sheets (*secondary structure*), three-dimensional conformation (*tertiary and quaternary structures*). Image adapted from *Estelle Levetin and Karen McMahon, Botany Visual Resource Library © 1996 The McGraw-Hill Companies, Inc. All rights reserved.*

The particular rearrangements of the *secondary structure* motifs defines the *tertiary structures* of the protein. Some proteins are made of multiple three-dimensional subunits. The arrangement of these subunits in the final functional configuration is called *quaternary structure* of the protein.

The encoded amino acids that form proteins are of twenty different types, i.e. present twenty different functional groups (R). The specific sequence of amino acids defines the protein univocally, and determines its three-dimensional structure(s). The three-dimensional structure of a protein is associated with its function. A typical example is the lock-and-key model for enzyme-ligand binding. Enzymes are

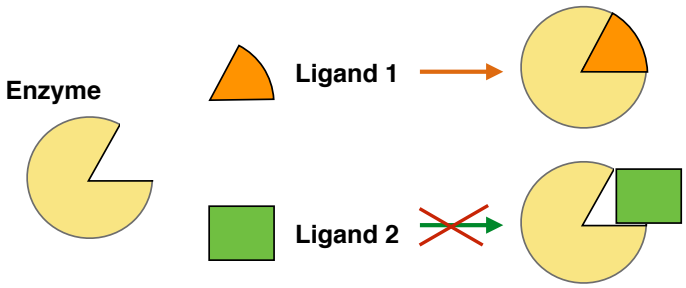


**Figure 1.4**: Schematic representation of the lock and key model for enzymes' specificity.

proteins which act as catalysts to accelerate a chemical reaction: the ligand binds to the enzyme and is transformed into product(s). The binding is usually very specific and can be pictured as complementary geometric shapes fitting exactly into one another (fig 1.4). If the enzyme's three-dimensional structure is altered, it can impede the enzyme from performing its function. Proteins are extremely sensitive to changes at atomistic level: even a single point mutation can have dramatic effects on the folded structure and even cause diseases [8]. Understanding how proteins find their three-dimensional functional conformation is thus of fundamental relevance.

A protein is a multiple-particle system. It can contain easily more than 500 amino acids (e.g. human serum albumin, the most abundant protein in human blood plasma has 585 [9]) and each amino acid consists, on average, of 20 atoms (the smallest is glycine with 10 and the largest is tryptophan with 27). The number of possible rearrangements of so many particles, i.e. the conformations of the protein, is therefore very large. Such number is reduced by the conformational restrictions imposed by the bonds that connect the atoms and the angle constraints, but remains too large so that finding the native state could be a random event. Were it the case, protein folding would take longer than the age of the universe [10] (*Levinthal paradox*), whereas it takes only micro-seconds to seconds *in vivo* and test-tubes [11], proving that folding is driven by the thermodynamics of the system.

The canonical view of protein folding is the folding funnel approach [12]: given a sequence of amino acids, there is a small number of conformations, which constitute

the native folded state of the protein; such conformations are much more energetically favorable than any other, i.e. the energy landscape has the shape of a deep funnel (fig. 1.5.a). Such view, however, cannot explain proteins which have more

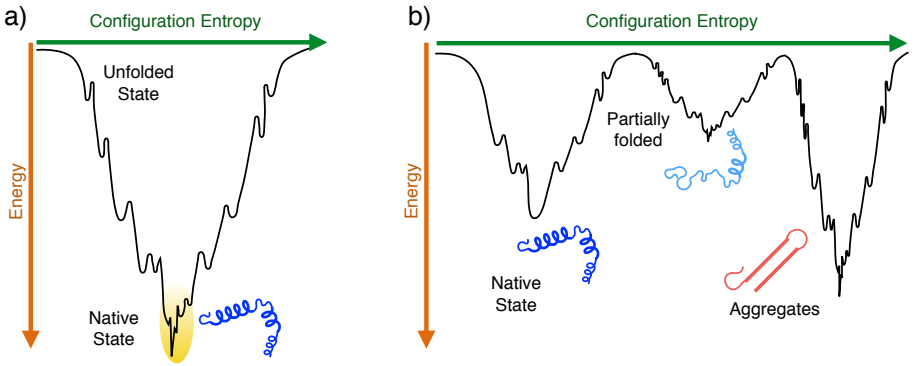


**Figure 1.5**: a) Schematic representation of the energy landscape in the folding funnel approach. b) Schematic representation of the multiple long-living states energy landscape.

than one long-living configuration (fig. 1.5.b). For instance, intrinsically disordered proteins (IDPs) [13, 14] can assume a variety of three-dimensional structures and interconvert quickly between them. Some of these structures perform functions in the cell, whereas others can cause diseases, such as amyloid formation in Alzheimer's disease [15] (amyloid-$\beta$) or in diabetes type two [16] (amylin islet polypeptide).
It is thus of essential relevance to understand not only which are the long-living states of the system, but also how the system interconverts between them. Over the last thirty years, more and more evidence has shown the importance of protein dynamics in relation to proteins' function [17, 18, 19, 20, 21]. Therefore, a complete understanding of proteins requires atomistically detailed models that capture both structural heterogeneity and dynamics.

Current laboratory experiments cannot provide a complete description of the heterogeneity of conformations at the temporal resolution relevant for proteins dynamics. Ensemble experiments, such as nuclear magnetic resonance (NMR), atomic force microscopy (AFM), and electron microscopy (EM), rely on the signal of multiple molecules, with the result that only ensemble averages can be estimated, rather than the behavior of single molecules. On the other hand, single molecule experiments, such as flourescence resonance energy transfer (FRET) or optical traps, cannot resolve very short timescales due to the effect of signal-noise ratio [22]. As a result, computer models are becoming a more and more used tool to complement experiments [23]. Computer models can provide an unambiguous description of the evolution of a bio-molecular system at atomistic detail and high temporal resolution. Moreover, observables derived by computer models can be directly compared to experimentally resolved features.

This thesis addresses results from a computational approach to investigate the

long-living conformations of bio-molecular systems and the kinetic network between them. Using advanced mathematical techniques, we aim at providing quantitative and statistically relevant models of the dynamics of the studied peptide systems.

## 1.1 Computational simulations of bio-molecules

In recent years, computational models have become a more and more accepted tool to study structural and dynamical properties of bio-molecular systems [24, 25, 26, 27, 28]. In 2013, the field of molecular simulations has also received the recognition of the Nobel Prize in Chemistry (Martin Karplus, Michael Levitt and Arieh Warshel for their development of "multiscale methods for complex systems").
There are two main classical simulation methods: Monte Carlo [29] simulations and Molecular Dynamics [30]. The Monte Carlo algorithm uses random sampling to infer thermodynamic information, but does not directly capture the dynamics of the system. In contrast, Molecular Dynamics (MD) simulations model the time evolution of a number $N$ of interacting particles. This thesis will focus on MD simulations of proteins and peptides.

### 1.1.1 Molecular Dynamics simulations of bio-molecules

Molecular Dynamics (MD) is an example of molecular simulations, which permits the study of complex, dynamical processes like those occurring in biological systems. In MD atoms are treated as classical point particles and their time-evolution is determined by numerically solving Newton's equation of motion. Particles interactions are described by an empirical, phenomenological force field, $U(\mathbf{R})$ (eq. 1.1 and fig. 1.6), which is a function of the positions of all particles $\mathbf{R}$. $U(\mathbf{R})$ is usually separated in a "bonded" and in a "non-bonded" part, where the interaction between particles is described by simple analytical forms, modulated by parameters. The parameters are chosen such that the empirical potential represents a good fit to $ab-initio$ calculations or reproduces experimental data.
Chemically different bonds are described by different parameter values. In most cases, the chosen functional form of the bonded terms is assumed to have little influence, as long as it is physically reasonable. On the contrary, the quality of a force field crucially depends on the parametrization. Non-bonded atoms interact via electrostatic and Van der Waals forces, which are treated as pairwise additive. The non-bonded terms are also modulated by parameters in order to reproduce physical properties of the system.

$$\begin{aligned} U = U_{\text{bonded}} + U_{\text{non-bonded}} = \\ = \sum U_{\text{bonds}} + \sum U_{\text{valence angles}} + \sum U_{\text{torsion angles}} + \\ \sum U_{\text{electrostatic}} + \sum U_{\text{Van der Waals}} + U_{\text{quantum corrections}} \end{aligned} \tag{1.1}$$

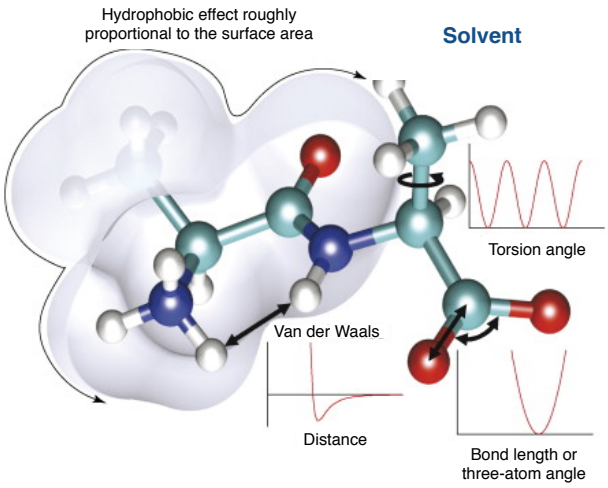


**Figure 1.6**: Example of MD force field adapted from ref. [31].

At the present stage, there are plenty of software packages which can perform a MD simulation [32, 33, 34, 35, 36] and many research groups have put effort into developing meaningful force fields [37, 38, 39, 40, 41, 42, 43] .

### 1.1.2 The Sampling Problem

Reaching biologically relevant timescales and sufficient statistics to characterize a system's behavior via simulations is an extremely challenging task. In fact, typical conformational changes of a protein span over a wide range of timescales, going from $ns$ to $s$. On the other hand, internal vibrations of a molecule occur at timescales of the order of femtoseconds ($10^{-15}s$). Despite the technological progress and the constantly increasing computational power, there is still an unfilled gap between biological and computationally achievable timescales, which is known as the *sampling problem* [44]. Consequently, a lot of effort has been placed into developing methods that could help to overcome this issue.

As an example, Replica Exchange Molecular Dynamics (REMD) [45, 46, 47, 48] takes advantage of the larger number of conformations accessible at higher temperature to enhance the sampling at the temperature of interest. Such method is effective in exploring the different conformations, but requires the information collected at all temperatures to correctly recover the kinetics.
Another approach is given by biased sampling. Examples are umbrella sampling [49, 50], metadynamics [51, 52, 53, 54] and ehnanced sampling [55, 56, 57]. The main difficulty of these approaches lies in discerning a correct way of biasing to recover the kinetics.
Coarse-grained models represent another method for accelerating simulations [58, 59], by considering group of atoms as one single dynamic unit. As a result an effective speed-up of the computation is provided, at the cost of a lower resolution; relative

movements of the atoms in one bead, and their effects on the overall dynamics, cannot in fact be captured by the model.

Another approach is to parallelize the computational cost over multiple independent trajectories, i.e. assuming an ensemble view of the dynamics [60]. Markov State Models (section 1.2) take advantage of this perspective to quantitatively model the dynamics of the system of interest.

## 1.2 Markov State Models

On top of the issue of reaching the timescales of biological interest, sufficient data to infer statistically relevant properties has to be collected. Moreover, even in the case of extremely long sampling, analyzing the enormous amount of data produced by MD simulations would not be a trivial task. The traditional approach of "looking" at the simulated trajectory in search of interesting events would be on the one hand unfeasible, and, on the other, misleading, as the observed properties carry no statistical relevance. Therefore, a quantitative, statistically relevant method to discern the interesting properties from the simulation data is required.

In MD the evolution of the simulated system is a deterministic function of the current and final states of the system itself. In probability theory and statistics, a stochastic process that depends only on the initial and final states, without requiring any prior knowledge of the history of the system, i.e. a memoryless-process, is identified as markovian. A discrete process that satisfies the markovian property can be modeled as a Markov State Model (MSM) [61, 62, 63, 64].

In MSMs the dynamics is represented as the chances of jumping, in a time $\tau$, between $n$ discrete states, which all together comprehend all the possible conformations of the system. The states are groups of conformations, whose dynamic behavior is treated as equivalent.

MSMs permit the computation of time independent quantities, such as equilibrium probabilities of the discrete states and the energy differences between them. Moreover, MSMs allow the identification of relevant (i.e. slow) structural changes and to associate them with timescales, which are directly comparable to experimentally measurable quantities. It is therefore possible to construct a bridge between experiments and computer simulations. MSMs also enable the measurement of quantities not immediately accessible via experiments, such as transition pathways and their probabilities [48, 65], which provide useful insight in understanding processes like protein folding. Finally, information extracted from the model itself, can be used to adaptively drive the model construction, in order to achieve statistical precision in a shorter amount of time [66].

### 1.2.1 Limitations of Markov Models

The dynamics of a perfectly specified system, as described in MD simulations (including solvent degrees of freedom and the velocities of all particles) is deterministic and therefore markovian by construction. In practice, however, schemes to maintain energy, temperature or pressure constant may introduce some stochasticity, which contradicts the markovian hypothesis.
Even in for those simulations where markovianity is truly respected, the dimensionality of the system, which depends linearly on the number of particles, makes it impractical to analyze the dynamics in full dimensional space. Thus, one usually operates in a reduced dimensional space, where the dynamics is projected onto a suitable, system-dependent, set of variables, denoted as reaction coordinates. However, the projected dynamics is not guaranteed to still fulfill the markovian property.

The issue of finding a good set of reaction coordinates to describe the dynamical processes is an entire branch of research. In the field of molecular simulations, the identification of the reaction coordinate is neither systematic nor consistent. Native contacts or root mean square deviation of atomic positions (RMSD) are commonly used as coordinates, nevertheless their suitability should always be verified [67]. Projection onto internal coordinates, such as principal component analysis (PCA) [68, 28, 69] or time-lagged independent component analysis (TICA) [70, 71] might help in the construction of the MSM, despite being of difficult interpretation. In this thesis, we use backbone dihedral angles as reaction coordinates. It has in fact been checked (ref. [67]) that for short peptides backbone dihedral angles capture the interesting dynamics appropriately.

Assuming that a good set of reaction coordinates has been chosen, the quality of the MSM crucially depends on the discretization, i.e. on the way conformations are grouped together. As kinetic vicinity of conformations is not known *a priori*, defining the MSM discrete states (i.e. microstates) is not trivial.
In fact, microstates should be small enough not to contain large energy barriers, and, at the same time, big enough to allow sufficient statistics, such that transition probabilities between each pair of states could be estimated accurately. Moreover, the number of possible states should not be too large for computational feasibility. The side effect of choosing a too coarse discretization is not only that the model would be less detailed, but also that memory effects can be introduced. For example, given a microstate that includes a barrier, trajectories entering into it from different sides of the barrier would have different dynamical behaviors, which would instead be treated as equivalent by the model. These memory effects can be reduced by increasing the lag time of the model, at the cost of a coarser time-resolution and lower statistics.
In recent years, it has been proved that MSM accuracy can be improved by using a discretization that well resembles the features of the underlying energy landscape [72, 61]. Such features are, however, unknown *a priori*, due to the high dimensionality

of the system. In this thesis we will tackle the issue of appropriate discretizations in chapter 4.

## 1.3 The Thesis

In this thesis we aim at highlighting the underlying hierarchical dynamics of short peptides. We exploit MSMs to analyze MD data, and chapter 2 is dedicated to summarizing the theoretical background common to the research presented in this dissertation.

Chapter 3 presents an application of MSMs to model the dynamics of an intrinsically disordered peptide: human islet amyloid polypeptide (hIAPP). The dynamics of hIAPP fragments shows inherent hierarchy, as the same long-living configurations that are found in the model of shorter fragments arises in the model of the longer sequences, albeit with different timescales.

The analysis presented in chapter 3 is based on a crisp definition of microstates, whose number, however, increases exponentially with the length of the sequence. Such approach is therefore computationally unfeasible for longer peptides and proteins. Following a novel description for conformational kinetics of peptides in terms of smooth basis functions, in chapter 4 we develop a novel set of transferable basis functions defined as combinations of residue-centered kinetic modes, which are obtained from kinetic models of terminally blocked amino acids. Such basis definitions has a straightforward interpretation of the dynamics in terms of the slow motions of the composing amino acids. This approach also identifies a possible path to describe more complex systems with a hierarchical approach.

Chapter 5 explores the differences induced by the MD force fields on the dynamic properties and MSMs are presented as a tool for comparison of different kinetic models. In light of these findings, chapter 6 presents a force-field-dependent library of all amino acids slowest processes, which constitute the residue-centered basis functions for the force field of choice.

Finally, the chapter 7 summarizes the most relevant results of this thesis, as well as some concluding remarks.

Chapter 2

# Theory

Throughout this chapter we present a more detailed description of the salient aspects of the theory of propagators (section 2.1), Markov State Models (section 2.2) and the Variational Approach to conformation dynamics (section 2.3). For further details refer to [73, 61, 74, 64, 75, 76, 77].

## 2.1 Continuous Markov Model

Lets consider a state space $\Omega$, whose elements $\mathbf{x} \in \Omega$ represent the dynamical states of the system at study. In the case of molecular systems in explicit solvent, $\mathbf{x}$ contains both the positions and the velocities of all particles (including solvent's) of the system. The time-evolution of the system (here for simplicity we consider time discrete) is the sequence of states $\{\mathbf{x}(t_1), \mathbf{x}(t_2), ..., \mathbf{x}(t_n)\}$ visited by the system at times $\{t_1, t_2, ..t_n\}$. Such sequence is a trajectory $\mathbf{x}(t) \in \Omega$. The trajectory $\mathbf{x}(t)$ is a Markov process if it fulfills the markovian property: the time-evolution of the system is memoryless. This means that the transition probability $p(\mathbf{x}, \mathbf{y}, \tau)$ of being in state $\mathbf{x}$ and going to state $\mathbf{y}$ in an amount of time $\tau$, only depends on the initial and final states and not on the history of the system. It can be described in form of an operator $\mathcal{P}(\tau)$, defined as:

$$p(\mathbf{x}, \mathbf{y}, \tau)\mathrm{d}\mathbf{y} = \mathcal{P}[\mathbf{x}(t+\tau) \in \mathbf{y} | \mathbf{x}(t) = \mathbf{x}] \qquad \mathbf{x}, \mathbf{y} \in \Omega; \tau \in \mathbb{R}_{0+}. \tag{2.1}$$

$\mathcal{P}(\tau)$ is referred to as the propagator and $\tau$ is a parameter of the model, known as the lag-time.
The trajectory obtained from a MD simulation (in full continuous state space) is a Markov process by construction, because the conformation $\mathbf{x}(t + dt)$ is a deterministic function of $\mathbf{x}(t)$. Modeling the time-evolution of the system at study in terms of a propagator constitutes the Markov model of the system's dynamics.

To better understand the concept of the propagator and what type of information can be derived from it, lets consider as an example a diffusion process in the one-dimension potential shown in fig. 2.1.a, which has been previously used in ref. [61]. The potential presents four minima, indicated in the figure as A, B, C, D.

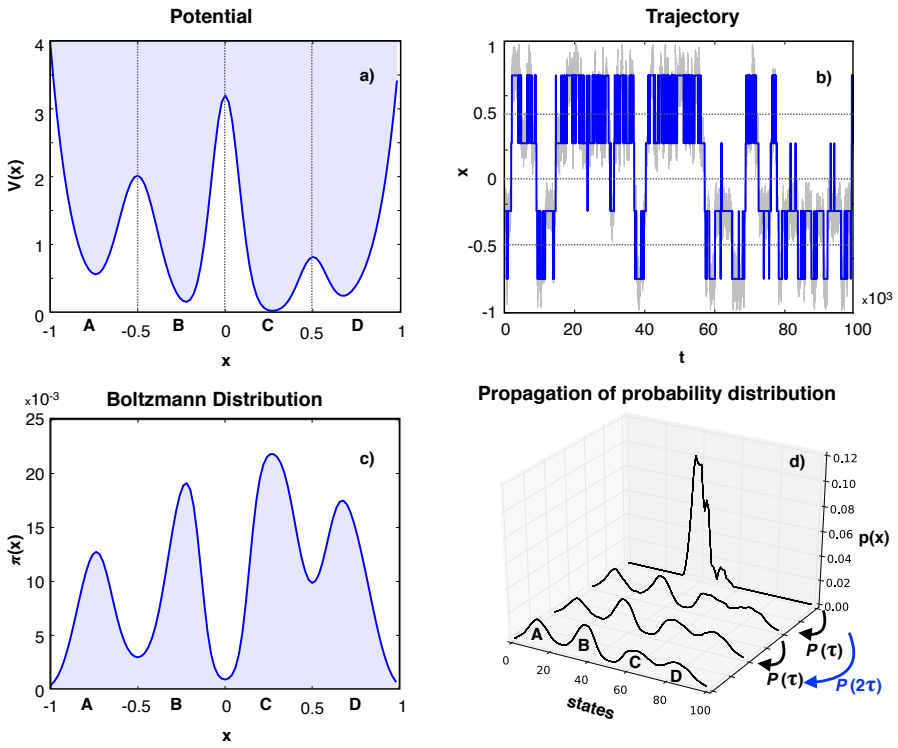


**Figure 2.1**: (a) Potential energy function with four metastable states: $V(x) = 4(x^8 + 0.8e^{-80x^2} + 0.2e^{-80(x-0.5)^2} + 0.5e^{-40(x+0.5)^2})$. (b) Trajectory of a jump process between $x$ and the orthogonal neighbors, i.e. from $x$ to $\{x, x+1, x-1\}$. $x$ is restricted in the range (-1,1) and the range is divided in 100 bins of width 0.1 (gray line) or in 4 bins: (-1,-0.5), [-0.5, 0), [0, 0.5), [0.5,1) (blue line). Probability of jumping is (in units of $k_B T$) $p(i,j) = \frac{1}{Z_i}\min\{1, exp(-(V(j) - V(i)))\}$ and $Z_i = \sum_i \min\{1, exp(-(V(j) - V(i)))\}$ (c) Boltzmann distribution relative to the potential in a. (d) Effect of the propagator in evolving forward in time the probability distribution $p(\mathbf{x})$.

An example trajectory is shown in fig. 2.1.b, where the value of $x$ are recorded over time (100000 time-steps). The trajectory oscillates in the range $x \in (-1, 1)$, and sharp jumps can be noticed whenever the trajectory overcomes the barrier at $x = 0$. Smaller jumps distinguish transitions across the second highest barrier, whereas transitions across the third barrier occur more frequently.
Looking at all the possible small variations of the reaction coordinate $x$ does not produce a clear and easily understandable picture of the dynamics. In this example, representing the dynamics as jumps between four different regions (vertical lines in fig. 2.1.a) approximates well the behavior of the trajectory. These are the regions where the system spends a long time, and are called metastable states. Within one region in fact $x$ oscillates quickly, whereas transitions out of the region occur more rarely. The dynamics of the system can thus be modeled as the four metastable states (roughly identifiable with the minima A, B, C, and D), and the transition probabilities between them. An equivalent simplification is what we seek for modeling the dynamics of complex molecular systems: identify the long-living

conformations of the molecule (metastable states) and the relaxation timescales of the corresponding conformational changes. Such type of information is encoded in the propagator.

The propagator transports the probability distribution forward in time (fig. 2.1.d) Lets assume that at time $t_0$ the probability distribution is concentrated around $x = 0$. The probability distribution at time $t_0 + \tau$ is thus described by:

$$p(\mathbf{x}(t_0 + \tau)) = \mathcal{P}(\tau)p(\mathbf{x}(t_0)). \tag{2.2}$$

If the dynamics fulfills the markovian property, the probability density of the system at time $t_0 + k\tau$ is a deterministic function of the probability density at time $t_0$. For example, the evolution of the probability density at $t_0 + 2\tau$ can be seen as:

$$p(\mathbf{x}(t_0 + 2\tau)) = \mathcal{P}(\tau)p(\mathbf{x}(t_0 + \tau)) = \mathcal{P}(\tau)[\mathcal{P}(\tau)p(\mathbf{x}(t_0))] = \mathcal{P}(2\tau)p(\mathbf{x}(t_0)). \tag{2.3}$$

Generalizing eq. 2.3 we obtain:

$$p(\mathbf{x}(t + k\tau)) = \mathcal{P}(k\tau)p(\mathbf{x}(t)) = [\mathcal{P}(\tau)]^k p(\mathbf{x}(t)), \tag{2.4}$$

where applying the propagator $k$ times is rewritten as:

$$\mathcal{P}(k\tau) = [\mathcal{P}(\tau)]^k. \tag{2.5}$$

Eq. 2.4 is denoted as the Chapman-Kolmogorov equation. As consequence of eq. 2.4, if the dynamic is truly markovian, the probability distribution of a simulated trajectory at time $t_0 + k\tau$ is equivalent to the one predicted by applying the propagator $k$ times to the initial probability distribution. Therefore, the fulfillment of eq. 2.4 can be used to prove the markovianity of the dynamics [28].

For $t \to \infty$, the probability distribution evolves towards its equilibrium distribution (fig. 2.1.c), that at constant temperature is given by the Boltzmann distribution [78]:

$$\pi(\mathbf{x}) = \mathcal{Z}(\beta)^{-1} \mathrm{e}^{-\beta\mathcal{H}(\mathbf{x})} \tag{2.6}$$

where $\mathcal{H}(\mathbf{x})$ is the Hamiltonian of the system , $\beta = 1/k_B T$ is the Boltzmann constant and $\mathcal{Z}(\beta) = \int \mathrm{e}^{-\beta\mathcal{H}(\mathbf{x})} \mathrm{d}\mathbf{x}$ is the partition function.
The Hamiltonian is given by the sum of the potential energy and the kinetic energy. In MD the kinetic energy depends on the velocities which are distributed according to the Maxwell distribution [79] and the potential energy is given by the empirical force field (eq. 1.1). It would thus be theoretically possible to compute $\pi(\mathbf{x})$ by simply integrating eq. 2.6. However, the state space of a biomolecular system is extremely vast and a direct evaluation of $\mathcal{Z}(\beta)$ is not possible.

A way to estimate $\pi(\mathbf{x})$ is via sampling, if the system is ergodic. This means that, for infinitely long trajectories, each state $\mathbf{x}$ will be visited infinitely often and the fraction of time the system will spend in each state will be proportional to its

equilibrium probability. If ergodicity holds, ensemble averages of an observable of interest $A$ can be computed as time averages:

$$\langle A \rangle = \frac{1}{T} \int_0^T A(\mathbf{x}(t)) \mathrm{d}t = \int A(\mathbf{x}) p(\mathbf{x}) \mathrm{d}x, \tag{2.7}$$

i.e. properties of the state space can be inferred via sampling.
Ergodicity implies that any state $\mathbf{x} \in \Omega$ can be reached from any other state in the state space. The physical process described by MD simulations is ergodic, however, in MD ergodicity can often only be assumed. In fact there is no way of knowing if the full state space has been explored and one relies on indirect arguments to verify the convergence of the simulations.

A property which is not necessary for the propagator, but implies profound analytical statements is *reversibility*. A system is reversible if it satisfies the *detailed balance* condition:

$$p(\mathbf{x}, \mathbf{y}, \tau)\pi(\mathbf{y}) = \pi(\mathbf{x})p(\mathbf{y}, \mathbf{x}, \tau) \ \forall \ \mathbf{x}, \mathbf{y}, \ \tau. \tag{2.8}$$

The detailed balance condition implies that the forwards and backwards transition probabilities between pairs of states are equal. In many cases the model obtained by MD simulations does not satisfy the detailed balance condition. The fulfillment of such property is, however, logically expected, as its breaking would indicate the unphysical scenario of a direction of the dynamics that allows the production of work from a system in equilibrium. Therefore, in practice, detailed balance is enforced in the construction of the model.

The properties we are interested in, i.e. metastable states and relaxation timescales of the transitions between them, are encoded in the eigenvalues and eigenfunctions of the propagator [60]. To better understand the interpretation of such quantities, lets first consider another operator, the *generator* $\mathcal{L}(\mathbf{x}(t))$. The generator is a time-continuous operator that represents the variation of the probability density with respect to time. Its effect can be expressed as a differential equation, similarly to a Fokker-Planck equation [80]:

$$\frac{\partial}{\partial t} p(\mathbf{x}(t)) = \mathcal{L}(\mathbf{x}) p(\mathbf{x}(t)). \tag{2.9}$$

The generator $\mathcal{L}(\mathbf{x}(t))$ is related to the propagator $\mathcal{P}(\tau)$ by the integration of eq. 2.9, hence:

$$\mathcal{P}(\tau) = \mathrm{e}^{\mathcal{L}(\mathbf{x})\tau}. \tag{2.10}$$

If the system is in equilibrium ($p(\mathbf{x}) = \pi(\mathbf{x})$), the variation of probability density in time is zero. From eq. 2.9:

$$\frac{\partial}{\partial t} \pi(\mathbf{x}) = \mathcal{L}(\mathbf{x})\pi(\mathbf{x}) = 0. \tag{2.11}$$

Therefore, the equilibrium distribution $\pi(\mathbf{x})$ is an eigenfunction $l_1(\mathbf{x})$ of the generator associated to the eigenvalue $k_1 = 0$. As a consequence, $l_1$ does not have negative entries. For ergodic systems the eigenvalue $k_1 = 0$ always exists and it is non-degenerate [81, 82], which implies the uniqueness of the equilibrium distribution (eq. 2.6). Moreover, all other eigenvalues are guaranteed to be smaller or equal to 0, i.e. $k_i \leq 0 \; \forall \; i$ (Perron-Froebenius Theorem [81, 82]).
For an isolated system, the generator is time independent, thus the temporal and the spatial parts of eq. 2.9 can be treated separately, in analogy to the solution of a time-dependent Schrödinger equation. Hence, the solution to eq. 2.9 is an exponential decay:

$$p(\mathbf{x}(t)) = \sum_i c_i \mathrm{e}^{k_i t} l_i(\mathbf{x}). \tag{2.12}$$

Eq. 2.12 indicates that an arbitrary probability density $p(\mathbf{x}(t = 0))$ can be seen as a superposition of modes given by the eigenfunctions of the generator $l_i(\mathbf{x})$, whose expansion coefficients, $c_i \mathrm{e}^{k_i t}$ are modulated in time and normalized such as that

$$\int p(\mathbf{x}(0))\mathrm{d}\mathbf{x} = \int \sum_i c_i l_i(\mathbf{x})\mathrm{d}\mathbf{x} = 1 \tag{2.13}$$

is assured.
From eq. 2.12 we notice that for $t \to \infty$ all processes disappear (their expansion coefficients tend to zero), with the exception of the process associated to $k_1 = 0$, i.e. the stationary distribution. The eigenvalues $k_i$ have thus the interpretation of a relaxation rate, as they indicate how fast a specific eigenfunction (or mode) will disappear. The modes with decaying constant close to zero indicate the slow dynamic processes of the system.
The dynamic processes are encoded in the sign-structure of the eigenfunctions $l_i(\mathbf{x})$. They assign negative or positive values to each state and represent the exchange of probability density between groups of states of different sign [60]. States to which the eigenfunction assigns a value of zero are not effected by the particular process described by the specific eigenfunction under consideration.
If detailed balance holds, eigenvalues and eigenfunctions of the generator are real valued [83, 84].
Often in literature instead of decaying rates, implied timescales are used, which are given by:

$$t_i = \frac{1}{k_i}. \tag{2.14}$$

Implied timescales are one of the most interesting kinetic properties captured by Markov models: they can be related to experimentally measurable quantities and thus link the experimental observations to conformational changes at the molecular level accessible by the simulations [28].

The eigenvalues and eigenfunctions of the propagator are linked to those of the generator. Applying the propagator to the $i^{th}$ eigenfunction of the generator one

obtains:

$$\mathcal{P}(\tau)l_i(\mathbf{x}) = \mathrm{e}^{\mathcal{L}(\mathbf{x})\tau}l_i(\mathbf{x}) \tag{2.15}$$

and expanding the exponential of an operator as:

$$\mathrm{e}^{\mathcal{L}(\mathbf{x})\tau} = \sum_{n=0}^{\infty}\frac{\tau^n}{n!}\mathcal{L}(\mathbf{x})^n, \tag{2.16}$$

we can rewrite eq. 2.15 as:

$$\begin{aligned}\mathcal{P}(\tau)l_i(\mathbf{x}) = \sum_{n=0}^{\infty}\frac{\tau^n}{n!}\mathcal{L}(\mathbf{x})^n l_i(\mathbf{x}) = \\ = \sum_{n=0}^{\infty}\frac{\tau^n}{n!}k_i^n l_i(\mathbf{x}) = \mathrm{e}^{k_i\tau}l_i(\mathbf{x}) = \lambda_i(\tau)l_i(\mathbf{x}).\end{aligned} \tag{2.17}$$

Eq. 2.17 indicates that the eigenfunctions of the propagator and those of the generator are the same and that the eigenvalues of the propagator $\lambda_i(\tau)$ are time dependent and related to those of the generator by:

$$\lambda_i(\tau) = \mathrm{e}^{k_i\tau}. \tag{2.18}$$

As a consequence, the spectrum of the propagator is bound from above by $\lambda_1(\tau) = \mathrm{e}^{k_1\tau} = 1$ [63], which is unique for ergodic systems, i.e.:

$$|\lambda_i(\tau)| \leq \lambda_1 = 1 \tag{2.19}$$

Hence, for ergodic systems, there is a unique stationary distribution $l_1(\mathbf{x}) = \pi(\mathbf{x})$. The eigenvalues of the propagator are related to the relaxation timescales of the systems by:

$$t_i = \frac{-\tau}{\ln(\lambda_i(\tau))}, \tag{2.20}$$

where $t_i$ is the implied timescale of the $i^{th}$ process. If detailed balance holds, then eigenvalues and eigenfunctions of the propagator are also real valued [83, 84].

An equivalent description to the one of the propagator for the time-evolution of the probability distribution $p(\mathbf{x})$ is given by the *transfer operator* $\mathcal{T}(\tau)$. The transfer operator is defined as:

$$p(\mathbf{x}(t+\tau)) = \mathcal{T}(\tau)p(\mathbf{x}(t)) = \frac{1}{\pi(\mathbf{x})}\int_{\mathbf{y}\in\Omega} p(\mathbf{y}.\mathbf{x})\pi(\mathbf{x})p(\mathbf{x}(t))\mathrm{d}\mathbf{y} \tag{2.21}$$

$\mathcal{T}(\tau)$ and $\mathcal{P}(\tau)$ share the same eigenvalues, and the eigenfunctions are related by:

$$l_i(\mathbf{x}) = \pi(\mathbf{x})r_i(\mathbf{x}), \tag{2.22}$$

where $l_i(\mathbf{x})$ is the $i^{th}$ eigenfunction of the propagator and $r_i(\mathbf{x})$ is the corresponding eigenfunction of the transfer operator (also called co-functions). $l_i(\mathbf{x})$ and $r_i(\mathbf{x})$

are orthogonal and normalized such that $\langle r_i(\mathbf{x}), l_j(\mathbf{x})\rangle_{\pi(\mathbf{x})} = \delta_{ij}$. The propagator eigenfunction $l_1(\mathbf{x})$, associated to $\lambda_1 = 1$, is the equilibrium distribution $\pi(\mathbf{x})$; consequently its co-funtion is $r_1(\mathbf{x}) = l_1(\mathbf{x})/\pi(\mathbf{x}) = \mathbf{1}$, i.e. it is a constant function on the entire state space $\Omega$. The dominant eigenfunctions of the transfer operator maintain the same sign structure possessed by the propagator eigenfunctions and therefore carry the same information about dynamical transitions. Thinking in terms of $\mathcal{T}(\tau)$ is useful for the construction of spatially and temporally discretized Markov models, which are discussed in section 2.2.

## 2.2 Discrete Markov Model

Despite MD on a continuous state space being markovian by construction, in practice, a dimensionality reduction is required. Such dimensionality reduction occurs in two steps:

- a projection onto a sub set of degrees of freedom, defined as reaction coordinates, which capture the relevant (i.e. slow) dynamics of the system;
- a discretization of the state space into a set $N$ of non-overlapping states ($S_i \cap S_j = 0$ for all $i \neq j$), whose union is the totality of the state space ($\cup_{i=1}^{n} S_i = \Omega$). The discrete states $S_i$ are called *microstates*. The trajectory is assumed to be at local equilibrium within one microstate, i.e. each point $\mathbf{x} \in S_i$ has an equivalent dynamic behavior.

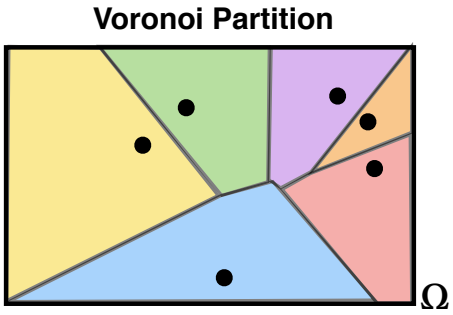


**Figure 2.2**: Schematic example of a Voronoi partition

A typical discretization is a *Voronoi partition* [85] (fig. 2.2), where a set of $n$ centers $\bar{\mathbf{x}}_i$, $i = 1, ..., n$ is defined and the set $S_i$ is the union of all the points $\mathbf{x}$ closer (according to some metric) to $\bar{\mathbf{x}}_i$ than to any of the other centers. The crucial point in the Voronoi partition is the choice of the centers, and there are multiple algorithms which aim at finding the most representative and robust set of centers $\bar{\mathbf{x}}_i$ [86, 87, 88].

The Markov model on a discretized state space is denoted as Markov State Model (MSM). In a MSM framework, the probability density of the system $\mathbf{p}(\mathbf{x}(t))$ is a

vector, whose elements represent the probabilities of finding the system in each microstate at time $t$. As the microstates constitute a full partition of the state space, the probability of finding the system at any microstate at time-step $t$ is equal to one.

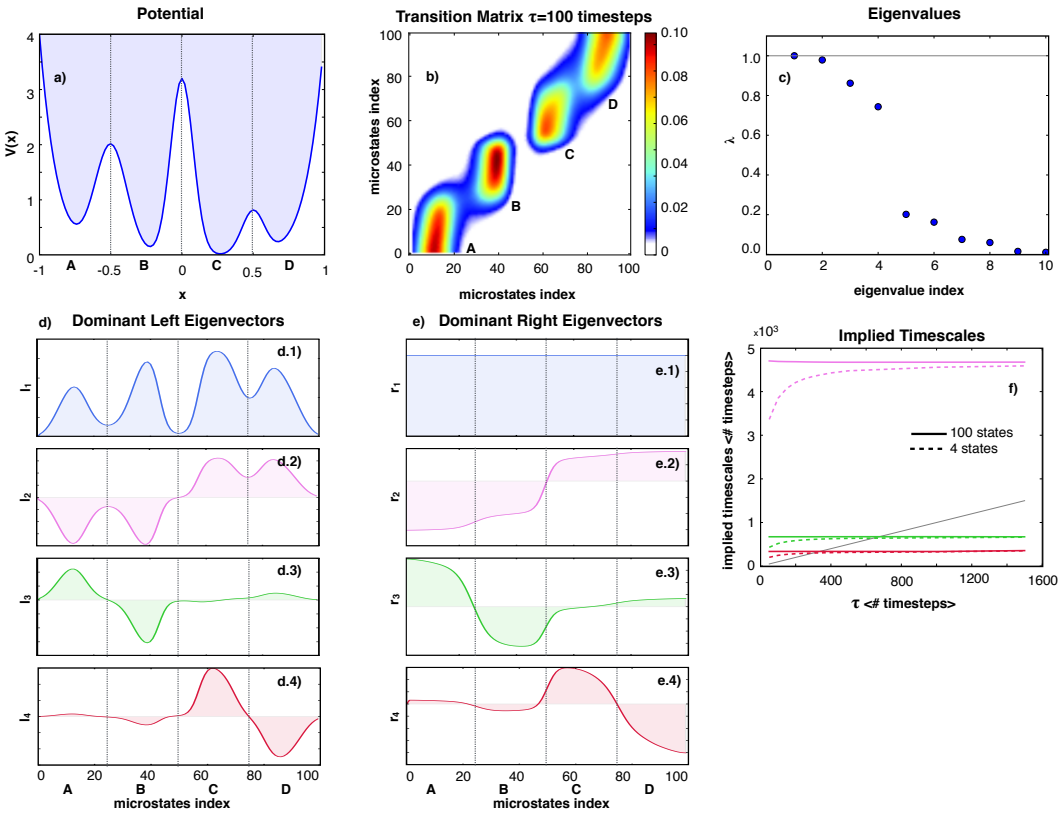


**Figure 2.3**: (a) Potential energy function with four metastable states as in fig. 2.1.a. (b) Transition matrix estimated for 100 microstates and lag-time $\tau = 100$ time-steps. (c) First 10 eigenvalues of the transition matrix in b. (d) Left dominant eigenvectors of the transition matrix in b. (e) Right dominant eigenvectors of the transition matrix in b. (f) Implied timescales estimated from a MSM with 100 states (solid line) or 4 states (dashed line).

The time-evolution of the probability density vector is given by:

$$\mathbf{p}(\mathbf{x}(t+\tau)) = \mathbf{T}(\tau)\mathbf{p}(\mathbf{x}(t)), \tag{2.23}$$

which is the discrete equivalent of eq. 2.1. $\mathbf{T}(\tau)$ is the *transition matrix* and is the discrete representation of the transfer operator. The elements of the transition matrix are the transition probabilities between pairs of microstates in an amount of

time $\tau$:

$$\begin{aligned} T_{ij} &= \mathbb{P}[\mathbf{x}(t+\tau) \in S_j | \mathbf{x}(t) \in S_i], \\ T_{ij} &\in \mathbb{R}\ T_{ij}(\tau) \geq 0\ \forall\ i, j, \tau, \\ \sum_{i}^{N} T_{ij}(\tau) &= 1\ \forall\ j, \tau, \end{aligned} \tag{2.24}$$

i.e. the transition matrix is a row stochastic matrix.

Going back to our four-well potential example, a visualization of the transition matrix is show in fig. 2.3.b. The state space $\Omega : \mathbf{x} \in (-1, 1)$ is subdivided in 100 bins of bin-width 0.1, and the entries of the transition matrix $T_{ij}$ are computed at a lag-time of $\tau = 100$ time-steps. Four regions of higher transition probability can be identified, which correspond to the four minima of the potential. The low-transition probability regions represent the barriers.
Given a MD trajectory, the elements $T_{ij}$ can be estimated by counting the number of times the trajectory is observed initially in state $S_i$ and at a time $\tau$ later in state $S_j$ ($K_{ij}$). Such transition counts between microstates are stored in form of a matrix, called *count matrix*, $\mathbf{K}(\tau)$. If multiple trajectories are available, the contribution to the count matrix of these trajectories can be computed independently. The elements of the transition matrix can be estimated from the count matrix according to:

$$T_{ij} = \frac{K_{ij}}{K_i(\tau)}, \tag{2.25}$$

where $K_{ij}$ is the number of transitions between states $S_i$ and $S_j$, and $K_i$ is the number of transitions originated in $S_i$. In theory, the transitions contributing to the count matrix should be independent, e.g. $\mathbf{x}(t_0) \to \mathbf{x}(t_0 + \tau), \mathbf{x}(t_0 + \tau) \to \mathbf{x}(t_0 + 2\tau)$ etc. This approach, however, has the consequence that a huge portion of simulation data does not contribute to the estimation of the model (fig. 2.4.a). In practice, one usually chooses a sliding window approach, which on one hand produces correlated counts, but on the other permits better statistics ( fig. 2.4.b).

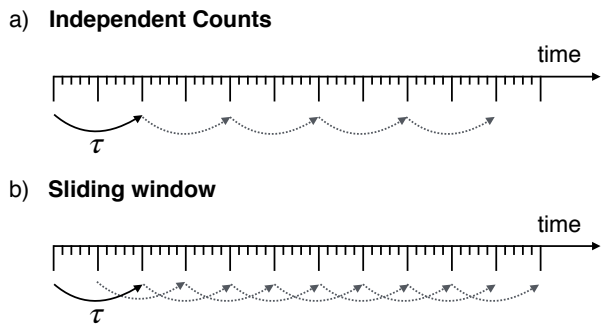


**Figure 2.4**: (a) Independent counts approach. (b) Sliding window approach.

In analogy to the theory of the transfer operator, the transition matrix eigenvalues are related to the relaxation timescales of the system, and the dynamic processes are encoded in the eigenvectors. Lets refer again to the transition matrix $\mathbf{T}(\tau)$ of fig. 2.3.b. The first ten eigenvalues of $\mathbf{T}(\tau)$ are shown in fig. 2.3.c. The first four eigenvalues have much higher values compared to the the remaining ones. Such separation is known as the spectral gap and identifies the dominant kinetic processes of the system. It is worth to notice that in a real bio-molecular system it is often difficult to distinguish the dominant processes from the fast-decaying ones, as there might be no evident spectral gap.
The eigenvalues $\lambda_i(\tau)$ of a transition matrix $\mathbf{T}(\tau)$ depend on the lag-time, which is a parameter of the model. Nonetheless, the associated timescales should be independent of the the lag-time at which the model is estimated.

$$t_i = -\frac{n\tau}{\ln(\lambda_{i,T(n\tau)})} = -\frac{n\tau}{\ln(\lambda^n_{i,T(\tau)})} = -\frac{n\tau}{n\ln(\lambda_{i,T(\tau)})} = -\frac{\tau}{\ln(\lambda_{i,T(\tau)})}, \tag{2.26}$$

where $\lambda_{i,T(n\tau)}$ is the $i^{th}$ eigenvalue of the model estimated at lag-time $n\tau$. In practice, however, the timescales are only constant in a limited range of lag-times, due to memory effects or limited statistics. This can be seen, for instance, in our example if instead of 100 microstates only four are used, corresponding to the four minima of the potential (fig. 2.3.f). The first three implied timescales estimated for the finest discretization (solid line) result constant for all values of $\tau$ between 50 and 1500 time-steps. The coarser discretization estimation instead converges to the same values only at longer lag-times. Assuming that an infinite amount of data is available, the implied timescales estimated from a transition matrix converge to their true value as the lag-time $\tau$ increases, because the approximation of local equilibrium within a microstate becomes progressively more accurate. On the other hand, if the lag-time is too big, the model becomes coarser (many processes have decayed) and the limited statistics has an effect on the quality of the model itself. In practice, one looks at the range of lag-times where the slowest relaxation timescales reach a plateau and selects a value of $\tau$ in such range for constructing the model. Moreover, the behavior of the dominant timescales with respect to $\tau$ can be used to check the quality of the model, as constant implied timescales indicate that the dynamics can be considered markovian. It should be noted, however, that the convergence of the implied timescales is not a complete test of markovianity, as the lag-time independence of the eigenvectors should also be verified.

The right eigenvectors of the transition matrix are the discrete equivalent of the eigenfunctions of the transfer operator. The first right eigenvector $\mathbf{r}_1(\mathbf{x})$ is constant over the visited microstates. The eigenvectors associated to the slowest timescales represent the dominant dynamic processes. They indicate the transfer of probability density between groups microstates of opposite sign and are almost constant within one metastable state. The left eigenvectors $\mathbf{l}_i(\mathbf{x})$ of the transition matrix are the discrete representation of the propagator eigenfunction. Fig.s 2.3.d and .e represent respectively the first four left and right eigenvectors of the transition matrix of fig.

2.3.b. As expected, the first right eigenvector is constant, whereas the first left eigenvector is the Boltzmann distribution. The second eigenvector (first dynamic process) corresponds to transitions across the barrier at $\mathbf{x} = 0$. The second and third dynamic processes represent the transitions A $\rightarrow$ B and C $\rightarrow$ D respectively.

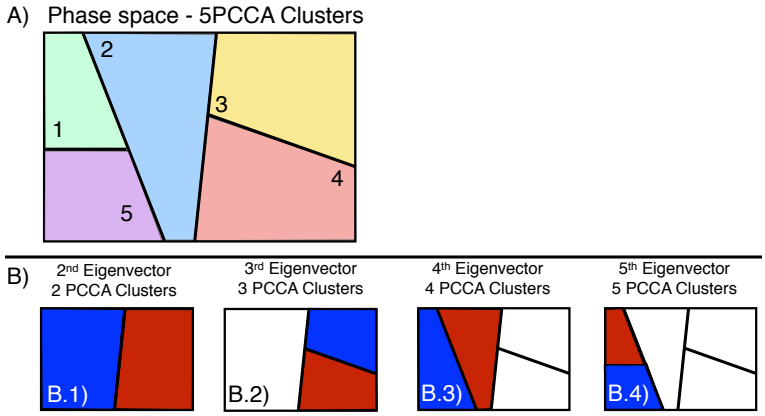


**Figure 2.5**: Schematic representation of PCCA. A) PCCA partition of the state space into 5 PCCA clusters. B) Hierarchy of the underlying energy landscape obtained by iterative use of the PCCA algorithm with increasing number of eigenvectors.

