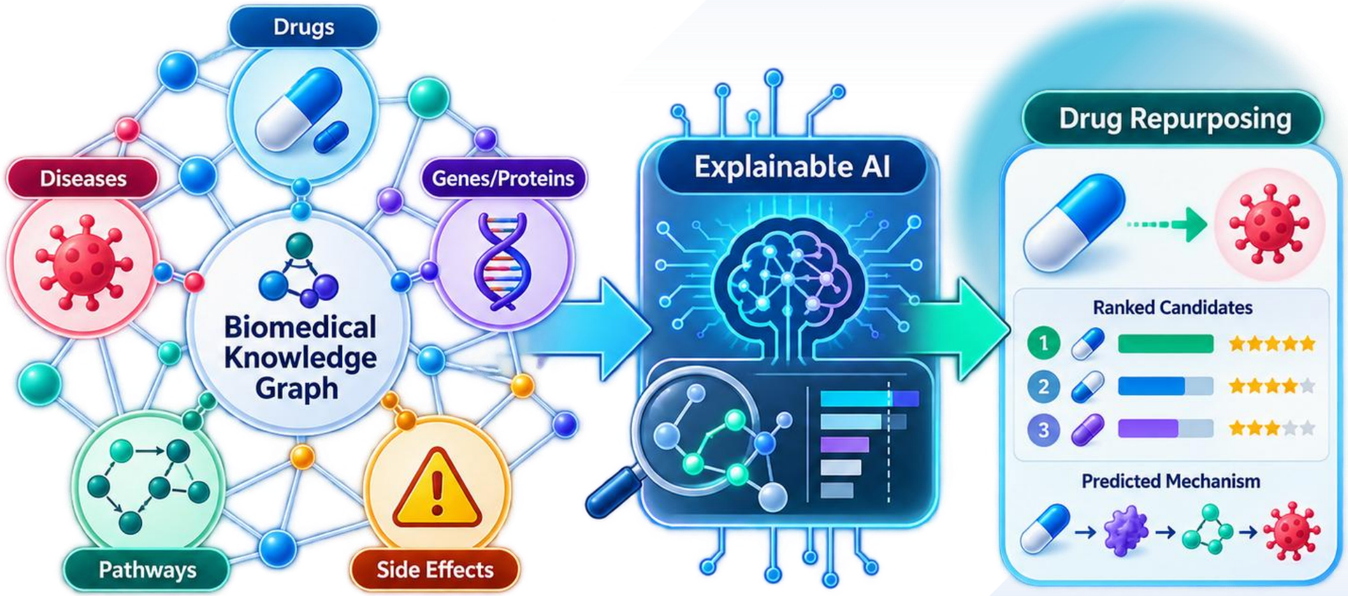
ACTION-BASED HYBRID KNOWLEDGE GRAPH LEARNING FOR DRUG REPURPOSING AND FUTURE DIRECTIONS TOWARD EXPLAINABLE AI

XXXII Symposium on Bioinformatics and Computer-Aided Drug Discovery (BCADD-2026)



Marzieh Khodadadi

Institute of Biochemistry and Biophysics,
Laboratory of Bioinformatics and drug design (LBD)
University of Tehran, Tehran, Iran

What is drug design process?

- Lengthy Process
- High Failure Rate
- Complexity of Underlying Biological
- Translational Failures Using Animal Models
- Patient Heterogeneity
- Collaboration Issues
- Clinical Holds

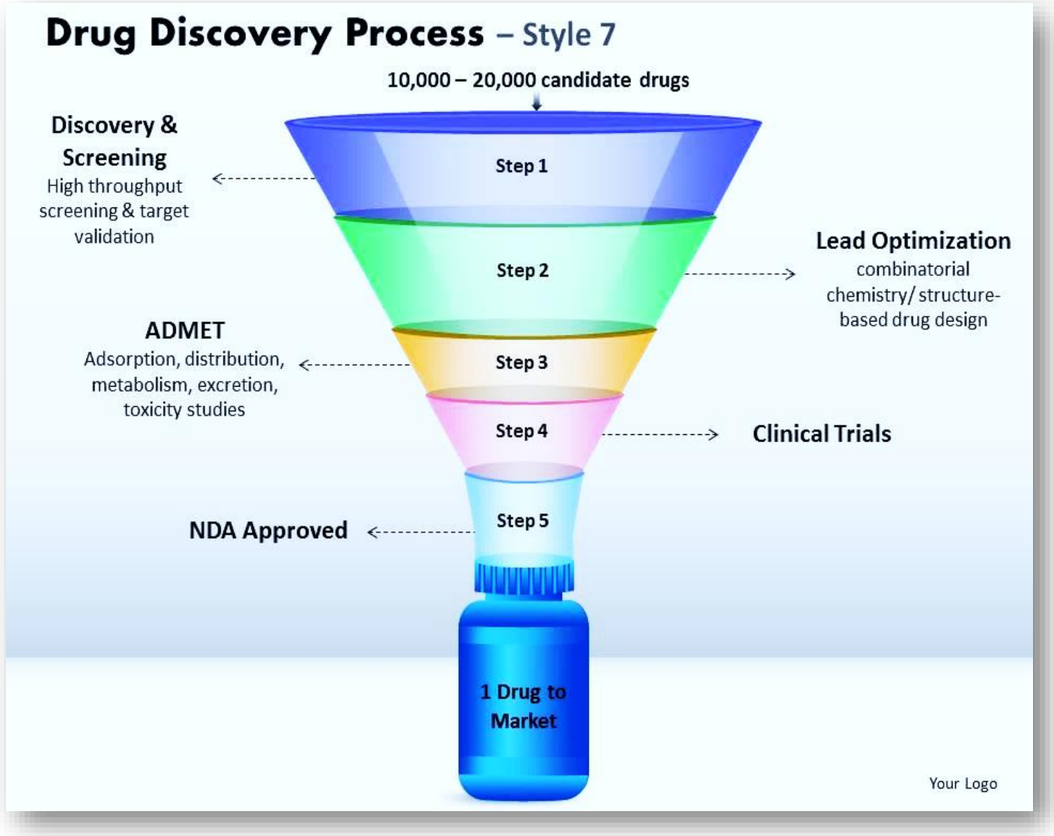
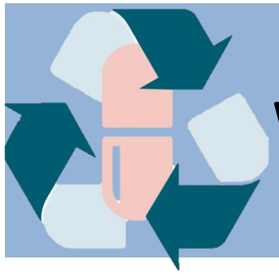


Figure (1): Overview of the Drug Discovery Process



What is Drug Repurposing?

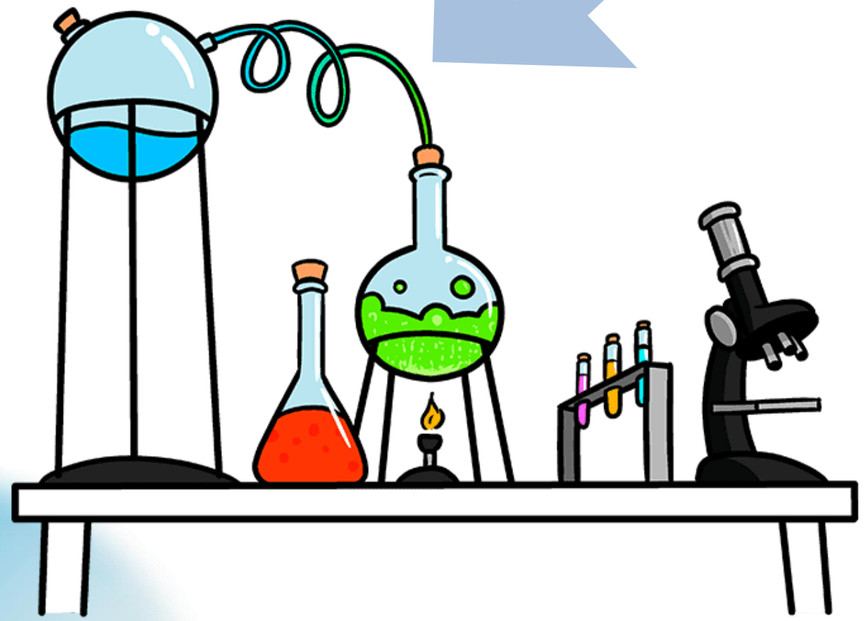
Drug repurposing is the process of discovering new indications (i.e., diseases or conditions) for already approved drugs

Rare Disease
Pandemic

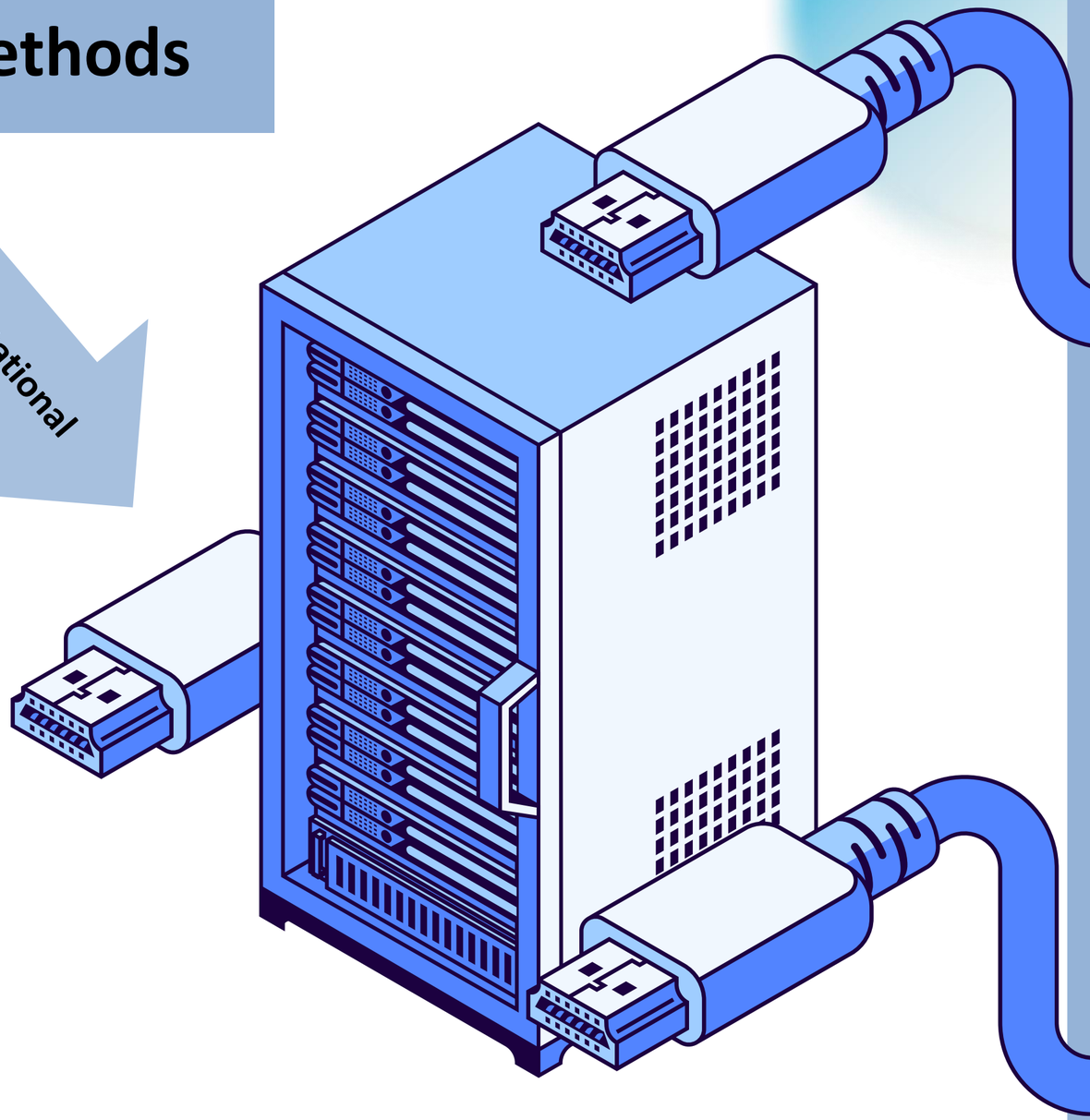
- Cost Reduction
- Time Savings
- Reuse Despite Past Issues
- Data Utilization
- Risk Minimization
- Direct Application to Preclinical Models
- Efficiency in Treating Rare Conditions
- Lower Costs for Clinical Trials

