k-Means NANI: an improved clustering algorithm for Molecular Dynamics simulations

Lexin Chen,[†,‡] Daniel R. Roe,[¶] Matthew Kochert,[§,‖] Carlos Simmerling,[§,‖,⊥] and Ramón Alain Miranda-Quintana[*,†,‡]

†*Department of Chemistry, University of Florida, FL, USA*

‡*Quantum Theory Project, University of Florida, FL, USA*

¶*Laboratory of Computational Biology, National Heart, Lung, and Blood Institute, National Institutes of Health, Bethesda, Maryland, USA*

§*Laufer Center for Physical & Quantitative Biology, Stony Brook University, Stony Brook 11794, USA*

‖*Department of Chemistry, Stony Brook University, Stony Brook 11794, USA*

⊥*Department of Applied Mathematics and Statistics, Stony Brook University, Stony Brook 11794, USA*

E-mail: quintana@chem.ufl.edu

Supporting Information Available

2D Datasets

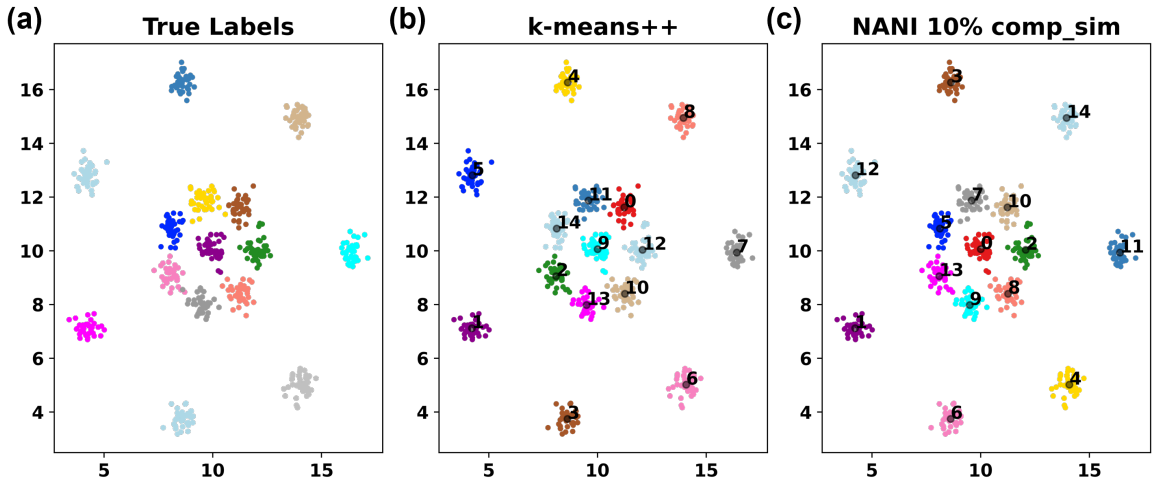

Figure S1: k-means NANI on a sample blob disk data. A different color represents a different cluster label. **(a)** True Labels. **(b)** k-means clustering with centroids initialized by k-means++. **(c)** k-means clustering with centroids initialized by k-means NANI.

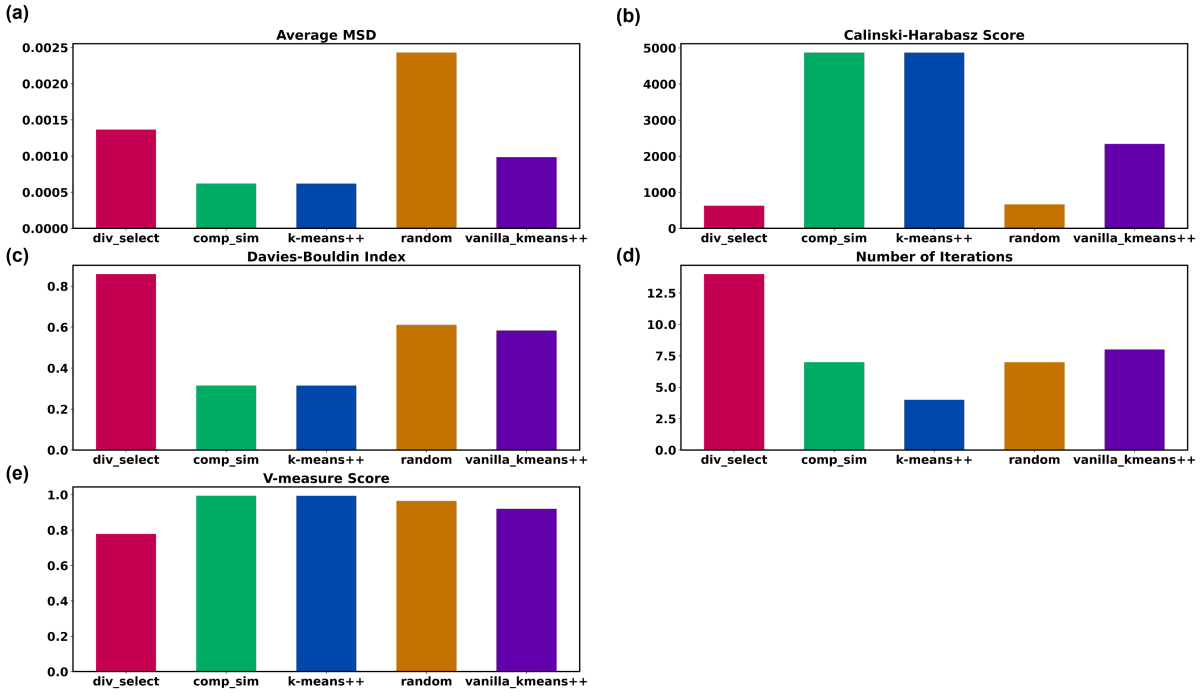

Figure S2: k-means NANI on a sample blob disk data. Summary of **(a)** average MSD, **(b)** Calinski-Harabasz score, **(c)** Davies-Bouldin index, **(d)** number of iterations, and **(e)** V-measure score using different seed selectors.

