Supporting Information for: Automated detection of many-particle solvation states for accurate characterizations of diffusion kinetics

Joseph F. Rudzinski,* Marc Radu, and Tristan Bereau

Max Planck Institute for Polymer Research, Mainz 55128, Germany

E-mail: rudzinski@mpip-mainz.mpg.de

Overview

In the following we present additional technical details and results to support the main text. To be concise, we employ the notation from the main text wherever convenient.

Methods

Weighted coordination numbers

The weighted coordination numbers (WCNs) are described in the main text. Fig. S1 presents the Gaussian weights for the A-A pairs.

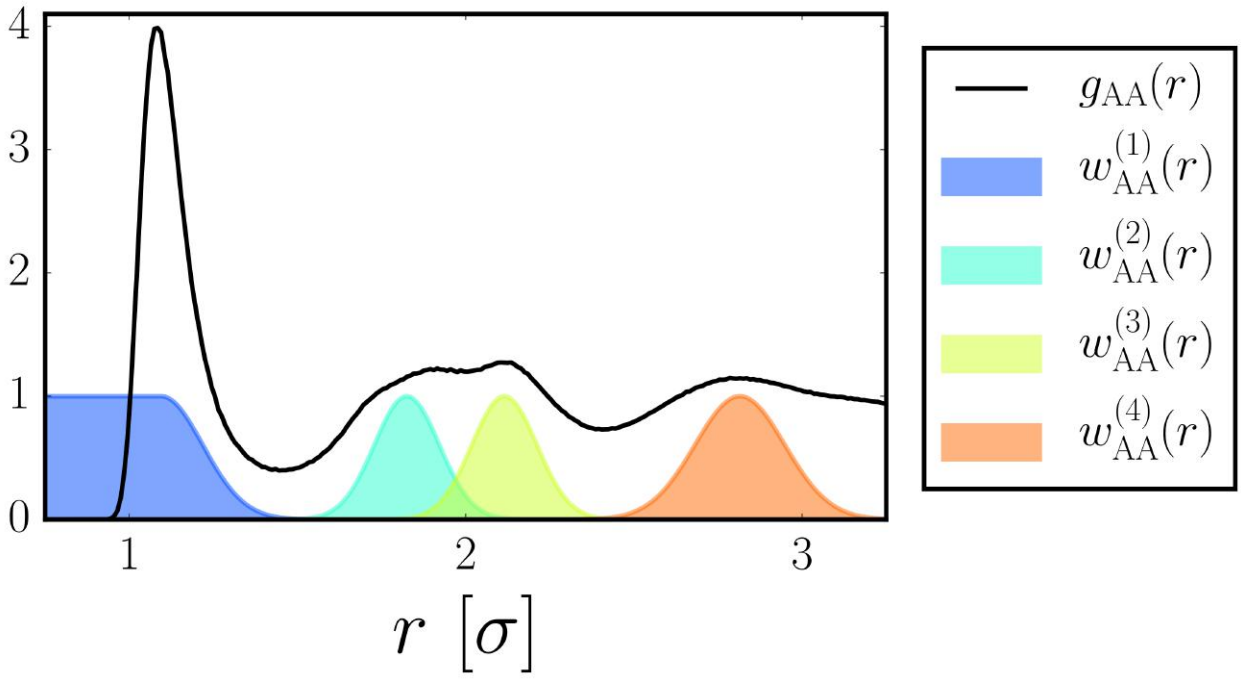

Figure S1: Gaussian weights (colored distributions) for determining the weighted coordination number, as described in the main text. Radial distribution function (solid black curve) between pairs of A particles is also presented.

Hidden-Markov-model filtering

Selection of hyperparameters

To construct Hidden Markov models (HMMs) for each particle type, surrounding particle type, and solvation feature, the number of hidden and emission state must be specified. For the models presented in the main text, we employed 20 emission states, corresponding to a simple discretization along the respective WCN feature. We performed an "implied timescale test" by constructing models with varying numbers of hidden states (from 2-6) and varying lag times (from 2-20 time steps, 1 step = 0.5 t^*). The results are presented in Figs. S2-S13. For a reasonable HMM, the timescales should converge with increasing lag time. One can make the following general observations from the implied timescale tests:

1. For too few hidden states, the implied timescales do not converge over the lag times considered.

2. For too many hidden states, the implied timescales demonstrate unstable behavior as the lag time is increased.

3. In the latter case, there is often a (small) regime at smaller lag times where the timescales are approximately constant.

These observations reflect the balance between having enough states to accurately characterize the underlying landscape and having few enough states such that there is enough sampling of the relevant transitions in the given data. To compromise between these two factors, and to avoid hand picking hyperparameters for individual features, we chose to employ four hidden states with a lag time of 8 steps for all models (indicated by vertical red lines in each figure). Note that the results presented in Figs. S2-S13 were obtained using a finer discretization of 25 emission states for each WCN. We later decreased this number to 20, which should slightly improve the robustness of the models for reasonably short lag times.

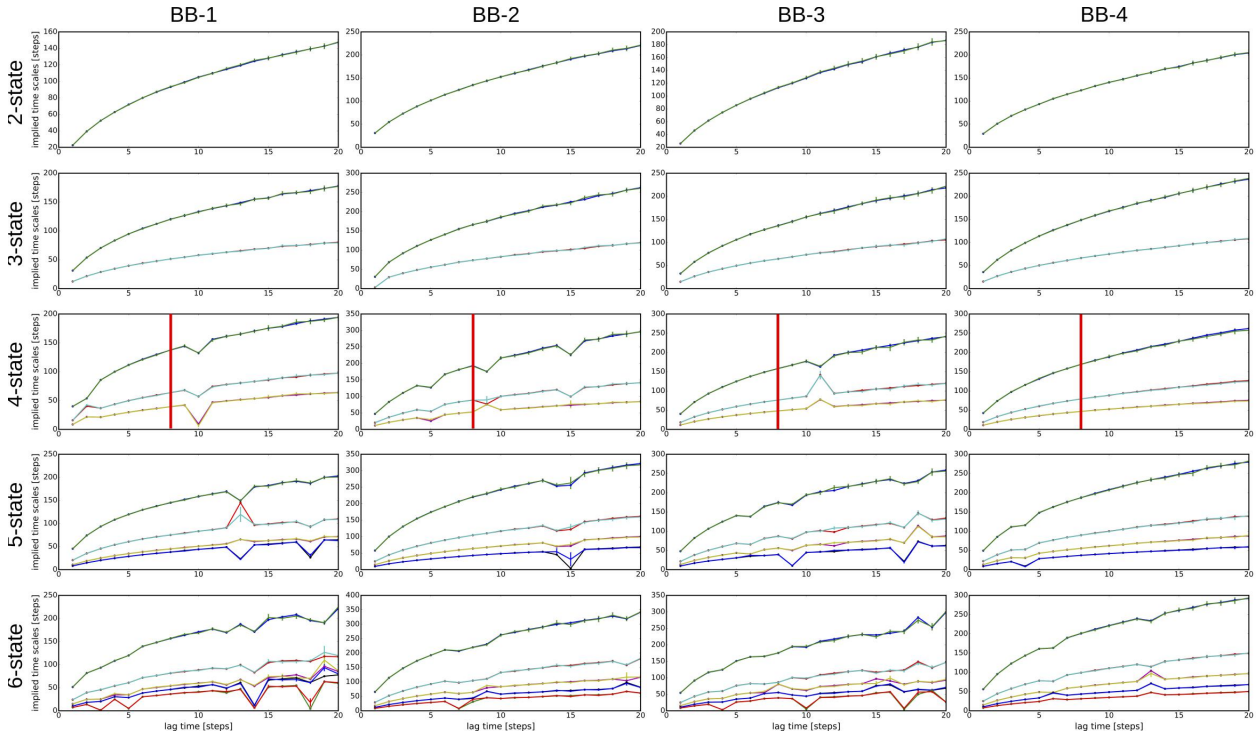

Figure S2: Implied timescale tests for HMMs constructed from WCN trajectories for features describing the solvation of B particles around a center B particle from the $T^* = 0.4$ simulations. Markers without error bars present the timescales determined from the maximum likelihood estimate with the corresponding lag time. Markers with errors bars present the average timescale from all samples obtained from the Bayesian error analysis. The columns correspond to particular solvation features, as described in the main text and shown in Fig. 2. The rows present results from different numbers of hidden states chosen for the HMM. The 4-state model with lag time equal to 8 steps (vertical red line) corresponds to the model used for filtering the WCN trajectories in all cases.

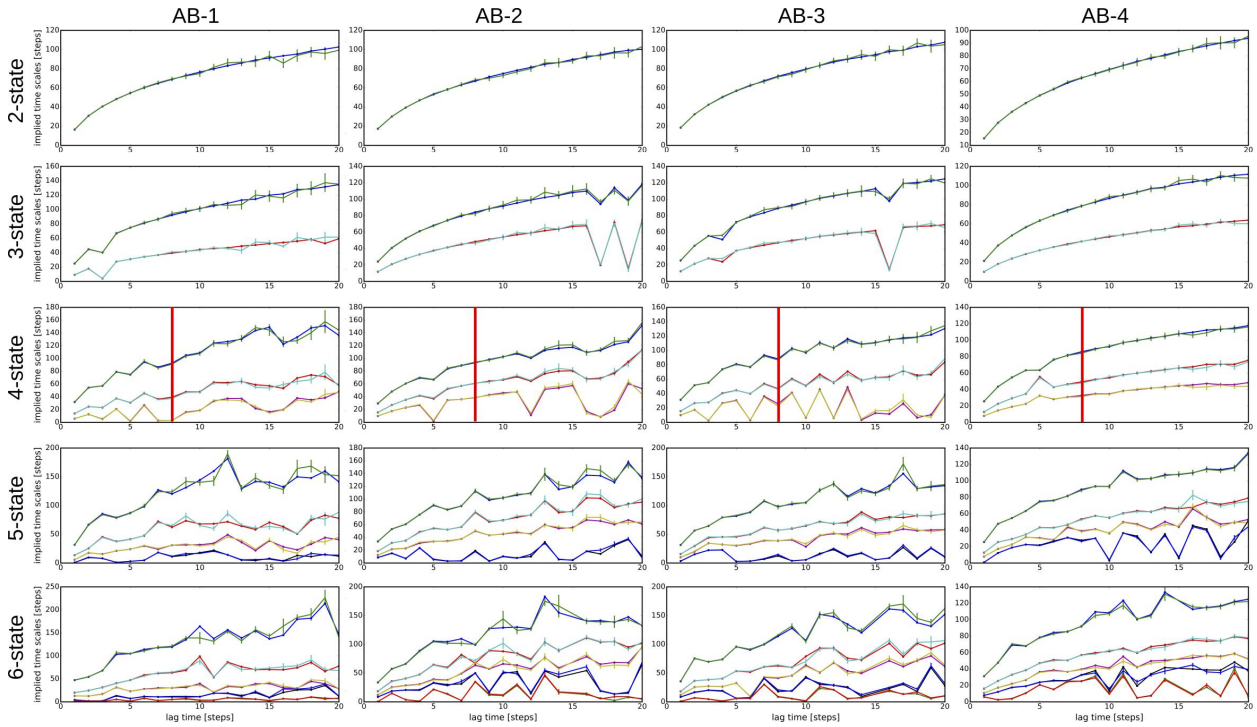

Figure S3: Implied timescale tests for HMMs constructed from WCN trajectories for features describing the solvation of A particles around a center B particle from the $T^* = 0.4$ simulations. Markers without error bars present the timescales determined from the maximum likelihood estimate with the corresponding lag time. Markers with errors bars present the average timescale from all samples obtained from the Bayesian error an alysis. The columns correspond to particular solvation features, as described in the main text and shown in Fig. 2. The rows present results from different numbers of hidden states chosen for the HMM. The 4-state model with lag time equal to 8 steps (vertical red line) corresponds to the model used for filtering the WCN trajectories in all cases.

