mdBIRCH for Fast, Scalable, Online Clustering of Molecular Dynamics Trajectories

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 mdBIRCH, an online clustering method that adapts the BIRCH CF-tree to molecular dynamics (MD) data by using a merge test calibrated directly to RMSD. Each arriving frame is routed to the nearest centroid and added only if the post-merge radius computed from the cluster feature remains within a user-supplied threshold. This keeps the average deviation to each cluster centroid bounded as the cluster grows and preserves a simple interpretation of resolution in physical units. We evaluate mdBIRCH on a β-heptapeptide and the HP35 system. We propose two protocols to make the threshold selection easier: (a) RMSD-anchored runs that use controlled structural edits to define interpretable operating points and (b) blind sweep that tracks how cluster count, occupancy, and coverage change with the threshold. In both systems, increasing the threshold reduces the number of clusters, concentrates coverage in high-occupancy states, and broadens within-cluster RMSD distributions. Furthermore, because decisions rely only on cluster summaries, mdBIRCH completely avoids the need for pairwise distance matrices, scales near-linearly with the number of frames on standard hardware, and naturally supports incremental operation. The method offers a practical combination of speed and interpretability for large-scale trajectory analysis.

1. INTRODUCTION

Clustering remains one of the most practical ways to make sense of large molecular dynamics (MD) trajectories.[1–3] By grouping structurally similar frames, one can pinpoint the dominant states or pick representatives for visualization or refinement.[3,4] However, one of the primary challenges is scalability. Pairwise distance matrices grow quadratically with the number of frames, so classic methods that rely on them such as hierarchical agglomerative clustering (HAC) or affinity propagation become costly in both time and memory as trajectories reach a few hundreds of thousands of frames.[2,5] In practical workflows, this often forces down-sampling (sieving) or other techniques that reduce the number of frames, useful for speed, but risk missing important structures.

In our recent work on Hierarchical Extended Linkage Method (HELM), we showed that one does not need to pay the full $O(N^2)$ price to get hierarchy-like results.[6] In HELM, we group the frames into microclusters and then construct the hierarchy over all subclusters using n-ary calculations. These n-ary similarity functions[7–14] let us evaluate linkage relationships between subclusters directly, without enumerating all pairwise distances, while keeping every frame assigned through its subcluster membership. The result is hierarchy-quality clustering that remains manageable, even at the scale of million-frame trajectories.

There are, however, common scenarios where a different tool would be more useful. First, many projects accumulate data over time: trajectories grow as simulations run, new replicas arrive, or additional systems are added.[15,16] Rebuilding the microclusters and the hierarchy is feasible but would still be a batch step. Second, researchers often want a hard, physically meaningful bound on cluster spread (i.e. keeping everything in a cluster within about a specified distance around its centroid).[17,18] Third, it is appealing in practice to tune clustering with only one interpretable parameter rather than a combination of a number of parameters such as final number of clusters, neighborhood radius, minimum number of points, linkage settings, and the like.

These needs motivate mdBIRCH. We adapt the Balanced Iterative Reducing and Clustering using Hierarchies (BIRCH)[19] framework to MD by using an RMSD-calibrated spread criterion inside the CF tree. In other words, we use BIRCH's streaming, memory bounded insertion of frames, but we decide the merges with a quantity that maps directly to the average distance of all points to the centroid. The result is a method that offers: (a) incremental operation, we can continue adding frames without re-running a global pipeline; (b) per-cluster bounds, every accepted merge preserves the user-set threshold (τ); and (c) a single-parameter control, cluster resolution is specified directly in physical units.

Moreover, mdBIRCH keeps outputs practical for analysis. Each leaf corresponds to a cluster constrained by a τ, so the medoids can be selected and real structures visualized immediately. Also, because decisions are local and rely only on cluster-feature statistics, the method scales near-linearly with the number of frames and never requires a pairwise distance matrix. In effect,

mdBIRCH, provides a physically grounded radius-bounded, streaming clustering of large-scale trajectories that can complement existing workflows.

