

BMI 702: Biomedical Artificial Intelligence

Foundations of Biomedical Informatics II, Spring 2024

Lecture 6: Foundations of geometric deep learning, graph representation learning, link prediction, node classification, graph clustering, graph classification, semi-supervised learning, label propagation, network medicine, disease modules and endotypes



HARVARD
MEDICAL SCHOOL

Marinka Zitnik
marinka@hms.harvard.edu

Responses to L5 Quick Check

Describe biomedical AI applications using the framework for fair AI

- Exam example of biomedical AI application is creation of a model that predicts whether patients will develop alzheimer's using patient EHR data.
- Data regulators would define the fairness criteria, which data should be used and what is the measure of similarity between users. For instance, how to determine the similarity between individuals within groups of black, white, asian etc. patients. Likely the health system or biomedical institution would ask as the regulator.
- Producers would take advice from data regulators and build a dataset that adhere to their requirements, defining the input and producing fair representation outputs. Likely to be research lab or clinic that has access to data from hospital and uses it as input.
- Users that build model that produces output prediction (i.e. prob that someone develops alzheimer's disease). This is likely clinic or research lab with technical skillsets.

Use a AI algorithm predicting heart disease risk as an example. The data regulator will determine the criteria of fairness. For example, if they use group fairness, then statistical parity or equal opportunity (eg. in terms of true positive rates) can be potential choices. Data producers will try to produce fair representations of data. Data users will take the data and train models to make predictions for some purpose. An expert in CDC or American Heart Association can be the data regulator who set the criteria. An architecture of a clinic EHR system can be the data producer and design the system to present fair data. A researcher from a lab will be the data user who use data for his/her research.

Responses to L5 Quick Check

Describe biomedical AI applications using the framework for fair AI

In the use of X-ray images to predict heart diseases:

The regulator would determine if certain attributes such as race or income level can be used, as they may introduce bias due to access to care. They will also determine the right fairness metric to use, for example by requiring group fairness across gender groups. Examples of regulators include regulatory boards or administrators

The producer are the ones who would collect and organise the chest x-ray images. To ensure fairness, they might perform survey sampling of collected x-ray images such that the dataset is representative of the population. They may also use image augmentation to obfuscate features that results in leakage of protected attributes such as body shapes that correspond to certain races. Examples include hospital departments or specialised teams in research labs who are responsible for data curation.

The data users are responsible for building fair ML models given the regulations and cleaned data. They will need to adhere to the fairness metrics and make efforts to clean or check for residual biases in their data and models. Examples include researchers in a research lab

In a health system or institution, a data regulator may be a research board like IRB or CMIO team prior to launch of an algorithm - they would determine which criteria constitute fairness and how individuals may be considered similar. The data producer creates the vectors of the people who are similar (they can be constructed based on feature learning or metric learning) - they could be data analysts in the institution. The data user trains the ML model on the vectors from the DP and they could be the computer scientists or informatician in the lab.

Responses to L5 Quick Check

Examples that require implementing individual fairness

When there are high heterogeneity of the samples in the model, ensuring individual fairness is necessary. For example, if we are using the EHR data to predict the cost of the health care of patients for insurance purposes, then ensuring individual fairness is important because the patient population could be very heterogenous and the number of features in the population is high. If we only consider the group fairness, the within group heterogeneity might also be high such that a probability that could apply to the group can't be apply on the individual level.

Guaranteeing individual fairness, which one could see as a stricter form of fairness than group, would be important for applications where underlying confounders are not well-known. A model to decide appointment/scheduling priorities for patients may fit into this group.

It is important to have individual fairness for everyone when determining waiting time appointments that may get determined by AI-based algorithms, this does not necessarily mean being treated the same though. here fairness is attained at the individual level.

Responses to L5 Quick Check

Examples that require implementing individual fairness

Cancer state prediction. Cancer is highly heterogeneous such that group fairness may not make sense in terms of broad categories of race or gender. Predictions should instead ensure that individuals with similar features such as gene mutations or transcriptional profiles receive similar predictions and can be treated similarly.

Guaranteeing individual fairness, which one could see as a stricter form of fairness than group, would be important for applications where underlying confounders are not well-known. A model to decide appointment/scheduling priorities for patients may fit into this group.

Responses to L5 Quick Check

Examples that require implementing group fairness

Group fairness would matter more for when underlying confounders are less significant or if computation complexity prevents ensuring individual fairness. Predicting the outcome of drug intake would be a good example.

Group fairness is necessary for applications that serves a broad variety of people (like the US population as a whole). An example would be the Framingham study, where longitudinal heart disease risk are measured for different groups (race, gender, ...) of people. In such a study, we want the disease risk to be fair across these different groups. Of course, if individual fairness can be achieved at the same time, it would be even better.

When we are trying to build a model to predict the behavior of mice from different species in the same behavioral task, where the genetic background (dominated by species) of mice is one of the main contributor to different behaviors in the task, the group fairness is necessary to be considered.

Responses to L5 Quick Check

Examples that require implementing group fairness

- A model that predicts risk of patients to help insurance systems determine insurance prices.
- These models can be unfair, with “black patients being considerably sicker than white patients at a given risk level.”
- This may be unfair for white individuals who pay more for being not as sick.
- Alternatively, if insurance rates are linked to ability to get medical treatments, this algorithm may be unfair to black people because they are sicker than white people at same level of health insurance.

- Another example is when AI is used to diagnose disease based on chest radiographs.
- Past examples show that under diagnosis in black patients can be common (Seyyed-Kalantari et al., 2021).
- Here, ensuring fairness is necessary to prevent this from happening.

Seyyed-Kalantari, L. et al. (2021) “Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations,” *Nature Medicine*, 27(12), pp. 2176–2182. Available at: <https://doi.org/10.1038/s41591-021-01595-0>.

Outline for today's class

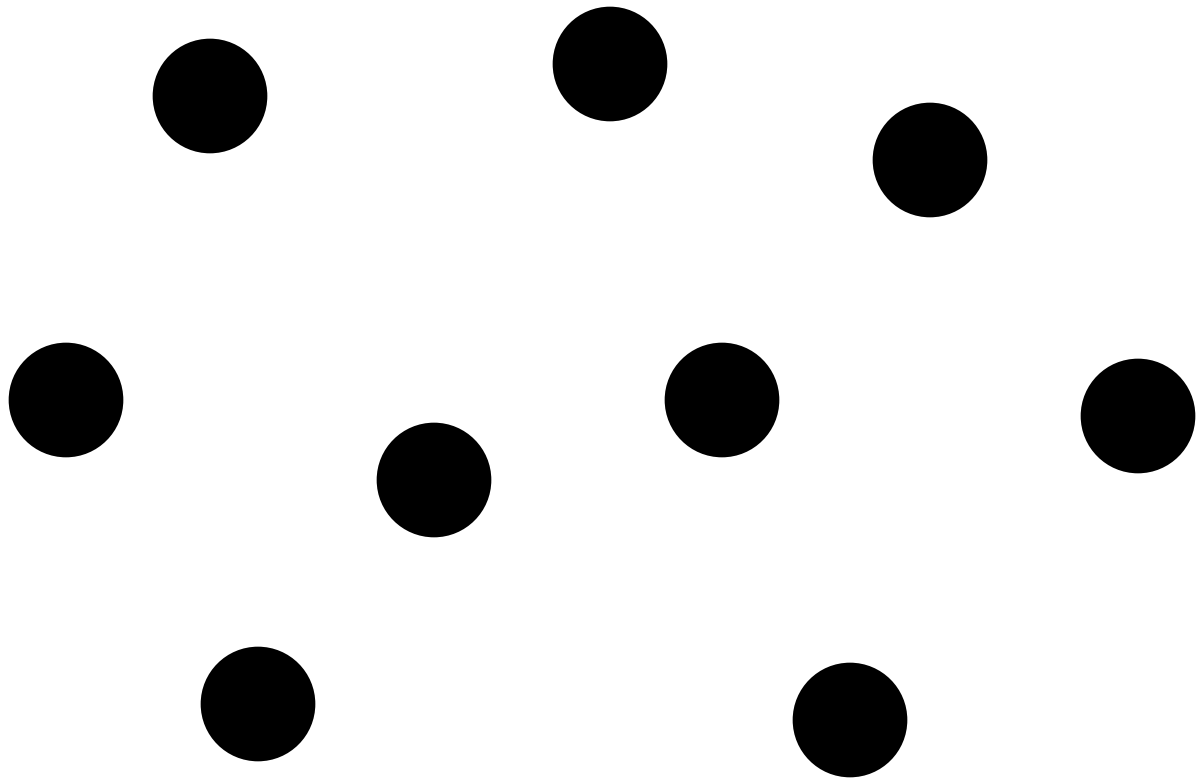
- **Foundations of network medicine**
- **Foundations of geometric deep learning**
 - Node classification, link prediction, graph clustering, graph classification
 - Semi-supervised learning and label diffusion
- **Genes-like-me: What does my gene do?**
- **Patients-like-me: Finding patients with similar genetic and phenotypic features**

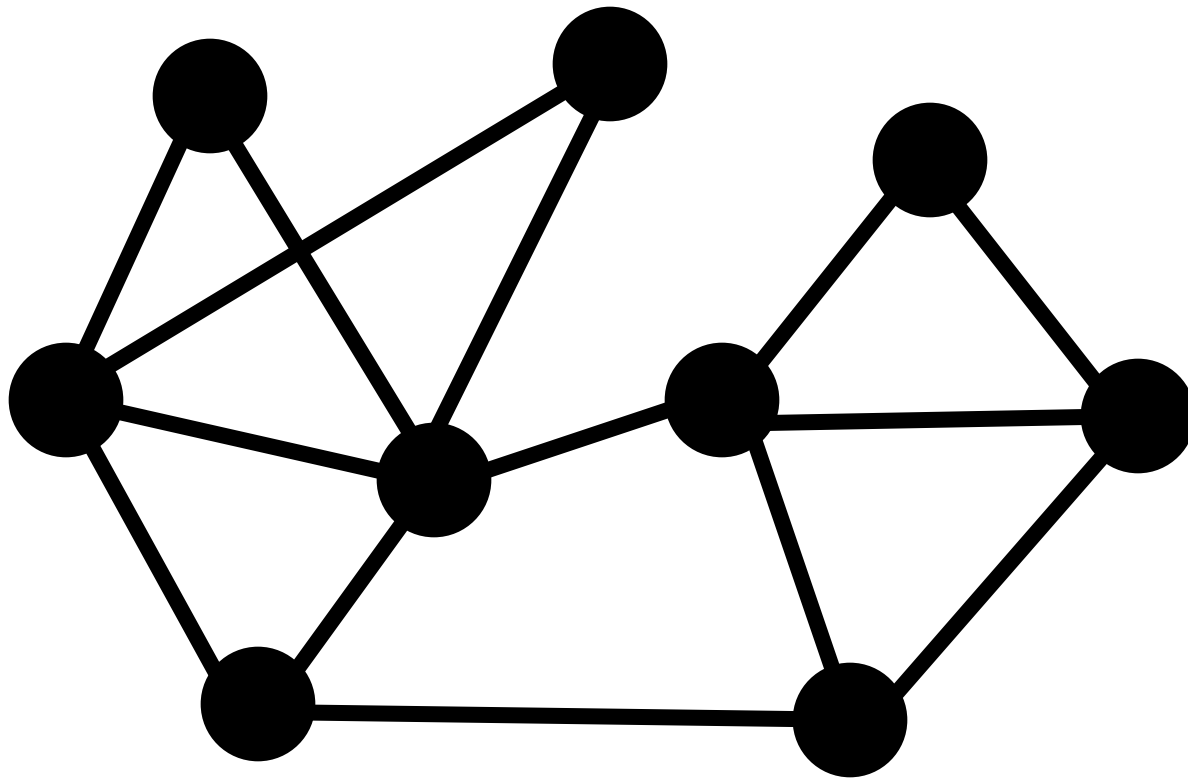
Foundations of network medicine

**What are networks/graphs?
Predictive modeling using graphs**

Why networks?

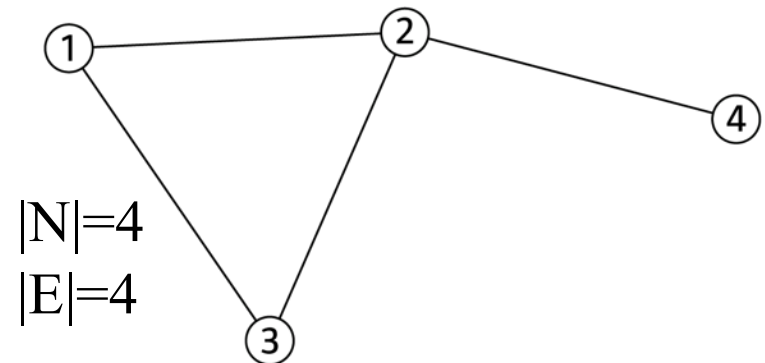
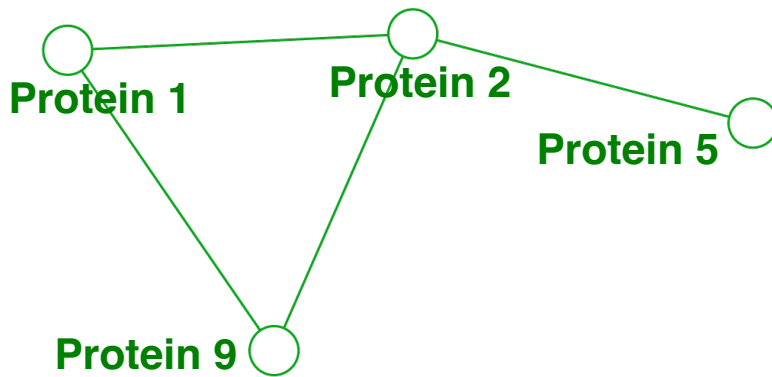
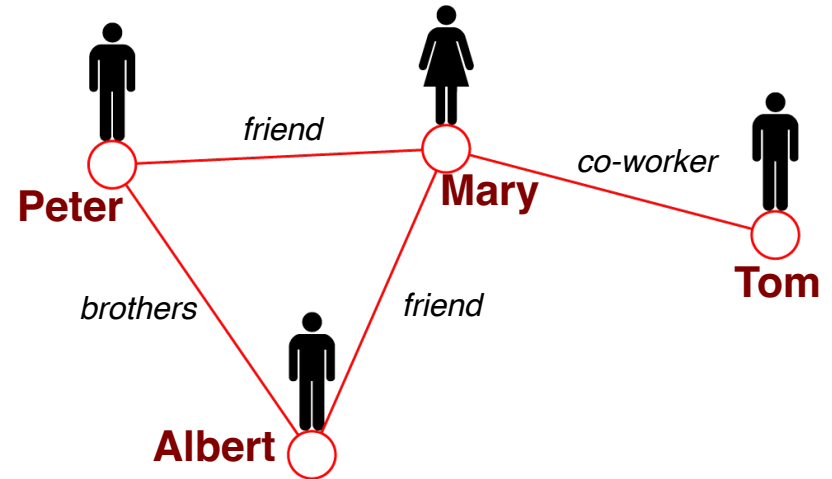
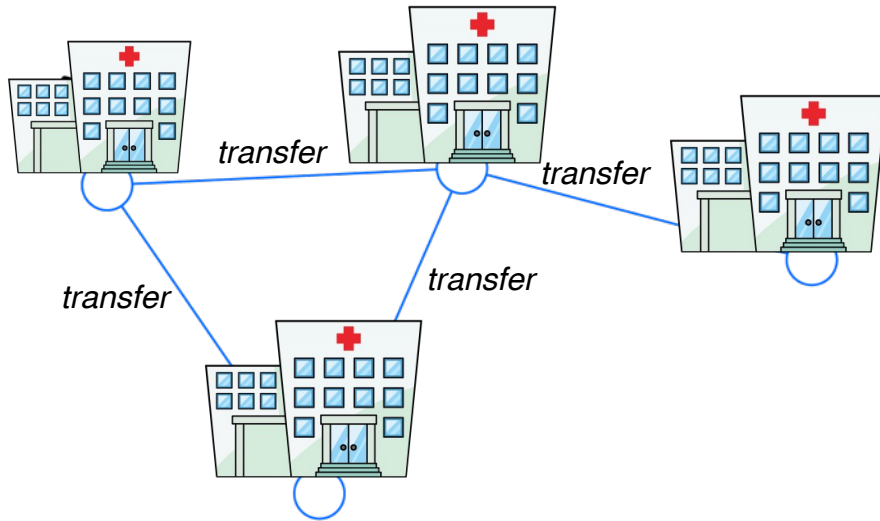
Networks are a general language for describing and modeling complex systems





Network!

General Mathematical Language



Why Networks? Why Now?

- **Question:** How are diseases and disease genes related to each other?
- **Findings:** Disease genes likely to interact and have similar expression

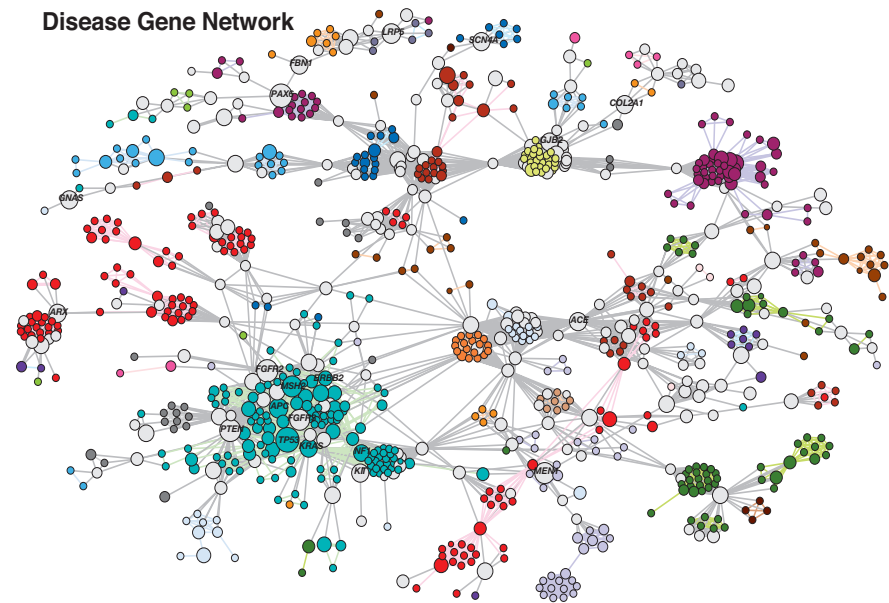
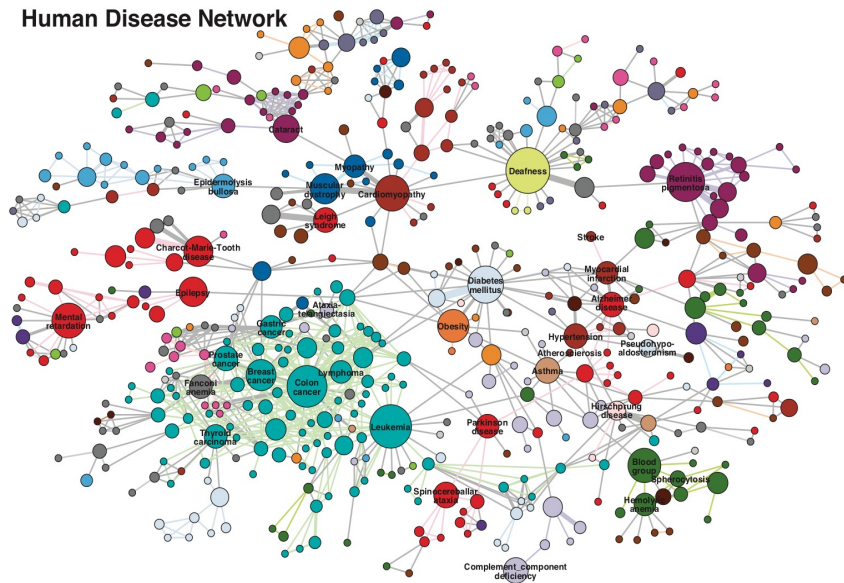


Image from: Goh et al. 2007. [The human disease network](#). PNAS.