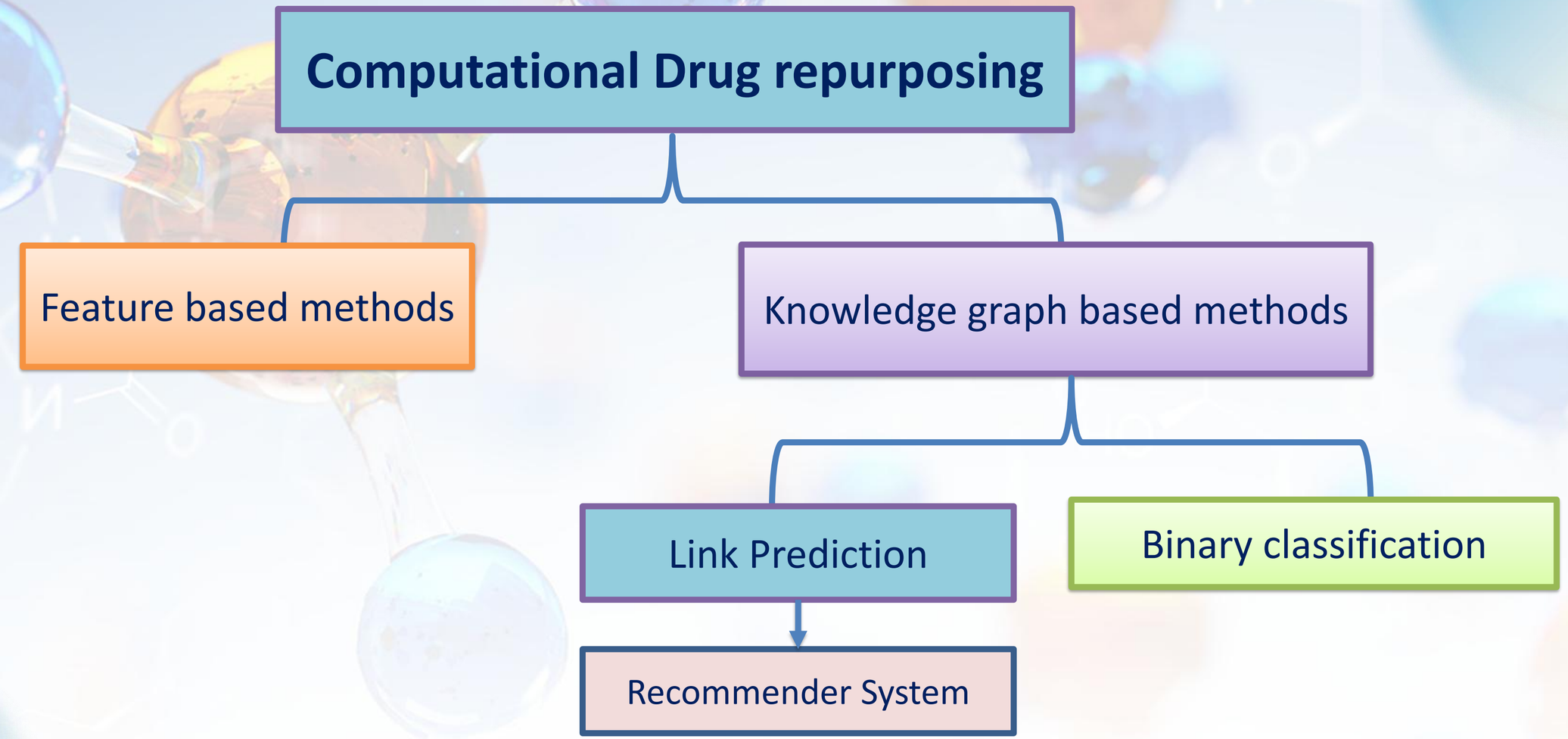
Drug Repurposing Methods

Experimental



Computational




Computational Drug repurposing
Feature based methods
Knowledge graph based methods
Link Prediction
Binary classification
Recommender System

Feature based Methods

DeepDRK

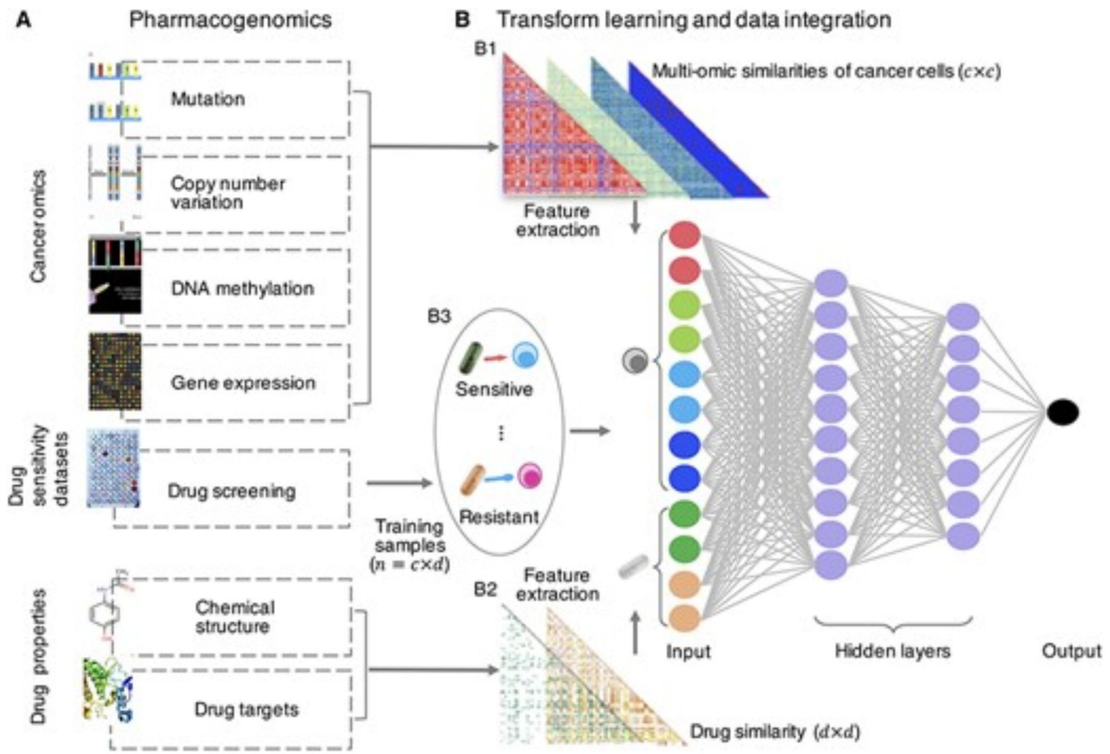


Figure (2): Overview of DeepDRK model
DOI: 10.1093/bib/bbab048

DeepLPI

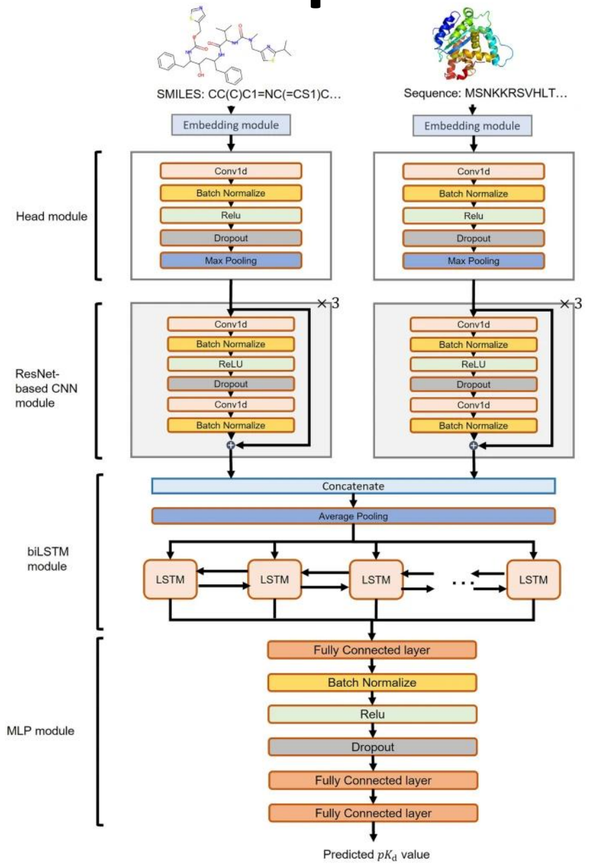


Figure (3): Overview of DeepLPI model
DOI: 10.1038/s41598-022-23014-1

Disadvantages of Feature-Based Modeling (Structure-Based)

- ⚠ Limited Representational Power
- ⚠ Poor Integration of Heterogeneous Data
- ⚠ Sparsity and Missing Information
- ⚠ Lack of Semantic Relationships
- ⚠ Scalability Issues
- ⚠ Static Modeling

Necessity for Knowledge Graph Based Modeling

- Integrates Multi-Source Biomedical Data
- Captures Semantic and Relational Context
- Supports Inference and Discovery

Advantages of Knowledge Graph-Based Approach

- ☑ Rich, Flexible Representation
- ☑ Improved Prediction and Interpretability
- ☑ Dynamic and Scalable
- ☑ Enables Advanced Techniques

Knowledge graph

A knowledge graph ***G=(V, E, R)*** is a **directed graph:**

•A set of **entities *V*** (nodes),

•A set of **relations *R***,

•A set of **directed edges *E*** linking entities via relations.