2. THEORY

Each aligned MD frame is represented as a vector $q \in \mathbb{R}^D$, with D as the number of dimensions/features. Each cluster j is summarized by a Cluster Feature (CF_j):

$$CF_j = \left(N_j, LS_j, SS_j\right) \tag{1}$$

$$LS_j = \sum_{i=1}^{N_j} q_i \tag{2}$$

$$SS_j = \sum_{i=1}^{N_j} q_i \cdot q_i \tag{3}$$

with centroid $c_j = LS_j/N_j$. The radius of a subcluster used in all merge decisions, is the mean squared deviation (MSD) about the centroid,

$$\mathrm{R}^2 = \frac{1}{N_j}\sum_{i=1}^{N}\left\|q_i - c_j\right\|^2 = \frac{SS_j}{N_j} - \frac{\left\|LS_j\right\|^2}{N_j^2}. \tag{4}$$

This quantity is computed exactly from CFs, so decisions reduce to vector additions and dot products, without the need for a pairwise distance matrix. For interpretability, we report this tightness as centroid MSD,

$$\mathrm{MSD} = R^2/D.$$

RMSD calibration. To make the threshold τ literal at cluster birth, we anchor it at two-frame nucleation. For aligned frames x and y with centroid $c = (x + y)/2$,

$$\mathrm{R}^2 = \frac{1}{4}\|x - y\|^2 = \frac{1}{4}D\ \mathrm{RMSD}(x, y)^2. \tag{5}$$

We therefore accept a tentative union with R'^2

$$\mathrm{R}'^2 = \frac{\mathrm{new}_{SS}}{\mathrm{new}_N} - \frac{\|\mathrm{new}_{LS}\|^2}{\mathrm{new}_N^2} \tag{6}$$

whenever the post-merge radius satisfies

$$R'^2 \leq \left(\frac{\tau^2}{4}\right)D \tag{7}$$

which is equivalent to: two frames nucleate a cluster if and only if the pairwise RMSD is at most τ, and the same bound is preserved as clusters grow. Full derivation can be found in Supporting

Information. Because the test is evaluated on post-merge CFs, the bound applies to the entire subcluster with respect to the updated centroid after each accepted insertion. This guarantee controls the average deviation to the centroid, which would be the natural summary for large MD ensembles.

Decision granularity and interpretation. Each arriving frame is routed to the nearest centroid through a CF-tree and tested with a post-merge radius; if the test fails, a new subcluster is created, otherwise the target cluster is updated. Two inputs control behavior: the threshold τ (cluster resolution) and the branching factor or the maximum number of subclusters in a node. No pairwise distance matrix is formed.

As an online method, mdBIRCH is also inherently sensitive to the order of data insertion. However, we made it a deliberate feature and not a limitation, as the method is designed to process trajectories as they are generated. This allows the cluster definitions to evolve with the simulation, in contrast to static batch methods that require a complete trajectory upfront to analyze frames in a global perspective. To the best of our knowledge, every single clustering method used to process MD simulations requires to have access to all of the data to be processed, before starting the clustering. mdBIRCH is unique in that it does not need to "see" all the conformations to start clustering them.

3. SYSTEMS

***β*-Heptapeptide.** Topology and trajectory files were retrieved from a public GitHub repository.[20,21] Frames 1-1000 were discarded as equilibration, and the remainder was aligned to the first frame. Clustering included a backbone-only representation (NB, Cα, C, O, H) restricted to residues Lys2 to Asp11, excluding terminal residues and all side chains, following Daura, et al.[22] In total, 6001 structures were included in the analysis.

HP35 (Nle/Nle mutant). The long-timescale trajectory from D.E. Shaw Research was used (305 μs at 360K; frames saved every 200 ps; ~1.52×10^6 frames).[23] Frames 1-5000 were discarded as equilibration, and the remainder was aligned to its first frame. Clustering used a back-bone heavy-atom definition consistent with previous HP35 studies: N of residue 1; N, Cα, and C of residues 2-34; and N of residue 35. To reduce data volume, every 10th frame was retained, yielding 151,104 total structures for clustering.

4. RESULTS AND DISCUSSION

Selecting a threshold

mdBIRCH exposes a single tolerance in RMSD that governs structural granularity. Because scales differ across systems and atom selections, we use two complementary strategies, applied to the studied systems: an RMSD-anchored operating points obtained from controlled modifications to a reference frame and a blind sweep of the thresholds to visualize how coarseness evolves.

