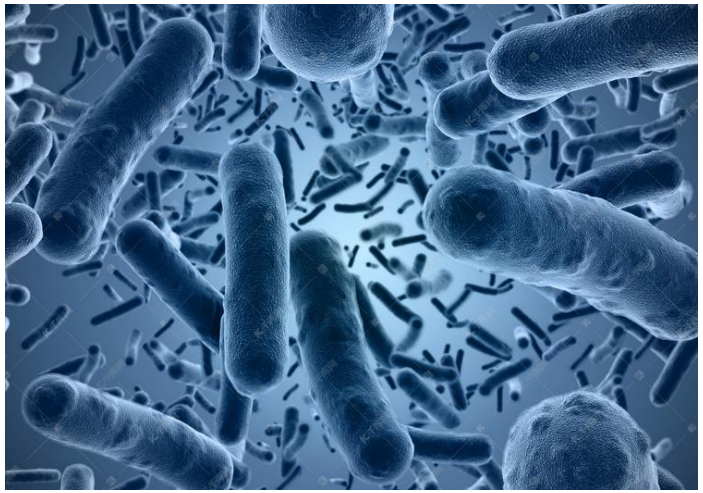
General Artificial Intelligence: Clustering DDI network

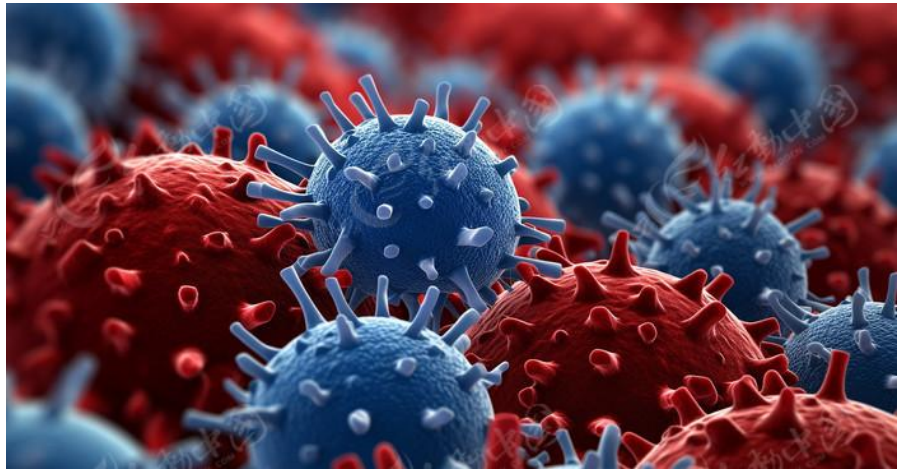
Ji Lv

Zhejiang Normal University

Introduction

- Why do doctors often request a blood test when you have a fever or a cold?
- This is not unnecessary testing — it is a critical first step in clinical diagnosis.
- The human body as an alarm system: Fever signals that the immune system is actively responding to pathogens.
- But who is the enemy?
  - ✓ Symptoms alone are insufficient to distinguish between viral and bacterial infections.





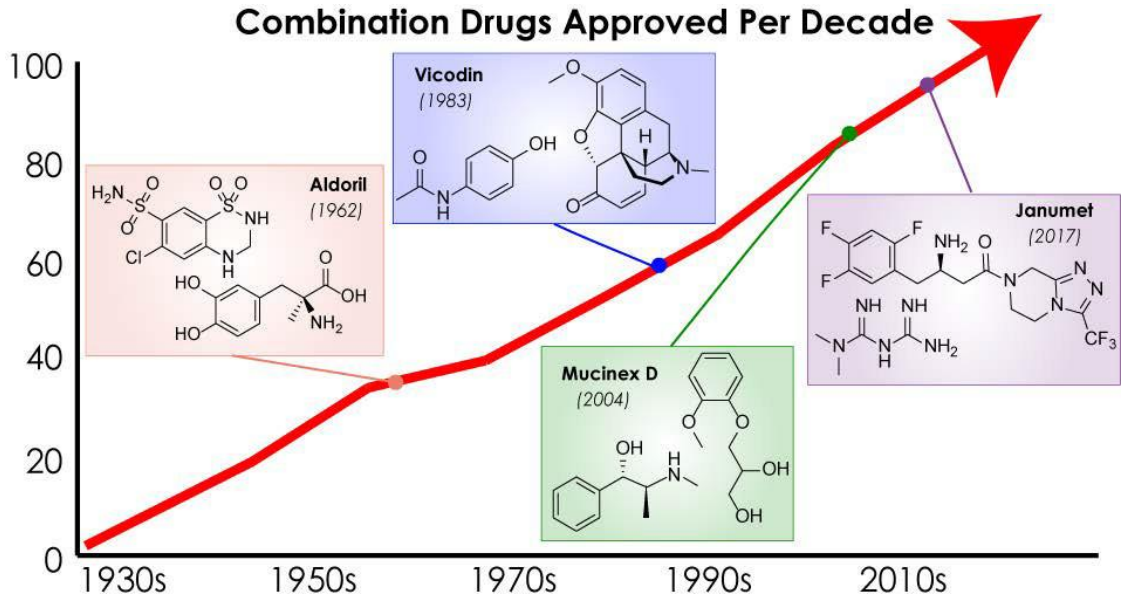
Introduction

- **Targeted Treatment: Specific Therapies for Bacterial vs. Viral Infections**
- **Antibiotics (e.g., cephalosporins, amoxicillin)**
  - ✓ **Effective against bacterial infections**
  - ✓ **Ineffective against viruses**
  - ✓ **Overuse may lead to antibiotic resistance**

- **Antiviral drugs (e.g., oseltamivir, ribavirin)**
  - ✓ **Inhibit viral replication**
  - ✓ **Help the immune system eliminate viruses**
  - ✓ **Symptomatic treatment is often sufficient (e.g., antipyretics, cough relief)**

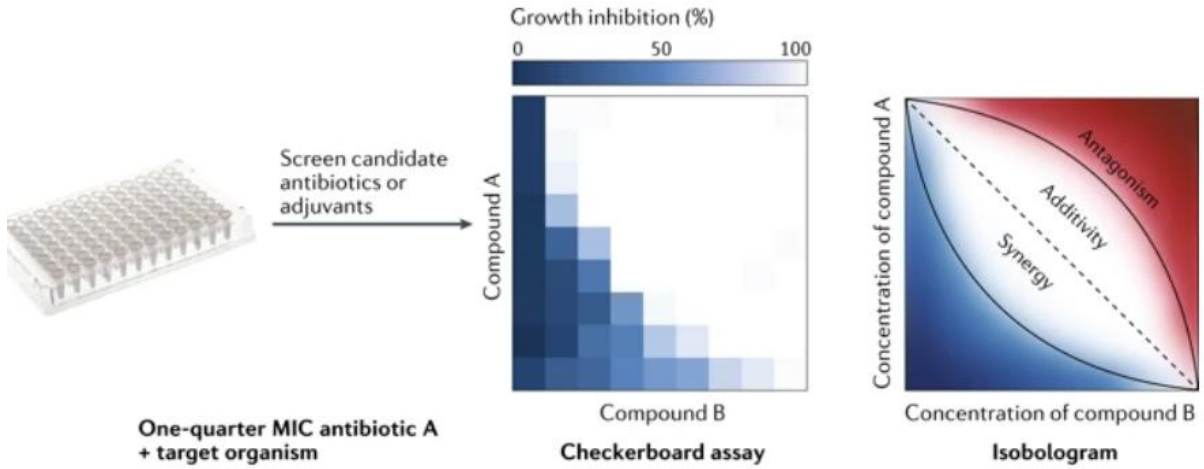
Combination Therapy

- In complex diseases (e.g., HIV and multidrug-resistant infections), monotherapy is often inadequate to achieve sustained therapeutic effects
- Combination therapy, involving multiple drugs,
  - ✓ Can enhance efficacy
  - ✓ Reduce resistance development
  - ✓ Improve clinical outcomes



Drug Combination

- Pharmacologically, drug combinations can be classified into:
  - ✓ **Synergistic effects** (enhanced combined effect)
  - ✓ **Antagonistic effects** (reduced effectiveness)
  - ✓ **Additive effects** (sum of individual effects)
- Identifying effective drug combinations requires:
  - ✓ Extensive laboratory experiments
  - ✓ Large-scale combinatorial testing
- This process is:
  - ✓ Time-consuming
  - ✓ Labor-intensive
  - ✓ Costly



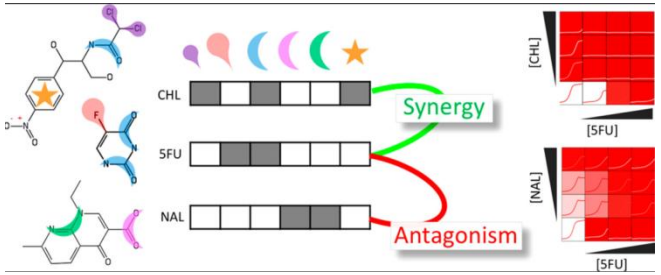
Drug Combination Prediction

Supervised Learning:

- ✓ Utilize labeled **datasets** and **features** (e.g., chemical structures)
- ✓ Input them are fed into machine learning models

Objective:

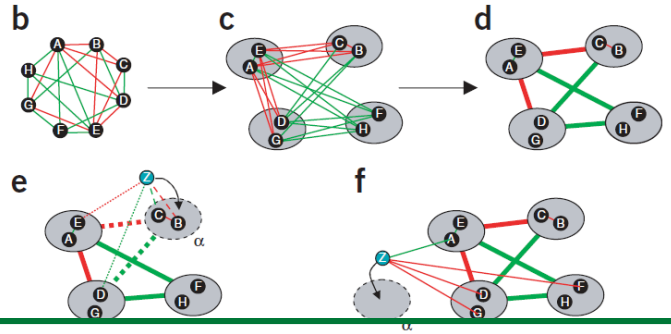
- ✓ **Classification** (e.g., synergy vs. antagonism)
- ✓ **Regression** (e.g., synergy scores)



J. Med. Chem., 2017, 60(9): 3902-3912.

Unsupervised Learning:

- ✓ Leverage **drug features or network structures**
- ✓ Apply clustering algorithms to group drugs or interactions



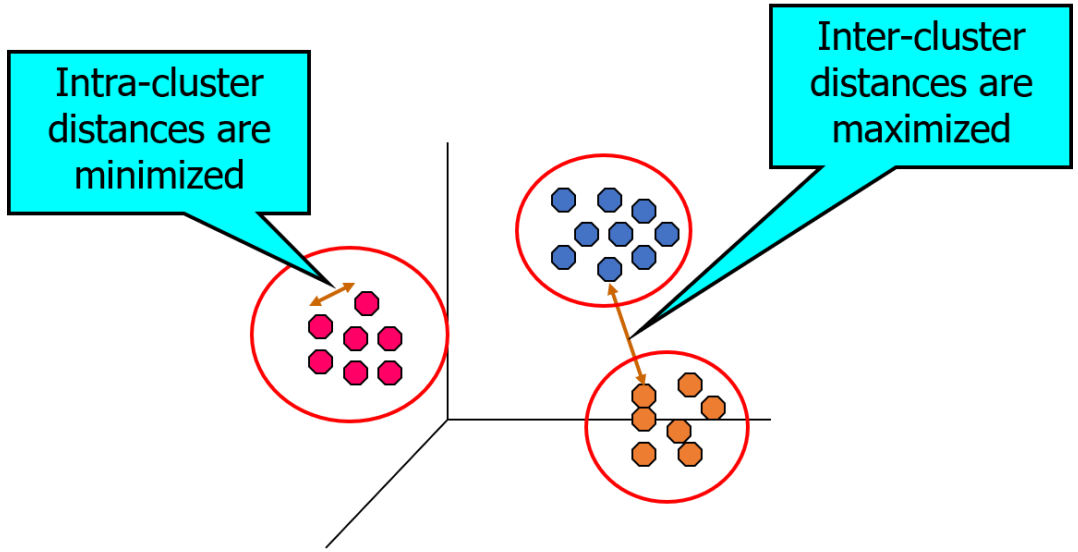
Nat. Genet., 2006, 38(4).

