Scaling *k*-Means for Multi-Million Frames: A Stratified NANI Approach for Large-Scale MD Simulations

Jherome Brylle Woody Santos,[1] Lexin Chen,[1] Ramón Alain Miranda-Quintana[1]

1. Department of Chemistry and Quantum Theory Project, University of Florida, Gainesville, Florida, 32611, USA

Email: quintana@chem.ufl.edu

ABSTRACT

We present improved *k*-means clustering initialization strategies for molecular dynamics (MD) simulations, implemented as part of the N-ary Natural Initiation (NANI) method. Two new deterministic seeding strategies—strat_all and strat_reduced—extend the original NANI approaches and dramatically reduce the clustering runtime while preserving the quality of clustering results. These methods also preserve NANI's reproducible partitioning of well-separated and compact clusters while avoiding the costly iterative seed selection procedures of previous implementations. Testing on the β-heptapeptide and the HP35 systems shows that these new flavors achieved Calinski–Harabasz (CH) and Davies–Bouldin (DB) scores comparable to the previous NANI variant, indicating that the efficiency gains come with no quality decrease. We also show how this new variant can be used to greatly speed up our previously proposed Hierarchical Extended Linkage Method (HELM). These enhancements extend the reach of NANI to accelerate large-scale MD analysis both in stand-alone *k*-means clustering and as a component of hybrid workflows, and remove a key barrier to routine, scalable, and reproducible exploration of complex conformational ensembles. The improved NANI implementation is accessible through our MDANCE package: https://github.com/mqcomplab/MDANCE.

1. INTRODUCTION

Clustering has become an important tool in molecular dynamics (MD) trajectory analysis. It is widely used to probe structure, dynamics, and energetics of biomolecular systems across time.[1–3] Among various clustering algorithms, *k*-means remains one of the most widely used due to its computational efficiency.[4] However, conventional *k*-means has some well-known drawbacks for its applications on MD data. In particular, the algorithm is highly sensitive to the initial centroid selection, and the common seeding schemes (random or *k*-means++ initialization) can perform poorly on complex and high-dimensional MD datasets, while also having great variability in the final results. *k*-means++ uses a probability based greedy initialization that improves upon random seeding, but being stochastic, it still cannot guarantee reproducible results.[5,6] This lack of reproducibility is also a serious issue for MD analysis as different runs can lead to different cluster assignments, which may hinder the extraction of meaningful physical insights from these trajectories.

To address these challenges, the *k*-means N-ary Natural Initiation (NANI) protocol was recently developed.[7] NANI introduces a new seeding approach by using efficient *n*-ary comparisons, which involve evaluating multiple objects simultaneously,[8–15] to identify high-density regions of the conformational space and select diverse, representative initial seeds. By design, the selected frames (medoids) are both dense and well-separated, allowing the *k*-means algorithm to generate well-partitioned clusters. More importantly, NANI is completely deterministic: it produces identical clusters and populations across repeated runs, removing stochastic variability and enhancing reproducibility. Previous benchmarks on peptides and proteins show that NANI clusters are compact, well-separated, and in agreement with known conformational states.

In its initial implementation, NANI provided two deterministic seeding modes: comp_sim and div_select. The former selects initial centers based on the diversity within the densest region of the dataset, while the latter maximizes diversity across the entire dataset. Both methods use only a subset of the trajectory frames (by default 10%) to identify candidate seeds, which have significantly improved clustering quality and reproducibility compared to random or *k*-means++ seeding. However, as the size of MD datasets continues to increase, even these strategies encounter performance bottlenecks due to the high computational cost of processing massive numbers of frames. In this approach, we found that the main computational bottleneck is the diversity selection algorithm, which can scale poorly to millions of frames.

In this application note, we introduce two new seeding strategies for NANI: strat_all and strat_reduced. These methods preserve NANI's deterministic character and robustness, while substantially accelerating the overall clustering process. We outline these stratified initialization approaches and demonstrate their performance on two representative MD systems: β-heptapeptide and villin headpiece (HP35). We use standard validation metrics like the Calinski-Harabasz index (CHI)[16] and Davies-Bouldin index (DBI)[17] to evaluate cluster compactness, separation, and consistency across the initialization schemes. The results show that strat_all and strat_reduced match the clustering performance of original NANI methods, while significantly reducing the computational time. A single *k*-means pass is ~1.5x faster for the β-heptapeptide, ~13x for 150 k HP35 frames, and nearly 45x for 1.5 M HP35 frames—speedups that directly translate to faster HELM clustering with no loss of accuracy.