For a high-dimentional molecular system, interpreting the eigenvectors in terms of conformational changes is not trivial. A dynamic process, in fact, is normally defined as a transition between the microstates to which the eigenvector assigns negative values and those to which it assigns positive values. Such approach, however, may lead to a confused interpretation, due to the arbitrary assignment of those microstates whose eigenvector entry is close to zero, and therefore are not representative of the conformational change.
An automatic approach to assign microstates to metastable macrostates is by Perron Cluster Cluster Analysis (PCCA) [60, 89], a method that exploits the eigenspectrum of the transition matrix to coarse-grain the MSM. In a PCCA the first M eigenvectors are used to map each of the microstates into M long-living states, in a crisp or fuzzy manner. A schematic representation of PCCA is shown in fig. 2.5. If five eigenvectors are provided, the state space is partitioned in five regions, corresponding to five metastable states of the system (fig. 2.5.A). However, this does not provide any information about the transitions between such metastable states. An iterative application of the method provides a picture of the hierarchy of the free-energy barriers of the system by splitting regions of the state space according to the eigenvectors sign(fig. 2.5.B). Giving as an input the first two dominant eigenvectors (fig. 2.5, B.1) the conformational space is split in two long-lived clusters along the highest energy barrier of the system. The blue and red coloring represent areas of the configurational space where the eigenvector has opposite sign, whereas the white area represent those microstates that do not participate to the dynamical process. When the third eigenvector is provided, the algorithm splits the right

cluster in two kinetically diverse states, based on the eigenvector sign (fig. 2.5, B.2). Analogously, the information yielded by the fourth eigenvector splits the left cluster (fig. 2.5, B.3) into two sub sets, and the fifth eigenvector splits the left-most state in two (fig. 2.5, B.4). Further iterations of the algorithm create more macrostates. The slowest process of the system can thus be interpreted as transition between the superimposition of clusters one, two and five on one side and of clusters three and four on the other. The second slowest process represents transitions between clusters three and four, whereas the third process represents transitions between cluster two and clusters one and five superimposed. Finally, the third slow process represents transitions between cluster one and cluster five. In recent years more efficient and robust versions of the algorithm have been implemented [90, 67, 91].

## 2.3 The Variational Approach to Conformation Dynamics

The transition matrix, together with the discretization, constitute the gist of the MSM. However, it has to be remarked that the dynamics of a MD-trajectory projected onto a reduced set of discretized degrees of freedom is not guaranteed to fulfill the markovian property. In fact, within a microstate, all elements are considered as belonging to the same local minimum and the probability of equilibrating back to the microstate instead of transitioning out of it is neglected. Therefore, the quality of the MSM depends crucially on the discretization.

The quality of the model can be improved if a discretization that finely separates into different states the transition barriers is used [72, 61]. This can be seen in our four-minima example, by comparing the dominant eigenvectors estimated with four-states discretization or 100-states discretization (fig. 2.6). However, the features of the energy landscape are not known *a priori* and a fine discretization of the full state space is unfeasible, due to the high-dimensionality of the state space itself. MSMs are therefore associated to a systematic error that depends on the discretization [72, 61].

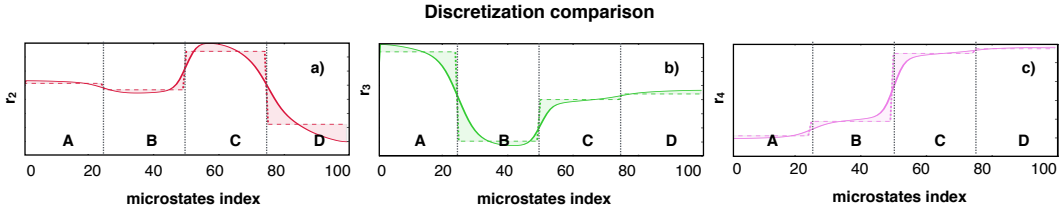


**Figure 2.6**: Comparison of the dominant right eigenvectors of the transition matrix in fig 2.3.b estimated with 4 or 100 states.

In recent years a new approach has been developed: instead of approximating

the transfer operator eigenfunctions on a crisp discretization (i.e. MSM), a linear combination of basis functions that maps elements of the state space to real values can be used [75, 76], in analogy to quantum mechanics. Such approach, known as variational approach to conformation dynamics (VAC), has the advantage of producing a precise model of the dynamics with a smaller number of basis functions than the number of microstates used in conventional MSM. The continuous functions, in fact, help in overcoming limitations provided by the crisp discretization. For instance, the basis functions can be designed as to mimic the conformational changes of the system [76]. Moreover, through appropriate selection of basis functions, previously acquired information or chemical intuition may be included in the model. It is also worth to notice that discrete states often lack a clear structural meaning, therefore the interpretation of the eigenvectors as conformational changes is not straightforward. If instead the basis functions carry some structural meaning, VAC can ease the interpretation of the model (chapter 4).

The VAC exploits the fact that the propagator is a self-adjoint operator with respect to the weighted scalar product ($\langle \mathcal{P}(\tau) f(\mathbf{x}), g(\mathbf{x}) \rangle_{\pi^{-1}} = \langle f(\mathbf{x}), \mathcal{P}(\tau) g(\mathbf{x}) \rangle_{\pi^{-1}}$) and has a bounded eigenvalue spectrum. Therefore a variational principle can be formulated:

$$\langle \varphi(\mathbf{x}), \mathcal{P}(\tau)\varphi(\mathbf{x}) \rangle_{\pi^{-1}} = \int_{\Omega} \varphi(\mathbf{x}) \pi^{-1}(\mathbf{x}) \mathcal{P}(\tau) \varphi(\mathbf{x}) \mathrm{d}x \leq \lambda_1 = 1, \tag{2.27}$$

where $\varphi(\mathbf{x})$ is a trial function normalized such that:

$$\langle \varphi(\mathbf{x}), \varphi(\mathbf{x}) \rangle_{\pi^{-1}} = \int_{\Omega} \varphi(\mathbf{x}) \pi^{-1}(\mathbf{x}) \varphi(\mathbf{x}) \mathrm{d}x = 1. \tag{2.28}$$

The equality in eq. 2.27 holds if and only if $\varphi(\mathbf{x}) = l_1(\mathbf{x})$. Finding the trial function $\varphi(\mathbf{x})$ that maximises the left term of eq. 2.27 is therefore a way to approximate $l_1(\mathbf{x})$. Such a procedure can be iteratively applied to approximate the other eigenfunctions, under the additional constraint that they will be orthogonal to the previous eigenfunctions.

The trial function $\varphi(\mathbf{x})$ can be linearly expanded in terms of a set of $M$ basis functions $\{\psi_i(\mathbf{x})\}_{i=1}^{M}$.

$$\begin{aligned} \varphi(\mathbf{x}) &= \sum_{i=1}^{M} a_i \psi_i(\mathbf{x}), \\ a_i &\in \mathbb{R}. \end{aligned} \tag{2.29}$$

The method of linear variation can be used to estimate the optimal coefficients $a_i$ that maximize the left term of eq. 2.27, while the basis functions, which are not required to be orthonormal, are kept constant. The normalization constraint eq. 2.28 is enforced using the method of Lagrange multipliers. The derivation of the method (appendix A) indicates that the optimal expansion coefficients $a_i$ can be obtained by solving the generalized eigenvalue problem (eq. 2.30)

$$\mathbf{C}(\tau)\mathbf{A} = \Lambda \mathbf{S} \mathbf{A}, \tag{2.30}$$

where the matrix elements of $\mathbf{A}$ are the expansion coefficients of the first N basis functions $\{\psi_i(\mathbf{x})\}$; $\Lambda$ is a diagonal matrix which has the variational estimates of the eigenvalues as diagonal elements; $\mathbf{S}$ is the overlap matrix:

$$S_{ij} = \langle \psi_i(\mathbf{x}), \psi_j(\mathbf{x}) \rangle, \tag{2.31}$$

and $\mathbf{C}(\tau)$ has the interpretation of time lagged correlation matrix:

$$C_{ij} = \langle \psi_i(\mathbf{x}), \mathbf{P}(\tau)\psi_j(\mathbf{x}) \rangle. \tag{2.32}$$

Due to the high dimensionality of the state space $\Omega$, the direct evaluation of $\mathbf{S}$ and $\mathbf{C}(\tau)$ is not feasible. They can, however, be estimated from a MD simulation trajectory of length $T$:

$$\begin{aligned} S_{ij} &= \lim_{t\to\infty} \widehat{corr}(\chi_i(\mathbf{x}), \chi_j(\mathbf{x}), \tau = 0) \\ &\approx \frac{1}{T} \sum_{t=1}^{T} \chi_j(\mathbf{x}(t))\chi_i(\mathbf{x}(t)); \end{aligned} \tag{2.33}$$

$$\begin{aligned} C_{ij} &= \lim_{t\to\infty} \widehat{corr}(\chi_i(\mathbf{x}), \chi_j(\mathbf{x}), \tau) \\ &\approx \frac{1}{T-\tau} \sum_{t=1}^{T-\tau} \chi_j(\mathbf{x}(t))\chi_i(\mathbf{x}(t+\tau)); \end{aligned} \tag{2.34}$$

where, in place of the basis functions $\{\psi_i(\mathbf{x})\}$, the corresponding co-functions $\{\chi_i(\mathbf{x})\}$ are used, which consist of the original basis functions weighted by $\pi^{-1}(\mathbf{x})$. Such a reformulation of the problem is equivalent to:

$$\begin{aligned} \pi^{-1}(\mathbf{x})\varphi(\mathbf{x}) &= \sum_{i=1}^{n} a_i \pi^{-1}(\mathbf{x})\psi_i(\mathbf{x}), \\ &= \sum_{i=1}^{n} a_i \chi_i(\mathbf{x}). \end{aligned} \tag{2.35}$$

If $\varphi(\mathbf{x})$ is the approximated propagator eigenfunction $l_i(\mathbf{x})$, $\pi^{-1}(\mathbf{x})\varphi(\mathbf{x})$ is the approximated $i^{th}$ eigenfunction of the transfer operator $r_i(\mathbf{x})$.

The VAC can therefore be used to numerically approximate the dominant eigenfunction eigenvalue pairs of the transfer operator. It is worth to notice that conventional MSM can be seen as a special case of VAC where the basis functions are chosen as step functions over the microstates.

Chapter 3

# Extensive Molecular Dynamics simulation and MSM analysis of IDPs dynamics

Intrinsically disordered peptides and proteins (IDPs) are a class of proteins which lack a fixed tridimensional structure. This does not mean that IDPs cannot assume stable structures, but rather that their state space covers a wide range of (partially) folded conformations. A specific tridimensional structure is selected in relation to specific conditions, like binding partner, environment, etc. Thanks to their flexibility, IDPs are involved in a series of functions and regulatory pathways. Misfolded IDPs can also be related to a number of diseases, such as Parkinson, Alzheimer's disease, and diabetes.

The challenge in characterizing IDPs dynamics lies in their extreme structural flexibility and in the consequent dynamic complexity. In this chapter we use MSM to characterize the dynamics of human islet amyloid polypeptide (hIAPP), a 37-residue long IDP. Human IAPP is related to type two diabetes disease, as hIAPP-rich deposits are found in over 95% of the patients [92, 93, 94]. hIAPP is known to form fibrils *in vitro*, and in particular residues 20-29 have been identified as amyloidogenic [95, 96, 97].
HIAPP can assume a variety of conformations, which can be stabilized by conformational selection upon contact with different binding partners. In this study we investigate and characterize the long-living configuration of the peptide by simulating and analyzing progressively longer hIAPP fragments around the amyloidogenic region (residues 23-27). Our study reveals a hierarchy in the dynamics. The long-living conformations identified in the model of the shorter fragments are also present the longer fragments.

**Hierarchy in the conformational ensemble of human islet amyloid polypeptide**

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

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

(Dated: 16 December 2015)

Human islet amyloid polypeptide (hIAPP) is an intrinsically disordered protein involved in glucose metabolism. The physiological $\alpha$-helical structure as well as the pathogenic amyloid fibril structure are thought to form via a conformational selection mechanism, in which the structure assembles around a conformation with some secondary structure content which is transiently sampled by the unstructured protein. We identify long-lived conformations, which might act as such nucleation points, in various sequence fragments by molecular-dynamics simulation and Markov state model analysis. More specifically we simulated the peptides FGAIL 23-27, NFGAIL 22-27, HSSNNF 18-23, ILSSTNV 26-32, HSSNNFGAIL 18-27, FGAILSSTNV 23-32, HSSNNFGAILSSTNV 18-32 and the full hIAPP 1-37 yielding a total simulation time of 57.8 $\mu$s. The conformations are matched across different fragments of the peptide sequence by comparing hydrogen bonds and backbone conformations to identify the hierarchy of interactions. A transiently formed $\alpha$-helix in FGAILS 23-28 is the likely nucleation point for the formation of the physiological structure. $\beta$-hairpins stabilized by local interactions are less frequently sampled as the chain length increases, suggesting the formation of pathogenic protofilaments do not require a preformed structure but possibly only an encounter complex between two hIAPP molecules.

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

Hierarchy in the conformational ensemble of human islet amyloid polypeptide

**Keywords:** intrinsically disordered protein, conformational dynamics, Markov state model, molecular-dynamics simulation

**Highlights:**

- Both, the physiological and the pathogenic structure of the intrinsically disordered protein hIAPP are thought to form around nucleation points with local secondary structure.
- We identify long-lived conformations in various sequence fragments by MD simulation and Markov state model analysis.
- The conformations are matched across different fragments of the peptide sequence by comparing hydrogen bonds and backbone conformations to identify the hierarchy of interactions.
- A transiently formed $\alpha$-helix in FGAILS 23-28 is the likely nucleation point for the formation of the physiological structure.
- $\beta$-hairpins stabilized by local interactions are less frequently sampled as the chain length increases, suggesting the formation of pathogenic protofilaments do not require a preformed structure but possibly only an encounter complex between two hIAPP molecules.

**Abbreviations:** hIAPP: human islet amyloid polypeptide; MSM: Markov State Model; MD: Molecular Dynamics; IDP: Intrinsically Disordered Peptide; NMI: Normalized Mutual Information; PCCA: Perron Cluster Cluster Analysis;

Hierarchy in the conformational ensemble of human islet amyloid polypeptide

**Graphical abstract:**

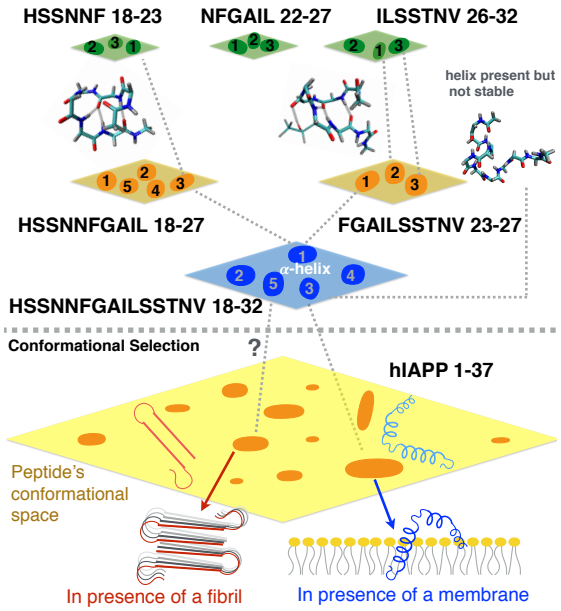


FIG. 1. Hierarchy of the dynamics

Hierarchy in the conformational ensemble of human islet amyloid polypeptide

## I. INTRODUCTION

Human islet amyloid polypeptide (hIAPP)[1], also known as amylin, is a 37-residue intrisically disordered peptide (IDP), which is synthesized in the pancreatic $\beta$-cells[2,3]. It is known to have multiple effects on glucose metabolism and homeostasis, yet the precise mechanism by which it influences these biochemical pathways has not been established[1]. Upon contact with biological membranes, hIAPP forms $\alpha$-helices[4–6]. Similar to other IDPs, hIAPP can also form $\beta$-sheets which aggregate into pathogenic amyloid deposits. In particular, these amyloid deposits are associated to type-2 diabetes[7–9]. Note that both structures, the physioligically active $\alpha$-helical structure and the pathogenic $\beta$-sheet structure, are stabilized by an interaction partner, whereas hIAPP is unstructured in solution.

The physiologically active state of hIAPP consists of two alpha helices (residues 7-17 and 21-28), connected by a kink (residues 18-20), and of a short $3_{10}$ helix (residues 33-35) (fig 2 A). The first $\alpha$-helix (residues 7-17) inserts into the membrane, while the rest of the structure is solvent-exposed[6,10].

The conformation of hIAPP within the amyloid fibril is not known precisely, and a variety of models for the protofilament have been proposed[11–16]. Residues 20-29 (SNFGAILSS) are especially critical for the formation of amyloid fibrils, which has been determined by comparing the sequences and the amyloid fibril propensities of IAPP variants of different mammals[17]. Interestingly, this fragment also contains the central $\alpha$-helix in the membrane-bound structure and therefore might act as a switch between the physiological conformation and the toxic protofilament. It has been proposed that residues 20-29 form a $\beta$-strand in solution and thus act as nucleation point for the formation of amyloid fibrils[18,19]. The model in Fig. 2 B is based on this idea and features an S-shaped conformation in which residues 22 to 27 form the central $\beta$-strand[12]. While the importance of residues 20-29 for the amyloid formation is undisputed, their role as a nucleation point is now being challenged.[13,17]. More recently proposed models were based on solid state NMR[13], X-ray crystallography[14], and on EPR experiments[16]. In all three models, the peptide chain assumes a U-shaped conformation, in which two $\beta$-strands are connected by a loop region which at least partially includes the critical region (Fig. 2 C, D). In the NMR-model, the loop region is assigned to residues 18 to 27, in the X-ray model to residues 20 to 23, and in the EPR model to residues 19 to 31. Note that in these models, the $\beta$-strands are not stabilized by intramolecular

Hierarchy in the conformational ensemble of human islet amyloid polypeptide

hydrogen bonds. Instead the fibril consists of two columns of anti-parallel hIAPP monomers packed against each other along the C-terminal $\beta$-strands. According to the NMR and the X-ray models, within the same column, the $\beta$-strands form hydrogen bonds with adjacent polypeptide chains, generating intermolecular, parallel, in-register $\beta$-sheets (Fig. 2 D.2). In the NMR-model[13] the intermolecular contacts are mediated by two serine residues in position 28 and 29, while in the X-ray model the contacts are formed by serine 29 and asparagine 31. Thus, exchanging these residues 28 and/or 29 by proline residues (as for example in rat IAPP) prevents the formation of intermolecular contacts and thus the formation of amyloid fibrils.

Additionally to the two folded structures, also the unstructured conformational ensemble has been characterized by experiment and by computational studies. These studies have shown that the conformational ensemble is not well described by a random-coil model but that the peptide transiently samples $\alpha$-helical as well as $\beta$-sheet structures[15,20–27]. Since the extraction of specific conformations from ensemble measurements is extremely difficult, molecular dynamics (MD) simulations have become an indispensable tool for the characterization of IDPs in solution[28–33]. Replica Exchange Molecular Dynamics (REMD) simulations of hIAPP, both coarse-grained[23] and atomistic, in either implicit[15,26] or explicit[22,25,27,31] solvent, have been used to explore the extent of the conformational ensemble. Three distinct conformational families of hIAPP monomer are found[15,22]: a $\beta$-sheet rich structure, a $\beta$-hairpin and a helix-coil structural family. Comparison to non-amyloidogenic sequences showed that $\beta$-hairpin conformations are only sampled by sequences which are capable of amyloid formation, whereas helical structures are transiently formed by all sequences[24–26].

Thus, a conformational selection mechanism is likely to be at work[31]: both the physiological $\alpha$-helical structure and the $\beta$-sheet structure are partially and transiently preformed in solution and can be stabilized upon contact with their binding partner (Fig 3). The residues which switch between these two structural sub-ensembles are likely to be located in the critical region from residue 20 to 29 for two reasons: ($i$) exchanges of amino acids in this regions strongly affects the amyloid propensity[17], and ($ii$) it contains the central $\alpha$-helix of the physiological conformation as well as part of the loop region in the various models of the conformation within the amyloid fibril.

Note that the complete structures are not preformed in solution, but that local secondary structure elements act as nucleation points around which the rest of the peptide can fold

Hierarchy in the conformational ensemble of human islet amyloid polypeptide

when the local secondary structure element is stabilized by an interaction partner (membrane or amyloid fibril). A prerequisite for a nucleation point is that the conformation has a long life-time, such that additional stabilizing contacts can be formed by the neighboring residues. The relative population of a conformation is of lesser importance. The energy landscape which corresponds to this mechanism is rather flat with multiple local minima representing conformations with local secondary structure elements (yellow plane in Figs. 3). Since the nucleation points only span short regions in the peptide, it is reasonable to assume that they are predominantly stabilized by interactions within these regions, and that the conformation of the nucleation point can be traced back to a conformation within a corresponding short peptide fragment. This argument justifies the relevance of studies of peptide fragments[20,34–36] (green planes in Figs. 3).

Yet, the conformational dynamics of the full hIAPP is not simply the sum of the conformational dynamics of its fragments. To investigate the hierarchy of the secondary structure elements, we perform all-atom explicit-solvent MD simulations of seven hIAPP fragments, which are centered around the critical region form residue 20 to 29. We use Markov State Models (MSM)[37–42] of the conformational dynamics to identify long-lived conformations and determine how the long-lived conformations of shorter fragments influence the stability of conformations in longer fragments. The results are compared to structural analysis of MD simulations of the full-length peptide hIAPP 1-37.

## II. RESULTS

### A. hIAPP fragments: disordered but not a random coil

All of the seven hIAPP fragments were disordered and highly flexible in the MD simulations, as expected for peptides with 15 or less residues. Fig. 4 compares the $\phi$-$\psi$-torsion angle distributions (Ramachandran plots) of corresponding residues in the hIAPP fragments (see also Fig.11 in the SI). On the level of a single amino acid residue, the backbone dynamics is essentially unrestrained with each of the residues occupying all of the canonical regions in the Ramachandran plot (Fig. 4 B). Moreover, the $\phi$-$\psi$-distributions of corresponding amino acid residues in different hIAPP fragments differ only marginally (Fig. 13 in the SI). Nonetheless, the dynamics of the hIAPP fragments cannot be described accurately by

Hierarchy in the conformational ensemble of human islet amyloid polypeptide

a random coil model, as all of the seven fragments formed several long-lived conformations. The probability of finding the peptide in one of these long-lived conformations varied from 1 to 23% between the different fragments (Fig. 6-10).

The apparent contradiction between unrestrained dynamics on the level of the single amino-acid residue and formation of long-lived conformations can be resolved by considering the properties of joint probability densities. The Ramachandran plots in Fig. 4 can be understood as the the probability density $p(\phi_i, \psi_i)$ of finding residue $i$ in a backbone conformation $\{\phi_i, \psi_i\}$. The joint probability density $p(\phi_i, \psi_i, \phi_j, \psi_j)$ represents the probability of finding residue $i$ in $\{\phi_i, \psi_i\}$ while simultaneously finding residue $j$ in $\{\phi_j, \psi_j\}$. If the two residues are fully uncorrelated, the joint probability density is given as the product of the marginal probability densities

$$p(\phi_i, \psi_i, \phi_j, \psi_j) = p(\phi_i, \psi_i) \cdot p(\phi_j, \psi_j)\,. \tag{1}$$

However, also other joint probability densities are possible which have the same marginal probability densities, $p(\phi_i, \psi_i)$ and $p(\phi_j, \psi_j)$, but cannot be constructed as their product

$$p'(\phi_i, \psi_i, \phi_j, \psi_j) \neq p(\phi_i, \psi_i) \cdot p(\phi_j, \psi_j)\,. \tag{2}$$

In these cases, the dynamics of residue $i$ and $j$ are correlated. The normalized mutual information (NMI) measures the difference between the actual joint probability density and the hypothetical uncorrelated density, and thus the degree to which the two residues are correlated (see section IV B). In contrast to the Pearson-correlation coefficient, the NMI also accounts for non-linear correlations[43].

Fig. 5 shows the pairwise NMI for the seven hIAPP fragments. Overall, the correlation between the amino-acid residues is small (NMI $< 0.1$ for most pairs) but clearly significant (NMI $> 0.01$). Throughout the chains, the strongest correlations are found between neighboring residues. Additionally, we find blocks of residues in which all residues are correlated, e.g. $S_{19}$-$F_{23}$ in HSSNNF 18-23, $L_{27}$-$N_{31}$ in ILSSTNV 26-32, $S_{19}$-$G_{24}$ in HSSNNFGAIL 18-27, $G_{24}$-$V_{31}$ in FGAILSSTNV 23-32, and $F_{23}$-$S_{28}$ in HSSNNFGAILSSTNV 18-32. The correlations indicate that conformational dynamics of these peptides deviates substantially from random coil dynamics.

Hierarchy in the conformational ensemble of human islet amyloid polypeptide

**B. The short fragments**

To identify long-lived conformations of the short fragments, we constructed Markov state models (MSMs) for the conformational dynamics of the fragments FGAIL 23-27, NFGAIL 22-27, HSSNNF 18-23, and ILSSTNV 26-32 and analyzed the eigenvectors of the MSM transition matrix using the PCCA+ algorithm[44]. The PCCA+ analysis additionally yields information on the hierarchy of the kinetic exchange procsesses between these long-lived conformations, and thus on the free-energy landscape of the peptides[40]. The timescale of the kinetic exchange processes can be calculated from eigenvalues of the MSM transition matrix (see section IV C). A similar landscape emerged for all four peptides: several states with low equilibrium population (small states) are identified which, in most cases, exhibit a specific pattern of backbone conformations in two or more consecutive amino acid residues. Each of these states is in kinetic exchange with a state with high equilibrium population and no discernible structural preference in the backbone torsion angles. There is no direct kinetic exchange between the small states. This corresponds to the situation depicted in Fig. 3: the large state represents the overall conformational ensemble (yellow plane) from which the molecule can transition into specific long-lived conformations (orange regions).

The timescales at which the long-lived conformations exchange with the overall conformational ensemble varies between tens and hundreds of nanoseconds. This confirms that the conformations are indeed stabilized compared to other conformations, and that the peptide chain is not a random coil.

Fig. 6 illustrates the long-lived conformations of the four short fragments, where we omitted the ensemble state. For FGAIL 23-27 (Fig. 6.A), only two conformations were separated from the ensemble state, each with a population of 2%. The state $C_1$ is characterized by a $\beta$-conformation in $I_{26}$ and a $L_\alpha$ conformation in $L_{27}$, whereas in $C_1$ $I_{26}$ is in $L_\alpha$ and $L_{27}$ in $\beta$. Neither of the states is stabilized by hydrogen bonds (SI Table I).

In NFGAIL 22-27 (Fig. 6.B,), we find two states, $C_2$ and $C_3$, which have a low population of 1% each and which are not stabilized by any hydrogen bonds. In state $C_1$ (5% equilibrium population), $F_{23}$ and $G_{24}$ are locked in the $L_\alpha$ conformation whereas the other residues are free to transition between the $\alpha$ and the $\beta$ conformation and, to some extent, the $L_\alpha$ conformation. The resulting conformation is almost a $\beta$-hairpin which is stabilized by hydrogen bonds from $N_{22}$ to $A_{25}$ and $I_{26}$ (SI Table II). The hydrogen bond between the

Hierarchy in the conformational ensemble of human islet amyloid polypeptide

carbonyl oxygen of $N_{22}$ and the backbone amide hydrogen of $A_{25}$ is the most populated one (formed in 13.83% of all frames in the $C_1$) and can be classified as a $i \leftarrow i+3$ backbone hydrogen bond, where $i$ is a residue number and the arrow indicates the direction of the hydrogen bond form hydrogen donor to acceptor. Analogously, the hydrogen bond between $N_{22}$ and $I_{26}$ (population within $C_1$: 10.61%) is classified as $i \leftarrow i+4$. Fig. 6.H shows that both hydrogen bonds can be formed simultaneously. Additionally, the side chain of $N_{22}$ acts as a hydrogen bond donor to the backbone carbonyl groups of $A_{25}$ and $I_{26}$ (5.26% and 6.46%, respectively).

In all three long-lived conformations of HSSNNF 18-23 (Fig. 6.C), $S_{20}$ is restricted to the $L_\alpha$-conformation (with some population in the $\beta$-conformation in $C_2$). Additionally either or both of the neighboring residues show an increased population in the $L_\alpha$-conformation. $C_1$ has the most clearly defined backbone structure. The predominant sequence of backbone conformations is $\{\alpha, \beta, L_\alpha, L_\alpha, \alpha, \beta\}$. This hairpin-conformation is stabilized by a the following sequence of backbone hydrogen bonds (Fig. 6.F and 6.I): carbonyl oyxgen of $S_{19}$ to amide hydrogen of $N_{22}$ (21.07%, $i \leftarrow i+3$), carbonyl oyxgen of $S_{19}$ to amide hydrogen of $F_{23}$ (19.99%, $i \leftarrow i+4$), and additionally amide hydrogen of $S_{19}$ to carbonyl oxygen of $F_{23}$ (11.44%, $i \rightarrow i+4$) (SI Table III). This is a class 3 $\beta$-hairpin[45] with three residues between the doubly hydrogen bonded residues $S_{19}$ and $F_{23}$, which is additionally stabilized the $i \leftarrow i+3$ hydrogen bond between $N_{22}$ and $S_{19}$. Although $C_2$ and $C_3$ have larger populations than $C_1$, their conformational sub-ensemble exhibits fewer and less populated hydrogen bonds. Neither of the two states is substantially stabilized by specific interactions.

For ILSSTNV 26-32 (Fig. 6.D), three long-lived conformations could be identified, $C_3$ covers 23% of the entire conformational ensemble and is a class 3 $\beta$-hairpin with three residues between the doubly hydrogen bonded residues $L_{27}$ and $N_{31}$ ($i \leftarrow i+4$: 28.65%, $i \rightarrow i+4$: 38.13%). The conformation is additionally stabilized by a $i \leftarrow i+3$ hydrogen bond from the amide hydrogen of $T_{30}$ to the carbonyl oxygen of $L_{27}$ and by a hydrogen bond from the side chain hydroxyl group of $T_{30}$ to the carbonyl oxygen of $S_{28}$ (SI Table IV). This hydrogen bond pattern is similar to $C_1$ in HSSNNF 18-23 and hence induces an analogous backbone conformation in residues $L_{27}$ to $N_{31}$: $\{\beta, L_\alpha, L_\alpha, \alpha, \beta\}$. By contrast, $C_1$ an $C_2$ have a low population. $C_1$ can be classified as a class 4 $\beta$-hairpin with four residues between the doubly hydrogen bonded residues $I_{26}$ and $N_{31}$ (hydrogen bonds $i \leftarrow i+5$ and $i \rightarrow i+5$). However, the most prominent hydrogen bond is formed between the side chain

Hierarchy in the conformational ensemble of human islet amyloid polypeptide

hydroxyl group of $T_{30}$ and the carbonyl group of $L_{27}$ (52.99%) which is accompanied by a hydrogen bond from the amide hydrogen of $T_{30}$ to the carbonyl group of $L_{27}$ ($C_1$ 10.07%). Both of these hydrogen bonds are also formed in $C_2$ (23.42% and 13.52%, respectively), but the interactions between $I_{26}$ and $N_{31}$ are missing.

To summarize, in three out of four peptide fragments we find a class 3 hairpin conformation ($C_1$ in NFGAIL 22-27, $C_1$ in HSSNNF 18-23, and $C_3$ in ILSSTNV 26-32). These conformations are stabilized by hydrogen bonds of type $i \leftarrow i{+}3$ and $i \leftarrow i{+}4$, where residues i+1 and 1+2 are in the $L_\alpha$ backbone conformation. In HSSNNF 18-23 and ILSSTNV 26-32, the conformation is additionally stabilized by a hydrogen bond of type $i \rightarrow i+4$ restraining the backbone conformation of residues $i$ to $i+4$ to $\{\beta, L_\alpha, L_\alpha, \alpha\}$. In ILSSTNV 26-32 we additionally find a class 4 $\beta$-hairpin.

### C. HSSNNFGAIL 18-27

The MSM analysis of the fragment HSSNNFGAIL 18-27 yields five long-lived conformations (Fig. 7). Remarkably in all five states, the formation of a specific backbone structure is limited to residues $H_{18}$ to $F_{23}$, whereas residues $G_{24}$ to $L_{27}$ can transition freely between the canonical backbone conformations. Only in state $C_2$, residue $I_{26}$ is restricted to the $L_\alpha$-region of the Ramachandran plot, whereas residue $A_{26}$ is restricted to the $\alpha/\beta$-region. Possibly, the flexible backbone torsion angles of $G_{24}$ prevent structure formation in the C-terminal part of the peptide.

$C_3$ has a very distinct backbone conformation in residues $H_{18}$ to $F_{23}$ ($\{\alpha, \beta, L_\alpha, L_\alpha, \alpha, \beta\}$) and can be classified as a class 3 $\beta$-hairpin with three residues between the doubly hydrogen bound residues $S_{19}$ and $F_{23}$ ($i \leftarrow i+4$: 33.36%, $i \rightarrow i+4$: 25.11%). The hydrogen bond between the amide hydrogen of $N_{22}$ and the carbonyl oxygen of $S_{19}$ ($i \leftarrow i+3$: 26.26%) additionally stabilizes the conformation. With these characteristics, $C_3$ can be related to $C_1$ in HSSNNF 18-23. $C_3$ in HSSNNFGAIL 18-27 is additionally stabilized by hydrogen bonds between $H_{18}$ and $G_{24}$ (5.85%) and between $H_{18}$ and $A_{25}$ ((5.33%), which cannot be formed in HSSNNF 18-23. We also observe a hydrogen bond between the side chain of $N_{22}$ and the carbonyl oxygen of $S_{20}$ (8.17%), which is not present in state $C_1$ of HSSNNF 18-23.

Interestingly, states $C_1$, $C_2$, $C_4$, and $C_5$ are not stabilized by hydrogen bonds. The only hydrogen bond with a relative population larger than 10% within the sub-ensemble of the

Hierarchy in the conformational ensemble of human islet amyloid polypeptide

respective state is found in state $C_2$ between the side chain of $S_{20}$ and the carbonyl oxygen of $A_{25}$. States $C_4$ and $C_5$ have corresponding states in HSSNNF 18-23. In state $C_4$, the predominant sequence of backbone conformations in residues $H_{18}$ to $N_{21}$ is $\{\beta, L_\alpha, L_\alpha, L_\alpha\}$, whereas residues $N_{22}$ and $F_{23}$ are flexible. A similar pattern is found in state $C_2$ of HSSNNF 18-23 (Fig. 6.C). Likewise the pattern $\{\beta/L_\alpha, \beta/L_\alpha, L_\alpha\}$ in residues $H_{18}$ to $S_{20}$ can be matched with state $C_3$ of HSSNNF 18-23 (Fig. 6.C). State $C_1$ and $C_2$ seem to be complements. In state $C_1$, residues $S_{19}$ to $F_{23}$ show an increased population in the $\beta/L_\alpha$-regions of the Ramachandran plot, whereas in state $C_2$ the same residues show an increased population of the $\alpha$-region. Additionally $I_{26}$ is restrained to the $\alpha/\beta$ region in state $C_1$, whereas in state $C_2$ it is restrained to the $L_\alpha$-region.

### D. FGAILSSTNV 18-27

FGAILSSTNV 23-32 has three long-lived conformations (Fig. 8 and 9). $C_1$ is a class 3 $\beta$-hairpin with three residues between the doubly hydrogen bound residues $L_{27}$ and $N_{31}$ ($i \leftarrow i+4$: 26.26%, $i \rightarrow i+4$: 17.79%). The conformation is additionally stabilized by a $i \leftarrow i+3$ backbone hydrogen bond between $L_{27}$ and $T_{30}$ (17.99%) and by a hydrogen bond from the hydroxyl group of the threonine side chain to the carbonyl group of $S_{28}$ (23.59%). $C_1$ is hence completely analogous to $C_3$ of ILSSTNV 26-32. Both conformations have the same backbone conformation in residues $L_{27}$ to $N_{31}$: $\{\beta, L_\alpha, L_\alpha, \alpha, \beta\}$. $I_{26}$ is flexible in ILSSTNV 26-32, but restrained to the $\alpha$-conformation FGAILSSTNV 23-32.

In $C_2$ of FGAILSSTNV 23-32 is a class 4 $\beta$-hairpin with four residues between the doubly hydrogen bound residues $I_{26}$ and $N_{31}$ ($i \leftarrow i+5$:18.57%, $i \rightarrow i+5$: 33.72%). It can be related to $C_1$ in ILSSTNV 26-33. As in this precursor, $C_2$ of FGAILSSTNV 23-32 is additionally stabilized by a very strong hydrogen bond between the side chain hydroxy group of $T_{30}$ and the carbonyl oxygen of $L_{27}$ (60.67%). The hydrogen bond from the amide hydrogen of $A_{25}$ to the carbonyl oxygen of $N_{31}$ is only present in the decamer because $A_{25}$ is absent in ILSSTNV 26-33. $C_2$ has the same backbone conformation in residues $I_{26}$ to $T_{30}$ as its precursor: $\{\beta, L_\alpha, L_\alpha, L_\alpha, \alpha\}$.

$C_3$ is again similar to a class 3 $\beta$-hairpin with three residues in the loop region. However the hairpin is shifted compared to $C_1$. The doubly hydrogen bound residues are $I_{26}$ and $T_{30}$ ($i \leftarrow i+4$: 23.89%, $i \rightarrow i+4$: 26.31%). Similar to the other class 3 hairpins, we also

Hierarchy in the conformational ensemble of human islet amyloid polypeptide

observe an $i \leftarrow i+3$ hydrogen bond between $I_{26}$ and $S_{29}$ (23.27%). The conformation is additionally stabilized by a hydrogen bond from the side chain hydroxy group of $S_{29}$ to the carbonyl oxygen of $L_{27}$ (37.92%) and by a hydrogen bond between the amide hydrogen of $A_{25}$ and the side chain amide group of $N_{31}$ (12.19%). This latter interaction might be the reason why $C_3$ does not have precursor in ILSSTNV 26-32. The backbone conformation in residues $A_{25}$ to $T_{30}$ is similar to $C_1$ but shifted by a residue: $\{\alpha, \beta, L_\alpha, L_\alpha, \alpha, \beta\}$.

The DSSP analysis of the trajectories (Fig. 6 and 8 in the SI) shows that, additionally to the $\beta$-hairpin structures identified by the MSM, the peptide can also assume a helical conformation, which involves residues $G_{24}$ to $L_{27}$ and occasionally extends to residue $N_{31}$ (Fig. 9). The lifetime of the conformation is, however, much shorter than those of $\beta$-hairpins. It is consequently not identified by the MSM analysis as a long-lived conformation.

### E. HSSNNFGAILSSTNV 18-32

HSSNNFGAILSSTNV 18-32 is a 15-residues long peptide with the amyloidogenic domain (FGAIL 23-27) in the center of the sequence. The chain length allows for the formation of multiple secondary structure elements and five long-lived conformations were identified by the MSM analysis (Fig. 10 and 11). $C_1$ consists of a $\beta$-hairpin-like structure element and a $\beta$-hairpin. The $\beta$-hairpin involves residues $I_{26}$ to $N_{31}$ which assume the backbone conformation $\{\alpha, \beta, L_\alpha, L_\alpha, \alpha\beta\}$. Its precursor is $C_1$ in FGAILSSTNV 18-27. That is, it is a class 3 $\beta$-hairpin with three residues between the doubly hydrogen bound residues $L_{27}$ and $N_{31}$ ($i \leftarrow i+4$: 28.32%, $i \rightarrow i+4$: 18.62%). As in the precursor, we additionally observe a $i \leftarrow i+3$ hydrogen bond between $L_{27}$ and $T_{30}$ (11.79%) and a hydrogen bond between the side chain hydroxy group of $T_{30}$ and $S_{28}$ (22.23 %). One additional $i \leftarrow i+5$ hydrogen bond is formed between $I_{26}$ and $N_{31}$ (33.58 %). The $\beta$-hairpin-like structure stretches from $H_{18}$ to $F_{23}$ and can be related to $C_3$ in HSSNNFGAIL 18-27. The $i \leftarrow i+4$ and $i \leftarrow i+3$ hydrogen bonds between the carbonyl oxygen of $S_{19}$ and the amide hydrogens of $F_{23}$ and $N_{22}$, respectively, are present 25.40% and 26.61% of the frames. But the $i \rightarrow i+4$ between $S_{19}$ and $N_{22}$ is missing because its formation is sterically prevented by a hydrogen bond from the side-chain hydroxy group of $S_{20}$ to the carbonyl oxygen of $H_{18}$ (13.38%).

In $C_2$, residues $H_{18}$ to $I_{26}$ assume a well-defined backbone conformation:
$\{\beta, \beta, L_\alpha, \alpha, \alpha, \beta, \text{G: lower right corner}, \alpha, \beta\}$. Residues $L_{27}$ to $V_{32}$ are flexible (Fig. 10). The

Hierarchy in the conformational ensemble of human islet amyloid polypeptide

conformation can however not be classified as one of the common secondary elements. It is stabilized by a $i \leftarrow i+7$ backbone hydrogen bond between $S_{20}$ and $L_{27}$ (40.58%) and by a $i \leftarrow i+4$ backbone hydrogen bond between $H_{18}$ and $N_{22}$ (27.38%). Additionally, the side-chain backbone hydrogen bonds from the backbone amide hydrogen of $H_{18}$ to the side chain amide oxygen of $N_{22}$ (13.71%) and from the side-chain hydroxy group of $S_{28}$ to the carbonyl oxygen of $S_{20}$ (12.74 %) are formed. This conformation does not have any precursors in the shorter fragments because none of the fragments is long enough to form the crucial hydrogen bond between $S_{20}$ and $L_{27}$.

$C_3$ is an $\alpha$-helical conformation, which stretches from $F_{23}$ to $S_{28}$. $S_{29}$ is in the $L_\alpha$ conformation. $C_3$ has an equilibrium population of 9%, which is in line with the DSSP[46] analysis of the trajectories (see Fig. 11, which shows that residues 23 to 28 form an $\alpha$-helix in 10% of the trajectory frames). The probability that the DSSP analysis assigns the $\alpha$-helical state to an individual residue in this region is roughly 30%. Correspondingly, the probability of finding these residues in a coil state is significantly lower than in any of the fragments discussed so far (Fig. 13). The $\alpha$-helix in $C_3$ is stabilized by a canonical $i \leftarrow i+4$ backbone hydrogen bond between $F_{23}$ and $L_{27}$ (26.59 %). The next canonical $i \leftarrow i+4$ hydrogen bond between $G_{24}$ and $S_{28}$ is also formed, albeit with a much lower probability: 8.88 %. Instead, a competing hydrogen bond between the side chain hydroxy group of $S_{28}$ and the carbonyl oxygen of $G_{24}$ is formed with 29.83% probability. The precursor of this conformation is the transient $\alpha$-helical conformation in FGAILSSTNV 18-27.

In $C_4$, only residues $I_{26}$ and $L_{27}$ have a distinct backbone conformation $\{\beta/L_\alpha, L_\alpha\}$, which is stabilized by a hydrogen bond from the side-chain hydroxy group of $S_{28}$ to the carbonyl oxygen of $I_{26}$(21.22 %). Additionally, there is a hydrogen bond between the side chains of $T_{30}$ and $N_{22}$(11.63%). Similarly, $C_5$ only has a distinct backbone conformation in residues $S_{19}$ to $N_{21}$: $\{\alpha/L_\alpha, L_\alpha, L_\alpha\}$, which is stabilized by a $i \leftarrow i+3$ backbone hydrogen bond between $N_{22}$ and $S_{19}$ (12.04%).

### F. hIAPP 1-37

For the full hIAPP 1-37 peptide, we obtained 10.8 $\mu$s of aggregated simulation time, which is sufficient to explore large part of the conformational space of hIAPP but not to construct a MSM and reliably identify long-lived conformations. The equilibrium $\{\phi-\psi\}$-distribution

Hierarchy in the conformational ensemble of human islet amyloid polypeptide

of all residues is shown in Fig. 13 in the SI. As in the fragments, the backbone dynamics is essentially unrestricted on the level of the individual residues. All residues, with exception of $I_{26}$ and $L_{16}$, explore the full Ramachandran plot. In $I_{26}$ and $L_{16}$, the $L_\alpha$ conformation is not visited. The mutual information (Fig. 12) exhibits two blocks. One block stretches from $C_2$ to $N_{21}$, the other encompasses the amyloidogenic region from $F_{23}$ to $S_{29}$. This latter block is also found in the mutual information plot of HSSNNFGAILSSTNV 18-32 (Fig. 5). The high mutual information among the residues $C_2$ to $C_7$ is due to a cystein bond between these two residues which strongly restricts the conformational flexibility of the residues 2 to 7.

The DSSP analysis of the trajectory confirms that hIAPP 1-37 is unstructured in solution but transiently visits a variety of secondary structures (Fig 12 and Fig. 12 in the SI). The most prominent secondary structure element is an $\alpha$-helix in the amyloidogenic region $F_{23}$ to $L_{27}$ which is visited in 42.4% of the total simulation time and in 13 out of 16 of the independent simulated runs. The probability that an individual residue within this region is found in the $\alpha$-helical state (as defined by the DSSP-analysis) ranges from 30 to 50%. This is particularly striking since all other residues are found in an unstructure coil state with 80% probability. The precursor of the $\alpha$-helix in residues 23 to 27 are quite clearly the conformation $C_3$ in HSSNNFGAILSSTNV 18-32 and the transiently visited $\alpha$-helix in FGAILSSTNV 23-32. Note however, that the remaining residues are not necessarily fully flexible when the $\alpha$-helix is formed, rather they assume a variety of distinct conformations. Compare for example Fig. 12.B and 12.D.

The DSSP analysis also shows that hIAPP forms several transient $\beta$-hairpin structures. Fig 12 A, C and D show three examples. Fig 12 A shows residues 6 or 7, 17 and 30 forming distant $\beta$-bridges, while the intermediate segments form $\alpha$-helices and turns. In C, three $\beta$-regions can be identified: residues 30-35 and 7-11, building simultaneously $\beta$-sheets and occasionally $\beta$-bridges, and residues 22-24 assuming transient $\beta$-structures intertwined by disordered configurations. Panel D presents residues 34-35 and 15-16 forming $\beta$-contacts, whilst residues 23-27 are in a $\alpha$-helix. However none of these $\beta$-hairpins can be related to the $\beta$-hairpins formed by the fragments.

Hierarchy in the conformational ensemble of human islet amyloid polypeptide

## III. DISCUSSION

The most critical caveat of MD studies of IDPs is the validity of the force field for these systems. While classical force fields have improved considerably in recent years with some force fields reaching experimental accuracy for folded proteins and for short peptides[47], it is less clear how well current force fields are suited to describe the conformational dynamics of disordered proteins. The balance between $\alpha$-helical structures and $\beta$-sheet structures can vary drastically between force fields[48]. Current force fields are also known to generate conformational ensembles which are too compact compared to experiment, which could be traced back to underestimated dispersion forces in common water models[49]. Moreover, the predicted dynamic properties (kinetic processes, implied timescales) of short peptides varies across force fields[50]. Nonetheless, several studies report good agreement with experimental for specific force fields[51]. AMBER ff99sb-ILDN, the force field we used in this study, has not been included in a systematic force field comparison for disordered proteins. However it performed very good in force field benchmarks for folded proteins[47], and the closely related force field AMBER ff99sb*-ILDN achieves a balance between $\alpha$-helical structures and $\beta$-structures in rat and human hIAPP which is in agreement with experiment[48]. Overall, we expect that the conformation predicted by our simulations as well as the relative order of their equilibrium populations and the relative order of the implied timescales for the kinetic exchange with the ensemble state are correct. The absolute values of the equilibrium populations and of the implied timescales are less reliable.

Our analysis reveals a hierarchy in the long-lived conformations from shorter to longer peptides (Fig. 14). In particular, the double-$\beta$-hairpin structure in $C_1$ of HSSNNFGAILSSTNV 18-32 can be related to precursors in the corresponding decamers HSSNNFGAIL 18-27 and FGAILSSTNV 23-32 and in shorter sequence fragments by matching hydrogen bond patterns as well as backbone conformations. The conformational ensemble of hIAPP 1-37, as simulated so far, does not contain a structure which can directly be matched with $C_1$ of HSSNNFGAILSSTNV 18-32. This might have two reasons. Either the corresponding conformation exists but has not yet been discovered by the simulation. Or the corresponding conformation is destabilized in the full peptide. $C_1$ of HSSNNFGAILSSTNV 18-32 is a nested conformation, which might be prone to sterical clashes if the peptide chain is prolonged. On the other hand, C- and N-termini of the 15-mer point towards different

sides of the peptide backbone. It is not obvious how precisely a sterical clash would arise.

