

clusterMD

Automated multidimensional clustering of molecular dynamics trajectories using HDBSCAN

Anastasios Melas • Nicholas M. Glykos

Department of Molecular Biology and Genetics · Democritus University of Thrace
MSc in Applied Bioinformatics and Data Analysis



Background and objective

Molecular dynamics simulations generate extensive conformational datasets that require robust methods to connect statistical patterns with molecular structures.

clusterMD automates density-based clustering of multidimensional Principal Component Analysis (PCA) and dihedral PCA (dPCA) projections.

The workflow applies HDBSCAN within the retained projection space, linking density-defined groups to population-derived free-energy visualizations, chronological summaries, and structural ensembles.

Analysis workflow

The **automated pipeline** integrates five core analytical stages: projection ingestion, density-based clustering, population and observed-label-change analysis, structural characterization, and visualization. Upstream **PCA or dPCA coordinates** are clustered via **HDBSCAN** while preserving the original trajectory-frame identifiers. Structural extraction, fitting, averaging, and optional downstream RMSD filtering are executed through the **external *carma*** program without altering the initial clustering assignments.

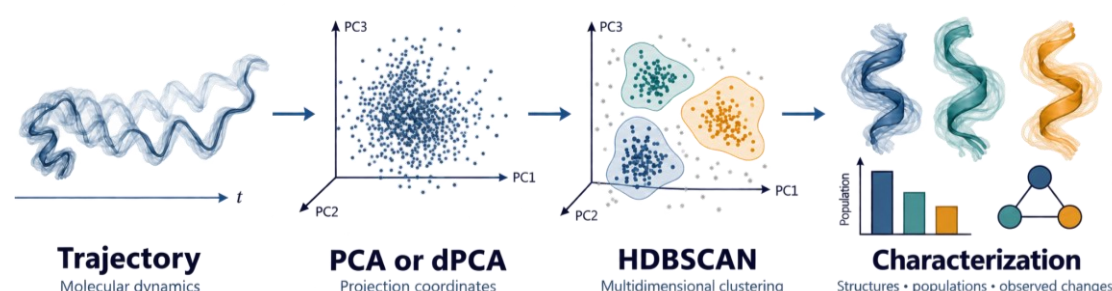
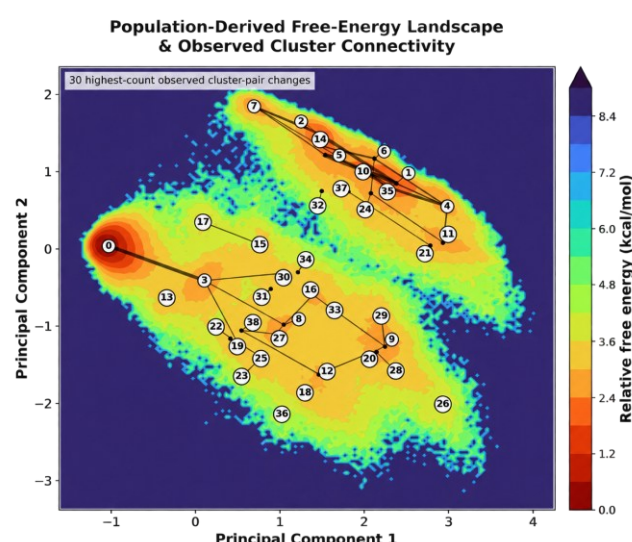


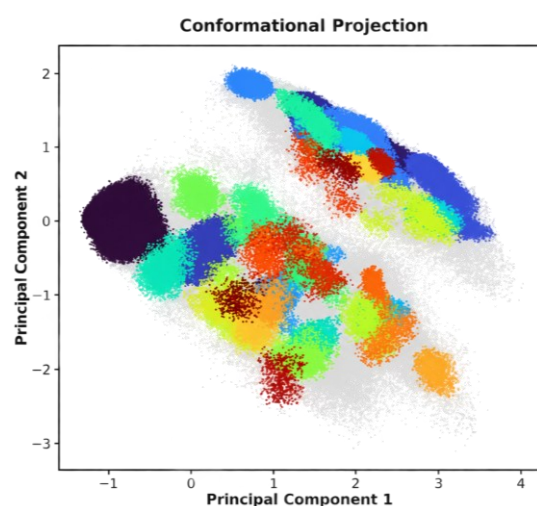
Figure 1. Automated analysis links projection coordinates with population summaries, molecular structures and observed label changes.

Free energy projection



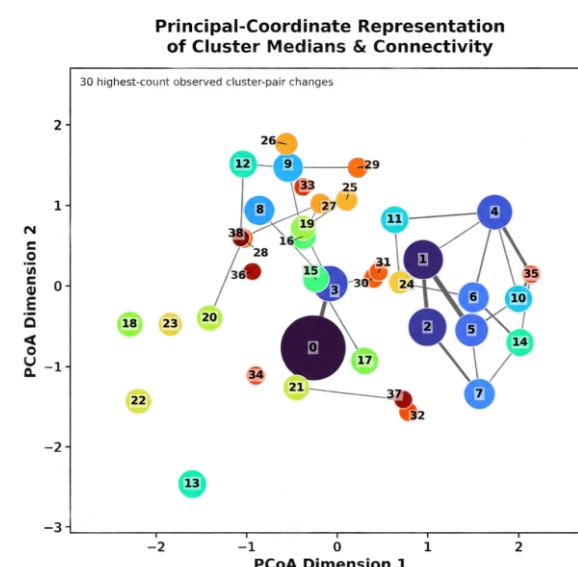
clusterMD converts sampled configuration densities into **population-derived relative free-energy estimates**. The two-dimensional landscape helps locate frequently sampled regions and examine their relationship to cluster assignments. Overlaid connections summarize observed label changes. This projection supports exploratory interpretation but does not establish kinetic barriers or fully describe separation in higher dimensions.

