



# Exploring the locking stage of NFGAILS amyloid fibrillation via transition manifold analysis

Andreas Bittracher<sup>1,a</sup> , Johann Moschner<sup>2</sup>, Beate Kokschr<sup>2</sup>, Roland Netz<sup>3</sup>, and Christof Schütte<sup>1,4</sup>

<sup>1</sup> Department of Mathematics, Freie Universität Berlin, Berlin, Germany

<sup>2</sup> Department of Chemistry, Freie Universität Berlin, Berlin, Germany

<sup>3</sup> Department of Physics, Freie Universität Berlin, Berlin, Germany

<sup>4</sup> Zuse Institute Berlin, Berlin, Germany

Received 13 April 2021 / Accepted 7 September 2021 / Published online 30 September 2021  
© The Author(s) 2021

**Abstract.** We demonstrate the application of the transition manifold framework to the late-stage fibrillation process of the NFGAILS peptide, a amyloidogenic fragment of the human islet amyloid polypeptide (hIAPP). This framework formulates machine learning methods for the analysis of multi-scale stochastic systems from short, massively parallel molecular dynamical simulations. We identify key intermediate states and dominant pathways of the process. Furthermore, we identify the optimally timescale-preserving reaction coordinate for the dock-lock process to a fixed pre-formed fibril and show that it exhibits strong correlation with the mean native hydrogen-bond distance. These results pave the way for a comprehensive model reduction and multi-scale analysis of amyloid fibrillation processes.

## 1 Introduction

Amyloid fibrils—long, well-ordered aggregates of short monomeric peptides—are long known to be associated with many neurodegenerative diseases such as Alzheimer’s or Parkinson’s [1,2]. More recent is the suggestion that it is mostly the half-formed, soluble oligomers that are pathogenic [3–5]. To develop new therapies that specifically intervene in the formation of amyloids, it is therefore vital to understand the process of both the initial oligomer formation, as well as the advanced fibril growth. Both have been studied extensively, and for an exhaustive review of recent advances, see [6].

### The dock-lock model of fibril elongation

Due to its microscale nature (in both size and duration), the initial formation of disordered and partially ordered nuclei from soluted monomers has mostly been studied by molecular dynamical simulations [7,8]. Mathematical models for the small-scale growth kinetics [9] as well as the conformational transitions [10] could be built, and

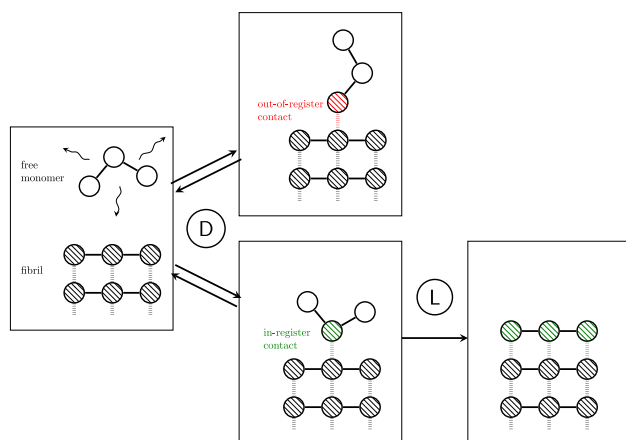
provided a decent understanding of the reaction pathways on an atomic level.

The later stages of fibril growth have been studied by both *in vitro* [11–13] and *in silico* [14] experiments (though the latter are limited due to the scale of the problem). With their help, comprehensive models for various aspects of the reaction could be built; see [15] for an overview. The so-called *dock-lock mechanism* [16] here is the prevalent model for the ordered elongation of already-formed fibrils. It describes the primary growth mechanism at realistic concentrations as the attachment of single peptides to either end of a “template” fibril [14]. Moreover, the monomer attachment is characterized by the existence of two stages, the “docking” stage in which the incoming monomer only loosely associates with the fibril template, forming only few, weak and thus reversible contacts, and the “locking” stage, in which the monomer undergoes internal re-configuration to form the native contacts across the whole length of the monomer. The process is illustrated in Fig. 1.

The validity of this model has been demonstrated in multiple computational experiments on various amyloid species. For example, the two stages could clearly be distinguished in long all-atom simulations of a  $A\beta_{16-22}$  fibril [17]. In the same work, it has also been found that the locking stage is the rate-limiting step, with typical durations of around 200 ns. In [18], a Markov state model of the combined dock-lock mechanism was constructed for the  $TRR_{105-115}$  amyloid. Based on multiple long all-atom simulations starting from multiple unbound states, intermediate and trapping states were identified, along with transition rates and dominant

**Supplementary Information** The online version contains supplementary material available at <https://doi.org/10.1140/epjb/s10051-021-00200-0>.

<sup>a</sup>e-mail: [bittracher@mi.fu-berlin.de](mailto:bittracher@mi.fu-berlin.de) (corresponding author)



**Fig. 1** Schematic illustration of the dock-lock mechanism. Incoming monomers first bind lightly to a pre-formed fibril template (the *docking phase*, D). Subsequently, more and more native contacts in the form of H bonds are formed in a process involving multiple intermediate steps, until the binding becomes (nearly) irreversible (the *locking phase*, L)

pathways<sup>1</sup>. However, the identified metastable intermediates have not been investigated for the existence of the characteristic native contacts that form with the progression of the locking stage (see Sect. 2.1). Finally, in [19], the locking stage of  $A\beta_{16-22}$  has been analyzed in isolation. Based on short simulations from multiple “milestones” along the progress of the locking stage, the expected formation and breakage times of all native contacts could be estimated.

Despite the successes in predicting *kinetic* aspects of the overall aggregation, the microscale *dynamical* aspects (e.g., the reaction pathways) of the locking stage have not been investigated extensively thus far. This represents a substantial deficiency, as it is the inter-peptide bonds formed during that stage that give many amyloids their extraordinary durability. However, this stage is also particularly challenging to analyze computationally due to its slowly equilibrating, rate-limiting, and highly metastable nature.

### Transition manifold system analysis

In this respect, the locking phase of amyloid aggregation resembles protein folding dynamics, another process that is notoriously hard to simulate and analyze due to its separation of scales. Recently, a novel machine learning approach for the analysis of such systems was developed by some of the authors, called the *transition manifold framework* (TMF) [20]. The goals of this approach are threefold:

1. The identification of dynamically meaningful, timescale-preserving reaction coordinates (RCs), i.e.,

<sup>1</sup> Note that “pathway” here merely means a specific sequence of discrete Markov states, and—unlike in TPT—does not describe a continuous curve through state space.