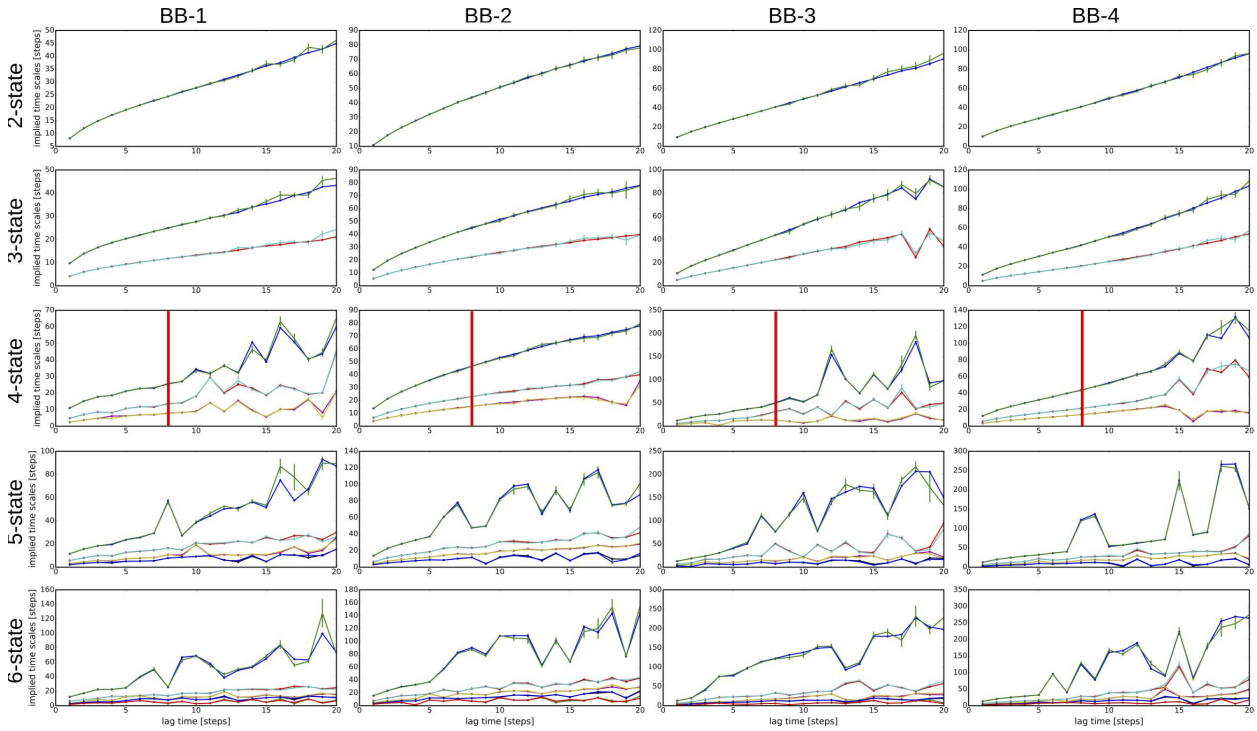

Figure S4: Implied timescale tests for HMMs constructed from WCN trajectories for features describing the solvation of B particles around a center B particle from the $T^* = 0.5$ simulations. Markers without error bars present the timescales determined from the maximum likelihood estimate with the corresponding lag time. Markers with errors bars present the average timescale from all samples obtained from the Bayesian error an alysis. The columns correspond to particular solvation features, as described in the main text and shown in Fig. 2. The rows present results from different numbers of hidden states chosen for the HMM. The 4-state model with lag time equal to 8 steps (vertical red line) corresponds to the model used for filtering the WCN trajectories in all cases.

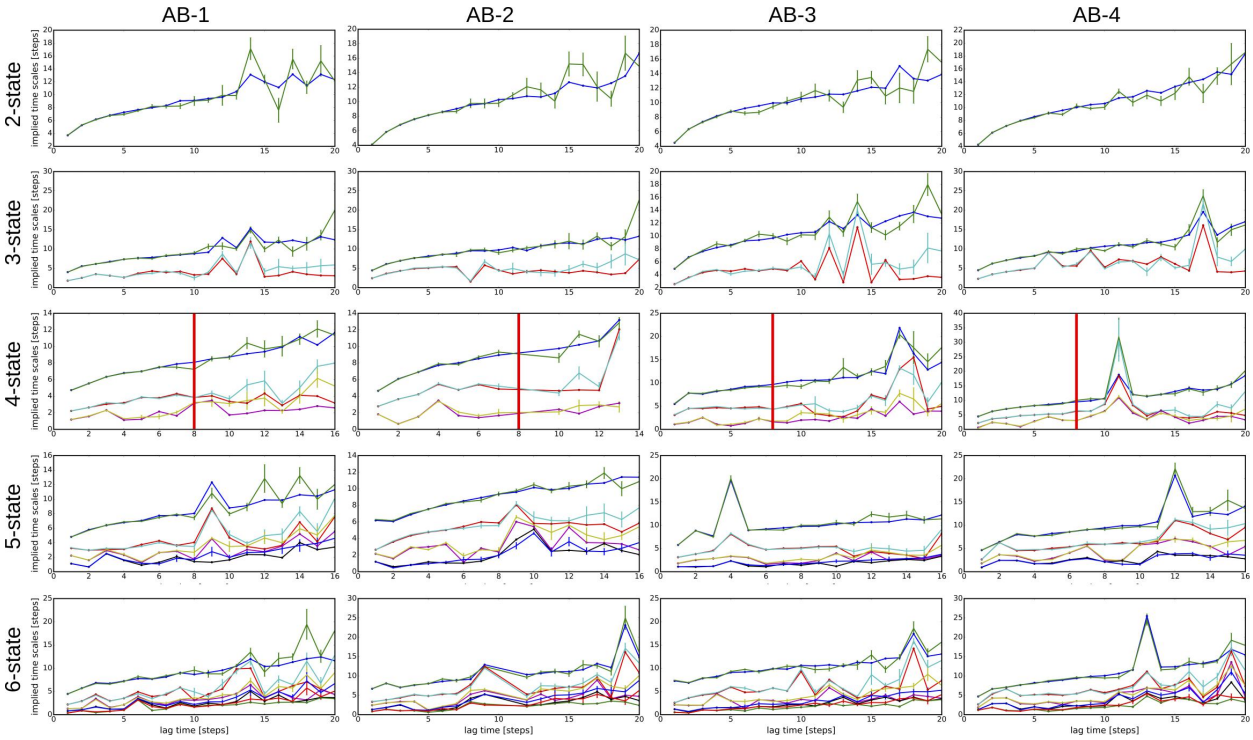

Figure S5: Implied timescale tests for HMMs constructed from WCN trajectories for features describing the solvation of A particles around a center B particle from the $T^* = 0.5$ simulations. Markers without error bars present the timescales determined from the maximum likelihood estimate with the corresponding lag time. Markers with errors bars present the average timescale from all samples obtained from the Bayesian error an alysis. The columns correspond to particular solvation features, as described in the main text and shown in Fig. 2. The rows present results from different numbers of hidden states chosen for the HMM. The 4-state model with lag time equal to 8 steps (vertical red line) corresponds to the model used for filtering the WCN trajectories in all cases.

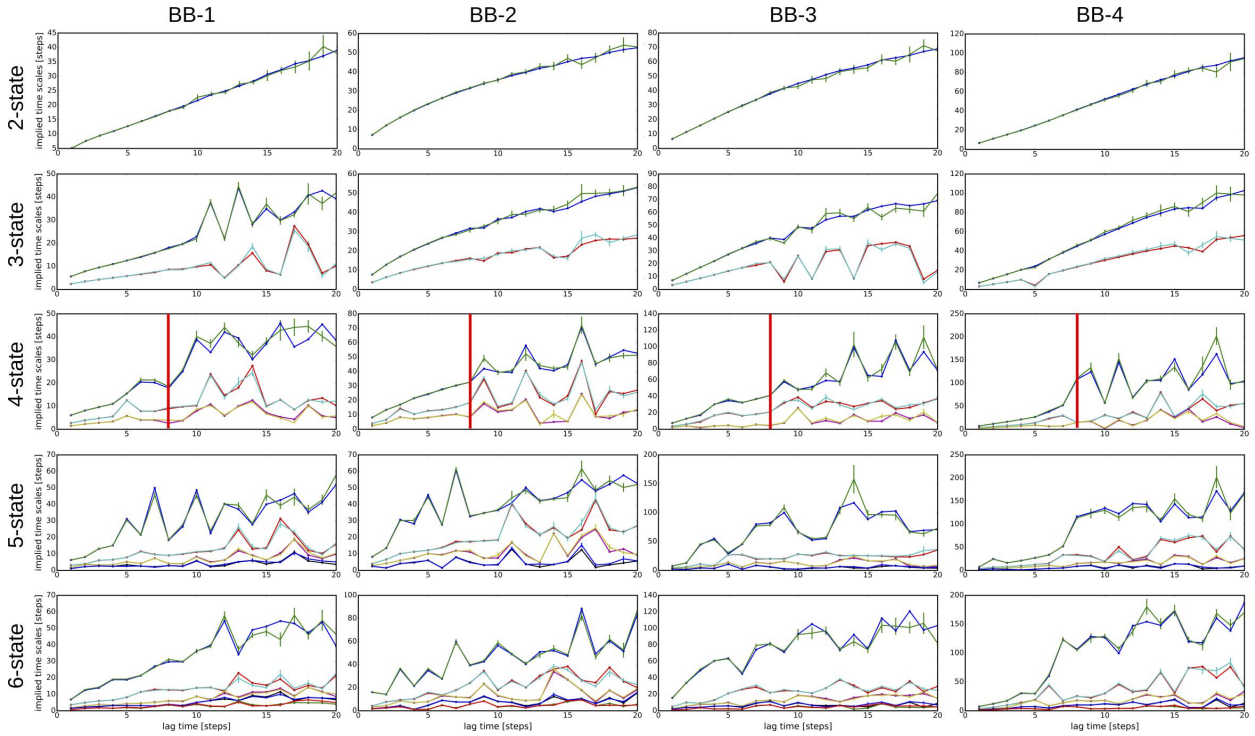

Figure S6: Implied timescale tests for HMMs constructed from WCN trajectories for features describing the solvation of B particles around a center B particle from the $T^* = 0.6$ simulations. Markers without error bars present the timescales determined from the maximum likelihood estimate with the corresponding lag time. Markers with errors bars present the average timescale from all samples obtained from the Bayesian error an alysis. The columns correspond to particular solvation features, as described in the main text and shown in Fig. 2. The rows present results from different numbers of hidden states chosen for the HMM. The 4-state model with lag time equal to 8 steps (vertical red line) corresponds to the model used for filtering the WCN trajectories in all cases.