2. THEORY

k-means clustering requires an initial set of *k* cluster centers (seeds), which are traditionally chosen at random or via *k*-means++. NANI introduced a deterministic approach by leveraging complementary similarity to rank the individual conformations, which are further pre-processed through a diversity picking algorithm. This approach ensures that the seeds are both from high-density regions and are diverse.

Complementary similarity relies on the principle of *n*-ary analysis by evaluating the change in overall dataset similarity when a particular frame/conformation is removed. A high complementary mean squared deviation (MSD), cMSD, indicates that its exclusion significantly reduces overall similarity. This identifies it as a highly representative and central conformation, most likely the medoid of the set. Building on this concept, the first NANI implementations introduced two modes for seed selection: comp_sim and div_select. In the comp_sim approach, one first computes the cMSD of each frame. A user-specified top fraction (e.g., 10%) of frames with the highest cMSD is retained as a pool of candidate medoids. From this group, *k* seeds are chosen using an iterative diversity maximization: starting from the dataset's medoid (most central frame) as the first seed, subsequent seeds are added one by one such that each new seed maximizes the distance to the already selected seeds. The div_select method is similar but it applies the diversity maximization over the entire dataset instead of doing an initial filtering of high-density frames. Both methods reliably produce well-spread representative medoids, but the iterative selection, which requires the repeated scanning of the remaining candidates, can be costly for large *N*.

Stratified Sampling Strategies: The latest version of NANI introduces two new deterministic initialization methods: strat_all and strat_reduced, which employ stratified sampling of the conformational ensemble to select initial medoids in a single pass, avoiding the iterative comparisons used in the original approach. These strategies leverage the observation that the *k* seeds should sample the full range of high-density conformations without the need for an iterative distance maximization. In both methods, we again exploit the complementary similarity scores as an efficient guide to the density distribution of the dataset.

strat_all: This method performs stratified sampling over the entire dataset. First, each frame's cMSD is computed in O(*N*) time. The frames are then sorted from highest to lowest cMSD (i.e., from the most to least important frames). Next, the sorted list is divided into P equal-sized strata (bins), where P is determined by a user-defined percentage of the dataset (for instance, 10% yields P = 0.1N). From each stratum, the top-ranked frame (highest cMSD) is selected as a representative. This yields P candidate medoids that are distributed across the diversity of the dataset. The first *k* representatives are then selected as the initial cluster medoids. By sampling in this stratified manner, strat_all ensures that the seeds include the most important conformations while also covering different regions of the space. And more importantly, there are no random steps made, given the same dataset and percentage parameter, strat_all will always produce the same set of seeds. The complexity is primarily in the calculation and sorting of the cMSD values, which are far more efficient than an exhaustive iterative selection.

strat_reduced: This variant combines comp_sim with stratified sampling. Initially, same as in comp_sim, we identify a subset of the data corresponding to the top fraction of highest density frames. Within this subset, cMSD values are recalculated, and frames are sorted and partitioned into a number of bins, based on the subset size. The first *k* representatives, one from each bin, are then selected as the initial medoids. This approach effectively filters out less important outliers prior to seed selection. The additional stratification within the subset reduces redundancy among the selected seeds and promotes greater structural diversity.

Both strat_all and strat_reduced skip the iterative addition loop of the earlier methods in favor of a one-time selection using the sorted indices. In the worst case, strat_all and strat_reduced perform O(*N*) work to compute the cMSD and O(*N* log *N*) to sort, whereas comp_sim and div_select can require O(*N* x *k*) complexity due to their reiterative nature. For typical MD clustering scenarios, stratified methods can be an order of magnitude faster at initialization.

Extensive comparisons between the original NANI methods and conventional initialization strategies (e.g., random, *k*-means++) have already been reported in our prior work,[7] where NANI was shown to outperform

these methods in both clustering quality and reproducibility. These analyses have also demonstrated that the comp_sim mode is significantly more computationally efficient and more robust than div_select. Therefore, for simplicity, we exclusively compare our new stratified sampling strategies (strat_all and strat_reduced) against the comp_sim baseline.

