Supporting information for: Uncovering large-scale conformational change in molecular dynamics without prior knowledge

Ryan L. Melvin,[†] Ryan C. Godwin,[†] Jiajie Xiao,[†] William G. Thompson,[†,¶] Kenneth S. Berenhaut,[‡] and Freddie R. Salsbury Jr.[*,†]

Department of Physics, Wake Forest University, Winston-Salem NC, USA, and Department of Mathematics & Statistics, Wake Forest University, Winston-Salem NC, USA

E-mail: salsbufr@wfu.edu

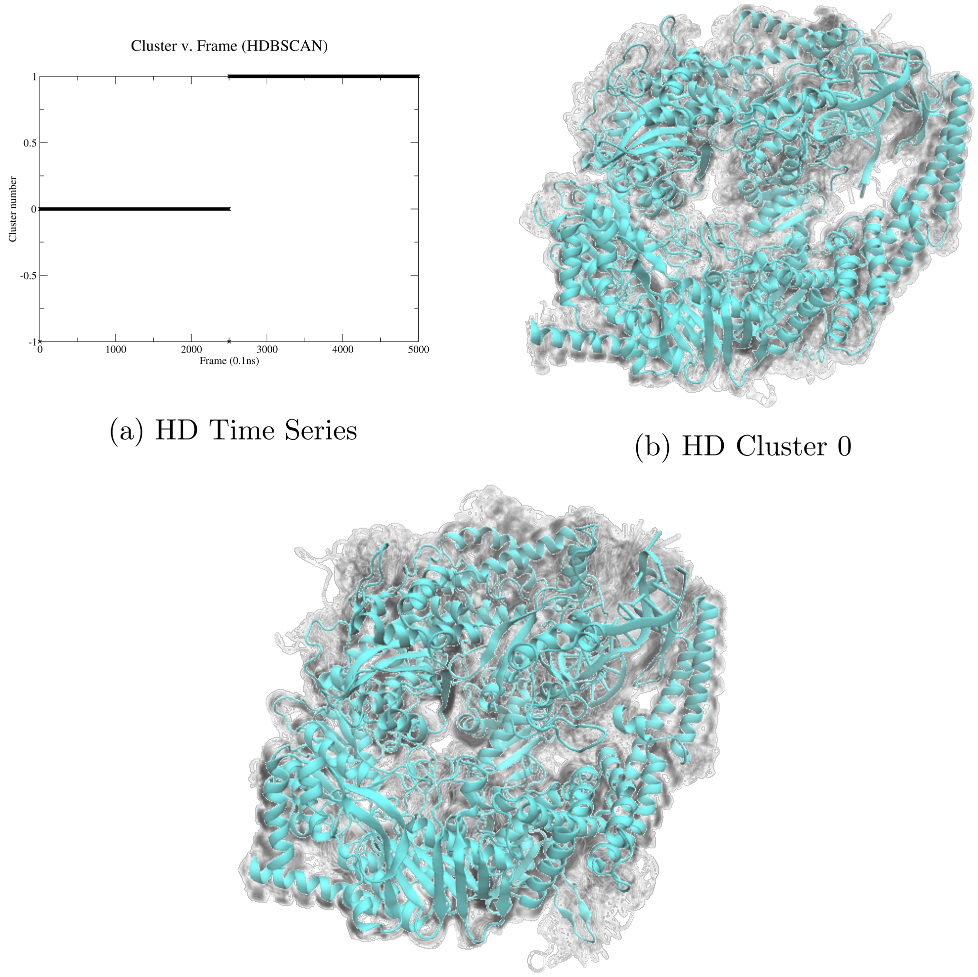
Phone: +1 (336) 758-4975. Fax: +1 (336) 758-6142

[*] To whom correspondence should be addressed

[†] Wake Forest University Department of Physics

[‡] Wake Forest University Department of Mathematics & Statistics

[¶] Current Address: Yale University Department of Physics, Yale University, New Haven, CT, USA

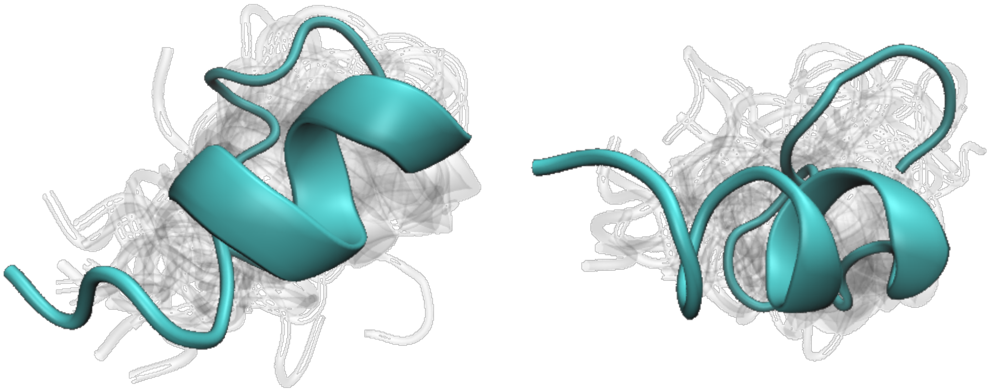


(a) HD Time Series

(b) HD Cluster 0

(c) HD Cluster 1

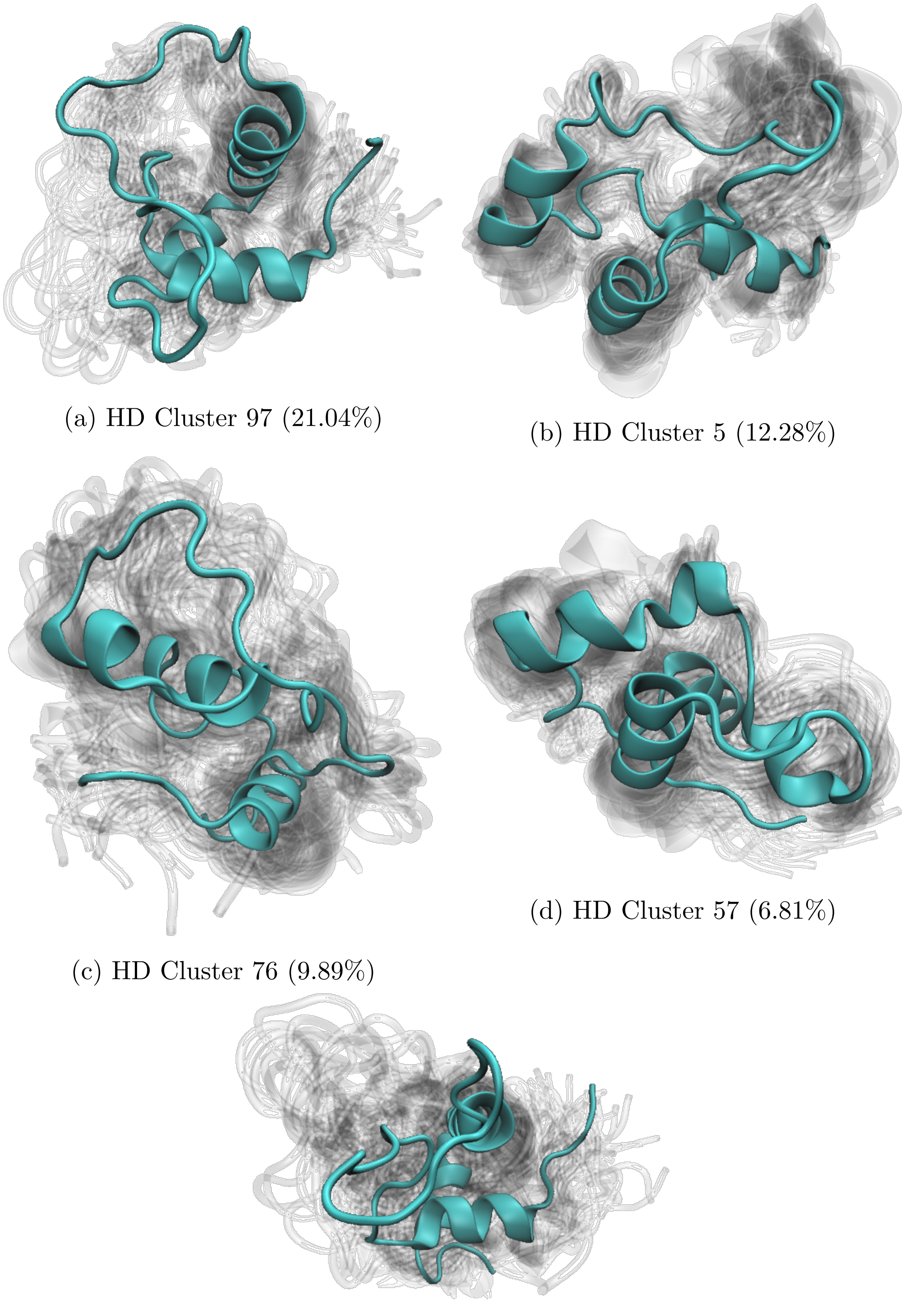
Figure S1: Clustering with HDBSCAN on alpha carbon atoms of MutSα complexed with cisplatinated DNA (a) yielded one cluster per trajectory, with the initial structure of each simulation labeled as noise. For comparison with Amorim-Hennig clustering of this system, presented in the paper proper, Figure 1, we (b-c) we visualize both of these clusters. Shadows are 50 evenly sampled frames from each cluster.



(a) A-H Cluster 0

(b) A-H Cluster 1

Figure S2: Visualizing the Amorim-Hennig clusters (a-b) of non-zinc-bound NEMO reveals a high level of variance within each cluster. Here we see the poor performance of Amorim-Hennig on unstable systems.