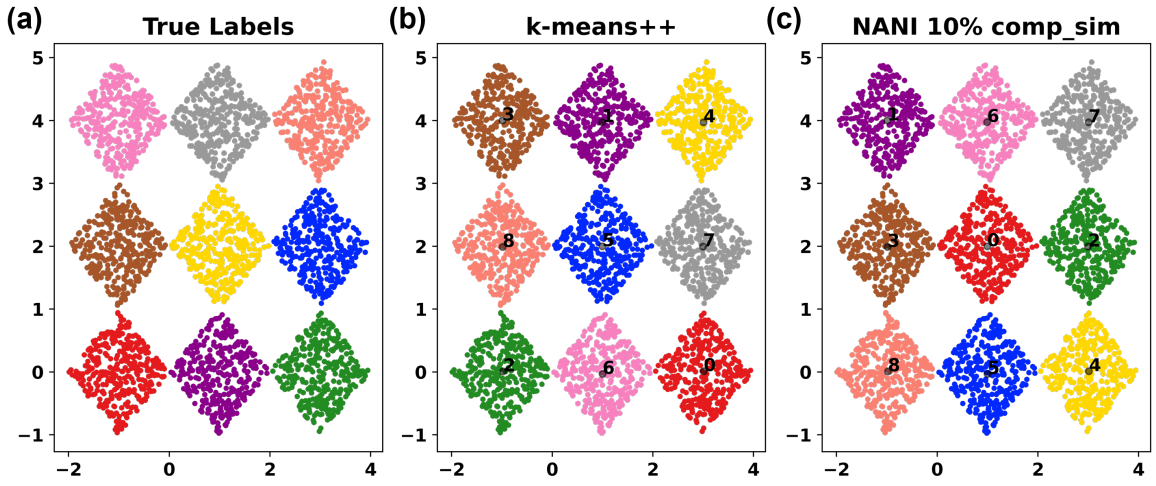

Figure S3: k-means NANI on a sample diamonds data. A different color represents a different cluster label. **(a)** True Labels. **(b)** k-means clustering with centroids initialized by k-means++. **(c)** k-means clustering with centroids initialized by k-means NANI.

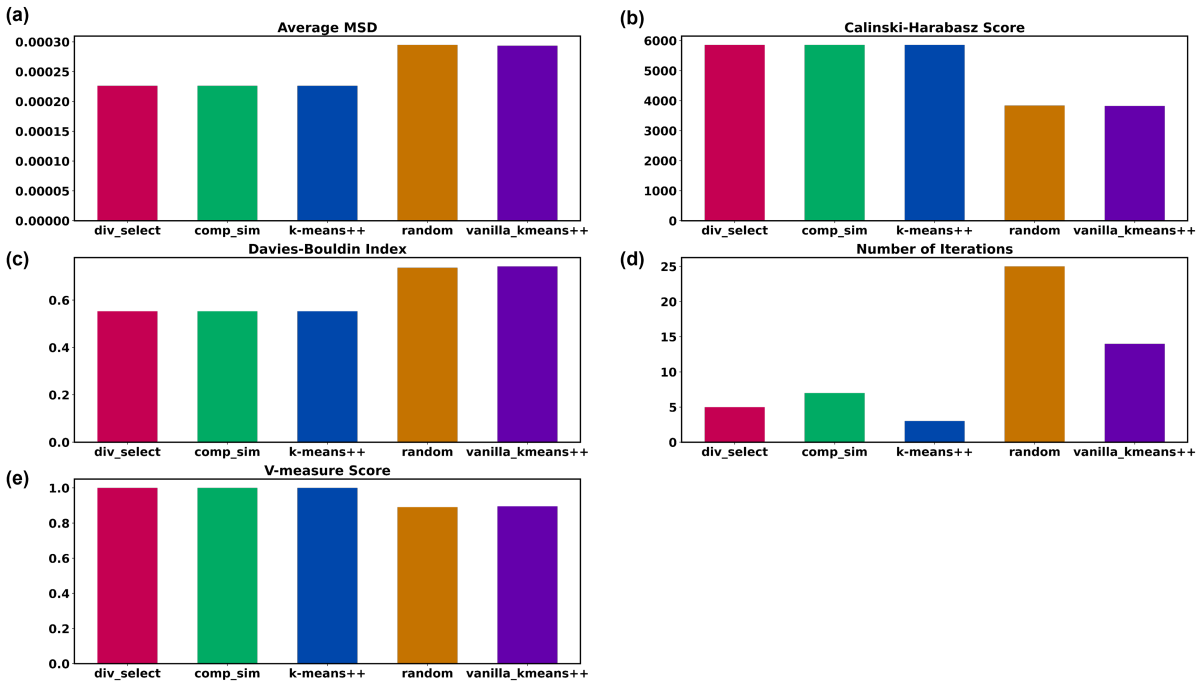

Figure S4: k-means NANI on a sample blob disk data. Summary of **(a)** average MSD, **(b)** Calinski-Harabasz score, **(c)** Davies-Bouldin index, **(d)** number of iterations, and **(e)** V-measure score using different seed selectors.

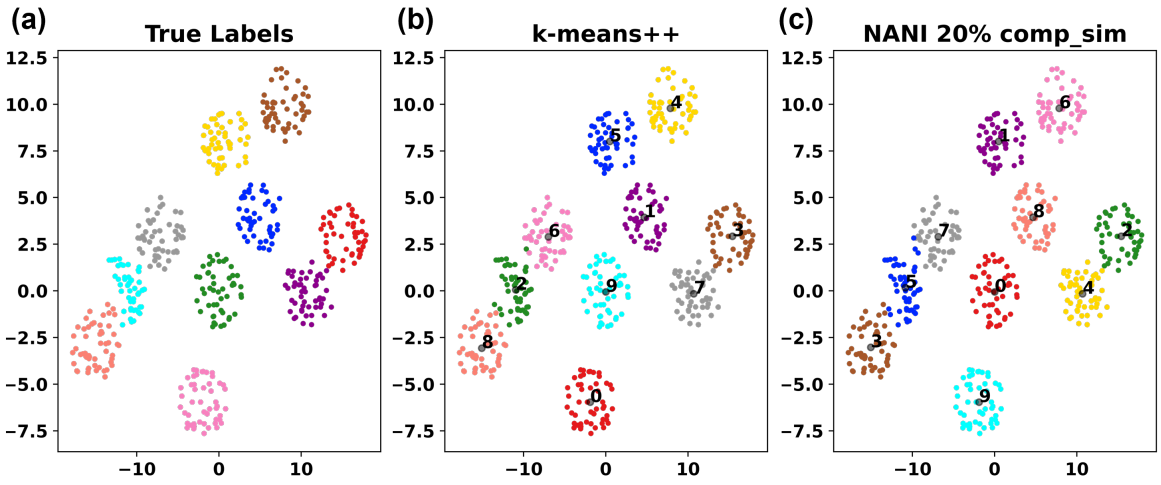

Figure S5: k-means NANI on a sample ellipses data. A different color represents a different cluster label. **(a)** True Labels. **(b)** k-means clustering with centroids initialized by k-means++. **(c)** k-means clustering with centroids initialized by k-means NANI.