Why Networks? Why Now?

- **Question:** How to simulate an eukaryotic cell?
- **Findings:** Simulations reveal molecular mechanisms of cell growth, drug resistance and synthetic life

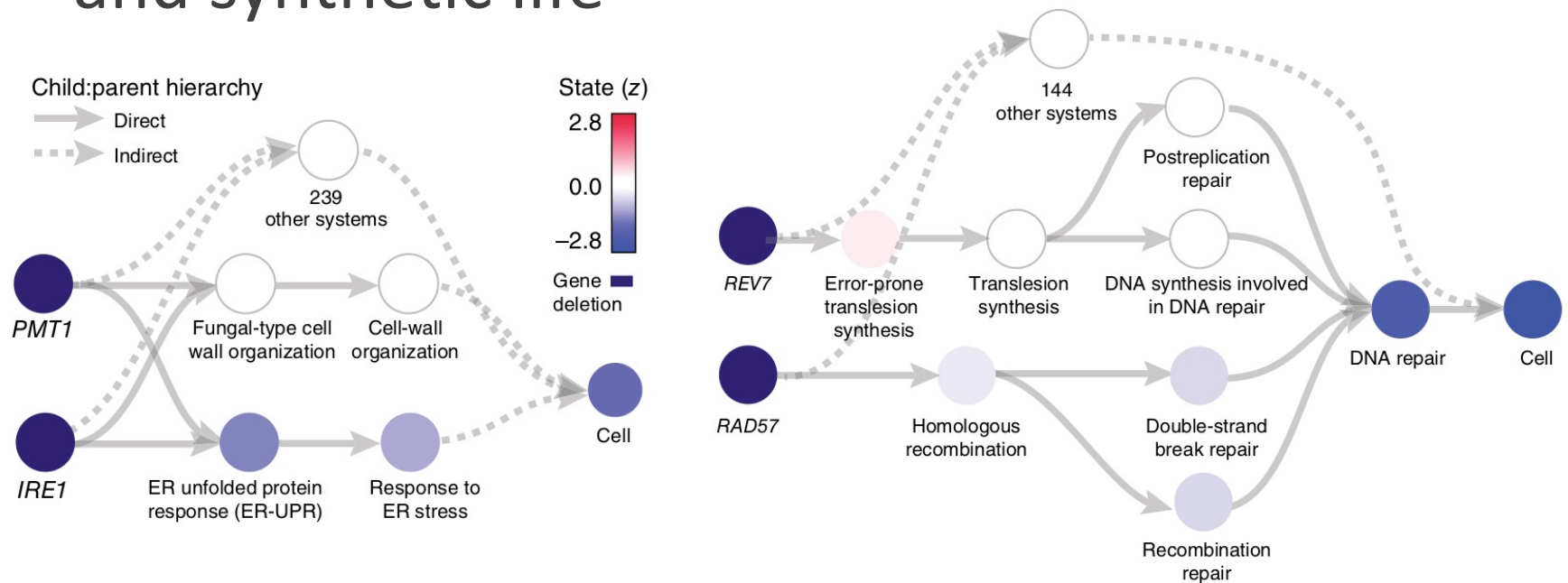


Image from: Ma et al. 2018. [Using deep learning to model the hierarchical structure and function of a cell.](#) *Nature Methods*.

Why Networks? Why Now?

- **Question:** How to model cancer heterogeneity?
- **Findings:** New cancer subtypes with distinct patient survival

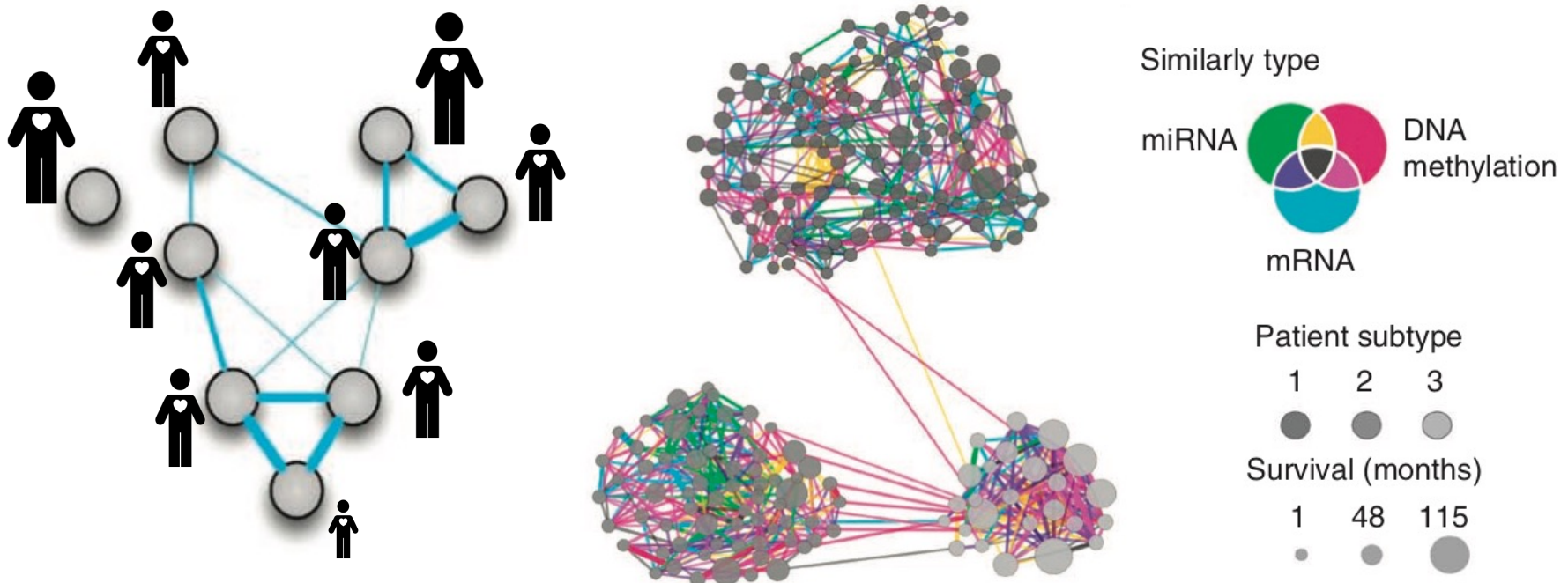


Image from: Wang et al. 2014. [Similarity network fusion for aggregating data types on a genomic scale](#). *Nature Methods*.

Why Networks? Why Now?

- **Question:** How to study ecological systems?
- **Findings:** Pollinators interact with flowers in one season but not in another, and the same flower species interact with both pollinators and herbivores

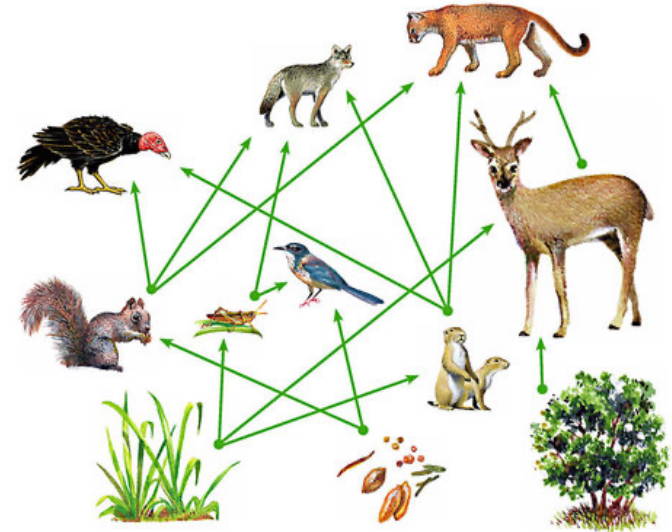
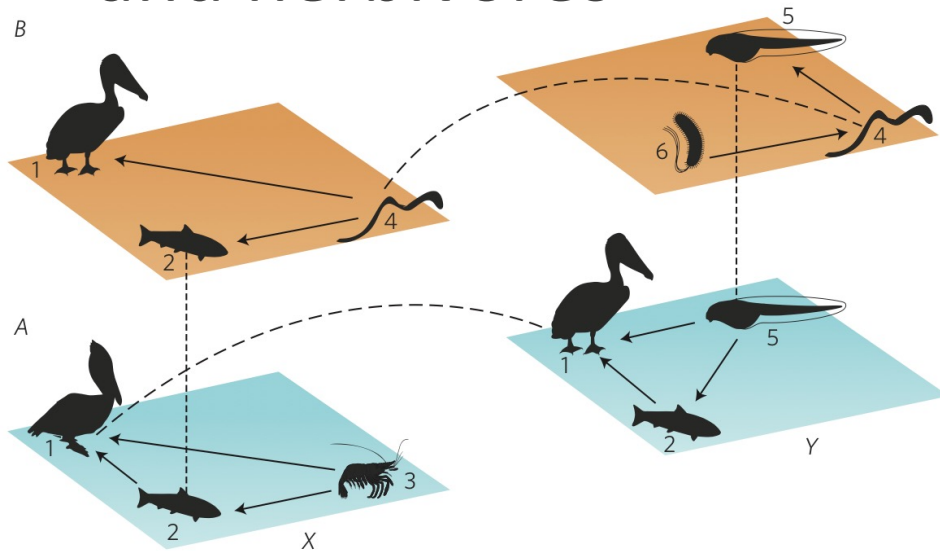


Image from: Pilosof et al. 2017. [The multilayer nature of ecological networks.](#) *Nature Ecology and Evolution.*

Why Networks? Why Now?

- **Question:** Do large, dense, and cosmopolitan areas support socioeconomic mixing and exposure among diverse individuals?
- **Findings:** Contrary to expectations, residents of large cosmopolitan areas have less exposure to a socioeconomically diverse range of individuals

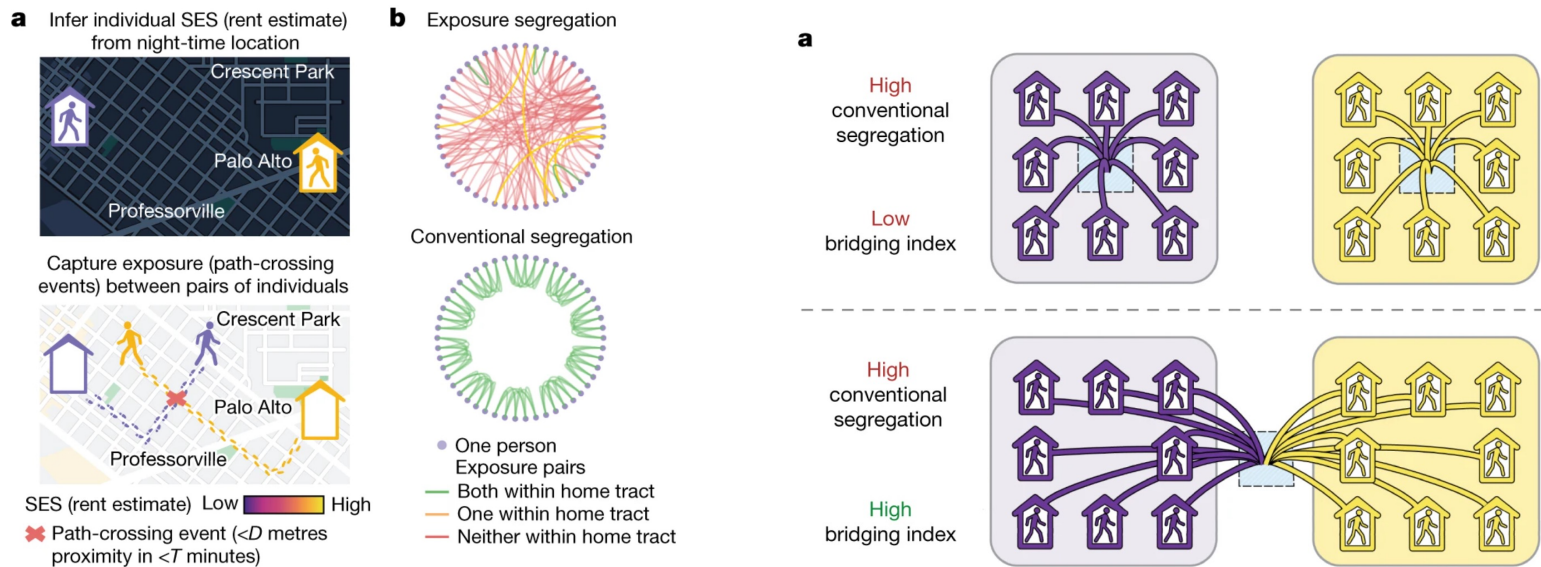


Image from: Nilforoshan et al. 2023. [Human mobility networks reveal increased segregation in large cities.](#) *Nature*.

Why Networks? Why Now?

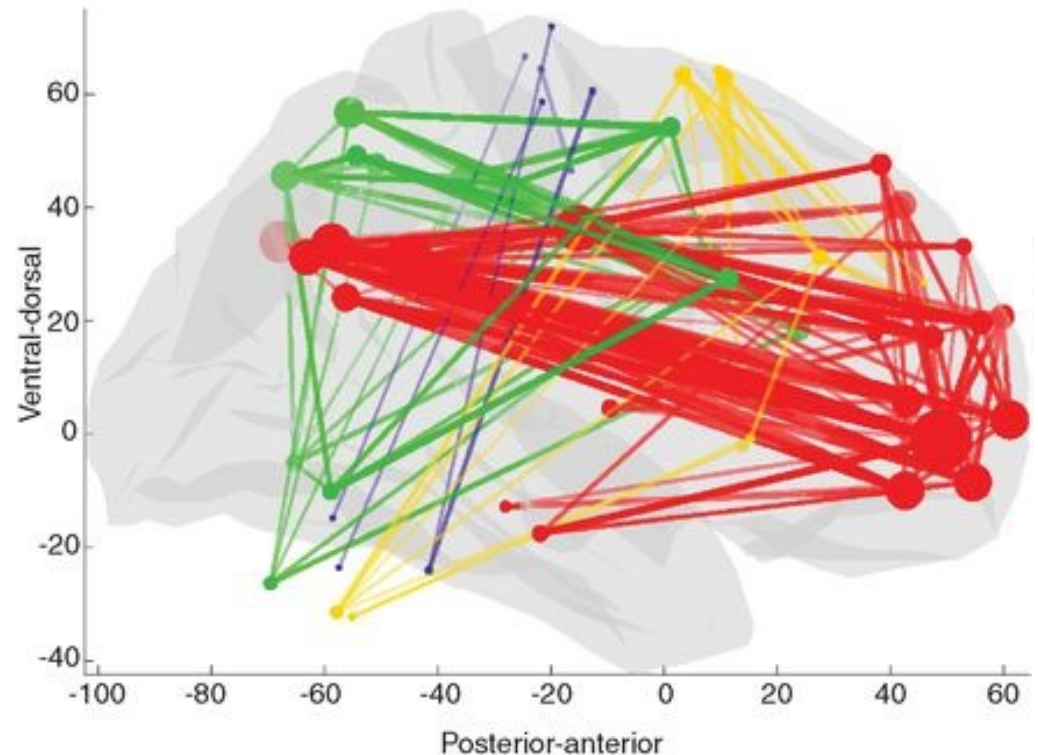
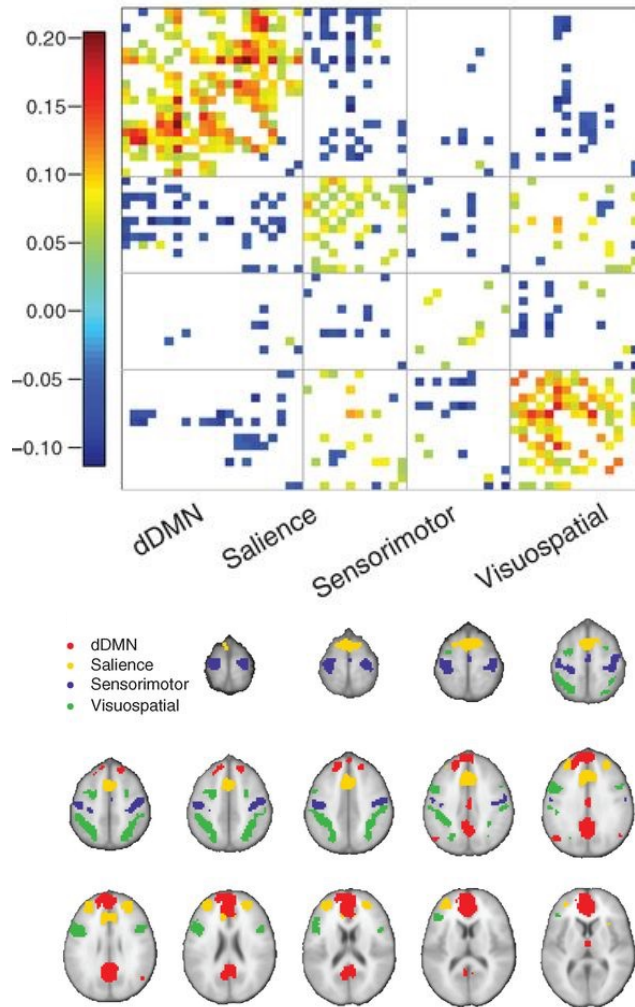
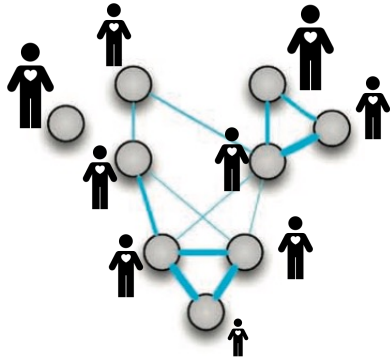
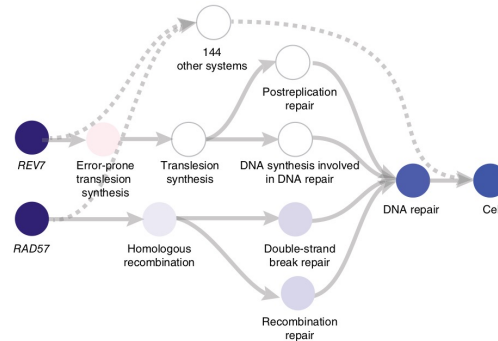


Image from: Richiardi et al. 2015. [Correlated gene expression supports synchronous activity in brain networks.](#) *Science*.

