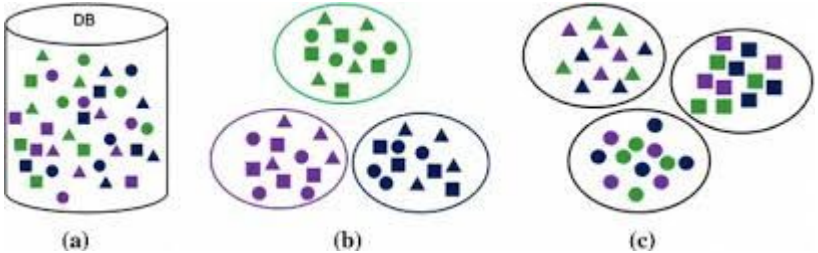
Lab 5: Clustering Techniques in Bioinformatics

Uncovering Patterns in Biological Data

Introduction to Clustering

Clustering - **unsupervised machine learning** technique used to:

- group similar data points
- reveal hidden patterns in complex biological datasets



Importance of Clustering in Bioinformatics

Clustering aids in identifying gene expression patterns, classifying patients for personalized medicine, and discovering potential drug targets.

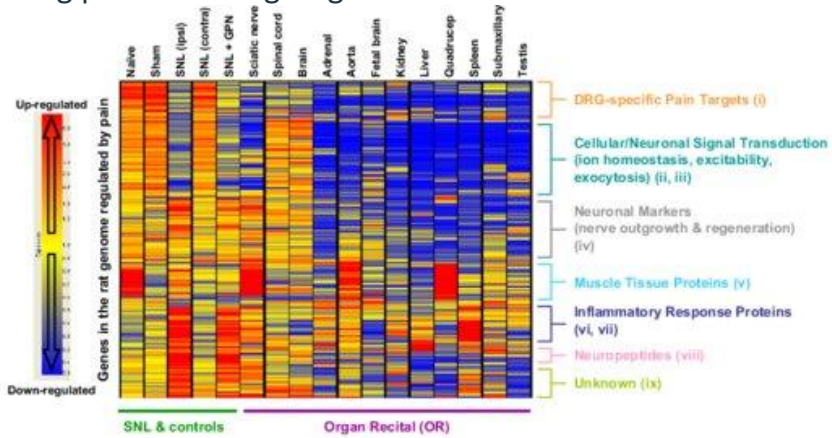


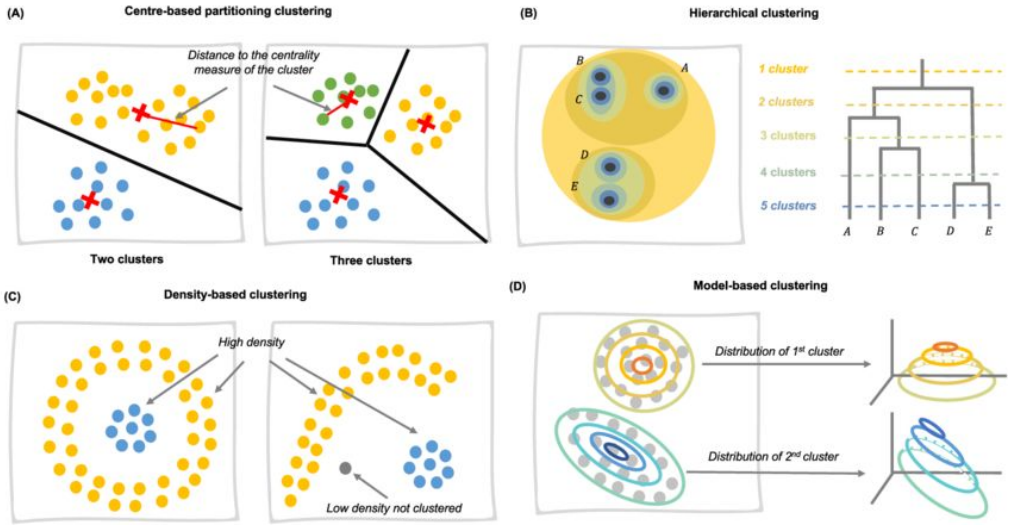
Fig. 1*. Heat map of gene expression data. Hierarchical clustering analysis was performed on 28 conditions

*Levin, Margaret & Jin, Jason & Ji, Rui-Ru & Tong, Jeifei & Pomonis, James & Lavery, Daniel & Miller, Scott & Chiang, Lillian. (2008). Complement activation in the peripheral nervous system following the spinal nerve ligation model of neuropathic pain. Pain. 137. 182-201. 10.1016/j.pain.2007.11.005.