Up to the 15-mer HSSNNFGAILSSTNV 18-32, the hierarchy of long-lived structures covers mostly $\beta$-hairpin conformations. This is not surprising since the backbone hydrogen bonds of $\beta$-hairpins are more stable than those of $\alpha$-helices and less easily attacked by surrounding water molecules. Hence, $\beta$-hairpins readily form in solution whereas $\alpha$-helices tend to be unstable in solution. Yet, the relative frequency of $\alpha$-helical structures increases as we prolong the sequence. The shortest sequence which exhibits $\alpha$-helical conformations is FGAILSSTNV 23-32 as demonstrated by the DSSP analysis (Figs. 6 and 8 in the SI) and by the slightly elevated $\alpha$-helix propensities of residues FGAI 23-26 (Fig. 13). However, not a single $\alpha$-helical conformation is sampled but rather a set of closely related conformations with some $\alpha$-helical content. Each of these conformations has a rather short life time and therefore no $\alpha$-helical structure is identified as a long-lived conformation. In the 15-mer, an $\alpha$-helix spanning from $F_{23}$ to $S_{28}$ is identified as long-lived conformation $C_3$ with an equilibrium population of 9%, which can be matched with the transiently occuring $\alpha$-helical conformations in FGAILSSTNV 23-32. Similar to FGAILSSTNV 23-32, the DSSP analysis of the 15-mer also shows several related $\alpha$-helical structures with short life times, such that the $\alpha$-helix propensity per residue reaches 30 to 50% in FGAIL 23-27 (Fig. 13). Note that the $\alpha$-helix in $C_3$ is stabilized by two canonical $\alpha$-helix hydrogen bonds and an additional hydrogen bond form the side chain of $S_{28}$ to carbonyl of $G_{24}$, thus compensating the entropy loss in $G_{24}$ by an additional enthalpic interaction. $C_3$ of HSSNNFGAILSSTNV 18-32 is a precursor to $\alpha$-helical structures sampled by the full peptide hIAPP 1-37. Since the $\alpha$-helix in FGAIL 23-27 directly matches the central $\alpha$-helix of the membrane-bound conformation of hIAPP (Fig. 2), this conformation is a likely candidate for a nucleation point of the physiological conformation which is stabilized by the interaction with a membrane.

To summarize, we identified a (possible) nucleation point for the physiological structure of hIAPP and trace its occurrence back to the conformational ensemble of a ten-residue fragment. We also find a hierarchy in the $\beta$-hairpin structures of different fragment lengths. However, the relative frequency with which the particularly conserved $\beta$-hairpins are sampled decreases with increasing chain length. In particular, no analogue of the double-$\beta$-hairpin structure in HSSNNFGAILSSTNV 18-32 has been sampled by the simulation of hIAPP 1-37. Instead the ensemble exhibits a range of $\beta$-hairpin structures with rather long loop regions. While it might be impossible to fully prevent the transient formation of $\beta$-hairpin

Hierarchy in the conformational ensemble of human islet amyloid polypeptide

structures in intrinsically disordered proteins, the formation of an extended $\beta$-hairpin which originates from a set of local stabilizing interaction is clearly hindered in hIAPP. Thus, the formation of pathological protofilaments does not seem to require an preformed structure but only an encounter complex between two hIAPP molecules. This is in line with the finding in Ref. 31 that many of the long-lived conformations in hIAPP exhibit extended hydrophobic interactions which might promote aggregation. The result might also explain, why the amyloid fibrils of hIAPP consist of $\beta$-strands which are stabilized by intermolecular hydrogen bonds rather than of a stack intramolecularly stabilized $\beta$-hairpins.

## IV. MATERIALS AND METHODS

### A. MD simulations

| system | # independent simulations | total simulation time |
|---|---|---|
| FGAIL 23-27 | 8 | 5 $\mu$s |
| NFGAIL 22-27 | 10 | 5 $\mu$s |
| HSSNNF 18-22 | 11 | 5 $\mu$s |
| ILSSTNV 26-32 | 5 | 5 $\mu$s |
| HSSNNFGAIL 18-27 | 9 | 4.7 $\mu$s |
| FGAILSSTNV 23-32 | 10 | 9.7 $\mu$s |
| | 30 | 3 $\mu$s |
| HSSNNFGAILSSTNV 18-32 | 10 | 9.6 $\mu$s |
| hIAPP-1-37 | 16 | 10.8 $\mu$s |

TABLE I. Number of independent replicas and total simulation time for each set up.

We performed all-atom molecular dynamics simulations of fragments of human islet amyloid polypeptide (hIAPP). Specifically we simulated the sequences : FGAIL (residues 23-27), NFGAIL (residues 22-27), HSSNNF (residues 18-23), ILSSTNV (residue 26-32), HSSNNFGAIL (residues 18-27), FGAILSSTNV (residues 23-32), HSSNNFGAILSSTNV (residues 18-32). The fragments were acetylated at the N-terminus and methylated at the C-terminus. We also simulated the full 37-residue hIAPP, where the C-terminal of the peptide was capped with an -$NH_2$ group, and a disulfide bond was present between cysteine residues $C_2$ and $C_7$.

Hierarchy in the conformational ensemble of human islet amyloid polypeptide

Starting structures where obtained from an NMR structure of hIAPP in an membrane environment (PDB ID: 2L86)[6]. The simulations were performed with the AMBER ff99SB-ILDN[52] force field in explicit water (TIP3P[53] water model), using the GROMACS simulation package[54] (versions 4.4.5 and 5.0.2). The NVT ensemble was applied, where the V-Rescale thermostat[55] 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, 5 to 16 structures for each system were selected randomly from the trajectory and used as starting conformations for independent simulation runs, yielding a total simulation time of more than 4.5 $\mu$s per system (Tab I). 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[56] 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[57] 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.

In the first 9.7 $\mu$s simulations of FGAILSSTNV 23-32, we encountered a sink state, i.e. a conformation which is entered but not left during the course of remaining trajectory. A sink state prevents the construction of a converged MSM. We therefore initiated 30 further simulations of 100 ps each from this conformation. 27 runs were started from the conformation where AILSSTN 25-31 are in $\{\alpha, \beta, \mathrm{L}_\alpha, \mathrm{L}_\alpha, \mathrm{L}_\alpha, \alpha, \beta\}$ respectively. The remaining 3 runs were started from equilibrated structures originated from the conformation in which ILSST 26-30 are in $\{\beta, \mathrm{L}_\alpha, \mathrm{L}_\alpha, \alpha, \beta\}$ conformation. Including these additional 3 $\mu$s simulation data into the MSM analysis yielded a converged model, in which the (former) sink state is identified as long-lived conformation $\mathrm{C}_3$ (Fig. 8).

### B. General analysis

Secondary structure analysis of the system was performed using the "dictionary of protein secondary structure" (DSSP)[46] definition - a standard method for secondary structure assignmen. The DSSP assignments were computed using the built-in function of MDTraj python-based software package[58].

To study the mutual dependencies of the backbone conformations of pairs of residues $i$ and

Hierarchy in the conformational ensemble of human islet amyloid polypeptide

$j$, we computed the normalized mutual information of the $\phi$-$\psi$-torsion angle distributions of these residues[59]. The backbone conformation of a residue $i$ is represented by a discrete random variable $x$, which is obtained by discretizing the $\phi$-$\psi$ torsion angle coordinates into a regular grid of $N_S = 36 \times 36 = 1296$ bins (36 bins per torsion angle, bin-width 10°). The probability distribution $p(x)$ is obtained as a normalized histogram in this state space from $\phi$-$\psi$-time series of residue $i$. The probability distribution of residue $j$ is denoted $p(y)$ and is obtained analogously. The joint probability density $p_x$ is the probability of finding residue one in state $x$, given that residue two is in state $y$. The normalized mutual information is a measure for the correlation between two residues.

The normalized mutual information of two discrete variables $x$ and $y$ (with $x, y \in 1, 2, ...N_s$) is defined as:

$$NMI(x,y) = \frac{1}{\min(H_x, H_y)} \sum_{x=1}^{N_s} \sum_{y=1}^{N_s} p(x,y) \log \left[ \frac{p(x,y)}{p(x)p(y)} \right] . \tag{3}$$

$p(x,y)$ is the joint probability distribution of the two variables, and $p(x)$ and $p(y)$ are the corresponding marginal probability distributions

$$\begin{aligned} p(x) &= \sum_{y=1}^{N_s} p(x,y) \\ p(y) &= \sum_{x=1}^{N_s} p(x,y) \, . \end{aligned} \tag{4}$$

$H_x$ and $H_y$ are the informational entropies associated to the marginal distributions

$$\begin{aligned} H_x &= -\sum_{x=1}^{N_s} p(x) \log \left[ p(x) \right] \\ H_y &= -\sum_{y=1}^{N_s} p(y) \log \left[ p(y) \right] \, . \end{aligned} \tag{5}$$

When estimated from MD data, the probability distributions are subject to statistical uncertainty. This noise induces are residual value of $NMI > 0$ even if two residues are fully uncorrelated. To determine this residual value and hence the significance level of the NMI-analysis, we estimated the NMI of two residues, G and A, of two independent simulations. Including a safety margin, the significance level was set to $NMI = 0.01$, i.e. any two residues with a $NMI < 0.01$ were considered fully uncorrelated.

Hierarchy in the conformational ensemble of human islet amyloid polypeptide

### C. Markov state models - overview

Markov state models (MSMs) approximate the deterministic dynamics in the full phase space of the molecule and the surrounding water molecules by a stochastic process $x_t$ which switches between $N$ discrete, non-overlapping conformational states. These so-called microstates are usually defined in terms of only a few internal coordinates. Thus, a MSM is an approach to reduce the dimensionality of the complex high-dimensional dynamics, such that the relevant features of the dynamics become humanly understandable. In this section we present the salient points of MSMs, which are covered in more details elsewhere[38,40,60,61].

Let $\Omega = \{S_1, S_2, .... S_N\}$ be the set of microstates on which the MSM is constructed (i.e. the state space of the MSM). In the model, one assumes that the dynamics in this state space are ergodic, and Markovian. Ergodicity implies that any microstate $S_i$ can be reached from any other microstate $S_j$. Markovianity implies that the probability of finding the molecule in state $S_j$ at time $t + \tau$ only depends on the state $S_i$ in which the system has been at time $t$. That is, the dynamics are determined by conditional probabilities

$$T_{ij}(\tau) = \mathbb{P}(x_{t+\tau} = j | x_t = i)\,. \tag{6}$$

Note that the transition probability $T_{ij}(\tau)$ does not imply that the molecule directly jumps from $S_i$ to $S_j$. It rather represents a probability which is calculated over all possible paths of length $\tau$ which connect the two states. Arranging these transition probabilities in a $N \times N$ matrix yields the transition matrix $\mathbf{T}(\tau)$. The lag time $\tau$ is a parameter of the model. The discretization, i.e. the choice of the microstates, and the transition matrix constitute the gist of the MSM.

The transition probabilities $T_{ij}(\tau)$ can be estimated directly from MD data, by counting the number of transitions between states

$$\hat{T}_{ij}(\tau) = \frac{C_{ij}(\tau)}{C_i(\tau)} = \frac{C_{ij}(\tau)}{\sum_{j=1}^{N} C_{ij}(\tau)}\,. \tag{7}$$

$C_{ij}(\tau)$ is the number of transitions $S_i \to S_j$ within time $\tau$, i.e. the number of all trajectory fragments of length $\tau$ which originate in $S_i$ and end in $S_j$. $C_i(\tau)$ is the number of transition from state $S_i$ to any other states, i.e. the number of all trajectory fragments which originate in $S_i$. In equilibrium MD simulation (i.e. no external forces) the dynamics are reversible, which means that, in the limit of infinite sampling, the number of transitions $S_i \to S_j$ is

equal to the number of transitions in the opposite direction $S_j \rightarrow S_i$,

$$C_{ij}(\tau) = C_{ji}(\tau)\,. \tag{8}$$

Reversibility can be enforced during the estimation of the transition matrix.

If the dynamics are ergodic and reversible, the transition matrix is decomposable in a complete set of real valued eigenvectors $\mathbf{r}_i$ and associated eigenvalues $\lambda_i(\tau)$.

$$\mathbf{T}(\tau)\mathbf{r}_i = \lambda_i(\tau)\mathbf{r}_i\,, \tag{9}$$

The eigenvectors and eigenvalues contain information on the conformational exchange processes in the system. $\mathbf{T}(\tau)$ has a bound eigenvalues spectrum

$$\lambda_1 = 1 \geq |\lambda_2(\tau)| \geq |\lambda_3(\tau)|... \tag{10}$$

where $\lambda_1 = 1$ is the largest eigenvalue by absolute value. It always exists, and is unique (non-degenerate) if the dynamics are ergodic. The eigenvector $\mathbf{r_1}$ associated to $\lambda_1 = 1$ is the stationary process. Eigenvectors associated to eigenvalues close to 1 (dominant eigenvalues) represent the slow kinetic processes of the system. They can be interpreted as mediating the conformational exchange between different long-lived conformations. The equilibration time $t_i$ of an exchange process $\mathbf{r}_i$ is connected to the corresponding eigenvalue $\lambda_i(\tau)$

$$t_i = -\frac{\tau}{\ln(\lambda_i(\tau))} \tag{11}$$

$t_i$ is called implied timescale or relaxation time. Note that, if the dynamics are indeed Markovian, the implied timescales do not vary with the lag time $\tau$ of the model. This can be used to test whether an MSM estimated from MD data is Markovian. A PCCA+ analysis (Perron Cluster Cluster Analysis +)[62] of the first $M$ eigenvectors yields fuzzy memberships to $M$ long-lived conformations for each microstate. Using these memberships, the conformational space can be dissected into long-lived conformations and free-energy barriers between them. A visual example of the PCCA+ assignment of microstates to long-lived macrostates is shown in Fig. 2 in the SI.

### D. Markov state models - construction and analysis

The microstates of all Markov state models were defined in terms of the $\phi$- and $\psi$-backbone torsion angles, which are known to be suitable coordinates for peptide systems[39,50,63,64].

Hierarchy in the conformational ensemble of human islet amyloid polypeptide

Time series of the backbone torsion angles of all systems were extracted from the simulated trajectories using the GROMACS command `g_rama`. Two-dimensional histograms of the $\{\phi-\psi\}$-distribution of each residue (Ramachandran plots) were constructed from the torsion angle time series (bin-widt $1^\circ \times 1^\circ$) and are shown in fig. 4 and in Fig. 13 in the SI.

For the construction of the microstates of FGAIL 23-27, NFGAIL 22-27, HSSNNF 18-23, and ILSSTNV 26-32, the $\{\phi-\psi\}$-space of each residue was discretized into three or five bins (exception six for glycine), such that each bins captures a maximum in the Ramachandran plot of the respective residue (See Fig. 1 A, C, B in the SI respectively). The conformation of a particular amino acid residue in the peptide chain is then represented by the corresponding bin index. The conformation of the peptide chain is represented by the combination of the bin indices of all amino acids. Each possible combination of bin indices is a microstate of the MSMs. This approach yields 486 possible microstates for FGAIL 23-27, 6144 for NFGAIL 22-27, 729 for HSSNNF 18-23 and 2187 for ILSSTNV 26-32. Of these microstate only a fraction are accessible at 300 K.

For longer peptide the Ramachandran-based discretization quickly leads to an untractable number of microstates. In FGAILSSTNV 23-27, we discretized residues $G_{24}$ to $N_{31}$ based on their Ramachandran plots, whereas the terminal residues $F_{23}$ and $V_{32}$ were discretized in only two states (Fig. 1 D in the SI). This yielded 26244 possible microstates, of which 3518 are visited in the simulations. For HSSNNFGAIL 18-27 and HSSNNFGAILSSTNV 18-32 this discretization did not yield a converged MSM. We therefore adopted a hierarchical approach. For HSSNNFGAIL 18-27, a residue based MSM was constructed for the subsegment SSNN 19-22 (see Fig. 4 in the SI). The fact that the MSM of the subsegment converges shows that the dynamics of the subsegment is largely uncorrelated from the dynamics of the remaining chain. This MSM was subjected to a PCCA+ analysis yielding four long-lived states within this segment. The conformation of the subsegment is then represented by the index of the corresponding long-lived state, whereas the conformations of the remaining residues are represented as before by residue-based bin indices. This yielded 5832 microstates, of which 1889 where visited by the trajectory. Similarly, the fragment HSSNNFGAILSSTNV 18-32, was divided into three subsegments HSSNN 18-22, FGAI 23-26 and LSSTNV 27-32, for which MSMs were constructed using the residue-based discretization (see Fig. 9 in the SI). A PCCA+ analysis of these MSMs identified 4 long-lived states for HSSNN 18-22 and LSSTNV 27-32, and 3 long-lived states for FGAI 23-26. Thus, the MSM of the HSSNNFGAILSSTNV

Hierarchy in the conformational ensemble of human islet amyloid polypeptide

18-32 was constructed using 48 possible microstates (44 actually visited).

The MD trajectory is projected onto the microstates of the system. The resulting microstate trajectory is the input for the Markov state model analysis which was performed using the EMMA software package[65] (see `pyemma.org`). With the command `mm_connectivity` the subset of connected microstates was generated. On this reduced set (option `-restrictToStates`), the actual transition matrices were estimated, using a sliding window algorithm (option `-slidingwindow`) and enforcing reversibility (option `-reversible`). From the analysis of the implied timescales associated to the eigenvalues (eq. 11) on a range of lag times spanning from 1 to 50 ns, a suitable lag time for the MSM construction for all models was found (10 ns for all systems; exceptions of NFGAIL 22-27, 5 ns, and FGAILSSTNV 23-32, 20 ns). The transition matrix was thus estimated using the command `mm_estimate` (using `-restrictToStates -slidingwindow -reversible` options) and further analyzed using `mm_transitionmatrixAnalysis` to extract the stationary density, the first eigenvalues and left and right eigenvectors. The PCCA+ analysis was performed using the `mm_pcca` command implemented in EMMA. Further processing of the eigenvalues and eigenvectors was implemented in Python[66].

## ACKNOWLEDGMENTS

The authors would like to thank Parthiv Patel, Elena Georgieva, and Dr Marieke Schor for the interesting discussions. F.V. acknowledges funding from the Dahlem research school. This research has been partially funded by Deutsche Forschungsgemeinschaft (DFG) through grant CRC 1114. The computer facilities of the Freie Universität Berlin (ZEDAT) are acknowledged for computer time.

## REFERENCES

[1]Westermark, P.; Andersson, A.; Westermark, G. T. *Physiol. Rev.* **2011**, *91*, 795–826.

[2]Clark, A.; Chargé, S. B.; Badman, M. K.; de Koning, E. J. *Acta Pathol Microbiol Immunol Scand* **1996**, *104*, 12–18.

[3]Westermark, P.; Wilander, E. *Diabetologia* **1983**, *24*, 342–346.

[4]Knight, J. D.; Hebda, J. A.; Miranker, A. D. *Biochemistry* **2006**, *45*, 9496–9508.

Hierarchy in the conformational ensemble of human islet amyloid polypeptide

[5]Patil, S. M.; Xu, S.; Sheftic, S. R.; Alexandrescu, A. T. *J. Biol. Chem.* **2009**, *284*, 11982–11991.

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

[7]Cooper, G. J.; Willis, A. C.; Clark, A.; Turner, R. C.; Sim, R. B.; Reid, K. B. *Proc. Natl. Acad. Sci. USA* **1987**, *84*, 8628–8632.

[8]Lorenzo, A.; Razzaboni, B.; Weir, G. C.; Yankner, B. A. *Nature* **1994**, *368*, 756–760.

[9]Jha, S.; Sellin, D.; Seidel, R.; Winter, R. *J. Mol. Biol.* **2009**, *389*, 907–920.

[10]Engel, M. F. M.; Yigittop, H.; Elgersma, R. C.; Rijkers, D. T. S.; Liskamp, R. M. J.; de Kruijff, B.; Höppener, J. W. M.; Antoinette Killian, J. *J. Mol. Biol.* **2006**, *356*, 783–789.

[11]Sumner Makin, O.; Serpell, L. C. *J. Mol. Biol.* **2004**, *335*, 1279–1288.

[12]Kajava, A. V.; Aebi, U.; Steven, A. C. *J. Mol. Biol.* **2005**, *348*, 247–252.

[13]Luca, S.; Yau, W.-M.; Leapman, R.; Tycko, R. *Biochemistry* **2007**, *46*, 13505–13522.

[14]Wiltzius, J. J. W.; Sievers, S. A.; Sawaya, M. R.; Cascio, D.; Popov, D.; Riekel, C.; Eisenberg, D. *Protein Sci.* **2008**, *17*, 1467–1474.

[15]Dupuis, N. F.; Wu, C.; Shea, J.-E.; Bowers, M. T. *J. Amer. Chem. Soc.* **2009**, *131*, 18283–18292.

[16]Bedrood, S.; Li, Y.; Isas, J. M.; Hegde, B. G.; Baxa, U.; Haworth, I. S.; Langen, R. *J. Biol. Chem.* **2012**, *287*, 5235–5241.

[17]Cao, P.; Abedini, A.; Raleigh, D. P. *Curr. Opin. Struct. Biol.* **2013**, *23*, 82–89.

[18]Westermark, P.; Engström, U.; Johnson, K. H.; Westermark, G. T.; Betsholtz, C. *Proc. Natl. Acad. Sci. USA* **1990**, *87*, 5036–5040.

[19]Betsholtz, C.; Christmanson, L.; Engström, U.; Rorsman, F.; Jordan, K.; O'Brien, T. D.; Murtaugh, M.; Johnson, K. H.; Westermark, P. *Diabetes* **1990**, *39*, 118–122.

[20]Jaikaran, E. T.; Higham, C. E.; Serpell, L. C.; Zurdo, J.; Gross, M.; Clark, A.; Fraser, P. E. *J. Mol. Biol.* **2001**, *308*, 515–525.

[21]Williamson, J. A.; Miranker, A. D. *Protein Sci.* **2007**, *16*, 110–117.

[22]Reddy, A. S.; Wang, L.; Singh, S.; Ling, Y. L.; Buchanan, L.; Zanni, M. T.; Skinner, J. L.; de Pablo, J. J. *Biophys. J.* **2010**, *99*, 2208–2216.

[23]Laghaei, R.; Mousseau, N.; Wei, G. *J. Chem. Phys. B* **2010**, *114*, 7071–7077.

[24]Andrews, M. N.; Winter, R. *Biophys. Chem.* **2011**, *156*, 43–50.

Hierarchy in the conformational ensemble of human islet amyloid polypeptide

[25]Miller, C.; Zerze, G. H.; Mittal, J. *J. Phys. Chem. B* **2013**, *117*, 16066–16075.

[26]Wu, C.; Shea, J.-E. *PLoS Comput. Biol.* **2013**, *9*, e1003211.

[27]Zerze, G. H.; Miller, C. M.; Granata, D.; Mittal, J. *J. Chem. Theory Comput.* **2015**, 150512141237002.

[28]Cecchini, M.; Curcio, R.; Pappalardo, M.; Melki, R.; Caflisch, A. *J. Mol. Biol.* **2006**, *357*, 1306–1321.

[29]De Simone, A.; Kitchen, C.; Kwan, A. H.; Sunde, M.; Dobson, C. M.; Frenkel, D. *Proc. Natl. Acad. Sci. USA* **2012**, *109*, 6951–6956.

[30]Knott, M.; Best, R. B. *PLoS Comput. Biol.* **2012**, *8*, e1002605.

[31]Qiao, Q.; Bowman, G. R.; Huang, X. *J. Amer. Chem. Soc.* **2013**, *135*, 16092–16101.

[32]Schor, M.; Mey, A. S. J. S.; Noé, F.; MacPhee, C. E. *J. Phys. Chem. Lett.* **2015**, *6*, 1076–1081.

[33]Stanley, N.; Esteban-Martín, S.; De Fabritiis, G. *Prog. Biophys. Mol. Biol.* **2015**, *119*, 47–52.

[34]Moriarty, D. F.; Raleigh, D. P. *Biochemistry* **1999**, *38*, 1811–1818.

[35]Goldsbury, C.; Goldie, K.; Pellaud, J.; Seelig, J.; Frey, P.; Müller, S. A.; Kistler, J.; Cooper, G. J.; Aebi, U. *J. Struct. Biol.* **2000**, *130*, 352–362.

[36]Azriel, R.; Gazit, E. *J. Biol. Chem.* **2001**, *276*, 34156–34161.

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

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

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

[40]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.

[41]Chodera, J. D.; Noé, F. *Curr. Opin. Struct. Biol.* **2014**, *25*, 135–144.

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

[43]Dionisio, A.; Menezes, R.; Mendes, D. A. *Phys A* **2004**, *344*, 326–329.

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

[45]Milner-White, E. J.; Poet, R. *Biochem J* **1986**, *240*, 289–292.

[46]Kabsch, W.; Sander, C. *Biopolymers* **1983**, *22*, 2577–637.

[47]Lindorff-Larsen, K.; Maragakis, P.; Piana, S.; Eastwood, M. P.; Dror, R. O.; Shaw, D. E. *PloS one* **2012**, *7*, e32131.

[48]Hoffmann, K. Q.; McGovern, M.; Chiu, C.-C.; de Pablo, J. J. *PloS one* **2015**, *10*, e0134091.

Hierarchy in the conformational ensemble of human islet amyloid polypeptide

[49]Piana, S.; Donchev, A. G.; Robustelli, P.; Shaw, D. E. *J. Phys. Chem. B* **2015**, *119*, 5113–5123.

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

[51]Lindorff-Larsen, K.; Trbovic, N.; Maragakis, P.; Piana, S.; Shaw, D. E. *J. Am. Chem. Soc* **2012**, *134*, 3787–3791.

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

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

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

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

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

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

[58]McGibbon, R. T.; Beauchamp, K. A.; Schwantes, C. R.; Wang, L.-P.; Hernandez, C. X.; Harrigan, M. P.; Lane, T. J.; Swails, J. M.; Pande, V. S. *MDTraj: a modern, open library for the analysis of molecular dynamics trajectories*; 2014; p 008896.

[59]Keller, B.; Gattin, Z.; van Gunsteren, W. F. *Proteins* **2010**, *78*, 1677–1690.

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

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

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

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

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

[65]Trendelkamp-Schroer, B.; Scherer, M. K.; Noé, F. EMMA: Emma's Markov Model Algorithms. Available at github.com/markovmodel/pyemma.

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

Hierarchy in the conformational ensemble of human islet amyloid polypeptide

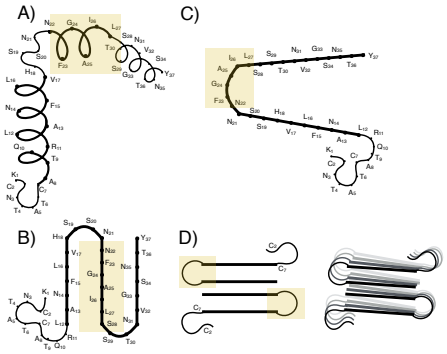


FIG. 2. Schematic representation of the suggested models. A) FGAIL 23-27 in a $\alpha$-helix. B) $\beta$-serpentine fold. C) $\beta$-hairpin with turn at FGAIL 23-27. D) Extended $\beta$-hairpin, with FGAIL23-27 in a $\beta$-strand configuration.

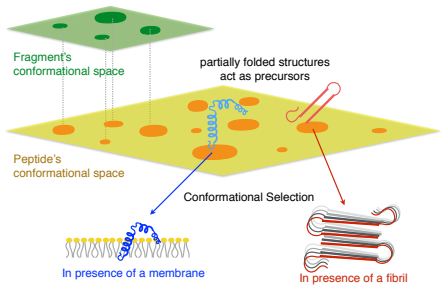


FIG. 3. Schematic representation of the configurational space of a fragment, with respect to the full peptide. The energy landscape is mainly flat with few minima that are slightly deeper than the rest. By lengthening the sequence, the configurational space is extended. Some of the more stable conformations can have a further increased or reduced stability. Specific conformations can be selected in presence of binding partners, such as membranes or oligomers.

Hierarchy in the conformational ensemble of human islet amyloid polypeptide

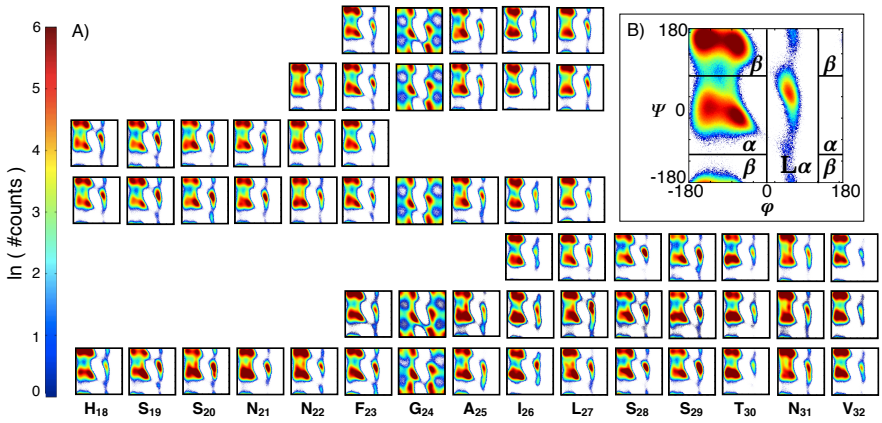


FIG. 4. A) Ramachandran distribution of all residues for each considered fragment. B) Example of Ramachandran plane of a capped amino acid and interpretation of the minima.

Hierarchy in the conformational ensemble of human islet amyloid polypeptide

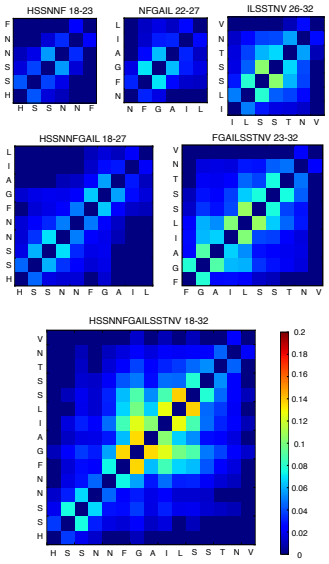


FIG. 5. Normalized Mutual Information (NMI) between each amino acid pair for each of the short fragments. FGAIL 23-27 presents very low NMI and is not included in the figure Groups of not-adjacent residues with a high NMI are involved secondary structure elements. Self NMI set to zero. A cutoff at 0.01 is used.

Hierarchy in the conformational ensemble of human islet amyloid polypeptide

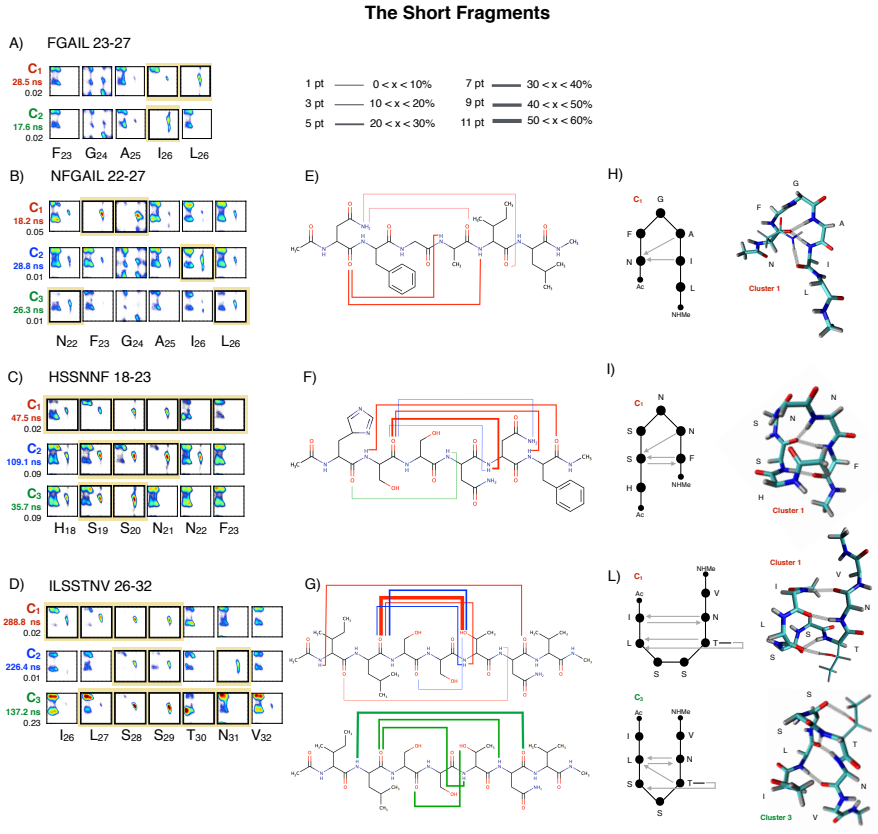


FIG. 6. Representation of the slow dynamics of the short fragments at $\tau$=10 ns (exception NFGAIL 22-27 $\tau$= 5 ns ). Structural characterization of the long-living states (clusters) via Ramachandran plots of each residue. Residues which undergo conformational changes are marked (A, B, C, D). The weight of each cluster and the timescale of the process are indicated. Hydrogen bonds with probability higher than 5% are shown (E, F, G). Pattern of hydrogen bonds with probability higher than 10% (the direction of the arrow goes from donor to acceptor) and most relevant hydrogen bonds marked in the structures of the corresponding cluster (H, I, L).

Hierarchy in the conformational ensemble of human islet amyloid polypeptide

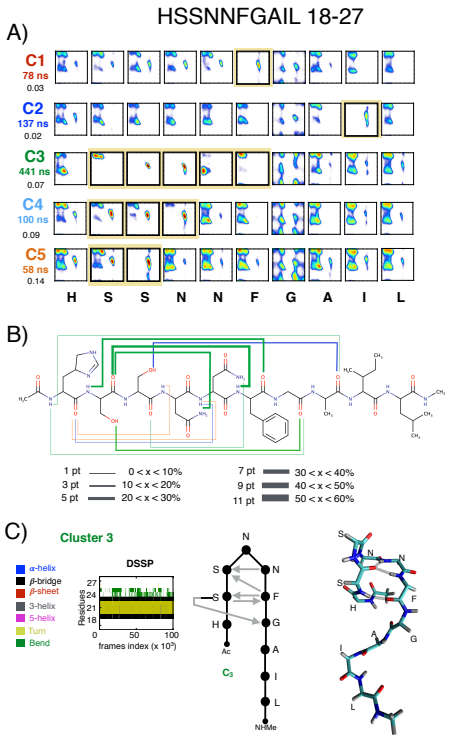


FIG. 7. Dynamics of HSSNNFGAIL 18-27 fragment at $\tau$= 10 ns. A) Structural characterization of the long-living states (clusters) via Ramachandran plots of each residue. Residues which undergo conformational changes are marked. The weight of each cluster and the timescale of the process are indicated. B) Hydrogen bonds with probability higher than 5%. C) Example structure and DSSP plot (resolution 100 frames) and pattern of hydrogen bonds with probability higher than 10% of cluster three (the direction of the arrow goes from donor to acceptor).

Hierarchy in the conformational ensemble of human islet amyloid polypeptide

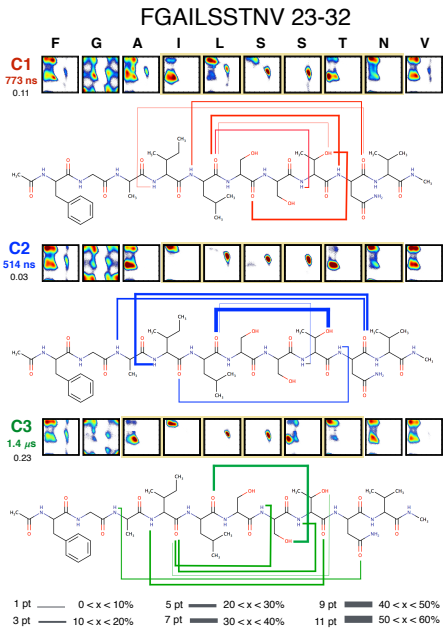


FIG. 8. Dynamics of FGAILSSTNV 23-32 fragment at $\tau$= 20 ns. Structural characterization of the long-living states (clusters) via Hydrogen bonds with probability higher than 5% and Ramachandran plots of each residue. Residues which undergo conformational changes are marked. The weight of each cluster and the timescale of the process are indicated.

Hierarchy in the conformational ensemble of human islet amyloid polypeptide

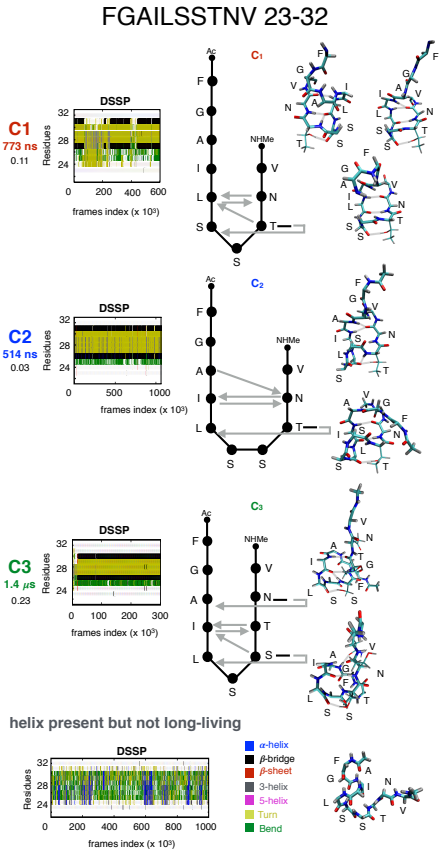


FIG. 9. Example structure and DSSP plot (resolution 100 frames) of clusters C1 - C3, pattern of hydrogen bonds with probability higher than 10% (the direction of the arrow goes from donor to acceptor) and corresponding structures.

Hierarchy in the conformational ensemble of human islet amyloid polypeptide

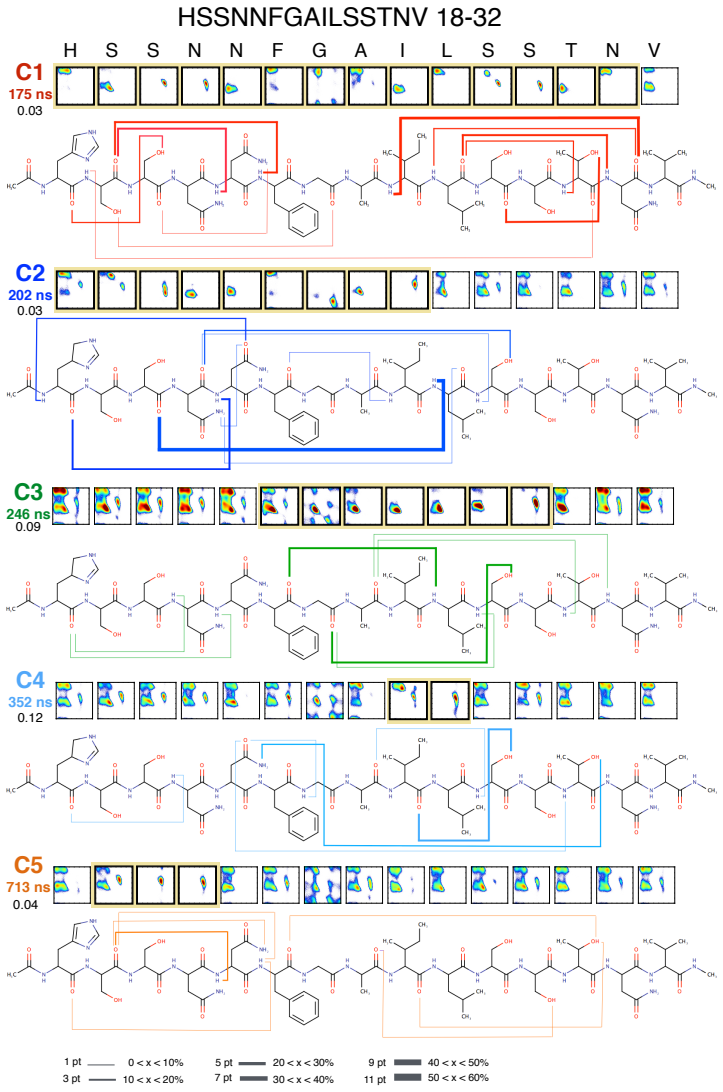


FIG. 10. Dynamics of HSSNNFGAILSSTNV 18-32 fragment at $\tau$= 10 ns. Structural characterization of the long-living states (clusters) via Hydrogen bonds with probability higher than 5 % and Ramachandran plots of each residue. Residues which undergo conformational changes are marked. The weight of each cluster and the timescale of the process are indicated.

Hierarchy in the conformational ensemble of human islet amyloid polypeptide

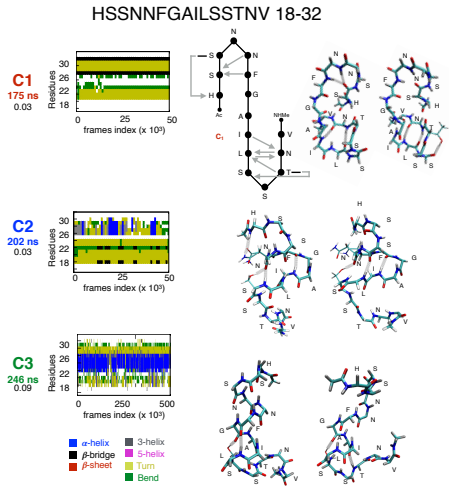


FIG. 11. Dynamics of HSSNNFGAILSSTNV 18-32 fragment at $\tau$= 10 ns. Example structures and DSSP plot (resolution 1000 frames) of clusters 1 and 3 and pattern of hydrogen bonds with probability higher than 10% for cluster 1(the direction of the arrow goes from donor to acceptor).

Hierarchy in the conformational ensemble of human islet amyloid polypeptide

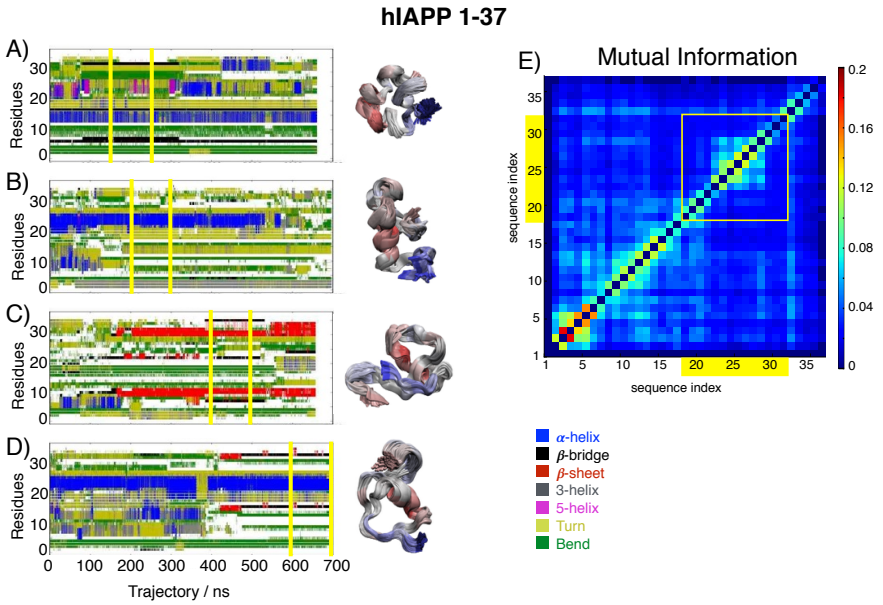


FIG. 12. A-D) DSSP plots of example runs of hIAPP 1-37 (resolution 0.1 ns). For each run a bundle of structures is shown. The trajectory region from which the structures were extracted is marked by yellow lines. E) Normalized Mutual Information (NMI). Self NMI set to zero and cutoff at 0.01.

Hierarchy in the conformational ensemble of human islet amyloid polypeptide

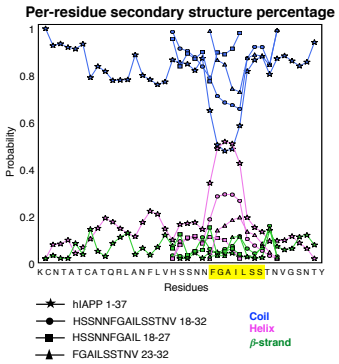


FIG. 13. Secondary structure probability per residue for hIAPP 1-37 (solid line) and HSSNNFGAILSSTNV 18-32 (dashed line, circular marks) HSSNNFGAIL 18-27 (dotted line square marks) and FGAILSSTNV 23-32 (dash-dotted line triangular marks)

Hierarchy in the conformational ensemble of human islet amyloid polypeptide

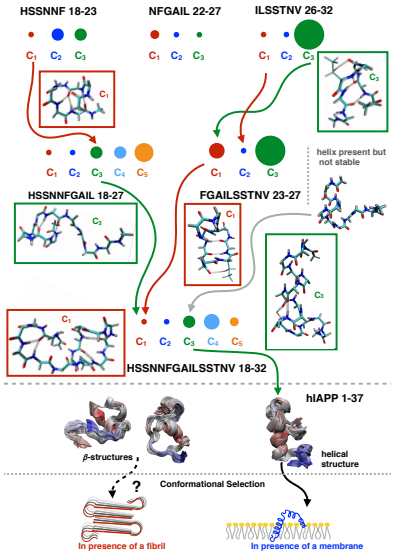


FIG. 14. Schematic representation of the hierarchy of the dynamics. The long-living configurations of the peptides' are identified with circles whose area is proportional to the equilibrium probability associated to the conformation by the model. Some of the long-living conformations of the shorter fragments are found in the long-lived conformations of the longer fragments (connected by arrows) and present the same hydrogen-bonds pattern.

**Hierarchy in the conformational ensemble of human islet amyloid polypeptide**

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

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

(Dated: 16 December 2015)

PACS numbers: Valid PACS appear here

Keywords: Suggested keywords

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

Hierarchy in the conformational ensemble of human islet amyloid polypeptide

## I. RESULTS

In this section we present additional results to the study.

### A. Discretization

The quality of a Markov State Model (MSM) crucially depends on how well the discretization captures the features of the energy landscape[1,2]. In this study we use backbone dihedral angles as reaction coordinates. The Ramachandran-plane of each residue is then discretized in two, three, five or six states. Fig 1 shows the stark boundaries in the four cases.

The $\{\phi - \psi\}$-distribution of an amino acid within a sequence has a typical distribution (fig 4 B in the main manuscript) with three minima: $\beta$, $\alpha$ and L$\alpha$. A three-states discretization (fig. 1 A) separates such minima in three distinguished states. Glycine, however, does not have a side-chain, thus its $\{\phi - \psi\}$-distribution is different, allowing for more of the configuration space to be populated. A six states discretization as in fig. 1 B separates the minima into different states. For the model of NFGAIL 22-27, a finer discretization of five states per Ramachandran-plane (fig. 1 C) produced a better converged model. In FGAILSSTNV 23-32 we used a two states discretization (fig. 1 D) to model the $\{\phi - \psi\}$-space of $F_{23}$ and $V_{32}$.

For systems with a longer chain a three-states per residue discretization may not produce a converged model. Sub-blocks of residues are therefore treated independently (sec I D and sec I F). Long-lived states of each sub-block are defined, and subsequently combined with a three-states per residue discretization of the remaining sequence to identify the state of the system.

### B. PCCA+

We model the dynamics by projecting it onto the backbone dihedral angles pairs of each residue. We applied the Perron Cluster Cluster Analysis (PCCA+) method to interpret the dynamics as exchanges between long-lived conformations. The PCCA+ method uses the dominant eigenvecotrs of the transition matrix to assign microstates to coarser sets[3,4]. By iteratively repeating the PCCA+ analysis with increasing number of eigenvectors, the

hierarchy of the free-energy barriers in the system can then be characterized. Fig. 2 shows a schematic representation of the application of the method. Providing the first four eigenvectors to the algorithm, four long-lived configurations are identified (fig 2, A). However, this does not provide any information on the relative hight of the free energy barriers between long-lived conformations. In order to interpret the dynamical processes as conformational probability density exchanges between the PCCA+ clusters, it is necessary to perform PCCA+ iteratively with increasing number of eigenvectors (fig. 2, B). Providing the first two dominant eigenvectors (fig. 2, B1) the conformational space is split in two long-lived clusters along the highest energy barrier of the system. When the third eigenvector is provided the algorithm splits the right cluster in two kinetically diverse states, based on the eigenvector sign (fig. 2, B2). The blue and red coloring represent areas of the configurational space where the eigenvector has opposite sign. The white area of the configurational space represent those microstates that do not participate to the dynamical process. Analogously, the information yielded by the fourth eigenvector splits the left cluster (fig. 2, B3) into two sub sets. Further iterations of the algorithm create more macrostates. Therefore, the slowest process of the system can be interpreted as transitions between the conformational state of the joint clusters one and two and the joint clusters three and four. Process two represents transitions between clusters one and two, whereas the third process represents transitions between cluster three and four. The main limitation of the PCCA+ method is given by the rather arbitral assignment of microstates that do not participate strongly in each dynamical process, leading to compounding error that is propagated with each iteration. The PCCA+ algorithm avoids the propagation of such error by considering multiple eigenvectors simultaneously[4,5]. In analogy to this example we can interpret the dynamics of the short fragments HSSNNF 18-23, NFGAIL 22-27 and FGAIL 23-27, ILSSTNV 26-32 and the longer fragments HSSNNFGAIL 18-27, FGAILSSTNV 23-32 and HSSNNFGAILSSTNV 18-32 as transition between PCCA+ clusters.

### C. FGAIL 23-27

FGAIL 23-27 is a pentamer of aliphatic amino acids. Therefore, it is extremely flexible and does not form secondary structure elements. This can be seen in fig. 3, where we plot the normalized mutual information (NMI) of each residue pair. Adjacent residues share the

highest NMI. $G_{24}$ has a slightly higher NMI than any other amino acid, probably due to the different backbone angles distribution typical of glycines.