RMSD-anchored from structural modifications

For both systems, we produced small to large, physical perturbations of representative frames (e.g., rotations or translations of motifs), restricted the atom set to match the trajectory alignment, and measured the RMSDs to the original reference (Fig. 1 and Fig. 2). We then ran mdBIRCH with each value as the threshold.

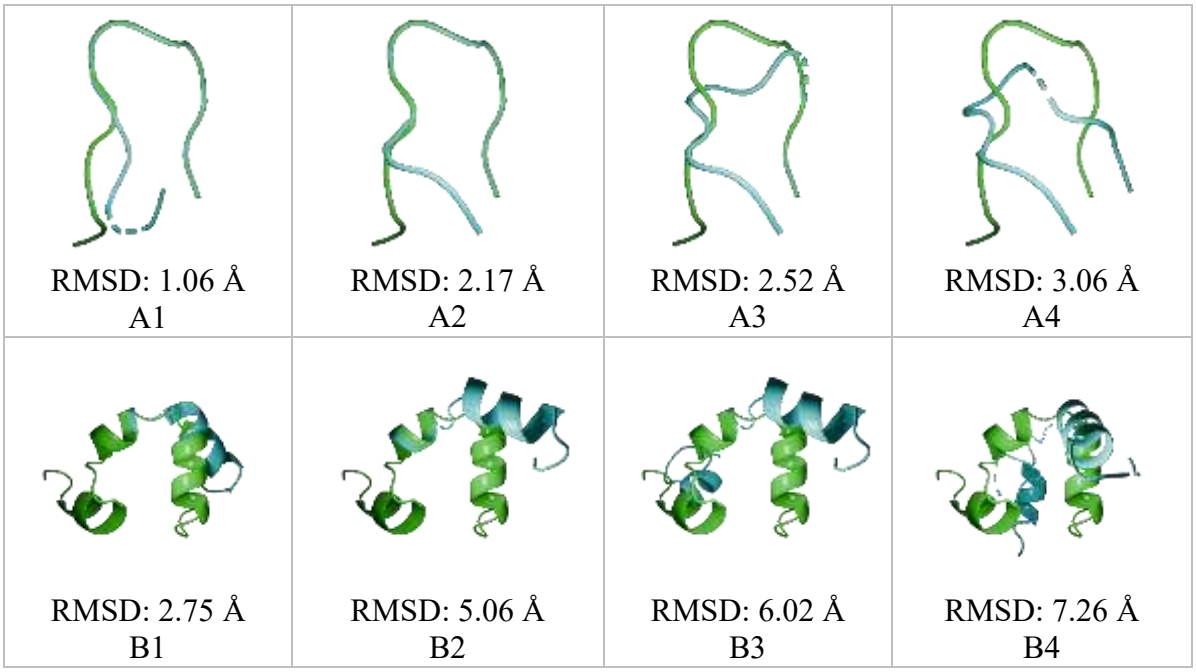


Figure 1. Edited reference overlays for the β-heptapeptide (A1-A4) and the HP35 (B1-B4) systems. Reference (green) versus controlled edits (cyan).

As τ increases, both systems show expected consolidation of states, but on different scales. For the β-heptapeptide (Table 1), raising this τ from 1.06 to 3.06 Å collapses the total number of clusters from 1229 to 4; the number of denser clusters (>10/>50/>100) converge to the same four dominant groups. The fraction of frames in the ten largest clusters also increases monotonically, indicating that at a threshold of ~3 Å, essentially all the frames are captured by a handful of states. For HP35 (Table 2) the absolute cluster counts are larger and the effective range is also shifted upwards, given the dataset has a larger number of frames. When τ grows from 2.75 to 7.26 Å, the total clusters decrease from 31363 to 73, while the top 10 coverage improves from 50.21% to 84.30%. The >50 and >100-frame peak at intermediate thresholds before final convergence to higher τ. Together, this shows that small τ values resolve fine-grained conformations, whereas higher values yield compact, high-occupancy structures.

Table 1. Clustering summary at thresholds from controlled structural edits for the β-heptapeptide system.

