The Journal of Chemical Physics

Automated Markov state models for molecular dynamics simulations of aggregation and self-assembly

Cite as: J. Chem. Phys. **150**, 115101 (2019); https://doi.org/10.1063/1.5083915
Submitted: 01 December 2018 . Accepted: 01 February 2019 . Published Online: 15 March 2019

Ushnish Sengupta, Martín Carballo-Pacheco, and Birgit Strodel

COLLECTIONS

Paper published as part of the special topic on Markov Models of Molecular Kinetics

Note: This article is part of the Special Topic “Markov Models of Molecular Kinetics” in J. Chem. Phys.

F This paper was selected as Featured

View Online Export Citation CrossMark

ARTICLES YOU MAY BE INTERESTED IN

Markov models of molecular kinetics: Generation and validation
The Journal of Chemical Physics **134**, 174105 (2011); https://doi.org/10.1063/1.3565032

Perspective: Identification of collective variables and metastable states of protein dynamics
The Journal of Chemical Physics **149**, 150901 (2018); https://doi.org/10.1063/1.5049637

Unsupervised machine learning in atomistic simulations, between predictions and understanding
The Journal of Chemical Physics **150**, 150901 (2019); https://doi.org/10.1063/1.5091842

AIP Publishing

Automated Markov state models for molecular dynamics simulations of aggregation and self-assembly

Cite as: J. Chem. Phys. **150**, 115101 (2019); doi: 10.1063/1.5083915
Submitted: 1 December 2018 • Accepted: 1 February 2019 •
Published Online: 15 March 2019

View Online Export Citation CrossMark

Ushnish Sengupta,[1,2] Martín Carballo-Pacheco,[1,3,4] and Birgit Strodel[1,5,a)]

AFFILIATIONS

[1] Institute of Complex Systems: Structural Biochemistry (ICS-6), Forschungszentrum Jülich, 52425 Jülich, Germany
[2] Department of Engineering, University of Cambridge, Cambridge CB2 1PZ, United Kingdom
[3] AICES Graduate School, RWTH Aachen University, Schinkelstraße 2, 52062 Aachen, Germany
[4] School of Physics and Astronomy, University of Edinburgh, Peter Guthrie Tait Road, Edinburgh EH9 3FD, United Kingdom
[5] Institute of Theoretical and Computational Chemistry, Heinrich Heine University Düsseldorf, 40225 Düsseldorf, Germany

Note: This article is part of the Special Topic "Markov Models of Molecular Kinetics" in J. Chem. Phys.
[a)] **Author to whom correspondence should be addressed:** b.strodel@fz-juelich.de

ABSTRACT

Markov state models have become popular in the computational biochemistry and biophysics communities as a technique for identifying stationary and kinetic information of protein dynamics from molecular dynamics simulation data. In this paper, we extend the applicability of automated Markov state modeling to simulation data of molecular self-assembly and aggregation by constructing collective coordinates from molecular descriptors that are invariant to permutations of molecular indexing. Understanding molecular self-assembly is of critical importance if we want to deepen our understanding of neurodegenerative diseases where the aggregation of misfolded or disordered proteins is thought to be the main culprit. As a proof of principle, we demonstrate our Markov state model technique on simulations of the KFFE peptide, a subsequence of Alzheimer's amyloid-β peptide and one of the smallest peptides known to aggregate into amyloid fibrils *in vitro*. We investigate the different stages of aggregation up to tetramerization and show that the Markov state models clearly map out the different aggregation pathways. Of note is that disordered and β-sheet oligomers do not interconvert, leading to separate pathways for their formation. This suggests that amyloid aggregation of KFFE occurs via ordered aggregates from the very beginning. The code developed here is freely available as a Jupyter notebook called TICAgg, which can be used for the automated analysis of any self-assembling molecular system, protein, or otherwise.

I. INTRODUCTION

Molecular dynamics (MD) simulations have become a fundamental tool for understanding the behavior of both biological and non-biological molecules at full atomic resolution.[1] However, extracting useful information from the increasing amount of MD data generated by powerful, state-of-the-art supercomputers is a big data challenge, which necessitates sophisticated post-processing techniques. Markov state models (MSMs) have recently gained traction in the computational biochemistry and physics community as a technique that can help us do precisely this. MSMs are network models that encode the system dynamics in a states-and-rates format; i.e., the molecular system at a given instant can exist in one amongst many possible states and it has a fixed probability of transitioning to other states, including itself, within a particular time interval. Notably, MSMs can be useful and versatile as a data analysis tool: they can be used for calculating quantities of interest that can be compared against experimental observables,[2–5] quantifying uncertainties in predictions,[6,7] gaining an intuitive understanding of the system,[8,9] driving efficient simulations by combining them with adaptive sampling,[10] and utilizing data from multiple short trajectories.[11,12]

In the past few years, great progress has been made toward automating the generation of MSMs. The release of user-friendly MSM libraries like MSMbuilder[13] and PyEMMA[14] has popularized MSMs amongst practitioners as a quick, unbiased, and informative post-processing technique for MD data. MSMs have been used to analyze a diverse range of biophysical problems including protein folding,[8,11] protein loop motion,[15] allosteric regulation,[16] protein-ligand association,[17] protein-protein association,[18,19] and monomer addition to amyloid fibrils.[20] However, a naive application of the typical automated MSM workflow to simulations of molecular self-assembly or aggregation would cause problems in the collective coordinate construction. The source of this problem lies in the existence of degenerate states in such simulations, i.e., identical oligomer conformations differing only by permutations of chemically identical molecules.[21] Since *a priori* knowledge of a system's reaction coordinates is typically unavailable, a crucial prerequisite for state decomposition in the automated MSM workflow is the construction of kinetically relevant collective coordinates from the data. This is typically performed by applying dimension reduction techniques such as time-lagged independent component analysis[22] (TICA) to generic internal coordinates such as interatomic distances.[23,24] This approach suits the philosophy of MSM automation because not only do manually chosen reaction coordinates require specialized knowledge about the system and user intervention, but in complicated free energy landscapes, they can often capture a limited amount of information.[25] However, simply applying this dimension reduction to intramolecular or intermolecular atomic distances in self-assembling systems leads to chemically identical, degenerate states occupying different positions in the free energy landscape. In this paper, we overcome this issue by introducing a modified TICA called TICAgg where the characterization of a molecule is invariant to the molecular indexing and thus immune to the degeneracy problem.

Molecular self-assembly or aggregation has received broad attention in the MD community because of its scientific relevance.[26–29] Pathological protein aggregation is a self-assembly problem that has received a lot of scrutiny from computational scientists as the aggregation of misfolded proteins is thought to be involved in the disease pathogenesis for diverse disorders, including Alzheimer's disease, Creutzfeldt-Jakob disease, Parkinson's disease, Huntington's disease, and type II diabetes.[30,31] The end product of this aggregation process is usually fibrils highly enriched in β-sheet content, termed amyloids. However, there is mounting evidence suggesting that it is not these end-stage amyloid fibrils but smaller, soluble oligomers formed at an earlier stage in the amyloid cascade that are principally responsible for the pathogenesis.[32–35] However, due to their high aggregation propensity and polydispersity, the preparation of appropriate samples of amyloid oligomers for high-resolution structural methods has proven to be difficult.[36] Modeling techniques such as MD simulations can help elucidate the mechanism of oligomer formation and identify metastable oligomer species[37] that can be new targets for the design of aggregation inhibitors.[38] Molecular self-assembly is of interest not just because of its role in disease: more often than not, molecular aggregates play important functional roles in biology. MD studies of such functional aggregates include simulations of bile salt aggregates which facilitate the metabolism of triglycerides,[39] a simplified model of virus capsid assembly,[40] coarse-grained MD simulations of the self-assembly of homotetrameric M2 channel protein from influenza A,[41] a minimal MD model of the formation of fibrin-like filament bundles,[42] simulations of aggregation and vesiculation of membrane proteins by curvature-mediated interactions,[43] and atomistic MD simulations of functional amyloid formation.[44] Nanotechnologists who seek to mimic nature in their designs and utilize self-assembly to create complex, useful materials have also turned to MD simulations as a possible source of design principles.[45]