### D. HSSNNFGAIL 18-27

The number of microstates depends exponentially on the length of the sequence. For the decamer HSSNNFGAIL 18-27, a discretization based on three states per residue leads to a non-converged model. We thus combine a sub-model of four residues (SSNN, i.e. residues 19 to 22) with a 3-states per residue discretization of the remaining sequence, as described in details in sec. IV D in the main manuscript. Here we show the analysis of the sub-model SSNN.

The Ramachandran plane of each residue in the segment 19-22 is discretized in three bins, based on the $\{\phi - \psi\}$- distribution features (fig. 4 A). Each bin-combination corresponds to a microstate of the system, i.e. $3^4 = 81$ possible microstates. The discretized time-series, obtained by projecting the simulation onto the discretization, is used as input for the MSM. At10 ns the slowest three timescales have reached a plateau (fig. 4 B), therefore we choose this value as the lag time for the MSM.

Using the PCCA+ algorithm, the microstates are divided into four clusters. Such PCCA+ clusters correspond to long-lived regions of the sub-model. A structural characterization of the clusters is shown in fig. 4 B in terms of the Ramachandran distribution of each considered residue and bundle of structures.

Cluster one, which accounts for 8% of the equilibrium distribution, shows a very distinctive structure: $S_{19}$ is confined in a $\beta$ conformation; $S_{20}$ and $N_{21}$ populate the L$\alpha$ region and $N_{22}$ populates the $\alpha$ minimum. This recurring pattern is the mark of a $\beta$-hairpin structure. Cluster two, which represent 12% of the equilibrium distribution, presents a preferences of residues 19 to 21 for the L$\alpha$ configuration. Cluster four, which represents 20% of the equilibrium distribution, is characterized by $S_{20}$ in the L$\alpha$ state. Cluster three resembles the equilibrium distribution and accounts for 60% of the equilibrium distribution. By looking at the splitting of the microstates into PCCA+ clusters, one can associate the slowest process, which occurs at c.a. 300 ns, to a conformational exchange between thr equilibrium (cluster four) and the $\beta$-hairpin (cluster one). The second slowest process, whose timescale is of c.a. 70 ns, is the transition between cluster four and configurations with $S_{20}$ in the L$\alpha$ state.

Hierarchy in the conformational ensemble of human islet amyloid polypeptide

The third process (c.a. 50 ns) is characterized by transitions between the equilibrium state and configurations with the residues 19 to 21 in a L$\alpha$ conformation.

To construct the model of the full system, the time-series of the backbone dihedral angles explored by residues 19-22 over the simulation is projected on the four PCCA+ clusters and combined to a three-states per residue discretization of the remaining sequence, as shown in the main manuscript.

Furthermore, we analyzed the secondary structure configurations assumed by the peptide HSSNNFGAIL 18-27 over the simulated trajectory (fig. 5). The 9 replicas have different lengths, spanning from 266 ns to 1 $\mu$s, with the exception of the eighth replica that is only 8 ns long and was thus excluded from the MSM construction. Fig. 5 A shows the DSSP analysis of each independent run at a time resolution of 1 ns. Residues 23-27, i.e the FGAIL sub-fragment, do not assume a helical structure. The main secondary structure element comprises residues 19 and 23 forming a $\beta$-bridge. It is interesting to compare the DSSP assignment to our Ramachandran-based discretization. In fig. 5 B, each residue at each time-step is color-coded according to the dihedral angle values (fig 1 A). The hairpin-like structure correspond to residues $S_{19}$ and $F_{23}$ in the $\beta$ state, residues 20-21 in the L$\alpha$ configuration, and residue $N_{22}$ in the $\alpha$ state. Such hairpin structure is found as long-lived by the full-system MSM. The other dynamical processes involve single residues, with exception of the third process, where residues 19-22 prefer the L$\alpha$ state (Cluster 4 in fig 7). The latter configuration, however, is identified as coil by the DSSP algorithm, pointing at the absence of stabilizing hydrogen bonds in the configuration.

### E. FGAILSSTNV 23-32

In analogy to the analysis of HSSNNFGAIL 18-27, we carried out a DSSP analysis of the secondary structure configurations the peptide FGAILSSTNV 23-32 assumes over time (fig. 6 A) at a time resolution of 1 ns. The FGAIL sub-fragment can assume helical structures, in agreement with the longer fragment HSSNNFGAILSSTNV 18-32. Such helical structures are, however, more transient than in HSSNNFGAILSSTNV 18-32 and therefore the corresponding long-lived state is not captured by our PCCA+ analysis. Moreover, $\beta$-structures between residues $I_{26}$-$T_{30}$, $L_{27}$-$N_{31}$ and $I_{26}$-$N_{31}$ are formed. The first two hairpin-like structures have the same dihedral angle pattern found in HSSNNFGAIL 18-27, where

Hierarchy in the conformational ensemble of human islet amyloid polypeptide

the residues forming the $\beta$-bridge are in a $\beta$-state and the residues in between are in the states $\{L\alpha, L\alpha, \alpha\}$. The $\beta$-structure between $I_{26}$-$N_{31}$ has a similar pattern, but the residues in the L$\alpha$ conformation are three instead of two. This can be noticed in the DSSP-like plots of fig. 6 B. Here, for each time-step, each residue is color-coded according to the backbone dihedral angle state (fig 1 A).

The first 10 $\mu$s of simulated data exhibit a sink state, i.e. a state which is accessed by the simulation but never left. More specifically, the sink state is the $\beta$-hairpin conformation between $I_{26}$-$N_{31}$ in replica 9. A MSM estimated from this simulation data will assign a very long life-time and consequently an extremely high equilibrium probability to this state. Fig 7 A shows a comparison of the relative equilibrium populations predicted by the MSM with the equilibrium populations estimated directly form the MD simulations. The equilibrium population of the sink state is grossly overestimated by the MSM. To construct a more realistic MSM one has to have a more realistic estimate of the state life-time. We hence started 27 short simulations (100 ns) from the sink state and three simulations from the closely related state in which residues $I_{26}$ and $T_{30}$ come in the $\beta$ conformation (fig 8). From the combined data set we constructed a MSM where we discretized the residues $G_{24}$ to $N_{31}$ into three states each and $F_{23}$ and $V_{32}$ into two states each. This yields converged implied timescales (fig. 14) for $\tau = 20$ ns (slightly larger than the other MSM) and an equilibrium distribution which is in agreement with populations estimated directly from the MD data (fig. 7 B). We checked the convergence of the eigenvectors by calculating the Euclidean distance between the eigenvector $l_i(\tau)$ and the corresponding reference eigenvector $l_i(\tau = 20ns)$ at different lag times. A distance of less than 0.1 indicates a well converged model ( Ref[6]).

### F. HSSNNFGAILSSTNV 18-32

As in HSSNNFGAIL 18-27, we use a hierarchical approach for the discretization of the conformational space of HSSNNFGAILSSTNV 18-32. We identified three different regions that we treated separately: *i*) residues 18-22 (HSSNN), *ii*) residues 23-26 (FGAI) and *iii*) residues 27-32 (LSSTNV). The segment 18-22 (block *i*) was discretized with a three-states per residue binning, i.e. $3^5 = 243$ (fig. 9 A). From the analysis of DSSP plots (fig 10), it results that the central region (FGAI 23-26), which constitutes block *ii*, has high propensity

Hierarchy in the conformational ensemble of human islet amyloid polypeptide

to form $\alpha$-helical structures. The Ramachandran plane of each of the residues in this region was thus binned into two states: one for the $\phi - \psi$ combinations corresponding to an $\alpha$-helix, and one for the remaining allowed combinations, for a total of $2^4 = 16$ possible states (fig. 9 B). Such two-states discretization, despite being less refined than three-states discretization, produces a better converged model. Moreover, it captures the interesting conformational change in this region, which corresponds to the formation of an $\alpha$-helix. Block *iii* also produces a converged model with the Ramachandran-plane of each residue being partitioned into three states, for a total of $3^6 = 729$ possible microstates (fig. 9 C).

By analyzing the dynamics of each block (fig. 9 D, E, F), we notice that blocks *i* and *iii* both present the slowest process at a timescale of c.a. 500 ns. Those processes are, however, independent and different. For block *i* the slowest process corresponds to the transition towards a conformation where residues $S_{20}$ and $N_{21}$ are in the L$\alpha$ minimum. Process two occurs at c.a. 150 ns and shows again $S_{20}$ in the L$\alpha$ configuration, whilst residues 21 and 22 are in the $\alpha$-helix state. Process three takes place at a similar timescale (c.a. 150 ns), but involves the exchange between $\alpha$ and $\beta$ configuration in residue $S_{19}$, whilst $S_{20}$ and $N_{21}$ are in the L$\alpha$ configuration. Block *iii* instead shows a slowest process characterized by $L_{27}$ populating the L$\alpha$ region. Process two (c.a 230 ns) presents $S_{29}$ in the L$\alpha$ configuration and $S_{28}$ populating both the $\beta$ and the L$\alpha$ minima. The third dynamic process occurs at c.a. 180 ns and involves residues 27 and 28 in a $\alpha$ state, combined with $S_{29}$ in the L$\alpha$ configuration. For both blocks *i* and *iii*, the dynamics was projected onto four PCCA+ states for further modeling.

The central fragment FGAI 23-26 shows faster processes compared to the other blocks. As only three PCCA+ states were used to construct the complete model, we present hereby only three long-lived conformations and two dynamic processes between them. As previously mentioned, such processes involve the formation of an $\alpha$-helix. The slowest movement involves residue $I_{26}$ assuming the $\alpha$-helix conformation (process one at c.a. 140 ns). The entire fragment forming an $\alpha$-helix, instead occurs at 60 ns.

The combination of the PCCA+ states of each sub-model forms the full-length peptide microstates, which constitute the input for the MSM, whose results are presented in fig. 10 in the main manuscript. It is worth to mention that the discretization is rather coarse; nonetheless, it is capable of identifying all the main secondary structure elements explored by the system.

Hierarchy in the conformational ensemble of human islet amyloid polypeptide

Furthermore, we analyzed the secondary structure configurations assumed by the peptide HSSNNFGIALSSTNV 18-32 over the simulated trajectory (fig. 10). Of the ten replicas, eight have a length of 1 $\mu$s and two are shorter (666 ns and 909 ns), for a total aggregated simulated time of c.a. 9.6 $\mu$s. Fig. 10 A shows the DSSP analysis of each independent run at a resolution of 1 ns. Despite the central part forming an $\alpha$-helix being the main secondary structure configuration, transient $\beta$-structures are also visited. In particular, one replica explores a conformation where a $\beta$-bridge is formed between residue $N_{22}$ and $V_{32}$, with the amyoloidogenic region (residues 23-27) mostly in a bend (fifth row). Such configuration is a possible precursor to a $\beta$-hairpin presenting the FGAIL 23-27 fragment in a turn. On the other hand, conformations with the FGAIL 23-27 fragment forming a $\beta$-strand are also visited (first row).

It is interesting to notice how the secondary structure elements evidenced by DSSP analysis are translated into residue-based $\{\phi - \psi\}$-discretization (fig. 10, B). Whereas the $\alpha$-helix, which is a local interaction, corresponds to the residues populating the $\alpha$ minimum of the Ramachandran-plane, $\beta$-structures affect residues which are far apart in the sequence. Multiple backbone dihedral angles states combinations can be thus associated to the same secondary structure feature. Moreover, residues assuming a stable L$\alpha$-conformation can be found in a variety of secondary structure elements according to the DSSP definition.

To further investigate how much of the state space is visited by the shorter fragments, we compared the equilibrium $\{\phi - \psi\}$-distributions of each residue of each fragment with those of the 15-mer HSSNNFGAILSSTNV 18-32 (fig. 11). Throughout the sequences, we notice a higher propensity of HSSNNFGAILSSTNV 18-32 residues for the $\alpha$ minimum. This is not surprising, given that the formation of an $\alpha$-helix is one of the main processes in the system. However, qualitatively all fragments agree in the $\{\phi - \psi\}$ equilibrium distributions, and all the three minima ($\alpha$, $\beta$ and L$\alpha$) are populated in each residue of each fragment. This indicates that the exploration of different structures in the fragments does not depend on the single residues but on the correlated dynamics of groups of residues.

### G. hIAPP

hIAPP is a 37-residue long intrinsically disordered peptide, i.e. its dynamics involves transitions between many different conformations. In order to construct a converged MSM

Hierarchy in the conformational ensemble of human islet amyloid polypeptide

and analyze the dynamics, more than the currently available 10.8 $\mu$s of data are thus needed. We present here a qualitative analysis of equilibrium properties.

We performed a DSSP analysis of the trajectories (fig. 12, and in the main manuscript fig. 12), which confirms that the peptide is intrinsically disordered in solution, but does not behave as a random coil. In addition, we investigated the equilibrium $\{\phi - \psi\}$-distribution of each residue in the sequence of hIAPP 1-37 (fig. 13). As the peptide is not capped on the N-terminus, the $\phi$-torsion angle of $K_1$ is not properly defined and $K_1$ is not included into the analysis. Every residue is highly flexible in its backbone conformations, i.e. another confirmation of the intrinsic disorder of the peptide. It is also worthy of notice that residues $L_{16}$ and $I_{26}$ do not populate the L$\alpha$ minimum. Additional simulations might explore these regions of the conformational space. The latter is particularly interesting, as it participates in the slow dynamic processes of the smaller fragments.

From the DSSP analysis of each independent simulation run (fig 12), we notice again an abundance of $\alpha$-helix at the central region. An helical propensity is also shown in other parts of the sequence. Residues 5-20 and 25-32 can assume $\alpha$-helical configurations, whereas residues 2-4 are mainly found in a 3-helix or a coil. Moreover, transient $\beta$-structures are also visited. However, which residues participate to the $\beta$-structure is not consistent throughout the simulations: different trajectories explore different $\beta$-structures. Distant residues forming a $\beta$-sheet or a $\beta$-bridge, separated by residues in a turn/bend configuration, suggest the formation of a $\beta$-hairpin-like structure. None of the independent runs explored configurations with three $\beta$-strands as in model B of fig. 2 of the main manuscript.

### H. Timescales

In this section we present the implied timescales at different lag times, obtained from the MSMs of each fragment. As described in section IV C in the main manuscript, the convergence of the implied timescales with respect to the lag time can be used as a confirmation of the Markovianity of the dynamics. Fig. 14 shows the implied timescales for the slowest processes for each system. At the lag time of model analysis (10 ns for each system, except NFGAIL 18-23 $\tau = 5$ ns and FGAILSSTNV $\tau = 20$ ns), the implied timescales have reached a plateau. The timescales of FGAILSSTNV 23-32 show the worst convergence. This is an effect of the sink state, which is not completely erased by the additional simulations. At the

Hierarchy in the conformational ensemble of human islet amyloid polypeptide

lag time of model construction ($\tau$=20 ns), however, the eigenvectors are independent of the lag time (fig: fig:FGAILSSTNV-sink).

## I. Hydrogen bond analysis

In this section we present the results of a hydrogen bond analysis performed with the VMD: visual molecular dynamics[7] software, to characterize the long-lived states of each fragment. The hydrogen bonds are defined by a donor-acceptor distance cut-off of 3 Angstrom and an angle cut-off of 20°. For each system we create separate trajectories containing those frames belonging to each PCCA+ cluster, i.e. the long-lived states, and calculate the relative population of each hydrogen bond using VMD. The results are listed in tables I-VII.

By comparing the hydrogen bonds of different fragments, a hierarchy of the dynamics is evinced. The same long-lived configurations, stabilized by the same hydrogen bonds patterns, are found in the shorter and in the longer fragments. The hierarchy of the hydrogen bonds is visualized in fig. 15. Reappearing bonds are coded in the same color.

Hierarchy in the conformational ensemble of human islet amyloid polypeptide

**FGAIL 23-27**:

| Cluster 1 | | | Cluster 2 | | |
|---|---|---|---|---|---|
| donor | acceptor | percentage | donor | acceptor | percentage |
| ILE26-Main | PHE23-Main | 0.41% | ILE26-Main | PHE23-Main | 0.72% |
| | | | LEU27-Main | PHE23-Main | 0.97% |
| | | | LEU27-Main | ALA25-Main | 1.08% |

TABLE I. Hydrogen bonds identified by VMD for each PCCA+ cluster of FGAIL 23-27. Hydrogen bonds with a relative population $> 0.4\%$ are listed.

**NFGAIL 22-27**:

| Cluster 1 | | | Cluster 2 | | |
|---|---|---|---|---|---|
| donor | acceptor | percentage | donor | acceptor | percentage |
| ALA25-Main | ASN22-Main | 13.83% | GLY24-Main | ASN22-Side | 0.76% |
| ASN22-Side | ALA25-Main | 5.26% | | | |
| ASN22-Side | ILE26-Main | 6.46% | | | |
| ILE26-Main | ASN22-Side | 1.16% | | | |
| GLY24-Main | ASN22-Side | 1.18% | | | |
| ILE26-Main | ASN22-Main | 10.61% | | | |
| ASN22-Main | ILE26-Main | 2.14% | | | |
| Cluster 3 | | | | | |
| donor | acceptor | percentage | | | |
| LEU27-Main | GLY24-Main | 1.22% | | | |
| ILE26-Main | PHE23-Main | 1.01% | | | |
| ASN22-Side | LEU27-Main | 0.99% | | | |

TABLE II. Hydrogen bonds identified by VMD for each PCCA cluster of NFGAIL 22-27. Hydrogen bonds with a relative population $> 1\%$ are listed.

Hierarchy in the conformational ensemble of human islet amyloid polypeptide

**HSSNNF 18-23**:

| Cluster 1 | | | Cluster 2 | | |
|---|---|---|---|---|---|
| donor | acceptor | percentage | donor | acceptor | percentage |
| ASN22-Main | SER19-Main | 21.07% | ASN22-Main | SER19-Main | 6.68% |
| PHE23-Main | SER19-Main | 19.99% | ASN22-Side | SER19-Main | 5.15% |
| SER19-Main | PHE23-Main | 11.44% | ASN22-Main | HIS18-Main | 4.91% |
| Cluster 3 | | | | | |
| donor | acceptor | percentage | | | |
| ASN21-Main | HIS18-Main | 5.51% | | | |

TABLE III. Hydrogen bonds identified by VMD for each PCCA+ cluster of HSSNNF 18-23. Hydrogen bonds with a relative population $> 5\%$ are listed.

Hierarchy in the conformational ensemble of human islet amyloid polypeptide

**ILSSTNV 26-32**:

| Cluster 1 | | | Cluster 2 | | |
|---|---|---|---|---|---|
| donor | acceptor | percentage | donor | acceptor | percentage |
| THR30-Side | LEU27-Main | 52.99% | THR30-Main | LEU27-Main | 13.52% |
| ASN31-Main | ILE26-Main | 5.61% | THR30-Side | SER28-Main | 7.56% |
| ILE26-Main | ASN31-Main | 15.28% | THR30-Side | LEU27-Main | 23.42% |
| THR30-Main | LEU27-Main | 10.07% | | | |
| Cluster 3 | | | | | |
| donor | acceptor | percentage | | | |
| THR30-Side | SER28-Main | 24.35% | | | |
| THR30-Main | LEU27-Main | 20.40% | | | |
| ASN31-Main | LEU27-Main | 28.65% | | | |
| LEU27-Main | ASN31-Main | 38.13% | | | |

TABLE IV. Hydrogen bonds identified by VMD for each PCCA+ cluster of ILSSTNV 26-32. Hydrogen bonds with a relative population $> 5\%$ are listed.

Hierarchy in the conformational ensemble of human islet amyloid polypeptide

**HSSNNFGAIL 18-27**:

| Cluster 1 | | | Cluster 2 | | |
|---|---|---|---|---|---|
| donor | acceptor | percentage | donor | acceptor | percentage |
| ALA25-Main | ASN22-Main | 2.95% | SER20-Side | ALA25-Main | 12.25% |
| ASN21-Main | ILE26-Main | 3.21% | ASN22-Main | HIS18-Main | 6.90% |
| Cluster 3 | | | Cluster 4 | | |
| donor | acceptor | percentage | donor | acceptor | percentage |
| ASN22-Main | SER19-Main | 26.26% | LEU27-Main | PHE23-Main | 2.25% |
| ASN22-Side | SER20-Main | 8.17% | ASN22-Main | SER19-Main | 2.69% |
| PHE23-Main | SER19-Main | 33.36% | | | |
| SER19-Main | PHE23-Main | 25.11% | | | |
| SER19-Side | GLY24-Main | 10.01% | | | |
| HIS18-Main | GLY24-Main | 5.85% | | | |
| HIS18-Main | ALA25-Main | 5.33% | | | |
| Cluster 5 | | | | | |
| donor | acceptor | percentage | | | |
| ASN21-Main | HIS18-Main | 5.63% | | | |
| ASN22-Main | HIS18-Main | 5.21% | | | |

TABLE V. Hydrogen bonds identified by VMD for each PCCA+ cluster of HSSNNFGAIL 18-27. Hydrogen bonds with a relative population $> 5\%$ are listed.

Hierarchy in the conformational ensemble of human islet amyloid polypeptide

**FGAILSSTNV 23-32**:

| Cluster 1 | | | Cluster 2 | | |
|---|---|---|---|---|---|
| donor | acceptor | percentage | donor | acceptor | percentage |
| THR30-Main | LEU27-Main | 17.99% | THR30-Side | LEU27-Main | 60.67% |
| THR30-Side | LEU27-Main | 7.04% | ALA25-Main | ASN31-Main | 22.58% |
| THR30-Side | SER28-Main | 23.59% | ILE26-Main | ASN31-Main | 33.72% |
| ASN31-Main | LEU27-Main | 26.26% | ASN31-Main | ILE26-Main | 18.57% |
| ILE26-Main | ASN31-Main | 8.81% | THR30-Main | LEU27-Main | 6.17% |
| LEU27-Main | ASN31-Main | 17.79% | | | |
| Cluster 3 | | | | | |
| donor | acceptor | percentage | | | |
| SER29-Main | ILE26-Main | 23.27% | | | |
| SER29-Side | LEU27-Main | 37.92% | | | |
| THR30-Main | ILE26-Main | 23.89% | | | |
| THR30-Side | ILE26-Main | 6.42% | | | |
| ALA25-Main | ASN31-Side | 12.19% | | | |
| ILE26-Main | THR30-Main | 26.31% | | | |

TABLE VI. Hydrogen bonds identified by VMD for each PCCA+ cluster of FGAILSSTNV 23-32. Hydrogen bonds with a relative population $> 5\%$ are listed.

Hierarchy in the conformational ensemble of human islet amyloid polypeptide

**HSSNNFNFGAILSSTNV 18-32**:

| Cluster 1 | | | Cluster 2 | | |
|---|---|---|---|---|---|
| donor | acceptor | percentage | donor | acceptor | percentage |
| LEU27-Main | ASN31-Main | 18.62% | ILE26-Main | PHE23-Main | 7.66% |
| THR30-Side | SER28-Main | 22.23% | SER28-Main | ASN21-Main | 6.40% |
| THR30-Main | LEU27-Main | 11.79% | SER28-Side | ASN21-Main | 12.74% |
| ASN31-Main | LEU27-Main | 28.32% | LEU27-Main | SER20-Main | 40.58% |
| ASN22-Main | SER19-Main | 26.61% | HIS18-Main | ASN22-Side | 13.71% |
| ILE26-Main | ASN31-Main | 33.58% | ASN21-Side | ASN22-Side | 7.03% |
| SER20-Side | HIS18-Main | 13.38% | ASN22-Main | HIS18-Main | 27.38% |
| PHE23-Main | SER19-Main | 25.40% | ASN21-Side | LEU27-Main | 7.17% |
| SER19-Side | GLY24-Main | 9.43% | | | |
| SER19-Main | THR30-Main | 9.69% | | | |
| ASN22-Side | SER20-Main | 5.19% | | | |
| Cluster 3 | | | Cluster 4 | | |
| donor | acceptor | percentage | donor | acceptor | percentage |
| LEU27-Main | PHE23-Main | 26.59% | SER28-Side | ILE26-Main | 21.22% |
| SER28-Side | GLY24-Main | 29.83% | ASN21-Main | HIS18-Main | 5.97% |
| SER28-Main | GLY24-Main | 8.88% | SER28-Main | ALA25-Main | 6.33% |
| THR30-Main | ALA25-Main | 9.45% | ASN22-Side | THR30-Side | 11.63% |
| ASN22-Main | HIS18-Main | 5.58% | THR30-Main | ASN22-Side | 9.58% |
| ASN21-Main | HIS18-Main | 5.22% | PHE23-Main | SER29-Side | 4.94% |
| ASN31-Main | ALA25-Main | 5.52% | GLY24-Main | ASN22-Side | 6.33% |
| Cluster 5 | | | | | |
| donor | acceptor | percentage | | | |
| ASN22-Main | SER19-Main | 12.04% | | | |
| THR30-Side | ILE26-Main | 7.62% | | | |
| THR30-Side | PHE23-Main | 7.55% | | | |
| ASN22-Side | SER19-Main | 7.15% | | | |
| LEU27-Main | PHE23-Main | 4.91% | | | |
| SER19-Side | GLY24-Main | 4.96% | | | |
| SER29-Side | ALA25-Main | 5.60% | | | |
| PHE23-Main | HIS18-Main | 5.83% | | | |
| PHE23-Main | SER19-Main | 5.14% | | | |

TABLE VII. Hydrogen bonds identified by VMD for each PCCA+ cluster of HSSNNFGAILSSTNV 18-32. Hydrogen bonds with a relative population $> 5\%$ are listed.

Hierarchy in the conformational ensemble of human islet amyloid polypeptide

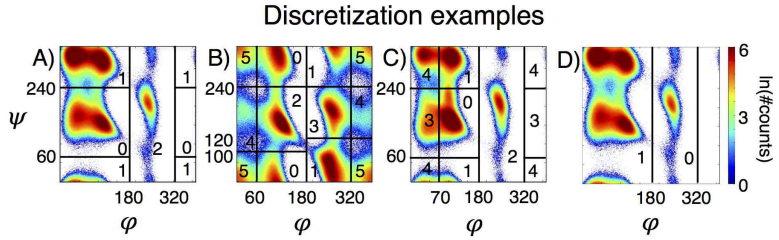


FIG. 1. Example of discretizations. A) Three-states discretization. B) Glycine six-states discretization. C) Five states discretization. D) Two states discretization

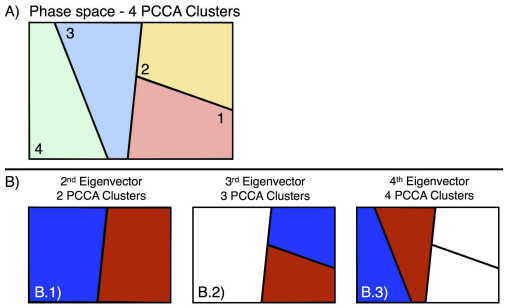


FIG. 2. Schematic representation of PCCA+. A) PCCA+ partition of the phase space into PCCA+ clusters. B) Hierarchy of the underlying energy landscape obtained by iterative use of the PCCA+ algorithm with increasing number of eigenvectors.

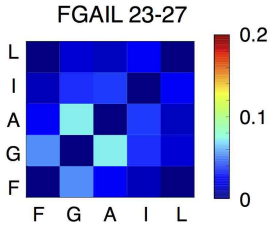


FIG. 3. Normalized Mutual Information (NMI) between each amino acid pair for FGAIL 23-27. Self NMI and NMI $< 0.01$ set to zero (significance level of the NMI analysis, see section IV B in the main manuscript).

Hierarchy in the conformational ensemble of human islet amyloid polypeptide

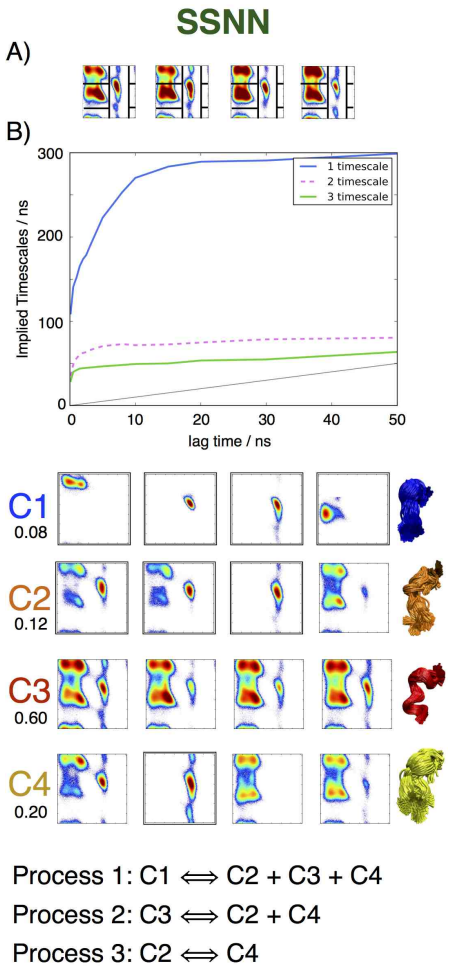


FIG. 4. Sub-model of HSSNNFGAIL at $\tau = 10$ ns. A) Discretization of the Ramachandran plane of the residues taken into consideration in the sub model. B) Timescales plots and interpretation of the dynamics as exchange of probabilities between long-lived conformations (clusters 1-4); for each cluster the Ramachandran plots of each residue and example structures are shown.

Hierarchy in the conformational ensemble of human islet amyloid polypeptide

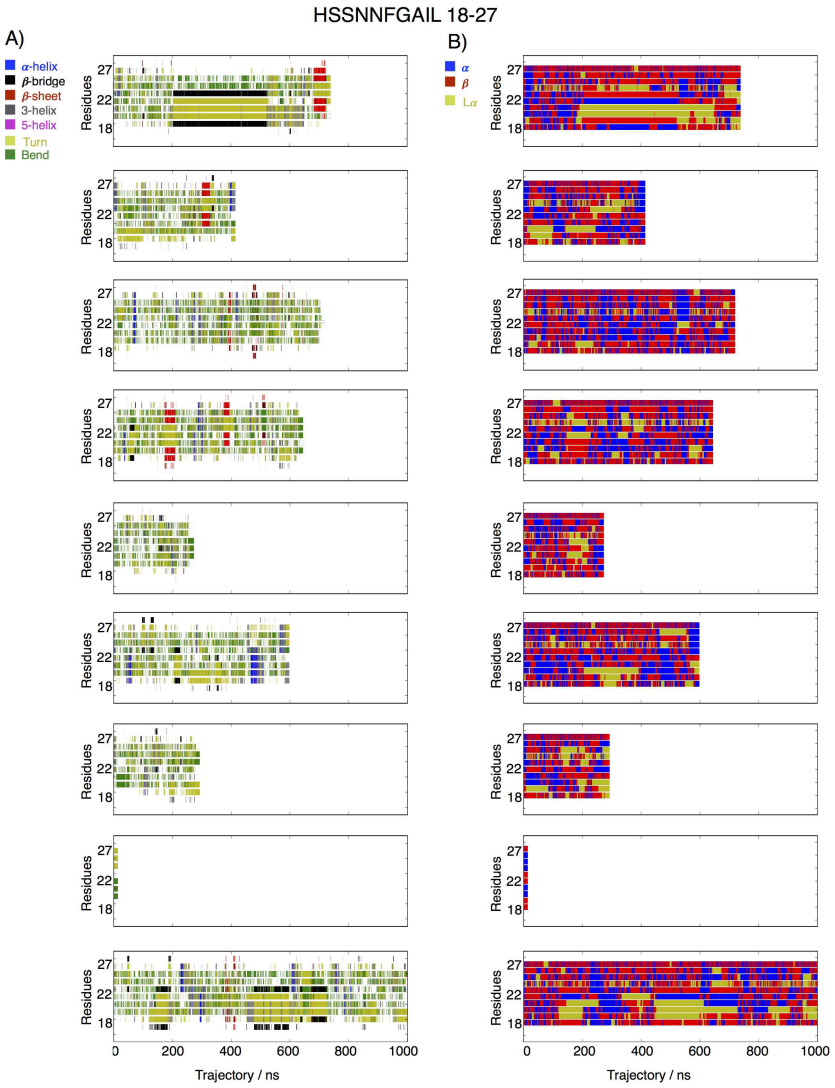


FIG. 5. 9 independent simulation runs of HSSNNFGAIL 18-27 A) DSSP plot, time resolution 1 ns:. B) $\{\phi - \psi\}$-state of each residue.

Hierarchy in the conformational ensemble of human islet amyloid polypeptide

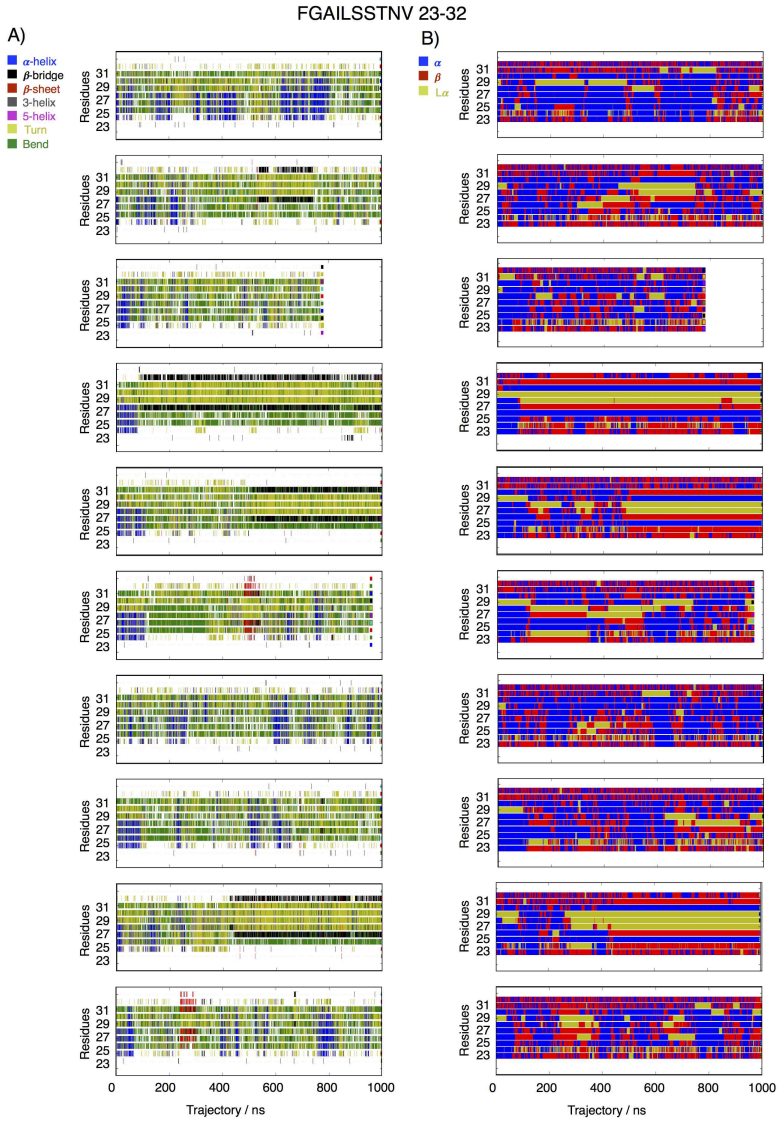


FIG. 6. 10 independent simulation runs of FGAILSSTNV 23-32: A) DSSP plot, time resolution 1 ns. B) $\{\phi - \psi\}$-state of each residue.

Hierarchy in the conformational ensemble of human islet amyloid polypeptide

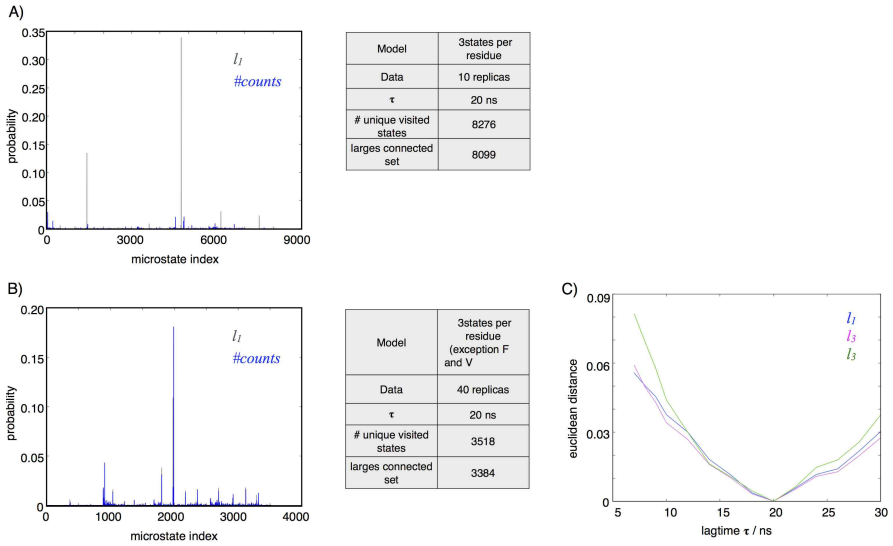


FIG. 7. A) Relative equilibrium population per microstate as predicted by a MSM with 3-states per residue (grey), compared with the relative equilibrium population estimated from 10 $\mu$s of simulation data (blue). B) Relative equilibrium population per microstate as predicted by a MSM with terminal residues discretized in 2 states and remaining sequence 3-states (grey), compared with the relative equilibrium population estimated from 13 $\mu$s of simulation data (blue). C) Convergence of the eigenvectors measured as the Euclidean distance $d(l_i(\tau), l_i(\tau_r ef))$, where $\tau_r ef = 20ns$ is the lag-time of the MSM in B.

Hierarchy in the conformational ensemble of human islet amyloid polypeptide

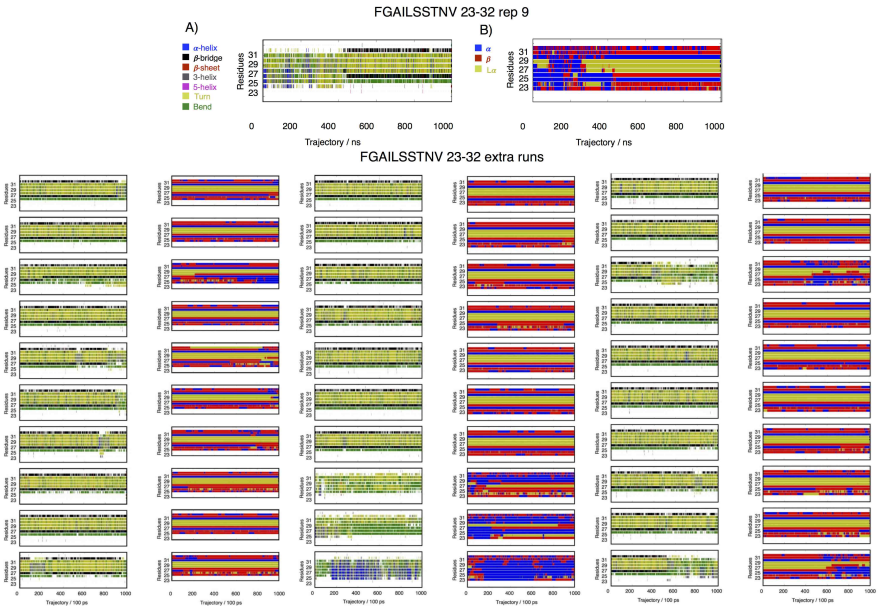


FIG. 8. 30 independent simulation runs of FGAILSSTNV 23-32: A) DSSP plot, timeresolution 100 ps. B) $\{\phi - \psi\}$-state of each residue.

Hierarchy in the conformational ensemble of human islet amyloid polypeptide

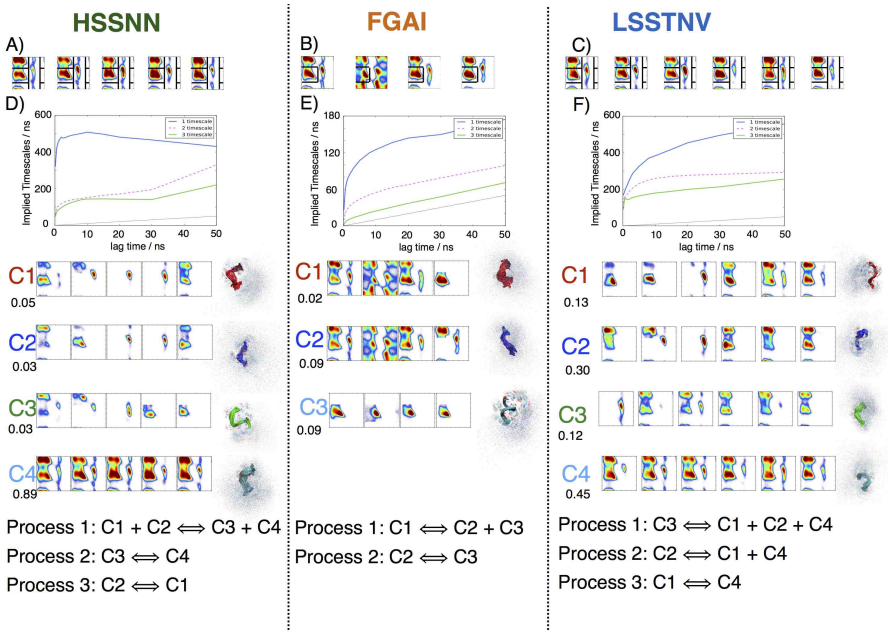


FIG. 9. Sub-models of HSSNNFGAILSSTNV at $\tau = 10$ ns. A, B, C) Discretization of the Ramachandran plane of the residues taken into consideration in each model. D, E, F) Timescales plots and interpretation of the dynamics as exchange of probabilities between long-lived conformation; for each conformation the Ramachandran plots of each residue and a bundle of structures are shown.

Hierarchy in the conformational ensemble of human islet amyloid polypeptide

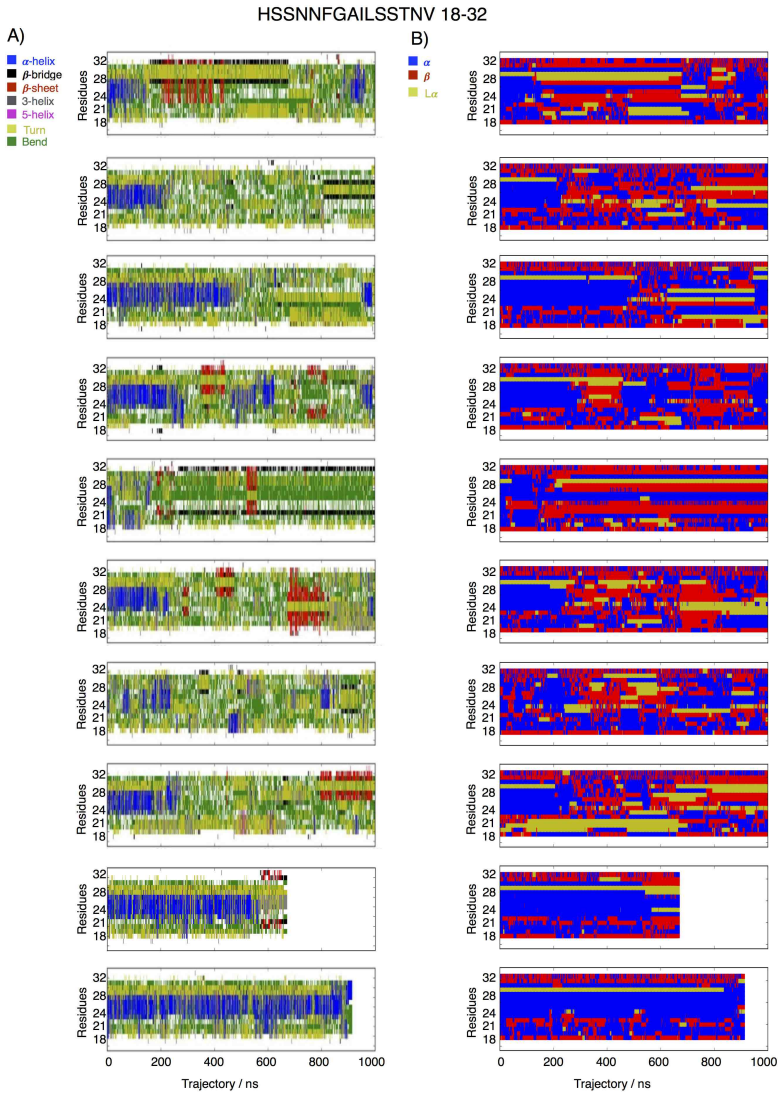


FIG. 10. 10 independent simulation runs of HSSNNFGAILSSTNV 18-32: A) DSSP plot, time resolution 1 ns. B) $\{\phi - \psi\}$-state of each residue.

Hierarchy in the conformational ensemble of human islet amyloid polypeptide

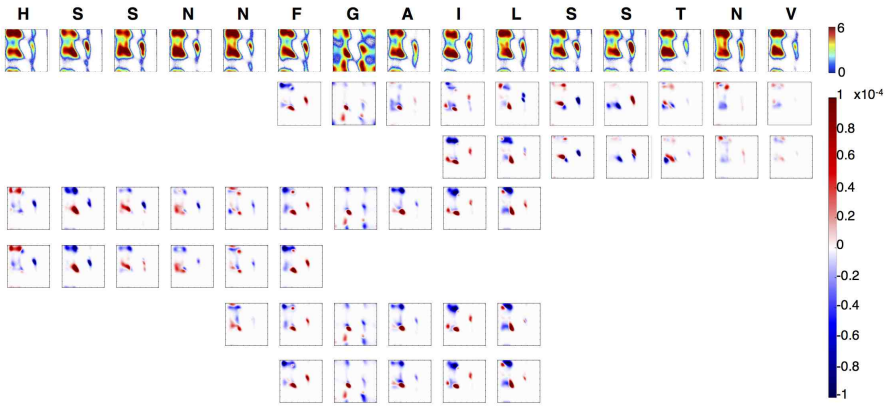


FIG. 11. HSSNNFGAILSSTNV 18-32 backbone dihedral angles distribution distribution (histogram of the logarithm of the counts) and difference plots of $\{\phi - \psi\}$-distributions of each residue for each fragment with respect to HSSNNFGAILSSTNV 18-32 distribution (histogram of the difference of the counts).

Hierarchy in the conformational ensemble of human islet amyloid polypeptide

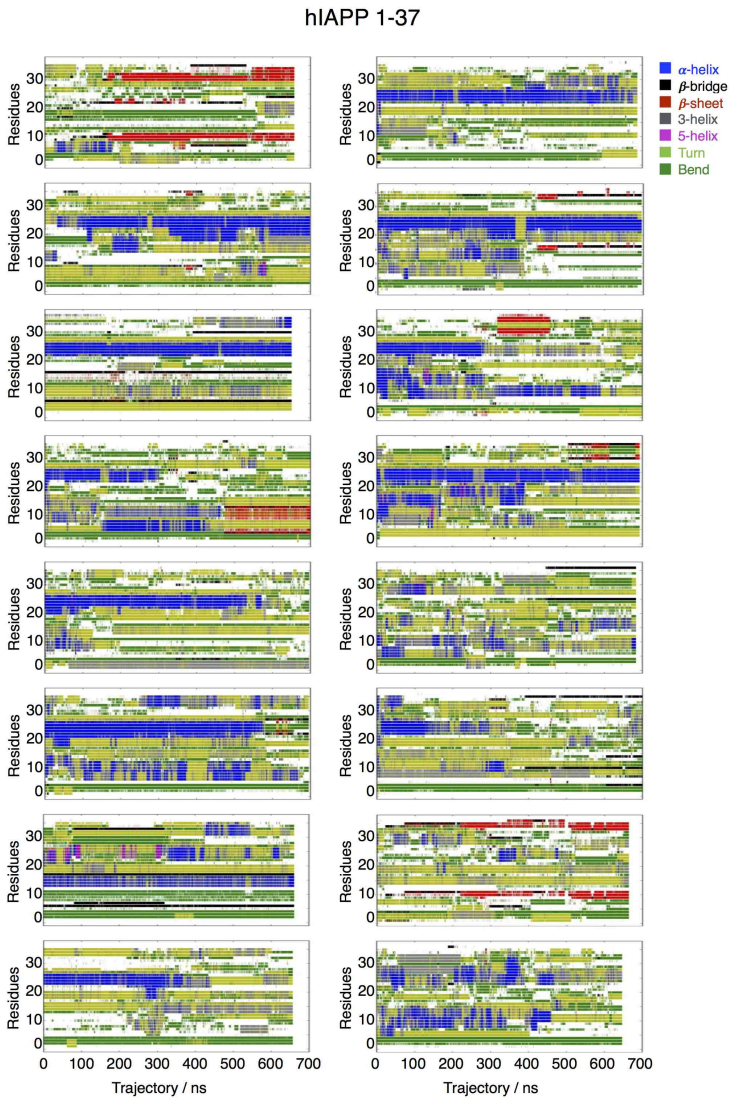


FIG. 12. DSSP analysis of the first 700 ns of the 16 independent simulation runs of HSSNNF-GAILSSTNV 18-32, time resolution 0.1 ns.

