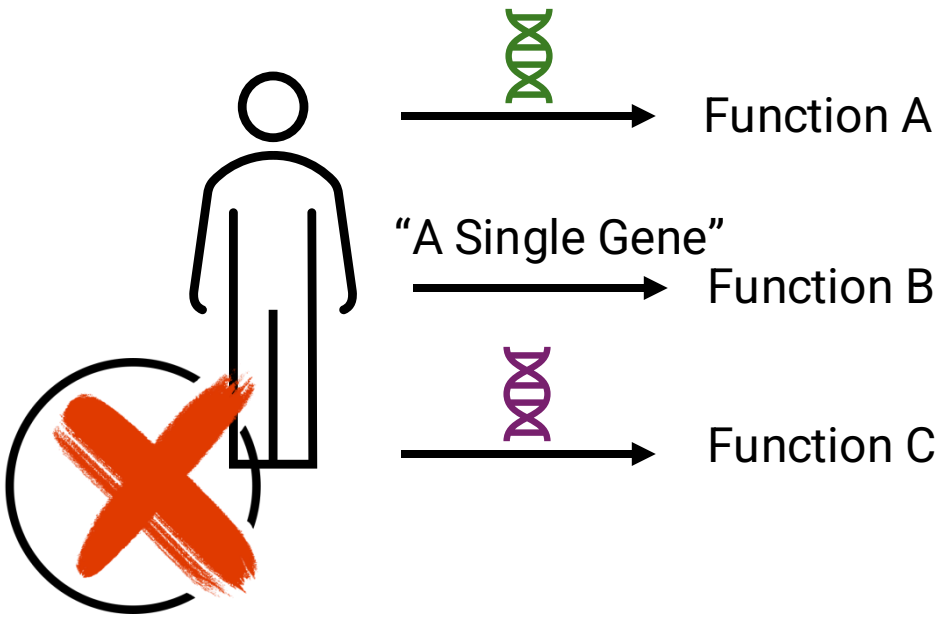
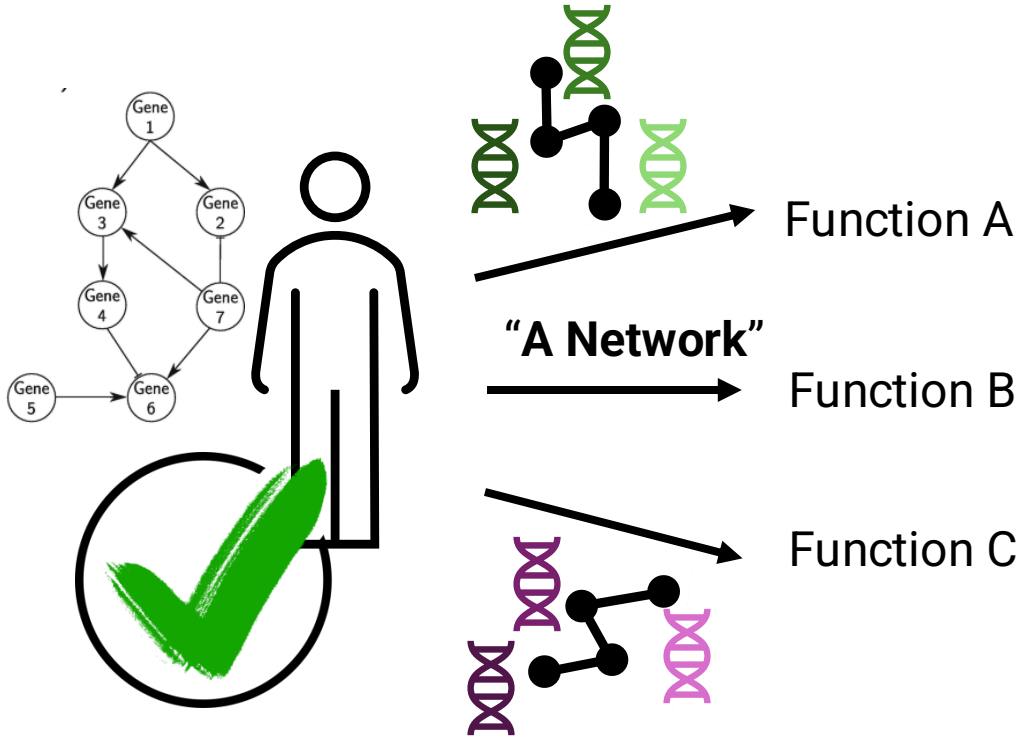
GeSubNet: Gene Interaction Inference for Disease Subtype Network Generation

Ziwei Yang, Zheng Chen*, Xin Liu*, Rikuto Kotoge, Peng Chen, Yasuko Matsubara, Yasushi Sakurai, Jimeng Sun

Background: Gene, Gene Network, and Biological Function

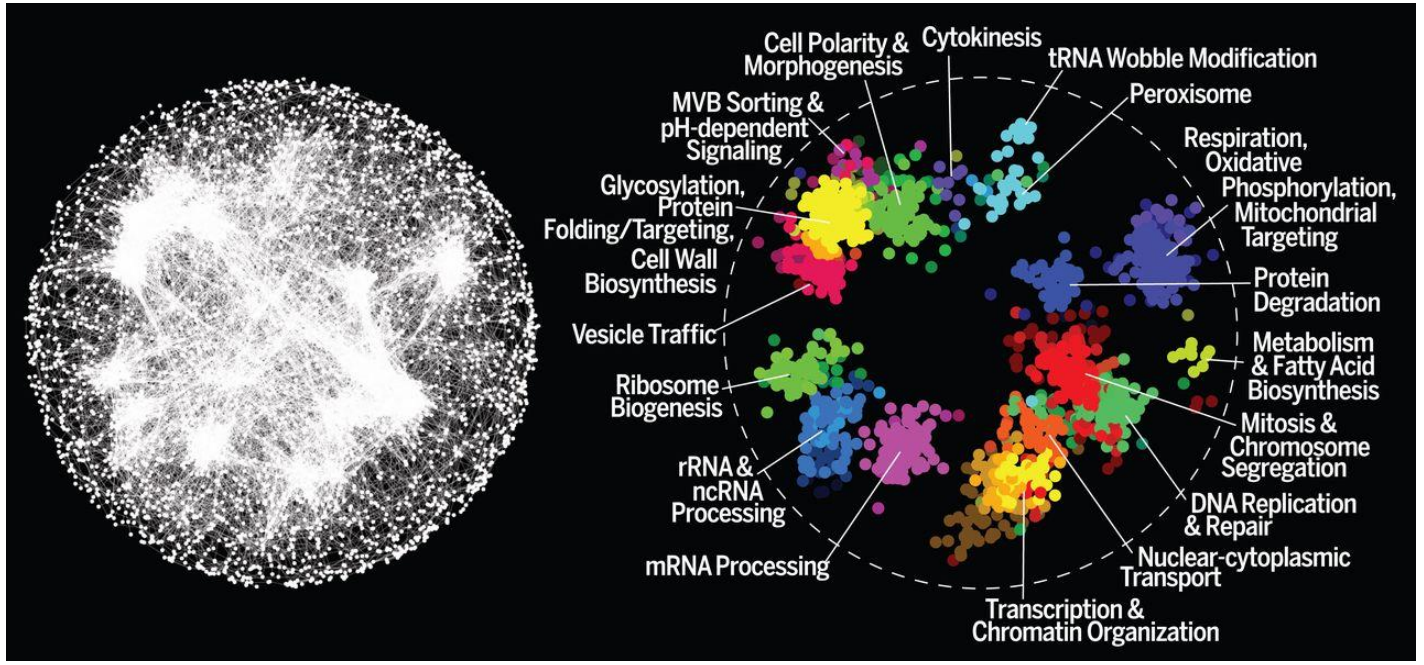
- The biological function of a living system is a result of many molecules interacting, it cannot be attributed to just a single gene.



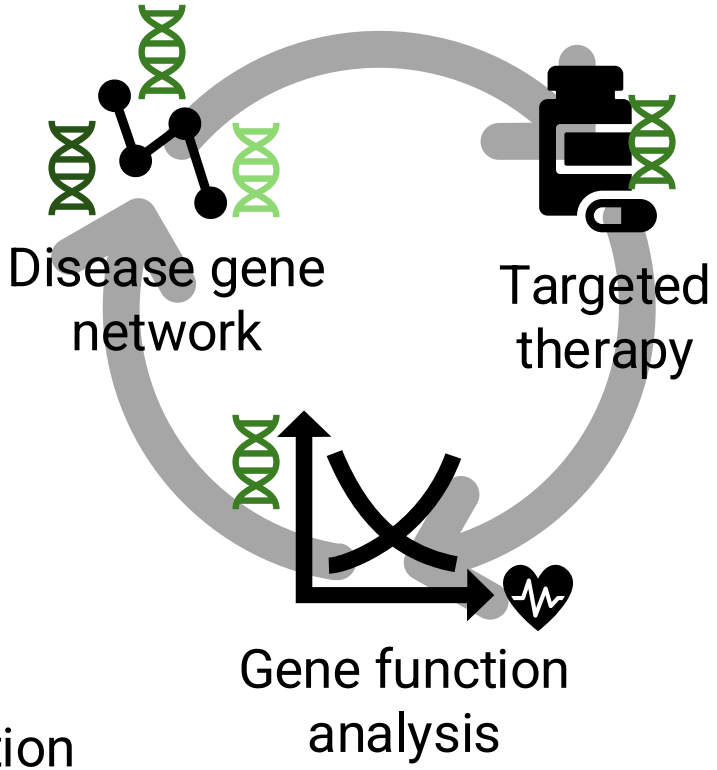


Background: Gene, Gene Network, and Biological Function

- The biological function of a living system is a result of many molecules interacting, it cannot be attributed to just a single gene.

