A kinetic ensemble of the Alzheimer's Aβ peptide

Thomas Löhr[1], Kai Kohlhoff[2], Gabriella T. Heller[1],
Carlo Camilloni[3], Michele Vendruscolo[1*]

[1]*Department of Chemistry, University of Cambridge, CB2 1EW, UK*
[2]*Google Research, Mountain View, CA 94043, USA*
[3]*Dipartimento di Bioscienze, Università degli Studi di Milano, 20133 Milano, Italy*

Abstract

The discovery that disordered proteins are widespread in the human proteome has prompted the quest for methods to characterize the conformational properties that determine their functional and dysfunctional behaviour. It has become customary to describe these proteins in terms of structural ensembles and free energy landscapes, which offer information about their conformations and thermodynamics. However, a major challenge is to generalize this description to 'kinetic ensembles', thereby also providing information on transition rates between states. Approaches based on the theory of stochastic processes can be particularly suitable for this purpose. Here, we develop a Markov state model and illustrate its application by determining a kinetic ensemble of the 42-residue form of the amyloid-β peptide (Aβ42), whose aggregation is associated with Alzheimer's disease. Using the Google Compute Engine, we generated 315 *μ*s all-atom, explicit solvent molecular dynamics trajectories, validated with nuclear magnetic resonance measurements. Using a probabilistic-based definition of conformational states in a neural network approach, we found that Aβ42 is characterized by inter-state transitions no longer than the microsecond timescale, exhibiting only fully unfolded or short-lived, partially-folded states. Our results illustrate how the use of kinetic ensembles offers an effective means to provide information about the structure, thermodynamics, and kinetics of disordered proteins.

Introduction

Proteins that are fully or partially disordered make up approximately one third of the human proteome, perform a variety of biological functions and are closely involved with many major human disorders[1–3]. The existence of disordered proteins is making it necessary to extend the structure-function relationship that has driven major advances in protein science in the last 50 years to a structure-dynamics-function relationship, in order to account for the essential role of structural disorder in determining the normal and aberrant behaviours of these proteins[1–4].

Because of their conformational heterogeneity, it is typically insufficient to characterize disordered proteins using one or a few specific structures, as it is standard for folded proteins[5]. Instead, it has become common to describe these proteins in terms of structural ensembles, which in turn are often represented through free energy landscapes[6–9], when the statistical weights of the states in the ensemble are available (for example, through the use of enhanced sampling techniques[6]). This is a powerful description, which concisely captures information about the structure and thermodynamics of disordered proteins. Here, we refer to these ensembles here as 'thermodynamic ensembles'. It has also been recently observed, however, that a more complete description should include information about the kinetics, which can be achieved by adding the transition rates between different conformational states[10]. We refer to these ensembles here as 'kinetic ensembles'. This task requires a characterisation of the kinetic properties of disordered proteins, which remains a challenging task, both experimentally and computationally.

In this work, we describe an approach to generate kinetic ensembles, which contain information about the molecular structures of proteins, the populations of their metastable states, and the transition rates between these different metastable states (**Figure 1A**). While acquiring structural and population information is already possible with molecular dynamics simulations alone, gaining interpretable kinetic information can be more effectively achieved by exploiting the theory of stochastic processes[11] during the analysis, and it is typically done using Markov state models[12,13] .

In contrast to thermodynamic ensembles, constructing kinetic ensembles requires a state decomposition. To define the states in these models, one can choose a set of features from a trajectory, such as backbone dihedral angles or the root-mean-square deviation to some reference structure, and then use a clustering algorithm to obtain a state assignment for each frame. One can then count the transitions between states and normalize these counts to obtain

a transition matrix[12,14]. An additional coarse-graining step can then be performed to obtain a more interpretable model with fewer states. Alternatively, clustering can be preceded with an additional dimensionality reduction step. One such example particularly relevant for building Markov models is time-lagged independent component analysis (TICA)[15,16] which projects conformations into a space where distances have some kinetic meaning. Clustering in this space thus has the potential of naturally preserving kinetic separation.

This particular model-building approach presents unique challenges for disordered proteins, due to the heterogeneous nature of their conformations. Transitions between metastable states are typically fast and not necessarily characterized by large variations in the overall shape of the protein. This is in contrast to folded proteins, where states can often be classified on structural properties alone[17–19]. For disordered proteins, even with suitable structural measures, because these transitions are fast, dividing this space into discrete areas is exceedingly difficult (**Figure 1B**), and many clustering algorithms may fail to consistently classify states. Consequently, it would be ideal to obtain a type of clustering, which we can describe as 'soft', where state assignments can have a probabilistic nature. In this approach, every frame is assigned a discrete probability distribution encoding its membership in a certain state. This information can be used to build a soft Markov state model, which can be interpreted in much the same way as one built from individual transition counts[20]. To optimize for kinetic separation and allow for probabilistic state assignments, we employed the VAMPNet technique with physical constraints[21,22] to construct a soft Markov state model (or Koopman model, see Methods). The resulting stochastic matrix allows us to compute the equilibrium distribution of the ensemble as well as the transition rates between metastable states.

We illustrate this approach to determine a kinetic ensemble of Aβ42, using 315 μs of explicit solvent molecular dynamics simulations. We chose Aβ42 because of its association with Alzheimer's disease, which is among the leading causes of death in the developed world, with no effective treatments available[23]. One of the characteristics of Alzheimer's disease is the formation of neurotoxic aggregates of Aβ42[24–29]. This complex process can be divided in to a set of nonlinearly coupled microscopic steps, for all of which the monomeric state of the protein plays a key role[30]. It is therefore of great importance to better understand the structural and kinetic properties of the monomeric state of Aβ42, as the dynamic properties in particular are key to determine the effects of small molecule drug candidates on the aggregation behaviour of this peptide[31,32]. Because of the ensemble-averaged nature of most experimental measurements of Aβ42, molecular simulations can provide a uniquely precise description of its kinetic ensemble. Previous computational studies to characterize Aβ42[33–38] have focused

mainly on the thermodynamic properties of this peptide, but less so on the kinetics of state transitions. It is therefore still debated as to whether this peptide ever adopts long-lived (i.e. more than hundreds of microseconds) states associated with its aggregation behaviour. We validate the kinetic ensemble using independent data obtained from nuclear magnetic resonance (NMR) experiments. To the best of our knowledge, this is the longest published simulation of any disordered protein in explicit solvent executed to date. The resulting kinetic ensemble provides the structural properties and the populations of the states of Aβ42 and the transition rates between them. It also allows us to make predictions of involved timescales and is especially informative in terms of the kinetics of secondary structure formation. We believe that this approach offers a unique perspective into the structure, thermodynamics, and kinetics of disordered proteins towards an increased understanding of their dynamic behaviour.

