Automated Clustering of Small Molecules

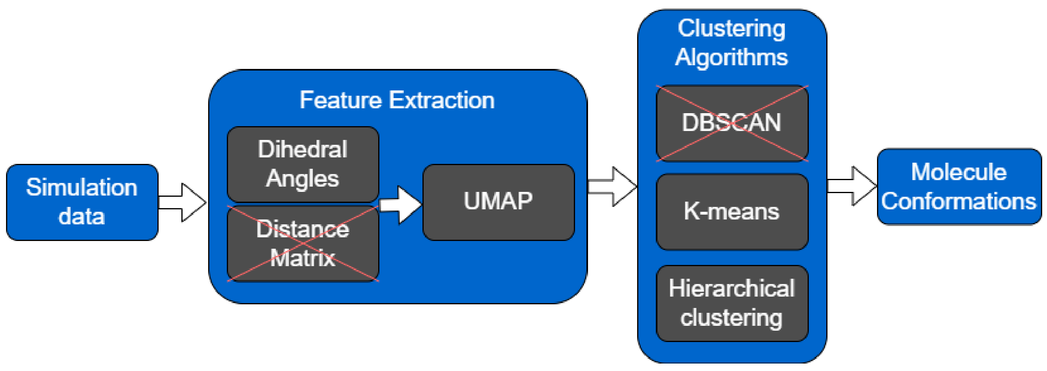
AISA Study Project

Universität Stuttgart
Artificial Intelligence Software Academy

Supervisor: Christian Pfaendner - PhD Researcher
christian.pfaendner@simtech.uni-stuttgart.de

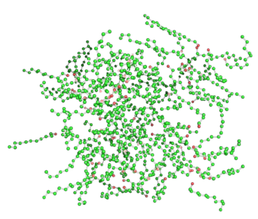
Mareike Harm
Tareq Abu El Komboz
Elena Holzäpfel

Workflow

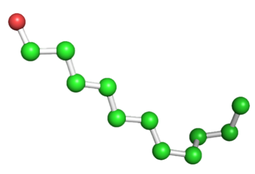


Motivation: The goal is to automatically determine distinct main molecular conformations (3D arrangements of atoms) representing a significant fraction of the population.

Data Exploration: We use data from an atomic simulation of 104 undecanol molecules with 6000 time steps.



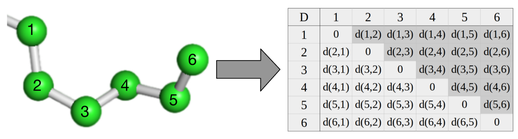
All molecules at the first timestep



One example molecule zoomed in

Feature Extraction is necessary to convert the molecular structures into values suitable for clustering algorithms. For finding different main conformations, only the non-hydrogen atoms are relevant.

Distance Matrix

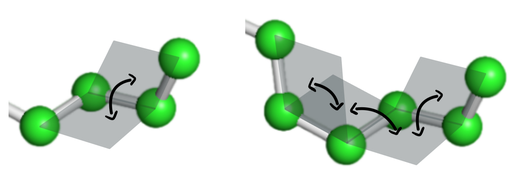


D	1	2	3	4	5	6
1	0	d(1,2)	d(1,3)	d(1,4)	d(1,5)	d(1,6)
2	d(2,1)	0	d(2,3)	d(2,4)	d(2,5)	d(2,6)
3	d(3,1)	d(3,2)	0	d(3,4)	d(3,5)	d(3,6)
4	d(4,1)	d(4,2)	d(4,3)	0	d(4,5)	d(4,6)
5	d(5,1)	d(5,2)	d(5,3)	d(5,4)	0	d(5,6)
6	d(6,1)	d(6,2)	d(6,3)	d(6,4)	d(6,5)	0

Distance matrix for 6 atoms

Challenge: Two different molecules that are symmetric result in the same distance matrix.

Dihedral Angles



Dihedral angles formed by chains of 4 atoms each

Challenge: The distance measure for dihedral angles must be circular. *Solution:* Define custom metrics, based on common metrics.

Uniform Manifold Approximation and Projection (UMAP) is a non-linear dimensionality reduction algorithm that helps to visualize high-dimensional data in a lower-dimensional space. Additionally, it can mitigate the curse of dimensionality in clustering.

Clustering Algorithms

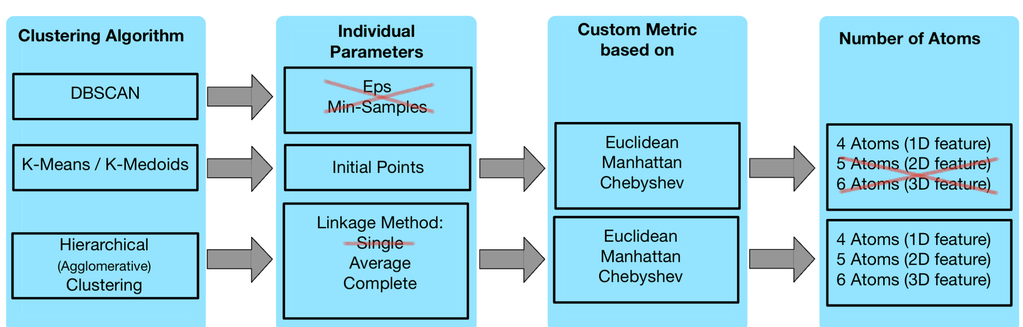
DBSCAN	K-means	Hierarchical (Agglomerative) Clustering
Identifies dense core samples and expands clusters from them	Minimizes within-cluster sum of squares	Clusters are recursively merged until all objects belong to one cluster
Limitations: Clusters with varying density	Limitations: Clusters of different sizes	Comparatively high run time
Number of clusters is automatically determined	Number of clusters must be known in advance	Distance threshold as argument possible instead of cluster number

https://scikit-learn.org/stable/_images/sphx_glr_plot_cluster_comparison_001.png

Application to our dataset

DBSCAN is very sensitive to the choice of parameters. Selecting the right parameters for our dataset proves to be challenging since the clusters vary drastically in density.

