BMI 702: Biomedical Artificial Intelligence

Foundations of Biomedical Informatics II, Spring 2023

Lecture 13: Label-efficient learning, few-shot learning, biomarker discovery, indication and contra-indication inference, drug repurposing, adverse event prediction

HARVARD
MEDICAL SCHOOL

Marinka Zitnik
marinka@hms.harvard.edu

L14: Data privacy, regulation and liability aspects of biomedical AI

Prof. Dr. Sara Gerke
Professor of Law, Penn State Dickinson Law
Research Fellow, Petrie-Flom Center for Health Law Policy, Biotechnology, and Bioethics Harvard Law School

- **Leading expert in ethical and legal challenges of AI and big data for health care and health law** in the US and Europe
- Comparative law and ethics of other issues at the cutting edge of medical developments, such as the clinical translation of stem cell research, biological products, such as somatic cells, tissues, and gene therapy, reproductive medicine, such as mitochondrial replacement techniques

Next week: May 8th, 2023

Responses to L12 Quick Check

Describe two challenges that models for generating molecular graphs need to address

Since not every molecular graph is chemically valid, we would have to focus on the functional group level. One approach to solving this is Using tree decomposition, which is the summarised molecular graphs in terms of functional groups at a junction tree level.
Also, since Intermediate steps can be invalid or very long (i.e., many node-by-node steps), it is not only challenging in training but also for validation.

One challenge is that the molecular graphs need to reflect the biological and chemical meaning of the compounds accurately. Invalid reflection can lead to errors in real-life application. Another challenge is to control the complexity of the molecular graphs. Too simple may not contain enough molecular information and too complex can cause the graphs too burdensome or repetitive.

1. You need to consider bond properties (edges). Different bonds and different bond structures interact in different ways depending on their location, structure, and neighboring groups.
2. You need to consider interactions between functional groups farther away than just a neighbor. It helps to consider interactions between the whole molecule, so you need to consider all atom interactions together.

Responses to L12 Quick Check

Describe two challenges that models for generating molecular graphs need to address

1. The set of all possible molecules is extremely vast and complex. As a result, it is very difficult to develop a model that can generate a wide range of molecules with desired properties.
2. The molecular properties of interest, such as solubility, bioactivity, and toxicity, are often difficult to quantify and represent mathematically.

1. It is hard to keep diversity and validation at the same time. If we want to purse diverse and complex chemical structure, it is highly probably that the structure is not chemically valid.
2. When we model for generating molecular graphs, we need to consider the spatial arrangement of atoms in a molecule. However, it will highly increase the dimensionality which makes modeling more challenging.

Responses to L12 Quick Check

What is the difference between traditional vs. neural fingerprint representations?

Neural graph fingerprints are generated with a neural network, which update atom features using only adjacent atoms, and use different weights for node degrees. While traditional finger prints usually compared to each other using the Tanimoto metric. Fingerprint features can each only be activated by a single fragment of a single radius. In contrast, neural fingerprint features can be activated by variations of the same structure, making them more interpretable, and allowing shorter feature vectors (ref: BMI702 slides).

The traditional fingerprint coded everything in binary setting as 1 and 0, which can varyy to hundreds of types, while neural fingerprint represents molecule as node and bond as edge to better understand their chemical properties, which is more effective at structuring drug.

Traditional fingerprint representations follow a rigid set of arbitrary rules, neural fingerprint representation allow neural networks to learn the representations automatically taking into account spatial, sequential, and chemical properties.

Traditional fingerprint representation is an encoding of individual atoms, where each element of a vector is a representation of a binary output of 1 if the answer is yes to the question. (e.g., is the molecule part of a benzene ring?) and 0 if the answer is no to the question. Therefore, the traditional fingerprints are manually inputted, and has no learning involved.

Outline for today's class

- High-throughput genetic and chemical perturbations
- Therapeutic use prediction, indication and contra-indication inference
- Drug repurposing

Words and genes share a correspondence:
their **meanings** arise from their **context**.

Gene perturbation measurements across diverse cell contexts
induce **semantics for genes**

(under the right approach)

“apple” is a **polysemic** word…

grow an apple

buy an apple

… whose **particular meaning** is resolved via **sentence context**.

Google

grow an apple

grow an apple **tree**

grow an apple **tree from seed**

grow an apple **tree in a pot**

grow an apple **tree indoors**

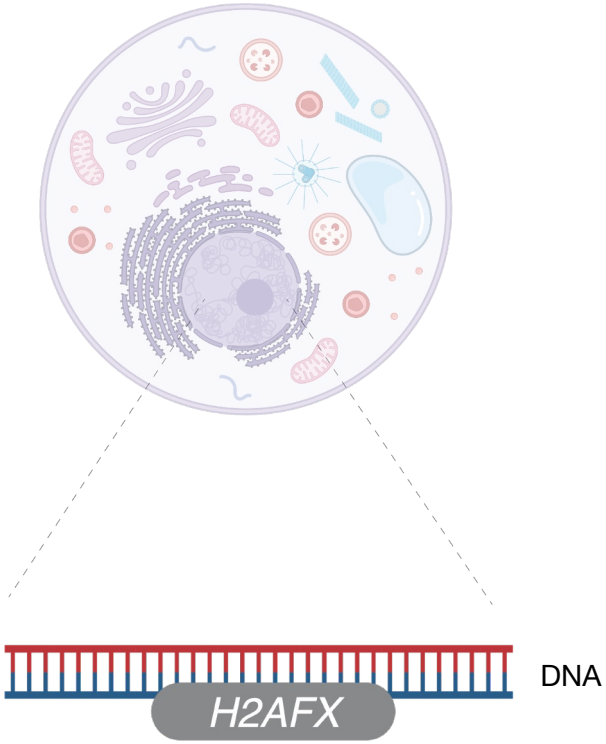
buy an apple

buy an apple **watch**

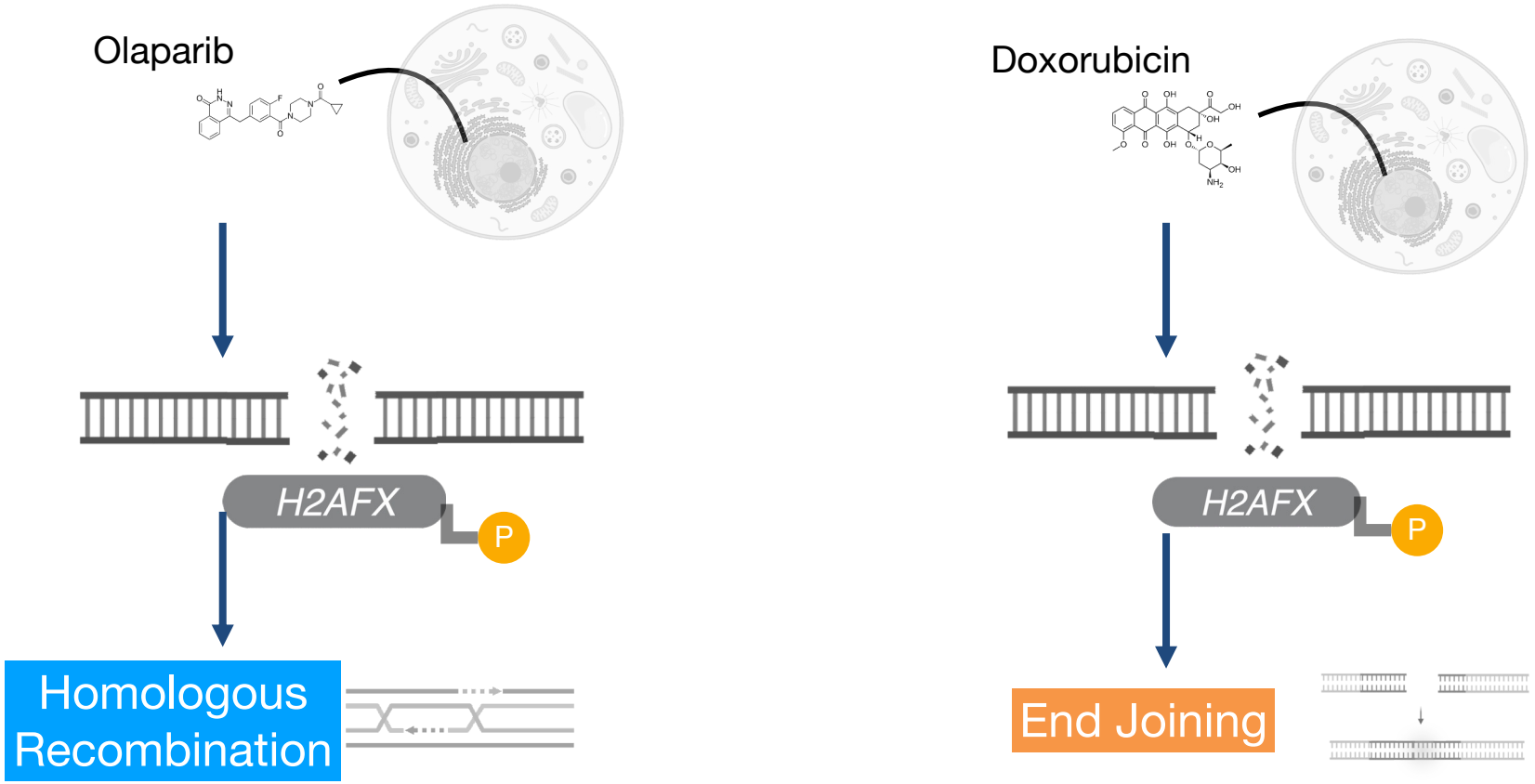
buy an apple **gift card**

buy an apple **tv**

H2AFX is a **pleiotropic** gene…



… whose **particular function** is resolved via **cell context.**



While unsupervised learning of word polysemy is **common**…

Data: corpus of sentence contexts

Approach: word embeddings w/ linear semantics

king - man + woman ≈ queen

unsupervised learning of gene pleiotropy is **unsolved**

Data: **?**

Approach: **?**

geneA - func1 + func2 ≈ geneB

Our goal for today

Unsupervised learning of gene pleiotropy with applications to therapeutic science

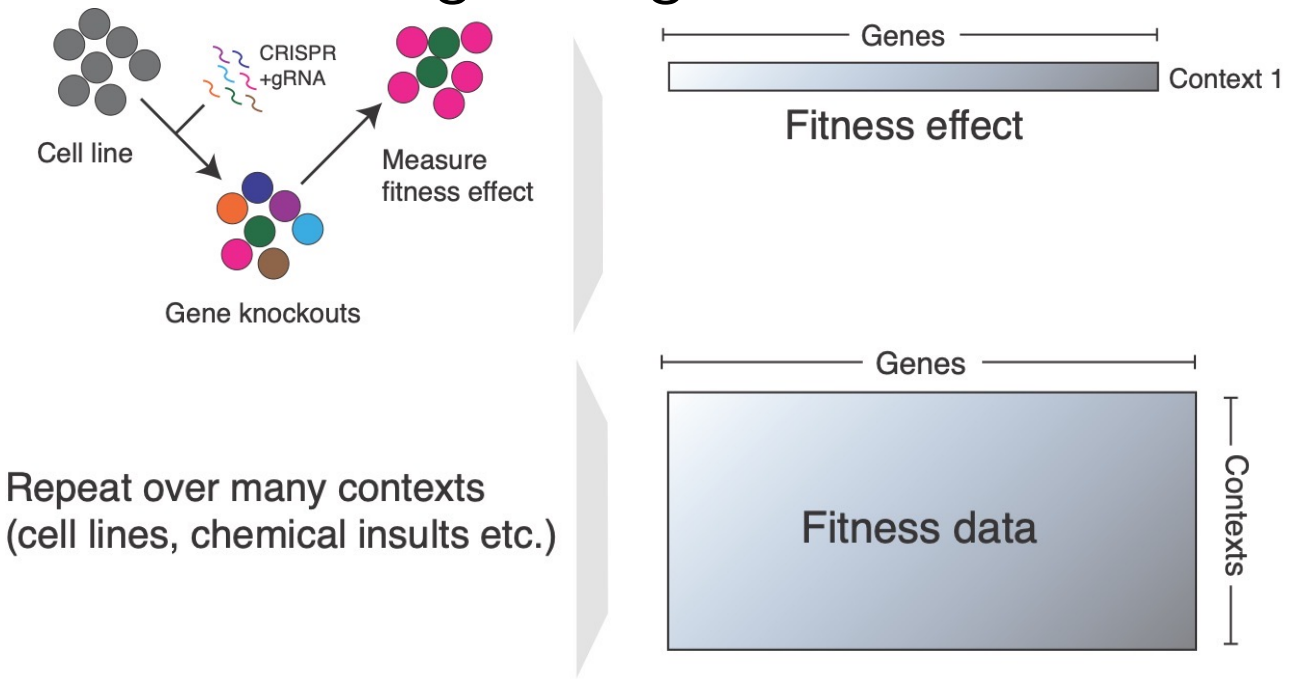
Data: ?