Goal:

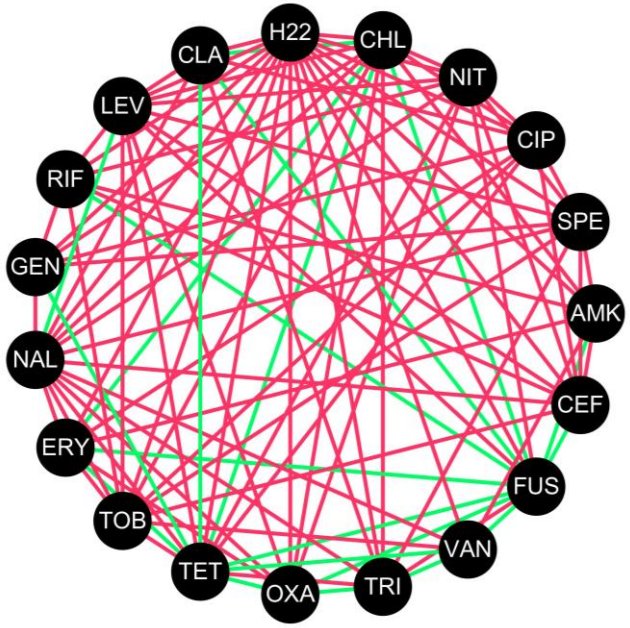
- ✓ Identify hidden patterns in drug interaction networks
- ✓ Explore interactions between clusters

Clustering

- **Clustering is an unsupervised learning task that groups samples based on their similarity or distance.**
- **Clustering relies on similarity measures or distance metrics**
- **Common choices include:**
  - ✓ **Euclidean distance**
  - ✓ **Cosine similarity**
  - ✓ **Jaccard similarity**
- **The choice of similarity or distance metric can significantly affect clustering results**

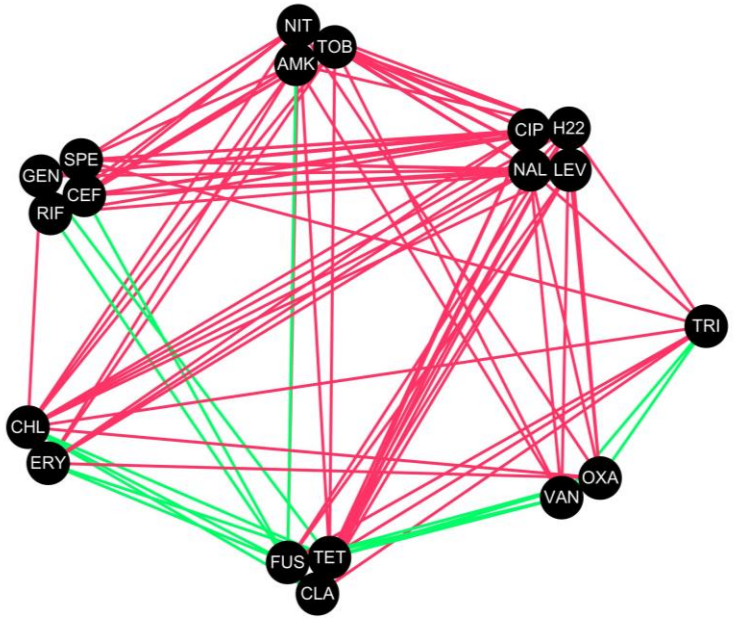


Clustering DDI network



18 Antibiotics

Clustering?



synergistic effects

antagonistic effects

(1) **Feature-based similarity**: Chemical, target-related features

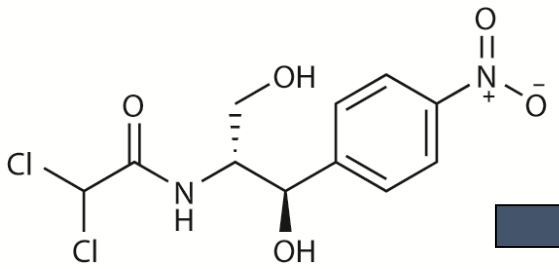
(2) **Network-based similarity**: Graph topology

Feature Selection in Clustering

- **Suppose we want to cluster students in this classroom:**
  - ✓ **Which attributes (features) can we use?**
- **Possible features include:**
  - ✓ **Gender**
  - ✓ **Hometown**
  - ✓ **Major**
  - ✓ **Hobbies**
  - ✓ **…**

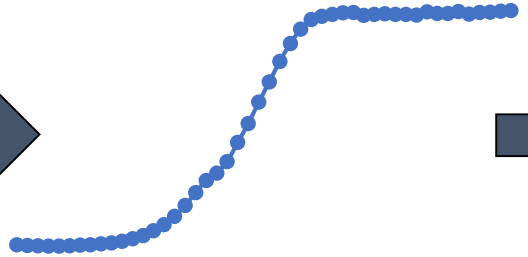
Drug Information

- **Drugs can be characterized across multiple biological levels:**
  - ✓ **Molecular level: chemical structure of small molecules**
  - ✓ **Target level: interactions with proteins or RNA/DNA (MoA)**
  - ✓ **Phenotypic level: observable cellular responses (e.g., growth inhibition)**
  - ✓ **Clinical level: therapeutic effects (e.g., ATC codes)**

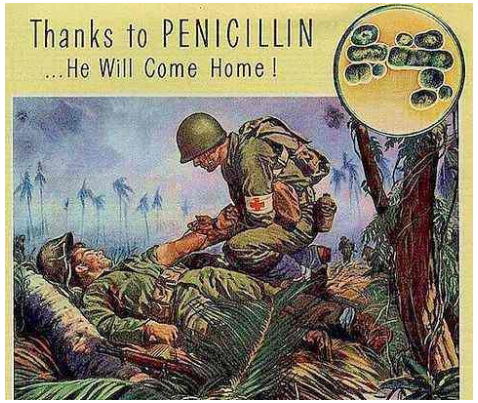


Chloramphenicol

***E. Coli* 70S ribosome**



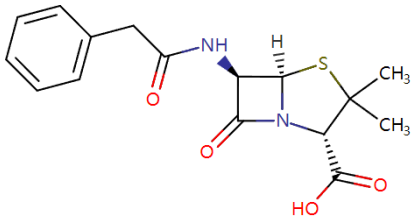
Bacteria growth

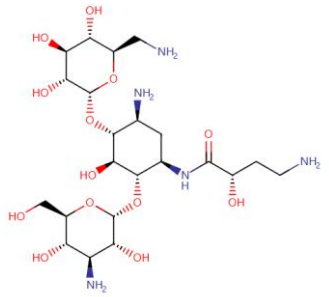


Treatment

Classification of Antibiotics by Chemical Structure

- **Antibiotics can be categorized based on their chemical structures:**
  - ✓ **β-lactams (e.g., penicillins, cephalosporins, cephamycins, carbapenems)**
  - ✓ **Aminoglycosides (e.g., gentamicin, amikacin, streptomycin, tobramycin)**
  - ✓ **Macrolides (e.g., erythromycin, clarithromycin, roxithromycin)**
  - ✓ **Tetracyclines (e.g., tetracycline, oxytetracycline, chlortetracycline)**
  - ✓ **Polypeptides (e.g., polymyxins, bacitracin, vancomycin)**
  - ✓ **…**



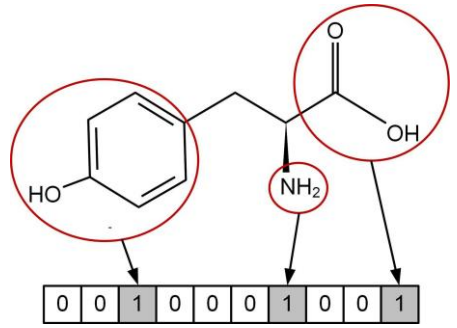


Certain combinations (β-lactams + aminoglycosides) exhibit synergistic effects

Nature Biotechnology,2020, 38, 1087–1096.

Structural Similarity

- SMILES strings can be retrieved from databases such as **PubChem** (https://pubchem.ncbi.nlm.nih.gov)
- Drugs can be represented using **molecular fingerprints**
  - ✓ e.g., Morgan fingerprints, MACCS keys
- These fingerprints encode chemical structures into
  - ✓ fixed-length binary vectors



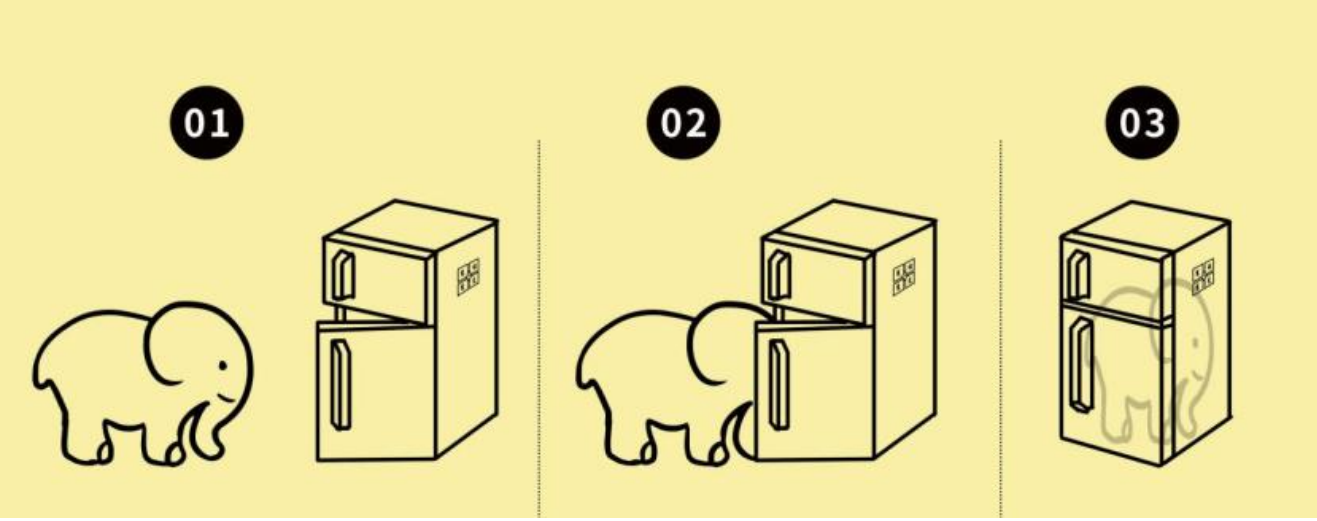
166 public MACCS keys

- **RDKit** (https://www.rdkit.org/docs/index.html)
  - ✓ Widely used for cheminformatics analysis
  - ✓ Enables generation of molecular fingerprints
- Structural similarity between drugs can be quantified using:
  - ✓ Dice similarity

$$S_s(A,B) = Dice(A,B) = \frac{|A| \cap |B|}{|A| \cup |B|}$$

Structural Similarity

- How can we use LLMs to help write code for similarity computation?
- A useful strategy is to **decompose a complex problem into smaller, manageable steps**
- Consider a classic question:
  - ✓ How can you put an elephant into a refrigerator?