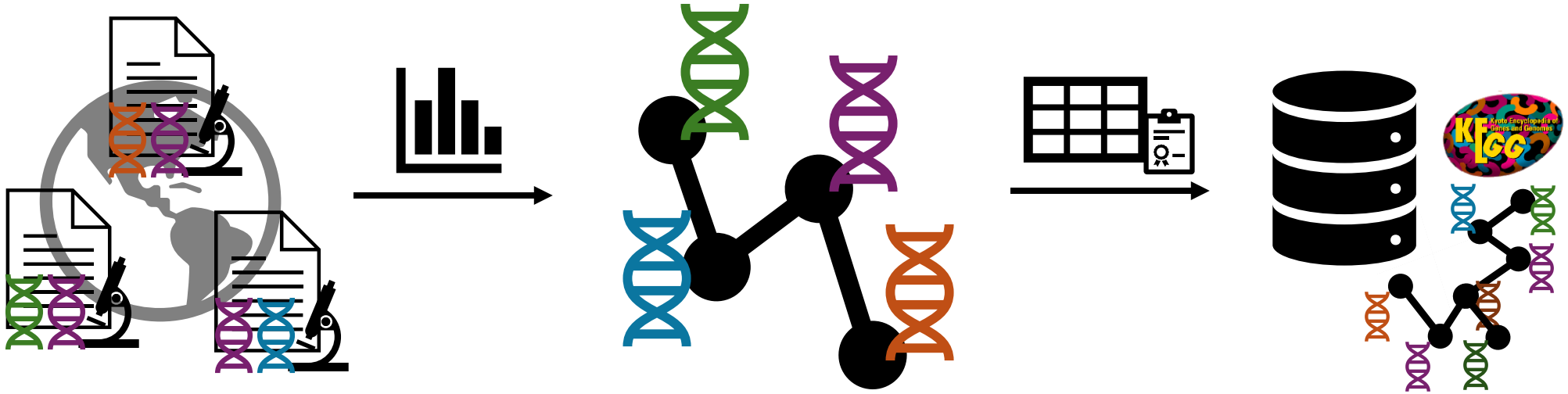


A gene interaction network maps a wiring diagram of cellular function



Background: Bio-knowledge Databases for Gene Network

- Biological knowledge bases integrate evidences on gene interactions, forming gene networks relate to many biological functions.



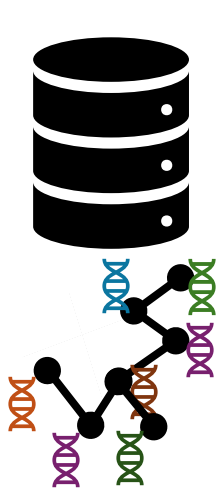
- Experimentally determined
- Protein homology
- Textmining
- …

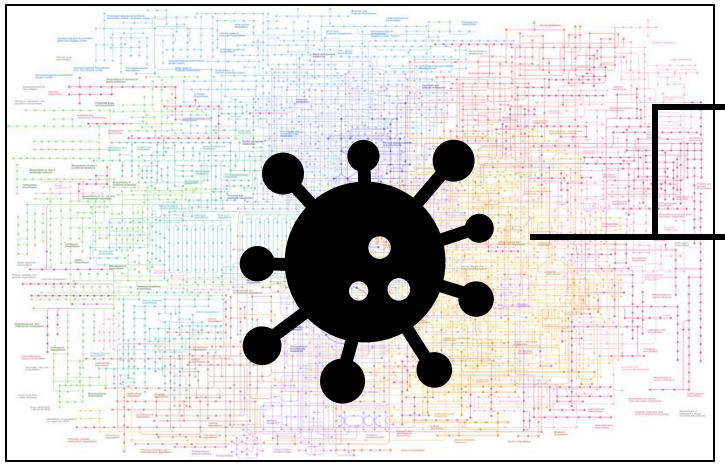
- Automated annotation
- Expert review
- …

A workflow for forming bio-knowledge databases for gene network

Problem: Gene Network Variations in Specific Disease Conditions

Available in current knowledge databases:
A fixed, general network

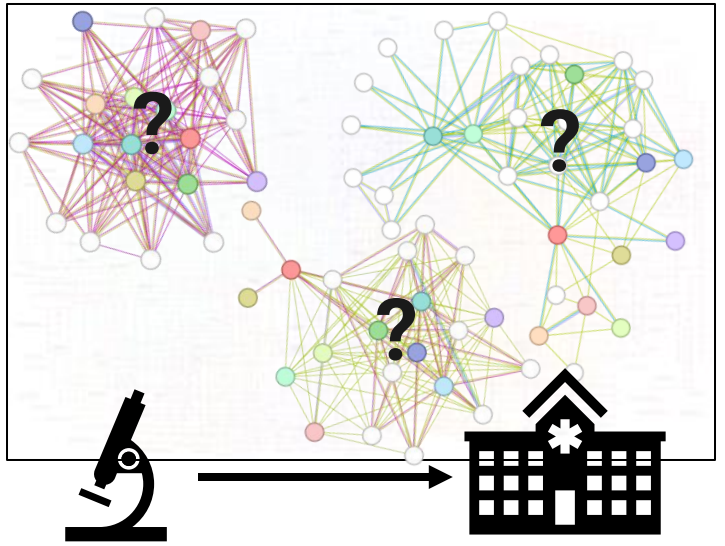




A disease type
(e.g., breast cancer)

Multiple disease subtypes
(e.g., BRCA-A,BRCA-B,…)

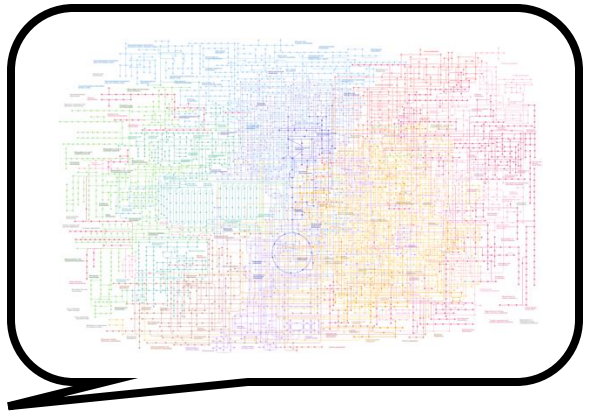
Necessary in practice but not available:
Sub-networks with variation



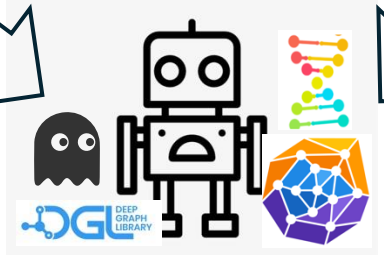
Problem: Gene Network Variations in Specific Disease Conditions

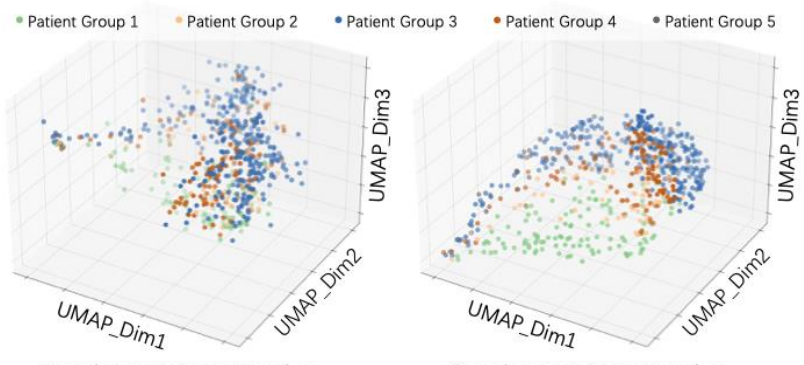
Patient sample groups

Experimental genetic data (e.g., high or low values of gene expression)