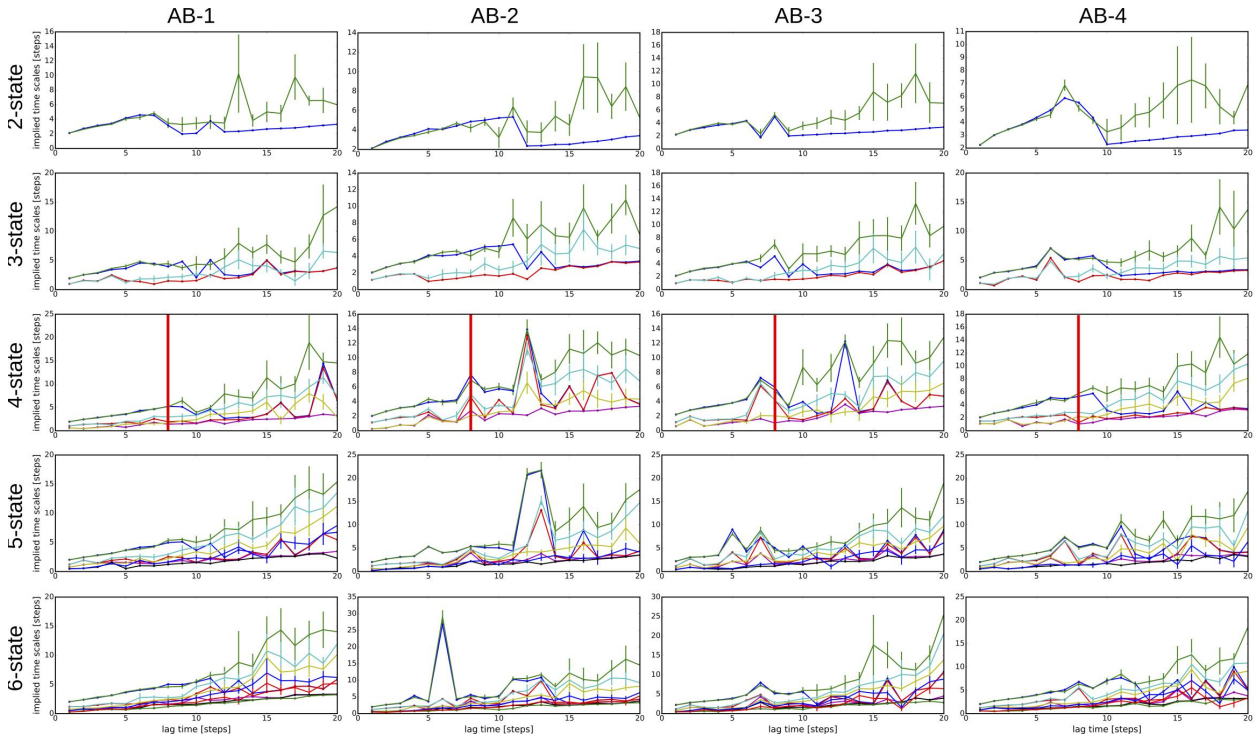

Figure S7: Implied timescale tests for HMMs constructed from WCN trajectories for features describing the solvation of A particles around a center B particle from the $T^* = 0.6$ simulations. Markers without error bars present the timescales determined from the maximum likelihood estimate with the corresponding lag time. Markers with errors bars present the average timescale from all samples obtained from the Bayesian error an alysis. The columns correspond to particular solvation features, as described in the main text and shown in Fig. 2. The rows present results from different numbers of hidden states chosen for the HMM. The 4-state model with lag time equal to 8 steps (vertical red line) corresponds to the model used for filtering the WCN trajectories in all cases.

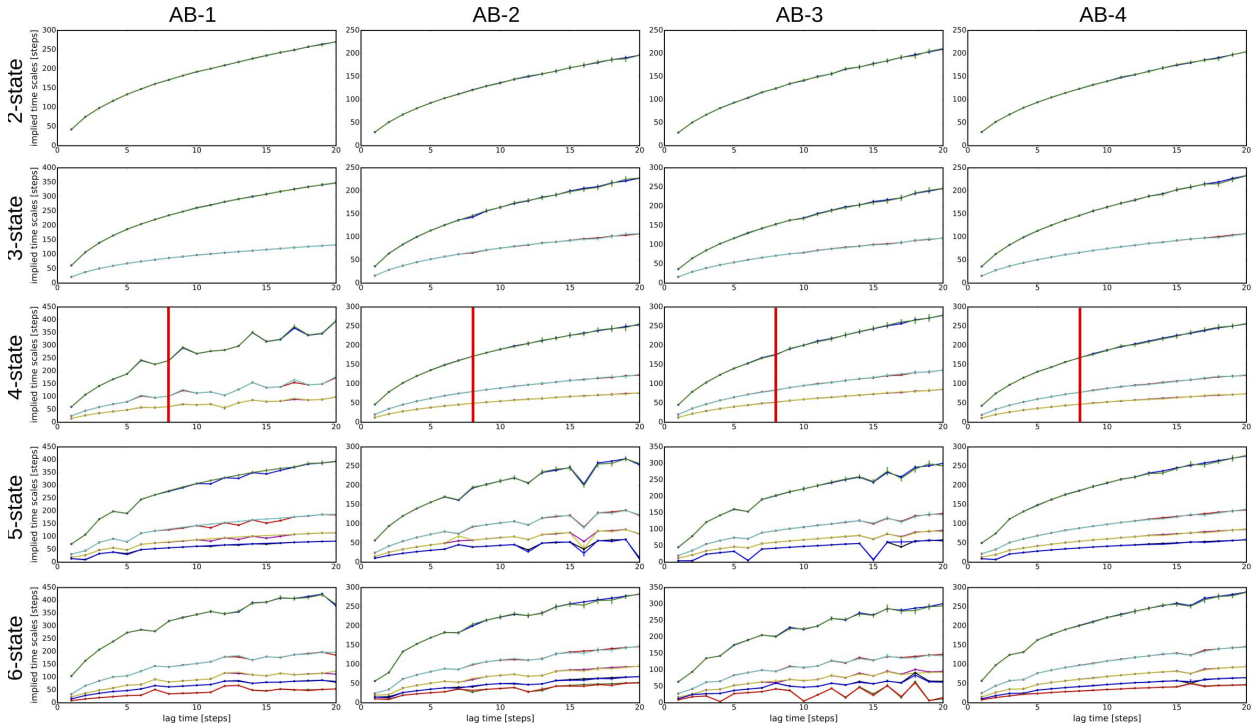

Figure S8: Implied timescale tests for HMMs constructed from WCN trajectories for features describing the solvation of B particles around a center A particle from the $T^* = 0.4$ simulations. Markers without error bars present the timescales determined from the maximum likelihood estimate with the corresponding lag time. Markers with errors bars present the average timescale from all samples obtained from the Bayesian error an alysis. The columns correspond to particular solvation features, as described in the main text and shown in Fig. 2. The rows present results from different numbers of hidden states chosen for the HMM. The 4-state model with lag time equal to 8 steps (vertical red line) corresponds to the model used for filtering the WCN trajectories in all cases.

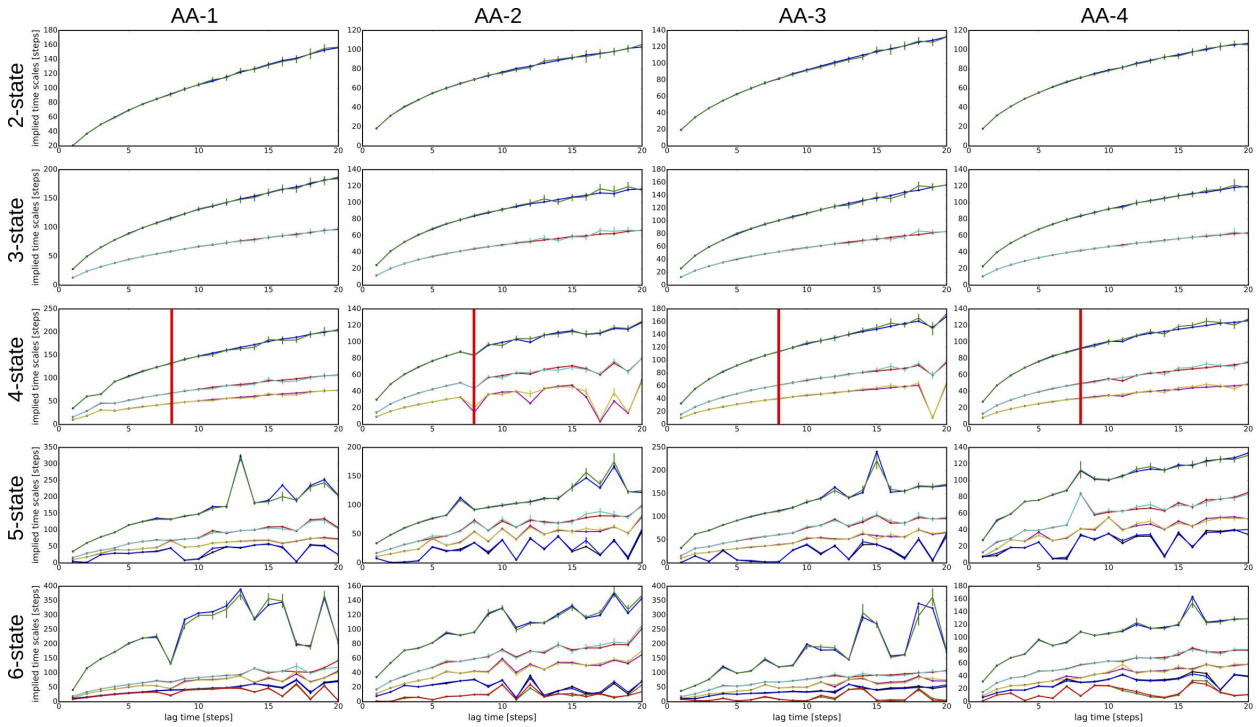

Figure S9: Implied timescale tests for HMMs constructed from WCN trajectories for features describing the solvation of A particles around a center A particle from the $T^* = 0.4$ simulations. Markers without error bars present the timescales determined from the maximum likelihood estimate with the corresponding lag time. Markers with errors bars present the average timescale from all samples obtained from the Bayesian error an alysis. The columns correspond to particular solvation features, as described in the main text and shown in Fig. S1. The rows present results from different numbers of hidden states chosen for the HMM. The 4-state model with lag time equal to 8 steps (vertical red line) corresponds to the model used for filtering the WCN trajectories in all cases.