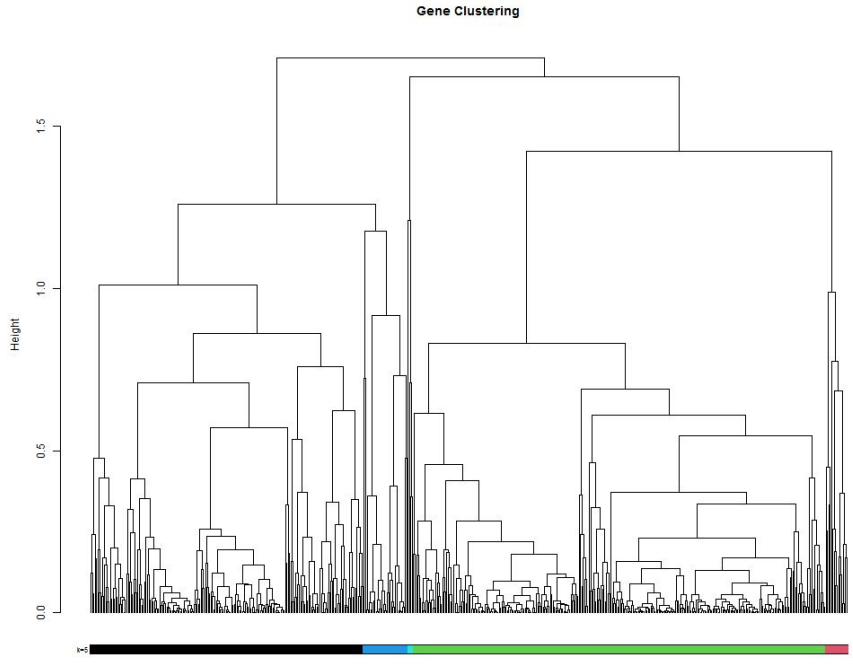
Types of Clustering Algorithms

Common clustering **methods** include:

- Hierarchical clustering
- K-means clustering
- Density-based clustering(DBSCAN)



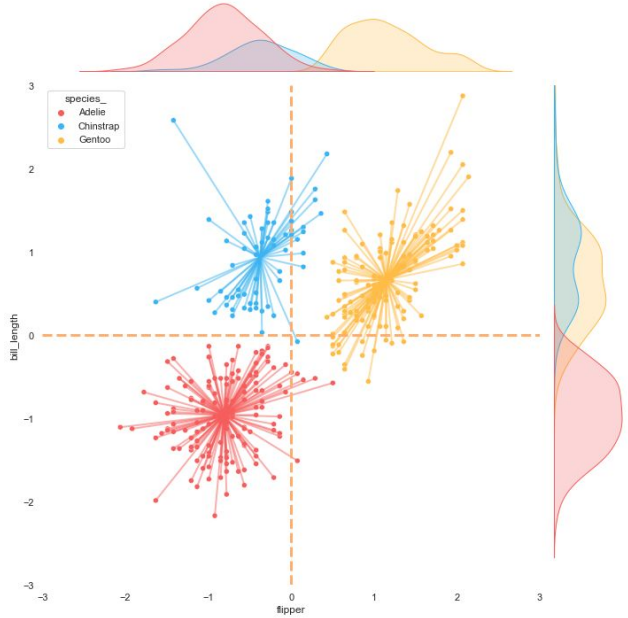
Hierarchical Clustering



Hierarchical clustering builds a tree-like structure (dendrogram) to represent nested clusters, useful for understanding data at multiple levels