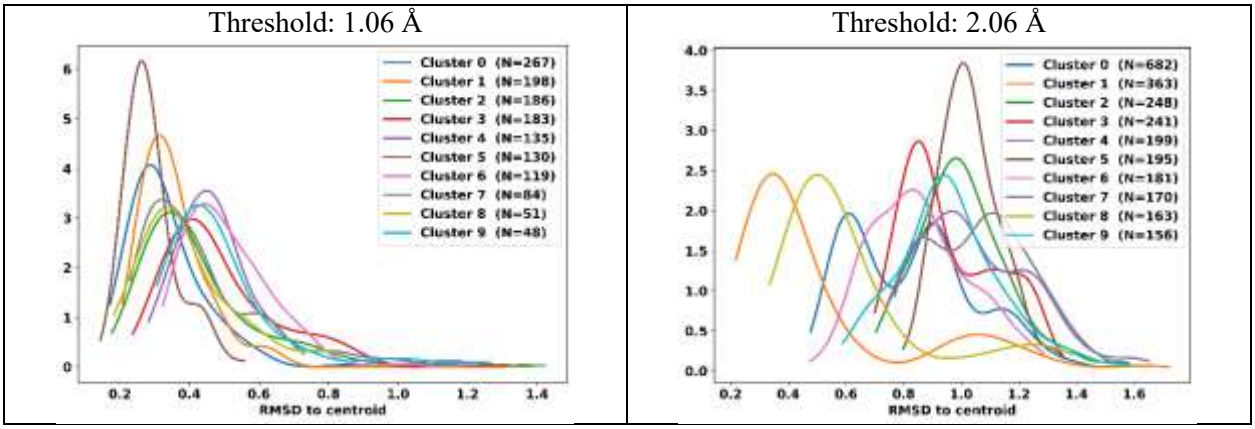
Threshold (τ)	Number of Clusters	Clusters >10 Frames	Clusters >50 Frames	Clusters >100 Frames	% Top 10 Clusters
1.06	1229	80	9	7	23.35
2.17	74	64	42	18	43.29

2.52	19	19	18	16	81.34
3.06	4	4	4	4	100.00

Table 2. Clustering summary at thresholds from controlled structural edits for the HP35 system.

Threshold (τ)	Number of Clusters	Clusters >10 Frames	Clusters >50 Frames	Clusters >100 Frames	% Top 10 Clusters
2.75	31363	383	100	60	50.21
5.06	8635	763	160	63	67.30
6.02	3022	532	200	95	71.30
7.26	73	64	60	51	84.30

Fig. 2 shows the kernel-density estimates of RMSD-to-centroid for the ten largest clusters at various τ for the β-heptapeptide system. As the τ increases, the modes shift right and broaden, consistent with the looser acceptance bound. At τ = 1.06 Å, the modes are tightly concentrated below 0.6 Å with limited tails, reflecting fine-grained states. At intermediate τ (B-C) the modes move toward ~0.4 to 1.2 Å regions and their tails increase in length, indicating consolidation of nearby conformations. At 3.06 Å, we observe a few high-occupancy clusters remaining. Their distributions are broader and partially overlapping, as expected for coarser groups. We also note here that individual RMSD values can exceed τ because mdBIRCH controls the average (post-merge) centroid RMSD per cluster, not the maximum; the observed smooth growth of the modal RMSD with the τ is consistent with that guarantee.



