SUPPORTING INFORMATION FOR THE MANUSCRIPT:
Amyloid β Fibril Elongation by Monomers Involves Disorder at the Tip

Marco Bacci, Jiří Vymětal, Maja Mihajlovic, Amedeo Caflisch, and Andreas Vitalis

S1 Supporting Methods and Tables

S1.1 NMR model and preparation of the initial structures

The reference initial structure for our molecular dynamics (MD) simulations was derived from the first conformation of the Aβ42 protofibril bundle determined by solution NMR by Lührs *et al.*[1] The structure is composed of five chains and features two characteristic and structurally different ends, called *odd* and *even* end respectively (see Fig. 1B in the main text). The ends cannot be interconverted and thus constitute two different structural templates for the addition of monomers during fibril growth. This lack of symmetry has been related to the unidirectional growth of amyloid fibrils seen in experiments.[2, 3] In the original paper[1] it is hypothesized that the *odd end* is the fast-growing one based on results from Aβ–analog peptide inhibitors of amyloid growth. Conversely, a recent *in silico* study[4] constructs a kinetic model to show that the binding to the *even end* should be favored.

Excluding the first 17 residues, the model by Lührs *et al.* comprises a *β strand – loop – β strand* architecture at the monomer level. Adjacent chains are engaged in two parallel in-registry β-sheets, with β1 encompassing residues 18-26 (N-terminal side) and β2 residues 31-42 (C-terminal β–sheet), which are further stabilized by staggered side chains contacts between the two sheets. This model has been widely studied in recent years through computer simulations[5, 6, 7, 8, 9, 10, 11, 12] and has proven to be rather stable under the action of modern force-fields,[13, 14, 15] which is not necessarily the case for other architectures.[16, 17] We manually added the missing disordered residues of the N-terminal parts to each of the five chains. These tails were subsequently equilibrated with 30 million Monte Carlo steps consisting of backbone and side chain dihedral angle pivot moves following standard methodology.[18] The conditions for this equilibration are not particularly important (310 K, partial charges and bonded potentials of the recent CHARMM36 force field,[19] ABSINTH implicit solvent model),[20] but it is crucial that the rest of the model was held completely rigid. The final conformation was solvated in a cubic box containing ~4.8×10^4 charmm-TIP3P water molecules and 0.15 M monovalent ions before being energy-minimized and relaxed through an NPT (1 atm, 310 K) simulation of ~1 ns, which was performed in GROMACS[21] and returned a final cubic box dimension of 113.7 Å per edge. We prepared a second initial structure from the previous one by manually removing about 20 protofibril-internal water molecules and by pulling apart a particularly long-lived salt bridge formed by the N-terminus of chain D and the C-termini of chains E and D with an external force. During the pulling, the rest of the protofibril was frozen but the solvent was mobile. We refer to this second initial structure by adding a * to the relevant simulations in Tables S1 and S2 below.

S1.2 Simulations

S1.2.1 General settings

All the simulations that are described below shared the following basic setup. We simulated in the canonical (NVT) ensemble, *i.e.*, at constant volume. The system was contained in a cubic periodic box and integrated with 2 fs integration time step at 310 K, coupled with the velocity-rescaling thermostat[22] and a decay time of 2 ps. All native bonds were constrained with SHAKE[23] at force-field reference values. We used generalized reaction-field corrections[24] to the Coulomb potential with a cutoff of 12 Å and a 16 fs update interval for the (buffered) neighbor lists. To alleviate an expected lack of stability on long timescales of the protofibril model due to the limited number

of chains we included, we biased the ϕ and ψ angles of residues 18-24, 26, 31, 32, 34-36 and 39-41 of chains B, C, D and E, *viz.* of all the chains except the terminal one at the *odd end* (chain A) with 2D Gaussian wells (see Fig. 1C-D in the main text) with a universal depth of 1 kcal/mol and residue-specific width parameters. These parameters were determined by a short preliminary simulation initiated from the aforementioned (first) starting structure. The residues not included are those in the N-terminus, the loop, or those that exhibited ϕ/ψ heterogeneity in the trial simulation for one or more of the chains. All simulations used the CHARMM36 force field,[19] and trajectories were appended every 2 ps.

S1.2.2 PIGS and conventional sampling simulations

The first initial conformation (see S1.1) was used as the starting point of an initial **P**rogress **I**ndex-**G**uided **S**ampling[25] or PIGS simulation. All the PIGS simulations were run with CAMPARI (http://campari.sourceforge.net/). As outlined briefly in the main text (see Methods) and in detail in the reference publication,[25] PIGS proceeds as a series of termination and reseeding cycles of multiple copies of a system evolving stochastically under the same Hamiltonian. The termination and reseeding decisions rely on a way to represent the system, and here we used a subset of backbone ψ angles of chain A for this purpose. A PIGS simulation will ultimately diversify a given representation, which corresponds to enhanced conformational exploration on the unbiased potential energy surface of the system. The choice of representation is what enables us to focus the exploration on a particular subset of the system without having to resort to a low-dimensional reaction coordinate. Here, we always employed 32 copies in a PIGS run and used the final snapshots of the 16 top-ranked copies as candidates for reseeding the final snapshots of the bottom-ranked 16 copies at 2 ps intervals. PIGS retains no memory beyond a single 2 ps stretch, and the ranking is high for copies occupying regions of phase space that feature low sampling density and are unique with respect to the other copies across an individual stretch.

As expected, once the copies of the simulated system have all diversified in the chosen representation, the reseeding rate drops (see Figure S2), and the PIGS protocol approaches conventional sampling. This behavior can be altered by changing the number of copies. Here, it proved impossible to increase this number far enough due to limitations in available computing resources. Alternatively, a change in representation provides an attractive means to focus on particular degrees of freedom as others may have lost their importance throughout the PIGS run. Ongoing research is devoted to the automatic weighting of degrees of freedom[26] to yield a sufficiently informative representation at any given reseeding point. Lastly, new PIGS runs can be created from intermediate points deemed interesting visually and/or in terms of low-dimensional projections. This is similar in spirit to trajectory swarm[27] or the recent WExplore methods.[28] We made extensive use of both of the latter approaches to enhance the exploration further, and these modifications are summarized in Tables S1 and S2. For example, the very first simulation mentioned at the beginning of this section is termed *PigsA* in Tables S1 and S2, and involved 32 replicas covering ~10.5 ns each.

Our entire data set can be roughly partitioned into two halves. The first half consists of the successive iteration of *PigsA* along with two additional PIGS runs, named *PigsB* and *PigsC*. These started from the same initial structure and served primarily to increase and evaluate robustness in terms of coverage of the conformational space in the vicinity of the starting conformation. The second half of the data set, *i.e.*, a further ~5 μs of cumulative simulation time include conventional sampling runs with GROMACS from both initial structures (*Gromacs* and *Gromacs** in Tables S1 and S2), three PIGS runs from the second initial structure (see S1.1), named *PigsA*, PigsB** and *PigsC**, and two conventional runs from two interesting structures that were identified during the successive iterations

of *PigsA*. These runs were termed *Gromacs0* and *Gromacs21*, and, like all GROMACS runs, they used 16 independent copies each and were able to emulate the same Gaussian well potentials used in all other PIGS runs thanks to the GROMACS built-in support for CMAP[29] and PLUMED.[30] The reasons behind this additional sampling effort were twofold: i) to increase statistical robustness; ii) to test for the influence of the particular initial condition we had used. One-to-one comparisons of results derived from just the first half or the entire data set are used for both tasks.

PigsA	Pigs0 (1)	Pigs0_24 (1)			
		Pigs0_29 (1)	Pigs0_29_27C (1)		
			Pigs0_29_27L (1)		
			Gromacs0 (1)		
		Pigs0_14C (1)			
		Pigs0_14L (1)			
	Pigs3 (1)				
	Pigs8 (1)				
	Pigs10 (1)				
	Pigs12 (1)				
	Pigs21 (1)	Pigs21LV (4)			
		Pigs21V (4)	Pigs21VL (8)	Pigs21VL_CL (1)	
				Pigs21VL_L (32)	
				Pigs21VL_Lc (2)	
				Pigs21VL_Lct (32)	Pigs21VL_Lct_C (1)
					Pigs21VL_Lct_L (1)
		Gromacs21 (1)			
	Pigs23 (1)	Pigs23_29C (1)	Pigs23_29C_16L (1)		
		Pigs23_29L (1)			
		Pigs23_4 (1)			
		Pigs23_9 (1)			
	Pigs25 (1)				
	Pigs27 (1)	Pigs27_0 (1)			
		Pigs27_25 (1)			
		Pigs27_25FC (1)			
		Pigs27_11 (1)	Pigs27_11m (4)	Pigs27_11m_25 (1)	
				Pigs27_11m_13 (1)	
			Pigs27_11m_fC (4)		
PigsB					
PigsC					
PigsA*					
PigsB*					
PigsC*					
Gromacs					
Gromacs*					

Table S1: Summary of the independent simulations we performed*. All PIGS simulations involved 32 copies, and all conventional (GROMACS) simulations employed 16 (independent) copies. Simulations are ordered from left to right in parent-child relationships, meaning that a simulation in column N was started from one of the replicas of the simulation in column N-1 (column 1 is leftmost and contains the simulations that were started from either one of the two initial structures (* as superscript refers to the second one). The precise number of parent replicas used to restart a specific run is annotated within parentheses. For example, run Pigs21LV was restarted from 4 final snapshots among the set of replicas of run PIGS21, and those were evenly distributed between the 32 structures needed to initiate a PIGS run. The simulations that form the "first half" of the data set mentioned in S1.3 are highlighted in blue.*

PigsA, PigsB, PigsC, PigsA*, PigsB*, PigsC* (0)					
Sim. Time [ns]	0 : 10.5				
Representation	19 20 21 22 23 24 26 31 32 34 35 36 39 40 41				
Pigs0 (1,4)					
Sim. Time [ns]	0 : 7	7 : 12.6	12.6 : 15.3	15.3 : 17.5	
Representation	31 32 34 35 36 39 40 41	19 20 21 22 23 24 26 31 32 34 35 36 39 40 41	17 18 19 20 21	17 18 19 20 21 39 40 41	
Pigs0_24 (0)					
Sim. Time [ns]	0 : 2.3				
Representation	17 18 19 20 21 30 31 32				
Pigs0_29 (0)					
Sim. Time [ns]	0 : 2.3				
Representation	17 18 19 20 21 30 31 32				
Pigs0_14C (0)					
Sim. Time [ns]	0 : 1.5	1.5 : 2.3			
Representation	30 32 36	30 32 36 40			
Pigs0_14L (0)					
Sim. Time [ns]	0 : 2.3				
Representation	17 18 19 20				
Pigs0_29_27C (0)					
Sim. Time [ns]	0 : 2.3				
Representation	34 35 36 39 40 41				
Pigs0_29_27L (1,1)					
Sim. Time [ns]	0 : 2.3	2.3 : 4.5			
Representation	17 18 19 20 21	20			
Pigs3 (0)					
Sim. Time [ns]	0 : 2.4	2.4 : 3.2			
Representation	19 20 21 22 23 24 26 31 32 34 35 36 39 40 41	17 18 19 20 21 32 34 35 36 39 40 41			
Pigs8 (2,1,1)					
Sim. Time [ns]	0 : 2.8	2.8 : 3.6	3.6 : 4.4	4.4 : 5.2	
Representation	19 20 21 22 23 24 26 31 32 34 35 36 39 40 41	17 18 19 20 21 30 31 32	18 20	17 18 20	
Pigs10 (0)					
Sim. Time [ns]	0 : 2.8				
Representation	19 20 21 22 23 24 26 31 32 34 35 36 39 40 41				
Pigs12 (0)					
Sim. Time [ns]	0 : 7	7 : 9.9			
Representation	31 32 34 35 36 39 40 41	19 20 21 22 23 24 26 31 32 34 35 36 39 40 41			
Pigs21 (2,4,4)					
Sim. Time [ns]	0 : 6.6	6.6 : 10.8	10.8 : 14.7		
Representation	19 20 21 22 23 24 26 31 32 34 35 36 39 40 41	17 18 19 20 21 32 34 35 36 39 40 41	24 34 35 36		
Pigs21LV (0)					
Sim. Time [ns]	0 : 2.3				
Representation	17 18 19 20 21 24				
Pigs21V (0)					
Sim. Time [ns]	0 : 1.5				
Representation	24				
Pigs21VL (3,1,1,1)					
Sim. Time [ns]	0 : 0.8	0.8 : 1.5	1.5 : 3.1	3.1 : 3.9	3.9 : 4.7
Representation	24	17 18 19 20 21 24	17 18 19 20 21	31 32	32

Pigs21VL_CL (0)					
Sim. Time [ns]	0 : 0.8				
Representation	17 18 19 31 32				
Pigs21VL_L (0)					
Sim. Time [ns]	0 : 2.3				
Representation	17 18 19 20 21				
Pigs21VL_Lc (0)					
Sim. Time [ns]	0 : 2.3				
Representation	17 34 35 36				
Pigs21VL_Lct (0)					
Sim. Time [ns]	0 : 1.4				
Representation	34 35 36				
Pigs21VL_Lct_C (0)					
Sim. Time [ns]	0 : 0.8	0.8 : 1.5			
Representation	34 35 36	31 32 34 35			
Pigs21VL_Lct_L (0)					
Sim. Time [ns]	0 : 2.3				
Representation	17 19 21				
Pigs23 (2,1,1)					
Sim. Time [ns]	0 : 0.8	0.8 : 1.3	1.3 : 2.1	2.1 : 2.9	
Representation	17 18 19 20 30 31 32 35 39 40 41	30 31 32 34 35 39	30 31 32 34 35 36	18 20 30 31 32 35	
Pigs23_29C (1,1)					
Sim. Time [ns]	0 : 0.8	0.8 : 1.5			
Representation	30 31 32 34 35 36	36 39 40 41			
Pigs23_29C_16L (0)					
Sim. Time [ns]	0 : 0.8				
Representation	17 18				
Pigs23_29L (0)					
Sim. Time [ns]	0 : 0.8				
Representation	17 18 19 20				
Pigs23_4 (0)					
Sim. Time [ns]	0 : 0.8				
Representation	18 20 30 31 32				
Pigs23_9 (0)					
Sim. Time [ns]	0 : 1.5				
Representation	18 20 31 32 35 36				
Pigs25 (1,1)					
Sim. Time [ns]	0 : 1.4	1.4 : 2.3			
Representation	19 20 21 22 23 24 26 31 32 34 35 36 39 40 41	17 19 21 24 32 34 35 36			
Pigs27 (0)					
Sim. Time [ns]	0 : 7.2	7.2 : 10			
Representation	19 20 21 22 23 24 26 31 32 34 35 36 39 40 41	17 18 19 20 21 32 34 35 36 39 40 41			
Pigs27_0 (0)					
Sim. Time [ns]	0 : 2.3	2.3 : 3.7			
Representation	34 35 36 39 40 41	39 40 41 (chain B)			
Pigs27_25 (0)					
Sim. Time [ns]	0 : 1.5				
Representation	17 18 19 20 34 35 36 39 40 41				
Pigs27_25FC (0)					
Sim. Time [ns]	0 : 0.8				
Representation	34 35 36 39 40 41				
Pigs27_11 (2,1,1)					
Sim. Time [ns]	0 : 2.8	2.8 : 5.8	5.8 : 6.6		
Representation	17 18 19 20 21 32 34 35 36 39 40 41	34 35 36 39 40 41	34 36		

Pigs27_11m (0)					
Sim. Time [ns]	0 : 3				
Representation	17 18 19 20 34 35 36 39 40 41				
Pigs27_11m_25 (0)					
Sim. Time [ns]	0 : 0.8	0.8 : 3.1			
Representation	17 18 19 20 34 35 36 39 40 41	39 40 41 (chain B)			
Pigs27_11m_13 (0)					
Sim. Time [ns]	0 : 0.8				
Representation	34 35 36 39 40 41				
Pigs27_11m_fC (0)					
Sim. Time [ns]	0 : 1.5	1.5 : 3.2			
Representation	34 35 36 39 40 41	39 40 41			
Conventional sampling					
Gromacs [ns]	60				
Gromacs* [ns]	60				
Gromacs0 [ns]	58				
Gromacs21 [ns]	58				

***Table S2: Representations and time extents of the simulations**. We list the ψ angles of the residues that were used as representations of the system in a specific time interval during a given PIGS run. Residues are identified by their position along the sequence of the Aβ42 monomer. For example, annotation '17 18' means that the ψ angles of LEU17 and VAL18 of chain A were used for the specified PIGS run and time window. The latter refers to the sampling time per copy. In addition, next to the name of a PIGS run, we report an annotation of the type $(N,X_1,..,X_N)$, where N indicates the number of manual reseedings that we performed during the relevant PIGS simulation and X_i informs on the number of independent replicas that were used for the manual reseeding i. For completeness, we also provide the simulation time of the conventional sampling runs at the end of the list where no representation was used. In this case, we always used 16 replicas per simulation. The simulations that form the "first half" of the data set are highlighted in blue.*