Results

Generation of molecular dynamics simulations on the Google Compute Engine. We performed explicit solvent molecular dynamics simulations in 5 rounds with 1024 trajectories each, using the fluctuation amplification of specific traits (FAST) approach[39] to accelerate the exploration of conformational space. Cloud architectures, such as the Google Compute Engine are especially well suited to generate Markov state models, as the individual simulations can be run independently from one another, with no communication between machines required[40] .

Standard Markov model techniques do not readily describe the Aβ42 ensemble. We first attempted to build discrete Markov state models using existing state-of-the-art techniques. We used time-lagged independent component analysis (TICA)[15,16] as a preliminary dimensionality reduction step with inter-residue nearest-neighbour heavy atom distances, followed by clustering with various algorithms, different numbers of input dimensions, and varying amounts of clusters. We used the cross-validated generalized matrix Rayleigh coefficient technique[41] to score the individual models (see Methods, **Figure S1**). Despite an extensive hyperparameter search, we were unable to construct a model with converging timescales (**Figure S2**). Nevertheless, we used the best-performing hyperparameters to build discrete coarse-grained hidden Markov state models[42]. These models, however, also suffered from non-converging timescales, suggesting the presence of significant problems in this particular approach for state-space discretization. We therefore decided to pursue an approach based on a probabilistic state description.

Soft Markov state models describe the Aβ42 ensemble. We next attempted to build a neural network for the disordered ensemble of Aβ42 with soft state assignments. To this end we employed the VAMPNet technique with physical constraints[21,22]. VAMPNet is a two-lobed neural network, taking as input two frames separated by a lag time τ, and yielding a soft state assignment vector for each frame. The loss function is given by the variational approach to Markov processes (VAMP) score[43], allowing the learning of a state decomposition without explicit state labelling. The dimensionality reduction and clustering steps are thus performed in a single procedure, allowing for highly non-linear state membership functions to be learned. A recent addition[22] to this approach allows the use of constraints to keep the elements in the learned transition matrix positive, and the model reversible. We adopted this framework but employed a self-normalising neural network architecture[44] to improve the hyperparameter search (see Methods). As input, we used inter-residue nearest-neighbour heavy atom distances, resulting in an input dimension of 780. We used 2 to 6 output states, finding that the 4-state model struck the best balance between interpretability, detail, and state assignment errors (**Figure S3**). Outputs with increased numbers of states resulted in larger errors and a lack of interpretability.

Computational validation of the Aβ42 kinetic ensemble. To construct a kinetic model, a suitable lag time τ between successive conformations along a trajectory needs to be chosen. This lag time should be small enough for the model to effectively resolve fast degrees of freedom, but large enough to not introduce a significant statistical error in the prediction of longer timescales. To choose this parameter and evaluate the general quality of the model it is helpful to visualize the longer timescales – a kinetic property of the model – as a function of the lag time itself. To this end, we assessed the dependence of the relaxation timescales, t_i, (also called implied timescales in this approach, see Methods) on the lag time τ (**Figure 2A**) and observed that a lag time $\tau = 12.5$ ns can resolve longer timescales well. To further validate this choice for the lag time, we used the Chapman-Kolmogorov test (**Figures S4, S8, and S9**), a stringent measure of the predictive abilities of our model. To evaluate the quality of sampling in the context of the final model, we rebuilt it with a varying number of trajectories (**Figure 2B**). The populations indeed converge to within the error of the model, as expected as we improve the sampling of each state. To understand how the choice of the number of states impacts the corresponding state assignments of individual frames, we illustrate the state decomposition as a tree (**Figure 2C**). States can be seen to be mostly consistent across different output sizes.

Experimental validation of the Aβ42 kinetic ensemble. We compared acquired chemical shifts[31] to values back-calculated from the simulations (**Figure S5A**). We found that the root-mean-square deviation between experiments and simulations is within the error of the forward model[45]. Additionally, we directly compared the distributions of the radius of gyration between the kinetic ensemble derived here and of a thermodynamic ensemble determined previously using metadynamic metainference[46,47] simulations[31] , using the same force field and experimental restraints in the form of chemical shifts (**Figure S5B**). The relaxation times of around 2 μs (**Figure 4B**) are also consistent with experimental NMR measurements[48]. However, acquiring more accurate data on state transitions of disordered proteins using NMR spectroscopy is limited by the vast structural variance of the proteins on short timescales and ensemble-averaged nature of the experiments[49].

Aβ42 states are structurally diverse. To characterize the local structural features and long-ranged self-interactions of the states in the kinetic ensemble, we calculated the α-helical and β-sheet content per residue (**Figures 3A, S6A** and **S7A**) and inter-residue contact probabilities (**Figures 3B, S6B** and **S7B**) in each state. We observed unique structural properties in each state and found that the C-terminal region of the peptide is generally more structured than the N-terminus. Specifically, each state is characterized by varying degrees of α-helix formation in localized areas, while the region between residues 10 and 20 generally remains completely disordered. The β-sheet content is also concentrated towards the C-terminal region with state 3 being a notable exception, partially forming an end-to-end contact. Moreover, many conformations identified by DSSP as being rich in β-sheet content are characterized more by strong backbone interactions and less by explicit β-like structures, as can be seen in the contact probability maps.

Aβ42 is characterized by transitions between states on the μs timescale. The overall kinetic properties of a protein are primarily given by the inter-state transition probabilities (**Figure 4A**), the slowest relaxation timescales (**Figure 4B**), and mean state lifetimes (**Figure 4C**). We observe mean first-passage times between ~1 and ~12 μs, with no observation of longer-lived folded states. Instead, we observe the formation of a disordered hub-like state 0, to which other structured states transition relatively quickly. On the other hand, transitions into structured states, i.e. direct formation of secondary structural motifs, are significantly slower. Transitions bypassing the hub-state are rare, such as transitions between states 1 to 2. The relationship between the relatively fast relaxation times (~2 μs) (**Figure 4B**) and the somewhat slower mean first-passage times (1-12 μs) has previously been investigated for unfolded states of small proteins[50,51]. It was observed that the lifetime of a state is approximately equal to the

timescale of relaxation to equilibrium (**Figure 4C**), thereby suggesting that the mean first-passage time into a state is mostly dependent on the population of the state itself.

Discussion

In this work we determined the kinetic ensemble of Aβ42 using a neural network approach. We observed that the choice of the number and quality of inputs is crucial. In particular, we observed a large increase in the VAMP2 scores upon switching from inter-residue Cα distance matrices to nearest-neighbour heavy atom distances. These results suggest that using more sophisticated and higher resolution features could further improve the state decomposition. One such method could be the use of convolutional layers[52–54], either acting on distance matrices or on the 3D protein structure itself.