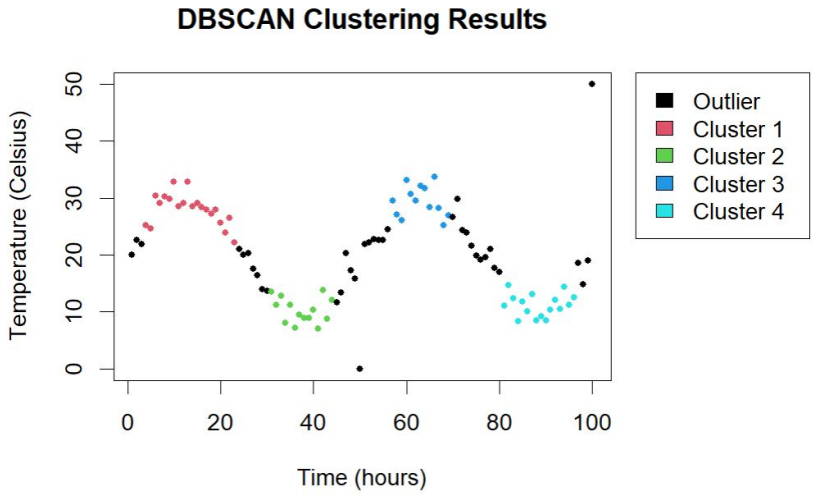
K-means Clustering

K-means clustering partitions data into K clusters by minimizing the variance within each cluster, effective for large datasets with well-defined clusters.



Density-Based Clustering (DBSCAN)

DBSCAN identifies clusters based on **data density**, effectively detecting clusters of various shapes and handling outliers.



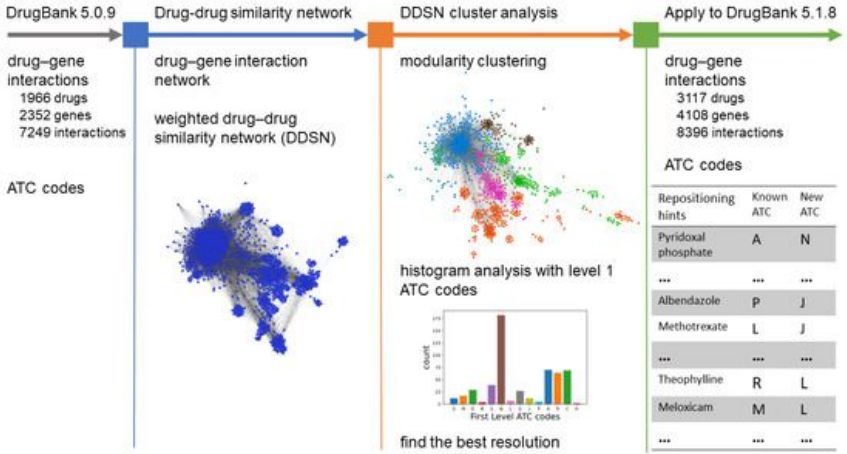
Case Study: Drug Repurposing Using Clustering

By constructing a **Drug-Drug Similarity Network** and applying the **Louvain algorithm**, we identified **clusters of drugs** with potential new therapeutic uses.

Our methodology involved:

- collecting drug-gene interaction data
- constructing a similarity network
- applying clustering algorithms
- validating clusters using ATC codes and literature.

ATC codes classify drugs based on therapeutic use and chemical characteristics, providing a framework for validating our clustering results



Groza, V., Udrescu, M., Bozdog, A., & Udrescu, L. (2021). Drug repurposing using modularity clustering in drug-drug similarity networks based on drug–gene interactions. *Pharmaceutics*, *13*(12), 2117.

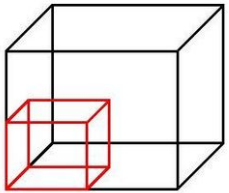
Intrinsic Dimensionality and the Curse of Dimensionality

As bioinformatics data grows in complexity, dimensions increase.

