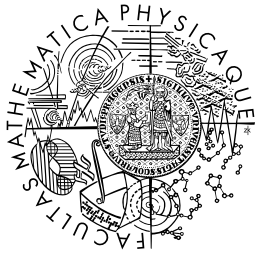


FACULTY
OF MATHEMATICS
AND PHYSICS
Charles University

MASTER THESIS

Jakub Kopko

Comparative Markov state analysis of APOE protein dynamics by neural networks

Department of Software and Computer Science Education

Supervisor of the master thesis: Jiří Sedlář
Study programme: Computer Science
Study branch: Artificial Intelligence

Prague 2023

I declare that I carried out this master thesis independently, and only with the cited sources, literature and other professional sources. It has not been used to obtain another or the same degree.

I understand that my work relates to the rights and obligations under the Act No. 121/2000 Sb., the Copyright Act, as amended, in particular the fact that the Charles University has the right to conclude a license agreement on the use of this work as a school work pursuant to Section 60 subsection 1 of the Copyright Act.

In ..... date ....

Author's signature

I would like to express my heartfelt gratitude to the remarkable individuals who contributed significantly to the completion of this thesis. Foremost, I am deeply thankful to Jiří Sedlář, my supervisor, whose unwavering guidance, expertise, and support were invaluable throughout this journey.

I am immensely grateful to Josef Šivic for his role as my advisor, providing invaluable insights and shaping the direction of my research. I extend my sincere appreciation to Petr Kouba for his collaborative spirit and shared enthusiasm in exploring new ideas and approaches.

I would also like to extend my gratitude to Sérgio Marques, Joan Planas-Iglesias, David Bednář, and Stanislav Mazurenko from the Loschmidt Laboratories at Masaryk University in Brno. Their expertise and contributions significantly enhanced the quality of this work.

Furthermore, I would like to acknowledge the generous support provided by the Visegrad Scholarship fund. Their funding enabled me to pursue this research and further my academic endeavors.

To Jiří Sedlář, Josef Šivic, Petr Kouba, Sérgio Marques, Joan Planas-Iglesias, David Bednář, Stanislav Mazurenko – I extend my deepest gratitude for their instrumental roles in this thesis. Their guidance, encouragement, and support have been invaluable, and I am truly honored to have collaborated with them on this important project.

Title: Comparative Markov state analysis of APOE protein dynamics by neural networks

Author: Jakub Kopko

Department: Department of Software and Computer Science Education

Supervisor: Jiří Sedlář, Czech Institute of Informatics, Robotics, and Cybernetics, Czech Technical University in Prague

Abstract: This thesis leverages the CoVAMPnet neural network architecture to analyze the dynamics of apolipoprotein E (APOE), a protein involved in the development of Alzheimer's disease. CoVAMPnet offers a versatile machine learning framework for extracting meaningful features from high-dimensional molecular dynamics data and constructing Markov state models to characterize protein conformational dynamics. By applying CoVAMPnet to APOE simulations, the thesis successfully captures the protein behavior by revealing its key conformational states and structural transitions. These findings provide new insights into the dynamics of APOE and its potential role in Alzheimer's disease. The thesis also investigates the influence of a small molecule drug candidate 3SPA on APOE's conformational behavior, shedding further light on its therapeutic possibilities. Overall, this work demonstrates CoVAMPnet's effectiveness in analyzing and comparing the dynamics of larger proteins in an interpretable manner, reinforcing its potential application for complex biomolecular studies.

Keywords: Machine learning for molecular dynamics, Neural networks, Variational approach to Markov processes, Markov state models, APOE protein

Contents

Introduction

Alzheimer's disease (AD) is a neurodegenerative disorder characterized by a progressive decline in cognitive function, memory loss, and behavioral changes. As the most prevalent form of dementia, AD affects over six million adults above 65 years old in the United States alone, and it is anticipated that this number will rise to nearly fourteen million by 2060 [1].

The immense burden of AD extends beyond the affected individuals themselves by impacting their caregivers and placing a significant financial strain on healthcare systems. The average lifetime cost of care for a patient with dementia was estimated in 2022 at 392,874 US dollars [2]. The development and testing of new drugs to combat AD incur substantial costs, making it imperative to deepen our understanding of the disease's underlying mechanisms and identify novel therapeutic strategies [2].

Among the various factors contributing to AD, one of the most influential genetic risk factors is the presence of the APOE4 variant of the polymorphic apolipoprotein E (APOE) [3]. Extensive research has been conducted on APOE, shedding light on its structure and function in recent years. Interestingely, APOE4 differs from the more common, neutral variant called APOE3 by just a single-point mutation. However, despite significant progress, we still lack a comprehensive understanding of how such a small change can increase the risk of this debilitating disease.

The APOE protein is contains flexible regions that significantly contribute to its functional properties and determine its interactions with other molecules. Due to its dynamic nature, studying APOE requires a comprehensive analysis of temporal data describing it. This task poses a challenge as it involves capturing the complex and intricate movements and conformational changes that occur within the protein. Gaining a deeper understanding of APOE's dynamic behavior is essential for unraveling its biological mechanisms and exploring potential therapeutic interventions.

This master's thesis employs the VAMPnet-based [4] neural network architecture CoVAMPnet [5] to investigate the dynamics of APOE, focusing on its association with the pathogenesis of Alzheimer's disease. Our research provides novel insights into the conformational dynamics of APOE and serves as a basis for investigating new ideas that may offer a deeper understanding of the involvement of APOE in the development and progression of Alzheimer's disease.

1. Background

1.1 Proteins and their structure

Proteins are vital macromolecules that play a central role in the functioning of living organisms. They are involved in a wide range of biological processes, including catalyzing chemical reactions, providing structural support, facilitating cellular communication, and serving as transporters and regulators [6].

The function of a protein is determined by its structure, i.e. the three-dimensional arrangement of its atoms. Proteins are composed of long chains of amino acids, that fold into specific shapes. There are 20 distinct types of amino acids that can be present in these chains. These individual amino acids are referred to as residues. The primary structure of a protein is the linear sequence of amino acids, while the secondary structure refers to local patterns such as alpha helices (see Fig. 1.1) and beta sheets. The tertiary structure is the overall three-dimensional arrangement of the protein. It is primarily dictated by the interactions between different amino acids. In some cases, proteins can have multiple chains, leading to a quaternary structure.

Proteins are often composed of domains, which can be identified using various methods. Sequence-based methods involve analyzing the amino acid sequence to detect conserved motifs and patterns. Structure-based methods rely on the three-dimensional structure of the protein obtained through X-ray crystallography, nuclear magnetic resonance (NMR), or cryogenic electron microscopy (cryo-EM). Hybrid methods integrate sequence-based and structure-based information, combining sequence similarity searches with structural data. Some protein domains can be also defined based on functional properties observed by other kinds of analysis [7].

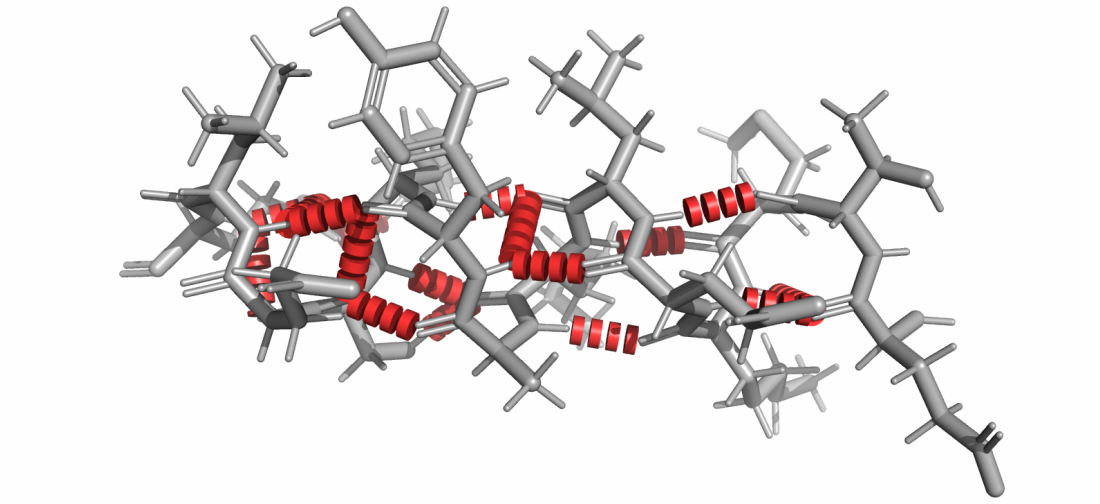
However, domain identification can be challenging for proteins with complex structure or dynamic behavior. To overcome these challenges, a combination of methods – including experimental data, computational predictions, and expert knowledge – is often required to identify the protein domains accurately. Moreover, the precise boundary between domains is often blurry, and the ultimate decision may be based on an expert's opinion [7].

In this work, the main focus will be on the full shape of the protein, i.e., its tertiary structure, but most of the changes will be related to changes on the level of secondary structures, specifically the alpha helices.

The alpha helix (see Fig. 1.1) is a secondary structure commonly found in proteins, characterized by a right-handed helical arrangement of amino acids. Within the alpha helix, each amino acid residue forms hydrogen bonds with its neighboring residues, creating a helical backbone. Specifically, the carbonyl oxygen of one residue forms a hydrogen bond with the amide hydrogen of the fourth residue downstream. This pattern repeats, resulting in a stable and tightly packed structure. The helix is further stabilized by van der Waals interactions between the side chains of adjacent residues, contributing to the overall stability and integrity of the alpha helical conformation [8]. The alpha helix plays a crucial role in maintaining the three-dimensional structure of proteins and is involved in diverse biological functions, including protein-protein interactions, membrane-

spanning regions, and DNA binding motifs.

(a) Cartoon view.



(b) Atomic view.

Figure 1.1: Visualization of the helical structure in proteins. Hydrogen bonds stabilizing the helical structure are shown in red.

The Define Secondary Structure of Proteins (DSSP) algorithm [9] is a widely used computational tool for the identification of secondary structures in proteins from their atomic coordinates. DSSP assigns to each residue a secondary structure element, e.g., alpha helix, beta sheet, or coil, based on hydrogen bonding patterns and geometric criteria. In this thesis, most of the observed results concern changes of local helical structure of the protein, so DSSP turned out to be an important tool in our analyses.

1.2 Protein dynamics and function

Protein structure plays a crucial role in determining the protein function. The three-dimensional arrangement of amino acids in a protein dictates its ability to

interact with other molecules and perform specific biological tasks. Equally important is the protein dynamics, which refers to the flexibility and motion within protein structures. The protein dynamics allows for conformational changes necessary for binding, catalysis, and regulation, which enable proteins to adapt and function in response to their environment. Understanding protein dynamics is essential for unraveling the full spectrum of protein functionality and designing effective therapeutic interventions [10].

Intrinsically disordered proteins (IDPs) are a class of proteins that lack a clearly defined structure in at least one of their domains. A full understanding of their functionality is usually impossible without a thorough investigation of their dynamics and flexibility by experimental and computational methods.

Let us consider the tau protein as an example of an IDP. Unbound, the tau protein lacks a stable 3D structure, embodying remarkable flexibility. However, its shape becomes fixed upon binding to nerve cell microtubules, stabilizing them in the process. In the context of Alzheimer's disease, the tau protein behavior becomes consequential as anomalous alterations of its structure lead to the protein aggregating into tangles, a hallmark of the disease [11].

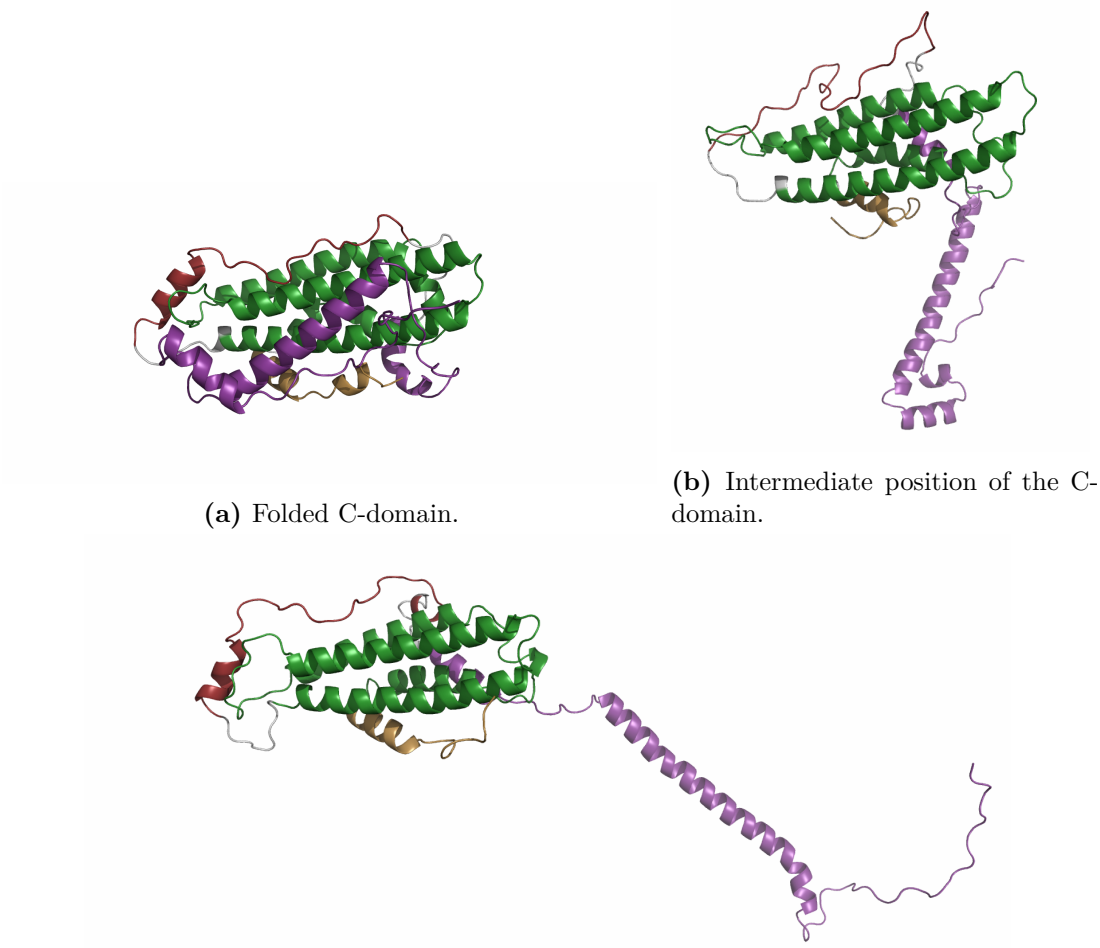
The importance of analyzing the properties of IDPs was highlighted in [12]. The authors make a strong case that understanding the dynamic variety of shapes these proteins can adopt is as vital as comprehending the static 3D structure of stable biomolecules. The paper showcases 17 Nobel Prize discoveries where IDPs played a significant role [12].

1.3 Apolipoprotein E and Alzheimer's disease

Apolipoprotein E (APOE) is a protein that plays a critical role in lipid metabolism and transport in the body. It is primarily produced by the liver and by astrocyte cells in the brain. APOE is involved in the regulation of cholesterol and triglyceride levels by interacting with lipoprotein particles, which are responsible for transporting fats in the bloodstream [13].

The APOE is also an IDP. It consists of a single chain with 299 residues. On a high level, we can identify at least four distinct domains: the N-terminus (residues 1-21), the 4-helix bundle (residues 24-161), the hinge region (residues 170-199), and the C-domain (residues 210-299) (see Fig. 1.2).

A dominant dynamical phenomenon observed in APOE is the unfolding of the C-domain (see Fig. 1.2). The C-domain is able to bind to several kinds of lipids, which would be impossible without its dynamical adaptability [14].



(a) Folded C-domain.

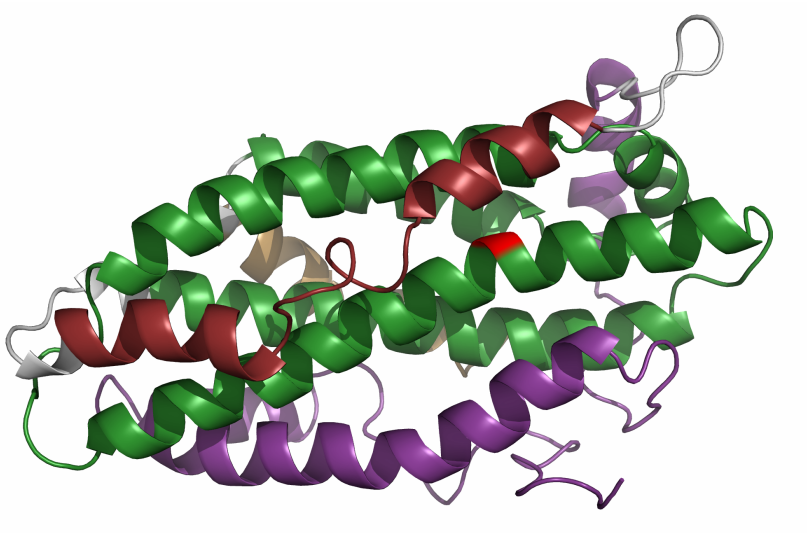
(b) Intermediate position of the C-domain.

(c) Unfolded C-domain.

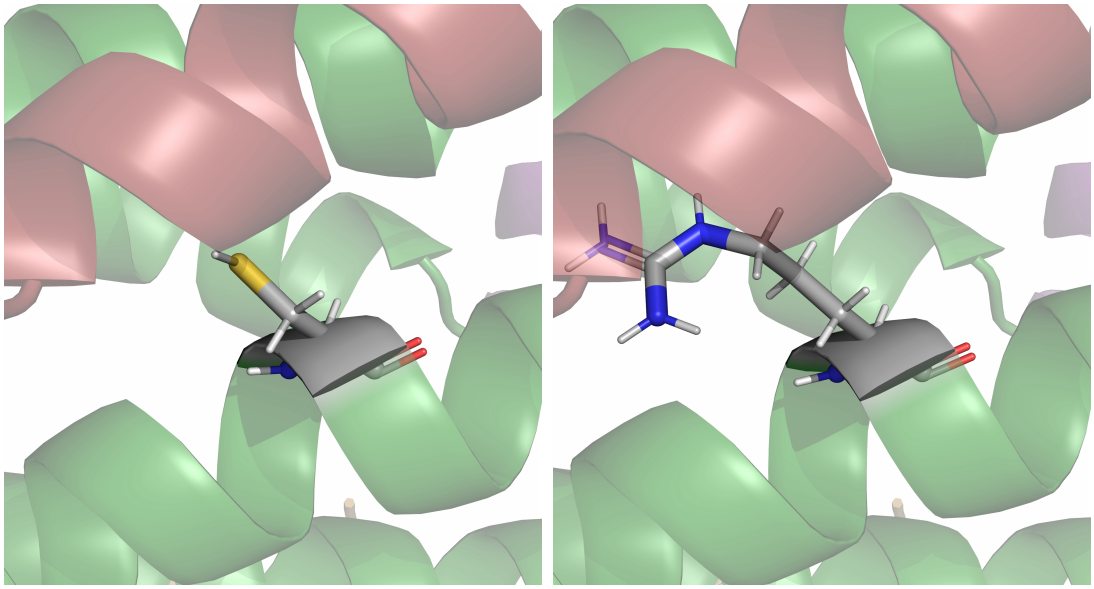
Figure 1.2: Several conformations of APOE3 showcasing the high flexibility of the C-domain (purple). Colors: light brown – N-terminus, dark green – 4-helix bundle, dark red – hinge region, purple – C-domain.

In humans, the APOE gene exhibits three predominant alleles: E2 (APOE2), E3 (APOE3), and E4 (APOE4). These alleles are associated with varying levels of Alzheimer's disease (AD) risk. APOE4, in particular, represents a major genetic risk factor, with the risk increasing up to 12-fold in individuals who are homozygous for this allele [3]. On the other hand, APOE2 reduces the risk of AD by nearly half and is linked to increased longevity [3].

In this thesis, we primarily examine the differences between APOE3 and APOE4. It is notable that these two variants differ merely by a single-point mutation. Specifically, the 112th residue contains cysteine in APOE3, but arginine in APOE4 (see Fig. 1.3).



(a) Location of the mutated residue 112 (bright red).



(b) Cysteine in position 112 in APOE3. **(c)** Arginine in position 112 in APOE4.

Figure 1.3: Location of C112R mutation. Notice the difference in size between cysteine and arginine. Colors: light brown – N-terminus, dark green – 4-helix bundle, dark red – hinge region, purple – C-domain.

Numerous studies have focused on understanding the impact of the APOE4 mutation on the brain function, leading to several hypotheses. One hypothesis suggests that the single mutation in APOE4 impairs its ability to effectively clear the amyloid beta protein from the brain, leading to increased aggregation of amyloid plaques, which are known to be neurotoxic [15]. However, it is important to note that APOE4 also damages the brain through mechanisms unrelated to the abeta protein. It was proven that APOE4 has detrimental influence on the brain-blood barrier [16]. Other hypotheses include impaired lipid clearance [17], interaction with the tau protein [18], or direct involvement through aggregation processes [19].

Establishing the root cause of the neuropathological influence of APOE4 becomes increasingly intricate when we acknowledge that each of these hypotheses is

backed by substantial evidence. It is important to recognize that these hypotheses are not mutually exclusive, as the available evidence suggests that the link between the types of APOE and the Alzheimer's disease may consist of multiple factors rather than a single isolated cause.

Given the multiple ways in which APOE4 can detrimentally affect the brain, it is evident that further exploration of its properties is of great relevance to medical research. This serves as the primary motivation for the work presented in this thesis, as unraveling the intricate mechanisms of APOE and its variants holds significant potential for advancing our understanding and developing targeted therapeutic strategies.

2. Related work

In this chapter, we give a concise summary of the methods used to gather data describing kinetic aspects of proteins and some of the basic methods of its analysis. We also provide a mathematical background for the methods we employed and the basic overview of the utilized neural network architecture. We also highlight relevant studies and literature about APOE, especially those related to our research hypothesis.

2.1 Molecular dynamics

Molecular dynamics (MD) is a discipline focused on simulating physical interactions at the molecular level, aiming to achieve precise and accurate representations of molecular systems. It can be classified into two main categories: quantum MD and classical MD. While quantum MD provides a detailed description of molecular behavior using quantum mechanics, it is often computationally intensive and time-consuming. In contrast, classical MD utilizes high-resolution classical physics simulations as an approximation to mimic quantum phenomena. This approach enables faster computations while striving to capture the essential features of molecular systems. By employing classical MD simulations, researchers can strike a balance between computational efficiency and maintaining a reasonable level of accuracy in modeling molecular dynamics [20].

MD simulations play a crucial role in the comprehensive understanding of protein behavior, supplementing the static pictures derived from techniques like crystallography or NMR. While these "static" methods provide invaluable insights into a protein's stable conformations, they cannot capture the dynamic nature of proteins, i.e., their constant conformational fluctuations. These dynamic changes are fundamental to a protein's function, including how it interacts with other molecules or responds to changes in its environment. MD simulations, by modeling the motions of atoms over time, can fill this gap in our knowledge. They allow us to observe transitions between different states, estimate their likelihood, and, importantly, understand the mechanistic details of these processes. Thus, MD simulations are an essential tool in the field of structural biology, providing a dynamic view that complements the static snapshots of other techniques.

In this thesis we analyze data obtained by classical MD simulations. Classical MD simulations are still computationally demanding, to the extent that dedicated supercomputers are employed to tackle some simulations [21]. Big molecular systems pose significant challenges that make them especially expensive to simulate. On top of that, the approximation of quantum effects by classical physics introduces inherent limitations. MD also requires a setting of hyperparameters, such as the simulation temperature, and naturally encounters the limitations of numerical approximation. Many processes can be simulated only in radically simplified versions. For example, in the case of bigger proteins, usually only a single biomolecule surrounded by water molecules is simulated. Such simplification does not account for the realistically complex environment in which multiple biomolecules can interact with each other simultaneously.

An important concept in MD are the timescales of the observed processes. It

is important to distinguish two factors that determine them. On one hand, some processes can truly span over the course of milliseconds, which is very long in the MD domain. This kind of "slow" dynamics is naturally difficult to observe in a simulation due to the time required to span the whole trajectory of the process of interest. An example of such a process is a slow binding of two big biomolecules. If the target conformation is known, one possible solution is introduction of a guiding force [22].

In this study, we have to deal with a different kind of "slow" dynamics, namely the rareness of the transitions. While a transition between two conformations may be quick, if its occurrence is rare, it will require proportionally more time to observe in a simulation. Hence we also refer to such dynamics as slow.

If the transitions between main conformations are rare, we call such states metastable. It means that a protein does not "easily" change its overall conformation. Interestingly, such metastable states are a proxy for free energy minima. The lower the free energy of a conformation, the more thermodynamically stable it is, and the more likely the protein is to adopt it. On the other hand, a protein becomes more unstable with higher free energy, so transitional states usually exhibit high free energy. That is why conformational landscapes generated by MD are often referred to as free energy landscapes, and rarely populated transitional areas are referred to as high energy barriers [23].

Note that we usually do not know the overall conformational landscape of the system of interest. Therefore, we would like to find out what are its main metastable states and the transitions between them.

2.2 Time-lagged Independent Component Analysis

Large amounts of MD simulations need to be generated to capture highly intricate free energy landscapes of protein systems. The free energy landscape embodies the dynamic nature of proteins and other biomolecules, depicting a multitude of energy barriers that reflect the convoluted behavior of these molecules.

However, navigating this vast configuration space presents a formidable challenge due to its high-dimensionality, where the number of interacting atoms can reach tens of thousands. This complexity necessitates the use of dimension reduction techniques designed to simplify the analysis of such complicated systems.

Time-lagged independent component analysis (tICA) [24] is a dimensionality reduction method commonly used for MD data. tICA is appreciated for its ability to account for the temporal dimensions of the data. It identifies the collective degrees of freedom that exhibit the strongest time-correlations for a given lag-time. This characteristic of tICA proves particularly beneficial when exploring how a protein shifts between different conformations over time. In addition, tICA serves as an effective preprocessing phase for representing the conformational dynamics of macromolecules in the setting of a discrete Markov process.

In our current research, we selected tICA as a method of choice due to this added temporal dimension it offers. By including time information from the input trajectory, tICA offers a more comprehensive perspective of the protein dynamics. This makes it particularly advantageous in the analysis of complex biomolecular

systems.

To formally estabilish tICA, we will follow the notation used in [25].