3. SYSTEMS

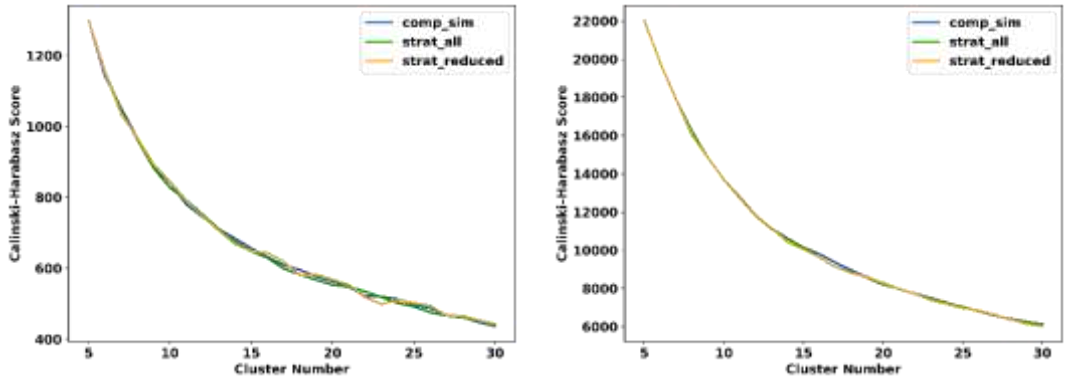
β-Heptapeptide. The topology and trajectory files were obtained from publicly available data and accessed via GitHub.[18,19] Clustering was performed using only the backbone atoms (N, Cα, C, O, H) of residues 2 to 11; excluding terminal residues and all side chains to reduce noise, consistent with the protocol described by Daura et al.[20] After discarding the first 1,000 frames as equilibration, the trajectory was aligned to the 1,000th frame. A total of 6,001 structures were used for clustering.

HP35. We analyzed the long-timescale trajectory of the Nle/Nle mutant of the C-terminal sub-domain of the villin headpiece (HP35) from D. E. Shaw Research.[21] The production run spans 305 μs at 360 K and contains ~1.52 million frames, saved every 200-ps intervals. The first 5,000 frames were discarded as an equilibration period, and the remaining frames were aligned to the 5,000th frame. Clustering was performed using the backbone heavy atoms, following conventions established in previous HP35 studies: the N atom of residue 1; the N, Cα, and C atoms of residues 2 to 34; and the N atom of residue 35. Sieving was applied to select every 10th frame, resulting in a more manageable dataset of 151,104 structures for clustering.

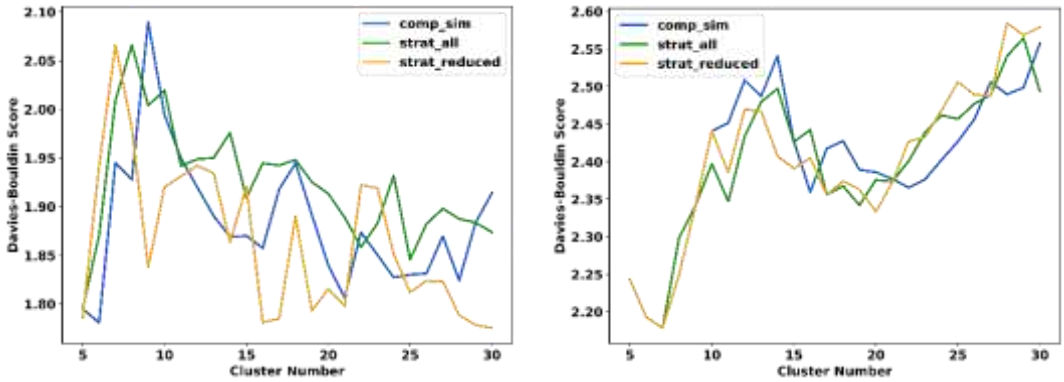
4. RESULTS & DISCUSSION

We evaluated the three NANI variants (comp_sim, strat_all, and strat_reduced) on the β-heptapeptide and the HP35 systems. Our initial test involves the screening on the number of clusters from 5 to 30: for each k we ran k-means with the calculated seeds and computed two standard internal validation metrics: Calinski–Harabasz (CHI) and Davies–Bouldin (DBI) indices. The CHI describes the ratio of the inter and intra-cluster dispersion; larger values signify tighter, better-separated clusters.[16] On the other hand, the DBI quantifies how well the clusters are both compact and mutually separated; lower values indicate greater inter-cluster separation and intra-cluster compactness.[17]

As shown in Fig.1, every seeding strategy within NANI yields nearly overlapping CHI curves on both systems. While the DBI curves show visual differences, they follow the same overall trend for all three initiators. These observations confirm that across the full screening window, the stratified NANI variants deliver clustering performance that matches the quality of the original NANI method.



A **B**



C **D**

Figure 1: Change in Calinski-Harabasz (A, B) and Davies-Bouldin (C, D) indices for A, C: β-heptapeptide and B, D: HP35 after *k*-means NANI screening from $k = 5$ to $k = 30$.