S1.3 Analysis

S1.3.1 Preamble

The data set we analyze is formed by many trajectories of variable length, ranging from 0.8 to 60 ns. A natural framework for treating ensembles of trajectories is that of *Markov state models* (MSMs),[31] the popularity of which has been growing considerably in the community in recent years.[32, 33] In theory, MSMs allow the computation of stationary distributions and implied timescales from ensembles of even very short trajectories by discretization (coarse-graining). The resultant network is able to describe a memoryless (Markovian) evolution in this space. Rather than attempting to achieve global equilibrium by means of few, extremely long trajectories, the fulfillment of local equilibrium becomes the convergence requirement for trajectory ensembles.[32, 34, 35, 36] However, MSMs have several pitfalls. For example, the approach requires discretization, and the usual workflow is a two-step process. First, molecular conformations are grouped into clusters at a fine structural resolution, followed by lumping states together. The second step often relies on spectral clustering algorithms and aims for a model with comparatively few states that are human-comprehensible. The level of scrutiny with regards to the first step is often unsatisfactory, however. The extent to which a memoryless stochastic process in a discrete space can correctly describe and interpret MD simulations is a separate issue. Few *a posteriori* Markovianity tests have been developed, but generally speaking Markovianity is more an assumption than a fully testable hypothesis for the analysis of MD simulations. Efforts to improve the suitability of the modeling are directed towards the optimization of the clustering parameters and the usage of larger lag times.[32, 37]

Transition path theory (TPT) builds on the framework of MSMs to return fluxes of probabilities between states, which can be used to compute kinetic rate constants and to extract representative transition pathways in order to gain mechanistic insights into molecular processes.[38] Two practically distinct but formally equivalent approaches have been developed, one that relies on rate matrices[39] and another that directly uses transition matrices, which we turn to.[40] Transition matrices are more straightforward to extract from a raw data set and preferred unless there are reasons preventing a clean definition of the connectivity between states, *e.g.*, in generalized ensemble simulations. The accurate inference of transition counts at different lag times from a heterogeneous set of PIGS and conventional simulations was implemented in the latest version of CAMPARI, which is freely available upon request. We linked the routines *EB13* and *MA48* of the linear algebra library HSL (*HSL. A collection of Fortran codes for large scale scientific computation. http://www.hsl.rl.ac.uk/*), which offers sparse matrix support, to CAMPARI in order to perform the spectral decomposition of the transition matrices and to solve the linear system in eq. S2 below. Users will have to obtain a copy of the HSL library independently. Other analyses were scripted in R or performed as built-in features within CAMPARI, *e.g.*, DSSP analysis,[41] principal component analysis,[42] cut-based free energy profiles,[43] *etc.*

S1.3.2 Clustering

As expressed above, the MSM framework mandates the definition of states, *viz*. groups of similar snapshots. To this aim, we used a tree-based clustering algorithm[44] implemented in CAMPARI, which produces clusters that are free of overlap and track local sampling density well. We represented molecular conformations by a set of 161 interatomic distances (D-RMSD), which are listed in Table S3 together with other clustering parameters. This choice provides an important advantage: it is not obviously correlated with the PIGS representations (dihedral angles). As a consequence, diversity in the clustering is not a direct consequence of PIGS. Furthermore, interatomic distances were also chosen in the study of Han and Schulten,[4] thus enabling comparison to their results. As is evident from Table S3, the set of interatomic distances captures not only the A-B interface but all interfaces (inter- and intramolecular ones), which provides information on the stability of the entire pentamer.

Region of the protofibril	Interfaces between adjacent chains (C_α to C_α at: A & B, B & C, C & D, D & E)	Adjacent residues (C_α to C_α intramol.)	Opposite residues (C_β to C_β or C_α to C_β for G25-I32, intramolecular)	
Involved atoms	V18,F19,F20,A21,E22,D23,V24,G25, S26,N27,L28,G29,A30,I31,I32,G33, L34,M35,V36,G37,G38,V39,V40,I41	F19-A21-A23-D23-G25-I32-L34-V36-G38-V40	F19–V40, A21-V36, D23-L34, G25-I32	
Clustering parameters				
Resolution at root	Resolution at leaves [Å]	Number of levels	Clusters (half data set ~2.6 million snapshots)	Clusters (full data set ~5.2 million snapshots)
8.0	1.0	15	2809	5655
7.9	0.9	15	4525	8654
7.8	0.8	15	7601	14119
7.7	0.7	15	13989	24404

Table S3: Definition of the interatomic distances used for the structural grouping, clustering parameters, and associated number of clusters*. For the definition of the two data sets, please refer to Section S1.3 and Tables S1 and S2.*

The selection of a set of degrees of freedom to group a set of snapshots into states is a fundamental but somewhat arbitrary choice that can contribute dramatically to the accuracy of any subsequent analysis.[44, 45] To dispel some of the concerns in this regard, Fig. S5 illustrates that the chosen metric partitions the data meaningfully also with respect to another metric, *viz.*, positional root mean square deviation (RMSD) with alignment based on the N and

O backbone atoms of all the Aβ42 residues from V18 to I41. Specifically, the D-RMSD-derived clusters retain structural homogeneity and tight radii also in terms of positional RMSD.

S1.3.3 Markov state models (MSMs)

For each of the different clustering resolutions listed in Table S3, we calculated the associated row-normalized transition matrix *T*, the element of which is defined as $T_{ij}(\tau) = c_{ij}(\tau) / \sum_{j=1}^{N} c_{ij}(\tau)$ where c_{ij} is the number of transitions from state *i* to state *j* at lag-time τ, and *N* is the number of clusters. Transitions were counted in a way that accounts for *all* connectivity changes introduced by the PIGS protocol and the manual changes summarized in Table S2. The element T_{ij} of the transition matrix expresses the conditional probability of jumping from state *i* to state *j* within a time-interval equal to the lag-time τ given that the system is initially in state *i*. The first left-eigenvector of *T*, which corresponds to the eigenvalue 1, informs on the steady-state probability π_i of each state, while the other eigenvectors contain information about the states involved in structural transitions (modes), the implied timescales of which can be computed as:

$$t_i(\tau) = -\tau / \ln \lambda_i(\tau) \qquad \text{(S1)}$$

In eq. S1, λ_i is the i^{th} eigenvalue that corresponds to the $(i\text{-}1)^{th}$ slowest transition, where we assume that the eigenvalues have been sorted in decreasing order. It is common to impose microscopic reversibility on the transition network, *i.e.*, $\pi_i T_{ij} = \pi_j T_{ji}$. Along with efforts to estimate statistical errors associated with predictions obtained from a particular MSM, advanced methods to impose this detailed balance condition have been developed.[32, 46] Here, we generally followed a simple protocol of symmetrizing the transition counts with one exception as noted in the next paragraph. For lag times larger than the saving frequency, it is possible to either extract multiple independent but correlated trajectories or to use the *sliding window* method, which simply adds all the individual count matrices to construct the maximum likelihood estimate of the transition matrix.

As a way to assess the appropriateness of MSMs, it is customary to look at the trends of the slowest implied timescales with lag time. A plateau with increasing lag time is a necessary but not sufficient condition for the network to be Markovian, *i.e.*, to fulfill $T(k\tau) = T^k(\tau)$. These analyses therefore allow making informed choices of both the lag time and the clustering resolution on which to perform the flux analysis and the extraction of the transition pathways. Fig. S6 shows the trends of the four slowest implied timescales for the MSMs at the four different clustering resolutions (Table S3). It is apparent that: i) the values plateau somewhat at increasing lag time in a way that does not depend strongly on clustering resolution; ii) the sliding-window result agrees well with the average of the results derived from individual trajectories at fixed lag time. Based on these data, we selected a resolution of 0.7 Å and a lag-time of 200 ps. To see whether a detailed balance-constrained maximum likelihood estimate of the transition matrix[46] would give results that are substantially different, we repeated the analysis in Fig. S6 for the sliding window case. Fig. S7 implies that the kinetic information in the networks is very similar for both symmetrization approaches, and we thus restricted all subsequent analysis to the much simpler one.

Finally, as a test of robustness, Fig. S8 provides the same results as Fig. S6 when considering only the first half of our data set (see Section S1.2.2 and Table S1). From Fig. S8, we would have decided to proceed with a resolution

of 0.8 Å and a lag-time of 200 ps for the extraction of the kinetic rates and transition paths, which is very similar to the choices based on Fig. S6. Indeed, the implied timescales are similar (within a factor of 2-3).

S1.3.4 Transition path theory (TPT)

As already stated, TPT connects naturally with MSMs by means of the underlying transition matrix. TPT allows one to extract kinetic and structural pathway information with the help of committor probabilities.[40] Pathways are defined by providing two sets of states as input, for example a disordered (D) and an ordered (O) set with the remainder of the states being part of the set of possible intermediates (I). The forward time (+) committor probability of any putative intermediate state *i*, $p^+_{fold_i}$, is defined as the probability that, by evolving a random walker according to *T* from state *i*, the ordered set is reached before the disordered one. Similarly, the backward (-) committor probability, $p^-_{fold_i}$, is formally the probability that a walker that reaches state *i* was last in the disordered state rather than in the ordered one. The values of $p^+_{fold_i}$ can be computed by solving the following set of equations:

$$-p^+_{fold_i} + \sum_{j \in I} T_{ij} p^+_{fold_j} = -\sum_{j \in O} T_{ij} \text{ for } i \in I \tag{S2}$$

Because we impose detailed balance (see S1.3.3), the values for $p^-_{fold_i}$ are simply equal to $1 - p^+_{fold_i}$. Once probabilities and committors were calculated for all states, we obtained the effective flux between states *i* and *j* as $f_{ij} = \pi_i p^-_{fold_i} T_{ij} p^+_{fold_j}$, the reactive flux as $f^+_{ij} = \max[0, f_{ij} - f_{ji}]$, and finally the total flux as:

$$F = \sum_{i \in D} \sum_{j \notin D} \pi_i T_{ij} p^+_{fold_j} \tag{S3}$$

From the total flux, the kinetic rate constant can be derived: $k_{DO} = F / (\tau \sum_{i=1}^{N} \pi_i p^-_{fold_i})$. If detailed balance holds, it is easy to see that transitions with nonzero reactive flux imply an increase in p^+_{fold}:

$$f^+_{ij} = \max[0, f_{ij} - f_{ji}] = \max[0, p^+_{fold,j} - p^+_{fold,i}] \tag{S4}$$

S1.3.5 Selection of reference states

The extraction of kinetic rate constants and transition pathways requires the definition of two sets of states, here an ordered (or locked) set and a disordered (or docked) set. As long as there is a separation of timescales for the transition versus local equilibration times, the precise definitions of sets in terms of clusters (states) have little impact on the final result. Here, we wish to characterize the locking step during elongation of the Aβ42 protofibril model by Lührs *et al.*[1] We defined the ordered set of states by identifying conformations that are an appropriate representation of the "relaxed" reference model under simulation conditions. The disordered set was meant to be as diversified and kinetically far away as possible from the starting structure.

Because we wanted to be able to estimate robustness by rerunning the analysis on the first half of the data alone, these sets were defined without consideration of the latter half of the data (see S1.2). We analyzed the

distribution of D-RMSD distances to the centroid representative of the cluster that contained the initial structure. This analysis is based on a clustering resolution of 0.8 Å (see S1.3.2 and Table S3). Fig. S21A shows the distances to the centroids of the clusters that were populated during *PigsA*, *PigsB*, and *PigsC* (Table S1). The associated histogram in Fig. S21B was used to set an upper threshold distance selecting a subset of possible candidates for the ordered set. Similarly, for the disordered set, we looked at the distances to the same centroid of all the clusters populated in all the other PIGS simulations of the first half of the data set, and this is reported in Figs. S21C-D. The pool of candidates falling above a lower threshold distance in this case came from three specific sets of PIGS runs, *i.e.*, *Pigs0*, *Pigs21*, and *Pigs27*, all including their respective derivatives (see Table S1). Candidates from these 3 runs were ultimately used to identify D2, D1, and D3, respectively (see main text and Fig. S11). To select a smaller number of states from the pool of candidates for the ordered set, we focused on the statistical weights of the candidate states and on the degree of homogeneity in their sampling extent with respect to runs *PigsA*, *PigsB*, and *PigsC*. Fig. S22 reports discovery times and sampling weights for the set of candidate states, and the four states depicted by their centroids in Fig. S10 were selected based on the data in Fig. S22. These four selected centroids also belong to four distinct clusters at the next coarser resolution for clustering, *viz.* 0.9 Å. In all cases, parent clusters in other groupings, in particular in the final one at 0.7 Å on the full data set, were identified as follows: we extracted the centroid snapshots of D1-D3 and the four members of the ordered set identified by the original grouping and simply found the corresponding clusters they belong to in the grouping in question.

S1.3.6 Flux decomposition

The net reactive probability flux from the disordered (D) to the ordered state (O) is known (S1.3.4) and can be iteratively decomposed into pathways. Random productive trajectories are those that start in D and reach O before crossing D again. Unfortunately, if the set of intermediates is finite, the number of possible productive trajectories will grow uncontrollably. It is therefore common to lump productive trajectories by the "reactive portion" of these trajectories, *i.e.*, by the sequences of states for which the committor probabilities, $p^+_{fold_{i,path}}$, increase monotonously. Then, the decomposition problem can be rephrased as a problem of finding shortest paths in a flux network where transitions between nodes *i* and *j* are allowed if $p^+_{fold_j} > p^+_{fold_i}$. The length of the corresponding edge (distance) is given by $d_{ij} = \ln\left[\left(\sum_{k=1}^{n_i} f^+_{ik}\right) / f^+_{ij}\right]$ [39]. Here, n_i is the number of states connected to node *i*. Shortest paths can be found conveniently with the help of Dijkstra's algorithm.[47]

The carried flux associated with any given pathway can be computed rigorously as $f^+_{path_{i_0,..,i_k,..,i_L}} = f^+_{i_0 i_1} \Pi_{k=1}^{L-1} (f^+_{i_k i_{k+1}} / (\sum_{j=1}^{n_{i_k}} f^+_{i_k j}))$, where i_0 belongs to D and *L* is the number of states that compose the pathway. In theory, it is possible to enumerate all possible pathways by subtracting their carried flux from any edge contributing to the pathway until all edges are exhausted. Unfortunately, this carried flux is generally several orders of magnitude smaller than the smallest (bottleneck) net flux of any edge that is on-pathway. This is because even in a network of just moderate complexity the net reactive flux of an edge, f^+_{ij}, results from contributions of very many distinct but often closely related pathways. This renders the direct decomposition approach computationally infeasible. Therefore, instead of removing the carried flux from all the edges of the strongest pathway, we subtracted a fraction of the net flux of the bottleneck edge (25% for the network relevant to the

sliding window case, which is the only one we used for the pathway decomposition) at each step of the iterative decomposition, which was terminated when 80% of the total flux was collected. This procedure differs from the literature precedent[4] in that only a fraction of the bottleneck net flux is subtracted rather than all of it. It allows us to avoid an oversimplification of the pathway picture that could occur if the same bottleneck is shared by several heterogeneous pathways.

S2 Supporting Figures

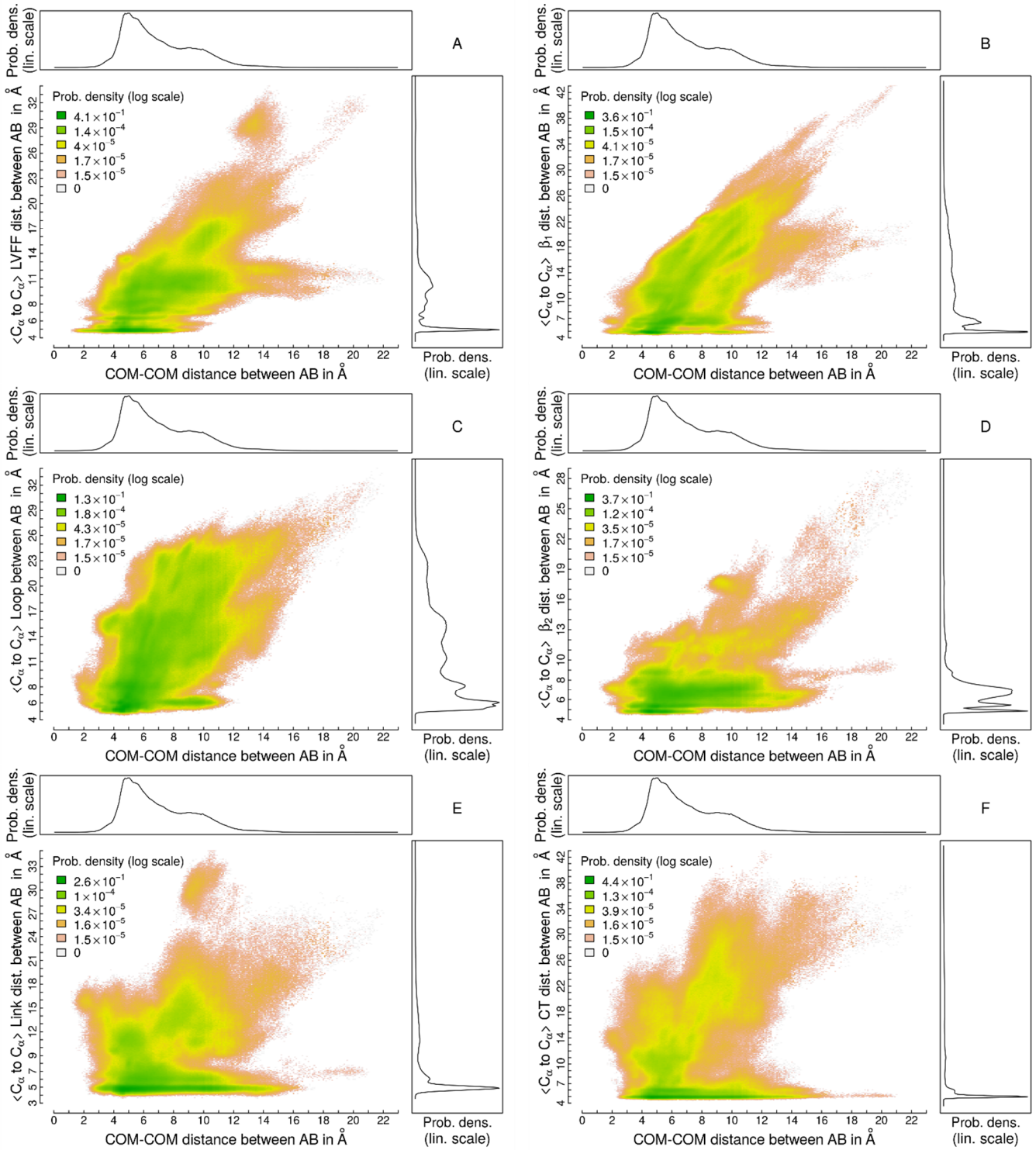

Figure S1: Two-dimensional histograms (log-scale) of mean segment distances and the molecular center of mass-distance for the A-B interface*. For the 6 segments indicated in Fig. 1D of the main text, we show two-dimensional histograms of the mean segment distance and the center of mass distance of chains A and B. The color scale differs per panel (legend is embedded). The different segments are indicated in the y-axis labels and proceed from N- to C-terminus in reading direction.*