Consider a d-dimensional trajectory $x(t) \in \mathbb{R}^d, t = 1, \ldots, T$ with Cartesian coordinates $x_1, \ldots, x_d$, which are assumed to be mean-free, i.e., the time average $\langle x(t) \rangle_t$ is zero.

tICA determines those "slowest" independent collective degrees of freedom $v_k \in \mathbb{R}^d, k = 1, \ldots, d$, onto which the projections $y_k(t) = v_k \cdot x(t)$ have the largest time-autocorrelation

$$\frac{\langle y_k(t) y_k(t+\tau) \rangle_t}{\langle y_k(t)^2 \rangle_t}$$

where τ is a chosen lag time.

This can be equivalently formulated using the time-lagged covariance matrix $C(\tau) = (\langle x_i(t) x_j(t+\tau)_t \rangle)_{ij} \in \mathbb{R}^{d \times d}$. Each degree of freedom v_k maximizes $v_k^T C(\tau) v_k / v_k^T C(0) v_k$ under the constraint that it is orthogonal to all previous degrees of freedom. Hence, the v_k are the solutions of the generalized eigenvalue problem $C(\tau) v_k = \lambda_k C(0) v_k$[25].

tICA aims to maximize the time-lagged autocorrelation along each component. This unique feature makes tICA particularly suitable for analyzing MD simulation data, where the slowest changing features (associated with high autocorrelation) are often of the most interest.

In our work we used tICA to visualize the energy landscapes of systems of interest in two dimensions. Interestingly, we observed well separated energy islands in all of them. This suggests the existence of significant conformational jumps between metastable conformations being a part of the analyzed data.

2.3 Markov state models

Markov state models (MSMs) are a powerful tool extensively employed in molecular dynamics simulations to characterize and analyze the dynamic behavior of molecular systems. Essentially, MSMs allow us to break down a complex molecular system into discrete states, and provide a way to study the transitions between them.

An MSM provides a probabilistic view of molecular dynamics, which gives us an understanding of how a molecule behaves over time. Each state in an MSM represents a specific set of conformations of the molecule, and the transitions between states depict the molecule's movement between them. In the most common scenario we want to build a MSM which states correspond to metastable conformations separated by high-energy barriers. This probabilistic model can yield both static and dynamic information about the molecular system.

From a static perspective, the properties of each state can provide intriguing insights into the structure and functionality of the molecular system. For instance, a specific state could reveal a protein's binding site, contributing valuable information to drug design efforts.

From a dynamic perspective, MSMs can shed light on the relative prevalence of certain states and the likelihood of transitions between the states. This knowledge is essential to understanding functional processes, like a protein's interaction with a ligand or lipid, and can also provide clues about the system's reaction to different

conditions or perturbations. Uncovering unknown transitional states may inspire development of new drugs addressing harmful transitions, for example.

However, constructing Markov state models can be quite challenging, particularly for systems characterized by high dimensionality and complex dynamics. The traditional process of building an MSM involves a multi-step pipeline [26], which includes :

- Choosing relevant features for analysis: This initial step involves deciding which aspects or properties of the molecular system should be included in the model. The choice of features can greatly impact the accuracy and interpretability of the resulting MSM.
- Performing spatial clustering: Once the features have been selected, the next step involves dividing the conformational space into discrete states based on spatial properties. This task can be complex due to the high dimensionality of the data and the need to ensure that each state is meaningfully distinct from the others. This step results in a data clustered into hundreds or thousands of microstates.
- Performing dynamical coarse-graining: The final step usually involves a process known as Perron Cluster-Cluster Analysis (PCCA) [27], which groups the microstates into macrostates based on their dynamical properties. This reduces the complexity of the model while preserving the essential dynamics of the system. The final number of states need to be carefully chosen to strike the right balance between expressiveness and interpretability.
- On top of clustering itself the transition rates between macrostates are also calculated.

2.3.1 Adaptive sampling

It is often impractical to observe rare transitions within a single simulation. To address this issue, enhanced sampling methods [28] have emerged as a common solution in protein dynamics research. These methods involve dynamically adjusting the simulation conditions to focus on the most relevant regions or events, allowing for more efficient exploration of the conformational space and enabling the observation of rare and important molecular events.

Adaptive sampling [29] aims to allocate computational resources more efficiently by focusing on regions of the conformational space that are most relevant to the scientific question at hand. It typically consists of several key stages. In the adaptive sampling regime, the full simulation can be divided into epochs consisting of several simulations. Usually simulations in one epoch are computed in a parallel manner.

The simulations in the first epoch are initalized with the known 3D structures of proteins, for example coming from the NMR or crystallography data, or some previously performed simulation. The simulations are run in the standard manner. This generates an initial dataset covering a broad conformational space.

Based on some arbitrarily chosen sampling criterion single or several frames are probabilistically chosen.A popular criterion is based on the traditional MSMs (see Sec. 2.3): after each epoch MSM is constructed from all the data obtained

so far. Then, based on properties of the constructed MSM frames are sampled. Specifically, frames are more likely to be chosen if they belong to macrostates that have been less explored [29].

Those frames are then used to initialize the next epoch of simulations. Due to the deterministic nature of the classical physics, a randomizing factor must be added, and it usually takes a form of randomly modifying the velocities of some of the atoms in the chosen frame. Next frame is computed and serves as the initial frame of the simulation. This process is iteratively performed with sampling done on frames from all the previous epochs. Adaptive sampling allows for a more targeted exploration while still allowing for the discovery of new regions. By biasing the simulation towards these regions, we are able to observe rare transitions in the radically shorter total simulation time compared to running a single simulation without adaptive sampling and often results in a more comprehensive exploration of the conformational space.

In this approach the most important is the level of detail the MSMs are built upon. In situations where we believe a very precise sampling is required to reach some interesting transitions, all information about the protein structure and position may be used. This is however time consuming and usually MSMs are constructed based on a very restricted representation, for example positions of residues or secondary structure in the region of interest [30].

Interestingely, the way MSM is built can also differ, for example authors of [31] created a highly efficient VAMPnet-based [4] architecture for constructing MSMs during adaptive sampling procedure.

It is important to remember, however, that bias introduced by the sampling criterion will almost always influence what conformations we will be able to observe in most realistic scenarios, because usually we are only able to sample a small part of the full conformational landscape [30].

2.4 Koopman theory

The Koopman theory is one of the most dominant frameworks in nonlinear dynamics, which was gaining in popularity in the recent years. This perspective leverages an infinite-dimensional linear operator, the Koopman operator, which acts on the space of all possible measurement functions of the system. Consequently, this theory enables prediction, estimation, and control of nonlinear systems using methods conventionally associated with linear systems [32].

The theory presented in this section is based on [33], but the notation in this and following sections is adapted to the notation used in [4].

Under some commonly used assumptions, molecular dynamics can theoretically be described as a Markov process $\{x_t\}$ in the state space Ω. The dynamics are fully characterized by a transition density $p_\tau(x, y)$. This density signifies the probability that a molecular dynamics trajectory at configuration x will transition to configuration y after a time lag τ. The Markov property allows sampling y from x alone, negating the need for previous time steps.

Although variables x_t can exhibit high non-linearity, the application of Koopman theory reveals their transformation into latent variables that evolve linearly on average.

Formally, transformations

$$\chi_0(x) = (\chi_{01}(x), \ldots, \chi_{0m}(x))^T$$

and

$$\chi_1(x) = (\chi_{11}(x), \ldots, \chi_{1m}(x))^T$$

exist such that the dynamics in these transformed variables are approximated by matrix K:

$$\mathbb{E}[\chi_1(x_{t+\tau})] \approx K^T \mathbb{E}[\chi_0(x_t)] \tag{2.1}$$

The accuracy of this approximation improves with an increase in the number of features ($m \to \infty$), eventually becoming exact [33]. However, even with a large lag time τ, a satisfactory approximation can be achieved with low-dimensional feature transformations[33]. We think about the χ transformations as one for the 'present' (χ_0) and one for the τ 'lag time in the future' (χ_1), what is justified by the fact that for a limited data first lag time points do not have a past and last lag time points do not have a future, but as we will see, in practice, to keep the results more interpretable and intuitive, only one transformation is used.

Equation 2.1 can be elucidated by considering $\{x_t\}$ as a discrete-state Markov chain. Here, if the feature transformation is defined by indicator functions – ($\chi_{0i} = 1$ when $x_t = i$ and 0 otherwise, and similarly with χ_{1i} and $x_{t+\tau}$), therefore m corresponds to the number of states in the chain – their expectation values are equivalent to the probabilities p_t and $p_{t+\tau}$ of the chain being in any given state. Subsequently, K mirrors the matrix of transition probabilities, denoted by $p_{t+\tau} = P(\tau)p_t$ [33].

2.5 Variational Approach to Markov Processes

Variational Approach to Markov Processes (VAMP) [33] adds a practical framework to the Koopman operator theory. While the Koopman operator provides a theoretical foundation for linearizing the evolution of complex dynamical systems, VAMP offers an efficient way to optimally approximate the Koopman operator in finite dimensions, thus enabling its use in practice.

The core of the VAMP theory suggests that the optimal finite-dimensional linear model is obtained when the subspaces spanned by χ_0 and χ_1 align with those of the top m left and right singular functions of the Koopman operator [33]. Given a certain feature transformation χ_0 and χ_1, we define the covariance matrices as:

$$C_{00} = \mathbb{E}_t[\chi_0(x_t)\chi_0(x_t)^T], \tag{2.2}$$

$$C_{01} = \mathbb{E}_t[\chi_0(x_t)\chi_1(x_{t+\tau})^T], \tag{2.3}$$

$$C_{11} = \mathbb{E}_{t+\tau}[\chi_1(x_{t+\tau})\chi_1(x_{t+\tau})^T], \tag{2.4}$$

where $\mathbb{E}_t[\cdot]$ and $\mathbb{E}_{t+\tau}[\cdot]$ are the averages over time points and lagged time points within and across trajectories, respectively. The optimal K that minimizes the least square error $\mathbb{E}_t[\chi_1(x_{t+\tau}) - K^T\chi_0(x_t)]^2$ is: $K = C_{00}^{-1}C_{01}$ [4].

Choosing appropriate transformations χ_0 and χ_1 is nontrivial. Consider an example where $\chi_0(x) = \chi_1(x) = 1(x)$; here, the least square error is zero for $K = [1]$, but the model yields no dynamic information [33].

To find a solution we need a central theorem of the VAMP theory [33]:

Theorem 1 (Optimal Approximation of Koopman Operator)**.** *Let K_τ be a Hilbert-Schmidt operator between the separable Hilbert spaces $L^2_{\rho_1}$ and $L^2_{\rho_0}$. The linear model 2.1 with the smallest modeling error in the Hilbert-Schmidt norm is given by $\chi_0 = (\chi_{0_1}, \ldots, \chi_{0_m})^T$, $\chi_1 = (\chi_{1_1}, \ldots, \chi_{1_m})^T$, and $K = diag(\sigma_1, \ldots, \sigma_m)$, i.e.,*

$$\mathbb{E}[\chi_{1_i}(x_{t+\tau})] = \sigma_i \mathbb{E}[\chi_{0_i}(x_t)], \quad for \quad i = 1, \ldots, m \tag{2.5}$$

subject to the constraint $dim(\chi_0), dim(\chi_1) \leq m$. The projected Koopman operator derived from 2.5 is

$$\hat{K}_\tau \chi_1 = \sum_{i=1}^{m} \sigma_i \langle \chi_1, \chi_{1_i} \rangle_{\rho_1} \chi_{0_i}, \tag{2.6}$$

where the singular value $\sigma_i > 0$ is the square root of the i-th largest eigenvalue of $K^*_\tau K_\tau$ or $K_\tau K^*_\tau$, the left and right singular functions χ_{0_i}, χ_{1_i} are the i-th eigenfunctions of $K^*_\tau K_\tau$ and $K_\tau K^*_\tau$ with

$$\langle \chi_{0_i}, \chi_{0_j} \rangle_{\rho_0} = 1_{i=j}, \quad \langle \chi_{1_i}, \chi_{1_j} \rangle_{\rho_1} = 1_{i=j}, \tag{2.7}$$

and the first singular component is always given by $(\sigma_1, \chi_{1_1}, \chi_{0_1}) = (1, \mathbb{1}, \mathbb{1})$ with $1(x) \equiv 1$.

VAMP introduces a useful scoring system, which also allows us to express the theorem above in a way useful for machine learning [4]. Given any two sets of linearly independent functions $\chi_0(x)$ and $\chi_1(x)$, we define their VAMP-2 score, denoted as $\hat{R}^2_{\chi_0,\chi_1}$, as follows:

$$\hat{R}^2_{\chi_0,\chi_1} = \left\| C_{00}^{-\frac{1}{2}} C_{01} C_{11}^{-\frac{1}{2}} \right\|_F^2 \tag{2.8}$$

where C_{00}, C_{01}, C_{11} are defined as in 2.4 and $\|A\|_F^2 = \frac{1}{n} \sum_{i,j} A_{ij}^2$ is the Frobenius norm of an $n \times n$ matrix A.

The maximum value of the VAMP-2 score is obtained when the top m left and right singular functions of the Koopman operator are contained in $\text{span}(\chi_{0_1}, \ldots, \chi_{0_m})$ and $\text{span}(\chi_{1_1}, \ldots, \chi_{1_m})$, respectively [4], and based on the theorem 1 we know that such a solution results in a smallest modelling error.

Consequently, we can optimize χ_0 and χ_1 by maximizing the VAMP-2 score. This way VAMP establishes a scoring system that can be used to construct a machine learning loss function enabling learning of the χ transformations directly from data [4].

2.6 VAMPnet

Neural networks, due to their capabilities as universal function approximators [34] and their potential for expressing strongly nonlinear functions are a natural choice for approximating the χ transformations (see Sec. 2.4). VAMPnets are neural networks that utilize VAMP scores as their loss functions [4].

VAMPnets utilize two parallel lobes to simultaneously process MD configurations at distinct times x_t and $x_{t+\tau}$. To enforce probabilistic interpretation, the last layer of the lobes utilizes softmax function. During training, given a batch of time-lagged pairs the network calculates the resulting covariance matrices (see 2.4) and differentiable VAMP-2 score (see 2.8) [4]. The network is trained by backpropagation [35].

For the purpose of simpler interpretation, a single basis set $\chi = \chi_0 = \chi_1$ is typically used. This is implemented as weight sharing between the lobes and training using the total gradient [4].

After training, we can evaluate the quality of the learned features and choose network's hyperparameters by computing the VAMP-2 validation score computed on data not used during training [4].

In study [36], VAMPnets were shown to yield superior results as compared to tICA, showing the ability to identify states that would otherwise remain undetected. This evidence underlines the significant utility and potential of VAMPnets in the realm of molecular dynamics.

2.7 CoVAMPnet

In this thesis, we employ the extension of VAMPnets proposed in [5]. This advanced version, referred to as CoVAMPnet (Comparative VAMPnet), incorporates two key innovations. The first innovation involves the alignment of states detected in different systems representing the same biomolecule in different conditions. These varied conditions may include changes in the simulation's physical parameters or, more significantly, the introduction of a potential drug candidate into the solvent. Alternatively, alignment can also be applied to different mutated variants of the same biomolecule. This alignment enables a comparative analysis of dynamic properties such as transition rates, particularly when some states can be considered analogous.

The second advancement involves the analysis of feature importance. CoVAMPnet works based on a matrix input, which is composed of distances between different residues in a specific time frame. Using this input, the method calculates average gradients [37] for each inter-residue distance. These gradients represent how changes in each of these distances can influence the classification of the molecule into a particular state. CoVAMPnet then visualizes calculated gradients in a form of feature importance matrices. Consequently, it is possible to identify specific regions of the protein, or pairs of residues, that significantly influence the classification into a particular state, including their impact, i.e. whether the proximity of the residues is a positive or negative factor. This innovation enables identification of regions critical for the dynamics of the protein [5].

2.8 Related research about APOE

2.8.1 The dynamics of APOE

The dynamics of APOE have garnered considerable attention in the scientific community, leading to multiple in-depth research studies. For instance, one research

project employed MD to analyze the interactions of APOE with the TREM2 protein, with the focus primarily on the hinge region and the C-domain [38]. In another study the focus was on lipid binding, where researchers differentiated how APOE3 and APOE4 interact with various types of lipids, again with a significant emphasis on the C-domain [14]. Interactions between APOE and amyloid beta have also been thoroughly investigated [39].

Of course, MD is not the only tool that allows us to understand dynamics more; one study utilized hydrogen/deuterium exchange and mass spectrometry rather than MD, to reveal new insights about the dynamical differences between APOE3 and APOE4 [40].

2.8.2 Exploring tramiprosate and 3SPA

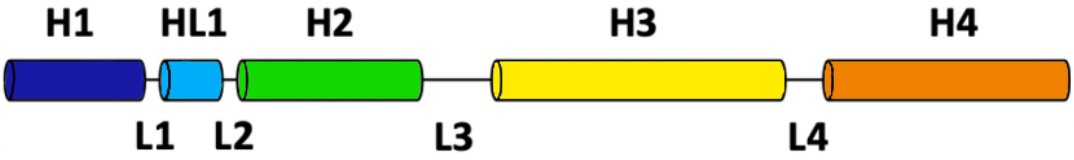
Tramiprosate, a drug initially developed for Alzeheimer's disease (AD) treatment, and its metabolite 3-sulfopropanoic acid (3SPA) have been studied for their potential therapeutic benefits in AD, particularly for patients carrying the APOE4/APOE4 genotype [41].

Our research investigates the influence of 3SPA on the dynamics of the APOE protein. It has been theorized that 3SPA may alter the behavior of APOE4, causing it to mimic the characteristics of the less harmful APOE3 isoform [19]. If this is the case, investigating a potential mechanism through which 3SPA exerts its therapeutic effects could bolster its promise as a potential treatment for AD.

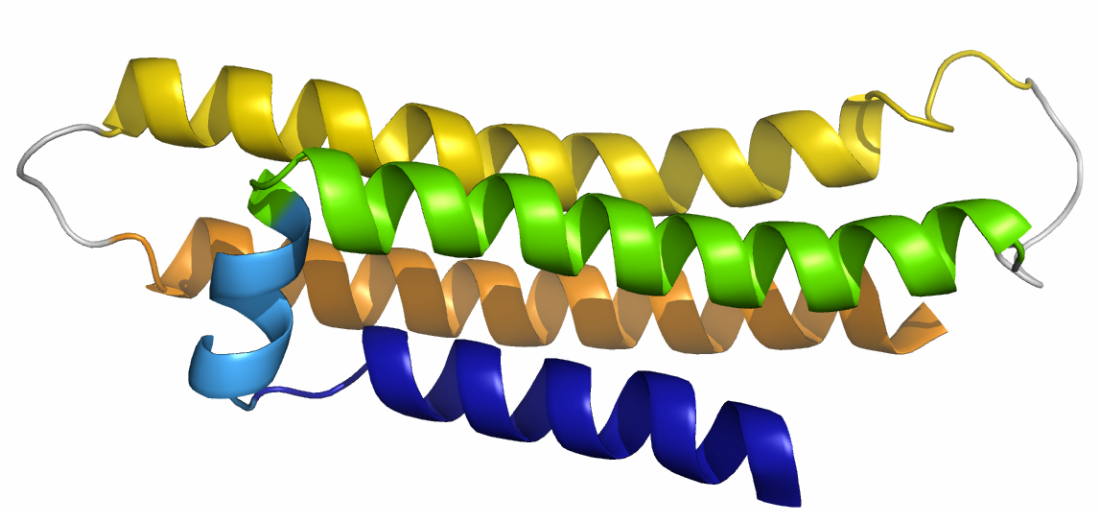
Interestingly, the statistically significant beneficial influence of tramiprosate was not observed in APOE3/APOE4 heterozygotes [42], which could suggest that the effect of 3SPA on those two types APOE differs. We investigate this as part of our work.

2.8.3 APOE oligomerization

Our work can be considered complementary to the recent publication by our colleagues from Loschmidt Laboratories at Masaryk University in Brno [19], who focused on investigating the properties of novel APOE dimers. To understand their work, it is required to understand that parts of the full-length APOE tend to degenerate in the brain. The most stable part of the protein is the so-called 4-helix bundle, which was the main focus of their research. The 4-helix bundle consists of four main helices (H1-H4) and loops connecting them, including the notable loop HL1 between H1 and H2, which often exhibits high helicality (see Fig. 2.1).



(a) 1D representation of the 4-helix bundle.



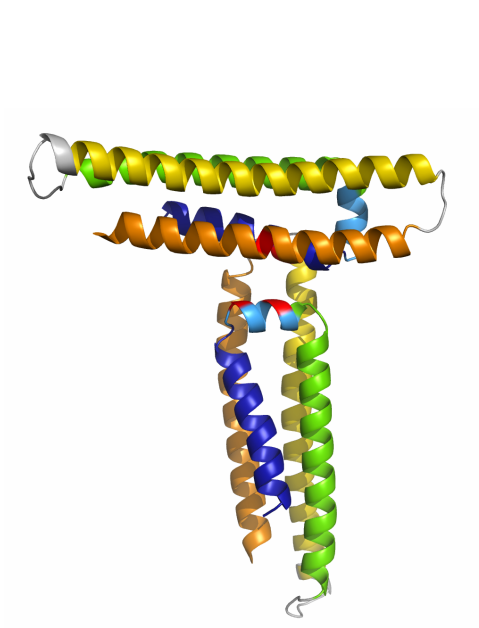
(b) 3D structure of the 4-helix bundle.

Figure 2.1: Representations of the 4-helix bundle. We will continue using this color-coding in the rest of the thesis: dark blue – H1, light blue – HL1, green – H2, yellow – H3, orange – H4 and grey – remaining regions.

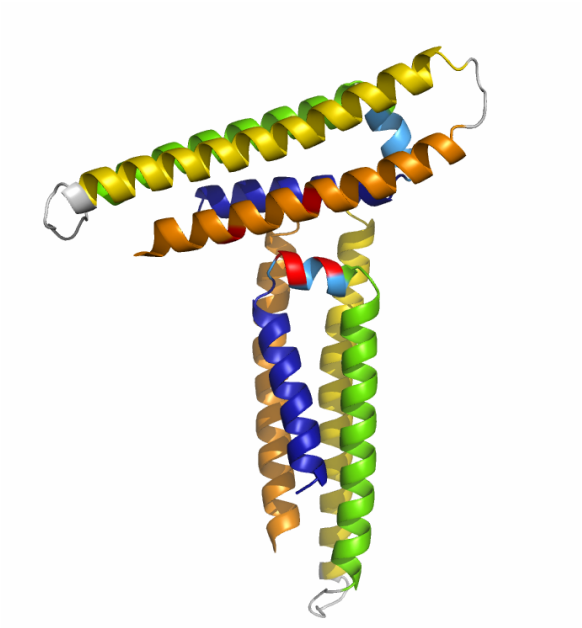
The study found out that such truncated APOE molecules create previously unknown types of dimers, differing slightly in shape between APOE3 and APOE4. APOE3 tends to create a so-called T-shaped dimer, while APOE4 creates an analogous but clearly altered V-shaped dimer (see Fig. 2.2). In these T-shaped and V-shaped dimers, we identify two distinct components: chain A, which corresponds to the "top of the T", and chain B, which corresponds to the "bottom of the T" (or their equivalents in the V-shaped dimer). The study included a thorough investigation of the most prominent properties of both dimers, with a focus of their differences. Moreover, using experimental and computational methods the authors analyzed the influence of 3SPA on the different APOE systems [19].

On top of that, thanks to yet unpublished insights from our collaborators, we also learned about the formation of another kind of dimer – the so-called parallel dimer – which, as the name suggests, involves two APOE molecules in a parallel position.

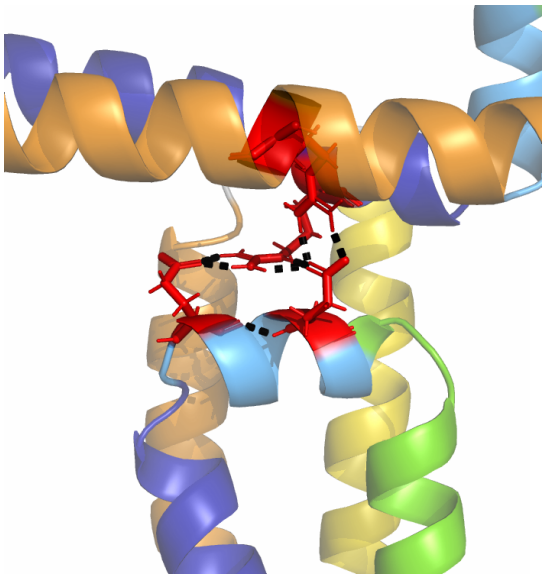
Their findings were in line with the APOE cascade hypothesis, which assumes that APOE is directly involved in the most fundamental molecular processes leading to AD.



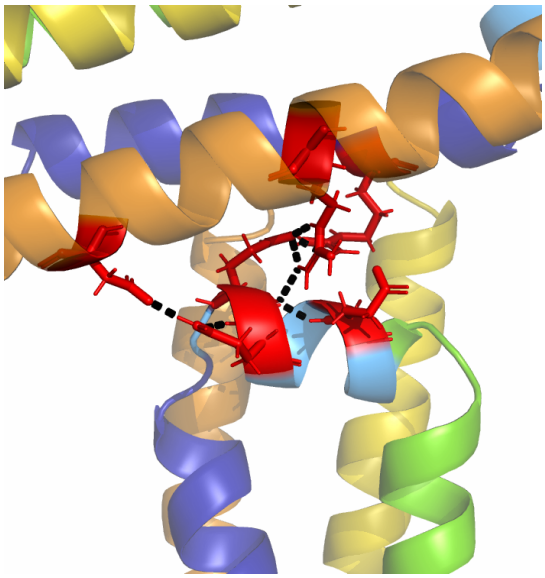
(a) T-shaped APOE3 dimer.



(b) V-shaped APOE4 dimer.



(c) Critical residues: chain A: R38 (H1), R145 (H4), chain B: E45 (HL1), E49 (HL1) .



(d) Critical residues - chain A: R38 (H1), R145 (H4), D153 (H4) chain B: E45 (HL1), E49 (HL1), Q46 (HL1).

Figure 2.2: Difference between T-shaped and V-shaped dimers. Red residues are critical for forming the self-association interface and black dashes represent polar interactions that stabilize it. Notice the additional interaction formed between D153 from top chain A and Q46 from bottom chain B in the V-shaped dimer (right), resulting in a different tilt of the chain A.

3. Data and models

3.1 Data

Our study was based on MD simulation data of four full-length monomeric APOE systems: APOE3 and APOE4, each simulated with and without the excess of the 3SPA. We will refer to these four systems as APOE3, APOE4, APOE3 + 3SPA, and APOE4 + 3SPA. The simulation data was kindly provided by our collaborators from Loschmidt Laboratories.