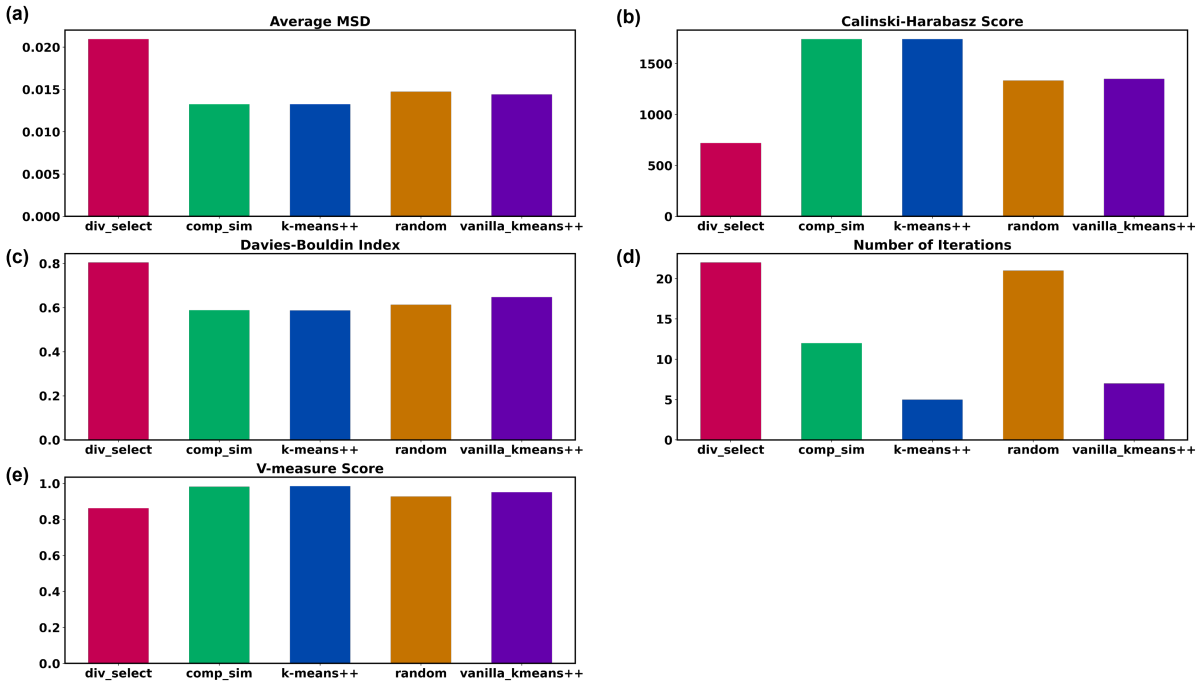

Figure S6: k-means NANI on a sample blob disk data. Summary of **(a)** average MSD, **(b)** Calinski-Harabasz score, **(c)** Davies-Bouldin index, **(d)** number of iterations, and **(e)** V-measure score using different seed selectors.

Peptides

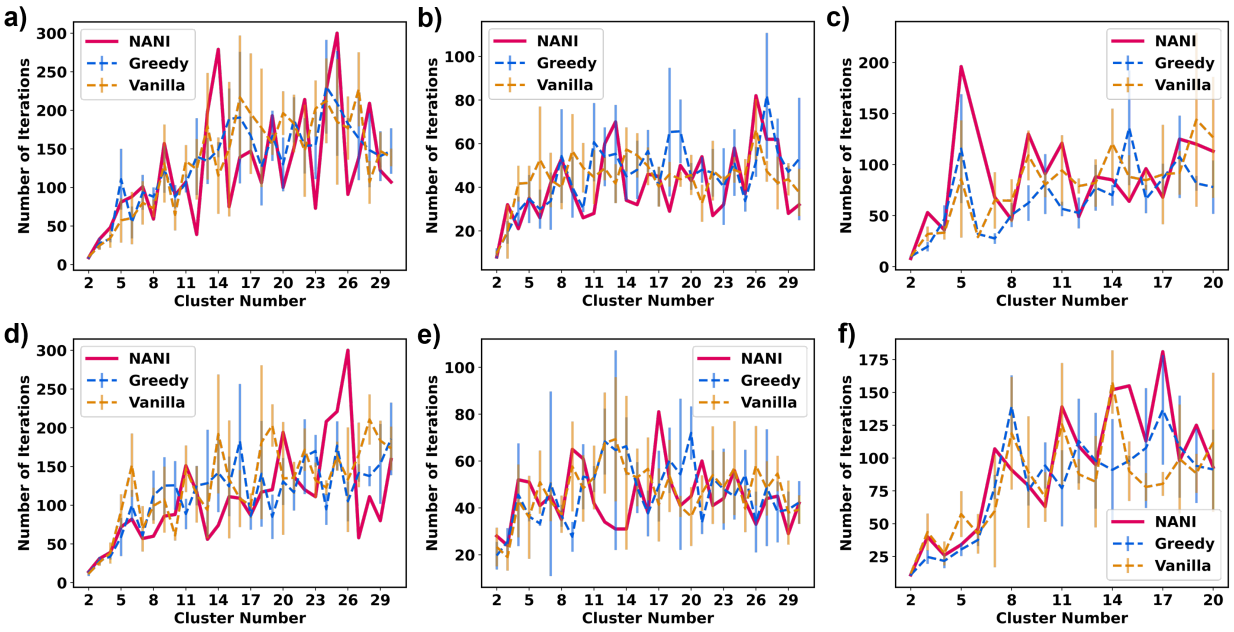

Figure S7: Number of Iterations of different seed selectors applied on three peptide systems. **(a)** R1Q hairpin single-reference aligned **(b)** β-heptapeptide single-reference aligned **(c)** Villin single-reference aligned **(d)** R1Q hairpin Kronecker aligned **(e)** β-heptapeptide Kronecker aligned **(f)** Villin Kronecker aligned