As we have shown in our previous studies, the CHI and DBI curves are valuable not only for comparing clustering methods (or initialization strategies in this case), but also for estimating the most meaningful cluster count for a given trajectory. We approach this in two ways. First, a global assessment of the curves or simply the selection of the absolute CHI maximum or DBI minimum across the screened range. Second, a local stability assessment of the curvatures of the metrics by evaluating their finite-difference second derivatives, the prominent extrema in these curvatures might be useful in identifying *k* values that are more stable than their immediate neighbors. As both indices are known to favor small cluster counts, the search was limited to $k \geq 5$.

In Table 1, we summarize the cluster counts suggested by the global (absolute DBI minimum, absolute CHI maximum) criteria for each seeding strategy. For the β-heptapeptide, the three modes: comp_sim, strat_all, strat_reduced, converge at $k = 5$ to 6 when the global measures are applied. The HP35 trajectory exhibits even stronger agreement: every global metric for the seeding scheme points to 5 to 7 clusters, mirroring the values reported in our original *k*-Means NANI study.

Table 1: Preferred numbers of clusters for β-heptapeptide and HP35 based on the global, DBI minimum and CHI maximum, metrics after the NANI screening.

Method/System	β-heptapeptide	HP35
minimum DBI / 2nd minimum DBI		
comp_sim	6/5	7/6
strat_all	5/25	7/6
strat_reduced	30/29	7/6
maximum CHI /2nd maximum CHI		
comp_sim	5/6	5/6
strat_all	5/6	5/6
strat_reduced	5/6	5/6

By contrast, the local-curvature tests (Table S2) frequently wander to much larger *k* values, especially for the β-heptapeptide, confirming earlier observations that second-derivative metric can over-react to minor fluctuations in the curves for these systems. Because the global metrics are both more conservative and more consistent across seeding modes, we retained them as the principal guide and fixed k = 6 for β-heptapeptide and $k = 7$ for HP35, a choice also supported by our previous analysis of this system. The

following section explores how these NANI clusters partition the conformational landscape of β-heptapeptide and HP35 and compare their properties as we vary the initialization schemes.

Table 2: Cluster populations and MSD for β-heptapeptide with $k = 6$ clusters obtained with the three NANI modes. Clusters are ordered in increasing population.

comp_sim		strat_all		strat_reduced	
Population	**MSD**	**Population**	**MSD**	**Population**	**MSD**
582	7.33	680	7.44	663	7.33
645	4.15	738	4.85	741	4.86
693	7.76	946	10.58	1007	10.62
1012	10.83	951	10.46	1067	10.39
1225	9.68	1235	9.13	1210	9.27
1844	10.35	1451	9.69	1313	9.51

To verify that strat_all and strat_reduced recover the same physical states as comp_sim at the optimal cluster count, we first analyzed the β-heptapeptide at $k = 6$ (Table 2). The three seedings strategies recover essentially the same partitioning, with the pronounced deviations confined to two clusters. The third cluster gains approximately 250-315 extra frames under the new methods, accompanied by a ~2.8 increase in MSD, indicating that they are slightly farther from the medoid. Conversely, the sixth cluster (largest), loses roughly the same number of structures, and its MSD drops by ~0.7, indicating that there is a tightening of its structural core. This trade-off suggests that the new initialization methods merely transferred/distributed a peripheral shell of frames from a diffuse, large cluster, into low population clusters. Although the medoids are not identical, we were still able to observe the important conformational states for this system. Fig. 2 shows the structural overlaps of the best representative structures from the six clusters of the β-heptapeptide system.