3.1.1 Molecular dynamics simulations

The initial three-dimensional structures of unbound APOE3 were acquired from the RCSB Protein Data Bank [43] (PDB entry 2L7B). Generated from NMR experiments, this entry consists of 20 unique models, from which the first model was selected as the starting point for our MD simulations. For APOE4, homology modeling was employed, using the corresponding FASTA sequence and the 2L7B entry of APOE3 as a template.

The simulations were implemented using the High Throughput Molecular Dynamics (HTMD) scripts [29]. Standard procedure of system protonation by simulation at physiological pH 7.4 and equilibration was performed. The starting point for our production MD runs were the systems resulting from the equilibration phase.

HTMD was utilized to conduct adaptive sampling of the conformational space of the proteins. The adaptive sampling process was guided by online-built MSMs (see Sec. 2.3) constructed using root-mean square deviation (RMSD) of the protein's alpha-carbon atoms relative to their positions in the initial structure. The RMSD is a measure used to calculate the average distance between the atoms (e.g., alpha-carbon atoms) of superimposed proteins. The initial 3D structures of both APOE3 and APOE4 contained a highly folded C-domain, so the RMSD criterion was chosen to heuristically increase the probability of observing unfolding of the whole protein.

To capture the trajectory, the coordinates of all atoms in the system were saved every 0.1 ns. For each system, at least 20 epochs were performed, each consisting of 10 individual MD runs. All individual simulations consisted of 1,000 frames, corresponding to an interval of 100 ns, with an exception of the first epoch of the APOE3 + 3SPA system, which consisted of 500 frames, i.e. 50 ns. Summing up the individual simulation times, the total simulation time was at least 20 μs per system. The details are described in Table 3.1.

System	#Epochs	#Simulations (total)	#Frames (total)
Free APOE3	20	200	200,000
Free APOE4	20	200	200,000
APOE3 + 3SPA	22	220	215,000
APOE4 + 3SPA	21	210	210,000

Table 3.1: Basic information about MD data of each system.

3.1.2 Data representation

In the training process, we used matrices of inter-residue distances for the 138 residues of the 4-helix bundle. Inter-residue distances establish an intrinsic coordinate system for molecular structures, providing a representation invariant to rotations and translations. Such a representation also fully captures the tertiary structure of the protein, at least on the residue level. We then adjusted these 138 x 138 matrices by excluding the distances up to the second closest neighbor for each residue, which correspond to the stable distances determined by peptide bonds and the zero distance to itself. In addition, since the matrices are symmetric, we considered only the elements above the main diagonal and flattened them into vectors of 9,180 values. Each of these vectors uniquely represents the structure at specific time frame. Data for each system was normalized before the training by subtracting the mean and dividing by the standard deviation.

Our analysis was specifically focused on the 4-helix bundle due to a couple of key factors. Firstly, the 4-helix bundle is the central element in the APOE dimerization hypothesis [19], and limiting our representation to it prevented C-domain movements from overshadowing the intricate dynamics of the 4-helix bundle. Secondly, we were aware of certain limitations in our model, which we address more thoroughly in Sec. 3.5.1. The models we derived do not necessarily depict the standalone behavior of the truncated 4-helix bundle, but rather represent its behavior influenced by the unfolding C-domain, which was not included in the analysis but exhibited clear correlation with the 4-helix bundle movements.

3.2 Model

In this study, we employed the VAMPnet implementation by Löhr et al. [44] with each system's model consisting of an ensemble of 20 models with identical architecture. The same model architecture successfully yielded a kinetic ensemble of amyloid beta [44] and was subsequently employed to analyze the influence of small molecules on amyloid beta [45]. The same approach was utilized in CoVAMPnet [5] to analyze MD simulations of amyloid beta, ultimately reproducing and extending the earlier results with innovations that significantly improved the interpretability of the findings. The states from the 20 models within an ensemble were aligned by a constrained k-means clustering algorithm [46], the same way as described in [5]. The employment of an ensemble of 20 models proved advantageous for obtaining error estimates through bootstrapping.

The VAMPnet model we used comprises two identical lobes, each featuring batch normalization followed by five fully connected layers of 256 neurons. An L2 regularizer is applied in each hidden layer to mitigate overfitting by adding a penalty proportional to the magnitude of the weights. The architecture employs physical constraints [47] to enforce non-negativity in the Koopman matrix, thereby preserving its probabilistic interpretation. These constraints also ensure the statistical reversibility of the learned MSM.

The implementation of the model includes a self-normalizing setup [48], which enhances the stability of the training process. Self-normalizing networks utilize specific types of activation functions, in this case the Scaled Exponential Linear Unit (SELU), that help to ensure that the outputs from neurons in the network

have a mean of zero and a standard deviation of one. This property helps the network to learn from the data more effectively and makes the training process more robust, improving the stability and the general performance of the model.

In conjunction with this, LeCun normal initialization of weights in the network is used. This initializer draws the weights from a truncated normal distribution centered in zero, with a standard deviation determined by dividing one by the number of input units in the weight matrix, and then calculating the square root of the result. It was proven to achieve good performance in conjunction with SELU [48].

3.3 Training the ensemble

We adhered to the same training strategy as employed for the amyloid beta models in [44]. We prepared 20 random data splits, each with a 90:10 training:validation ratio. For each split, we trained three models and chose the best-performing one according to the VAMP-2 score to be included in our ensemble. This step enabled us to minimize the variance of the results, which we deemed unacceptable without this measure.

Training of the χ model (see Sec. 2.4) was performed using batches of $10,000$ frame pairs and the Adam minimizer with a learning rate of 0.05, $\beta_1 = 0.99$ and $\epsilon = 0.0001$. Overfitting was addressed through early stopping, i.e., the training was stopped when the VAMP-2 [33] validation score did not increase by at least 0.001 over the previous twenty epochs.

After training, we had performed the temporal validation with implied timescales [49] and Chapman-Kolmogorov (CK) test [50]. Both these validation methods were used to ensure the quality of the obtained models. The implied timescales plot reveals the relaxation times of the system's dynamical processes, while the CK test validates the Markovianity of the model by comparing the model's predictions of longer time events with the actual observations.

The implied timescales test [49] is a common approach to determine whether a MSM accurately captures the temporal behavior of a system. Implied timescales give us an indication of the timescale on which the system transitions between states. They are defined as follows:

$$t_i(\tau) = -\frac{\tau}{\log|\lambda_i(\tau)|} \tag{3.1}$$

where τ is the lag time and $\lambda_i(\tau)$ is the i^{th} eigenvalue of the transition matrix K estimated with lag time τ. If the implied timescales are approximately constant over a long range of lag times, the MSM is said to be Markovian. Usually, the smallest lag time for which time scales become constant is chosen for the analysis to capture the dynamics of system in the finest detail while preserving the Markovianity of the model. Based on the implied timescales plots after initial trainings with different lag times, we decided to use 12.5 ns lag timefor the analysis of our systems. Fig. 3.1 shows the implied timescales plot for the APOE3 system; the implied timescales plots for the other systems are available in the Appendix (see Fig. A.3).

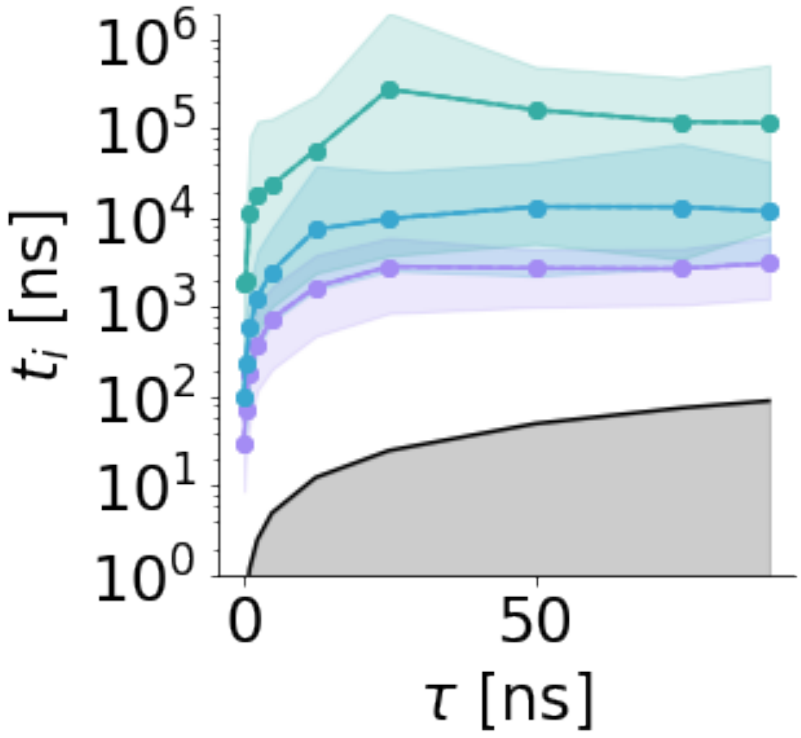


Figure 3.1: Implied timescales of the free APOE3 system. The vertical axis corresponds to the timescales computed according to equation 3.1 for lag time τ values on the horizontal axis.

The Chapman-Kolmogorov (CK) test [50] is another method to validate the Markov property of an MSM. It compares the predicted transitions of the MSM over a certain period of time with the observed transitions over the same period. For an MSM to be valid, it must satisfy the Chapman-Kolmogorov equation:

$$K(n\tau) = K^n(\tau) \tag{3.2}$$

where K^n is the transition matrix propagated for n steps (prediction), and $K(n\tau)$ is a transition matrix estimated with lag time $n\tau$ (observation). If the predicted and observed transitions agree, the MSM satisfies the Markov property.

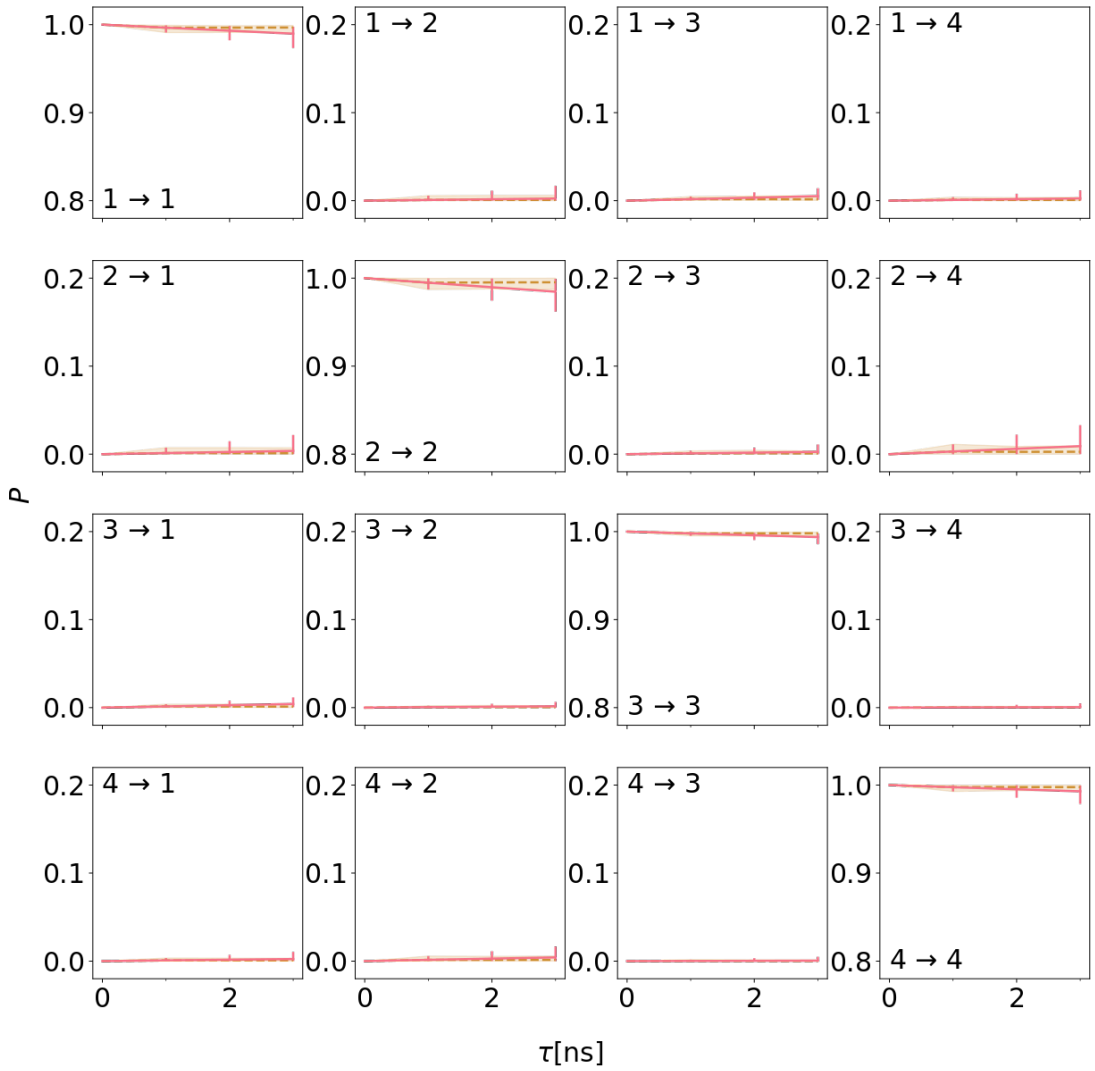


Figure 3.2: Chapman-Kolmogorov test for the free APOE3 system. The horizontal axis corresponds to the lag time τ and the vertical axis to the probability P of transition. The brown dash line marks the values estimated from observed transitions at a given lag time τ, while the red line corresponds to predictions based on propagating the Koopman operator estimated at 12.5 ns. Rows and columns correspond to states. In other words, the plot at position (i, j) represents the probability of transitioning from state i to state j.

As can be seen from the CK test plots (see Fig. 3.2, Fig. A.4, Fig. A.5, Fig. A.6), we have managed to obtain models that yield satisfactory results as implied timescales converged with time and remained relatively constant after the utilized lag time and the CK-test plots show good agreement between long-term model predictions and actual observations. These promising outcomes highlight the robustness and validity of our models.

We experimented with different numbers of states for each system. However, due to the resource-intensive nature of the training process, combined with limited resources and satisfactory preliminary results, we did not carry out a comprehensive hyperparameter search. The final number of states for each system was determined by the quality of the implied timescales and the interpretability of the results.

3.4 Statistics and visualizations

In our study, we used the first two time-lagged Independent Components (tIC) determined by tICA to provide a two-dimensional representation of each system. The systems were transformed using the corresponding tICA transformation, with the density determined via Gaussian kernel smoothing. Interestingly, this technique revealed a clear separation of several high-density areas observable in the two-dimensional space.

This effect was particularly noteworthy when contrasted with the results from the smaller, fully unstructured amyloid beta. Being an unstructured peptide with rapid conformational changes, the tICA landscape of amyloid beta was notably shallow. In contrast, our APOE systems exhibited more pronounced transitions between main conformational states, suggesting a better separation of detected metastable conformations.

We examined varying numbers of states for each system, utilizing the quality of implied timescales, compatibility of the tICA projection of detected clusters with the density plots of tICA, and interpretability of results as key determinants for choosing the optimal number of states for each system. We concluded that 4 states were most suitable for APOE3 and APOE4 + 3SPA, while 3 states were best for APOE3 + 3SPA and APOE4.

We overlaid MSM graphs onto the tICA density plots (see e.g. Fig. 4.1). The circles represent the states and their size corresponds to the equilibrium probability estimated from the Koopman matrix. The arrows represent the transitions between states and their thickness is proportional to the probability of transition at the used lag time of 12.5 ns. The transition probabilities are stated in the arrow labels, with percentage (%) used as a unit of measure.

We provide a visualization depicting the temporal evolution of each system (see e.g. Fig. 4.2). For this, we sorted the simulations according to the time elapsed since the initial frame of the first epoch, treating the sampled frame as immediate past for the seeded simulations, effectively concatenating simulations from different epochs. Upon visual examination, it was usually straightforward to discern the time correlation of specific states within the total simulation. For example certain states surfaced only around the first epoch's initial frames, while others emerged only after some time had elapsed. These plots were used to choose the order of the states, enhancing their interpretability. Smaller numbers are assigned to states that dominate earlier in the overall timeline. This allows us to quickly figure out the initial and final states of our simulations and makes it easier to compare states between different systems.

Since most changes in the systems occurred on the level of secondary structure, plots showing the average level of helicality for each state as computed by DSSP algorithm [9] were probably the most important tool in our analysis (see e.g. Fig. 4.3).

To get more insight about higher-level geometric structures of the obtained states, we plotted the average residue contact maps for each state, utilizing a cutoff threshold of 0.8 nm (see e.g. Fig. 4.3). The residue contact map is a matrix-based representation that outlines the spatial proximity of residues in a protein. Each element of the matrix corresponds to a pair of residues in the protein, and is set to one if the distance between the two residues is smaller than

a predefined cutoff threshold, and zero otherwise .

To gain some intuition about the overall changes in the shape of APOE between states, we visualized several representative frames for each Markov state (see e.g. Fig. 4.5). We applied a probabilistic approach to avoid selecting frames from a single trajectory, which would frequently happen if frames with the highest probability for a given state were selected. To this end, we calculated a probability distribution over all frames of a given system, weighted according to the probability of assignment to a given state. After drawing representative frames according to this distribution, we eliminated frames with less than a 50% probability of belonging to that state. We expect this method to select frames adequately representing the state, not solely concentrating on its center but also potentially reducing the occurrence of highly transitional frames.

In the visualizations, we opted for varying numbers of representative frames for different states and systems. This approach was taken to prevent outliers from masking the overall characteristic property of a state, which could happen in particular when visualizing the loss of helical structure between two states. Since the helix is visualized as a broad band, one outlier could effectively obscure all the narrow coils representing the loss of structure in a given state. Moreover, due to locality of most changes, we only presented short ranges of residues to see them in more detail. All 3D visualizations in this thesis were rendered in PyMol [51].

For the gradient analysis by CoVAMPnet [5], we used 1,000 frames. We selected frames based on a uniform distribution over all frames. However, the quality of the feature importance matrices varied across the systems. For example, while gradient matrices for APOE3 allowed for a quick and unambiguous identification of important features, the results for APOE4 were consistently uninformative.

Due to the relatively low reliability of the Koopman operators for our data, which we discuss below in Sec. 3.5.1, we decided to support our estimates of the equilibrium distributions of states with the population counts (see Fig. A.1). The populations are calculated as the number of frames classified to each state by hard assignment divided by the total number of frames in a given system. Koopman operators, although central for our analysis, are difficult to interpret visually, and were thus included only in the Appendix (see Fig. A.2).

3.5 Limitations of our study

3.5.1 Full-length APOE modeling attempts

Our original goal was to train a neural network utilizing data from the full spectrum of APOE protein residues. To be more precise, this involved inter-residue distances between all pairs of 299 residues of the full-length APOE protein. However, the network encountered difficulties in generating a coherent Koopman operator. A recurrent issue was the emergence of eigenvalues exceeding the maximum theoretical limit of 1. This is likely due to the limited dataset and the inherent randomness of the neural network. Despite various attempts to train the network using different lag times, the results remained largely unchanged.

To overcome these difficulties, we modified our approach by excluding the C-domain and the N-terminus from the analysis and eliminating motions from highly flexible protein sections that could potentially skew the Koopman operator. Even though less than half of the residues were excluded, 4-helix bundle is significantly more stable. This effectively reduced the conformational landscape of the protein. This deliberate refinement improved the network's learning ability and minimized the risk of overfitting, potentially unmasking obscured properties. On top of that, 4-helix bundle is of crucial importance for the leading hypotheses about APOE dimerization, including hypothesis of our collaborators from Loschmidt Laboratories [19], therefore it was a natural choice for more detailed analysis.

However, it is crucial to underscore the inherent limitations of our study. The total amount of data necessary to fully capture the conformational landscape of the APOE, whether we are dealing with the whole protein or just the 4-helix bundle, remains unclear. As far as we know, no other studies have applied VAMPnets to a system of comparable size and complexity with such a short total simulation time. The conformational shifts of APOE are undeniably rare, so while the observed clusters can offer intriguing insights into the metastable states of APOE, we must take the estimates of dynamical properties with caution.

We observed that our data may be too far from equilibrium to accurately reflect it. This conclusion is based on several observations, including the scarcity of transitions between states and the significant correlation between state transitions and the unwinding of the flexible C-domain. Another factor casting doubt on the reliability of the dynamic element of our conclusions is the implied timescales test. While our models passed this test by showing converging timescales, the implied time of the slowest process exceeded the total simulation length for each system, suggesting that the theoretical number of transitions that were observed in the data is extremely small. Situations leading to those kind of observations were discussed in detail in [52].

However, despite these challenges, we believe our study yielded meaningful clustering. Our models passed the implied timescales and CK tests, and agreed with the tICA landscape. Even if we cannot truly interpret our results as a model of the APOE dynamics at equilibrium, it nevertheless gives us insights into the overall flexibility of the 4-helix bundle. Furthermore, the most significant validation came from the replication of our collaborators' results, who used traditional methods for estimating Markov state models or directly observed the evolution of features of interest.

3.5.2 CoVAMPnet limitations

On top of that, we were also limited by the computational expenses when computing the feature importance matrices implemented in CoVAMPnet. Free APOE3 presented a highly interpretable, sensible matrix. While APOE3 + 3SPA and APOE4 + 3SPA offered decent results, we believe there's room for substantial improvement. Conversely, the free APOE4's matrix was of subpar quality and lacked interpretability (see Fig. 4.11).

This might be attributed to the sparse data, causing the network to fixate on less crucial features. Further, we based these matrices on merely 1,000 time frames, a fraction of the frames used in the original CoVAMPnet paper, which

utilized 10,000 frames. Moreover, our analyzed 4-helix bundle system is approximately three times larger than the amyloid beta system in the original study. We found that increasing the number of frames for gradient computation could substantially improve the results. However, due to the time-intensive nature of the process and other pressing research priorities, we deemed these matrices acceptable for this master's thesis.

3.5.3 Emergent symmetry of CoVAMPnet gradient analysis

Interestingly, CoVAMPnet often identifies nearly perfectly opposite feature importance matrices for two detected states. This symmetry was even visible in some matrices from the original paper (see Fig. 3. in [5]). We noted similar symmetry in our matrices, despite some numerical differences. This intriguing pattern could likely be mathematically explained. It could be advantageous for the network to identify states in such a high-contrast manner, akin to contrastive learning [53], considering it processes two frames concurrently and utilizes loss which involves measuring similarity between them in the form of covariance matrices.

3.5.4 Limitations of adaptive sampling and VAMPnet based analysis

We want to highlight a specific challenge when applying VAMPnets to simulations generated with adaptive sampling. This issue arose while trying to estimate the model for the full-length protein, where we noticed that many transitions occur between two epochs of simulations for the analyzed system. This is possible because the frames selected by adaptive sampling as "seeds" to initialize the simulations in the next epoch are not technically the first frames of those simulations – the frames immediately following are. The goal of adaptive sampling is to increase the likelihood of interesting slow transitions occurring from the chosen seed frames, which results in higher probability of the seed frames to represent such an uncertain conformation. We observed that even our states computed for the 4-helix bundle of each system strongly correspond to a particular level of unfolding of the protein – usually one state corresponds to a completely folded, and one to a completely unfolded C-domain, which position, as we mentioned before, dominates the value of the RMSD used for constructing MSMs during adaptive sampling. We conclude that due to this strong correlation it simply became more likely for the network to "agree" with the sampling criterion, resulting in transitions happening between epochs.

Despite VAMPnets working with soft assignments, which one might expect to introduce some uncertainty in transitioning between different states, we still deem transitional pairs of frames valuable. This is particularly relevant for systems with slow dynamics or rare transitions, as these pairs of frames provide VAMPnets with most of the information about these transitions. This is especially the case if the transition is sharp, i.e., if we observe a large change in the calculated probability distribution for two frames being one lag time apart - which is more likely to

happen if there are distinctly separated energy islands in the conformational landscape.

Furthermore, we would like to point out that this phenomenon leads to some issues of practical nature. To better understand the system one may try to calculate the matrix that represents the transitions according to the hard assignment, which in this case might lead to misleadingly low values, due to transitions "hidden" between epochs. That actually happened during our initial attempts with the full-length APOE. Moreover, identifying transitional trajectories becomes less straightforward, as they become more likely to span across multiple epochs. Such trajectories can provide especially useful visualizations, including videos, which help us understand system's movements.

3.5.5 Limitations of modeling slow dynamics with short simulations

Another more fundamental issue with estimating dynamical properties based on many short simulations is the way a lag time determines the number of frame pairs that can be used for MSM estimation. When dealing with a single long simulation, the effective number of frames is simply

$$total\_simulation\_length - lag\_time$$

In the case of multiple short simulations, such as in the case of adaptive sampling, however, it becomes

$$total\_simulation\_length - total\_number\_of\_simulations * lag\_time$$

. In our case we had around 200 simulations per systems, so we "lost" 200 times more pairs than we would in the case of single trajectory. This problem grows proportionally to the utilized lag time, which in turn grows if the dynamics of the systems of interest has slower dynamics. This scenario, which often presents itself in larger and more complex systems, consistently garners significant research interest and continues to be a frontier area of study [54]. In our case the utilized lag time turned out to be acceptably small, but we would like to point out this inherent limitation of adaptive sampling techniques combined with time-lagged methods for systems with slow dynamics.

3.5.6 Concatenting trajectories

To address these limitation, we propose an approach that involves artificially concatenating the trajectories. In this method, instead of treating each trajectory from the adaptive sampling regime as separate, we consider sampled frames as the historical context for seeded simulations. This gives every simulation in subsequent epochs a well defined past, enabling us to generate time-lagged pairs spanning between epochs.

This way, the effective number of pairs would become approximately

$$total\_simulation\_length - number\_of\_simulations\_per\_epoch * lag\_time$$

On top of that, we could even utilize lag times surpassing lengths of individual trajectories. We believe this idea holds a potential to increase applicability

of time-lagged methods for systems distinguishing slow dynamics. We plan to implement and test it in the nearest future.

4. Analysis of the free APOE dynamics

In this chapter, we thoroughly analyze and discuss our primary findings concerning the free APOE3 and free APOE4 systems. The focus will primarily be on the potential correlations with the hypotheses surrounding APOE dimerization proposed by our collaborators [19].