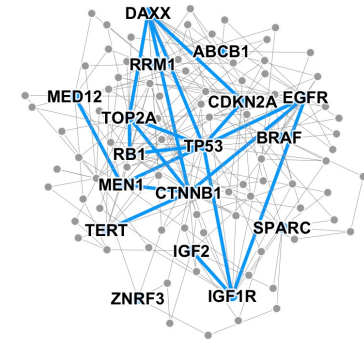
Many Data are Networks



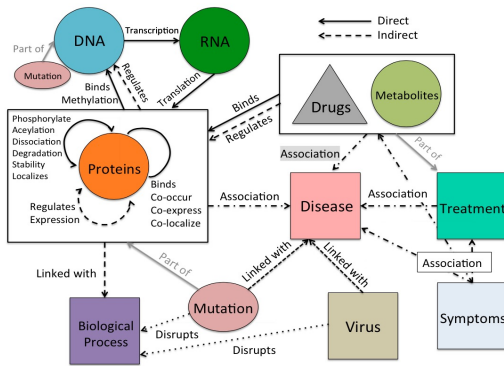
Patient networks



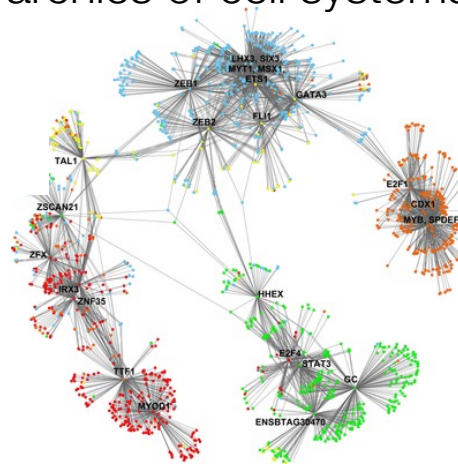
Hierarchies of cell systems



Disease pathways



Biomedical knowledge graphs



Gene interaction networks



Cell-cell similarity networks

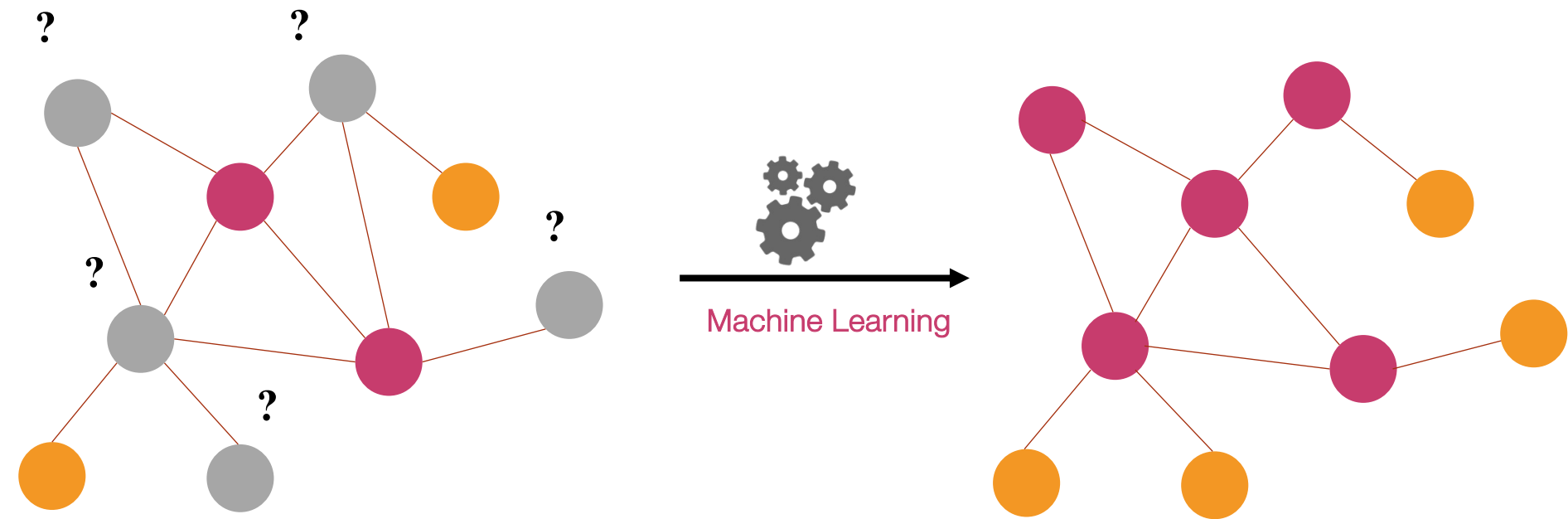
Evolution of Resilience in Protein Interactomes Across the Tree of Life, *PNAS*, 2019; MARS: Discovering Novel Cell Types across Heterogeneous Single-Cell Experiments, *Nat Methods*, 2020; Leveraging the Cell Ontology to Classify Unseen Cell Types, *Nat Commun*, 2021; Identification of Disease Treatment Mechanisms through the Multiscale Interactome, *Nat Commun*, 2021; Network Medicine Framework for Identifying Drug Repurposing Opportunities for COVID-19, *PNAS*, 2021; Population-Scale Patient Safety Data Reveal Inequalities in Adverse Events Before and During COVID-19 Pandemic, *Nat Comput Science*, 2021 20

Predictive and Generative Modeling

- **Predict a type of a given node**
 - Node classification
- **Predict whether two nodes are linked**
 - Link prediction
- **Identify densely linked clusters of nodes**
 - Community detection, module detection
- **How similar are two nodes/networks**
 - Network similarity
- **Design graphs with desirable properties**
 - Generative modeling and molecular design

This topic will be covered in M6: Generative AI

Node Classification



Node Classification: Example

Classifying the function of proteins in the interactome!

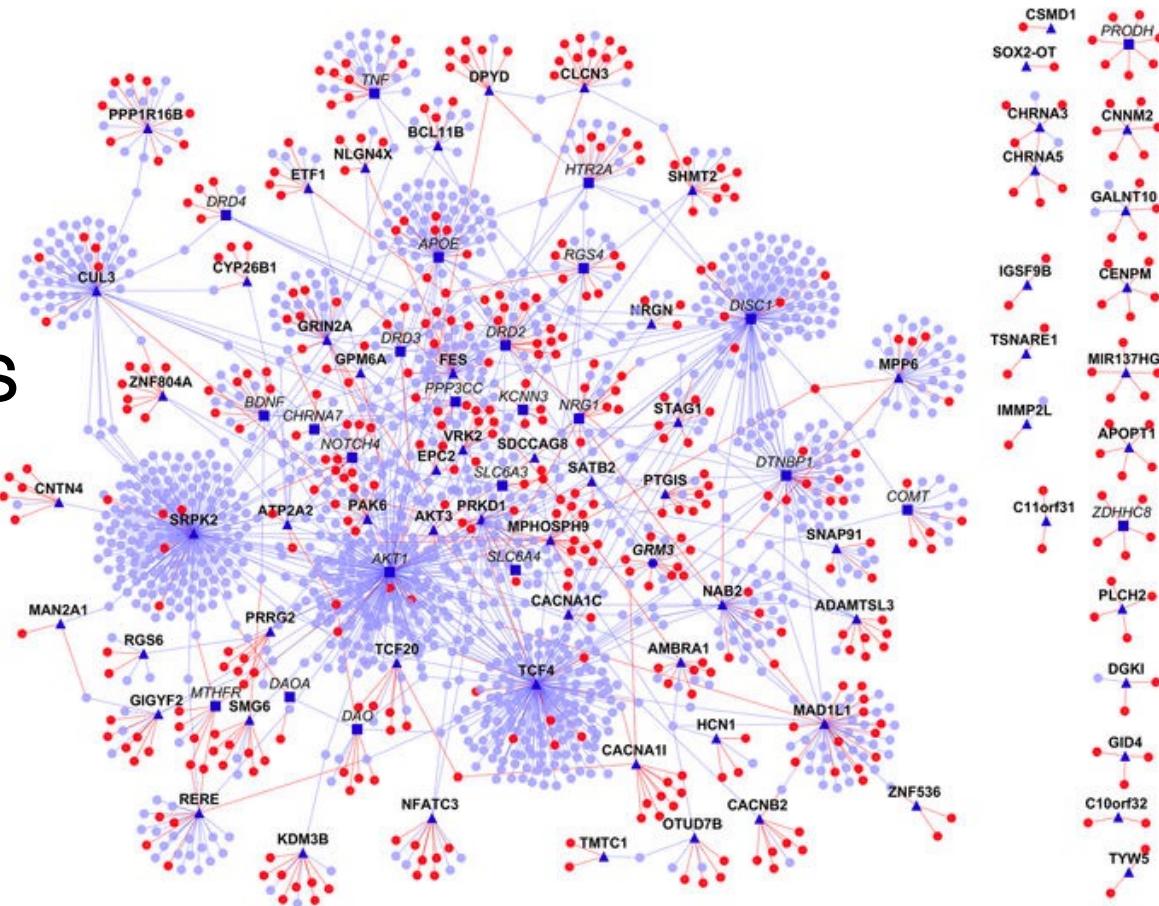
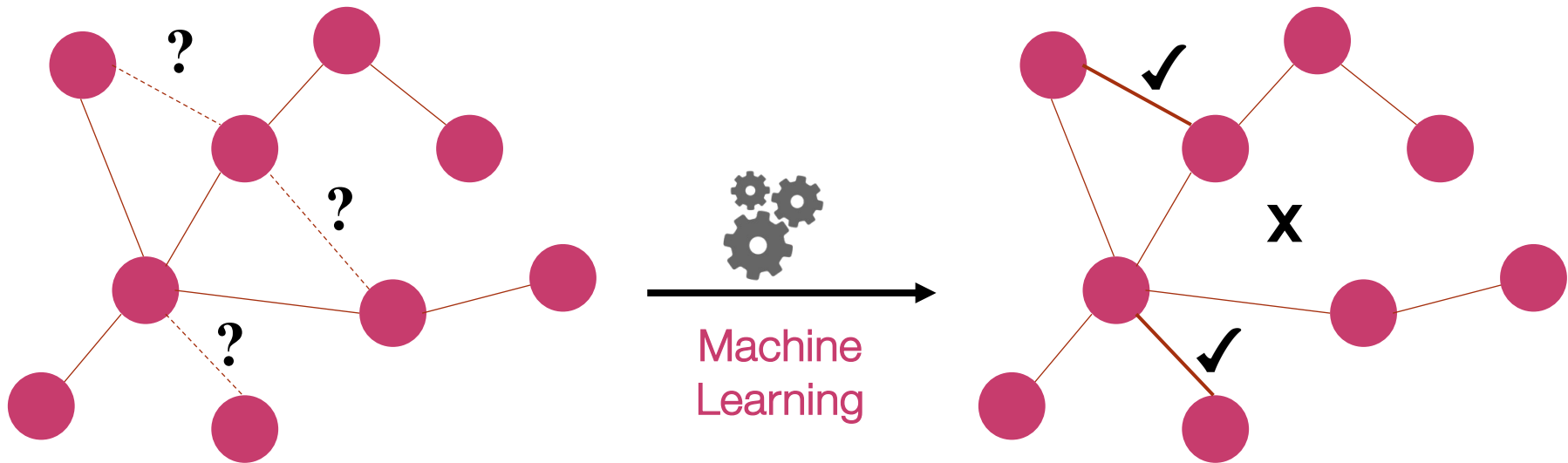


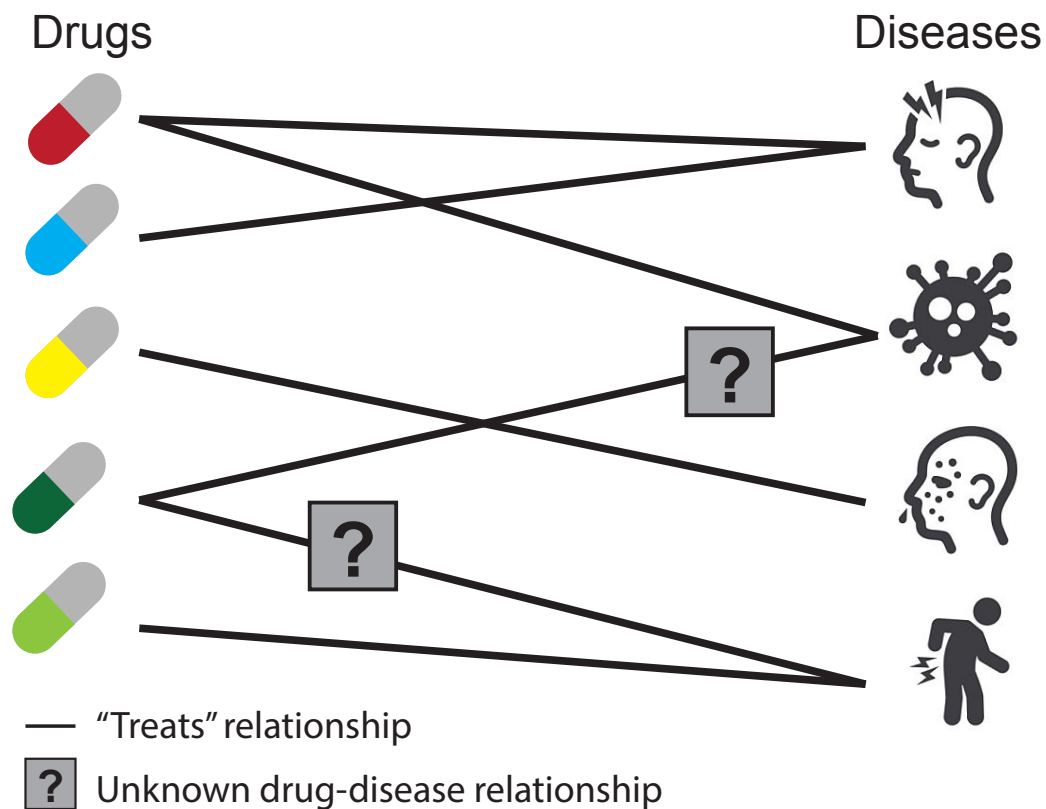
Image from: Ganapathiraju et al. 2016. [Schizophrenia interactome with 504 novel protein-protein interactions](#). *Nature*.

Link Prediction

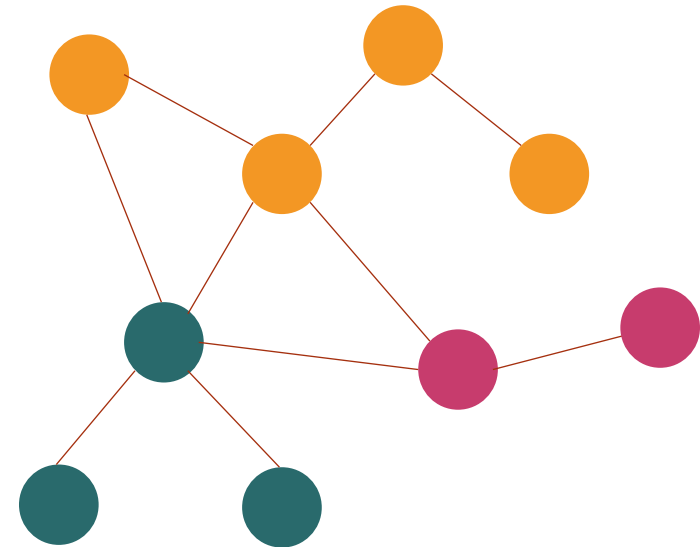
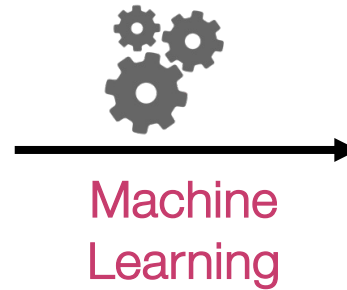
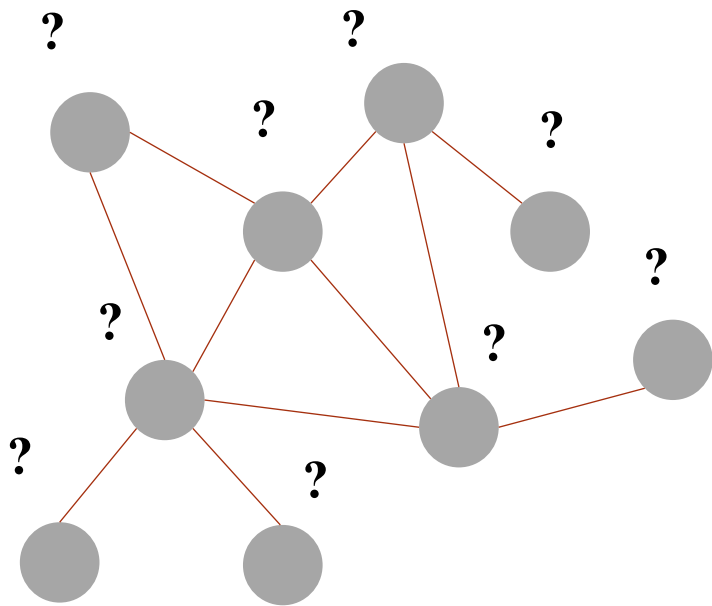


Link Prediction: Example

Predicting which diseases a new molecule might treat!



Community Detection



Community Detection: Example

Identifying
disease proteins
in the
interactome!