Our study is not the first to construct MSMs for computational simulations of molecular self-assembly. However, previous work often did not include atomistic detail and relied on manual, coarse state decompositions where oligomers were differentiated based on distance cutoffs,[46,47] undirected graphs,[48] or asphericity parameters of aggregates.[49] Atomistic details were indeed considered in a recent study of the dimerization of the amyloid-β(1-40) peptide[50] where the authors extended principal component analysis (PCA) based clustering to resolve the degeneracy problem by retaining only the minimum of permutable inter-residue distances for constructing MSMs. While this approach works well for a dimer, it is unlikely to generalize to larger oligomers and systems composed of monomers and oligomers of different sizes, where retaining only the minimum of many permutable distances would disregard most of the data. Another possible approach would be the usage of atom-centric symmetry functions which are used in machine learning approaches to computational chemistry for taking into account the permutation of like atoms.[51] However, the evaluation of a large number of these functions for every frame of an MD trajectory would be cost-prohibitive. Moreover, the collective coordinates obtained from such an approach might also be hard to interpret. Our molecule-agnostic approach to Markov state modeling circumvents the degeneracy issue in dimension reduction, considers both inter- and intra-molecular atomistic details to characterize oligomers, and is broadly applicable to all varieties of self-assembly simulations: systems can contain any number of monomers, multiple oligomer-forming species, or even both oligomeric and non-oligomeric chemical species. Our publicly available Jupyter notebook TICAgg thus reduces the effort of analyzing aggregation simulations.

As a proof of concept, we apply our TICAagg-based Markov state modeling to study the aggregation of the KFFE peptide, a sequence derived from Alzheimer's amyloid-β peptide known to aggregate on its own into amyloid fibrils.[52] The motivation for studying short amyloidogenic protein fragments stems from the hypothesis that their aggregation pathways will capture the essential features of those of their parent proteins.[53] The tendency of KFFE to aggregate is a result of the hydrophobicity and β-sheet propensity of the two phenylalanine residues in the peptide core[54] and the oppositely charged lysine and glutamate residues at the two ends attracting each other. This causes KFFE peptides to preferentially line up against each other in an antiparallel fashion so that they can form salt bridges as well as reduce the exposure of hydrophobic residues to the solvent.[55] The self-assembly of this peptide has already been studied using MD simulations that employed either a coarse-grained description of the peptide,[56–58] an atomistic model in implicit solvent,[59] or explicit solvent.[60] In the current work, KFFE is our computationally cheap minimal and already well-studied model for more biologically relevant aggregating proteins in order to demonstrate the effectiveness of our Markov state modeling technique.

II. METHODS

A. Overview of the automated Markov state modeling workflow

Here, we briefly introduce Markov state modeling; for a more thorough description of the theory underlying the calculation of MSMs from MD data, the reader is referred to some recent review papers.[61–63] The typical MSM analysis is initiated by defining molecular descriptors or "features" (e.g., distances between atoms, dihedral angles, or contacts), from which collective coordinates will be constructed. These features are then computed for each frame in the simulation trajectories, thus transforming the Cartesian coordinate trajectories into feature vectors. The next step is to conduct a linear transformation on these feature vectors for dimension reduction using TICA,[23,24] which can identify a maximally slow subspace from the feature space by maximizing the autocorrelation of the reduced collective coordinates.[64] TICA can capture the slow, chemically relevant transitions in our system and is therefore preferable to the more commonly used PCA for the construction of kinetic models since the latter maximizes the variance in the reduced coordinates but pays no importance to kinetic information. Only the components corresponding to the largest autocorrelations are retained.

After obtaining a convenient low-dimensional representation of our data, we use k-means clustering[65] to decompose the free energy landscape into hundreds of discrete "microstates" such that each frame in our trajectories can be assigned to one of these microstates. The discretized trajectories thus obtained are used to estimate an MSM of the microstates. This model can be used to calculate quantities of interest; however, it is too granular to provide a simple, intuitive picture of the system dynamics. That is achieved by coarse-graining the MSM into a Hidden Markov Model (HMM) with a few metastable states, using robust Perron cluster analysis[66] (PCCA+). PCCA+ is a fuzzy version of the spectral algorithm for partitioning graphs that assigns each microstate a probability of belonging to a metastable macrostate. Finally, whether our model satisfies the Markovian assumptions can be verified with a Chapman-Kolmogorov test. The MSM workflow is graphically summarized as a flowchart in Fig. 1.

In this work, the Markov state models were built with a Jupyter notebook using PyEMMA.[14] The *mdtraj* library[67] was employed for handling protein coordinates, and the Jupyter widget *nglview* was used to integrate structural visualization into the notebook, enabling a smooth modeling workflow.

B. Modifying dimension reduction to deal with degeneracy in self-assembly

As discussed previously, the degeneracy of oligomeric states hinders the straightforward application of dimension reduction to inter-atomic distances of such systems. This problem is visualized in Fig. 2 for a toy system containing two identical molecules with four atoms each. The two distances shown by arrows are an example of a pair of permutable distances, i.e., distances that switch places if the two molecules are permuted. If we naively used these two distance values as our feature vectors and fed them into TICA, it might happen that at some later point of time in the simulations, these two chemically identical molecules switch positions. In that case, these distances will also switch their values. From a physicochemical point of view, the two conformations are equivalent since the molecules are indistinguishable. However, due to our choice of inter-atom distance-based features, these degenerate configurations would be represented in the feature vectors by different pairs of numbers, which differ by a permutation and, therefore, after applying TICA, by different points in the free energy landscape.

To avoid this, we use features that are invariant to molecular indexing as molecular indices of copies of the same

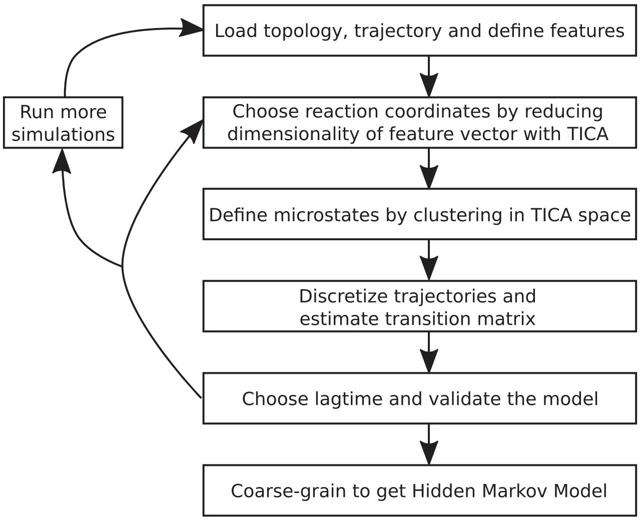
Load topology, trajectory and define features
Choose reaction coordinates by reducing dimensionality of feature vector with TICA
Define microstates by clustering in TICA space
Discretize trajectories and estimate transition matrix
Choose lagtime and validate the model
Coarse-grain to get Hidden Markov Model
Run more simulations

FIG. 1. The workflow for building MSMs from MD trajectories.