*Integrals (1D histograms) are shown for both axes with a linear scale at the top and right of each panel. Here and wherever applicable, the statistical weight of each snapshot is derived from the steady state of the final network model (see "**Construction of Markov model**" in the main text and S1.3.2-S1.3.3 above).*

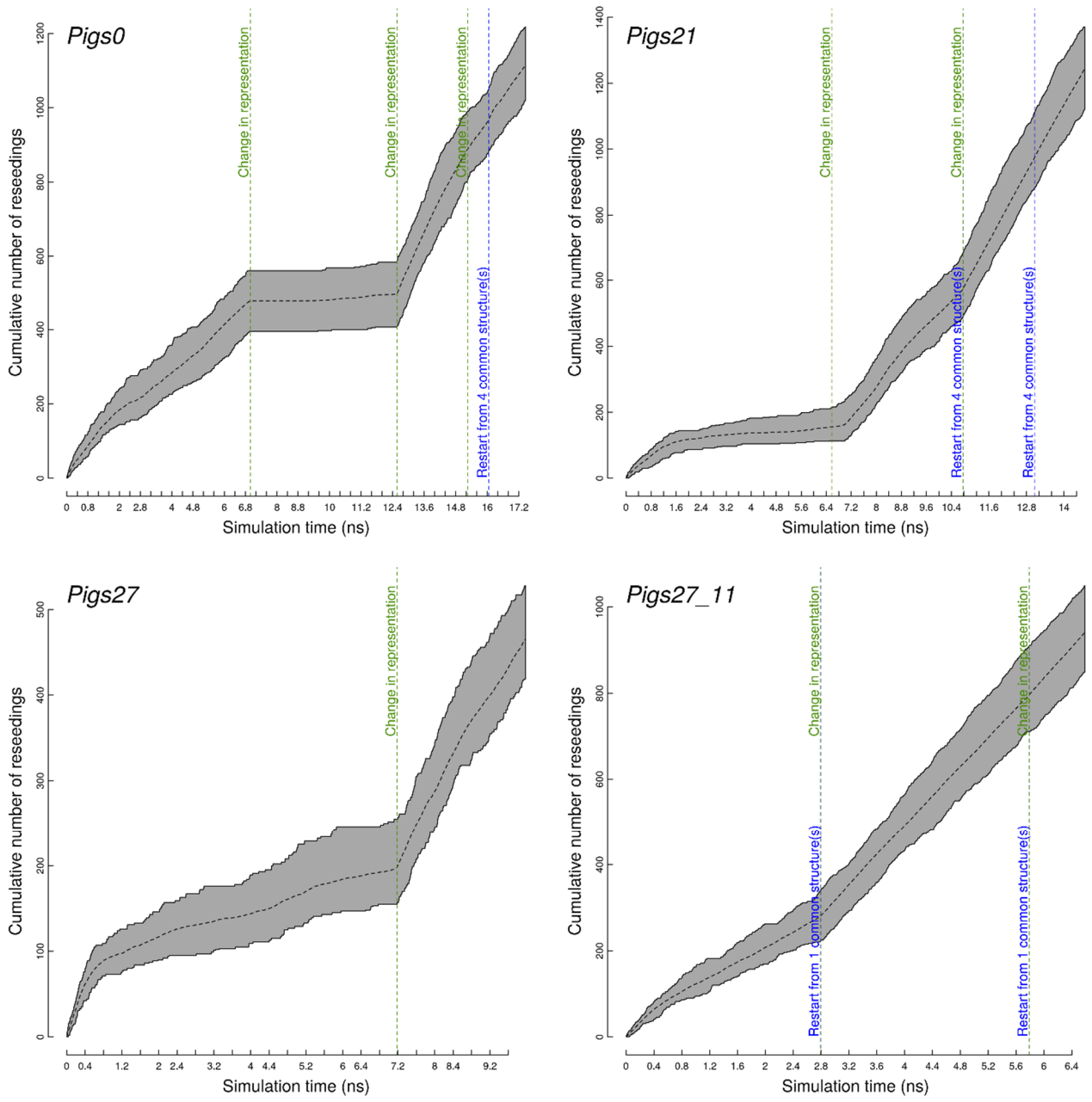

Figure S2: Time evolution of the reseeding rate for selected PIGS simulations*. In each panel, the gray envelope is defined by the maximum and minimum number of reseedings per replica at any given time, and the dashed black line within the envelope denotes the average number of reseedings per replica. Vertical dashed lines mark the time points when a change in representation (green), a manual reseeding from specific structures (blue), or both occurred (see Table S2).*

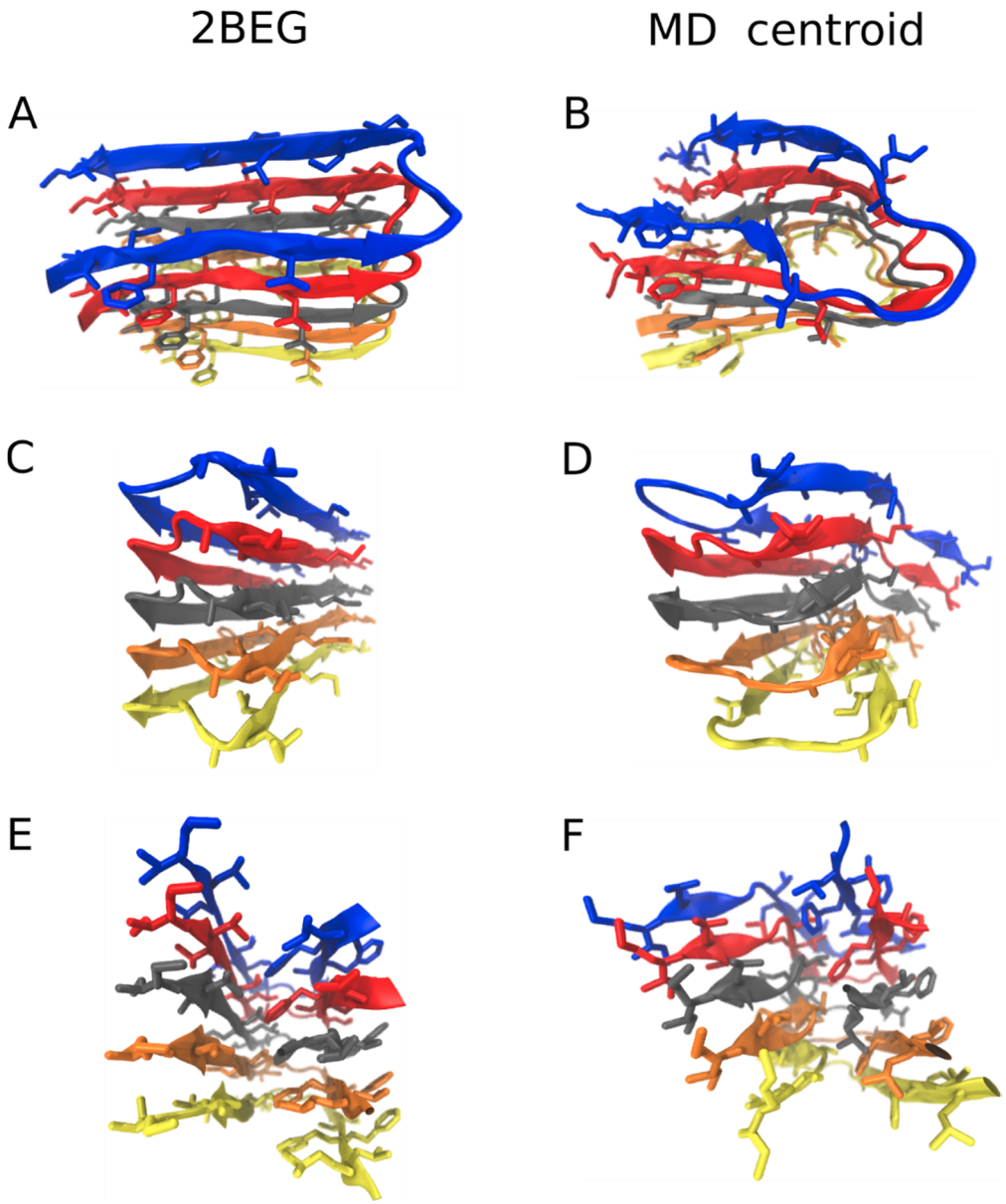

Figure S3: Visual comparison of NMR structure[1] and relaxed MD structure*. Chains are shown as cartoon representations with the N-termini missing in 2BEG being truncated. Coloring is by chain and the same as in Fig. 1B in the main text (chain A is blue). The MD centroid is the central snapshot extracted from a large cluster populated by both GROMACS and GROMACS* simulations (see Table S1). In order to identify it, we followed a protocol similar to that in Figs. S21-S22.* ***A-B****. Views of the odd end (β1 at the bottom). Note the increase in twist of the parallel β-sheet assembly.* ***C-D****. Views of the loop region (β1 is left). The intramolecular strand-to-strand distance widens during MD suggesting the influx of water (compare Fig. 4 in the main text). This is accompanied by a loss of regularity.* ***E-F****. Views of the C-terminal end (β1 is right). The short sheet formed by the CT segments is bent in MD relative to β2 while it is straight in 2BEG. This part is of course heavily occluded from access by the N-termini. Note the increase in stagger (see Fig. S20 below). All molecular graphics here and in subsequent figures were rendered with VMD[48] and Tachyon (http://jedi.ks.uiuc.edu/~johns/tachyon/).*

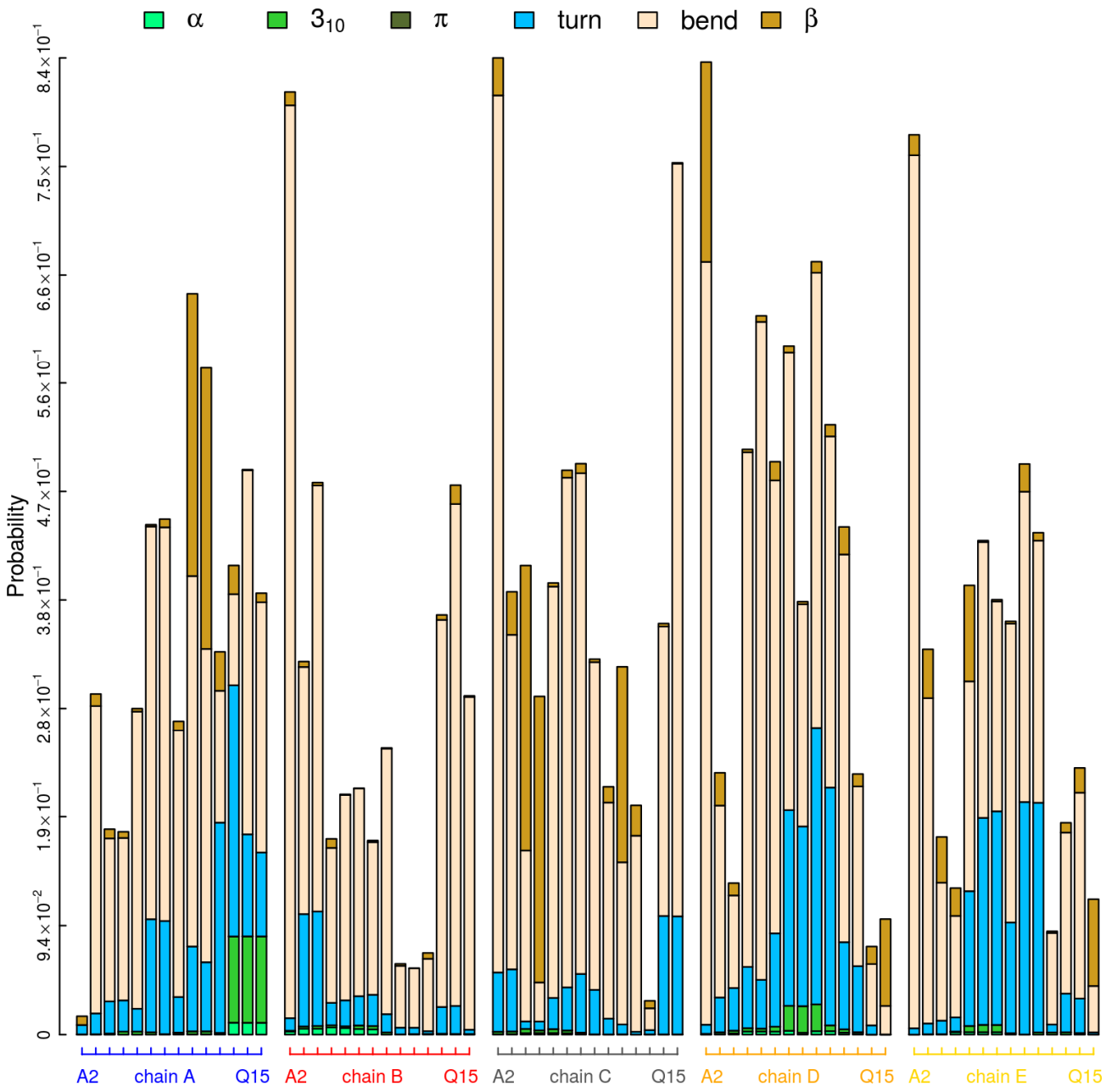

Figure S4: Secondary structure content for N-terminal residues by DSSP.[41] *For each residue from A2 to Q15 of each chain, we plot secondary structure assignments as stacked bars reflecting per-residue probabilities, which were determined by standard DSSP analysis. The most probable assignment is almost always either bend, turn, or no assignment (the latter corresponding to whatever is missing from the total bar height toward a value of 1.0). Canonical secondary structure occurs seldom and is distributed randomly across sequence and chains.*