Approach: ?

geneA - func1 + func2 ≈ geneB

Data

Use **gene perturbation effect measurements** for inferring biological functions

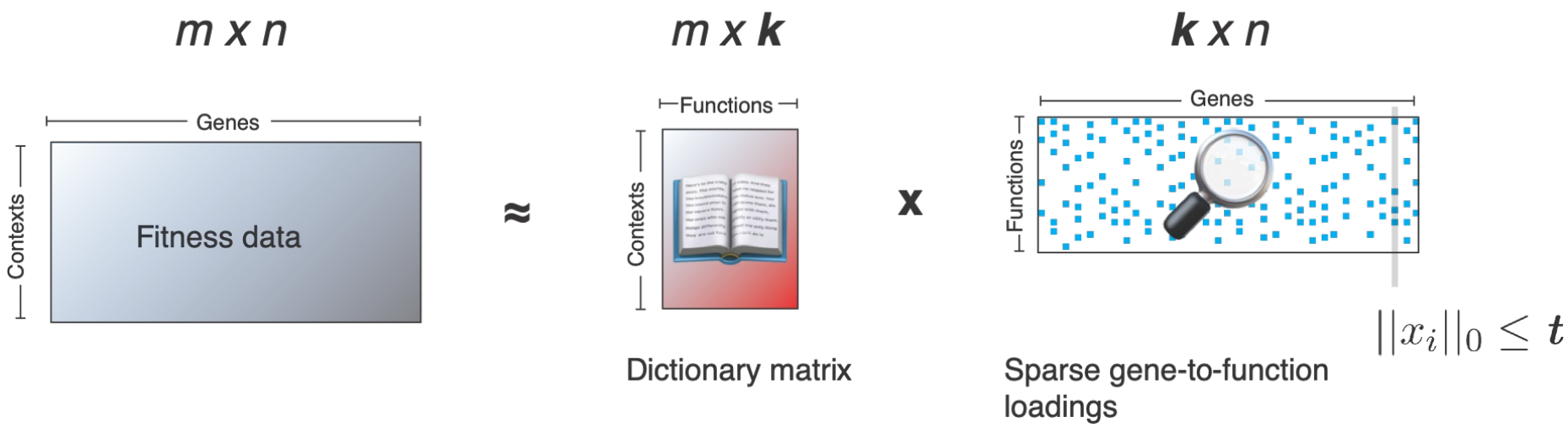


Why perturbation datasets? Alternative data types:

- **Transcriptomics:** gene co-expression is necessary but not sufficient for co-function
- **Protein-protein interactions:** direct interactions are not necessary for co-function

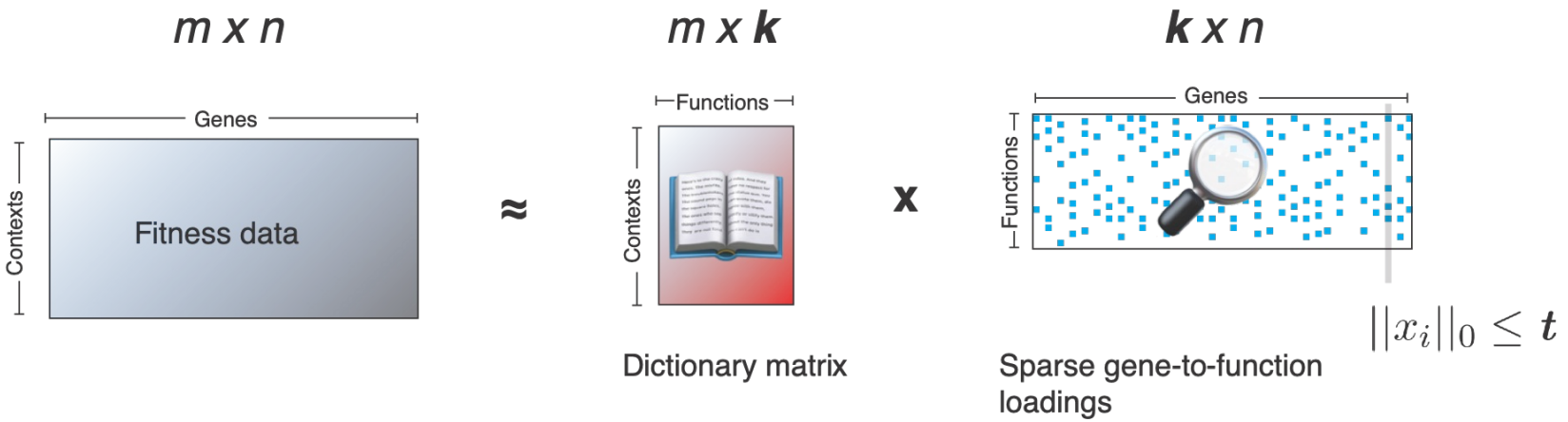
Approach: Webster

- **Low-dimensional vector embeddings** that satisfy three criteria:
  - Sparse
  - Latents are biologically meaningful
  - Account for redundancy between cell contexts

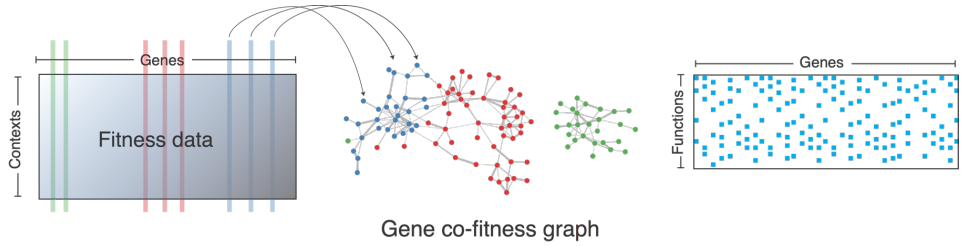


Approach: Webster

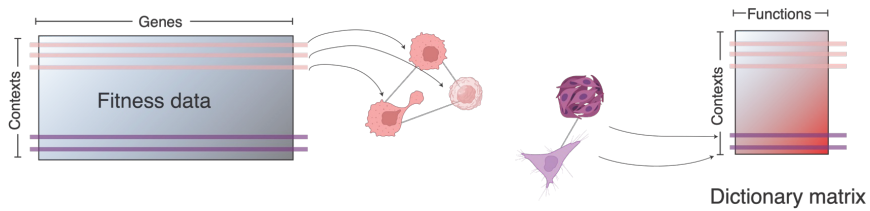
Webster learns a dictionary matrix that **sparsely** approximates gene effects…



1 … while preserving interpretable relationships between genes



2 … and accounting for redundancies between cell contexts

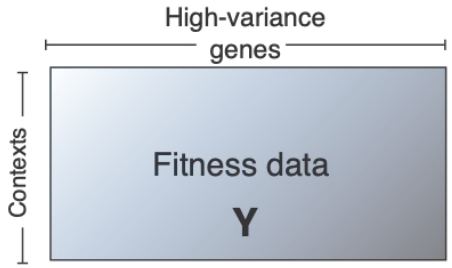


Overview of Webster

Preprocessing

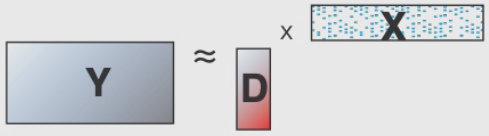
Raw fitness data

Standardize cell lines
Center gene effects
Filter genes by variance



Graph-regularized dictionary learning
Objectives

Reduce dimensionality

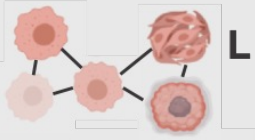


$$\|Y - DX\|_F^2$$

Preserve gene similarity

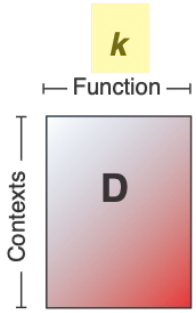
$$Tr(XL_cX^T)$$

Preserve cell context similarity

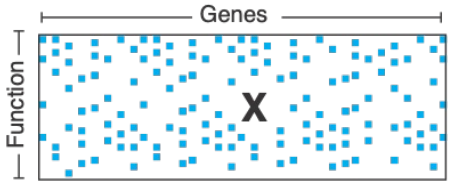


$$Tr(D^T LD)$$

Output



Dictionary matrix

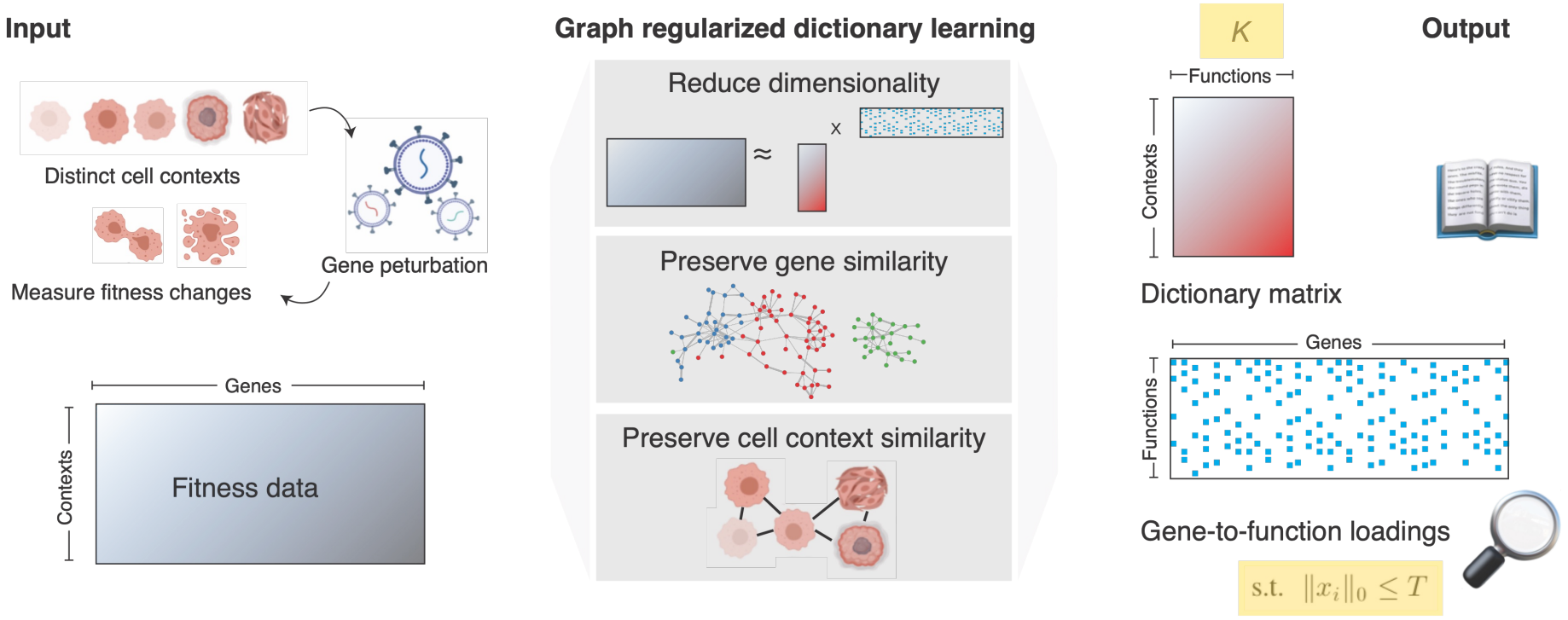


Gene-to-function loadings

$$||x_i||_0 \leq t$$

= key hyperparameters

Its key parameters are dictionary size (K) and sparsity on loadings (T)

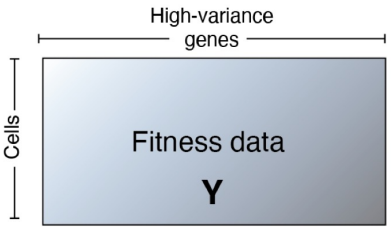


Model optimization

Preprocessing

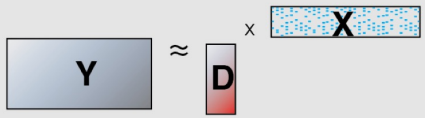
Raw fitness data

Standardize cell lines
Filter genes by variance



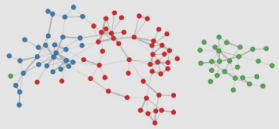
Graph-regularized dictionary learning
Objectives

Reduce dimensionality



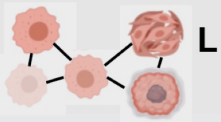
$$\|Y - DX\|_F^2$$

Preserve gene dependency manifold



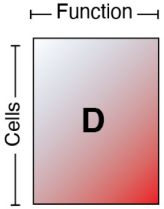
$$Tr(XL_cX^T)$$

Preserve cell state manifold

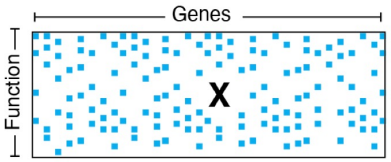


$$Tr(D^T LD)$$

Output



Dictionary matrix



Gene-to-funciton loadings

$$\text{s.t.} \quad \|x_i\|_0 \leq T$$

Parameters

$$\arg\min_{D,X} \|Y - DX\|_F^2 + \alpha Tr(D^T LD) + \beta Tr(XL_cX^T) \quad \text{s.t.} \quad \|x_i\|_0 \leq T \quad \forall i.$$

k	=latent dimension size
L	=cell Laplacian (num neighbors, metric)
Lc	=gene Laplacian (num neighbors, metric)
α	=weight of cell Laplacian
β	=weight of gene Laplacian
T	=sparsity