Hierarchy in the conformational ensemble of human islet amyloid polypeptide

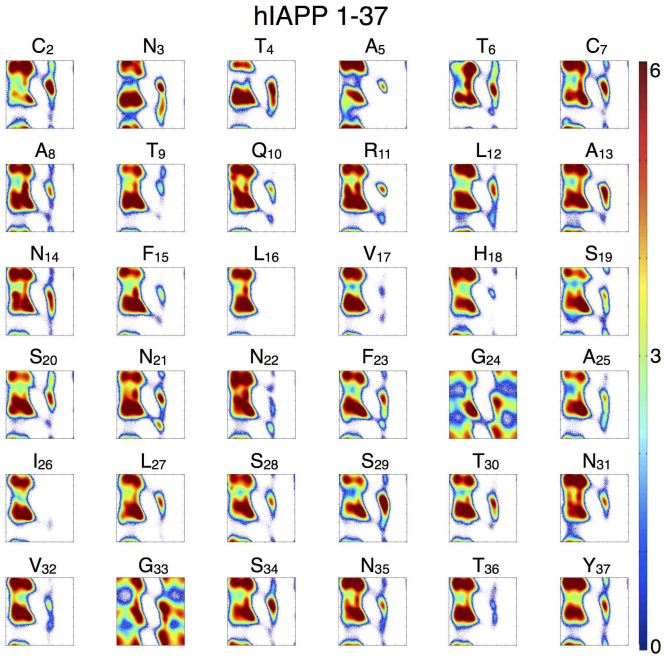


FIG. 13. Ramachandran plots of each residue in hIAPP (histogram of the logarithm of the counts). Residue $K_1$ is not constrained on the $\phi$ backbone angle and is thus not shown.

Hierarchy in the conformational ensemble of human islet amyloid polypeptide

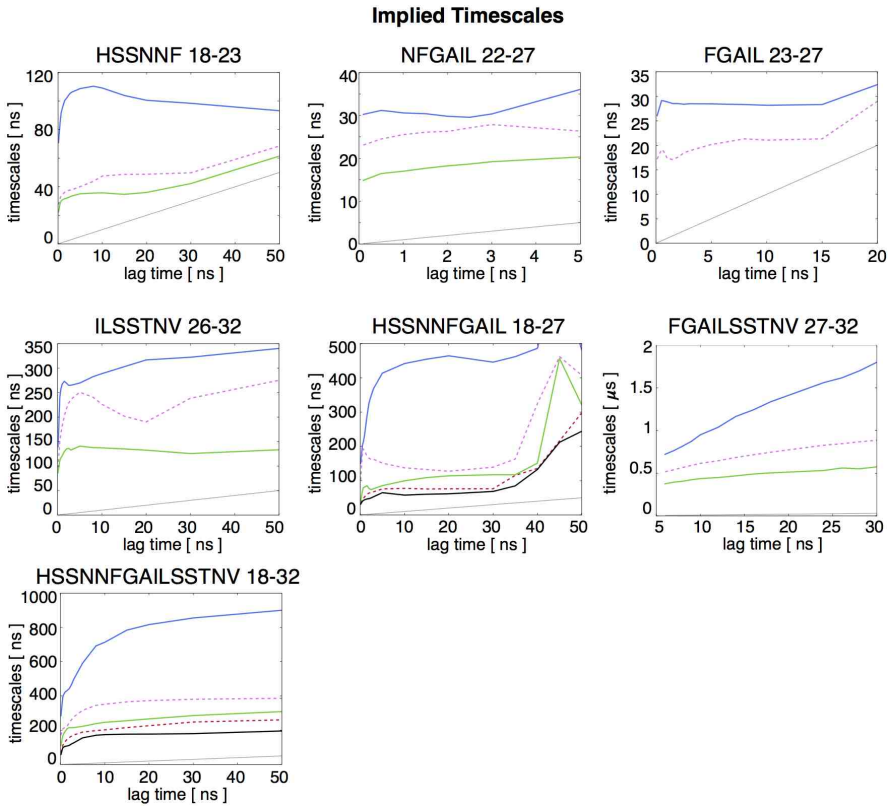


FIG. 14. Implied timescales as a function of the lag times for the MSMs of each peptide.

Hierarchy in the conformational ensemble of human islet amyloid polypeptide

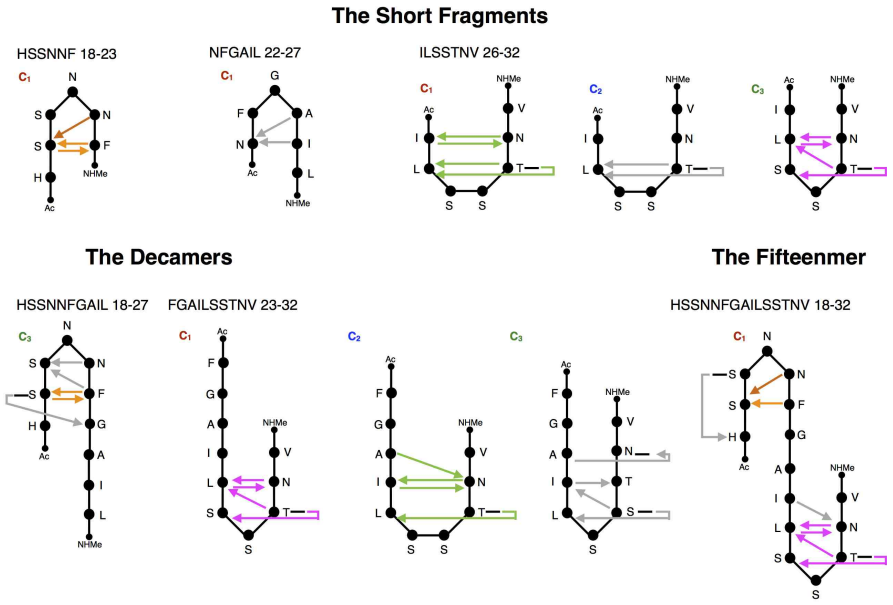


FIG. 15. Most probable (>10%) hydrogen bonds for the long-lived states of the fragments. The direction of the arrow goes from donor to acceptor. The color highlights the the same hydrogen bonds patter.

## REFERENCES

[1]M. Sarich, F. Noé, and C. Schütte, "On the Approximation Quality of Markov State Models," Multiscale Model. Sim. **8**, 1154–1177 (2010).

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

[3]C. Schütte, A. Fischer, W. Huisinga, and P. Deuflhard, "A Direct Approach to Conformational Dynamics based on Hybrid Monte Carlo," J. Comput. Phys. , 146–168 (1999).

[4]P. Deuflhard and M. Weber, "Robust Perron cluster analysis in conformation dynamics," Linear Algebra Appl. **398**, 161–184 (2005).

[5]F. Noé, I. Horenko, C. Schütte, and J. C. Smith, "Hierarchical analysis of conformational

dynamics in biomolecules: transition networks of metastable states." J. Chem. Phys. **126**, 155102 (2007).

[6]F. Vitalini, A. S. J. S. Mey, F. Noé, and B. G. Keller, "Dynamic properties of force fields," J. Chem. Phys. **142**, 084101 (2015).

[7]W. Humphrey, A. Dalke, and K. Schulten, "VMD: visual molecular dynamics." J. Mol. Graphics Modell. **14**, 33–38, 27–28 (1996).

## 3.3 hIAPP 1-37 extended simulations

As mentioned in the previous section, 10.8 $\mu$s of aggregated simulation time of hAIPP 1-37 resulted insufficient for the construction of a converged MSM. Therefore, we decided to extend the simulations up to over 21 $\mu$s. In this section, we present the analysis of such extended simulations. For simulation details please refer to sec. IV A in the manuscript.

### 3.3.1 Convergence of the Simulations

Backbone dihedral angles of an amino acid in a sequence have a very characteristic distribution which can be represented in a Ramachandran plane (fig. 1.2). For extremely flexible systems, such as IDPs, all regions of the Ramachandran plane are expected to be populated.
Using the GROMACS command `g_rama`, we extracted the $\{\phi - \psi\}$-timeseries for each amino acid in the sequence of the extedended simulations. In fig. 3.1, we plot, the per-residue $\{\phi-\psi\}$-distribution, as logarithm of the counts (bins of 1°). Residue $K_1$ does not present a standard backbone $\phi$ angle, therefore it is not included. By comparing it with fig 13 in the supporting information, we notice that new configurations have been explored, where residue $I_{26}$ populates the L$\alpha$ minimum. On the contrary, $L_{16}$ does not present any $\{\phi - \psi\}$-combinations corresponding to the L$\alpha$ minimum. This is not expected to be the consequence of structural property of hIAPP 1-37, because of the high flexibility of the IDP. This thus hints at a not-convergence of the simulations.

It is not trivial, however, to estimate the convergence of MD simulations. Usual convergence checks relay on the variations of properties estimated from the trajectory being within a user-defined cut-off. These are indirect measures and do not necessarily implicate that the full configurational space has been visited by the simulation.
For instance, in fig 3.2.A, we show the variation of the average end-to-end distance with increasing simulation time. The end-to-end distance is computed between the $C_\alpha$ of residue $K_1$ and the $C_\alpha$ of residue $Y_{37}$. After the first 10 million time-steps the average end-to-end distance is converged to c.a. 1.6 nm and there is no dramatic improvement in the confidence interval over the subsequent 11 million time-steps. Analogously the average radius of gyration (fig 3.2.B), which measures the level of compactness of the molecule, is converged after the first one million time-steps to a value of c.a. 1 nm. Fig. 3.2.C shows the convergence of the average number of residues in a $\alpha$-helix. After 10 million time-steps the average number of residues in a $\alpha$-helix is of about three residues, and its standard deviation does not improve with increasing simulation time. Despite these three parameters hinting at a convergence of estimated properties of the MD simulations, in practice these properties show only few aspects of the structural complexity of the system. In fact, to the same value of radius of gyration can correspond multiple structures, characterized by a

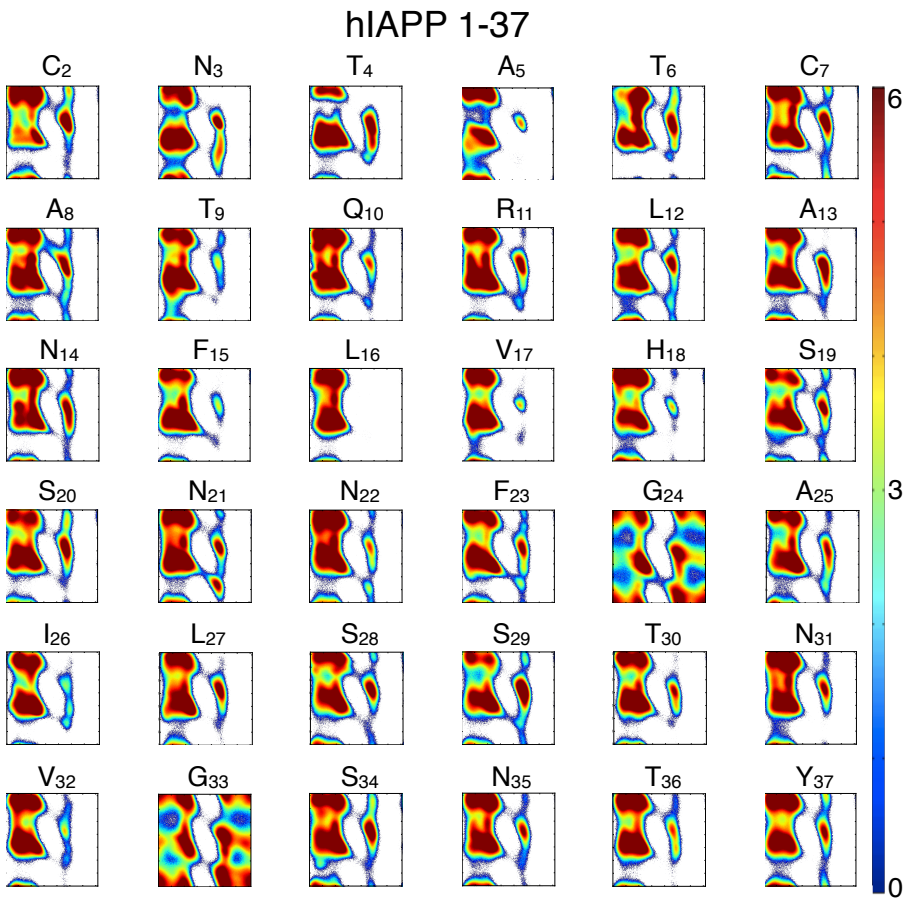


**Figure 3.1**: Per-residue $\{\phi - \psi\}$-distribution as logarithm of the one-degree-grid histogram of the counts over the trajectory.

different number of residues in a $\alpha$-helix configuration (fig. 3.2D).

A much more meaningful measure is the number of uniquely visited microstates, as it is directly related to the exploration of the state space. As explained in sec. IV D in the paper, each configuration explored by the simulations is mapped into a numbered-string, corresponding to the combinations of backbone angles of each residue ($\alpha = 0$, $\beta = 1$, L$\alpha$=2). The mapped string is the microstate associated to the configuration. Such three-states per-residue characterization is sufficiently refined for a meaningful description of different configurations, without being computationally unfeasible. The unique configurations explored by the extended trajectory are 114903. Fig. 3.3.A shows the number of unique states (normalized with respect to the total number of unique states) visited with increasing simulation time. It is thus evident that the extra simulations are exploring new parts of the configurational space and that convergence is not yet reached.

However, the newly discovered states might be of small relevance, if their equilib-

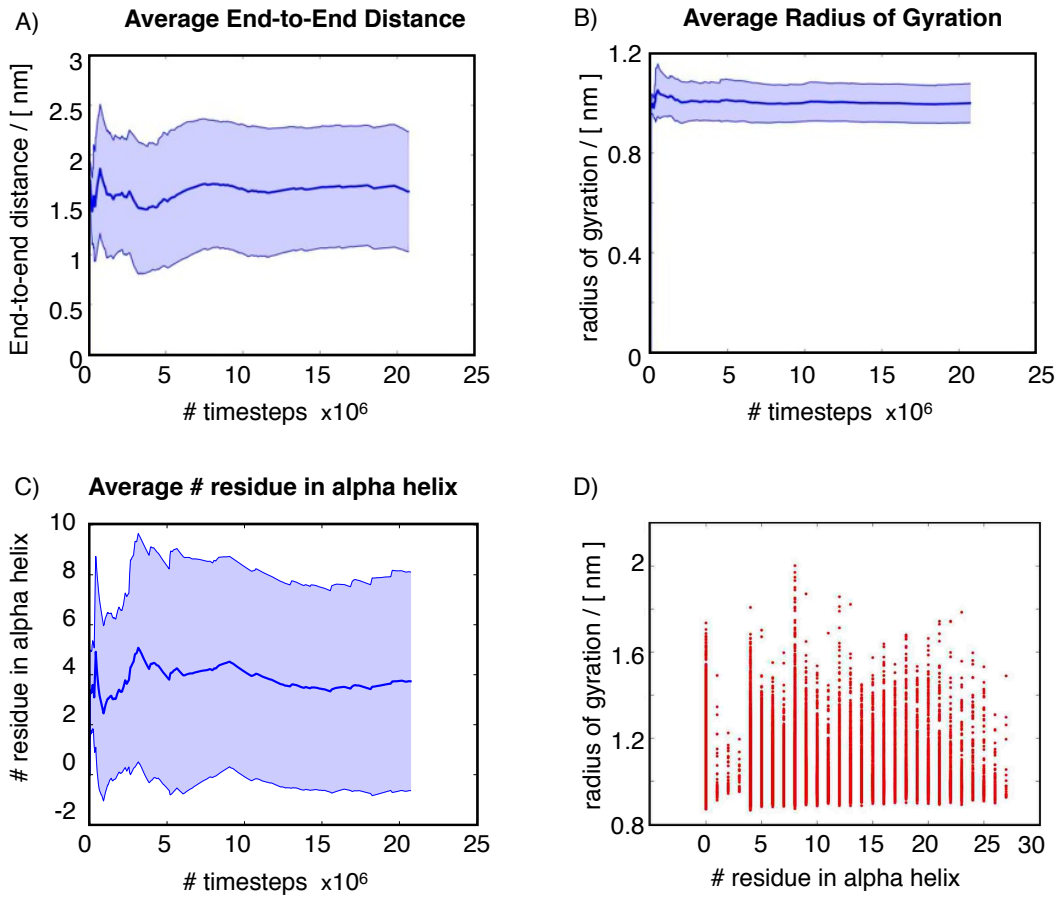


**Figure 3.2**: A) Average end-to-end distance with respect to simulation time. B) Average Radius of gyration with respect to simulation time. C) Average number of residues in a $\alpha$-helix with respect to simulation time. D) Radius of Gyration with respect to number of amino acids in a $\alpha$-helix.

rium probability is very small. To verify it, we computed the weight of each unique state by counting the number of visits throughout the trajectory. Note that this would be equivalent to the equilibrium probability of the state if the simulations are converged. Using such weights we evaluate the percentage of equilibrium probability visited with increasing simulation time. Fig 3.3.B shows an almost constant increase in the portion of equilibrium distribution visited in time, i.e. the contribution adduced by each new explored state is equivalently relevant.

This can be explained by looking at the weights of each state (fig 3.4.A). Only 84 out of 114903 states have a probability of over 0.1%, and only one state is 1% probable. Therefore, each new visited state adds a small and almost comparable contribution to the total equilibrium probability. Fig 3.4.B shows the seven states, whose probability is over 0.5%. Four of those structures present the FGAIL 23-27 fragment in the $\alpha$-helix configuration, whereas the remaining three show the helical

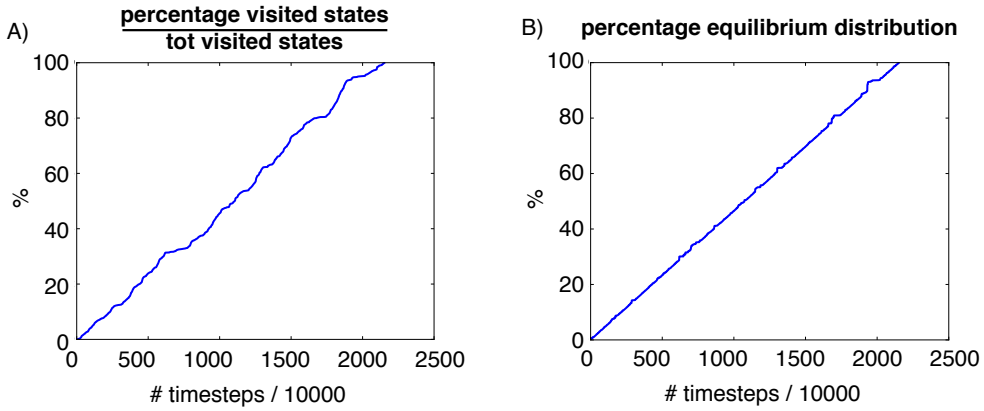


**Figure 3.3**: A) Percentage of unique visited states with respect to simulation time. B) Percentage of equilibrium distribution with respect to simulation time.

motive shifted along the sequence. The preference of the FGAIL 23-27 fragment for the helical motif had already arisen in the previous work and is confirmed in the extended simulations analysis. As shown in fig 3.4.C, residue $N_{21}$- $L_{27}$ have a significant probability of being in a $\alpha$-helix, with residues $F_{23}$ - $A_{25}$ showing a higher helical propensity than random coil. In comparison with fig 12 in the manuscript, such propensity is not as pronounced.

This indicates that the extended simulations explore new configurations where FGAIL 23-27 is not mainly helical. It is thus interesting to check wether the configuration space of the fragment is converged. On the line of argument followed for fig 3.3, we evaluate the percentage of unique states visited by the fragment with increasing simulation time (fig 3.5.A). After 5 $\mu$s of simulation, i.e. in the $5^{th}$ independent run, a new portion of the configuration space is discovered and after 17 $\mu$s the percentage of unique states visited is at convergence. Obviously this is not a full proof of complete exploration of the configuration space, as it cannot be demonstrated that there is not an ulterior energy barrier yet to be overcome. This issue is, however, common to all simulation studies and of non-trivial solution.
Fig 3.5.B shows the percentage of equilibrium distribution (relative to the fragment in question) visited throughout the trajectory, where with equilibrium distribution we intend the count of the visits of each unique state of the fragment. Comparing fig 3.5.B and fig 3.5.A, we evidence that the states visited after the first 10 $\mu$s of simulation contribute marginally to the equilibrium distribution. Or better, the states visited in the second half of the simulations are visited only a few times by the trajectory, either because their true equilibrium probability is small, or because of limited statistics.
Fig 3.5.C shows the probability of each unique state explored by the FGAIL 18-23 fragment. As expected, such probability distribution is peaked around the helical state, which is populated 50% of the time. On the contrary, a state where all residues assume a $\beta$-sheet conformation is only 0.7% probable.

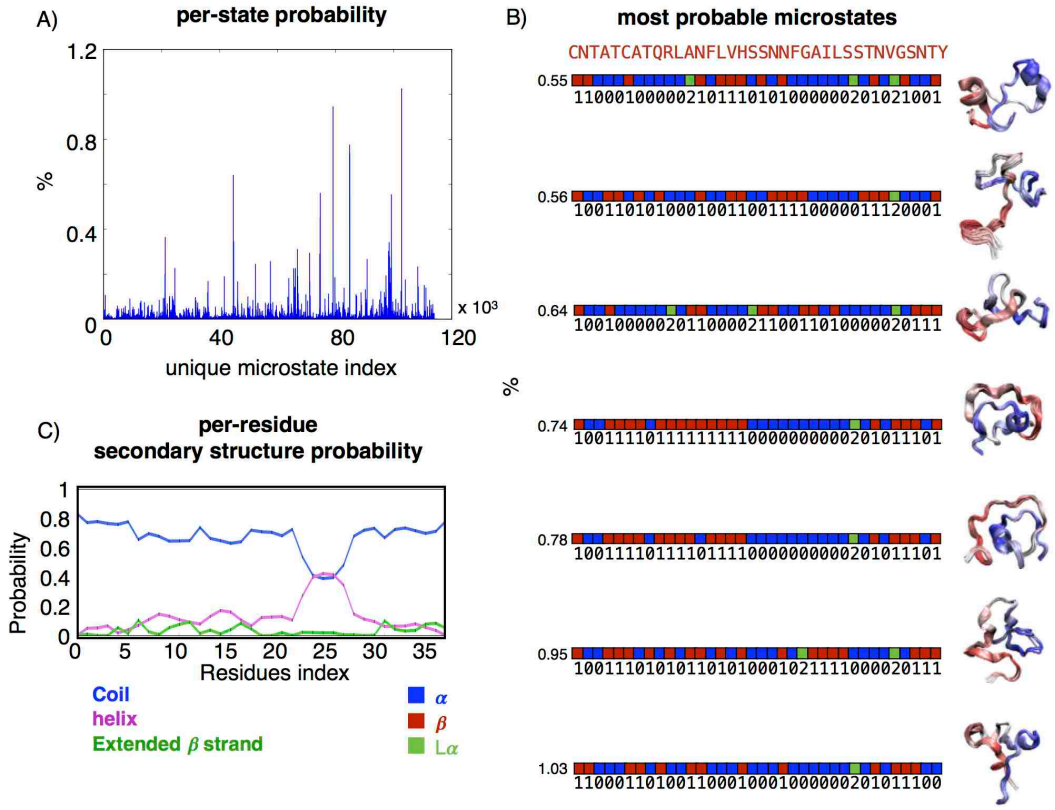


**Figure 3.4**: A) Per state equilibrium probability. B) Seven most probable states as color-coded strings and corresponding structures. C) Secondary structure probability per residue.

Despite the results for FGAIL 23-27 configurational space exploration, fig 3.3 demonstrates that the simulations are not yet converged for the full-length system. It is therefore interesting to evaluate at which segment-length the convergence check breaks. We investigate this in fig 3.6, where segments of progressively increasing length are taken into consideration. Segments NFGAILS 22-28 and NNFGAILSS 21-29 both reach the 100% of unique visited states in 17 $\mu$s and their percentage of equilibrium distribution is also at convergence in the same amount of simulation time. However, segment HSSNNFGAILSSTNV 18-32 presents a behavior in line with the full-length system, i.e. the number of uniquely visited states is constantly increasing with extended simulation time and the weight of each newly discovered state is comparable.

From the analysis of the convergence of the simulation it is thus clear that we cannot claim to have explored the full configurational space, and therefore we do not have sufficient data for the construction of a meaningful MSM. As future prospective, we plan to further extend the simulations, to achieve sufficient statistics of the rare transitions for inferring a meaningful model of the dynamics.

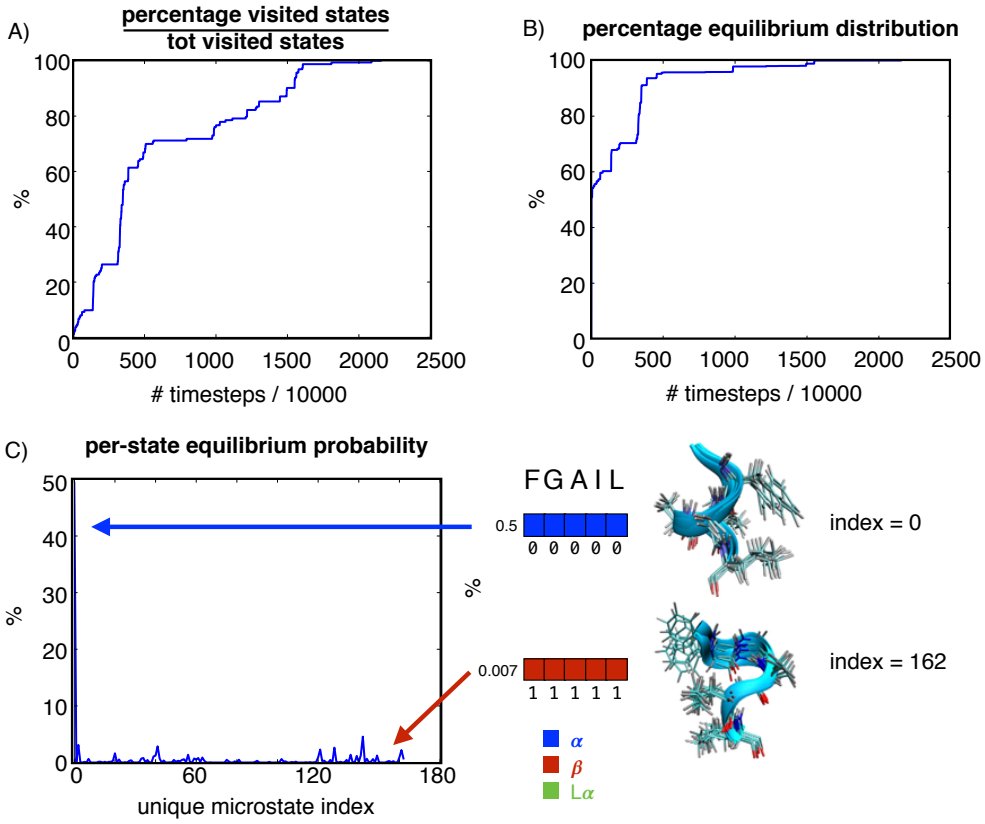


**Figure 3.5**: A) Percentage of unique visited states by the FGAIL 23-27 fragment with respect to simulation time. B) Percentage of equilibrium distribution of the FGAIL 23-27 fragment with respect to simulation time. C) Per state equilibrium probability of the FGAIL 23-27 fragment. The states corresponding to full $\alpha$-helix and full $\beta$-sheet are marked and corresponding structures are shown.

### 3.3.2 Analysis of the Extended Simulations

Despite the not-convergence of the simulation data, it is possible to deduce interesting properties of the system. For instance, the correlation between the backbone dihedral angles distributions of residue-pairs can be estimated via the normalized mutual information (NMI).

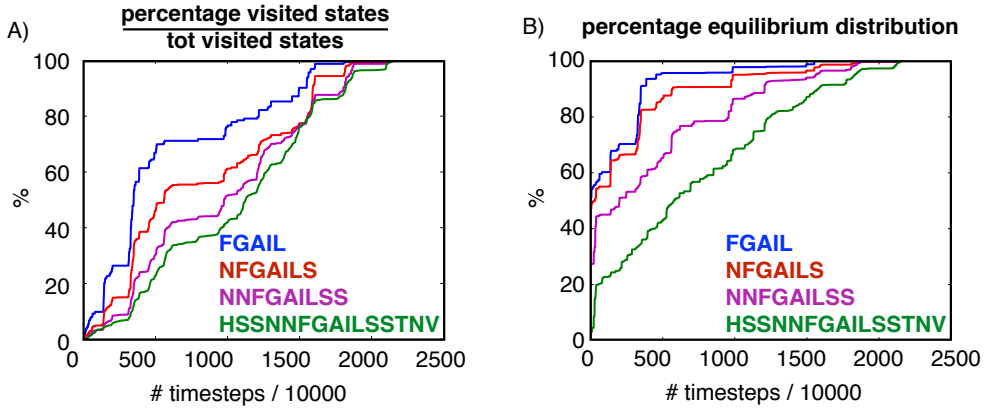


**Figure 3.6**: A) Percentage of unique visited states with respect to simulation time. B) Percentage of equilibrium distribution with respect to simulation time.

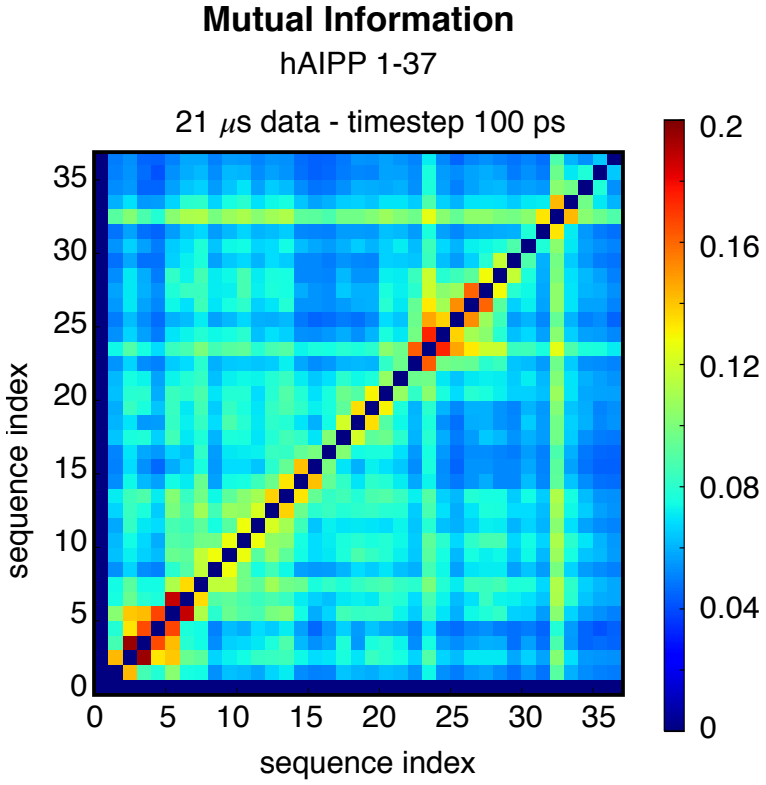


**Figure 3.7**: Mutual information. Plot obtained using 21 $\mu$s and 100 ps downsampling.

NMI is a measure of the difference between the real joint probability density the case of uncorrelated distributions. It can be used to highlight the formation of transient secondary structures. For further discussion of the NMI please refer to sec. IV B in the manuscript. Fig. 3.7 shows the pair-wise NMI for the extended simulations. Note that, for computational tractability, the simulation data has been down-sampled by a factor of 100 time-steps. Such downsampling has an

effect on the NMI as the entropy associated to the independent variables is smaller. As a consequence, the NMI between two residues has a higher value the bigger the downsampling factor, which explains the higher pair-wise correlation of fig. 3.7 with respect to fig 11 E in the manuscript. Qualitatively however fig. 3.7 and fig 11 E in the manuscript are in agreement, and the secondary structures elements evidenced by both analysis are the same: (*i*) a strong correlation between residues one to seven, induced by the cysteine bond between residues $C_2$ and $C_7$; (*ii*) the formation of helical structures induces high NMI values in the regions 7-19 and 23-29; (*iii*) residues 5-14 show long-distance communication with residues 23-34; (*iv*) $G_{24}$ and $G_{33}$ present strong correlation with the distribution of all the other residues in the sequence, which is to be related to the peculiar glycine $\{\phi - \psi\}$-distribution.

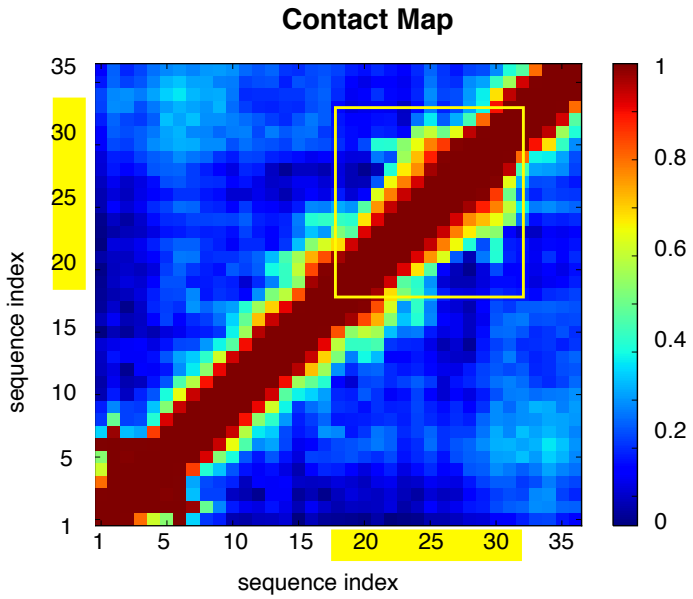


**Figure 3.8**: Contact Map. Average contact probability throughout the trajectory. Two residues are defined in contact if the C$\alpha$-C$\alpha$ distance is smaller than 1 nm.

The long-range interaction identified by the NMI are also found in the contact map analysis (fig. 3.8). In this analysis we consider two residues in contact if their C$\alpha$-C$\alpha$ distance is smaller than 1 nm. For each time-step a binary contact map is evaluated using the `md.compute_contacts` built-in function of MDTraj python-based software package[98]. Subsequently we compute the average contact map throughout the simulation data. As shown in fig. 3.8, neighbor residues are in contact, as well as residues 1-7 due to the cysteine bond. The high propensity of FGAIL 23-27 to form $\alpha$-helices is shown in the high contact probability of residues 23-27. Moreover, configurations with residues 5-10 in proximity of residues 34-37 are found in 40% of the data. Residue 30 is also likely to form contacts with residues 21-27.

Another standard methodology to analyze protein configuration is to assign configurations to structure elements via the Define Secondary Structure of Proteins algorithm (DSSP) [99]. The DSSP algorithm assigns a secondary structure element to each residue in the sequence, based on hydrogen bonds formation. Multiple types of secondary structures are distinguished: $3_{10}$ (grey in this analysis), $\alpha$ (blue) and $\pi$ (magenta) helices, corresponding to repetitive sequences of hydrogen bonds between residues respectively three, four or five positions apart in the sequence; $\beta$-sheets (red) corresponds to strands of residues forming the typical hydrogen pattern; $\beta$-bridges (black) are localized $\beta$-sheets hydrogen bonds; turns (light green) are single hydrogen bonds typical of helices; bends (dark green) are regions with high curvatures and the only element not to depend on hydrogen bonds. Figs 3.9-3.12 show DSSP plots of all replicas in sniplets of 100 ns and resolution of 100 ps. It can be noted that multiple configurations are explored, spanning between helical and $\beta$ structures. Which are the residues involved in the $\beta$ structures, however, is not a constant, indicating that the simulations visit varied $\beta$ structures This is another confirmation of the necessity of extra data.
As already shown in the previous study, it is of interest to compare standard DSSP plots and the three-states per residue discretization time-series (fig 12 in the supporting information). The discretization is only based on dihedral angles and does not consider the stabilizing effect of hydrogen bonds. Fig. 3.13 presents DSSP-like plots per replica, based on the three-states discretization of each residue. The residues of the FGAIL 23-27 fragment are often in $\{\phi, \psi\}$-combinations that correspond to a $\alpha$-helix, in agreement with fig. 3.5.C. The residues involved in the bend structures identified in the shorter fragments are also present in the full-length system. Occasionally those bend structures are substituted by more stable $\beta$-bridges connected by a loop. The simulations also explore other long range $\beta$-bridges, not observed in the short fragments.

### 3.3.3 Conclusive Reamarks

In this section we have analyzed extended simulations of hAIPP 1-37. We have accumulated an aggregated simulation time of over 21 $\mu$s. The exploration of the configurational space is, however, only partial, hinting at the necessity of collecting additional data. We have here presented different possible indirect measurements of the convergence of the simulations, pointing out the limitation of using structural properties that are not directly linked to the configurational space. We have also showed that the number and statistical weight of the configurations visited by the simulations is a better measure for testing the convergence of the simulations.

Despite the limited convergence of the simulations, important findings can still be obtained. HAIPP 1-37 is an IDP, but is not a random coil in solution: it explores a variety of conformations that could be stabilized upon contact with different binding partners. We have used the DSSP algorithm to investigate which conformations are assumed by the system. These conformations involve $\beta$-structures, which could act

as precursors to amyloid formation, as well as helical structures, which the peptide is known to assume in presence of membranes. A conformational selection mechanism could thus stabilize hAIPP 1-37, according to the binding partner.
The FGAIL 23-27 fragment, which is the core of the so-called amyloidogenic region, has a preference for the the helical configuration, but is also involved in the formation of long-range $\beta$-structures. It is worth to notice though that the preference of the helical configuration might be an effect of the force field, as it is known that different force fields produce different secondary structure distributions [100].

As the simulations are not converged, it is impossible to construct a meaningful model of the dynamics. In general it is not trivial to construct a dynamical model of an IDP. IDPs conformational ensemble is in fact extremely vast and obtaining sufficient statistics to model the transitions between conformations requires much data. It is our plan to extend further our data set, in order to build a model of the dynamics of hAIPP 1-37.