- low-dimensional observables of the full system that are associated with slow phenomena.
2. The identification of dominant pathways associated with the slow phenomena. By following these pathways, “artificial reactive trajectories” can be constructed without the need to simulate the full system.
3. Gaining a visual impression of the essential dynamical structure. This is achieved naturally by the approach, as its algorithms are based on embedding the abstract “backbone” of the dynamics into Euclidean space.

Unlike other methods with similar goals, the algorithms of the TMF require short, local simulation bursts, without the need to ever simulate a full-binding event, instead of long equilibrated trajectories.

Applied to the NTL9 protein folding process [21], the TMF was able to identify chemically interpretable optimal RCs, as well as key folding milestones with an accuracy comparable to neural network-based Markov state modeling techniques [21]. Note however that, unlike those *kinetic* model reduction methods, the main focus of the TMF is less the accurate reproduction of system statistics (although the transition manifold reaction coordinate is specifically constructed to preserve those), but to provide chemically intuitive and interpretable insight into the system’s effective behavior.

This gives reason to hope that similar insights can be gained for the locking phase of amyloid aggregation. In this article, we therefore demonstrate the application of the various transition manifold methods to such a system, specifically to the NFGAILS heptameric amyloid. The individual steps will however be described as universal as possible, so that this work should be seen not so much as a quantitative analysis of this specific molecular system, but more as general instructions for the analysis of a wide range of amyloid species. We motivate the framework and associated methods from a general mathematical/dynamical point of view and discuss the theoretical and numerical requirements on the underlying system. We demonstrate the data collection via MD simulation, apply the various machine learning methods, and finally give an interpretation of the results aimed toward a qualitative understanding of the locking process.

## 2 Methods

### 2.1 Characteristics of NFGAILS amyloid fibrils

The NFGAILS heptapeptide is a highly amyloidalogenic fragment of the human islet amyloid polypeptide (hIAPP), whose soluble oligomers play a role in type II diabetes. It was experimentally confirmed to form fibrils characterized as *steric zippers* [22]. That is, the individual monomers are arranged in beta-sheets that are internally stabilized by hydrogen bonds, and the beta-sheets in turn are stacked face-to-face, stabilized



the various machine learning techniques to the samples. The general workflow is as follows:

1. Sample starting points  $x_k \in \mathbb{R}^{3N}$  uniformly from the admissible state space
2. Estimate the transition densities  $p^t(x_k, \cdot)$  by parallel simulation (Monte Carlo sampling)
3. Compute pair-wise distances between densities (based on the estimates)

$$D_{ij} = d(p^t(x_i, \cdot), p^t(x_j, \cdot))$$

for a “meaningful” distance metric  $d$  (see below).

4. Apply embedding and manifold learning algorithms to  $D$ .

The distance metric  $d$  needs to be meaningful in the sense that it can reliably distinguish densities, but is not overly sensitive to small deviations. Generic choices include the maximum mean discrepancy, Wasserstein metric, or Kullback–Leibler divergence. The reliance on statistical distances as opposed to state space distances can therefore be seen as another advantage of our approach, independent of the dynamical interpretation. It is a well-known phenomenon that the pair-wise distance between randomly drawn points becomes constant—and hence meaningless—with growing state space dimension [28].

Finally, empirical approximation of the densities  $p^t(x_k, \cdot)$  requires the ability to simulate the system up to time  $t$ . Unlike the generation of long MD trajectories, however, sampling the  $p^t(x_k, \cdot)$  is trivially parallelizable, and thus well suited for distributed computing architectures. Moreover, as explained above, only relatively short simulations are necessary.

### 3 Computational setup

The specific molecular system used in our experiments consists of a pre-formed NFGAILS heptamer fibril and an incoming monomer. One monomer consists of 53 heavy atoms, the overall system hence of 424 heavy atoms. Simulation is performed in aqueous solution at 310K. For details on the MD parameters, see the supplementary information (SI).

The locking phase of each of the six possible initial docking contacts described in Sect. 2.1 (three for each the “even” and “odd” configuration) essentially corresponds to a separate molecular system, with its own transition manifold. Hence, also for longer peptides with more native contacts, each starting state of interest needs to be investigated separately. “Interesting” states would typically be those with the highest probability to be formed during the docking phase, but could also be selected based on specific chemical expert knowledge. Hence, we limit our investigations to the two outer, i.e., most exposed contacts LEU–PHE and PHE–LEU of the “even” configuration. However, the

experimental setup is valid for the remaining contacts, as well.

To facilitate the subsequent analysis, we impose two artificial restrictions on the binding process:

1. The heavy atoms of the fibril core are restrained to their crystal configuration. This way, only the motion of the monomer atoms is relevant for the subsequent analysis.
2. The initial native contact is prevented from breaking. This prevents the monomer from dissociating from the fibril and thus ending the locking phase. As such trajectories are not part of any successful docking pathways, they represent wasted computational effort.

The restraints are realized by imposing a strong harmonic potential on the respective atom positions.

As these restraints leave the heptamer fibril essentially motionless (except of fast, low-amplitude vibrations in the restraint potential), the system effectively consists only of the 53 heavy atoms of the incoming monomer. Hence, we will consider only the degrees of freedom of the monomer in the transition manifold analysis. We therefore have  $N = 53$  atoms which we consider in cartesian coordinates, leading to a  $3 \cdot 53 = 159$ -dimensional state space. Moreover, due to the fixed position of the template fibril in space, no global translational or rotational movement can occur in the incoming monomer which normally would have to be removed by alignment to some reference structure.

#### 3.1 Sampling of the reaction space

The first step of the transition manifold algorithm now consists of sampling starting points  $x_k$  from configuration space. The sampled states should roughly cover the full range of the reaction, i.e., contain states that are “freshly docked”, “almost locked”, and everything in between, to obtain a dense covering of the transition manifold. Note that for this, it is *not* necessary to sample the admissible state space densely, as one point on the transition manifold corresponds to many points in state space, and in theory, one of these points is sufficient to mark the transition manifold in an embedding. The number of required sample points hence scales (linearly) with the size and complexity<sup>3</sup> of the transition manifold, and not with the dimension of the state space.

For creating the random samples, we use a *heat sampling* approach: we consider a configuration with all native contacts between the monomer and fibril intact, i.e., the bound state. We then restrain the initial contact (as well as the heptamer fibril), and simulate the system at very high temperature at which the unrestrained contacts break. The resulting trajectory will

<sup>3</sup> By complexity of the transition manifold, we mean its dimension (which can vary locally), the number of sections or “arms”, the number of junctions etc.

explore all of the admissible state space, but no bonds will be formed due to the high temperature. The same technique has previously been applied [19] to generate the “milestone states” for a Markov model analysis of the A $\beta_{16-22}$  amyloid. Like in [19], we used a temperature of 1000 K for the heat sampling, but were able to reduce the simulation length from 50 ns to 20 ns, due to the smaller size of NFGAILS.