Presenting all results for such a multifaceted analysis is by itself quite a daunting task. Understanding the full scope of the observed changes directly from particular plots alone would be difficult, therefore, we have adopted a consistent layout across all systems to simplify navigation. Each system's introduction consists of the most essential numerical findings.

First, we present MSM graphs superimposed on tICA density plots (Fig. 4.1, Fig. 4.8, Fig. 5.1, Fig. 5.8). A visualization of each system's temporal evolution based on the plots presenting simulations sorted by the time elapsed from the first frames of the first epoch is provided. Frames are colored according to hard assignment of a MSM state (Fig. 4.2, Fig. 4.9, Fig. 5.2, Fig. 5.9). Following are average secondary structure plots and contact maps of each state. They are accompanied by equilibrium populations calculated based on the corresponding Koopman operators (Fig. 4.3, Fig. 4.10, Fig. 5.3, Fig. 5.10). Next, CoVAMPnet's feature importance matrices are presented (Fig. 4.4, Fig. 4.11, Fig. 5.4, Fig. 5.11). We provided cues in forms of arrows in the colors of corresponding subdomains and circles highlighting the regions of high importance where it was possible.

For every system, we included a table outlining the most notable differences among the identified states. As previously mentioned, while our analysis focuses on the 4-helix bundle, we observed a clear correlation between the position of the C-domain and the detected states. This is why our tables provide information not just about the key segments of the 4-helix bundle, but also about the overall position of the C-domain associated with each state (Tab. 4.1, Tab. 5.1, Tab. 4.2, Tab. 5.2).

We recommend to first read a detailed description of the observed changes in the following subsections, including their possible links to APOE dimerization, and only then coming back to those high-level plots. We believe such approach is the quickest way to obtain a better insight into APOE's dynamical behavior.

4.1 Examination of free APOE3

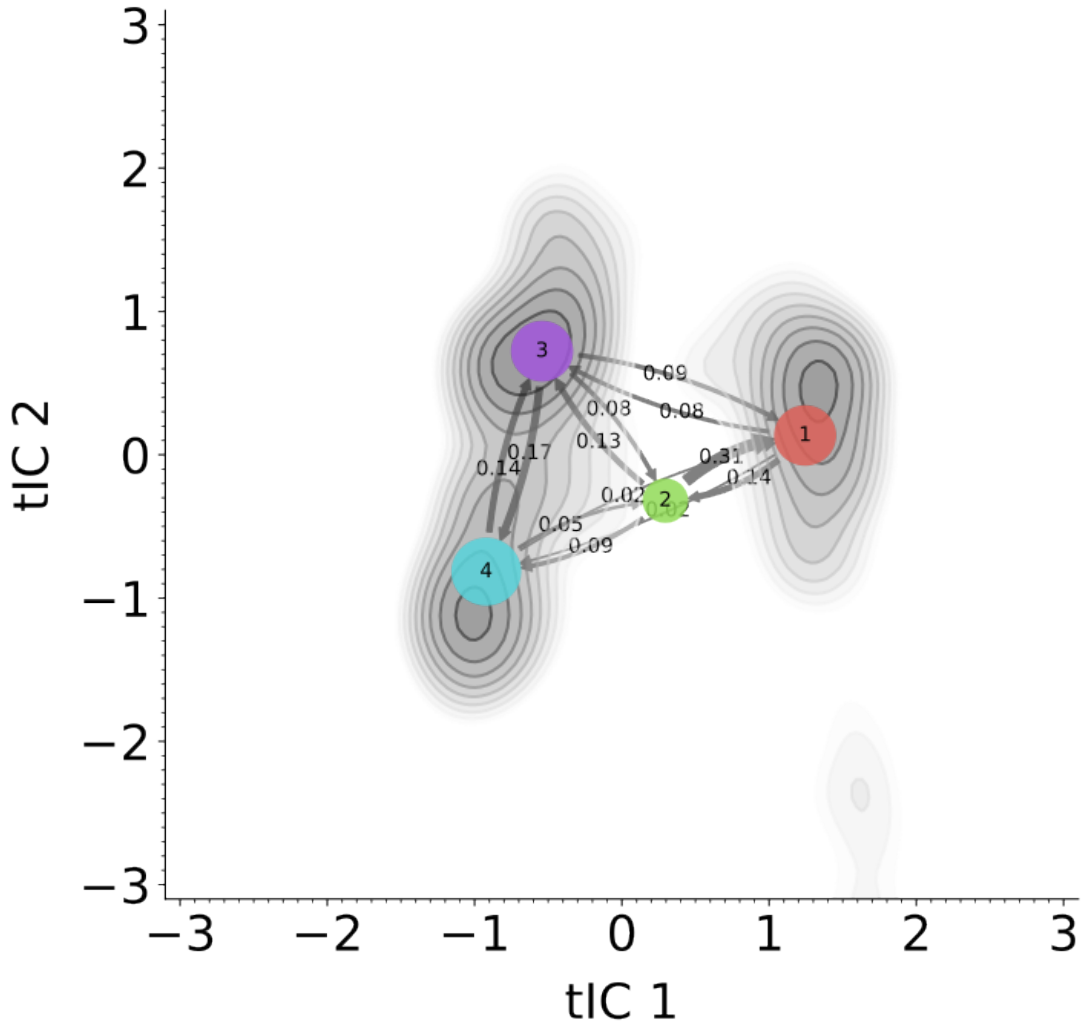


Figure 4.1: tICA density plot and the MSM graph representation of the free APOE3 system. Darker shades of grey correspond to higher density. Circles correspond to states and are accordingly numbered. Size of the circle is proportional to the probability of a corresponding state at equilibrium. Arrows represent probability of transitions between states, with 1% used as a unit of measure.

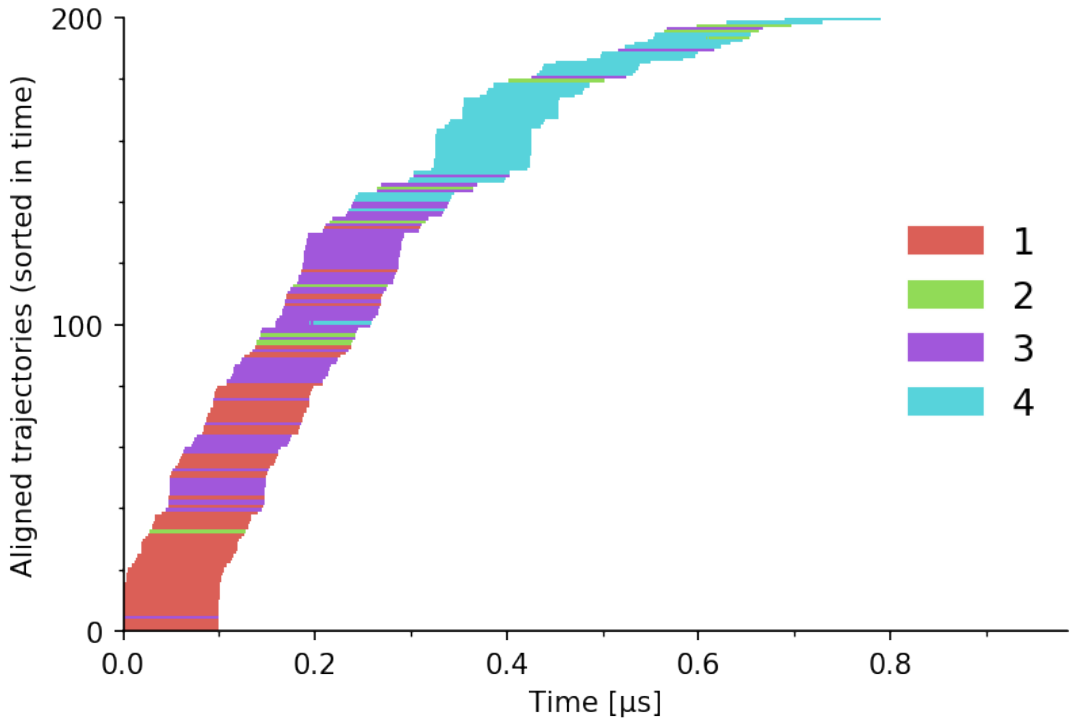


Figure 4.2: Temporal evolution of the free APOE3 system. Simulations were sorted according the the time elapsed from the initial frames of first epochs (x-axis), treating epochs from which frames were sampled as a past of newly initialized simulations. Frames were colored according the their hard classification to a state, whose numbers are visible in the legend on the right.

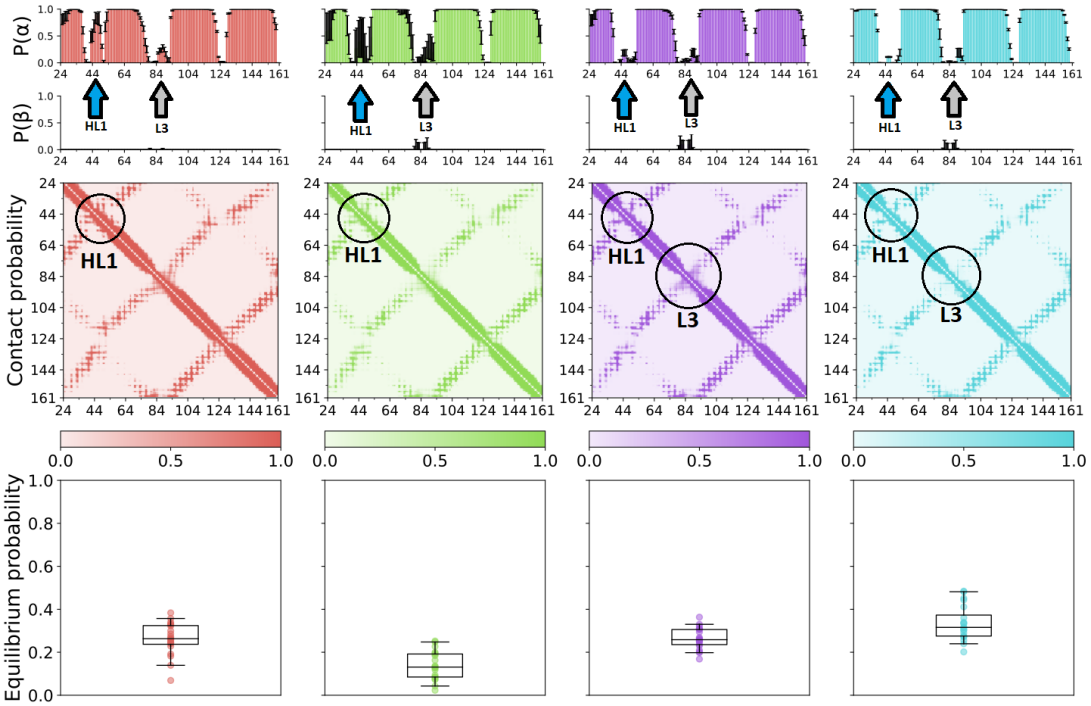


Figure 4.3: Average secondary structure, contact maps and equilibrium probability of states obtained for the free APOE3 system.

The use of feature importance matrices, generated in the CoVAMPnet pipeline, proved to be essential for this system. They facilitated the swift identification of two regions in the protein—HL1 and L3—that exert significant influence on its dynamics. HL1 and L3 display the largest differences among the different states.

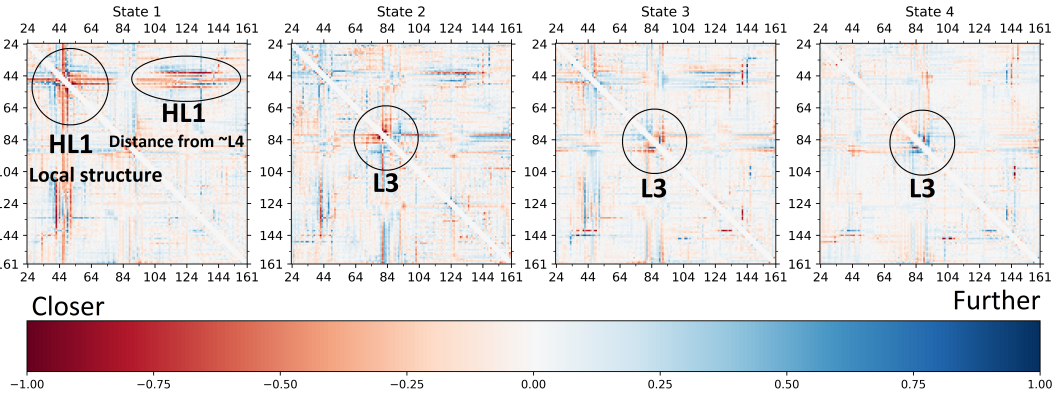


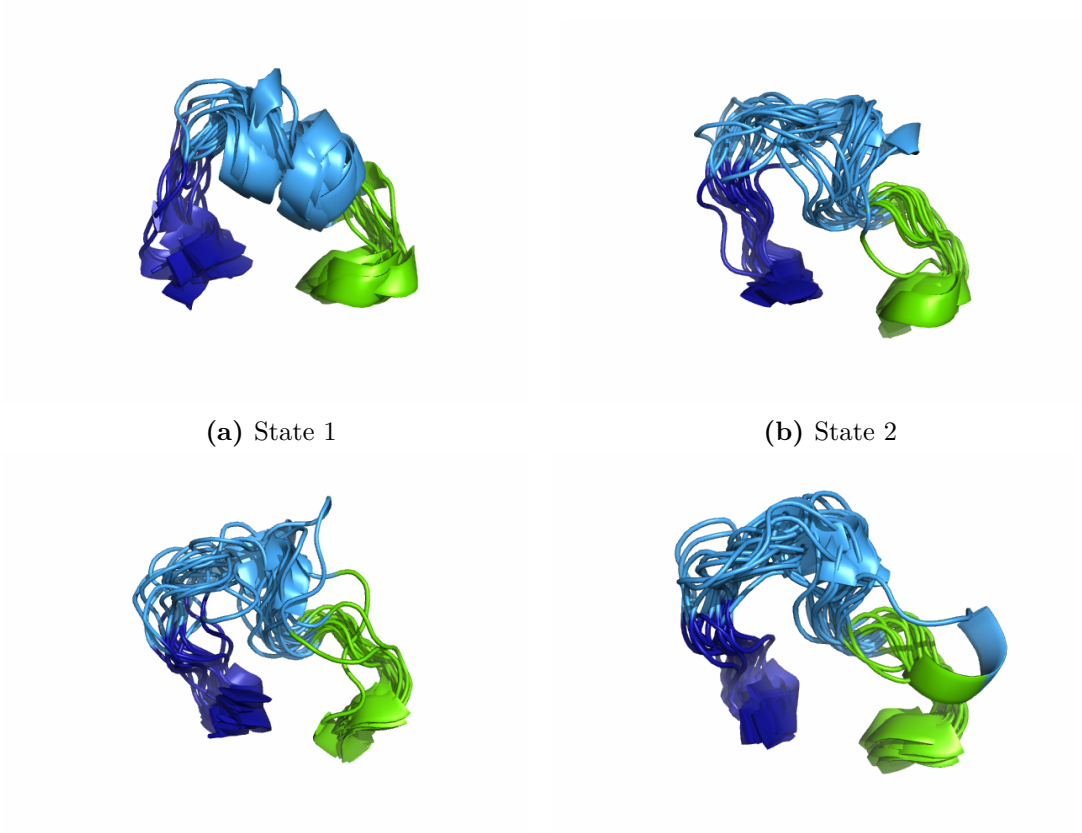
Figure 4.4: Feature importance matrices obtained by the 4-state CoVAMPnet model for the free APOE3 system. State 1 exhibits high importance of the distances of residues belonging to HL1, while other states seem to be more focused on the L3 area.

Subdomain	State 1	State 2	State 3	State 4
HL1	Structured	Intermediate	Unstructured	Unstructured
L3	Structured	Structured	Structured	Unstructured
C-domain	Folded	Mixed	Intermediate	Unfolded

Table 4.1: Structure of the most important subdomains and position of the C-domain observed in different states of the free APOE3 system.

4.1.1 Free APOE3 dynamics is dominated by the structural changes in HL1

The structural integrity of the HL1 area is the most distinctive feature among the identified states. The protein experiences a noticeable loss of structure in this region as the C-domain unwinds (see Fig. 4.5).



(a) State 1 **(b)** State 2

(c) State 3 **(d)** State 4

Figure 4.5: Structure of HL1 in different states of free APOE3 system. Residues 39-57, 20 representative frames. Dark blue – H1, light blue – HL1, green – H2. Notice that HL1 maintains a high degree of structure in state 1, while it exhibits a total absence of structure in other states.

State 1 is characterized by a highly structured HL1, whereas state 2 depicts an intermediate state marked by a high variance of helicality and position in the tICA landscape. HL1 in states 3 and 4, by contrast, can be regarded as unstructured (see Fig. 4.3). The absence of structure observed in states 3 and 4 distinguishes the APOE3 system from the other systems under investigation. While a degree of structure loss is apparent in the other systems, none display as substantial a loss as observed in these two states.

Importance of the HL1 region was clear when looking at the computed feature importance matrices (see Fig. 4.4). We can see that state 1 is characterized by the high intensity of the entries corresponding to the proximity of the residues in the HL1 to the rest of the protein. The region of high intensity in the left circle roughly corresponds to the secondary structure of HL1. The other region in the elongated ellipsis on the right indicates relatively small distance between HL1 and residues of the H3, L4 and the beginning of H4. It is worth noting that L4 is right next to HL1 in the overall 3D structure of the 4-helix bundle.

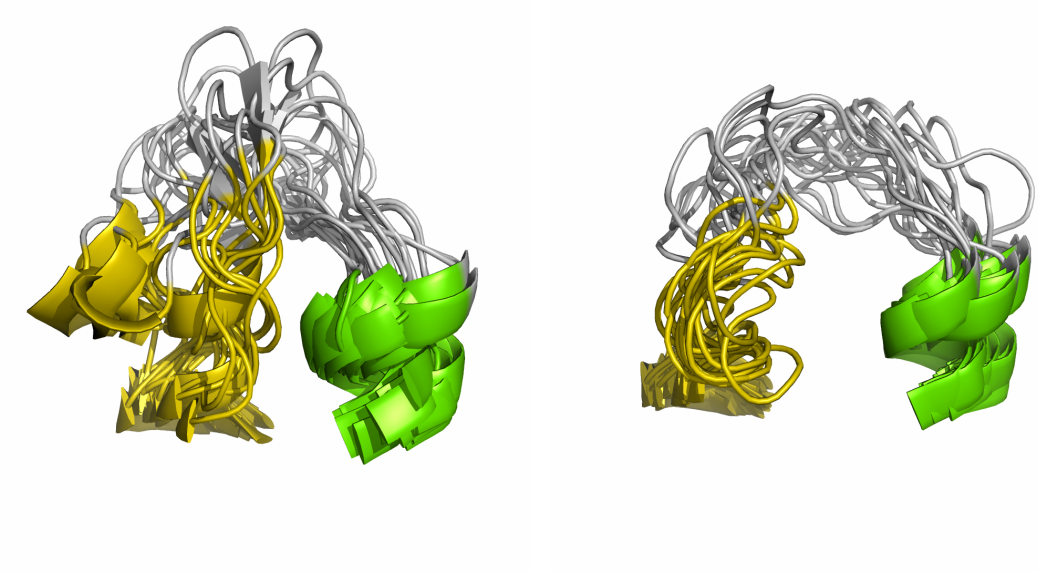
Considering temporal ordering of states (see Fig. 4.2), we can suspect that as the whole protein was unfolding, HL1 was losing its helical structure and simultanously drifting away from the L4 and the residues surrounding it in H3 and H4. According to the visual inspection this drifting is mostly a consequence of the unwinding of HL1, and most likely is of low significance.

These findings bear particular significance in the context of the T-shaped dimer hypothesis. Given that HL1 contains two critical residues of the B chain involved in self-association interface formation (see Fig. 2.2), it would be plausible to suggest that the high flexibility of HL1 in this system plays a role in forming a highly stable interface. Due to the findings related to this protein simulated in the presence of 3SPA, however, such a straightforward interpretation seems unlikely, as discussed in the next chapter.

It is also important to note that our collaborators did not report analogous changes in the simulations of the dimers of interest [19]. Therefore, additional investigation is necessary to reconcile these observations and comprehend the potential impact of this structural change on the dimerization process.

4.1.2 Role of L3 flexibility in free APOE3 dynamics

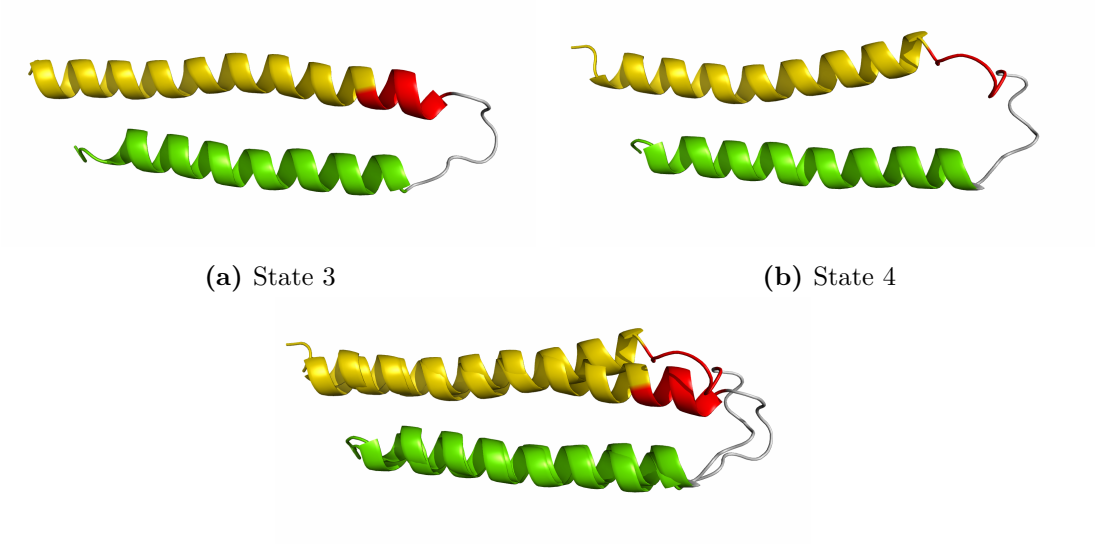
Furthermore, we conducted a detailed investigation of the differences between states 3 and 4 in the APOE3 system. These states showed distinct characteristics in the second time-lagged independent component. The feature importance matrix highlighted structural changes in the L3 region, for which the states appeared to be complementary (see Fig. 4.4). High intensity blue region in the center of the feature importance matrix for state 4 suggested that higher distance between residues belonging to the L3 area increases the probability of classifying the state as state 4. This was subsequently confirmed through average contact maps generated for these clusters. Notably, in state 4, the slight blur indicating relative proximity of residues observed in state 3 disappeared. From the secondary structure plots and visual inspection it was also evident that some residues at the beginning of H3 were drifting apart and losing their helical structure (see Fig. 4.6).



(a) State 3 **(b)** State 4

Figure 4.6: Structure around L3 in state 3 and state 4 of free APOE3 system. Residues 76-97, 20 representative frames. Yellow – H3, grey – L3, green – H2. Notice the slight unwinding of H3 in state 4.

These changes were also reported in [19]. Among the 8 PDB structures of APOE3 dimers the authors analyzed, two of them exhibited conformations more similar to the V-shape dimer characteristic of APOE4 and displayed similar changes in this region. These changes do not involve residues indicated in the creation of dimer interfaces. However, it is plausible that these alterations have important implications through long-range effects on protein behavior. Quoting the authors of the paper : "The unwinding of the beginning of helix H3 (residues 89–91) probably led to a weakening of the interaction between helices H2 and H3 and, consequently, to conformational changes similar to those caused by the C112R substitution." [19]. One of such changes involved the orientation of W34 residue. This atomic scale phenomenon was not visible in the residue level representations processed by VAMPnets, but analyzing the possible correlation of such changes with the states obtained by our clustering could be an interesting next step in our analyses. Another thing that will require further discussion is the coexistence of this change with the unstructured HL1. It is possible that both of those changes in conjunction – unstructured HL1 and weaker interaction between H2 and H3 – strengthen the interactions leading to a formation of a dimer.



(a) State 3

(b) State 4

(c) Both states aligned

Figure 4.7: Unwinding of the beginning of H3 in free APOE3. Residues 89-94 are depicted in red.

Some of the frames associated with state 4 show substantial unwinding at the start of H3 (see Fig. 4.7). We observe a structural loss for residues 89-94, which is a slightly more pronounced than the reported unwinding of residues 89-91 [19], but essentially represents the same transformation.

Intriguingly, the close alignment between the VAMPnet-based clustering and tICA lends itself to a more intuitive interpretation of the computed independent components. The primary differentiation occurs in the first time-lagged independent component, seemingly corresponding to the structural degradation in HL1. This is inferred by examining the 3D structure, the feature importance, and the state location within the projection. The second component appears to segregate states 3 and 4 based on the structure of L3, specifically the unwinding of residues ranging from 89 to 94 (see Fig. 4.1).

The identification of subtle changes in protein conformations highlights the sensitivity of VAMPnet [4] in detecting even the small local conformational changes. By utilizing gradient extraction methods, we were able to effectively identify the most relevant conformational changes without introducing bias from preconceived expectations. This capability of CoVAMPnet [5] to identify crucial conformational changes and regions of high flexibility reaffirms its potential, particularly for researchers without specialized expertise in biochemistry.

4.2 Free APOE4

We used 3 states in our Markov state model to describe the dynamics of the free APOE4 system - they turned out to be in perfect agreement with the high-density islands of tICA landscape (see Fig. 4.8), and using higher number of states led to less informative results with higher variance. Therefore we believe that for this system tICA was a sufficient method of extracting information about the dynamics of the protein, at least based on the provided data.