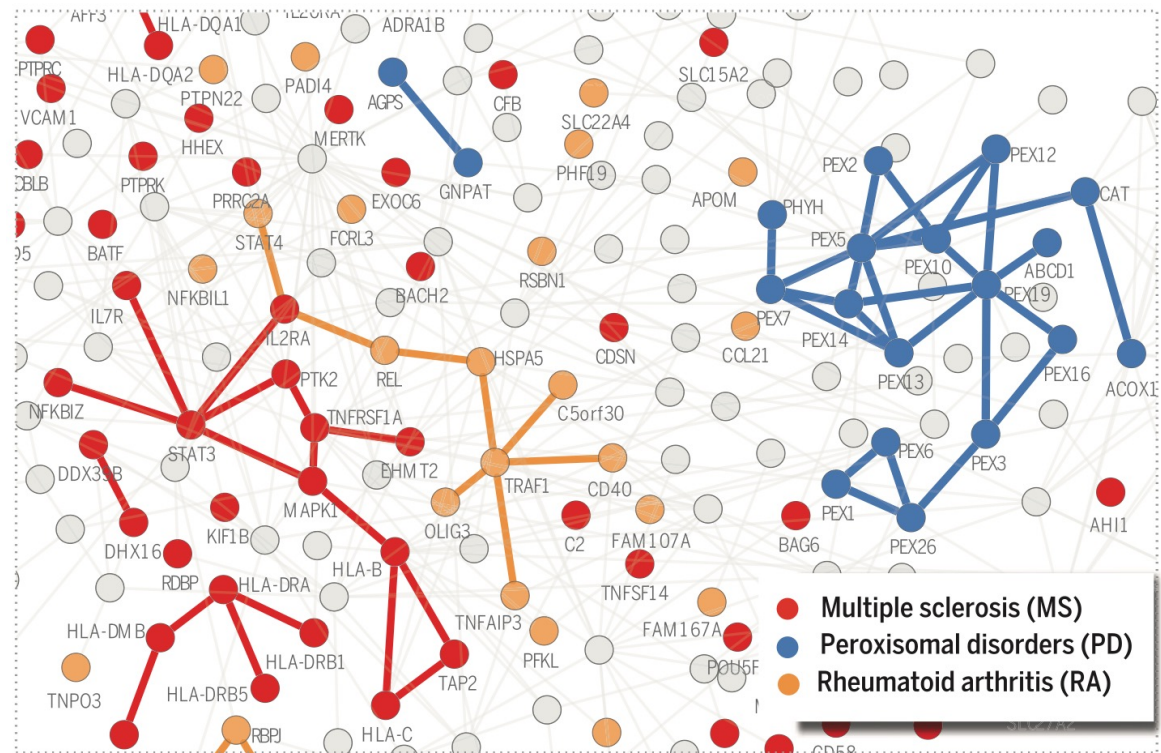
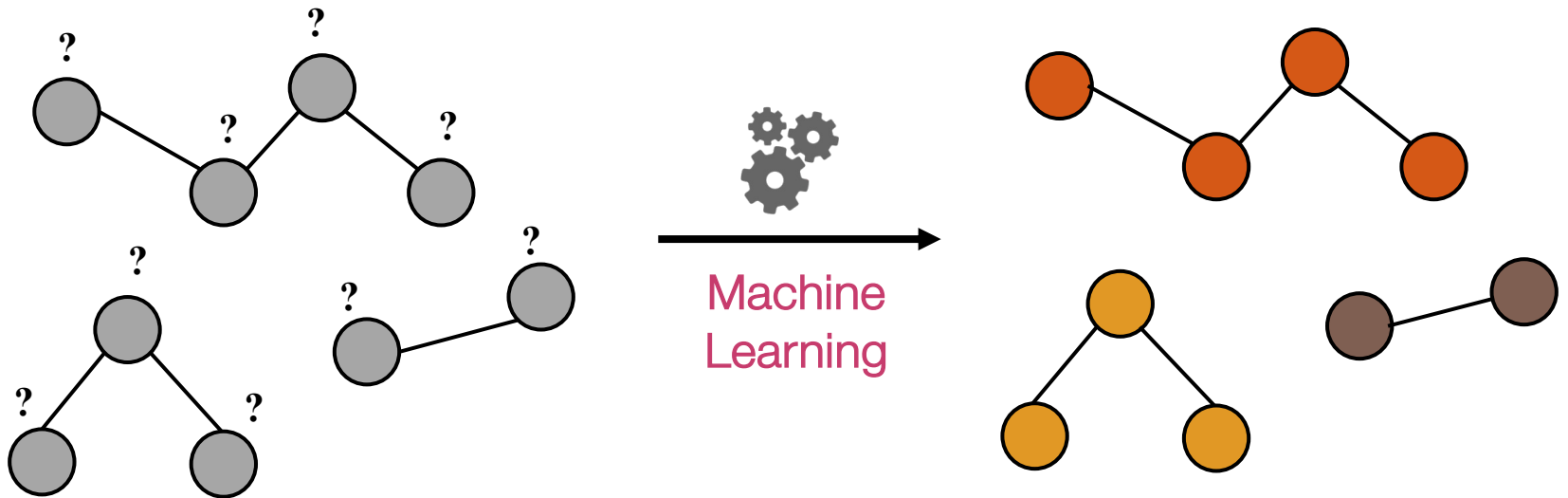


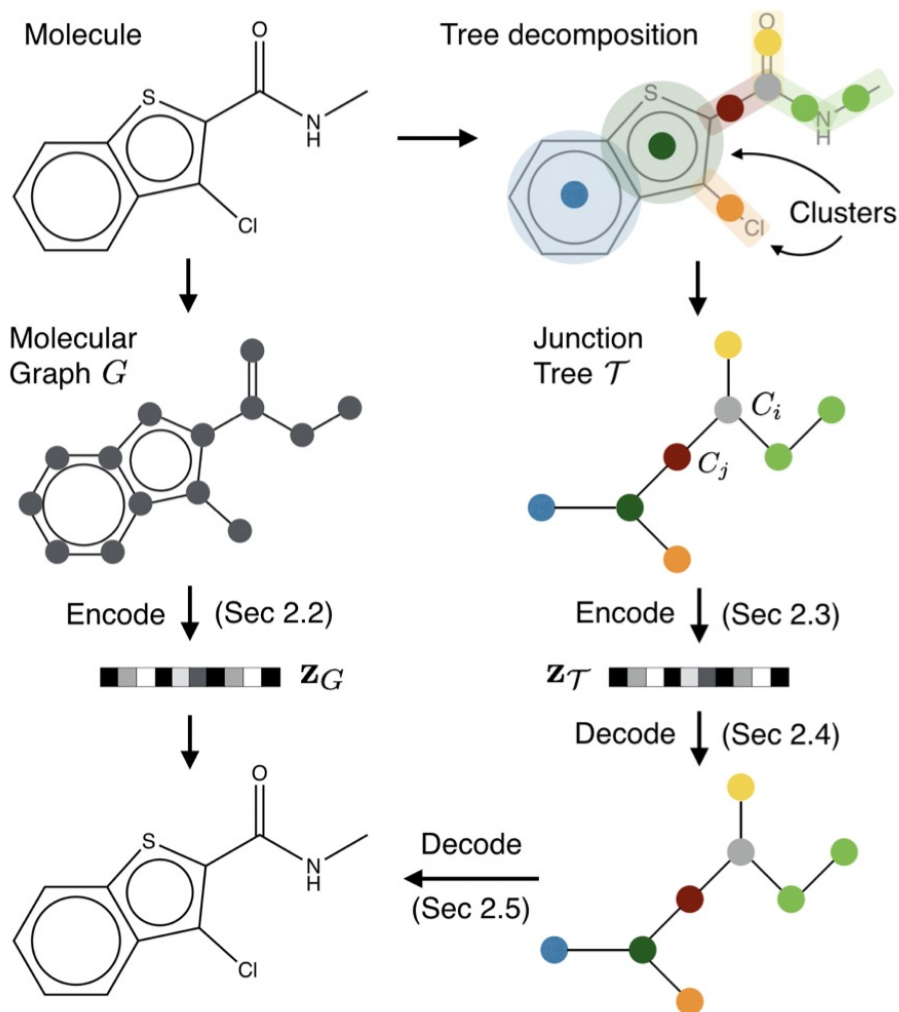
Image from: Menche et al. 2015. [Uncovering disease-disease relationships through the incomplete interactome](#). *Science*.

Graph Classification



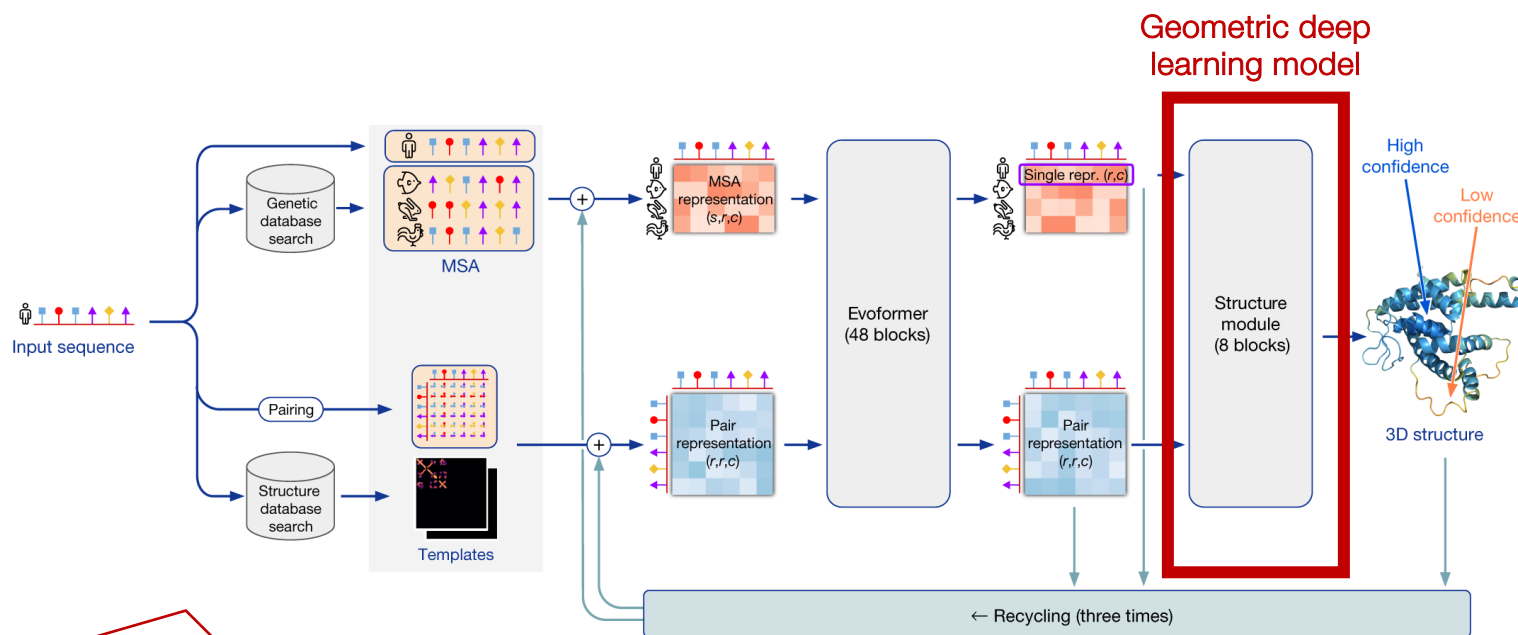
Graph Classification: Example

Designing new small molecule compounds to treat a disease!



Generative Modeling and Design

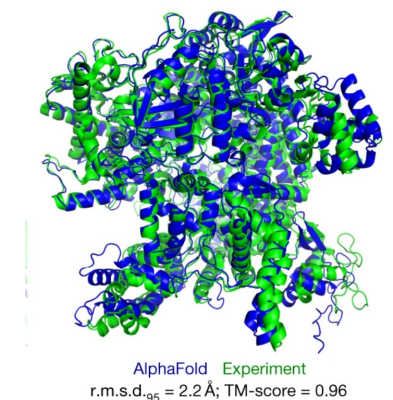
Geometric deep learning underlies several breakthroughs, including AlphaFold for protein structure prediction



Geometric deep learning is receiving increasing interest in biology, chemistry, and medical sciences as a new tool for molecular design and optimization

AlphaFold Network

- What drives accurate protein structure prediction?
 - Novel neural architecture based on the **evolutionary**, **physical** and **geometric** constraints of protein structures
- Input:
 - Primary AA sequence of a given protein
 - Aligned sequences of homologues
- Output:
 - Predicted 3D coordinates of all heavy atoms in a protein

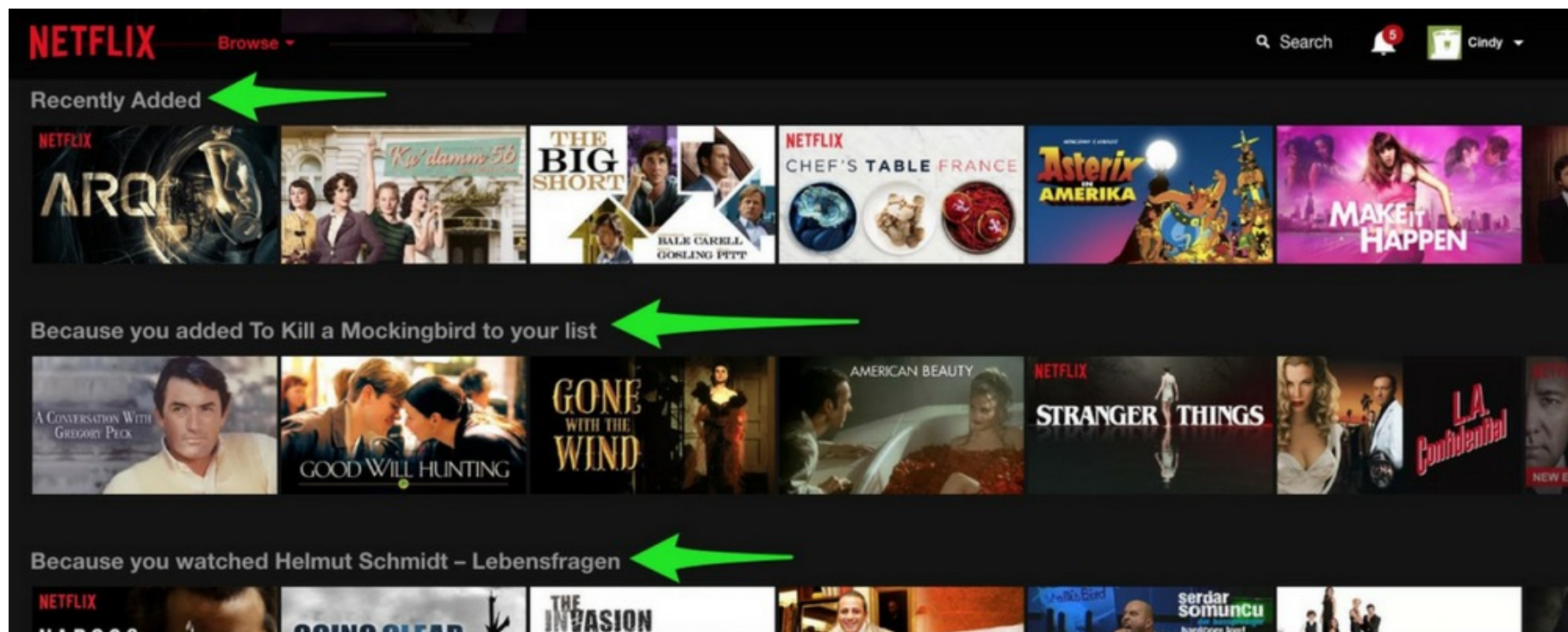


Genes-like-me

What does my gene do? Give me more genes like these

Recommender Systems

Consider user x : Find set S of other users whose ratings are “similar” to x ’s ratings; Estimate x ’s preference based on ratings in S

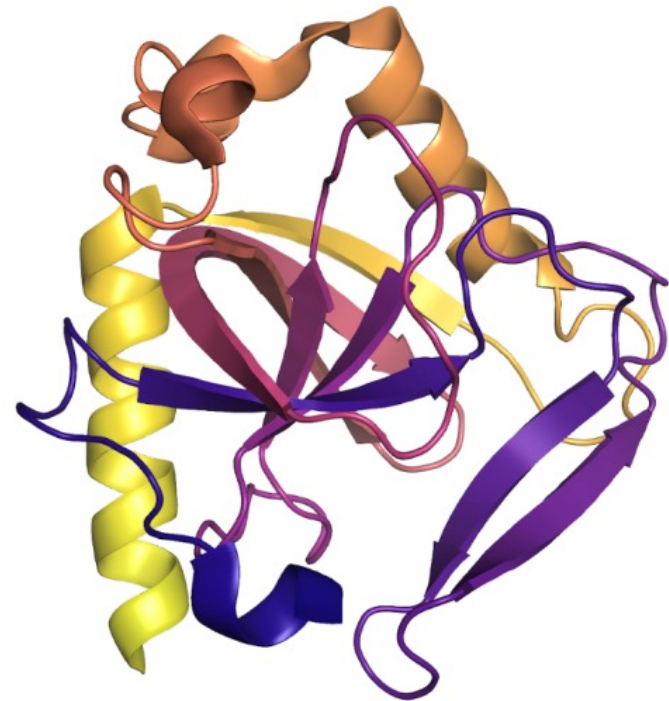


Recommender Systems in Biology

“Give me more
movies like
this one”

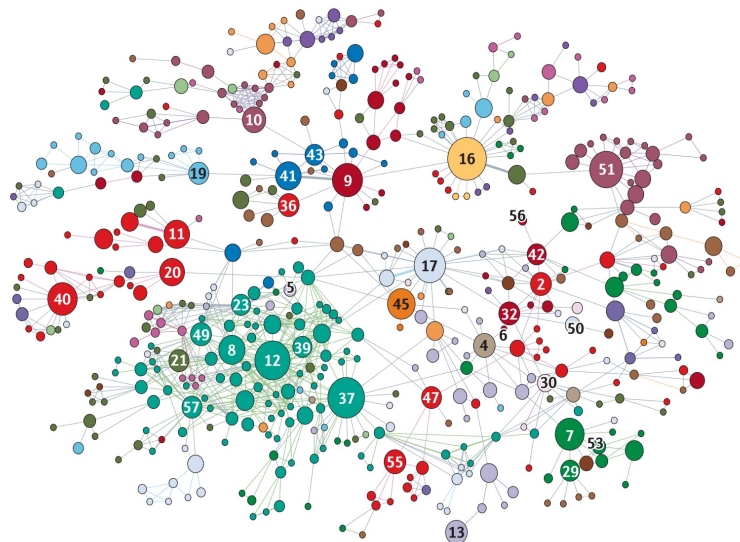
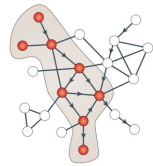


“Give me more
proteins like
this one”



Biological Rationales

- **Local hypothesis:** Proteins involved in the same disease have an increased tendency to interact with each other
- **Disease module hypothesis:** Cellular components associated with disease tend to cluster in the same network neighborhood



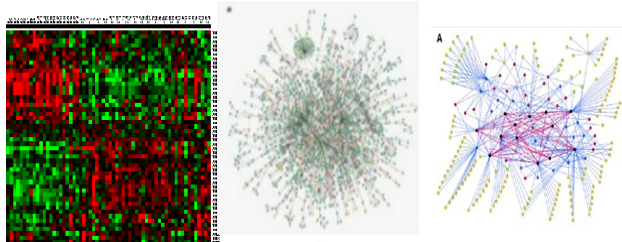
① Aldosteronism	②① Epilepsy	④② Myocardial infarction
② Alzheimer's disease	②① Fanconi's anaemia	④③ Myopathy
③ Anaemia, congenital deserythropoietic	②② Fatty liver	④④ Nucleoside phosphorylase deficiency
④ Asthma	②③ Gastric cancer	④⑤ Obesity
⑤ Ataxia-telangiectasia	②④ Gilbert's syndrome	④⑥ Paraganglioma
⑥ Atherosclerosis	②⑤ Glaucoma 1A	④⑦ Parkinson's disease
⑦ Blood group	②⑥ Goitre congenital	④⑧ Pheochromocytoma
⑧ Breast cancer	②⑦ HARP syndrome	④⑨ Prostate cancer
⑨ Cardiomyopathy	②⑧ HELLP syndrome	④⑩ Pseudohypoadosteronism
⑩ Cataract	②⑨ Haemolytic anaemia	④⑪ Retinitis pigmentosa
⑪ Charcot–Marie–Tooth disease	③⑩ Hirschprung disease	④⑫ Schizoaffective disorder
⑫ Colon cancer	③⑪ Hyperbilirubinaemia	④⑬ Spherocytosis
⑬ Complement component deficiency	③⑫ Hypertension	④⑭ Spina bifida
⑭ Coronary artery disease	③⑬ Hypertension diastolic	④⑮ Spinocerebellar ataxia
⑮ Coronary spasm	③⑭ Hyperthyroidism	④⑯ Stroke
⑯ Deafness	③⑮ Hypoadosteronism	④⑰ Thyroid carcinoma
⑰ Diabetes mellitus	③⑯ Leigh syndrome	④⑱ Total iodide organification defect
⑱ Enolase-β deficiency	③⑰ Leukaemia	④⑲ Trifunctional protein deficiency
⑲ Epidermolysis bullosa	③⑱ Lymphoma	④⑳ Unipolar depression
	④⑰ Mental retardation	
	④⑱ Muscular dystrophy	

Recommender Systems in Biology

- **“What does my gene do?”**
 - **Goals:** Determine a gene’s function based on who it interacts with – **“guilty-by-association” principle**
- **“Give me more genes like these”**
 - **Goals:**
 - Find more multiple sclerosis genes
 - Find new ciliary genes
 - Find members of a proteasome complex, etc.