From the resulting trajectory, we then sample the desired number  $n$  of starting points, separately for each of the two docking contacts. As the subsampling method, we apply the k-means clustering algorithm with  $k = n$  to the high-temperature trajectory. Note, however, that the purpose of this step is not to find clusters in the trajectory, but instead to exploit the fact that the centroids generated by k-means are evenly spaced across the whole data range. This generates starting points that more uniformly cover the admissible state space compared to, for example, simple random subsampling. This trick of using k-means as a subsampling method is also commonly used in the construction of Markov state models [29].

For the number of starting points, we found that choosing  $n = 192$  leads to a clear image of the transition manifold in the latter embeddings. For longer peptides, with more complex transition pathways lined with non-native intermediate states, this number will grow accordingly.

### 3.2 Parallel simulation

In the next step, the transition densities  $p^t(x_k, \cdot)$  associated with each test point  $x_k$  need to be approximated by Monte Carlo sampling. The number of samples required to approximate the density up to a given error tolerance hereby scales with the variance of the density [30]. This variance will be small, as  $p^t(x_k, \cdot)$  is non-zero only in a small portion of state space (recall that the simulation time  $t$  is only long enough for the system to equilibrate locally). This holds independently of the system size, and hence,  $M$  is essentially independent of the peptide length.

As there is no practical way for us to estimate this variance prior to sampling, we will justify our choice a posteriori: if a clear low-dimensional structure is visible in the final embedding, the number of samples has been sufficient; otherwise, more samples have to be created. We will see that  $M = 32$  samples produce a reasonably clear embedding of the transition manifold.

As explained in Sect. 2.2, the parameter  $t$  must fall between the fast and slow timescales. The estimation of these time scales is the only step in our algorithm that requires (limited) expert chemical knowledge. We can expect the elastic bond- and valence-angle vibrations to belong to the fast process and be irrelevant for the locking dynamics. The equilibration of these vibrations occurs on the picosecond time scale. Moreover, the residual side-chains may contain quickly equilibrating torsion angle rotations, which fall on time scales of a few hundred picoseconds.

The slow processes on the other hand will consist of the backbone configurational changes that are associated with the formation of the remaining native contacts. In [19], the longest formation time of a single native contacts in the A $\beta_{16-22}$  amyloid has been found to be on the order of 6 ns. As NFGAILS and A $\beta_{16-22}$  are of comparable size, we take 6 ns as an estimate for the slow timescale. In conclusion,  $t$  should be chosen on a timescale of several hundred picoseconds. To exactly characterize the slow and fast degrees of freedom, we will perform our experiments for  $t = 0.1$  ns,  $t = 0.4$  ns, and  $t = 1$  ns, and compare the results.

The sampling is now realized by performing  $M = 32$  MD simulations for each of the  $n = 192$  test points, each simulation with different random momenta and a different random seed on the heat bath. Hence, overall,  $n \cdot M = 6144$  simulations need to be performed for each of the two initial contacts we consider. Simulations were performed on a 1536 core compute cluster (32 Intel Xeon 9242 CPUs) using the Gromacs molecular dynamics package [31], which allows easy parallelization of multiple runs of the same system via the `multidir` option. The overall runtime for one contact was 14 h. The resulting GROMACS structure files of the simulation end points (for the three lag times mentioned above) are available in the SI.

### 3.3 Transition manifold analysis

In this section, we describe the various steps of the transition manifold analysis that are performed on the simulation data. The transition manifold data analysis was performed using the special-purpose pyTMRC (Python Transition Manifold Reaction Coordinate) package [32]. The completion time for all the steps described in this section was less than 5 min on a 4-core laptop. Two Jupyter notebooks, implementing the analysis for the LEU-PHE and the PHE-LEU initial contact, respectively, can be found in the SI. To reproduce our results, download the pyTMRC package, download and extract the end point data, and execute all cells in the notebooks.

#### Pair-wise distances

In a first step, the samples are used to estimate the relative position of the transition densities to each other (in density space), i.e., computation of the distance matrix  $D \in \mathbb{R}_+^{n \times n}$ . For the statistical distance (called  $d$  in Sect. 2.2), we use the maximum mean discrepancy (MMD) [33], which, as the name suggests, measures the discrepancy between two densities by computing the mean of a class of test functions applied to the densities, and choosing the maximum distance between the means. More precisely, we define the distance  $d$  as

$$d(p^t(x_i, \cdot), p^t(x_j, \cdot)) := \sup_{f \in \mathcal{F}} \left| \mathbb{E}_{x \sim p^t(x_i, \cdot)}[f(x)] - \mathbb{E}_{x \sim p^t(x_j, \cdot)}[f(x)] \right|,$$

where the class of test functions  $f$  is generated by the so-called kernel function  $k : \mathbb{R}^{3N} \times \mathbb{R}^{3N} \rightarrow \mathbb{R}$ .

$$\mathcal{F} = \text{span} \{k(x, \cdot), x \in \mathbb{R}^{3d}\}.$$

For the kernel  $k$ , we use a Gaussian kernel of bandwidth  $\sigma = 5000$ . The bandwidth was optimized manually to produce the clearest image of the transition manifold under the MDS embedding (see the next section). The MMD has been shown to both analytically and numerically preserve the distance structure of the transition manifold [34]. Moreover, its estimation from samples of the compared densities is straight-forward.

### Euclidean embedding

To visualize the low-dimensional structure of the transition manifold that is encoded in  $D$ , we use the *multi-dimensional scaling* (MDS) algorithm [27, 35]. MDS constructs a set of  $n$  points in Euclidean space of selectable dimension (in our case, two-dimensional), so that the pair-wise distances between those points approximate  $D$  optimally. More precisely, MDS implicitly constructs an embedding of the densities, i.e., a map  $\mathcal{E} : L^1(\mathbb{R}^{3N}) \rightarrow \mathbb{R}^2$

$$\mathcal{E} : p^t(x_i, \cdot) \mapsto z_i \in \mathbb{R}^2, \quad i = 1, \dots, n,$$

so that the Euclidean distances between the embedded points, i.e.,  $\|z_i - z_j\|_2$ , optimally approximate the distance  $D_{ij}$ , for all pairs  $i, j = 1, \dots, n$ . Note that domain of  $\mathcal{E}$  is the infinite-dimensional space of absolutely integrable functions  $L^1(\mathbb{R}^{3N})$ , which includes probability densities. The points  $z_i$  then serve as the Euclidean representation of the densities  $p^t(x_i, \cdot)$ .

We specifically use the implementation of MDS provided by the Python package Scikit-learn [36]. Besides the distance matrix  $D$ , it does not require additional input parameters.