Applications to three screens of gene perturbation effects

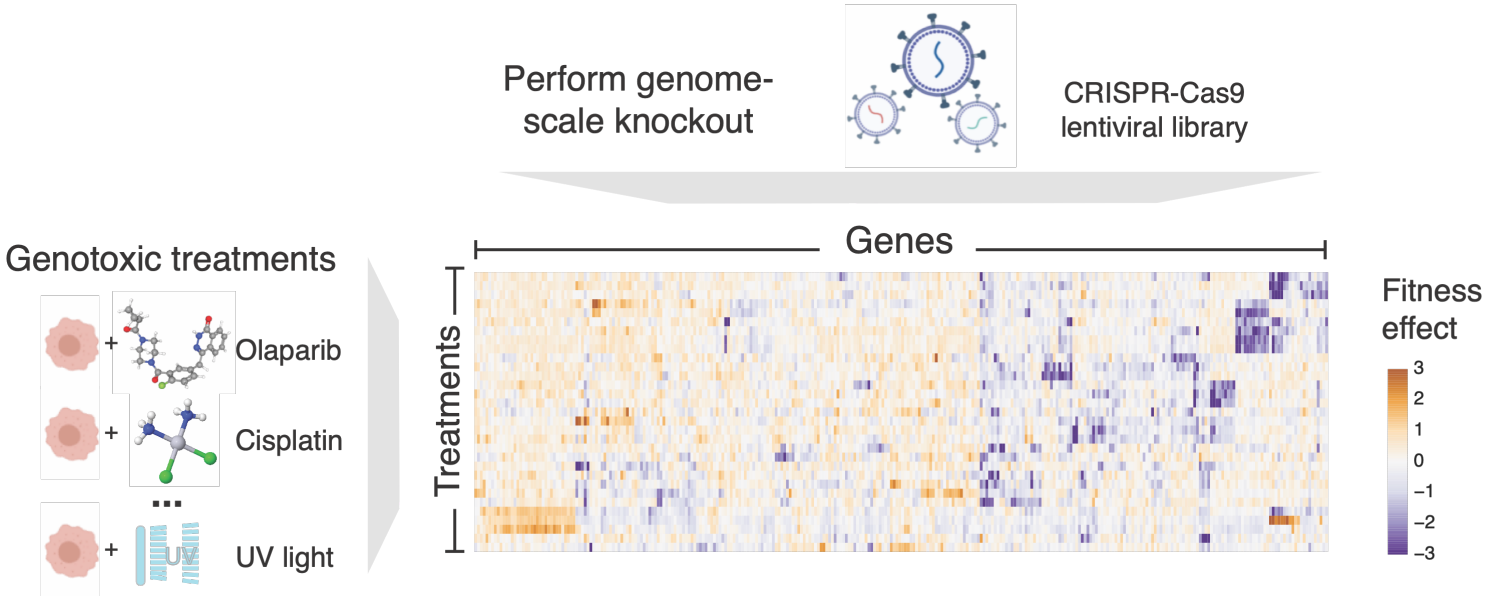
1) Genotoxic screens

2) Cancer fitness screens

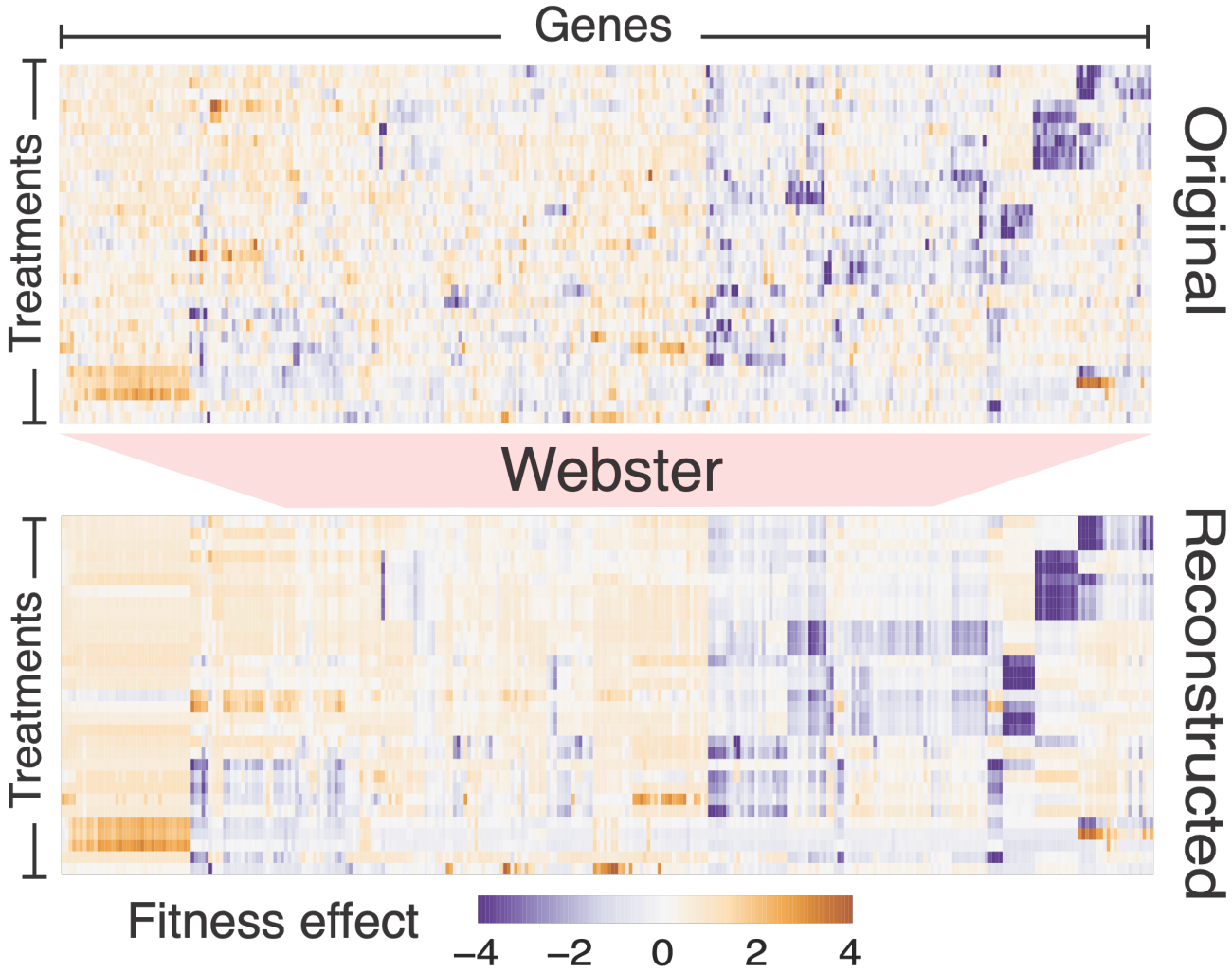
3) Compound sensitivity screens

Part 1: Genotoxic screens

Olivieri et al. 2020: fitness effect of gene knockout in presence of genotoxins

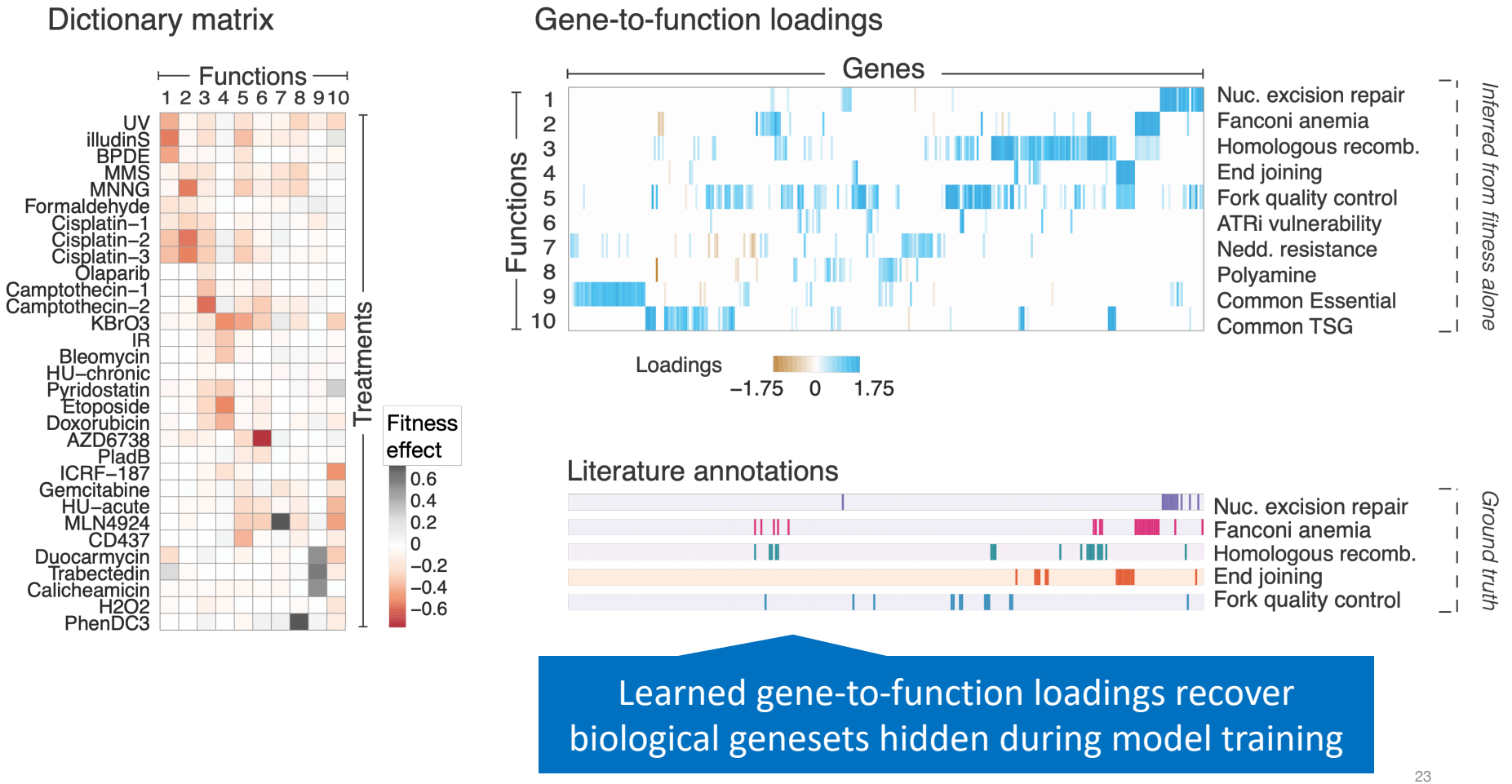


Webster approximates the input data matrix…



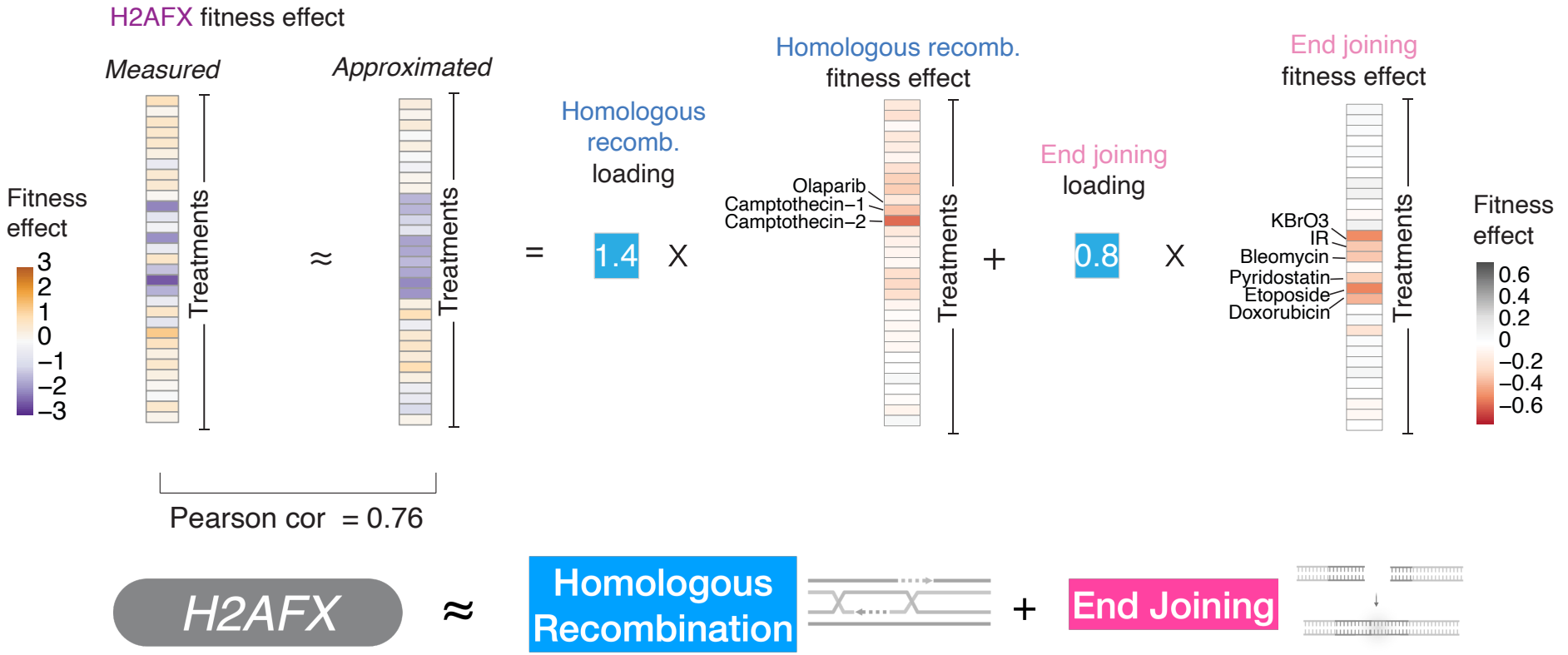
k=10
t=2

… as a product between a dictionary matrix and a loadings matrix



Learned gene-to-function loadings recover biological genesets hidden during model training