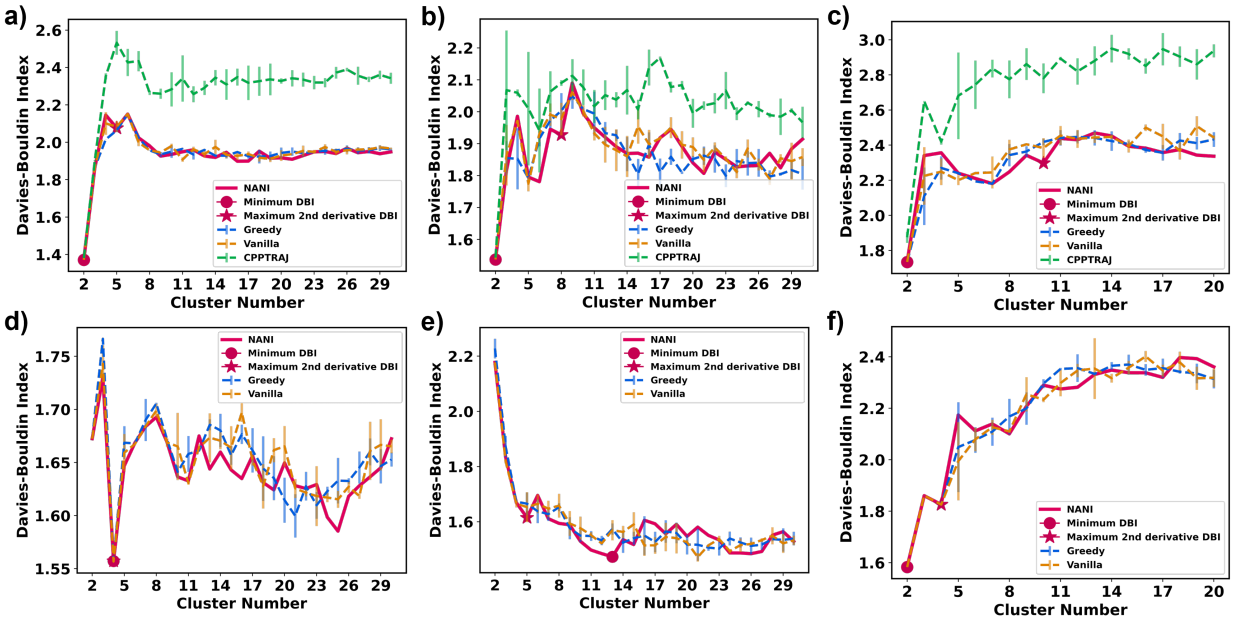

Figure S8: Davies-Bouldin Index of different seed selectors applied on three peptide systems starting at 2 clusters. **(a)** R1Q hairpin single-reference aligned **(b)** β-heptapeptide single-reference aligned **(c)** Villin single-reference aligned **(d)** R1Q hairpin Kronecker aligned **(e)** β-heptapeptide Kronecker aligned **(f)** Villin Kronecker aligned

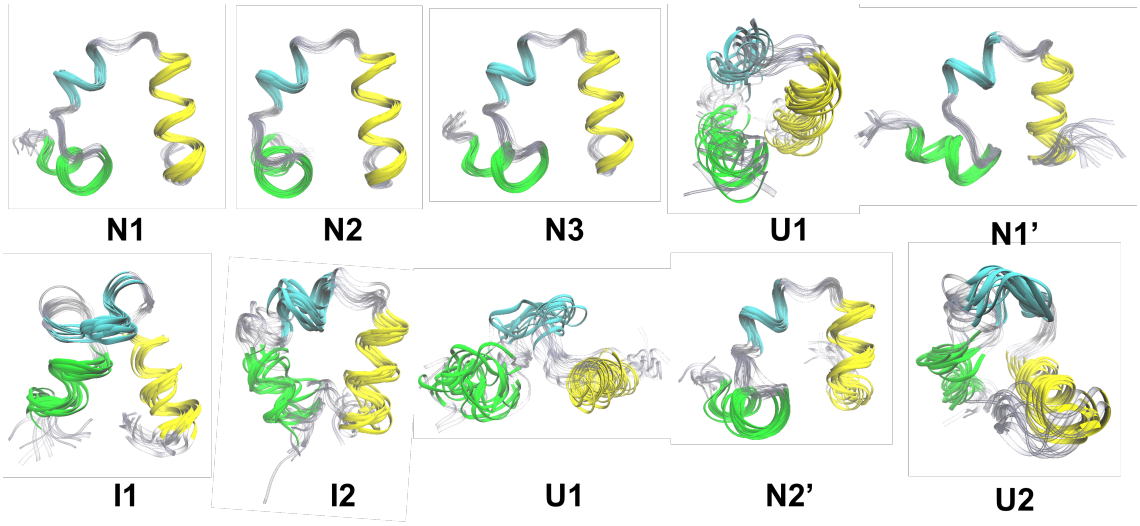

Figure S9: Structural overlaps for Villin in four states: folded (N), partially folded (N'), intermediate (I), and unfolded (U). Helix 1, 2, and 3 are in green, cyan, and yellow, respectively. Clustering was done using single-reference alignment and ten clusters were the maximum 2nd derivative DB determined from DB plots in Fig. 7c.