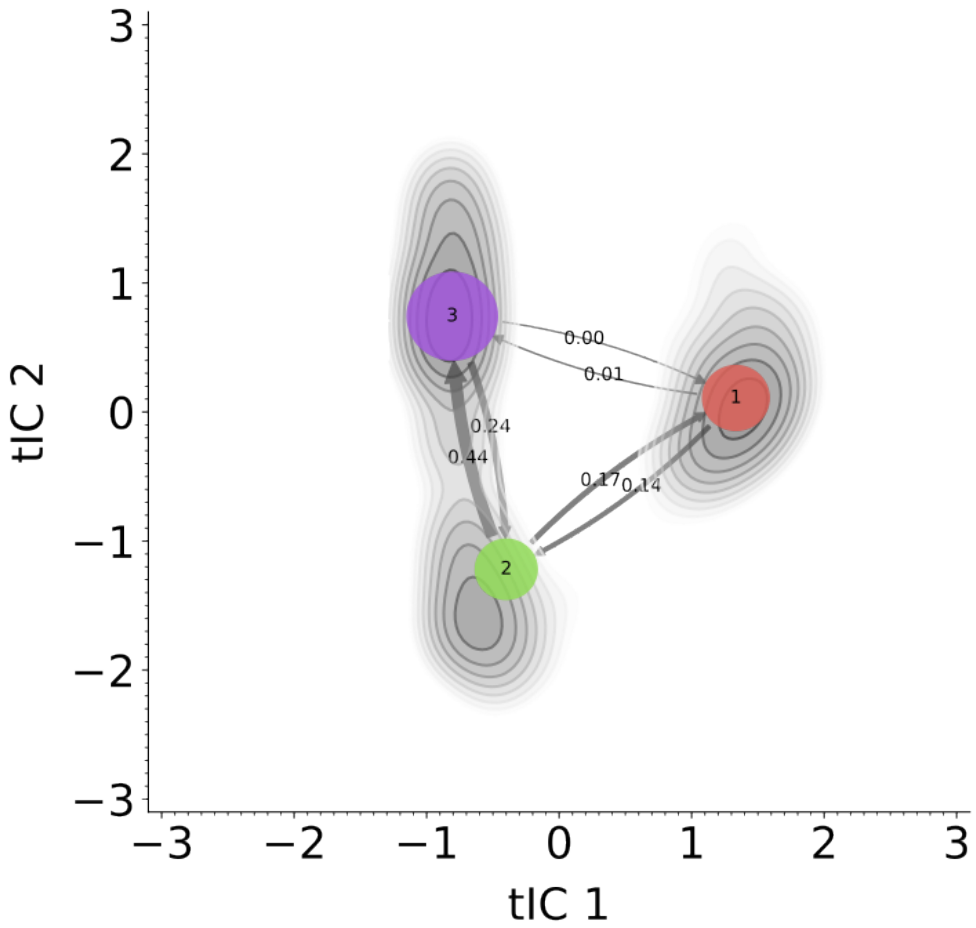


Figure 4.8: tICA density plot and the MSM graph representation of the free APOE4 system. Darker shades of grey correspond to higher density. Circles correspond to states and are accordingly numbered. Size of the circle is proportional to the probability of a corresponding state at equilibrium. Arrows represent probability of transitions between states, with 1% used as a unit of measure.

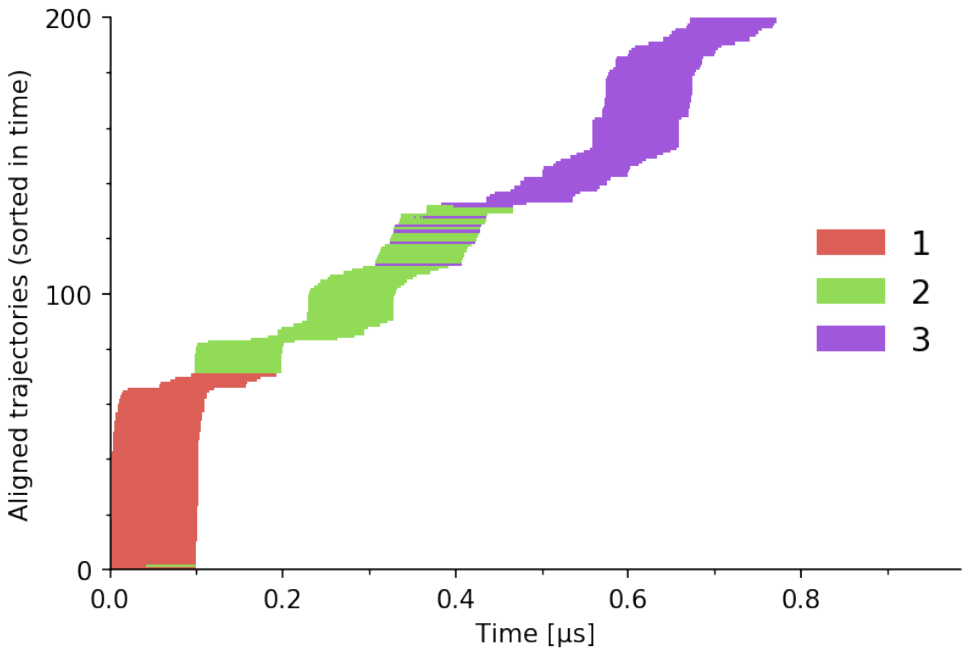


Figure 4.9: Temporal evolution of the free APOE4 system. Simulations were sorted according the the time elapsed from the initial frames of first epochs (x-axis), treating epochs from which frames were sampled as a past of newly initialized simulations. Frames were colored according the their hard classification to a state, whose numbers are visible in the legend on the right.

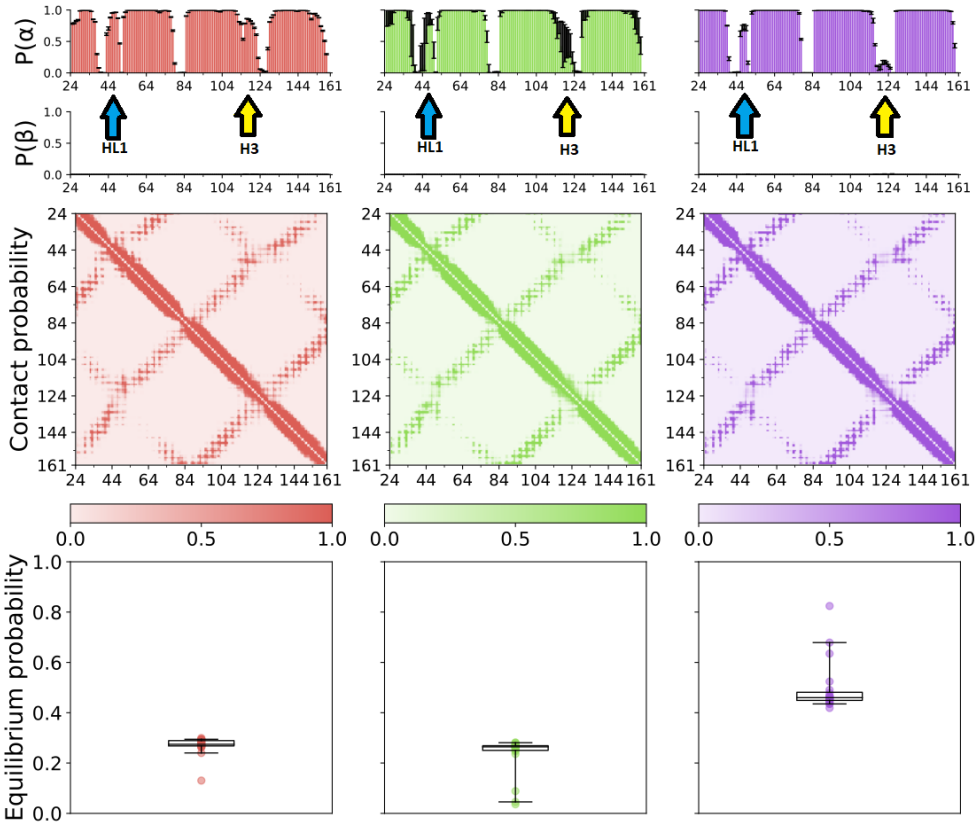


Figure 4.10: Average secondary structure, contact maps and equilibrium probability of states obtained for the free APOE4 system.

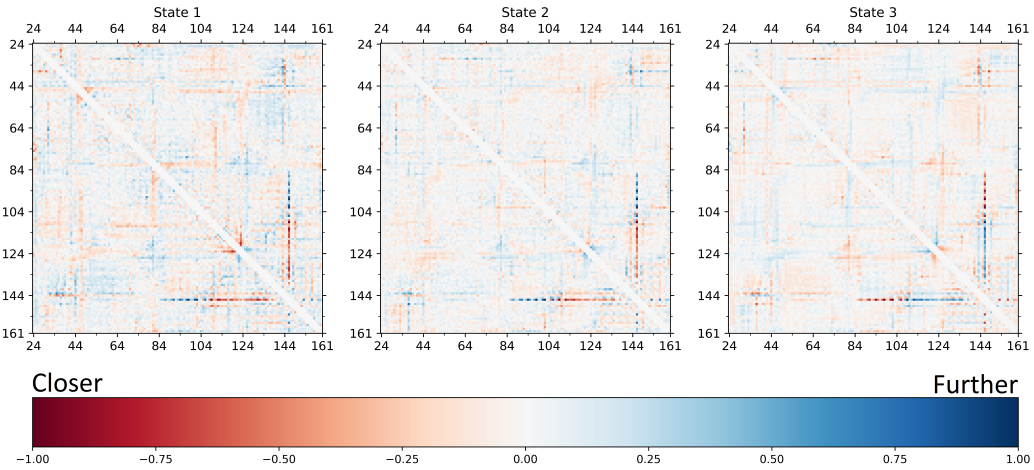


Figure 4.11: Feature importance matrices calculated for free APOE4 system. Compared to other system they turned out to be completely uninformative.

Subdomain	State 1	State 2	State 3
HL1	Structured	Structured	Less structured
H3	Bent, structured	Intermediate	Untructured
C-domain	Folded	Intermediate	Unfolded

Table 4.2: Structure of the most important subdomains observed in different states of the free APOE4 system.

4.2.1 Free APOE4 dynamics is dominated by the unwinding of H3

Most visually striking difference between different states in this system focus on the end of H3, specifically residues 120-124 (see Fig. 4.12). Over the course of the simulations (see Fig. 4.9), we observed a significant unwinding of this region, which was already described by our collaborators. As they indicated, those changes involve residue 123, which is involved in creating interactions with W39/T42 residues and forming the V-shaped dimer [19].

(a) State 1

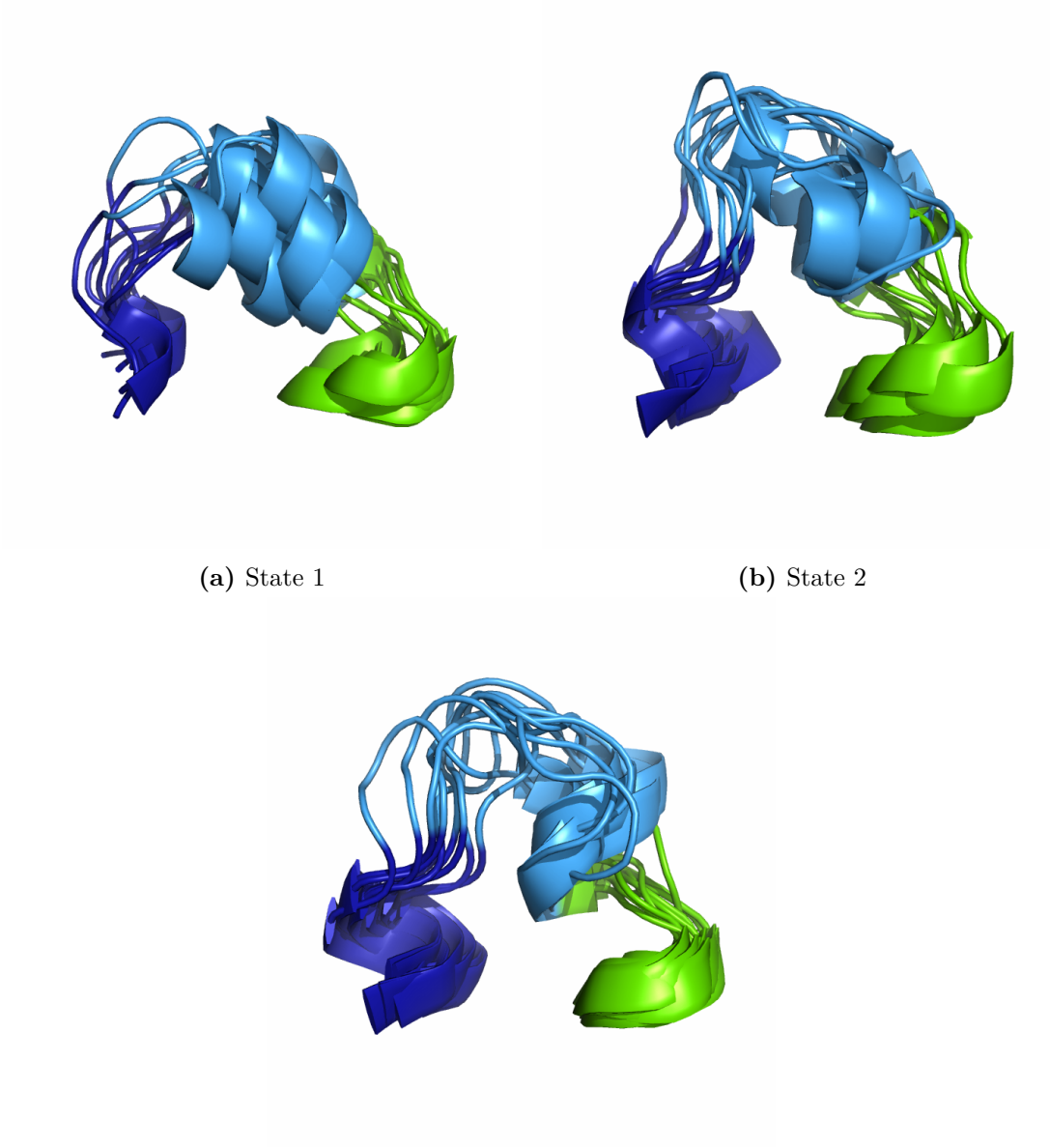
(b) State 2

(c) State 3

Figure 4.12: Structure of H3 in different states of free APOE4 system. Residues 116-139, 5 representative frames. Yellow – H3, grey – L4, orange – H4. Notice the progressive loss of helical structure between residues 117-122 when the system was evolving towards state 3.

4.2.2 States represent changes in the HL1 domain

Additionally, similar to APOE3, the structure of the HL1 domain in APOE4 exhibited changes across different states. However, unlike the APOE3 system, none of the states in APOE4 exhibited a fully unstructured loop in the HL1 region. This suggests that the HL1 domain in APOE4 retains a higher level of structural stability. As in all systems, higher level of structure was observed in state 1, so in the state with on average most folded C-domain. As the simulations run and transitioned towards state 2 and 3, we observed unwinding of the residues 44-47, and the residues 45-51 remained relatively stable (see Fig. 4.13).



(a) State 1

(b) State 2

(c) State 3

Figure 4.13: Structure of HL1 in different states of free APOE4 system. Residues 39-57, 10 representative frames. Dark blue – H1, light blue – HL1, green – H2. We can observe a gradual loss of HL1 structure, but less severe than in the free APOE3 system (see Fig. 4.5).

Like in the case of APOE3, this alteration could potentially impact the final shape of the dimer, leading to a shape more characteristic of the APOE4 V-shaped dimer. It is possible that this is linked to better exposure of the residue Q46, which forms a critical interaction with D153 in chain A (see Fig. 2.2). This interaction causes a tilt of chain A, skewing it and resulting in the formation of a V-shape.

5. Analysis of the effect of a small molecule drug candidate on the APOE protein dynamics

In this chapter we focus on describing the changes we observed in simulations performed with the 3SPA introduced to the solvent. We put the emphasis on the differences visible in the dynamical behavior of the protein under the influence of this drug candidate, as compared to the free APOE systems. We also elaborate on their possible connection to the processes leading to APOE aggregation.

5.1 APOE3 with 3spa

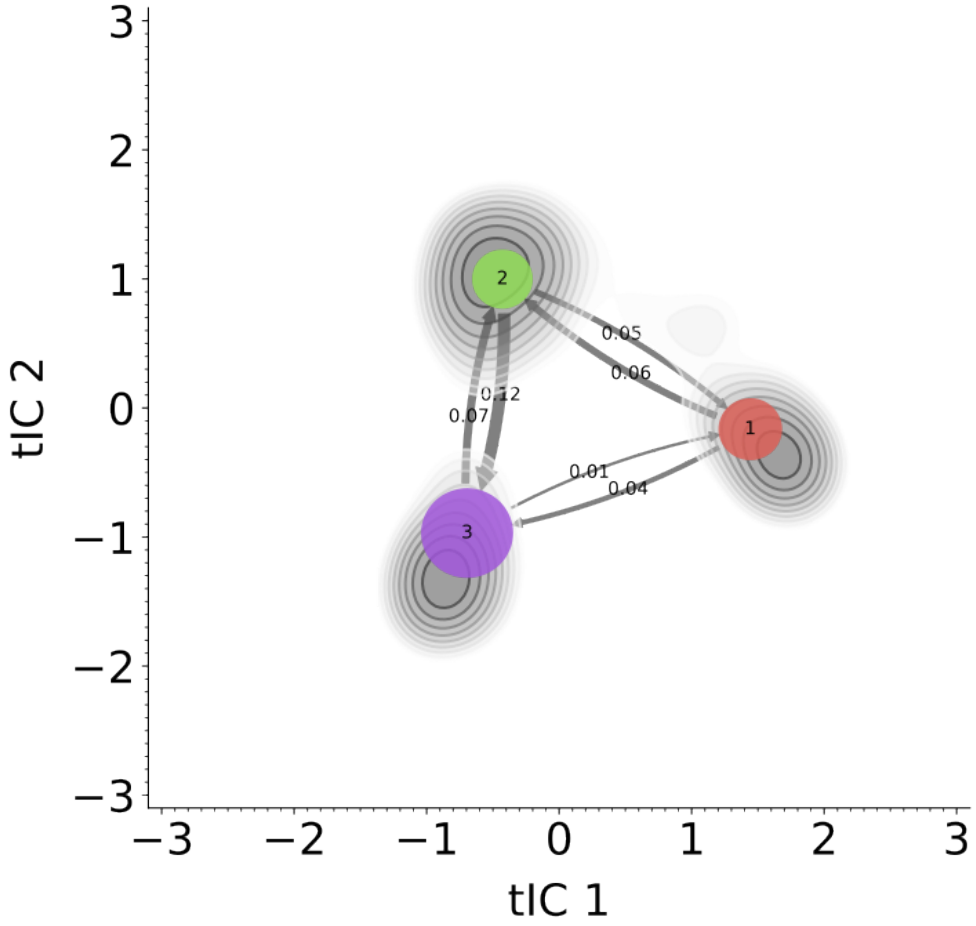


Figure 5.1: tICA density plot and the MSM graph representation of the APOE3 + 3SPA system. Darker shades of grey correspond to higher density. Circles correspond to states and are accordingly numbered. Size of the circle is proportional to the probability of a corresponding state at equilibrium. Arrows represent probability of transitions between states, with 1% used as a unit of measure.

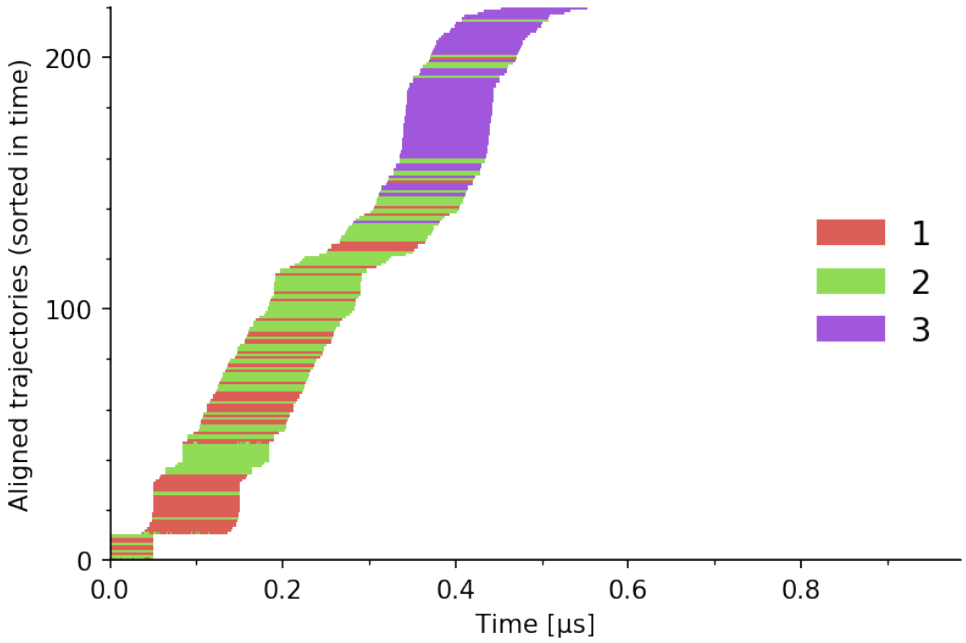


Figure 5.2: Temporal evolution of the APOE3 + 3SPA system. Simulations were sorted according the the time elapsed from the initial frames of first epochs (x-axis), treating epochs from which frames were sampled as a past of newly initialized simulations. Frames were colored according the their hard classification to a state, whose numbers are visible in the legend on the right.

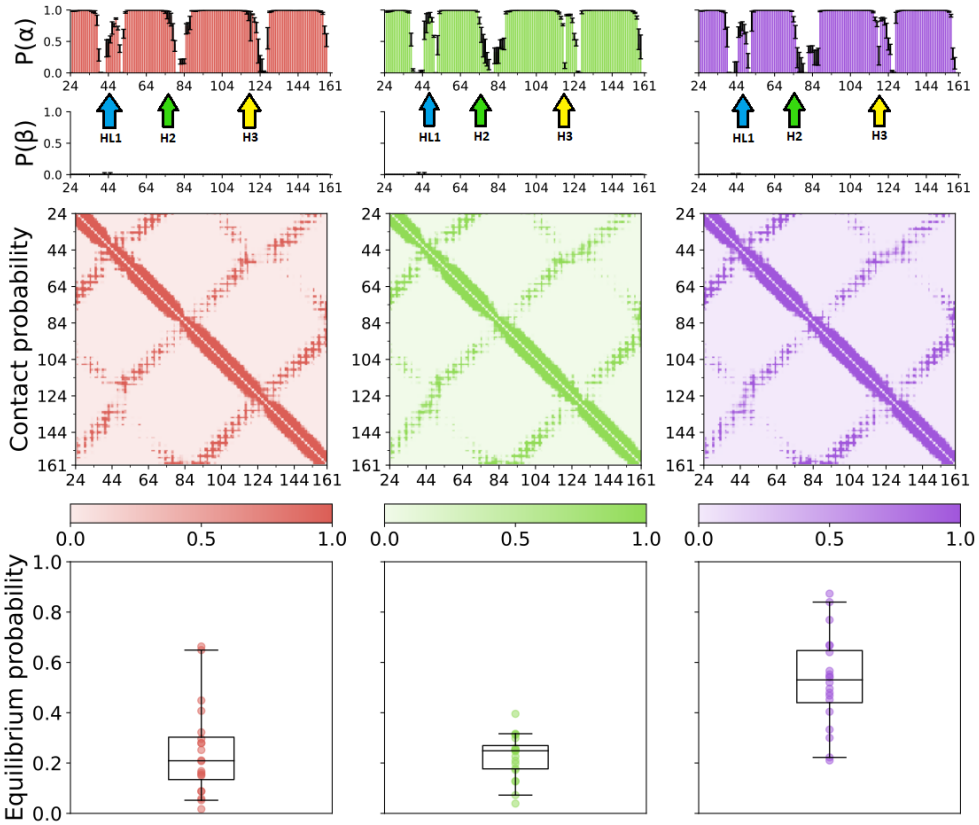


Figure 5.3: Average secondary structure, contact maps and equilibrium probability of states obtained for the APOE3 + 3SPA system.

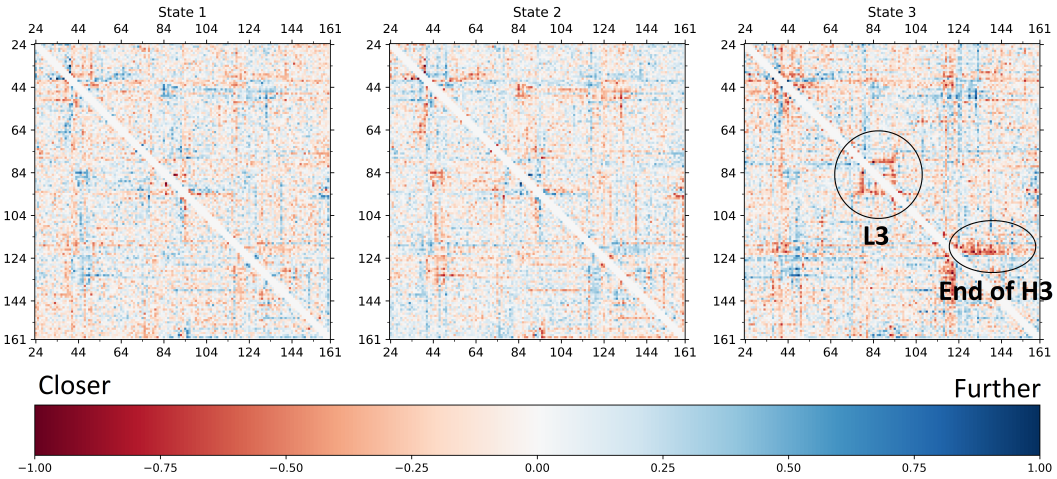


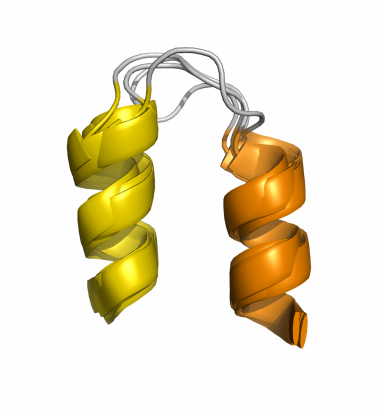
Figure 5.4: Feature importance matrices obtained by the 3-state CoVAMPnet model for the APOE3 + 3SPA system. State 1 and state 2 look very noisy, but matrix for state 3 has drawn our attention the highly relevant features - residues at the end of H2 and H3.

Subdomain	State 1	State 2	State 3
HL1	Structured	Less structured	Less structured
L3	Structured	Less structured	Less structured
H3	Straight	Bent	Bent
C-domain	Folded	Intermediate	Unfolded

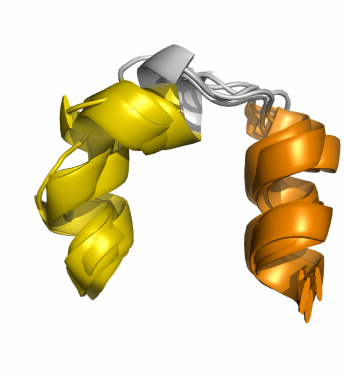
Table 5.1: Structure of the most important subdomains observed in different states of the APOE3 + 3SPA system.