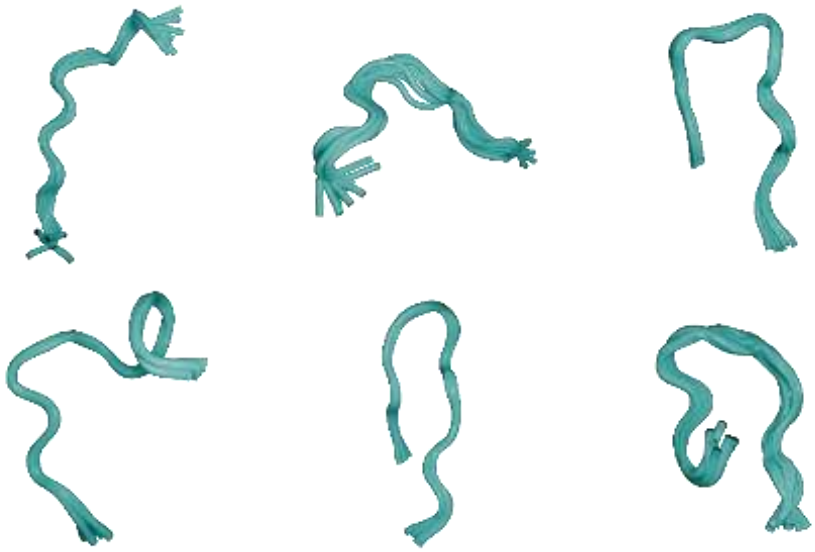


Figure 2: Overlaps of best representative structures from the six clusters of β-heptapeptide after performing NANI with strat_reduced.

Table 3: Cluster populations and MSD for HP35 with $k = 7$ clusters obtained with the three NANI modes. Clusters are ordered in increasing population.

comp_sim		strat_all		strat_reduced	
Population	**MSD**	**Population**	**MSD**	**Population**	**MSD**
5577	80.44	5569	80.33	5577	80.44
8123	91.31	8085	91.37	8121	91.31
11929	76.01	11950	76.01	11927	76.02
12096	46.55	12108	46.70	12100	46.55
13274	72.29	13260	72.29	13274	72.29
32815	14.10	32824	14.11	32815	14.10
67291	9.27	67309	9.28	67291	9.27

We also examined the cluster populations and MSDs for HP35 at $k = 7$. As Table 3 shows, the seven clusters are virtually unchanged by the choice of seed: each cluster's population differs by fewer than 40 structures out of ~150 000 frames (< 0.03 %), and the MSDs vary by at most 0.15. Fig. 3 shows the best representative frames from each cluster produced by strat_reduced, illustrating the clear separation and compactness of the resulting clusters. Upon inspection, we found that the same representative structures were selected by comp_sim and strat_all as well. In other words, all three initialization methods produced identical medoid frames at $k = 7$. Together, these results demonstrate that the new stratification methods carve out the same conformational states, with the same structural heterogeneity as the original protocol, but at a fraction of the computational cost.

Figure 3: Overlaps of best representative structures from the seven clusters of HP35 after performing NANI with strat_reduced. Helices 1, 2, and 3 are colored green, cyan, and yellow, respectively.

Having established that the two stratified protocols reproduce the comp_sim clusters at $k = 7$, we next quantified their speed advantage by benchmarking a full k-means pass at $k = 60$. The results in Table 4 show that the stratified seeders already save approximately 30% of the runtime on a small β-heptapeptide trajectory and scale far better on the larger HP35 dataset, where the time spent on the complete cycle dramatically decreased from almost 30 min to ~160 s, an order-of-magnitude improvement. In practical terms, the stratified seeders deliver clusters of similar quality at about one-tenth the cost on a 150,000-frame trajectory. The payoff becomes more dramatic at larger scales: k-means clustering of the full 1.5 million HP35 frames at $k = 60$ using comp_sim took ~30 hours whereas strat_all and strat_reduced now completes the same task in ~40 min, almost a 45-fold speedup.

Table 4: Runtime (s) to complete a single *k*-means pass ($k = 60$) with each NANI seeding strategy.

seeding strategy	β-heptapeptide 6001 frames	HP35 ~150,000 frames	HP35 ~1,500,000 frames
comp_sim	6.37	1995.48	109200.04
strat_all	4.23	158.99	2192.19
strat_reduced	4.48	183.39	2386.92

Integration in the HELM workflow: In the Hierarchical Extended Linkage Method (HELM)[22] pipeline, NANI provides a deterministic *k*-means pre-clustering that groups frames into small representative clusters before the hierarchical merging. In our previous runs, this pre-clustering, performed with the comp_sim variant of NANI, dominated the computational cost. On the 151k HP35 frames, the NANI part consumed ~33 min, accounting for 98 % of HELM's total runtime, creating a clear bottleneck. Replacing comp_sim with either strat_all or strat_reduced eliminates this challenge. A complete HELM run now finishes in ~3 min, a full order-of-magnitude speed-up (Table S2).

We saw that the significant speed-up does not compromise on clustering quality. While each seeding strategy produced a different set of initial *k*-means cluster labels, running HELM on these different inputs yielded highly similar results. As we show on Fig. 4, the CHI and DBI at each merging step in the HP35 system follow the same trends for all three initialization methods. Fig. 5 presents the overlaps of the best representative frames when $k = 6$. All HELM runs were done with trimming (retaining clusters with MSD < 10) and merged using the inter/intra linkage. This improvement now makes fully deterministic clustering feasible for the multi-million frame trajectories typical of modern large-scale MD simulations.