(a) HD Cluster 97 (21.04%)

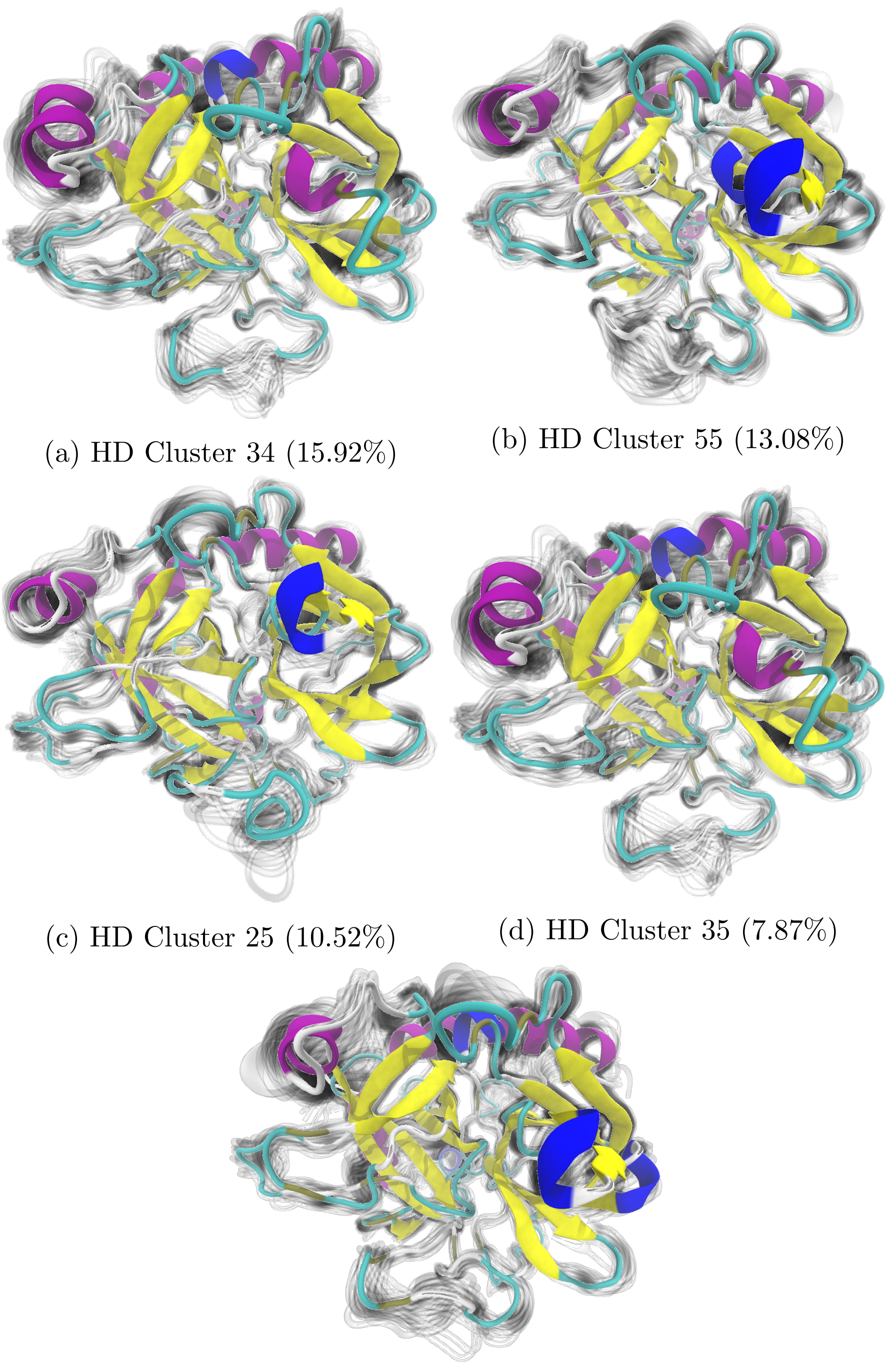
(b) HD Cluster 5 (12.28%)

(c) HD Cluster 76 (9.89%)

(d) HD Cluster 57 (6.81%)

(e) HD Cluster 107 (5.71%)

Figure S3: Visualizing the top 6 HDBSCAN clusters of Villin headpiece by population revealed that HDBSCAN had found stable folding intermediates.



(a) HD Cluster 34 (15.92%)

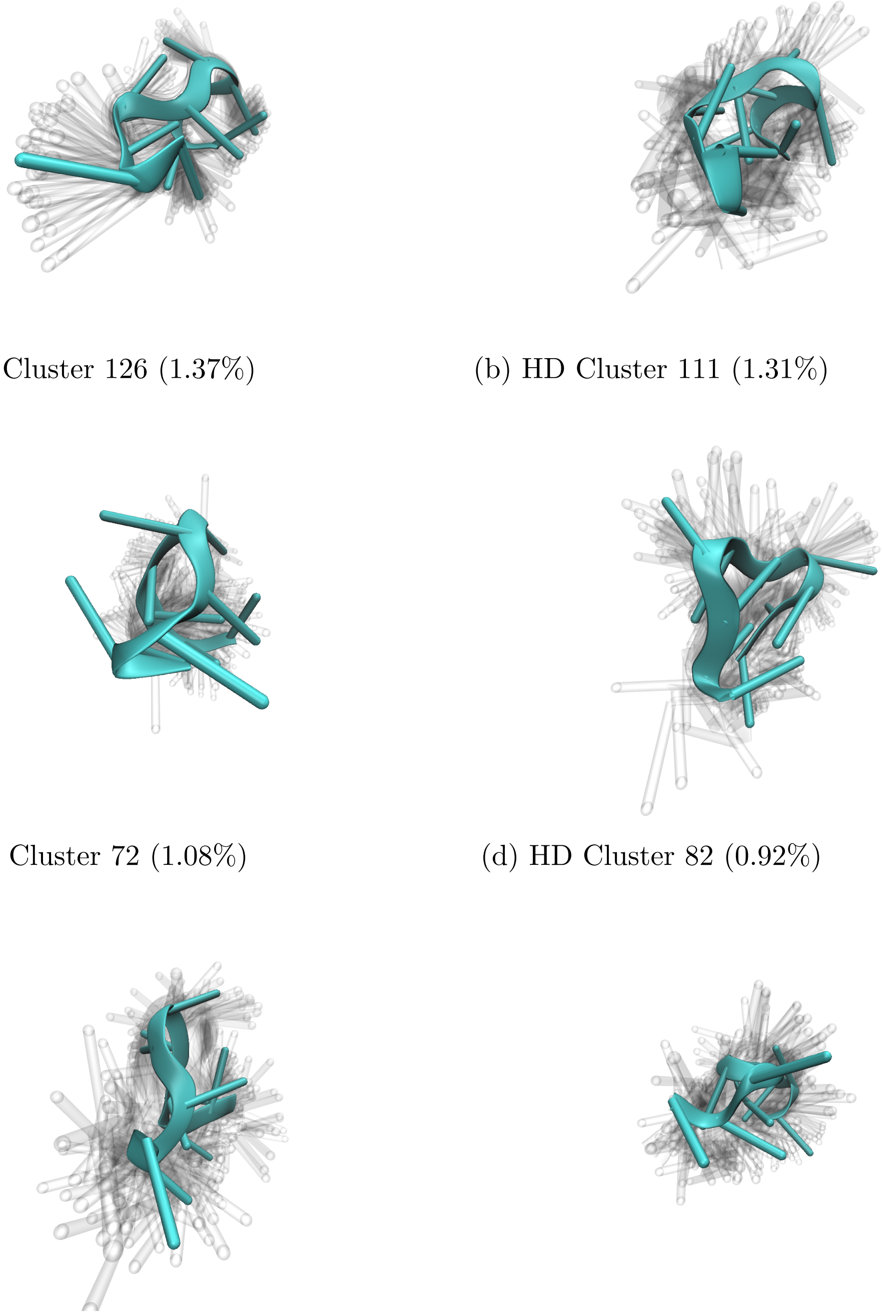
(b) HD Cluster 55 (13.08%)

(c) HD Cluster 25 (10.52%)

(d) HD Cluster 35 (7.87%)

(e) HD Cluster 57 (7.78%)

Figure S4: Visualizations of representative structures of the most populated clusters of Thrombin from HDBSCAN show that the structural differences between the dominant clusters mainly occur at the flexible gamma loop and the light chain termini, likely due to the high variation in position from their mobility.



(a) HD Cluster 126 (1.37%)

(b) HD Cluster 111 (1.31%)

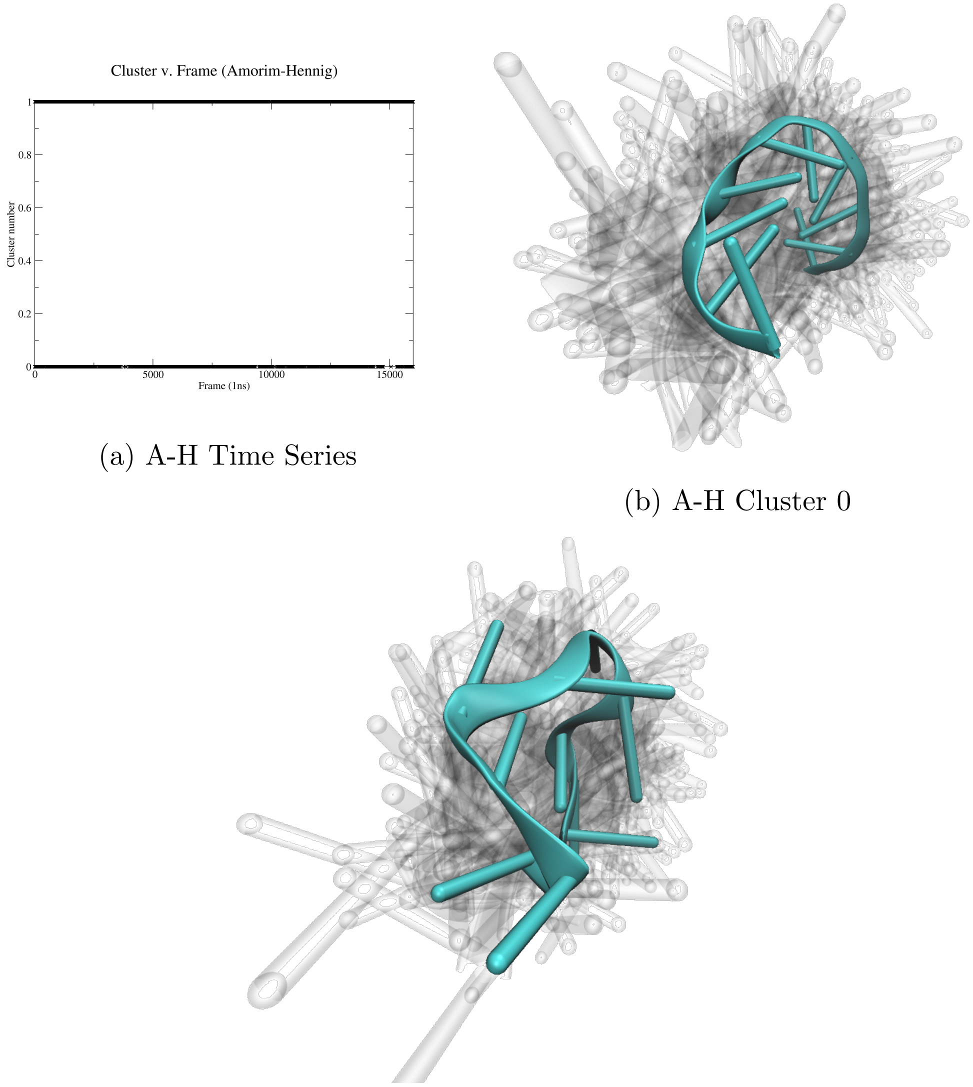
(c) HD Cluster 72 (1.08%)

(d) HD Cluster 82 (0.92%)

(e) HD Cluster 70 (0.89%)

(f) HD Cluster 73 (0.84%)

Figure S5: On a trajectory concatenated from four simulations of F10 in the presence of 150mMNaCl, HDBSCAN yields mostly noise (-1) and 126 clusters, the largest of which comprises 1.37% of the trajectory frames, indicating an unstable system. Here we (a-f) visualize the top 6 clusters by population with shadows as 50 evenly sampled frames from the given cluster.