“What Does My Gene Do?”

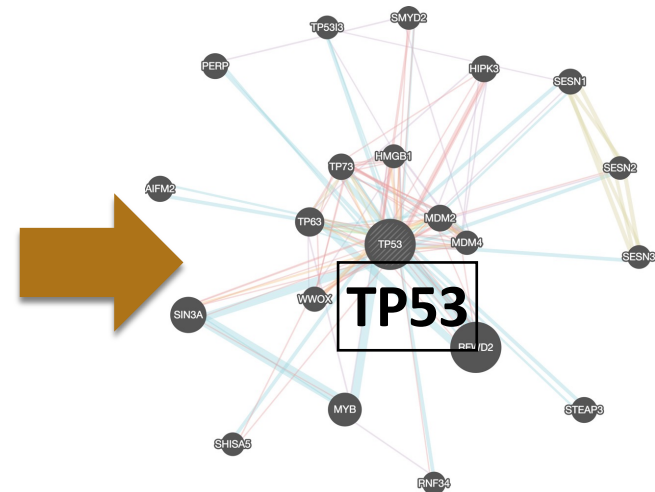
Networks



Query gene

TP53

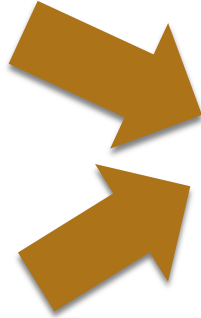
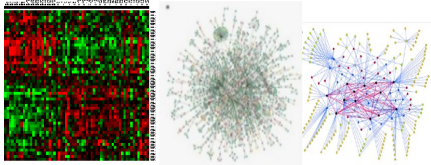
Find set N of other genes whose interactions are “similar” to $TP53$'s interactions



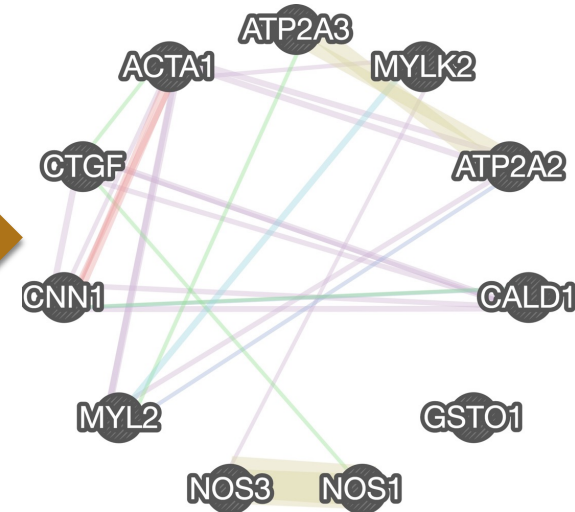
Prediction using guilty-by-association principle: Estimate $TP53$'s function in the cell based on functions of genes in N

“Give Me More Genes Like These”

Networks



Gene recommender system



Query genes

ATP2A2
ATP2A3
NOS1
CNN1
GSTO1
NOS3
CNN2
MYLK2
CALD1
ACTA1
MYL2

Networks

- Co-expression
- Shared protein domains
- Physical interactions
- Pathway
- Co-localization
- Genetic interactions

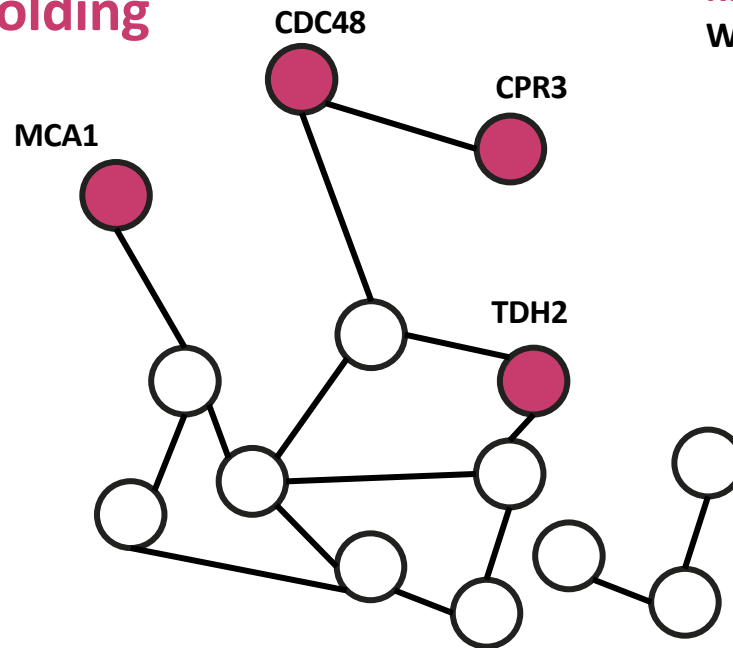
Functions

- muscle system process
- muscle contraction
- regulation of system process
- regulation of muscle system process
- heart contraction

Finding “Guilty Associates”

- Predict gene functions using **guilty-by-association**:

Protein folding



Red: Genes involved in protein folding
White: Genes with unknown function

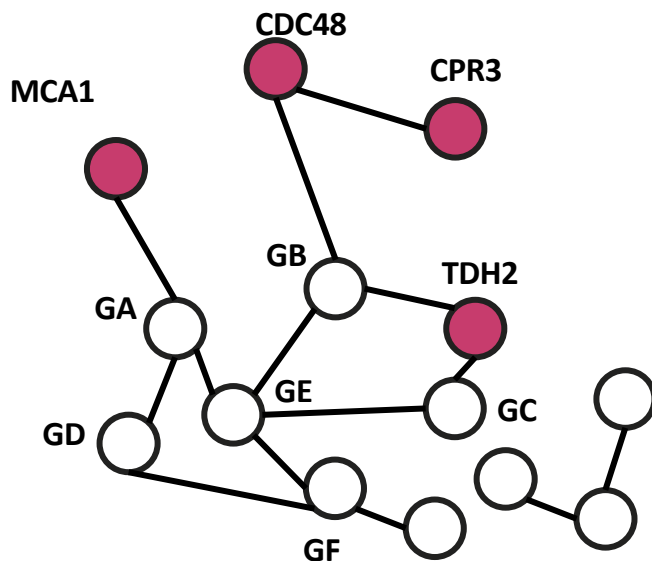
- What other genes participate in “protein folding”?

“Guilty Associates” Problem

- Let W be a $n \times n$ (weighted) adjacency matrix over n genes
- Let $\mathbf{y} = \{-1, 0, 1\}^n$ be a vector of **labels**:
 - 1: **positive** gene, known to be involved in a gene function/biological process
 - -1: **negative** gene
 - 0: **unlabeled** gene
- **Goal**: Predict which **unlabeled** genes are likely **positive**

“Guilty Associates” Approach

- **Approach:** Learn a vector of discriminant scores f , where f_i is **likelihood** that node i is positive
- **Example:**



$$y = [1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0]$$

W = (weighted) adjacency matrix

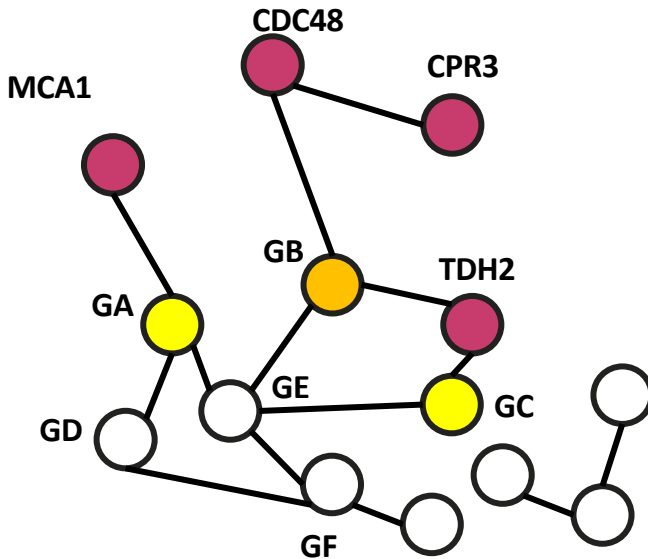
$$f = ?$$

Approach 1: Neighbor Scoring

- Node score f_i is weighted sum of the labels of i 's **direct neighbors**:

$$f_i = \sum_{j=1}^n W_{ij} y_j$$

- Example:**



$$f_{GA} = W_{GA,MCA1} \cdot y_{MCA1}$$

$$f_{GB} = W_{GB,CDC48} \cdot y_{CDC48} + W_{GB,TDH2} \cdot y_{TDH2}$$

$$f_{GC} = W_{GC,TDH2} \cdot y_{TDH2}$$

Red: Positive nodes

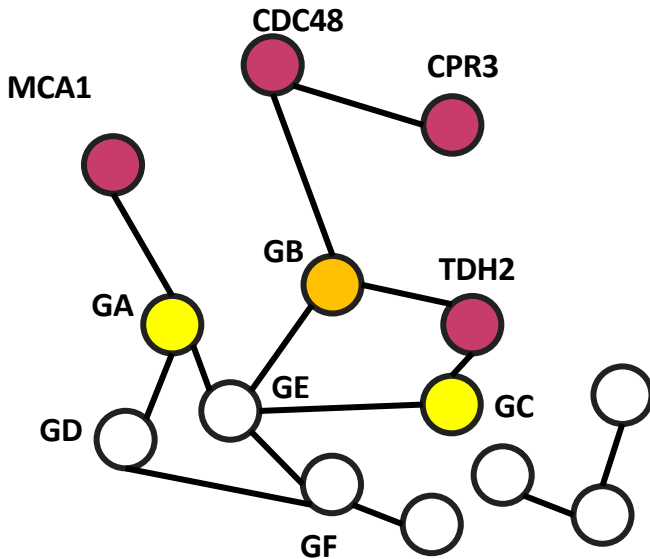
White: $f_i = 0$

Approach 1: Neighbor Scoring

- Node score f_i is weighted sum of the labels of i 's **direct neighbors**:

$$f_i = \sum_{j=1}^n W_{ij} y_j$$

- Example:**



$$f_{GA} = W_{GA,MCA1} \cdot y_{MCA1}$$

$$f_{GB} = W_{GB,CDC48} \cdot y_{CDC48} + W_{GB,TDH2} \cdot y_{TDH2}$$

$$f_{GC} = W_{GC,TDH2} \cdot y_{TDH2}$$

- One half of GC's neighbors are positives
- One third of GA's neighbors are positives
- But:** $f_{GC} = f_{GA}$ (if W is binary)

Weighted Neighbors

- Normalize matrix W by node degrees:

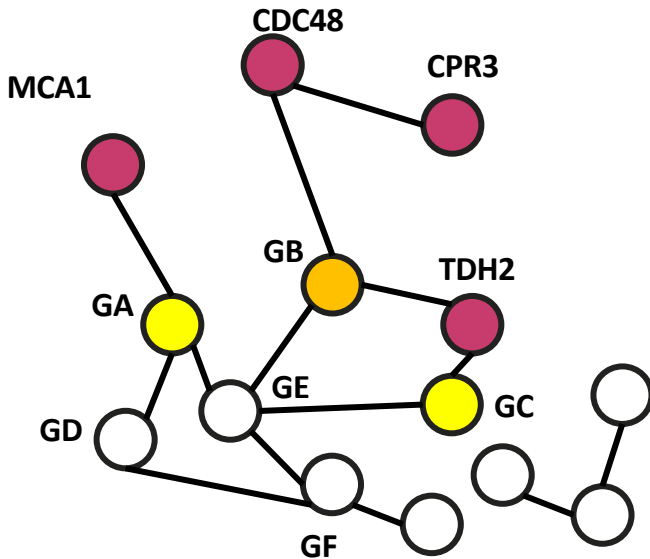
$$f_i = \frac{1}{d_i} \sum_{j=1}^n W_{ij} y_j, \quad d_i = \sum_j W_{ij}$$

Matrix notation:

$$f_i = D^{-1} W y$$

$$D = \text{diag}(d)$$

- Example:



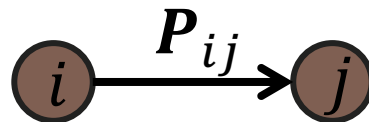
$$f_{GA} = \frac{1}{3} W_{GA, MCA1} \cdot y_{MCA1}$$

$$f_{GB} = \frac{1}{3} (W_{GB, CDC48} \cdot y_{CDC48} + W_{GB, TDH2} \cdot y_{TDH2})$$

$$f_{GC} = \frac{1}{2} W_{GC, TDH2} \cdot y_{TDH2}$$

Random Walks

- Matrix $\mathbf{P} = \mathbf{D}^{-1}\mathbf{W}$ is known as **Markov transition matrix**
 - \mathbf{D} is a diagonal matrix with diagonal elements d_i
 - \mathbf{P} is **a row stochastic matrix**, $\sum_j P_{ij} = 1$
- Row i is a probability distribution over **random walks** starting at node i

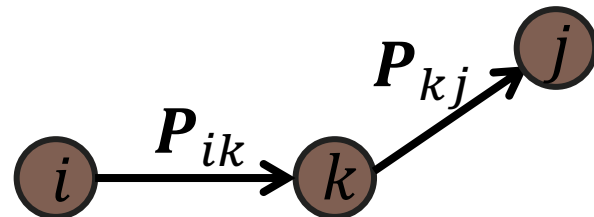


- P_{ij} is probability of a **random walker following a link from node i to node j**

Indirect Neighbor Scoring

- Use **random walks** to include **indirect neighbors** in computations
- **Idea:** Extend **direct neighbor scoring** formula $f = D^{-1}W\mathbf{y} = P\mathbf{y}$ to include **2-hop neighbors**
- Probability of a random walk of length **two** between node i and node j is:

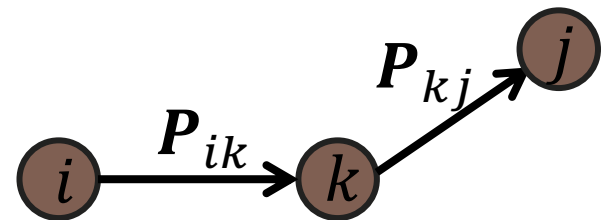
$$[P^2]_{ij} = \sum_{k=1}^n P_{ik}P_{kj}$$



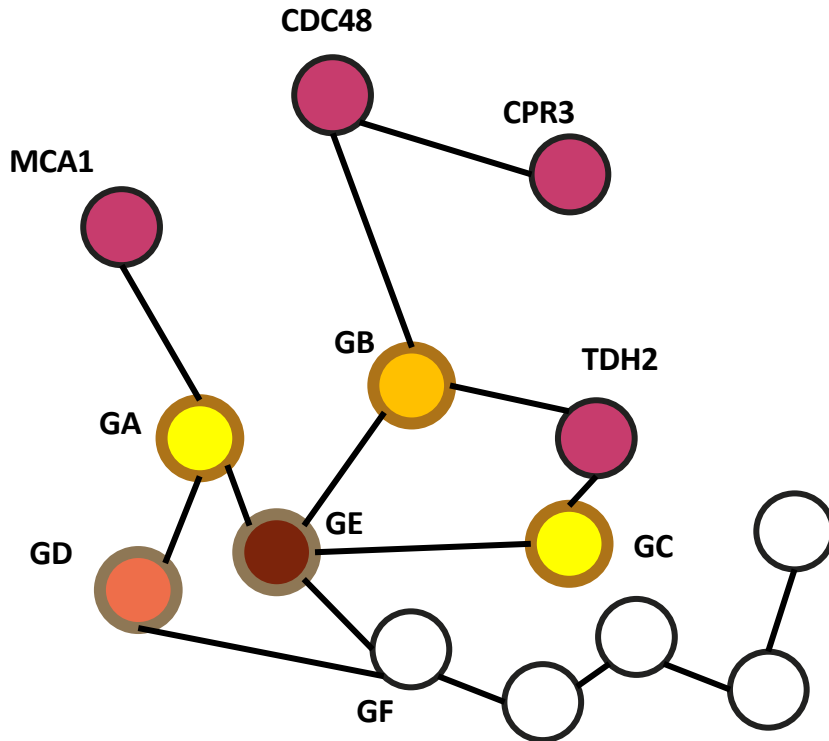
Approach 2: 2-Hop Neighbors

- Consider **2-hop neighbors** when calculating node score f_i as:

$$f_i = \underbrace{\sum_{j=1}^n P_{ij} \mathbf{y}_j}_{\text{Direct neighbors}} + \underbrace{\sum_{j=1}^n [P^2]_{ij} \mathbf{y}_j}_{\text{2-hop neighbors}}$$