Our study also identifies the importance of developing robust mathematical tools to handle the Koopman matrix and corresponding error estimations. There have been numerous developments in the Markov model literature for these problems, such as the use of Bayesian methods to estimate both state discretization and statistical sampling uncertainties[55] or the use of additional experimental data to improve the transition matrix estimation[56]. These tools are, however, all based on the availability of discrete transition counts, and there are no direct analogs for Koopman models so far.

Aβ42 has been studied extensively using molecular simulations in the context of thermodynamic ensembles, commonly using enhanced sampling techniques such as replica-exchange molecular dynamics or variants thereof[33–38]. These and related studies suggest that this peptide is extremely sensitive to the choice of force field and water model[57] as well as the amount of sampling[58]. These studies used higher simulation temperatures, making a quantitative comparison with our study difficult. The work by Lin *et al.*[34] is especially notable as the authors used a Markov state model to analyze their 200 μs simulation. They, however, did not report on the kinetic properties of the system and focused on ensemble-averaged structural features instead. Meng *et al.* also identified a large disordered population of 72 % in their analysis[33], consistent with state 0 in our analysis. Rosenman *et al.* found many highly diverse clusters with populations of ~5 %, comprising both extended and locally structured conformations[35]. Sgourakis *et al.* used a spectral clustering technique based on contact maps to identify many locally structured conformations[36].

Our results are particularly interesting when viewed in context of the kinetic hub model of protein folding[59]. In this model, the native state of a folded protein is the most populated state, quickly reachable from other partially or completely unfolded states. This model can be seen as the kinetic complement of the thermodynamic funnel concept of protein folding[60]. We may therefore propose an analogous model for disordered proteins, i.e. a kinetic counterpart to the recently proposed inverted funnel concept[38]. Aβ42 clearly adopts a highly populated disordered state, with slow transitions into less populous, more folded states. We can thus view this kinetic geometry as an 'inverted kinetic hub'.

Conclusions

We have presented a general strategy to characterize the structure, thermodynamics, and kinetics of disordered proteins through kinetic ensembles. To illustrate the application of this approach, we generated a kinetic ensemble of Aβ42, a disordered peptide closely associated with Alzheimer's disease. Our results show that the use of soft Markov state models enables one to find a suitable state-space discretization to generate kinetic ensembles, a problem that has been challenging to date. Our findings also suggest that soft Markov models and clustering approaches based on neural networks could be applicable to other systems characterized by fast state transitions and conformational heterogeneity.

Methods

Simulation details. All simulations were carried out with GROMACS 2018.1[61] using 1,024 individual Virtual Machine instances on Google Compute Engine. All instances used the compute-optimized n1-highcpu-8 machine type, each configured with 8 Intel Haswell CPU cores, 7.2 GB of RAM, and 100 GB of disk space. 1,024 Aβ42 starting conformations were chosen from a previous metadynamics-biased molecular dynamics simulation by sample-weighted *k*-Means clustering on the space of backbone dihedral angles and picking a random conformation from each of the 1024 clusters. Each conformation was solvated separately in a rhombic dodecahedron box with a volume of 358 nm^3 using between 11,711 and 11,751 water molecules. The system was minimized using the steepest descent algorithm to a target force of less than 1000 kJ / (mol / nm). Equilibration was performed over a time range of 500 ps in the NVT ensemble using the Bussi thermostat[62] and 500 ps in the NPT ensemble using Berendsen pressure coupling[63] while applying a position restraint on all heavy atoms.

Production simulations were carried out at 278 K using the CHARMM22* force field[64] and the TIP3P water model[65] using a 2 fs timestep in the NVT ensemble. Electrostatic interactions were modelled using the Particle-Mesh-Ewald approach[66] with a cut-off for the short-range interactions of 1.2 nm. Constraints were applied on all bonds with the LINCS algorithm[67] using a matrix expansion on the order of 4 and 1 iteration per step. Simulations were carried out with the initial 1,024 starting conformations, the resulting trajectories were then used with the fluctuation amplification of specific traits (FAST) approach[39] to choose 1,024 new starting structures using Cα-distance matrices as a selection criterion. The clustering for FAST was conducted by first performing time-lagged independent component analysis (TICA)[15,16] and using the *k*-Means algorithm in this reduced space to create 128 clusters. This procedure was repeated three times to yield a total of 5,119 trajectories, with an aggregated simulation time of 315 µs. The shortest and longest trajectory lengths were 9.75 and 90.5 ns respectively. The trajectories were subsampled to 250 ps time steps to yield 1,259,172 frames.

Neural network. State decomposition was performed using the VAMPNet approach with physical constraints.[21,22] The two-lobed neural network takes as input a pair of frames separated by some lag time and yields a state assignment vector as output. We chose inter-residue nearest-neighbour heavy atom distance matrices with first- and second-degree off-diagonals removed as input (780 dimensions) and used between 2 and 8 output nodes with softmax activation. The architecture of the χ layers of the neural network follows the self-normalizing setup described by Klambauer et al.[44], using scaled-exponential linear units, normal LeCun weight initialization, and alpha dropout. The constrained part is composed of two additional layers u and S which are trained with and without the χ layers, successively, using the VAMP-e score[22]. Error estimates were obtained through bootstrap aggregation (bagging), i.e. by training 20 independent neural networks with independently randomized and shuffled 9:1 train-test splits. The model was implemented using Keras 2.2.4[68] with the Tensorflow 2.1.0[69] backend. The neural network parameters were chosen through two successive random grid searches with scikit-optimize 0.5.2[70], first using a coarse grid spanning a large parameter space, then a finer grid over a local space around the optimum parameters. We found the best parameters to be a layer width of 256 nodes, a depth of 5 layers, an L2 regularization strength of 10^{-8} and no dropout. Training was performed on a single Google Compute Engine instance with an NVidia Tesla V100 GPU, 12 Haswell cores, and 78 GB of RAM. Training of the χ model was performed using batch sizes of 10000 frame pairs and the Adam minimizer with a learning rate of 0.05, $\beta_2 = 0.99$ and $\varepsilon = 0.0001$. Overfitting was addressed through early stopping, i.e. training was stopped when the VAMP validation set score did not increase by at least 0.001 over the previous 5 epochs. The constraint layers u

and S as well as the full network including the χ layers were trained with the Adam minimizer with a learning rate of 0.0005. The constraint layers were trained with batches the size of the full dataset, with early stopping with no improvement in the loss after 10 epochs.