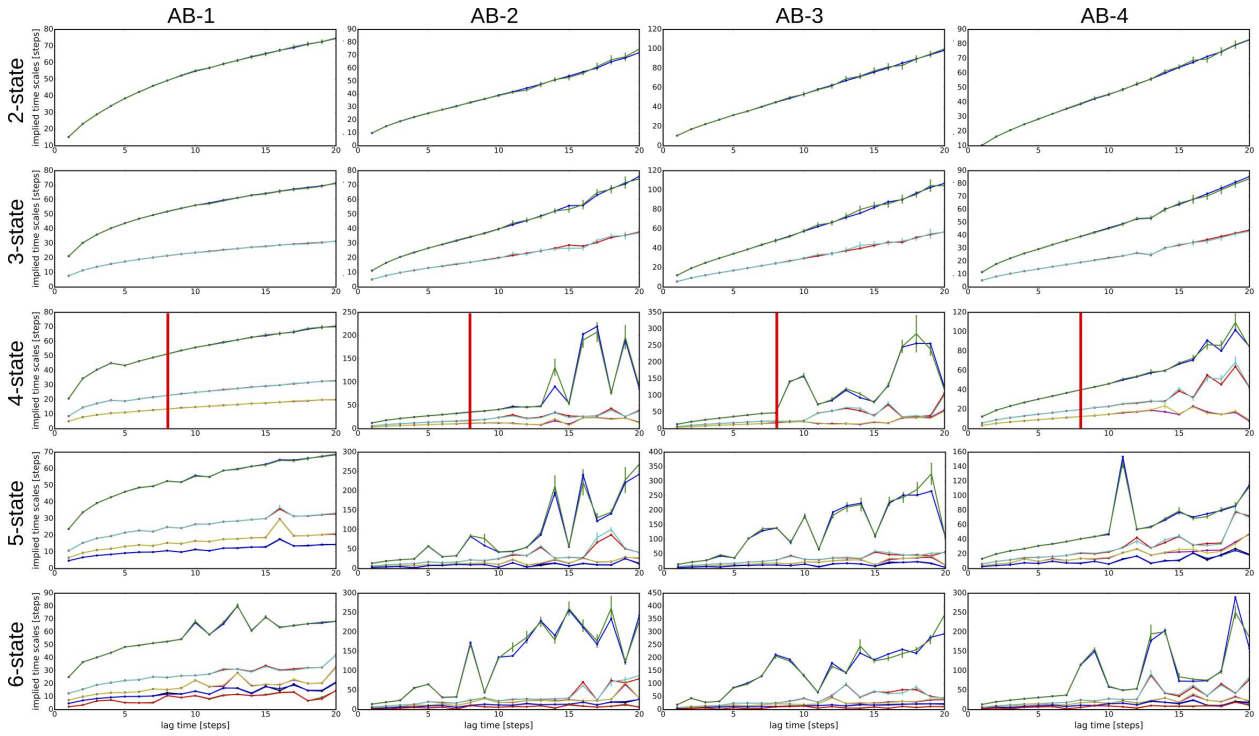

Figure S10: Implied timescale tests for HMMs constructed from WCN trajectories for features describing the solvation of B particles around a center A particle from the $T^* = 0.5$ simulations. Markers without error bars present the timescales determined from the maximum likelihood estimate with the corresponding lag time. Markers with errors bars present the average timescale from all samples obtained from the Bayesian error an alysis. The columns correspond to particular solvation features, as described in the main text and shown in Fig. 2. The rows present results from different numbers of hidden states chosen for the HMM. The 4-state model with lag time equal to 8 steps (vertical red line) corresponds to the model used for filtering the WCN trajectories in all cases.

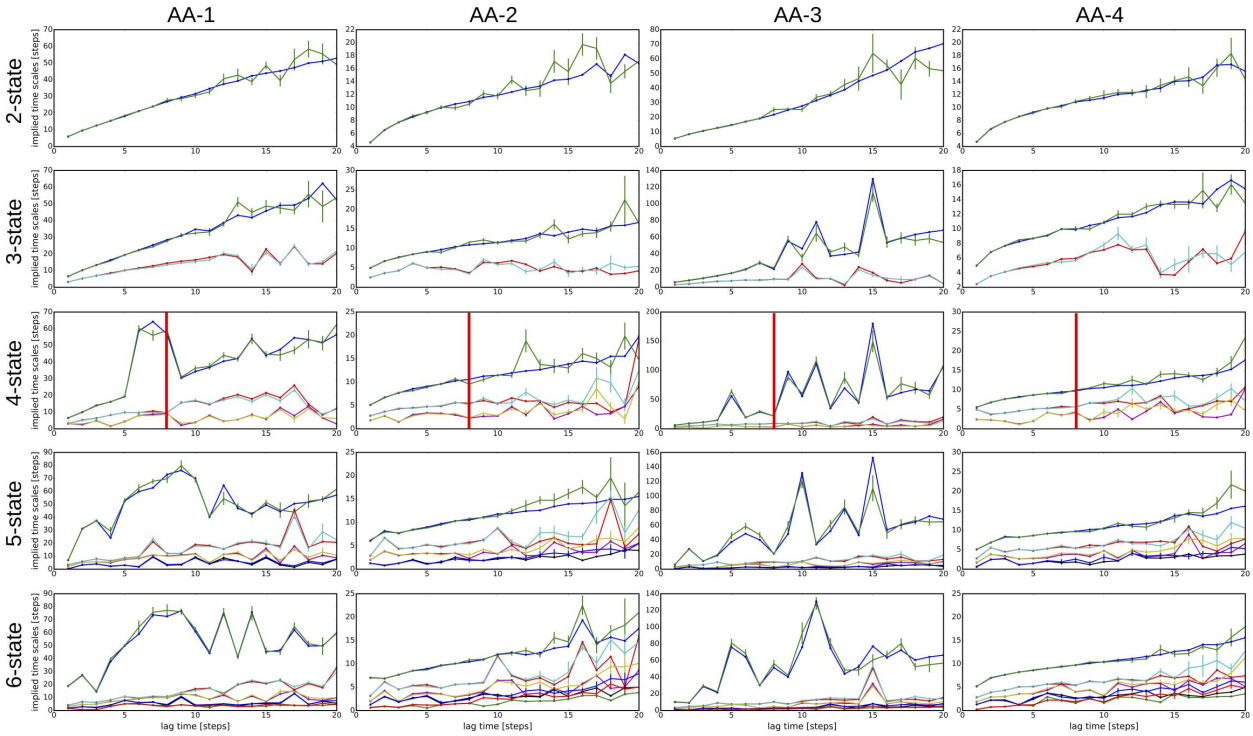

Figure S11: Implied timescale tests for HMMs constructed from WCN trajectories for features describing the solvation of A particles around a center A particle from the $T^* = 0.5$ simulations. Markers without error bars present the timescales determined from the maximum likelihood estimate with the corresponding lag time. Markers with errors bars present the average timescale from all samples obtained from the Bayesian error an alysis. The columns correspond to particular solvation features, as described in the main text and shown in Fig. S1. The rows present results from different numbers of hidden states chosen for the HMM. The 4-state model with lag time equal to 8 steps (vertical red line) corresponds to the model used for filtering the WCN trajectories in all cases.

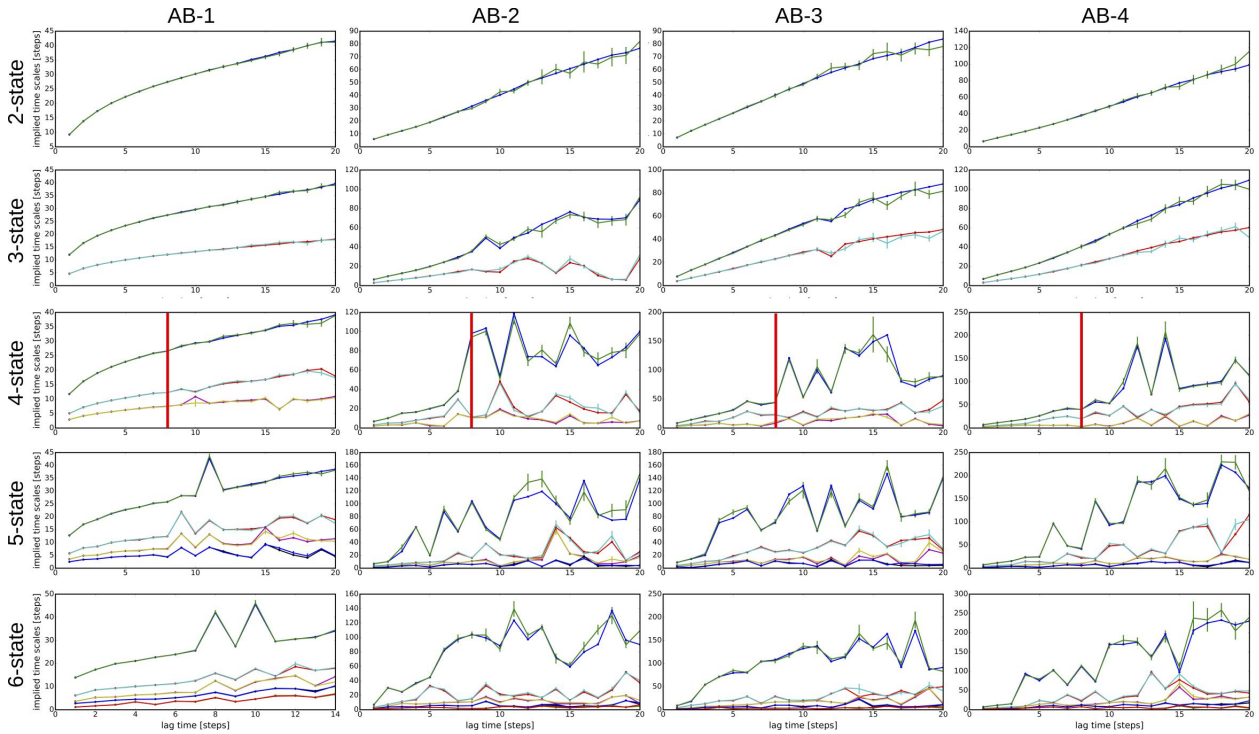

Figure S12: Implied timescale tests for HMMs constructed from WCN trajectories for features describing the solvation of B particles around a center A particle from the $T^* = 0.6$ simulations. Markers without error bars present the timescales determined from the maximum likelihood estimate with the corresponding lag time. Markers with errors bars present the average timescale from all samples obtained from the Bayesian error an alysis. The columns correspond to particular solvation features, as described in the main text and shown in Fig. 2. The rows present results from different numbers of hidden states chosen for the HMM. The 4-state model with lag time equal to 8 steps (vertical red line) corresponds to the model used for filtering the WCN trajectories in all cases.