5.1.1 APOE3 with 3SPA exhibits unique bending of H3

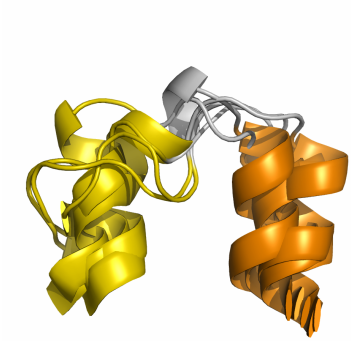
The most notable change with respect to the free APOE3 we observed is the bending of H3 around residue 118. While state 1 is characterized by a straight H3, states 2 and 3 have a bent H3, which is visible in the one-point loss of structure around residue 118. State 3 also encompasses frames with reduced helicality in this area, specifically around residues 118-125 (see Fig. 5.5).



(a) State 1



(b) State 2



(c) State 3

Figure 5.5: Structure of H3 and L4 in different states of free APOE3+3SPA system. Residues 116-139, 5 representative frames. Yellow – H3, grey – L4, orange – H4. Notice the bending of H3 in state 2 and unwinding of H3 in state 3.

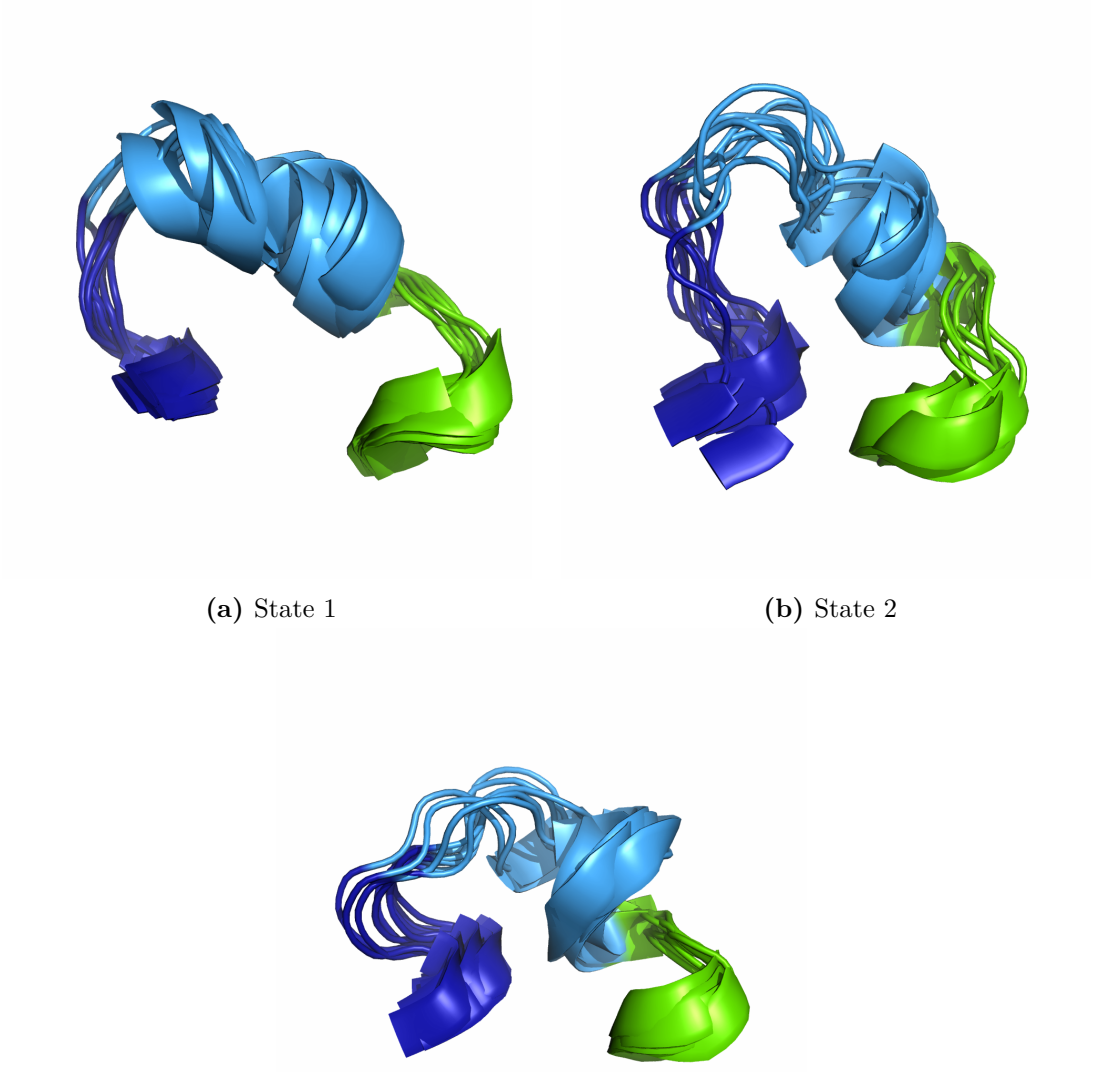
We currently do not see a straightforward interpretation of this change in relation to the T-shaped dimers, as it does not directly involve any residues contributing to the self-association interfaces. However, as proven even in the discussed publication [19], even small conformational changes can lead to long range, domino-like effects of profound consequences for protein interactions. To quote the authors: "Interestingly, while all three ApoE isoforms share the same self-association interface, the pathological ApoE4 isoform differs from the ApoE2 and ApoE3 isoforms by the angle between the two interacting NTDs. We demonstrate that this angular difference is a consequence of a "domino-like effect" of the C112R substitution, starting with the loss of the R61-E109 interaction, leading to destabilization of the H3 helix and re-orientation of Q123. " [19].

Another important observation is the fact that this bending and unwinding resembles conformational changes we observed for the free APOE4 system (see Fig. 4.12). We will elaborate on this in the next subsection.

5.1.2 Loss of HL1 structure looks reduced in the presence of 3SPA

As in previous systems describing free APOE3 and free APOE4, we could observe different levels of structural integrity in HL1 subdomain between different states.

HL1 was the most structured in the state 1. With time, the beginning of the HL1 became less structured, but contrary to the free APOE3 simulations it never became fully unstructured (see Fig. 5.6). Interestingly, this change is virtually identical with behavior of free APOE4 described previously (see Fig. 4.13). As stated in the previous chapter, this is one of the regions of very high importance for the T-shaped dimer, and according to data obtained by our collaborators from Loschmidt Laboratories, 3SPA increased the APOE3 propensity to aggregate as a T-shaped dimer [19].



(a) State 1

(b) State 2

(c) State 3

Figure 5.6: Structure of HL1 in different states of APOE3 + 3SPA system. Residues 39-57, 10 representative frames. Dark blue – H1, light blue – HL1, green – H2. Higher structural integrity of state 1 is clearly apparent.

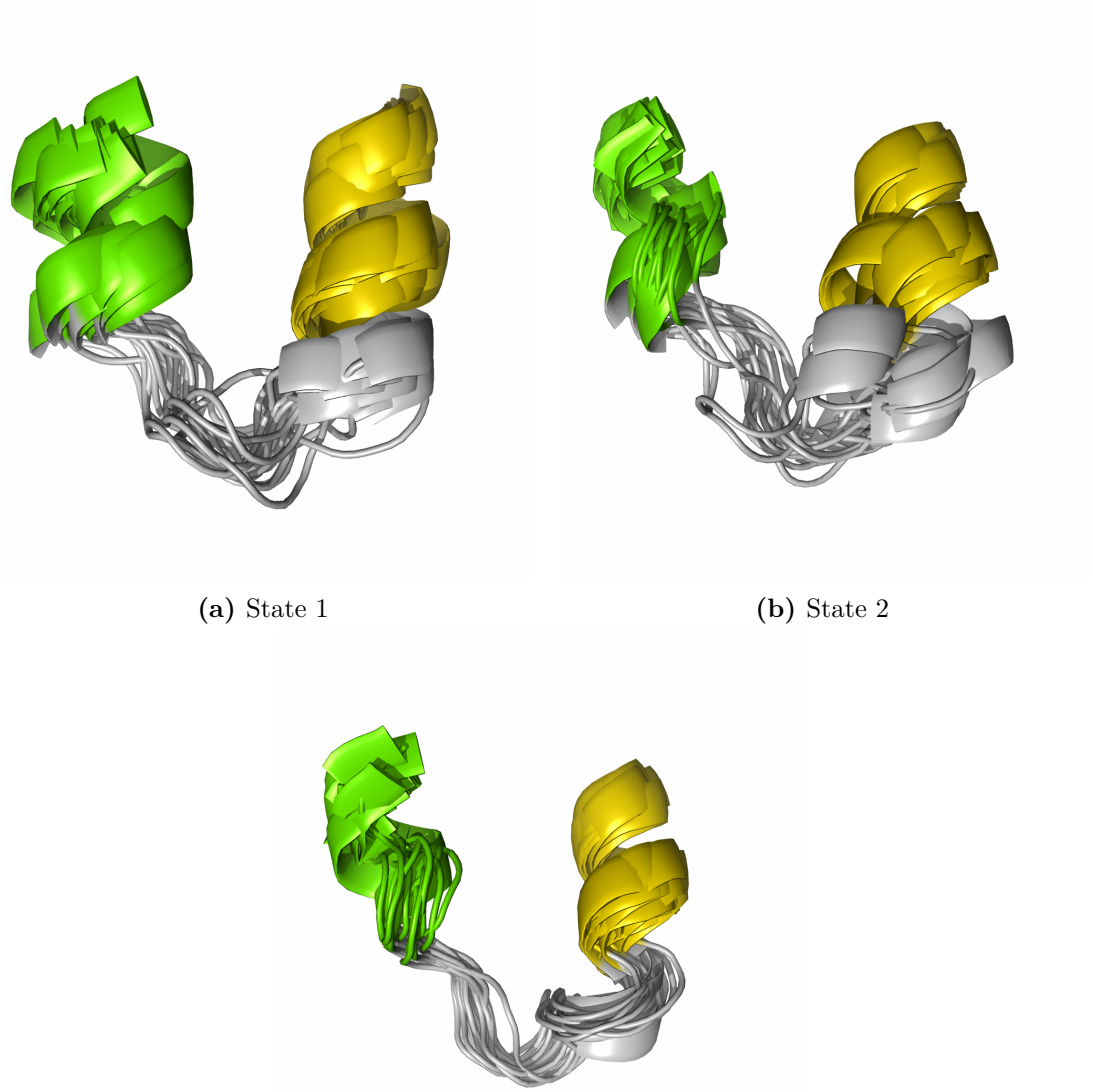
It seems like the idea that increased flexibility of HL1 in the free APOE3 system supports creation of T-shaped dimers leads to a contradiction. 3SPA which

made the HL1 in APOE3 more similar to that of APOE4 actually increased aggregation into the desired T-shaped dimer. There are several possible explanations for that. Possibilities to consider include the changes in HL1 being irrelevant for dimer behavior, these properties being significant but inadequately captured due to limited data, or our residue-level analysis obscuring atomic-level properties such us specific orientations of residues.

The most interesting explanation, however, is related to the fact that 3SPA was found to only have a positive medical effect on patients with the APOE4/APOE4 genotype [41], [42]. As described in the next section, 3SPA really has effect on the conformational dynamics of APOE4 that could be considered positive. Considering the previously described bending and unwinding of the end of H3 and the influence on the structure of HL1 in APOE3, it may be concluded that 3SPA induces more APOE4-like conformations on APOE3. If that truly is the case, 3SPA may actually cause a negative effect on APOE3, and the lack of effect on APOE3/APOE4 patients might be the effect of both positive effect on APOE4 and negative effect on APOE3 cancelling each other out.

5.1.3 3SPA introduces unwinding of the H2 helix near the L3 loop in APOE3

Another significant change we observed was the unwinding of helices around the L3 loop. This is the same region of importance as we observed for free APOE3. Contrary to the simulation without the drug candidate, however, we also observed some progressive loss of structure on the side of H2, specifically residues 75-79. APOE3 with 3SPA also suffered from some loss structure at the beginning of H3, but at a lesser degree (see Fig. 5.7).



(a) State 1

(b) State 2

(c) State 3

Figure 5.7: Structure of H2 and L3 in different states of APOE3 + 3SPA system. Residues 76-96, 20 representative frames. Yellow – H3, grey – L3, green – H2. Notice the progressive loss of structure of the green H2 and the grey L3.

We can consider these changes to be somewhat analogous to the L3 changes observed in free APOE3 (see Fig. 4.6). Increased unwinding of the end of H2 helix seems to be a significant effect of 3SPA. As mentioned in the previous chapter, unwinding of the H3 resulted in conformations leading to more APOE4-like, V-shaped dimers. According to research [19], 3SPA increased the APOE3 propensity to form T-shaped dimers. It could be therefore possible that this minor stabilization of the H3 increases the probability to form more APOE3-like, T-shaped dimer. If the V-shaped dimerization is responsible for the neuropathological effects of APOE4, then this finding suggests a possible explanation for the therapeutic effects of 3SPA observed in clinical experiments. However, confirming this hypothesis demands more rigorous investigation.

5.2 APOE4 with 3SPA

Some of our most interesting findings concern the simulation of APOE4 in the presence of 3SPA.

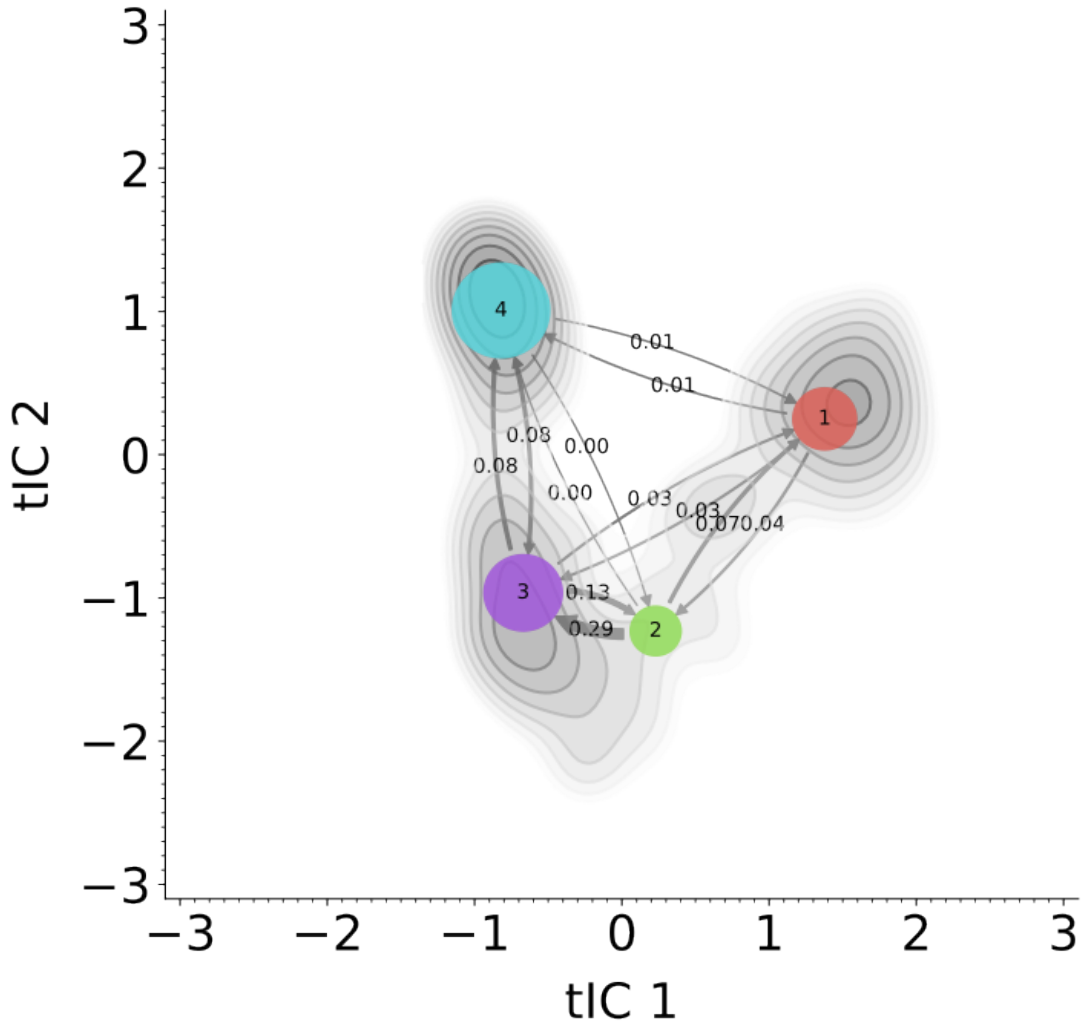


Figure 5.8: tICA density plot and the MSM graph representation of the APOE4 + 3SPA system. Darker shades of grey correspond to higher density. Circles correspond to states and are accordingly numbered. Size of the circle is proportional to the probability of a corresponding state at equilibrium. Arrows represent probability of transitions between states, with 1% used as a unit of measure.

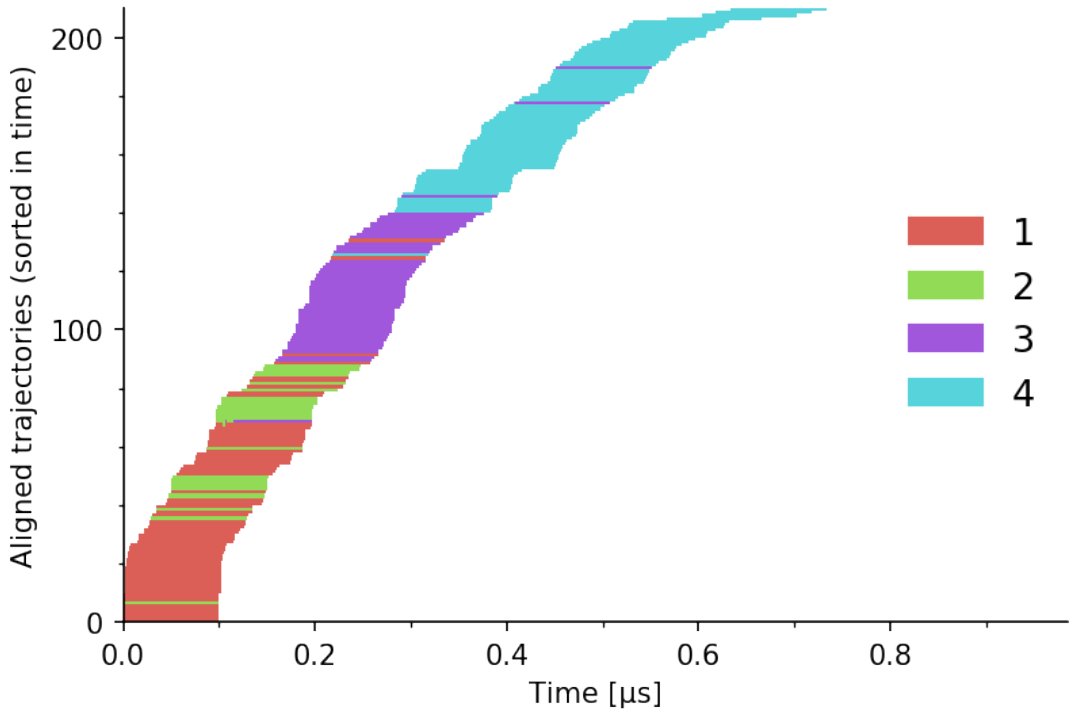


Figure 5.9: Temporal evolution of the APOE4 + 3SPA system. Simulations were sorted according the the time elapsed from the initial frames of first epochs (x-axis), treating epochs from which frames were sampled as a past of newly initialized simulations. Frames were colored according the their hard classification to a state, whose numbers are visible in the legend on the right.

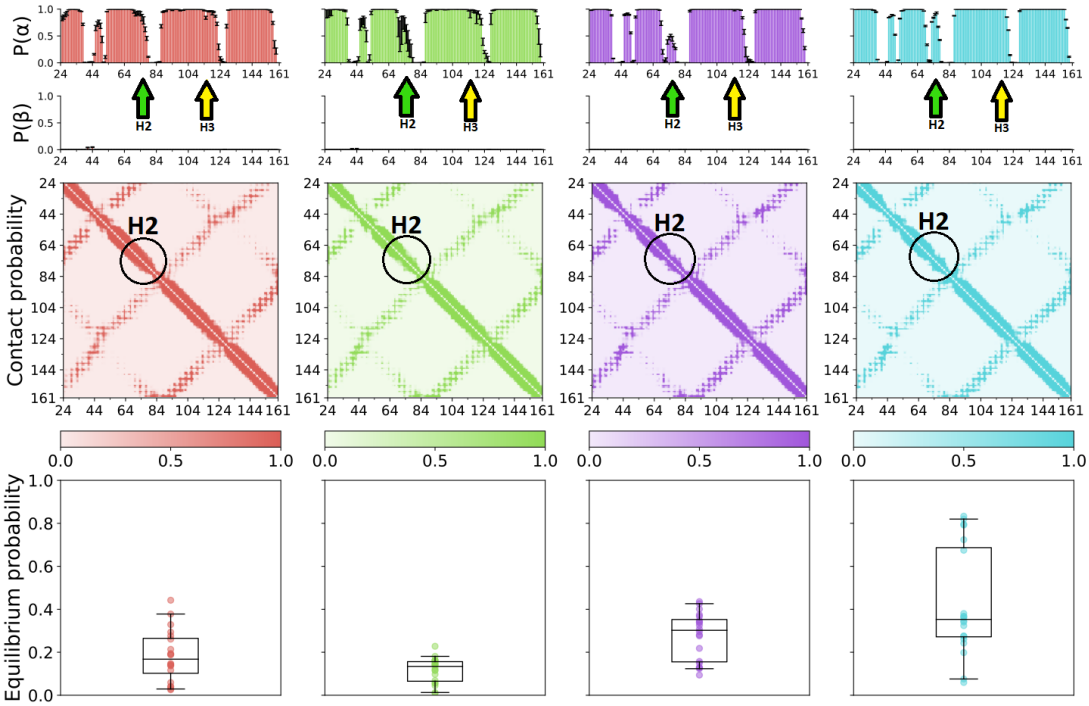


Figure 5.10: Average secondary structure, contact maps and equilibrium probability of states obtained for the APOE4 + 3SPA system.

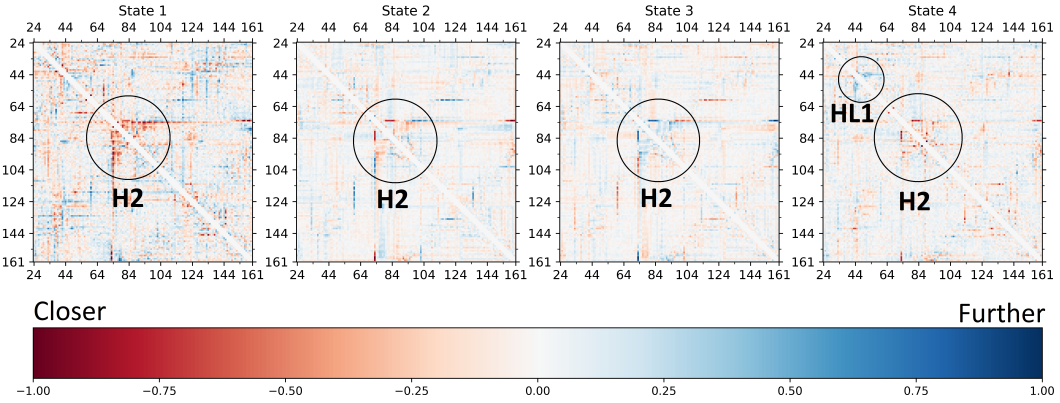


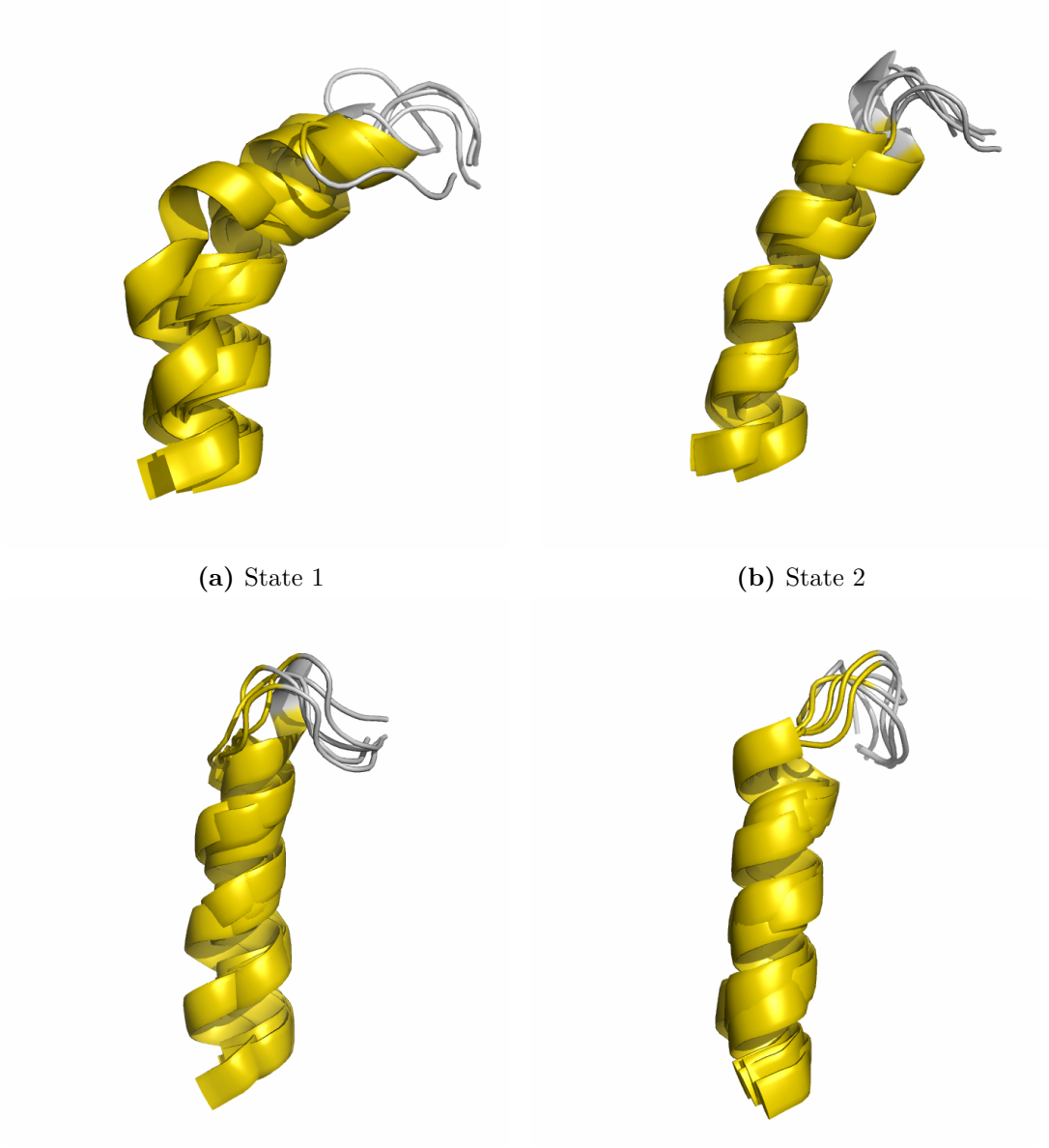
Figure 5.11: Feature importance matrices obtained by the 4-state CoVAMPnet model for the APOE4 + 3SPA system. Circles in the center of the matrices uderscore the importance of changes in the H2, while the small circle in the state 4 also suggest higher relevance of the HL1 structure.