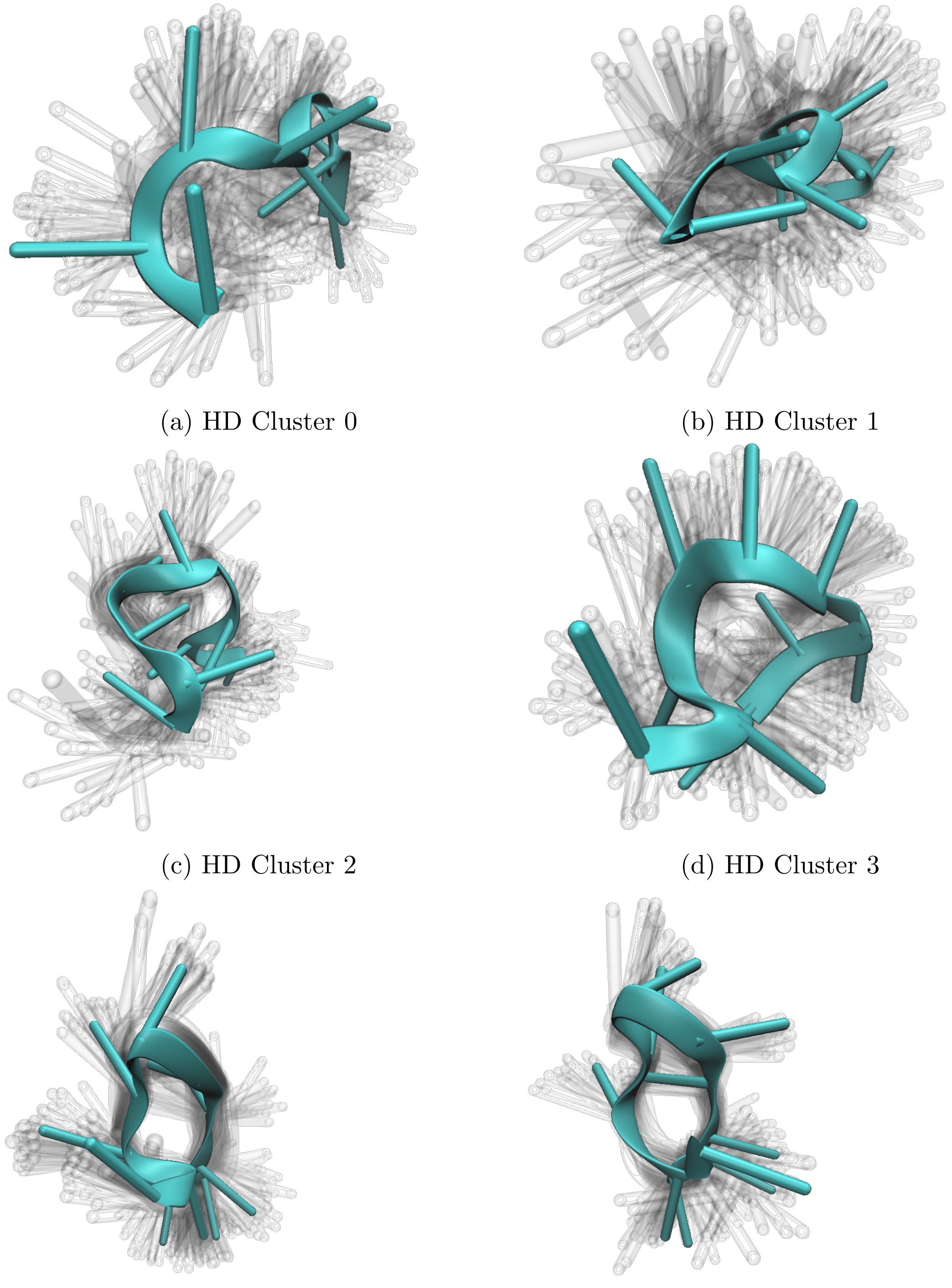


(a) A-H Time Series

(b) A-H Cluster 0

(c) A-H Cluster 1

Figure S6: On a trajectory concatenated from four simulations of F10 in the presence of 150mMNaCl, (a) Amorim-Hennig places all structures into two clusters within a high level of variance (b-c) within each cluster.Shadows are 50 evenly sampled frames from the given cluster.



(a) HD Cluster 0

(b) HD Cluster 1

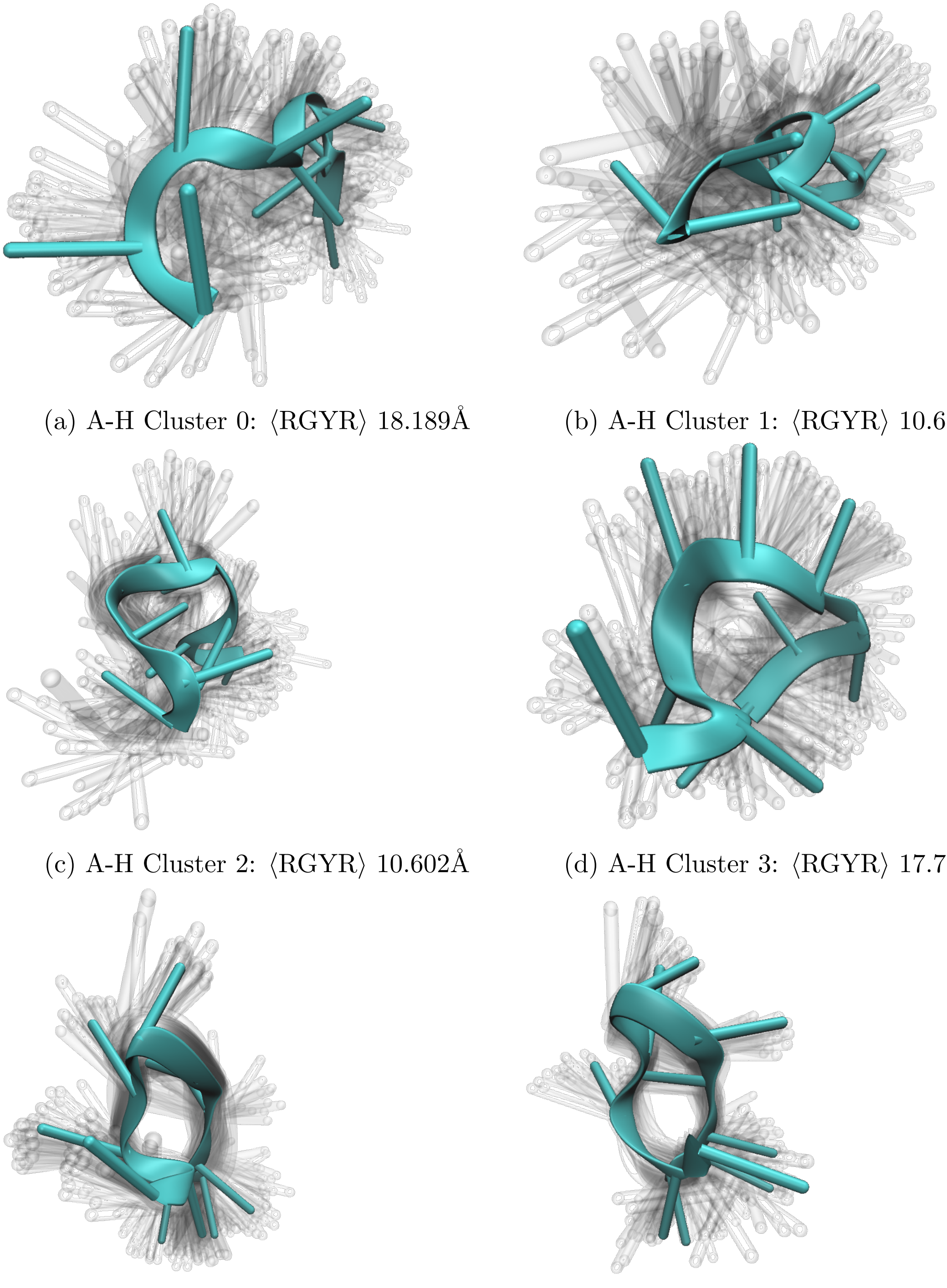
(c) HD Cluster 2

(d) HD Cluster 3

(e) HD Cluster 4

(f) HD Cluster 5

Figure S7: On a trajectory concatenated from four simulations, HDBSCAN primarily split up the individual simulations, indicating that in each of the concatenated simulation F10 finds a different stable conformations. (a-f) Visualizing these conformations with shadows as 50 evenly sampled frames reveals little variance in the nucleic acid backbone position within each cluster.



