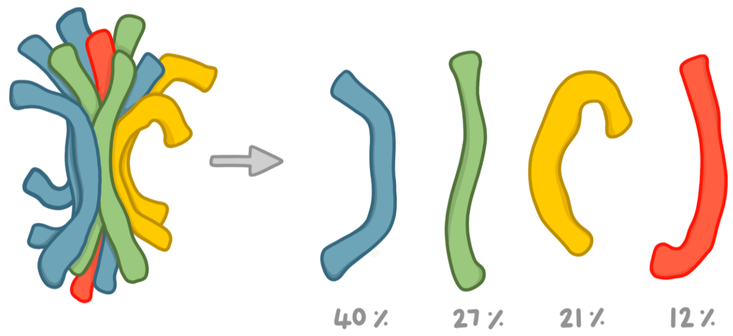
CLUSTERMOL

Background

Molecular Dynamics (MD) simulations of carbohydrate molecules output trajectories which contain details of the movement of the molecule over time. These trajectories can be used to expose the dominant conformations of molecules of interest – but how does one analyze them when a single trajectory can contain upwards of tens of thousands of snapshots in time?

Objective

Clustering is a machine learning technique that partitions the data into distinct groups – perfect for separating conformations in MD trajectories! Our objective was to compare different clustering algorithms and methods to find those that can cluster flexible carbohydrate molecules which may have many conformations.



Finding conformations
By using clustering algorithms, we can find and separate unique conformations!

What We Did

We developed a framework that allowed for automated clustering jobs and explored the following algorithms:

Dimensionality Reduction

UMAP
An algorithm that projects given data onto a lower dimensional manifold.

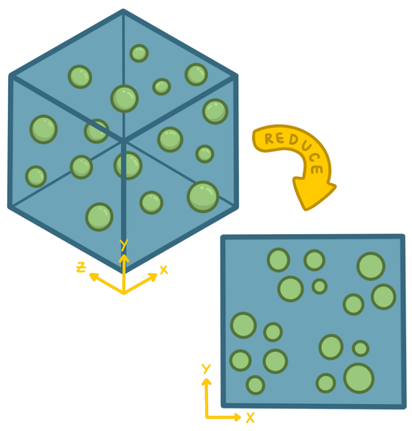
Clustering Algorithms

Hierarchical
An algorithm that generates clusters by iteratively merging smaller clusters.

Quality Threshold
An algorithm that finds clusters with a guaranteed quality specified by the user.

HDBSCAN
A density-based algorithm which can detect and classify outliers and noise.

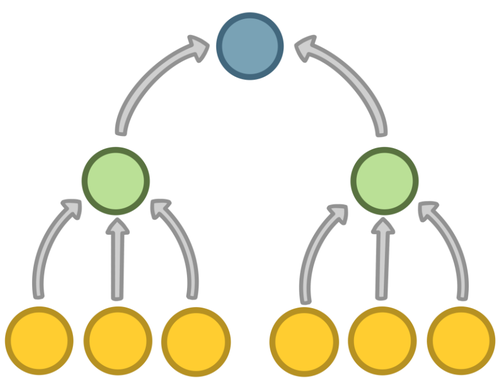
iMWK-Means
A deterministic k-means variant which does not require a cluster count.



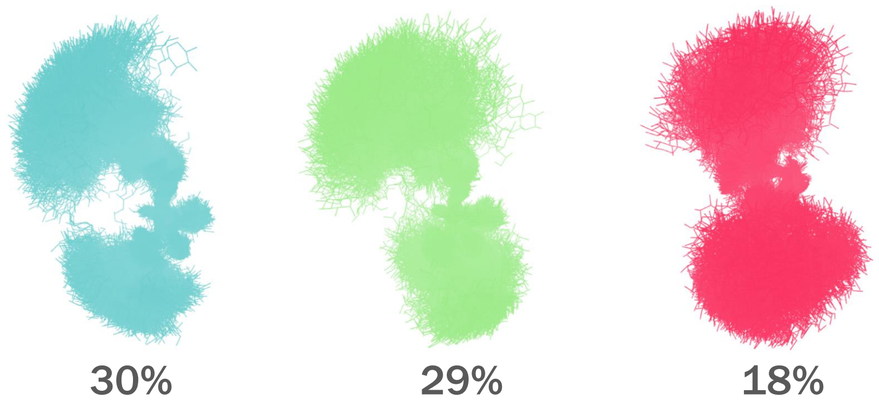
Dimensionality Reduction
UMAP can reduce the dimensionality of MD trajectories. This makes them easier to work with, especially for algorithms that struggle with high dimensional data!

What We Found

- The Quality Threshold algorithm can cluster highly flexible molecules. This further attests to its existing popularity within the domain of clustering MD data.
- Hierarchical clustering was unable to produce reasonable clusters due to its inability to classify noise separately.
- HDBSCAN can produce well-formed clusters of flexible molecules. However, unlike iMWK-Means, it is unable to detect subtly different conformations in stable molecules.
- The use of UMAP as a preprocessing step allows HDBSCAN to better cluster high-dimensional data, and therefore MD trajectories.



Clustering Paradigms
Different clustering algorithms follow different paradigms. The Hierarchical algorithm gets its name from the hierarchical tree it produces to find clusters!



Meningoccocal W
After clustering, we render superimposed frames with VMD to visualize the results! These are the three most dominant conformations of meningococcal W found by UMAP + HDBSCAN.

Nicholas Limbert
niclimbert@gmail.com

Team
Wen Kang Lu
wklu99@gmail.com

Robyn McKenzie
robynkate@gmail.com

University of Cape Town • iYunivesithi yaseKapa • Universiteit van Kaapstad

Department of Computer Science

Supervisor
Associate Professor
Michelle Kuttel
michelle.kuttel@uct.ac.za

Illustrations by Louzanne Swart
louzanneswart@gmail.com