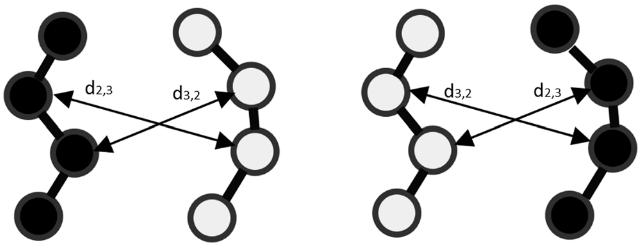

FIG. 2. Two chemically identical molecules switching positions does not lead to a physicochemical change in the system. However, due to the indexing of the molecules during a simulation, this leads to a permutation of the molecular descriptors as indicated here for the inter-atom distances $d_{2,3}$ and $d_{3,2}$. Without considering the possibility of permutation, the two identical dimers shown left and right would be identified as different conformations as $d_{2,3}$ and $d_{3,2}$ exchange with each other upon permutation of the two molecules. We resolve this problem by identifying $d_{2,3}$ and $d_{3,2}$ as permutable distances and sorting them before processing them by TICA. In the example shown here, this would imply that the distances are passed in the order ($d_{3,2}$, $d_{2,3}$) to TICA for the dimer shown on the left and as ($d_{2,3}$, $d_{3,2}$) for the one on the right.

molecule in self-assembly simulations are essentially artificial. Our approach is to sort each set of permutable distances and define these sorted distance values as our feature vector, which we then feed into TICA. This removes the dependence of our features on the molecular indexing and makes them permutation-invariant. The sorting procedure can be applied to both intra- and inter-molecular distances for an oligomeric system. The resulting TICA approach is called TICAgg, which is publicly available at https://github.com/loswald/TICAgg as a Python notebook. The TICAgg notebook is also capable of handling systems with more than one species of oligomer-forming molecules. In these general cases, the user needs to define lists of permutable sets of atoms and the sorting procedure is handled accordingly. It should be further noted that our sorted TICA dimension reduction method is not only capable of handling degeneracies arising from multiple monomers of the same molecule but also those due to internal symmetries of the molecule. Finally, it is also possible in TICAgg to only retain the k smallest values in every set of permutable distances, should the system become too large and the TICA calculations become too expensive. The basic idea of TICAgg is illustrated in the simplified pseudocode below:

```
Input:
coordinates of atoms for $N$ trajectory frames
a list of $k$ permutable sets containing $m$ atoms each:
{{$a_{11}$, $a_{21}$, ..., $a_{m1}$}, {$a_{12}$, $a_{22}$, ..., $a_{m2}$}, ..., {$a_{1k}$, $a_{2k}$,
..., $a_{mk}$}} (for an oligomeric system as in our study, $k$
is the oligomer size, $m$ is the number of atoms in each
monomer whose coordinates are used for the feature vec-
tor, and the indexing is consistent such that $a_{iq}$ from
monomer $q$ is chemically the same atom as $a_{ir}$ from mono-
mer $r$)
Output:
matrix $D$ consisting of feature vectors to be used as
input for the dimension reduction algorithm
```

```
Algorithm:
initialize matrix $D_{\text{intra}}$ for sorted intra-set distances
for each frame $f$ = 1, ..., $N$:
    initialize vector for distances $D_{\text{intra},f}$
    for each pair ($i$, $j$) of atoms:
        for each set $p$ = 1, ..., $k$:
            compute distance between ($a_{ip}$, $a_{jp}$)
        sort the distances
        append the sorted distances to $D_{\text{intra},f}$
    append assembled vector $D_{\text{intra},f}$ as the last row of
    $D_{\text{intra}}$
initialize matrix $D_{\text{inter}}$ for sorted inter-set distances
for each frame $f$ = 1, ..., $N$:
    initialize vector for distances $D_{\text{inter},f}$
    for each pair ($i$, $j$) of atoms:
        for each pair ($q$, $r$) of sets:
            compute distance between ($a_{iq}$, $a_{jr}$)
        sort the distances
        append the sorted distances to $D_{\text{inter},f}$
    append assembled vector $D_{\text{inter},f}$ as the last row of
    $D_{\text{inter}}$
concatenate $D_{\text{intra},f}$ and $D_{\text{inter},f}$ columnwise to obtain $D$
```

A potential weak point of our sorting strategy could be that it might be sensitive to perturbations of a trajectory, which would lead to discontinuities in the evolution of the feature vector by changing the order of a list. However, the order of distances can only be permutated when the permutable distances are extremely close in value. Therefore, even if the order of distances is reversed, this can only occur due to a small perturbation, which does not lead to any or substantial changes in the actual list of permutation-invariant distances that we feed to TICA. Thus, a slightly perturbed trajectory would be mapped to a path that is very close to the original trajectory in the transformed feature space, keeping the discontinuity impact negligible. For studying KFFE, we used interatomic distances between the backbone atoms of the peptide for constructing both intermolecular and intramolecular features. As it has been noted[68] that backbone dihedral angles are often better suited for describing the conformational ensemble of monomeric peptides compared to interatomic distances, we tested the effect of using dihedral angles on Markov state modeling for the KFFE monomer. The resulting HMM was identical to the one obtained with Cartesian coordinates. However, should for peptides larger than KFFE dihedral angles become more appropriate for describing their dynamics, the same sorting approach could be applied to the sines and cosines of dihedral angles to construct the intramolecular features. In a recent opinion piece,[69] it was argued that TICA may not always be superior to PCA for constructing MSMs since the slowest motions are not necessarily the most important ones when describing conformational transitions. To this end, we tested for the KFFE dimer whether we would benefit from using PCA for constructing HMMs for the aggregation of this peptide. However, we found that, unlike the TICA projections, the dynamics of KFFE dimerization is not well-resolved in the first two principal components. Nonetheless, it should be noted that for our innovation, which is the construction of permutation-invariant input features, there is no restriction imposed on the choice of dimension reduction algorithm. Therefore, in principle, our sorting approach pairs perfectly well with any sensible dimension reduction method.

C. Molecular dynamics simulations

Oligomerization of the KFFE peptide was investigated stepwise, starting from simulations of the monomer and going up to tetramers. The starting conformations of the trajectories at each step were taken from the metastable states discovered by the Markov state modeling performed during the previous step; i.e., simulations of the dimer were started from the metastable states of the monomer, the trimer simulations were started from metastable dimer plus monomer states, and the tetramer simulations were initiated from dimer plus dimer states as well as trimer plus monomer states. We proceed in stages like this as we wish to obtain equilibrated structures for each oligomer size, including the sampling of encounter complexes. This way we aim to circumvent the problem that in MD simulations of peptide aggregation the peptide concentration is usually 2-3 orders magnitude higher than under *in vitro* conditions.[26] The internal dynamics of a peptide or an oligomer is always in competition with their aggregation with other peptides and oligomers, so if we introduce peptides at a high concentration in our simulation box, the peptides could all aggregate together in a clump before the different species (monomer, dimer, trimer, etc.) have had time to relax. This is in contrast to a more dilute solution where collisions are much less frequent, giving rise to encounter times on the millisecond to second time scale.[26] As the computational length- and time scale limitations do not allow the modeling of peptide aggregation at concentrations found under *in vitro* let alone *in vivo* conditions, we instead assume that diffusion-limited encounter complex formation has already taken place and concentrate on simulating the transition from encounter complexes to stable oligomer states.