(a) A-H Cluster 0: ⟨RGYR⟩ 18.189Å

(b) A-H Cluster 1: ⟨RGYR⟩ 10.606Å

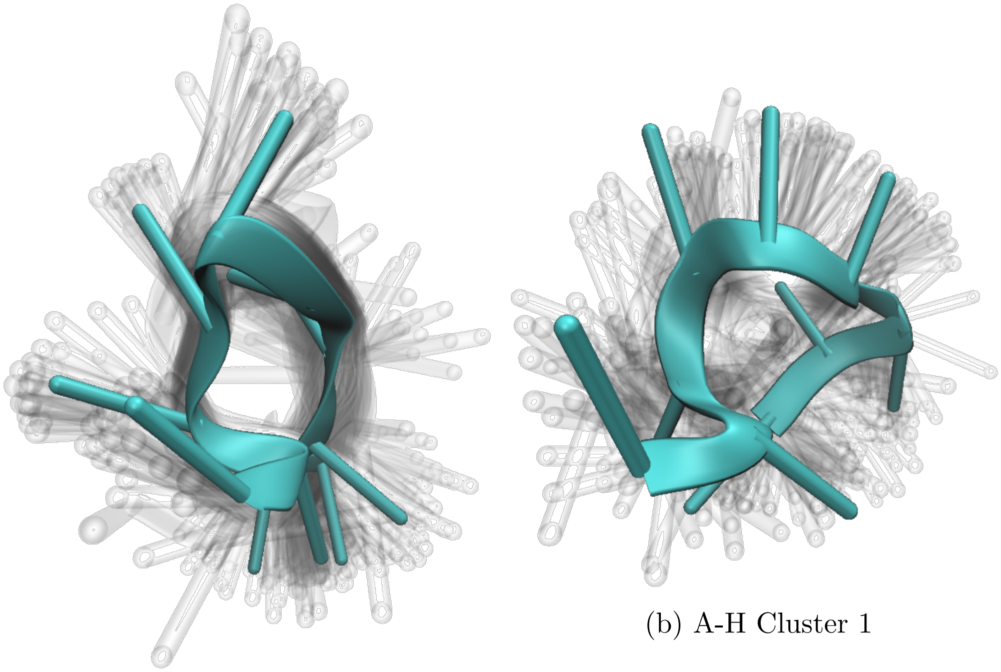
(c) A-H Cluster 2: ⟨RGYR⟩ 10.602Å

(d) A-H Cluster 3: ⟨RGYR⟩ 17.774Å

(e) A-H Cluster 4: ⟨RGYR⟩ 15.181Å

(f) A-H Cluster 5: ⟨RGYR⟩ 13.281Å

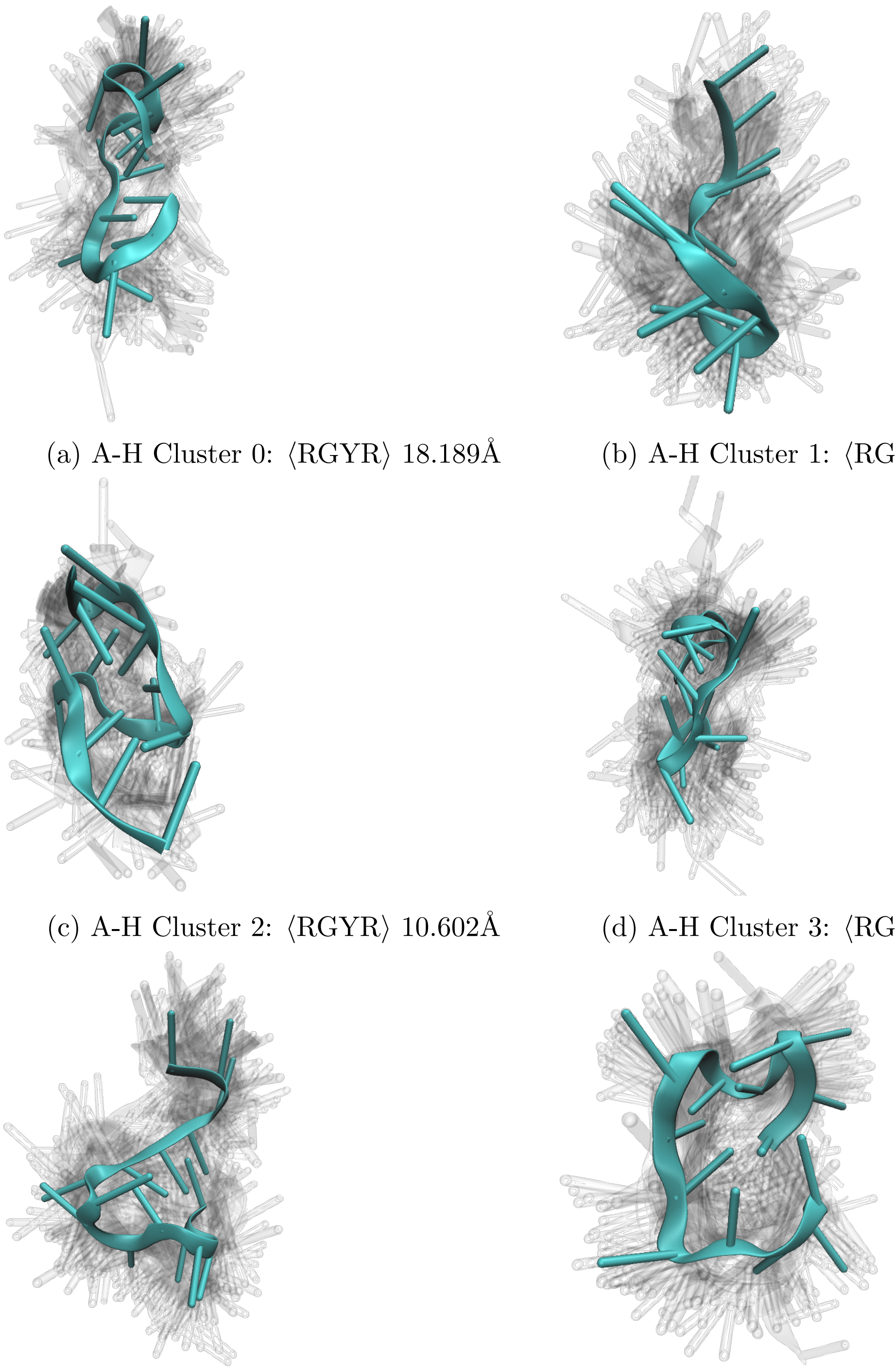
Figure S8: On a trajectory concatenated from four simulations, HDBSCAN primarily split up the individual simulations, indicating that in each of the concatenated simulation F10 finds a different stable conformations. (a-f) Visualizing these conformations with shadows as 50 evenly sampled frames reveals little variance in the nucleic acid backbone position within each cluster.



(b) A-H Cluster 1

(a) A-H Cluster 0

Figure S9: Amorim-Hennig split two the four concatenated trajectories of F10 in 150mM CaCl$_2$ crisply into one cluster each. These conformation ensembles exhibit little uncertainty (shadow width), indicating the structures are highly stable.



(a) A-H Cluster 0: ⟨RGYR⟩ 18.189Å

(b) A-H Cluster 1: ⟨RGYR⟩ 10.606Å

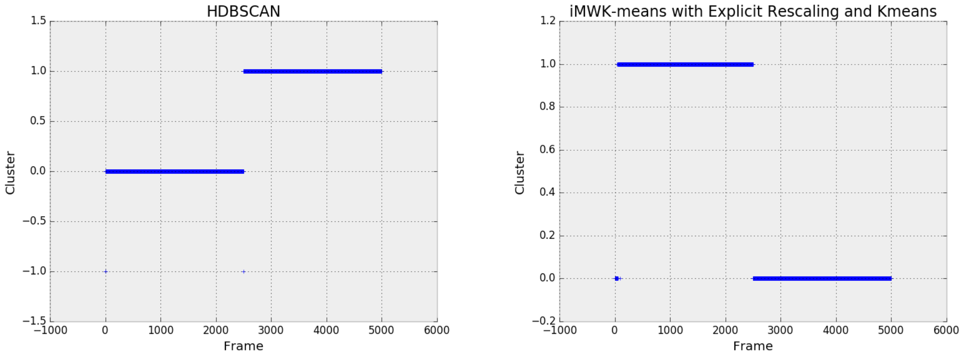
(c) A-H Cluster 2: ⟨RGYR⟩ 10.602Å

(d) A-H Cluster 3: ⟨RGYR⟩ 17.774Å

(e) A-H Cluster 4: ⟨RGYR⟩ 15.181Å

(f) A-H Cluster 5: ⟨RGYR⟩ 13.281Å

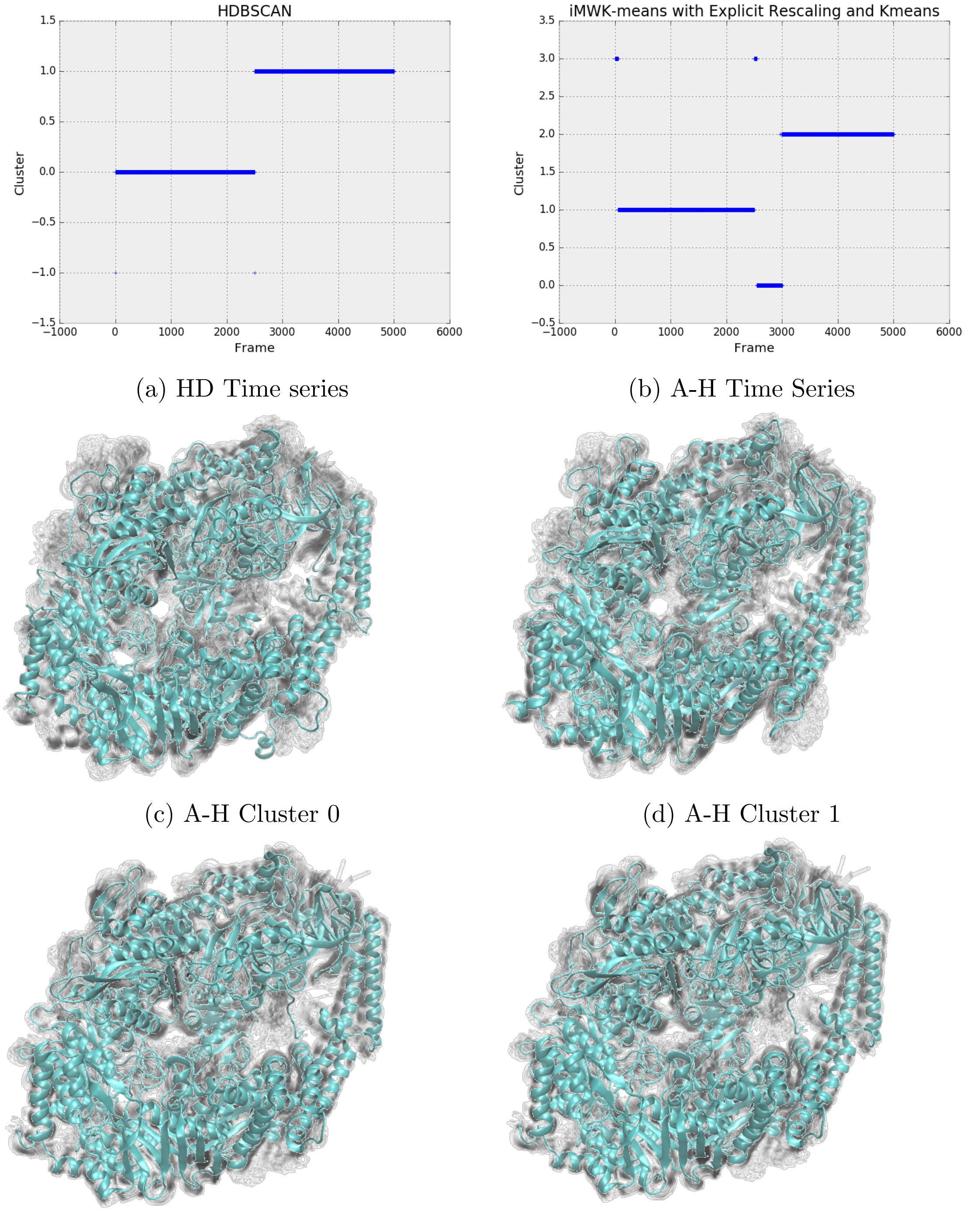
Figure S10: Amorim-Hennig divided this trajectory of unbound 15-TBA into what appear upon visualization to be compactness-based bins. Calculating the average RGYR of each of these clusters bolsters this assumption.