Structural Similarity

- How can we perform clustering using chemical structure information?
- Workflow
  - Load data
    - ✓ Read the Excel file to obtain compound names and SMILES strings
  - Molecular representation
    - ✓ Use RDKit to convert SMILES into molecular objects
    - ✓ Compute MACCS fingerprints
  - Similarity computation
    - ✓ Calculate pairwise Tanimoto similarity
  - Clustering
    - ✓ Apply Spectral Clustering
  - Output results
    - ✓ Represent clustering results as a dictionary

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	CID	Drug	Abbreviati	Target	Pathway	SMILES	ATCcode						
2	37768	Amikacin	AMK	rpsL	Protein synthesis, 30S	C1[C@@H]([C	S01AA21, D06AX12, J01GB06						
3	441199	Cefoxitin	CEF	mrcA, mrcB, dac	Cell wall	CO[C@@]1([C	J01DC01						
4	5959	Chloramp	CHL	rplP	Protein synthesis, 50S	C1=CC(=CC=	D06AX02, S02AA01, D10AF03, S01AA01, G01AA05, J01BA01, S03AA08						
5	2764	Ciprofloxa	CIP	gyrA, parC	DNA gyrase	C1CC1N2C=C	J01RA11, S03AA07, J01MA02, S02AA15, S01AE03						
6	84029	Clarithron	CLA	rplD, rplV	Protein synthesis, 50S	CC[C@@H]1[	A02BD12, A03BD07, J01FA09						
7	12560	Erythromy	ERY	rplD, rplV	Protein synthesis, 50S	CC[C@@H]1[	J01FA01, D10AF02, S01AA17						
8	3000226	Fusidic ac	FUS	fusA	Elongation factor—protei	C[C@H]1[C@	S01AA13, D09AA02, J01XC01, D06AX01,						
9	3467	Gentamici	GEN	rpsL	Protein synthesis, 30S	CC(C1CCC(C(	S01AA11, S02AA14, D06AX07, S03AA06, J01GB03						
10	149096	Levofloxac	LEV	gyrA, parC	DNA gyrase	C[C@H]1COC	J01MA12, A02BD10, J01RA05, S01AE05						
11	4421	Nalidixic a	NAL	gyrA	DNA gyrase	CCN1C=C(C(=	J01MB02						
12	6604200	Nitrofuran	NIT	ydbK, nfsA, rpsJ	Multiple mechanisms	C1C(=O)NC(=	J01XE01						
13	6196	Oxacillin	OXA	dacB, ftsI	Cell wall	CC1=C(C(=N	J01CR50						
14	135398735	Rifampicin	RIF	rpoB	RNA synthesis	C[C@H]1/C=C	J04AM02						
15	15541	Spectinon	SPE	rpsL	Protein synthesis, 30S	C[C@@H]1CC	J01XX04						
16	54675776	Tetracyclir	TET	rpsG, rpsN	Protein synthesis, 30S	C[C@@]1([C@	S02AA08, A02BD02, A01AB13, S01AA09, J01AA20, J01RA08, S03AA02, D06AA0						
17	36294	Tobramyc	TOB	rpsL	Protein synthesis, 30S	C1[C@@H]([C	S01AA12, J01GB01						
18	5578	Trimethop	TRI	folA	Folic acid biosynthesis	COC1=CC(=C	J01EE03, J04AM08						
19	14969	Vancomyc	VAN	ddpX	Cell wall	C[C@H]1[C@	J01XA01, A07AA09, S01AA28						

Structural Similarity

Example Prompt for LLM-Assisted Code Generation:

- ✓ **"Write a Python program to compute the structural similarity between compounds and perform clustering analysis using spectral clustering. Please use libraries such as NumPy, RDKit, pandas, and scikit-learn. The workflow should include the following steps:**

  1. Read an Excel file to obtain the compound names and SMILES strings.

  2.Use RDKit to convert SMILES into molecular objects and compute MACCS fingerprints.

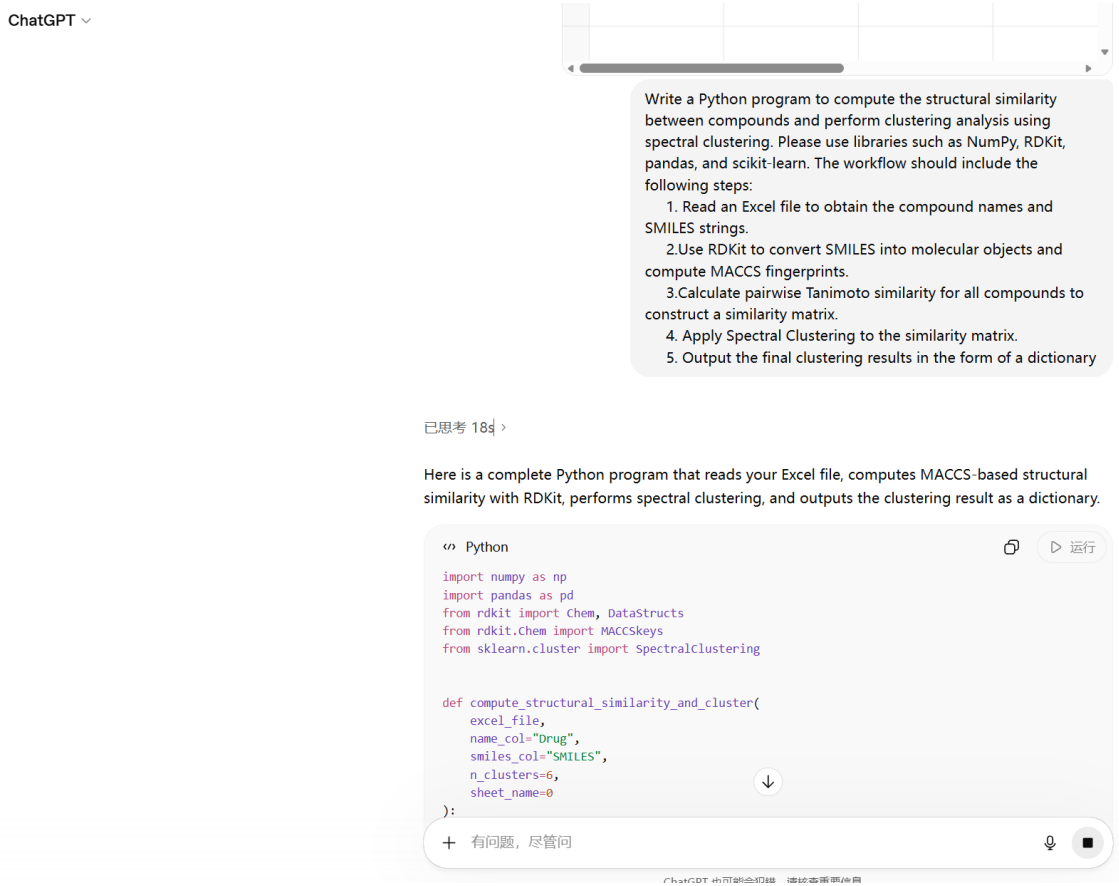
  3.Calculate pairwise Tanimoto similarity for all compounds to construct a similarity matrix.

  4. Apply Spectral Clustering to the similarity matrix.

  5. Output the final clustering results in the form of a python dictionary."

Structural Similarity

- For example, we can use ChatGPT to assist with code writing and debugging.



Structural Similarity

- Use Spyder to and run Python programs

cs.py

```python
import numpy as np
import pandas as pd
from rdkit import Chem, DataStructs
from rdkit.Chem import MACCSkeys
from sklearn.cluster import SpectralClustering


def structural_similarity_clustering(
    excel_file,
    abbr_col="Abbr",      # 药物缩写列
    smiles_col="SMILES",
    n_clusters=6
):
    # 1. 读取Excel
    df = pd.read_excel(excel_file)

    if abbr_col not in df.columns or smiles_col not in df.columns:
        raise ValueError(
            f"必须包含列: {abbr_col}, {smiles_col}\n当前列: {list(df.columns)}"
        )

    df = df[[abbr_col, smiles_col]].dropna()

    # 2. SMILES → Mol → MACCS指纹
    abbrs = []
    fps = []

    for _, row in df.iterrows():
        abbr = row[abbr_col]
        smiles = row[smiles_col]

        mol = Chem.MolFromSmiles(smiles)
        if mol is None:
            continue

        fp = MACCSkeys.GenMACCSKeys(mol)

        abbrs.append(abbr)
        fps.append(fp)

    n = len(fps)

    # 3. 构建Tanimoto相似度矩阵
```

Structural Similarity

- **Clustering Results of 18 Antibiotics**

Clusters = {

0: ['CHL', 'NIT'],

1: ['CEF', 'FUS', 'OXA', 'RIF', 'TET', 'VAN'],

2: ['CLA', 'ERY'],

3: ['NAL', 'SPE', 'TRI'],

4: ['CIP', 'LEV'],

5: ['AMK', 'GEN', 'TOB']}

Structural Similarity

- **Clustering quality can be evaluated using edge purity**
  - ✓ **Edge purity measures the consistency of interactions within clusters**

$$edge\ purity = \frac{1}{N}\sum_{i}\sum_{j}\max(l_{ij}, r_{ij})$$

$$i = 1, 2, \ldots, k-1$$

$$j = i+1, i+2, \ldots, k$$

- l_{ij} : number of **synergistic interactions** between cluster i and j
- r_{ij} : number of **antagonistic interactions** between cluster i and j
- k: total number of clusters
- N: total number of drug combinations

Structural Similarity

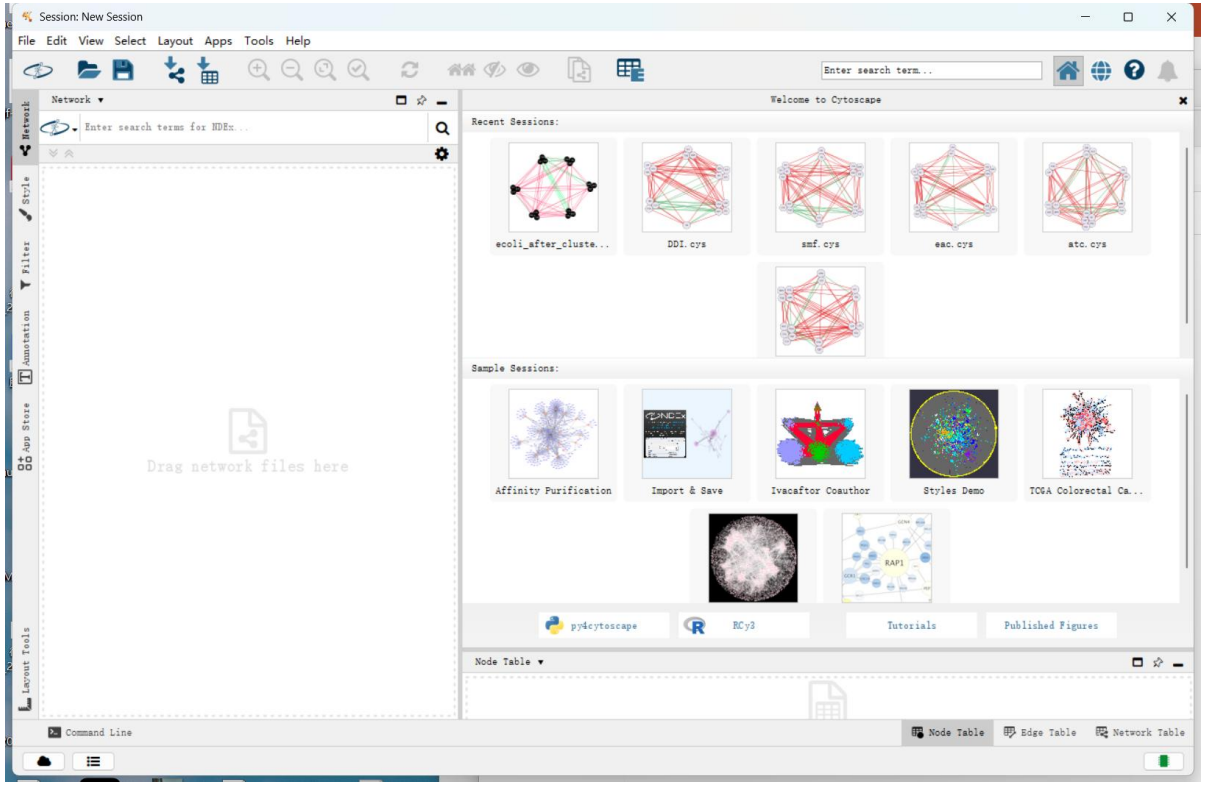
- The code for computing edge purity has already been provided.
- You only need to replace the clustering results.

```python
import numpy as np
from collections import Counter


def search(x):
    for i in range(len(drug)):
        if drug[i]==x:
            cluster_type = cluster_result[i]
    return cluster_type

def group_search(x):
    group_list =[]
    for i in range(len(GGI)):
        if GGI[i] ==x:
            group_list.append(alpha[i][2])
    return group_list


clusters = {"0": ['CHL', 'NIT'],
            "1": ['CEF', 'FUS', 'OXA', 'RIF', 'TET', 'VAN'],
            "2": ['CLA', 'ERY'],
            "3": ['NAL', 'SPE', 'TRI'],
            "4": ['CIP', 'LEV'],
            "5": ['AMK', 'GEN', 'TOB']}


drug = np.loadtxt('drug_19.txt',dtype=str,delimiter='    ')
alpha = np.loadtxt('./combination_cytospace_19.txt',dtype=str,delimiter='    ')

cluster_result = []
for i in range(len(drug)):
    for j in range(len(clusters)):
        if str(drug[i]) in clusters[str(j)]:
            cluster_result.append(j)

GGI = []
for i in range(len(alpha)):
    drug1_ID = search(alpha[i][0])
    drug2_ID = search(alpha[i][1])
```