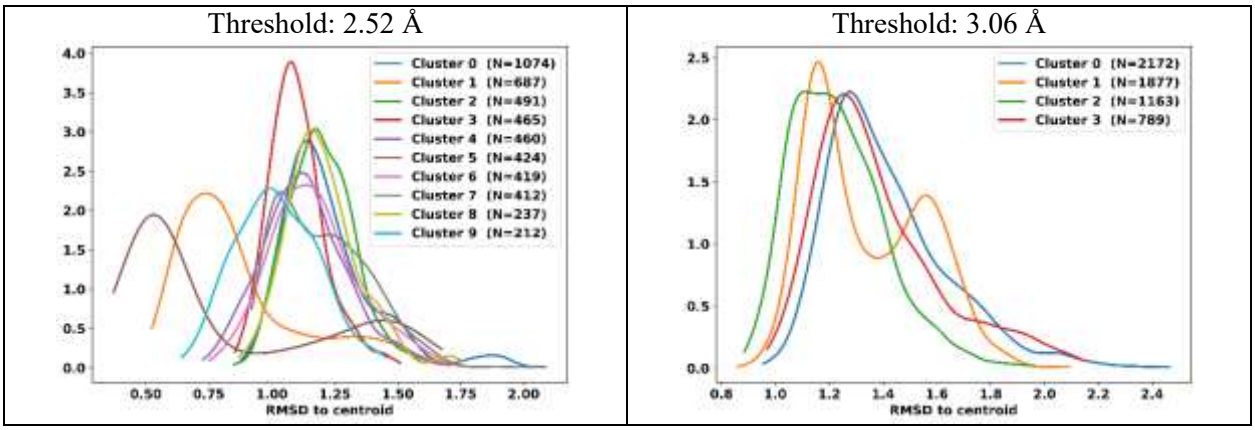


Figure 2. Per-cluster RMSD-to-centroid distributions at controlled-modification thresholds of the β-heptapeptide system. Curves show kernel densities of RMSD to the cluster centroid for the ten largest clusters with the legend reporting cluster sizes (*N*).

For the case of HP35 (Fig. 3), we see the same monotonic pattern as the β-heptapeptide but on a higher scale. The distributions migrate to higher RMSDs and become wider as clusters absorb more conformational variability. At the lowest τ, the clusters have narrower peaks near 1 Å. For the intermediate thresholds, peaks shift to 1 to 2 Å with heavier tails. Lastly, at 7.26 Å, a small number of dominant clusters again display broad, partially overlapping densities.