Multidimensional assignments



HDBSCAN identifies **density-defined groups** within the retained multidimensional projection space without requiring a predefined cluster count. Colors map these assignments onto a two-dimensional plot, where apparently overlapping groups may remain distinguishable through additional components. **clusterMD** reports each group's population and retains unassigned configurations as noise rather than forcing every frame into a cluster.

Observed label changes



The **network** represents cluster medians through **principal-coordinate analysis**, with node sizes indicating populations. Edges summarize undirected changes between assigned labels along the analyzed sequence. Because noise frames are omitted from counting, connections may span intervening unassigned configurations. Frame-level diagnostic inspection supports exploration of these events; the counts describe observed connectivity rather than kinetic rates.

Structural extraction and RMSD filtering

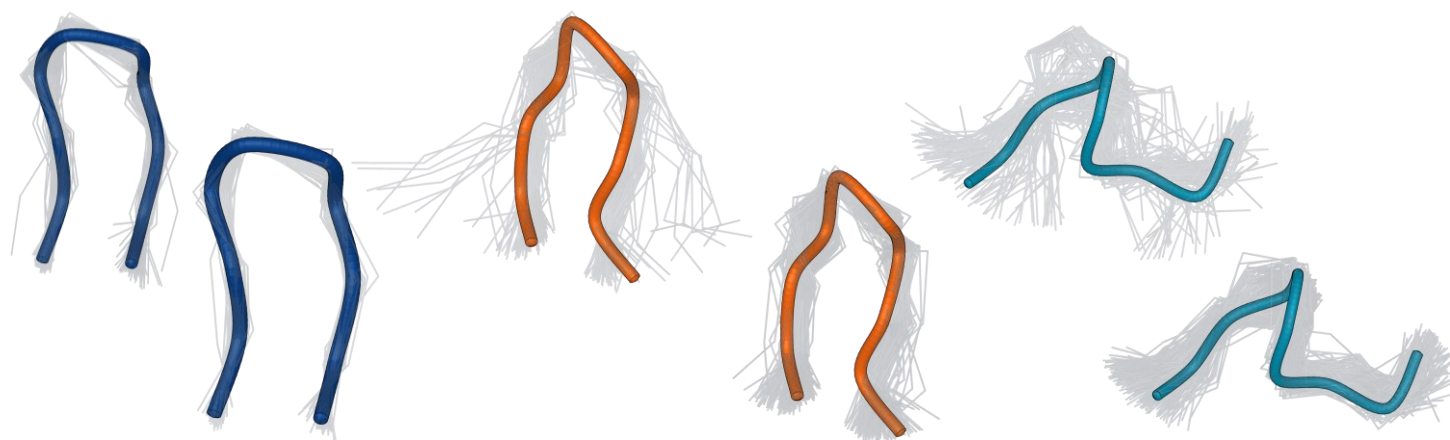


Figure 5. CLN025 clusters 0, 3 and 19 before and after RMSD filtering, with gray ensemble traces and colored representative frames.

clusterMD maps cluster assignments back to trajectory frames to **extract molecular ensembles and representative structures**. Gray backbone traces illustrate variation within each displayed group, while colored tubes highlight representative frames. Optional downstream **RMSD filtering** selects structural subsets relative to fitted ensemble averages, allowing comparison of the original and filtered ensembles while preserving the original HDBSCAN assignments and populations.

Conclusions

clusterMD minimizes reliance on two-dimensional visualizations by clustering directly in the retained multidimensional space. The interpretation of outputs remains strictly dependent on the underlying trajectory sampling quality, the selected coordinate representation, and the applied HDBSCAN parameters. Observed label changes function as descriptive summaries rather than absolute kinetic measurements; reliable kinetic interpretations require independent validation and physical time analysis.

Application to a larger trajectory

CDP-1 is a 14-residue antimicrobial peptide with approximately 7.5 million trajectory frames.

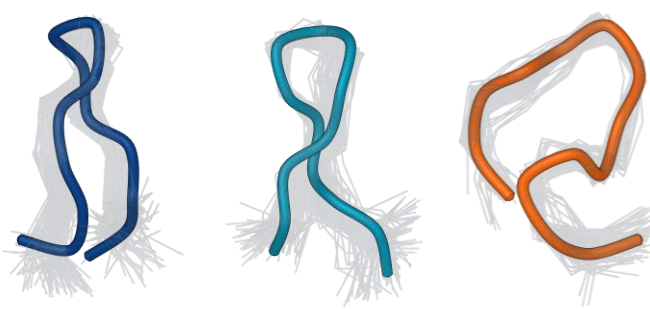


Figure 6. Application to **CDP-1** demonstrates the same characterization workflow on another molecular dynamics trajectory. Extracted ensembles and highlighted representatives connect groups in projection space with their molecular conformations. These outputs allow users to inspect structural differences between clusters and variation within them, illustrating how **clusterMD** supports consistent analysis across molecular systems.

Selected references

- Altis et al. (2008). J Chem Phys 128, 245102.
- Campello et al. (2013). PAKDD, 160–172.
- Glykos (2006). J Comput Chem 27, 1765–1768.
- Koukos and Glykos (2013). J Comput Chem 34, 2310–2312.
- McInnes et al. (2017). J Open Source Softw 2, 205.