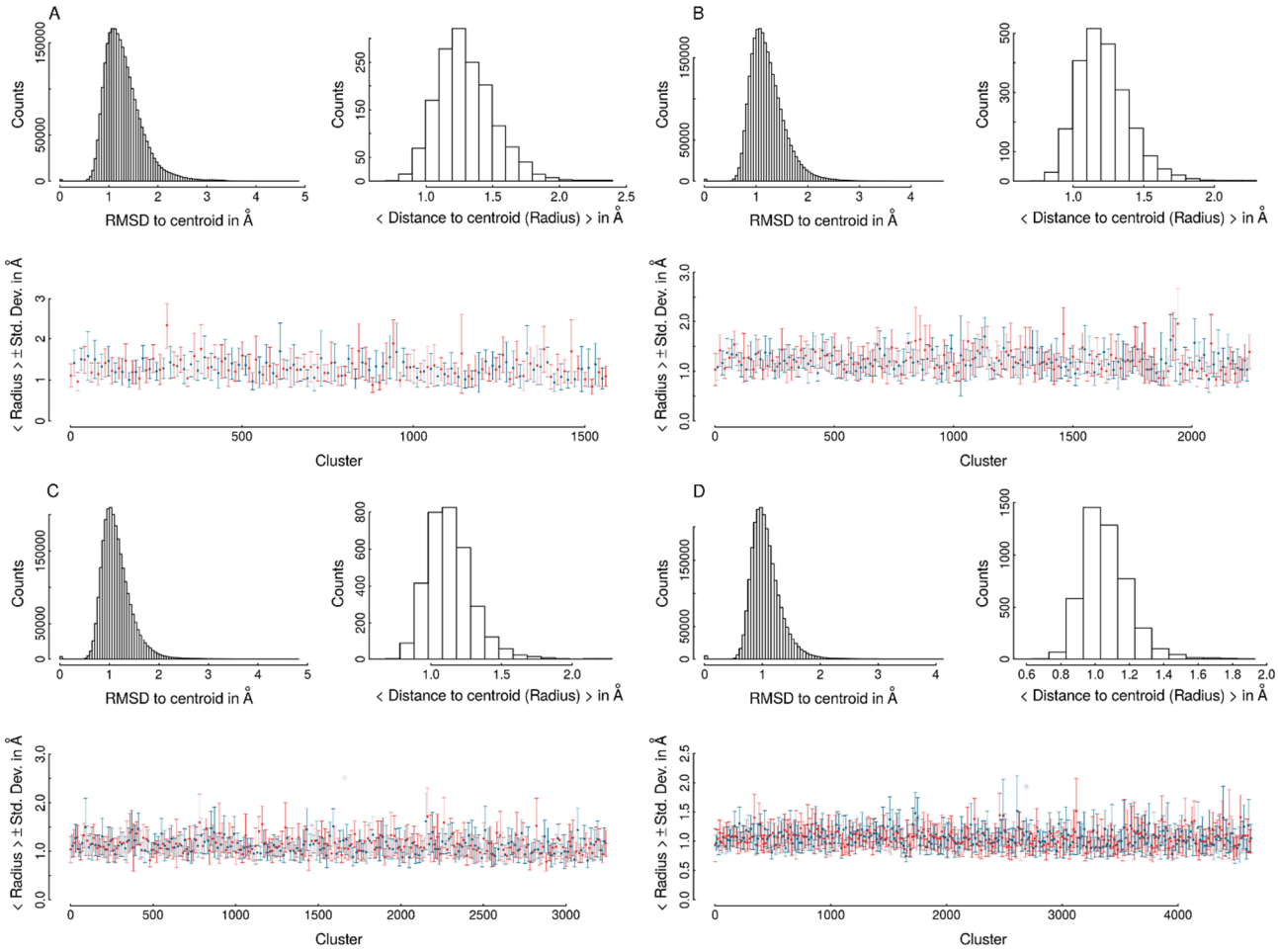

Figure S5: Cross-check of distance metric for clustering against coordinate RMSD (see S1.3.2)*. There are four main panels corresponding to different size threshold for clusters calculated using a D-RMSD metric. Each panel has three subpanels as follows. The upper left plots show the distribution of the distances of snapshots to the relevant centroids of the clusters they belong to. Here, as in all other plots, the grouping is made according to the D-RMSD metric but the distances are positional RMSD values after pairwise alignment, thus informing on the distribution of RMSD values within the clusters. The upper right panels show the distributions of cluster radii in the RMSD metric. Cluster radius is defined as the mean snapshot-to-centroid distance in a cluster (singles excluded). Lastly, the bottom panels provide individual values for the means and standard deviations of snapshot-to-centroid distances within the largest clusters plotted until a cluster size of 100. Colors alternate for improved readability. The means are the aforementioned cluster radii.* ***A-D****. Results for clustering size thresholds of 1.0 Å, 0.9 Å, 0.8 Å, and 0.7 Å, respectively. Source data are the first half only (blue in Table S1).*

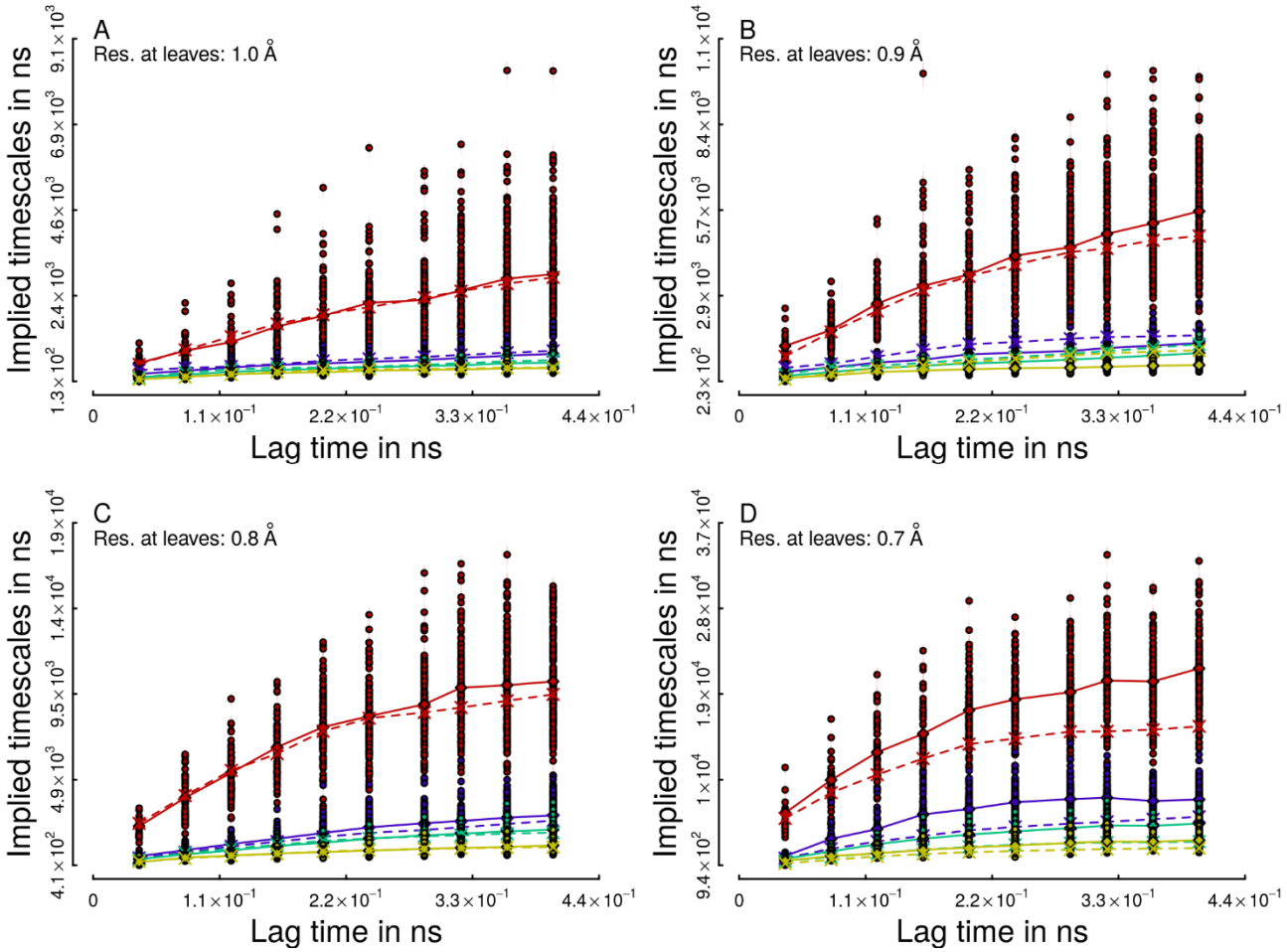

Figure S6: Trends of the four slowest implied timescales as a function of lag time at different clustering resolutions for the whole data set*. Dots mark the values of the timescales derived from the independent but correlated trajectories that can be extracted at any given lag time, which are as many as the multiplicity of the lag time with respect to the saving frequency. Colors indicate the different modes (red, blue, cyan and yellow). Solid lines trace the average values of these distributions. Dashed lines connect the results from the sliding window approach. Lines are purely meant as a guide to the eye. Panels* ***A-D*** *show the results for clustering resolutions of 1.0Å, 0.9Å, 0.8Å, and 0.7Å, respectively.*

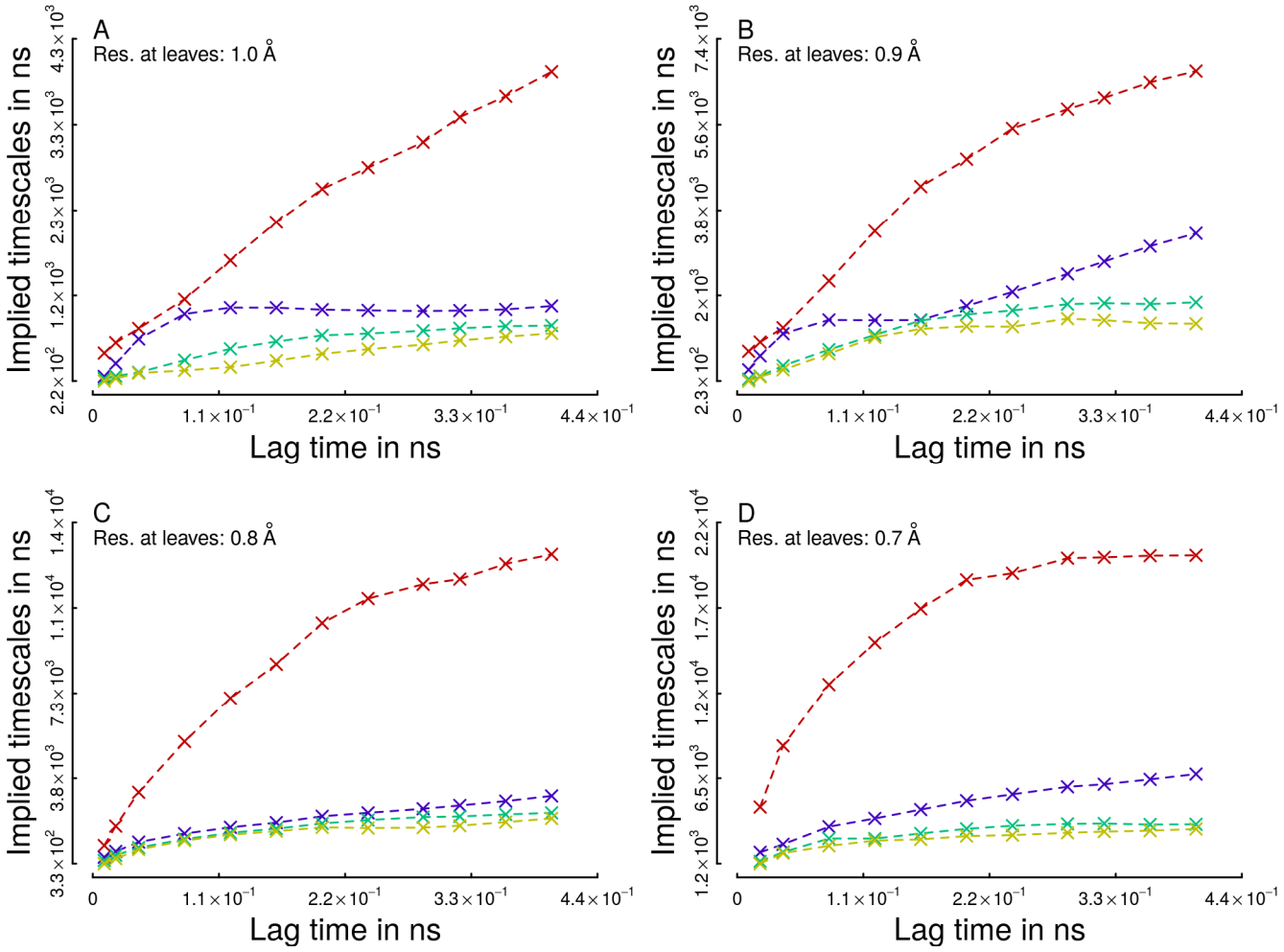

Figure S7: Trends of the four slowest implied timescales as a function of lag time at different clustering resolutions with an alternative symmetrization of the count matrix. *This figure is identical to Fig. S6 except that count matrices were symmetrized by the procedure introduced in Bowman et al.,[46] and that only the results for the sliding window case are shown. This symmetrization attempts to maximize the likelihood of observing the count matrix given an estimate of the transition matrix under the constraint of obeying detailed balance. This is itself an iterative procedure with two disadvantages: 1) convergence can be slow which was inconvenient for the large number of networks we constructed; 2) the definition of likelihood implies Markovianity such that the optimality is not easy to assert in general. Importantly, the results for the sliding window case are fundamentally similar to those in Fig. S6. The only noteworthy differences are more curvature changes for coarse resolutions and a small general increase in timescales (well within a factor of 2). In consequence, we utilized the simple symmetrization procedure throughout.*

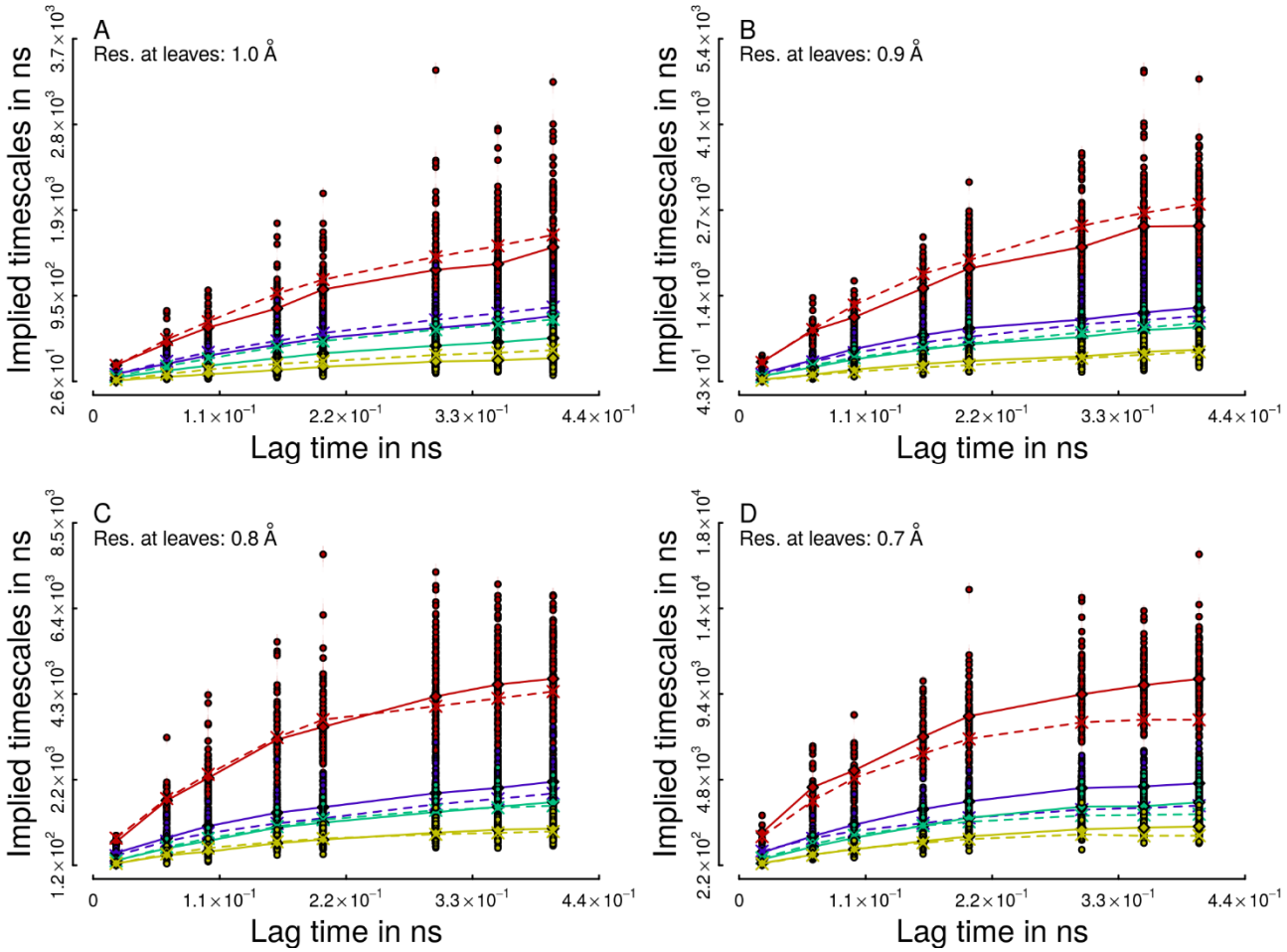

Figure S8: Trends of the four slowest implied timescales as a function of lag time at different clustering resolutions for the first half of the data set*. This figure is identical to Fig. S6 except that only the first half of the data are considered (see S1.2).*