Latents inferred by the model recapitulate pleiotropy *without prior knowledge*

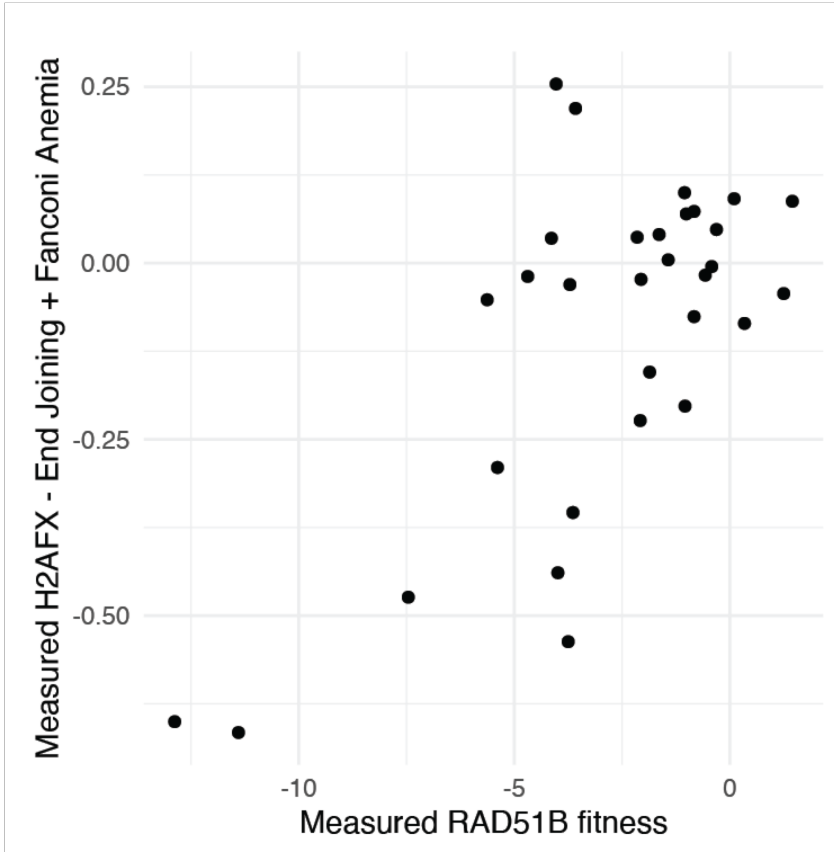


(hidden during model training!)

Latents are biologically meaningful

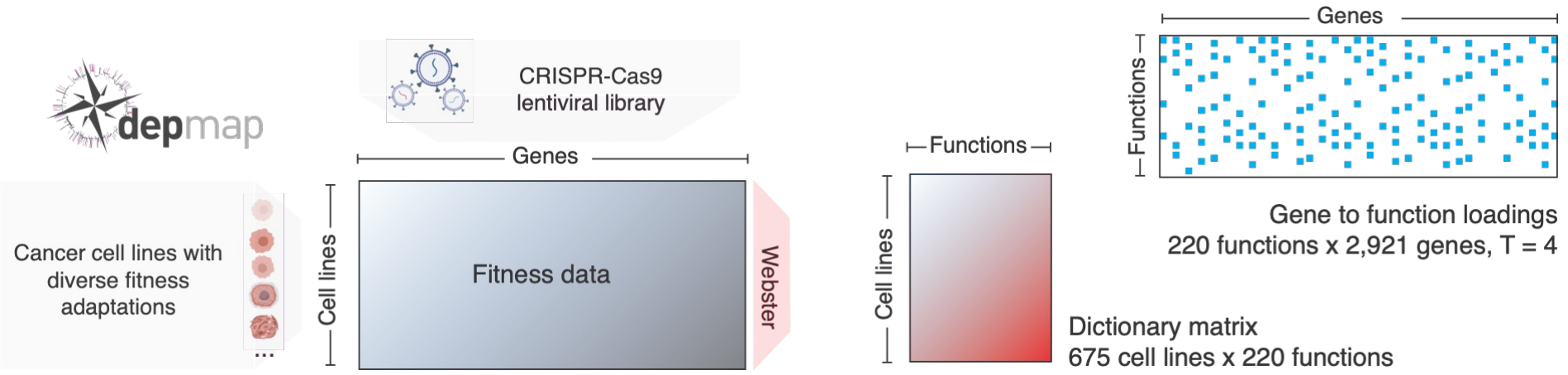
geneA - func1 + func2 ≈ geneB

H2AFX - End Joining + Fanconi Anemia ≈ RAD51B



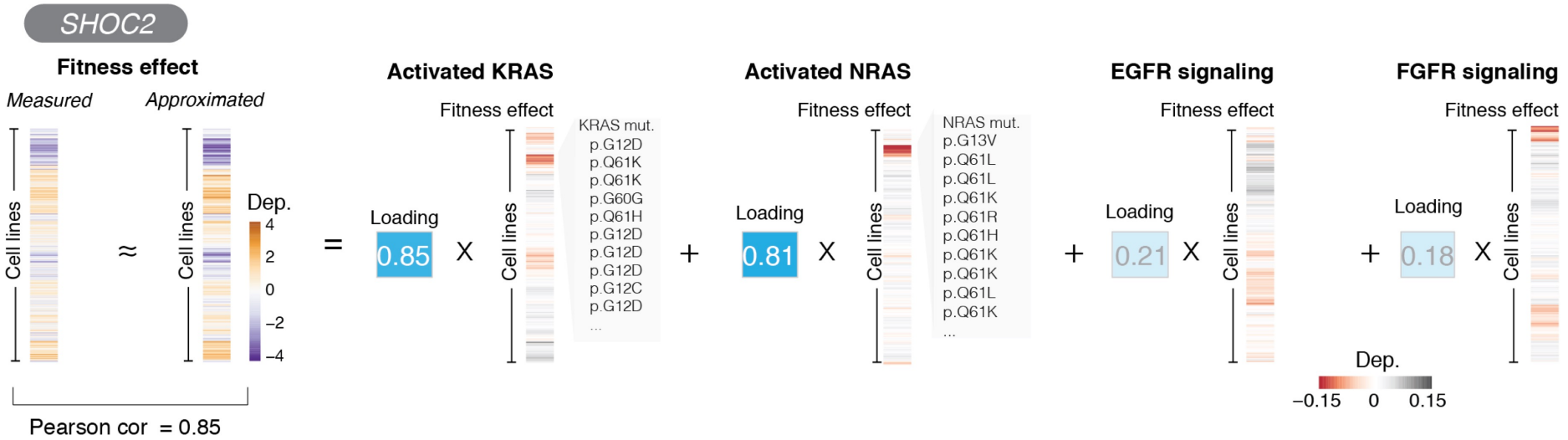
= cell context (treatment)

Part 2: Cancer fitness screens

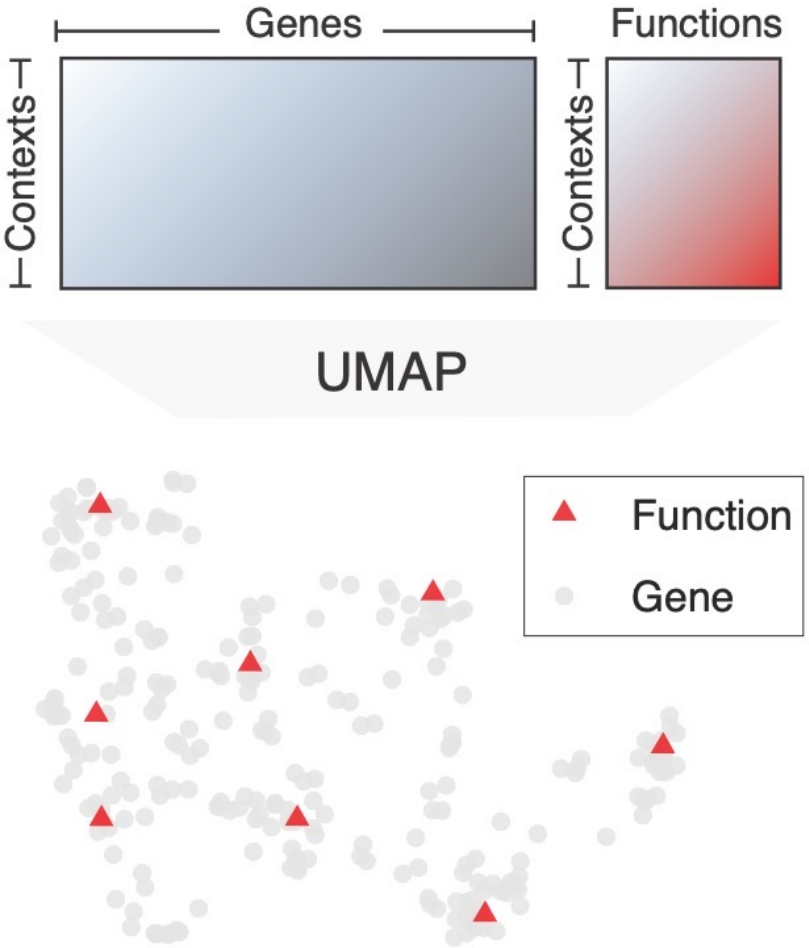


Pleiotropic genes obey linear semantics in the latent space

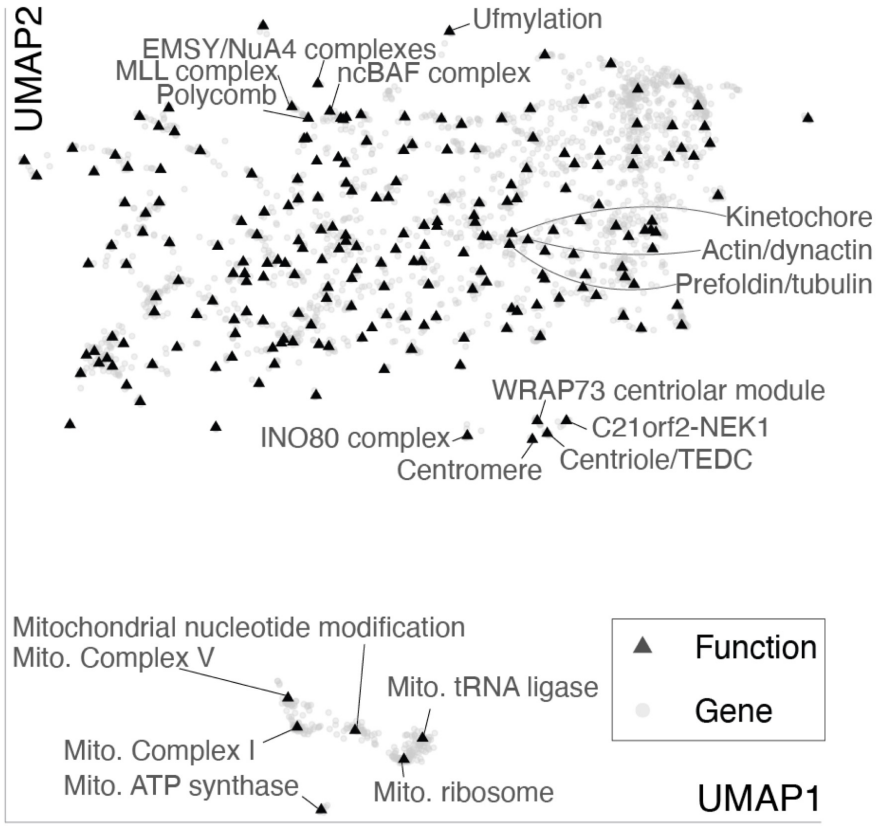
SHOC2 ≈ Activated KRAS + Activated NRAS + EGFR Signaling + FGFR Signaling