Analysis. The output of a single neural network is a state assignment vector $\chi(\boldsymbol{x_t})$ each frame x_t of the simulation. The ensemble averaged value of an observable $A(x_t)$ for a state i is therefore an average weighted by the state assignment for T time steps

$$\langle A_i \rangle = \left(\sum_t^T \chi_i(x_t) \right)^{-1} \sum_t^T \chi_i(x_t) A(x_t)$$

Ensemble averaged quantities can be calculated by first computing a weight w_t for each frame of the simulation based on the state assignment $\chi(\boldsymbol{x_t})$ at time t and the equilibrium distribution π

$$\mathrm{w_t} = \frac{\langle \chi(\mathrm{x_t}) | \pi \rangle}{\sum_{\mathrm{t=1}}^{\mathrm{T}} \langle \chi(\mathrm{x_t}) | \pi \rangle}$$

The ensemble averaged observable $\langle A \rangle$ can then be calculated as the weighted average

$$\langle A \rangle = \sum_{t=1}^{T} w_t A(x_t)$$

To estimate the error, we sort each state assignment vector, as multiple trained neural networks do not necessarily conserve the order of identified states. To do so, we calculate the mean inter-residue nearest-neighbour heavy atom distance matrix for each identified state and sort the states based on the lowest root-mean-square deviation between these matrices. We also ensure that the sorting is unique, i.e. we are never duplicating state assignments. The Koopman matrix K is then estimated directly from the constrained neural network.

Internal model validation is performed with the Chapman-Kolmogorov test (**Figures S4,S8,S9**)

$$K(n\tau) \approx K^n(\tau)$$

i.e. we expect a model estimated at some lag time τ to behave the same way as one estimated at a multiple $n\tau$ of it. The correct lag time τ for the model is estimated by plotting the relaxation (also known as implied) timescales t_i (**Figure 2A**)

$$t_i = \frac{-\tau}{\log \lambda_i}$$

where λ_i is the ith eigenvalue of the Koopman matrix K

$$K v_i = \lambda_i v_i$$

As the Koopman matrix $\boldsymbol{K}$ is row-stochastic, the largest eigenvalue is 1, and its associated eigenvector is the stationary distribution π of the system. The state lifetimes $\bar{t}_i$ are given by

$$\bar{t}_i = \frac{-\tau}{\log K_{ii}}$$

where K_{ii} are the diagonal entries of the Koopman matrix.

Chemical shifts were back calculated with CamShift[45] as implemented in PLUMED 2.4.1[71,72] using the full trajectories, averaged as described above and compared to previously recorded NMR data[31].

Conventional Markov state model. Markov state models based on discrete states were constructed as follows: Time-lagged independent component analysis was performed with inter-residue nearest-neighbour heavy atom distances as input, using a lag time of 5 ns. Various clustering methodologies were then evaluated using the 5-fold cross-validated generalized matrix Rayleigh quotient score using every 50th frame of the full trajectory (25183 frames) and a lag time of 12.5 ns[41]. Evaluation was performed over the number of time-lagged independent components, clusters, and algorithm (minibatch k-Means[73,74], minibatch k-Medoids[75], Gaussian mixture model) using MSMBuilder[76] (**Figure S1**). 100 Markov models with the highest scoring parameters were then sampled and the timescales evaluated for both the conventional Bayesian (**Figure S2A**)[12,77] and hidden Markov model cases (**Figure S2B-D**)[42].

Errors. Observable errors for each state were calculated by taking a trajectory sample based on the training data for a neural network sample. All statistics were then calculated on the bootstrap sample, and the errors reported in the figures show the 95th percentile of this sample unless noted otherwise.

Data availability. The datasets generated and analysed during the current study are available from a Google cloud-bucket at `gs://kinetic_ensemble_paper/`.

Code availability. The full analysis code in the form of Jupyter notebooks with example data is available from `https://github.com/vendruscolo-lab/ab42-kinetic-ensemble`.

Acknowledgements

GTH is supported by the Rosalind Franklin Research Fellowship at Newnham College. The authors would like to thank Dr. Steven Kearnes for useful feedback and suggestions for the manuscript and Dr. Massimilliano Bonomi for useful discussions and suggestions. The authors

also acknowledge support from Google and the Google Accelerated Science team for providing access to the Google Cloud Platform for simulations and analysis.

References

1. Heller, G. T., Sormanni, P. & Vendruscolo, M. Targeting disordered proteins with small molecules using entropy. *Trends Biochem. Sci.* **40**, 491–496 (2015).
2. Uversky, V. N. *et al.* Unfoldomics of human diseases: linking protein intrinsic disorder with diseases. *BMC Genomics* **10**, S7 (2009).
3. Babu, M. M., van der Lee, R., de Groot, N. S. & Gsponer, J. Intrinsically disordered proteins: regulation and disease. *Curr. Opin. Struct. Biol.* **21**, 432–440 (2011).
4. Sormanni, P. *et al.* Simultaneous quantification of protein order and disorder. *Nat. Chem. Biol.* **13**, 339–342 (2017).
5. Berman, H. M. *et al.* The Protein Data Bank. *Nucleic Acids Res.* **28**, 235–242 (2000).
6. Bonomi, M., Heller, G. T., Camilloni, C. & Vendruscolo, M. Principles of protein structural ensemble determination. *Curr. Opin. Struct. Biol.* **42**, 106–116 (2017).
7. Lindorff-Larsen, K., Best, R. B., DePristo, M. A., Dobson, C. M. & Vendruscolo, M. Simultaneous determination of protein structure and dynamics. *Nature* **433**, 128–132 (2005).
8. Boehr, D. D., Nussinov, R. & Wright, P. E. The role of dynamic conformational ensembles in biomolecular recognition. *Nat. Chem. Biol.* **5**, 789–796 (2009).
9. Fraser, J. S. *et al.* Accessing protein conformational ensembles using room-temperature X-ray crystallography. *Proc. Natl. Acad. Sci.* **108**, 16247–16252 (2011).
10. Bonomi, M. & Vendruscolo, M. Determination of protein structural ensembles using cryo-electron microscopy. *Curr. Opin. Struct. Biol.* **56**, 37–45 (2019).
11. Van Kampen, N. G. *Stochastic Processes in Physics and Chemistry*. (Elsevier, 2007). doi:10.1016/B978-0-444-52965-7.X5000-4.
12. Prinz, J.-H. *et al.* Markov models of molecular kinetics: Generation and validation. *J. Chem. Phys.* **134**, 174105 (2011).
13. Chodera, J. D. & Noé, F. Markov state models of biomolecular conformational dynamics. *Curr. Opin. Struct. Biol.* **25**, 135–144 (2014).
14. Husic, B. E. & Pande, V. S. Markov State Models: From an Art to a Science. *J. Am. Chem. Soc.* **140**, 2386–2396 (2018).