The monomer simulations were started with the peptide in extended conformation. The system was introduced in a dodecahedral box, large enough so that the distance between the peptide and the box walls is 1.2 nm. The system was then solvated and minimized using the steepest descent algorithm. No ions were introduced to simulate the environmental conditions at which KFFE aggregation is maximized.[52] The system was then equilibrated using a 0.1-ns NVT and a 0.1-ns NPT simulation. Then, five production runs of 2 μs were performed. During the production runs, the temperature was kept constant at 310 K using the Nosé-Hoover algorithm[70] with a time constant of 0.5 ps. Pressure was kept constant at 1 bar using the Parrinello-Rahman barostat[71] with a compressibility of 4.5 $\times 10^{-5}$ bar^{-1} and a time constant of 2 ps. Protein bonds were constrained using the LINCS (LINear Constraint Solver) algorithm[72] with an order of 2 and the internal water dynamics was constrained using the SETTLE algorithm.[73] The time step for integrating the equations of motion was 2 fs. Electrostatic interactions were calculated using the particle mesh Ewald algorithm with a short-range cutoff of 1 nm and a Fourier spacing of 0.12 nm. Van der Waals interactions were calculated with a cutoff of 1 nm. Simulations were performed using the AMBER99SB*-ILDN force field[74] with the TIP4P-Ew water model,[75] a combination that has been shown to model accurately intrinsically disordered proteins[76] and protein aggregation.[77,78] Simulations for dimers, trimers, and tetramers were performed with the same parameters, but with dodecahedral boxes of 5.0, 5.0, 5.6, and 6.0 nm edge length, respectively. For the dimer, we simulated 10 trajectories of 2 μs each and for the trimer and tetramer, 10 trajectories of 3 μs each, amounting to 90 μs calculated simulation time spent on KFFE. All simulations were performed using Gromacs 5.1.2.[79]

III. RESULTS AND DISCUSSION

We use the results for the KFFE monomer to guide through the information obtained at each step of the MSM procedure, whereas we will concentrate on the final MSMs for the oligomers. We used the sorted distances between backbone atoms as described above as input to the TICA dimension reduction for each system. The first two collective coordinates were kept for the subsequent Markov state modelling, which we found to be sufficient for describing the dynamics of the monomer and also oligomers.

A. Dynamics of the KFFE monomer

We first study the dynamics of the KFFE monomer. The free energy landscape in Fig. 3(a) shows that the monomer adopts mainly three conformations corresponding to three minima. The next step in the MSM procedure is the k-means clustering producing 250 microstates for the monomer, which are indicated as dots representing the centroids of the microstates, while their colors are indicative of the free energy minimum they have the highest probability of belonging to. The microstate-based MSM is then used to determine the lagtime for calculating the HMM based on the convergence of the implied time scales for the slowest processes. The latter can be directly determined from the Markov model eigenvalues for different lagtimes. The results for this procedure are shown in Fig. 3(b), revealing that the slowest processes converge for a lagtime of 250 time steps corresponding to 2.5 ns, which was chosen for coarse-graining the MSM into a HMM. The transition matrix of the resulting 3-macrostate HMM is represented as a network diagram in Fig. 3(c), where the sizes of the circles represent the relative contribution of each macrostate to the stationary distribution and the thickness of each arrow is proportional to the probability of the transition it represents. The Markovianity of this MSM was checked using the Chapman-Kolmogorov test. The results of this test in Fig. 3(d) reveal that Markovianity is guaranteed.

From the analysis of these data, we find that the KFFE monomer adopts three main conformations: a β-strand, a tight U, and an extended U conformation, which co-exist in a dynamic equilibrium. The U states are stabilized by electrostatic interactions between the oppositely charged glutamate and lysine residues, which are in competition with the intrinsic β-strand promoting tendencies of the two phenylalanine

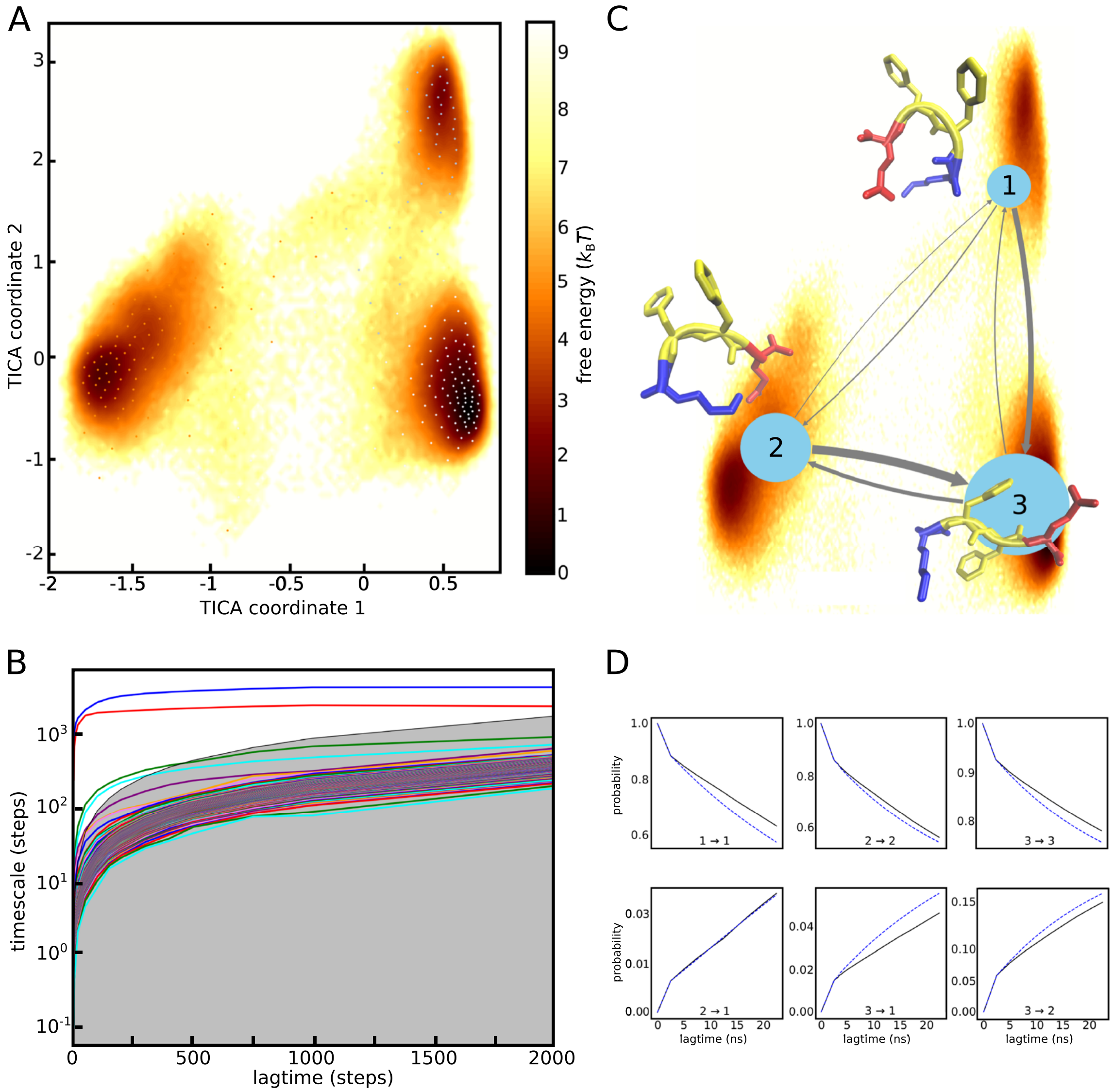

FIG. 3. Markov state model results for the KFFE monomer. (a) Free energy (in $k_B T$, see the color scale on the right) plotted along the first two TICA coordinates. The microstates as obtained from k-means clustering are shown as dots using different colors to indicate their assignment to one of the three free energy minima. (b) Convergence of the slowest implied time scales with increasing lagtime, based on which 250 time steps were chosen as lagtime for coarse-graining the MSM. (c) A network diagram of the HMM obtained for the KFFE monomer, which is overlaid on the corresponding free energy landscape. The circles represent the stable macrostates, where the area of the circles correlates with the population of the corresponding state. The arrows indicate transitions between the states, with the line thickness correlating with the transition probability. Representative KFFE structures (K: blue, F: yellow, E: red) are presented next to the corresponding node: a tight U (state 1), an extended U (state 2), and a β-strand (state 3). (d) The Chapman-Kolmogorov test for the transitions between the three HMM states. The transition probabilities estimated from the simulation data are shown in black, and the ones predicted by the HMM model are shown in blue.