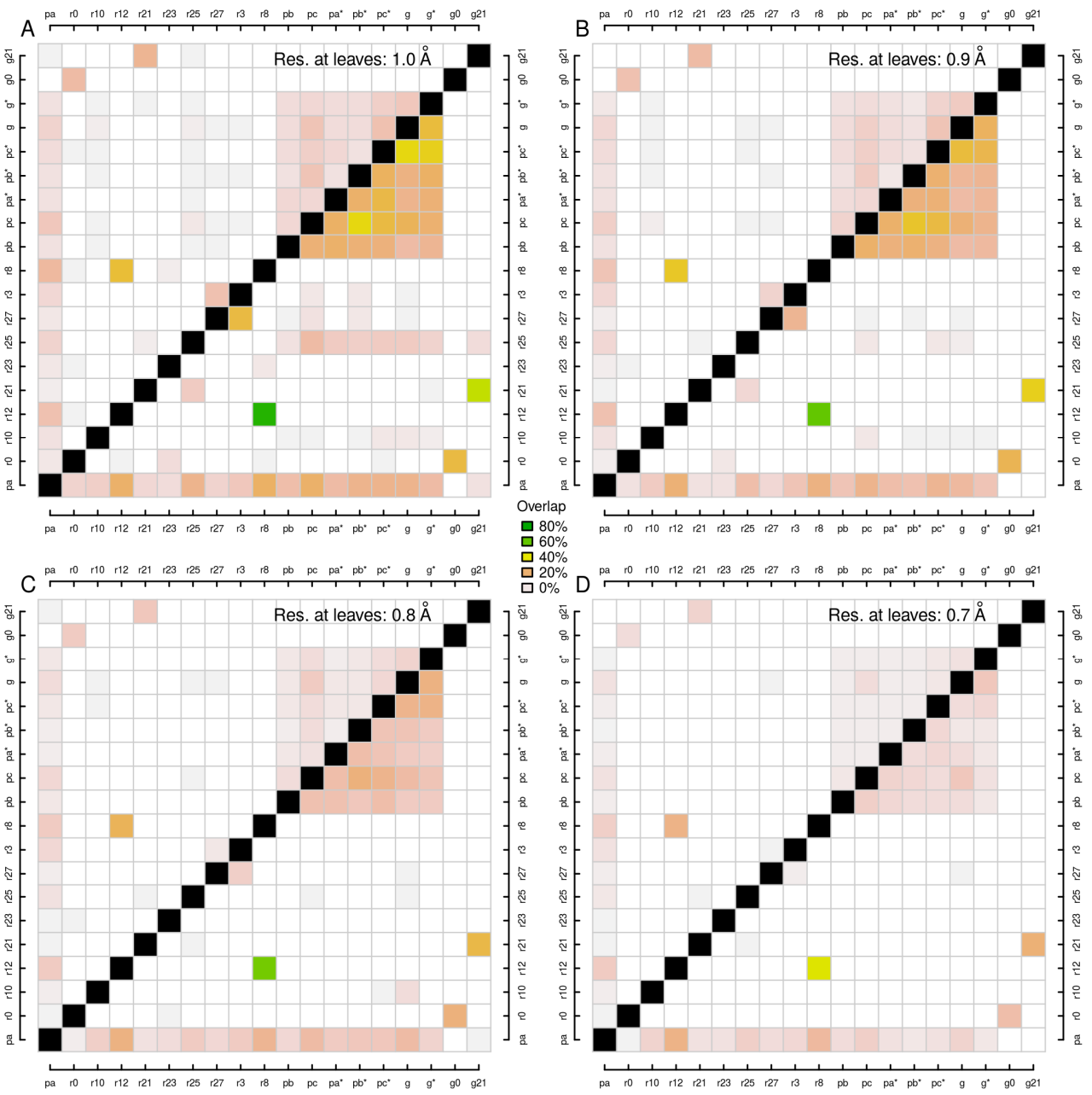

Figure S9: Cluster overlap between sets of runs. *For the same four resolutions displayed in Figs. S6-S8, we show the mutual similarity of runs by two cluster overlap measures. The upper left half matrix counts the number of clusters containing snapshots from both runs in question and divides this number by the total number of clusters. The lower right half matrix instead sums up the total Markov model weight of all clusters containing snapshots from both runs in question. The labels on the axes are short-hand notations of the terminology in Table S1: "g" stands for "Gromacs", "pa" for "PigsA", "pb*" for PigsB*", "r21" for "Pigs21", and so on. For the analysis, each category includes all its children as listed in Table S1. The color legend in the middle applies to all panels.* ***A-D****. Data for resolutions of 1.0Å, 0.9Å, 0.8Å, and 0.7Å, respectively. In general, overlap decreases as the number of clusters increases (finer resolution), which is expected. The mutual overlap of all the PIGS runs starting directly from the NMR structure is apparent. The conventional (non-PIGS) simulations agree strongly with their respective starting structures despite being longer in simulation time (e.g., Gromacs21 and Pigs21, see Table S2), indicating that PIGS is absolutely essential for diversification on the ns timescale.*

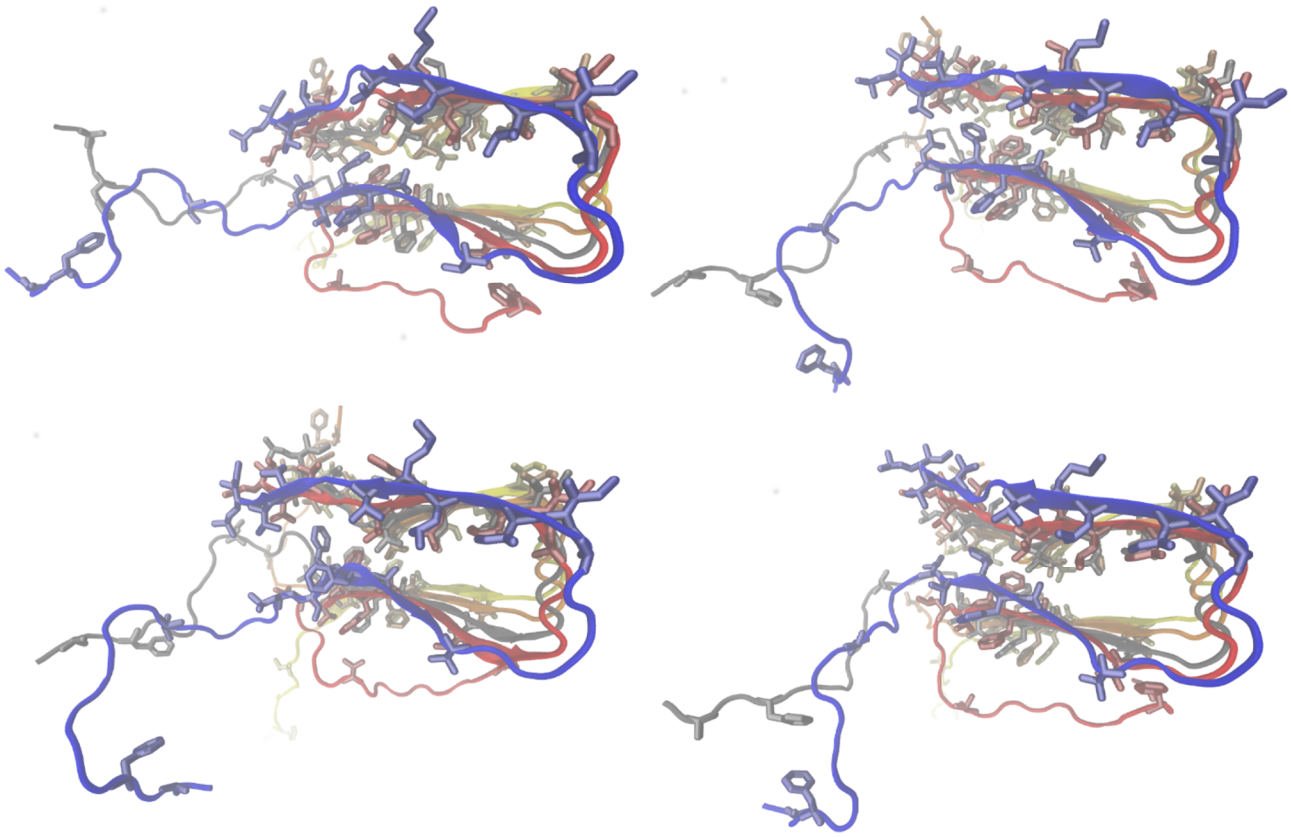

Figure S10: Centroids of the four clusters that constitute the ordered (locked) set. *Please refer to section S1.3.5 for information on how these were identified. Backbone conformations are shown as Cartoons with the same chain coloring used throughout. Sidechains of hydrophobic residues are shown as sticks.*

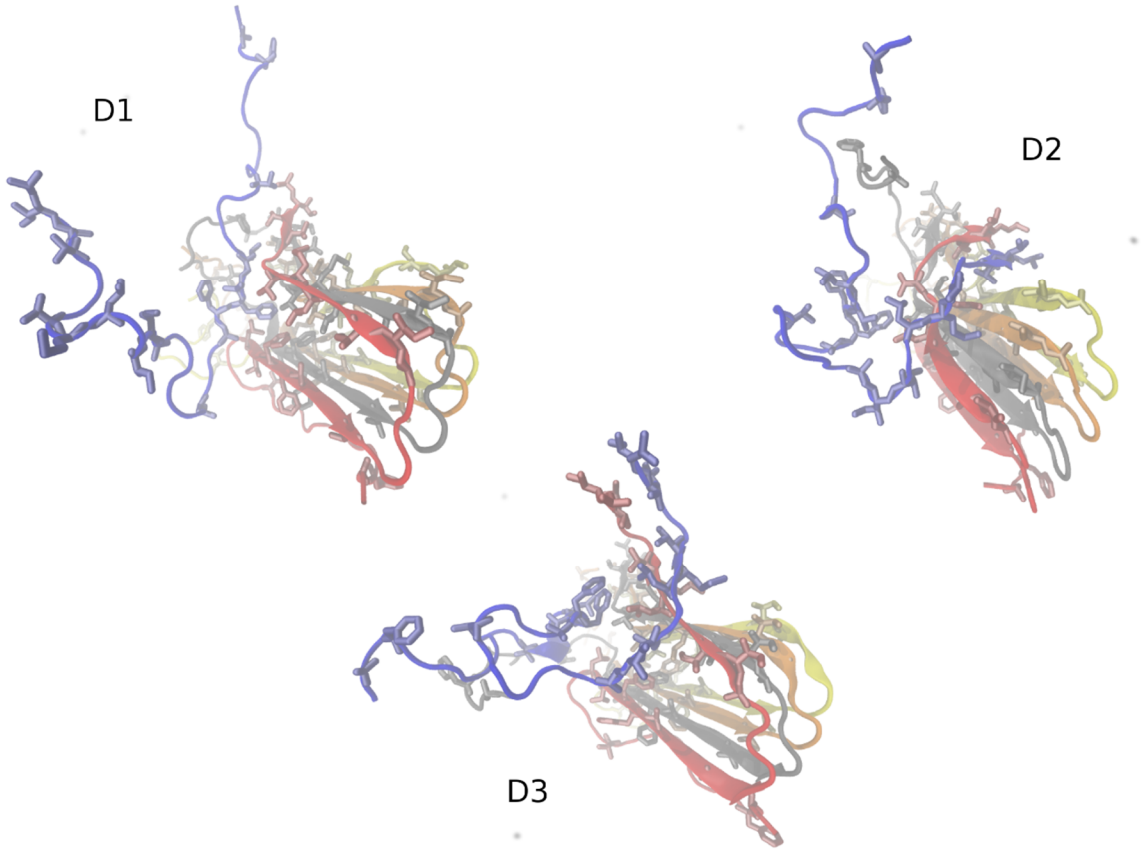

Figure S11: Centroids of the three clusters that made up the disordered set, D1-D3. *Please refer to section S1.3.5 for information on how these were identified. Backbone conformations are shown as Cartoons with the same chain coloring used throughout. Sidechains of hydrophobic residues are shown as sticks.*

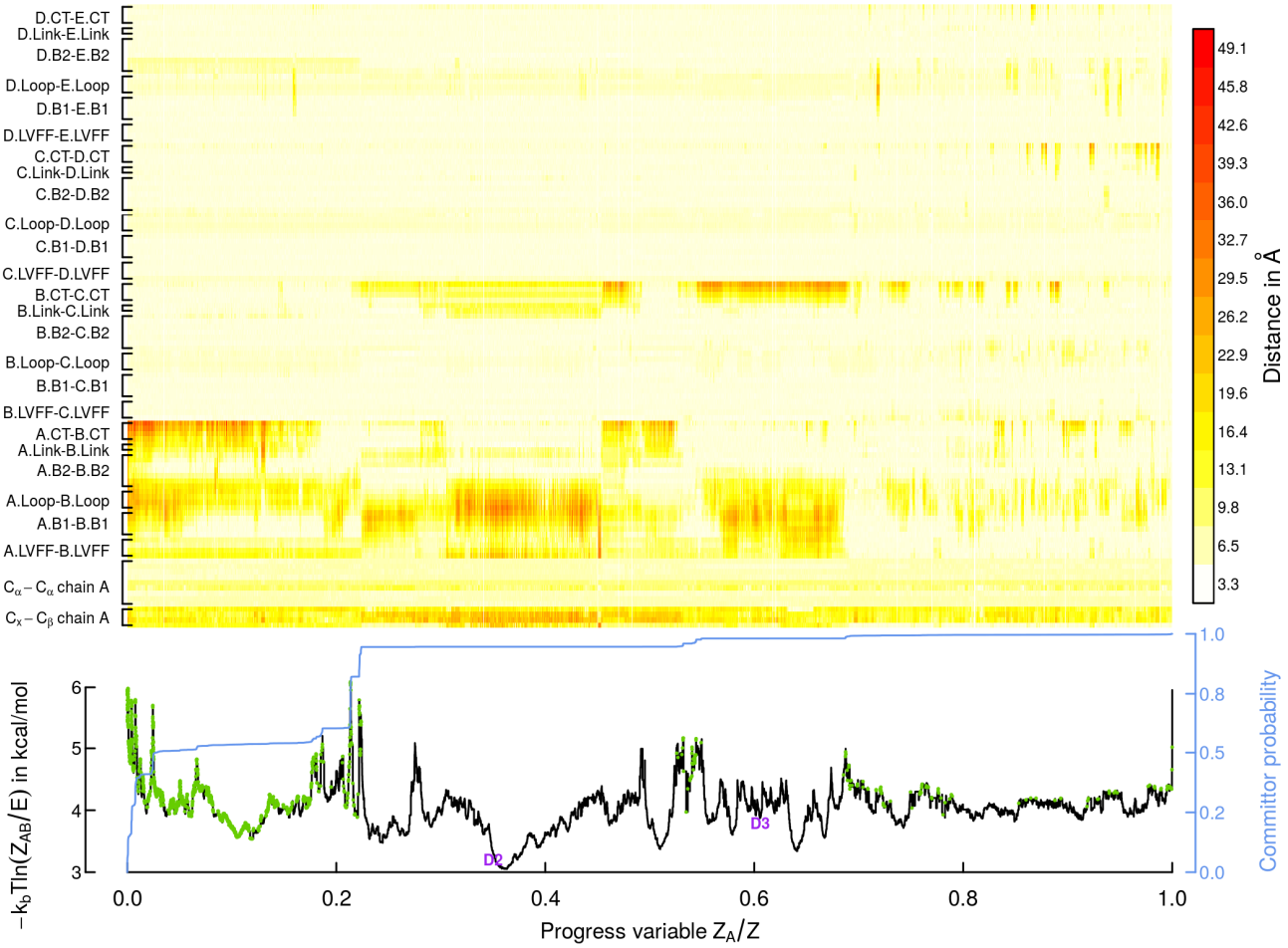

Figure S12: Cut-based free energy profile based on p^{+}_{fold} ***values for D1 with structural annotation****. Briefly, the calculated forward committor probabilities (see S1.3.4) were sorted. For each cluster, its* $p^{+}_{fold_i}$ *value defines a cutting surface in the network with two states, one with higher and one with lower committor probabilities. Direct transitions crossing the respective surface for each cluster are counted. These are the* Z_{AB} *values plotted logarithmically on the y-axis as pseudo free energies (black curve). The x-axis position of a cluster corresponds to the cumulative steady state probability of itself and all clusters with lower committor probabilities.*[43] *The profile is annotated fourfold. First, on-pathway intermediates for the locking of D1 are highlighted in green on the black curve. Second, purple labels indicate the positions of the other two docked states, D2 and D3. These are obviously not intermediates as they are off-pathway. Third, the blue line plots the actual committor probabilities (right y-axis). Fourth, the color map on top reports in-registry distances for all intermolecular interfaces (text legend on the left, color legend on the right). In addition, selected intramolecular distances for chain A are shown as well (bottom of the color map, where the set of distances is the same as in Table S3 but for chain A alone). This figure is complementary to Fig. 7 in the main text because: i) the whole data set is represented; ii) the scaling on the x-axis is by population and not by committor probability increment. This implies that the x-axis here is* ***not*** *a pathway variable of any kind.*