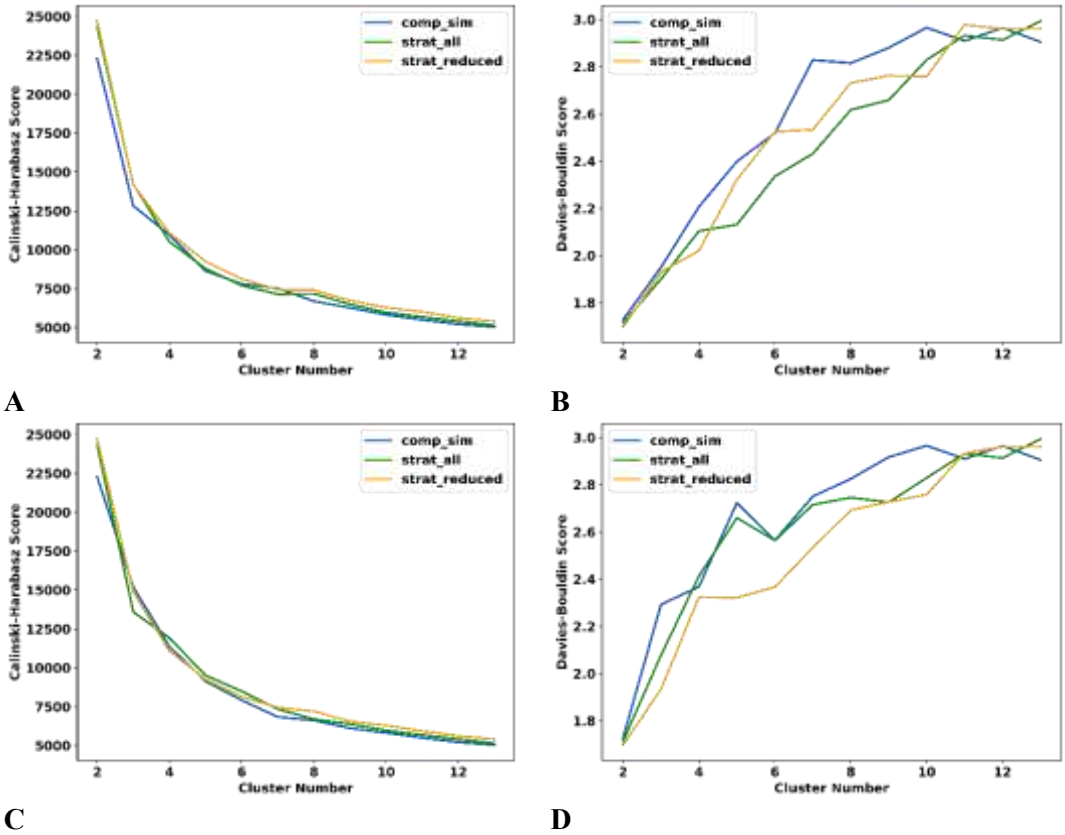


Figure 4: Change in Calinski-Harabasz (A, B) and Davies-Bouldin (C, D) indices for the HP35 simulation after trimming (retaining clusters with MSD < 10) the initial NANI clusters using the inter merge (A, C) and the intra merge (B, D).

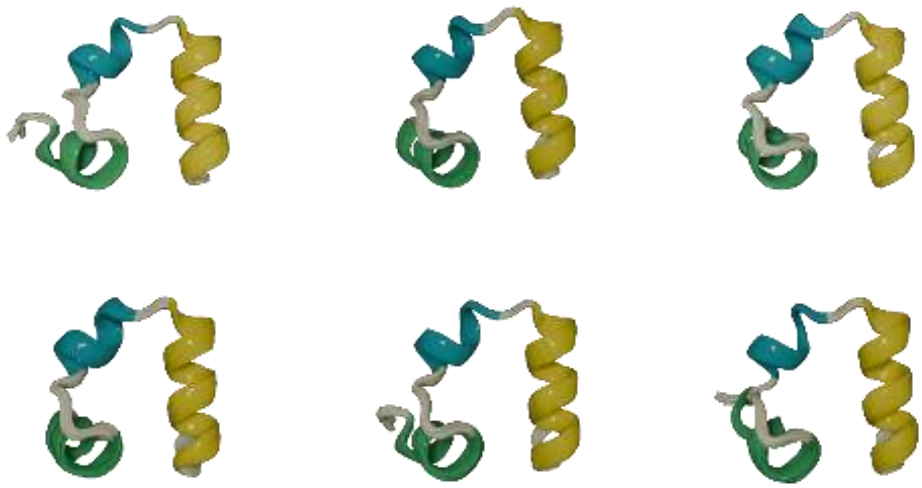


Figure 5: Overlaps of best representative structures from the six clusters of HP35 after performing HELM from 60 NANI strat_reduced clusters using the intra merge with trimming.