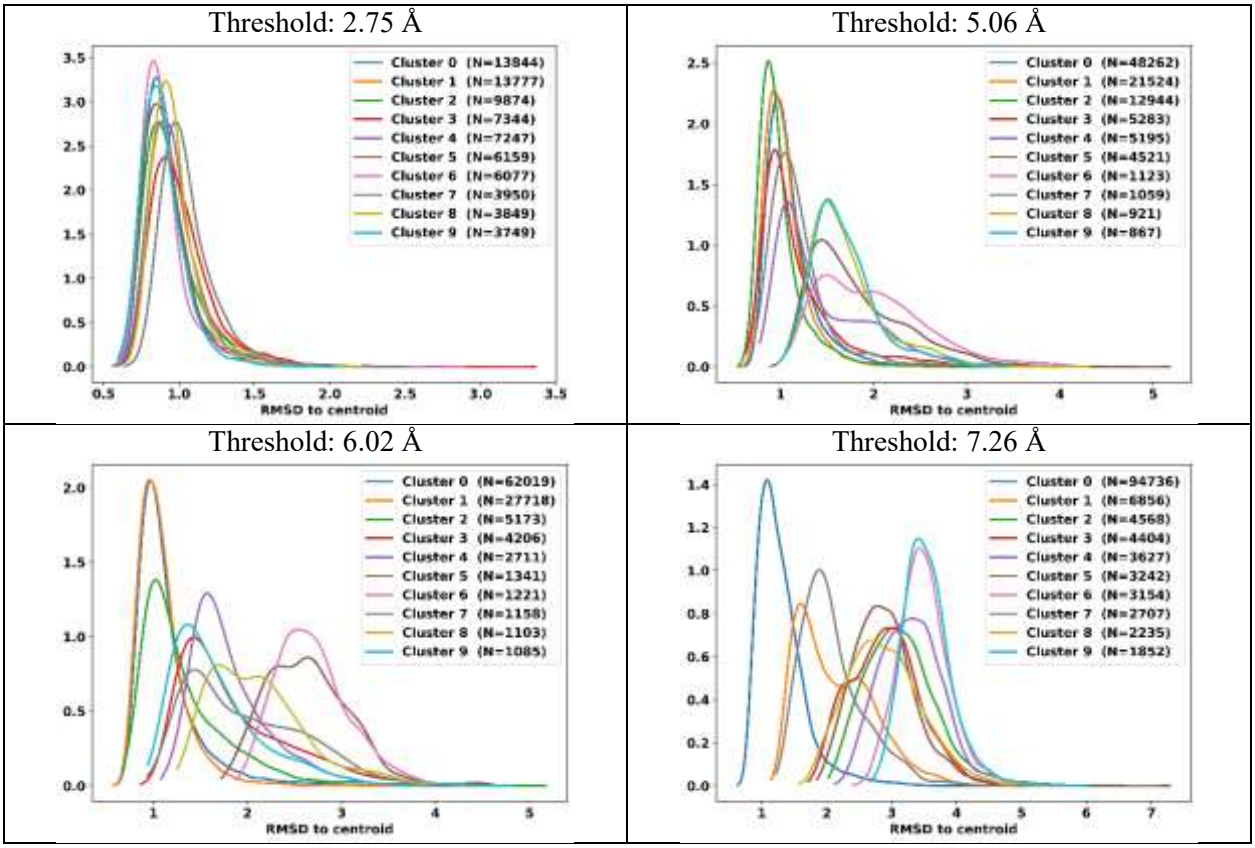


Figure 3. Per-cluster RMSD-to-centroid distributions at controlled-modification of the HP35 System. Curves show kernel densities of RMSD to the cluster centroid for the ten largest clusters with the legend reporting cluster sizes (*N*).

Taken together, we see a consistent, system-specific response to increasing τ. Total cluster counts decrease; well-occupied clusters consolidate, and within-cluster spreads broaden. These establish physically interpretable operating points and validate that the post-merge bound translates into the observed cluster tightness.

Blind threshold sweeps

To complement the RMSD-anchored operating points, we performed blind sweeps over RMSD ranges in fixed increments. At each τ, we again recorded the total number of clusters, clusters with >10/>50/>100 frames, and the fraction of frames in the ten largest clusters. This protocol maps how granularity evolves across the full τ range, without assuming a reference edit.

Across both systems, the blind sweeps exhibit the same behavior as τ increases, total cluster count decreases monotonically, coverage by the largest clusters rises toward unity, and the occupancy filters reveal consolidation of small groups into more well-populated states. The absolute values and the τ scale at which consolidation become pronounced is system dependent: lower for β-heptapeptide and higher for HP35, but the qualitative behavior is the same. Smaller τ yields fine-grained resolution, while higher τ collapses the trajectory to a few broad clusters.

Table 3. Clustering summary at thresholds from blind τ sweep for the β-heptapeptide system.

Threshold	Total Clusters	Clusters >10 Frames	Clusters >50 Frames	Clusters >100 Frames	% Top 10 Clusters
0	6001	0	0	0	0.17
0.5	3433	44	3	0	7.12
1	1404	69	8	6	21.48
1.5	557	153	13	6	25.25
2	163	113	31	14	35.24
2.5	26	25	23	20	68.44
3	5	5	5	5	100.00
3.5	1	1	1	1	100.00
4	1	1	1	1	100.00
4.5	1	1	1	1	100.00
5	1	1	1	1	100.00

Table 4. Clustering summary at thresholds from blind τ sweep for the HP35 system.

Threshold	Total Clusters	Clusters >10 Frames	Clusters >50 Frames	Clusters >100 Frames	% Top 10 Clusters
0	151105	0	0	0	0.01

1	131815	1	0	0	0.05
2	46869	460	152	115	19.20
3	28061	374	101	53	54.60
4	17209	609	92	48	62.64
5	8801	739	162	69	67.39
6	3099	581	195	92	70.56
7	189	124	97	84	76.06
8	12	12	12	12	98.31
9	1	1	1	1	100.00
10	1	1	1	1	100.00

These sweeps provide a system specific map of how granularity changes with τ. The trends make sweeps useful for verifying that a chosen τ behaves sensibly and for identifying a reasonable operating range when no prior anchor is available. But then again, choosing a threshold from a sweep alone can still be ambiguous without system context, and in online/streaming use, we often do not know a good initial τ range at all. In practice, we therefore use RMSD-anchored runs as the primary way to set the threshold, and (optionally) employ sweeps to confirm and fine-tune that choice.