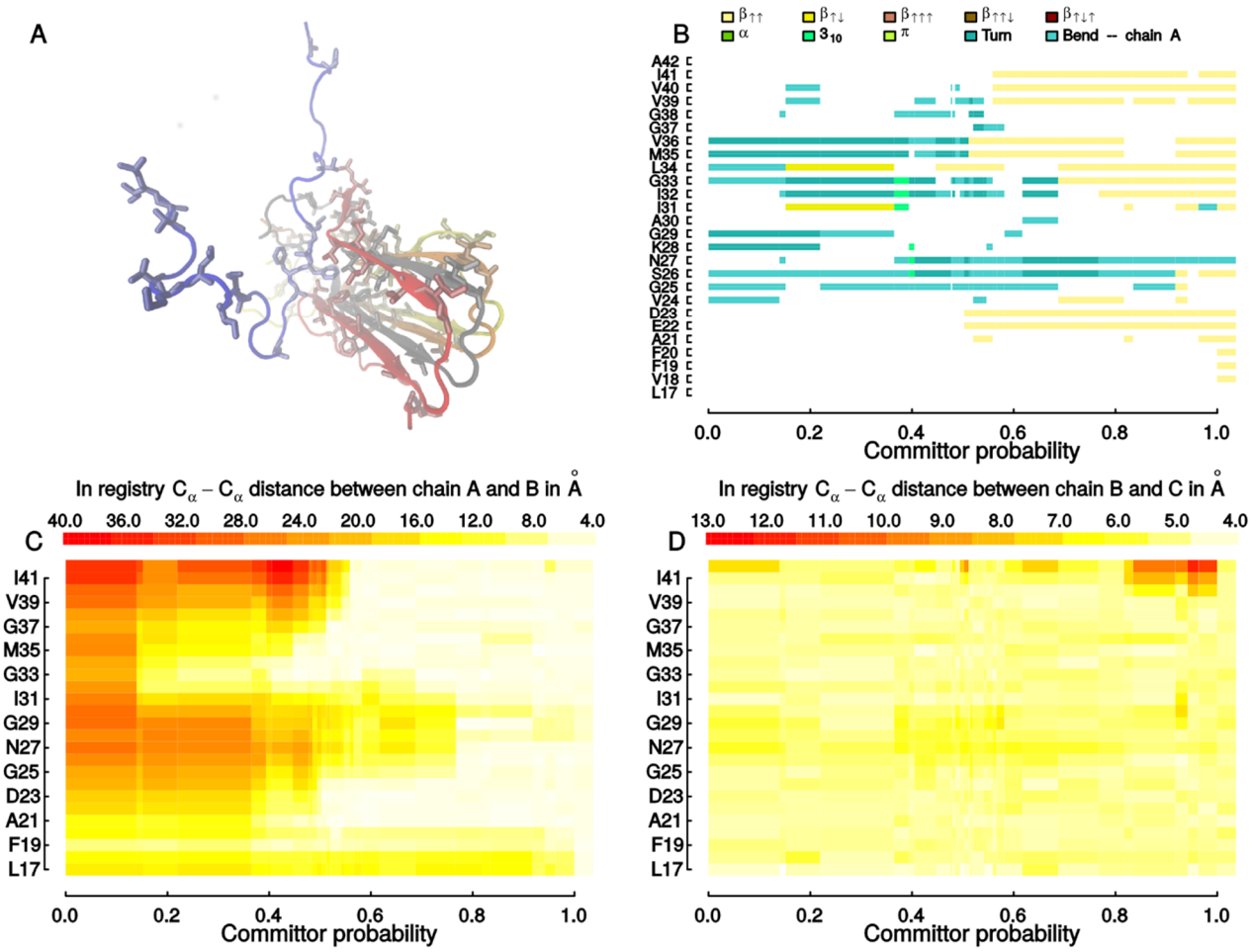

Figure S13: Structural progression of the single strongest locking pathway for D1*. This figure is very similar to Fig. 7 in the main text (please refer to the caption of Fig. 7 for details). The only difference is that here only those clusters contributing to the strongest of all locking pathways determined by pathway decomposition (see S1.3.6) are shown. As a result, the spacing in committor probability is often large. As these values are chosen as the left boundary for plotting, the interval from 0.96 to 1.0 corresponds to the last on-pathway intermediate preceding the locked state. For clarity, the locked state itself is added as a bar for committor probabilities beyond 1.0 (there is no significance to its width).*

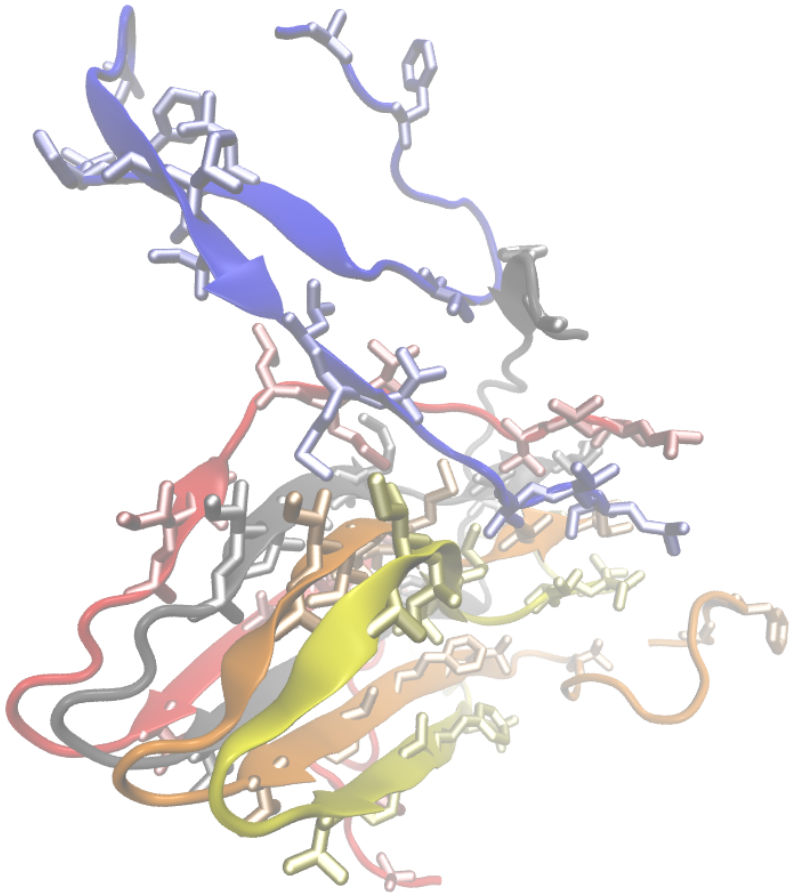

Figure S14: Visual illustration of a laterally bound and β-hairpin-containing docked conformer related to D2*. Backbone conformations are shown as cartoons with the same chain coloring used throughout. Sidechains of hydrophobic residues are shown as sticks. This conformation was extracted from the simulation labeled Pigs0_29_27L in Table S2. It is kinetically close to D2 (* p^{+}_{fold} *= 0.12 and a mean-first passage time to D2 that is only 13% of that of the locked state), i.e., the major barrier is between D2 and the locked state.*

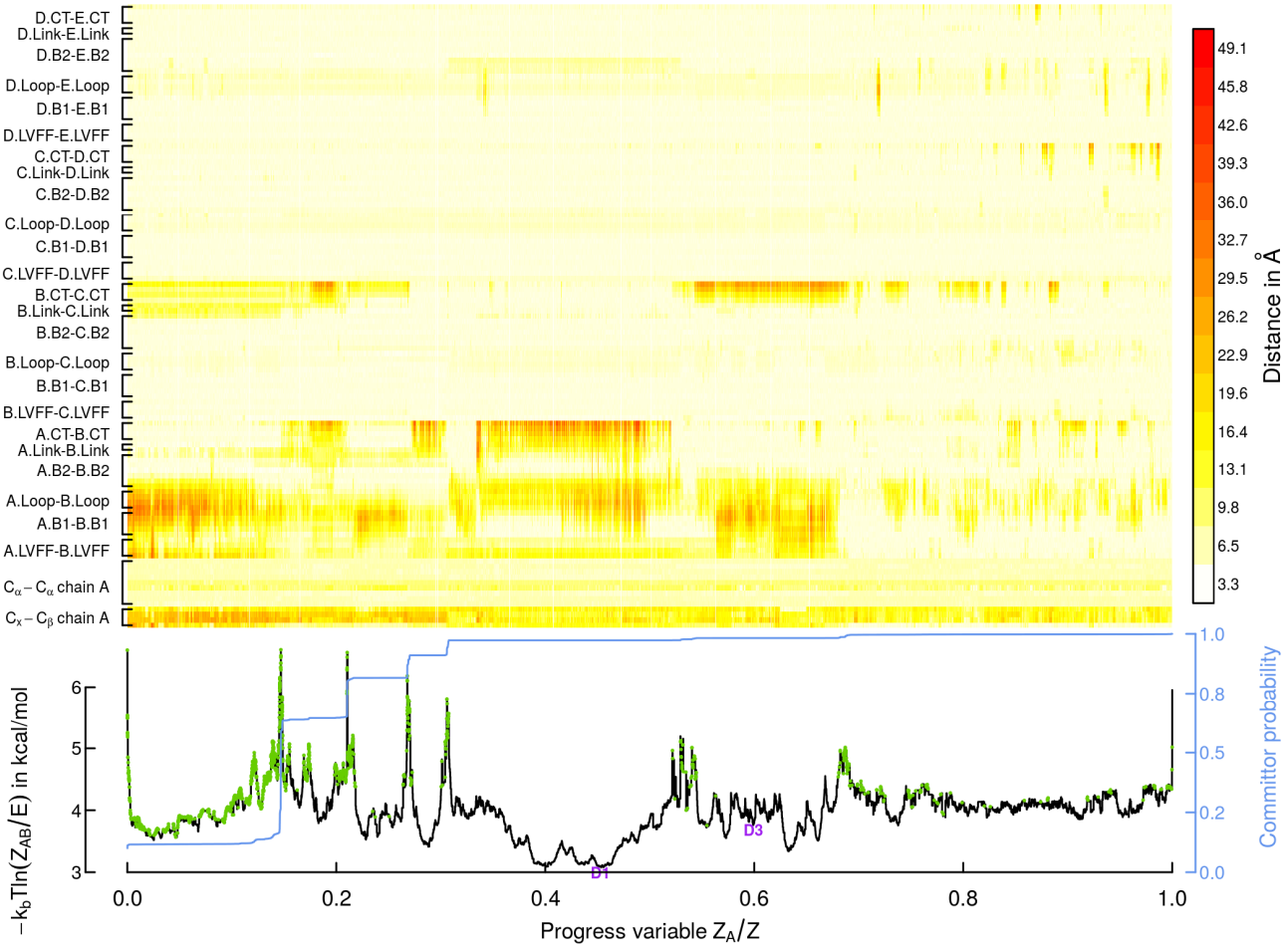

Figure S15: Cut-based free energy profile based on p^{+}_{fold} ***values for D2 with structural annotation***. *This figure is identical to Fig. S12 except that data for D2 are shown. Therefore, it is complementary to Fig. 8 in the main text.*

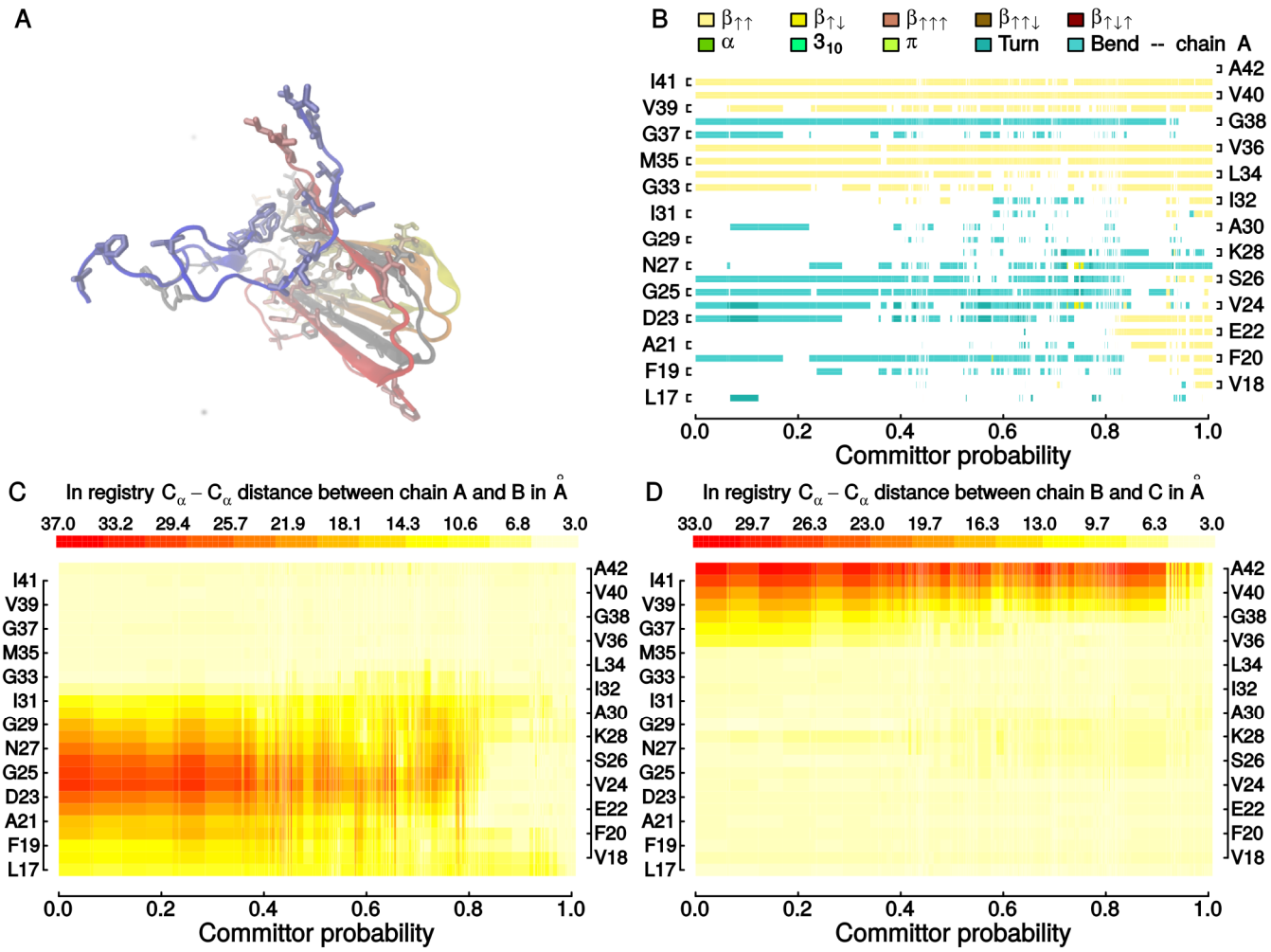

Figure S16: Structural progression of locking pathway for D3*. This figure is identical to Fig. 7 in the main text except that data for D3 are shown. Please refer to the caption of Fig. 7 for plotting details.*

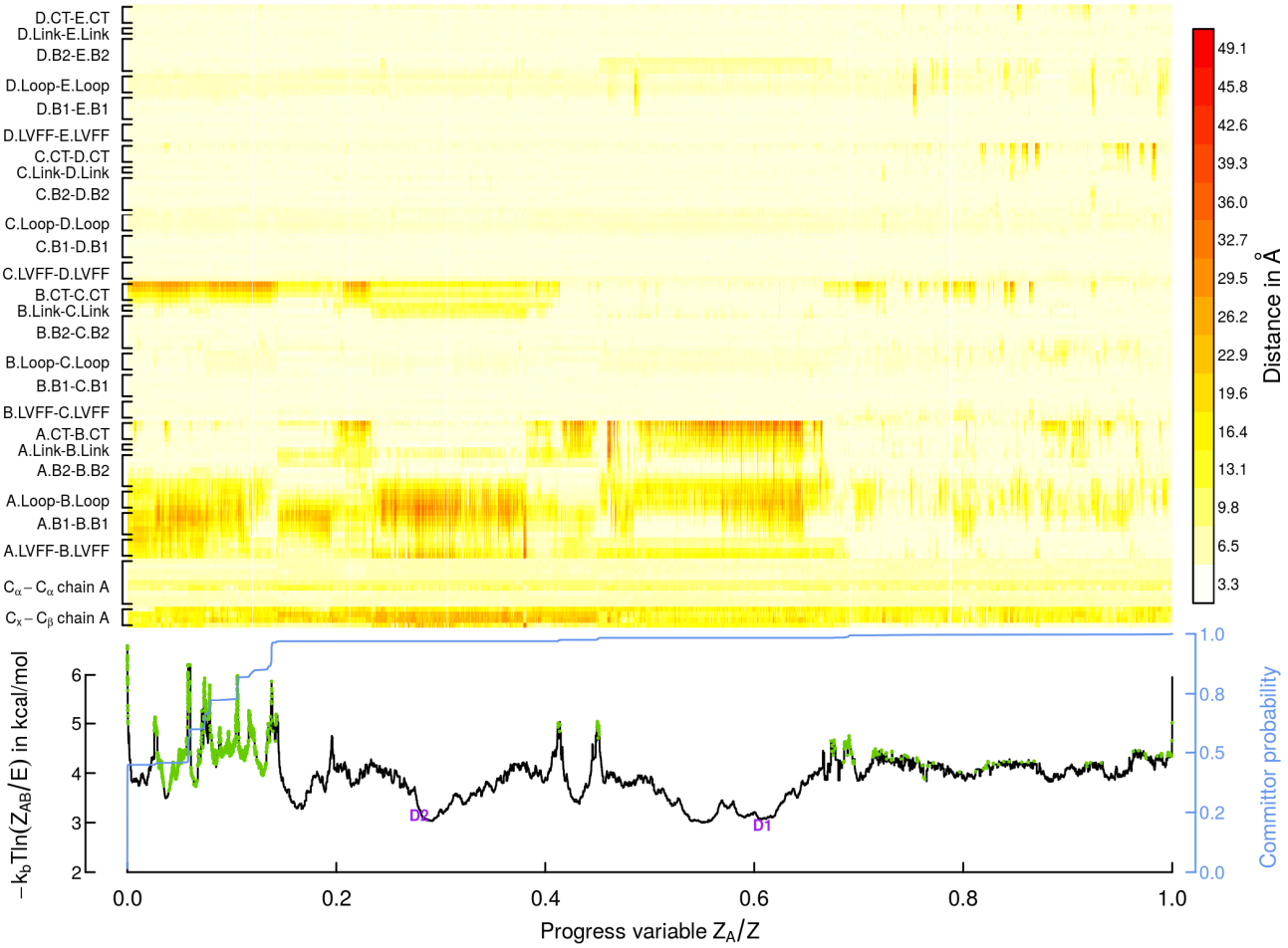

Figure S17: Cut-based free energy profile based on p^{+}_{fold} ***values for D3 with structural annotation****. This figure is identical to Fig. S12 except that data for D3 are shown. It is thus complementary to Fig. S16.*