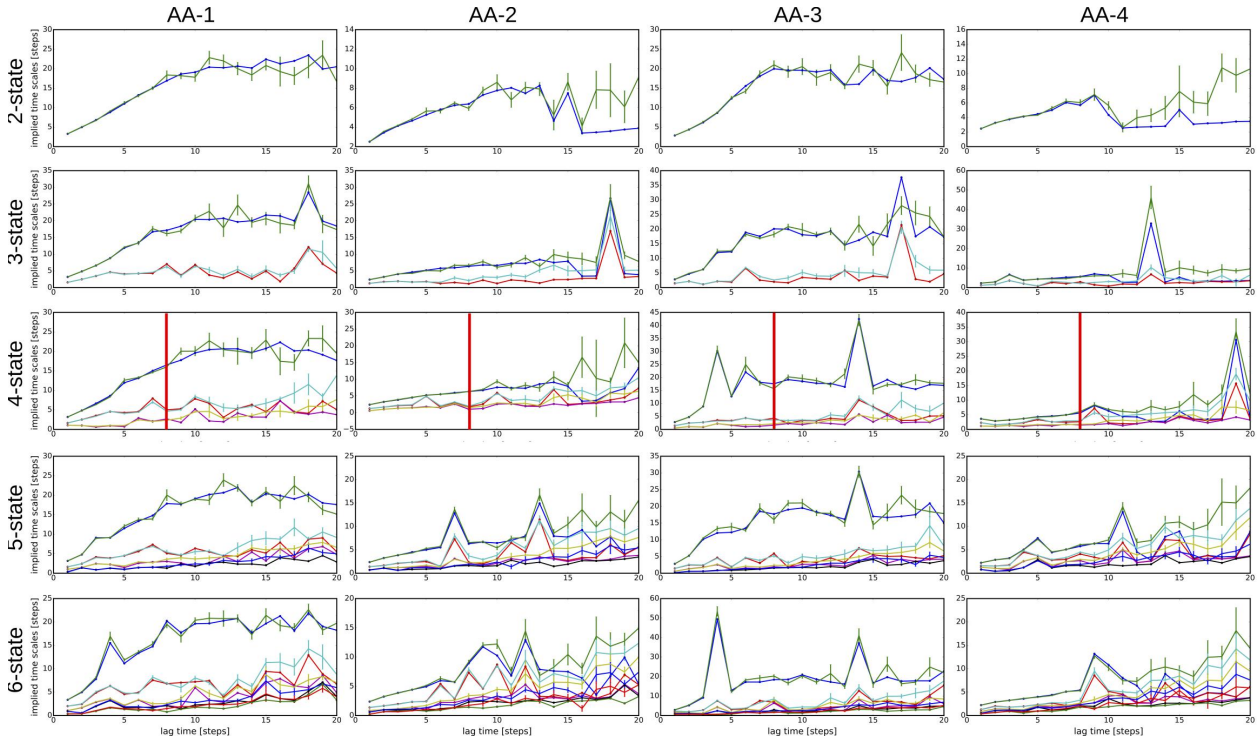

Figure S13: Implied timescale tests for HMMs constructed from WCN trajectories for features describing the solvation of A particles around a center A particle from the $T^* = 0.6$ simulations. Markers without error bars present the timescales determined from the maximum likelihood estimate with the corresponding lag time. Markers with errors bars present the average timescale from all samples obtained from the Bayesian error an alysis. The columns correspond to particular solvation features, as described in the main text and shown in Fig. S1. The rows present results from different numbers of hidden states chosen for the HMM. The 4-state model with lag time equal to 8 steps (vertical red line) corresponds to the model used for filtering the WCN trajectories in all cases.

Cross validation of HMMs

We checked the accuracy of the HMM filtering by cross-validation for the models built from trajectories of A particles. Each HMM was parametrized from the trajectories of 1000 particles. Since there are a total of 4000 A particles in the system, we constructed four different HMMs for each feature. We then compared the resulting filtered trajectories, counting the fraction of frames in which 1. all models predicted the same hidden state, 2. three out of four models predicted the same hidden states, 3. at most two models predicted the same hidden state, or 4. each model predicted a different hidden state.

Fig. S14 presents this analysis for each AA and AB feature and each temperature. In all cases, there are a negligible number of frames where all four models predict different states. At the lowest temperature, the model construction is very robust, with most features demonstrating consistent predictions for the majority of frames. The are a few exceptions in which an outlier model exists, i.e., three out of four models predict consistently for the majority of frames. As the temperature is increased, the consistency of the predictions falls significantly, especially for the AA features. This likely explains the inability of the procedure to generate multi-state diffusion models for B particles at $T = 0.6$. However, quite remarkably, despite these inconsistencies in the model predictions, the diffusion constant for A particles is very accurately described by the resulting models at high temperatures. These models are likely rescued by the semi-robust predictions of the AB features. It appears that the potential inaccuracy of the HMM for the AA features result in predictions which yield little to no correlations with the more accurate predictions of the AB features. As a consequence, these discrepancies are effectively removed during the dimensionality reduction. We also note that for the A-particle filtering, we determined the predicted state by considering all four models. In case of an indeterminate prediction (i.e., two predictions for one state and two predictions for another), we flipped a coin to determine the predicted state.

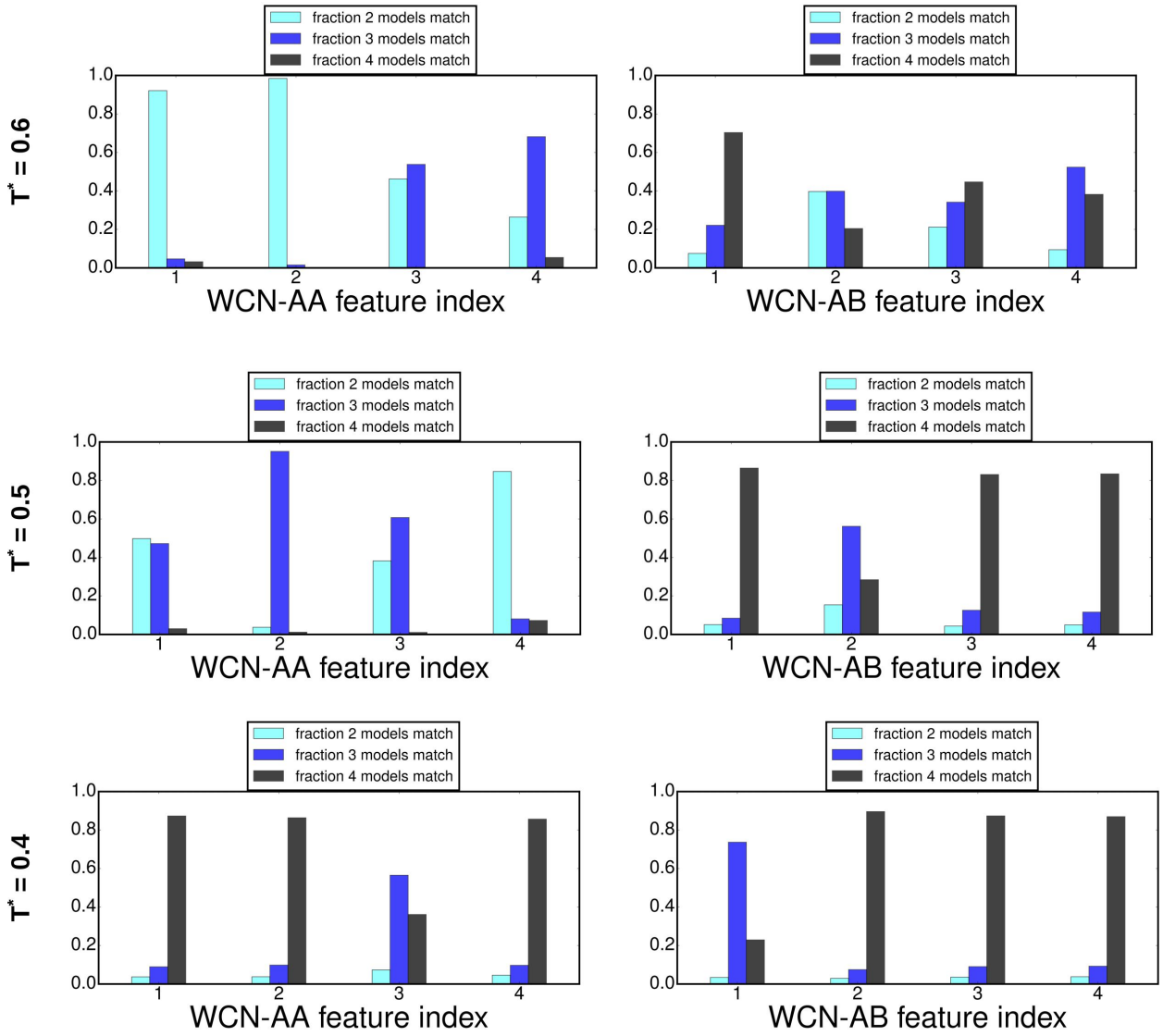

Figure S14: Cross validation of HMMs constructed from trajectories of A particles. The x-axis indicates the feature index for a particle pair type, as defined in Figs. 2 and S1. The y-axis quantifies the fraction of frames in which 4 (black), 3 (blue), or 2 (cyan) models predict the same hidden state for each individual feature.

To further analyze the potential errors arising from the HMMs, we quantified the distribution of segment lengths in cases where the predictions from the four models were indeterminate. In particular, we defined an indeterminate trajectory segment as a set of consecutive frames where two of the models predicted a particular hidden state and the other two models predicted a different hidden state. Figs. S15-S17 present the distributions for each of the AA and AB features for each temperature. Each panel of these figures quantifies the fraction of indeterminate segments as a function of length. We note that total number of indeterminate frames represents a very small fraction of total frames, except for the AA features at $T = 0.6$. These figures demonstrate that in all cases the majority of indeterminate segments have length 5 or less. This result is most likely highlighting the uncertainty of transitions between states during the HMM filtering. Recall, the emission signal (WCN trajectory) is related probabilistically to the hidden states in the context of the HMM.

In light of these results and considering the subsequent processing of trajectories via dimensionality reduction, clustering, coarse-graining, and coring, it is not so surprising that the apparent discrepancies in the predicted states of the HMMs do not propagate to the predictions of the diffusion constant. First, the dimensionality reduction removes highly uncertain or uncorrelated features in terms of the HMM prediction. The coarse-graining of the initial microstates identifies the relevant minima of the underlying landscape. Finally, the coring procedure effectively ignores any differences in the HMM predictions that have to due with the precise placement of the transition between these minima (i.e., the metastable states). Ultimately, these factors remove any potential artifacts of the HMM filtering from the resulting kinetic model.