Time and Scaling

Moreover, we measured wall-time as a function of the number of frames in a single pass (CPU only; branching factor = 50, τ = 8). As shown in Fig. 4, the runtime grows smoothly and close to linearly up to 300k frames, with some variance across repeats. This behavior is consistent with the mdBIRCH implementation; each insertion is routed to the nearest centroid through a CF-tree and performs a single post-merge check, so the number of centroid comparisons is bound by the branching factor and no single pairwise distance calculation is made. In practical terms, mdBIRCH processes hundreds of thousands of frames on a single CPU core in seconds. We want to highlight that even this impressive performance is the worst case scenario for mdBIRCH. That is, in this benchmark we illustrated how mdBIRCH would scale if the data is processed after the whole MD simulation has been run. In a more practical scenario, mdBIRCH will be coupled to the MD engine, in a way that every new frame generated will pass through the tree before the next MD time-step has been completed. This makes mdBIRCH effectively instantaneous, able to provide a full analysis of the trajectory immediately after the last conformation has been produced.

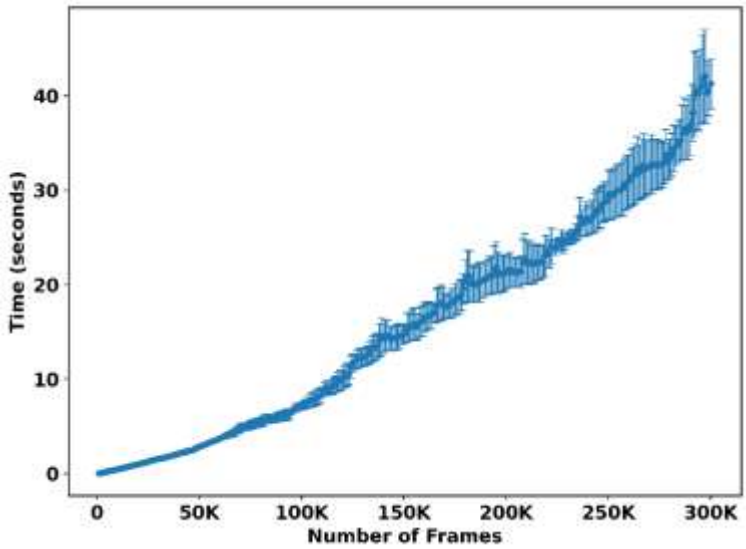


Figure 4. Wall time versus the number of frames with the HP35 system.

5. CONCLUSIONS

We set out to keep online clustering of trajectories simple and physically interpretable. mdBIRCH achieves this by combining a BIRCH CF-tree with a merge decision calibrated in RMSD. For each new insertion we evaluate the candidate cluster after the hypothetical inclusion of the new frame; if the merge satisfies the bound implied by the threshold, the frame is accepted, and the CFs are updated. This provides a clear guarantee that as the clusters evolve, the average distance to the centroid of each cluster remains within the chosen tolerance. This holds throughout the single pass and aligns with the summary that is useful for large ensembles.

On both the β-heptapeptide and HP35 systems, the threshold behaves as an intuitive control for granularity. Using RMSD-anchored operating points from controlled structural edits and blind threshold sweeps, we showed consistent trends: increasing threshold reduces the number of clusters, concentrates coverage in the largest states, and broadens within-cluster RMSD distributions. The anchored runs provide interpretable choices while the sweeps map how granularity evolves over threshold ranges. Because decisions rely on post-merge checks, mdBIRCH is memory-bounded and fast, enabling dense sweeps and routine checks without the need for expensive calculations.

mdBIRCH blends the fast partitioning of *k*-means-like approaches with the ability to recognize noise of density-based methods, and the possibility of dissecting the hierarchical structure of the data. Remarkably, all of this can be achieved at the lowest time and memory cost, with essentially no overhead while the simulation is running. With no need to discard precious MD frames, and the possibility of having the full clustering results immediately after the simulation concluded, mdBIRCH offers key advantages over more traditional methods.

ACKNOWLEDGEMENTS

We 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 mdBIRCH 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. *Computational Intelligence and Neuroscience* **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. *The Journal of Chemical Physics* **2010**, *132* (7), 074110. https://doi.org/10.1063/1.3301140.