15. Pérez-Hernández, G., Paul, F., Giorgino, T., De Fabritiis, G. & Noé, F. Identification of slow molecular order parameters for Markov model construction. *J. Chem. Phys.* **139**, 015102 (2013).
16. Schwantes, C. R. & Pande, V. S. Improvements in Markov State Model Construction Reveal Many Non-Native Interactions in the Folding of NTL9. *J. Chem. Theory Comput.* **9**, 2000–2009 (2013).
17. Bowman, G. R. An Overview and Practical Guide to Building Markov State Models. in *An Introduction to Markov State Models and Their Application to Long Timescale Molecular Simulation* (eds. Bowman, G. R., Pande, V. S. & Noé, F.) vol. 797 7–22 (Springer Netherlands, 2014).
18. Kohlhoff, K. J. *et al.* Cloud-based simulations on Google Exacycle reveal ligand modulation of GPCR activation pathways. *Nat. Chem.* **6**, 15–21 (2014).
19. Voelz, V. A., Bowman, G. R., Beauchamp, K. & Pande, V. S. Molecular Simulation of ab Initio Protein Folding for a Millisecond Folder NTL9(1–39). *J. Am. Chem. Soc.* **132**, 1526–1528 (2010).
20. Klus, S. *et al.* Data-Driven Model Reduction and Transfer Operator Approximation. *J. Nonlinear Sci.* **28**, 985–1010 (2018).
21. Mardt, A., Pasquali, L., Wu, H. & Noé, F. VAMPnets for deep learning of molecular kinetics. *Nat. Commun.* **9**, 5 (2018).
22. Mardt, A., Pasquali, L., Noé, F. & Wu, H. Deep learning Markov and Koopman models with physical constraints. *ArXiv191207392 Phys.* (2019).
23. Cummings, J., Lee, G., Ritter, A., Sabbagh, M. & Zhong, K. Alzheimer's disease drug development pipeline: 2019. *Alzheimers Dement. Transl. Res. Clin. Interv.* **5**, 272–293 (2019).
24. Hardy, J. & Allsop, D. Amyloid deposition as the central event in the aetiology of Alzheimer's disease. *Trends Pharmacol. Sci.* **12**, 383–388 (1991).
25. Glenner, G. G. & Wong, C. W. Alzheimer's disease and Down's syndrome: Sharing of a unique cerebrovascular amyloid fibril protein. *Biochem. Biophys. Res. Commun.* **122**, 1131–1135 (1984).
26. Hardy, J. A. & Higgins, G. A. Alzheimer's disease: the amyloid cascade hypothesis. *Science* **256**, 184–185 (1992).
27. Knowles, T. P. J., Vendruscolo, M. & Dobson, C. M. The amyloid state and its association with protein misfolding diseases. *Nat. Rev. Mol. Cell Biol.* **15**, 384–396 (2014).
28. Knowles, T. P. J. *et al.* An analytical solution to the kinetics of breakable filament assembly. *Science* **326**, 1533–1537 (2009).

29. Jack, C. R. *et al.* NIA-AA Research Framework: Toward a biological definition of Alzheimer's disease. *Alzheimers Dement.* **14**, 535–562 (2018).
30. Cohen, S. I. A., Vendruscolo, M., Dobson, C. M. & Knowles, T. P. J. From macroscopic measurements to microscopic mechanisms of protein aggregation. *J. Mol. Biol.* **421**, 160–171 (2012).
31. Heller, G. T. *et al.* Small molecule sequestration of amyloid-β as a drug discovery strategy for Alzheimer's disease. *bioRxiv* 729392 (2019) doi:10.1101/729392.
32. Michaels, T. C. T. *et al.* Thermodynamic and kinetic design principles for protein aggregation inhibitors. *bioRxiv* 2020.02.22.960716 (2020) doi:10.1101/2020.02.22.960716.
33. Meng, F. *et al.* Highly disordered amyloid-β monomer probed by single-molecule FRET and MD simulation. *Biophys. J.* **114**, 870–884 (2018).
34. Lin, Y.-S., Bowman, G. R., Beauchamp, K. A. & Pande, V. S. Investigating how peptide length and a pathogenic mutation modify the structural ensemble of amyloid beta monomer. *Biophys. J.* **102**, 315–324 (2012).
35. Rosenman, D. J., Connors, C. R., Chen, W., Wang, C. & García, A. E. Aβ monomers transiently sample oligomer and fibril-like configurations: Ensemble characterization using a combined MD/NMR approach. *J. Mol. Biol.* **425**, 3338–3359 (2013).
36. Sgourakis, N. G. *et al.* Atomic-Level Characterization of the Ensemble of the Aβ(1–42) Monomer in Water Using Unbiased Molecular Dynamics Simulations and Spectral Algorithms. *J. Mol. Biol.* **405**, 570–583 (2011).
37. Nasica-Labouze, J. *et al.* Amyloid β Protein and Alzheimer's Disease: When Computer Simulations Complement Experimental Studies. *Chem. Rev.* **115**, 3518–3563 (2015).
38. Granata, D. *et al.* The inverted free energy landscape of an intrinsically disordered peptide by simulations and experiments. *Sci. Rep.* **5**, 15449 (2015).
39. Zimmerman, M. I. & Bowman, G. R. FAST Conformational Searches by Balancing Exploration/Exploitation Trade-Offs. *J. Chem. Theory Comput.* **11**, 5747–5757 (2015).
40. Hellerstein, J. L., Kohlhoff, K. J. & Konerding, D. E. Science in the Cloud: Accelerating Discovery in the 21st Century. *IEEE Internet Comput.* **16**, 64–68 (2012).
41. McGibbon, R. T. & Pande, V. S. Variational cross-validation of slow dynamical modes in molecular kinetics. *J. Chem. Phys.* **142**, 124105 (2015).
42. Noé, F., Wu, H., Prinz, J.-H. & Plattner, N. Projected and hidden Markov models for calculating kinetics and metastable states of complex molecules. *J. Chem. Phys.* **139**, 184114 (2013).
43. Wu, H. & Noé, F. Variational approach for learning Markov processes from time series data. *ArXiv170704659 Math Stat* (2017).