Fixed Gene network in databases

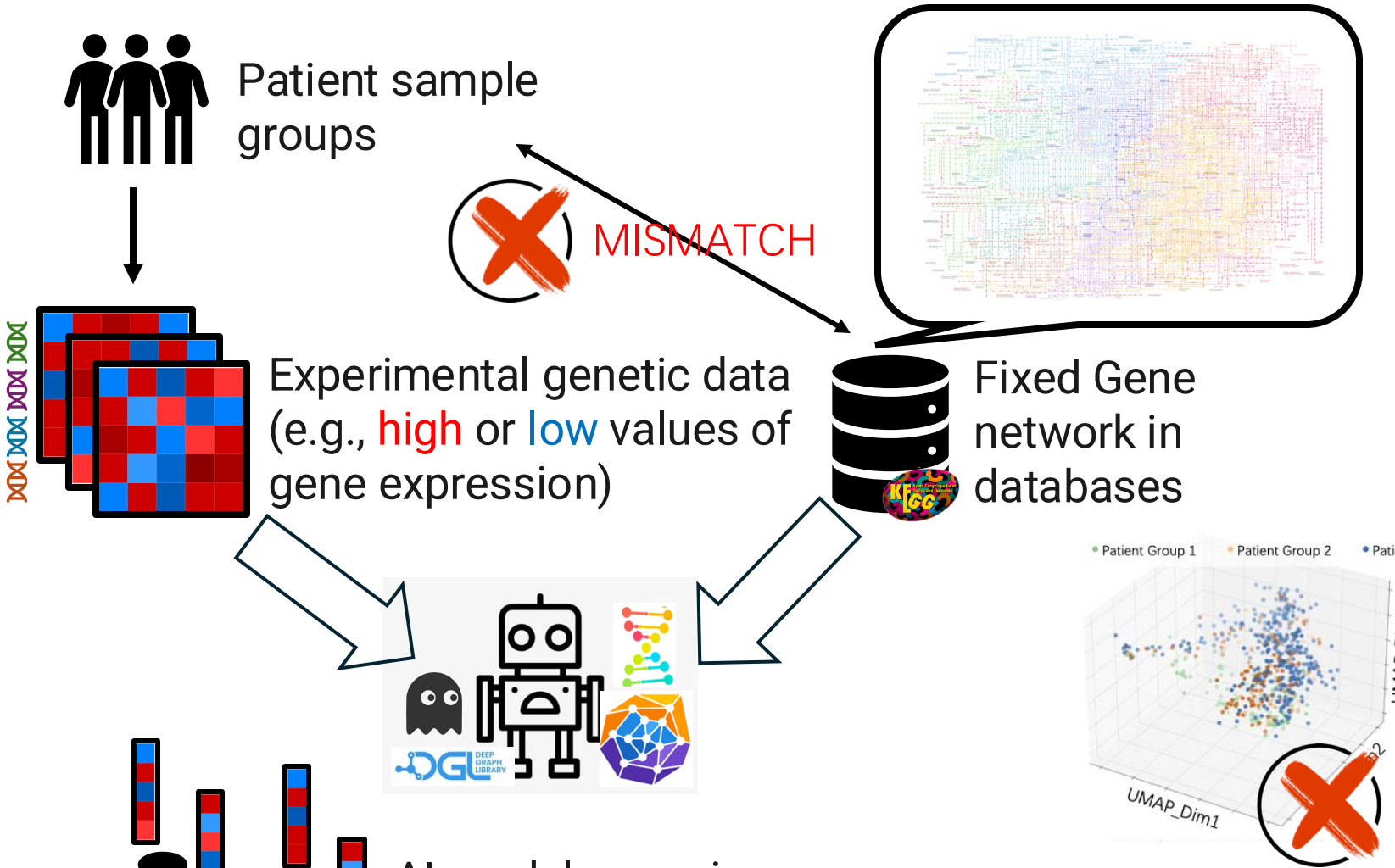




AI model: mapping genetic data to network

Downstream tasks (e.g., patient clustering/ classification)

Problem: Gene Network Variations in Specific Disease Conditions



Problem: Gene Network Variations in Specific Disease Conditions

Patient sample groups

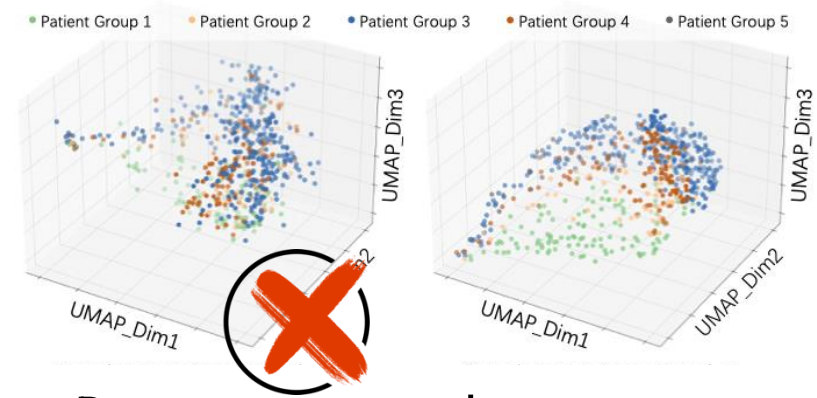
MISMATCH

Experimental genetic data (e.g., high or low values of gene expression)

Fixed Gene network in databases

Match, not available

Sub-networks of different patient groups

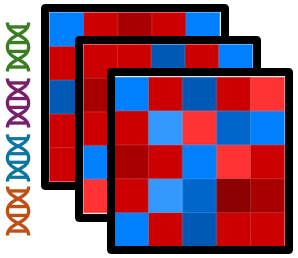


AI model: mapping genetic data to network

Downstream tasks (e.g., patient clustering/ classification)

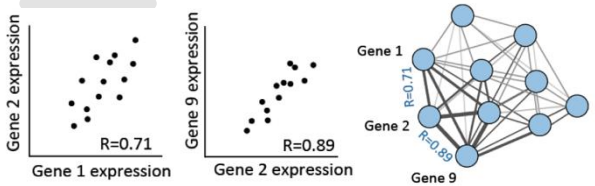
Motivation & Idea

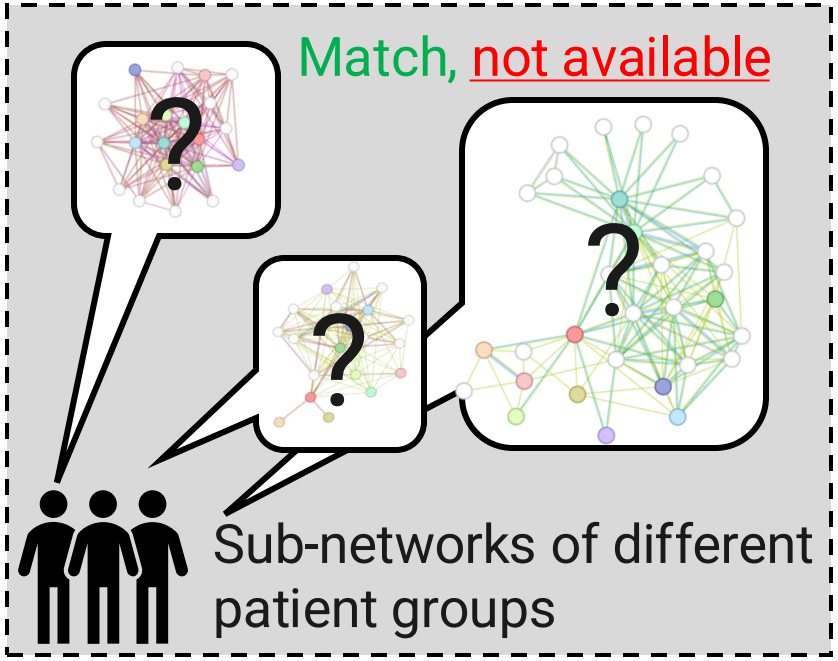
Patient sample groups



Experimental genetic data (e.g., high or low values of gene expression)

Current Solution A (Classical bioinformatics)

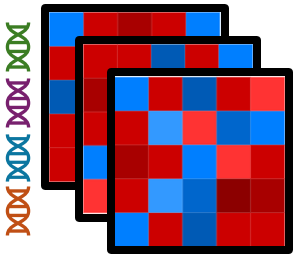




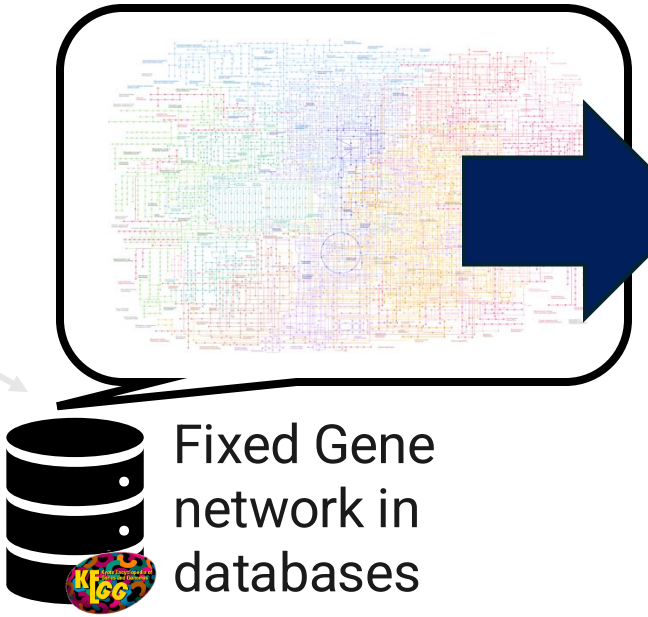
- Challenge in retrieving gene networks: General networks **mismatch** with subtype-specific variations.
- Experimental **gene expression data** are key to understanding **real gene links**.
- Current methods: find **correlations between genes** and build sub-networks (e.g., co-expression, Pearson correlation).