(a) HD Time series

(b) A-H Time Series

Figure S11: Both HDBSCAN and Amorim-Hennig split two concatenated simulations of MutSα in the presence of flouridated DNA into the individual simulations.



(a) HD Time series

(b) A-H Time Series

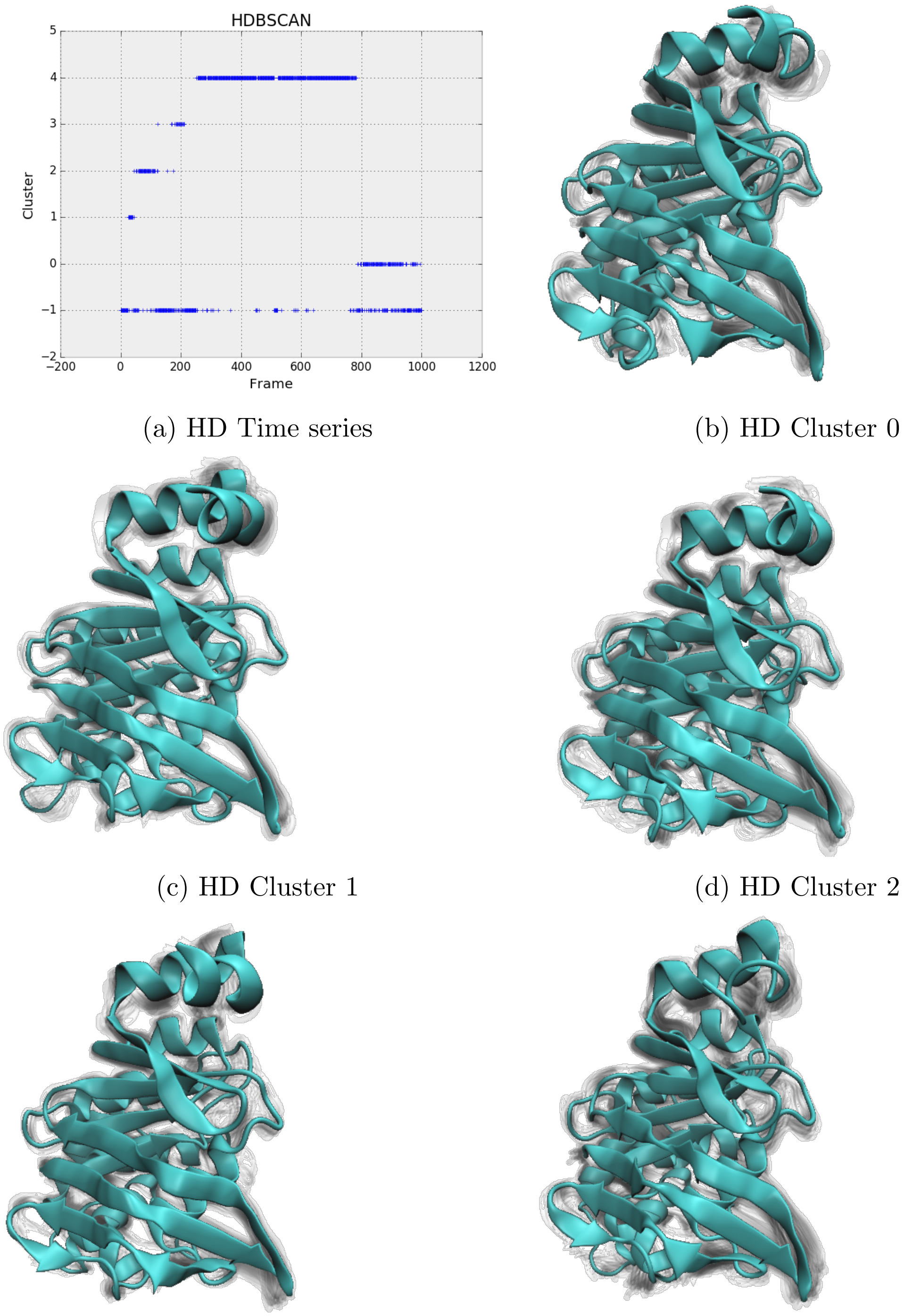
(c) A-H Cluster 0

(d) A-H Cluster 1

(e) A-H Cluster 2

(f) A-H Cluster 3

Figure S12: (a) HDBSCAN splits a trajectory of MUTSα in the presence of mismatched DNA into the two concatenated simulations comprising it. (b) Amorim-Hennig, though, (c-f) finds 4 states.



(a) HD Time series

(b) HD Cluster 0

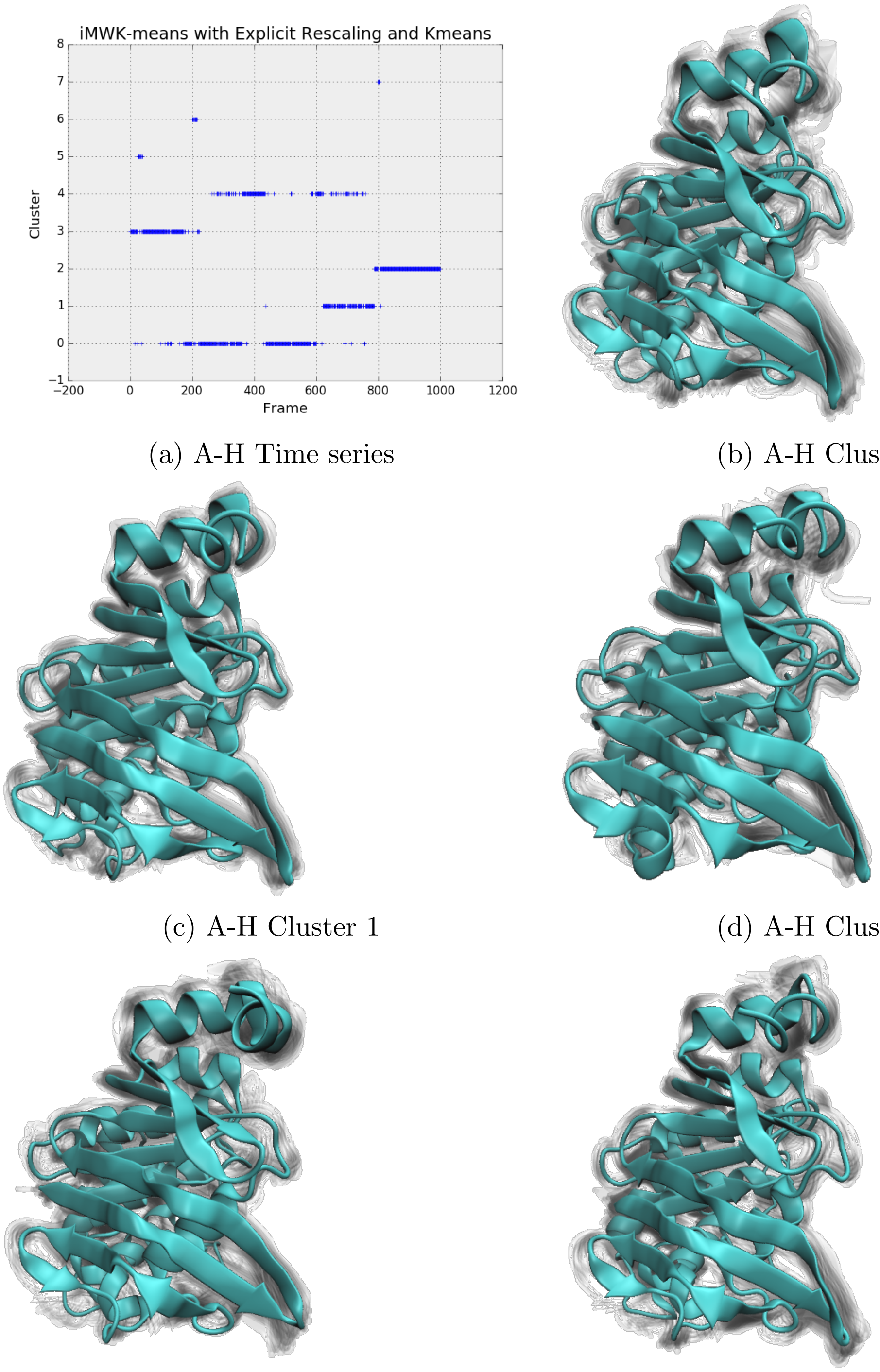
(c) HD Cluster 1

(d) HD Cluster 2

(e) HD Cluster 3

(f) HD Cluster 4

Figure S13: HDBSCAN finds 5 clusters in this 1 microsecond simulation of homology-modeled SufC from *Bacillus subtilis*.