Each edge is a **triple**:

(h , r , t)

where:

•***h*** (head entity) and ***t*** (tail entity) belong to ***V***

•***r*** (relation) belongs to ***R***

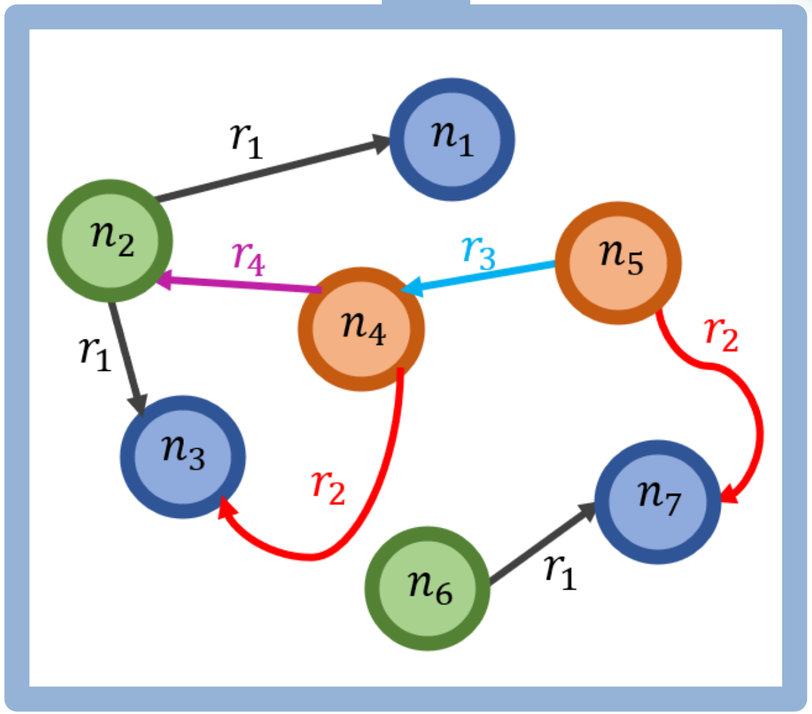


Figure (4): Knowledge Graph Illustrating Multi-Relational Entity Connections

Knowledge graph based methods

EKGDR

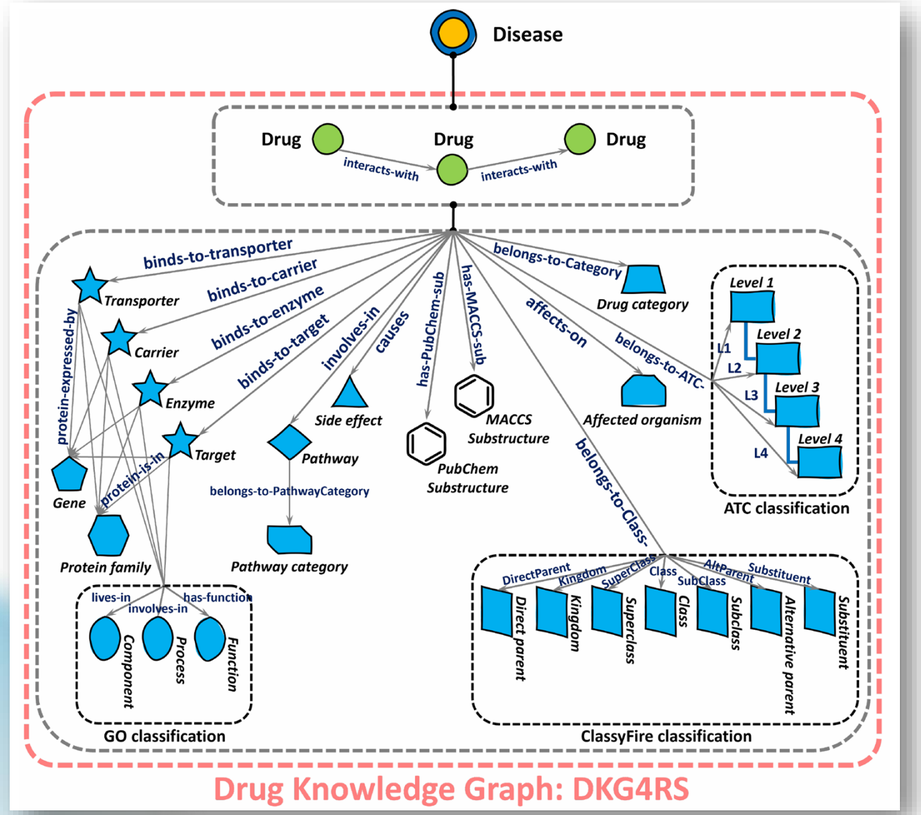


Figure (5): Overview of EKGDR model
https://doi.org/10.1021/acs.jcim.3c01925

DrugRep-HeSiaGraph

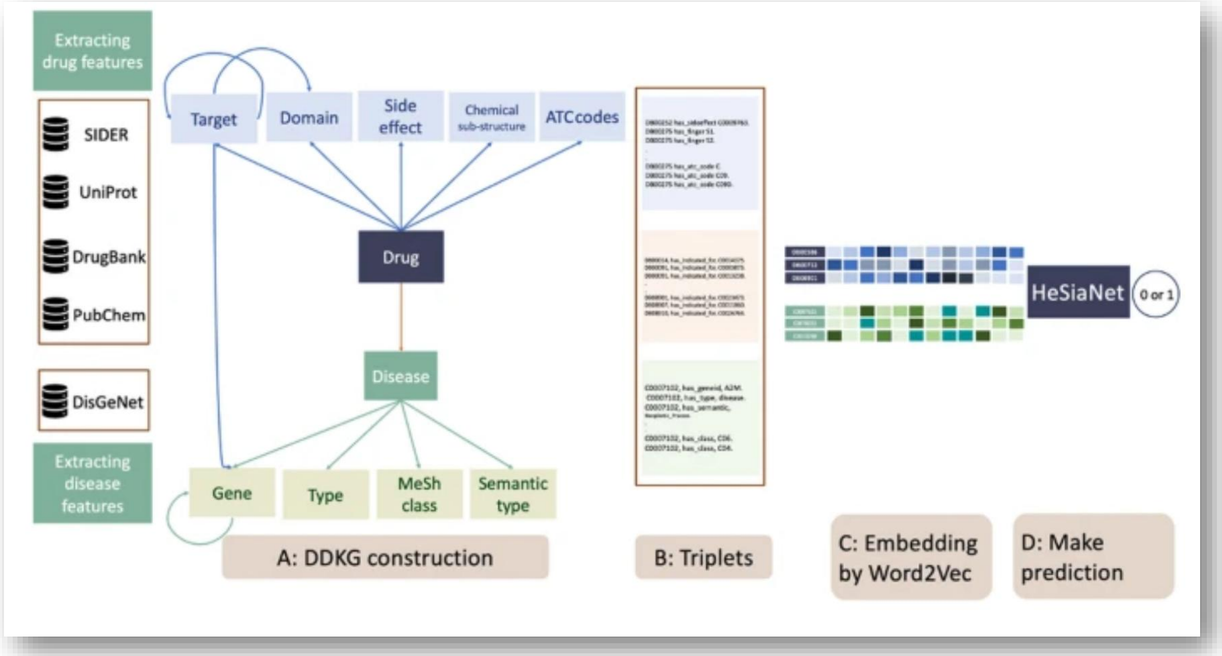
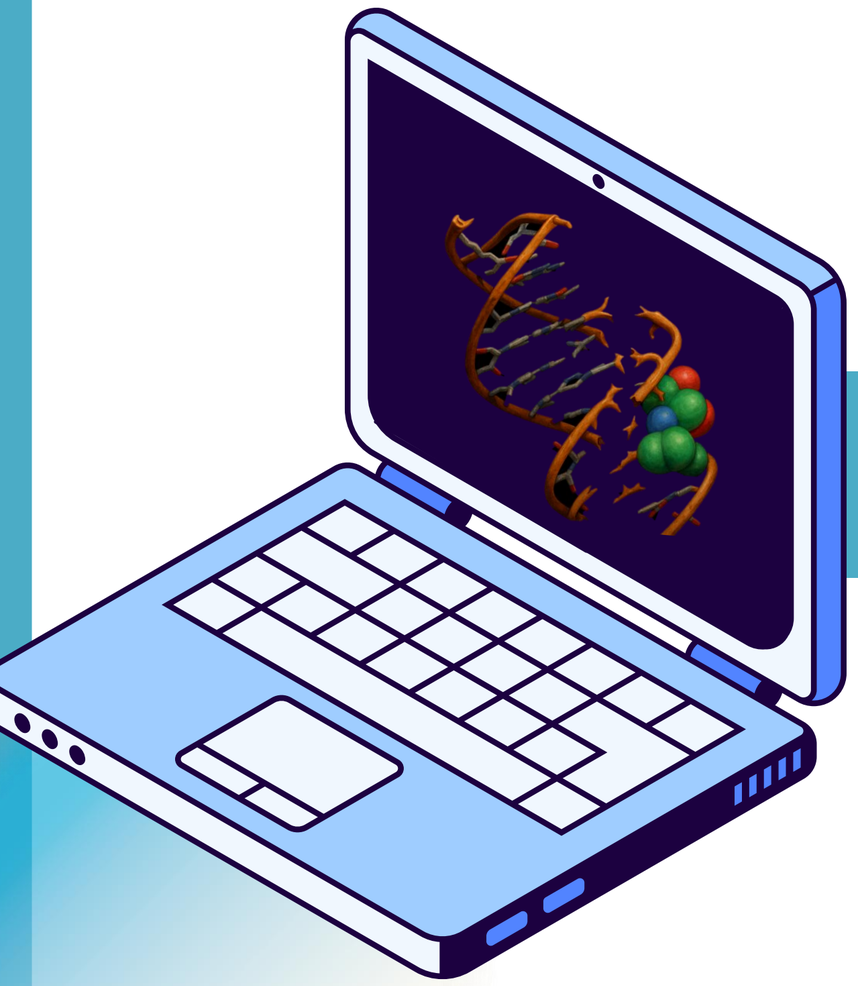


Figure (6): Overview of DrugRep-HesiaGraph model
https://doi.org/10.1186/s12859-023-05479-7