Motivation & Idea

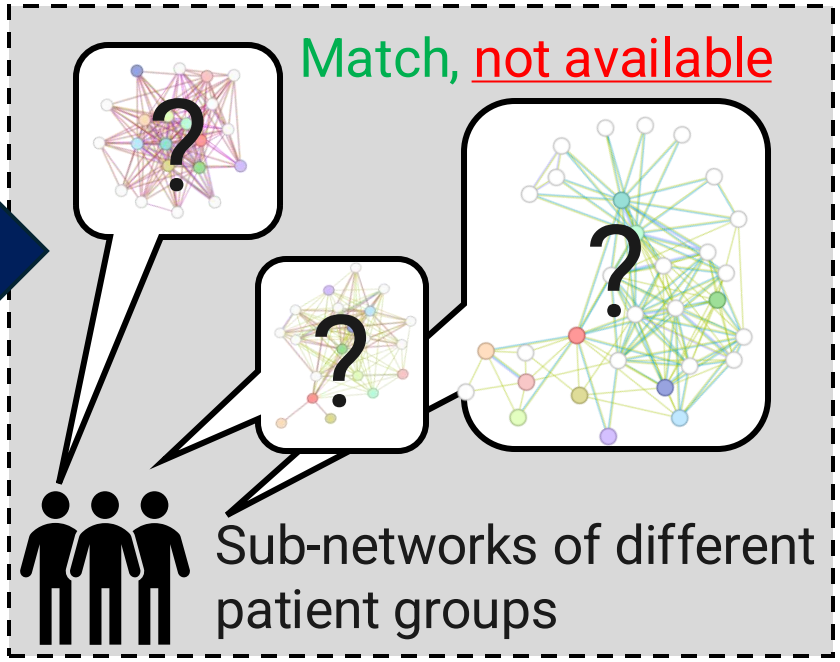
Current Solution B
(Recent popular GNN models)



Experimental genetic data (e.g., high or low values of gene expression)



Fixed Gene network in databases

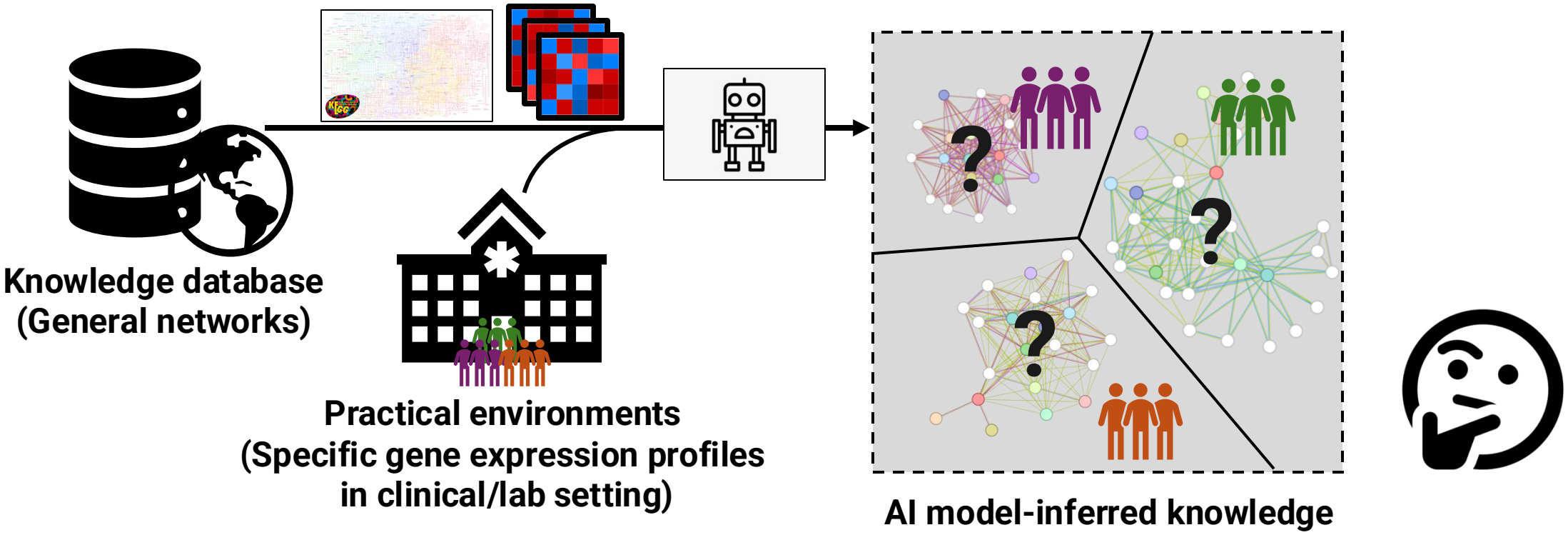


Sub-networks of different patient groups

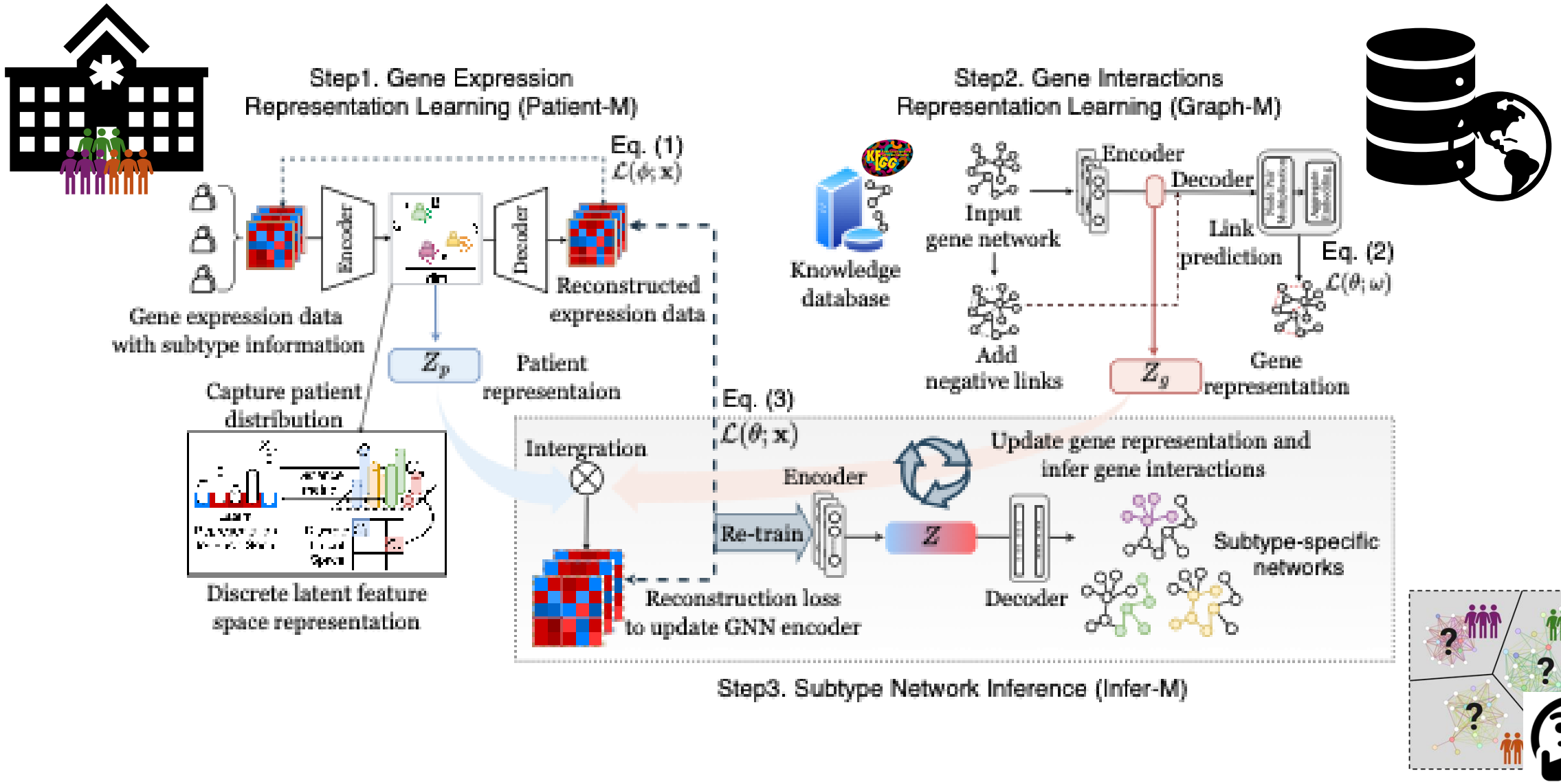
- Current graph learning methods: use both knowledge databases and experimental data, but only predict links on **general networks**.
- The patient **group-specific gene expression patterns** are not properly modeled.

Motivation & Idea

How can we integrate learning from general knowledge databases with disease subtype-specific experimental data to create targeted knowledge graphs?



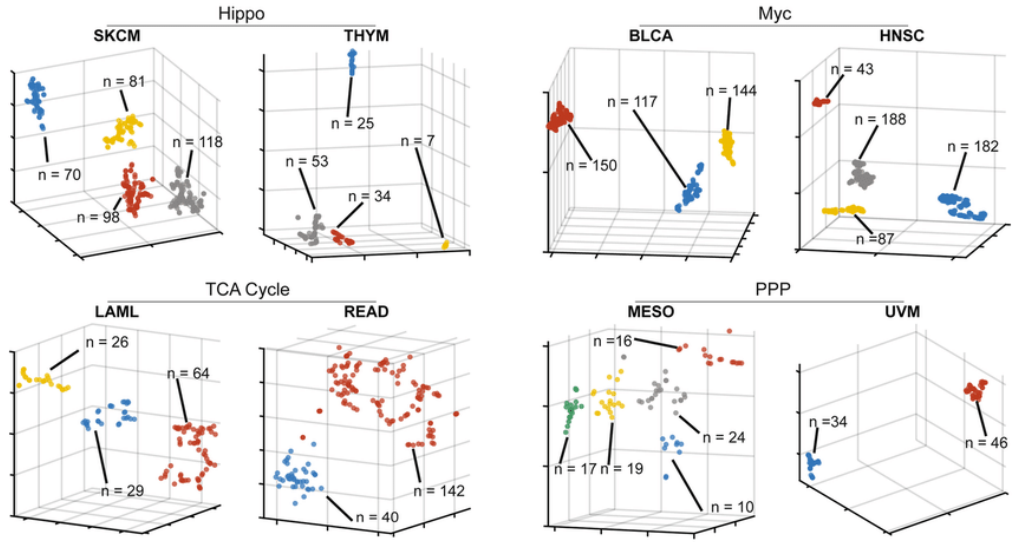
Our Contribution: Gene Interaction Inference for Disease Subtype Network Generation



