Executive Summary — K-Means Application on AWS COVID-19 Molecular Structure & Therapeutics Hub

Date: October 02, 2025

# Overview

This project implements a K-Means clustering application in Python (developed in PyCharm) to group molecular candidates from the COVID-19 Molecular Structure and Therapeutics Hub (MolSSI) on the Registry of Open Data on AWS. The goal was to reveal structure in basic physicochemical/docking-like features that may correspond to chemotypes or binding behavior.

# Dataset & Features

A representative CSV containing ligand descriptors was used. Key numeric features included predicted affinity (kcal/mol), molecular weight (MW), logP, topological polar surface area (TPSA), hydrogen bond donors/acceptors (HBD/HBA), and rotatable bonds. A target column (e.g., Mpro, Spike RBD) was retained for reference but excluded from clustering.

# Method

The pipeline auto-detects numeric columns, standardizes them, and fits K-Means for k in [2..8]. The best k is selected using silhouette score, and we provide a PCA 2D visualization of cluster separation. Outputs include clustered CSV, per-cluster summary means, and diagnostic plots.

# Results & Insights

The best number of clusters on the demonstration run was k=3 (silhouette=0.121). Clusters showed distinct profiles across affinity and physicochemical space. For example, one group contained lower-MW, lower-TPSA ligands with slightly weaker predicted affinity; another grouped higher-MW ligands with higher HBA and rotatable bonds.

# Challenges & Mitigations

• Data volume and heterogeneity: The MolSSI hub contains large datasets; we mitigated by selecting a compact subset and focusing on core descriptors.  
• Feature selection: We avoided bias by starting with general physicochemical features and scaling all numeric columns.  
• Cluster interpretability: We added per-cluster summary tables and a PCA plot for stakeholder-friendly visuals.

# Conclusion

The K-Means application provides a quick, reproducible way to explore molecular space within the COVID-19 Hub. With modest tuning and expanded descriptors (e.g., fingerprints), the workflow can help guide chemotype identification and prioritization for downstream screening.

# How to Reproduce (PyCharm)

1) Install dependencies from requirements.txt in a PyCharm virtual environment.  
2) Download a CSV from the MolSSI Hub (Registry of Open Data on AWS) with at least 2 numeric descriptors.  
3) Run: python kmeans\_app.py --csv yourfile.csv --kmin 2 --kmax 8 --outprefix results  
4) Review: results\_clusters.csv, results\_summary.csv, results\_silhouette.png, results\_pca2\_scatter.png

# References (APA-style)

Pedregosa, F., Varoquaux, G., Gramfort, A., et al. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825–2830.  
MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, 1, 281–297.  
AWS Registry of Open Data. COVID-19 Molecular Structure and Therapeutics Hub (MolSSI).