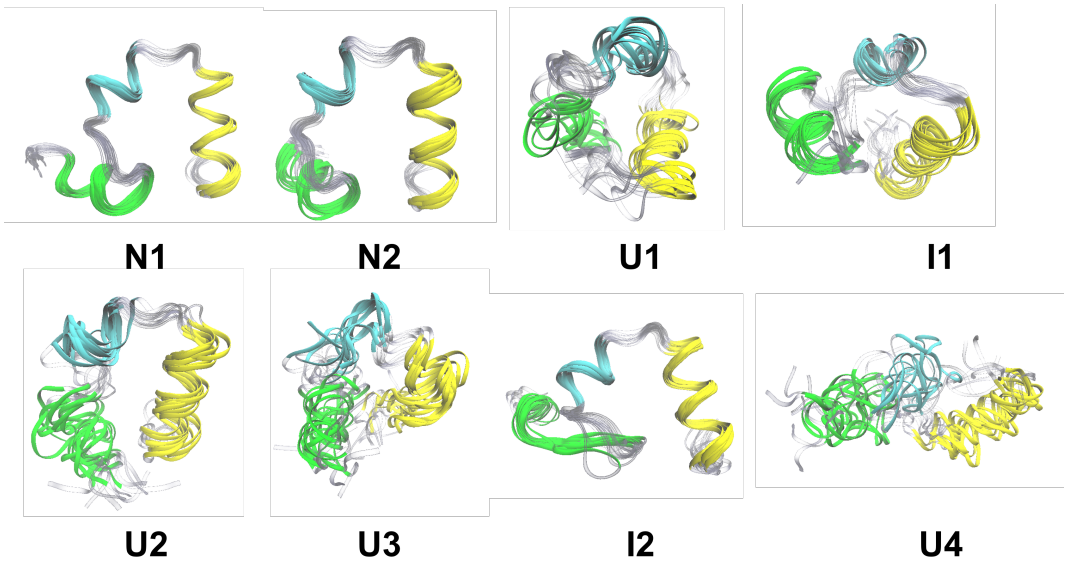

Figure S10: Structural overlaps for Villin in three states: folded (N), intermediate (I), and unfolded (U). Helix 1, 2, and 3 are in green, cyan, and yellow, respectively. Clustering was done using Kronecker alignment and eight clusters were the minimum DB determined from DB plots in Fig. 7f.

Protein

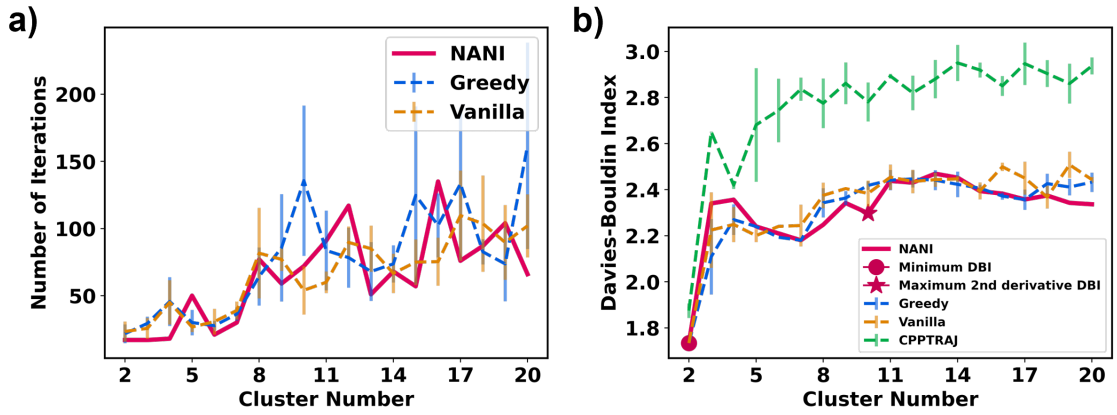

Figure S11: Indicators of different seed selectors applied on NuG2. **(a)** Number of iterations **(b)** Davies-Bouldin index starting at 2 clusters

Internal Coordinates

In order to showcase the generality of NANI we performed a proof-of-principle study of a well-studied model system (alanine dipeptide) using internal coordinates (e.g., two dihedral angles) instead of Cartesian coordinates. Internal coordinates are particularly attractive because they could help alleviate the problem of picking a reference to align with respect to. That is, even if an arbitrary set of internal coordinates could still not be enough to univocally set a frame of reference, it is certainly possible to select a complete set of internal coordinates that would eliminate the redundant rotational and translational degrees of freedom associated with the Cartesian coordinates. In this case, we can bypass to align to a reference, which removes an extra layer of bias in the clustering protocol. This, however, comes with the price of having to carefully handle the distance relations in internal coordinate space. In the

particular case of dihedral coordinates, we must be careful with the periodicity conditions (angles 0 and 359 are closer than angles 0 and 90). We incorporated these considerations in the NANI analysis and, as shown in Fig.S12, we were able to reproduce the PES structure of the alanine dipeptide, containing 6 main clusters, as reported in a plethora of previous studies.

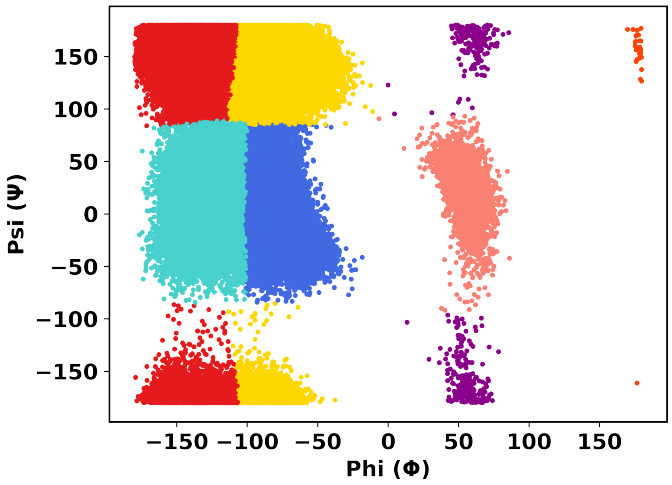

Figure S12: 2D representation of k-means NANI clustering applied on the internal coordinates (dihedral angles) of alanine dipeptide.

Cluster Representatives Over Multiple Runs

The non-deterministic nature of standard k-means methods affects not only the total number of clusters in the ensemble, but posses difficulties identifying a robust set of cluster representatives, since these can also largely vary from one run to the other.