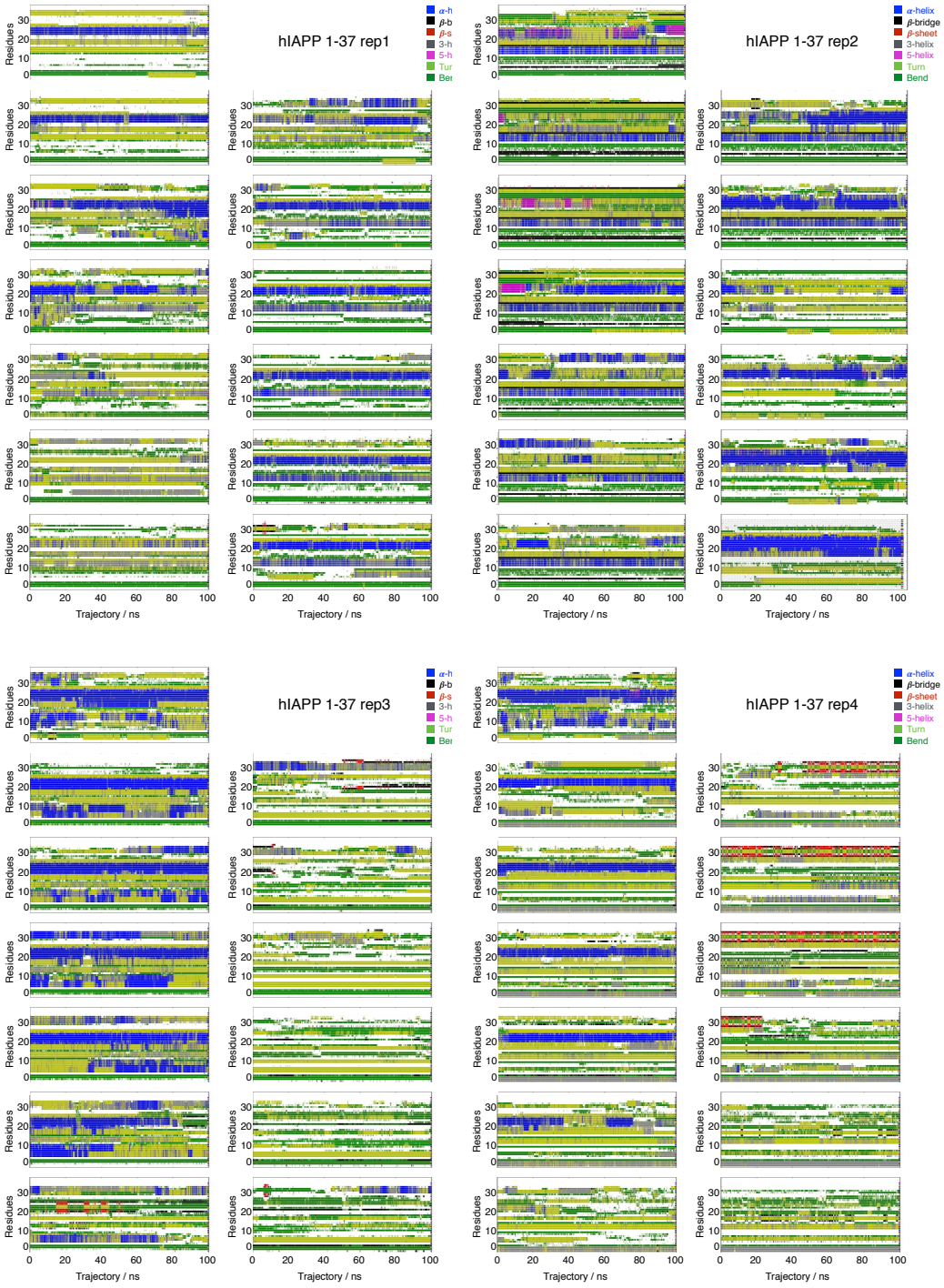


**Figure 3.9**: DSSP plots of independent runs 1-4 in sniplets of 100 ns.

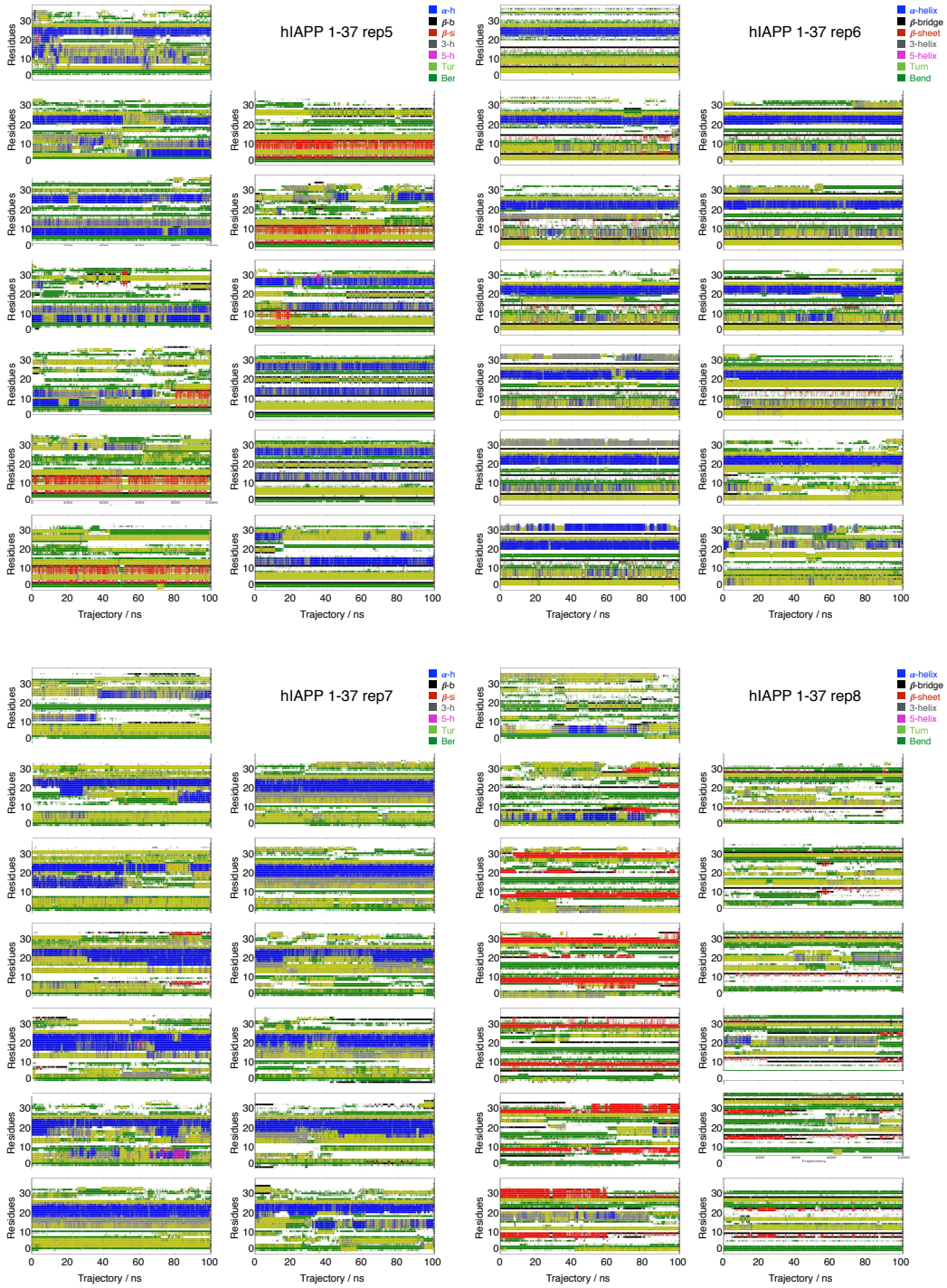


**Figure 3.10**: DSSP plots of independent runs 5-8 in sniplets of 100 ns.

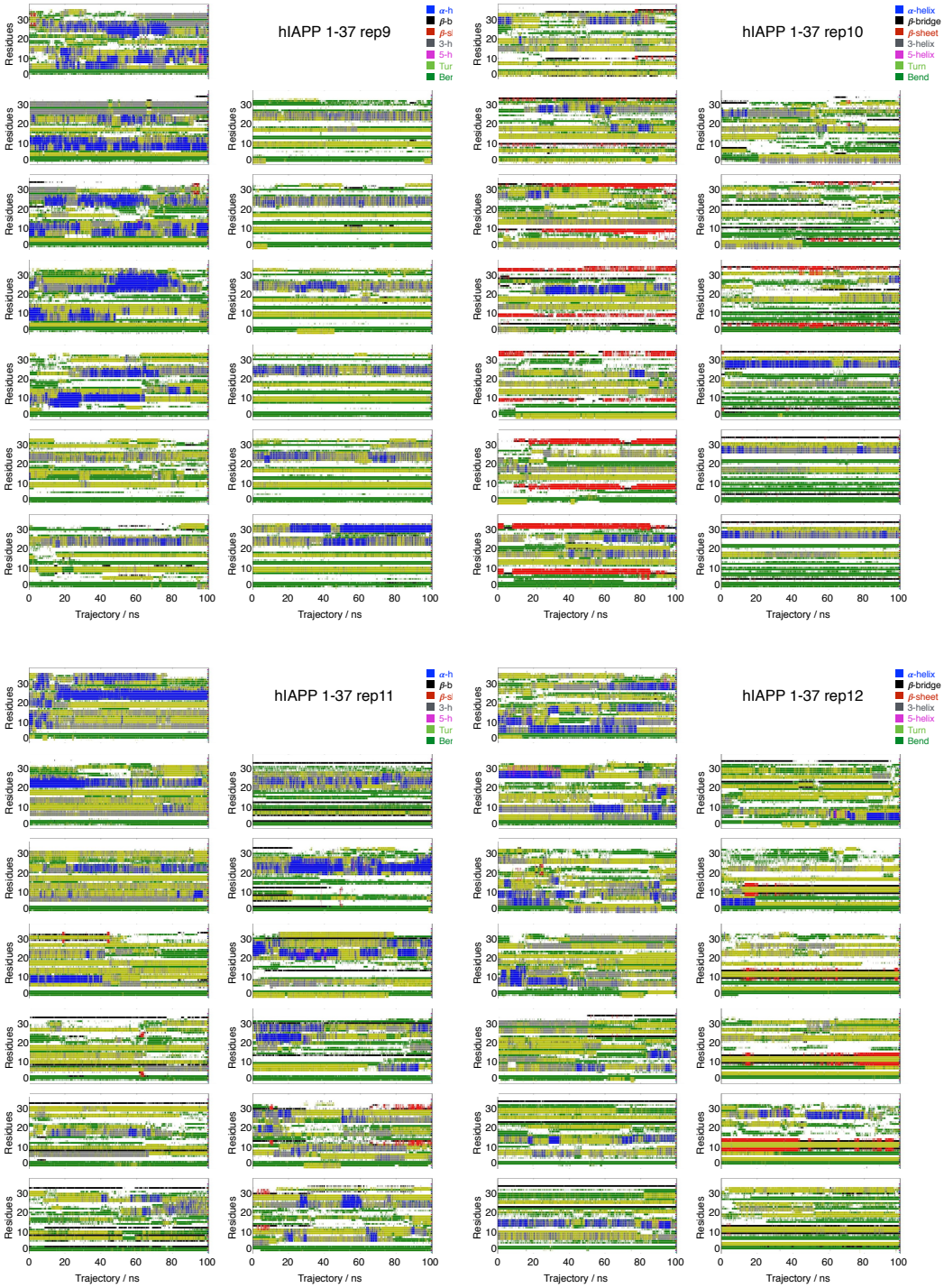


**Figure 3.11**: DSSP plots of independent runs 9-12 in sniplets of 100 ns.

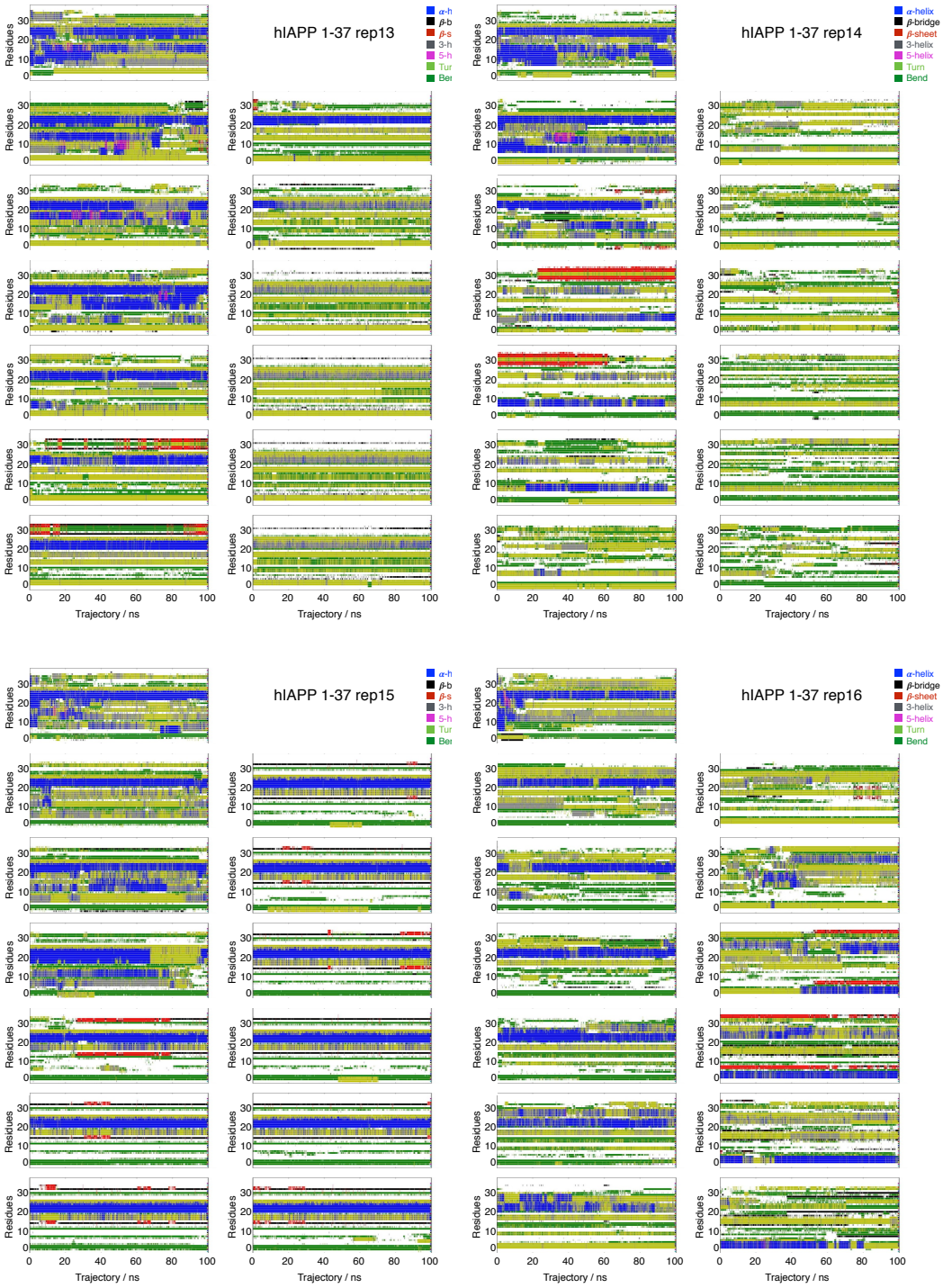


**Figure 3.12**: DSSP plots of independent runs 13-16 in sniplets of 100 ns.

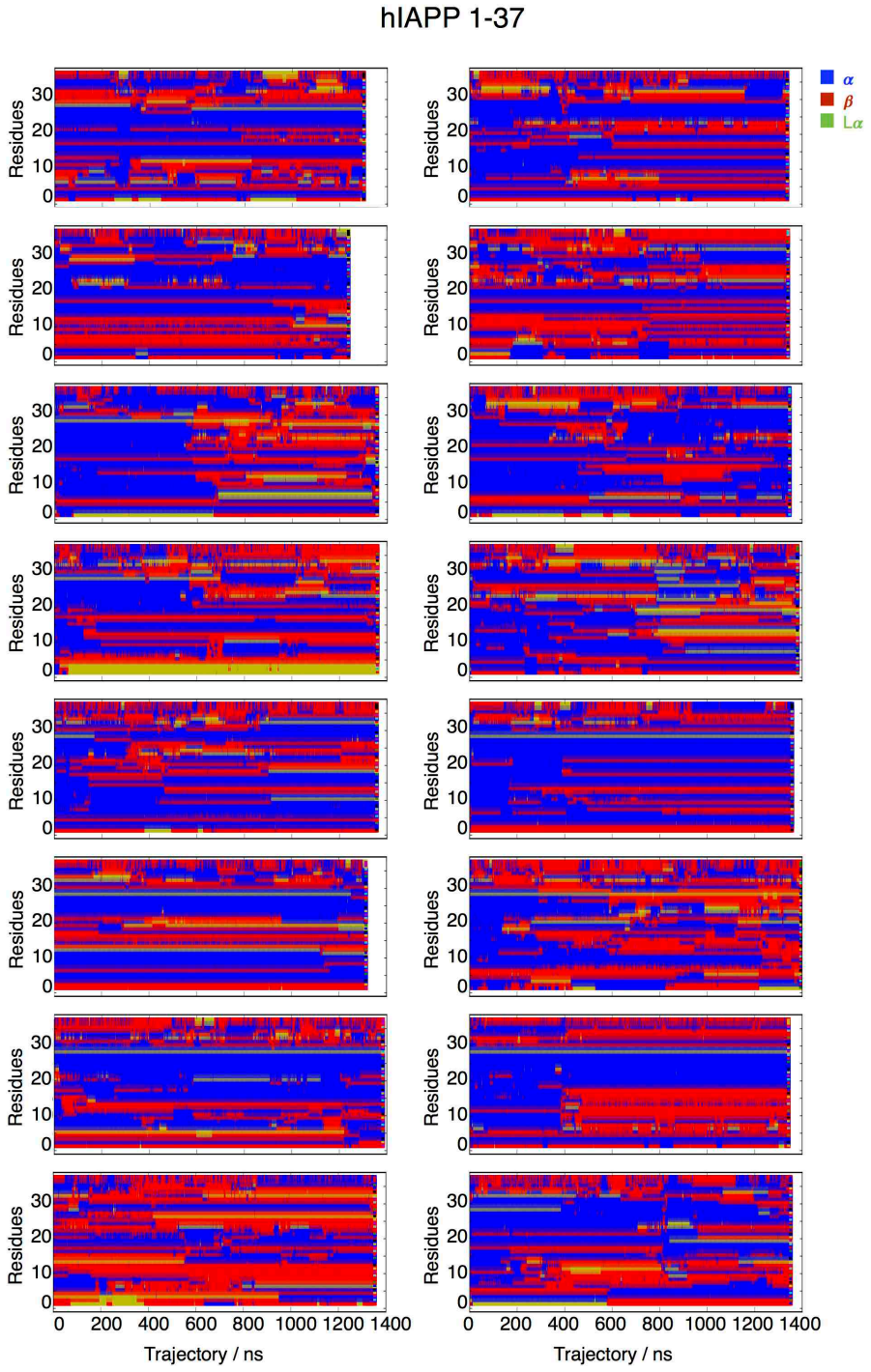


**Figure 3.13**: DSSP-like plots for the 3-states discretization.

Chapter 4

# Basis Set Description of Peptides' Dynamics

The previous chapter focused on small peptides, whose intrinsic disorder resulted in non-trivial kinetic models construction. Substantial user-interaction was therefore necessary for the construction and interpretation of a converged MSM, especially for the longer sequences, where a regular Ramachandran-based grid resulted into a computationally unfeasible number of states. The hierarchical states definition, despite providing a useful description of the dynamics in the presented cases, relies on considerable arbitrariness. How to combine residue dynamic, without being driven by prior knowledge of the system and experience, remains a subjective choice, which might introduce errors in the model. Therefore, it becomes evident the need of overcoming the traditional MSM, in favor of simulation data independent methods, capable of identifying the interesting dynamic modes without relying on the definition of an excessive number of discrete states.

The recently introduced variational approach to conformation dynamics (VAC) [75, 76] has opened up a route to the construction of kinetic models based on basis functions, in place of crisp-states as in standard MSM. VAC describes a general approach for the combination of basis functions, so to find the best approximation of the true eigenfunctions of the propagator given the basis set. VAC also allows to systematically control the approximation quality of the model, by varying the basis set size. For further details please refer to section 2.3.
A crucial step in the application of VAC is the definition of the basis functions. If the basis functions model faithfully the features of the energy landscape, only a small number of basis functions would be relevant to approximate each of the dominant eigenfunctions of the propagator. Therefore, a good model can be obtained with a smaller number of basis functions than discrete states in a standard MSM. Coupling VAC with optimization algorithms, such as a Tensor-Train approach [101], hints at the applicability of the method for large size systems, by automatically define the optimal sub-set of basis functions.

In this chapter we introduce and test a basis set for a variational description of peptides' kinetics. The proposed basis set is constructed by combining local residue-centered kinetic modes that are obtained from a library of kinetic models of terminally blocked amino acids. Such residue-centered kinetic modes are system-independent, therefore the basis functions depend only on the sequence. Moreover,

such definition of the basis functions allows for a direct interpretation of the slow kinetic modes, without an additional clustering in the space of the dominant eigenfunctions. Additionally, changes in the conformational kinetics due to point mutations can be directly quantified, as the basis functions definition allows for direct model comparison.

F.Vitalini, F. Noé and B. G. Keller; A Basis Set for Peptides for the Variational Approach to Conformational Kinetics; Journal of Chemical Theory and Computation, 11, 3992-4004; 2015.

dx.doi.org/10.1021/acs.jctc.5b00498

F.Vitalini, F. Noé and B. G. Keller; A Basis Set for Peptides for the Variational Approach to Conformational Kinetics; Journal of Chemical Theory and Computation, 11, 3992-4004; 2015.

http://dx.doi.org/10.1021/acs.jctc.5b00498

Chapter 5

# Dynamic Properties dependance on Force Fields

The significance of a dynamic model derived form MD data depends on to what extent the MD simulations are representative of the true dynamics. Current MD force fields are not parametrized against dynamic properties, however, the recent developments in *in silico* computations permit not only to calculate equilibrium populations of conformations, but also to compute the transition rates between such conformations. Assessing the reliability of MD force fields in capturing dynamic properties is therefore called into question.

In this chapter MSM are used to compare dynamic models of test systems simulated with different force fields. Representative of each of the major force field families are used to evaluate how do the dynamic properties of capped amino acids and test peptides differ in simulations, where all parameters except the force field are identical.
A significant dependance of timescales and conformational changes is evinced by the analysis, suggesting that dynamic properties should be taken into consideration in the development of future force fields. Moreover, we propose MSM based on a regular discretization of the backbone dihedral angle space as a tool for inferring and comparing dynamic properties of force fields.

F.Vitalini, A. S. J. S. Mey, F. Noé and B. G. Keller; Dynamic Properties of Force Fields; Journal of Chemical Physics, 142, 0804101; 2015.

http://dx.doi.org/10.1063/1.4909549

F.Vitalini, A. S. J. S. Mey, F. Noé and B. G. Keller; Dynamic Properties of Force Fields; Journal of Chemical Physics, 142, 0804101; 2015.
http://dx.doi.org/10.1063/1.4909549

Chapter 6

# Basis Set Library of commonly used Force Fields

---

The accuracy of the basis set introduced in chapter 4 depends also on how well the residue-centered basis functions represent the dynamic modes within the peptide sequence. Given that there is quite a discrepancy between the dynamics simulated by different force fields, it is important to match each simulation with the residue-centered basis functions of the corresponding force fields. Hereby we introduce the simulations of the twenty encoded amino acids in combinations with the force fields introduced in chapter 5. Such data has been made publicly available on a `ftp` repository. Finally, a library of the residue-centered basis functions for the twenty encoded amino acids and the different force fields combinations is presented.

F.Vitalini, F. Noé and B. G. Keller; Molecular Dynamics simulation data of the twenty encoded amino acids in different force fields; Data Br. 7 (2016) 582?590, 10.1016/j.dib.2016.02.086.

http://dx.doi.org/10.1016/j.dib.2016.02.086

## 6.2 Residue Centered Basis Functions Library

In this section we present a force field dependent library of the residue-centered basis functions introduced in chapter 4. For each residue, a MSM was constructed (lag time $\tau$=50 ps for all residues in all force fields except Glycine in AMBER03 and all residues in GROMOS43a1 where $\tau$=20 ps). The lag time is chosen such as the first three eigenvectors represent stable and clearly defined processes. The first left eigenvector (equilibrium distribution $\pi$) and the first three right eigenvectors (RBVs) are shown (figs. 6.1 to 6.5 for AMBER99SB-ILDN; figs. 6.6 to 6.10 for AMBER03; figs. 6.11 to 6.15 for CHARMM27; figs. 6.16 to 6.20 for OPLS-AA; figs. 6.21 to 6.25 for GROMOS43a1).

For further details on the MD simulations and the MSM construction, refer to the Method section in chapter 4.

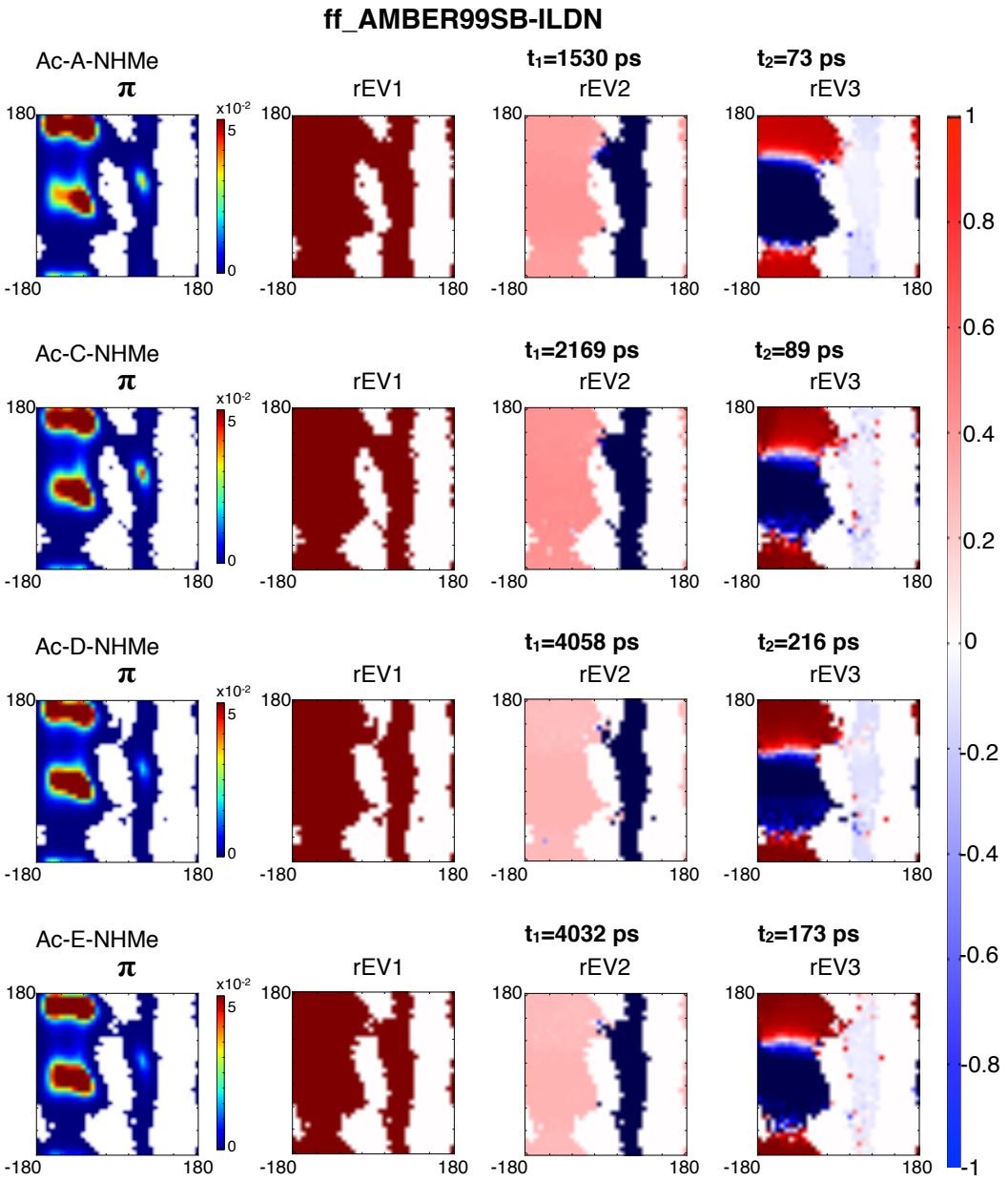


**Figure 6.1**: First left eigenvector ($\pi$), corresponding to the equilibrium distribution, and right eigenvectors 1-3, corresponding to the RBVs associated to the amino acid.

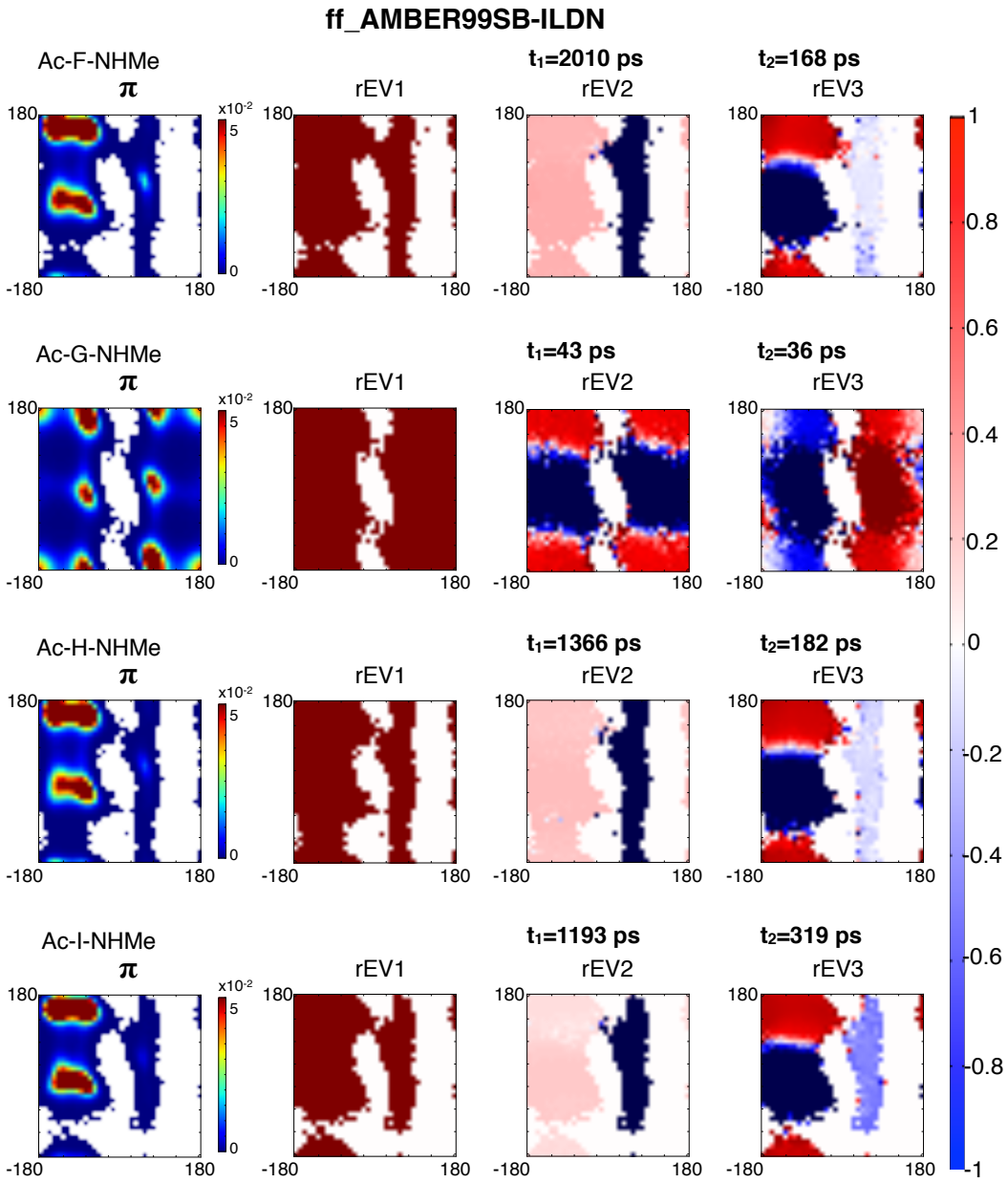


**Figure 6.2**: First left eigenvector ($\pi$), corresponding to the equilibrium distribution, and right eigenvectors 1-3, corresponding to the RBVs associated to the amino acid.

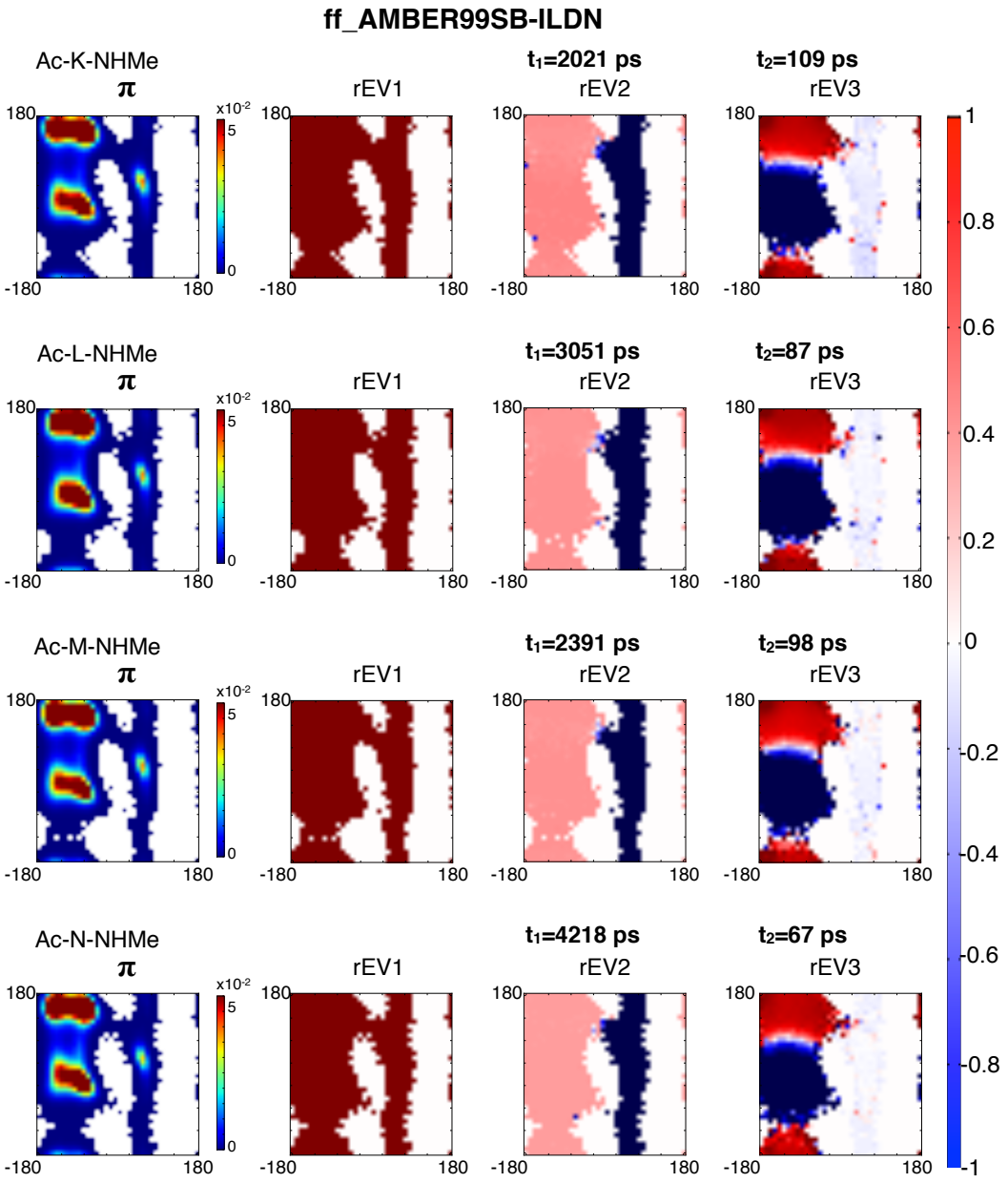


**Figure 6.3**: First left eigenvector ($\pi$), corresponding to the equilibrium distribution, and right eigenvectors 1-3, corresponding to the RBVs associated to the amino acid.

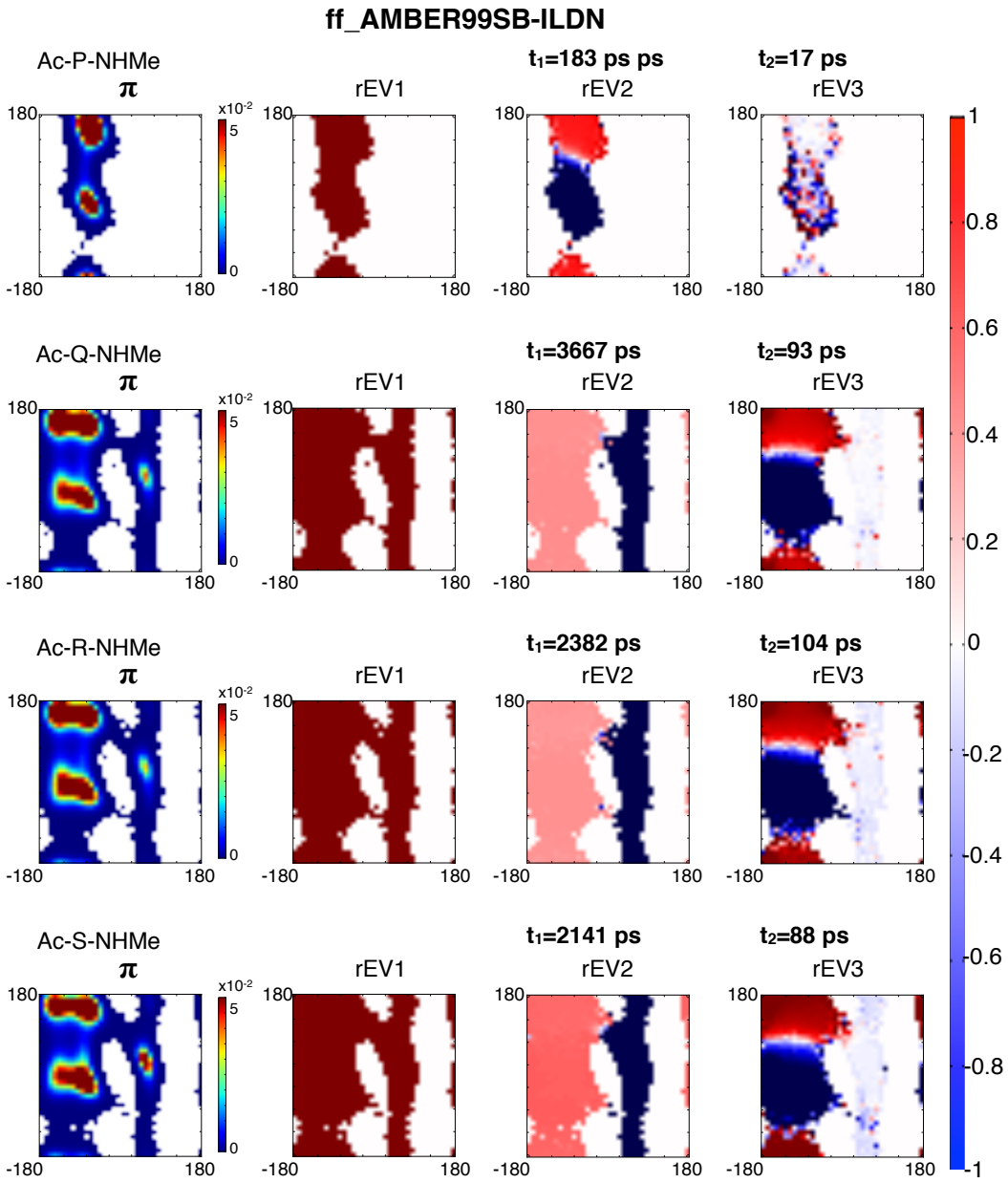


**Figure 6.4**: First left eigenvector ($\pi$), corresponding to the equilibrium distribution, and right eigenvectors 1-3, corresponding to the RBVs associated to the amino acid.

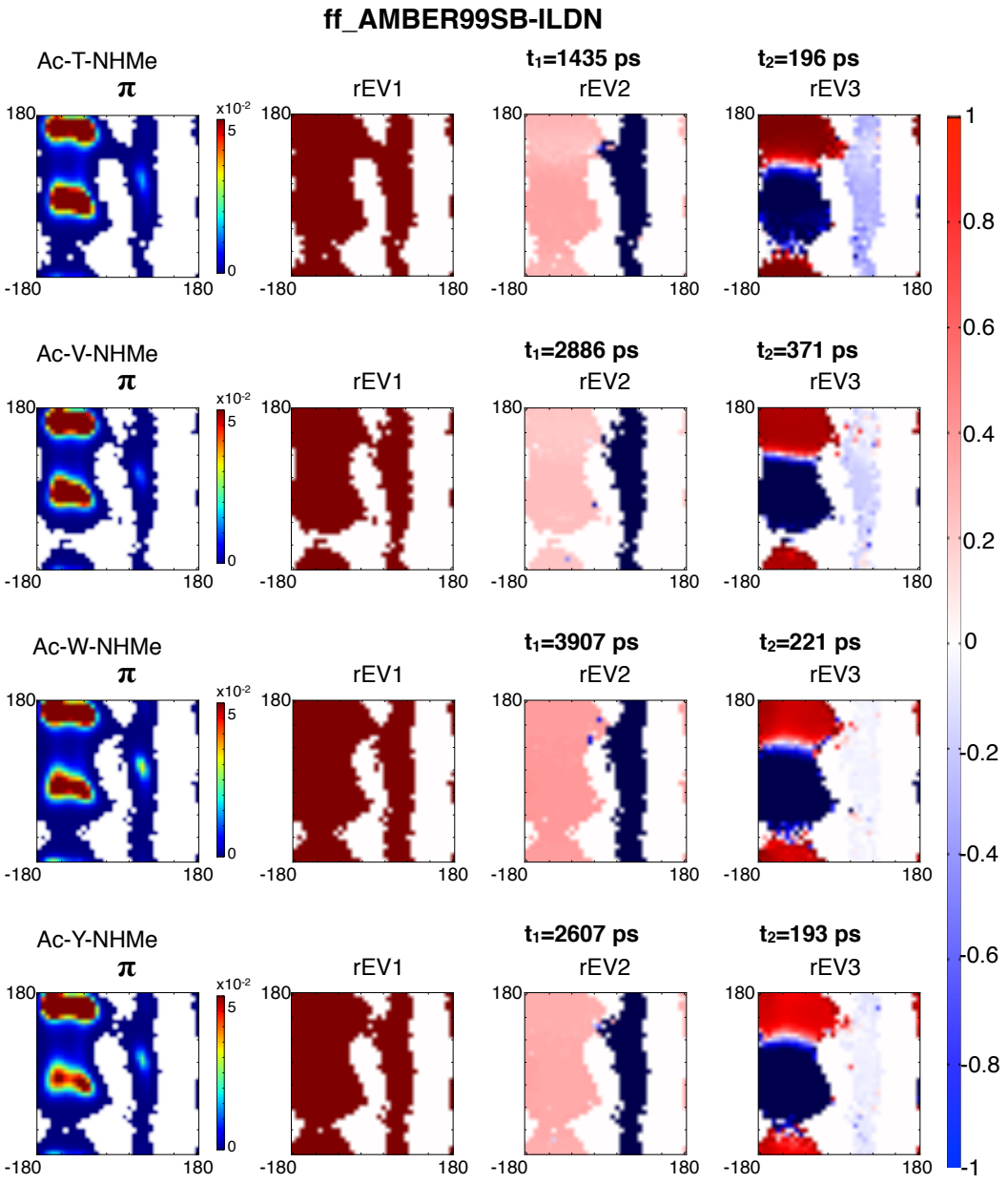


**Figure 6.5**: First left eigenvector ($\pi$), corresponding to the equilibrium distribution, and right eigenvectors 1-3, corresponding to the RBVs associated to the amino acid.

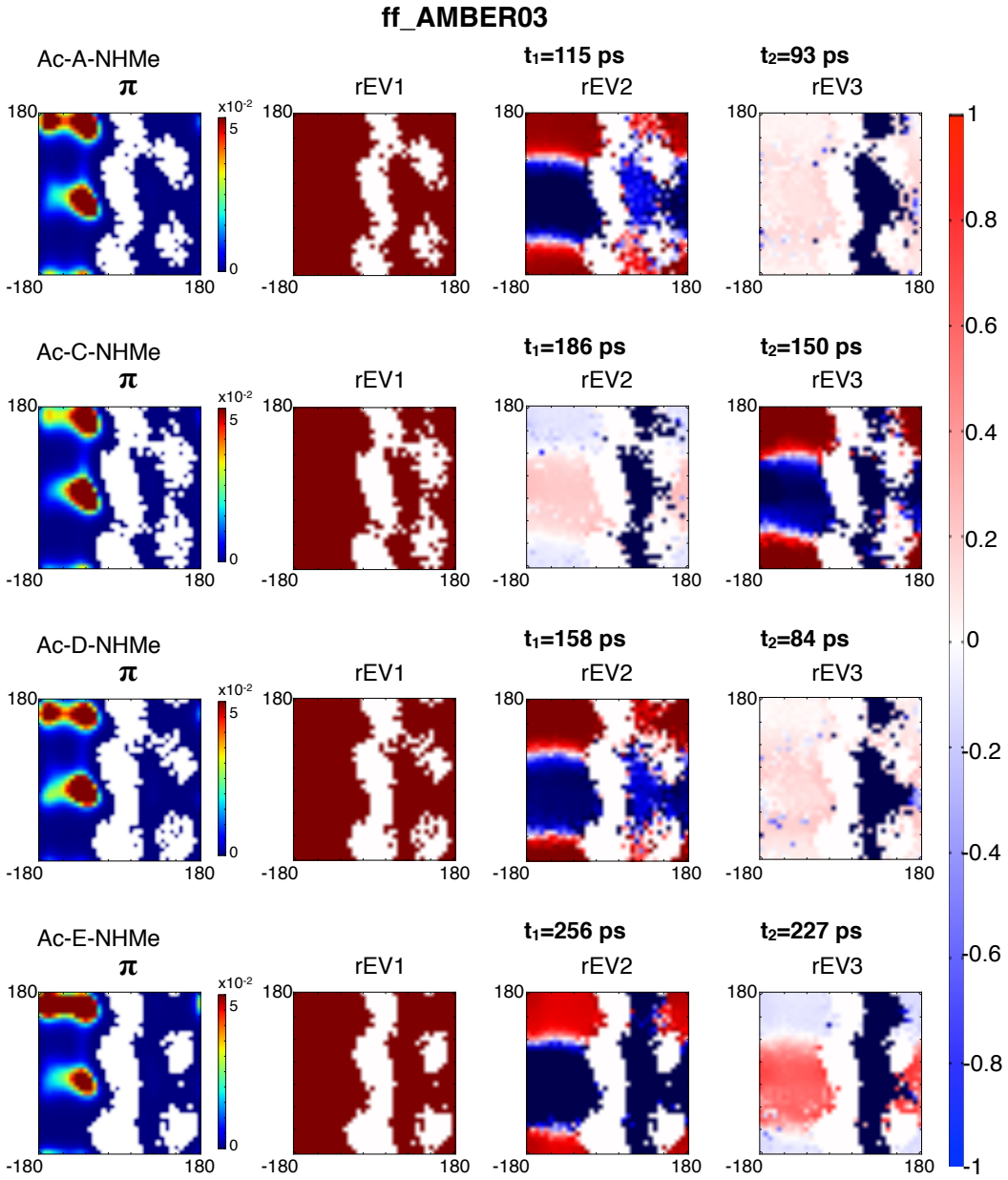


**Figure 6.6**: First left eigenvector ($\pi$), corresponding to the equilibrium distribution, and right eigenvectors 1-3, corresponding to the RBVs associated to the amino acid.

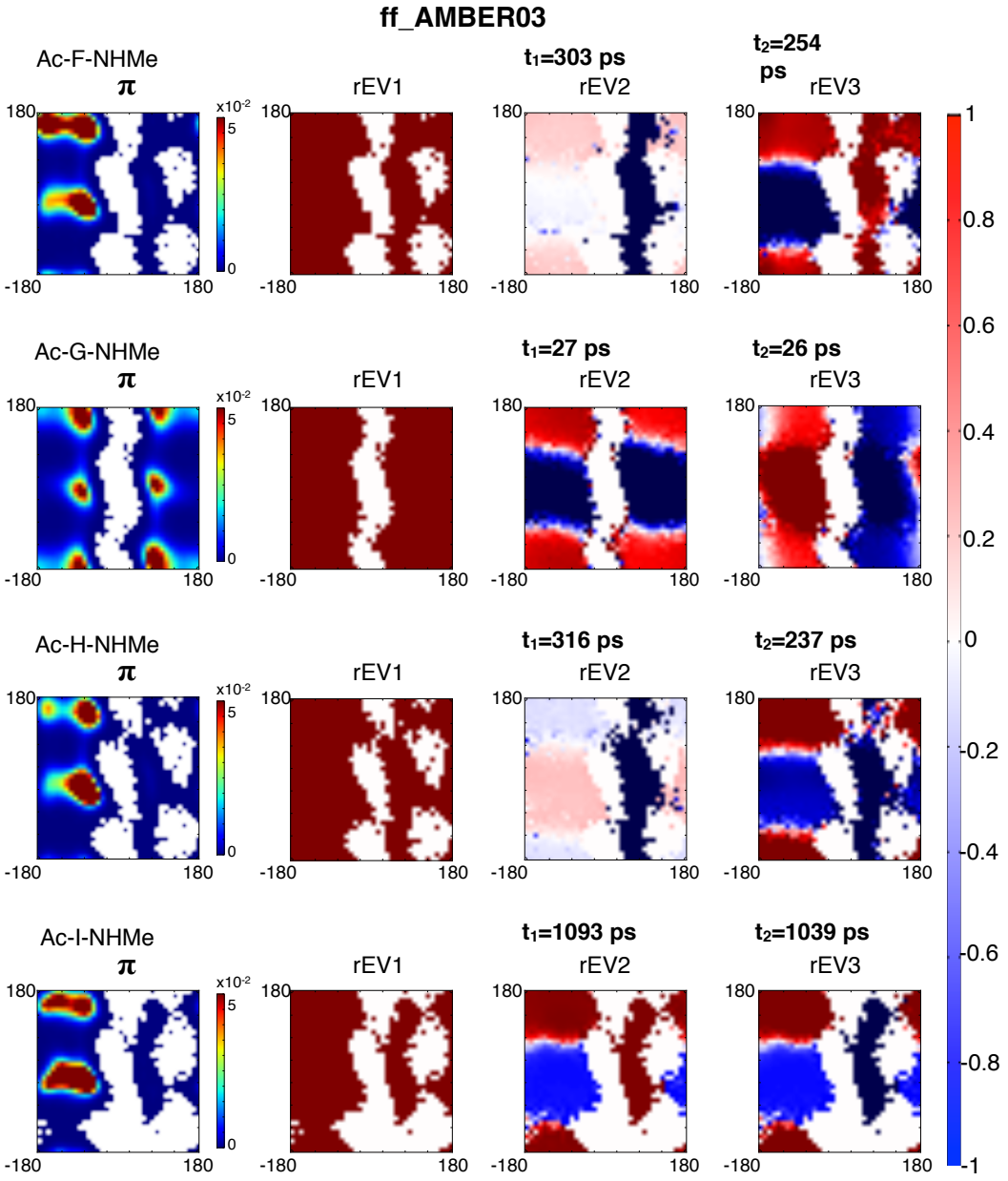


**Figure 6.7**: First left eigenvector ($\pi$), corresponding to the equilibrium distribution, and right eigenvectors 1-3, corresponding to the RBVs associated to the amino acid.

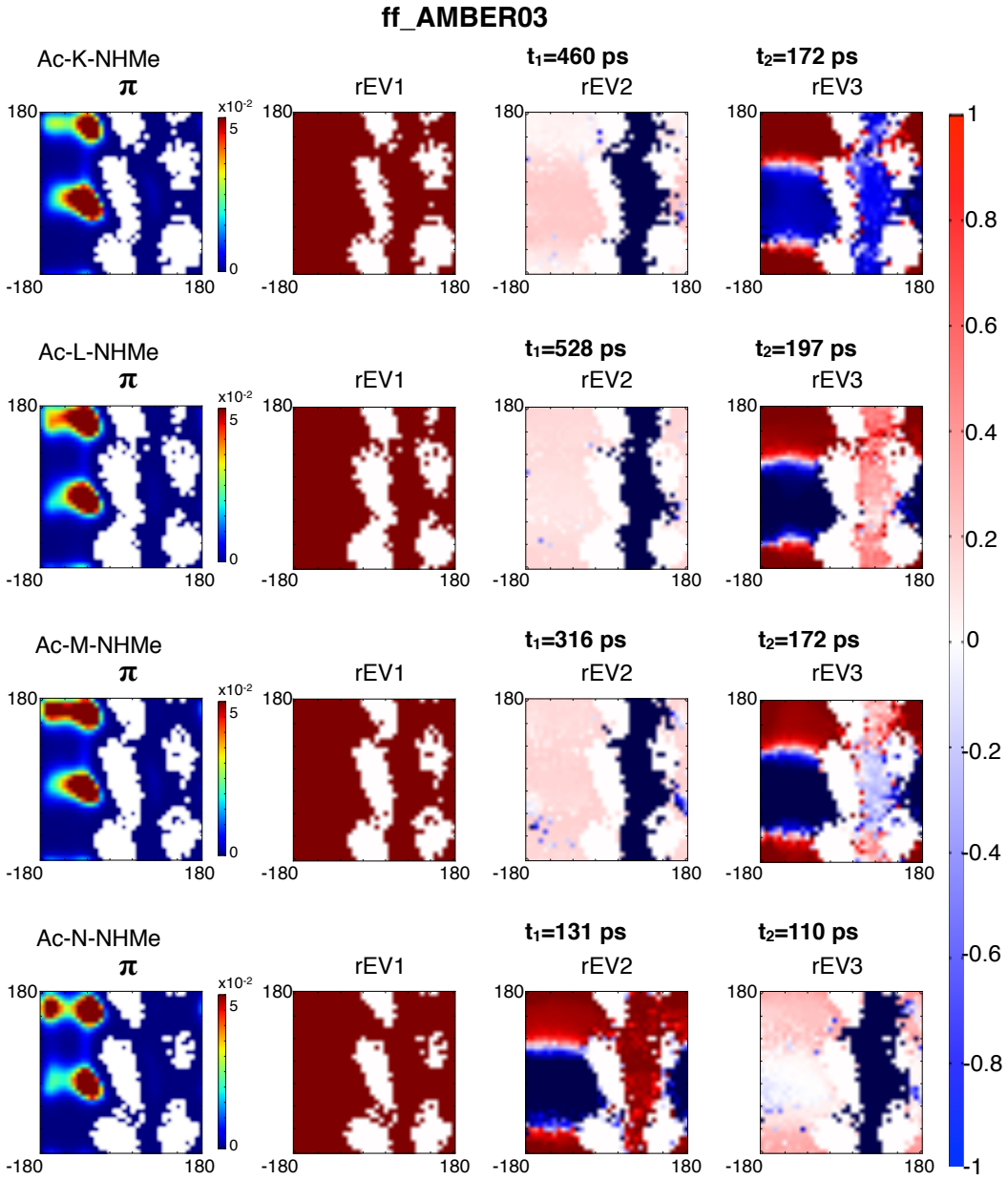


**Figure 6.8**: First left eigenvector ($\pi$), corresponding to the equilibrium distribution, and right eigenvectors 1-3, corresponding to the RBVs associated to the amino acid.

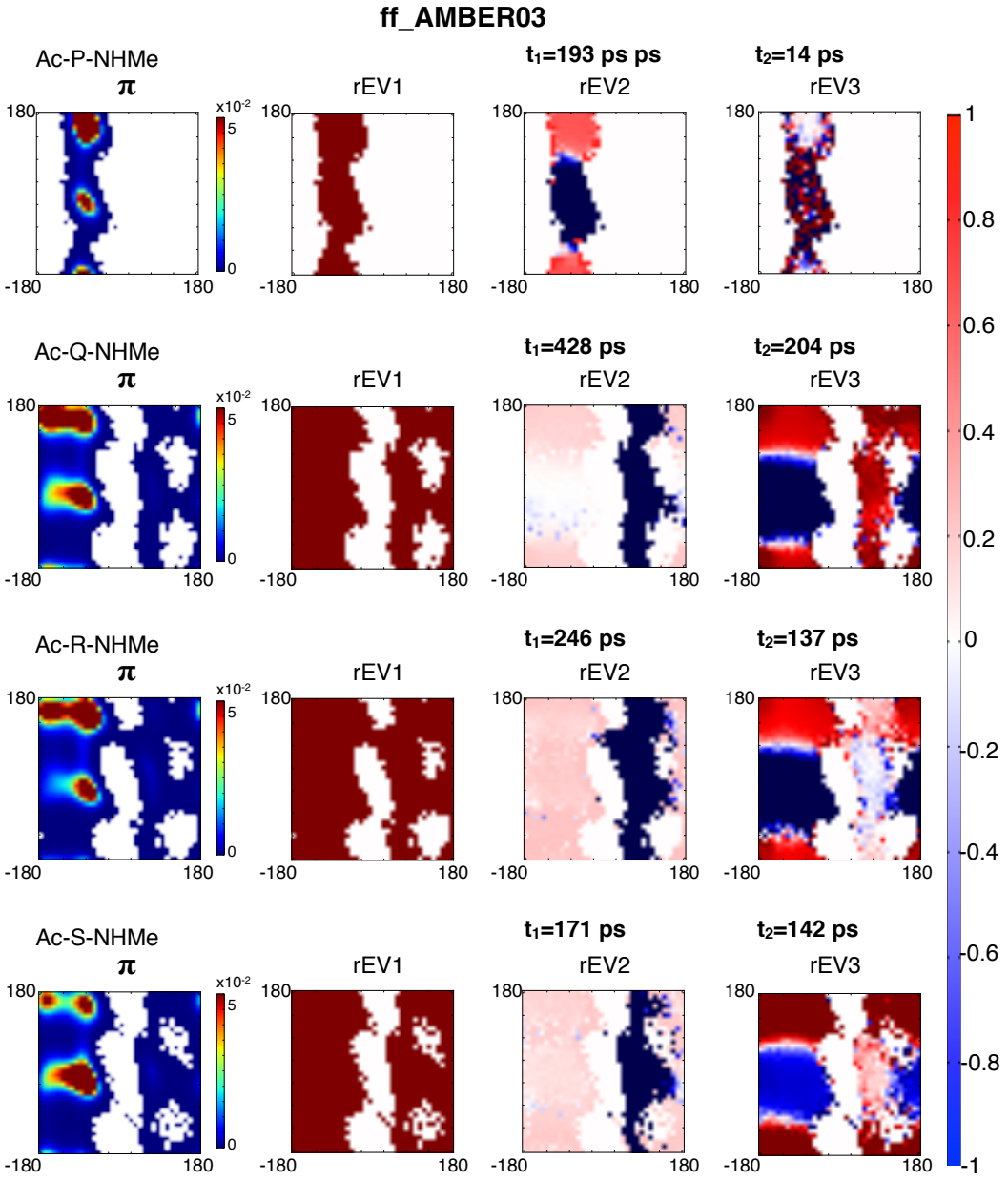


**Figure 6.9**: First left eigenvector ($\pi$), corresponding to the equilibrium distribution, and right eigenvectors 1-3, corresponding to the RBVs associated to the amino acid.