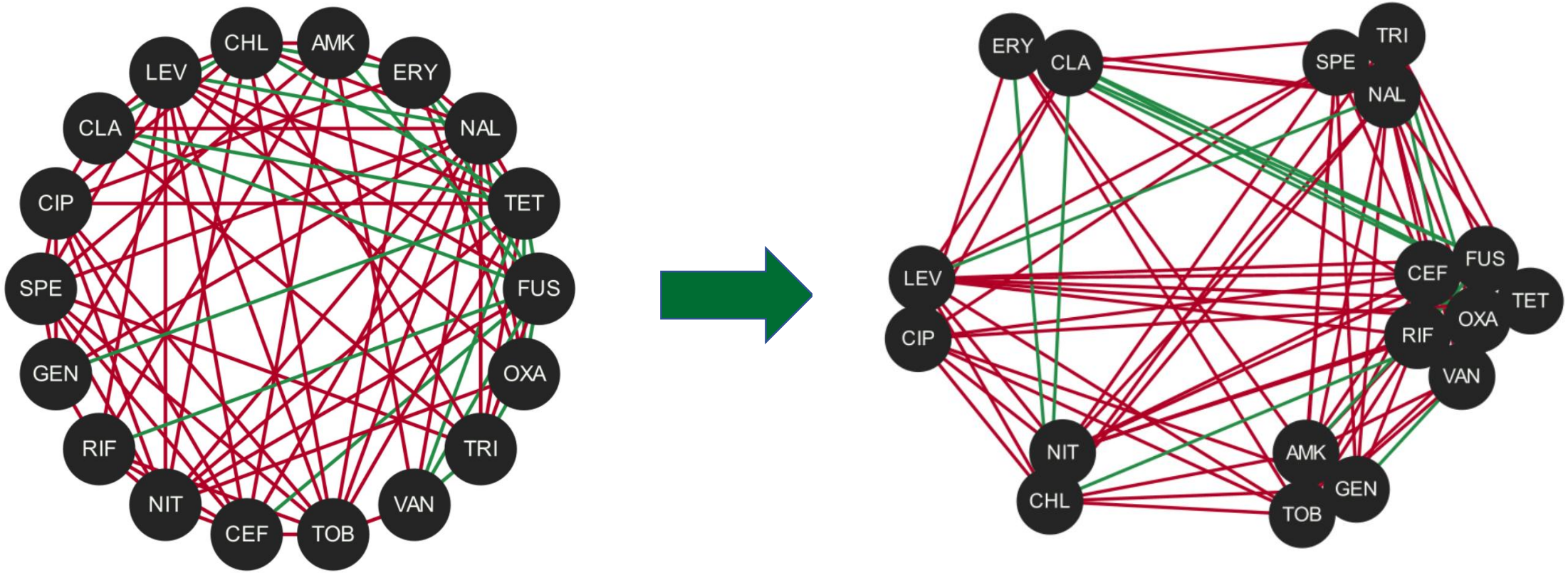
edge purity = 0.8857

Structural Similarity

- Clustering results can be visualized using Cytoscape (https://cytoscape.org)
- Cytoscape provides an intuitive way to explore
  - ✓ Network structure
  - ✓ Cluster assignments
  - ✓ Interaction patterns



Structural Similarity



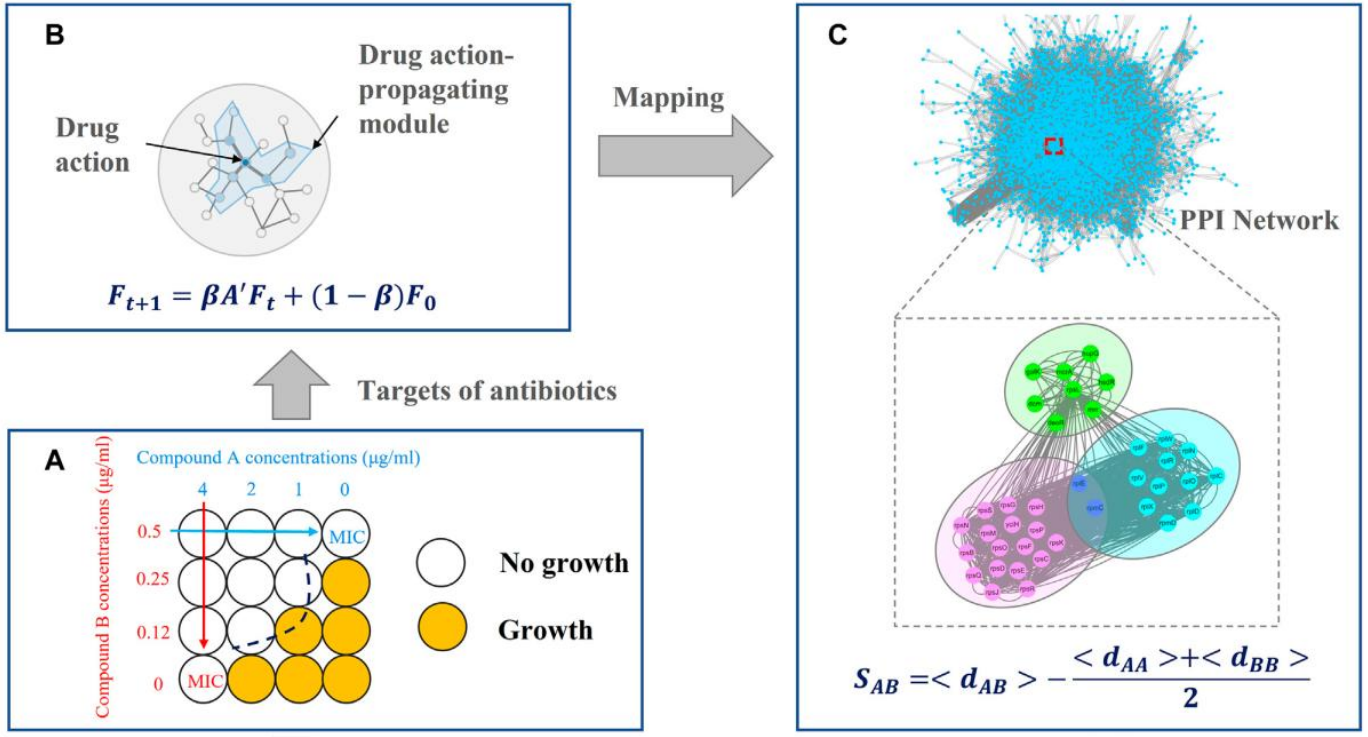
Pharmacological Similarity

- **Antibiotics can also be categorized based on their mechanisms of action :**
  - ✓ **Inhibition of DNA or RNA synthesis (e.g., quinolones, rifamycins)**
  - ✓ **Inhibition of protein synthesis**
    - ✓ **Targeting the 30S ribosomal subunit (e.g., tetracyclines, aminoglycosides)**
    - ✓ **Targeting the 50S ribosomal subunit (e.g., macrolides)**
  - ✓ **Inhibition of cell wall synthesis (e.g., β-lactams, glycopeptides)**
  - ✓ **Disruption of folate metabolism (e.g., trimethoprim, sulfonamides)**
  - ✓ **…**

Nature Biotechnology,2020, 38, 1087–1096.

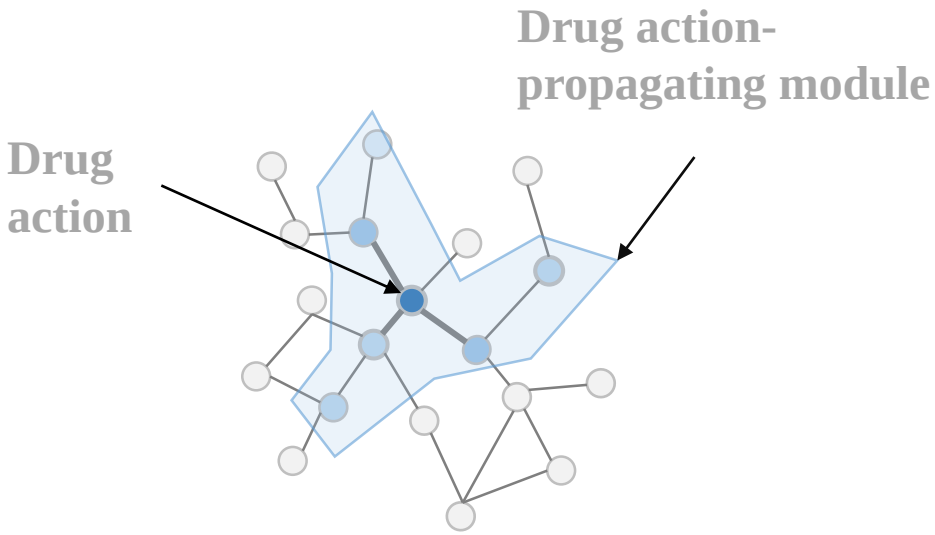
Pharmacological Similarity

- How to compute **pharmacological similarity**?
  - ✓ Pharmacological similarity can be characterized from a **network perspective**
  - ✓ Drug targets can be retrieved from **DrugBank** (https://go.drugbank.com)
  - ✓ PPI networks can be obtained from **STRING** (https://string-db.org)