Example: 2-Hop Neighbors



$$P = D^{-1}W$$

$$f_i = \underbrace{\sum_{j=1}^n P_{ij} y_j}_{\text{Direct neighbors}} + \underbrace{\sum_{j=1}^n [P^2]_{ij} y_j}_{\text{2-hop neighbors}}$$

$$f_{GA} = P_{GA,MCA1} \cdot y_{MCA1}$$

$$f_{GE} = P_{GE,MCA1}^2 \cdot y_{MCA1} + P_{GE,TDH2}^2 \cdot y_{TDH2} + P_{GE,CDC48}^2 \cdot y_{CDC48}$$

- Direct neighbor of a positive gene
- 2-hop neighbor of a positive gene

Red: Positive genes

White: $f_i = 0$

$[P^2]_{ij} > 0$ if there is a walk of length 2 between i and j

Beyond 2-Hop Neighbors

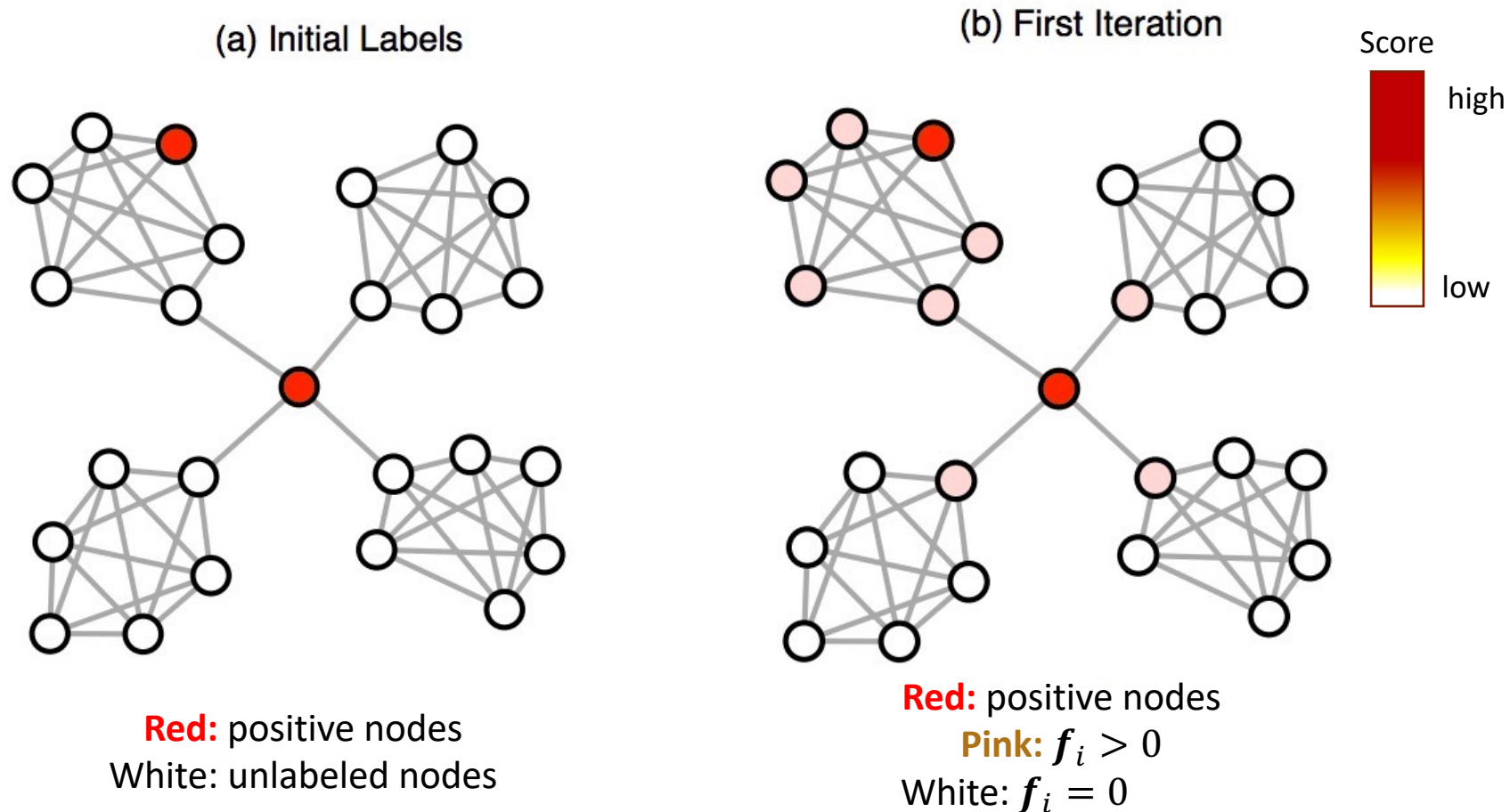
- **This approach** can be **extended** to include nodes at **distance r** (usually $r < 4$):
 - $[P^r]_{ij}$ = Probability of a walk from i to j in **r steps**
- Increasing r beyond 2 sometimes results in degradation of prediction performance
 - [Chua et al., Bioinformatics 2006; Myers et al., Genome Biology 2005, Cowen et al., Nature Reviews 2017]
- **Next:** Use **random walks** propagate **labels** throughout the network

Beyond 2-Hops: Label Propagation

- Label propagation **generalizes** neighborhood-based approaches by **considering random walks of all possible lengths**
- The algorithm can be derived as:
 1. Iterative diffusion process [Zhou et al., NIPS 2004]
 2. Solution to a specific convex optimization task [Zhou et al., NIPS 2004, Zhu et al., ICML 2003]
 3. Maximum a posteriori (MAP) estimation in Gaussian Markov Random Fields [Rue and Held, Chapman & Hall, 2005]
- **Next:** Derivation based on **diffusion**

Label Propagation: Intuition

Intuition: Diffuse labels through edges of the network



Diffusion Process: Idea

- **Diffusion** is defined as an **iterative process** [Zhou et al., NIPS 2004]
- **Diffuse labels through network edges:**
 - Start with initial label information, $f_i^{(0)} = y_i$
 - In each iteration, node i receives **label information from its neighbors** and also **retains some of its initial label**
 - λ specifies **relative amount** of label information from i 's neighbors and its initial label
- Finally: Label for each unlabeled node is set to be the class (-1 or 1) of which it has **received most information**

Diffusion Process: Formally

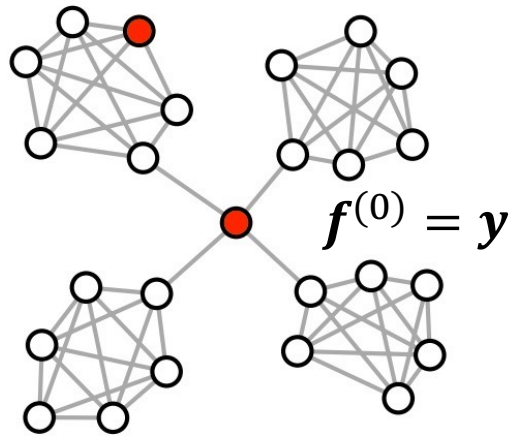
- **Diffusion process** is defined as **iteration**:
 - At iteration $r = 0$, define $\mathbf{f}_i^{(0)} \leftarrow \mathbf{y}_i$
 - At iteration $r + 1$, the score for node i is **weighted average** of the scores for i 's neighbors in iteration r , and i 's initial label:

$$\mathbf{f}_i^{(r+1)} \leftarrow (1 - \lambda)\mathbf{y}_i + \lambda \sum_{j=1}^n \mathbf{w}_{ij} \mathbf{f}_j^{(r)}$$

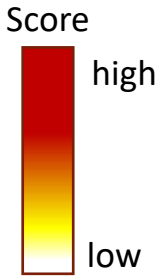
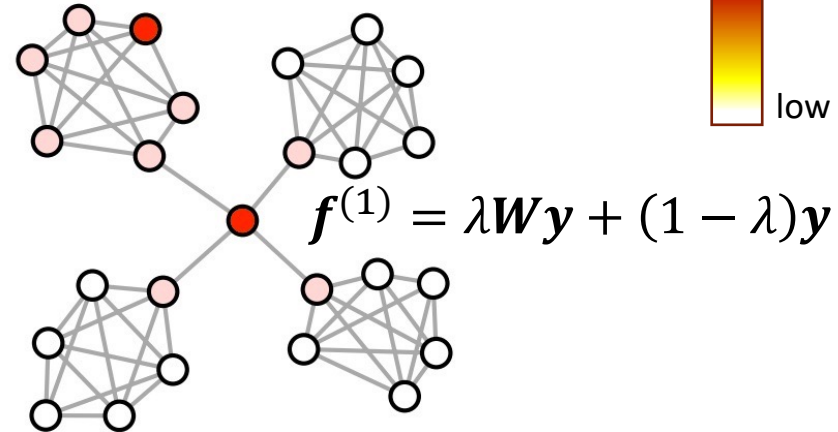
$0 < \lambda < 1$ is model parameter

Diffusion Process: Example

(a) Initial Labels

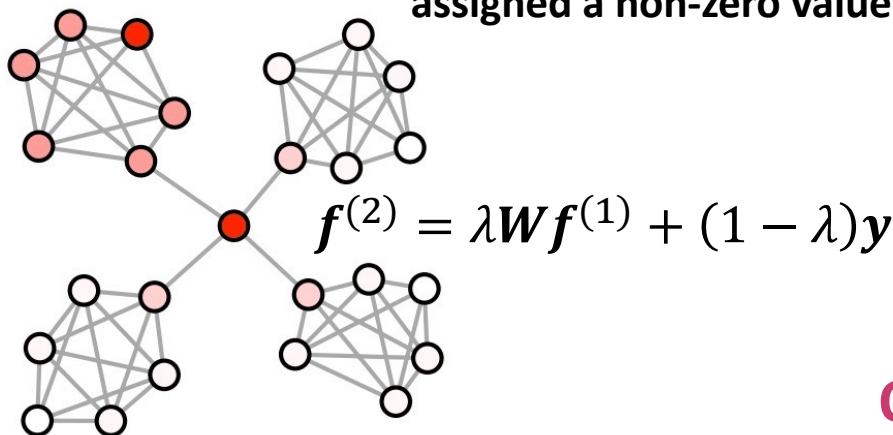


(b) First Iteration



(c) Second Iteration

All nodes within 2 hops are assigned a non-zero value



Red: positive nodes
Pink: $f_i > 0$
White: $f_i = 0$

Question: How many iterations?

Convergence Condition

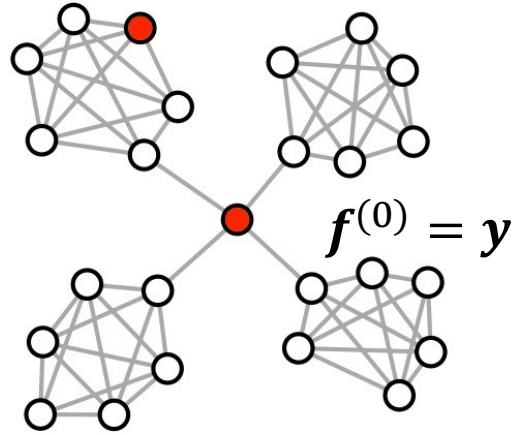
- If all **eigenvalues** of W are in range $[-1, 1]$, then the sequence $f^{(r)}$ converges to:

$$f = (1 - \lambda) \sum_{r=0}^{\infty} (\lambda W)^r y$$

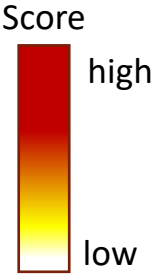
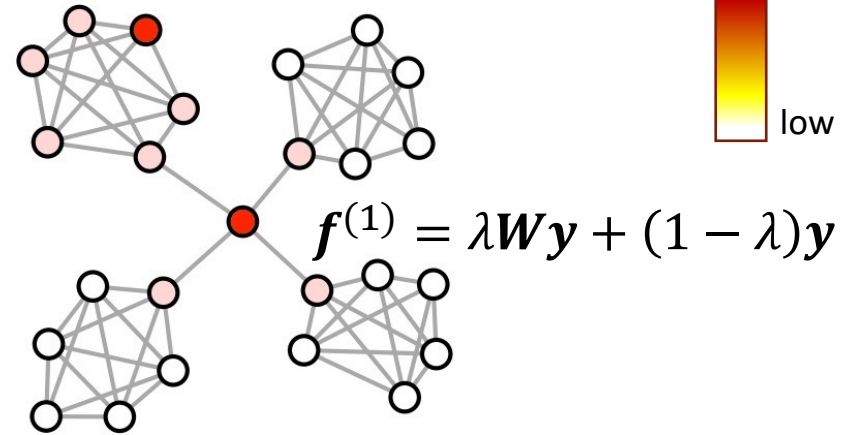
- $[W^r]_{ij} > 0$ if a walk of length r between i and j
- Weight λ^r decreases with increasing distance
- \Rightarrow Discriminant scores f are **weighted sum of walks of all lengths between nodes**
- \Rightarrow **High value** f_i : i is connected to positively labeled nodes with **many short walks**

Diffusion Process: Example

(a) Initial Labels

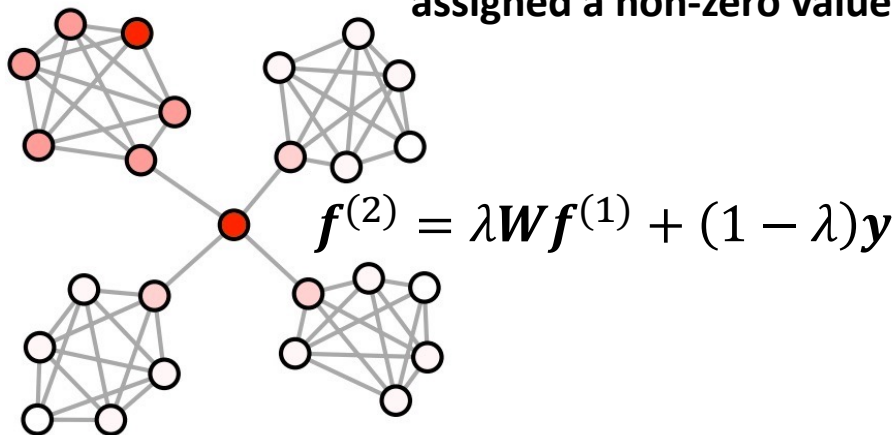


(b) First Iteration

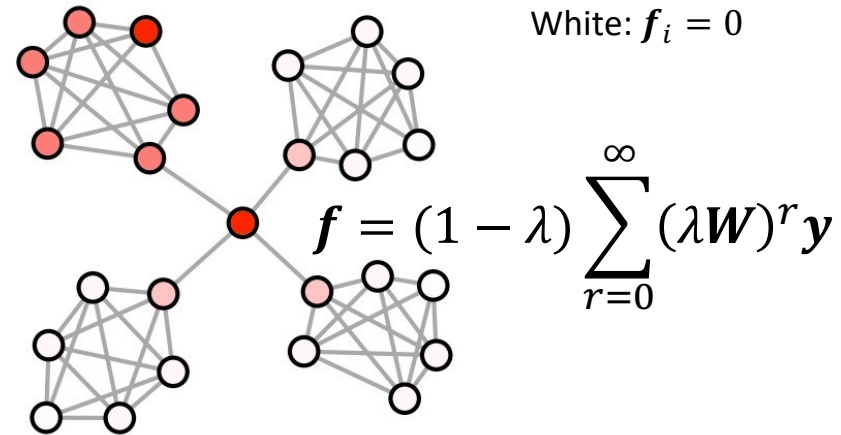


(c) Second Iteration

All nodes within 2 hops are assigned a non-zero value



(d) Final Scores



Red: positive nodes
Pink: $f_i > 0$
White: $f_i = 0$

Does the Process Always Converge?