(a) A-H Time series

(b) A-H Cluster 0

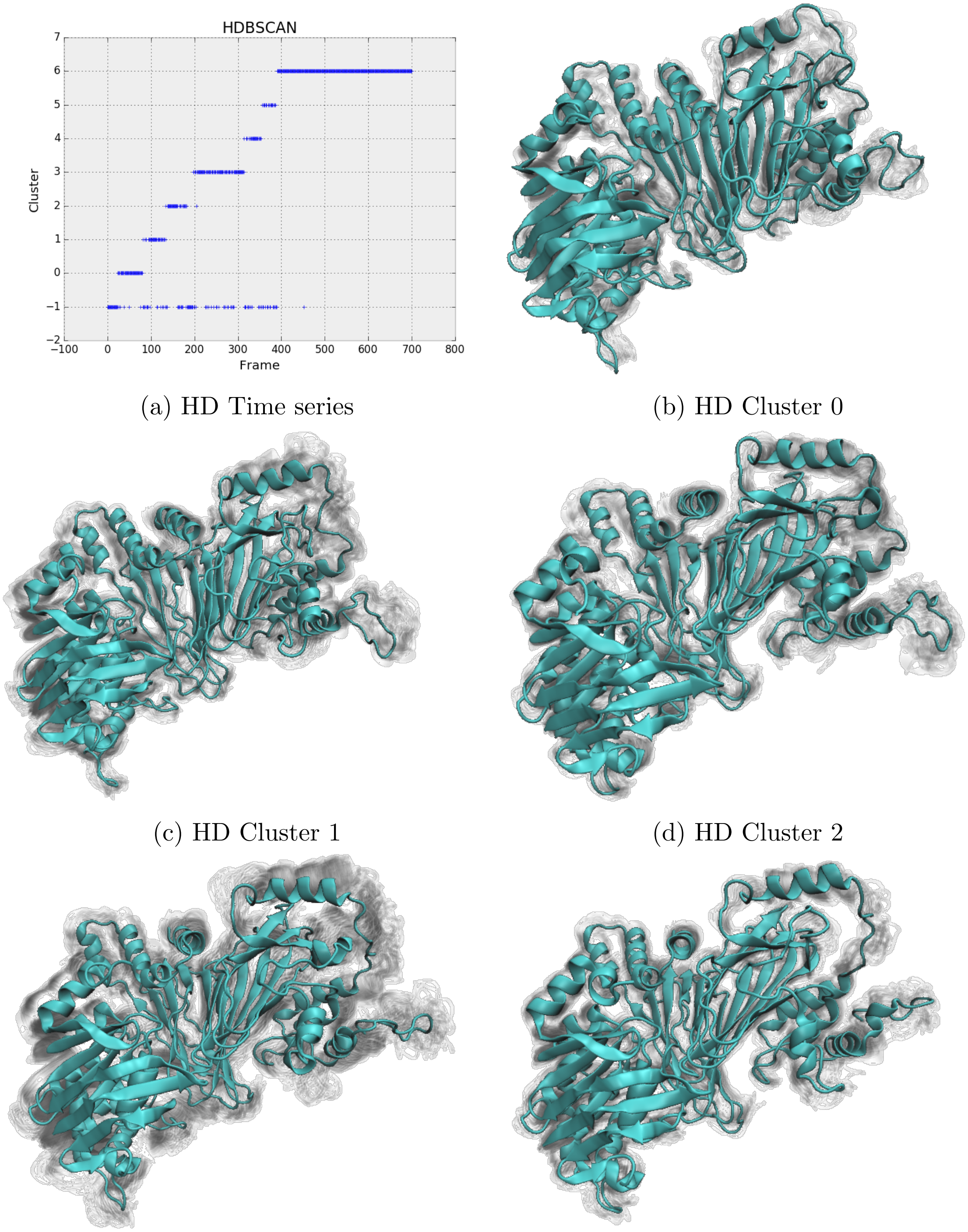
(c) A-H Cluster 1

(d) A-H Cluster 2

(e) A-H Cluster 3

(f) A-H Cluster 4

Figure S14: Amorim-Hennig finds 8 clusters in this 1 microsecond simulation of homology-modeled SufC from *Bacillus subtilis*.



(a) HD Time series

(b) HD Cluster 0

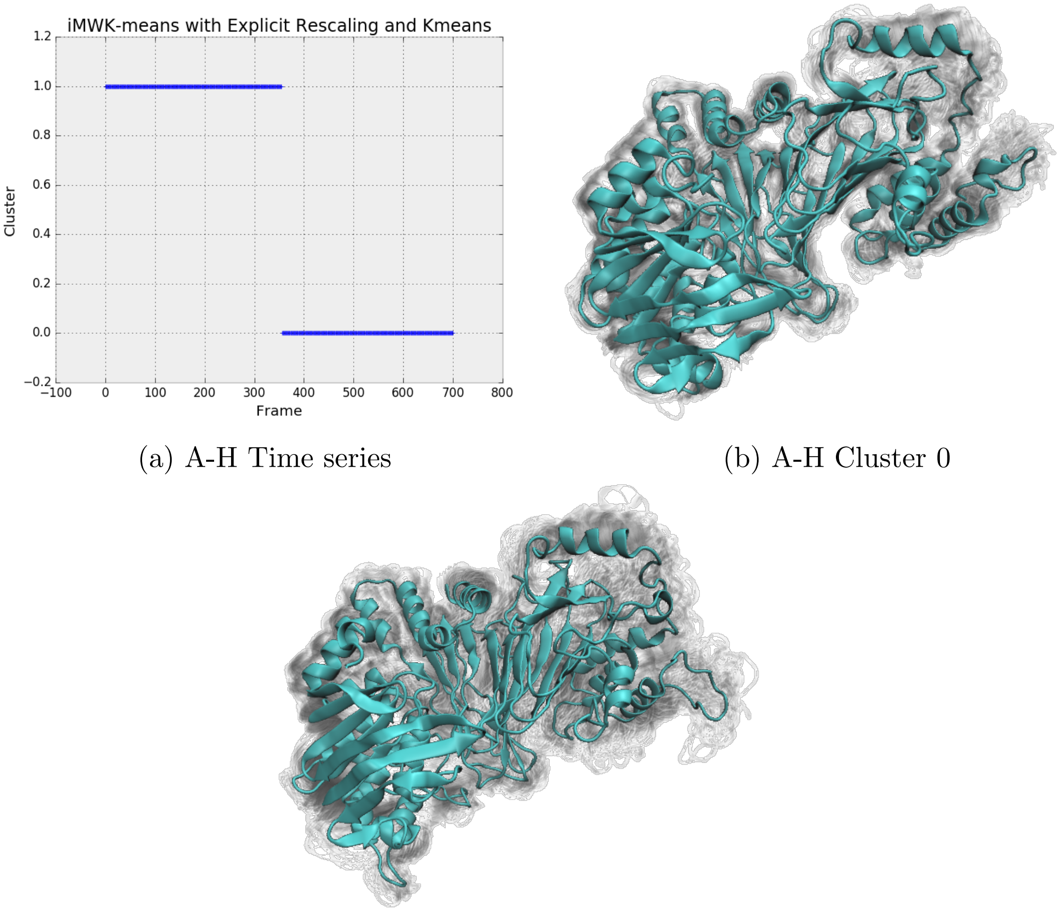
(c) HD Cluster 1

(d) HD Cluster 2

(e) HD Cluster 3

(f) HD Cluster 4

Figure S15: HDBSCAN finds 7 clusters in this 1 microsecond simulation of docked homology-models of SufC and SufD from *Bacillus subtilis*.

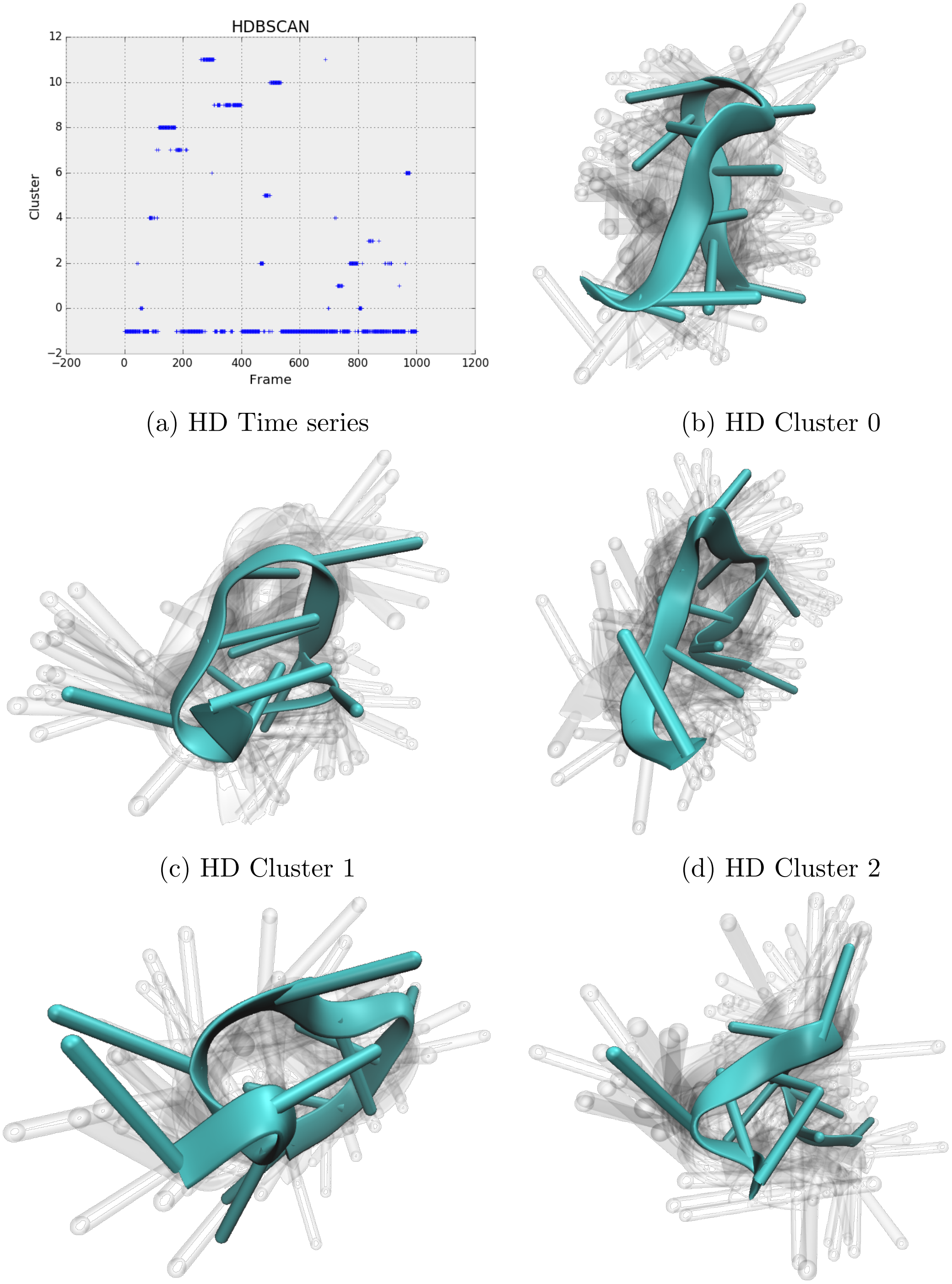


(a) A-H Time series

(b) A-H Cluster 0

(c) A-H Cluster 1

Figure S16: Amorim-Hennig finds 2 clusters in this 1 microsecond simulation of docked homology-models SufC and SufD from *Bacillus subtilis*.