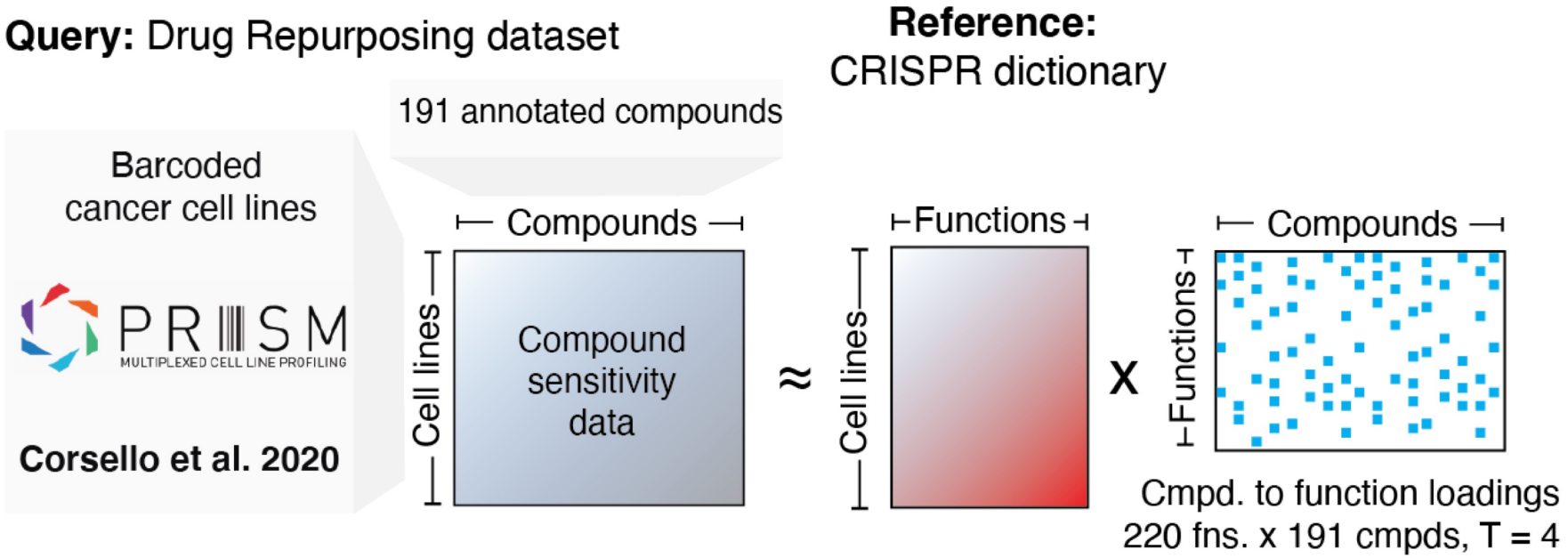
Joint embedding space of genes and functions



It captures interpretable processes in cancer

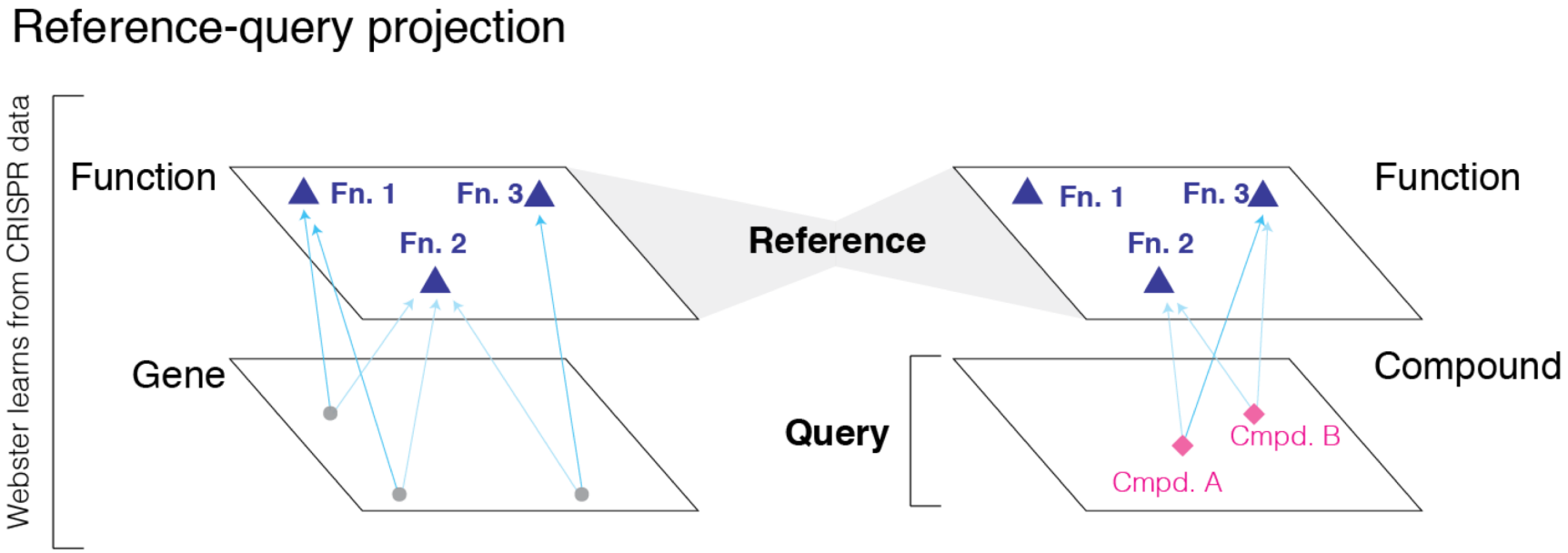


Part 3: Compound sensitivity screens



Modeling compound sensitivity profiles as mixtures of functions learned from CRISPR

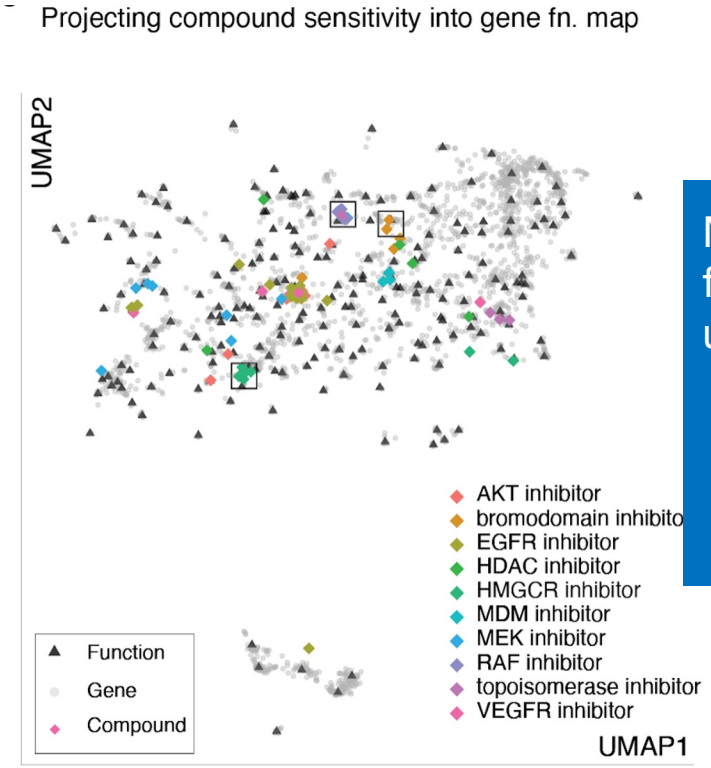
Modeling compounds as mixtures of latent functions

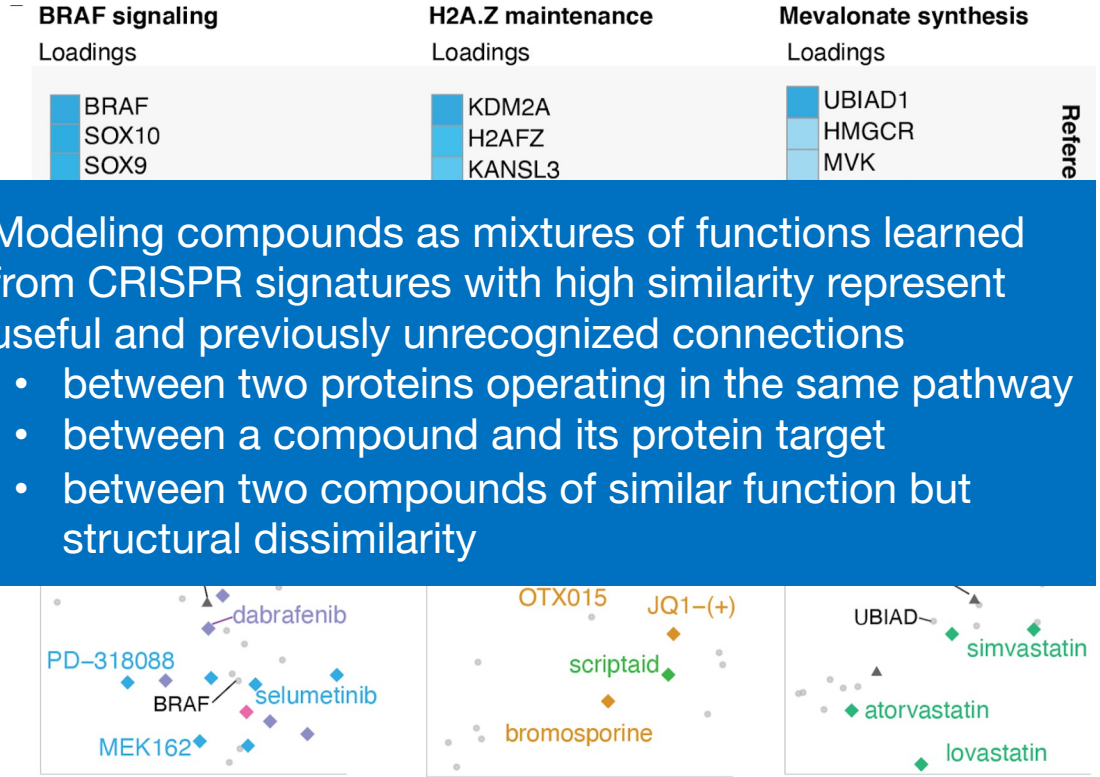


- Modeling compounds as mixtures of functions learned from CRISPR signatures with high similarity represent useful and previously unrecognized connections
  - between two proteins operating in the same pathway
  - between a small-molecule and its protein target
  - between two small-molecules of similar function but structural dissimilarity
- Such a catalog of connections can serve as a functional look-up table of compounds to predict sensitivity and genotoxic profiles and to inform therapeutic use

Compounds' mechanisms of action

Compounds are embedded nearby gene functions, reflecting their mechanism of action





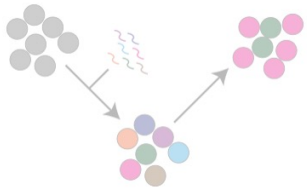
Modeling compounds as mixtures of functions learned from CRISPR signatures with high similarity represent useful and previously unrecognized connections

- between two proteins operating in the same pathway
- between a compound and its protein target
- between two compounds of similar function but structural dissimilarity

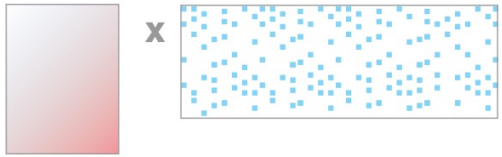
Key takeaways

- Analogously to word semantics, genes can be modeled as **distributions over latent bio functions**
  - **Sparse learning** is an effective strategy for learning bio functions from high-dimensional chemical and genetic perturbations
  - New perturbations can be **projected** into learned space