### Reaction coordinate computation

Next, we seek the “best” one-dimensional parametrization of the low-dimensional structure encoded in  $D$ . Pulled back onto the starting points, this will then become our final reaction coordinate. Again, we have multiple options in choosing the error metric. For preserving the distances in  $D$  directly, the one-dimensional MDS embedding gives the optimal result. However, due to its higher robustness to outliers and good performance in previous computations [21], we here use the diffusion maps method. Its parametrization optimally preserves the so-called *diffusion distance* between the points underlying the matrix  $D$ , which characterizes closeness by a high transition probability in an artificially constructed Markov jump process between the points (not to be confused with the original molecular dynamical process). This process, a discretized heat diffusion, contains a scale parameter  $\tau$  controlling the velocity of the diffusion, which we choose as  $\tau = 20$

(optimized manually to achieve an even parametrization of the structure observed in the MDS embedding).

### Shortest locking pathway

Finally, we discuss how the transition manifold embedding can be used to identify transition pathways and artificial trajectories between two states  $x_A$  and  $x_B$  on the transition manifold. There is no single, universally accepted concept of an “optimal” transition pathway between two states, and many proposed definitions with differing objectives and physical interpretations exist [37–39]. The TMF proposes another such pathway, namely the *geodesic* between  $p^t(x_A, \cdot)$  and  $p^t(x_B, \cdot)$  on the transition manifold  $\mathbb{M}$ . This is the shortest differentiable curve  $\Gamma$  in the metric space  $L^1$  that starts in  $p^t(x_A, \cdot)$ , ends in  $p^t(x_B, \cdot)$ , and does not leave  $\mathbb{M}$ . As each point  $p^t(x, \cdot) \in \mathbb{M}$  corresponds to exactly one starting point  $x \in \mathbb{R}^{3d}$ , we can “pull back”  $\Gamma$  to a “traditional” transition pathway  $\gamma$  in state space by setting  $\gamma(x) := \Gamma(p^t(x, \cdot))$ . Note that, while  $\gamma$  has a clear interpretation within the TMF, its interpretation in terms of more intuitive dynamical concepts such as transition probabilities or minimum energy pathways is still outstanding. For further discussion on the link between the transition manifolds and transition path theory, see [21].

As our data consist only of discrete samples close to  $\mathbb{M}$ , we take a heuristic approach for the numerical computation of  $\gamma$ . We construct a weighted, complete graph  $G = (V, E, W)$  with nodes  $V = \{x_1, \dots, x_n\}$  and edges  $E = \{(x_i, x_j) \mid i, j = 1, \dots, n\}$ . For the weight matrix  $W \in \mathbb{R}^{n \times n}$ , we take the squared maximum mean discrepancy

$$W_{ij} = D_{ij}^2.$$

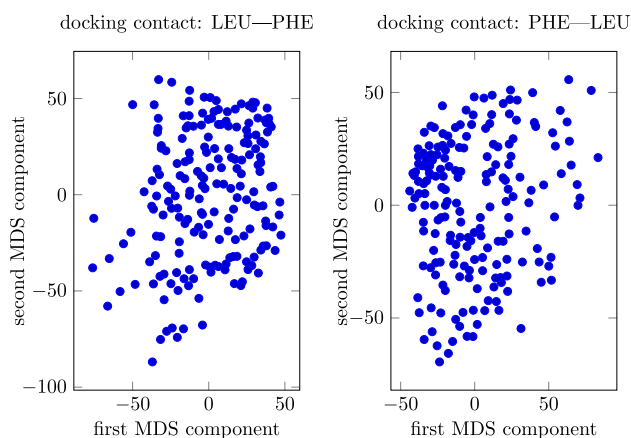
The squaring compresses small, local distances, and further increases already long distances. The discrete shortest path in  $G$  between the nodes  $x_A, x_B$  thus tends to take small steps instead of large jumps, and thus is encouraged to follow the transition manifold. Thus, we can take this discrete shortest path as a heuristic approximation of  $\gamma$ .

## 4 Results and discussion

### 4.1 Starting points

We first examine the distribution of the 192 (for each initial contact) starting points  $x_k$  in state space. This does *not* utilize the information of the parallel simulations, but only illustrates the result of the heat sampling and k-means clustering strategy.

For this, we apply MDS to starting points  $x_k \in \mathbb{R}^{159}$  themselves (not the densities  $p^t(x_k, \cdot)$ ), i.e., seek points



**Fig. 4** Two-dimensional MDS embedding of the sampled start points. The slightly denser region visible for the PHE–ILE contact corresponds to configurations close to the final locked state. Besides that, no structure that could indicate the essential dynamics of the locking process can be recognized

$z_k \in \mathbb{R}^2$ , such that

$$\|x_i - x_j\|_2 \approx \|z_i - z_j\|_2, \quad i, j = 1, \dots, n.$$

The points optimally fulfilling this requirement are shown in Fig. 4. We see that the embedded points are spaced quite evenly, and no clear low-dimensional structure is visible (with the exception of one slight clustering in the case of the PHE–ILE contact). This indicates that

1. the starting points are spaced evenly in the admissible state space (by Euclidean distance), and
2. the identification of transition pathways or other dynamical features based purely on the location of the embedded starting points is *not* possible.

An even distribution of the starting points is important for the subsequent analysis, as clusters of starting points would lead to artificial clusters on the embedded transition manifold that do not contain dynamical information, so this represents an optimal situation. Recall however that any statement about the high-dimensional starting points based on their relative Euclidean distance should be handled with care.

Gromacs structure files containing the starting points are provided in the SI, so that the reader may inspect them visually using molecular rendering software.

## 4.2 MDS-embedded transition manifold

Figure 5 shows the MDS embedding of the transition densities  $p^t(x_k, \cdot)$ ,  $i = k, \dots, 192$ , i.e., the points  $z_k$  from the *Euclidean embedding* subsection of Sect. 3.3. The embeddings are shown for both the two locking processes associated with the LEU–PHE and the PHE–LEU initial contact, and for the three lag times

$t = 0.1$  ns,  $t = 0.4$  ns and  $t = 1$  ns. The relative mean deviation of the pair-wise distances between the embedded points from the distances in  $D$  ranges from 0.2% to 3.5%. The illustrations therefore give a faithful impression of the true relative locations of the transition densities.