Subdomain	State 1	State 2	State 3	State 4
H2 (70-73)	Structured	Intermediate	Unstructured	Unstructured
H2 (74-80)	Structured	Intermediate	Less structured	Structured
H3	Bent	Less bent	Even less bent	Straight
C-domain	Folded	Folded	Intermediate	Unfolded

Table 5.2: Structure of the most important subdomains observed in different states of the APOE4 + 3SPA system.

5.2.1 3SPA prevents the loss of structure in the H3 subdomain in APOE4

One notable change was the prevention of structure loss in the bent H3 subdomain observed in free APOE4. In the case of free APOE4, we observed gradual degradation of helical structure of residues 115-125 (see Fig. 4.12). In contrast to the free system, the presence of 3SPA led to a gradual unbending of the helix, with state 4 exhibiting a fully straight H3 (see Fig. 5.12).



(a) State 1 **(b)** State 2

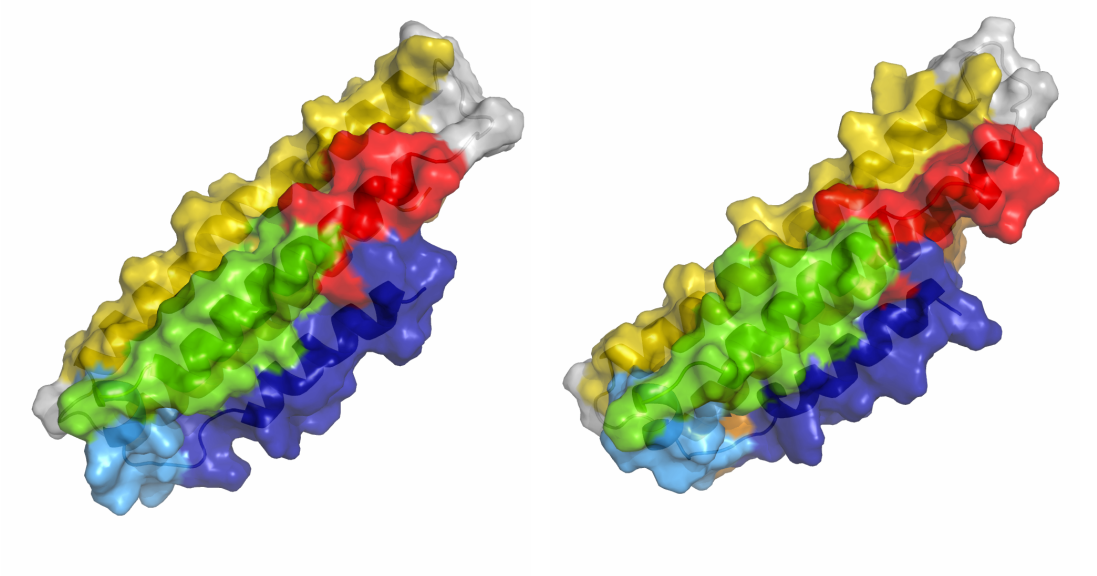
(c) State 3 **(d)** State 4

Figure 5.12: Structure of H3 in different states of APOE4+3SPA system. Residues 109-129, 5 representative frames. Notice the gradual unbending of the helix.

This particular change was already reported [19] and is considered as a very positive and promising effect of 3SPA: the straight H3 resembles the H3 from the free APOE3 system. Quoting the authors: "Interestingly, 3SPA modulates the structural features of APOE4, e.g., the conformation of helix H3 and the orientation of W34 towards resembling APOE3. It has been previously shown that small-molecule structure correctors can modify the aberrant conformation of APOE4 and abolish its detrimental effects in cultured neurons." [19]. If the shape of H3 was of crucial importance for the neurodegenerative effect of the APOE4, regardless of the hypotheses about the oligomerization processes, then it would prove that 3SPA serves as a good corrector.

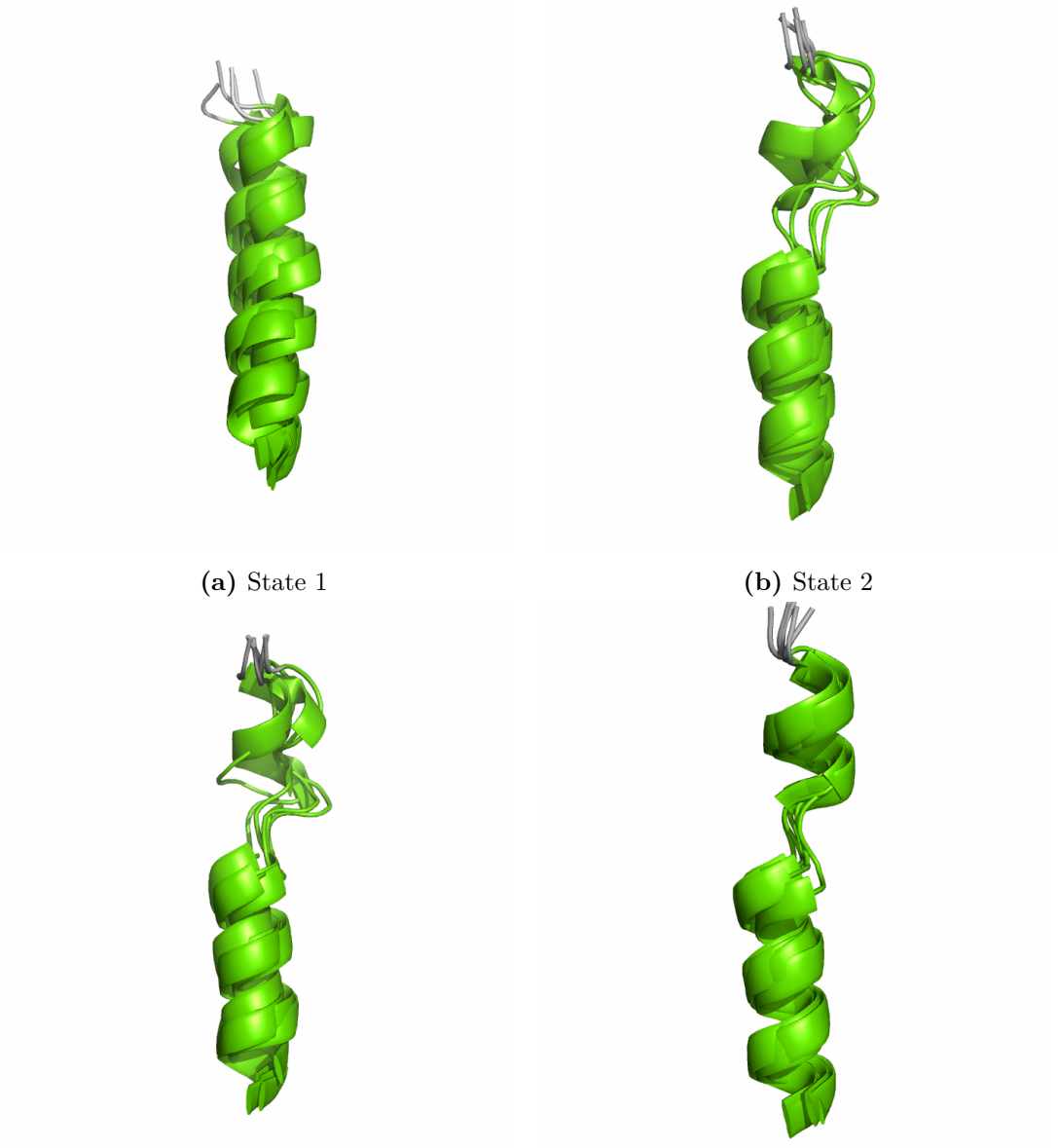
5.2.2 3SPA led to a loss of structure in H2 in APOE4

One distinct feature we detected with our analysis was the significant loss of structure in the H2. To be more specific, we observed a progressive loss of structure in residues 70-80 going from state 1 to state 3. Interestingly, state 4 saw an increase in structure for residues 74-80 at the end of H2, but the residues 70-73 looked consistently unstructured (see Fig. 5.14). Those changes were reflected in the altered surface of the protein (see Fig. 5.13). We consider this to be one of the most interesting findings, as according to yet unpublished observations from our collaborators from Loschmidt Laboratories, several residues in that area were identified to play a role in the creation of parallel dimers. APOE4 has a significantly higher propensity to form parallel dimer than APOE3, but this propensity was severly reduced in the presence of 3SPA. We could therefore hypothesize, that introduction of 3SPA disrupts the structural integrity of H2, which in turn results in weaker interactions between APOE4 molecules, reducing its propensity to create parallel dimers. Another important observation is that this change was clearly correlated with the unbending of the H3 (see Fig. 5.12), which was not previously reported. As a relatively small change, it could have been obscured in the analysis of the full-length APOE, reinforcing our belief that restricting it to the 4-helix bundle provided unique advantages.



(a) State 1 **(b)** State 4

Figure 5.13: Differences in protein surface between state 1 and state 4 of APOE4 + 3SPA. Only one frame per state was used for this visualization. Residues 70-80 are colored in red. Notice how big is the change of the surface for those residues.



(a) State 1 **(b)** State 2

(c) State 3 **(d)** State 4

Figure 5.14: Structure of H2 in different states of APOE4+3SPA system. Residues 61-82, 5 representative frames. Interestingly, after losing a significant amount of the structural integrity when transitioning from state 1 to state 2 and state 3, system seems to partially regain it in state 4, as indicated by shape and good alignment of frames in this state.

The interpretation of observed conformational changes is still a part of our ongoing discussions with our collaborators from Loschmidt Laboratories who posses in-depth knowledge about proteins and development of small molecule drugs. Their insight will provide the motivation to investigate most promising findings from this thesis more thoroughly.

Conclusion

Our study offers several significant insights, including previously unobserved structural changes in the 4-helix bundle of the APOE protein. In total, we analyzed four systems which allowed us to get insight into the dynamics of APOE3 and APOE4 simulated with and without the excess of 3SPA. The most important achievements of our analysis could be summarized as follows:

- We comprehensively analyzed flexibility of the HL1 across various systems, noting that the HL1 of free APOE3 demonstrated the most unstructured conformations among all the systems studied.
- We uncovered novel structural transformations at the end of H2 of the APOE4 + 3SPA system and their correlation with the straightening of H3.
- For the APOE3 + 3SPA system, we witnessed an unwinding of the end of H2 and an increased stability at the beginning of H3, as compared with the free APOE3 system.
- We identified that 3SPA introduces two changes in APOE3 that interestingly make APOE3 + 3SPA more similar to free APOE4, namely a bend and unwinding in the H3 helix and an increased stability of the HL1 subdomain. These findings could potentially elucidate the lack of 3SPA's positive effect on APOE3/APOE4 patients, as previous works were centered around how 3SPA induces more APOE3-like conformations on APOE4. We would like to put forward the idea that the opposite process also occurs.

All of these observations could provide an insight into the protein's oligomerization process, which is thought to be a precursor to Alzheimer's disease.

Additionally, we successfully validated several findings of our colleagues at Loschmidt Laboratories [19], who employed more traditional methods:

- The increased flexibility around L3 in APOE3, resulting in a higher unwinding of the beginning of H3.
- The stabilizing effect of 3SPA on APOE4, leading to a straightening of H3, hence making it more akin to APOE3.

VAMPnet simplified the creation of models that fulfilled the necessary implied timescales and CK test parameters, based on a representation encapsulating full complexity of APOE's 3D structure, as informed by the inter-residue distances. The use of feature importance matrices in CoVAMPnet allowed for an easy identification of the protein's most significant flexible regions, further showcasing the potential of this machine learning pipeline. To our knowledge, neither CoVAMPnet nor VAMPnet have been successfully utilized on a protein of similar size yet.

Our study also serves as a valuable case study by offering another example of the application of VAMPnets to highly flexible proteins with complex conformational landscapes, and highlighting the limitations of MD in this context. It shows that obtaining sufficient data to estimate the MSM that describes the

protein's equilibrium behavior remains a significant bottleneck. However, it also demonstrates that it is possible to derive informative MSMs, even if they cannot fully capture the system's equilibrium behavior, and even if some compromises on the processed data are necessary to manage the task's complexity – such as constraining the input representation, as we ultimately did by focusing on 4-helix bundle. Thus, our work contributes to a deeper understanding of the challenges involved in estimating MSMs with limited data.

Future work

In the future, we intend to further delve into the implications of the discovered conformational states, possibly extending the analysis with methods we did not apply yet. We believe that our observations possess the potential to deepen our understanding of the molecular foundations of Alzheimer's disease and enhance our comprehension of the impact of 3SPA on APOE, especially on its oligomerization process. Ultimately, this could lead to new avenues in the development of drug candidates for Alzheimer's disease.

We also aim to use the same pipeline to the simulations of the earlier described APOE dimers, with the objective of identifying the critical conformations during the formation of T-shaped, V-shaped, and parallel dimers. Due to the considerably higher complexity of this system, it will likely necessitate more computational resources. This may prompt us to consider modifications to the neural network architecture or the training regime. Notably, as dimer simulations exhibit permutational symmetry, architectures like Graph Neural Networks [55] [56] might be required. This could potentially offer a chance to employ their inductive bias to enable the network to concentrate on critical interactions between the two chains of the dimer, and perhaps pave the way for the creation of new architectures specialized in analyzing the dynamics of such intricate systems modeling interactions of multiple biomolecules. Moreover, we would like to implement and test the "concatenated trajectory" for time-lagged methods which we described in section 3.5.6.

Bibliography

[1] Kumar B Rajan, Jennifer Weuve, Lisa L Barnes, Elizabeth A McAninch, Robert S Wilson, and Denis A Evans. Population estimate of people with clinical Alzheimer's disease and mild cognitive impairment in the United States (2020–2060). *Alzheimer's & dementia*, 17(12):1966–1975, 2021.

[2] Alzheimer's Association. 2023 Alzheimer's disease facts and figures. *Alzheimers Dement*, 19(4):1598–1695, March 2023.

[3] Chia-Chen Liu, Chia-Chan Liu, Takahisa Kanekiyo, Huaxi Xu, and Guojun Bu. Apolipoprotein E and Alzheimer disease: risk, mechanisms and therapy. *Nat Rev Neurol*, 9(2):106–118, January 2013.

[4] Andreas Mardt, Luca Pasquali, Hao Wu, and Frank Noé. VAMPnets for deep learning of molecular kinetics. *Nature Communications*, 9(1):5, January 2018.

[5] Sérgio M. Marques, Petr Kouba, Anthony Legrand, Jiri Sedlar, Lucas Disson, Joan Planas-Iglesias, Zainab Sanusi, Antonin Kunka, Jiri Damborsky, Tomas Pajdla, Zbynek Prokop, Stanislav Mazurenko, Josef Sivic, and David Bednar. Effects of Alzheimer's Disease Drug Candidates on Disordered Aβ42 Dissected by Comparative Markov State Analysis (CoVAMPnet). *bioRxiv*, 2023.

[6] A. Kessel and N. Ben-Tal. *Introduction to Proteins: Structure, Function, and Motion*. Chapman & Hall/CRC Mathematical and Computational Biology, 2018.

[7] Yan Wang, Hang Zhang, Haolin Zhong, and Zhidong Xue. Protein domain identification methods and online resources. *Computational and Structural Biotechnology Journal*, 19:1145–1153, 2021.

[8] A. I. Dragan, C. Crane-Robinson, and P. L. Privalov. Thermodynamic basis of the α-helix and DNA duplex. *European Biophysics Journal*, 50(5):787–792, Jul 2021.

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

[10] Li-Quan Yang, Peng Sang, Yan Tao, Yun-Xin Fu, Ke-Qin Zhang, Yue-Hui Xie, and Shu-Qun Liu. Protein dynamics and motions in relation to their functions: several case studies and the underlying mechanisms. *J Biomol Struct Dyn*, 32(3):372–393, March 2013.

[11] Rodrigo Medeiros, David Baglietto-Vargas, and Frank M LaFerla. The role of tau in Alzheimer's disease and related disorders. *CNS Neurosci Ther*, 17(5):514–524, June 2010.

[12] Lolita Piersimoni, Marina Abd El Malek, Twinkle Bhatia, Julian Bender, Christin Brankatschk, Jaime Calvo Sánchez, Guy W Dayhoff, Alessio Di Ianni, Jhonny Oscar Figueroa Parra, Dailen Garcia-Martinez, Julia Hesselbarth, Janett Köppen, Luca M Lauth, Laurin Lippik, Lisa Machner, Shubhra Sachan, Lisa Schmidt, Robin Selle, Ioannis Skalidis, Oleksandr Sorokin, Daniele Ubbiali, Bruno Voigt, Alice Wedler, Alan An Jung Wei, Peter Zorn, Alan Keith Dunker, Marcel Köhn, Andrea Sinz, and Vladimir N Uversky. Lighting up Nobel prize-winning studies with protein intrinsic disorder. *Cell Mol Life Sci*, 79(8):449, July 2022.

[13] Yadong Huang and Robert W Mahley. Apolipoprotein E: structure and function in lipid metabolism, neurobiology, and Alzheimer's diseases. *Neurobiol Dis*, 72 Pt A:3–12, August 2014.

[14] Carl Frieden, Hanliu Wang, and Chris M. W. Ho. A mechanism for lipid binding to apoE and the role of intrinsically disordered regions coupled to domain–domain interactions. *Proceedings of the National Academy of Sciences*, 114(24):6292–6297, 2017.

[15] Philip B Verghese, Joseph M Castellano, Kanchan Garai, Yinong Wang, Hong Jiang, Aarti Shah, Guojun Bu, Carl Frieden, and David M Holtzman. ApoE influences amyloid-β (Aβ) clearance despite minimal apoE/Aβ association in physiological conditions. *Proc Natl Acad Sci U S A*, 110(19):E1807–16, April 2013.

[16] Axel Montagne, Daniel A. Nation, Abhay P. Sagare, Giuseppe Barisano, Melanie D. Sweeney, Ararat Chakhoyan, Maricarmen Pachicano, Elizabeth Joe, Amy R. Nelson, Lina M. D'Orazio, David P. Buennagel, Michael G. Harrington, Tammie L. S. Benzinger, Anne M. Fagan, John M. Ringman, Lon S. Schneider, John C. Morris, Eric M. Reiman, Richard J. Caselli, Helena C. Chui, Julia TCW, Yining Chen, Judy Pa, Peter S. Conti, Meng Law, Arthur W. Toga, and Berislav V. Zlokovic. Apoe4 leads to blood–brain barrier dysfunction predicting cognitive decline. *Nature*, 581(7806):71–76, May 2020.

[17] G William Rebeck. The role of APOE on lipid homeostasis and inflammation in normal brains. *J Lipid Res*, 58(8):1493–1499, March 2017.

[18] Tristan Williams, Alejandra Jolie Ruiz, Angelica Maria Ruiz, Quan Vo, Wangchen Tsering, Guilian Xu, Karen McFarland, Benoit I. Giasson, Patrick Sullivan, David R. Borchelt, and Paramita Chakrabarty. Impact of APOE genotype on prion-type propagation of tauopathy. *Acta Neuropathologica Communications*, 10(1):57, Apr 2022.

[19] Michal Nemergut, Sérgio M. Marques, Lukas Uhrik, Tereza Vanova, Marketa Nezvedova, Darshak Chandulal Gadara, Durga Jha, Jan Tulis, Veronika Novakova, Joan Planas-Iglesias, Antonin Kunka, Anthony Legrand, Hana Hribkova, Veronika Pospisilova, Jiri Sedmik, Jan Raska, Zbynek Prokop, Jiri Damborsky, Dasa Bohaciakova, Zdenek Spacil, Lenka Hernychova, David Bednar, and Martin Marek. Domino-like effect of C112R mutation on ApoE4

aggregation and its reduction by Alzheimer's disease drug candidate. *Molecular Neurodegeneration*, 18(1):38, Jun 2023.

[20] Scott A Hollingsworth and Ron O Dror. Molecular dynamics simulation for all. *Neuron*, 99(6):1129–1143, September 2018.

[21] David E. Shaw, Peter J. Adams, Asaph Azaria, Joseph A. Bank, Brannon Batson, Alistair Bell, Michael Bergdorf, Jhanvi Bhatt, J. Adam Butts, Timothy Correia, Robert M. Dirks, Ron O. Dror, Michael P. Eastwood, Bruce Edwards, Amos Even, Peter Feldmann, Michael Fenn, Christopher H. Fenton, Anthony Forte, Joseph Gagliardo, Gennette Gill, Maria Gorlatova, Brian Greskamp, J.P. Grossman, Justin Gullingsrud, Anissa Harper, William Hasenplaugh, Mark Heily, Benjamin Colin Heshmat, Jeremy Hunt, Douglas J. Ierardi, Lev Iserovich, Bryan L. Jackson, Nick P. Johnson, Mollie M. Kirk, John L. Klepeis, Jeffrey S. Kuskin, Kenneth M. Mackenzie, Roy J. Mader, Richard McGowen, Adam McLaughlin, Mark A. Moraes, Mohamed H. Nasr, Lawrence J. Nociolo, Lief O'Donnell, Andrew Parker, Jon L. Peticolas, Goran Pocina, Cristian Predescu, Terry Quan, John K. Salmon, Carl Schwink, Keun Sup Shim, Naseer Siddique, Jochen Spengler, Tamas Szalay, Raymond Tabladillo, Reinhard Tartler, Andrew G. Taube, Michael Theobald, Brian Towles, William Vick, Stanley C. Wang, Michael Wazlowski, Madeleine J. Weingarten, John M. Williams, and Kevin A. Yuh. Anton 3: Twenty microseconds of molecular dynamics simulation before lunch. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, SC '21, New York, NY, USA, 2021. Association for Computing Machinery.

[22] J Schlitter, M Engels, and P Krüger. Targeted molecular dynamics: a new approach for searching pathways of conformational transitions. *J Mol Graph*, 12(2):84–89, June 1994.

[23] Alessandro Laio and Michele Parrinello. Escaping free-energy minima. *Proc Natl Acad Sci U S A*, 99(20):12562–12566, September 2002.

[24] L. Molgedey and H. G. Schuster. Separation of a mixture of independent signals using time delayed correlations. *Phys. Rev. Lett.*, 72:3634–3637, Jun 1994.

[25] Steffen Schultze and Helmut Grubmüller. Time-lagged independent component analysis of random walks and protein dynamics. *Journal of Chemical Theory and Computation*, 17(9):5766–5776, 2021. PMID: 34449229.

[26] Brooke E. Husic and Vijay S. Pande. Markov state models: From an art to a science. *Journal of the American Chemical Society*, 140(7):2386–2396, 2018. PMID: 29323881.

[27] Peter Deuflhard and Marcus Weber. Robust Perron cluster analysis in conformation dynamics. *Linear Algebra and its Applications*, 398:161–184, 2005. Special Issue on Matrices and Mathematical Biology.

[28] Yi Isaac Yang, Qiang Shao, Jun Zhang, Lijiang Yang, and Yi Qin Gao. Enhanced sampling in molecular dynamics. *The Journal of Chemical Physics*, 151(7):070902, 08 2019.

[29] S. Doerr, M. J. Harvey, Frank Noé, and G. De Fabritiis. HTMD: High-Throughput Molecular Dynamics for molecular discovery. *Journal of Chemical Theory and Computation*, 12(4):1845–1852, 2016. PMID: 26949976.

[30] Maxwell I. Zimmerman, Justin R. Porter, Xianqiang Sun, Roseane R. Silva, and Gregory R. Bowman. Choice of adaptive sampling strategy impacts state discovery, transition probabilities, and the apparent mechanism of conformational changes. *Journal of Chemical Theory and Computation*, 14(11):5459–5475, 2018. PMID: 30240203.

[31] Diego E. Kleiman and Diwakar Shukla. Active Learning of the Conformational Ensemble of Proteins Using Maximum Entropy VAMPnets. *Journal of Chemical Theory and Computation*, 0(0):null, 2023. PMID: 37027313.

[32] Steven L. Brunton, Marko Budišić, Eurika Kaiser, and J. Nathan Kutz. Modern Koopman theory for dynamical systems, 2021.

[33] Hao Wu and Frank Noé. Variational approach for learning Markov processes from time series data, 2019.

[34] G. Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals and Systems*, 2(4):303–314, Dec 1989.

[35] David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. Learning representations by back-propagating errors. *Nature*, 323(6088):533–536, Oct 1986.

[36] Kirill A. Konovalov, Ilona Christy Unarta, Siqin Cao, Eshani C. Goonetilleke, and Xuhui Huang. Markov state models to study the functional dynamics of proteins in the wake of machine learning. *JACS Au*, 1(9):1330–1341, 2021.

[37] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *Workshop at International Conference on Learning Representations*, 2014.

[38] Zhenhua Mai, Wenyan Wei, Haibin Yu, Yongze Chen, Yongxiang Wang, and Yuanlin Ding. Molecular recognition of the interaction between ApoE and the TREM2 protein. *Transl Neurosci*, 13(1):93–103, April 2022.

[39] Jinghui Luo, Jean-Didier Maréchal, Sebastian Wärmländer, Astrid Gräslund, and Alex Perálvarez-Marín. In silico analysis of the apolipoprotein E and the amyloid beta peptide interaction: misfolding induced by frustration of the salt bridge network. *PLoS Comput Biol*, 6(2):e1000663, February 2010.

[40] Richard Y-C. Huang, Kanchan Garai, Carl Frieden, and Michael L. Gross. Hydrogen/Deuterium Exchange and Electron-Transfer Dissociation Mass

Spectrometry Determine the Interface and Dynamics of Apolipoprotein E Oligomerization. *Biochemistry*, 50(43):9273–9282, 2011. PMID: 21899263.