Data: high-dimensional gene perturbation measurements

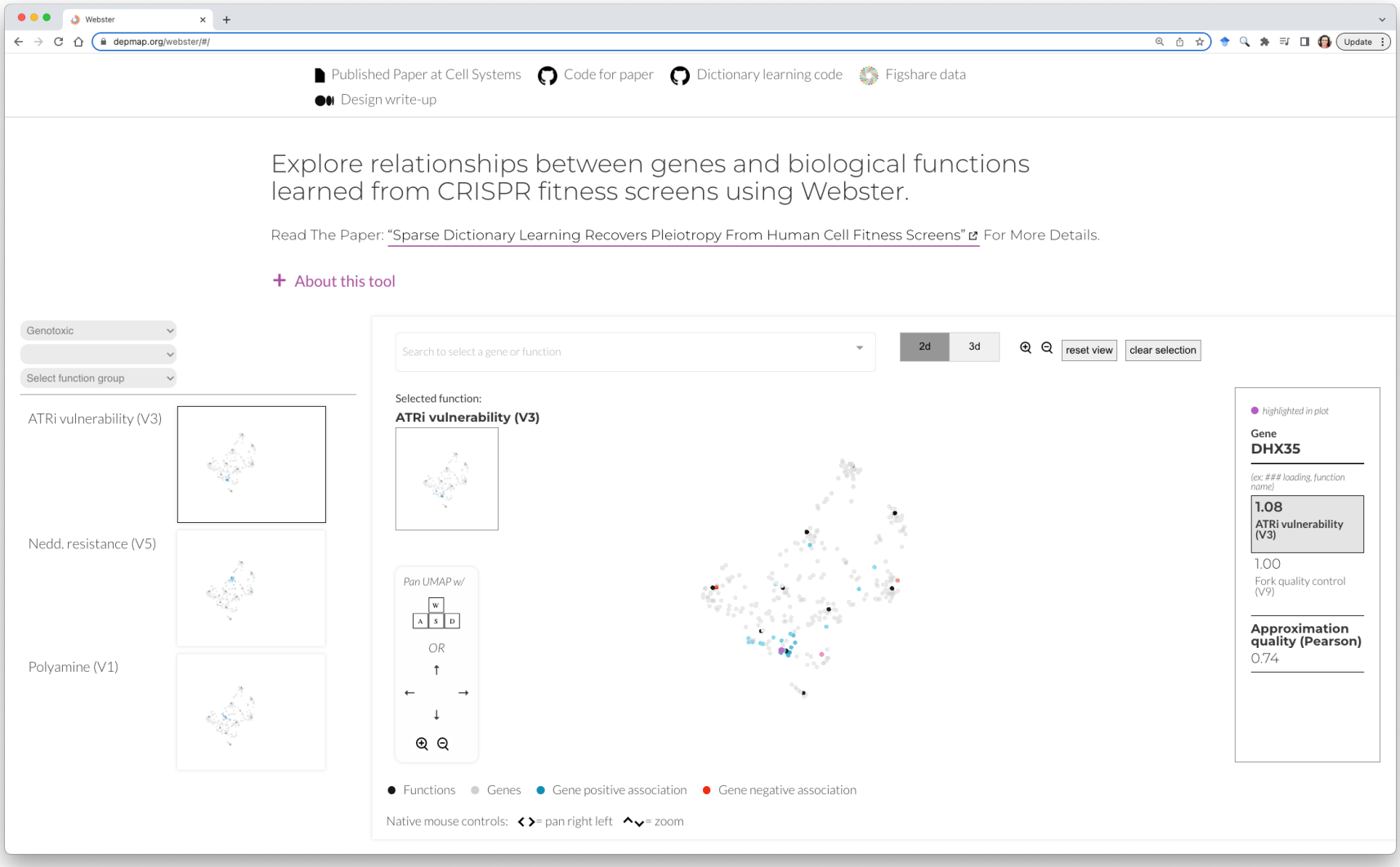


Approach: sparse approximation embeddings



geneA - func1 + func2 ≈ geneB

https://depmap.org/webster



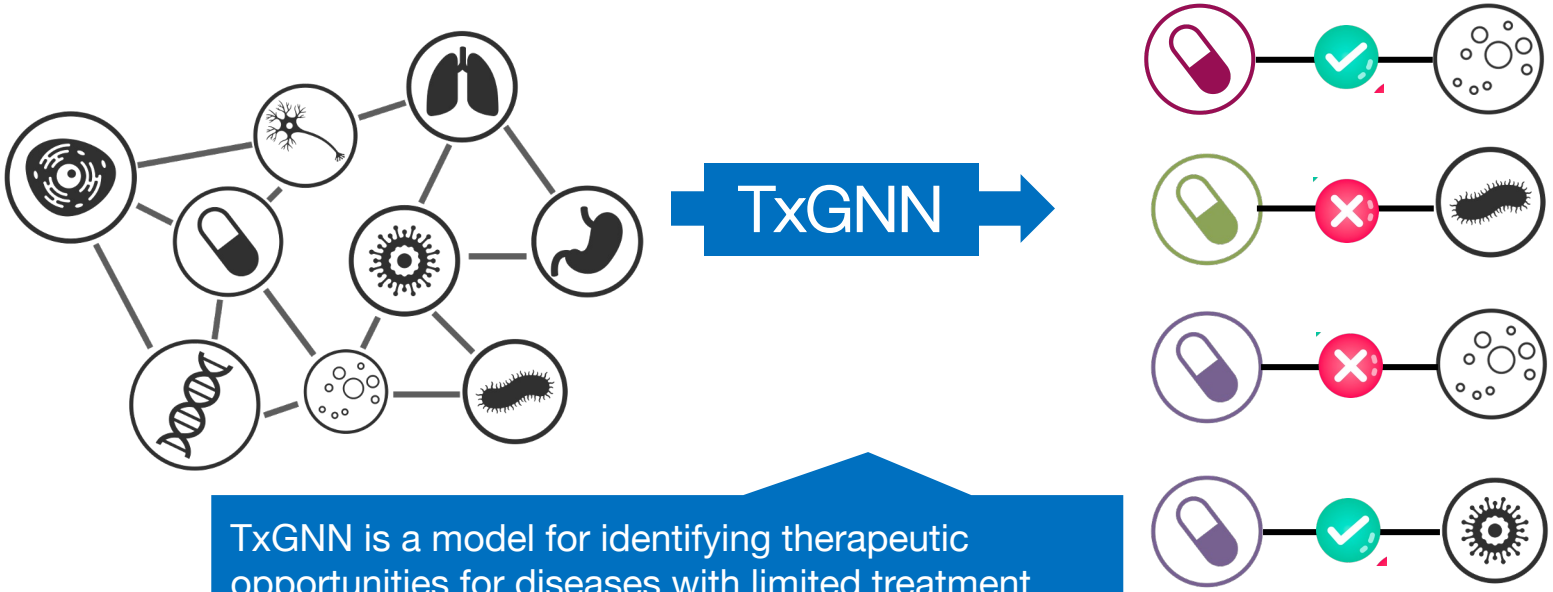
Outline for today's class

- High-throughput genetic and chemical perturbations
- Therapeutic use prediction, indication and contra-indication inference
- Drug repurposing

Therapeutic use prediction

Comprehensive knowledge graph of 17,080 clinically-recognized diseases

Process various therapeutic tasks, such as indication and contraindication prediction, in a unified formulation

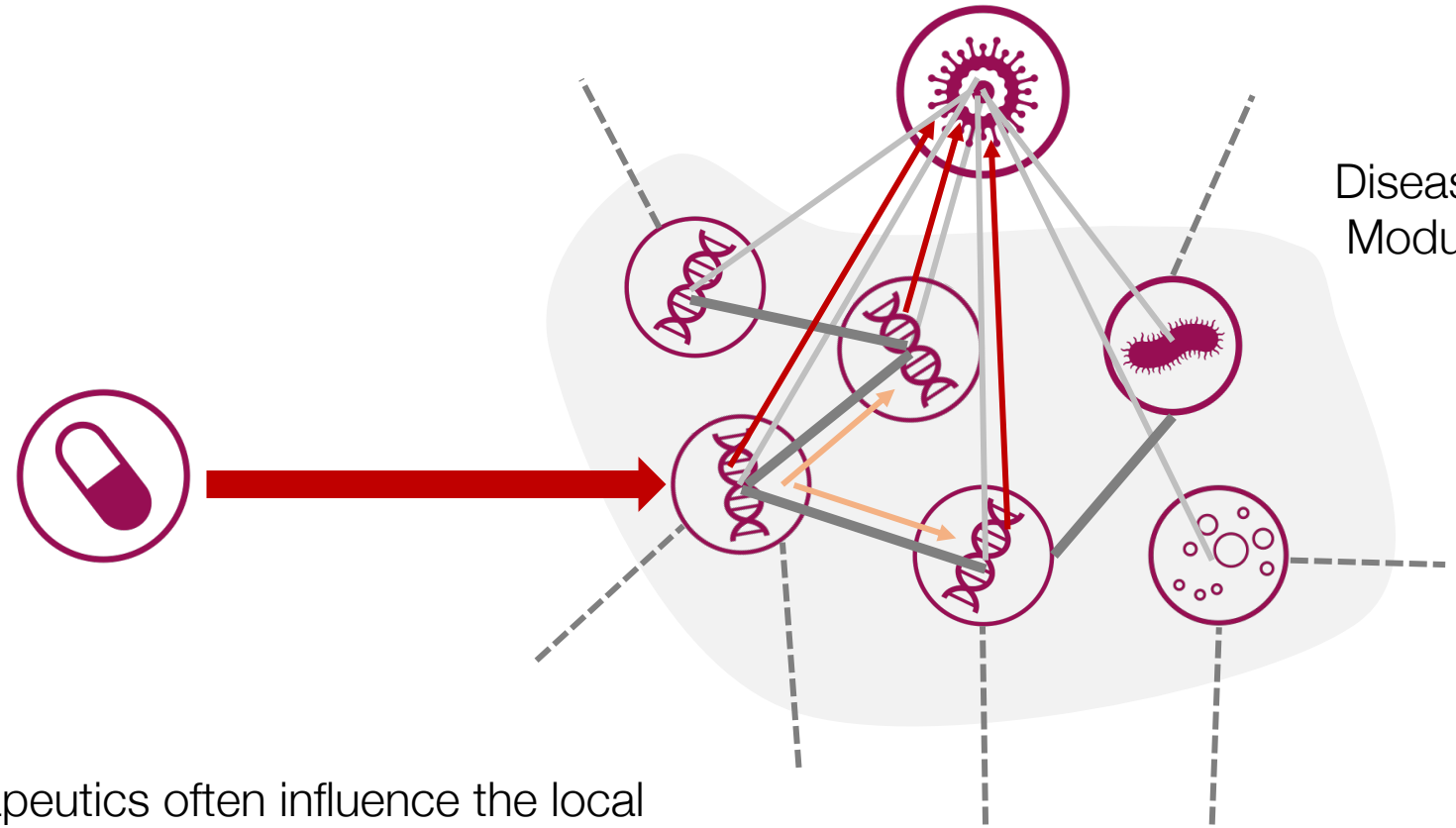


TxGNN is a model for identifying therapeutic opportunities for diseases with limited treatment options and molecular understanding. It is a graph neural network pre-trained on a comprehensive knowledge graph of 17,080 clinically-recognized diseases and 7,957 therapeutic candidates

Applications:
Drug repurposing/virtual screening
Understanding disease mechanisms
Understanding treatment effects

Zero-Shot Prediction of Therapeutic Use with Geometric Deep Learning and Clinician Centered Design, medRxiv, 2023

TxGNN: Mechanistic view of therapeutic effects



Therapeutics often influence the local biological system of disease-associated agents to create therapeutics effects

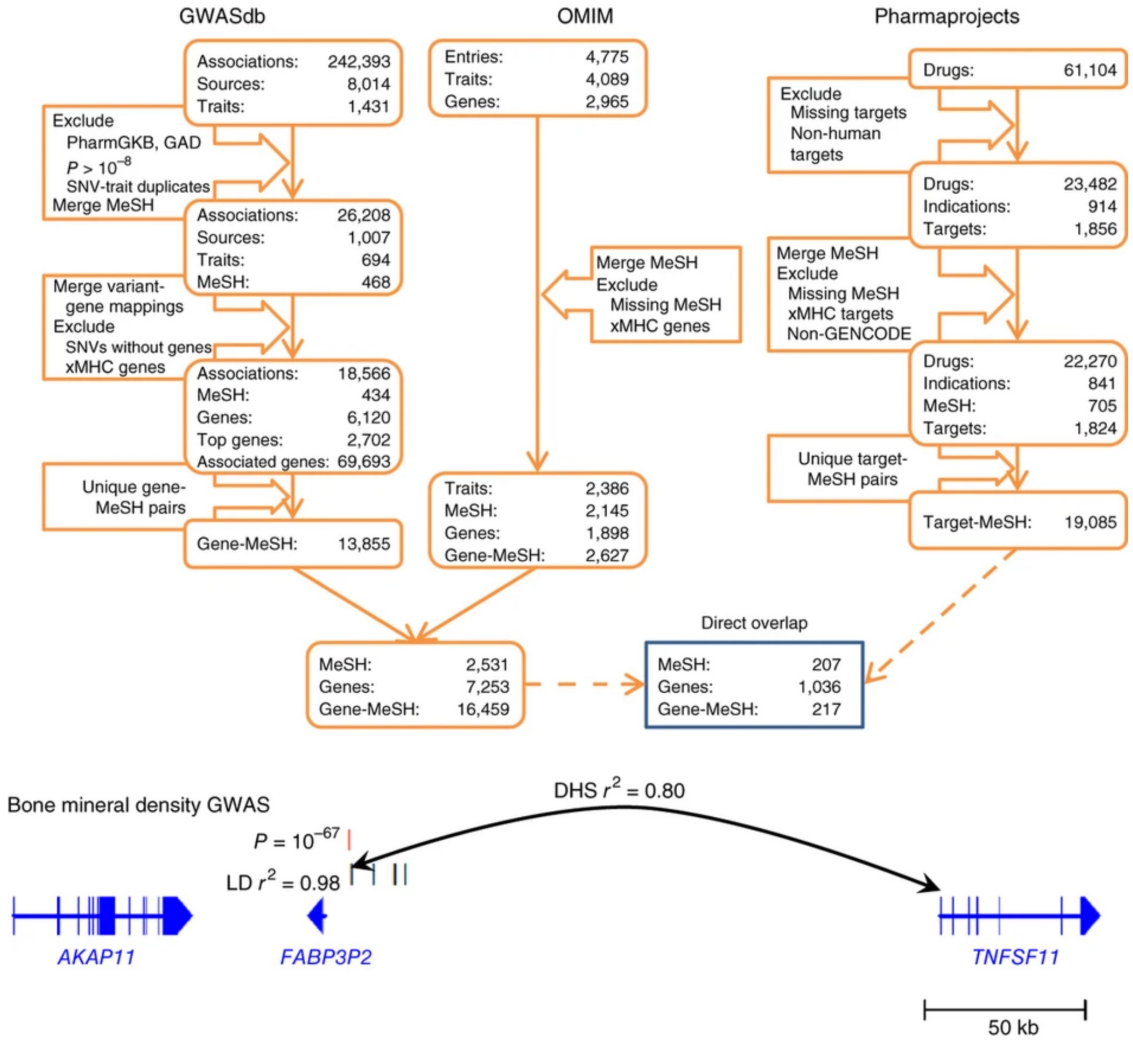
Li et al., Graph Representation Learning in Biomedicine and Healthcare. Nature Biomedical Engineering, 2022

TxGNN: Mechanistic view of therapeutic effects

- Growing insight into genes that influence human disease may affect how drug targets and indications are selected
- Questions:
  - How well the current archive of genetic evidence predicts drug mechanisms?
  - Can using the growing wealth of human genetic data to select the best targets and indications have a measurable impact on the successful development of new drugs?