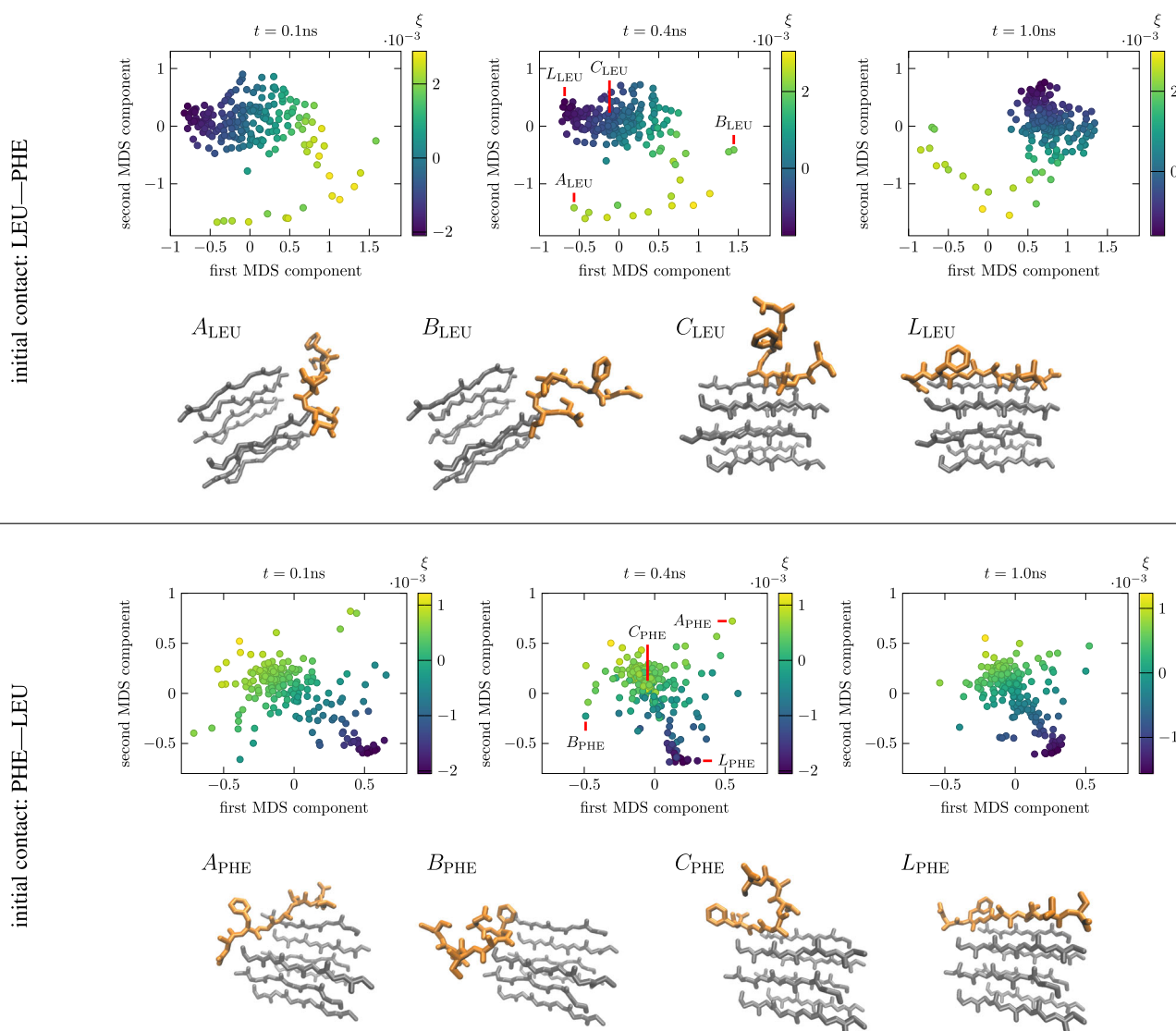
In each of the six plots, we see structures consisting of certain low-dimensional extensions as well as clusters of points. These structures are Euclidean representations of the respective transition manifold. Even though the structures are clearly not “manifolds” in the strict topological sense, we will continue to refer to them as such.

We observe that for both contacts, the high-level structure of the transition manifold is similar across all lag times. From that we conclude that the least important fast degrees of freedom are already equilibrated after 100 picoseconds, and that all remaining degrees of freedom are all relevant for the locking process.

We also observe that the transition manifold gets more compact with increasing  $t$ , i.e., the distance between the embedded points decreases. This is consistent with the fact that for increasing  $t$ , the distances between the transition densities also decrease, as they equilibrate toward the system’s invariant density for  $t \rightarrow \infty$ . Importantly, the structures seem to lose some fine detail at  $t = 1.0$  ns, such as the branch-like extensions perpendicular to the structure’s main axis. We will see soon that these branches are associated with important folding motion of the backbone, and hence, a choice of  $t = 1.0$  ns should be considered too long. For the following detailed analysis, we therefore only consider the embedding at lag time  $t = 0.4$  ns.

Next, we investigate certain key points of the manifold. We denote the end points of the branch-like extensions by  $A_{LEU}$ ,  $B_{LEU}$  for the LEU–ASN and  $A_{PHE}$ ,  $B_{PHE}$  for the PHE–ILE initial contact. We denote by  $C_{LEU}$  and  $C_{PHE}$ , a point in the center of the respective main cluster. Finally, we denote for LEU–PHE the point at the far end of the main cluster by  $L_{LEU}$ , and for PHE–LEU the point at the far end of the small, secondary cluster by  $L_{PHE}$ . Their molecular configurations are shown in Fig. 5 beneath the embedding plots. The exact choice of all these points is of minor importance, as their neighbors all have very similar structure and essentially differ only in their side-chain configuration.

$L_{LEU}$  and  $L_{PHE}$  correspond to the respective locked state, with all three native contacts intact. The points  $A_{LEU}$ ,  $B_{LEU}$  and  $A_{PHE}$ ,  $B_{PHE}$  are furthest away from  $L_{LEU}$  and  $L_{PHE}$ , respectively, in the topology of the transition manifold. Hence, they can be interpreted as the points that are “dynamically maximally different” from  $L_{LEU}/L_{PHE}$ . Indeed, as we see by looking at their molecular configurations, the monomer is maximally outstretched and pointing away from the fibril in different directions.  $C_{LEU}$  and  $C_{PHE}$  correspond to configurations where the second contact ALA–GLY already has been formed. Their central position on the transition manifold therefore indicates their role of “intermediate points” of the locking process. Overall, the dynamical role of individual points can be inferred very well from their (relative) position in the embedding space.



**Fig. 5** Scatter plots: Two-dimensional MDS embedding of the transition densities for the two initial contacts and the three lag times. The coloring represents the one-dimensional reaction coordinate computed by diffusion maps. Molecular structures: configurations of certain selected extremal points of the embedded structure

### 4.3 Shortest locking pathways

Figure 6 shows the shortest pathways (as defined in Sect. 3.3) between selected extremal states and the locked state for both initial contacts, specifically between  $A_{LEU}$  and  $L_{LEU}$  as well as between  $B_{PHE}$  and  $L_{PHE}$ . Pathways from the extremal points  $B_{LEU}$  and  $A_{PHE}$  are qualitatively similar, but have been omitted due to space limitations. The corresponding Gromacs files are however provided in the SI.