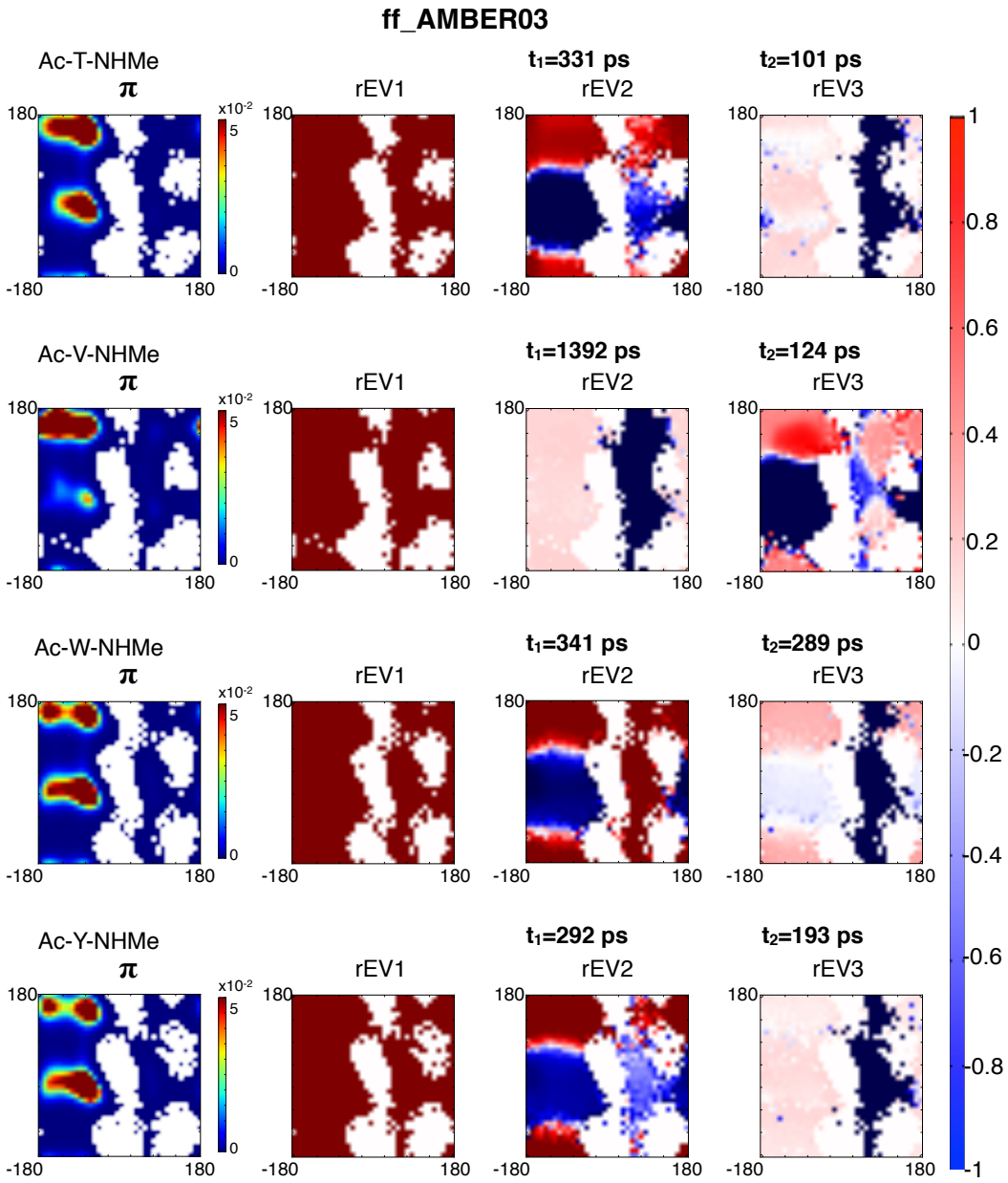


**Figure 6.10**: First left eigenvector ($\pi$), corresponding to the equilibrium distribution, and right eigenvectors 1-3, corresponding to the RBVs associated to the amino acid.

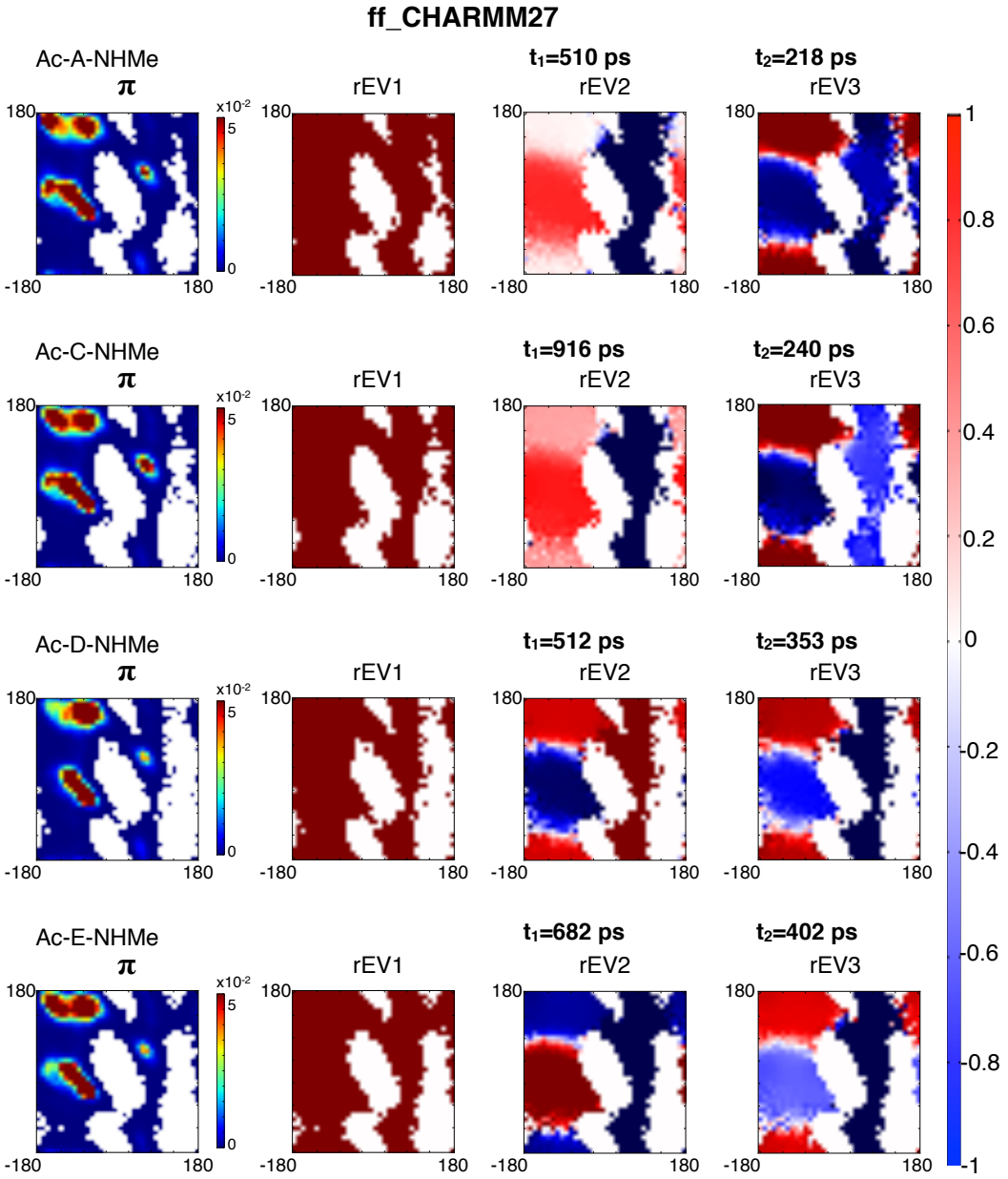


**Figure 6.11**: First left eigenvector ($\pi$), corresponding to the equilibrium distribution, and right eigenvectors 1-3, corresponding to the RBVs associated to the amino acid.

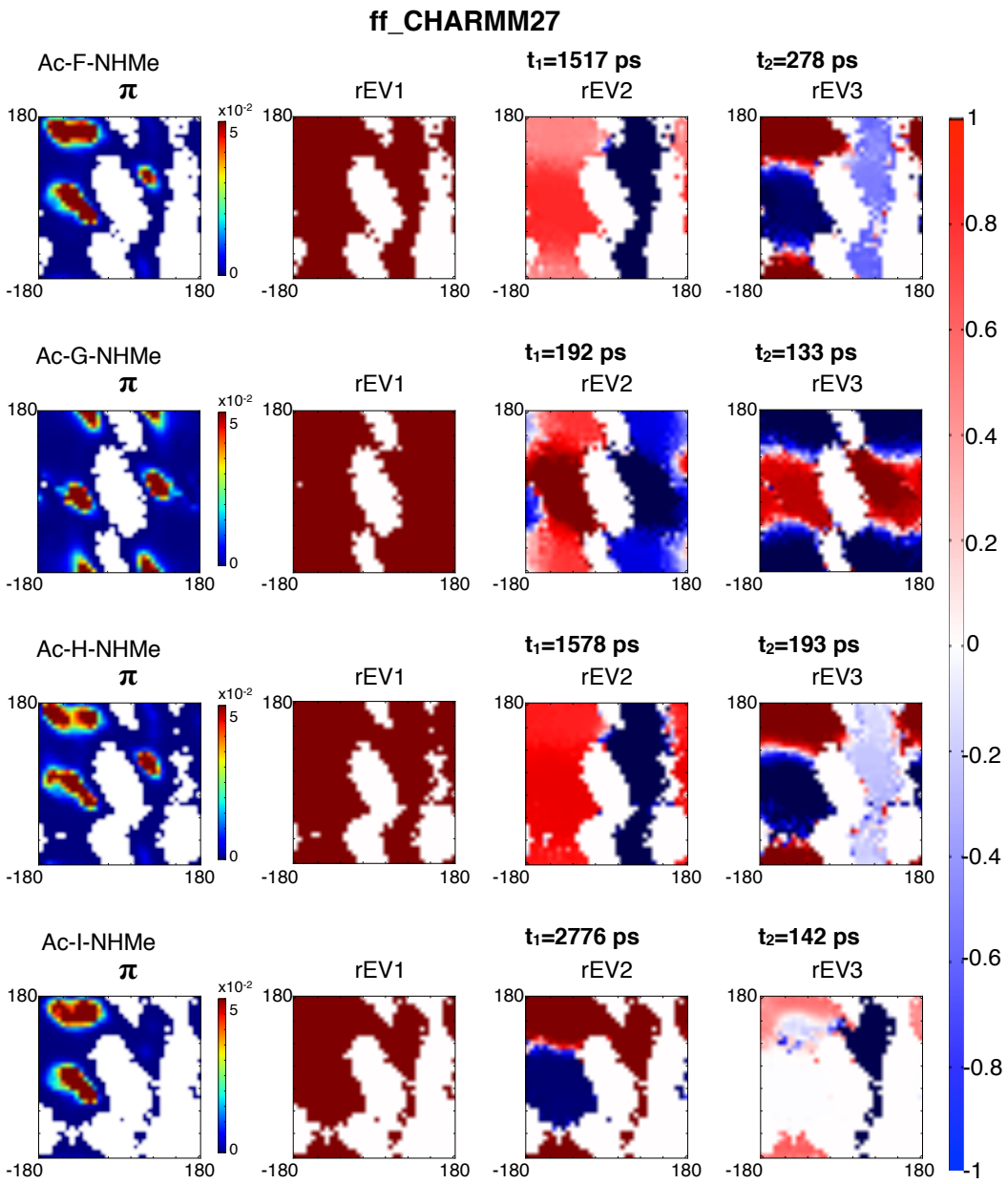


**Figure 6.12**: First left eigenvector ($\pi$), corresponding to the equilibrium distribution, and right eigenvectors 1-3, corresponding to the RBVs associated to the amino acid.

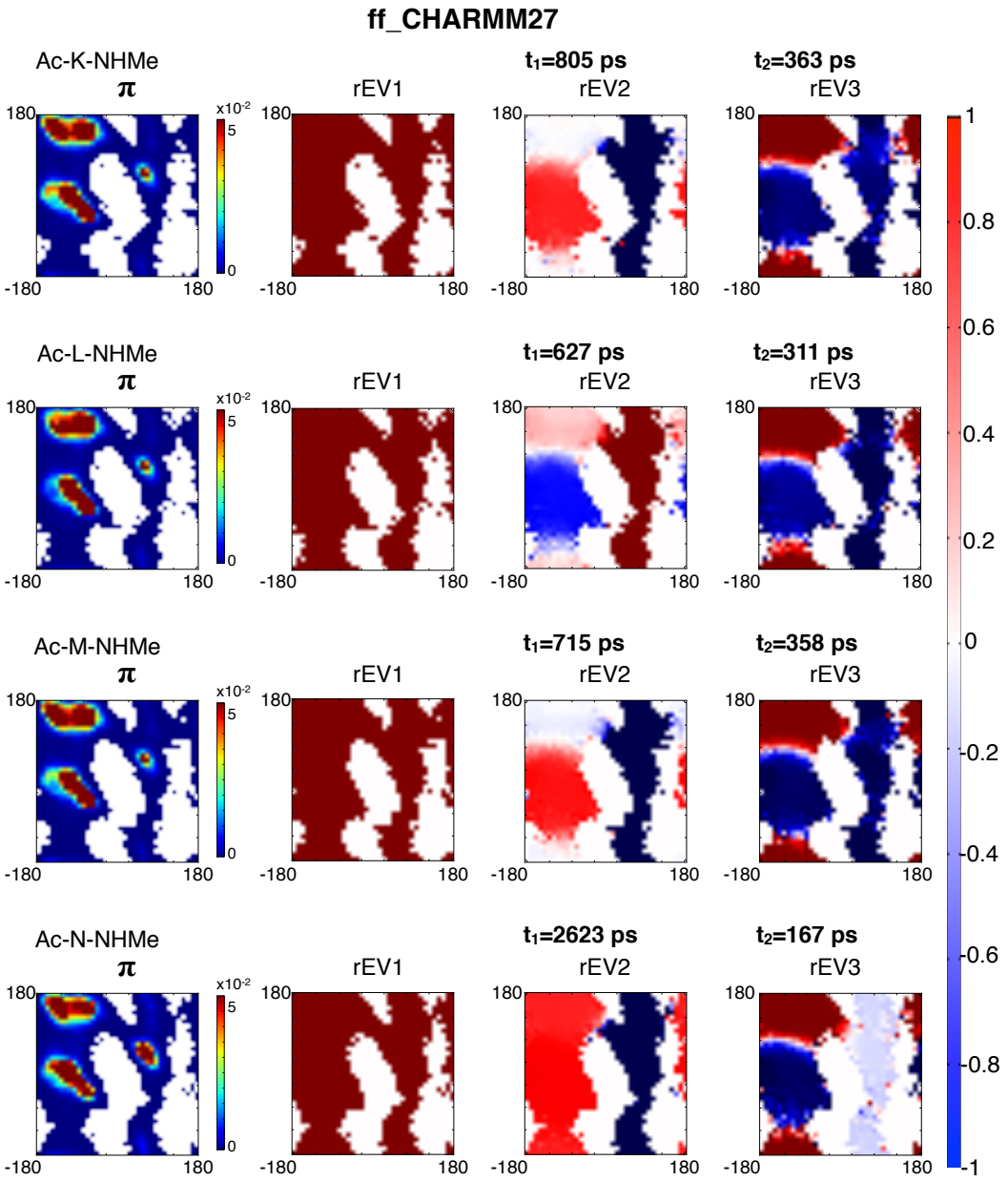


**Figure 6.13**: First left eigenvector ($\pi$), corresponding to the equilibrium distribution, and right eigenvectors 1-3, corresponding to the RBVs associated to the amino acid.

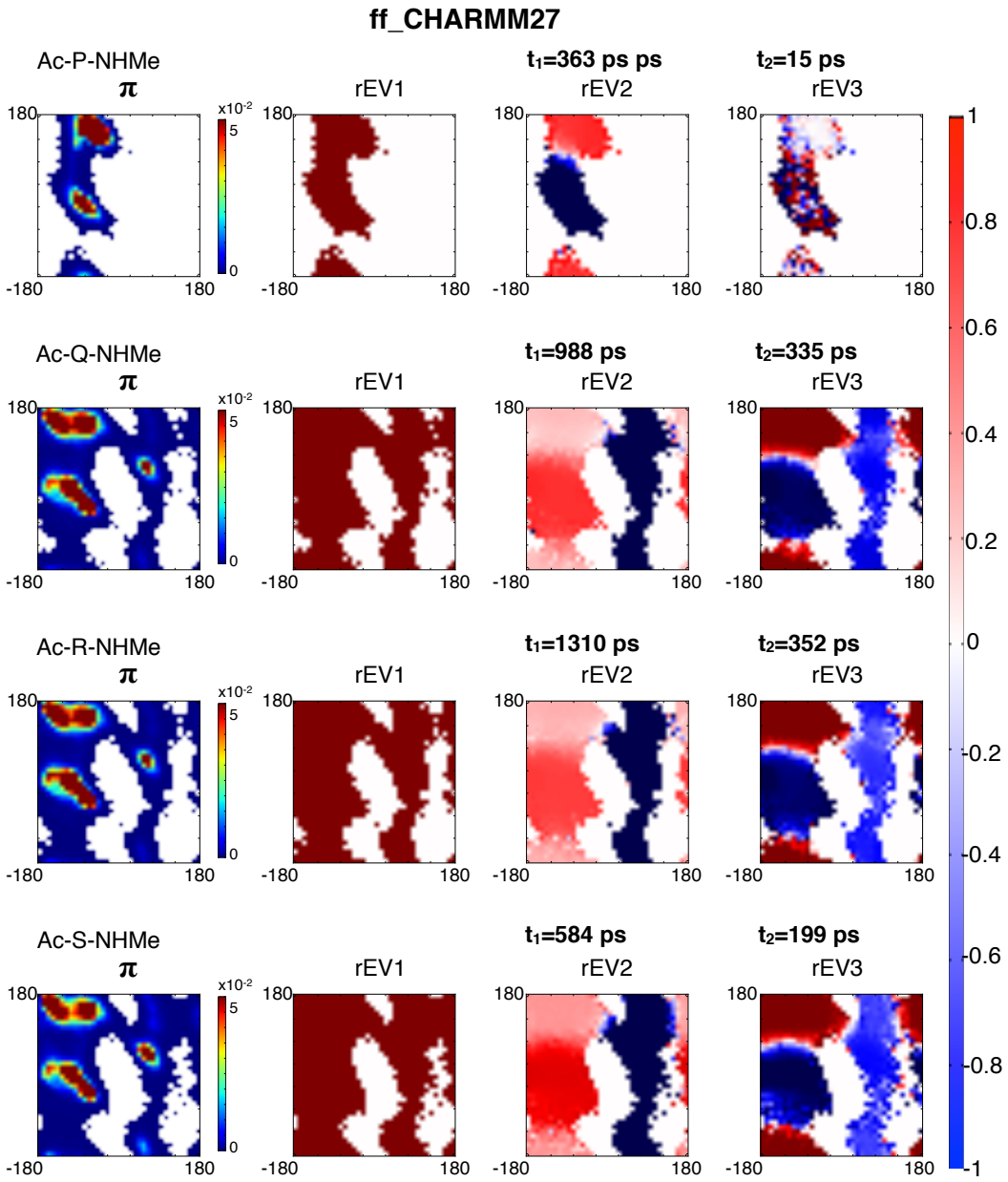


**Figure 6.14**: First left eigenvector ($\pi$), corresponding to the equilibrium distribution, and right eigenvectors 1-3, corresponding to the RBVs associated to the amino acid.

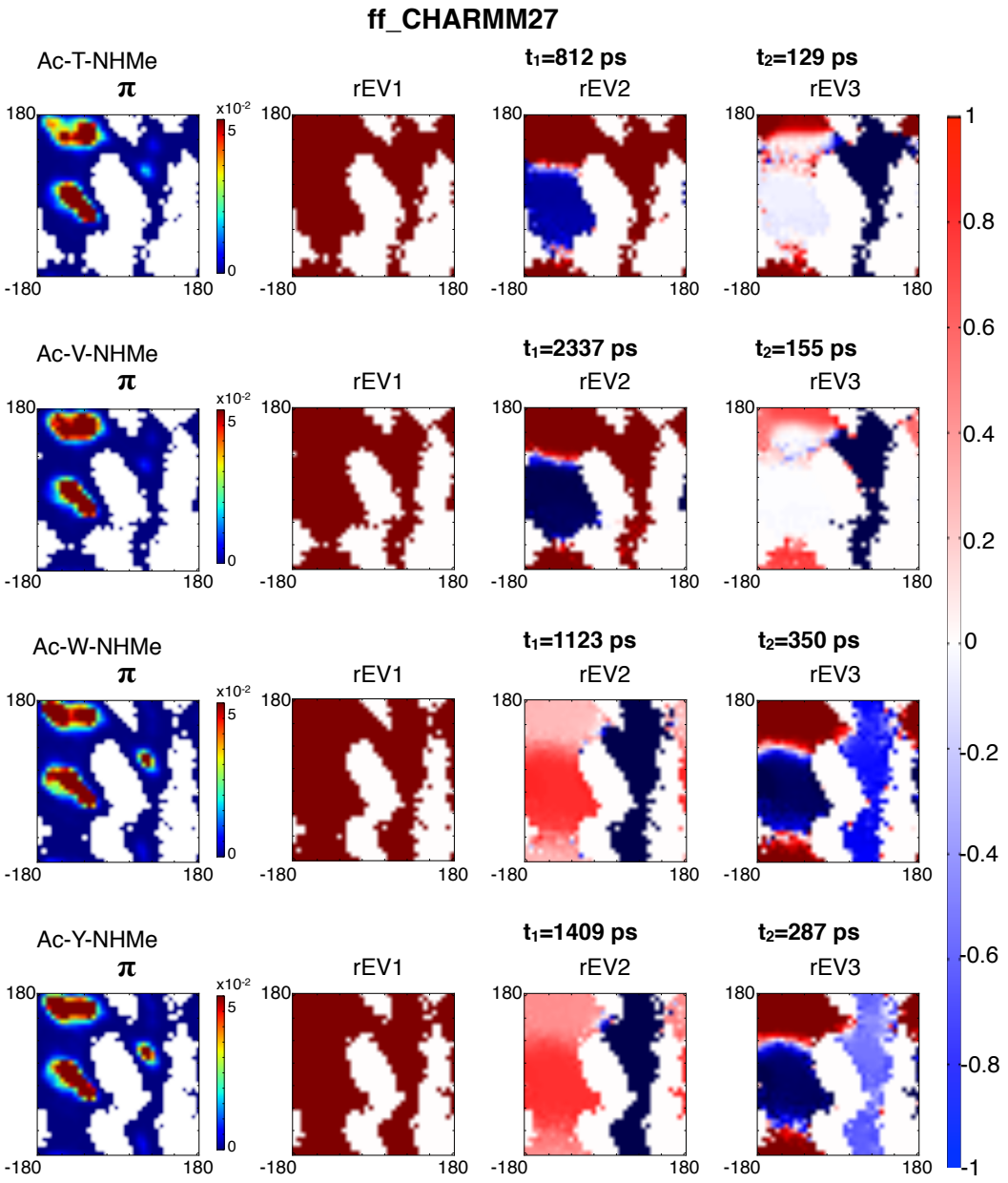


**Figure 6.15**: First left eigenvector ($\pi$), corresponding to the equilibrium distribution, and right eigenvectors 1-3, corresponding to the RBVs associated to the amino acid.

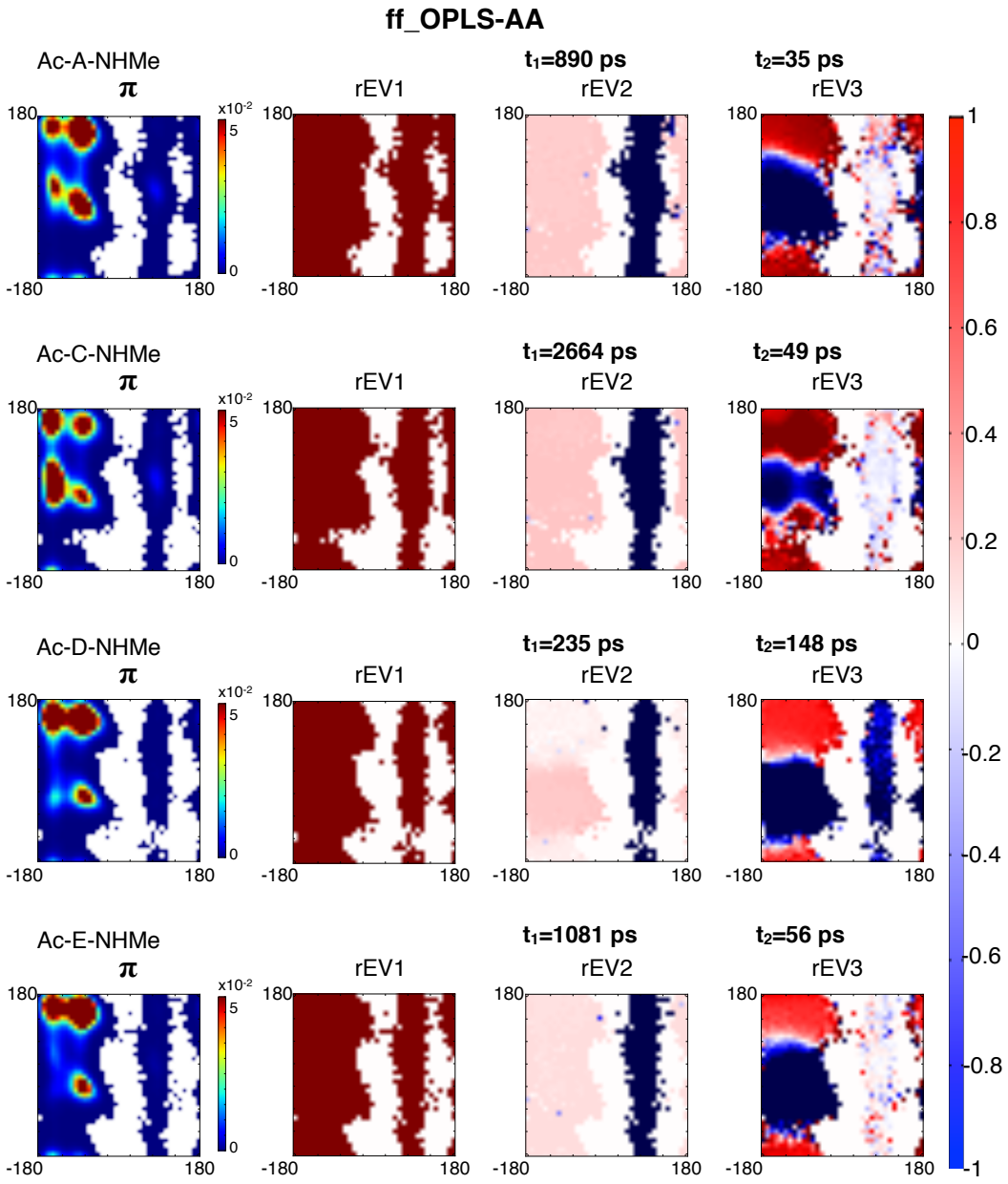


**Figure 6.16**: First left eigenvector ($\pi$), corresponding to the equilibrium distribution, and right eigenvectors 1-3, corresponding to the RBVs associated to the amino acid.

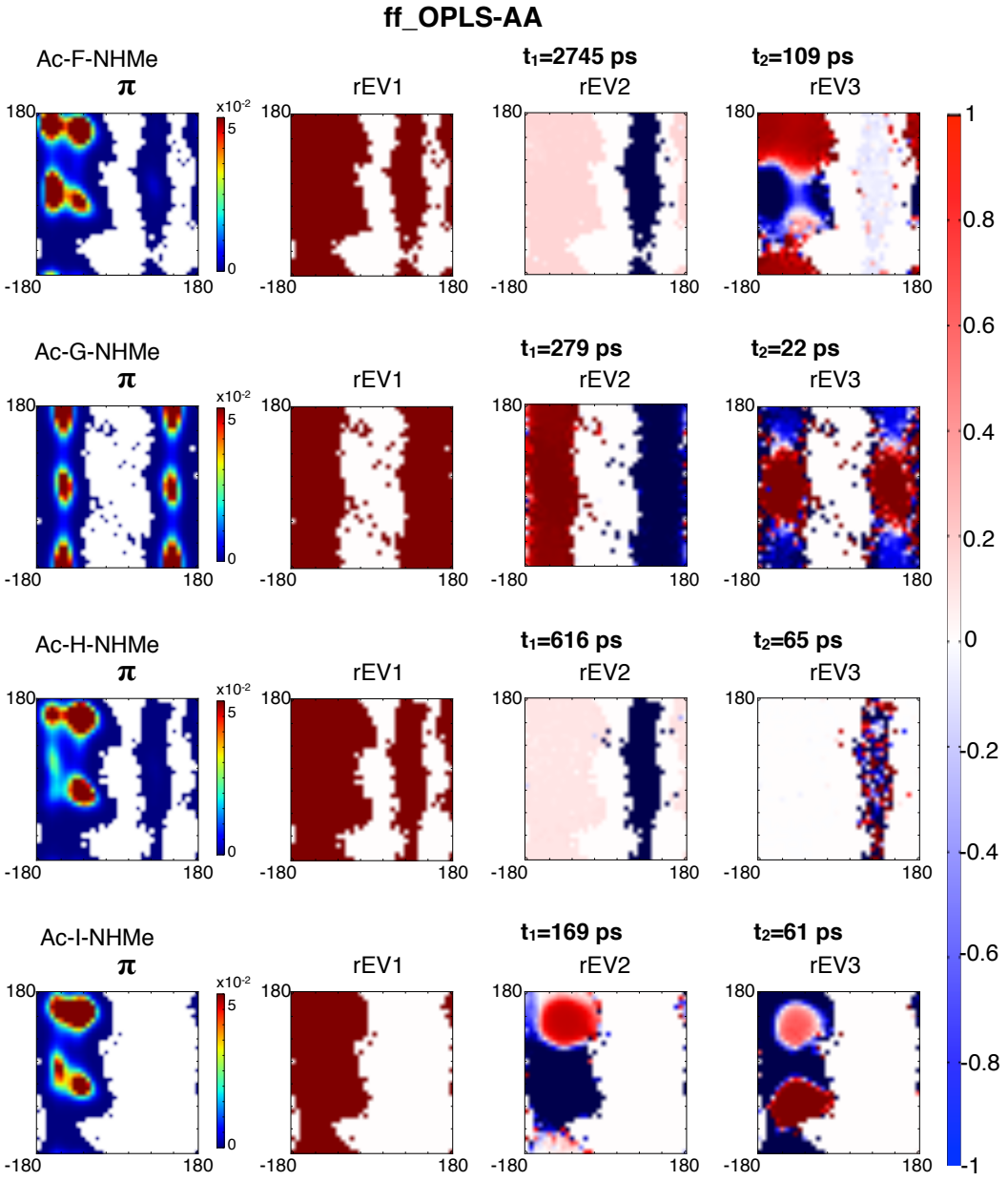


**Figure 6.17**: First left eigenvector ($\pi$), corresponding to the equilibrium distribution, and right eigenvectors 1-3, corresponding to the RBVs associated to the amino acid.

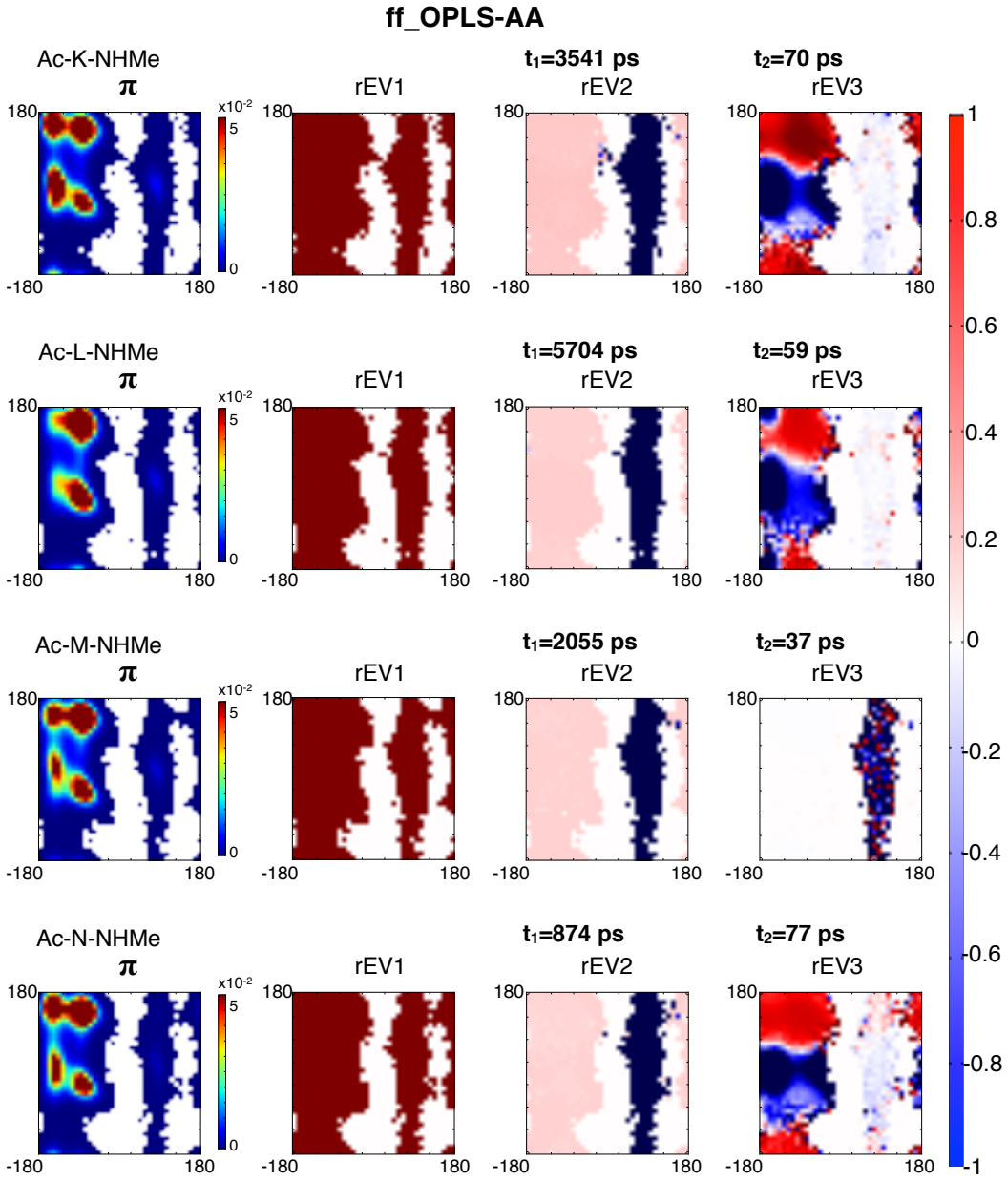


**Figure 6.18**: First left eigenvector ($\pi$), corresponding to the equilibrium distribution, and right eigenvectors 1-3, corresponding to the RBVs associated to the amino acid.

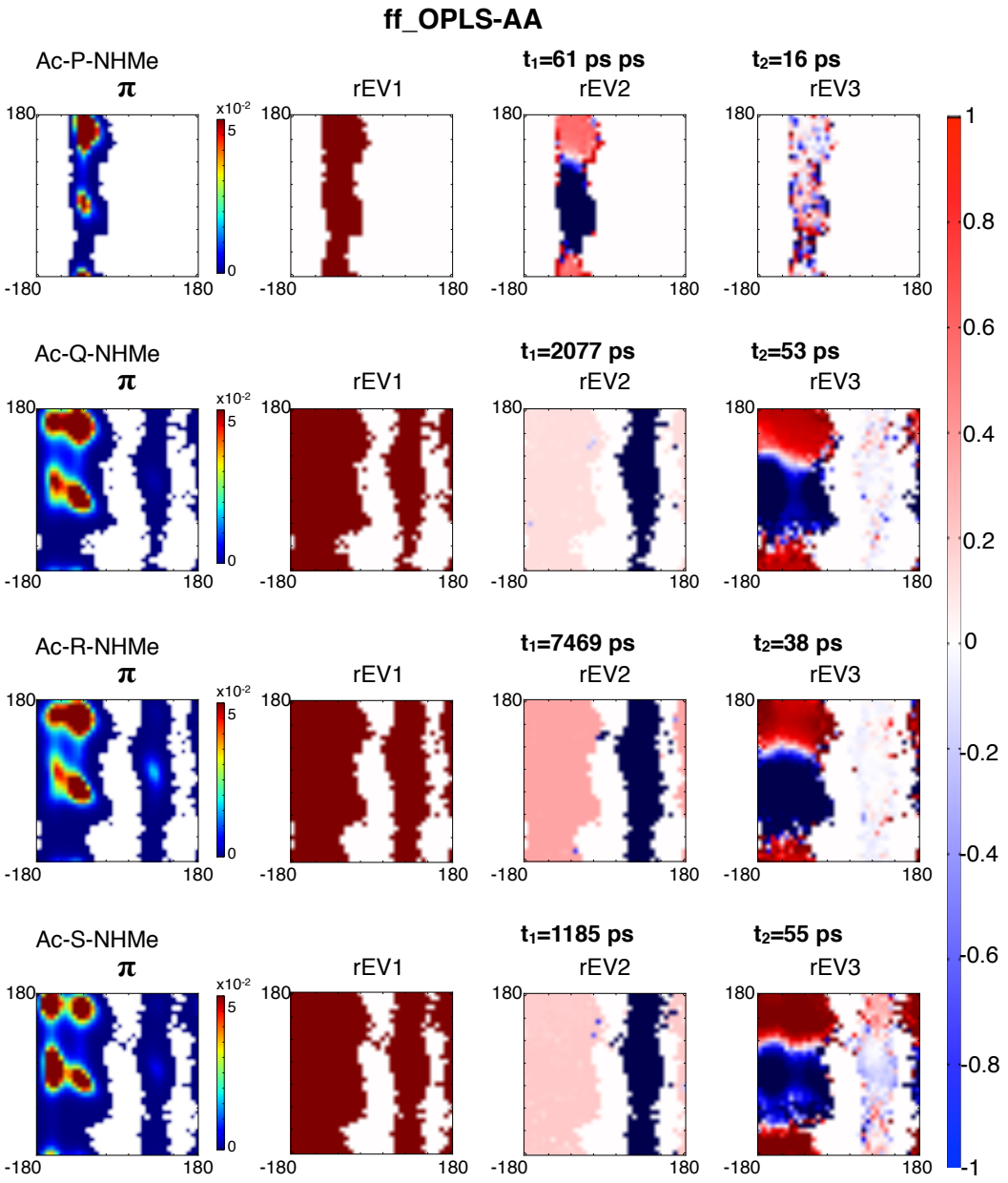


**Figure 6.19**: First left eigenvector ($\pi$), corresponding to the equilibrium distribution, and right eigenvectors 1-3, corresponding to the RBVs associated to the amino acid.

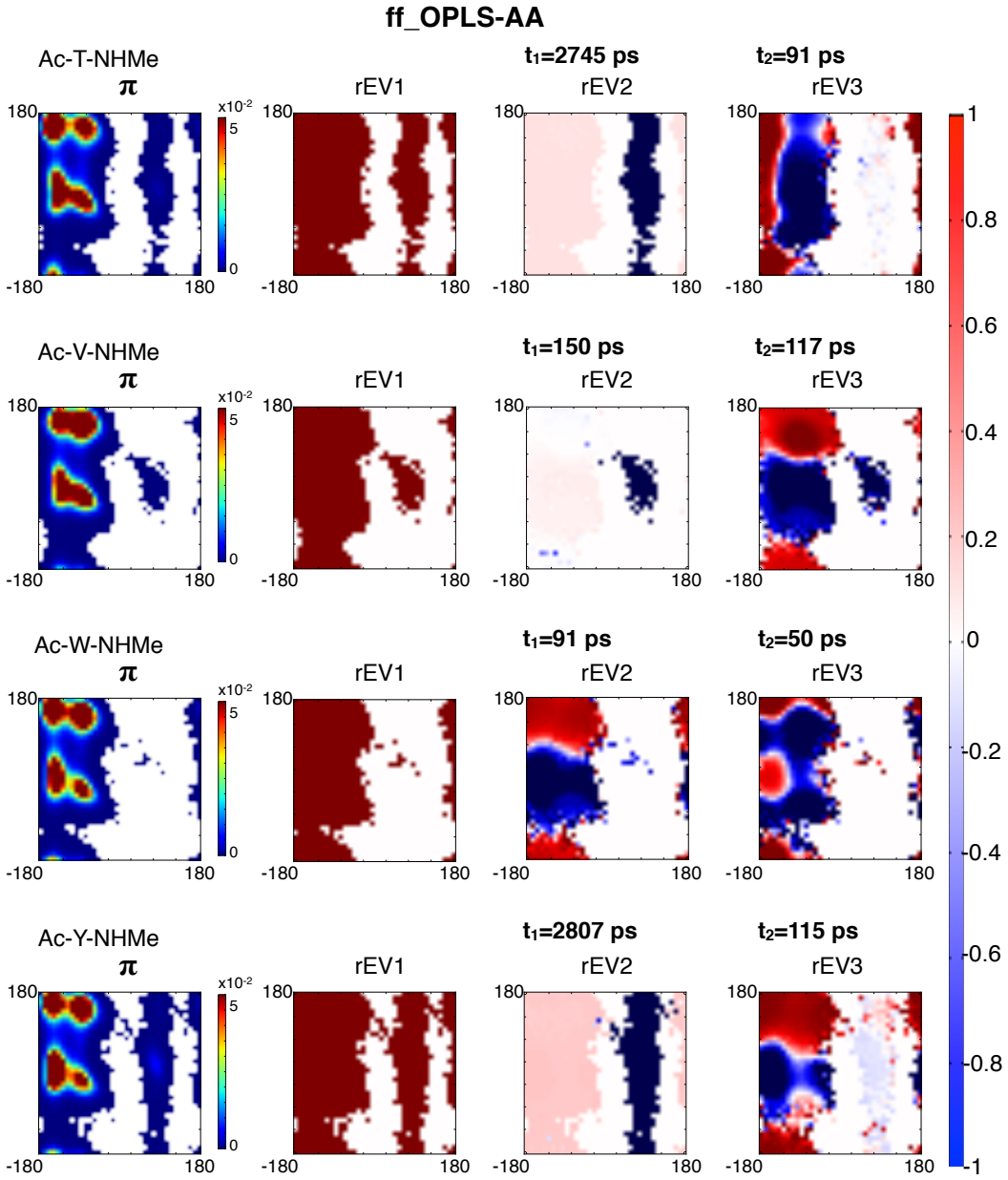


**Figure 6.20**: First left eigenvector ($\pi$), corresponding to the equilibrium distribution, and right eigenvectors 1-3, corresponding to the RBVs associated to the amino acid.

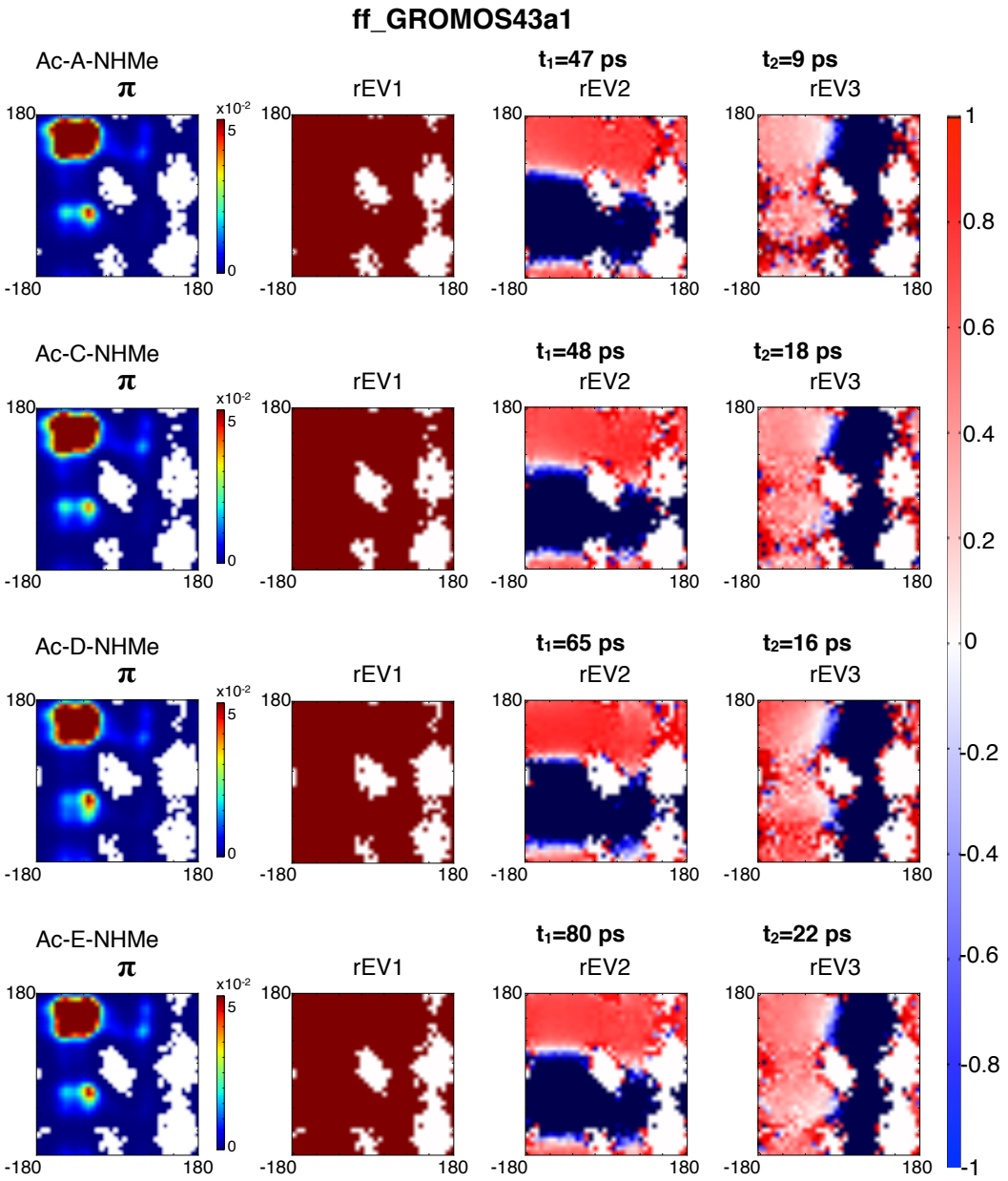


**Figure 6.21**: First left eigenvector ($\pi$), corresponding to the equilibrium distribution, and right eigenvectors 1-3, corresponding to the RBVs associated to the amino acid.

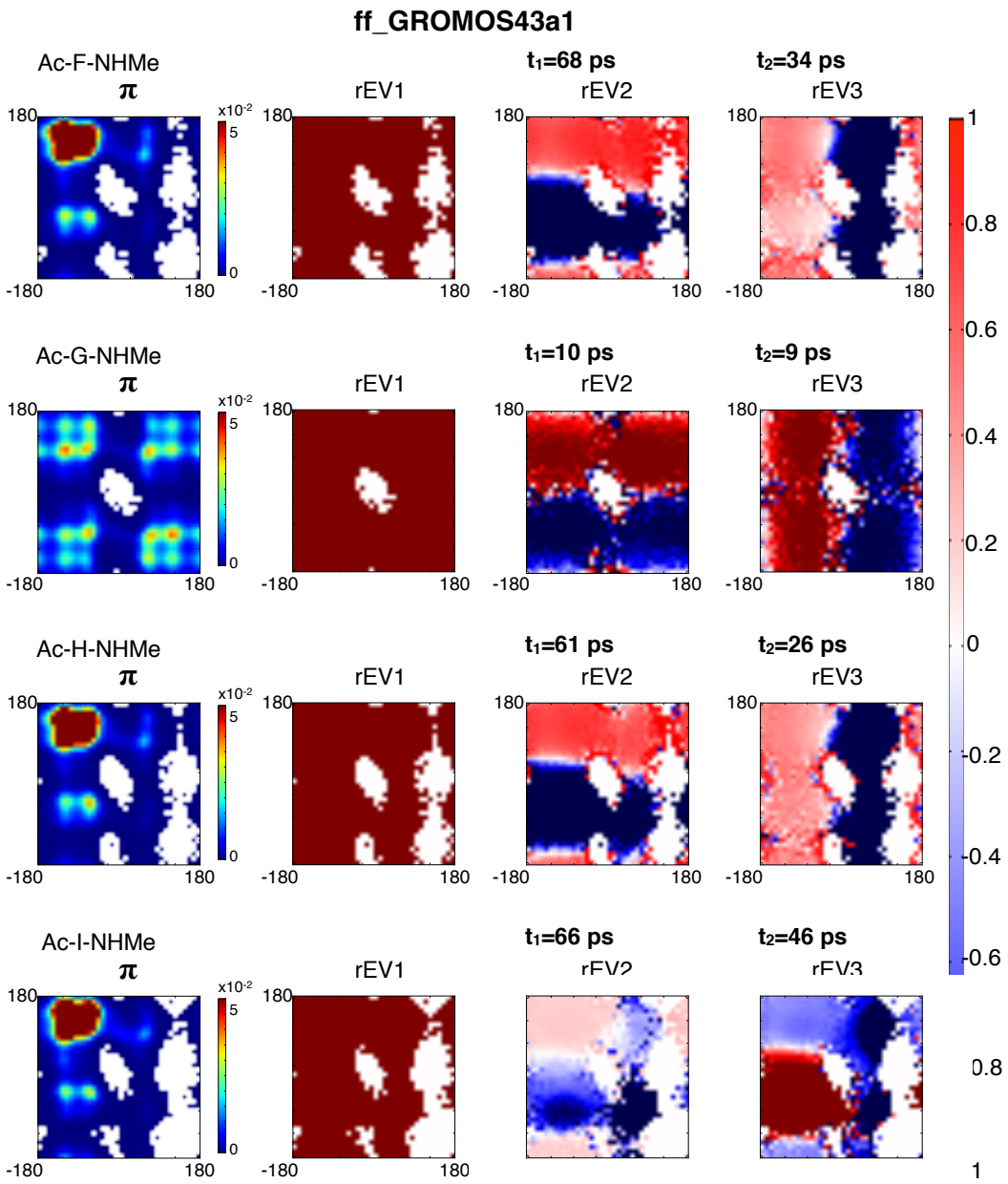


**Figure 6.22**: First left eigenvector ($\pi$), corresponding to the equilibrium distribution, and right eigenvectors 1-3, corresponding to the RBVs associated to the amino acid.