5. CONCLUSION

We have presented two deterministic, stratified *k*-means initialization strategies—strat_all and strat_reduced— that improve on the efficiency of *k*-means NANI clustering for MD simulations. These new variants address a key limitation of the previous NANI implementations: the high computational cost of iterative seed selection on large datasets. By replacing that step with a one-pass stratified sampling based on complementary similarity, we retain the core strengths of the method: determinism, robustness, and physical interpretability, while dramatically reducing runtime. Our benchmark across systems demonstrate that these improvements do not sacrifice clustering quality, as evidenced by the consistent CHI and DBI scores as well as cluster populations that are in agreement with those from prior NANI runs.

Importantly, we showed how these improvements substantially improve the efficiency of HELM. The original HELM pipeline leveraged NANI pre-clustering to group frames into small clusters before the hierarchical merging. However, as noted in our prior work, the NANI step had become the computational bottleneck, particularly on large-scale trajectories. By replacing comp_sim with strat_all or strat_reduced, we removed this bottleneck entirely. Clustering runtimes fell by more than an order of magnitude while producing hierarchies that preserved the biophysically interpretable clusters.

Again, the success of these new stratified NANI methods lies not only in their significant acceleration of the seed selection process in *k*-means clustering but also how they enable scalable downstream hierarchical workflows. By replacing greedy initialization with deterministic stratification, our methods remove one of the key barriers to utilizing hybrid protocols like HELM at large scale. As such, strat_all and strat_reduced extend NANI's impact beyond standalone use, positioning it as an important tool for advancing routine, scalable, and reproducible analysis of complex conformational ensembles in molecular simulations.

SUPPORTING INFORMATION

Second-derivative cluster count analysis for β-heptapeptide and HP35, structural overlaps of representative cluster conformations generated using different NANI seeding strategies and HELM merging schemes, and detailed runtime benchmarks for NANI and HELM processing (PDF).

ACKNOWLEDGEMENTS

RAMQ, LC, and JBWS thank support from the National Institute of General Medical Sciences of the National Institutes of Health under award number R35GM150620.

DATA AND SOFTWARE AVAILABILITY

The NANI code can be found here: https://github.com/mqcomplab/MDANCE.

Conflict of Interest: The authors declare no competing financial interests.

REFERENCES

(1) De Paris, R.; Quevedo, C. V.; Ruiz, D. D.; Norberto De Souza, O.; Barros, R. C. Clustering Molecular Dynamics Trajectories for Optimizing Docking Experiments. *Comput. Intell. Neurosci.* **2015**, *2015*, 1–9. https://doi.org/10.1155/2015/916240.

(2) 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. https://doi.org/10.1021/ct700119m.

(3) Keller, B.; Daura, X.; Van Gunsteren, W. F. Comparing Geometric and Kinetic Cluster Algorithms for Molecular Simulation Data. *J. Chem. Phys.* **2010**, *132* (7), 074110. https://doi.org/10.1063/1.3301140.

(4) Xu, D.; Tian, Y. A Comprehensive Survey of Clustering Algorithms. *Ann. Data Sci.* **2015**, *2* (2), 165–193. https://doi.org/10.1007/s40745-015-0040-1.

(5) Wang, S.; Chang, T.-H.; Cui, Y.; Pang, J.-S. Clustering by Orthogonal NMF Model and Non-Convex Penalty Optimization. *IEEE Trans. Signal Process.* **2021**, *69*, 5273–5288. https://doi.org/10.1109/TSP.2021.3102106.

(6) Arthur, D.; Vassilvitskii, S. K-Means++: The Advantages of Careful Seeding. In *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*; SODA '07; Society for Industrial and Applied Mathematics: USA, 2007; pp 1027–1035.

(7) Chen, L.; Roe, D. R.; Kochert, M.; Simmerling, C.; Miranda-Quintana, R. A. K-Means NANI: An Improved Clustering Algorithm for Molecular Dynamics Simulations. *J. Chem. Theory Comput.* **2024**, *20* (13), 5583–5597. https://doi.org/10.1021/acs.jctc.4c00308.