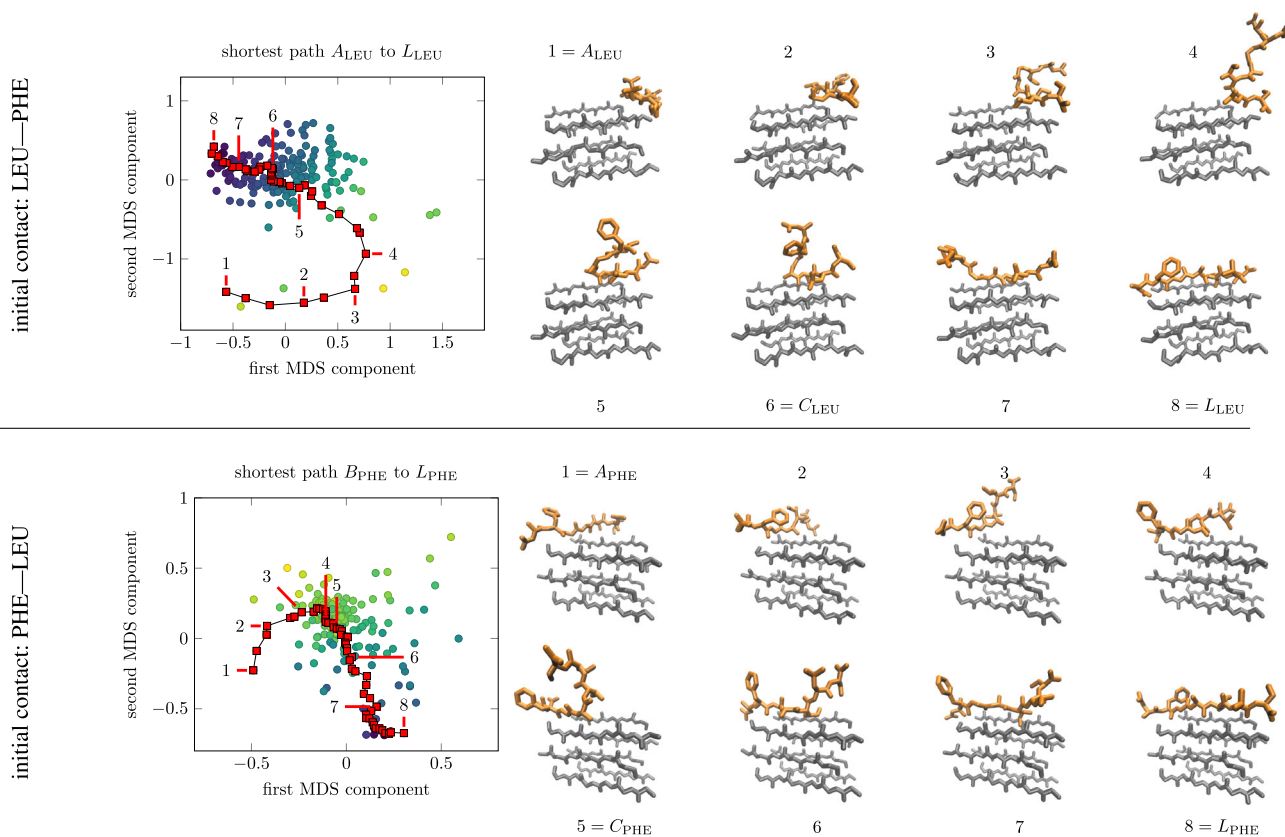
We see that the pathways indeed follow the low-dimensional structure identified by the MDS embedding. Checkpoints chosen equidistantly along the pathway show a continuous folding of the monomer onto the fibril core. Specifically, we first observe for both initial contacts first the straightening of the monomer, then the formation of the second contact ALA—GLY, and

finally the formation of the third contact. We emphasize that these “artificial locking trajectories” did not require the simulation of a full locking trajectory, but are “stitched together” from short, independent simulations, most of which are far from the locked state. In that, their construction is similar to replica exchange MD [40], a well-known technique for accelerating barrier crossing events in complex systems.

### 4.4 Reaction coordinates

The color gradient in Fig. 5 indicates the value of the diffusion maps reaction coordinate  $\xi$  in the start points. While the structure is clearly not one-dimensional,  $\xi$  can be seen as the “best possible one-dimensional embedding” of the transition manifold, i.e., the reaction coordinate that preserves the long time scales best





**Fig. 6** Shortest pathways on the transition manifold from  $A_{LEU}$  to  $L_{LEU}$  (top) and from  $B_{PHE}$  to  $L_{PHE}$  (bottom). Renderings of the molecular structure illustrate the locking progress along the pathway

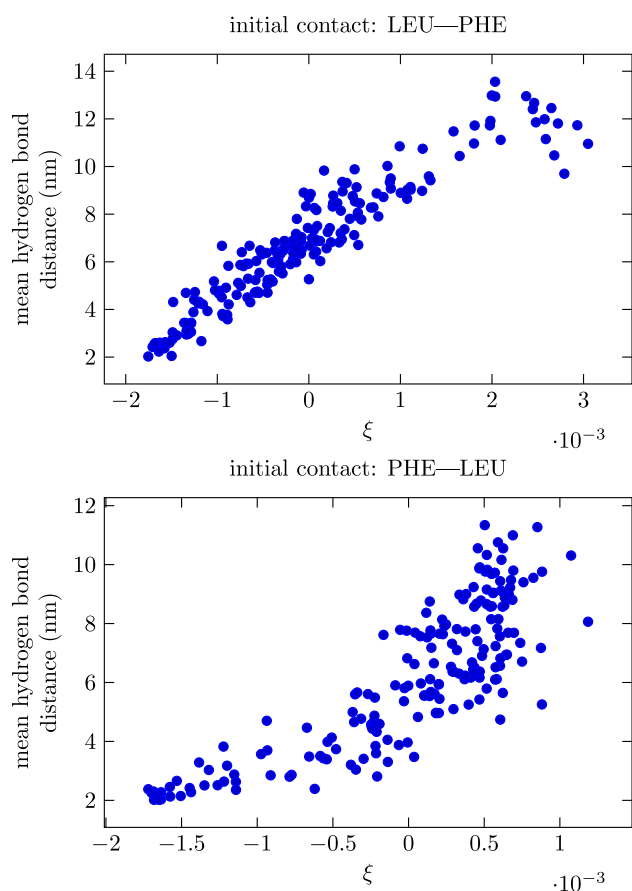
out of all one-dimensional reaction coordinates. We see that for both initial contacts,  $\xi$  takes its minimum in the locked state  $L$ , and is close to its maximum in both the branch end points  $A$  and  $B$ . Hence, this one-dimensional reaction coordinate cannot distinguish the two locking pathways identified in the previous subsection, but instead seems to indicate the (state space) distance to  $L$ . Indeed, Fig. 7 indicates a strong correlation of  $\xi$  to the mean hydrogen-bond distance, i.e., the mean distance of the heavy atoms involved in the three native contacts between monomer and fibril. This is a well-known heuristic reaction coordinate used in a wide variety of biochemical reactions [41]. Our result is therefore able to verify the mean hydrogen-bond distance as a dynamically optimal reaction coordinate for the amyloid locking process.