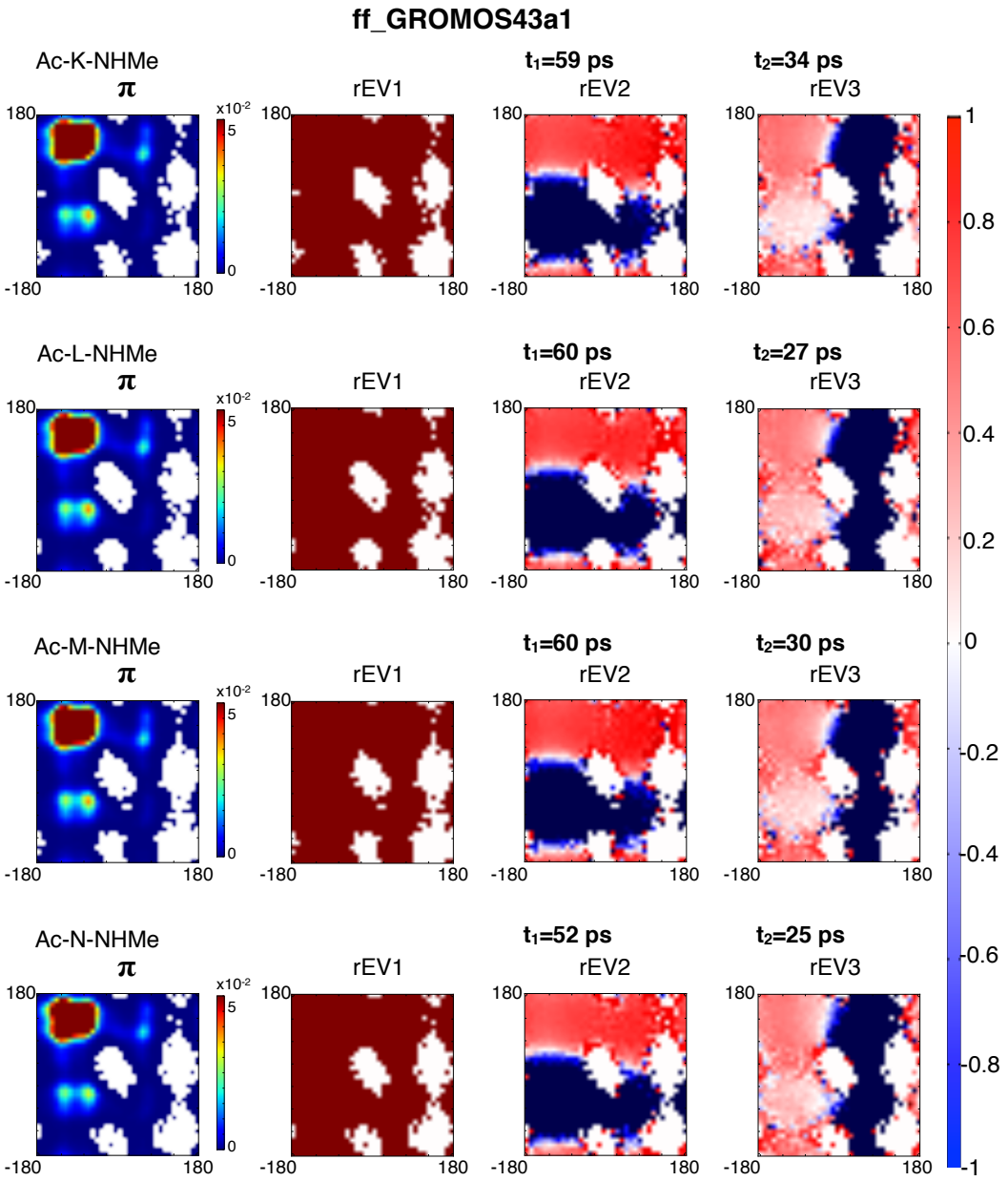


**Figure 6.23**: First left eigenvector ($\pi$), corresponding to the equilibrium distribution, and right eigenvectors 1-3, corresponding to the RBVs associated to the amino acid.

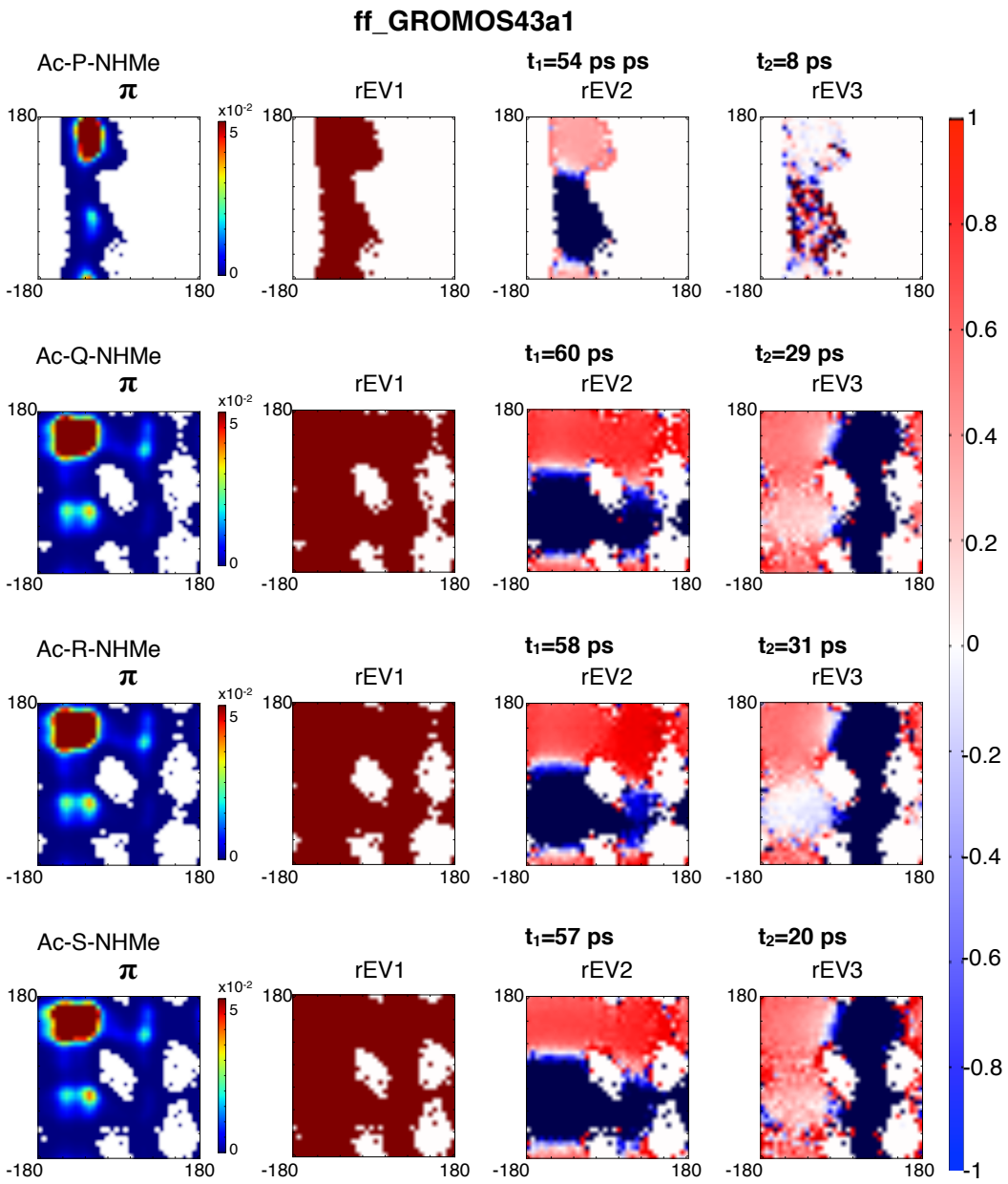


**Figure 6.24**: First left eigenvector ($\pi$), corresponding to the equilibrium distribution, and right eigenvectors 1-3, corresponding to the RBVs associated to the amino acid.

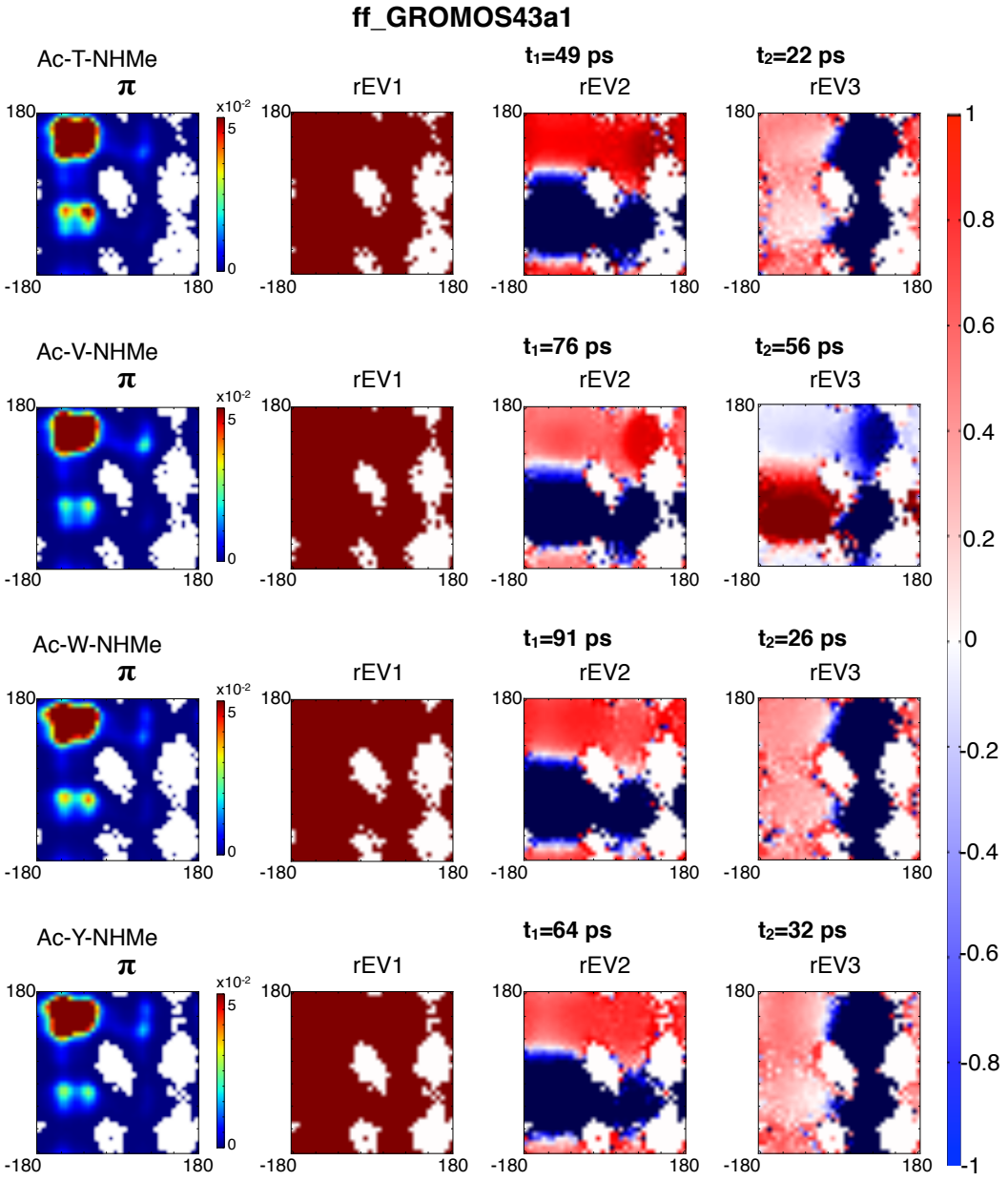


**Figure 6.25**: First left eigenvector ($\pi$), corresponding to the equilibrium distribution, and right eigenvectors 1-3, corresponding to the RBVs associated to the amino acid.

Chapter 7

# Conclusions

Within this thesis, we have exploited MD simulations to gain an atomistic representation of the time evolution of peptides. MD simulations produce an enormous amount of high dimensionality data; therefore, obtaining a quantitative and humanly understandable picture of the dynamics is not trivial. A classic "look and see" approach is bound to be unsuccessful, as it cannot provide a quantitative and statistically meaningful picture. We use MSMs to model the dynamics as transitions between a (small) number of long-living configurations.

In chapter 3 MSMs have been applied to model the dynamic properties of human amylin polipeptide, a 37-mer long intrinsically disordered peptide (IDP), related to type-2 diabetes. Projecting the dynamics onto backbone dihedral angles, we could asset the relationship between slow dynamic modes of short and longer hIAPP fragments, uncovering the hierarchy of the underlying dynamics. The same metastable configurations assumed by short peptide fragments can in fact be found in the model of the longer sequences, albeit with different timescales. The presence of long-living configurations hints at a conformational selection process taking place: in fact a configuration has to be sufficiently stable to meet to a specific binding partner and be stabilized.
This study remarks the benefits of MSMs as a tool to describe IDPs dynamics. IDPs' high flexibility makes it non-trivial to extract structural information from experiments. For example, on the one hand, it is challenging to crystallize an IDP, while on the other it is not representative of the variety of conformations that an IDP can assume. In addition, the small relative equilibrium populations and the fast relaxation times of the partially structured configurations of an IDP are difficult to measure in ensemble experiments. MD simulations in combination with MSM analysis can be extremely useful at identifying the long-living configurations of the system at atomistic resolution, and at estimating the timescales of the associated conformational changes.

The quality of a MSMs depends sensitively on the discretization. At a given lag-time, for fixed simulation data and reaction coordinates, a states definition that finely discretizes the barriers of the energy landscape produces a better model [72, 61]. The features of a high-dimensional energy landscape are, however, unknown *a priori* and a fine discretization of the full configuration space is computationally unfeasible. Therefore, in recent years the novel variational approach to conformation dynamics (VAC) [75, 76], which overcomes the crisp-states definition in favor of

continuous basis functions, has been put forward. VAC exploits the variational principle together with the method of linear variations to gain the best approximation of the true propagator eigenfunctions and eigenvalues in terms of the user-defined basis functions. The methods also permits a systematic control over the discretization error, by increasing the size of the basis set.
In chapter 4 we have presented and validated a specific basis set for the application of VAC, optimized to model the kinetics of peptides. The basis functions are defined as combinations of residue-dependent dynamic modes, which are pre-parametrized from kinetic models of terminally blocked amino acids. Such basis set definition depends only on the peptide sequence and a library of the residue-based functions has been presented in chapter 6.
The main advantage of such basis set relies in the straight-forward interpretation of the dynamic modes in terms of single amino acids conformational changes, without requiring any subsequent projection onto the space of eigenvectors. The analogous interpretation of the residue-centered basis functions renders model comparison straight-forward, which proves particularly useful in comparing the effects of point-like mutations. The code and the library of residue-centered basis functions have been made public in form of a python-package for further applications of the method (`github.com/markovmodel/variational`).
At the current state, the main limitation for the application of the residue-derived basis functions for the variational model is given by the basis set size, which increases as $3^N$, with N number of residues in the sequence. Even if only a small number of basis functions is necessary for a good approximation of the slowest dynamic processes of the bio-molecular system, how to identify these functions *a priori* is still an open question. Recent developments of the method [101] aim at finding automatically the optimal subset of basis functions for the application of VAC.
Future development of the method would work towards the definition of basis functions specific for secondary structure elements, such as $\alpha$-helix, $\beta$-sheet or turns. In a hierarchical fashion, such elements could be combined with residue-centered basis functions to extend the applicability of the method to larger peptides and proteins.

The effect of the MD force fields on the time-evolution of bio-molecular systems have also been investigated (chapter 5). The reliability of dynamic models based on MD simulations depends on how well the empirical force fields capture the dynamic properties of the system. Dynamic properties, however, are usually not taken into account in the parametrization of the force fields themselves.
To asset to which extent the force field effects the dynamics, we simulated two capped aliphatic amino acids and two test peptides with equivalent set-up options, except for the force field. The results confirm a strong dependence of the dynamic properties on the force field of choice. In the case of capped amino acids, where the expected slow processes are known, the main effect is in the order of the dynamic modes and in the associated timescales values, which can vary up to an order of magnitude. For peptides, where the entity of the slow dynamic modes is unknown, the force fields show discrepancies in both the process type and the timescales, which

can also vary up to an order of magnitude. We have therefore suggested the consideration of dynamic properties in the development of new force fields and we have proposed MSM as a tool for comparing dynamic models and relate to experiments.

Given the strong force-field dependance of the dynamics also at the amino acid level, the basis functions introduced in chapter 4 have to be force field specific. Therefore, chapter 6 presents the residue-dependent dynamic modes of the twenty encoded amino acids in five different commonly used force fields: AMBER ff-99SB-ILDN[38] , AMBER ff-03[39], OPLS-AA/L[40], CHARMM27[41] and GROMOS43a1 [42, 43]. The library can be easily extended to newly developed force fields for future applications of the method. The simulation data set has also been made available as a useful test-bed for methods and force fields development (`ftp://bdg.chemie.fu-berlin.de/`).

Appendix A

# Derivation of the Variational Principle for a transfer operator

In ref. [75, 76] it is shown that for a self-adjoint operator whose spectrum is bounded, a Variational Principle can be stated.
Here we prove that, due to reversibility, the Propagator $\mathcal{P}(\tau)$ is a self-adjoint operator with respect to the weighted scalar product, thus:

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

where with scalar product weighted with respect to the equilibrium distribution $\langle\cdot|\cdot\rangle_{\pi^{-1}}$we intend:

$$\langle f|g\rangle_{\pi^{-1}} = \int_\Omega g(\mathbf{x})^\star \pi^{-1}(\mathbf{x}) f(\mathbf{x}) \mathrm{d}\mathbf{x}, \tag{A.2}$$

where $g(\mathbf{x})^\star$ stands for complex conjugate of $g(\mathbf{x})$.

Eq. A.1 can be rewritten as:

$$\langle f|\mathcal{P}(\tau)|g\rangle_{\pi^{-1}} = \int_\Omega [\int_\Omega \mathbf{p}(\mathbf{x},\mathbf{y},\tau) f(\mathbf{x})\mathrm{d}\mathbf{x}] \frac{1}{\pi(\mathbf{y})} g(\mathbf{y})\mathrm{d}\mathbf{y}, \tag{A.3}$$

where $\mathbf{p}(\mathbf{x},\mathbf{y},\tau)$ is the probability of transitioning to $\mathbf{y}$, given that the system was in $\mathbf{x}$ a lag-time $\tau$ earlier.
If detailed balance holds:

$$\mathbf{p}(\mathbf{x},\mathbf{y},\tau) = \mathbf{p}(\mathbf{y},\mathbf{x},\tau)\frac{\pi(\mathbf{y})}{\pi(\mathbf{x})} \tag{A.4}$$

Thus:

$$\begin{aligned}
\langle f|\mathcal{P}(\tau)|g\rangle_{\pi^{-1}} &= \int_\Omega [\int_\Omega \mathbf{p}(\mathbf{y},\mathbf{x},\tau)\frac{\pi(\mathbf{y})}{\pi(\mathbf{x})} f(\mathbf{x})\mathrm{d}\mathbf{x}] \frac{1}{\pi(\mathbf{y})} g(\mathbf{y})\mathrm{d}\mathbf{y} \\
&= \int_\Omega [\int_\Omega \mathbf{p}(\mathbf{y},\mathbf{x},\tau) g(\mathbf{y})\mathrm{d}\mathbf{y}] \frac{1}{\pi(\mathbf{x})} f(\mathbf{x})\mathrm{d}\mathbf{x} \\
&= \langle g|\mathbf{P}(\tau)|f\rangle_{\pi^{-1}}
\end{aligned} \tag{A.5}$$

which proves that the propagator is a self-adjoint operator with respect to the weighted scalar product defined in eq. A.2.

This has the consequence of the eigenfunctions of $\mathcal{P}(\tau)$ being a full basis set for the Hilbert space of square-integrable functions.

The Variational Principle for the propagator states that, for any trial function $|f\rangle$, normalized as $|f| = \sqrt{\langle f|f\rangle_{\pi^{-1}}}$, the following inequality holds:

$$\langle f|\mathcal{P}(\tau)|f\rangle_{\pi^{-1}} = \int_X f(\mathbf{x})\pi^{-1}(\mathbf{x})\mathcal{P}(\tau)f(\mathbf{x})\mathrm{d}\mathbf{x} \leq 1 \tag{A.6}$$

Given any trial function $|f\rangle$, it can be linearly expanded using a basis of M basis functions $\{|\psi_i\rangle\}_{i=1}^{M}$ as:

$$|f\rangle = \sum_{i=1}^{M} a_i|\psi_i\rangle \tag{A.7}$$

According to the Method of Linear Variation, the coefficient $a_i$ are varied while the basis functions are kept constant, in order to maximize:

$$\langle f|\mathcal{P}(\tau)|f\rangle_{\pi^{-1}} = \int_X f(\mathbf{x})\pi^{-1}\mathcal{P}(\tau)f(\mathbf{x})\mathrm{d}\mathbf{x} \tag{A.8}$$

As $|f\rangle = \sum_i^M a_i|\psi_i\rangle$

$$\begin{aligned} 1 \geq \langle f|\mathcal{P}(\tau)|f\rangle_{\pi^{-1}} = \\ \langle\sum_i^M a_i\psi_i|\mathcal{P}(\tau)|\sum_j a_j\psi_j\rangle_{\pi^{-1}} = \\ \sum_{i,j=1}^{M} a_i a_j\langle\psi_i|\mathcal{P}(\tau)|\psi_j\rangle_{\pi^{-1}} = \\ \sum_{i,j=1}^{M} a_i a_j C_{i,j} \end{aligned} \tag{A.9}$$

Where $\mathbf{C}$ is interpretable as time-lagged correlation matrix.

To maximize $\langle f|\mathcal{P}(\tau)|f\rangle_{\pi^{-1}}$

$$\begin{aligned} &\frac{\partial}{\partial a_k}\langle f|\mathcal{P}(\tau)|f\rangle_{\pi^{-1}} = 0 \\ &\frac{\partial}{\partial a_k}\sum_{i,j=1}^{M} a_i a_j C_{i,j} = 0 \ \forall k = 1...M-1 \end{aligned} \tag{A.10}$$

with the normalization condition $\langle f|f\rangle_{\pi^{-1}} = \sum_{i,j=1}^{M} a_i a_j\langle\psi_i|\psi_j\rangle\pi^{-1} = \sum_{i,j=1}^{M} a_i a_j S_{i,j}$, where $\mathbf{S}$ has the meaning of overlap matrix.

To include this constraint in the optimization problem, Lagrange Multipliers are used.

$$\begin{aligned}\mathcal{L} &= \sum_{i,j=1}^{M} a_i a_j \langle \psi_i | \mathcal{P}(\tau) | \psi_j \rangle_{\pi^{-1}} - \lambda [\sum_{i,j=1}^{M} a_i a_j \langle \psi_i | \psi_j \rangle_{\pi^{-1}} - 1] \\ &= \sum_{i,j=1}^{M} a_i a_j C_{i,j} - \lambda [\sum_{i,j=1}^{M} a_i a_j S_{i,j}]\end{aligned} \tag{A.11}$$

The variational problem is then reduced to:

$$\mathcal{L} = \sum_{i=1}^{M} a_i C_{i,j} - \lambda \sum_{i=1}^{M} a_i S_{ij} = 0 \tag{A.12}$$

Which can be rewritten in a matrix form as:

$$\mathbf{Ca} = \lambda \mathbf{Sa} \tag{A.13}$$

where $\mathbf{a}$ is the vector of the expansion coefficients.

To estimate $C_{ij}$ we define the co-functions of $\psi$ as $\psi_i(x) = \pi^{-1}\varphi_i(x)$. Consequently:

$$\begin{aligned}C_{ij}(\tau) &= \langle \psi_i | \mathcal{P}(\tau) | \psi_j \rangle_{\pi^{-1}} \\ &= \langle \varphi_i \pi | \mathcal{P}(\tau) | \pi \varphi_j \rangle_{\pi^{-1}} \\ &= \iint \varphi_i(z) p(z, y, \tau) \pi(y) \varphi_j(y) dy dz\end{aligned} \tag{A.14}$$

which can be interpreted as a time-lagged cross-correlation between $\varphi_i$ and $\varphi_j$.

$$corr(\varphi_i, \varphi_j, \tau) = \iint \varphi_i(z) \mathbf{p}(x_{t+\tau} = z | x_t = y) \varphi_j(y) \mathbf{p}(x_t = y) dz dy \tag{A.15}$$

In the limit of $T \to \infty$ it can be estimated from the trajectory as:

$$\begin{aligned}\widehat{corr}(\varphi_i, \varphi_j, \tau) &= \frac{1}{T-\tau} \int_0^{T-\tau} \varphi_j(x_t) \varphi_i(x_{t-\tau}) dt \\ &= \frac{1}{T-\tau} \sum_{t=1}^{T-\tau} \varphi_j(x_t) \varphi_i(x_{t-\tau}) \Delta t.\end{aligned} \tag{A.16}$$

Analogously $S_{ij}$ can be estimated from the trajectory as:

$$\begin{aligned}S_{ij} &= \langle \psi_i | \psi_j \rangle_{\pi^{-1}} \\ &= \int_0^T \varphi_j(x_t) \varphi_i(x_t) dt \\ &= \sum_{t=1}^{T} \varphi_j(x_t) \varphi_i(x_t) \Delta t.\end{aligned} \tag{A.17}$$

## Acknowledgments

There is a lot of people that I would like to thank for being some how involved in me succeeding in completing this PhD. I do hope not to forget anyone worth mentioning.

I should probably start chronologically to thank my family and closest long-time friends for always being there even when we where all scattered around the world. Thanks Mom and Dad for all you made so that I could accomplish my goals. Thanks Andrea, Gisella, Zia Antonietta for being my family. Thanks to Gaia, the newest member, to be goy of us all. Thanks to Lucia and Silvia, you are my closest and longest-time friends, more than that actually, you are probably more like sisters to me and I would have not have accomplished anything in my life, had I not had you in it.

Thanks to my university friends and professors for accompanying me in the journey of learning to love science. Special thanks to Prof. Luisi for inspiring me as a scientist and a person. I do owe you the choice of looking into the field of bio-physics. Thanks to my Californian friends and colleagues for witnessing my first steps into the world of Molecular Dynamics. Many thanks to Prof John Chodera for introducing me to what would become my PhD (without you my life would have probably ended somewhere else entirely) and for being such an enthusiastic scientist. More should be like you and everyone would love science (and cocktails).

Thanks to my Berlin friends, without whom I would have probably ended up crazy in the last four years. Here the list should be very long, but if you belong to it you know it. Thanks to my GoT-fans fellows for the nights spent watching and discussing. And thanks to my fellow Port&Cheese Society members for rising up the inner nerd in me to its best.

Thanks to my colleagues and friends in the math and chemistry departments for the scientific help and support during this PhD. Thanks to the Healthy Wednesday crew, I had some extremely nice meals in my last two years of PhD. Special thanks to my TC-boyz for the laugh that should never be missing in a PhD.

Many thanks to my supervisors Prof. Frank Noè and Prof. Bettina Keller for giving me the opportunity of accomplishing a doctorate adventure. Thank you Bettina for actually teaching me how to be a researcher, you have been a supervisor at 360-degrees.

Final and thanks to Pio for being there with me, support me and love me in this four years, and to Toni for being my other awkward half of the brain.

If you have recognized yourself in more than one thank-you note, well it just means I owe you more than just one “Thank you" only.

# References


[1] D Whitford. *Proteins: Structure and Function.* Wiley, 2005.

[2] G M Cooper. *The Cell: A Molecular Approach.*, chapter 2: The Central Role of Enzymes as Biological Catalysts. Sinauer Associates, 2nd edition, 2000.

[3] A.J.F. Griffiths, S.R. Wessler, R.C. Lewontin, and S.B. Carroll. *Introduction to Genetic Analysis*, chapter 7: DNA: Structure and Replication, pages 283–390. W. H. Freeman and Company, 2008.

[4] A R Dinasarapu, B Saunders, I Ozerlat, K Azam, and S Subramaniam. Signaling gateway molecule pages–a data model perspective. *Bioinformatics*, 27(12):1736–1738, June 2011.

[5] D Sadava, D. M. Hillis, H. C. Heller, and M. Berenbaum. *Life, the Science of Biology.* Macmillan Publishers, 9th edition, 2009.

[6] L. Pauling. *The nature of the chemical bond.* Cornell University Press, 3rd edition, 1960.

[7] G.N. Ramachandran, C. Ramakrishnan, and V. Sasisekharan. Stereochemistry of polypeptide chain configurations. *J. Mol. Biol.*, 7(1):95–99, July 1963.

[8] J. E. Guerois, R.and Nielsen and L. Serrano. Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations. *J. Mol. Biol.*, 320(2):369–387, July 2002.

[9] A. Farrugia. Albumin usage in clinical medicine: tradition or therapeutic? *Transfusion medicine reviews*, 24(1):53–63, January 2010.

[10] C. Levinthal. Are there pathways for protein folding? *J Med Phys*, 65:44–45, 1969.

[11] C. B. Anfinsen. Principles that Govern the Folding of Protein Chains. *Science*, 181(4096):223–230, July 1973.

[12] J. N. Onuchic and P. G Wolynes. Theory of protein folding. *Curr. Opin. Struct. Biol.*, 14(1):70–75, February 2004.

[13] H. J. Dyson and P. E. Wright. Intrinsically unstructured proteins and their functions. *Nat. Rev. Mol. Cell Biol.*, 6(3):197–208, March 2005.

[14] V. N. Uversky, C. J. Oldfield, and A. K. Dunker. Intrinsically disordered proteins in human diseases: introducing the D2 concept. *Annu. Rev. Biophys.*, 37:215–246, January 2008.

[15] M P. Murphy and H. LeVine. Alzheimer's disease and the amyloid-beta peptide. *J. Alzheimers Dis.*, 19(1):311–323, January 2010.

[16] P. Westermark, A. Andersson, and G. T. Westermark. Islet amyloid polypeptide, islet amyloid, and diabetes mellitus. *Physiol. Rev.*, 91(3):795–826, July 2011.

[17] P. E. Wright and H. J. Dyson. Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. *J. Mol. Biol.*, 293(2):321–331, October 1999.

[18] R. Ishima and D. A. Torchia. Protein dynamics from NMR. *Nature Struct Biol*, 7(9):740–743, September 2000.

[19] M. Karplus and J. A. McCammon. Dynamics of proteins: elements and function. *Annu. Rev. Biochem.*, 52:263–300, January 1983.

[20] A. J. Wand. Dynamic activation of protein function: a view emerging from NMR spectroscopy. *Nature Struct Biol*, 8(11):926–931, November 2001.

[21] K. Henzler-Wildman and D. Kern. Dynamic personalities of proteins. *Nature*, 450(7172):964–972, December 2007.

[22] N. Forns, S. de Lorenzo, M. Manosas, K. Hayashi, J.M. Huguet, and F. Ritort. Improving Signal/Noise Resolution in Single-Molecule Experiments Using Molecular Constructs with Short Handles. *Biophys. J.*, 100(7):1765–1774, April 2011.

[23] G. Adam, J. Buša, and M. Hnatič. *Mathematical Modeling and Computational Science: International Conference, MMCP 2011, Stará Lesná, Slovakia, July 4-8, 2011, Revised Selected Papers*, chapter Proteins studied by Computer Simulations, pages 56–65. Springer Science & Business Media, 2012.

[24] C. D. Snow, E. J. Sorin, Y. M. Rhee, and V. S. Pande. How Well Can Simulation Predict Protein Folding Kinetics and Thermodynamics? *Annu. Rev. Biophys.*, 34(1):43–69, June 2005.

[25] J. D. Chodera, W. C. Swope, J. W. Pitera, and K. A. Dill. Long Time Protein Folding Dynamics from short Time Molecular Dynamics simulations. *Multiscale Model. Simul.*, 5(4):1214–1226, 2006.

[26] W. F. van Gunsteren, D. Bakowies, R. Baron, I. Chandrasekhar, M. Christen, X. Daura, P. Gee, D. P. Geerke, A. Glättli, P. H. Hünenberger, M. a Kastenholz, C. Oostenbrink, M. Schenk, D. Trzesniak, N. F. a van der Vegt, and H. B. Yu. Biomolecular modeling: Goals, problems, perspectives. *Angew. Chem. Int. Ed. (English)*, 45(25):4064–4092, June 2006.

[27] W. F. van Gunsteren, J. Dolenc, and A. E. Mark. Molecular simulation as an aid to experimentalists. *Curr. Opin. Struct. Biol.*, 18(2):149–153, April 2008.

[28] F. Noé, E. Schütte, C.and Vanden-Eijnden, L. Reich, and T. R. Weikl. Constructing the equilibrium ensemble of folding pathways from short off-equilibrium simulations. *Proc. Natl. Acad. Sci. USA*, 106(45):19011–19016, November 2009.

[29] D. P Landau and K. Binder. *A Guide to Monte Carlo Simulations in Statistical Physics*. Cambridge university press, 2nd edition, 2005.

[30] J. M. Haile. *Molecular Dynamics Simulations: Elementary Methods*. Wiley, 1997.

[31] F. E. Boas and P. B. Harbury. Potential energy functions for protein design. *Curr. Opin. Struct. Biol.*, 17(2):199–204, April 2007.

[32] D. Van Der Spoel, E. Lindahl, B. Hess, G. Groenhof, A. E. Mark, and H. J. Berendsen. Gromacs: Fast, flexible and free. *J. Comp. Chem*, 26(1701–1718), 2005.

[33] R. Salomon-Ferrer, D. A. Case, and R. C. Walker. An overview of the Amber biomolecular simulation package. *Wiley Interdiscip Rev Comput Mol Sci*, 3(2):198–210, March 2013.

[34] J. C. Phillips, R. Braun, W. Wang, J. Gumbart, E. Tajkhorshid, E. Villa, C. Chipot, R. D. Skeel, L. Kale, and K. Schulten. Scalable molecular dynamics with namd. *J. Comp. Chem.*, 16(26):1781–1802, December 2005.

[35] P. Eastman, M. S. Friedrichs, J. D. Chodera, R. J. Radmer, C. M. Bruns, J. P. Ku, K. A. Beauchamp, T. J. Lane, L. Wang, D. Shukla, T. Tye, M. Houston, T. Stich, C. Klein, M. R. Shirts, and V. S. Pande. OpenMM 4: A Reusable, Extensible, Hardware Independent Library for High Performance Molecular Simulation. *J. Chem. Theory Comput.*, 9(1):461–469, January 2013.

[36] M. J. Harvey, G. Giupponi, and G. De Fabritiis. ACEMD: Accelerating Biomolecular Dynamics in the Microsecond Time Scale. *J. Chem. Theory Comput.*, 5(6):1632–1639, June 2009.

[37] R. B. Best and G. Hummer. Optimized molecular dynamics force fields applied to the helix-coil transition of polypeptides. *J. Phys. Chem. B*, 113(26):9004–9015, July 2009.

[38] K. Lindorff-Larsen, S. Piana, K. Palmo, P. Maragakis, J. L. Klepeis, R. O. Dror, and D. E. Shaw. Improved side-chain torsion potentials for the Amber ff99SB protein force field. *Proteins*, 78(8):1950–1958, June 2010.

[39] Y. Duan, C. Wu, S. Chowdhury, M. C. Lee, G. Xiong, W. E. I. Zhang, R.g Yang, P. Cieplak, R. A. Y. Luo, T. Lee, J. Caldwell, J. Wang, and P. Kollman. A Point-Charge Force Field for Molecular Mechanics Quantum Mechanical Calculations. *J. Comput. Chem.*, 24(16):1999–2012, December 2003.

[40] G. A. Kaminski, R. A. Friesner, J. Tirado-Rives, and W. L. Jorgensen. Comparison with Accurate Quantum Chemical Calculations on Peptides. *J. Phys. Chem. B*, 2(105):6474–6487, 2001.

[41] A. D. MacKerell, N. Banavali, and N. Foloppe. Development and current status of the CHARMM force field for nucleic acids. *Biopolymers*, 56(4):257–265, 2001.

[42] X. Daura, A. E. Mark, and W. F. Van Gunsteren. Parametrization of Aliphatic CH n United Atoms of GROMOS96 Force Field. *J. Comp. Chem.*, 19(21):535–547, 1998.

[43] W. R. P. Scott, P. H. Hu, I. G. Tironi, A. E. Mark, S. R. Billeter, J. Fennen, A. E. Torda, T. Huber, P. Kru, and W. F. Van Gunsteren. The GROMOS Biomolecular Simulation Program Package. *J. Phys. Chem. A*, 103:3596–3607, 1999.

[44] J. B. Clarage, T. Romo, B. K. Andrews, B. M. Pettitt, and G. N. Phillips. A sampling problem in molecular dynamics simulations of macromolecules. *Proc. Natl. Acad. Sci. USA*, 92(8):3288–3292, April 1995.

[45] D. J. Earl and M. W. Deem. Parallel tempering: Theory, applications, and new perspectives. *Phys. Chem. Chem. Phys.*, 7(23):3910, November 2005.

[46] R. M. Neal. Sampling from multimodal distributions using tempered transitions. *Stat. Comput.*, 6(4):353–366, December 1996.

[47] Y. Sugita and Y. Okamoto. Replica exchange molecular dynamics method for protein folding simulation. *Chem. Phys. Lett.*, 314:141–151, January 1999.

[48] J.H. Prinz, J. D. Chodera, V. S. Pande, W. C. Swope, J. C. Smith, and F. Noé. Optimal use of data in parallel tempering simulations for the construction of discrete-state Markov models of biomolecular dynamics. *J. Chem. Phys.*, 134(24):244108, June 2011.

[49] S. Kumar, J. M. Rosenberg, D. Bouzida, R. H. Swendsen, and P. A. Kollman. THE weighted histogram analysis method for free-energy calculations on biomolecules. I. The method. *J. Comp. Chem.*, 13(8):1011–1021, October 1992.

[50] G.M. Torrie and J.P. Valleau. Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling. *J. Comput. Phys.*, 23(2):187–199, February 1977.

[51] G. Bussi, F. L. Gervasio, A. Laio, and M. Parrinello. Free-energy landscape for beta hairpin folding from combined parallel tempering and metadynamics. *J. Am. Chem. Soc.*, 128(41):13435–13441, October 2006.

[52] B. Ensing, M. De Vivo, Z. Liu, P. Moore, and M. L. Klein. Metadynamics as a tool for exploring free energy landscapes of chemical reactions. *Acc. Chem. Res.*, 39(2):73–81, February 2006.

[53] F. L. Gervasio, A. Laio, and M. Parrinello. Flexible docking in solution using metadynamics. *J. Am. Chem. Soc.*, 127(8):2600–2607, March 2005.

[54] V. Spiwok, P. Lipovová, and B. Králová. Metadynamics in essential coordinates: free energy simulation of conformational changes. *J Phys Chem B.*, 111(12):3073–3076, March 2007.

[55] A. Mitsutake, Y. Mori, and Y. Okamoto. Enhanced sampling algorithms. *Methods Mol Biol*, 924:153–195, January 2013.

[56] D. R. Roe, C. Bergonzo, and T. E. Cheatham. Evaluation of enhanced sampling provided by accelerated molecular dynamics with Hamiltonian replica exchange methods. *J. Phys. Chem. B*, 118(13):3543–3552, April 2014.

[57] T. Schlick. Molecular dynamics-based approaches for enhanced sampling of long-time, large-scale conformational changes in biomolecules. *F1000 Biol Rep*, 1(51), July 2009.

[58] A. Chakrabarty and T. Cagin. Coarse grain modeling of polyimide copolymers. *Polymer*, 51(12):2786–2794, May 2010.

[59] A. Y. Shih, A. Arkhipov, P. L. Freddolino, and K. Schulten. Coarse grained protein-lipid model with application to lipoprotein particles. *J. Phys. Chem. B*, 110(8):3674–3684, March 2006.

[60] C. Schütte, A. Fischer, W. Huisinga, and P. Deuflhard. A Direct Approach to Conformational Dynamics based on Hybrid Monte Carlo. *J. Comput. Phys.*, 151(1):146–168, May 1999.

[61] J.H. Prinz, H. Wu, M. Sarich, B. G. Keller, M. Senne, M. Held, J. D. Chodera, C. Schütte, and F. Noé. Markov models of molecular kinetics: generation and validation. *J. Chem. Phys.*, 134(17):174105, May 2011.

[62] V. S. Pande, K. Beauchamp, and G. R. Bowman. Everything you wanted to know about Markov State Models but were afraid to ask. *Methods*, 52(1):99–105, September 2010.

[63] N. Singhal, C. D. Snow, and V. S. Pande. Using path sampling to build better Markovian state models: predicting the folding rate and mechanism of a tryptophan zipper beta hairpin. *J. Chem. Phys.*, 121(1):415–425, July 2004.

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

[65] T. S. van Erp, D. Moroni, and P. G. Bolhuis. A novel path sampling method for the calculation of rate constants. *J. Chem. Phys.*, 118(17):7762, April 2003.

[66] G. R. Bowman and V. S. Pande. Protein folded states are kinetic hubs. *Proc. Natl. Acad. Sci. USA*, 107(24):10890–10895, June 2010.

[67] F. Noé, I. Horenko, C. Schütte, and J. C. Smith. Hierarchical analysis of conformational dynamics in biomolecules: transition networks of metastable states. *J. Chem. Phys.*, 126(15):155102, April 2007.

[68] A. Amadei, A. B. Linssen, and H. J. Berendsen. Essential dynamics of proteins. *Proteins*, 17(4):412–425, December 1993.

[69] F. Noé, S. Doose, I. Daidone, M. Löllmann, M. Sauer, J. D. Chodera, and J. C. Smith. Dynamical fingerprints for probing individual relaxation processes in biomolecular dynamics with simulations and kinetic experiments. *Proc. Natl. Acad. Sci. USA*, 108(12):4822–4827, March 2011.

[70] G. Pérez-Hernández, F. Paul, T. Giorgino, Gianni De Fabritiis, and F. Noé. Identification of slow molecular order parameters for Markov model construction. *J. Chem. Phys.*, 139(1):015102, July 2013.

[71] C. R. Schwantes and V. S. Pande. Improvements in Markov State Model Construction Reveal Many Non-Native Interactions in the Folding of NTL9. *J. Chem. Theory Comput.*, 9(4):2000–2009, April 2013.

[72] M. Sarich, F. Noé, and C. Schütte. On the Approximation Quality of Markov State Models. *Multiscale Model. Simul.*, 8(4):1154–1177, January 2010.

[73] W. C. Swope, J. W. Pitera, and F. Suits. Describing Protein Folding Kinetics by Molecular Dynamics Simulations. 1. Theory. *J. Phys. Chem. B*, 108(21):6571–6581, May 2004.

[74] B. G. Keller, P. Hünenberger, and W. F. van Gunsteren. An Analysis of the Validity of Markov State Models for Emulating the Dynamics of Classical Molecular Systems and Ensembles. *J. Chem. Theory Comput.*, 7(4):1032–1044, 2011.

[75] F. Noé and F. Nüske. A Variational Approach to Modeling Slow Processes in Stochastic Dynamical Systems. *Multiscale Model. Simul.*, 11(2):635–655, June 2013.

[76] F. Nüske, B. G. Keller, G. Pérez-Hernández, A. S. J. S. Mey, and F. Noé. Variational Approach to Molecular Kinetics. *J. Chem. Theory Comput.*, 10(4):1739–1752, April 2014.

[77] J. D Chodera and F. Noé. Markov state models of biomolecular conformational dynamics. *Curr. Opin. Struct. Biol.*, 25:135–144, April 2014.

[78] J. W. Gibbs. *Elementary Principles in Statistical Mechanics.* New York: Charles Scribner's Sons., 1902.

[79] L. Boltzmann. *Über das Wärmegleichgewicht zwischen mehratomigen Gasmolekülen*, pages 397–418. Number 63. Wiener Berichte, 1871.

[80] N. G. van Kampen. *Stochastic Processes in Physics and Chemistry.* Elsevier, Amsterdam, 4th edition edition, 2006.

[81] O. Perron. Zur Theorie der Matrices. *Mathematische Annalen*, 64(2):248–263, jun 1907.

[82] G. Frobenius. *Über Matrizen aus nicht negativen Elementen.* Walter De Gruyter Incorporated, 1912.

[83] C. Schütte and W. Huisinga. Biomolecular conformations can be identified as metastable sets of molecular dynamics. In P. Ciarlet and C. Bris, editors, *Special Volume*, volume X, pages 699 – 744, 2003.

[84] Christof Schütte and Marco Sarich. *Metastability and Markov State Models in Molecular Dynamics: Modeling, Analysis, Algorithmic Approaches.* A copublication of the AMS and the Courant Institute of Mathematical Sciences at New York University, 2013.

[85] G. Voronoi. Nouvelles applications des paramètres continus à la théorie des formes quadratiques. . *Reine Angew. Mat*, 133:97–178, 1907.

[86] T. F. Gonzalez. Clustering to minimize the maximum intercluster distance. *Theoretical Computer Science*, 38:293–306, 1985.

[87] H. Steinhaus. Sur la division des corps matériels en parties. *Bull. Acad. Pol. Sci., Cl. III*, 4:801–804, 1957.

[88] J. A. Hartigan. *Clustering Algorithms.* John Wiley & Sons, Inc., feb 1975.

[89] P. Deuflhard, W. Huisinga, A. Fischer, and Ch. Schütte. Identification of almost invariant aggregates in reversible nearly uncoupled Markov chains. *Linear Algebra Appl.*, 315(1-3):39–59, August 2000.

[90] P. Deuflhard and M. Weber. Robust Perron cluster analysis in conformation dynamics. *Linear Algebra Appl.*, 398:161–184, March 2005.

[91] S. Röblitz and M. Weber. Fuzzy spectral clustering by PCCA+: application to Markov state models and data classification. *Adv Data Anal Classif*, 7(2):147–179, may 2013.

[92] A. Clark, G. J. Cooper, C. E. Lewis, J. F. Morris, A. C. Willis, K. B. Reid, and R. C. Turner. Islet amyloid formed from diabetes-associated peptide may be pathogenic in type-2 diabetes. *Lancet*, 2(8553):231–234, August 1987.

[93] P. Westermark, E. Wilander, G. T. Westermark, and K. H. Johnson. Islet amyloid polypeptide-like immunoreactivity in the islet B cells of type 2 (non-insulin-dependent) diabetic and non-diabetic individuals. *Diabetologia*, 30(11):887–892, November 1987.

[94] C. Oosterwijk, J. W. M. Höppener, K. L. van Hulst, and C. J. M. Lips. Pancreatic islet amyloid formation in patients with noninsulin-dependent diabetes mellitus. *Int. J. Pancreatol.*, 18(1):7–14, August 1995.

[95] P. Cao, A. Abedini, and D. P. Raleigh. Aggregation of islet amyloid polypeptide: from physical chemistry to cell biology. *Curr. Opin. Struct. Biol.*, 23(1):82–89, February 2013.

[96] P. Westermark, U. Engström, K. H. Johnson, G .T. Westermark, and C. Betsholtz. Islet amyloid polypeptide: pinpointing amino acid residues linked to amyloid fibril formation. *Proc. Natl. Acad. Sci. U.S.A.*, 87(13):5036–5040, July 1990.

[97] C. Betsholtz, L. Christmanson, U. Engström, F. Rorsman, K. Jordan, T. D. O'Brien, M. Murtaugh, K. H. Johnson, and P. Westermark. Structure of cat islet amyloid polypeptide and identification of amino acid residues of potential significance for islet amyloid formation. *Diabetes*, 39(1):118–122, January 1990.

[98] R. T. McGibbon, K. A. Beauchamp, C. R. Schwantes, L.-P. Wang, C. X. Hernandez, M. P. Harrigan, T. J. Lane, J. M. Swails, and V. S. Pande. MDTraj: a modern, open library for the analysis of molecular dynamics trajectories. Technical report, September 2014.

[99] C. Kabsch, W .and Sander. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22(12):2577–2637, December 1983.

[100] K. Q. Hoffmann, M. McGovern, C. C. Chiu, and J. J. de Pablo. Secondary Structure of Rat and Human Amylin across Force Fields. *PloS one*, 10(7):e0134091, jan 2015.

[101] F. Nüske, R. Schneider, F. Vitalini, and F. Noé. Variational tensor approach for approximating the rare-event kinetics of macromolecular systems. *submitted to Journal of Chemical Physics*, 2015.