(a) HD Time series

(b) HD Cluster 0

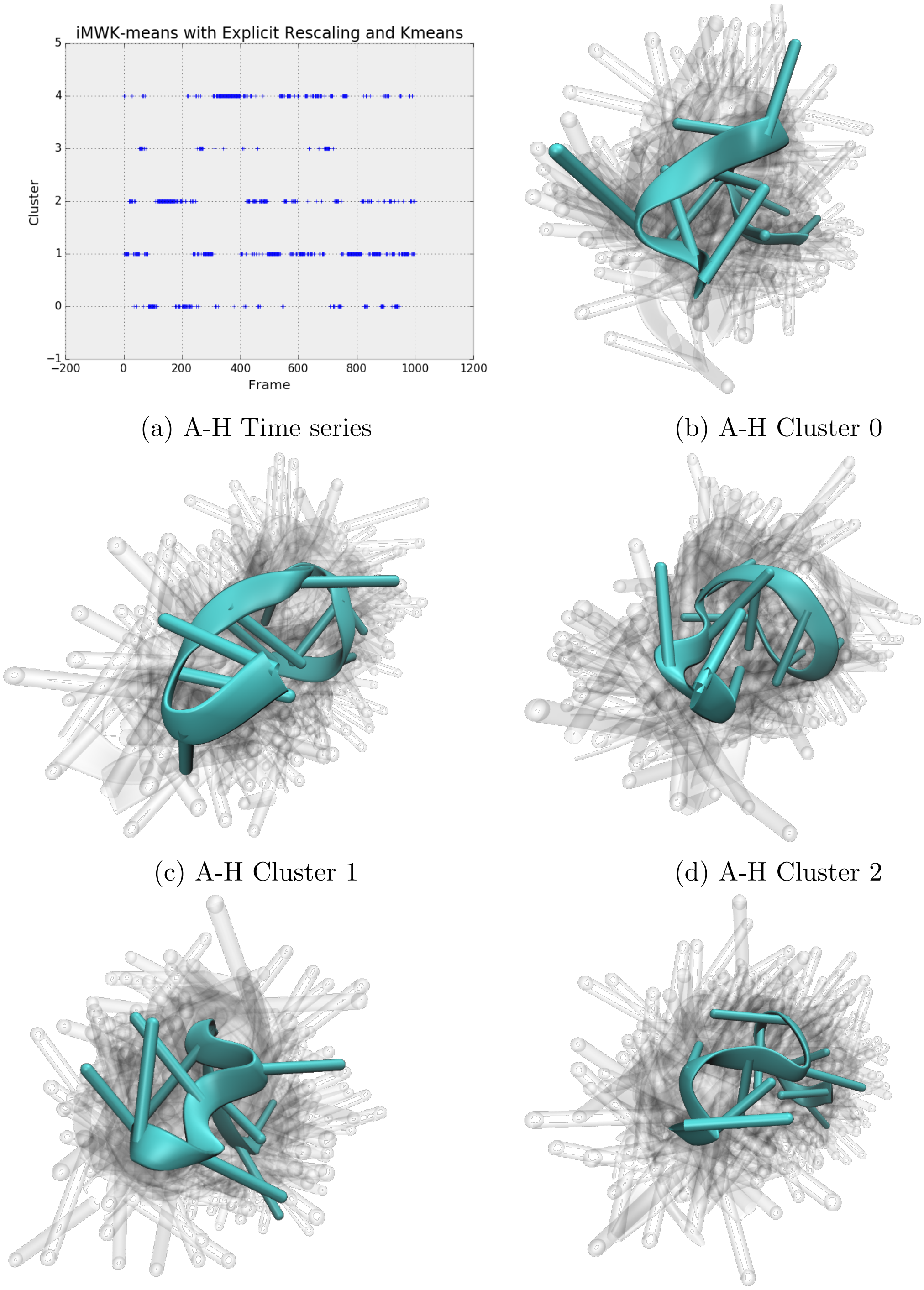
(c) HD Cluster 1

(d) HD Cluster 2

(e) HD Cluster 3

(f) HD Cluster 4

Figure S17: HDBSCAN finds 12 clusters in this 1 microsecond simulation of F10 in 150mM $MgCl_2$.



(a) A-H Time series

(b) A-H Cluster 0

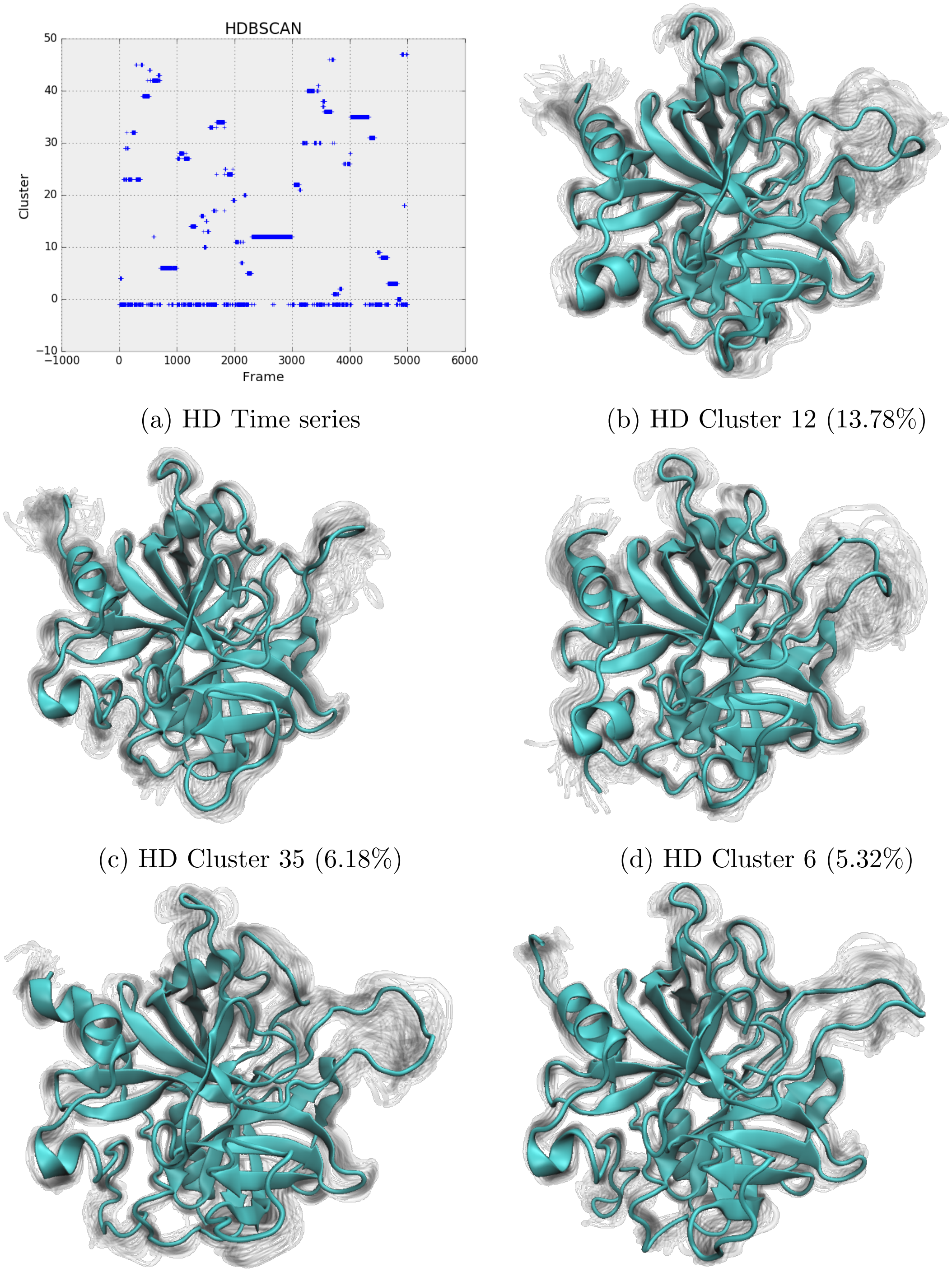
(c) A-H Cluster 1

(d) A-H Cluster 2

(e) A-H Cluster 3

(f) A-H Cluster 4

Figure S18: Amorim-Hennig finds 5 clusters in this 1 microsecond simulation of F10 in 150mM $MgCl_2$.