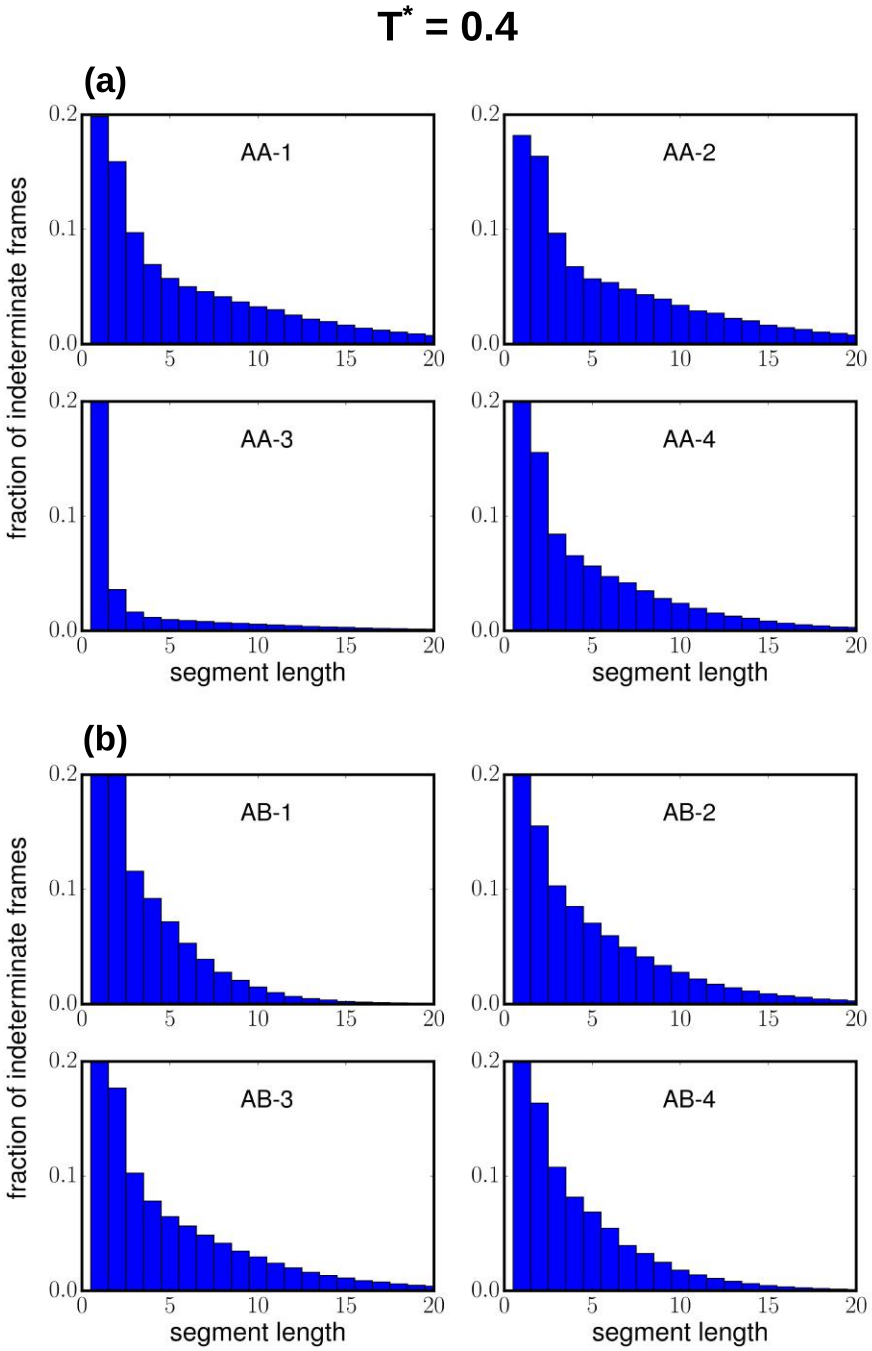

Figure S15: Cross validation of HMMs constructed from trajectories of A particles. The x-axis denotes the length of a trajectory segment (i.e., a consecutive set of frames). The y-axis quantifies the fraction of indeterminate frames with a corresponding indeterminate segment length.

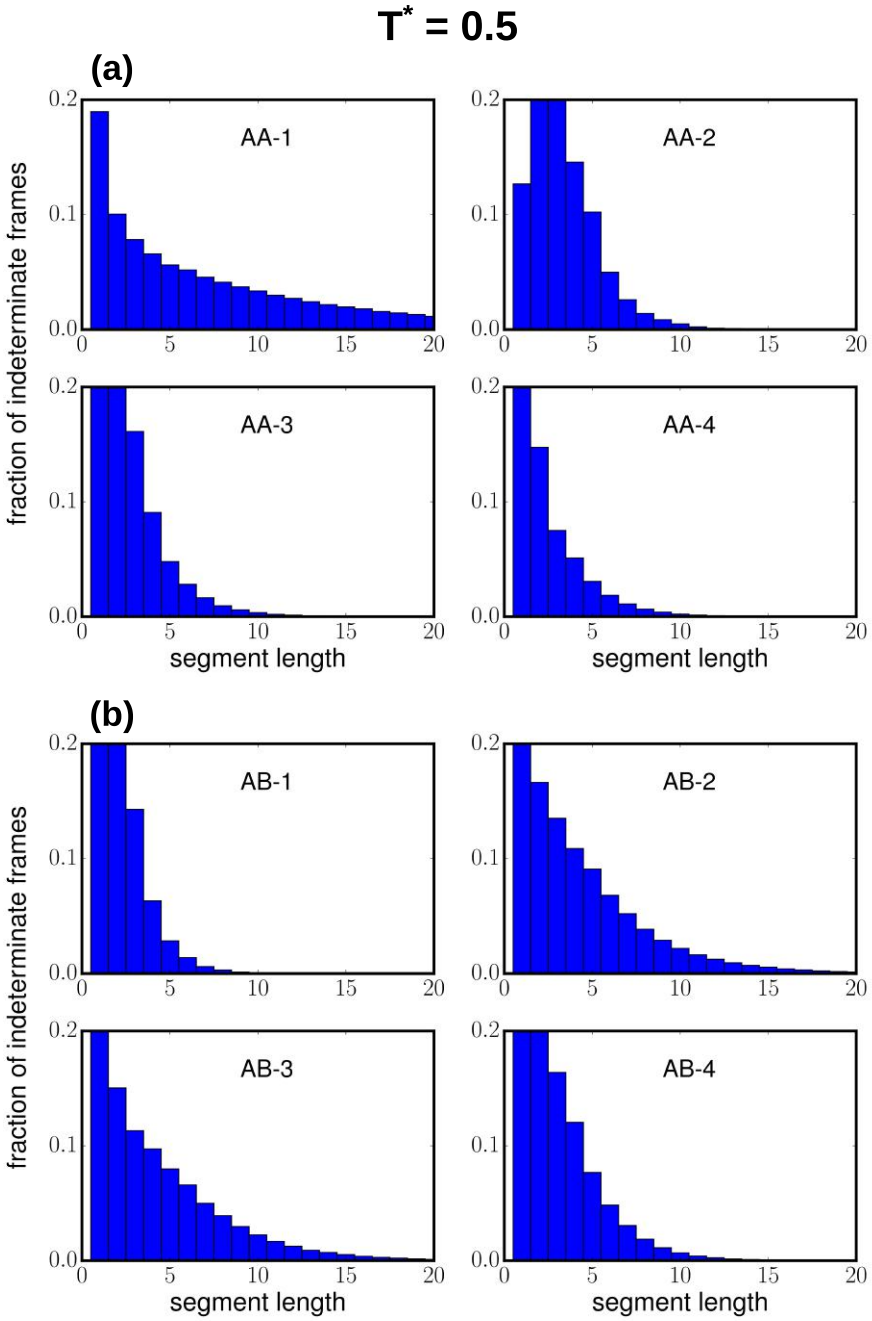

Figure S16: Cross validation of HMMs constructed from trajectories of A particles. The x-axis denotes the length of a trajectory segment (i.e., a consecutive set of frames). The y-axis quantifies the fraction of indeterminate frames with a corresponding indeterminate segment l ength.

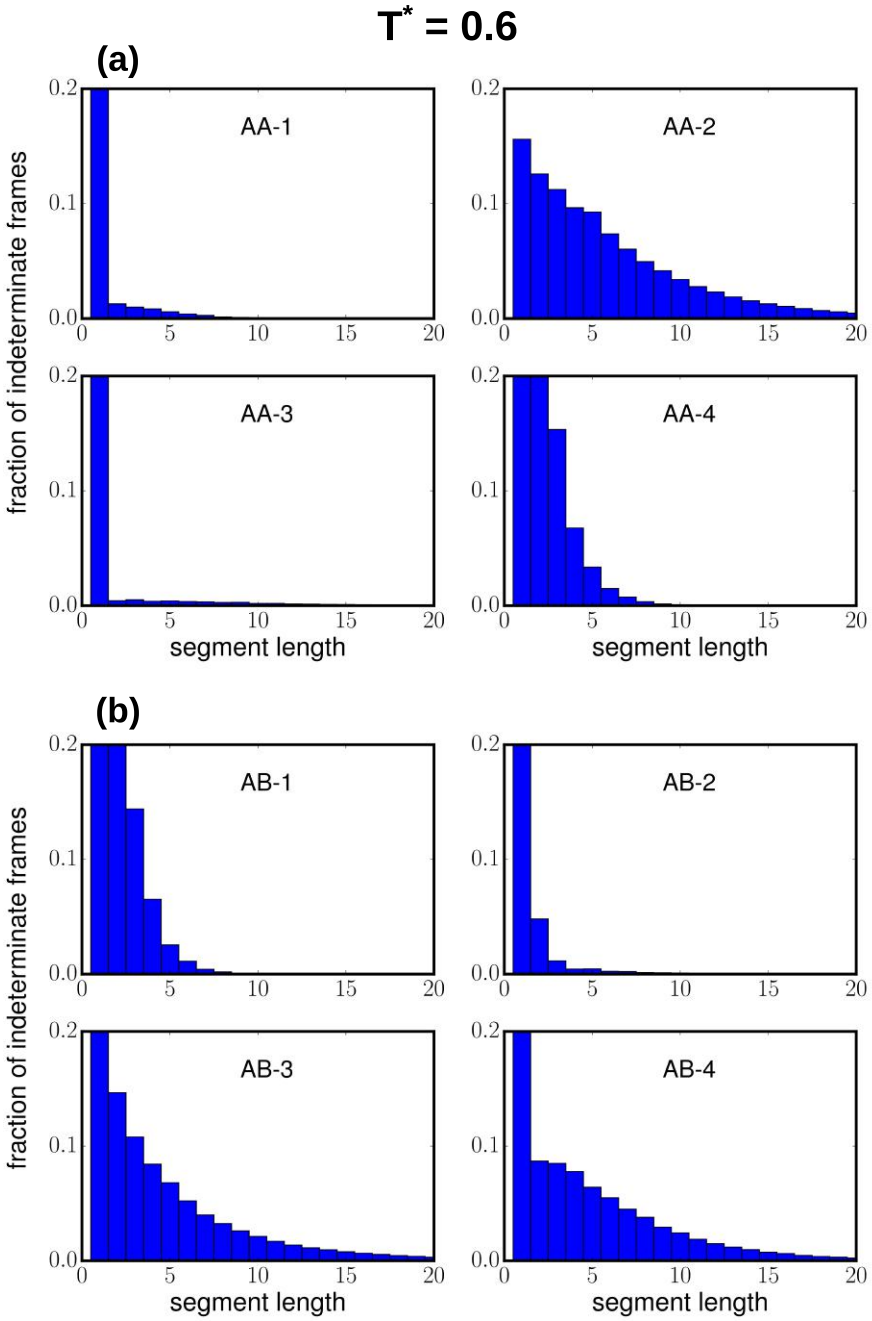

Figure S17: Cross validation of HMMs constructed from trajectories of A particles. The x-axis denotes the length of a trajectory segment (i.e., a consecutive set of frames). The y-axis quantifies the fraction of indeterminate frames with a corresponding indeterminate segment l ength.

Dimensionality reduction and clustering

Principal component analysis and k-means clustering

As described in the main text, we performed principal component analysis on the filtered trajectories to obtain a 2-D representation of configuration space for each particle type. Similar to Fig. 4 in the main text, Fig. S15 presents these free-energy landscapes, along with the 50 clusters generated using the k-means clustering algorithm. We were not particularly careful in the placement of these clusters, since our plan was to coarse-grain this representation into just a few metastable states.