Pharmacological Similarity

- In this study, we assume that when a drug targets specific proteins, its effect propagates through the protein–protein interaction (PPI) network
- The propagation process results in a subnetwork centered around drug targets
  - ✓ This subnetwork is defined as **Drug Action-Propagating Module (DAPM)**
- Only nodes satisfying the following condition are included: $F_i^* \geq 0.0065$



$$F_{t+1} = \beta A' F_t + (1-\beta) F_0$$

$$\lim_{t\to\infty} F_t = (I - \beta A')^{-1}$$

Pharmacological Similarity

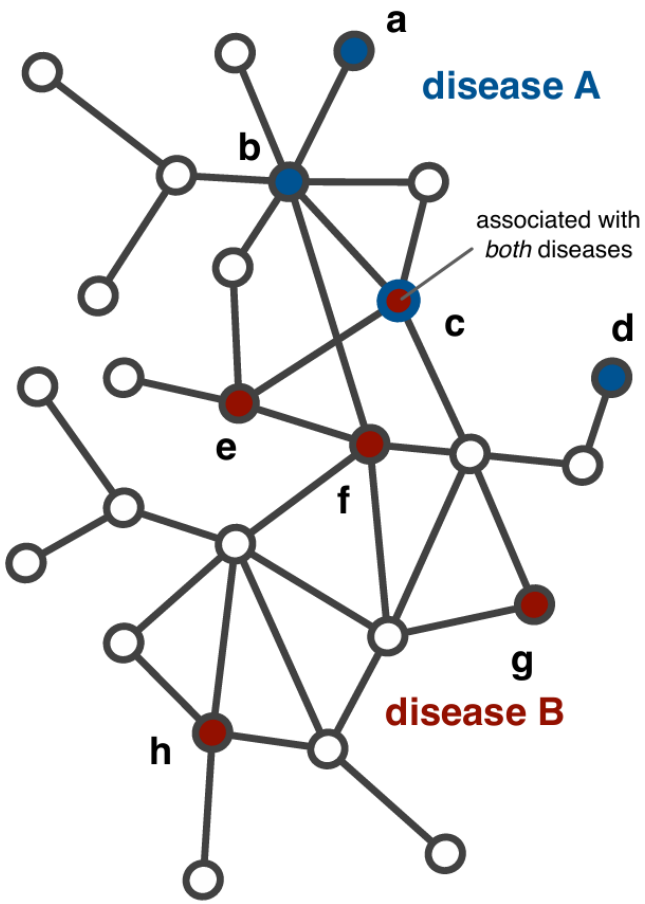
- **NetworkX for Network Analysis**
  - ✓ **NetworkX is a widely used Python library for complex network analysis**
- **Network proximity can be used to quantify pharmacological similarity between drugs**
  - ✓ **Let drug A and drug B act on two corresponding modules in a PPI network**

$$S_{AB} = < d_{AB} > - \frac{< d_{AA} > + < d_{BB} >}{2} \qquad < d_{AB} >= \frac{1}{|D_A| + |D_B|} \sum_{y \in D_A} min_{x \in D_B} d(x, y)$$

- $<d_{AA}>$ and $<d_{BB}>$: the average shortest path lengths within module A and module B, respectively
- $<d_{AB}>$: the average shortest path length between the two modules
- $d(x, y)$: the shortest path length between node x and node y

- **Normalization**
  - ✓ **Since $S_{AB} \in$ [-1,1]** → $S_{Pha}(A, B) = \frac{1}{1 + e^{-S_{AB}}}$

Pharmacological Similarity



shortest distances *within* disease **A**:

a: 1
b: 1
c: 1
d: 3

$$\langle d_{AA} \rangle = \frac{3}{2}$$

shortest distances *within* disease **B**:

c: 1
e: 1
f: 1
g: 2
h: 2

$$\langle d_{BB} \rangle = \frac{7}{5}$$

shortest distances *between* diseases **A** & **B**:

a: 2
b: 1
c: 0
d: 3

c: 0
e: 1
f: 1
g: 2
h: 3

$$\langle d_{AB} \rangle = \frac{13}{9}$$

resulting separation:

$$s_{AB} = \frac{13}{9} - \frac{1}{2}\left(\frac{3}{2} + \frac{7}{5}\right) = -\frac{1}{180}$$

$$S_{AB} = < d_{AB} > - \frac{< d_{AA} > + < d_{BB} >}{2}$$

$$< d_{AB} > = \frac{1}{|D_A| + |D_B|} \sum_{y \in D_A} min_{x \in D_B} d(x, y)$$

Science **346**, 1480–1486 (2014).

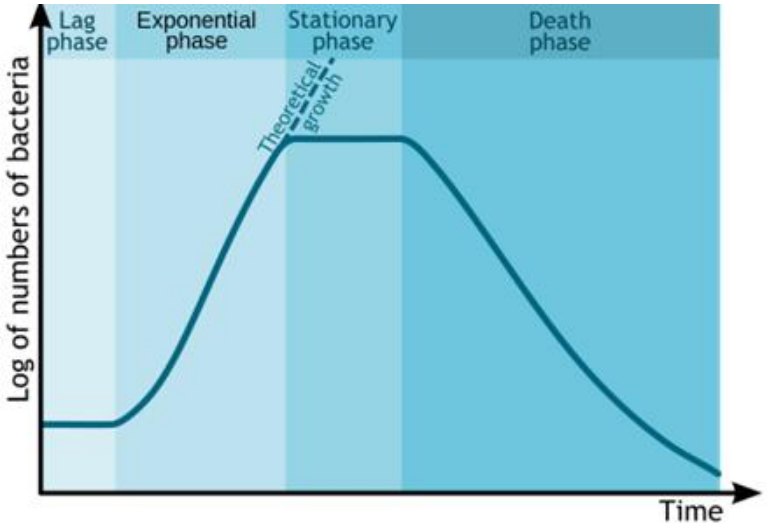
Nat Commun **10**, 1197 (2019)

Pharmacological Similarity

- **How to Design a Prompt for Computing Pharmacological Similarity?**
  - ✓ **To effectively use an LLM, we need to clearly define the task and break it into structured steps**
- **Task definition**
  - ✓ **pharmacological similarity between drugs**
- **Input specification**
  - ✓ **drug targets, PPI networks**
- **Method description**
  - ✓ **network proximity based on shortest paths**
- **Expected output**
  - ✓ **Pyhton dictionary**

Phenotypic Similarity

- The bacterial growth curve consists of four distinct phases:
  - ✓ Lag phase
  - ✓ Exponential (log) phase
  - ✓ Stationary phase
  - ✓ Death phase



- From Bacterial Growth to Phenotypic Similarity Measure
  - ✓ Under favorable conditions, bacteria grow exponentially
  - ✓ When antibiotics are introduced, bacterial growth can be inhibited

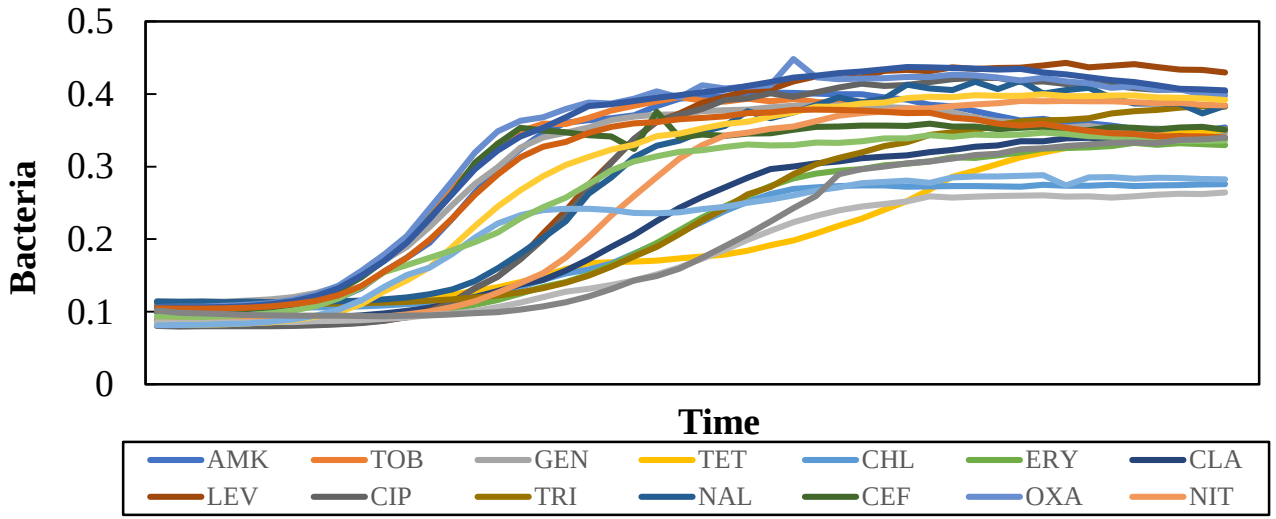
Phenotypic Similarity

- **Classification of Antibiotics Based on Bactericidal and Bacteriostatic Effects**
- **Bactericidal agents (active during growth phase)**
  - ✓ **penicillins, cephalosporins, imipenem, vancomycin, rifamycins**
- **Bactericidal agents (active during stationary phase)**
  - ✓ **aminoglycosides, polymyxins, bacitracin**
- **Fast-acting bacteriostatic agents**
  - ✓ **macrolides, tetracyclines, lincosamides**
- **Slow-acting bacteriostatic agents**
  - ✓ **sulfonamides, capreomycin**

Nature Biotechnology,2020, 38, 1087–1096.

Phenotypic Similarity

- **Based on bacterial growth curves, antibiotics can be classified into:**
  - ✓ **Bacteriostatic agents (inhibit growth)**
  - ✓ **Bactericidal agents (kill bacteria)**



- **We can quantify phenotypic similarity**

$$S_{Phe}(A,B) = \exp\left(-\frac{|\mathbf{x_A} - \mathbf{x_B}|^2}{2\sigma^2}\right)$$

- x_A, x_b: bacterial growth curve of drugs A and B, respectively.
- $|x_A - x_b|$: distance between two growth curves
- σ: parameter

Shuyuan Cao, Guixia Liu*, Xiangrun Zhou, Ji Lv*. *Quantitative Biology*, **2025.**

Phenotypic Similarity

- **How to Design a Prompt for Computing Phenotypic Similarity?**
  - ✓ **To effectively use an LLM, we should formulate the task as astructured, step-by-step problem**
- **Task definition**
  - ✓ **Compute phenotypic similarity between drugs**
- **Input specification**
  - ✓ **Bacterial growth curves**
- **Method description**
  - ✓ **Distance between growth curves**
  - ✓ **Gaussian kernel for similarity computation**
- **Expected output**
  - ✓ **Pyhton dictionary**

Therapeutic Similarity

- According to the Anatomical Therapeutic Chemical (ATC) classification system
- ATC codes can be obtained DrugBank (https://go.drugbank.com)
- Drugs are categorized into 14 major groups (e.g., anti-infectives for systemic use, respiratory system)
- Each drug can be associated with one or multiple ATC codes
- For example, erythromycin is associated with multiple ATC codes:
  - ✓ J01FA01
  - ✓ D10AF02
  - ✓ S01AA17
  - ✓ D01AF52

  - ✓ Level 1: Anatomical main group (J)
  - ✓ Level 2: Therapeutic subgroup (J01)
  - ✓ Level 3: Pharmacological subgroup (J01F)
  - ✓ Level 5: Chemical substance (J01FA01)
- Hierarchical Representation of ATC Codes
- The ATC system is a multi-level hierarchical classification

Therapeutic Similarity

- **From ATC codes to Similarity**
- **ATC codes provide a hierarchical feature representation**
- **Drugs sharing the same prefix (e.g., J01F)**
  - ✓ **are more similar at a given level**
- **Compare drugs using 3-level ATC codes**
  - ✓ **Enables computation of $S_T^k(A,B)$**

$$S_T^k(A,B) = \frac{|ATC_k(A) \cap ATC_k(B)|}{|ATC_k(A) \cup ATC_k(B)|}$$

Therapeutic Similarity

- **Example: AMK vs TOB**
  - ✓ **AMK (Amikacin)**
    - **S01AA21, D06AX12, J01GB06, J01RA06**
  - ✓ **TOB (Tobramycin)**
    - **S01AA12, J01GB01**
- **Step 1: Extract ATC prefixes (e.g., first 3 characters)**
  - ✓ **AMK: S01A, D06A, J01G, J01R**
  - ✓ **TOB: S01A, J01G**
- **Step 2: Compute similarity**
  - ✓ **Intersection: {S01A, J01G} → size = 2**
  - ✓ **Union: {S01, D06, J01, J01R} → size = 4**

$$S_T^3(\mathrm{AMK}, \mathrm{TOB}) = \frac{2}{4} = 0.5$$

Therapeutic Similarity

- **How to Design a Prompt for Computing Therapeutic Similarity?**
  - ✓ **To effectively use an LLM, we need to clearly define the task and break it into structured steps**
- **Task definition**
  - ✓ **Therapeutic Similarity between drugs**
- **Input specification**
  - ✓ **ATC codes**
- **Method description**
  - ✓ **Jaccard similarity**
- **Expected output**
  - ✓ **Pyhton dictionary**