- **Problem:** The infinite sum converges only if all eigenvalues of W are in $[-1, 1]$, i.e., $\rho(W) \leq 1$
- **Solution:** Normalize W before diffusion:

- **Symmetric** normalization:

$$S = D^{-1/2} W D^{-1/2} \quad D = \text{diag}(d)$$

- Signal is spread in a **breadth-first search** manner
- **Asymmetric** normalization:

$$P = D^{-1} W$$

Exact Solution at Convergence

- If $\rho(W) \leq 1$, use **Taylor expansion** to compute **exact solution for label propagation**:

$$\mathbf{f} = (1 - \lambda) \sum_{r=0}^{\infty} (\lambda \mathbf{S})^r \mathbf{y}$$



$$\mathbf{f} = (1 - \lambda)(\mathbf{I} - \lambda \mathbf{S})^{-1} \mathbf{y}$$

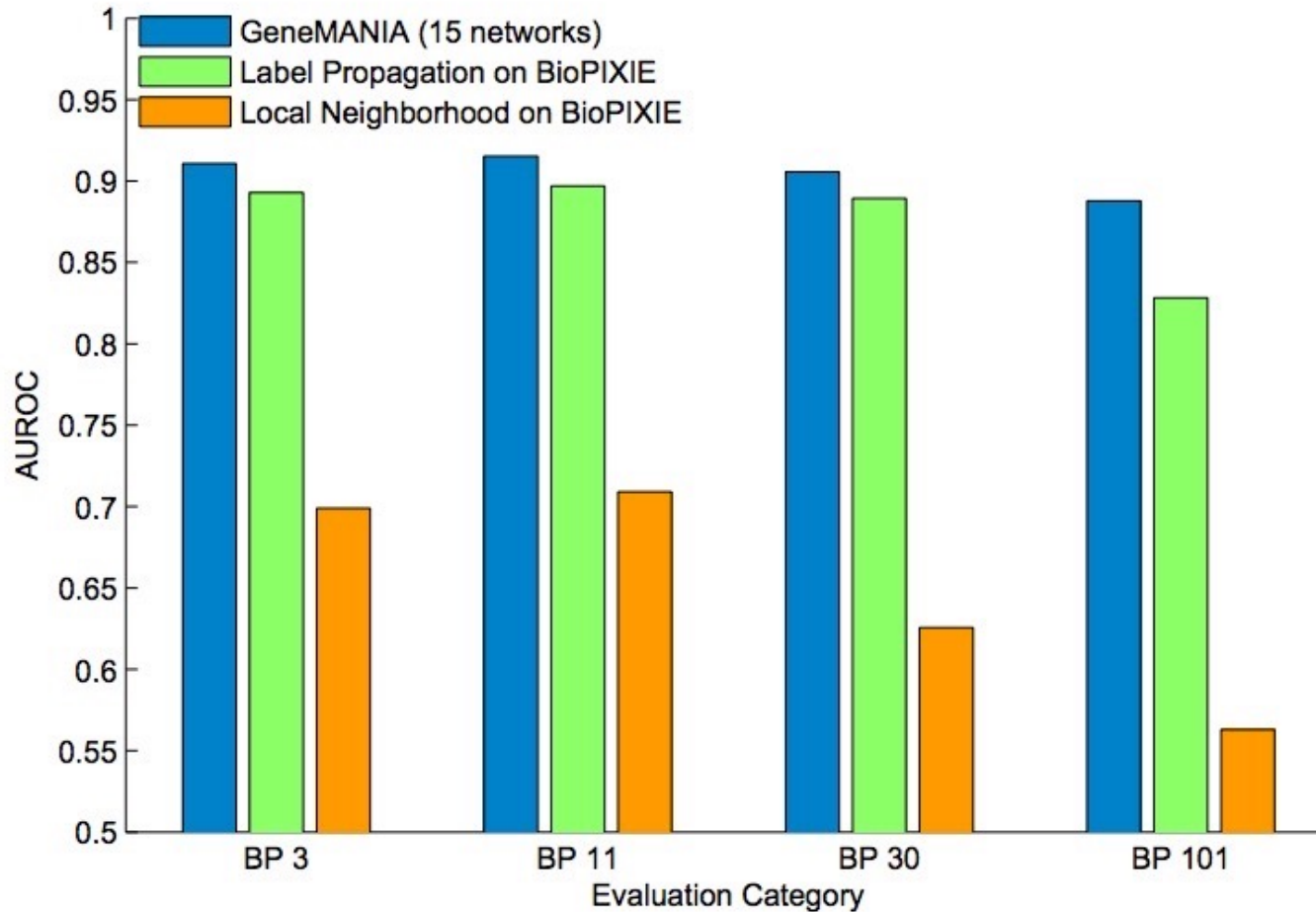
Taylor expansion, sum of geometric series:

$$(\mathbf{I} - \mathbf{A})^{-1} = \sum_{r=0}^{\infty} \mathbf{A}^r$$

Function Prediction: Setup

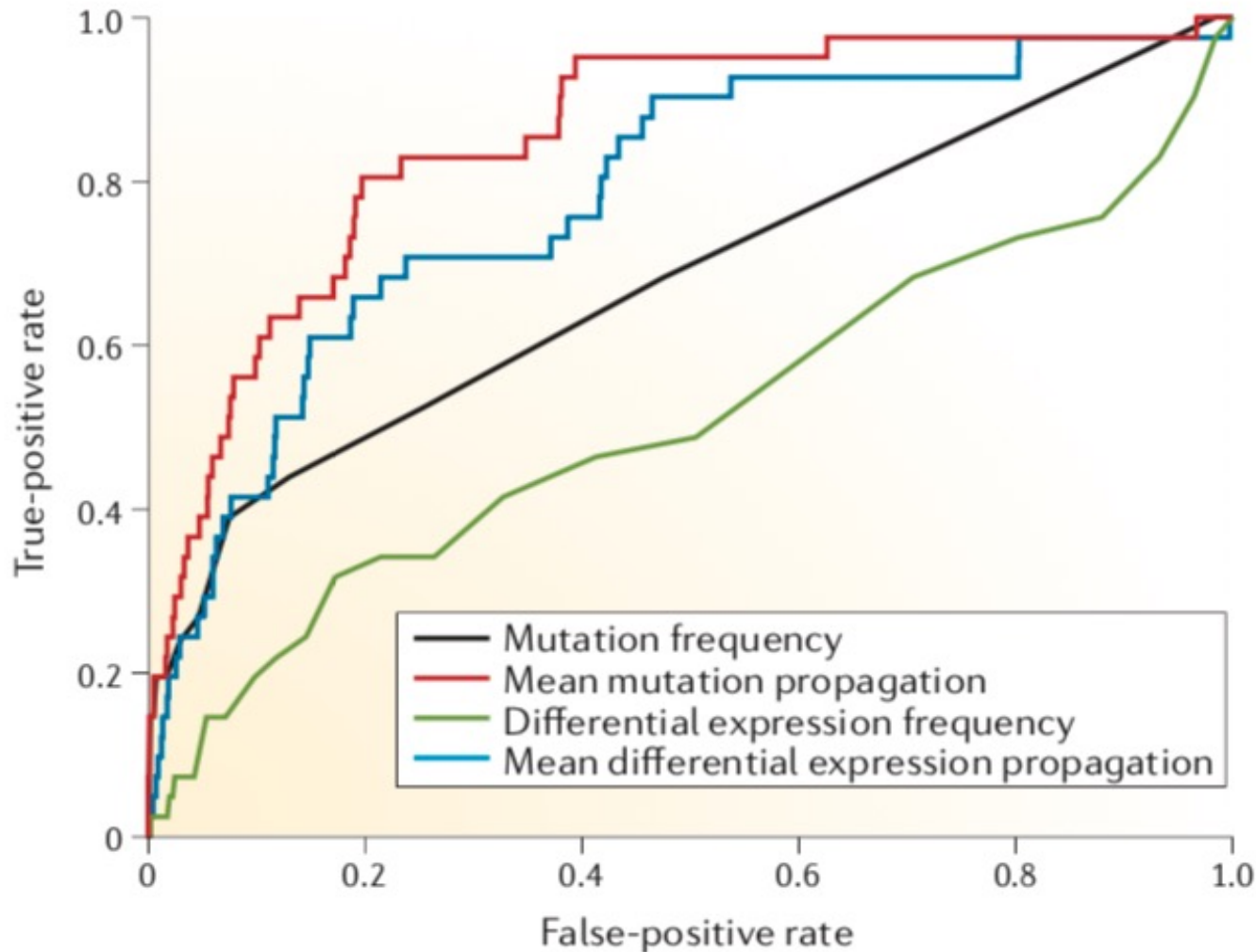
- **Multi-label node classification:** Node (gene) has 0+ labels (functions):
 1. For each label learn a **separate vector f** :
 - **High value of f_i :** i is connected to many labeled nodes through **many short walks** → **i likely has the label**
 2. Train: Observe a fraction of nodes and their labels
 3. Test: Predict functions for the remaining nodes
- Select optimal value for λ using **cross-validation**

Function Prediction: Results



Label propagation outperforms neighborhood scoring methods

Function Prediction: Results



Network propagation variants outperform their frequency-based counterparts (compare the **blue** curve to the **green** curve, and the **red** curve to the **black** curve)

GeneMANIA Tool (genemania.org)

Query list:

The screenshot displays the GeneMANIA web interface. On the left, a table lists various biological functions with their corresponding FDR and Coverage values. The central part of the image shows a complex network diagram with nodes representing genes and edges representing interactions. On the right, a sidebar shows a list of query genes and a network filter panel.

Function	FDR	Coverage
<input type="checkbox"/> DNA recombination	3.29e-36	22 / 151
<input type="checkbox"/> reciprocal DNA recombination	1.32e-22	12 / 35
<input type="checkbox"/> reciprocal meiotic recombination	1.32e-22	12 / 35
<input type="checkbox"/> meiotic nuclear division	3.33e-22	14 / 84
<input type="checkbox"/> meiotic cell cycle	4.53e-22	14 / 87
<input type="checkbox"/> meiosis I	9.47e-21	12 / 50
<input type="checkbox"/> structure-specific DNA binding	4.58e-19	14 / 142
<input type="checkbox"/> cellular process involved in reproduction	9.19e-17	14 / 207
<input type="checkbox"/> double-stranded DNA binding	9.00e-16	11 / 84
<input type="checkbox"/> nuclear division	1.59e-15	14 / 257
<input type="checkbox"/> organelle fission	5.38e-15	14 / 282
<input type="checkbox"/> double-strand break repair	1.86e-14	11 / 112
<input type="checkbox"/> ATPase activity	1.59e-13	12 / 197
<input type="checkbox"/> double-strand break repair via homologous recombination	1.64e-13	9 / 55
<input type="checkbox"/> recombinational repair	1.70e-13	9 / 56
<input type="checkbox"/> DNA-dependent ATPase activity	1.70e-13	9 / 56
<input type="checkbox"/> mismatch repair	2.52e-12	7 / 22
<input type="checkbox"/> single-stranded DNA binding	9.04e-12	8 / 50
<input type="checkbox"/> regulation of DNA recombination	8.76e-11	7 / 35
<input type="checkbox"/> ribonucleoside monophosphate catabolic process	5.88e-10	9 / 139
<input type="checkbox"/> purine nucleoside monophosphate catabolic process	5.88e-10	9 / 139
<input type="checkbox"/> purine ribonucleoside monophosphate catabolic process	5.88e-10	9 / 139
<input type="checkbox"/> ATP catabolic process	5.88e-10	9 / 137

Query list:

- MRE11A
- RAD51
- MLH1
- MSH2
- DMC1
- RAD51AP1
- RAD50
- MSH6
- XRCC3
- PCNA
- XRCC2

Networks Filter:

- Predicted: 50.16%
- Physical interactions: 13.72%
- Shared protein domains: 12.86%
- Co-expression: 10.67%
- Pathway: 8.85%
- Co-localization: 3.74%

Quick Check

<https://forms.gle/mHGeVkk7yZu2oUhr7>

BMI 702: Biomedical Artificial Intelligence

Foundations of Biomedical Informatics II, Spring 2024

Foundations of geometric deep learning, graph representation learning, link prediction, node classification, graph clustering, graph classification, semi-supervised learning, label propagation, network medicine, disease modules.

Course website and slides: <https://zitniklab.hms.harvard.edu/BMI702>

* Indicates required question

First and last name *

Your answer

Harvard email address *

Your answer

Think of another network example in biology or medicine that was not covered in today's lecture. What are nodes? How are edges defined? What predictive or generative tasks can be meaningfully defined on your network? *

Your answer

In class, we introduced the guilty-by-association approach (i.e., direct neighbor scoring, indirect neighbor scoring, label propagation) through gene function prediction. Can you think of a different biomedical problem where the same approach can be helpful? *

Your answer

Submit

Clear form

Patients-like-me

Finding patients with similar genetic and phenotypic features

Diagnostic Odysseys

- Over 7,000 rare diseases, each affects < 200,000 patients in the US
 - Most diseases are phenotypically heterogeneous
 - Front-line clinicians might lack disease experience, resulting in expensive clinical workups for patients across multiple years
 - Diagnosis often requires a specialist, sub-specialist, or multi-disciplinary referrals
- On average, the long search for a **rare disease diagnosis takes 5 to 7 years, 4 up to 8 physicians, and 2 to 3 misdiagnoses**
- Diagnostic delay is so pervasive that it leads to problems for patients:
 - Undergoing **redundant testing and procedures**
 - Substantial delay in obtaining disease-appropriate management and **inappropriate therapies**
 - **Irreversible disease progression**—time window for intervention can be missed leading to disease progression

Can AI help shorten diagnostic odysseys for rare disease patients?

AI-assisted Medical Diagnosis

- Deep learning models trained (via supervised learning) on large labeled datasets can achieve **near-expert clinical accuracy for common diseases**
- Existing models require **labeled datasets with thousands of diagnosed patients per disease**:
 - Diabetic retinopathy: deep neural net on 128 K retinal images
 - Skin lesions: deep neural net on 129 K clinical images of skin cancers
 - Childhood diseases: deep neural net on 1 M pediatric patient visits

The challenge with rare diseases is fundamental — **datasets are three orders of magnitude smaller than in other uses of AI for medical diagnosis**

Needed is an entirely new approach to making AI-based rare disease diagnosis possible. This is for two primary reasons:

- Rare disease diagnosis cannot simply be solved by recruiting/labeling more patients because of high disease heterogeneity and low disease prevalence
- Rare disease diagnosis cannot be solved by supervised deep learning because the models cannot extrapolate to novel genetic diseases and atypical disease presentations

Graph Learning Approach

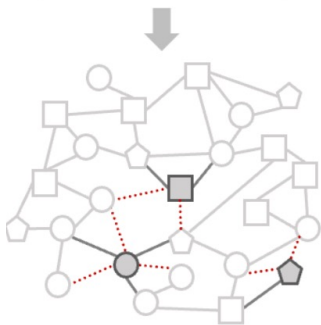
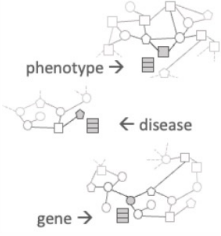
1 Embed Biomedical Knowledge

Sample biomedical knowledge nodes (unrelated to patients)



KG	# Types	Count
Nodes	7	105,220
Edges	15	1,678,274

Embed knowledge graph entities



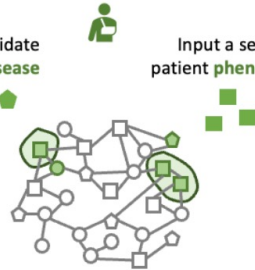
Self-supervised learning via link prediction on the rest of knowledge graph.

2 Embed Rare Disease Patient Information

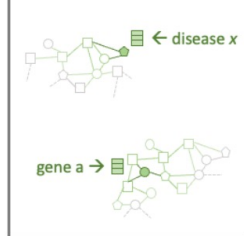
Input candidate gene or disease



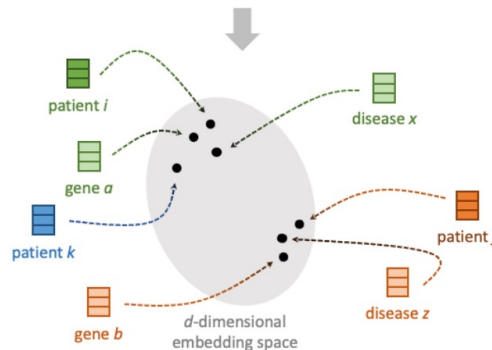
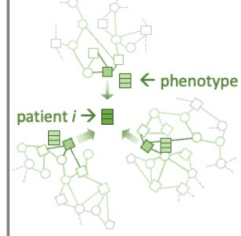
Input a set of patient phenotypes



Embed candidate gene or disease



Embed & aggregate patient phenotypes

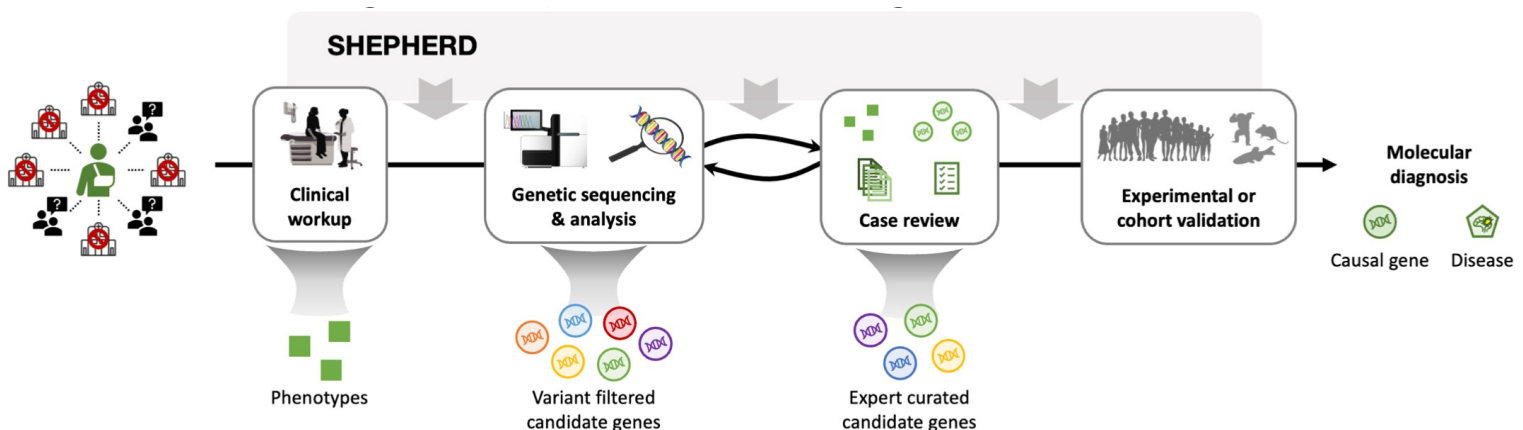


Embed patient closer to the correct gene, disease, or patients with the same gene/disease, and farther from the incorrect gene, disease, or patients with a different gene/disease.