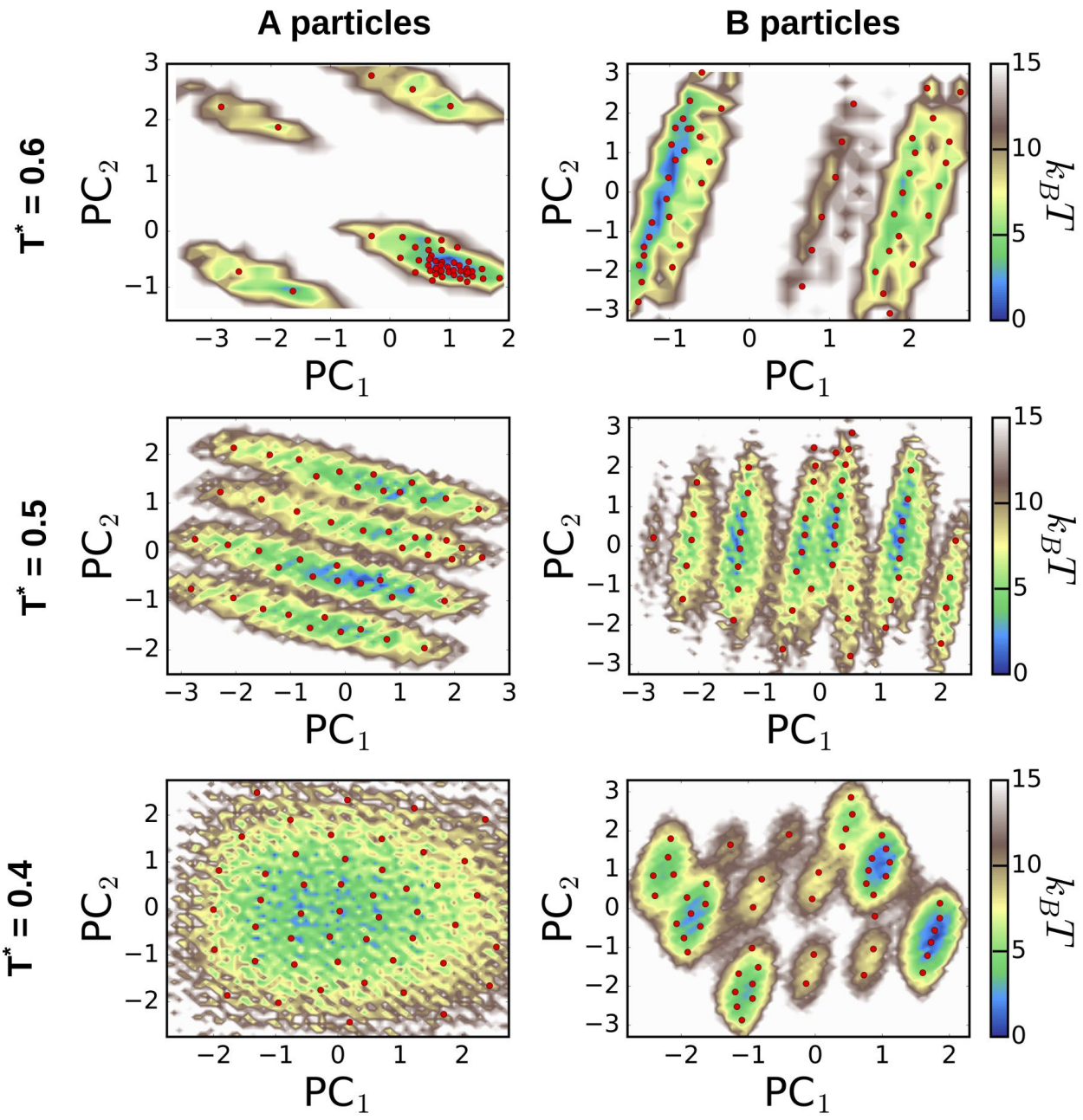

Figure S18: FESs for each temperature and particle type

Results

Characterization of solvation states

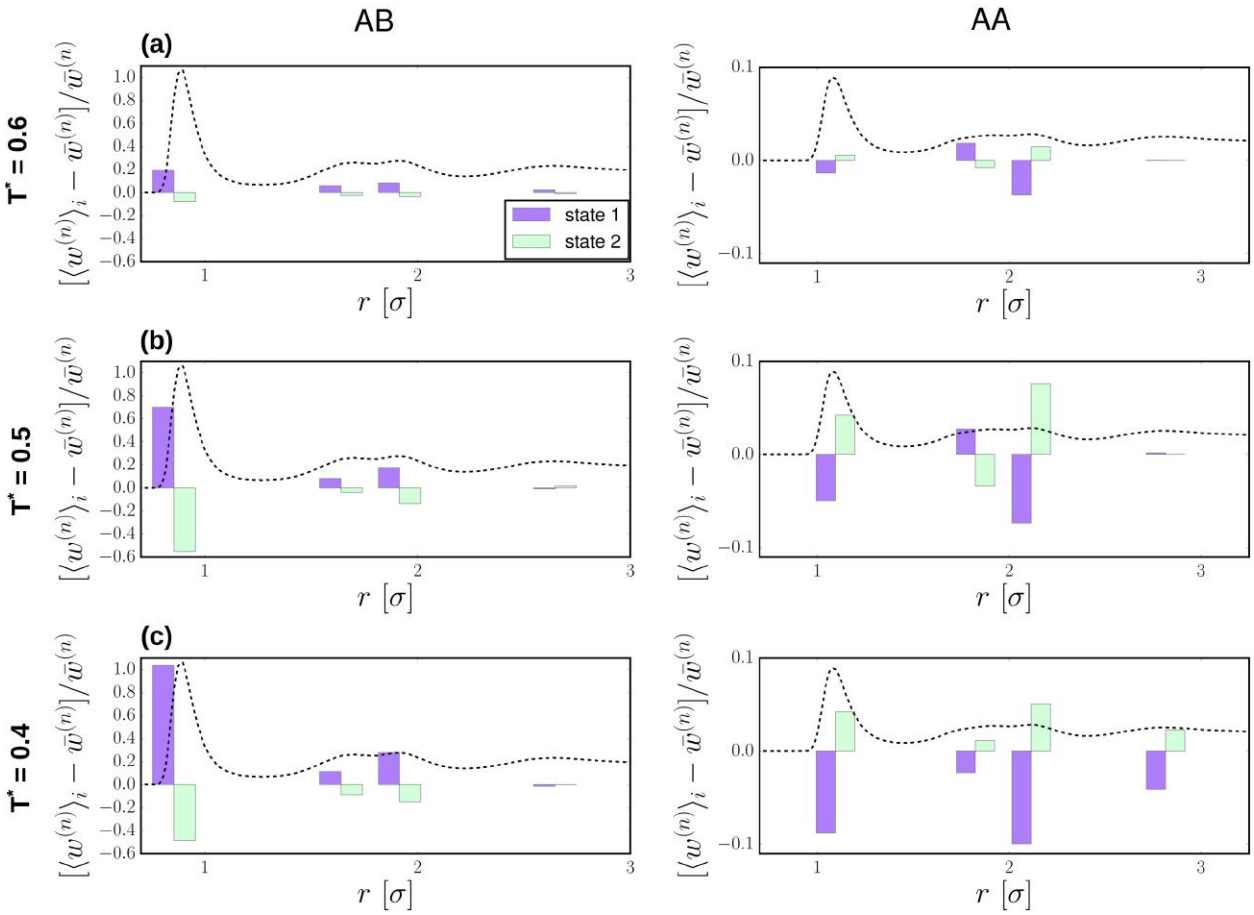

Figure S19: Characterization of a 2-state representation of solvation for A particles. The transparent bars quantify the difference in the average WCN within a state relative to the ensemble average.

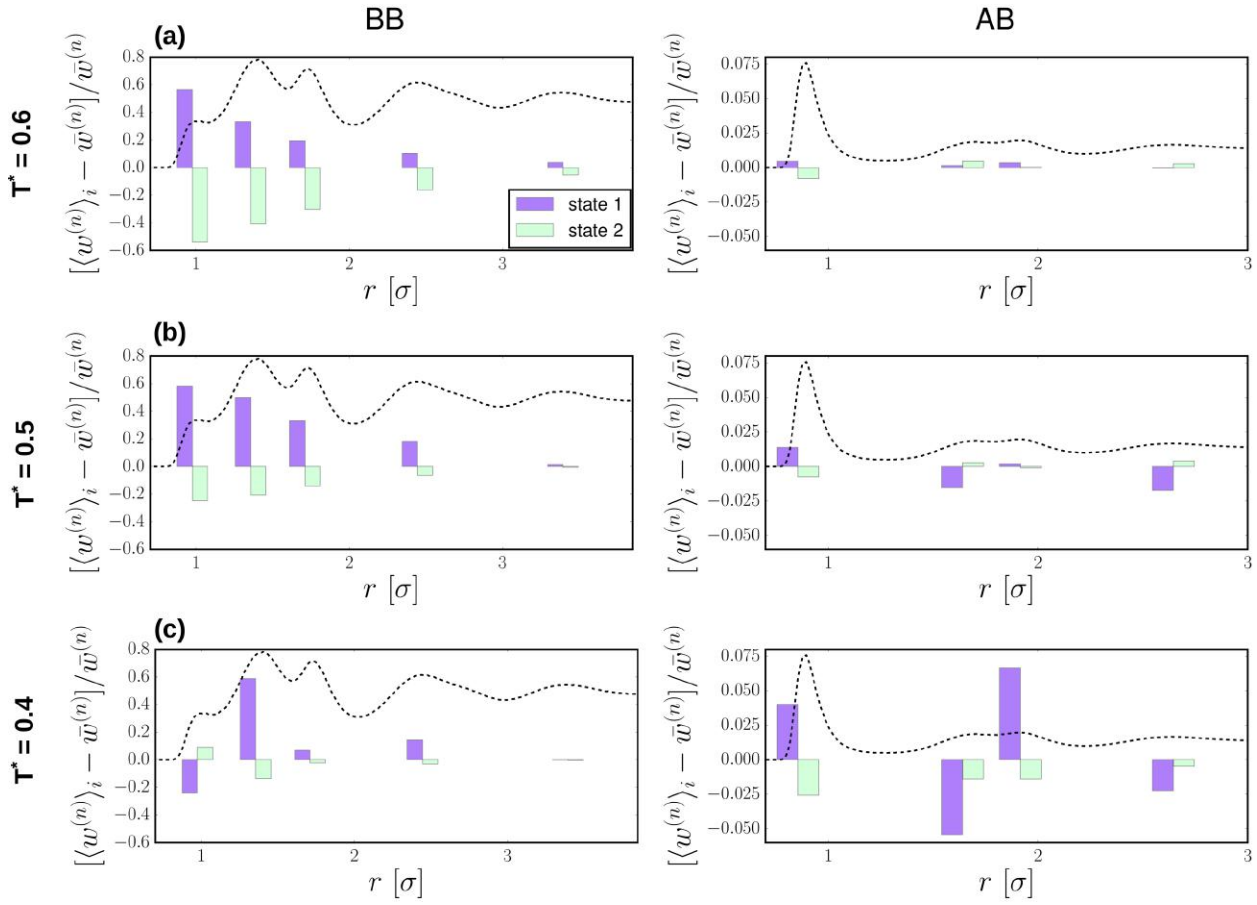

Figure S20: Characterization of a 2-state representation of solvation for B particles. The transparent bars quantify the difference in the average WCN within a state relative to the ensemble average.