Nelson et al., The support of human genetic evidence for approved drug indications. Nature Genetics, 2015

TxGNN: Mechanistic view of therapeutic effects

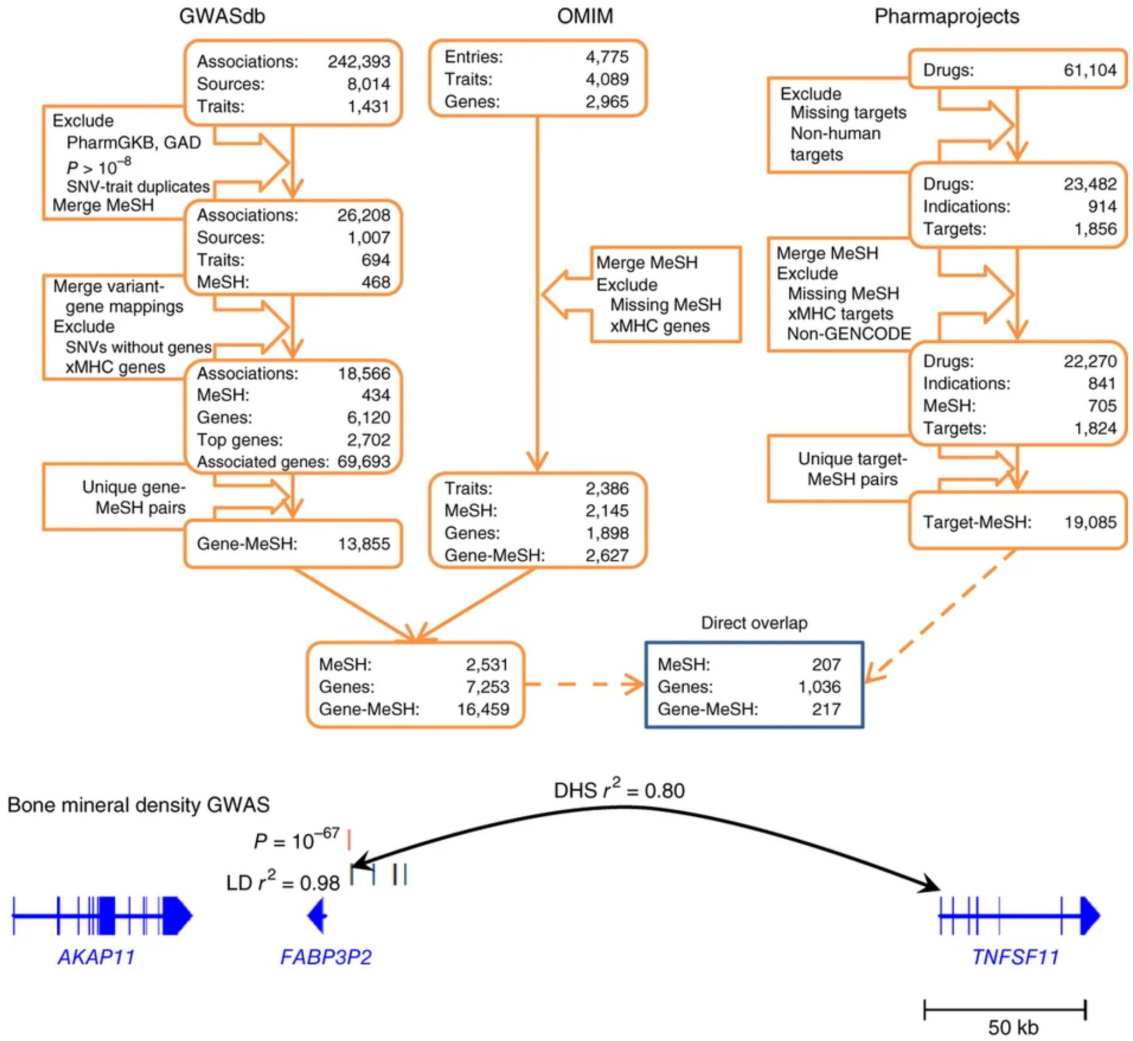


Summary of each data resource and the key filtering and processing steps applied to create the final set of gene-trait and drug target–indication combinations investigated in this study. GWASdb sources correspond to unique PubMed IDs or other unique data sources given for each association. GAD, Genetic Association Database

Approach to mapping genetically associated variants to genes. Example illustrated with the bone mineral density GWAS association with rs9533090 (depicted in red). Of five SNPs in strong LD with rs9533090 (r2 ≥ 0.8), one falls within a DNase I–hypersensitive site (DHS) that was found to have a signal correlated with the DHS of the TNFSF11 gene transcription start site

Nelson et al., The support of human genetic evidence for approved drug indications. Nature Genetics, 2015

TxGNN: Mechanistic view of therapeutic effects



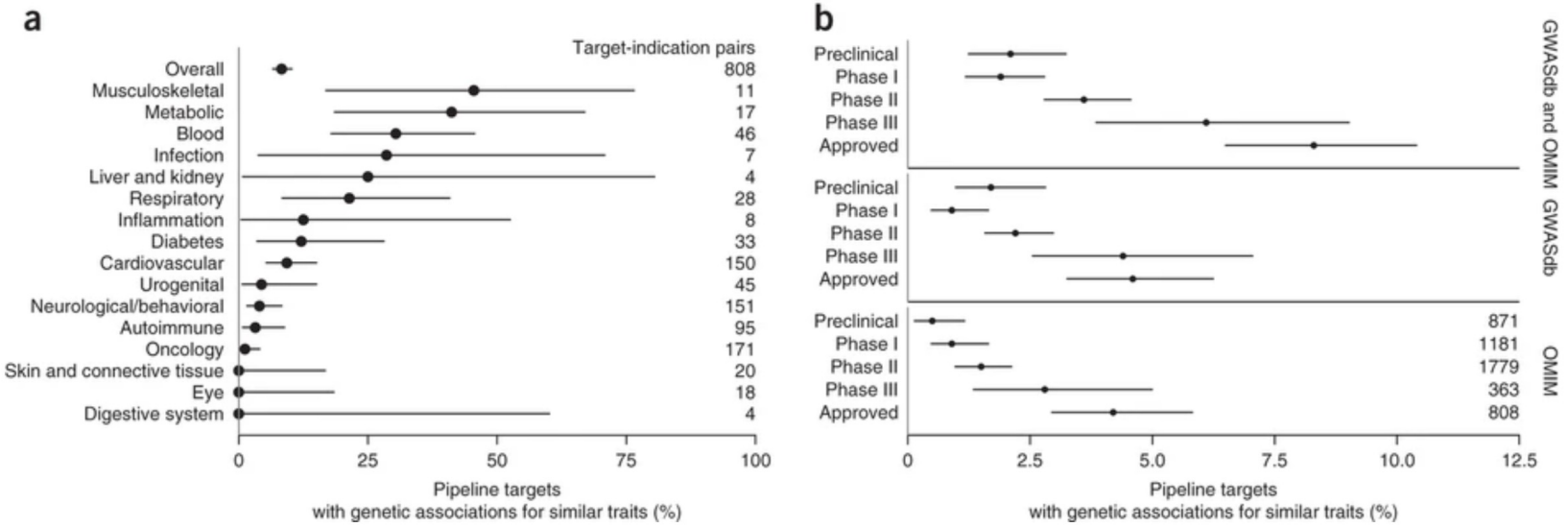
Summary of each data resource and the key filtering and processing steps applied to create the final set of gene-trait and drug target–indication combinations investigated in this study. GWASdb sources correspond to unique PubMed IDs or other unique data sources given for each association. GAD, Genetic Association Database

Approach to mapping genetically associated variants to genes. Example illustrated with the bone mineral density GWAS association with rs9533090 (depicted in red). Of five SNPs in strong LD with rs9533090 (r2 ≥ 0.8), one falls within a DNase I–hypersensitive site (DHS) that was found to have a signal correlated with the DHS of the TNFSF11 gene transcription start site

Nelson et al., The support of human genetic evidence for approved drug indications. Nature Genetics, 2015

TxGNN: Mechanistic view of therapeutic effects

Percentage of target-indication pairs for drugs approved in the United States or the European Union overlapping with gene-trait combinations

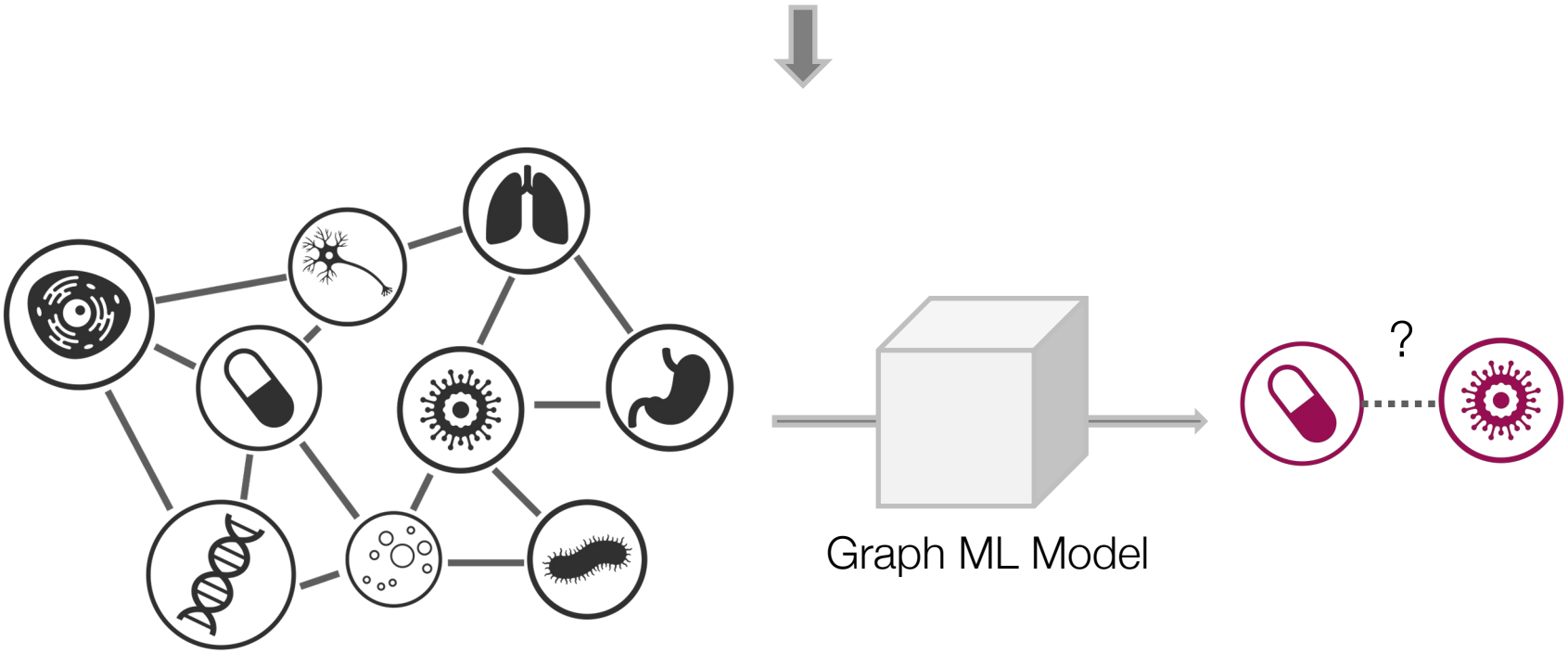


- Among well-studied indications, the proportion of drug mechanisms with direct genetic support increases significantly across the drug development pipeline, from 2.0% at the preclinical stage to 8.2% among mechanisms for approved drugs, and varies dramatically among disease areas.
- Selecting genetically supported targets could double the success rate in clinical development