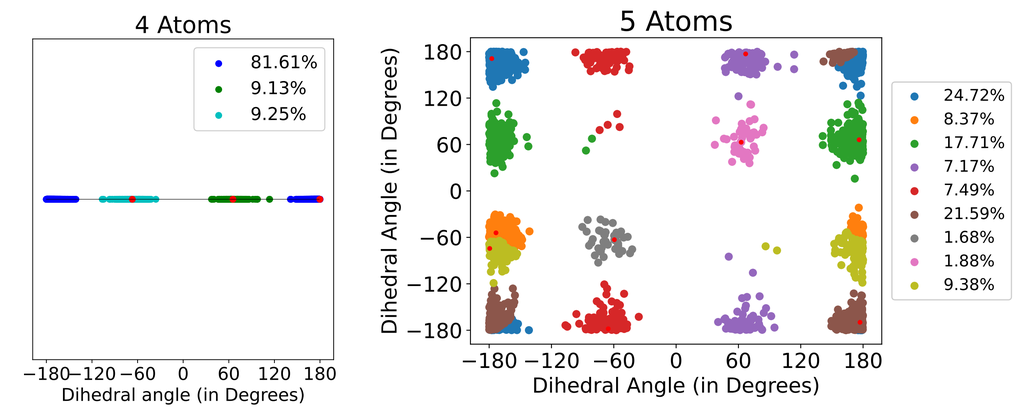
K-medoids is conceptually similar to K-means with the main difference being that the centroids are the median data point instead of the mean.



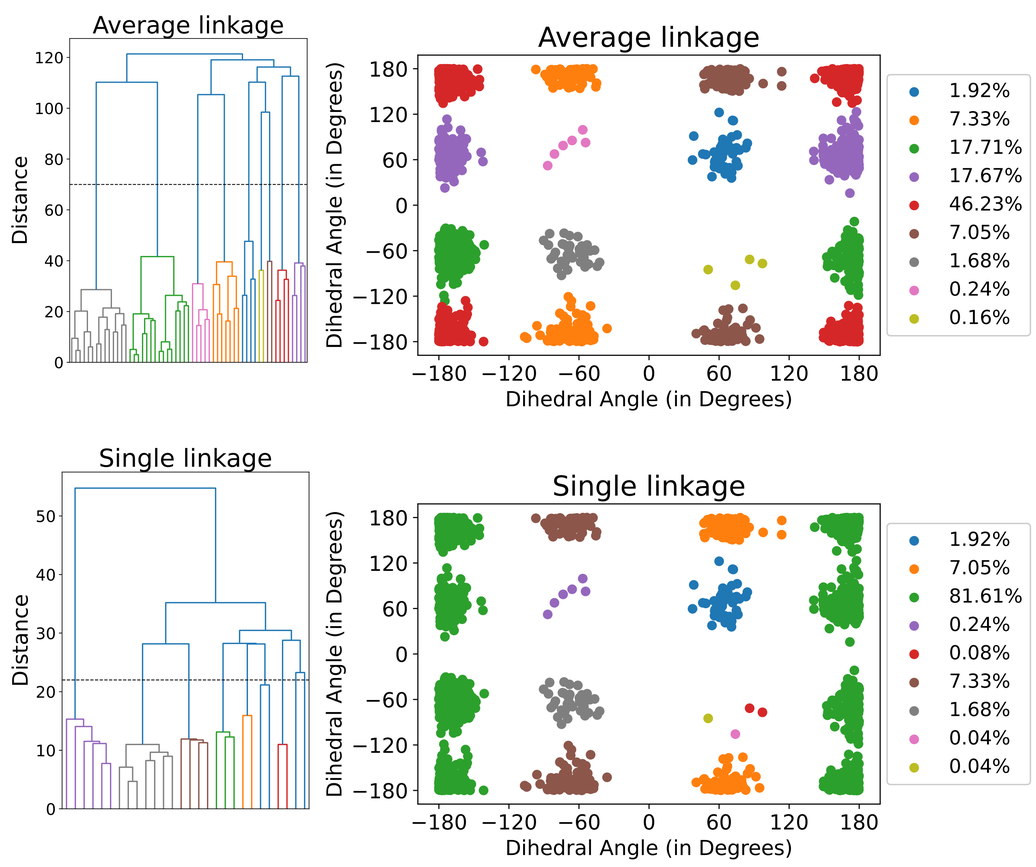
Clustering Results

The following clustering results were performed with 2496 datapoints (24 time steps) and a modified Chebyshev distance. Either the first four or five non-hydrogen atoms were used, yielding 1D or 2D datapoints.

K-medoids



Hierarchical Clustering



Conclusion

- Hierarchical Clustering in combination with average and complete linkage methods yields the best clustering results
- K-Medoids does not perform as well in higher dimensions due to strongly varying cluster sizes
- DBSCAN is not suitable for automating our task, as the clusters vary widely in density
- Distance-matrix is not useful as descriptor, because two different conformations can lead to the same representation

Future Work

- Automate determination of the most suitable number of cluster via: Reachability Graphs, Gaussian Mixture Model
- Evaluate proposed clustering algorithms on different molecules and with higher number of atoms (UMAP might be effective here)

SimTech - AISA