Method

Graphical Abstract

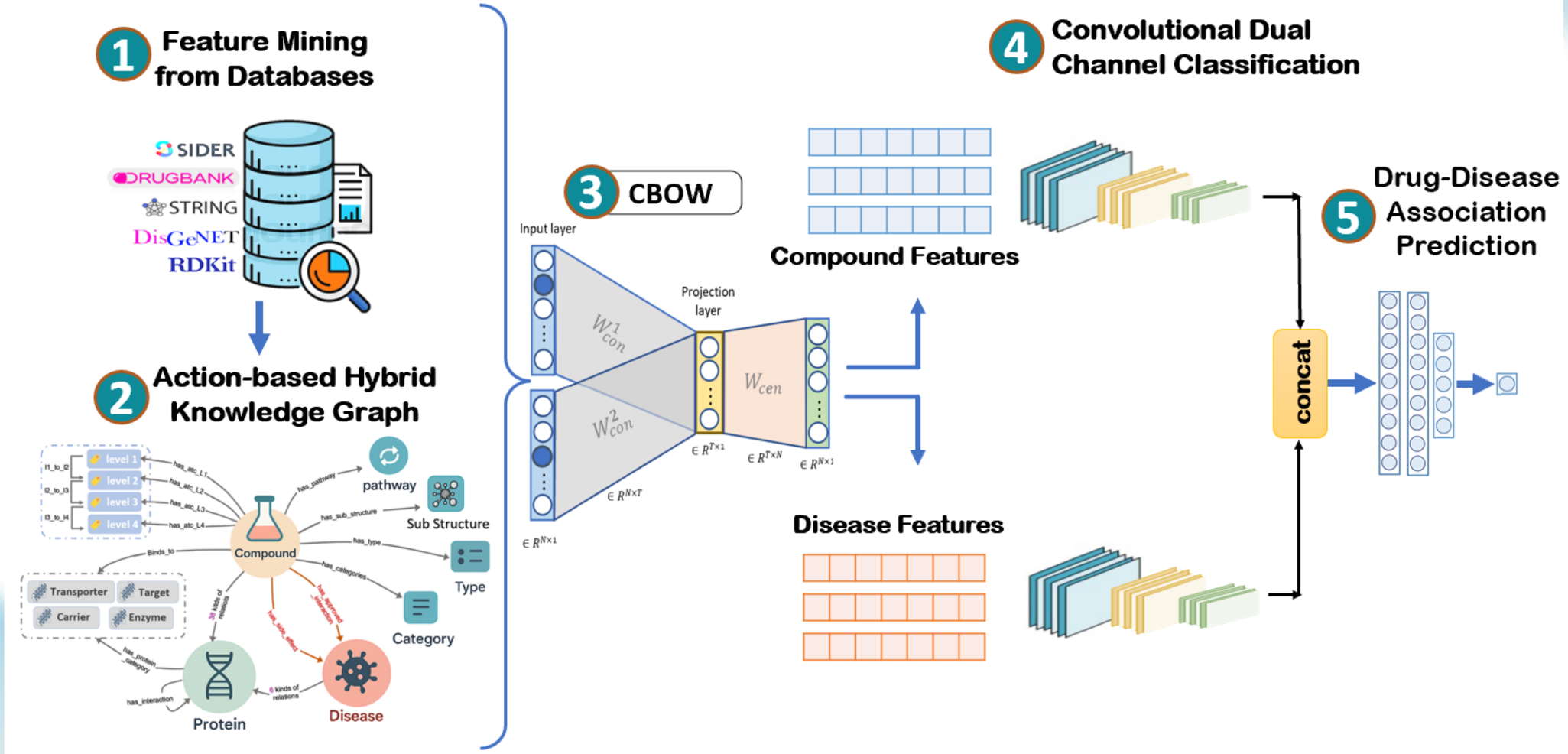


Figure (7): Graphical abstract of ConvAHKG model

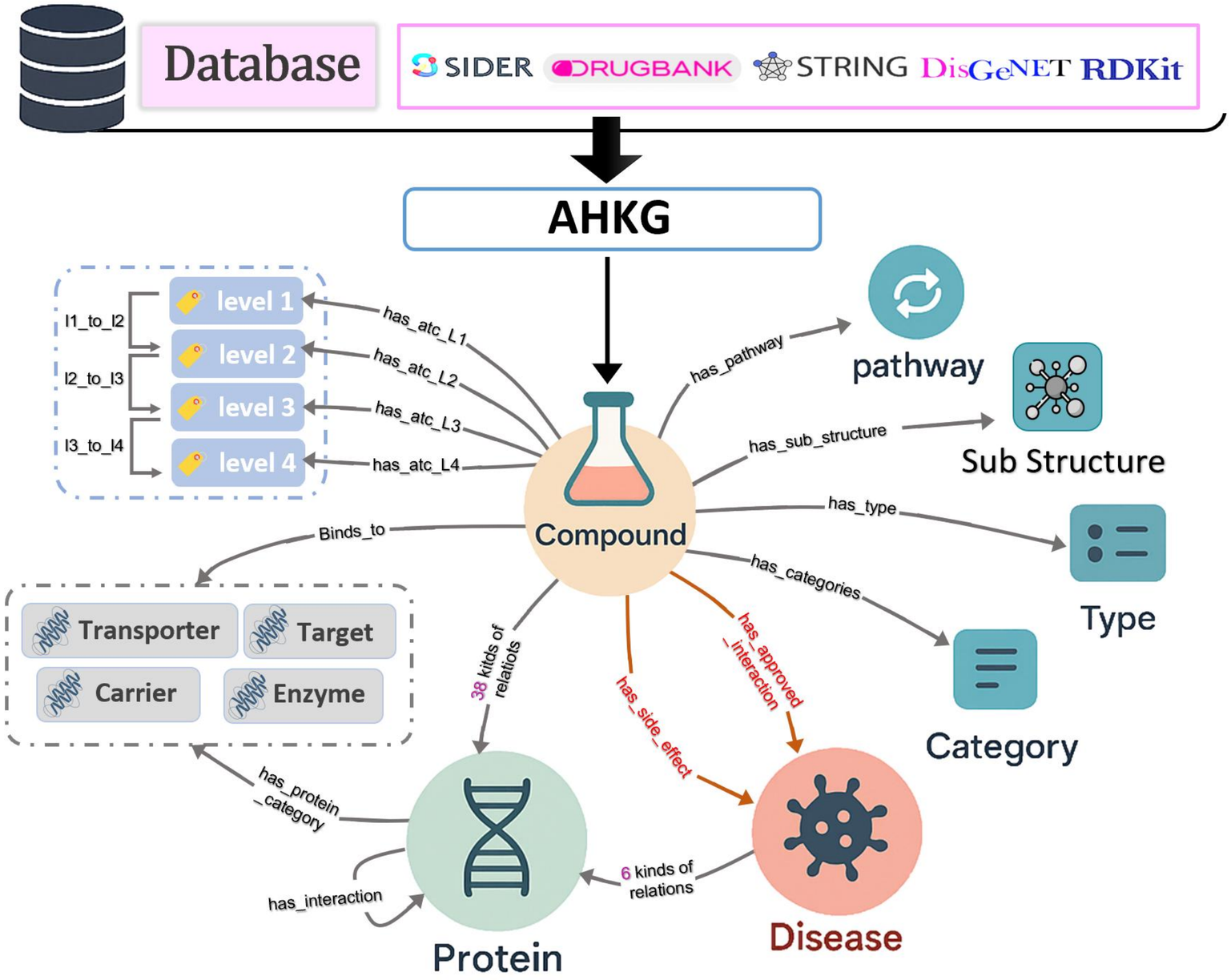


Figure (8): AHKG is a heterogeneous hybrid graph constructed for drug repurposing, integrating multiple biomedical databases to model complex relationships among compounds, proteins, and diseases.

Compound-Protein Interactions

- **Effect-Based Classification**:
  - Agonist, Binding, Inhibition, etc.

- **Bond Type Classification**:
  - **Target Bond**: Alters protein function.
  - **Enzyme Bond**: Modifies enzyme activity.
  - **Carrier Bond**: Binds transport proteins.
  - **Transporter Bond**: Affects membrane transport proteins.

STRING: Protein-protein interactions (threshold > 0.8)

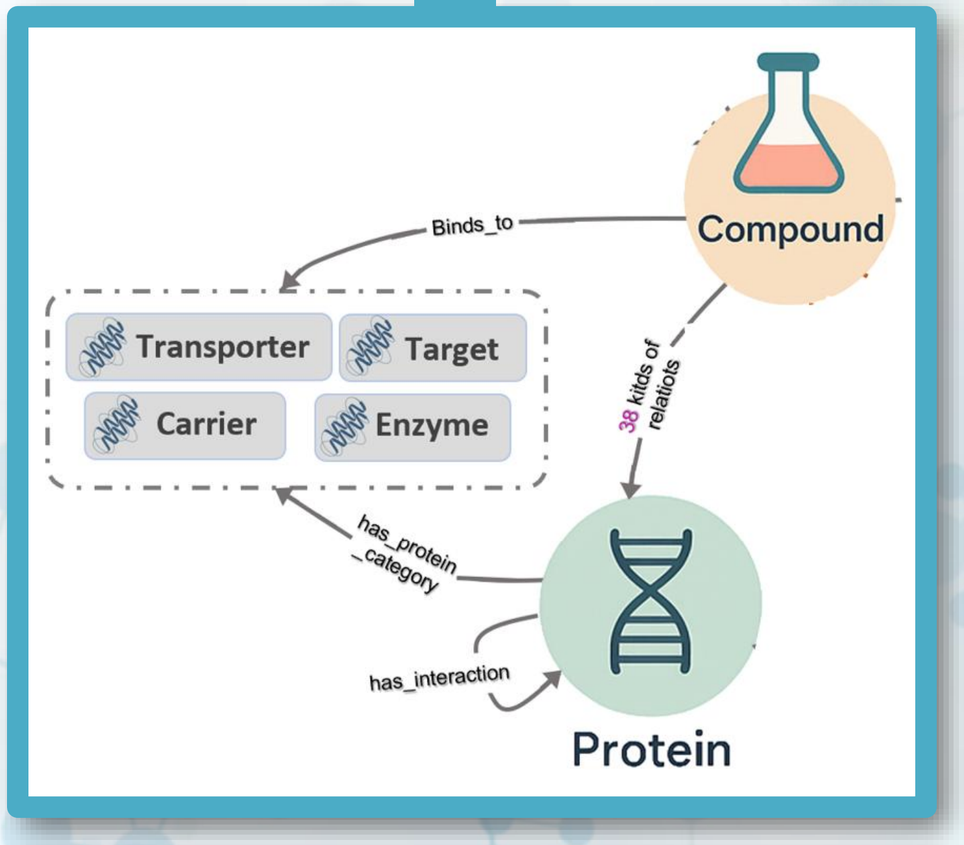


Figure (9): Compound-Protein Interactions in AHKG

Disease-Protein Interactions

- Uses multiple **relations** such as:
  - Biomarkers
  - Altered Expression
- Total of **6 relation types** in the graph

Compound-Disease Relationships

- Two primary relationship types:
  - ***"Has Side Effect"***
  - ***"Has Approved Interaction"***