[41] Sagrario Manzano, Luis Agüera, Miquel Aguilar, and Javier Olazarán. A review on tramiprosate (homotaurine) in Alzheimer's disease and other neurocognitive disorders. *Frontiers in Neurology*, 11, 2020.

[42] S Abushakra, A Porsteinsson, B Vellas, J Cummings, S Gauthier, J A Hey, A Power, S Hendrix, P Wang, L Shen, J Sampalis, and M Tolar. Clinical benefits of tramiprosate in Alzheimer's disease are associated with higher number of APOE4 alleles: The "APOE4 Gene-Dose effect". *J Prev Alzheimers Dis*, 3(4):219–228, 2016.

[43] Helen M. Berman, John Westbrook, Zukang Feng, Gary Gilliland, T. N. Bhat, Helge Weissig, Ilya N. Shindyalov, and Philip E. Bourne. The Protein Data Bank. *Nucleic Acids Research*, 28(1):235–242, 01 2000.

[44] Thomas Löhr, Kai Kohlhoff, Gabriella T. Heller, Carlo Camilloni, and Michele Vendruscolo. A kinetic ensemble of the Alzheimer's Aβ peptide. *Nature Computational Science*, 1(1):71–78, Jan 2021.

[45] Thomas Löhr, Kai Kohlhoff, Gabriella T. Heller, Carlo Camilloni, and Michele Vendruscolo. A small molecule stabilizes the disordered native state of the Alzheimer's Aβ peptide. *ACS Chemical Neuroscience*, 13(12):1738–1745, 2022. PMID: 35649268.

[46] J. B. MacQueen. In L. M. Le Cam and J. Neyman, editors, *Proc. of the fifth Berkeley Symposium on Mathematical Statistics and Probability*, 1967.

[47] Andreas Mardt, Luca Pasquali, Frank Noé, and Hao Wu. Deep learning Markov and Koopman models with physical constraints. In Jianfeng Lu and Rachel Ward, editors, *Proceedings of The First Mathematical and Scientific Machine Learning Conference*, volume 107 of *Proceedings of Machine Learning Research*, pages 451–475. PMLR, 20–24 Jul 2020.

[48] Günter Klambauer, Thomas Unterthiner, Andreas Mayr, and Sepp Hochreiter. Self-normalizing neural networks. *CoRR*, abs/1706.02515, 2017.

[49] William C. Swope, Jed W. Pitera, and Frank Suits. Describing protein folding kinetics by molecular dynamics simulations. 1. theory. *The Journal of Physical Chemistry B*, 108(21):6571–6581, 2004.

[50] N.G. VAN KAMPEN. Chapter iv - markov processes. In N.G. VAN KAMPEN, editor, *Stochastic Processes in Physics and Chemistry (Third Edition)*, North-Holland Personal Library, pages 73–95. Elsevier, Amsterdam, third edition edition, 2007.

[51] Schrödinger, LLC. The PyMOL molecular graphics system, version 1.8. November 2015.

[52] Anton V Sinitskiy and Vijay S Pande. Theoretical restrictions on longest implicit time scales in Markov state models of biomolecular dynamics. *J Chem Phys*, 148(4):044111, January 2018.

[53] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. A simple framework for contrastive learning of visual representations. *CoRR*, abs/2002.05709, 2020.

[54] John Strahan, Adam Antoszewski, Chatipat Lorpaiboon, Bodhi P Vani, Jonathan Weare, and Aaron R Dinner. Long-Time-Scale predictions from Short-Trajectory data: A benchmark analysis of the Trp-Cage miniprotein. *J Chem Theory Comput*, 17(5):2948–2963, April 2021.

[55] Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *CoRR*, abs/1609.02907, 2016.

[56] Tian Xie, Arthur France-Lanord, Yanming Wang, Yang Shao-Horn, and Jeffrey C. Grossman. Graph dynamical networks for unsupervised learning of atomic scale dynamics in materials. *Nature Communications*, 10(1):2667, Jun 2019.

List of Figures

List of Tables

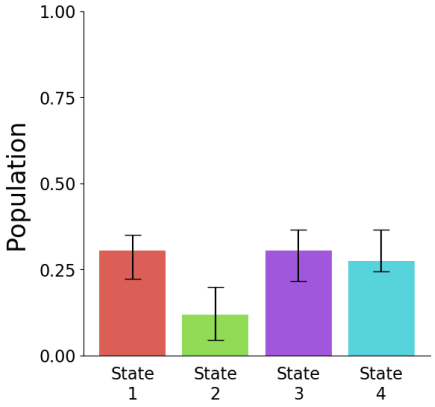
List of Abbreviations

Abbreviation	Explanation
AD	Alzheimer's disease
APOE	Apolipoprotein E
IDP	Intrinsically disordered protein
MD	Molecular dynamics
MSM	Markov state model
tICA	Time-lagged independent component analysis
3SPA	3-sulfopropanoic acid

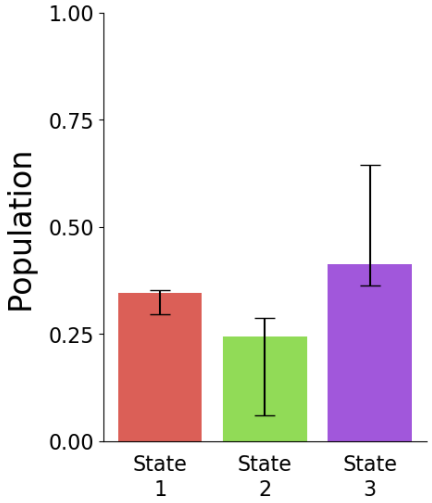
A. Appendix

A.1 Populations of states calculated according to hard assignments

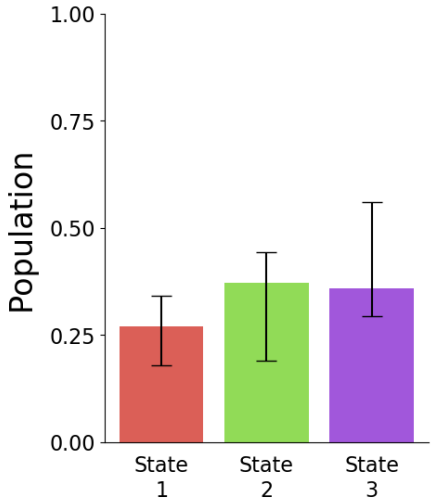
In a situation when we can consider the estimated Koopman operator to be reliable, equilibrium probabilities are more informative - they inform us not only about the number of frames observed in a given state, but about the hypothetical distribution of frames we would observe infinitely sampling from the equilibrium. In our case, however, the reliability of the operators is arguable, therefore we provide these hard assignment populations based estimates.



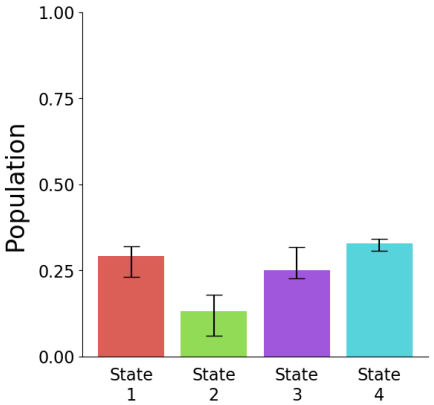
(a) Free APOE3



(b) Free APOE4



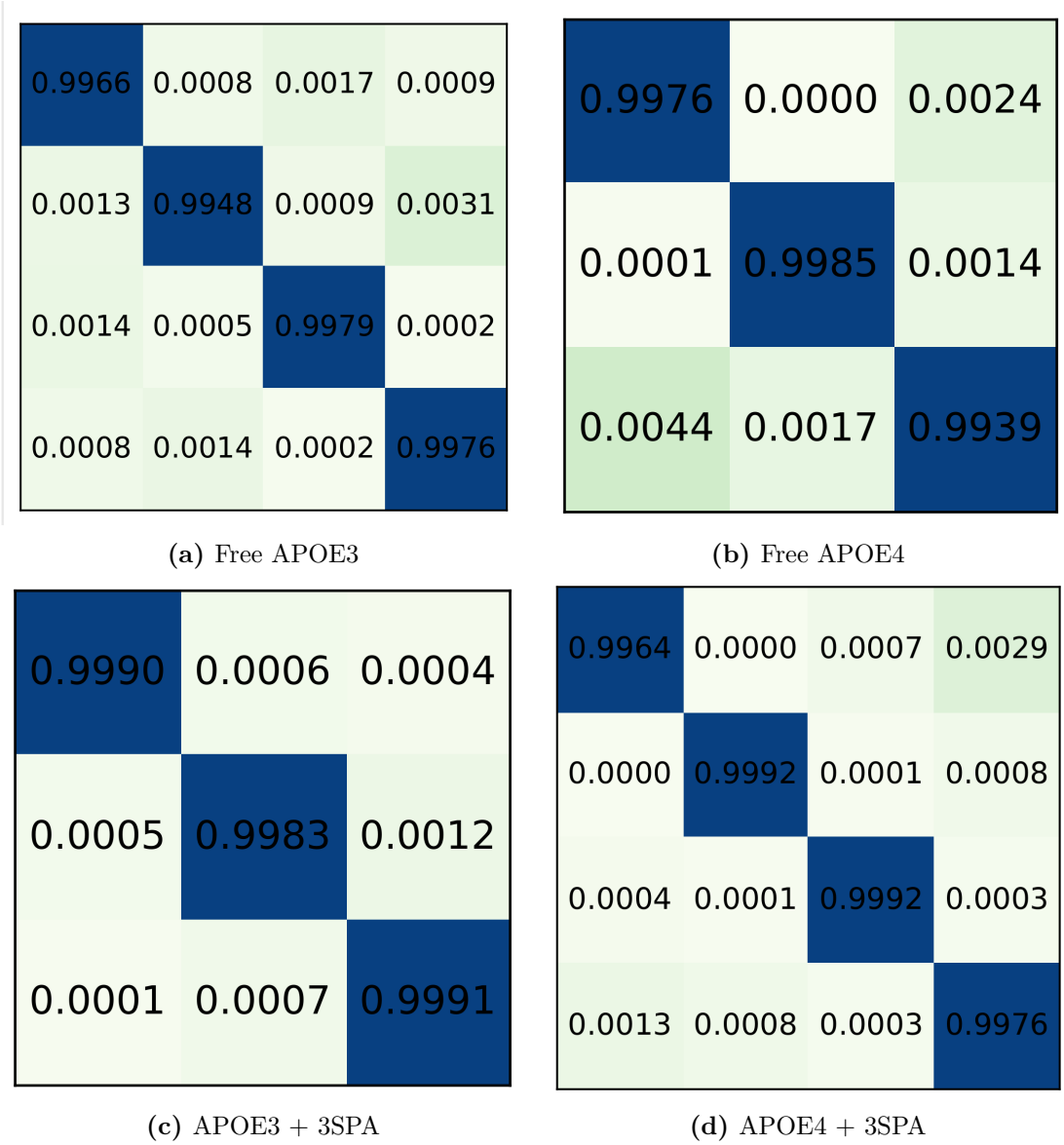
(c) APOE3 + 3SPA



(d) APOE4 + 3SPA

Figure A.1: Populations of all analyzed systems.

A.2 Koopman operators



(a) Free APOE3

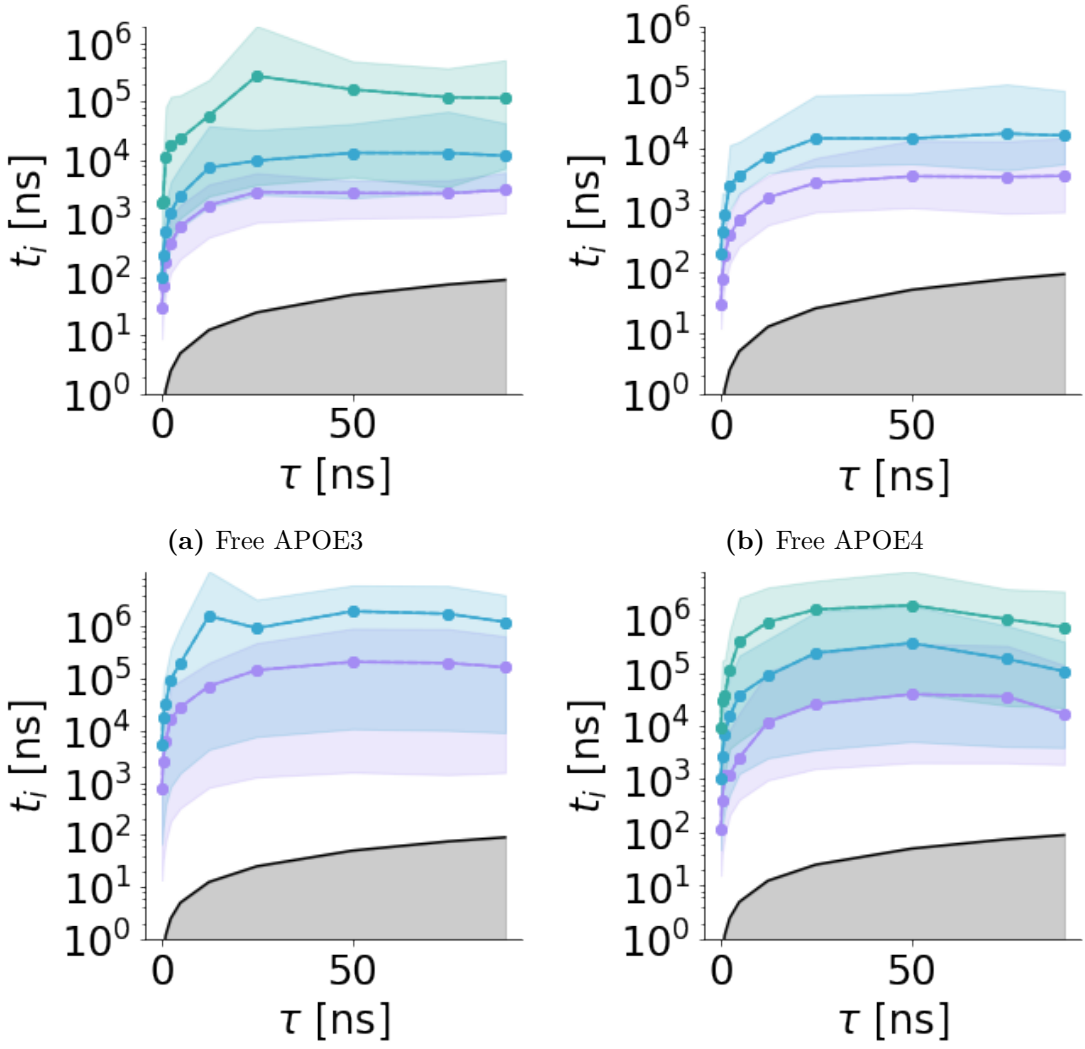
(b) Free APOE4

(c) APOE3 + 3SPA

(d) APOE4 + 3SPA

Figure A.2: Koopman operators of all analyzed systems.

A.3 Implied timescales



(a) Free APOE3

(b) Free APOE4

(c) APOE3 + 3SPA

(d) APOE4 + 3SPA

Figure A.3: Implied timescales of all analyzed systems.

A.4 CK tests

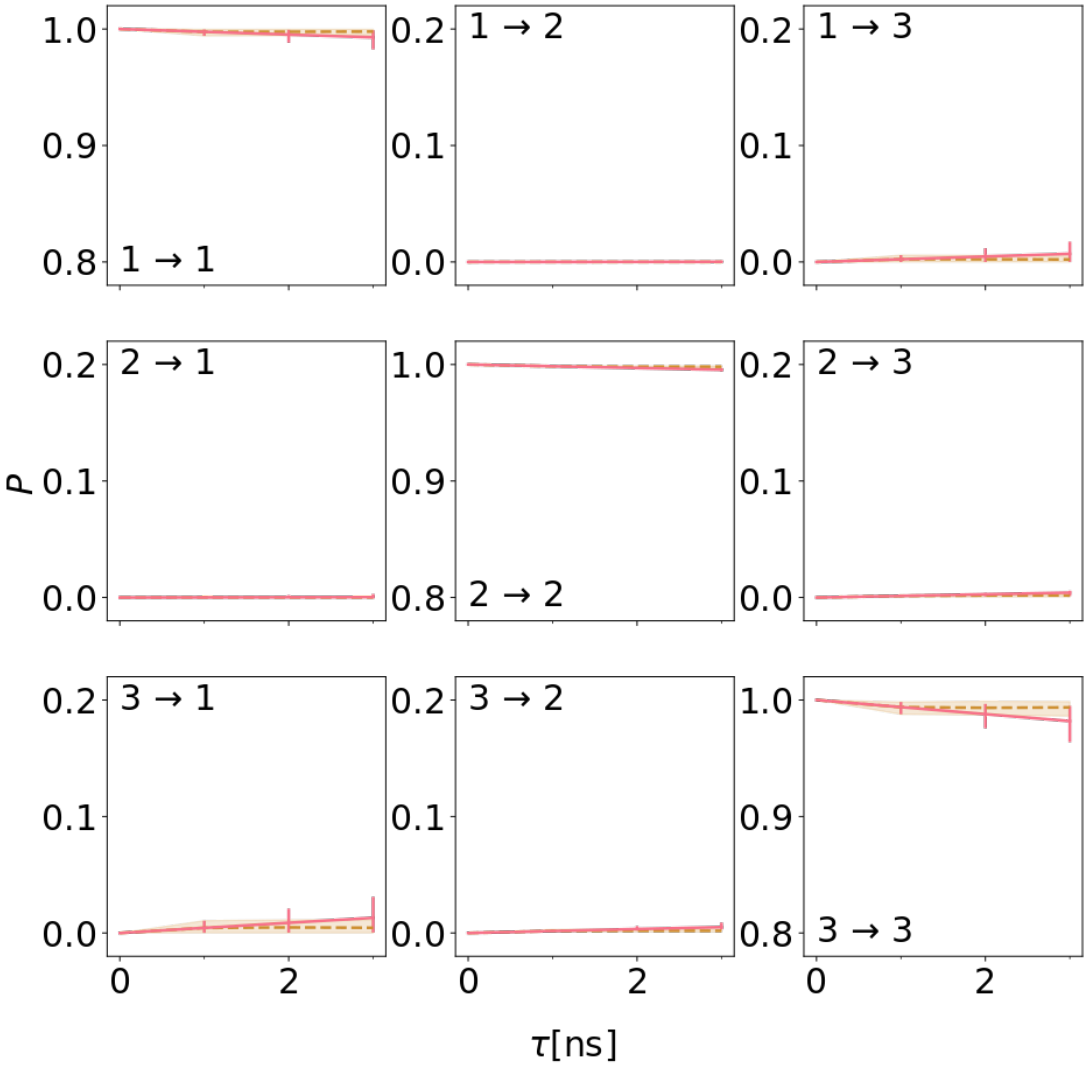


Figure A.4: Chapman-Kolmogorov test of the free APOE4 system.

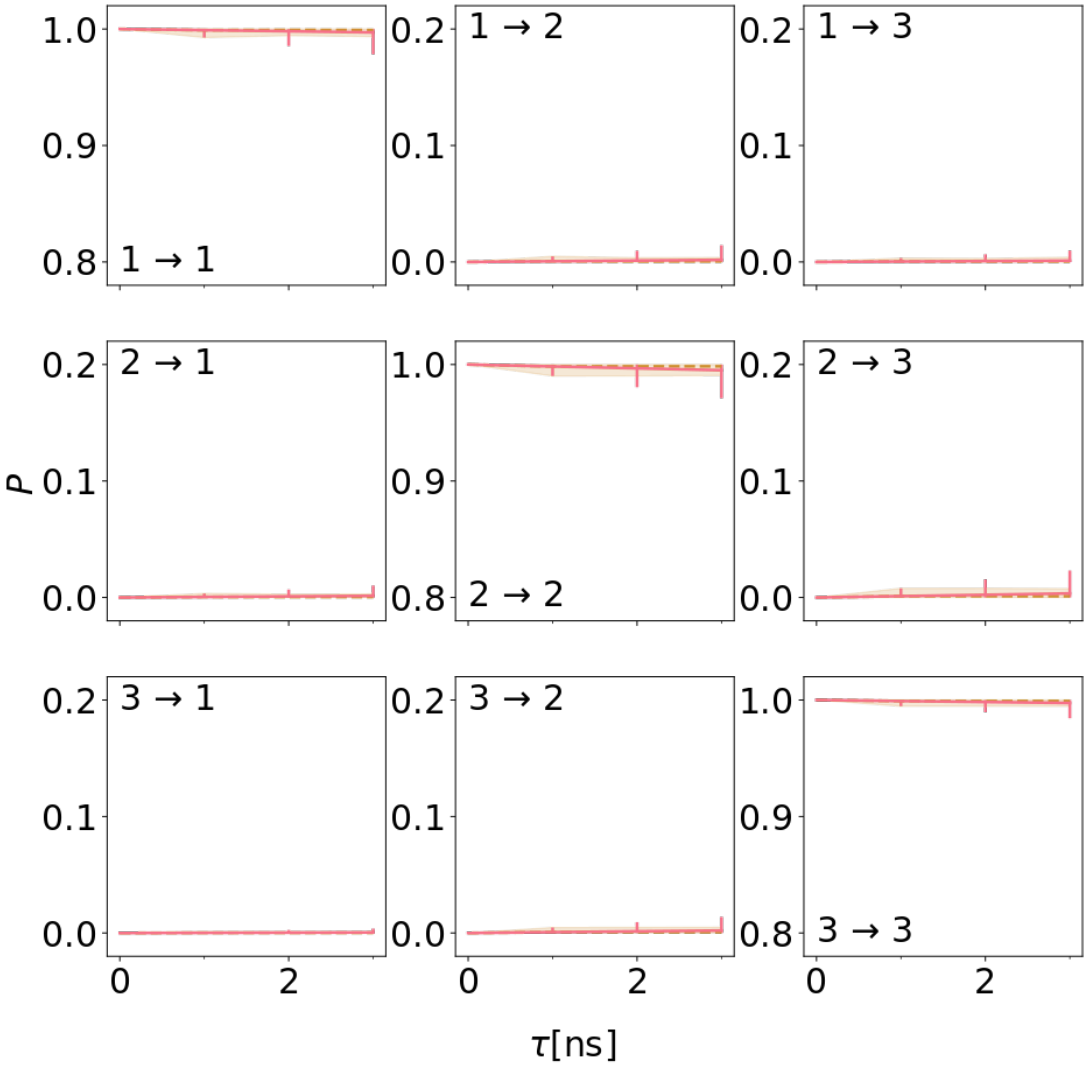


Figure A.5: Chapman-Kolmogorov test of the APOE3 + 3SPA system.

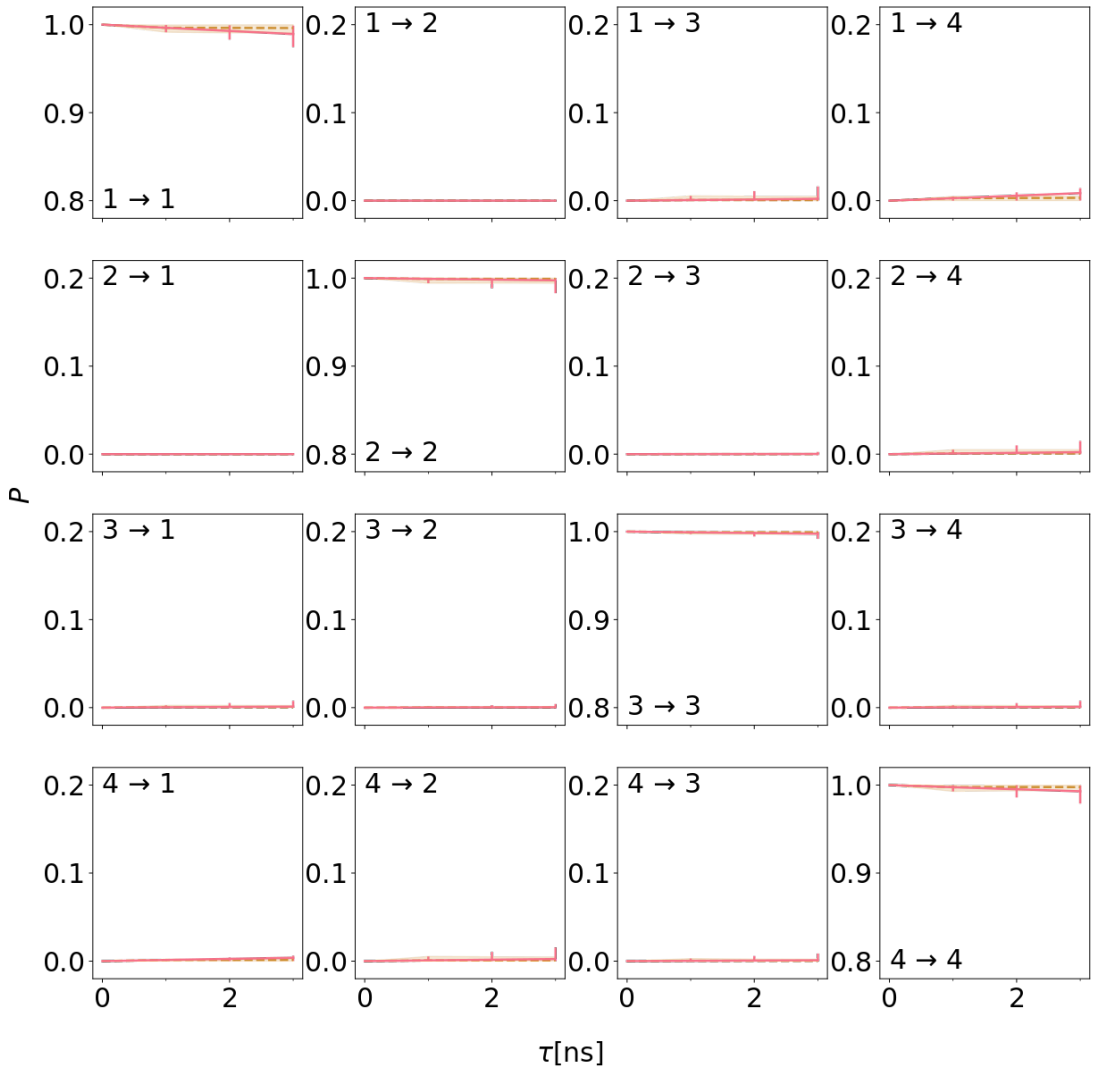


Figure A.6: Chapman-Kolmogorov test of the APOE4 + 3SPA system.