residues. The low free energy barriers mean that the slowest transition in the MSM corresponds to a time scale of ~38 ns, which is a typical time scale for conformational transitions in a small peptide.[80] The β-strand has the lowest free energy and dominates with a 62% probability the stationary distribution. The extended U state has intermediate stability, with its free energy ~1 kcal/mol above the β-strand and occupies 28% of the stationary distribution. Finally, the tight U state is the least stable and has a 10% presence in the stationary distribution. Curiously, the two U-shaped states tend not to directly interconvert and the β-state acts as an intermediate between these, which is evident from the thickness of the arrows between the states in the HMM model shown in Fig. 3(c).

Our results for the monomer are in agreement with experimental observations[52] from circular dichroism spectroscopy of freshly dissolved (and presumably monomeric) KFFE, which revealed the co-existence of a β-strand structure with random structures. Furthermore, our findings are in qualitative agreement with previous MD simulations of KFFE,[55,59,60] which usually also produced the β-strand, tight, and extended U conformations as the three stable structures for the KFFE monomer. However, depending on the protein force field and solvent model used, different orders of stability of these three conformations were identified, which was discussed in detail by Strodel and Wales.[59] They found that combining Charmm19[81] with either the generalized Born model GB1[82] or the empirical implicit solvent model EEF1[83] favors the β-strand over the two U conformations, while the other two Charmm/implicit solvent model combinations prefer one of the two U conformations. It was argued that this might result from the overstabilization of electrostatic interactions between amino acid residues sometimes found in implicit solvent. Though also Bellesia and Shea, who employed an explicit water model in conjunction with OPLS/AA, identified the extended U as more stable than the β-strand. They also observed only rare transitions from either of the U-shaped conformations to the β-strand, while the HMM in Fig. 3(c) clearly shows that with AMBER99SB*-ILDN and the TIP4P-Ew water model the preferred interconversion occurs from the extended U to the β-strand. Nonetheless, our studies agree on having identified the same three conformations that the KFFE monomer can adopt.

B. Dynamics of KFFE dimerization

The simulations of the KFFE dimer reveal that the three monomeric conformations aggregate in almost every possible combination, giving rise to numerous possible dimer configurations, which has already been observed previously.[55,59,60] The dimers themselves are not very persistent, and many aggregation and dissociation events are observed in our simulations. This is surprising since a common criticism of force fields is that they overestimate protein-protein interaction energies and thus overstabilize protein aggregation.[78,84] The frequent association and dissociation events in combination with the many possible combination patterns of monomer conformations lead to a complex free energy landscape with a diversity of aggregates, making these simulations a challenging testbed for our automatic MSM approach. The HMMs generated by TICAgg are successful at providing us with a clear look into the dynamics of the system; they identify the metastable states and their interconversion rates along the aggregation pathway. The HMM for the dimer is shown in Fig. 4, while the free energy landscape with the microstates from k-means clustering, convergence of the

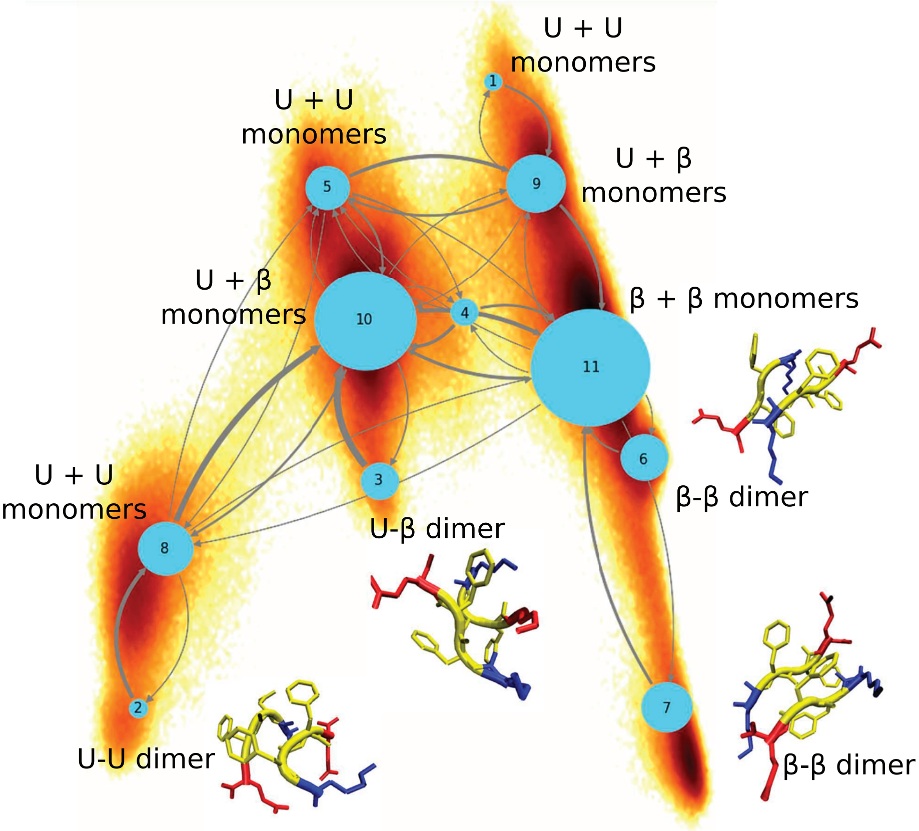

FIG. 4. A network diagram of the HMM model obtained for the KFFE dimer, which is overlaid on the corresponding free energy landscape. States 2, 3, 6, and 7 represent dimeric states, while all other states are different combinations of monomers. Representative structures of the dimers are shown. An explanation of the colors is given in Fig. 3.

slowest implied time scale and the results from the Chapman-Kolmogorov test are provided in the supplementary material (Figs. S1–S3).

From the dimer HMM represented by its network diagram in Fig. 4, one can see that the two flavors of β-β dimers are the most stable dimers (states 6 and 7, 7% and 8% contribution to the stationary distribution), followed by the mixed U-β dimer (state 3, 4%) and the U-U dimer (state 2, 1%). Each dimer is kinetically closest to the corresponding dissociated monomers, which dominate the configurational space with an overall population of 80%. In other words, the rate of dissociation is considerably larger than the rate of association for all dimers, which is directly visible from the thickness of the arrows between the monomeric and dimeric states. Consequently, the dimers are rather unstable and frequently dissociate into their respective monomers, which can change conformations rapidly and re-associate into a different kind of dimer. The network diagram also indicates that there are no direct interconversions between the different dimeric species, i.e., dimers are only formed via monomer assembly. As expected, the stable dimeric species detected in our simulations are aligned in an anti-parallel fashion (Fig. 4), resulting from the electrostatic attraction between the oppositely charged termini and the hydrophobic interactions between the cores. It should be noted that even though the β-strands are aligned antiparallel in the β-β dimers, proper β-sheets were not formed due to the lack of sufficient hydrogen bonds (H-bonds) between the backbones of the two peptides. In the previous simulations of the KFFE dimer,[55,59,60] a perfect β-sheet was also not observed apart from the simulation that employed the Charmm19/EEF1 force field/solvent model combination.