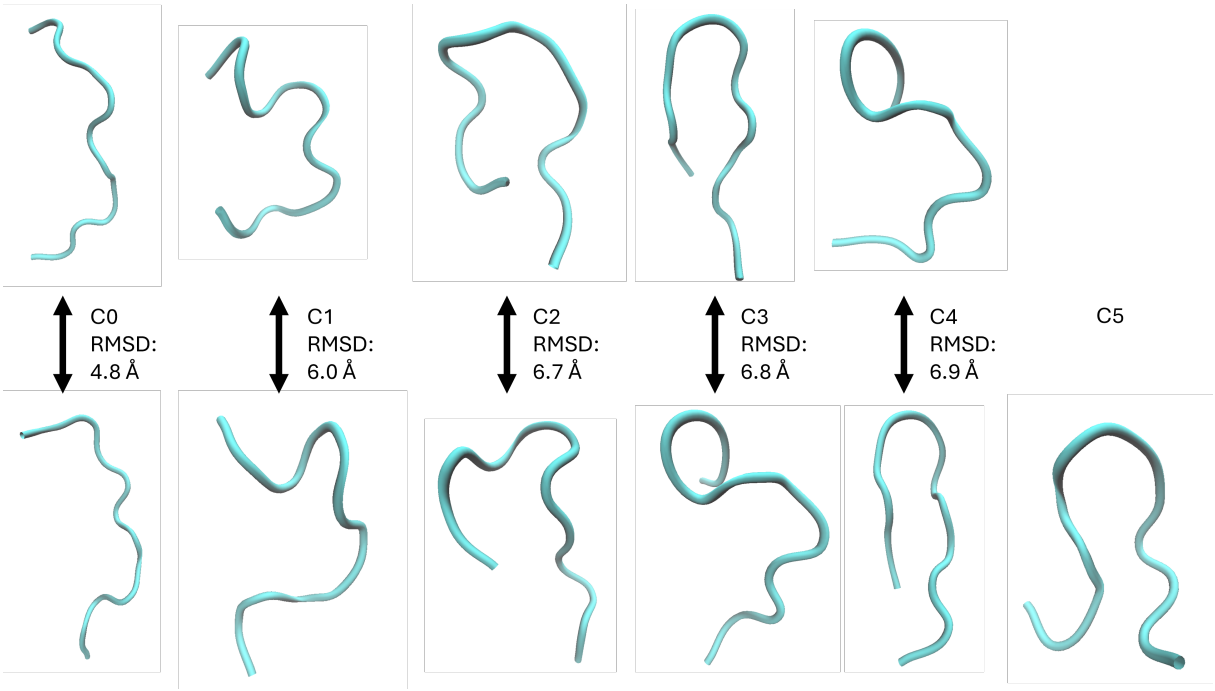

Figure S13: Structural overlaps for the cluster medoids of β-heptapeptide over two independent replicates of the CPPTRAJ flavor of k-means++.

NANI vs. Hierarchical Agglomerative Clustering

Although the main goal of this manuscript is to compare NANI to other k-means alternatives, we also want to showcase a brief comparison with other very popular tool in the MD community: Hierarchical Agglomerative Clustering (HAC). In particular, we considered two variants of HAC, using complete or average linkage criteria, respectively. First, regarding the quality clustering metrics, we see that NANI greatly outperforms the average and complete HAC when it comes to the CH index. Also, NANI offers results of equivalent quality to those of the average HAC when one considers the DB index, with both these methods greatly surpassing the complete HAC. Only when it comes to the silhouette index, average HAC outperforms NANI. However, NANI is completely unmatched if we consider the resources re-

quired to perform the clustering. Both HAC methods demand the construction of a pairwise RMSD matrix, which requires an extra 11 GB of memory. This step is not needed in NANI, which makes it a much more memory-efficient approach. Moreover, the complete HAC took 2.8 DAYS to cluster the HP35 trajectory, with the average HAC requiring even more time, at 3.8 DAYS. On the other hand, the full NANI scan was completed in 85 SECONDS.

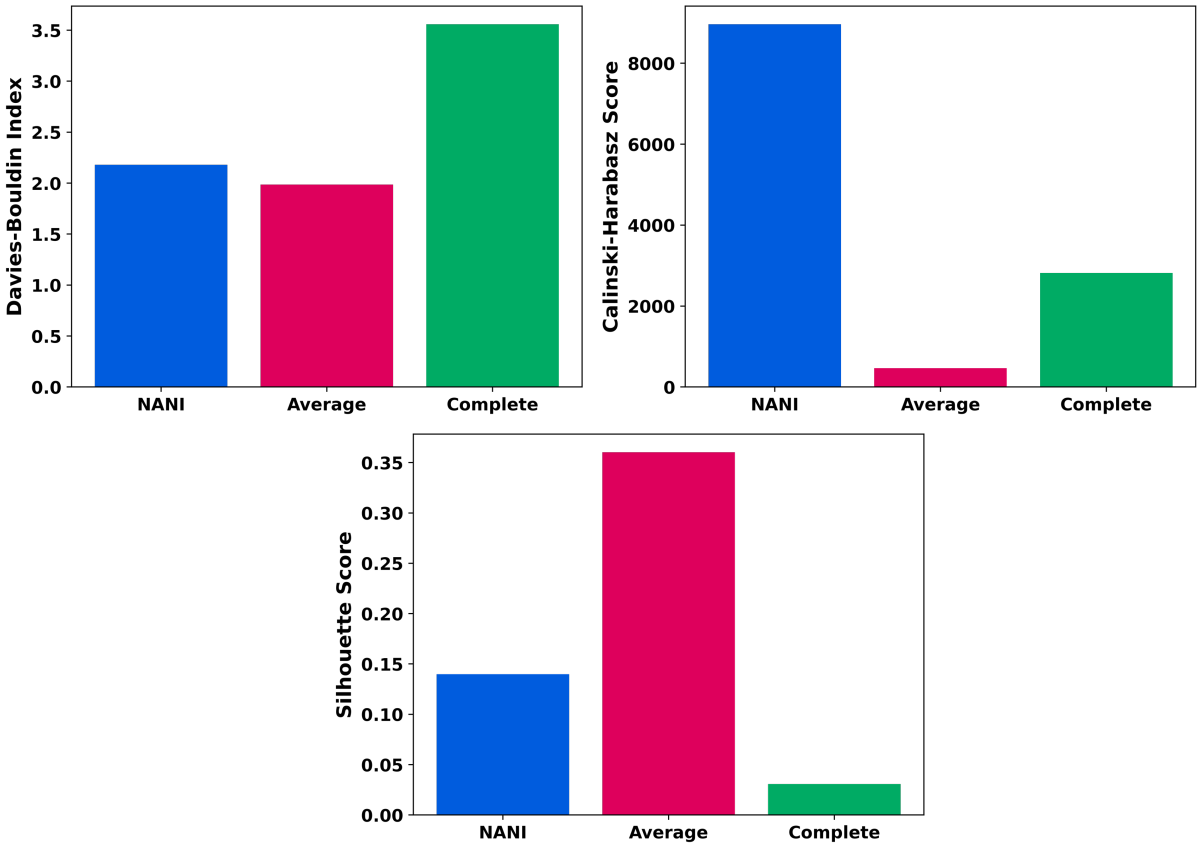

Figure S14: DBI, CHI, and Silhouette results for k=7 clusters from the HP35 simulation obtained with NANI (blue), average HAC (red), and complete HAC (green).