(4) Knaggs, M. H.; Salsbury, F. R.; Edgell, M. H.; Fetrow, J. S. Insights into Correlated Motions and Long-Range Interactions in CheY Derived from Molecular Dynamics Simulations. *Biophysical Journal* **2007**, *92* (6), 2062–2079. https://doi.org/10.1529/biophysj.106.081950.

(5) Sittel, F.; Stock, G. Perspective: Identification of Collective Variables and Metastable States of Protein Dynamics. *The Journal of Chemical Physics* **2018**, *149* (15), 150901. https://doi.org/10.1063/1.5049637.

(6) Chen, L.; Brylle Woody Santos, J.; 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.

(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) Pérez, K. L.; Jung, V.; Chen, L.; Huddleston, K.; Miranda-Quintana, R. A. BitBIRCH: Efficient Clustering of Large Molecular Libraries. *Digital Discovery* **2025**, *4* (4), 1042–1051. https://doi.org/10.1039/D5DD00030K.

(9) Chen, L.; Leung, J. M. G.; Zsigmond, K.; Chong, L. T.; Miranda-Quintana, R. A. SHINE: Deterministic Many-to-Many Clustering of Molecular Pathways. Biophysics February 8, 2025. https://doi.org/10.1101/2025.02.07.636541.

(10) Chen, L.; Mondal, A.; Perez, A.; Miranda-Quintana, R. A. Protein Retrieval via Integrative Molecular Ensembles (PRIME) through Extended Similarity Indices. *J. Chem. Theory Comput.* **2024**, *20* (14), 6303–6315. https://doi.org/10.1021/acs.jctc.4c00362.

(11) 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.
(12) 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 Cheminform* **2021**, *13* (1), 32. https://doi.org/10.1186/s13321-021-00505-3.
(13) 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 Cheminform* **2021**, *13* (1), 33. https://doi.org/10.1186/s13321-021-00504-4.
(14) 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.
(15) Yu, I.; Feig, M.; Sugita, Y. High-Performance Data Analysis on the Big Trajectory Data of Cellular Scale All-Atom Molecular Dynamics Simulations. *J. Phys.: Conf. Ser.* **2018**, *1036*, 012009. https://doi.org/10.1088/1742-6596/1036/1/012009.
(16) Beltrán, D.; Hospital, A.; Gelpí, J. L.; Orozco, M. A New Paradigm for Molecular Dynamics Databases: The COVID-19 Database, the Legacy of a Titanic Community Effort. *Nucleic Acids Research* **2024**, *52* (D1), D393–D403. https://doi.org/10.1093/nar/gkad991.
(17) Deng, D. DBSCAN Clustering Algorithm Based on Density. In *2020 7th International Forum on Electrical Engineering and Automation (IFEEA)*; IEEE: Hefei, China, 2020; pp 949–953. https://doi.org/10.1109/IFEEA51475.2020.00199.
(18) Christen, M.; Hünenberger, P. H.; Bakowies, D.; Baron, R.; Bürgi, R.; Geerke, D. P.; Heinz, T. N.; Kastenholz, M. A.; Kräutler, V.; Oostenbrink, C.; Peter, C.; Trzesniak, D.; Van Gunsteren, W. F. The GROMOS Software for Biomolecular Simulation: GROMOS05. *J Comput Chem* **2005**, *26* (16), 1719–1751. https://doi.org/10.1002/jcc.20303.
(19) Zhang, T.; Ramakrishnan, R.; Livny, M. BIRCH: An Efficient Data Clustering Method for Very Large Databases. *SIGMOD Rec.* **1996**, *25* (2), 103–114. https://doi.org/10.1145/235968.233324.
(20) González-Alemán, R. Graph-Based Approach to the Quality Threshold Clustering of Molecular Dynamics. https://github.com/LQCT/BitQT.
(21) 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.
(22) 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.
(23) Piana, S.; Lindorff-Larsen, K.; Shaw, D. E. Protein Folding Kinetics and Thermodynamics from Atomistic Simulation. *Proc. Natl. Acad. Sci. U.S.A.* **2012**, *109* (44), 17845–17850. https://doi.org/10.1073/pnas.1201811109.