(a) HD Time series

(b) HD Cluster 12 (13.78%)

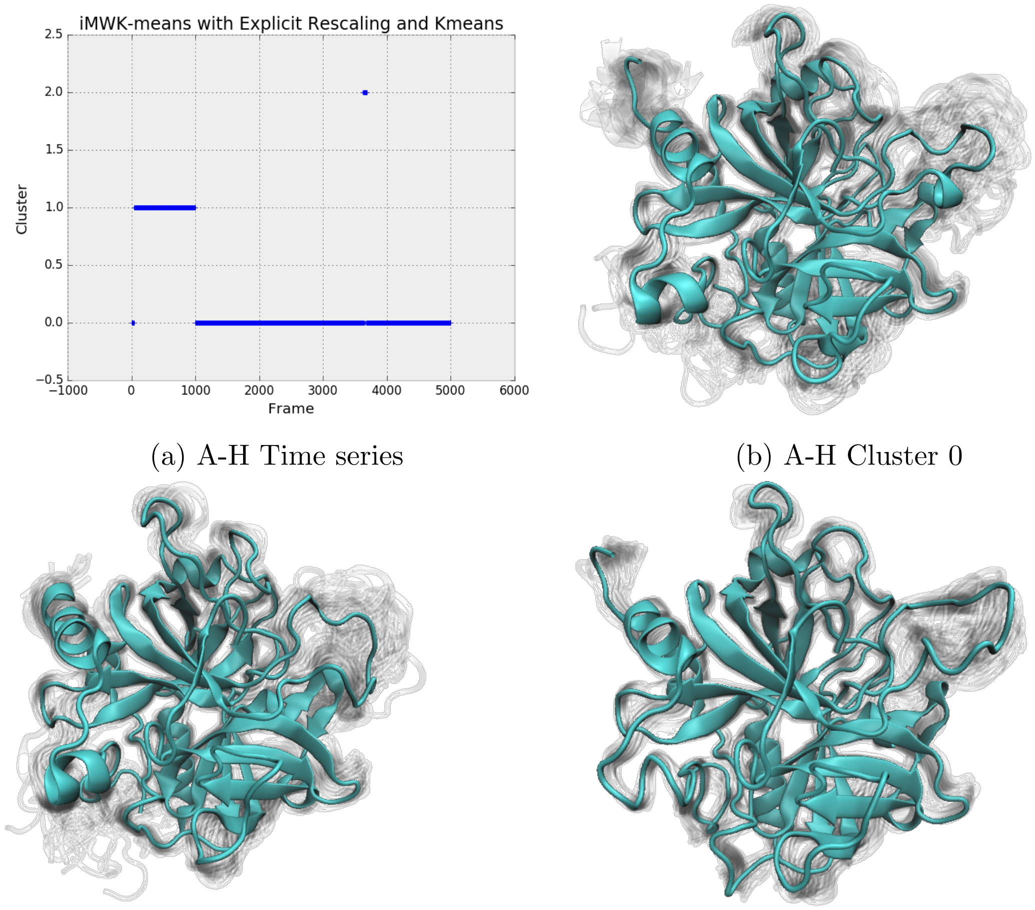
(c) HD Cluster 35 (6.18%)

(d) HD Cluster 6 (5.32%)

(e) HD Cluster 3 (3.02%)

(f) HD Cluster 40 (2.58%)

Figure S19: HDBSCAN finds 48 clusters in this 5 microsecond trajectory of unbound thrombin in 150mM KCl.



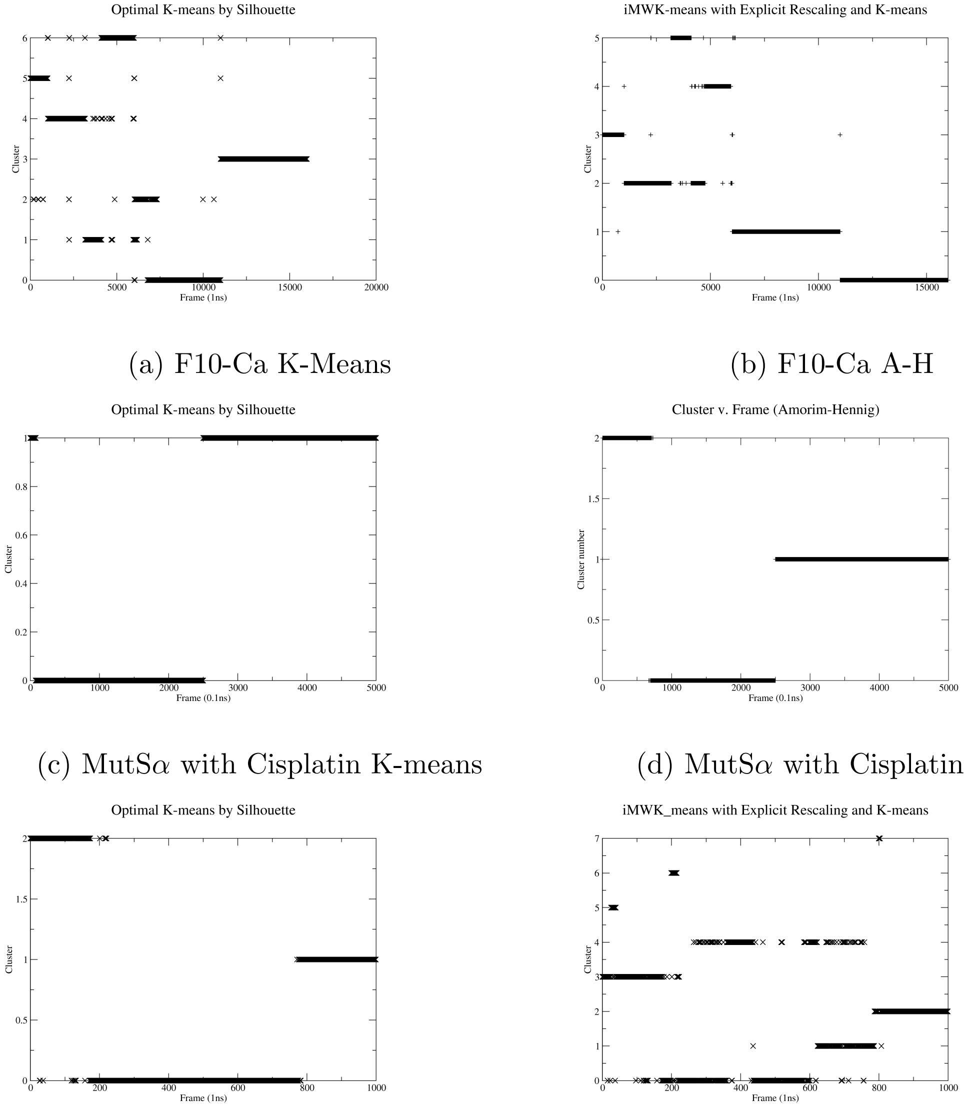
(a) A-H Time series

(b) A-H Cluster 0

(c) A-H Cluster 1

(d) A-H Cluster 2

Figure S20: Amorim-Hennig finds 4 clusters in this 5 microsecond simulation of unbound thrombin in 150mM KCl.



(a) F10-Ca K-Means

(b) F10-Ca A-H

(c) MutSα with Cisplatin K-means

(d) MutSα with Cisplatin A-H

(e) SufC K-Means

(f) SufC A-H

Figure S21: We compare the time series of K-Means clusterings that maximize silhouette values. Here we perform K-Means clusterings with with values of k between 2 and 20 (inclusive). We report the time series for the value of k with the highest silhouette score. Aside each of these, we plot the A-H time series for comparison. We see similar clusterings for F10 in the presence of calcium (a-b). However, A-H has one fewer clusters and has more temporally grouped clusters. That is, whereas A-H mostly split the concatenated trajectories into its substituent 4 individual trajectories, K-Means maximizing silhouette indicated more overlap in the trajectories. For MutSα in the presence of cisplatinated DNA, we see that A-H indicates one additional cluster, appearing to split K-means cluster 1 into two clusters. For the SufC protein, we see A-H indicating 5 more clusters than K-Means. SufC is a highly stable system, indicating that A-H is finding finer details in this system. From these comparisons, and those in Supporting Information Figure S21, we see that A-H has a preference for more tight clusters over K-Means maximizing the silhouette score.

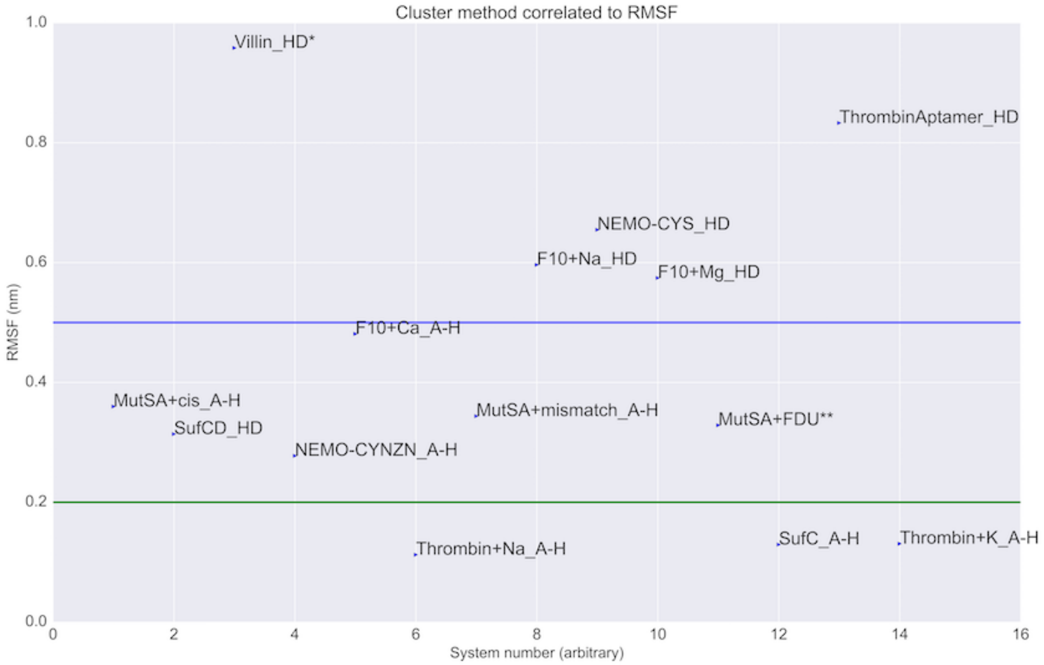


Figure S22: For systems with an average RMSF less than 2Å, we consistently observed distinct conformational changes across Amorim-Hennig clusters and gained little information beyond deciding the stability of the system from HDBSCAN clusters). We observed that HDBSCAN provided more meaningful clusters for polymers with an average RMSF larger than 5Å (Figure 9), which is likewise consistent with our conclusion that HDBSCAN is best for more systems with higher structural variance. Systems with average RMSFs between 2Å and 5Å had no clear pattern. In this figure, simulated systems are assigned an arbitrary system number (x-axis) and labeled in the format *system short name_most informative clustering algorithm*. *Our simulation for Villin headpiece was one of folding. The system is expected to undergo large conformational shifts, as it transitions from unfolded to folded. Above we presented Amorim-Hennig clusters for Villin headpiece due to their correlation to RGYR; however, it was HDBSCAN that found the stable folding intermediates. **On MutSα exposed to FdU-substituted DNA, the two clustering methods gave essentially the same result, with only a few initial simulation frames labeled differently.