44. Klambauer, G., Unterthiner, T., Mayr, A. & Hochreiter, S. Self-Normalizing Neural Networks. *ArXiv170602515 Cs Stat* (2017).
45. Kohlhoff, K. J., Robustelli, P., Cavalli, A., Salvatella, X. & Vendruscolo, M. Fast and Accurate Predictions of Protein NMR Chemical Shifts from Interatomic Distances. *J. Am. Chem. Soc.* **131**, 13894–13895 (2009).
46. Löhr, T., Jussupow, A. & Camilloni, C. Metadynamic metainference: Convergence towards force field independent structural ensembles of a disordered peptide. *J. Chem. Phys.* **146**, 165102 (2017).
47. Bonomi, M., Camilloni, C. & Vendruscolo, M. Metadynamic metainference: Enhanced sampling of the metainference ensemble using metadynamics. *Sci. Rep.* **6**, 31232 (2016).
48. Roche, J., Shen, Y., Lee, J. H., Ying, J. & Bax, A. Monomeric Aβ 1–40 and Aβ 1–42 Peptides in Solution Adopt Very Similar Ramachandran Map Distributions That Closely Resemble Random Coil. *Biochemistry* **55**, 762–775 (2016).
49. Konrat, R. NMR contributions to structural dynamics studies of intrinsically disordered proteins. *J. Magn. Reson.* **241**, 74–85 (2014).
50. Dai, W., Sengupta, A. M. & Levy, R. M. First Passage Times, Lifetimes, and Relaxation Times of Unfolded Proteins. *Phys. Rev. Lett.* **115**, 048101 (2015).
51. Levy, R. M., Dai, W., Deng, N.-J. & Makarov, D. E. How long does it take to equilibrate the unfolded state of a protein? *Protein Sci.* **22**, 1459–1465 (2013).
52. Schütt, K. T., Sauceda, H. E., Kindermans, P.-J., Tkatchenko, A. & Müller, K.-R. SchNet – A deep learning architecture for molecules and materials. *J. Chem. Phys.* **148**, 241722 (2018).
53. Fout, A., Byrd, J., Shariat, B. & Ben-Hur, A. Protein Interface Prediction using Graph Convolutional Networks. 10.
54. Boomsma, W. & Frellsen, J. Spherical convolutions and their application in molecular modelling. in *Advances in Neural Information Processing Systems 30* (eds. Guyon, I. et al.) 3433–3443 (Curran Associates, Inc., 2017).
55. Chodera, J. D. & Noé, F. Probability distributions of molecular observables computed from Markov models. II. Uncertainties in observables and their time-evolution. *J. Chem. Phys.* **133**, 105102 (2010).
56. Olsson, S., Wu, H., Paul, F., Clementi, C. & Noé, F. Combining experimental and simulation data of molecular processes via augmented Markov models. *Proc. Natl. Acad. Sci.* **114**, 8265–8270 (2017).
57. Gerben, S. R., Lemkul, J. A., Brown, A. M. & Bevan, D. R. Comparing atomistic molecular mechanics force fields for a difficult target: a case study on the Alzheimer's amyloid β-peptide. *J. Biomol. Struct. Dyn.* **32**, 1817–1832 (2014).

58. Robustelli, P., Piana, S. & Shaw, D. E. Developing a molecular dynamics force field for both folded and disordered protein states. *Proc. Natl. Acad. Sci.* **115**, E4758–E4766 (2018).
59. Bowman, G. R. & Pande, V. S. Protein folded states are kinetic hubs. *Proc. Natl. Acad. Sci.* **107**, 10890–10895 (2010).
60. Dill, K. A. & Chan, H. S. From Levinthal to pathways to funnels. *Nat. Struct. Biol.* **4**, 10–19 (1997).
61. Abraham, M. J. *et al.* GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* **1–2**, 19–25 (2015).
62. Bussi, G., Donadio, D. & Parrinello, M. Canonical sampling through velocity-rescaling. *J. Chem. Phys.* **126**, 014101 (2007).
63. Berendsen, H. J. C., Postma, J. P. M., van Gunsteren, W. F., DiNola, A. & Haak, J. R. Molecular dynamics with coupling to an external bath. *J. Chem. Phys.* **81**, 3684–3690 (1984).
64. MacKerell, A. D. *et al.* All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J. Phys. Chem. B* **102**, 3586–3616 (1998).
65. Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W. & Klein, M. L. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* **79**, 926–935 (1983).
66. Essmann, U. *et al.* A smooth particle mesh Ewald method. *J. Chem. Phys.* **103**, 8577–8593 (1995).
67. Hess, B. P-LINCS: A Parallel Linear Constraint Solver for Molecular Simulation. *J. Chem. Theory Comput.* **4**, 116–122 (2008).
68. Chollet, F. Keras. (2015).
69. Abadi, M. *et al.* TensorFlow: A System for Large-Scale Machine Learning. in 265–283 (2016).
70. Head, T. *et al. Scikit-Optimize/Scikit-Optimize: V0.5Rc1*. (Zenodo, 2018). doi:10.5281/ZENODO.1157319.
71. Tribello, G. A., Bonomi, M., Branduardi, D., Camilloni, C. & Bussi, G. PLUMED 2: New feathers for an old bird. *Comput. Phys. Commun.* **185**, 604–613 (2014).
72. PLUMED consortium. Promoting transparency and reproducibility in enhanced molecular simulations. *Nat. Methods* **16**, 670–673 (2019).
73. Arthur, D. & Vassilvitskii, S. K-means++: The Advantages of Careful Seeding. in *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms* 1027–1035 (Society for Industrial and Applied Mathematics, 2007).

74. Sculley, D. Web-scale k-means clustering. in *Proceedings of the 19th international conference on World wide web - WWW '10* 1177 (ACM Press, 2010). doi:10.1145/1772690.1772862.
75. de Hoon, M. J. L., Imoto, S., Nolan, J. & Miyano, S. Open source clustering software. *Bioinformatics* **20**, 1453–1454 (2004).
76. Beauchamp, K. A. *et al.* MSMBuilder2: Modeling conformational dynamics on the picosecond to millisecond scale. *J. Chem. Theory Comput.* **7**, 3412–3419 (2011).
77. Trendelkamp-Schroer, B., Wu, H., Paul, F. & Noé, F. Estimation and uncertainty of reversible Markov models. *ArXiv150705990* (2015).