## 5 Conclusions and future work

Application of the transition manifold system analysis framework to NFGAILS locking phase has revealed important microscopic details about the monomer-fibril binding process. Using statistical sampling and embedding techniques, we were able to learn the structure of the effective dynamics directly in transition

probability space, and could extract from that structure dominant reaction pathways and optimal reaction coordinates. The different reaction pathways reveal a common, chemically plausible binding mechanism that specifies the order in which the native contacts are formed. This is consistent with the observed strong correlation of our optimal reaction coordinate to the mean native hydrogen-bond distance.

While our results are mostly qualitative in nature, they form the basis for further quantitative investigations. We are planning to compute the free energy profile along the identified reaction pathways, and with it the binding and unbinding rates. This will however require longer MD simulations that are able to overcome the energy barriers along the pathways. Also, we will construct a comprehensive one-dimensional reduced model of the locking process in the state space of our reaction coordinate. Specifically, we will estimate the parameters of a generalized Langevin equation using a novel data-driven technique that has already been applied successfully to various other biomolecular systems [41–44]. Finally, we will investigate to what extent the transition manifold framework can be applied to the other stages of the amyloid fibrillation process, such as the docking phase or the early condensation phase.



**Fig. 7** Comparison of the computed transition manifold reaction coordinate ( $\xi$ ) to the mean hydrogen-bond distance. The almost linear dependency indicates a strong correlation between the parameters (especially for the LEU–PHE initial contact)

**Acknowledgements** This research has been funded by Deutsche Forschungsgemeinschaft (DFG) through grant CRC 1114 “Scaling Cascades in Complex Systems”, Project Number 235221301, Project B03 “Multilevel coarse graining of multiscale problems”.

The authors thank Philip Loche and Cihan Ayaz for their help with the molecular dynamical simulations as well as helpful discussions and input. The authors thank Vedat Durmaz for providing the initial fibril configuration files.

## Author contributions

AB proposed the initial idea. AB and CS designed the methodology. JM contributed chemical insight and helped designing Figs. 2 and 3. AB set up and performed the computations and analyzed the data. AB wrote the paper and created the figures. AB, RN, and CS discussed the results and refined the paper. RN, BK, and CS supervised the project and acquired the funding.

**Funding** Open Access funding enabled and organized by Projekt DEAL.

**Data Availability Statement** This manuscript has data included as electronic supplementary material. The online version of this article contains supplementary material, which is available to authorized users.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

## References

1. C.M. Dobson, Protein misfolding, evolution and disease. *Trends Biochem. Sci.* **24**(9), 329–332 (1999)
2. J.W. Kelly, The alternative conformations of amyloidogenic proteins and their multi-step assembly pathways. *Curr. Opin. Struct. Biol.* **8**(1), 101–106 (1998)
3. H. John, J.S. Dennis, The amyloid hypothesis of Alzheimer’s disease: progress and problems on the road to therapeutics. *Science* **297**(5580), 353–356 (2002)
4. R. Kaye, E. Head, J.L. Thompson, T.M. McIntire, S.C. Milton, C.W. Cotman, C.G. Glabe, Common structure of soluble amyloid oligomers implies common mechanism of pathogenesis. *Science* **300**(5618), 486–489 (2003)
5. M. Bucciantini, G. Calloni, F. Chiti, L. Formigli, D. Nosi, C.M. Dobson, M. Stefani, Prefibrillar amyloid protein aggregates share common features of cytotoxicity. *J. Biol. Chem.* **279**(30), 31374–31382 (2004)
6. G. Grasso, A. Danani, Molecular simulations of amyloid beta assemblies. *Adv. Phys. X* **5**(1), 1770627 (2020)
7. C. Wu, H. Lei, Y. Duan, Formation of partially ordered oligomers of amyloidogenic hexapeptide (NFGAIL) in aqueous solution observed in molecular dynamics simulations. *Biophys. J.* **87**(5), 3000–3009 (2004)
8. H.-H. Tsai, M. Reches, C.-J. Tsai, K. Gunasekaran, E. Gazit, R. Nussinov, Energy landscape of amyloidogenic peptide oligomerization by parallel-tempering molecular dynamics simulation: Significant role of Asn ladder. *Proc. Natl. Acad. Sci.* **102**(23), 8174–8179 (2005)
9. C. Wu, H. Lei, Y. Duan, The role of Phe in the formation of well-ordered oligomers of amyloidogenic hexapeptide (NFGAIL) observed in molecular dynamics simulations with explicit solvent. *Biophys. J.* **88**(4), 2897–2906 (2005)
10. U. Sengupta, M. Carballo-Pacheco, B. Strodel, Automated Markov state models for molecular dynamics