Assignment 1

- **Required Tasks**
  1. **Compute Drug Similarity from Multiple Perspectives**
     - ✓ **Structural similarity**
     - ✓ **Pharmacological similarity**
     - ✓ **Phenotypic similarity**
     - ✓ **Therapeutic similarity**
  2. **Clustering Analysis**
     - ✓ **Apply Spectral Clustering**
     - ✓ **Cluster 18 antibiotics into 6 groups**
     - ✓ **Perform clustering based on each similarity type**
  3. **Evaluation and Visualization**
     - ✓ **Evaluate clustering quality using Edge Purity**
     - ✓ **Visualize clustering results using Cytoscape**

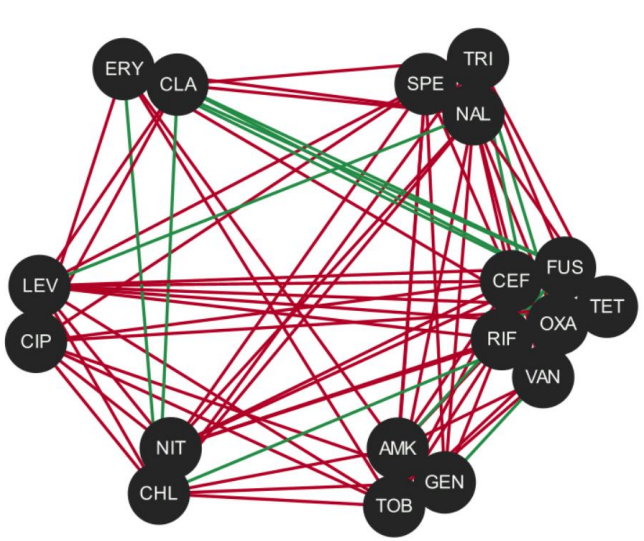
Assignment 1

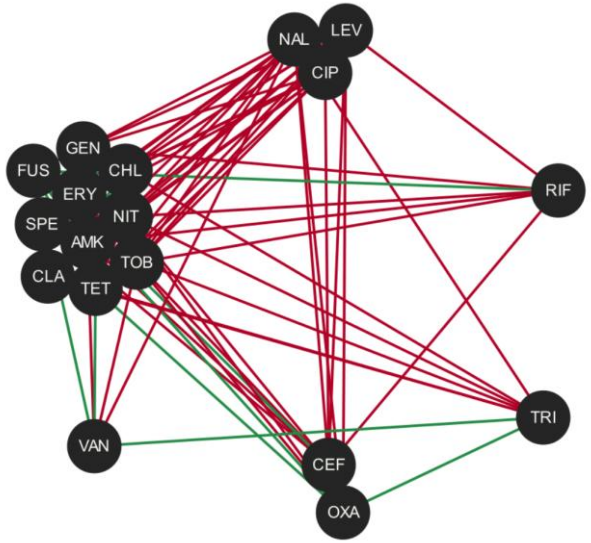
- **Optional Tasks**
  1. **Explore different numbers of clusters**
     - ✓ **Determine the optimal cluster number**
  2. **Try different clustering algorithms**
     - ✓ **e.g., K-means, hierarchical clustering**
  3. **Explore additional drug similarity features**
     - ✓ **Literature review or interaction with LLMs**

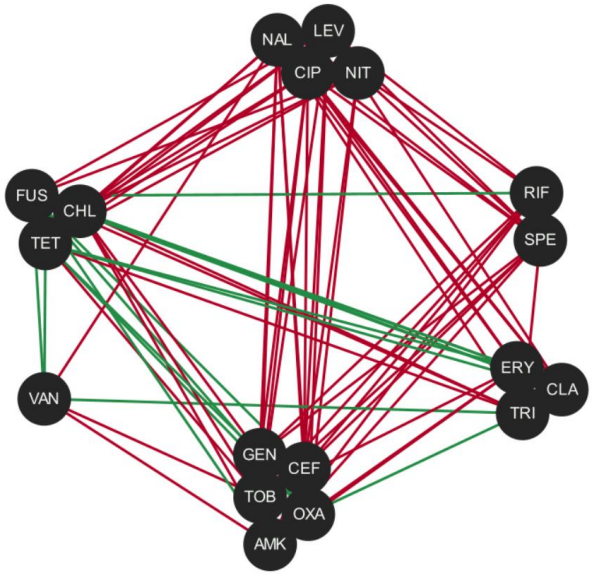
Thank you

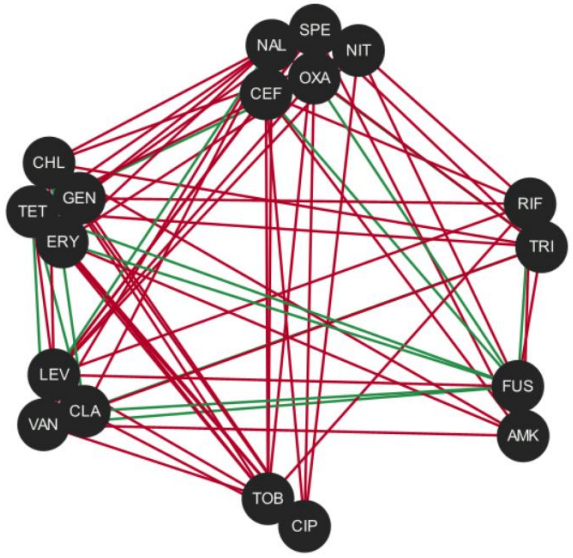
Integrating Multi-Source Drug Information

- Assignment 1, we performed clustering based on four types of drug information:
  - ✓ Chemical structure
  - ✓ Mechanism of action
  - ✓ Bacterial growth curves
  - ✓ ATC codes
- The clustering results show both: Common patterns and Distinct patterns
- How can we integrate multi-source drug information to obtain better clustering results?



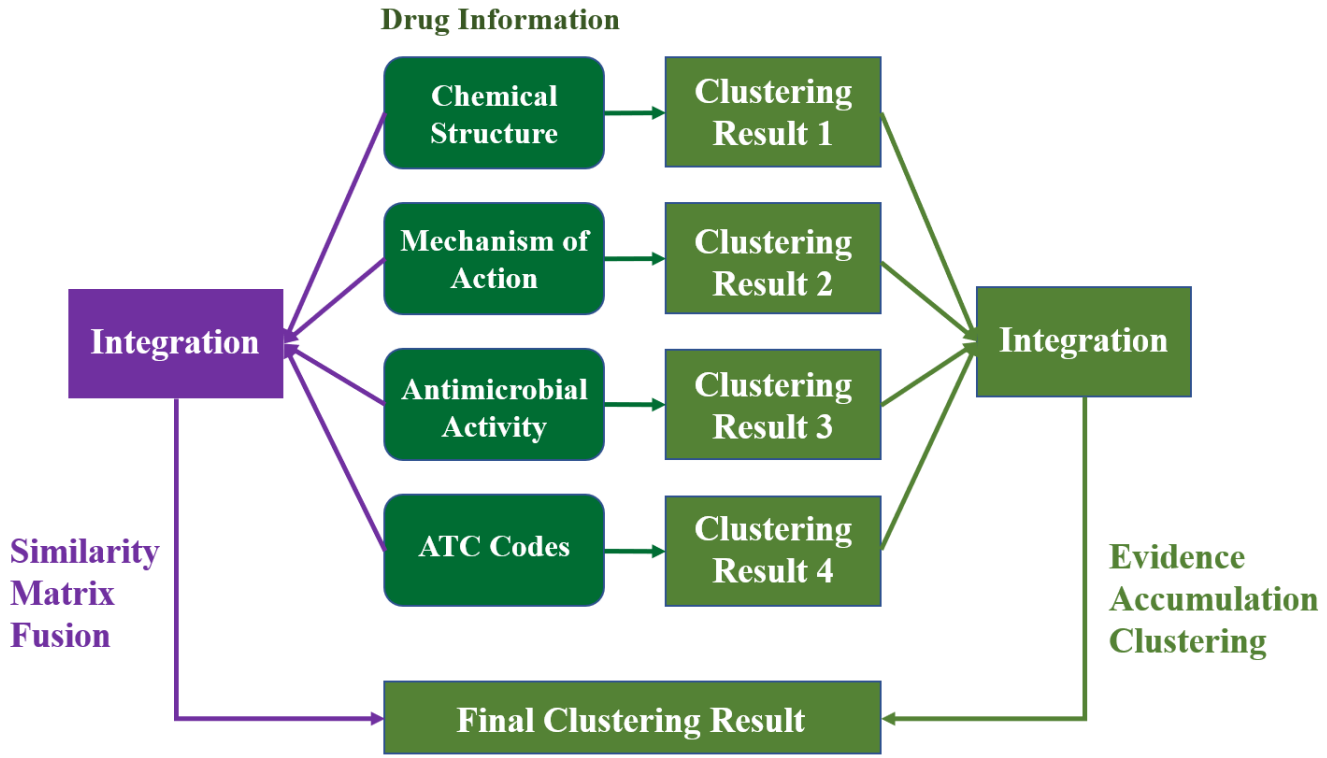






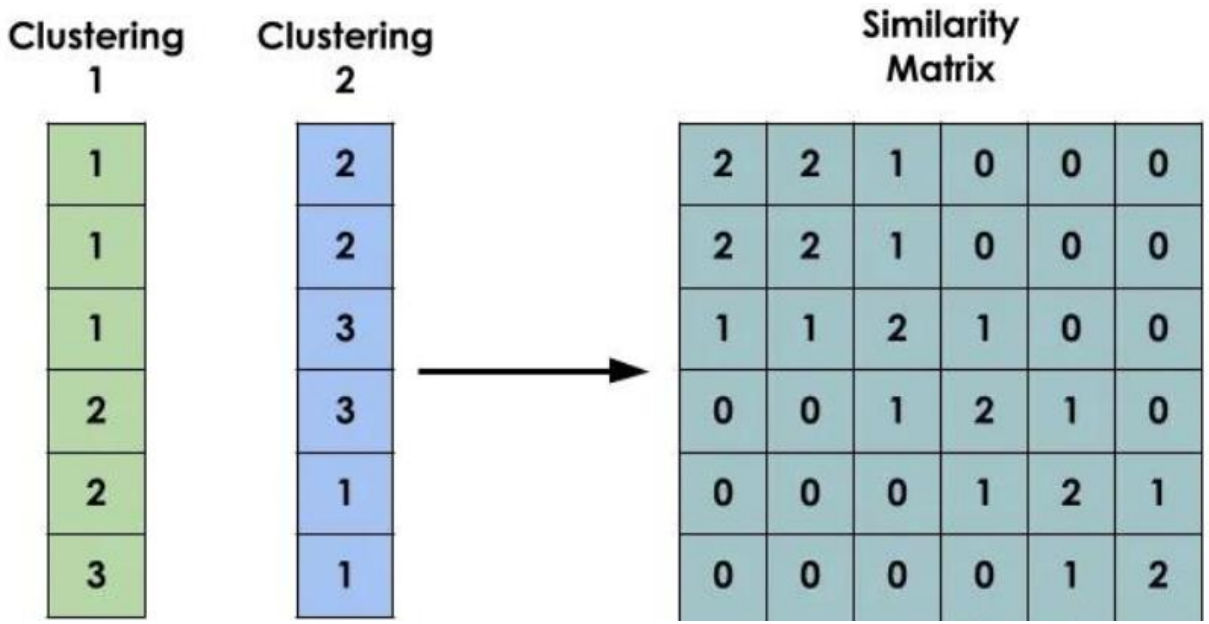
Integrating Multi-Source Drug Information