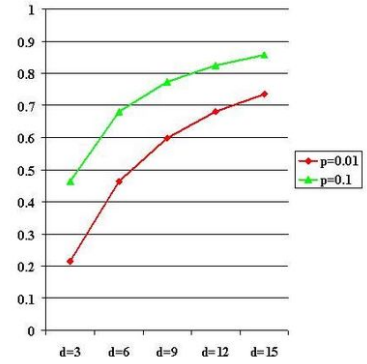
High-dimensionality, common in gene expression and multi-omics data, makes points appear uniformly distant, reducing clustering effectiveness - '**CURSE OF DIMENSIONALITY**'

(I) 50% of each dimension is sufficient to cover 25% of a 2-dimensional space

(II) 50% of each dimension is only sufficient to cover 12.5% of a 3-dimensional space



(III) A proportion $p^{1/d}$ of each dimension is needed to cover a proportion p of a d-dimensional space. The graph below plots $p^{1/d}$ vs. d for p = 1% and p = 10%.

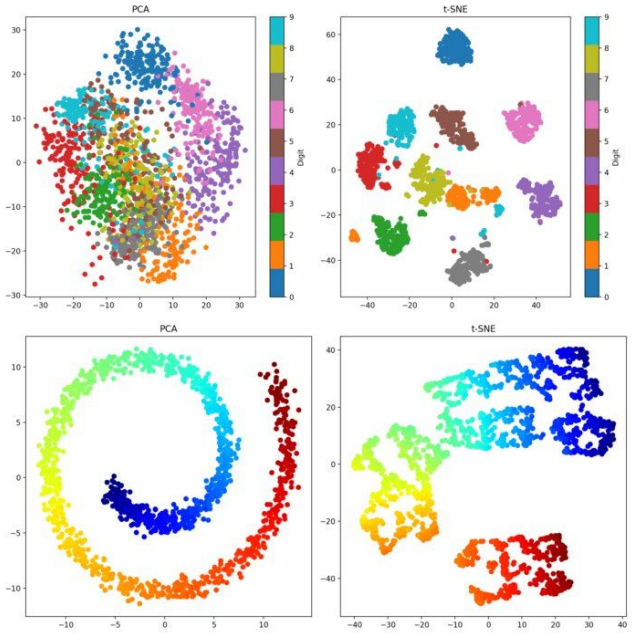


Dimensionality Reduction Techniques: PCA and t-SNE

PCA (Principal Component Analysis): Reduces high-dimensional data by capturing major variance in fewer dimensions.

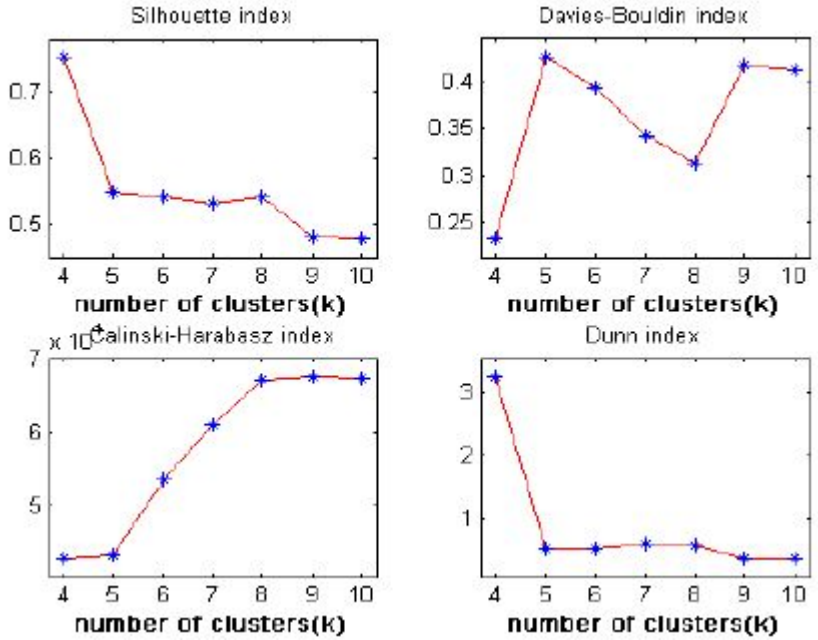
t-SNE (t-distributed Stochastic Neighbor Embedding): Visualizes high-dimensional relationships in 2D or 3D, preserving clusters.