- simulations of aggregation and self-assembly. *J. Chem. Phys.* **150**(11), 115101 (2019)
11. W. Hoffmann, K. Folmert, J. Moschner, X. Huang, H. von Berlepsch, B. Koks, M.T. Bowers, G. von Helden, K. Pagel, NFGAIL amyloid oligomers: the onset of beta-sheet formation and the mechanism for fibril formation. *J. Am. Chem. Soc.* **140**(1), 244–249 (2018)
  12. J. Moschner, V. Stulberg, R. Fernandes, S. Huhmann, J. Leppkes, B. Koks, Approaches to obtaining fluorinated  $\alpha$ -Amino acids. *Chem. Rev.* **119**(18), 10718–10801 (2019)
  13. S. Chowdhary, J. Moschner, D.J. Mikolajczak, M. Becker, A.F. Thunemann, C. Kastner, D. Klemczak, A.K. Stegemann, C. Bottcher, P. Metrangolo, R.R. Netz, The impact of halogenated phenylalanine derivatives on NFGAIL amyloid formation. *ChemBioChem* **21**(24), 3544 (2020)
  14. T. Gurry, C.M. Stultz, Mechanism of amyloid- $\beta$  fibril elongation. *Biochemistry* **53**(44), 6981–6991 (2014)
  15. A.J. Dear, G. Meisl, T.C.T. Michaels, M.R. Zimmermann, S. Linse, T.P. J. Knowles, The catalytic nature of protein aggregation. *J. Chem. Phys.* **152**(4), 045101 (2020)
  16. W.P. Esler, E.R. Stimson, J.M. Jennings, H.V. Vinters, J.R. Ghilardi, J.P. Lee, P.W. Mantyh, J.E. Maggio, Alzheimer's disease amyloid propagation by a template-dependent dock-lock mechanism. *Biochemistry* **39**(21), 6288–6295 (2000)
  17. P.H. Nguyen, M.S. Li, G. Stock, J.E. Straub, D. Thirumalai, Monomer adds to preformed structured oligomers of A $\beta$ -peptides by a two-stage dock-lock mechanism. *PNAS* **104**(1), 111–116 (2007)
  18. M. Schor, A.S.J.S. Mey, F. Noé, C.E. MacPhee, Shedding light on the dock-lock mechanism in amyloid fibril growth using Markov state models. *J. Phys. Chem. Lett.* **6**(6), 1076–1081 (2015)
  19. Z. Jia, A. Beugelsdijk, J. Chen, J.D. Schmit, The Levinthal problem in amyloid aggregation: sampling of a flat reaction space. *J. Phys. Chem. B* **121**(7), 1576–1586 (2017)
  20. A. Bittracher, P. Koltai, S. Klus, R. Banisch, M. Dellnitz, C. Schütte, Transition manifolds of complex metastable systems: theory and data-driven computation of effective dynamics. *J. Nonlinear Sci.* **28**(2), 471–512 (2017)
  21. A. Bittracher, R. Banisch, C. Schütte, Data-driven computation of molecular reaction coordinates. *J. Chem. Phys.* **149**(15), 154103 (2018)
  22. M.R. Sawaya, S. Sambashivan, R. Nelson, M.I. Ivanova, S.A. Sievers, M.I. Apostol, M.J. Thompson, M. Balbirnie, J.J.W. Wiltzius, H.T. McFarlane, A.O. Madsen, C. Riek, D. Eisenberg, Atomic structures of amyloid cross- $\beta$  spines reveal varied steric zippers. *Nature* **447**(7143), 453–457 (2007)
  23. R. Akter, P. Cao, H. Noor, Z. Ridgway, L.-H. Tu, H. Wang, A.G. Wong, X. Zhang, A. Abedini, A.M. Schmidt, D.P. Raleigh, Islet amyloid polypeptide: structure, function, and pathophysiology. *J. Diabetes Res* **2016**, 2798269 (2016)
  24. A.B. Soriaga, S. Sangwan, R. Macdonald, M.R. Sawaya, D. Eisenberg, Crystal structures of IAPP amyloidogenic segments reveal a novel packing motif of out-of-register beta sheets. *J. Phys. Chem. B* **120**(26), 5810–5816 (2016)
  25. J.D. Schmit, Kinetic theory of amyloid fibril templating. *J. Chem. Phys.* **138**(18), 185102 (2013)
  26. R.R. Coifman, S. Lafon, Diffusion maps. *Appl. Comput. Harmonic Anal.* **21**(1), 5–30 (2006)
  27. F.W. Young, *Multidimensional Scaling: History, Theory, and Applications* (Psychology Press, Hove, 2013)
  28. M. Ledoux, The concentration of measure phenomenon. No. 89. American Mathematical Society (2005)
  29. M.K. Scherer, B. Trendelkamp-Schroer, F. Paul, G. Pérez-Hernández, M. Hoffmann, N. Plattner, C. Wehmeyer, J.-H. Prinz, F. Noé, PyEMMA 2: a software package for estimation, validation, and analysis of markov models. *J. Chem. Theory Comput.* **11**(11), 5525–5542 (2015)
  30. M.H. Kalos, P.A. Whitlock, *Monte Carlo Methods* (Wiley, Hoboken, 2009)
  31. H.J.C. Berendsen, D. van der Spoel, R. van Drunen, GROMACS: a message-passing parallel molecular dynamics implementation. *Comput. Phys. Commun.* **91**(1), 43–56 (1995)
  32. A. Bittracher, M. Mollenhauer, PyTMRC. <https://github.com/abittracher/pytmrc>, commit 5b0b52e (2020)
  33. A. Gretton, K. Borgwardt, M. Rasch, B. Schölkopf, A. Smola, A kernel two-sample test. *J. Mach. Learn. Res.* **13**, 723–773 (2012)
  34. A. Bittracher, S. Klus, B. Hamzi, P. Koltai, C. Schütte, Dimensionality reduction of complex metastable systems via kernel embeddings of transition manifolds. *J. Nonlinear Sci.* **31**(1), 3 (2020)
  35. J.B. Kruskal, Nonmetric multidimensional scaling: a numerical method. *Psychometrika* **29**(2), 115–129 (1964)
  36. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: machine learning in python. *J. Mach. Learn. Res.* **12**(85), 2825–2830 (2011)
  37. K. Fukui, Formulation of the reaction coordinate. *J. Phys. Chem.* **74**(23), 4161–4163 (1970)
  38. L. Maragliano, A. Fischer, E. Vanden-Eijnden, G. Ciccotti, String method in collective variables: minimum free energy paths and isocommittor surfaces. *J. Chem. Phys.* **125**(2), 024106 (2006)
  39. E. Vanden-Eijnden, M. Venturoli, Revisiting the finite temperature string method for the calculation of reaction tubes and free energies. *J. Chem. Phys.* **130**(19), 194103 (2009)
  40. P. Liu, B. Kim, R.A. Friesner, B.J. Berne, Replica exchange with solute tempering: a method for sampling biological systems in explicit water. *PNAS* **102**(39), 13749–13754 (2005)
  41. Cihan A, Lucas T, Florian NB, Julian K, Jan OD, Roland RN, Non-Markovian modeling of protein folding. Submitted (2021)
  42. Julian Kappler, Jan O. Daldrop, Florian N. Brünig, Moritz D. Boehle, Roland R. Netz, Memory-induced acceleration and slowdown of barrier crossing. *J. Chem. Phys.* **148**(1), 014903 (2018)

43. J. Kappler, F. Noé, R.R. Netz, Cyclization and relaxation dynamics of finite-length collapsed self-avoiding polymers. *Phys. Rev. Lett.* **122**(6), 067801 (2019)
44. B. Kowalik, J.O. Daldrop, J. Kappler, J.C.F. Schulz, A. Schlaich, R.R. Netz, Memory-kernel extraction for different molecular solutes in solvents of varying viscosity in confinement. *Phys. Rev. E* **100**(1), 012126 (2019)