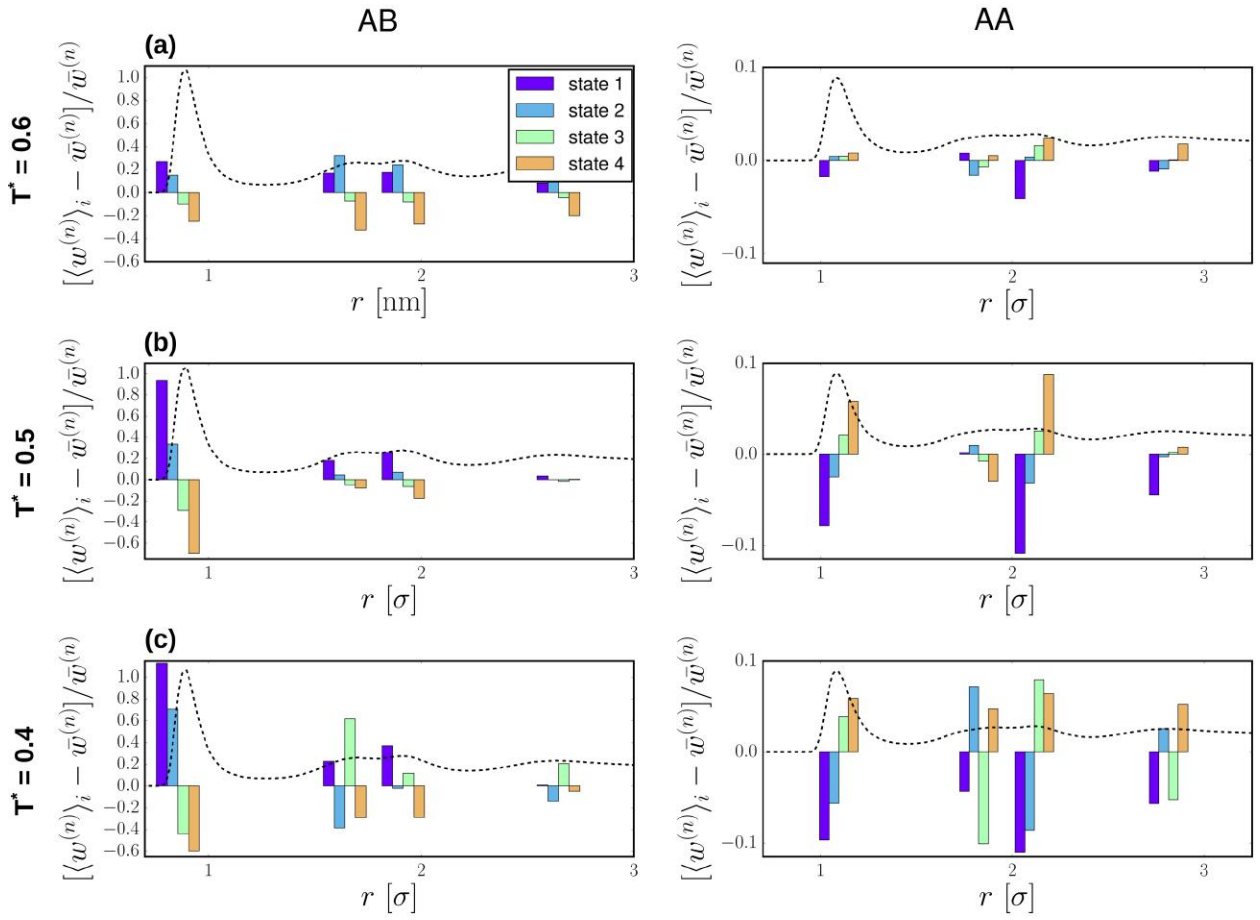

Figure S21: Characterization of a 4-state representation of solvation for A particles. The transparent bars quantify the difference in the average WCN within a state relative to the ensemble average.

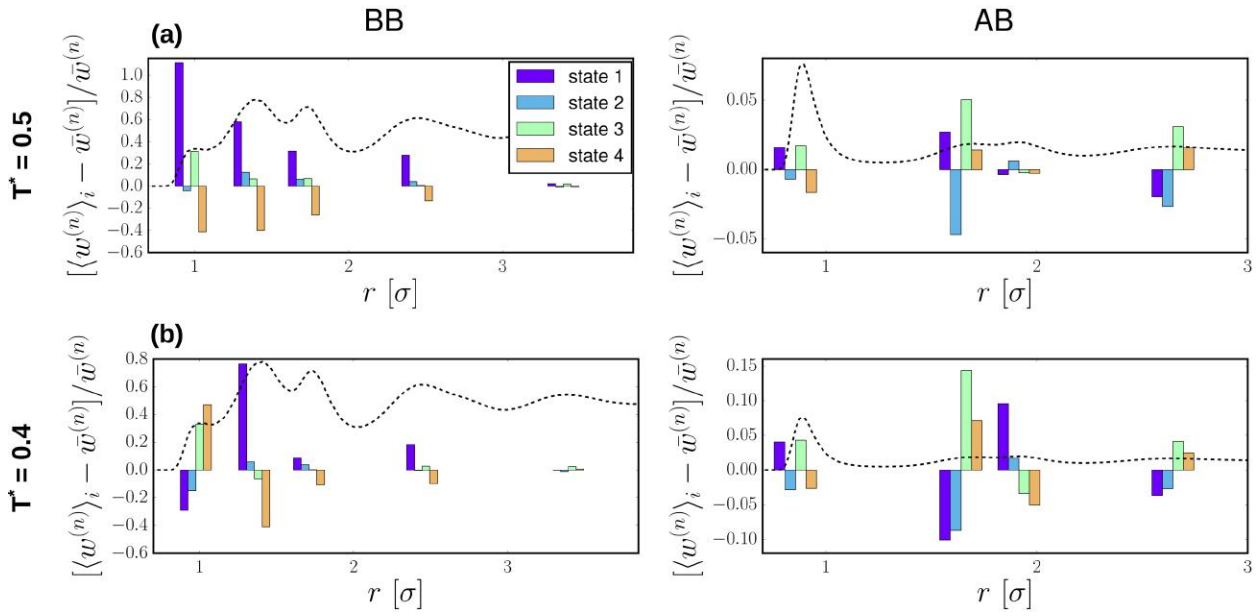

Figure S22: Characterization of a 4-state representation of solvation for B particles. The transparent bars quantify the difference in the average WCN within a state relative to the ensemble average.

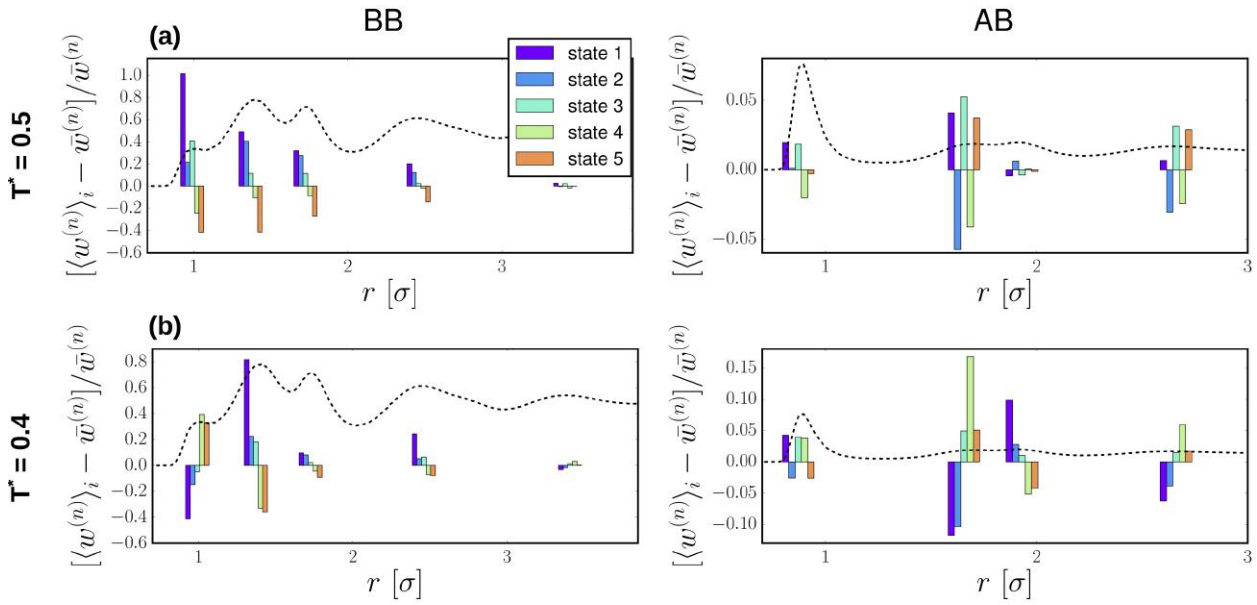

Figure S23: Characterization of a 5-state representation of solvation for B particles. The transparent bars quantify the difference in the average WCN within a state relative to the ensemble average.

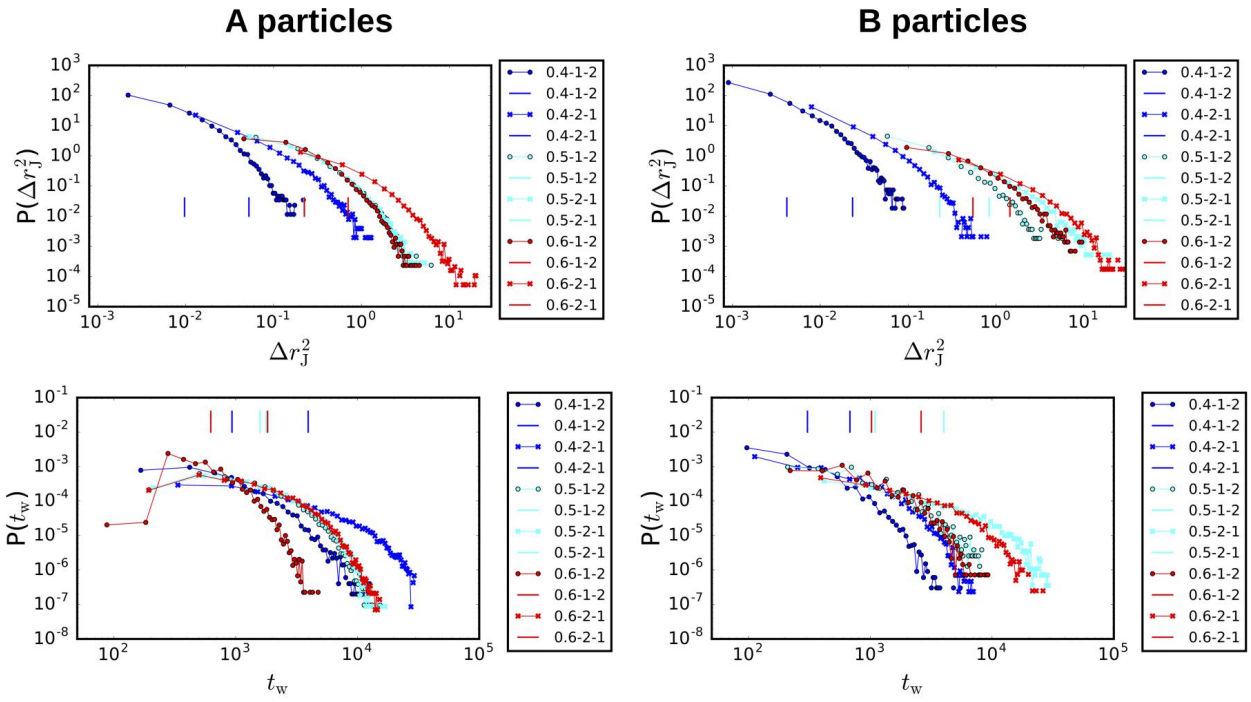

Figure S24: Jump length and waiting time distributions for the 2-state models. In the legend, the first number indicates the reduced temperature, while the second and third numbers indicate the starting and ending states, respectively, for the jump.