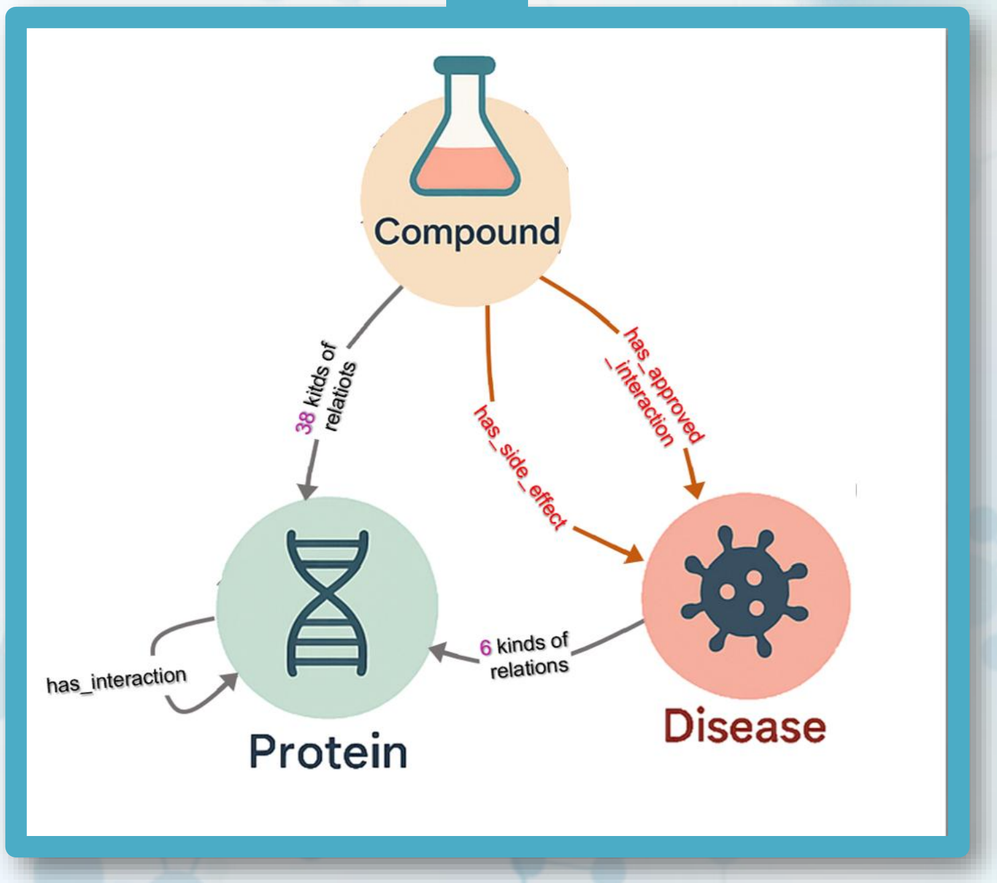


Figure (10): Compound-Protein- disease Interactions in AHKG

Classification of Drug Categories

ATC Classification in AHKG

•**Four-Level Hierarchy**:

- **Level 1**: Anatomical category.
- **Level 2**: Therapeutic subcategory.
- **Level 3**: Pharmacological subcategory.
- **Level 4**: Chemical characteristics.

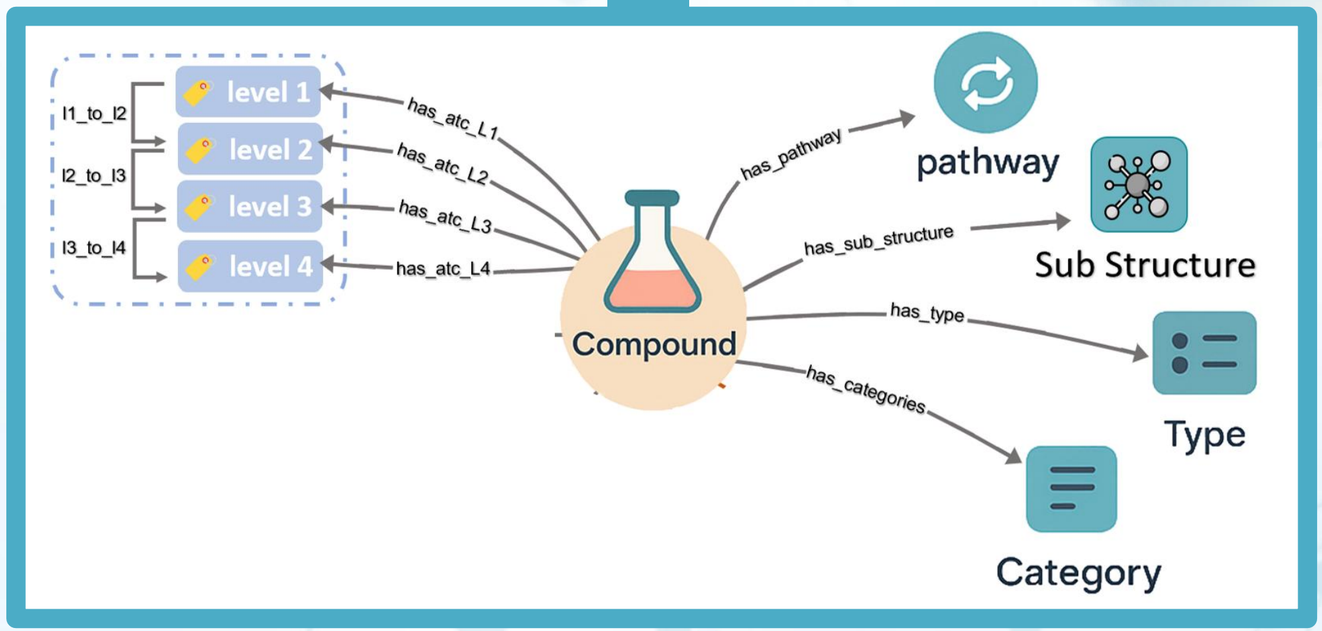


Figure (11): Compound Features

Chemical Substructures

•Uses **Morgan Fingerprints** for molecular analysis.

•**RDKit toolkit** extracts **1024-bit substructures**.

•Each substructure is represented as a **separate node** in the graph.

Method

Summary of Datasets with Node Types, Sizes, and Database Sources

Node Types	Remarks	Node Size	Database
Compound		1509	RepoDB
Disease		1229	RepoDB
Protein	The protein-protein interactions were derived from the STRING database.	2995	Drug Bank
Compound_Type	Including small molecules and biotech products	2	Drug Bank
Compound_Category		2722	Drug Bank
ATC Code	The ATC code consists of four levels: anatomical main group, therapeutic main group, pharmacological subgroup, and chemical subgroup.	1389	Drug Bank
Side Effect		352	SIDER
Pathway		650	Drug Bank
Sub_Structure	Morgan fingerprints consisting of 1024 substructures.	1024	RDKit
Bond Types	Enzymes, Transporters, Receptors, Target.	4	Drug Bank

Table (1): Summary of datasets with node types, sizes, and database sources



Methods Overview

- **Objective:** Repurpose drugs using a knowledge graph framework.
- **Challenges:** Binary classification in imbalanced datasets.

Approach:

- Feature extraction through embeddings.
- Classification of drug-disease pairs using dual convolutional channels.
- Handling class imbalance with weighted binary cross-entropy loss.



A Feature extraction

- **Technique:** Word2Vec algorithm
- **Goal:** Embed entities & relationships into a vector space.
- **Process:**
- Convert triples *(h, r, t)* into sentence-like structures.

h+" " +r+" " +t+"."

- Learn vector representations using Continuous CBOW
- Predict center word based on context in the KG