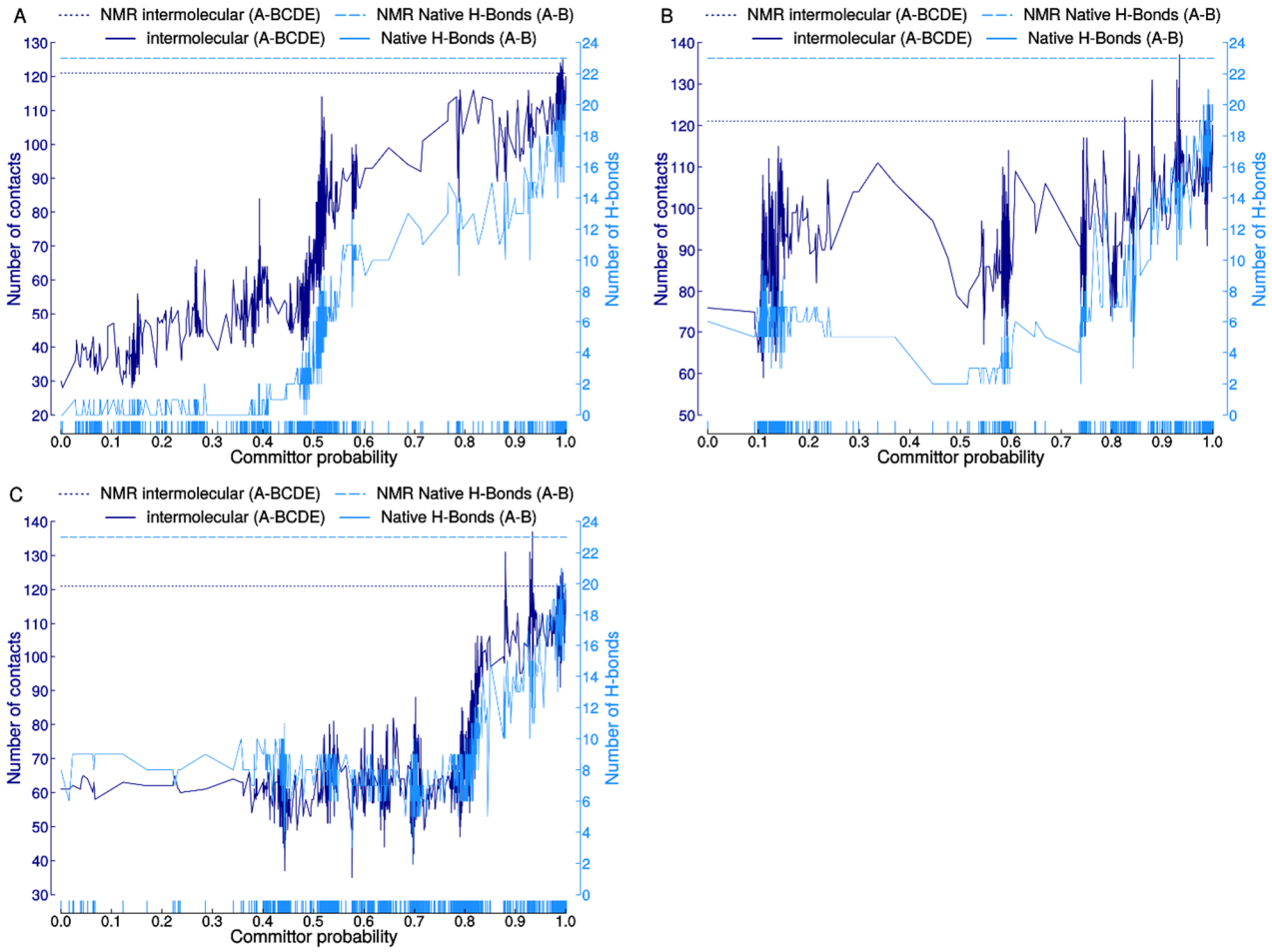

Figure S18: Committor probabilities annotated with the formation of β-hydrogen bonds and intermolecular contacts for D1-D3. *We plot the number of hydrogen bonds and the number of intermolecular contacts between chain A and the rest of the assembly as a function of the (+) committor probabilities, p_{fold} (see S1.3.4-S1.3.6), for different locking pathways. Vertical segments at the bottom highlight the p_{fold} values of the clusters that are actually on this locking pathway (similar to Figs. 7, 8, and S16). Areas with low bar density are likely barrier regions. All plotted quantities refer to the representative snapshot (centroid) of the cluster in question, which is taken as a consensus representation. Dashed and dotted horizontal lines mark the number of (native) hydrogen bonds (light blue) and intermolecular contacts (dark blue) for the initial NMR structure, respectively. Hydrogen bonds were determined using standard DSSP[41] analysis with the default energy threshold of -0.5 kcal/mol and keeping only the strongest one per donor or acceptor. They can be native or nonnative but exclude the N-termini (residues 1 to 16). An intermolecular contact was counted every time the shortest distance from any atom of a residue of chain A to an atom of any other chain was smaller than 5.0 Å, again with the exclusion of the N-terminal residues.* ***A.*** *Trends of native hydrogen bonds and intermolecular contacts for the pathways starting in the disordered state D1.* ***B.*** *Same as A for D2.* ***C.*** *Same as A for D3. Three important observations emerge. First, the number of hydrogen bonds can be very low even for large values of p_{fold} (e.g., only 4 are formed for p_{fold} values between 0.7 and 0.8 in B). Second, in areas of p_{fold} where significant transitions occur (particularly visible in A and C), the complete "zipping up" of the β-sheets appears to occur at larger values of p_{fold} relative to contact formation (they should be superimposed if β-hydrogen bonds drive contact formation). Third, long stretches of p_{fold} values occur that seem to show significant changes only in contacts but not in β-hydrogen bonds (e.g., 0.0-0.5 in A, 0.0-0.7 in B). It is a caveat that important changes that approximately preserve the total number reported are masked in this analysis (see C). This issue is more pronounced for contacts than for hydrogen bonds (since the absolute numbers are larger).*

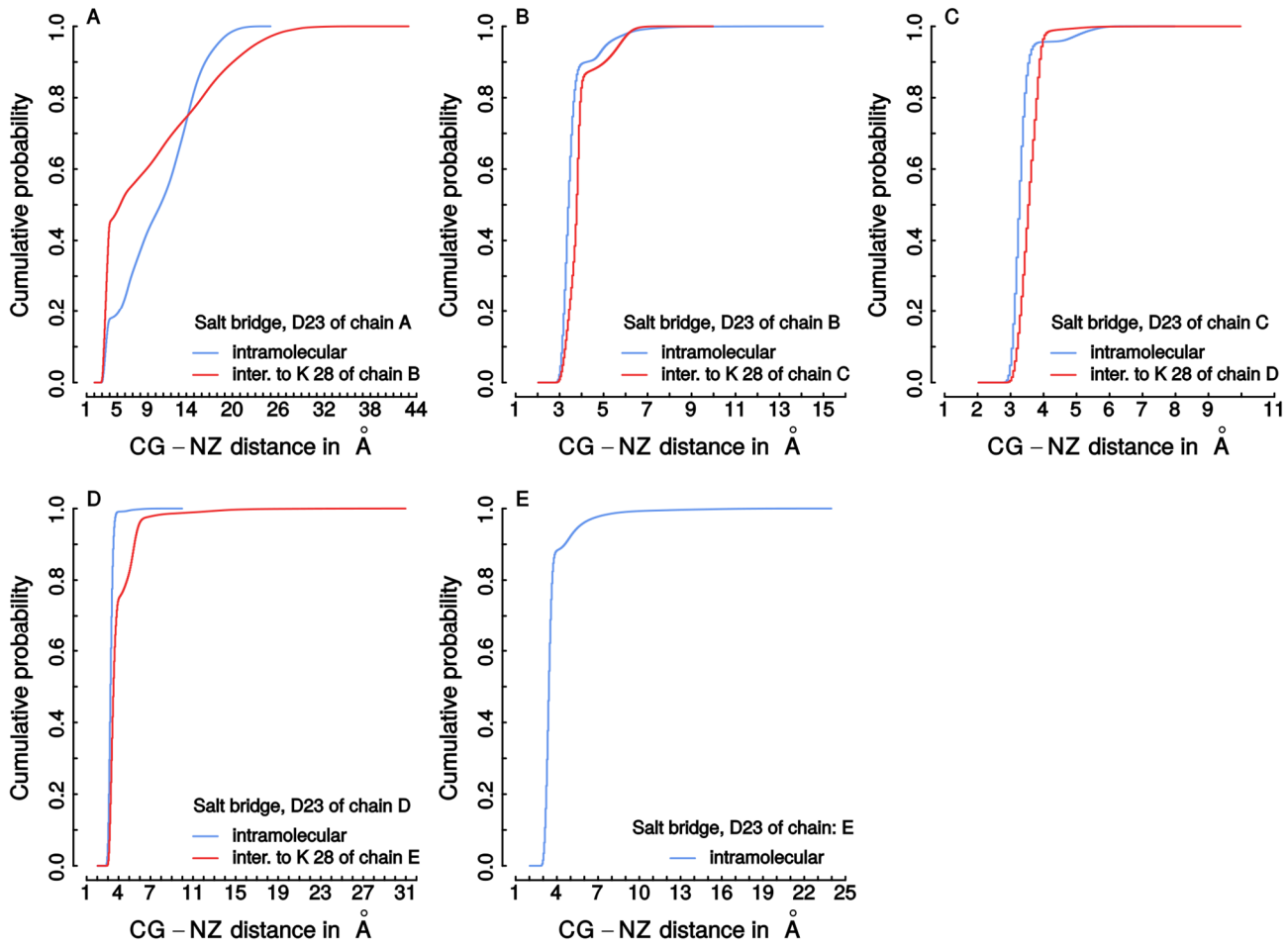

Figure S19: Cumulative distribution functions for salt bridge distances at all interfaces*. We recorded the distances between K28 N_ζ and D23 C_γ atoms both intramolecularly and intermolecularly, and the corresponding MSM-weighted cumulative distribution functions are plotted. Note the variable scale on the x-axes.*

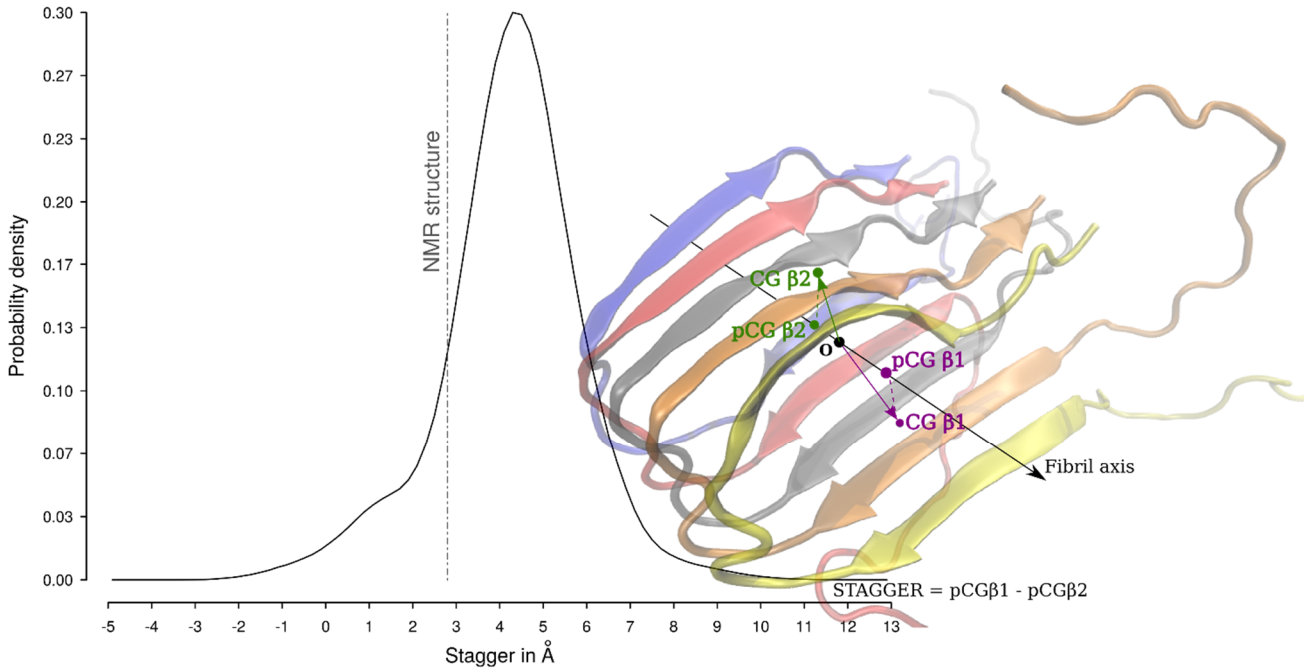

Figure S20: Distribution of the stagger of the β1- and β2-sheets along the protofibril axis. *By defining the protofibril axis at each time point as the vector joining the center of geometry of chain C with the center of geometry of chain D, we projected the centers of geometry of the β1 and β2 sheets onto the protofibril axis and calculated the absolute difference in Å between the two. In the Cartoon, the centers of geometry are labeled "CG β1" and "CG β2," and the respective projections onto the protofibril axis are "pCG β1" and "pCG β2." A positive value in the histogram is indicative of a stagger of the same type as the one in the reference NMR model (PDB 2BEG).[1] The center of geometry of β1 was simply determined as the average value of the C_α coordinates of residues A21 and E22 of chains B to E, while the center of geometry of β2 was given by the average C_α coordinates of L34 and M35 of the same chains. The average over all four points for a single chain gave rise to the centers of geometry used to define the protofibril axis (based on chains C and D). The increase in average stagger is due to the deformation of the less twisted NMR structure upon relaxation in the simulations (compare Figs. S3E and S3F).*

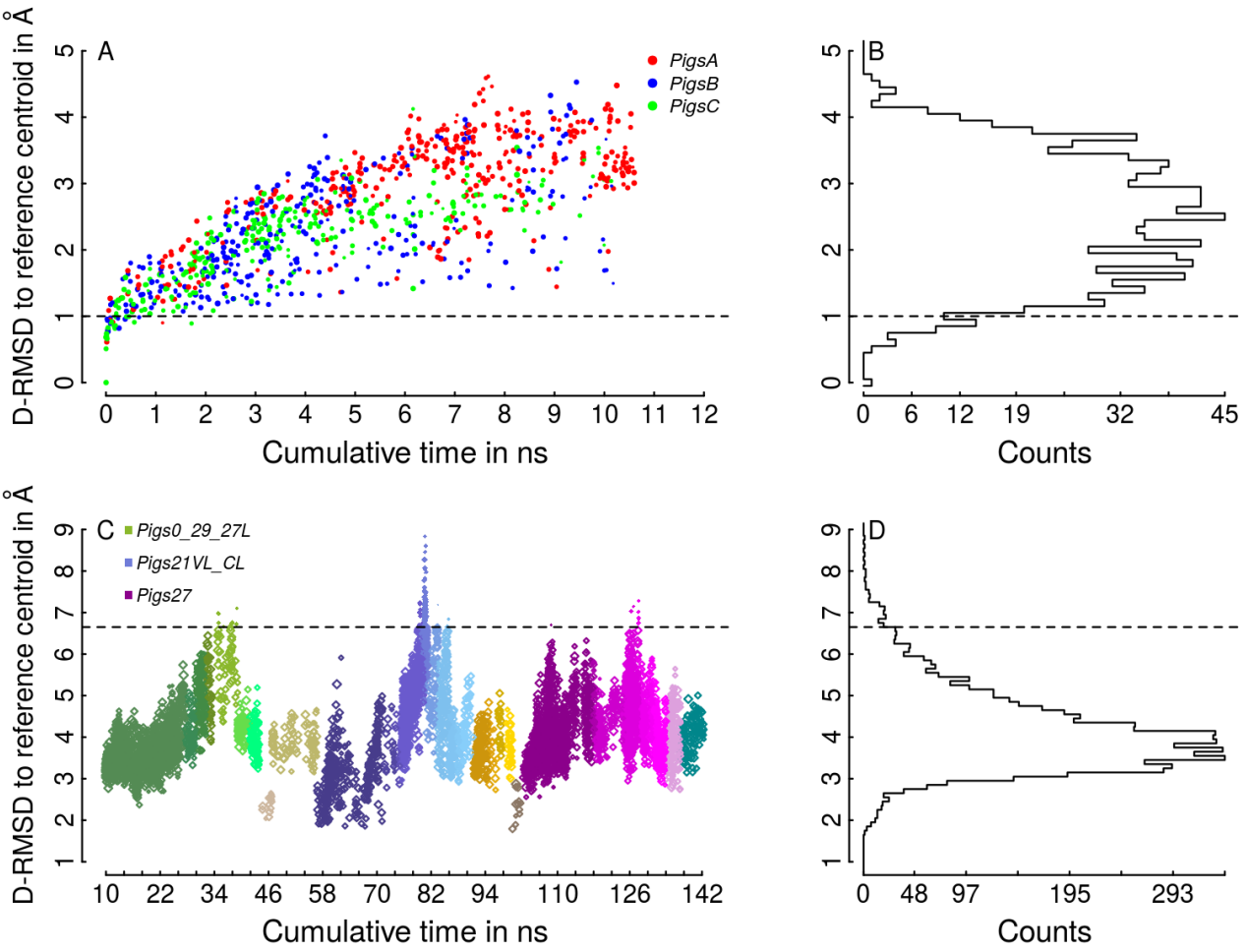

Figure S21: Selection of candidates for the ordered and disordered sets by distance thresholds. ***A***. *Clusters sampled in runs PigsA, PigsB, and PigsC are identified by their centroids and ordered along the x-axis according to the first time of discovery. The color code identifies the run encountering a given cluster first. The D-RMSD distance between the centroid of the cluster that contains the starting structure and the various clusters is reported on the y-axis. The upper threshold distance used to obtain a pool of plausible candidates for the ordered set is shown as a dashed line.* ***B***. *Histogram of the distances in A used to justify the choice of threshold.* ***C***. *For derived PIGS runs (individual runs as in Table S1 are differentiated by color), we plot the same centroid-to-centroid distance as in A as a function of cumulative time. Cumulative time means that within each run, progression is with sampling time per copy, but that runs are simply concatenated. The lower threshold distance used to obtain a pool of plausible candidates for the disordered set is shown as a dashed line.* ***D***. *The same as B but for the data in C.*