Further discussion on alignment

Given the central role of alignment on clustering, we performed some extra tests in order to showcase the generality of NANI. First, we considered the uniform multiple alignment proposed in the same paper that introduced the Kronecker alignment discussed in the main text. This is a much simplified version of the Kronecker version, that assumes a uniform relevance for all the involved coordinates, with no individual weights depending on their variances. However, as shown in Fig. S15, single-reference NANI consistently outperforms the referee's uniform alignment suggestion across the different quality clustering indices. In particular, for both DB and CH, single-reference NANI is better for all possible numbers of clusters. This supports our previous findings that the single-reference alignment method is still a valuable tool at the time of performing the clustering.

To conclude, we include an extreme case of single-reference alignment. In Fig. S17 we present a NANI scan after aligning to the most dissimilar frame from the original reference chosen in the main paper (see Fig. 8 in the main text). This is meant to provide the most strenuous test on the single-reference step. As can be seen in these new results, the quality of the final clustering, as quantified by the DBI, is relatively similar to that of the original single-reference method (and, remarkably, similar to the uniform alignment shown in Fig. S15). Moreover, the optimum number of clusters also closely follow the original reference values presented in the paper. This is not to say that the reference frame selection is not a critical step (it absolutely is!), but merely to show that the NANI results and performance are quite robust, even after a drastic change in single-reference selection.

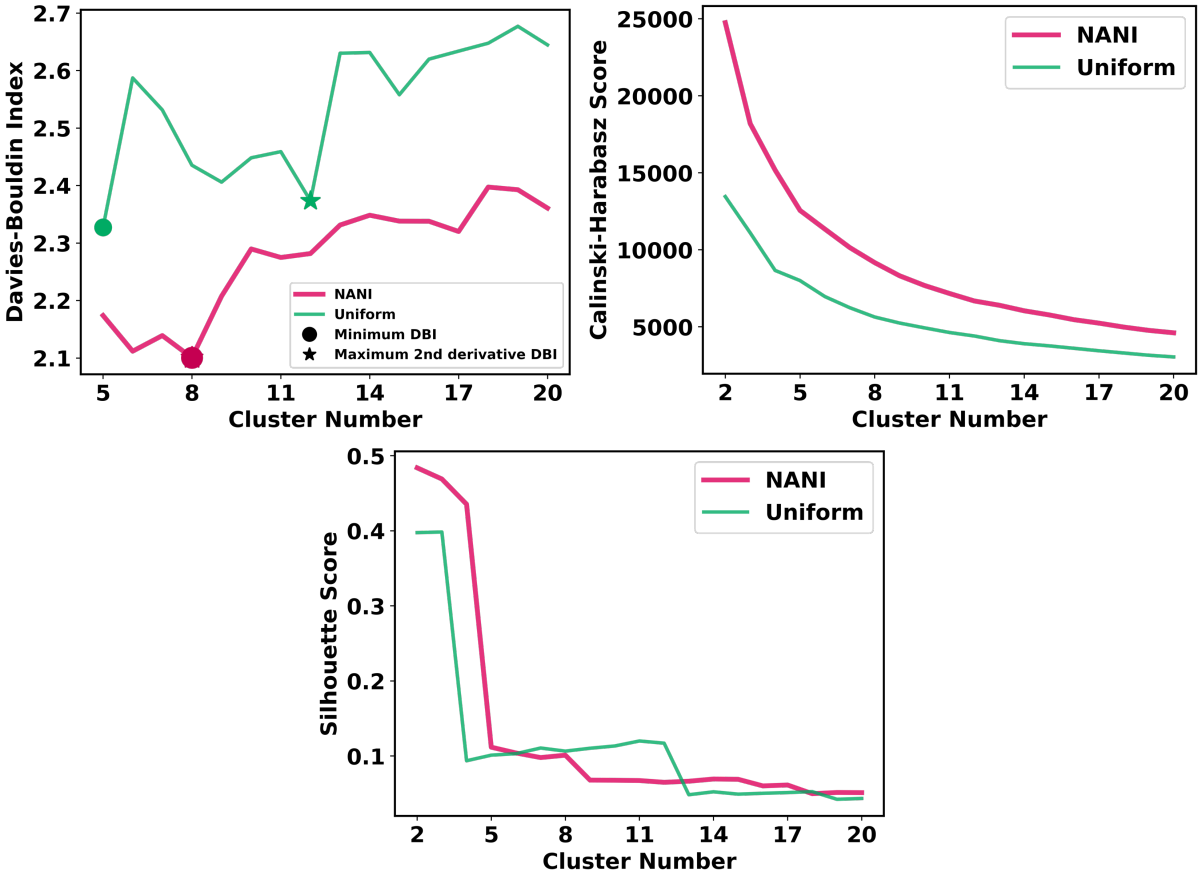

Figure S15: DBI, CHI, and Silhouette results for the HP35 simulation obtained with single-reference NANI (red) and uniform alignment (green).