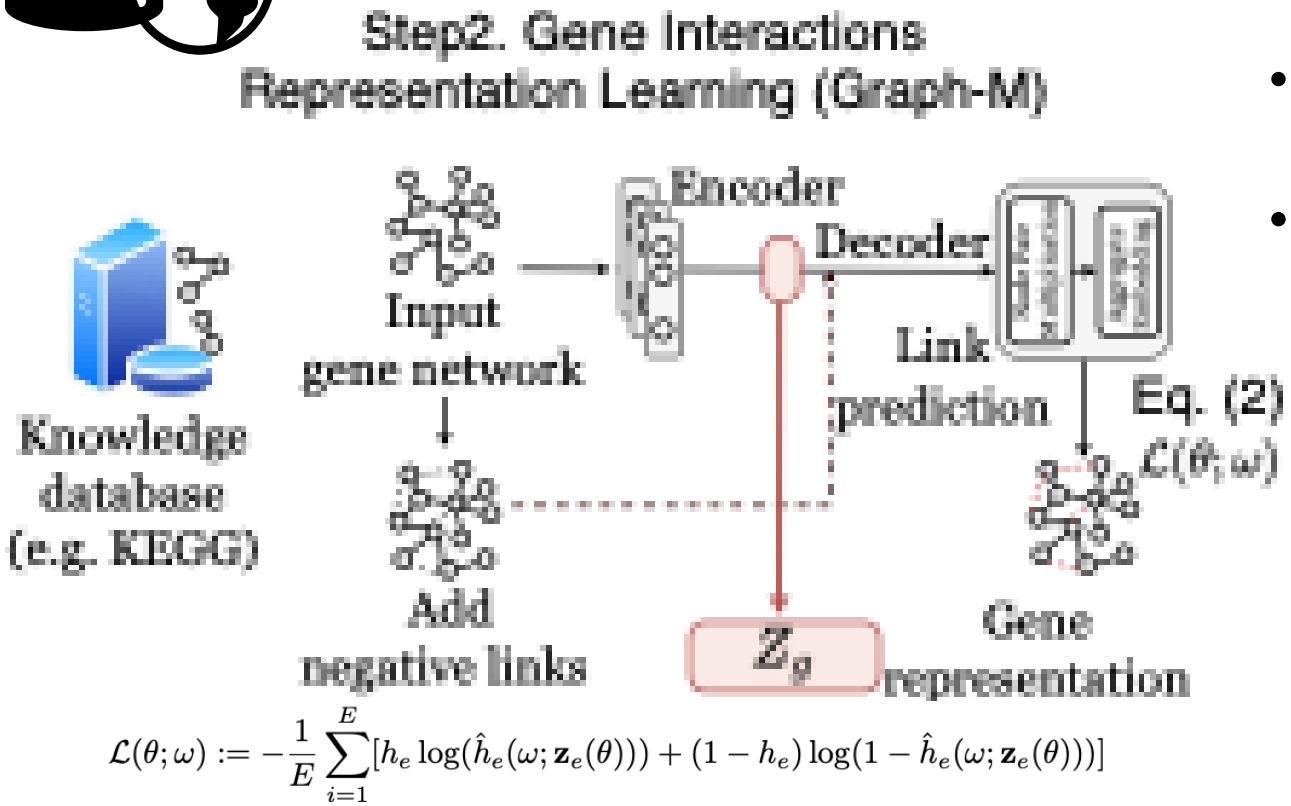
Our Contribution: Gene Interaction Inference for Disease Subtype Network Generation

Patient-M:

- This module sets up a patient sample subtyping task.
- Projects patient gene expression profiles into a latent representation Z_p.
- Z_p can **distinguish patients of different subtypes** (learning from their gene expression patterns).



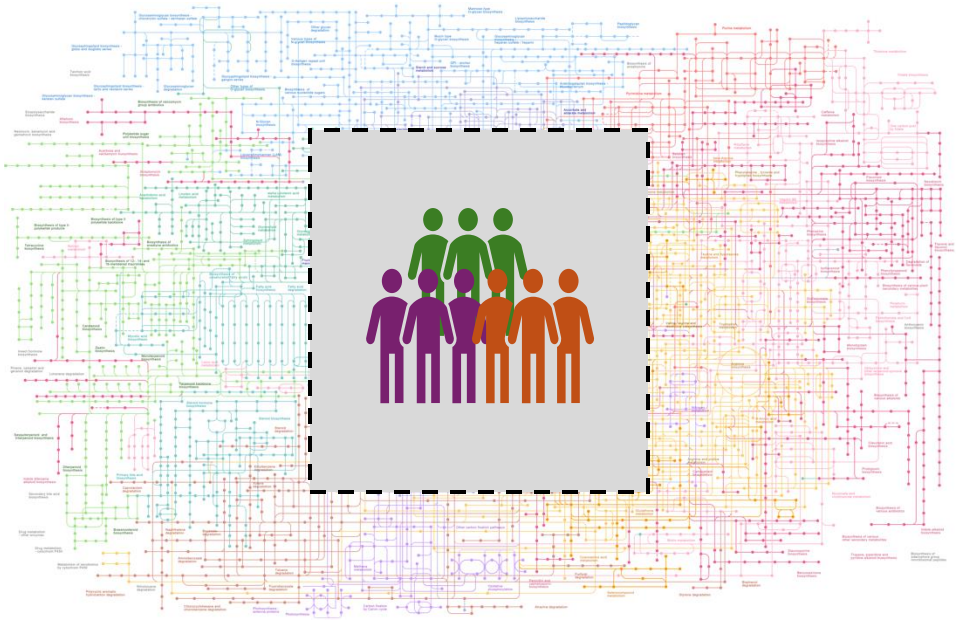
Our Contribution: Gene Interaction Inference for Disease Subtype Network Generation



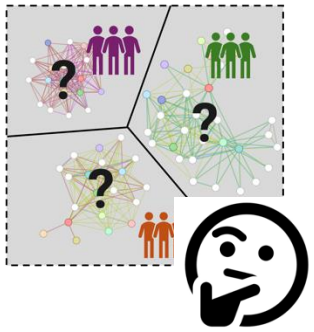
$$\mathcal{L}(\theta;\omega) := -\frac{1}{E}\sum_{i=1}^{E}[h_e \log(\hat{h}_e(\omega;\mathbf{z}_e(\theta))) + (1-h_e)\log(1-\hat{h}_e(\omega;\mathbf{z}_e(\theta)))]$$

Graph-M:

- This module sets up a link prediction task on the general gene network.
- Gene expression data as GNN node embedding to learn Z_g.
- Z_g contains **prior structure information** on the general network from database.

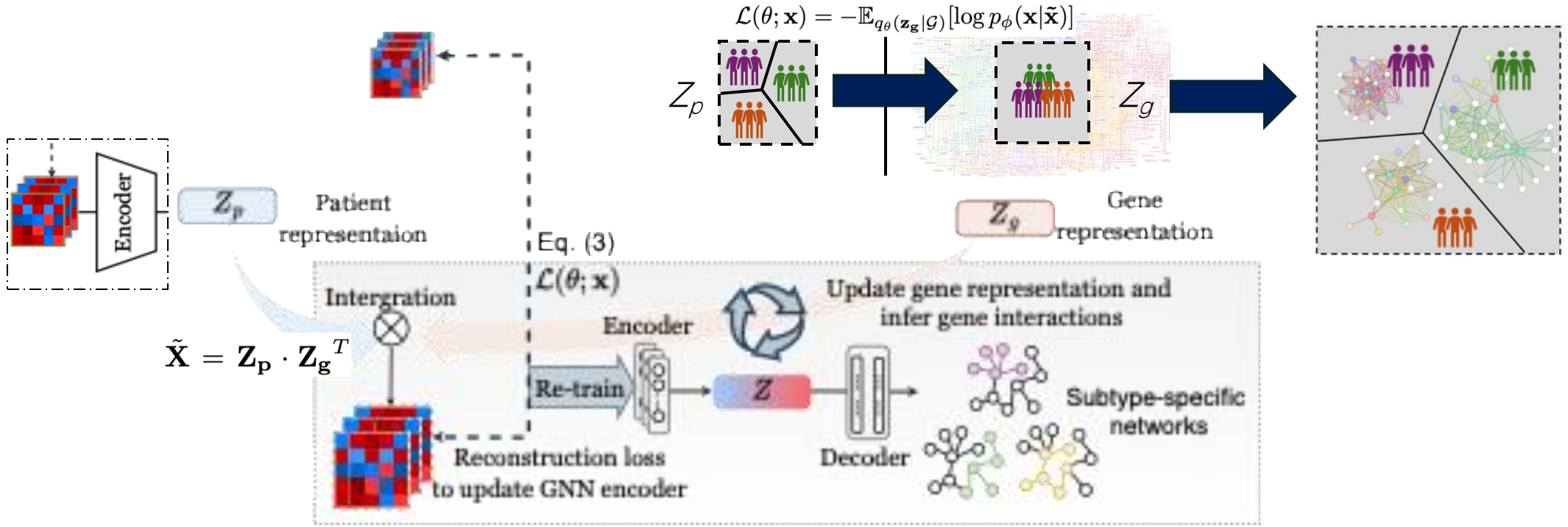


Our Contribution: Gene Interaction Inference for Disease Subtype Network Generation



Infer-M:

- This module integrates information from both Patient-M and Graph-M to **optimize the prior cancer network** and generate subtype-specific networks.
- The objective function uses Graph-M's **graph generation** capabilities, **conditioned on the patient separation** loss in Patient-M.