Purpose: Reduces data complexity, making clustering more effective.

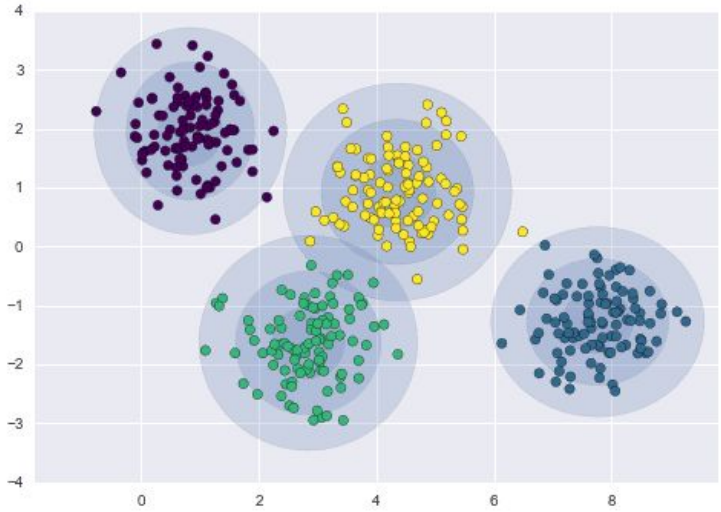


Validity Indices for Cluster Quality

- **Silhouette Score**: Measures how well-separated clusters are.
- **Dunn Index**: Assesses compactness and separation of clusters.
- **Biological Relevance**: External metrics (e.g., ATC codes) and literature ensure clusters reflect meaningful biological patterns.



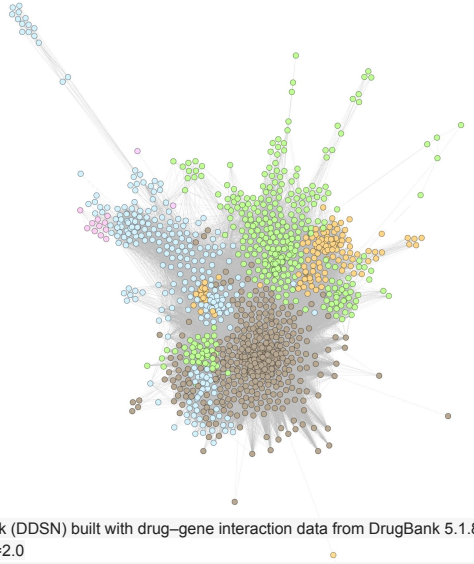
Handling Uncertainty with Probabilistic and Fuzzy Clustering



- **Gaussian Mixture Models (GMM)**: Allows overlapping clusters by modeling each cluster as a Gaussian distribution.
- **Fuzzy C-means Clustering**: Assigns probabilities to data points, allowing membership in multiple clusters.
- **Ideal for**: Complex, overlapping biological data, such as gene expression profiles.

Applications and Future Directions

- **Drug Repurposing**: Identifies relationships in drug networks, suggesting new therapeutic uses.
- **Gene Co-expression Networks**: Reveals gene modules across multiple data layers.
- **Next Steps**: Multi-omics integration for a systems-level view of biological processes.



Drug–drug similarity network (DDSN) built with drug–gene interaction data from DrugBank 5.1.8, clustered using modularity classes for resolution λmax =2.0
The brown nodes represent drugs in cluster $C0$(479 drugs), green nodes represent drugs in cluster $C1$ (346 drugs), light blue nodes represent drugs in cluster $C2$(270 drugs), orange nodes represent drugs in cluster $C3$(129 drugs), and pink nodes represent drugs in cluster $C4$(12 nodes).