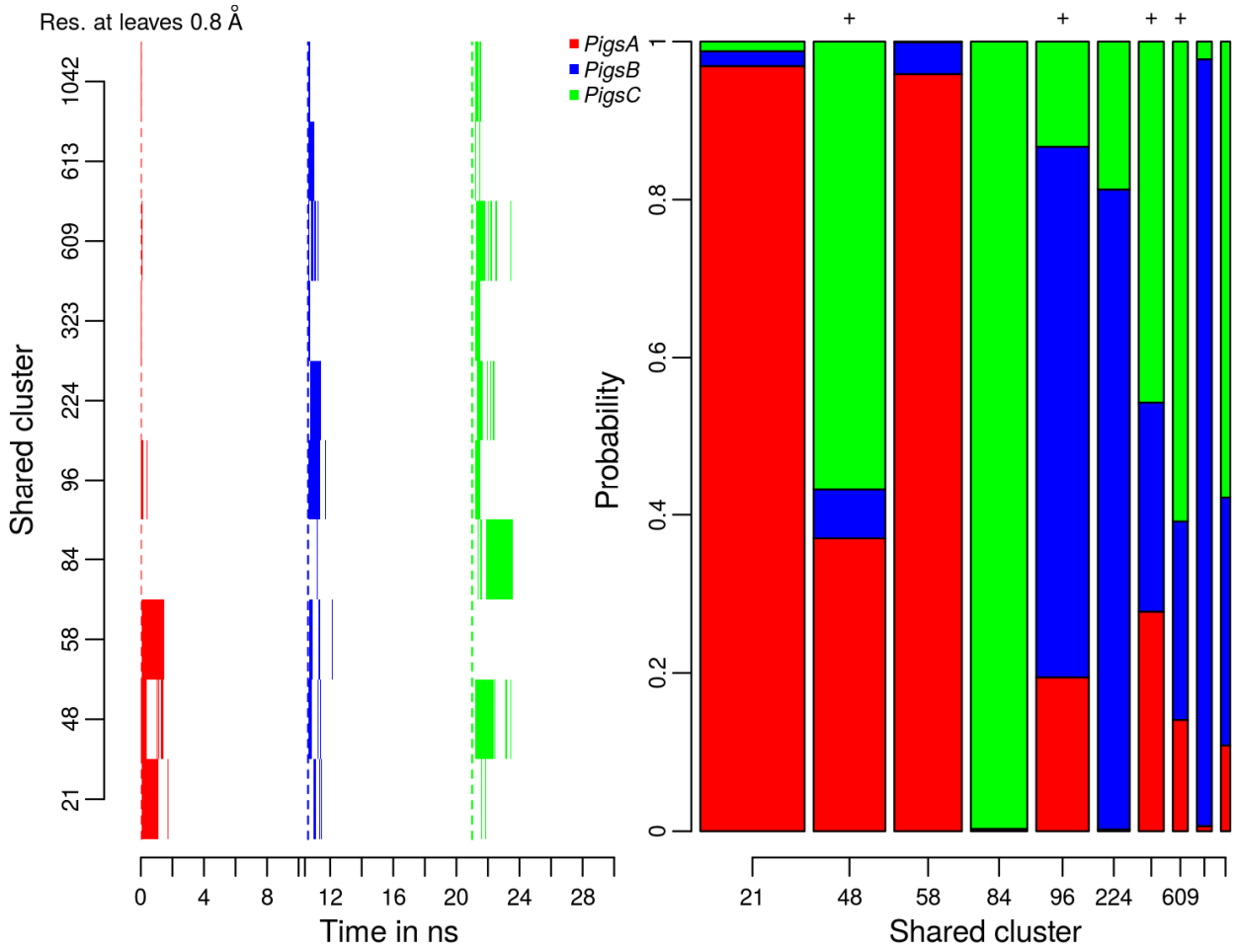

Figure S22: Final selection of the states that compose the ordered set. ***A***. *Time trace of occurrence of the snapshots that constitute the candidate clusters containing snapshots from all of the 3 initial PIGS runs considered, i.e., PigsA, PigsB, and PigsC. Time refers to the sampling time per copy* ***B***. *Statistical weight of the clusters in A (widths of the bars) resolved by homogeneity across runs (the height of the colored rectangles is proportional to the number of snapshots contributed by a specific PIGS run to a cluster). Clusters we selected for the ordered set are highlighted with "+" signs on top of the bars.*

S3 Supporting References

(1) Lührs, T.; Ritter, C.; Adrian, M.; Riek-Loher, D.; Bohrmann, B.; Döbeli, H.; Schubert, D.; Riek, R., 3D structure of Alzheimer's amyloid-β (1–42) fibrils. *Proc. Natl. Acad. Sci. USA* **2005,** *102* (48), 17342-17347.

(2) Ban, T.; Hoshino, M.; Takahashi, S.; Hamada, D.; Hasegawa, K.; Naiki, H.; Goto, Y., Direct observation of Aβ amyloid fibril growth and inhibition. *J. Mol. Biol.* **2004,** *344* (3), 757-767.

(3) Kellermayer, M. S. Z.; Karsai, Á.; Benke, M.; Soós, K.; Penke, B., Stepwise dynamics of epitaxially growing single amyloid fibrils. *Proc. Natl. Acad. Sci. USA* **2008,** *105* (1), 141-144.

(4) Han, W.; Schulten, K., Fibril elongation by Aβ17–42: Kinetic network analysis of hybrid-resolution molecular dynamics simulations. *J. Am. Chem. Soc.* **2014,** *136* (35), 12450-12460.

(5) Kumar, A.; Srivastava, S.; Tripathi, S.; Singh, S. K.; Srikrishna, S.; Sharma, A., Molecular insight into amyloid oligomer destabilizing mechanism of flavonoid derivative 2-(4′ benzyloxyphenyl)-3-hydroxy-chromen-4-one through docking and molecular dynamics simulations. *J. Biomol. Struct. Dyn.* **2016,** *34* (6), 1252-1263.

(6) Kuang, G.; Murugan, N. A.; Tu, Y.; Nordberg, A.; Ågren, H., Investigation of the binding profiles of AZD2184 and Thioflavin T with amyloid-β(1–42) fibril by molecular docking and molecular dynamics methods. *J. Phys. Chem. B* **2015,** *119* (35), 11560-11567.

(7) Hernández-Rodríguez, M.; Correa-Basurto, J.; Benitez-Cardoza, C. G.; Resendiz-Albor, A. A.; Rosales-Hernández, M. C., *In silico* and *in vitro* studies to elucidate the role of Cu^{2+} and galanthamine as the limiting step in the amyloid beta (1–42) fibrillation process. *Prot. Sci.* **2013,** *22* (10), 1320-1335.

(8) Rodriguez-Rodriguez, C.; Rimola, A.; Rodriguez-Santiago, L.; Ugliengo, P.; Alvarez-Larena, A.; Gutierrez-de-Teran, H.; Sodupe, M.; Gonzalez-Duarte, P., Crystal structure of thioflavin-T and its binding to amyloid fibrils: Insights at the molecular level. *Chem. Commun.* **2010,** *46* (7), 1156-1158.

(9) Ma, B.; Nussinov, R., Polymorphic C-terminal β-sheet interactions determine the formation of fibril or amyloid β-derived diffusible ligand-like globulomer for the Alzheimer Aβ42 dodecamer. *J. Biol. Chem.* **2010,** *285* (47), 37102-37110.

(10) Miller, Y.; Ma, B.; Nussinov, R., Zinc ions promote Alzheimer Aβ aggregation via population shift of polymorphic states. *Proc. Natl. Acad. Sci. USA* **2010,** *107* (21), 9490-9495.

(11) Autiero, I.; Langella, E.; Saviano, M., Insights into the mechanism of interaction between trehalose-conjugated beta-sheet breaker peptides and Aβ(1-42) fibrils by molecular dynamics simulations. *Mol. Biosyst.* **2013,** *9* (11), 2835-2841.

(12) Lemkul, J. A.; Bevan, D. R., Destabilizing Alzheimer's Aβ42 protofibrils with morin: Mechanistic insights from molecular dynamics simulations. *Biochemistry* **2010,** *49* (18), 3935-3946.

(13) Masman, M. F.; Eisel, U. L. M.; Csizmadia, I. G.; Penke, B.; Enriz, R. D.; Marrink, S. J.; Luiten, P. G. M., *In silico* study of full-length amyloid β 1−42 tri- and penta-oligomers in solution. *J. Phys. Chem. B* **2009,** *113* (34), 11710-11719.

(14) Lemkul, J. A.; Bevan, D. R., Assessing the stability of Alzheimer's amyloid protofibrils using molecular dynamics. *J. Phys. Chem. B* **2010,** *114* (4), 1652-1660.

(15) Zheng, J.; Jang, H.; Ma, B.; Tsai, C.-J.; Nussinov, R., Modeling the Alzheimer Aβ17-42 fibril architecture: Tight intermolecular sheet-sheet association and intramolecular hydrated cavities. *Biophys. J.* **2007,** *93* (9), 3046-3057.

(16) Xi, W.; Wang, W.; Abbott, G.; Hansmann, U. H. E., Stability of a recently found triple-β-stranded Aβ1–42 fibril motif. *J. Phys. Chem. B* **2016,** *120* (20), 4548-4557.

(17) Alred, E. J.; Phillips, M.; Berhanu, W. M.; Hansmann, U. H. E., On the lack of polymorphism in Aβ-peptide aggregates derived from patient brains. *Prot. Sci.* **2015,** *24* (6), 923-935.

(18) Vitalis, A.; Pappu, R. V., Methods for Monte Carlo simulations of biomacromolecules. *Annu. Rep. Comput. Chem.* **2009,** *5*, 49-76.

(19) Best, R. B.; Zhu, X.; Shim, J.; Lopes, P. E. M.; Mittal, J.; Feig, M.; MacKerell, A. D., Optimization of the additive CHARMM all-atom protein force field targeting improved sampling of the backbone φ, ψ and side-chain χ1 and χ2 dihedral angles. *J. Chem. Theory Comput.* **2012,** *8* (9), 3257-3273.
(20) Vitalis, A.; Pappu, R. V., ABSINTH: A new continuum solvation model for simulations of polypeptides in aqueous solutions. *J. Comput. Chem.* **2009,** *30* (5), 673-699.
(21) Pronk, S.; Páll, S.; Schulz, R.; Larsson, P.; Bjelkmar, P.; Apostolov, R.; Shirts, M. R.; Smith, J. C.; Kasson, P. M.; van der Spoel, D.; Hess, B.; Lindahl, E., GROMACS 4.5: A high-throughput and highly parallel open source molecular simulation toolkit. *Bioinformatics* **2013,** *29* (7), 845-854.
(22) Bussi, G.; Donadio, D.; Parrinello, M., Canonical sampling through velocity rescaling. *J. Chem. Phys.* **2007,** *126* (1), 014101.
(23) Ryckaert, J. P.; Ciccotti, G.; Berendsen, H. J. C., Numerical integration of the Cartesian equations of motion of a system with constraints: Molecular dynamics of *n*-alkanes. *J. Comput. Phys.* **1977,** *23*, 327-341.
(24) Tironi, I. G.; Sperb, R.; Smith, P. E.; van Gunsteren, W. F., A generalized reaction field method for molecular dynamics simulations. *J. Chem. Phys.* **1995,** *102* (13), 5451-5459.
(25) Bacci, M.; Vitalis, A.; Caflisch, A., A molecular simulation protocol to avoid sampling redundancy and discover new states. *Biochim. Biophys. Acta* **2015,** *1850* (5), 889-902.
(26) Blöchliger, N.; Caflisch, A.; Vitalis, A., Weighted distance functions improve analysis of high-dimensional data: Application to molecular dynamics simulations. *J. Chem. Theory Comput.* **2015,** *11* (11), 5481-5492.
(27) Pan, A. C.; Sezer, D.; Roux, B., Finding transition pathways using the string method with swarms of trajectories. *J. Phys. Chem. B* **2008,** *112* (11), 3432-3440.
(28) Dickson, A.; Brooks III, C. L., WExplore: Hierarchical exploration of high-dimensional spaces using the weighted ensemble algorithm. *J. Phys. Chem. B* **2014,** *118* (13), 3532-3542.
(29) Mackerell, A. D.; Feig, M.; Brooks, C. L., Extending the treatment of backbone energetics in protein force fields: Limitations of gas-phase quantum mechanics in reproducing protein conformational distributions in molecular dynamics simulations. *J. Comput. Chem.* **2004,** *25* (11), 1400-1415.
(30) Bonomi, M.; Branduardi, D.; Bussi, G.; Camilloni, C.; Provasi, D.; Raiteri, P.; Donadio, D.; Marinelli, F.; Pietrucci, F.; Broglia, R. A.; Parrinello, M., PLUMED: A portable plugin for free-energy calculations with molecular dynamics. *Comput. Phys. Commun.* **2009,** *180* (10), 1961-1972.
(31) Pande, V. S.; Beauchamp, K.; Bowman, G. R., Everything you wanted to know about Markov State Models but were afraid to ask. *Methods* **2010,** *52* (1), 99-105.
(32) Prinz, J.-H.; Wu, H.; Sarich, M.; Keller, B.; Senne, M.; Held, M.; Chodera, J. D.; Schütte, C.; Noé, F., Markov models of molecular kinetics: Generation and validation. *J. Chem. Phys.* **2011,** *134* (17), 174105.
(33) Chodera, J. D.; Noé, F., Markov state models of biomolecular conformational dynamics. *Curr. Opin. Struct. Biol.* **2014,** *25*, 135-144.
(34) Chodera, J. D.; Swope, W. C.; Pitera, J. W.; Dill, K. A., Long-time protein folding dynamics from short-time molecular dynamics simulations. *Multiscale Model. Simul.* **2006,** *5* (4), 1214-1226.
(35) Noé, F.; Fischer, S., Transition networks for modeling the kinetics of conformational change in macromolecules. *Curr. Opin. Struct. Biol.* **2008,** *18* (2), 154-162.
(36) Huang, X.; Bowman, G. R.; Bacallado, S.; Pande, V. S., Rapid equilibrium sampling initiated from nonequilibrium data. *Proc. Natl. Acad. Sci. USA* **2009,** *106* (47), 19765-19769.
(37) Buchete, N. V.; Hummer, G., Coarse master equations for peptide folding dynamics. *J. Phys. Chem. B* **2008,** *112* (19), 6057-6069.
(38) Metzner, P.; Schütte, C.; Vanden-Eijnden, E., Transition path theory for Markov jump processes. *Multiscale Model. Simul.* **2009,** *7* (3), 1192-1219.
(39) Berezhkovskii, A.; Hummer, G.; Szabo, A., Reactive flux and folding pathways in network models of coarse-grained protein dynamics. *J. Chem. Phys.* **2009,** *130* (20), 205102.
(40) Noé, F.; Schütte, C.; Vanden-Eijnden, E.; Reich, L.; Weikl, T. R., Constructing the equilibrium ensemble of folding pathways from short off-equilibrium simulations. *Proc. Natl. Acad. Sci. USA* **2009,** *106* (45), 19011-19016.

(41) Kabsch, W.; Sander, C., Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **1983,** *22* (12), 2577-2637.
(42) Jolliffe, I., *Principal component analysis*. Wiley Online Library: 2005.
(43) Krivov, S. V.; Karplus, M., One-dimensional free-energy profiles of complex systems: Progress variables that preserve the barriers. *J. Phys. Chem. B* **2006,** *110* (25), 12689-12698.
(44) Vitalis, A.; Caflisch, A., Efficient construction of mesostate networks from molecular dynamics trajectories. *J. Chem. Theory Comput.* **2012,** *8* (3), 1108-1120.
(45) Shao, J.; Tanner, S. W.; Thompson, N.; Cheatham, T. E., Clustering molecular dynamics trajectories: 1. Characterizing the performance of different clustering algorithms. *J. Chem. Theory Comput.* **2007,** *3* (6), 2312-2334.
(46) Bowman, G. R.; Beauchamp, K. A.; Boxer, G.; Pande, V. S., Progress and challenges in the automated construction of Markov state models for full protein systems. *J. Chem. Phys.* **2009,** *131* (12), 124101.
(47) Dijkstra, E. W., A note on two problems in connexion with graphs. *Numer. Math.* **1959,** *1* (1), 269-271.
(48) Humphrey, W.; Dalke, A.; Schulten, K., VMD: Visual molecular dynamics. *J. Mol. Graph.* **1996,** *14* (1), 33-38.