Step3. Subtype Network Inference (Infer-M)

Experiments & Results: Disease Subtype Network Generation

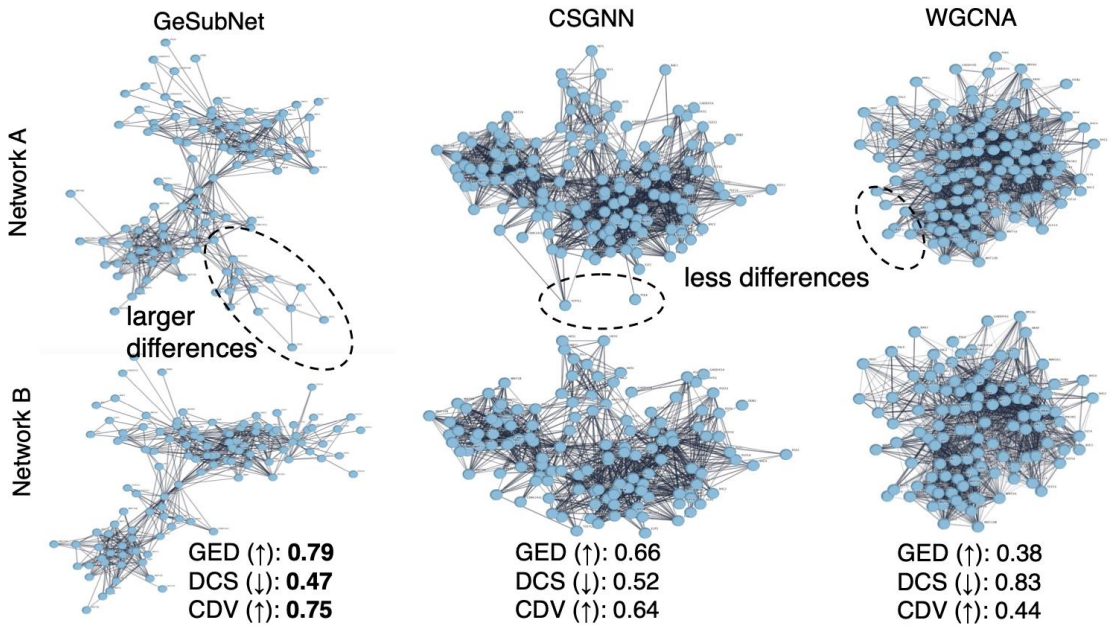
Graph Data for Cancer-Type Network

Cancer	Data Source				#Node	#Edge (Type I)	#Edge (Type II)
	KE	ST	Int	Mona			
BRCA	✓	✓	✓	✓	146	289	579
GBM	✓	✓		✓	102	75	128
LGG	✓	✓	✓		103	206	139
OV	✓	✓			109	46	95

Gene Expression Data of Cancer Subtype

Cancer	Raw Transcriptomics			Gene Expression Matrix	
	#Gene	#Patient	#Group	Sample size	Feature size
BRCA	19537	638	5	{320, 124, 119, 54, 21}	11327
GBM	17455	416	5	{125, 111, 80, 68, 32}	11273
LGG	16245	451	3	{213, 151, 87}	11124
OV	17226	291	4	{81, 76, 68, 66}	11324

Cancer Sub-network Generation

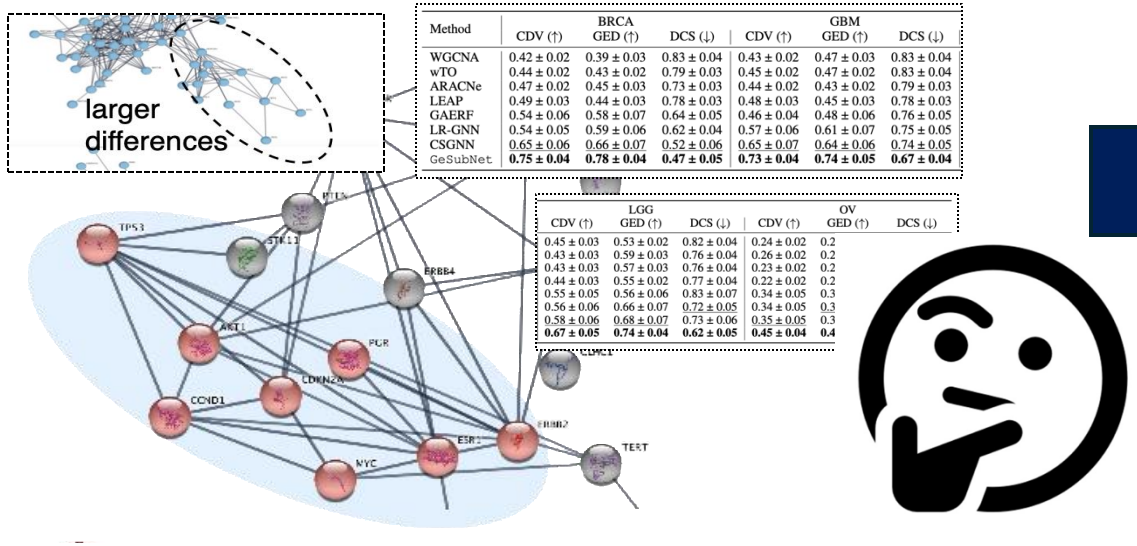


Generated Sub-network Attribute Ccomparison

Method	BRCA			GBM			LGG			OV		
	CDV (↑)	GED (↑)	DCS (↓)	CDV (↑)	GED (↑)	DCS (↓)	CDV (↑)	GED (↑)	DCS (↓)	CDV (↑)	GED (↑)	DCS (↓)
WGCNA	0.42 ± 0.02	0.39 ± 0.03	0.83 ± 0.04	0.43 ± 0.02	0.47 ± 0.03	0.83 ± 0.04	0.45 ± 0.03	0.53 ± 0.02	0.82 ± 0.04	0.24 ± 0.02	0.25 ± 0.03	0.83 ± 0.04
wTO	0.44 ± 0.02	0.43 ± 0.02	0.79 ± 0.03	0.45 ± 0.02	0.47 ± 0.02	0.83 ± 0.04	0.43 ± 0.03	0.59 ± 0.03	0.76 ± 0.04	0.26 ± 0.02	0.25 ± 0.03	0.83 ± 0.04
ARACNe	0.47 ± 0.02	0.45 ± 0.03	0.73 ± 0.03	0.44 ± 0.02	0.43 ± 0.02	0.79 ± 0.03	0.43 ± 0.03	0.57 ± 0.03	0.76 ± 0.04	0.23 ± 0.02	0.25 ± 0.03	0.81 ± 0.03
LEAP	0.49 ± 0.03	0.44 ± 0.03	0.78 ± 0.03	0.48 ± 0.03	0.45 ± 0.03	0.78 ± 0.03	0.44 ± 0.03	0.55 ± 0.02	0.77 ± 0.04	0.22 ± 0.02	0.24 ± 0.03	0.84 ± 0.04
GAERF	0.54 ± 0.06	0.58 ± 0.07	0.64 ± 0.05	0.46 ± 0.04	0.48 ± 0.06	0.76 ± 0.05	0.55 ± 0.05	0.56 ± 0.06	0.83 ± 0.07	0.34 ± 0.05	0.36 ± 0.06	0.82 ± 0.06
LR-GNN	0.54 ± 0.05	0.59 ± 0.06	0.62 ± 0.04	0.57 ± 0.06	0.61 ± 0.07	0.75 ± 0.05	0.56 ± 0.06	0.66 ± 0.07	0.72 ± 0.05	0.34 ± 0.05	0.37 ± 0.06	0.82 ± 0.07
CSGNN	0.65 ± 0.06	0.66 ± 0.07	0.52 ± 0.06	0.65 ± 0.07	0.64 ± 0.06	0.74 ± 0.05	0.58 ± 0.06	0.68 ± 0.07	0.73 ± 0.06	0.35 ± 0.05	0.35 ± 0.06	0.80 ± 0.05
GeSubNet	**0.75 ± 0.04**	**0.78 ± 0.04**	**0.47 ± 0.05**	**0.73 ± 0.04**	**0.74 ± 0.05**	**0.67 ± 0.04**	**0.67 ± 0.05**	**0.74 ± 0.04**	**0.62 ± 0.05**	**0.45 ± 0.04**	**0.44 ± 0.04**	**0.75 ± 0.04**

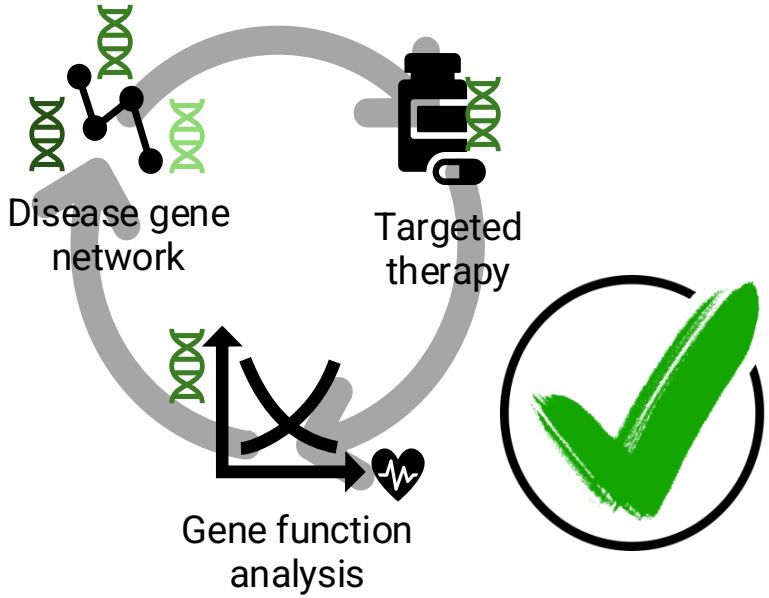
CDV: Coefficient of Degree Variation; GED: Graph Edit Distance; DCS: DeltCon Similarity