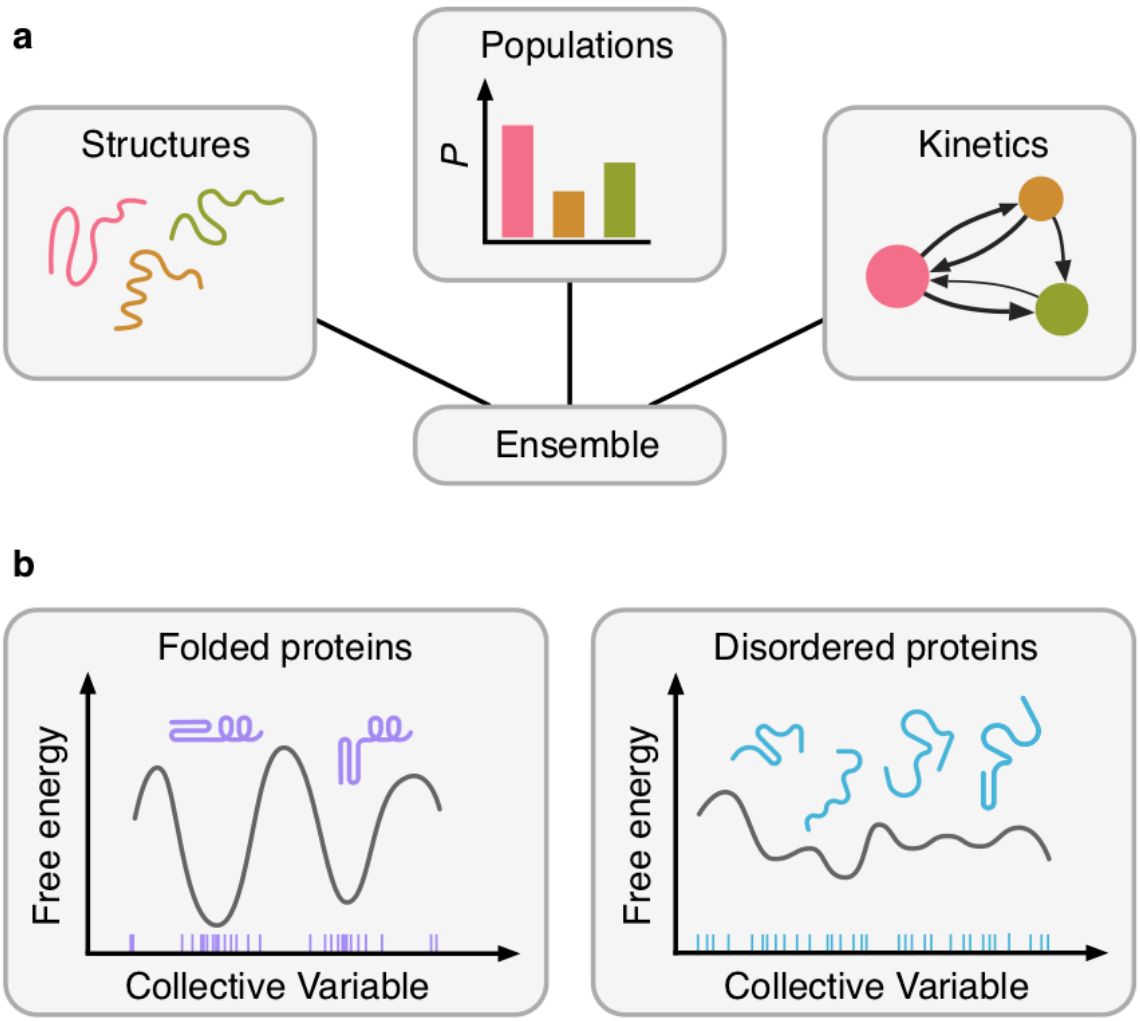

Figure 1. **Schematic illustration of a kinetic ensemble of a protein. (a)** A kinetic ensemble consists of three components: (1) structures, (2) their statistical weights (populations), and (3) interconversion rates between groups of related structures (states). Structural ensembles have only the first component and thermodynamic ensembles only the first two components. **(b)** The free energy landscape of disordered proteins is flatter and more heterogeneous than that of folded proteins, making a state decomposition difficult, and thus the determination of a kinetic ensemble challenging.

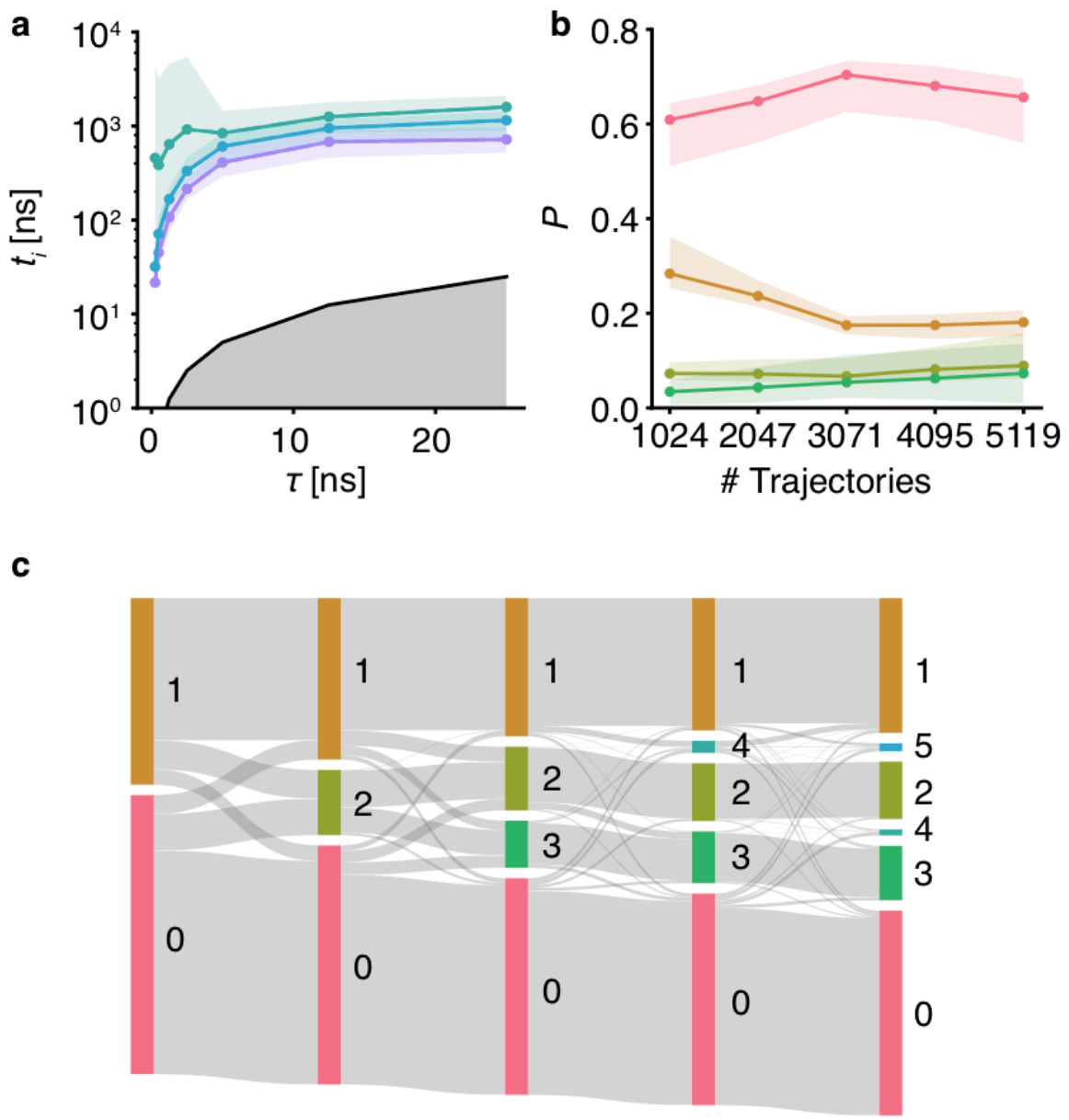

Figure 2. Determination of the states in the kinetic ensemble of Aβ42. (a) Dependence of the three longest relaxation timescales t_i (green, cyan and purple, respectively) on the lag time τ. These timescales are derived from the eigenvalues of the transition matrix (see Methods). The grey shading indicates the timescales for which the Koopman model can no longer resolve the relaxation timescales. **(b)** Dependence of the equilibrium distribution (i.e. the occupation probabilities of the 4 states) on the number of trajectories used to build the kinetic ensemble as a 4-state model. **(c)** Shift of equilibrium distributions and state assignments over different output sizes. Grey lines indicate state decomposition across multiple models with varying numbers of states, showing consistency across different output sizes. Shaded areas indicate 95th percentiles of the bootstrap sample of the mean.