- To improve clustering performance, we can integrate multiple types of drug information by
  - ✓ Consensus Clustering
  - ✓ Similarity Matrix Fusion



Consensus Clustering

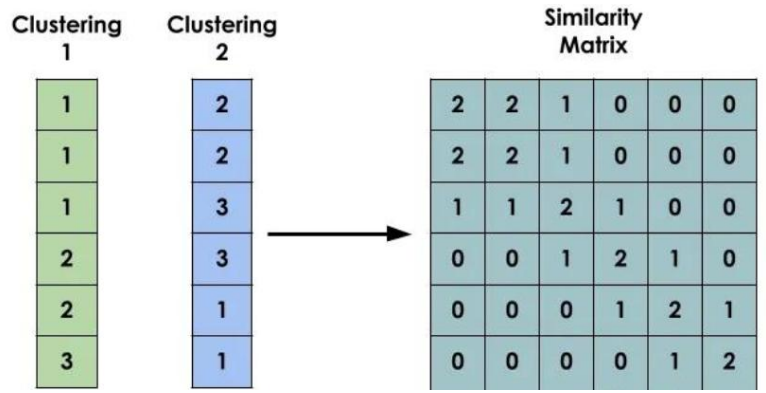
- Perform clustering separately on each drug information
- Then combine clustering results to obtain a consensus clustering
- Methods:
  - ✓ Voting-based clustering
  - ✓ Co-association matrix
  - ✓ Ensemble clustering



IEEE Transactions on Pattern Analysis and Machine Intelligence, **2005**; *27*(6): 835-850.

Consensus Clustering

- How Can We Use LLMs to Help Write Code for Consensus Clustering?
- Consensus clustering is a relatively complex task
  - ✓ It is more effective to ask an LLM to solve it step by step



- Task Decomposition
  - ✓ Step 1: Load clustering results from different similarity types
  - ✓ Step 2: Construct a co-association matrix
  - ✓ Step 3: Measure how often two drugs are assigned to the same cluster
  - ✓ Step 4: Apply a clustering algorithm to the co-association matrix
  - ✓ Step 5: Output the final consensus clustering results

Similarity Matrix Fusion

- **Combine similarity matrices from different drug information:**
  - ✓ **Structural similarity**
  - ✓ **Pharmacological similarity**
  - ✓ **Phenotypic similarity**
  - ✓ **Therapeutic similarity**

- **Methods:**
  - ✓ **Average**
  - ✓ **Weighted sum of similarity matrices**
  - ✓ **Similarity Network Fusion (SNF)**

Ji Lv, et al. *Computers in Biology and Medicine*, **2023**, 162, 107088.

Similarity Matrix Fusion

Each similarity matrix not only preserves its own information, but can also incorporate information from other similarity matrices

$$\mathbf{S_v^{t+1}} = \sum_{k \neq v} \alpha_k \mathbf{S'_k S_v^t} + (1 - \sum_{k \neq v} \alpha_k)\mathbf{S_v^0}$$

- k: number of similarity matrices.
- α_k: weight of the k-th similarity matrix, all similarity matrices are considered equally important $\alpha_k = 0.25$

Normalization of Similarity Matrix

$$\boldsymbol{S'_k} = \mathbf{D_k^{-1/2} S_k D_k^{-1/2}}$$

- $\mathbf{D_k}$: diagonal matrix
- $D_k(i,i)$: sum of the i-th row of $\mathbf{S_k}$
- Reduce the influence of high-weight nodes

Ji Lv, et al. *Computers in Biology and Medicine*, **2023**, 162, 107088.

Similarity Matrix Fusion

- **After 20–30 iterations, the similarity matrix, S_v^t converges to a stable solution**
- **Final Fused Similarity Matrix**

$$S_F = \frac{\sum_v \widetilde{S_F}}{m} \qquad \widetilde{S_F} = D^{-1/2} S_F D^{-1/2}$$

- S_F : fused similarity matrix
- m: number of similarity matrices
- $D = \mathbf{diag}(S_{11}, S_{22}, \dots)$

- **Iterative updates allow each similarity matrix to**
  - ✓ **exchange information with others**
  - ✓ **progressively refine its representation**

Ji Lv, et al. *Computers in Biology and Medicine*, **2023**, 162, 107088.

Similarity Matrix Fusion

- **How Can We Use LLMs to Help Write Code for Similarity Matrix Fusion?**
- **Similarity matrix fusion is a multi-step task**
  - ✓ **It is more effective to guide an LLM with a structured workflow**

- **Task Decomposition**
  - ✓ **Step 1: Load multiple similarity matrices from different modalities**
  - ✓ **Step 2: Normalize each similarity matrix**
  - ✓ **Step 3: Assign weights to different matrices**
  - ✓ **Step 4: Fuse the matrices into a unified similarity matrix**
  - ✓ **Step 5: Apply clustering or further analysis to the fused matrix**

Assignment 2

- **Required Tasks**
  - ✓ **Integrate the clustering results from Assignment 1 using a consensus clustering algorithm**
  - ✓ **Integrate the four similarity matrices from Assignment 1 using a similarity matrix fusion method, and perform clustering based on the fused similarity matrix**
  - ✓ **Evaluate the performance of the two ensemble clustering approaches using edge purity**

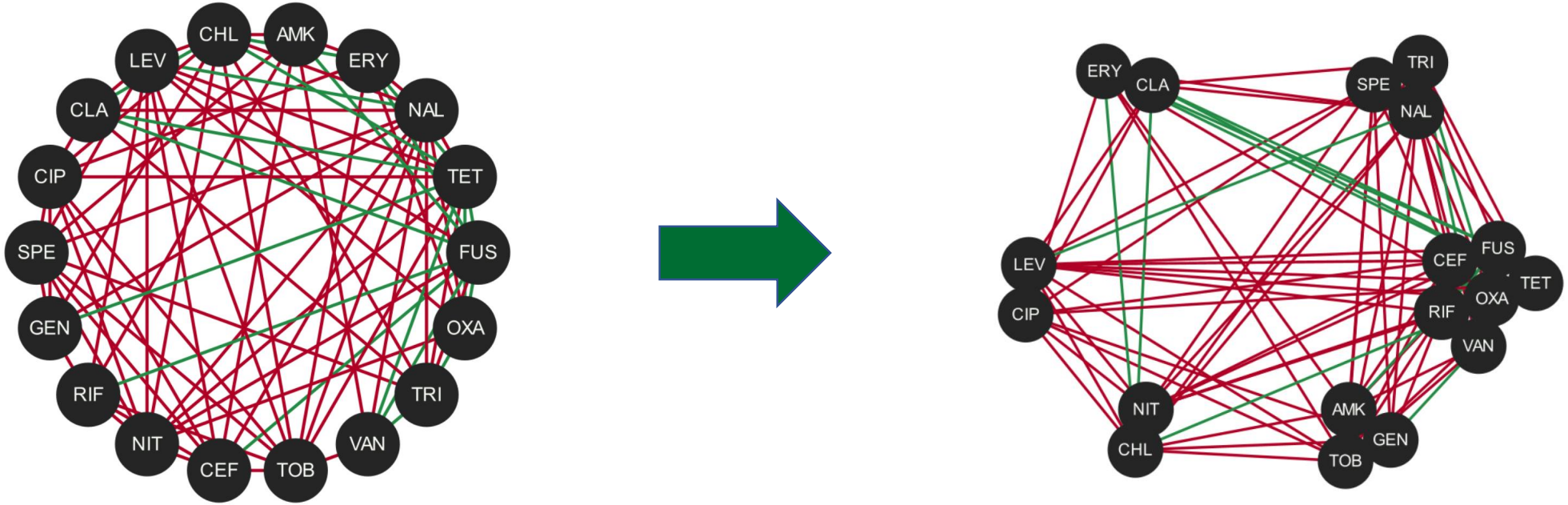
Assignment 2

☐ Optional Tasks

- ✓ **Conduct an ablation study to evaluate how removing one type of drug information (e.g., chemical structure) affects clustering performance, thereby identifying the most important drug feature or feature combination**
- ✓ **Assign three new antibiotics (kanamycin, penicillin G, and roxithromycin) to the clustered DDI network, and predict their interactions with the 18 antibiotics included in the clustering analysis**

Drug Similarity in DDI Networks

- In Assignments 1 and 2, we focused on clustering based on **drug-specific information**
  - ✓ e.g., chemical structure, mechanism of action, phenotypic response, ATC codes
- In this assignment, we shift our focus to **network topology**
- We compute drug similarity in DDI networks



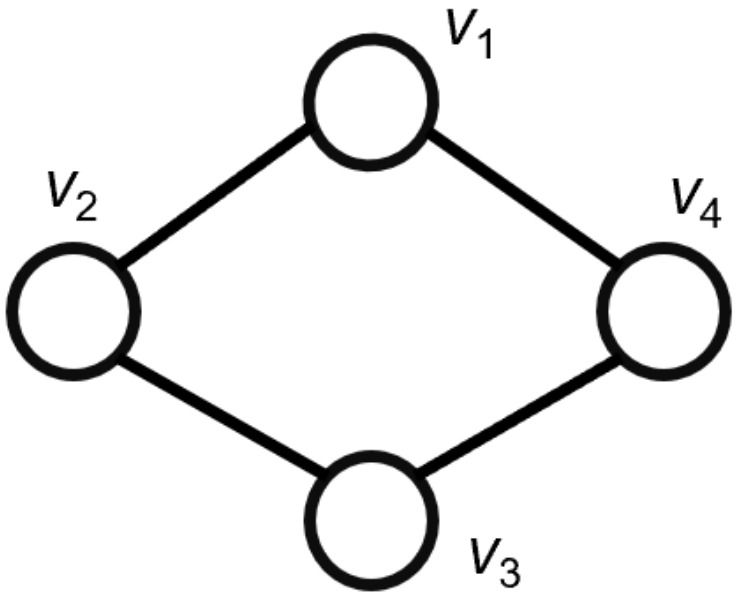
Ji Lv, et al. *Quantitative Biology*, 2024, 12(2), 164-172.

Drug Similarity in DDI Networks

- **In Unsigned Networks**
  - ✓ **If two nodes share similar neighbors, they are considered similar**
  - ✓ **drug similarity can be defined based on neighbor overlap**

$$S_{ij} = \frac{|N(i) \cap N(j)|}{\sqrt{|N(i)||N(j)|}}$$

- **This similarity is based on the local structure of the network**



If nodes i and j share a common neighbor k:

$$a_{ik}a_{kj} = 1$$

Therefore

$$S_{ij} = \frac{\sum_k a_{ik}a_{kj}}{\sqrt{\sum_k a_{ik}^2}\sqrt{\sum_k a_{jk}^2}} \qquad S_{ij} \in [0,1]$$

Cosine Similarity

Drug Similarity in DDI Networks

- However, DDI networks are signed networks
- Edges can represent
  - ✓ Synergistic effects (+)
  - ✓ Antagonistic effects (−)
- Challenge: Traditional similarity measures ignore edge signs
- Signed Neighborhood Decomposition
  - ✓ The neighbors of a node can be divided into two disjoint sets:
  - ✓ $N(i)^+$: neighbors with positive (synergistic) interactions
  - ✓ $N(i)^-$: neighbors with negative (antagonistic) interactions

Ji Lv, et al. *Quantitative Biology*, 2024, 12(2), 164-172.

