Clustermol

We aim to investigate and compare the performance of the various clustering algorithms to identify which are effective in clustering trajectories of flexible molecules, specifically, carbohydrate molecules. In addition, we investigate the use of dimensionality reduction techniques on the trajectory data and examine whether this improves clustering results.

An effective clustering algorithm is one that generates high-quality clusters with good partitioning. This means that the clusters have high intra-cluster similarity and low inter-cluster similarity. Intra-cluster similarity indicates how well a cluster is formed and how similar the objects within a specific cluster are. The inter-cluster similarity is used to determine how dissimilar different clusters are from one another. This is also referred to as cluster separation.

We aim to address the following research question: which clustering algorithms are most effective in clustering MD simulation data of highly flexible molecules? Each member has also outlined their individual research questions which pertain specifically to their contribution.

Nicholas Limbert - HIERARCHICAL | QT

- Which hierarchical clustering method of Ward, maximum, single and average linkage produces the most effective clustering results of MD trajectory data?
- Is the Quality Threshold Algorithm able to effectively cluster MD simulation data from highly flexible molecules?

Wen Kang Lu - HDBSCAN | UMAP

- Is HDBSCAN suitable for clustering MD trajectories?
- Does the use of a dimensionality reduction technique, UMAP, improve HDBSCAN clustering performance?

Robyn McKenzie - HDBSCAN | IMWK-MEANS

- Given that HDBSCAN and iMWK-Means are both able to handle noise in the data, do these algorithms perform well when clustering trajectories of flexible molecules?
- Is the one algorithm more suitable for MD data than the other?

Please refer to the poster, website, or one (or all!) of our papers for conclusions.