Experiments & Results: Biological Meaning Evaluation & Review



Method	BRCA CDV (↑)	BRCA GED (↑)	BRCA DCS (↓)	GBM CDV (↑)	GBM GED (↑)	GBM DCS (↓)
WGCNA	0.42 ± 0.02	0.39 ± 0.03	0.83 ± 0.04	0.43 ± 0.02	0.47 ± 0.03	0.83 ± 0.04
wTO	0.44 ± 0.02	0.43 ± 0.02	0.79 ± 0.03	0.45 ± 0.02	0.47 ± 0.02	0.83 ± 0.04
ARACNe	0.47 ± 0.02	0.45 ± 0.03	0.73 ± 0.03	0.44 ± 0.02	0.43 ± 0.02	0.79 ± 0.03
LEAP	0.49 ± 0.03	0.44 ± 0.03	0.78 ± 0.03	0.48 ± 0.03	0.45 ± 0.03	0.78 ± 0.03
GAERF	0.54 ± 0.06	0.58 ± 0.07	0.64 ± 0.05	0.46 ± 0.04	0.48 ± 0.06	0.76 ± 0.05
LR-GNN	0.54 ± 0.05	0.59 ± 0.06	0.62 ± 0.04	0.57 ± 0.06	0.61 ± 0.07	0.75 ± 0.05
CSGNN	0.65 ± 0.06	0.66 ± 0.07	0.52 ± 0.06	0.65 ± 0.07	0.64 ± 0.06	0.74 ± 0.05
GeSubNet	**0.75 ± 0.04**	**0.78 ± 0.04**	**0.47 ± 0.05**	**0.73 ± 0.04**	**0.74 ± 0.05**	**0.67 ± 0.04**

LGG CDV (↑)	LGG GED (↑)	LGG DCS (↓)	OV CDV (↑)	OV GED (↑)	OV DCS (↓)
0.45 ± 0.03	0.53 ± 0.02	0.82 ± 0.04	0.24 ± 0.02	0.2	
0.43 ± 0.03	0.59 ± 0.03	0.76 ± 0.04	0.26 ± 0.02	0.2	
0.43 ± 0.03	0.57 ± 0.03	0.76 ± 0.04	0.23 ± 0.02	0.2	
0.44 ± 0.03	0.55 ± 0.02	0.77 ± 0.04	0.22 ± 0.02	0.2	
0.55 ± 0.05	0.56 ± 0.06	0.83 ± 0.07	0.34 ± 0.05	0.3	
0.56 ± 0.06	0.66 ± 0.07	0.72 ± 0.05	0.34 ± 0.05	0.3	
0.58 ± 0.06	0.68 ± 0.07	0.73 ± 0.06	0.35 ± 0.05	0.3	
0.67 ± 0.05	**0.74 ± 0.04**	**0.62 ± 0.05**	**0.45 ± 0.04**	**0.4**	

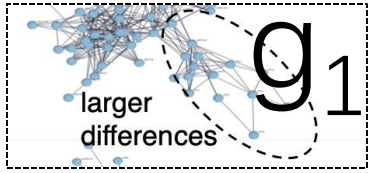


Sub-network structure differences (graph metrics)

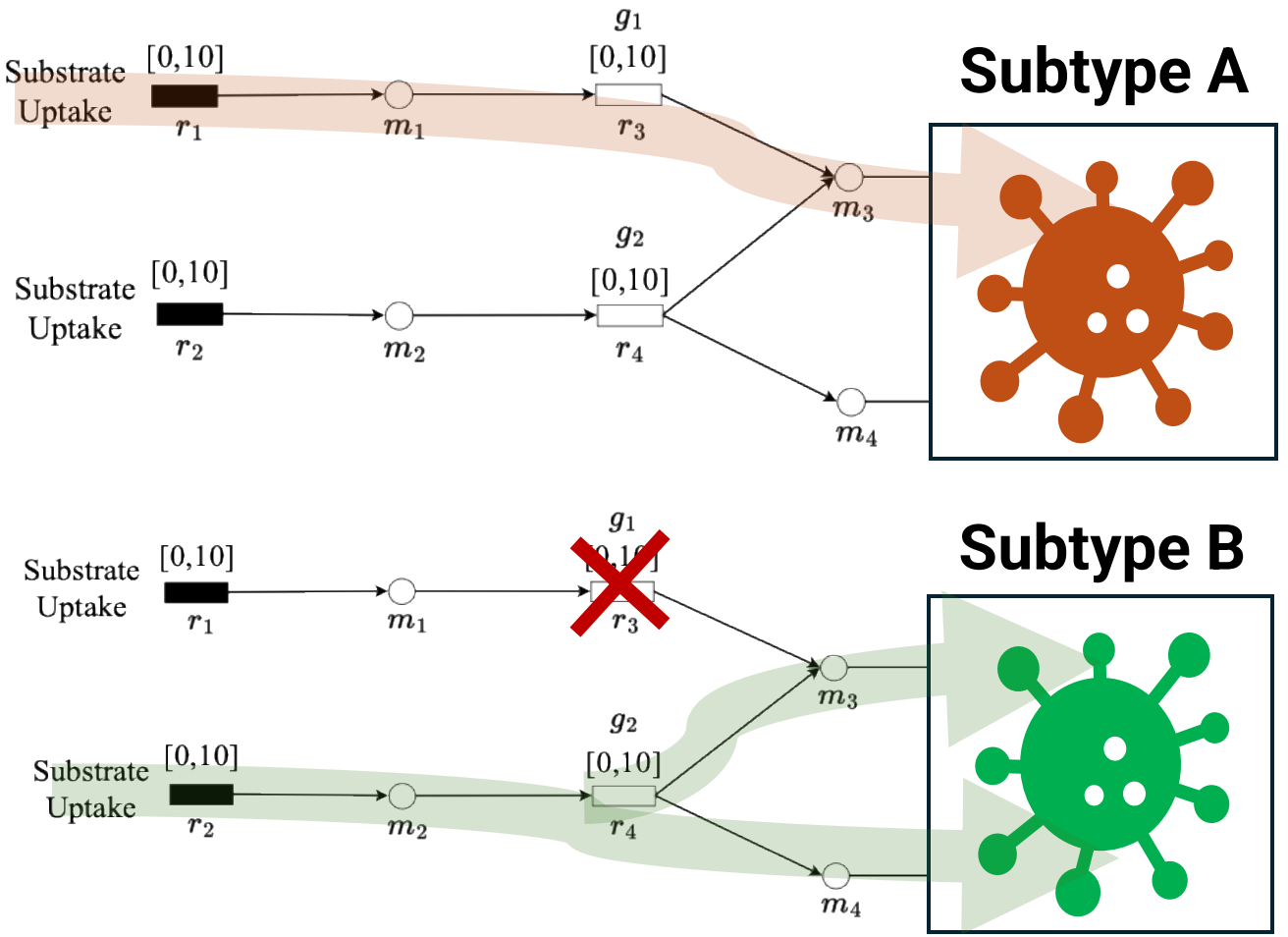
Biological significance behind the structure differences (biological metrics)

Simulated targeted knockout of variation gene nodes

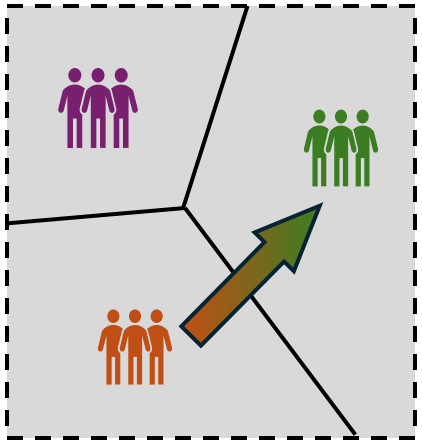
Original Subtype A Network



Subtype A Network with g_1 knockout



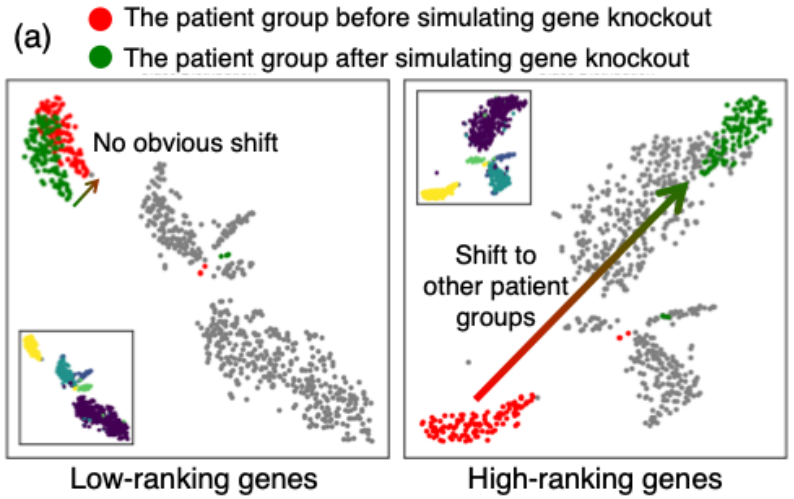
Network variations lead to **patient distribution shifts**?