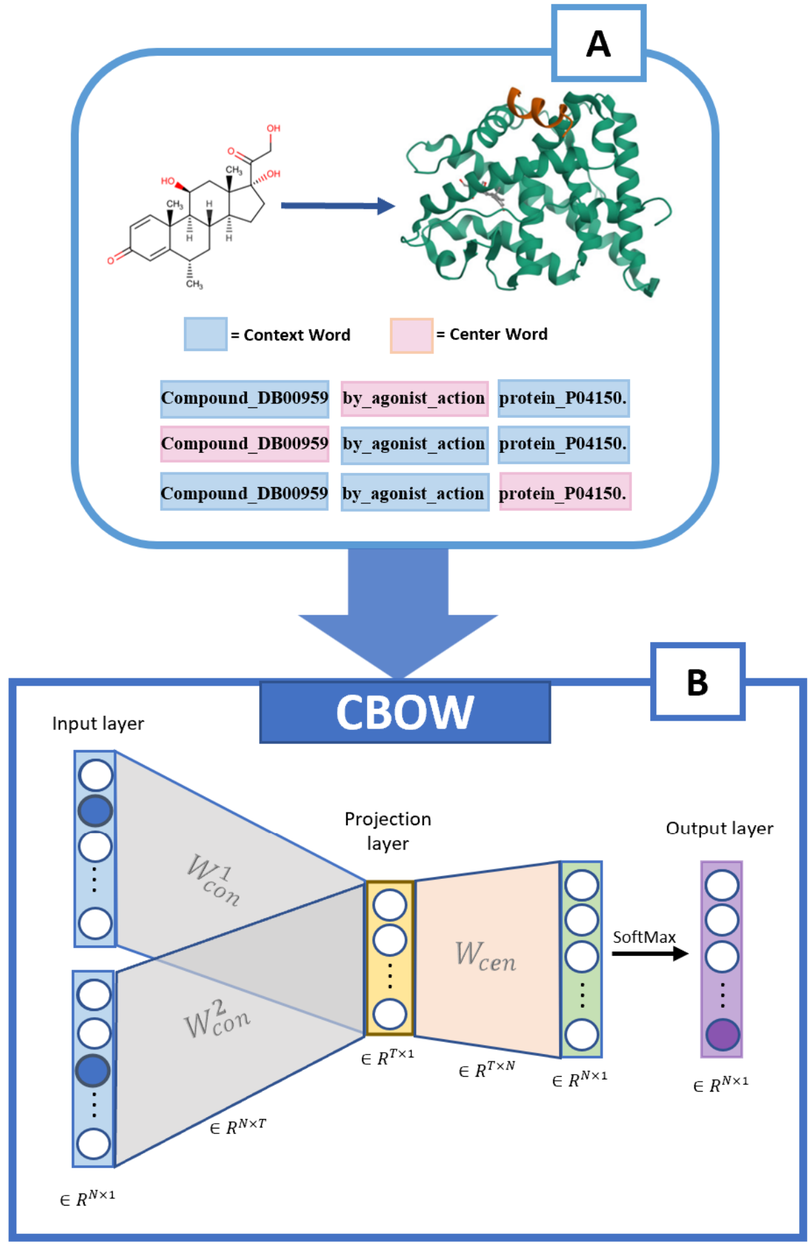


Figure (12): Feature Extraction Workflow

B Binary Classification



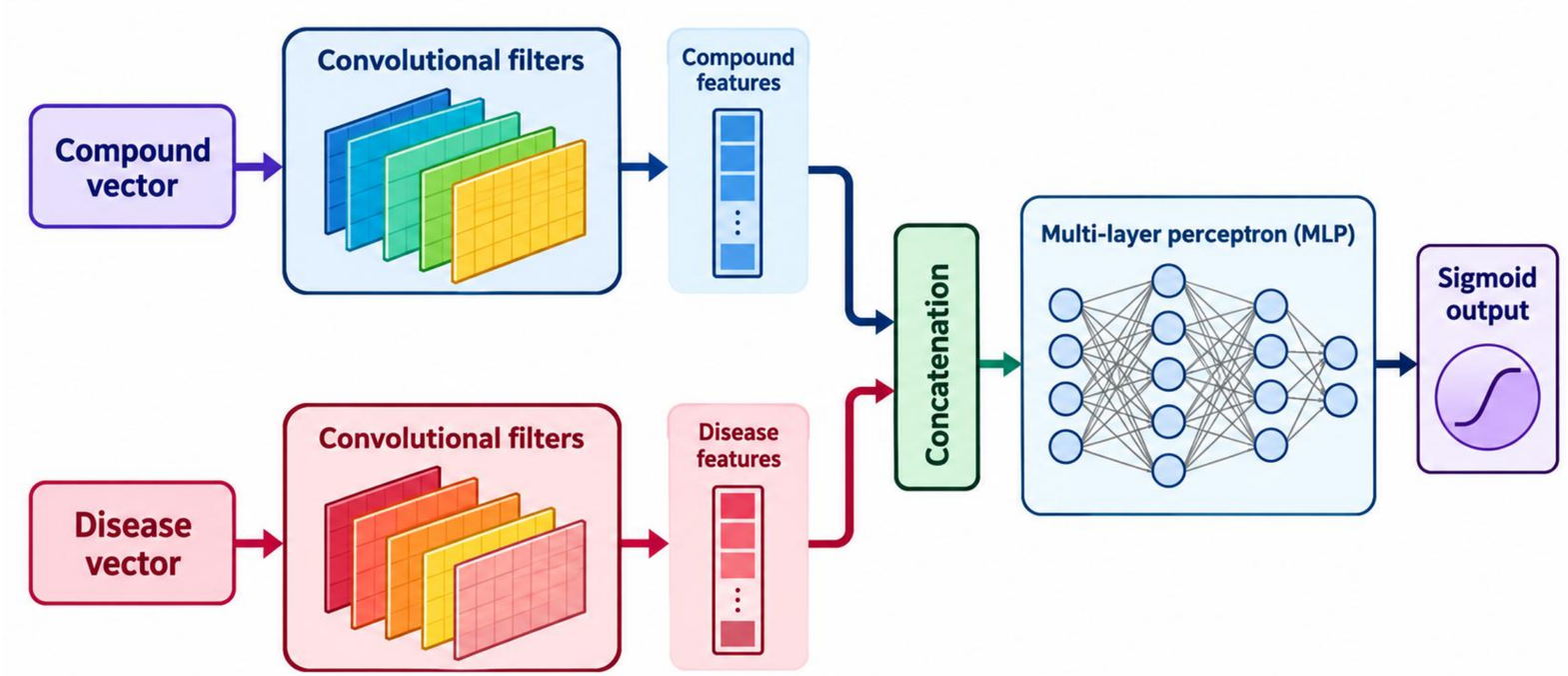


Figure (13): dual-channel 1D convolutional neural network Workflow

C Unbalanced data handling

Positive Pairs → 6657
Negative Pairs →21120

$$\mathrm{BCE}(y_{\mathrm{true}}, y_{\mathrm{pred}}) = -y_{\mathrm{true}} \log(y_{\mathrm{pred}}) - (1 - y_{\mathrm{true}}) \log(1 - y_{\mathrm{pred}})$$

$$\omega_{\mathrm{pos}} = \frac{\sum_{i=1}^{N}(1 - y_{\mathrm{true},i})}{\sum_{i=1}^{N} y_{\mathrm{true},i}} \quad \text{and} \quad \omega_{\mathrm{neg}} = 1$$

$$\mathrm{WBCE\ Loss}(y_{\mathrm{true}}, y_{\mathrm{pred}}) = -\omega_{\mathrm{pos}} y_{\mathrm{true}} \log(y_{\mathrm{pred}}) - (1 - y_{\mathrm{true}}) \log(1 - y_{\mathrm{pred}})$$



Results

Performance of Imbalance Handling Techniques

Results

- **Data Splitting:** 90% training, 10% testing
- **Cross-Validation:** 10-fold stratified K-fold
- **Class Imbalance Techniques:**
  - Random Oversampling *(ROS)*
  - Random Undersampling *(RUS)*
  - Oversampling using SMOTE *(OSMOTE)*
  - SMOTE with Undersampling *(SORU)*
  - Weighted Binary Cross-Entropy *(WBCE)*

Technique	AUC-ROC	AUC-PR	Accuracy	Precision	F1 Score
ROS	0.9743	0.9511	0.9438	0.8947	0.8811
RUS	0.9715	0.9446	0.9388	0.8875	0.8698
OSMOTE	0.9691	0.9456	0.941	0.8838	0.8758
SORU	0.9716	0.9491	0.9366	0.8592	0.8694
WBCE	**0.9812**	**0.9656**	**0.9573**	**0.944**	**0.9074**

Table (2): Performance Metrics for Imbalanced Data Handling Techniques

Performance Comparison of Different Embedding Methods

Model	AUC-ROC	AUC-PR	Overall Recall	Accuracy	Precision	F1 Score	Time (s)
ComplEx	0.9731	0.9499	**0.8526**	0.9460	0.9160	0.8832	2038
RotatE	0.9699	0.9434	0.8316	0.9445	0.9294	0.8778	2354
TransE	**0.9770**	**0.9556**	0.8436	0.9474	0.9303	0.8849	1224
Word2Vec	0.9750	0.9535	0.8423	**0.9525**	**0.9541**	**0.8947**	**272**

Table (3): Performance comparison of different embedding methods.

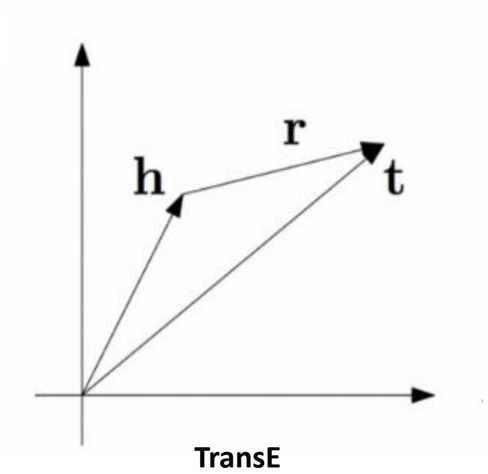
Results

Models Compared:

- **Word2Vec** (Our method)
- **ComplEx**, **RotatE**, **TransE** (KG-based embeddings)

Evaluation Setup:

- Same dataset, 600-dim embeddings, identical computational environment
  - **ComplEx:** $f(h,r,t) = Re(\langle h, r, \overline{t} \rangle)$
  - **RotatE:** $h \circ r \approx t$
  - **TransE:** $h + r \approx t$



TransE

Feature Impact on AHKG Performance

Results

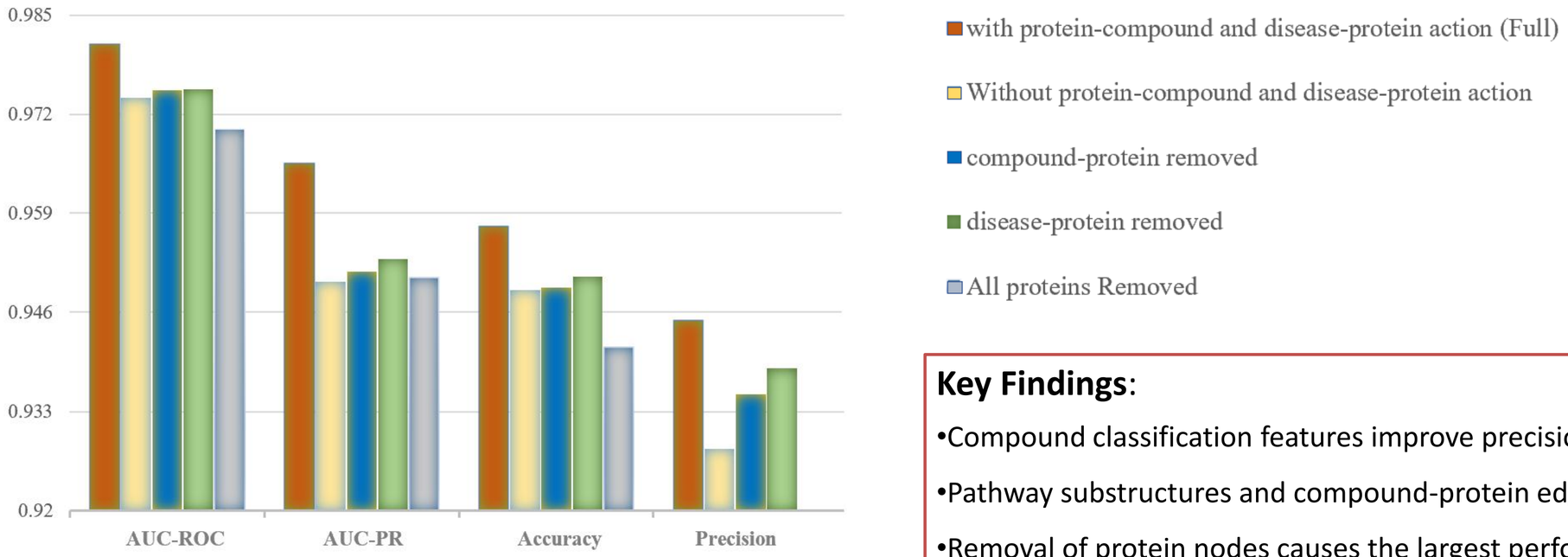


Figure (15): Impact of Protein and Action Relationships on AHKG Performance

Key Findings:

- Compound classification features improve precision and accuracy
- Pathway substructures and compound-protein edges are critical
- Removal of protein nodes causes the largest performance drop