C. Aggregation into larger KFFE oligomers

The results for the KFFE trimer and tetramer are similar to those for the dimer. During the MD simulations, numerous dissociation and re-association events were sampled. Similar to the dimer, a large number of combinations of U conformations, and β-strands can aggregate to form a variety of different trimers or tetramers, which are, however, not very stable. The detailed results for the trimer and tetramer are visible in the supplementary material (Figs. S4–S11). They confirm the findings obtained for the dimer as the direct interconversion between different trimer or tetramer configurations seem to not be preferred. Instead, the addition of a monomer or dimer to the corresponding lower oligomer seems to be the dominant aggregation mechanism. The largest time scales for the dimer, trimer, and tetramer HMMs are about 50, 300, and 400 ns, respectively. Although there are hardly any large free energy barriers between adjacent macrostates in the energy landscape of KFFE oligomerization, the increasing number of states with the introduction of further monomeric units results in slower transitions between (non-adjacent) states. Among the trimers, some degree of metastability is achieved by the β-β-U trimers (states 1, 2 and 7 in Fig. S7), which result from the assembly of a β-β dimer and a U monomer. Other trimers of notable stability are β-β-β and U-U-U trimers (state 3 in Fig. S7 and state 9 in Fig. S11, respectively). While all the tetramers tend to disintegrate quickly, the two tetramers with β-strands, one consisting of four β-strands (state 11 in Fig. S11) and the other one comprising two β-strands and two U-shaped units (state 8 in Fig. S11), are slightly more persistent than the U-rich tetramers (state 1 in Fig. S11). Similar to the dimer, no perfect β-sheet formed despite the antiparallel alignment of three or four β-strands in the β-rich trimers and tetramers. The structures of KFFE dimers, trimers, or tetramers are yet to be experimentally determined. It is currently unclear whether these oligomers serve as nuclei for the further formation of amyloid fibrils or whether they are only kinetic intermediates. The relevance of these results to the fibrillization mechanism depends on whether the critical nucleus is smaller than or equal to four KFFE peptides. The relative instability of the oligomers and lack of proper β-sheet formation that we observed seem to suggest that the fibrillization nucleus might actually be larger.

IV. CONCLUSION

A tool to construct automated Markov state models from MD simulations of molecular self-assembly and aggregation is presented here. Molecular aggregation is the cause of a wide-ranging variety of diseases but can also play important functional roles in our body, and self-assembled nanostructures have many promising applications. Our PyEMMA-based Jupyter notebook TICAgg reduces the burden of constructing a MSM for self-assembly simulations to selecting a handful of parameters. The main challenge to automation was the degeneracy present in an oligomeric system. We solved this problem during dimension reduction by sorting the inter-atom distances that are permutable. The modified TICA that we have developed here can be used as a standalone dimension reduction technique for the construction of good collective coordinates in systems with degeneracies arising from multiple identical molecules or symmetry or be combined with MSMs as we have done.

We tested our technique on extensive simulations of aggregating KFFE peptides, a small peptide that has been used in the literature as a minimal model to investigate protein aggregation.[52] Similar to previous atomistic simulation studies,[55,59,60] we found that the KFFE monomer has three metastable configurations: two U shaped states and one β-strand state. These monomeric configurations can assemble in all possible combinations to form a diverse variety of oligomers. TICAgg successfully identified the metastable aggregates, and in combination with the calculation of HMMs, the relevant aggregation pathways were revealed. A schematic showing the dominant oligomerization pathways starting from monomers and leading up to tetramers discovered from our HMMs is shown in Fig. 5. Trimers are formed via the association of a dimer and a monomer, while tetramers evolve from both a trimer plus a monomer and a dimer plus a dimer. Especially β-β and U-U dimers contribute to the latter association pathway. The most important conclusion from Fig. 5 is that there is no direct interconversion between different oligomer configurations of the same aggregate size. This leads

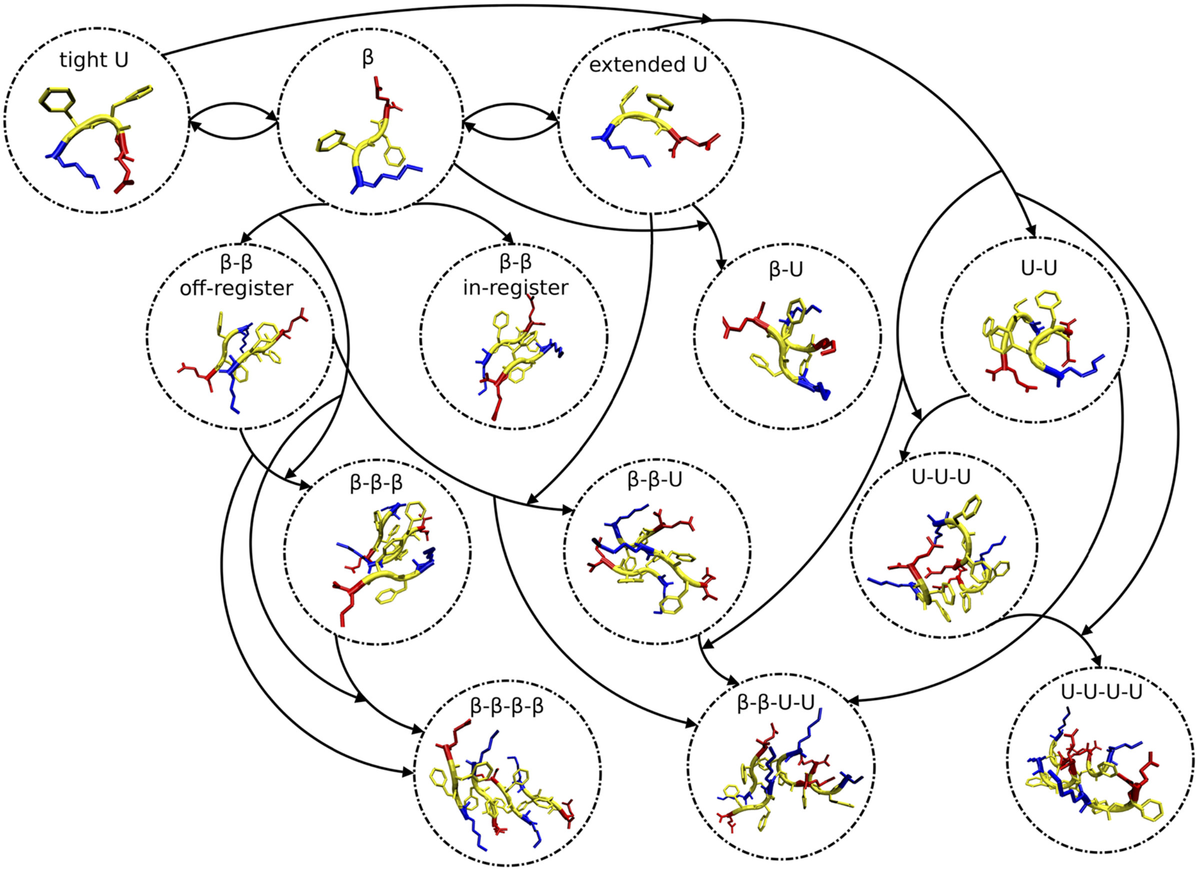

FIG. 5. A scheme summarizing the major KFFE oligomerization pathways discovered from the HMMs of the monomer (top row), dimer (second row), trimer (third row), and tetramer (bottom row). No considerable conversions between oligomers of the same size are observed. The major pathways for the different oligomer types occur via association between monomers and/or smaller oligomers, leading to separate pathways for the formation of β-sheets (on the left side of the scheme), completely disordered oligomers (right side), and mixed β/disordered oligomers (middle).