Drug Similarity in DDI Networks

- **Signed Common Neighbors**
- **Same-sign neighbors:**

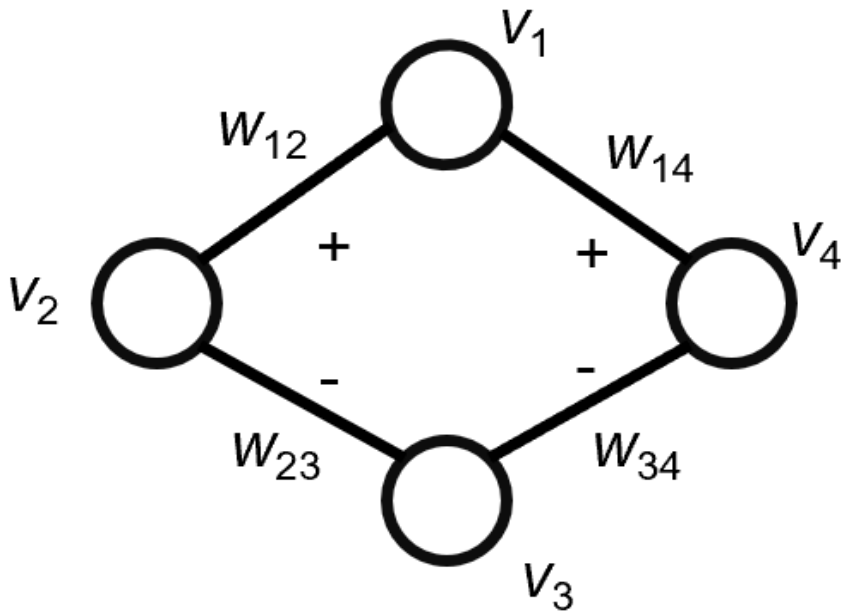
$$S_{ij}^{c} = (N(i)^{+} \cap N(i)^{+}) \cup (N(i)^{-} \cap N(i)^{-})$$

- **Opposite-sign neighbors:**

$$S_{ij}^{i} = (N(i)^{+} \cap N(i)^{-}) \cup (N(i)^{-} \cap N(i)^{+})$$

- **Signed Similarity**

$$S_{ij} = \frac{|S_{ij}^{c}| - |S_{ij}^{i}|}{\sqrt{|N(i)||N(j)|}}$$



- **If nodes i and j share a common neighbor k:**
  - ✓ **Same sign:** $a_{ik}a_{kj} = 1$
  - ✓ **Opposite sign:** $a_{ik}a_{kj} = -1$

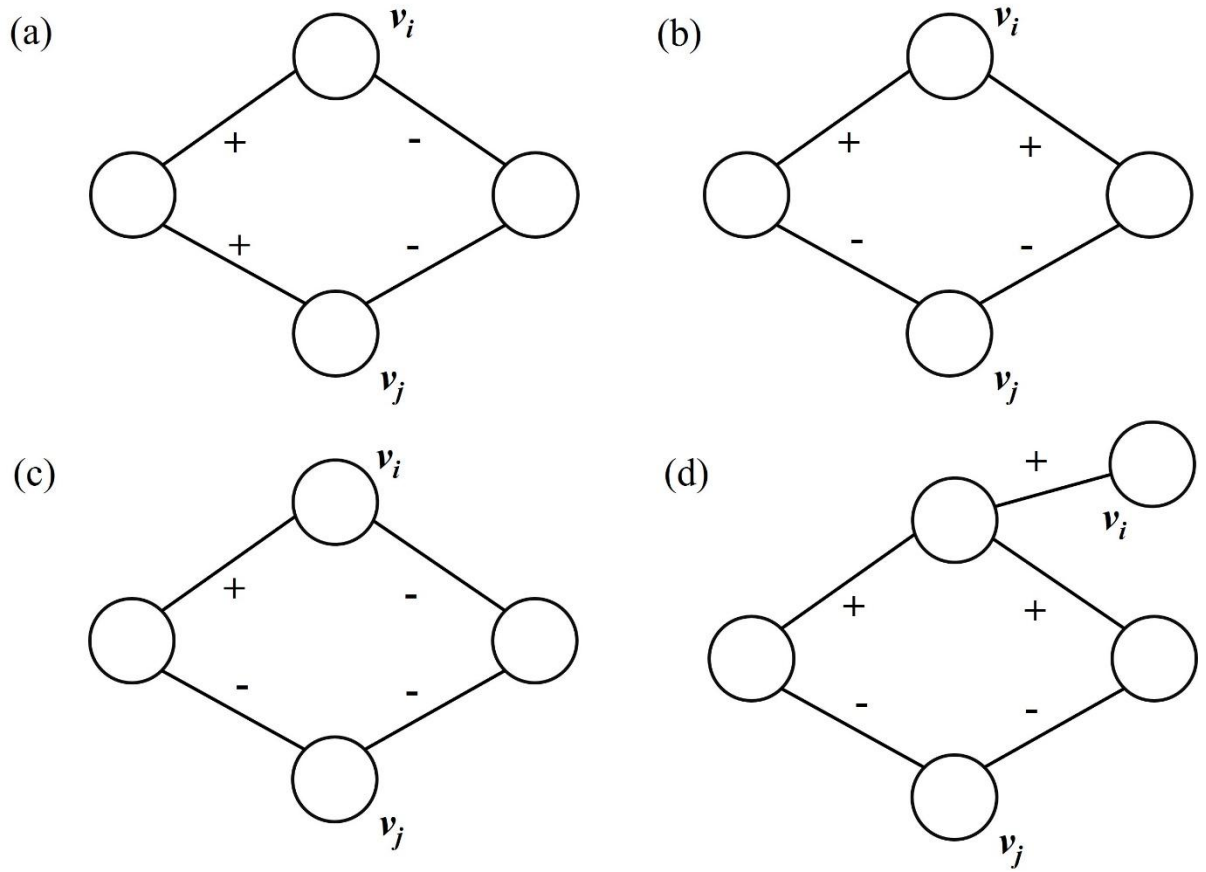
$$\boldsymbol{S_{ij}} = \frac{\sum_k \boldsymbol{a_{ik}a_{kj}}}{\sqrt{\sum_k \boldsymbol{a_{ik}^2}}\sqrt{\sum_k \boldsymbol{a_{jk}^2}}} \qquad S_{ij} \in [-1,1]$$

Cosine Similarity

Ji Lv, et al. *Quantitative Biology*, 2024, 12(2), 164-172.

Drug Similarity in DDI Networks

$S_{ij}^{signed} = 1$ $S_{ij}^{signed} = -1$



$S_{ij}^{signed} = 0$ $S_{ij}^{signed} = 0$

Ji Lv, et al. *Quantitative Biology*, 2024, 12(2), 164-172.

Drug Similarity in DDI Networks

- ☐ **How Can We Use LLMs to Help Write Code for Network-Based Similarity?**
- ☐ **Network-based similarity is a multi-step task**
  - ✓ **It is more effective to guide an LLM with a clear and structured workflow**
- ☐ **Task Decomposition**
  - ✓ **Step 1: Construct the DDI network**
    - **Nodes represent drugs**
    - **Edges represent drug–drug interactions**
  - ✓ **Step 2: Represent the network as an adjacency matrix**
  - ✓ **Step 3: Compute pairwise similarity between drugs**
  - ✓ **Step 4: Output the final similarity matrix for downstream clustering**

Ji Lv, et al. *Quantitative Biology*, 2024, 12(2), 164-172.

Assignment 3

☐ **Required Tasks**

1. **Compute drug similarity based on DDI information**
2. **Analyze the correlation between network-based similarity and the four similarity measures introduced in Assignment 1**
   - ✓ **structural similarity**
   - ✓ **pharmacological similarity**
   - ✓ **phenotypic similarity**
   - ✓ **therapeutic similarity**
3. **Explore the characteristics of DDI network clustering and discuss its potential applications in predicting the mechanisms of action of compounds**

Assignment 3

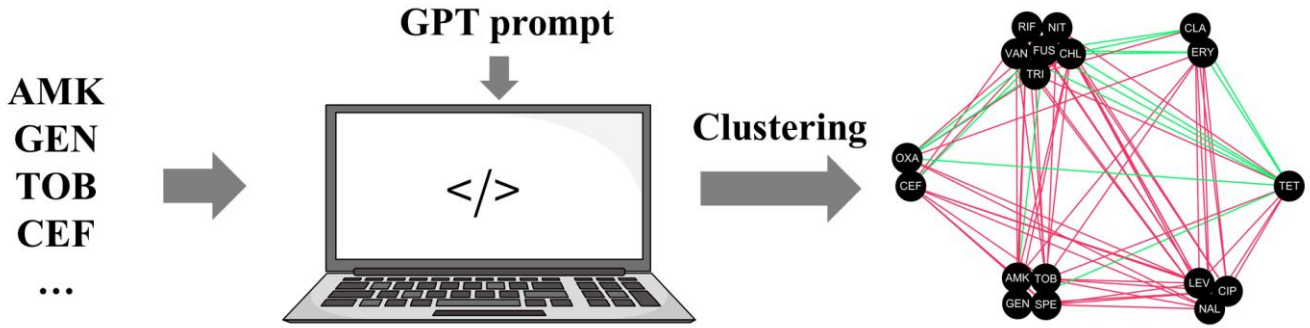
☐ **Optional Task**

1. **Apply the clustering integration methods introduced in Assignment 2 to integrate multi-species DDI information and obtain more robust clustering results**

Thank you

LLM-Based Clustering of DDI Networks

- In Assignments 1, 2, and 3, we clustered DDI networks usingdrug information and DDI information
- However, these workflows typically involve multiple steps: **data collection, similarity computation, clustering analysis, result interpretation**
- As a result, the overall process can be time-consuming and labor-intensive
- Large language models (LLMs) provide a new opportunity
- LLMs, such as ChatGPT and DeepSeek, have demonstrated strong capabilities in
  - ✓ natural language understanding
  - ✓ code generation
  - ✓ scientific problem-solving



Clustering drug-drug interaction network using large language models

LLM-Based Clustering of DDI Networks

- **Key Challenge**: How to design an effective prompt for clustering tasks?
- A Good Prompt Consists of Four Components: **Background, Task Description, Constraints, Output Format**

Prompt: There are 18 antibiotics. Dividing these antibiotics into 6 groups according to their mechanism of action. Dividing protein synthesis inhibitors into two classes (30S/50S ribosomal inhibitors). Output in abbreviated and Python dictionary format.

Answer from LLM:

```
{
    "Cell Wall Synthesis Inhibitors": ["VAN", "OXA", "CEF"],           # Vancomycin, Oxacillin, Cefoxitin
    "30S Ribosomal Inhibitors": ["AMK", "GEN", "TOB", "TET", "SPE"],  # Aminoglycosides, Tetracycline, Spectinomycin
    "50S Ribosomal Inhibitors": ["CHL", "CLA", "ERY", "FUS"],          # Chloramphenicol, Macrolides, Fusidic acid
    "DNA Synthesis Inhibitors": ["CIP", "LEV", "NAL", "NIT"],          # Quinolones, Nitrofurantoin
    "RNA Synthesis Inhibitor": ["RIF"],                                # Rifampicin
    "Folate Pathway Inhibitor": ["TRI"]                                # Trimethoprim
}
```

Ji Lv, et al. *Current Bioinformatics*, **2026**.

Assignment 4

- **Required Tasks**
- **Using LLMs (e.g., ChatGPT, DeepSeek), perform:**

1. **Clustering based on**
   - ✓ **chemical structure**
   - ✓ **mechanism of action**
   - ✓ **bacterial growth curves**
   - ✓ **ATC codes**
2. **Multi-source integration for clustering**
3. **Network-based clustering using DDI topology**
4. **Compare LLM performance with traditional algorithms**

Assignment 4

☐ **Optional Task**

1. **Ensemble of LLMs**
   - ✓ **Combine outputs from multiple LLMs**
2. **Retrieval-Augmented Generation (RAG)**
   - ✓ **Incorporate external knowledge bases**

Thank you