Feature Impact on AHKG Performance

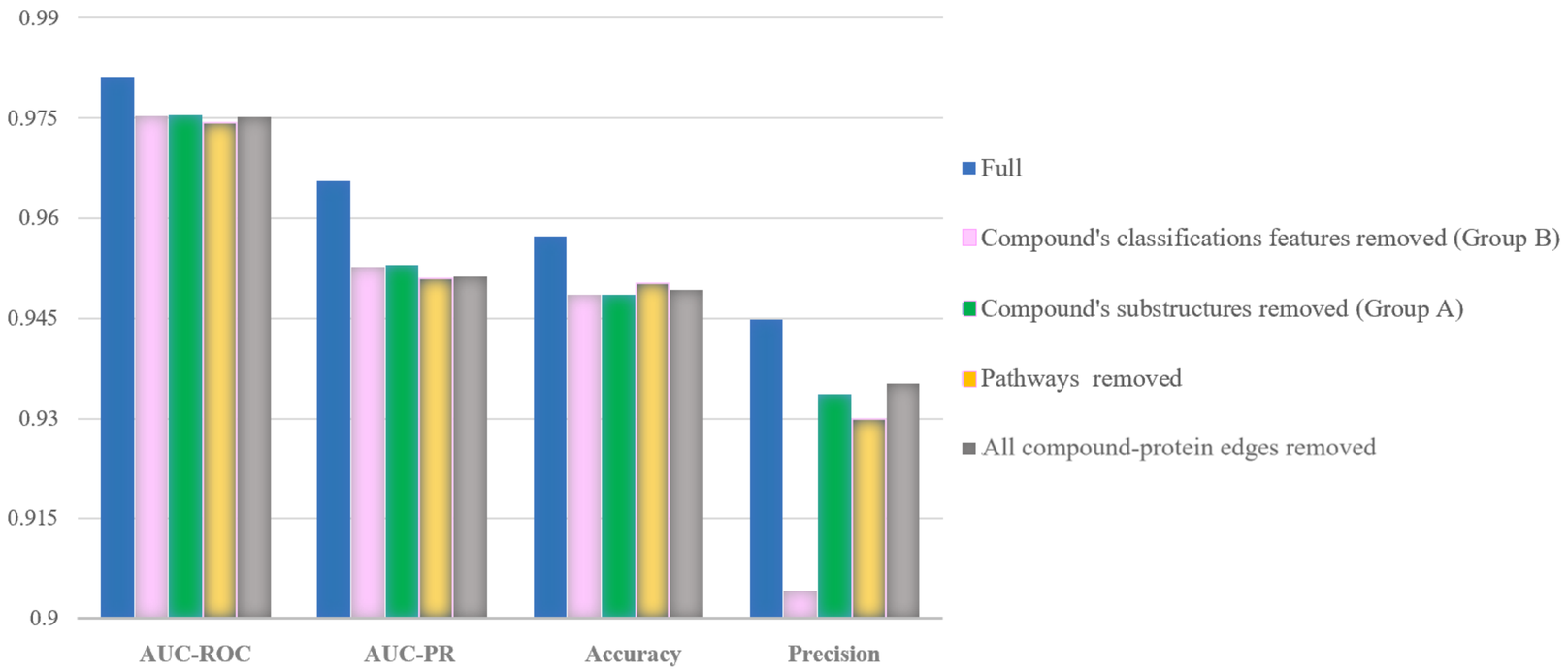


Figure (16): Impact of Compound Features on AHKG Performance

T-SNE Feature Representation

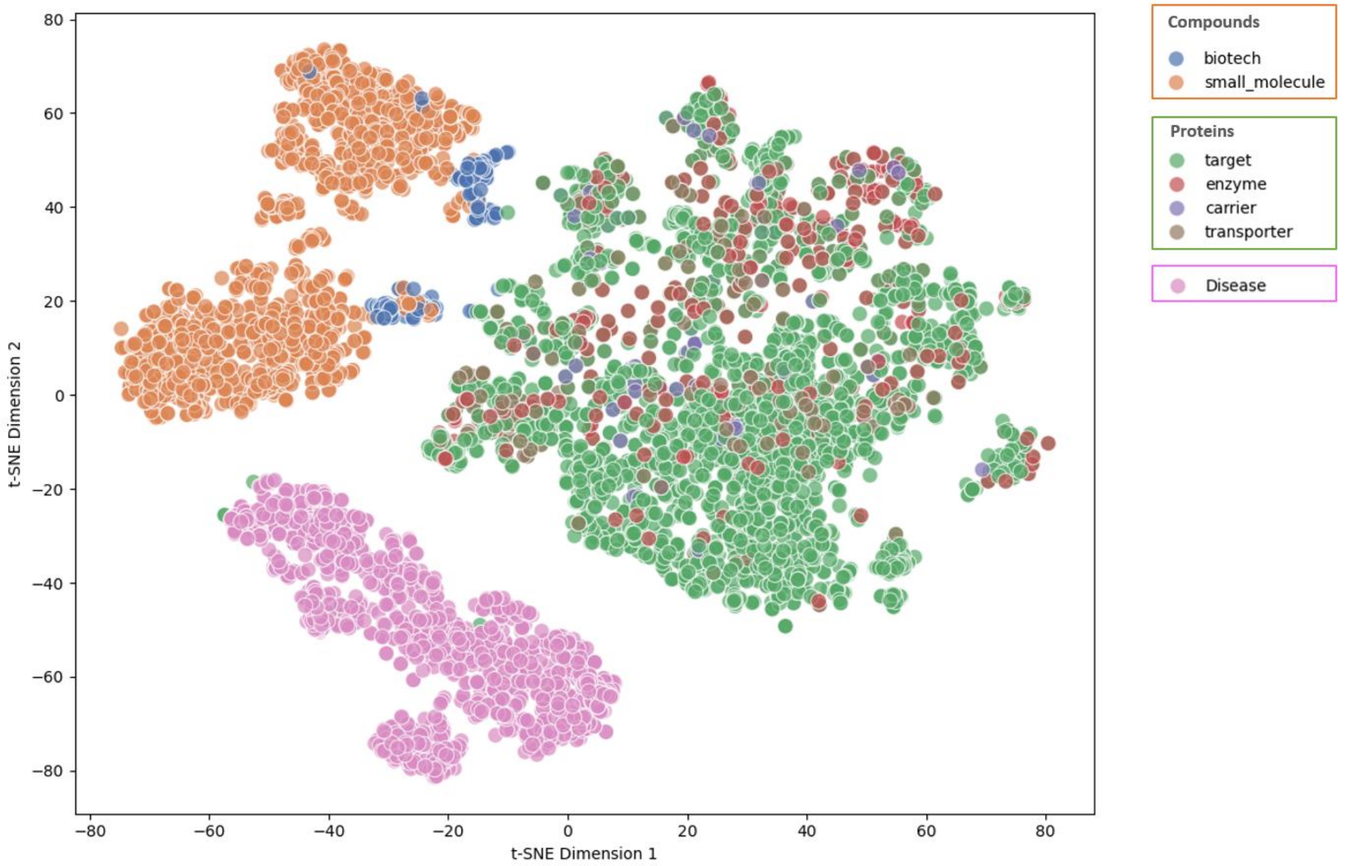
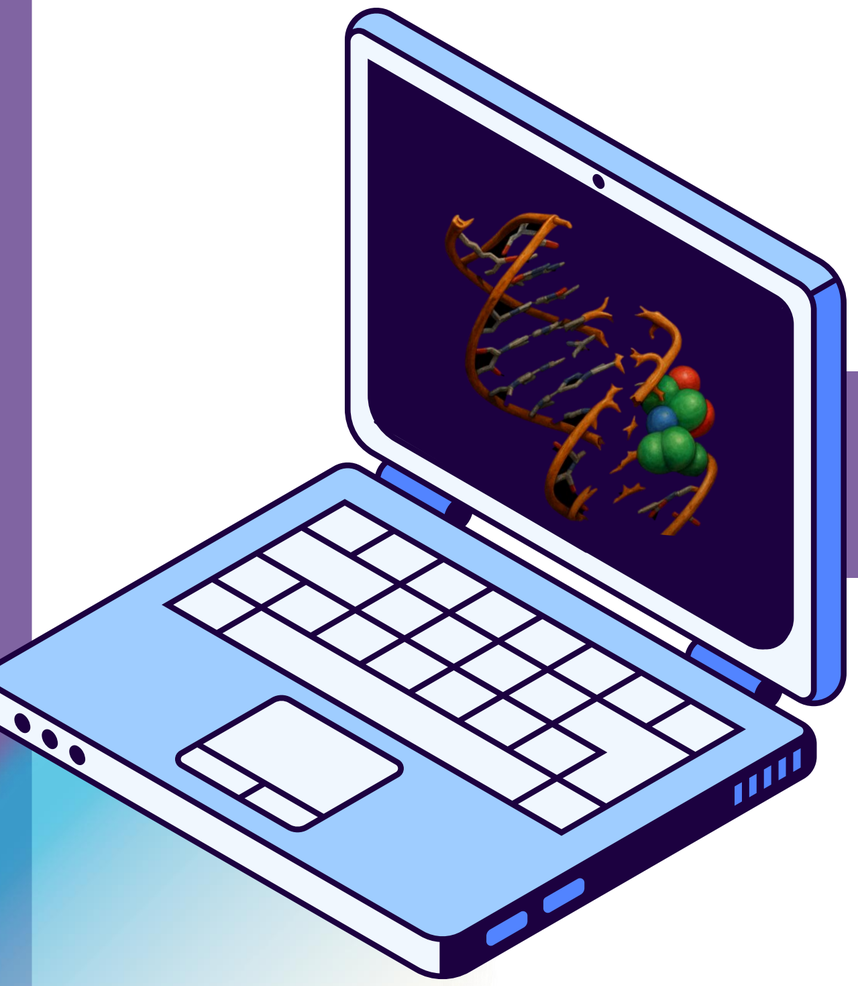


Figure (17): t-SNE Visualization of Diseases, Compounds, and Proteins in AHKG

Results



Case Study

Case Study

Rank	Compound	Reference
1	Trastuzumab	✓
2	Obiltoxaximab	-
3	Raxibacumab	-
4	Sebelipase alfa	✓
5	Sinecatechins	-
6	Aflibercept	✓
7	Protamine sulfate	✓
8	Benzylpenicillin	-
9	Denileukin diftitox	✓
10	Dihydrotachysterol	✓

Table (4): Summary of 10 Top-Ranked Compounds for NSCLC

Types of Lung Cancer:

- Non-Small Cell Lung Cancer (NSCLC)
- Small Cell Lung Cancer (SCLC)

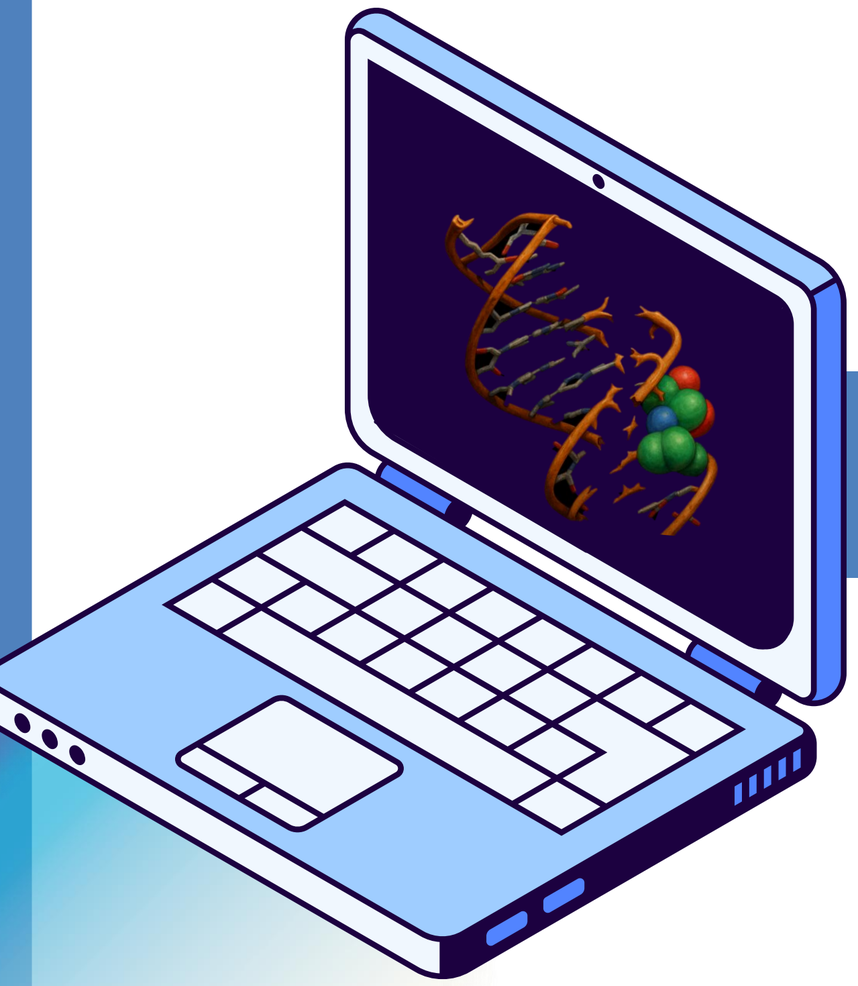
Lung Cancer Prevalence:

- Accounts for 13% of all cancer cases globally
- 2.1 million new diagnoses and 1.8 million deaths annually

Leading Cause of Cancer Deaths:

- #1 cause of cancer deaths in men
- #2 cause of cancer deaths in women
- 85-90% of lung cancer cases are attributed to tobacco smoking
- Rising tobacco consumption, especially in Asia, is a key factor