to different pathways for the formation of β-strand oligomers, which are the most likely fibrillization precursors, and of non-β-strand oligomers, also known as disordered oligomers. A common hypothesis in the amyloid field is that aggregation into amyloid fibrils is initiated by hydrophobic collapse into disordered aggregates, followed by their structural transition into ordered β-sheets. However, our results in Fig. 5 seem to show that this is not the case as the disordered oligomers do not convert to ordered β-rich oligomers (and vice versa). The aggregation pathway leading to disordered oligomers can thus be expected to be off-pathway with respect to amyloid aggregation.[26] We argued that the hydrophobic collapse commonly seen in MD simulation studies of amyloid peptide aggregation is most likely a consequence of the 2–3 orders higher peptide concentration compared to the *in vitro* and *in vivo* situations, which gets further enhanced by the overstabilization of inter-protein interactions by most force fields.[78] We avoided the concentration problem by studying each oligomer state individually. The conclusions may be different for larger peptides such as full-length Aβ. Thus, future studies should test whether the separation between the pathways for aggregation into β-sheets and disordered oligomers holds true for amyloid peptides other than KFFE.

Further work with TICAgg could be both applied and methodological. It can be used to study more biologically relevant peptides, such as Aβ, or other self-assembling systems of interest. Recent advances being made in Markov state modeling can also be incorporated into the tool, such as the newly developed TRAMMBAR estimator[19] that allows building MSMs from enhanced sampling MD data such as Hamiltonian replica exchange or novel dimension reduction techniques like SparseTICA,[85] which constructs sparse collective coordinates that are easier to interpret.

SUPPLEMENTARY MATERIAL

See supplementary material for detailed results obtained during Markov modeling of the KFFE dimer, trimer, and tetramer. For each of these oligomers, plots of (i) the free energy surface including the microstates, (ii) the implied time scales, (iii) the results from the Chapman-Kolmogorov test, and (iv) the hidden Markov model with representative conformations for the most stable oligomer states are shown.

ACKNOWLEDGMENTS

The authors gratefully acknowledge the computing time granted by JARA-HPC from RWTH Aachen University under Project No. JARA0095. M.C.-P. and B.S. acknowledge financial support from the Deutsche Forschungsgemeinschaft (German Research Foundation, DFG) through Grant No. GSC 111.

REFERENCES

[1] R. O. Dror, R. M. Dirks, J. P. Grossman, H. Xu, and D. E. Shaw, Annu. Rev. Biophys. **41**, 429 (2012).

[2] J.-H. Prinz, B. Keller, and F. Noé, Phys. Chem. Chem. Phys. **13**, 16912 (2011).

[3] B. G. Keller, A. Kobitski, A. Jäschke, G. U. Nienhaus, and F. Noé, J. Am. Chem. Soc. **136**, 4534 (2014).

[4] S. Olsson, H. Wu, F. Paul, C. Clementi, and F. Noé, Proc. Natl. Acad. Sci. U. S. A. **114**, 8265 (2017).

[5] S. Olsson and F. Noé, J. Am. Chem. Soc. **139**, 200 (2017).

[6] B. Trendelkamp-Schroer, H. Wu, F. Paul, and F. Noé, J. Chem. Phys. **143**, 174101 (2015).

[7] N. Djurdjevac, M. Sarich, and C. Schütte, Multiscale Model. Simul. **10**, 61 (2012).

[8] G. R. Bowman and V. S. Pande, Proc. Natl. Acad. Sci. U. S. A. **107**, 10890 (2010).

[9] A. Sirur, D. De Sancho, and R. B. Best, J. Chem. Phys. **144**, 075101 (2016).

[10] G. R. Bowman, D. L. Ensign, and V. S. Pande, J. Chem. Theory Comput. **6**, 787 (2010).

[11] F. Noe, C. Schutte, E. Vanden-Eijnden, L. Reich, and T. R. Weikl, Proc. Natl. Acad. Sci. U. S. A. **106**, 19011 (2009).

[12] B. Zagrovic, C. D. Snow, M. R. Shirts, and V. S. Pande, J. Mol. Biol. **323**, 927 (2002).

[13] K. A. Beauchamp, G. R. Bowman, T. J. Lane, L. Maibaum, I. S. Haque, and V. S. Pande, J. Chem. Theory Comput. **7**, 3412 (2011).

[14] M. K. Scherer, B. Trendelkamp-Schroer, F. Paul, G. Pérez-Hernández, M. Hoffmann, N. Plattner, C. Wehmeyer, J.-H. Prinz, and F. Noé, J. Chem. Theory Comput. **11**, 5525 (2015).

[15] Q. Liao, Y. Kulkarni, U. Sengupta, D. Petrović, A. J. Mulholland, M. W. van der Kamp, B. Strodel, and S. C. L. Kamerlin, J. Am. Chem. Soc. **140**, 15889 (2018).

[16] U. Sengupta and B. Strodel, Philos. Trans. R. Soc., B **373**, 20170178 (2018).

[17] N. Plattner and F. Noé, Nat. Commun. **6**, 7653 (2015).

[18] N. Plattner, S. Doerr, G. De Fabritiis, and F. Noé, Nat. Chem. **9**, 1005 (2017).

[19] F. Paul, C. Wehmeyer, E. T. Abualrous, H. Wu, M. D. Crabtree, J. Schöneberg, J. Clarke, C. Freund, T. R. Weikl, and F. Noé, Nat. Commun. **8**, 1095 (2017).

[20] M. Schor, A. S. J. S. Mey, F. Noé, and C. E. MacPhee, J. Phys. Chem. Lett. **6**, 1076 (2015).

[21] P. H. Nguyen, M. S. Li, and P. Derreumaux, J. Chem. Phys. **140**, 094105 (2014).

[22] L. Molgedey and H. G. Schuster, Phys. Rev. Lett. **72**, 3634 (1994).

[23] G. Pérez-Hernández, F. Paul, T. Giorgino, G. De Fabritiis, and F. Noé, J. Chem. Phys. **139**, 015102 (2013).

[24] C. R. Schwantes and V. S. Pande, J. Chem. Theory Comput. **9**, 2000 (2013).

[25] A. Altis, M. Otten, P. H. Nguyen, R. Hegger, and G. Stock, J. Chem. Phys. **128**, 245102 (2008).

[26] M. Carballo-Pacheco and B. Strodel, J. Phys. Chem. B **120**, 2991 (2016).

[27] A. Morriss-Andrews and J.-E. Shea, Annu. Rev. Phys. Chem. **66**, 643 (2015).

[28] S. Whitelam and R. L. Jack, Annu. Rev. Phys. Chem. **66**, 143 (2015).

[29] F. Sterpone, S. Melchionna, P. Tuffery, S. Pasquali, N. Mousseau, T. Cragnolini, Y. Chebaro, J.-F. St-Pierre, M. Kalimeri, A. Barducci, Y. Laurin, A. Tek, M. Baaden, P. H. Nguyen, and P. Derreumaux, Chem. Soc. Rev. **43**, 4871 (2014).

[30] T. P. J. Knowles, M. Vendruscolo, and C. M. Dobson, Nat. Rev. Mol. Cell Biol. **15**, 496 (2014).