(8) Chen, L.; Smith, M.; Roe, D. R.; Miranda-Quintana, R. A. Extended Quality (eQual): Radial Threshold Clustering Based on *n* -Ary Similarity. *J. Chem. Inf. Model.* **2025**, *65* (10), 5062–5070. https://doi.org/10.1021/acs.jcim.4c02341.

(9) Miranda-Quintana, R. A.; Bajusz, D.; Rácz, A.; Héberger, K. Extended Similarity Indices: The Benefits of Comparing More than Two Objects Simultaneously. Part 1: Theory and Characteristics†. *J. Cheminformatics* **2021**, *13* (1), 32. https://doi.org/10.1186/s13321-021-00505-3.

(10) Miranda-Quintana, R. A.; Rácz, A.; Bajusz, D.; Héberger, K. Extended Similarity Indices: The Benefits of Comparing More than Two Objects Simultaneously. Part 2: Speed, Consistency, Diversity Selection. *J. Cheminformatics* **2021**, *13* (1), 33. https://doi.org/10.1186/s13321-021-00504-4.

(11) López-Pérez, K.; Kim, T. D.; Miranda-Quintana, R. A. iSIM: Instant Similarity. *Digit. Discov.* **2024**, *3* (6), 1160–1171. https://doi.org/10.1039/D4DD00041B.

(12) Chen, L.; Leung, J. M. G.; Zsigmond, K.; Chong, L. T.; Miranda-Quintana, R. A. SHINE: Deterministic Many-to-Many Clustering of Molecular Pathways. *J. Chem. Inf. Model.* **2025**, *65* (10), 4775–4782. https://doi.org/10.1021/acs.jcim.5c00240.
(13) Chang, L.; Perez, A.; Miranda-Quintana, R. A. Improving the Analysis of Biological Ensembles through Extended Similarity Measures. *Phys. Chem. Chem. Phys.* **2022**, *24* (1), 444–451. https://doi.org/10.1039/D1CP04019G.
(14) Rácz, A.; Mihalovits, L. M.; Bajusz, D.; Héberger, K.; Miranda-Quintana, R. A. Molecular Dynamics Simulations and Diversity Selection by Extended Continuous Similarity Indices. *J. Chem. Inf. Model.* **2022**, *62* (14), 3415–3425. https://doi.org/10.1021/acs.jcim.2c00433.
(15) Rácz, A.; Dunn, T. B.; Bajusz, D.; Kim, T. D.; Miranda-Quintana, R. A.; Héberger, K. Extended Continuous Similarity Indices: Theory and Application for QSAR Descriptor Selection. *J. Comput. Aided Mol. Des.* **2022**, *36* (3), 157–173. https://doi.org/10.1007/s10822-022-00444-7.
(16) Calinski, T.; Harabasz, J. A Dendrite Method for Cluster Analysis. *Commun. Stat. - Theory Methods* **1974**, *3* (1), 1–27. https://doi.org/10.1080/03610927408827101.
(17) Davies, D. L.; Bouldin, D. W. A Cluster Separation Measure. *IEEE Trans. Pattern Anal. Mach. Intell.* **1979**, *PAMI-1* (2), 224–227. https://doi.org/10.1109/TPAMI.1979.4766909.
(18) González-Alemán, R. Graph-Based Approach to the Quality Threshold Clustering of Molecular Dynamics. https://github.com/LQCT/BitQT.
(19) González-Alemán, R.; Hernández-Castillo, D.; Caballero, J.; Montero-Cabrera, L. A. Quality Threshold Clustering of Molecular Dynamics: A Word of Caution. *J. Chem. Inf. Model.* **2020**, *60* (2), 467–472. https://doi.org/10.1021/acs.jcim.9b00558.
(20) Daura, X.; Conchillo-Solé, O. On Quality Thresholds for the Clustering of Molecular Structures. *J. Chem. Inf. Model.* **2022**, *62* (22), 5738–5745. https://doi.org/10.1021/acs.jcim.2c01079.
(21) Piana, S.; Lindorff-Larsen, K.; Shaw, D. E. Protein Folding Kinetics and Thermodynamics from Atomistic Simulation. *Proc. Natl. Acad. Sci.* **2012**, *109* (44), 17845–17850. https://doi.org/10.1073/pnas.1201811109.
(22) Chen, L.; Santos, J.B.W.; Gaza, J.; Perez, A.; Miranda-Quintana, R. A. Hierarchical Extended Linkage Method (HELM)'s Deep Dive into Hybrid Clustering Strategies. *J. Chem. Inf. Model.* **2025**. https://doi.org/10.1021/acs.jcim.5c00539.