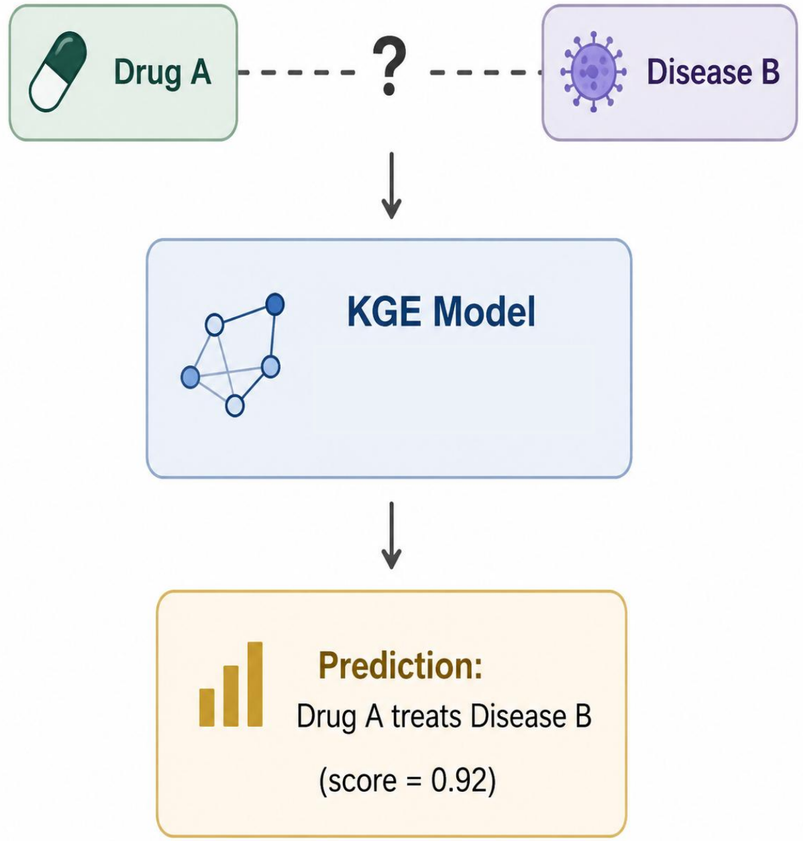


Conclusion

Future

Why Explainability Matters in KG-Based Drug Repurposing?

Knowledge graphs (KGs) enable AI models to discover hidden drug–disease relationships by integrating heterogeneous biomedical knowledge. However, many high-performing KG embedding models operate as **black boxes**, providing predictions without revealing the biological mechanisms behind them.





Future

Path-Based Explanation

Identifies meaningful biological paths connecting drugs and diseases.

Methods:

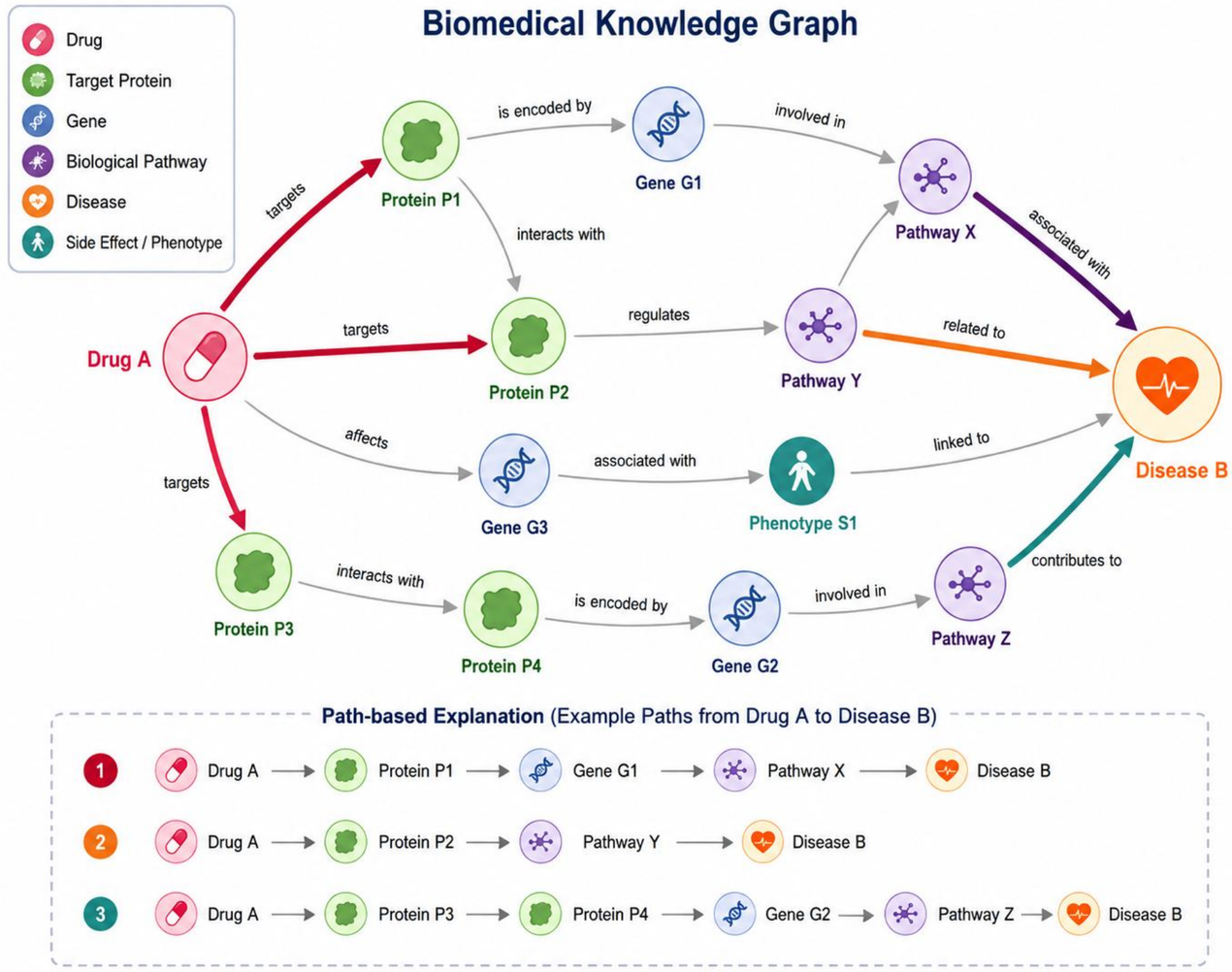
- Predefined biological paths (drug, protein, pathway,disease)
- Path mining

Strength:

✓ Human-interpretable biological reasoning

Limitation:

- Depends on KG completeness
- Multiple possible paths require ranking



Future

Sub Graph-Based Explanation

Identifies the most important local subgraph around a drug–disease prediction, highlighting key entities and relationships that drive the model's decision.

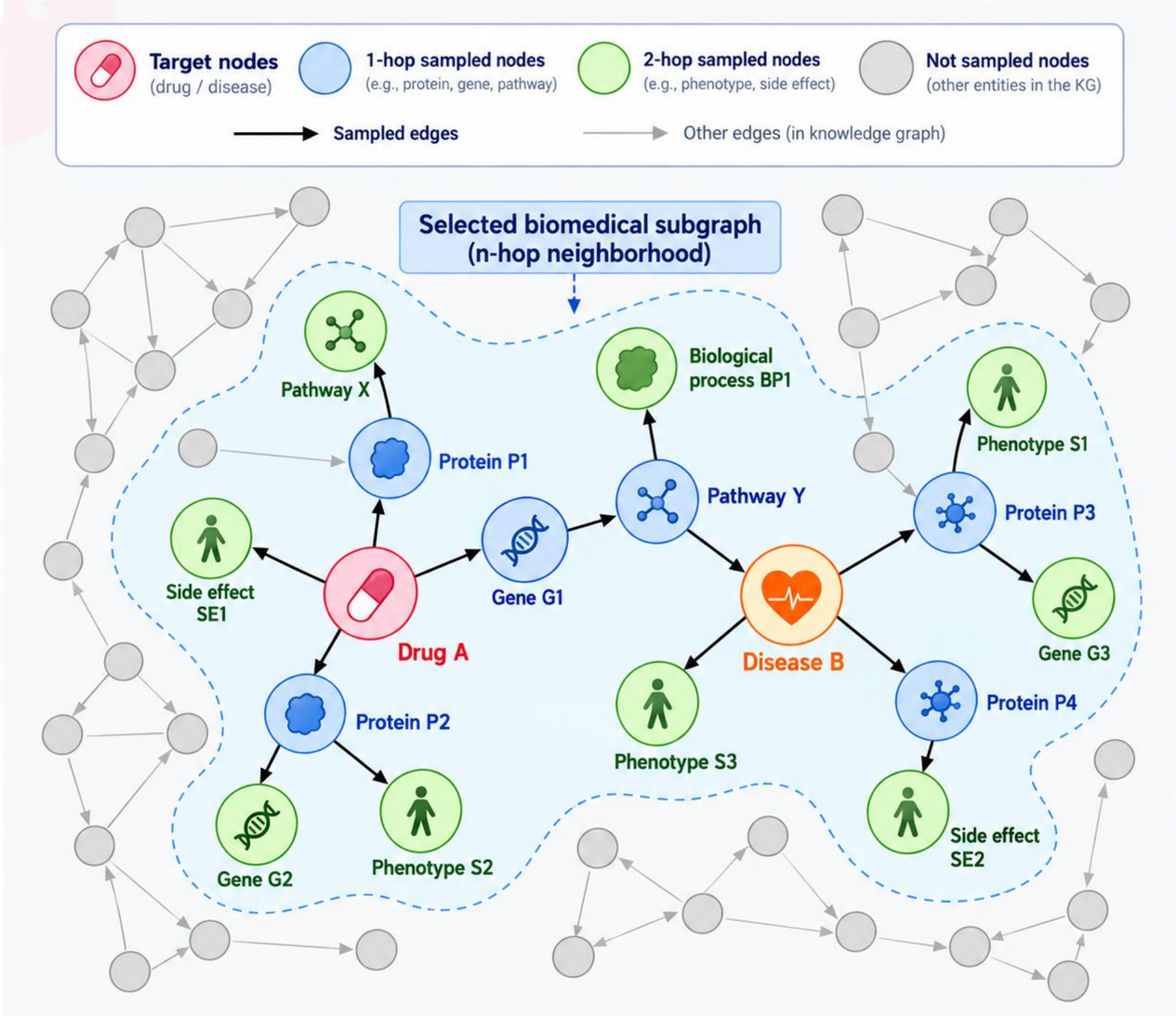
Methods:

- Node and edge importance scoring
- Graph masking / perturbation

Strength:

✓ Captures local structural evidence and improves interpretability **Limitation:**

- Important subgraphs may still be large or noisy.
- Results can depend on the explainer design.



Thank You!

Email : Marzieh.Khodadadi.a@gmail.com

Khodadadi AghGhaleh, M., Abedian, R., Zarghami, R., Fotuhi Siahpirani, A., & Gharaghani, S. (2026). ConvAHKG:Action-based hybrid knowledge graph with adual-channel convolutional approach fordrugrepurposing. *Scientific Reports*. https://doi.org/10.1038/s41598-026-38656-8