Small Molecule Clustering and Analysis

Presented by

TEAM – HEHE

MEMBERS
G SRI CHARAN TEJA REDDY
P V SATYA VENKATESH

1. Introduction

1.1 Background:

The project focused on exploring a benchmarking dataset derived from the ZINC Clean Leads collection, containing 4,591,276 small molecules. The dataset underwent extensive refinement to ensure the inclusion of diverse yet chemically relevant molecules.

1.2 Objectives:

Apply clustering techniques to identify meaningful groupings within the small molecule dataset.
Assess the quality and interpretability of clusters using visualization and silhouette analysis.
Explore chemical diversity within and between clusters.
Lay the foundation for potential applications in drug discovery, compound library design, and chemical informatics.

2. Data Overview:

2.1 Dataset Refinement:

The initial dataset underwent rigorous filtering based on molecular weight, rotatable bonds, XlogP, and medicinal chemistry filters. The resulting dataset comprised 1,936,962 molecules, forming the basis for subsequent analyses.

2.2 Preprocessing:

Molecular structures were represented in SMILES notation and processed using the RDKit library. Morgan fingerprints were employed as molecular descriptors, and PCA was applied to reduce dimensionality while preserving key features.

3. Clustering Analyses:

3.1 K-Means Clustering:

3.1.1 Number of Clusters: 5

3.1.2 Visualization:

A 2D scatter plot visualized the distribution of molecules, with distinct clusters identified. Centroids, representing cluster centers, were marked in red.

3.1.3 Silhouette Analysis:

An average silhouette score of [insert score] indicated [high/moderate/low] quality clustering. Silhouette plots provided insights into the cohesion and separation of clusters.

3.2 Hierarchical Clustering:

3.2.1 Number of Clusters: 5

3.2.2 Visualization:

A 2D scatter plot illustrated the hierarchical clustering results, offering an alternative perspective on molecular relationships.

3.3 Chemical Diversity Analysis:

Molecular fingerprints facilitated a chemical diversity analysis. The diversity within each cluster and across all clusters was assessed, contributing to a comprehensive understanding of the dataset.

4. Findings and Insights:

4.1 Clustering Insights:

K-means clustering revealed clusters with distinct structural features. Hierarchical clustering highlights relationships based on hierarchical structures. Chemical diversity analysis showcased the rich diversity of molecular structures within and between clusters.

4.2 Potential Applications:

The identified clusters and chemical diversity insights hold potential applications in:

Drug Discovery: Clusters may represent cphemically similar compounds with varying bioactivities, aiding in drug target exploration.
Compound Library Design: Diverse clusters offer opportunities for designing compound libraries that cover a broad chemical space.
Chemical Informatics: Insights into molecular relationships can enhance structure-activity relationship (SAR) predictions and virtual screening.

5. Future Directions:

5.1 Further Algorithm Exploration:

Explore alternative clustering algorithms (e.g., DBSCAN, spectral clustering) to compare their efficacy in capturing different aspects of molecular diversity.

5.2 Fingerprinting Methods:

Investigate the impact of different fingerprinting methods and parameters on clustering results, considering descriptors beyond Morgan fingerprints.

5.3 Scalability:

Extend the analysis to larger datasets to assess the scalability of clustering approaches for real-world applications.

5.4 Application-Specific Analyses:

Conduct more granular analyses tailored to specific applications, such as predicting bioactivity within clusters.

6. Conclusion:

The project successfully applied clustering techniques to a refined dataset of small molecules, providing valuable insights into structural relationships and chemical diversity. The findings lay the groundwork for applications in drug discovery and compound library design. Further research avenues include algorithm exploration, method refinement, and scalability assessments.