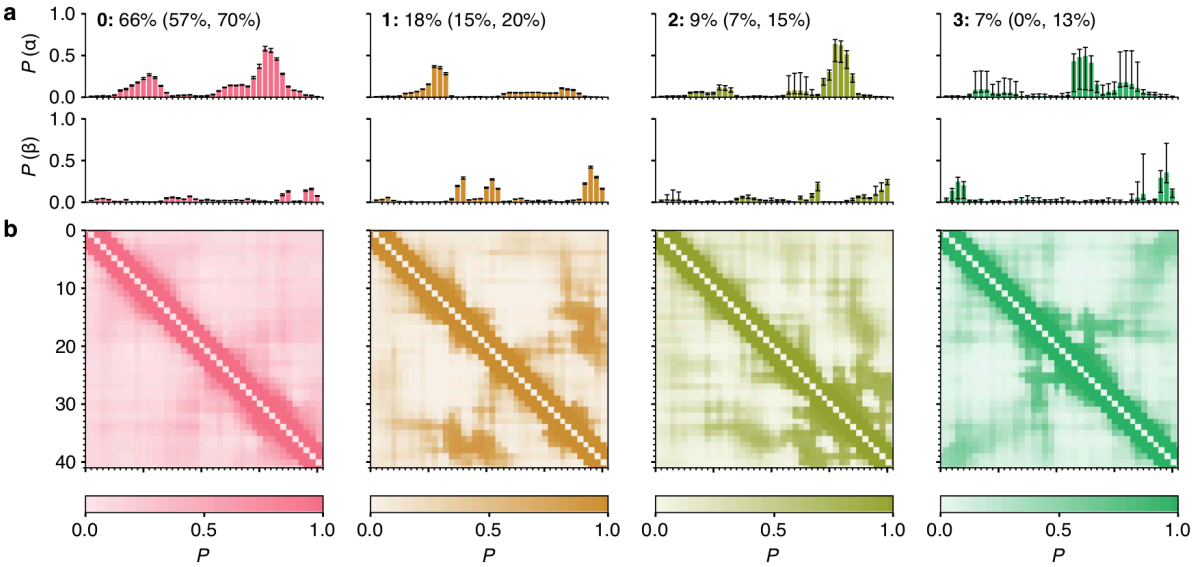

Figure 3. **Structural properties of Aβ42 in the kinetic ensemble. (a)** Populations of α-helical and β-sheet content per residue in each of the four states in the kinetic ensemble, as calculated using DSSP. The equilibrium percentage of each state is given above, with the 95th percentile in parentheses. **(b)** Contact probability maps of each of the four states with a cut-off of 0.8 nm. Error bars indicate 95th percentiles of the bootstrap sample of the mean.

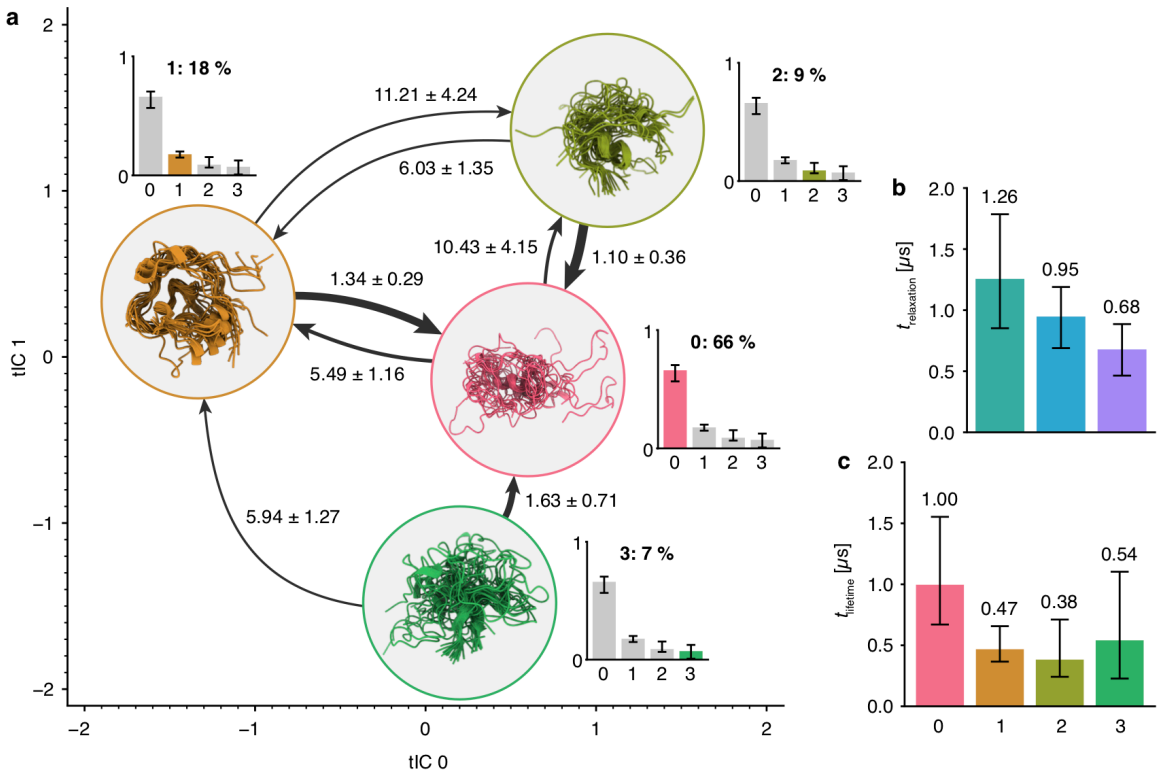

Figure 4. Populations and mean first-passage times in the kinetic ensemble of Aβ42. (a) Mean first-passage times (inverse transition rates) and their standard deviations in microseconds between states in the kinetic ensemble; thicker arrows correspond to faster transitions. The state location is projected on to the two slowest time-independent coordinates (tICs) and the structures shown are 20 high-weight conformations aligned on the most prominent secondary structure motifs (see **Figure 3A**). **(b)** Slowest relaxation timescales of the 4-state model. **(c)** Mean lifetime of each state in the 4-state model. Error bars indicate 95th percentiles of the bootstrap sample of the mean.