[31] J. Nasica-Labouze, P. H. Nguyen, F. Sterpone, O. Berthoumieu, N. V. Buchete, S. Coté, A. De Simone, A. J. Doig, P. Faller, A. Garcia, A. Laio, M. S. Li, S. Melchionna, N. Mousseau, Y. Mu, A. Paravastu, S. Pasquali, D. J. Rosenman, B. Strodel, B. Tarus, J. H. Viles, T. Zhang, C. Wang, and P. Derreumaux, Chem. Rev. **115**, 3518 (2015).

[32] C. A. Ross and M. A. Poirier, Nat. Med. **10**, S10 (2004).

[33] W. L. Klein, J. Alzheimer's Dis. **33**, S49 (2012).

[34] M. D. Kirkitadze, G. Bitan, and D. B. Teplow, J. Neurosci. Res. **69**, 567 (2002).

[35] R. Kayed and C. A. Lasagna-Reeves, J. Alzheimer's Dis. **33**, S67 (2012).

[36] L. Nagel-Steger, M. C. Owen, and B. Strodel, ChemBioChem **17**, 657 (2016).

[37] B. Barz, Q. Liao, and B. Strodel, J. Am. Chem. Soc. **140**, 319 (2018).

[38] M. Dong, H. Li, D. Hu, W. Zhao, X. Zhu, and H. Ai, ACS Chem. Neurosci. **7**, 599 (2016).

[39] F. Mustan, A. Ivanova, G. Madjarova, S. Tcholakova, and N. Denkov, J. Phys. Chem. B **119**, 15631 (2015).

[40] D. C. Rapaport, J. E. Johnson, and J. Skolnick, Comput. Phys. Commun. **121**, 231 (1999).

[41] T. Carpenter, P. J. Bond, S. Khalid, and M. S. P. Sansom, Biophys. J. **95**, 3790 (2008).

[42] Y. Yang, R. B. Meyer, and M. F. Hagan, Phys. Rev. Lett. **104**, 258102 (2010).

[43] B. J. Reynwar, G. Illya, V. A. Harmandaris, M. M. Müller, K. Kremer, and M. Deserno, Nature **447**, 461 (2007).

[44] M. Carballo-Pacheco, A. E. Ismail, and B. Strodel, J. Phys. Chem. B **119**, 9696 (2015).

[45] D. Roccatano, *Micro Nanomanufacturing* (Springer International Publishing, Cham, 2018), Vol. II, pp. 123–155.

[46] N. W. Kelley, V. Vishal, G. A. Krafft, and V. S. Pande, J. Chem. Phys. **129**, 214707 (2008).

[47] C. T. Leahy, A. Kells, G. Hummer, N.-V. Buchete, and E. Rosta, J. Chem. Phys. **147**, 152725 (2017).

[48] M. R. Perkett and M. F. Hagan, J. Chem. Phys. **140**, 214101 (2014).

[49] X. Zeng, B. Li, Q. Qiao, L. Zhu, Z.-Y. Lu, and X. Huang, Phys. Chem. Chem. Phys. **18**, 23494 (2016).

[50] Y. Cao, X. Jiang, and W. Han, J. Chem. Theory Comput. **13**, 5731 (2017).

[51] G. Imbalzano, A. Anelli, D. Giofré, S. Klees, J. Behler, and M. Ceriotti, J. Chem. Phys. **148**, 241730 (2018).

[52] L. Tjernberg, W. Hosia, N. Bark, J. Thyberg, and J. Johansson, J. Biol. Chem. **277**, 43243 (2002).

[53] M. Balbirnie, R. Grothe, and D. S. Eisenberg, Proc. Natl. Acad. Sci. U. S. A. **98**, 2375 (2001).

[54] F. Bemporad, Protein Sci. **15**, 862 (2006).

[55] G. Bellesia and J. E. Shea, Biophys. J. **96**, 875 (2009).

[56] G. Wei, N. Mousseau, and P. Derreumaux, Biophys. J. **87**, 3648 (2004).

[57] A. Melquiond, G. Boucher, N. Mousseau, and P. Derreumaux, J. Chem. Phys. **122**, 174904 (2005).

[58] A. Melquiond, N. Mousseau, and P. Derreumaux, Proteins: Struct., Funct., Bioinf. **65**, 180 (2006).

[59] B. Strodel and D. J. Wales, J. Chem. Theory Comput. **4**, 657 (2008).

[60] A. Baumketner and J.-E. Shea, Biophys. J. **89**, 1493 (2005).

[61] B. E. Husic and V. S. Pande, J. Am. Chem. Soc. **140**, 2386 (2018).

[62] V. S. Pande, K. Beauchamp, and G. R. Bowman, Methods **52**, 99 (2010).

[63] J. D. Chodera and F. Noé, Curr. Opin. Struct. Biol. **25**, 135 (2014).

[64] F. Noé and F. Nüske, "A variational approach to modeling slow processes in stochastic dynamical systems," SIAM Multiscale Model. Simul. **11**, 635–655 (2013).

[65] T. F. Gonzalez, Theor. Comput. Sci. **38**, 293 (1985).
[66] S. Kube and M. Weber, J. Chem. Phys. **126**, 024103 (2007).
[67] R. T. McGibbon, K. A. Beauchamp, M. P. Harrigan, C. Klein, J. M. Swails, C. X. Hernández, C. R. Schwantes, L. P. Wang, T. J. Lane, and V. S. Pande, Biophys. J. **109**, 1528 (2015).
[68] M. Ernst, F. Sittel, and G. Stock, J. Chem. Phys. **143**, 244114 (2015).
[69] F. Sittel and G. Stock, J. Chem. Phys. **149**, 150901 (2018).
[70] S. Nosé, J. Chem. Phys. **81**, 511 (1984).
[71] M. Parrinello and A. Rahman, J. Appl. Phys. **52**, 7182 (1981).
[72] B. Hess, H. Bekker, H. J. C. Berendsen, and J. G. E. M. Fraaije, J. Comput. Chem. **18**, 1463 (1997).
[73] S. Miyamoto and P. A. Kollman, J. Comput. Chem. **13**, 952 (1992).
[74] K. Lindorff-Larsen, S. Piana, K. Palmo, P. Maragakis, J. L. Klepeis, R. O. Dror, and D. E. Shaw, Proteins: Struct., Funct., Bioinf. **78**, 1950 (2010).
[75] H. W. Horn, W. C. Swope, J. W. Pitera, J. D. Madura, T. J. Dick, G. L. Hura, and T. Head-Gordon, J. Chem. Phys. **120**, 9665 (2004).
[76] M. Carballo-Pacheco and B. Strodel, Protein Sci. **26**, 174 (2017).
[77] A. M. Fluitt and J. J. De Pablo, Biophys. J. **109**, 1009 (2015).
[78] M. Carballo-Pacheco, A. E. Ismail, and B. Strodel, J. Chem. Theory Comput. **14**, 6063 (2018).
[79] M. J. Abraham, T. Murtola, R. Schulz, S. Páll, J. C. Smith, B. Hess, and E. Lindah, SoftwareX **1–2**, 19 (2015).
[80] B. Strodel and D. J. Wales, Chem. Phys. Lett. **466**, 105 (2008).
[81] X. Zhu, P. E. M. Lopes, and A. D. MacKerell, Wiley Interdiscip. Rev.: Comput. Mol. Sci. **2**, 167 (2012).
[82] B. N. Dominy and C. L. Brooks, J. Phys. Chem. B **103**, 3765 (1999).
[83] T. Lazaridis and M. Karplus, Proteins: Struct., Funct., Genet. **35**, 133 (1999).
[84] D. Petrov and B. Zagrovic, PLoS Comput. Biol. **10**, e1003638 (2014).
[85] R. T. McGibbon, B. E. Husic, and V. S. Pande, J. Chem. Phys. **146**, 044109 (2017).