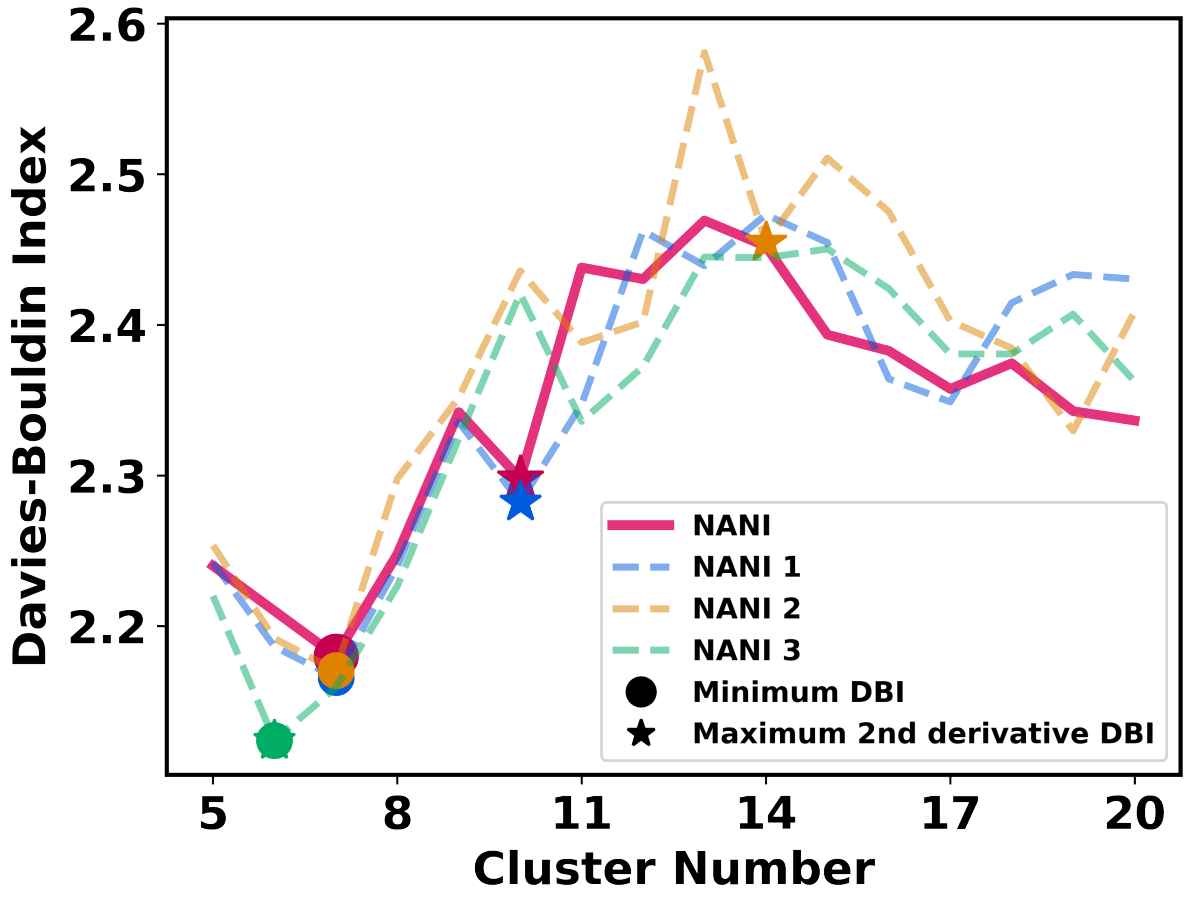

Figure S16: Different NANI scans for the HP35 trajectory. In continuous line, the original NANI scan presented in the main text. Discontinuous lines present the results after 3 random samples of 80% of the original frames.

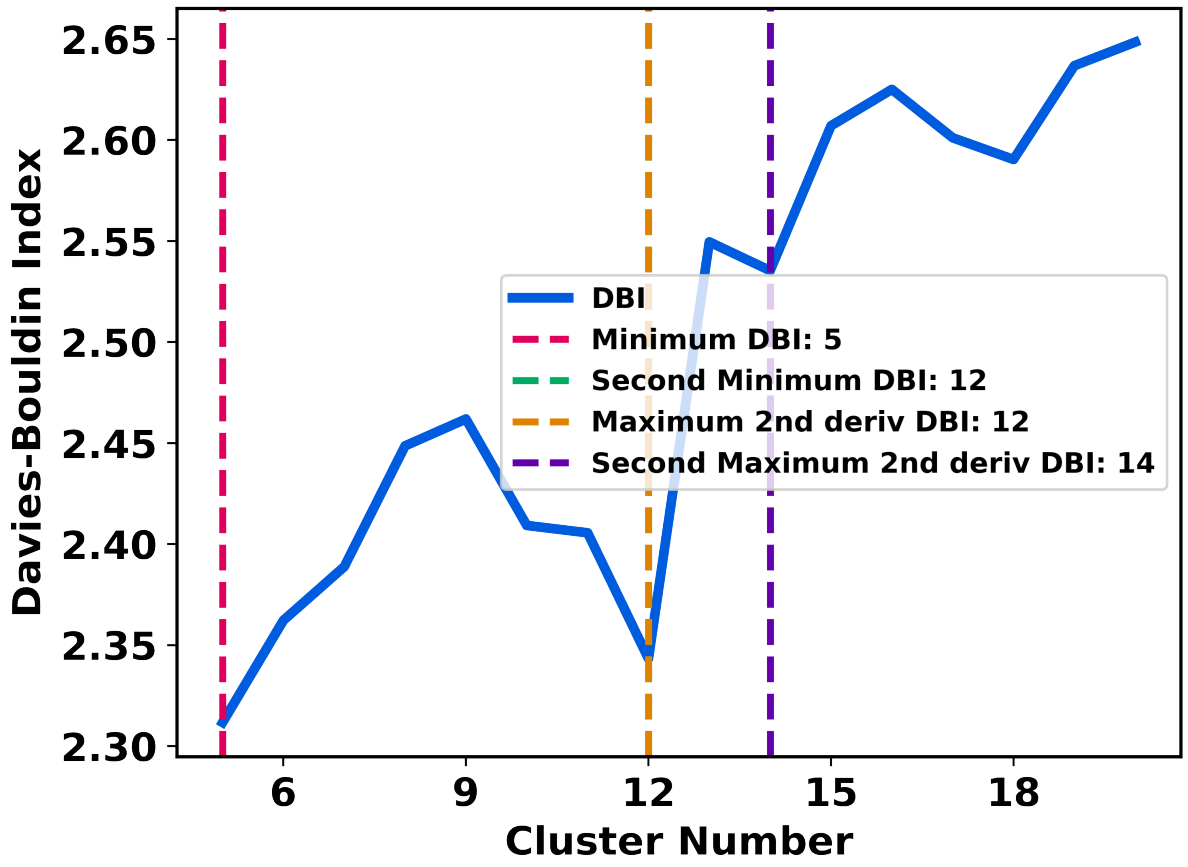

Figure S17: NANI DBI scan for the HP35 simulation, after aligning to the most distant frame from the one used as reference in Fig. 8 of the main text.