- Step 1: Incorporate knowledge of known phenotype, gene, and disease relationships via GNN
 - Knowledge-guided learning is achieved by self-supervised pre-training on our precision-medicine knowledge graph
- Step 2: Pre-trained GNN from Step 1 is fine-tuned using synthetic patients
 - Training exclusively on synthetic rare disease patients without the use of any real-world labeled cases
 - Synthetic patients used for training are created using an adaptive simulation approach
 - Realistic rare disease patients with varying numbers of phenotypes and candidate genes

Downstream Diagnostic Tasks

- Three diagnostic tasks:
 - **Causal gene discovery:** Given a patient's set of phenotypes and a list of genes in which the patient has mutations, **prioritize genes** harboring mutations that cause the disease (phenotypes)
 - **Patients-like-me:** Given a patient, **find other patients** with similar genetic and phenotypic features suitable for clinical follow-up
 - **Characterization of novel diseases:** Given a patient's phenotypes, **provide an interpretable NLP name** for the patient's disease based on its similarity to each disease in the KG



Experimental Setup

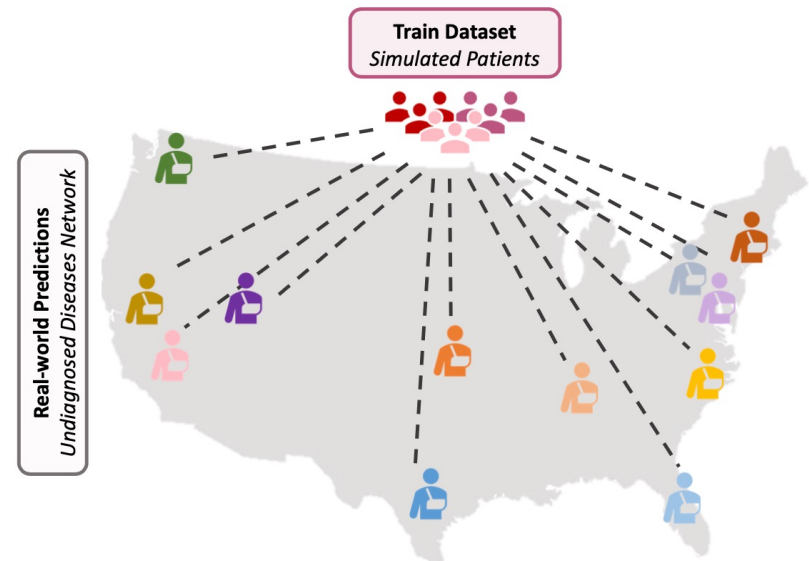
<https://undiagnosed.hms.harvard.edu>

SHEPHERD's model training:

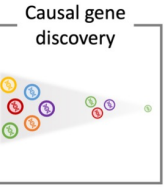
- 36K synthetic patients

SHEPHERD's model evaluation

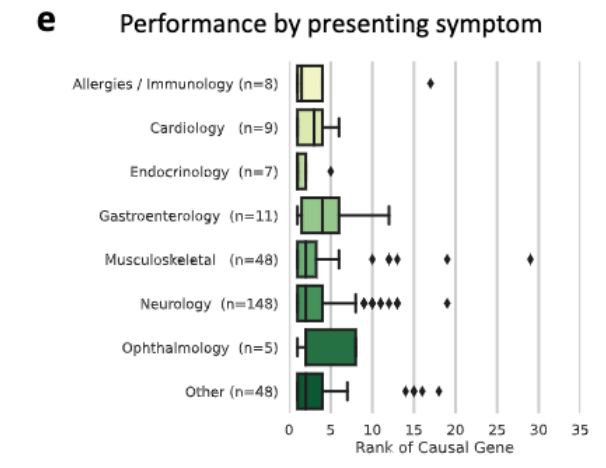
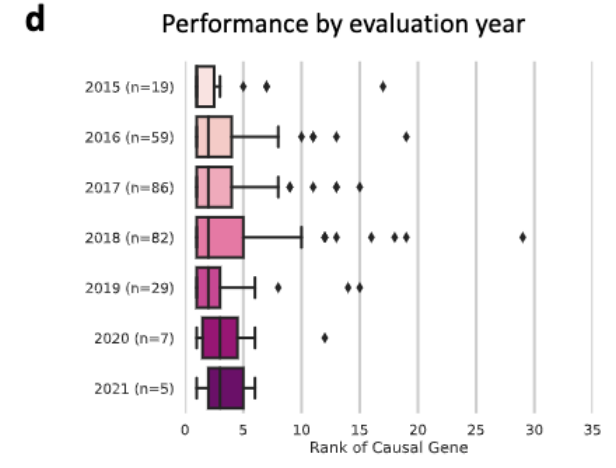
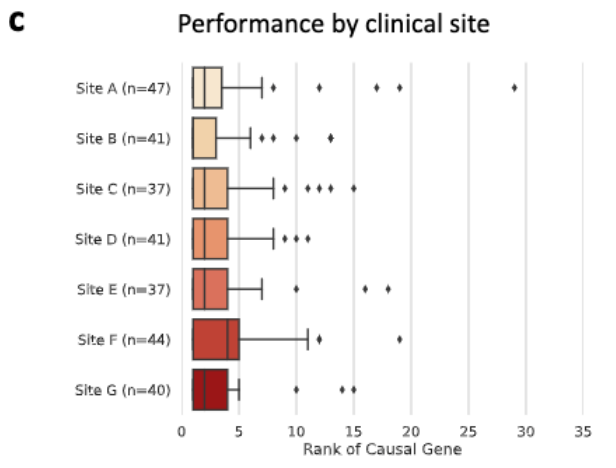
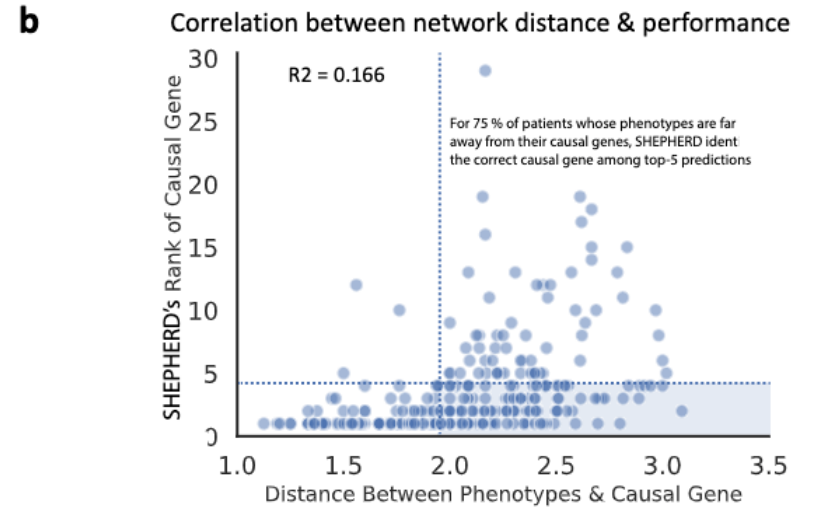
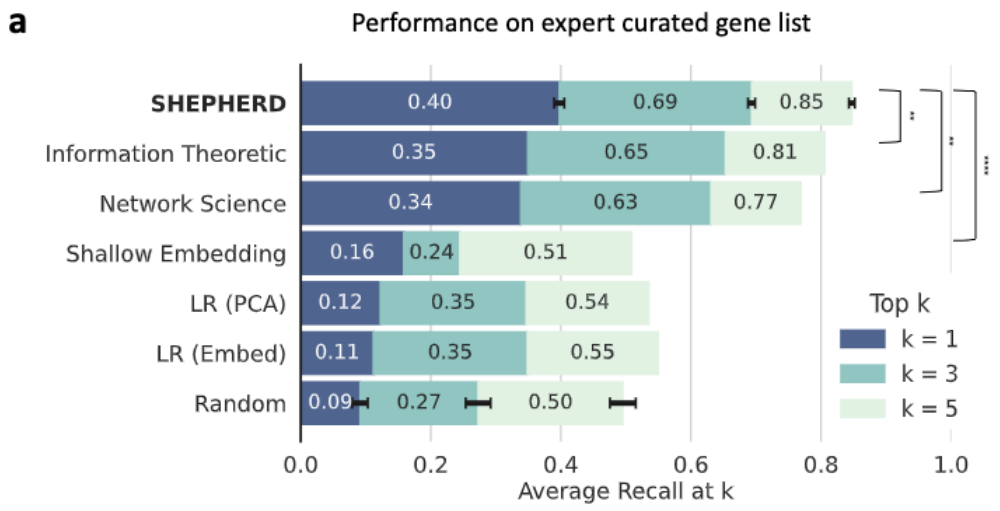
- UDN patient cohort:** 465 rare disease patients with labeled diagnoses, spanning 299 diseases
 - 79% of genes and 83% of diseases are represented in only a single patient
- MyGene2 patient cohort:** 146 rare disease patients, spanning 55 diseases



Patient dataset	Train cohort	Validation cohort	Test cohort
Simulated	N = 36,224	N = 6,400	---
UDN	---	---	N = 465
MyGene2	---	---	N = 146



Results: Causal Gene Ranking





Results: Patients-Like-Me

UMAP plot of SHEPHERD's embedding space of all simulated (circle), UDN (up-facing triangle), and MyGene2 (down-facing triangle) patients colored by their Orphanet disease category

a **Patient: UDN-P3** *Patient Card*
Causal gene: RPS6KA3
Disease: Coffin-Lowry syndrome

Patient Rank	Gene	Disease
1	<i>GRIA3</i>	X-linked intellectual disability due to <i>GRIA3</i> anomalies
2	<i>RPS6KA3</i>	Coffin-Lowry syndrome
3	<i>THOC2</i>	X-linked intellectual disability-short stature-overweight syndrome
4	<i>AP1S2</i>	Fried syndrome
5	<i>SMS</i>	Syndromic X-linked intellectual disability Snyder type

Patient: UDN-P4 *Patient Card*
Causal gene: CAPN1
Disease: autosomal recessive spastic paraplegia type 76

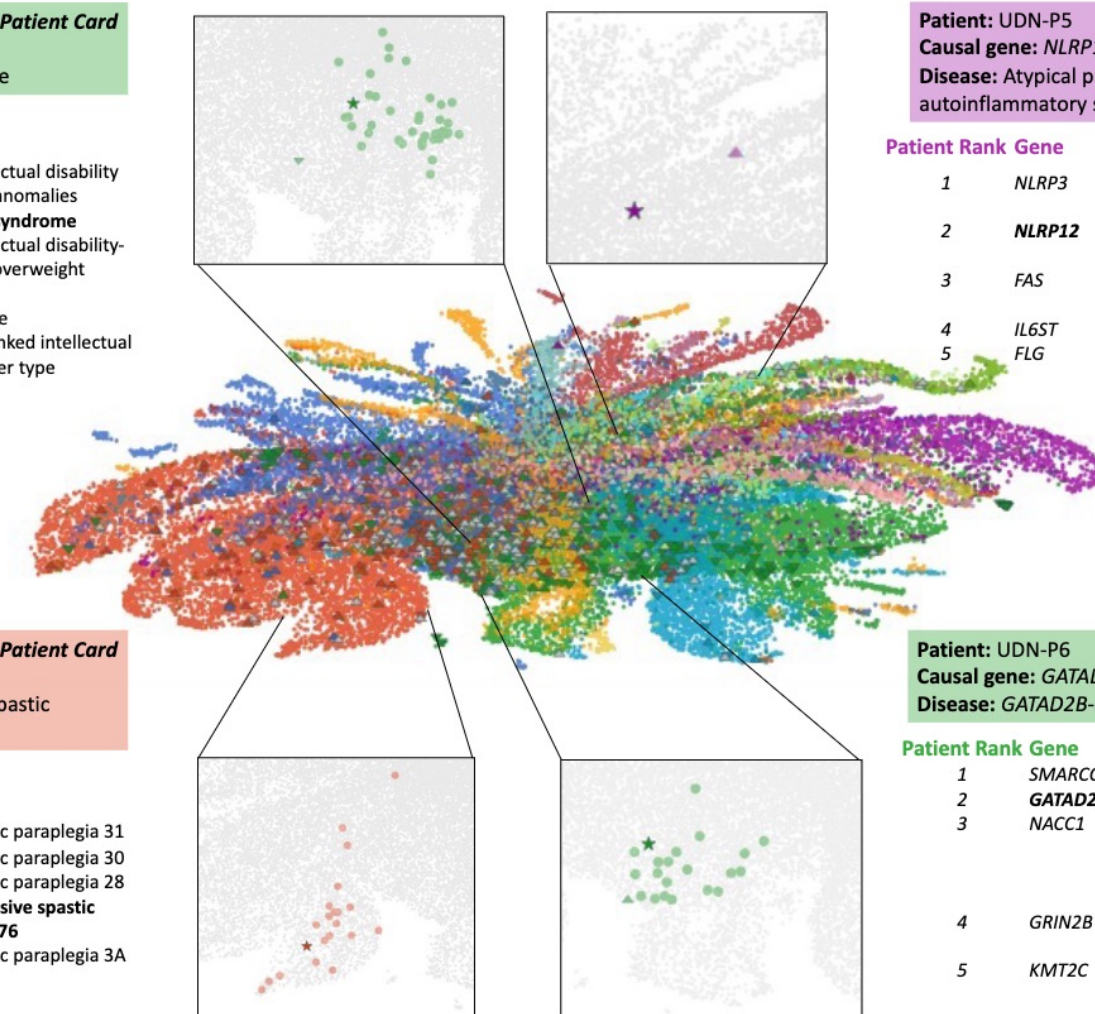
Patient Rank	Gene	Disease
1	<i>REEP1</i>	hereditary spastic paraplegia 31
2	<i>KIF1A</i>	hereditary spastic paraplegia 30
3	<i>DDHD1</i>	hereditary spastic paraplegia 28
4	<i>CAPN1</i>	autosomal recessive spastic paraplegia type 76
5	<i>MTPAP</i>	hereditary spastic paraplegia 3A

Patient: UDN-P5 *Patient Card*
Causal gene: NLRP12, RAPGEFL1
Disease: Atypical presentation of familial cold autoinflammatory syndrome

Patient Rank	Gene	Disease
1	<i>NLRP3</i>	Familial cold-induced autoinflammatory syndrome 1
2	<i>NLRP12</i>	Familial cold-induced autoinflammatory syndrome 2
3	<i>FAS</i>	autoimmune lymphoproliferative syndrome type 1
4	<i>IL6ST</i>	GP130-deficient hyper-IgE syndrome
5	<i>FLG</i>	atopic dermatitis 2

Patient: UDN-P6 *Patient Card*
Causal gene: GATAD2B
Disease: GATAD2B-associated syndrome

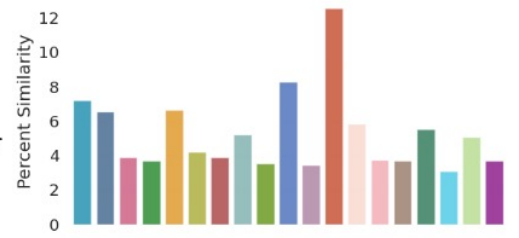
Patient Rank	Gene	Disease
1	<i>SMARCC2</i>	Coffin-Siris syndrome 8
2	<i>GATAD2B</i>	GATAD2B-associated syndrome
3	<i>NACC1</i>	neurodevelopmental disorder with epilepsy, cataracts, feeding difficulties, and delayed brain myelination syndrome
4	<i>GRIN2B</i>	intellectual disability, autosomal dominant 6
5	<i>KMT2C</i>	Kleefstra syndrome



Results: New Disease Naming

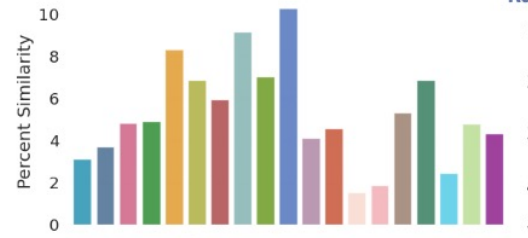
a Rank Disease

- 1 AR limb-girdle muscular dystrophy type 2B
- 2 GNE myopathy
- 3 MYH7-related late-onset scapulothoracic muscular dystrophy
- 4 Emery-Dreifuss muscular dystrophy 2, AD
- 5 AR limb-girdle muscular dystrophy type 2G

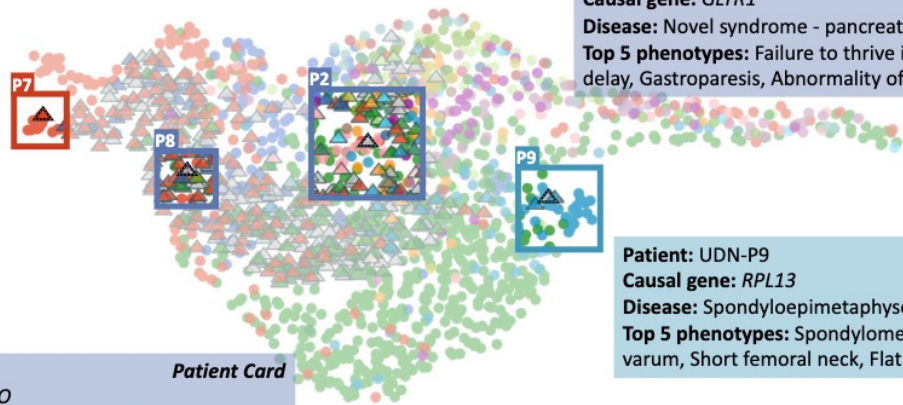


Rank Disease

- 1 Methylmalonic aciduria & homocystinuria type cblF
- 2 Neonatal hemochromatosis
- 3 Homozygous 11P15-p14 deletion syndrome
- 4 ALG8-CDG
- 5 Congenital anemia



Patient: UDN-P7
Causal gene: *SGCA*
Disease: AR limb-girdle muscular atrophy type 2D
Top 5 phenotypes: Toe walking, Calf muscle pseudohypertrophy, Elevated serum creatine kinase, Proximal muscle weakness, Generalized muscle weakness

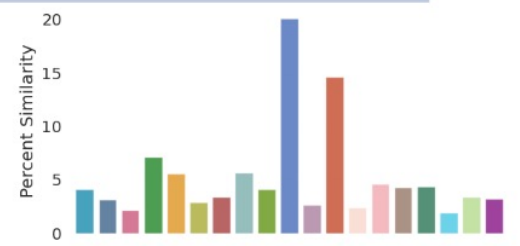


Patient: UDN-P2
Causal gene: *GLYR1*
Disease: Novel syndrome - pancreatic insufficiency & malabsorption
Top 5 phenotypes: Failure to thrive in infancy, Global developmental delay, Gastroparesis, Abnormality of vision, Duodenal atresia

Patient: UDN-P9
Causal gene: *RPL13*
Disease: Spondyloepimetaphyseal dysplasia, Isidor-Toutain type
Top 5 phenotypes: Spondylometaphyseal dysplasia, Genu varum, Short femoral neck, Flat glenoid fossa, Platyspondyly

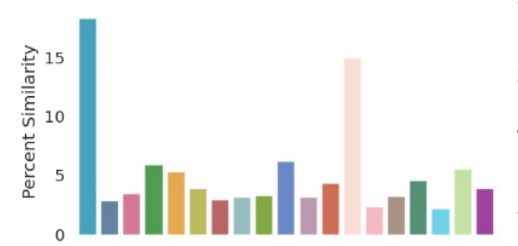
Rank Disease

- 1 Combined oxidative phosphorylation deficiency 39
- 2 Hypomyelinating leukodystrophy-20
- 3 Pyruvate dehydrogenase E3-binding protein deficiency
- 4 Intellectual disability-epilepsy-extrapyramidal syndrome
- 5 Combined oxidative phosphorylation defect type 27



Rank Disease

- 1 Multiple epiphyseal dysplasia type 1
- 2 Progressive pseudorheumatoid arthropathy of childhood
- 3 Multiple epiphyseal dysplasia type 5
- 4 Metaphyseal chondrodysplasia, Spahr type
- 5 Multiple epiphyseal dysplasia



Take-away Messages

- SHEPHERD overcomes limitations of standard machine learning:
 - Model inputs as **KG subgraphs** (i.e., clinic-genetic subgraphs of patients)
 - Use **self-supervised pre-training on biomedical knowledge**
 - Train the model on a large cohort of **synthetic patients**
- SHEPHERD generalizes to novel phenotypes, genes, and diseases:
 - Performs well on patients whose **subgraphs are of varying size**
 - Performs well on **diagnosing patients with novel diseases**
- Implications:
 - Implications for **generalist models applicable across diagnostic process**
 - New opportunities to **shorten the diagnostic odyssey for rare disease**
 - Implications for using **deep learning on medical datasets with very few labels**

**First deep learning approach for individualized diagnosis
of rare genetic diseases**

Graph learning approach is not only helpful but necessary