Experiments & Results: Biological Meaning Evaluation & Review

Simulated Knockout of Variation Genes

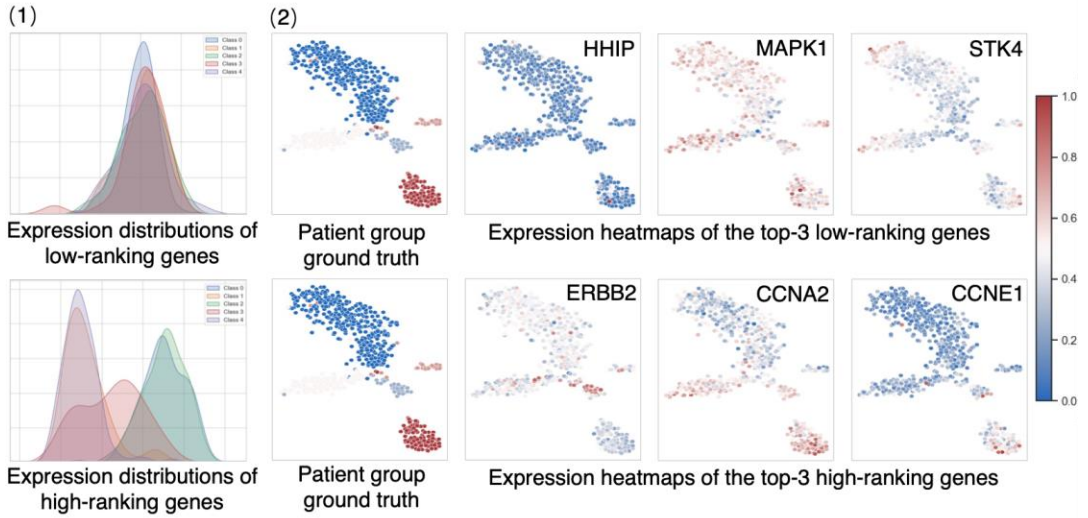
(a)



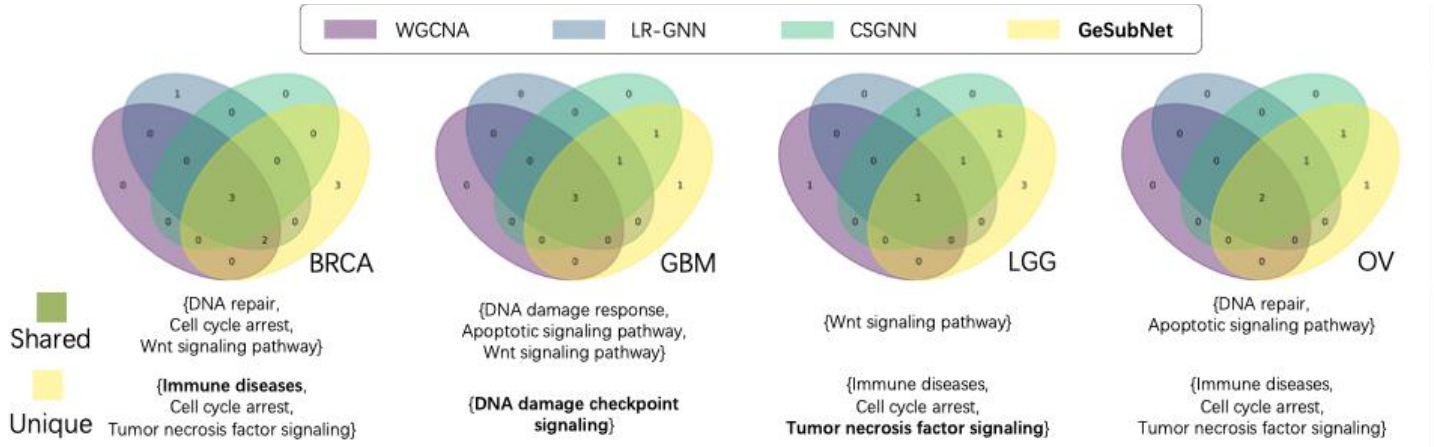
(b) Table: Shift rates (Δ_{SR}) on knocking out high- and low-ranking genes across different methods

Method	Metric: Δ_{SR}	
	High-ranking genes (↑)	Low-ranking genes (↓)
WGCNA	0.24	0.22
wTO	0.24	0.24
ARACNe	0.23	0.25
LEAP	0.28	0.24
GAERF	0.32	0.28
LR-GNN	0.23	0.23
CSGNN	0.34	0.29
GeSubNet	**0.83**	**0.12**

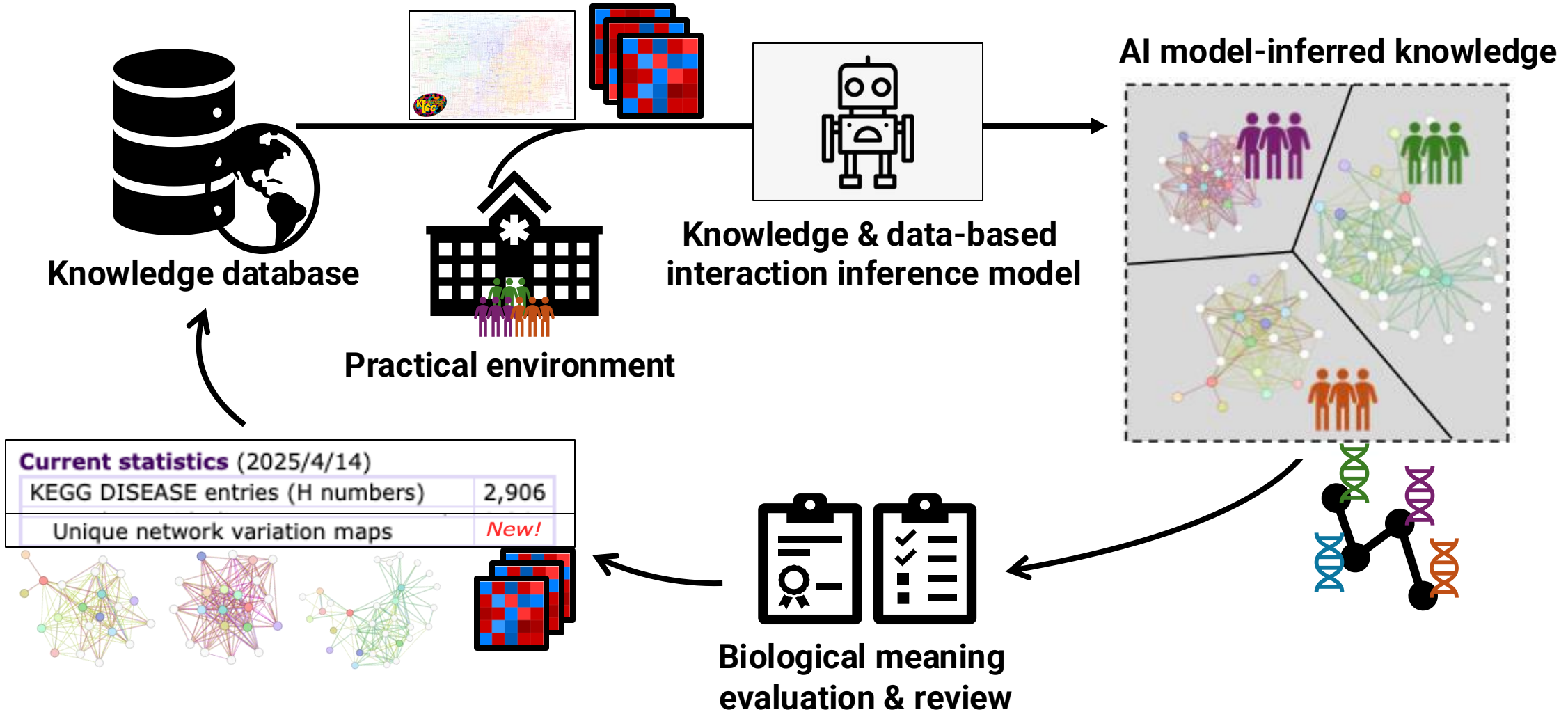
Network-variation Genes Expression Pattern



Biological function topics in different sub-networks



Summary: Gene Interaction Inference for Disease Subtype Network Generation

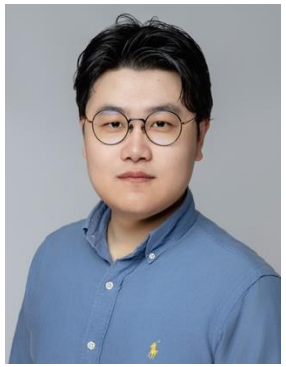


Thank you!

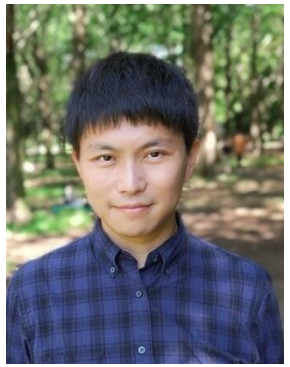
Resource

Paper

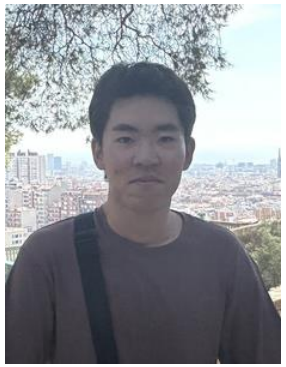
Thank you!



Zheng Chen



Xin Liu



Rikuto Kotoge

Peng Chen

Yasuko Matsubara

Yasushi Sakurai

Jimeng Sun