Nelson et al., The support of human genetic evidence for approved drug indications. Nature Genetics, 2015

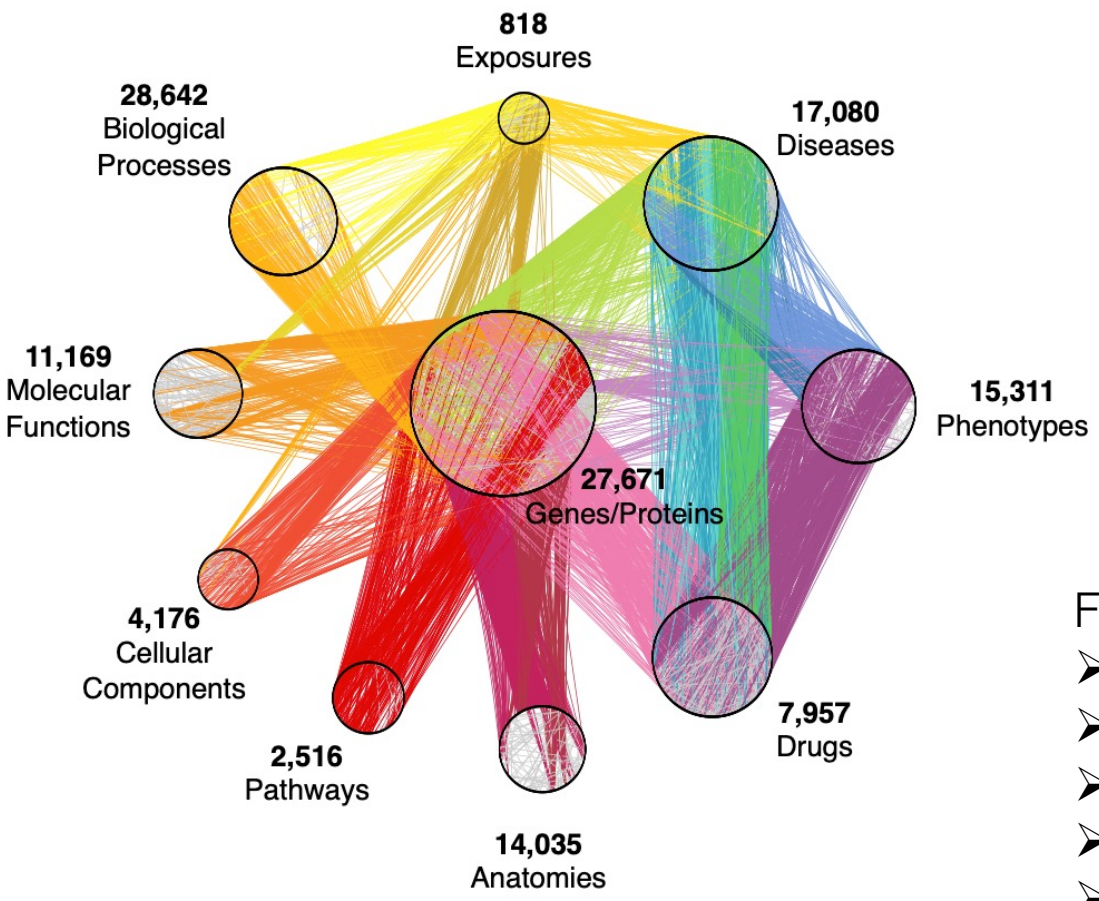
TxGNN

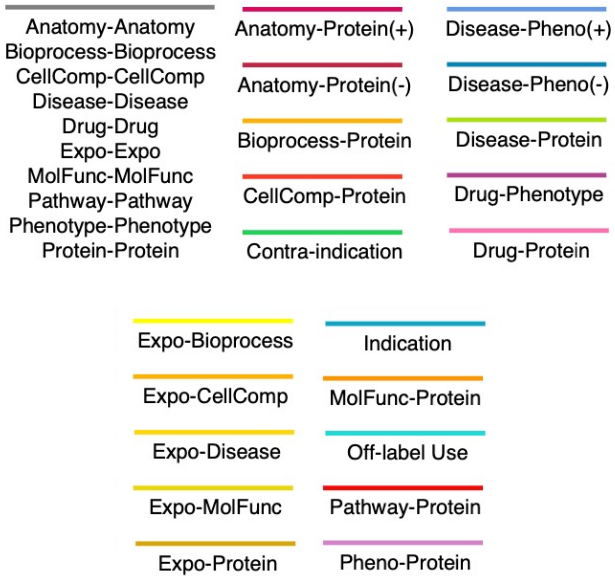
To model this mechanistic view, we need to ground the model in known mechanisms of diseases and treatment effect



Zero-Shot Prediction of Therapeutic Use with Geometric Deep Learning and Clinician Centered Design, medRxiv, 2023

Dataset: PrimeKG





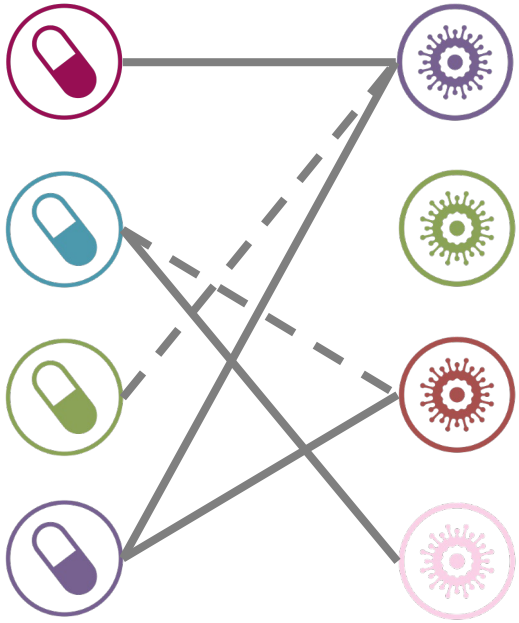
From 17 data resources.

- 10 node types.
- 29 relation types.
- 100K nodes.
- 4 millions connections.
- 9,388 indications from 1,361 diseases and 1,801 drugs.

Building a Knowledge Graph to Enable Precision Medicine, Scientific Data, 2023

Seeting: Baseline approach

Random split across known drug-disease pairs



Train Drug-Disease Pair

Test Drug-Disease Pair

Treatment candidate

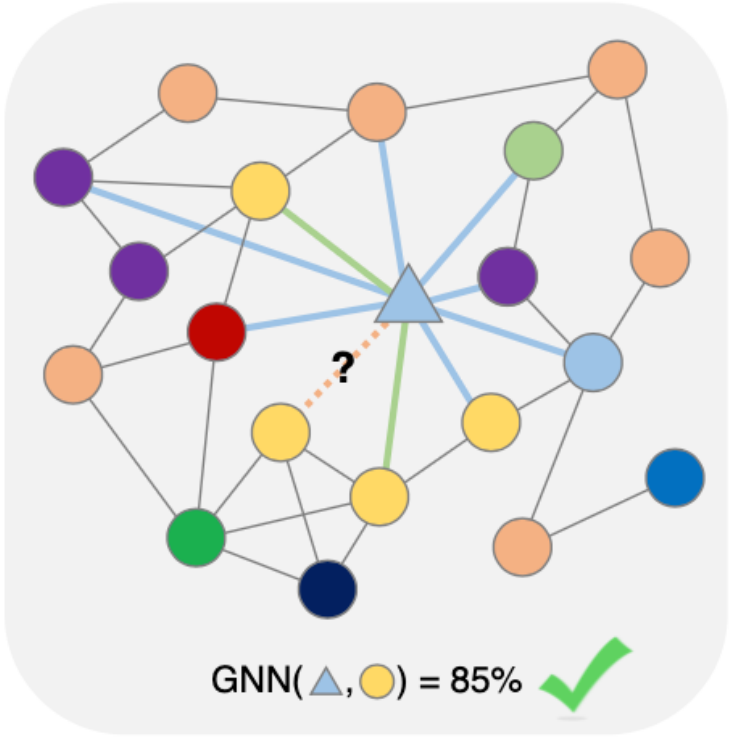
Molecular underpinnings

Existing treatments

Target disease

Drug

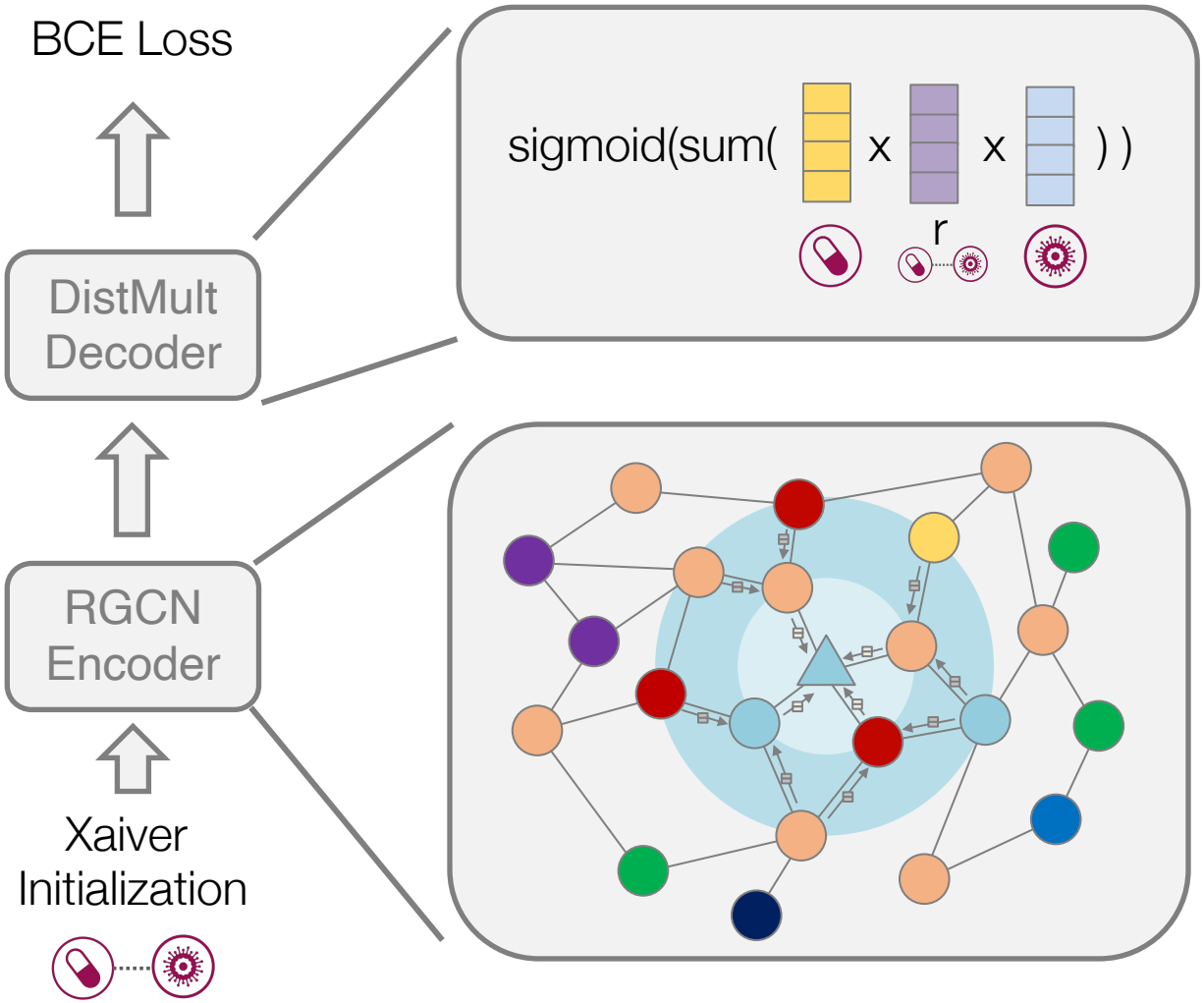
Other node types

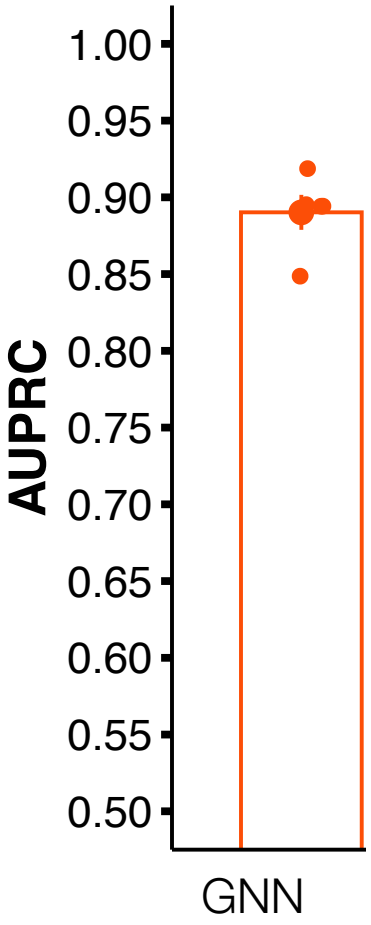


Scenario A

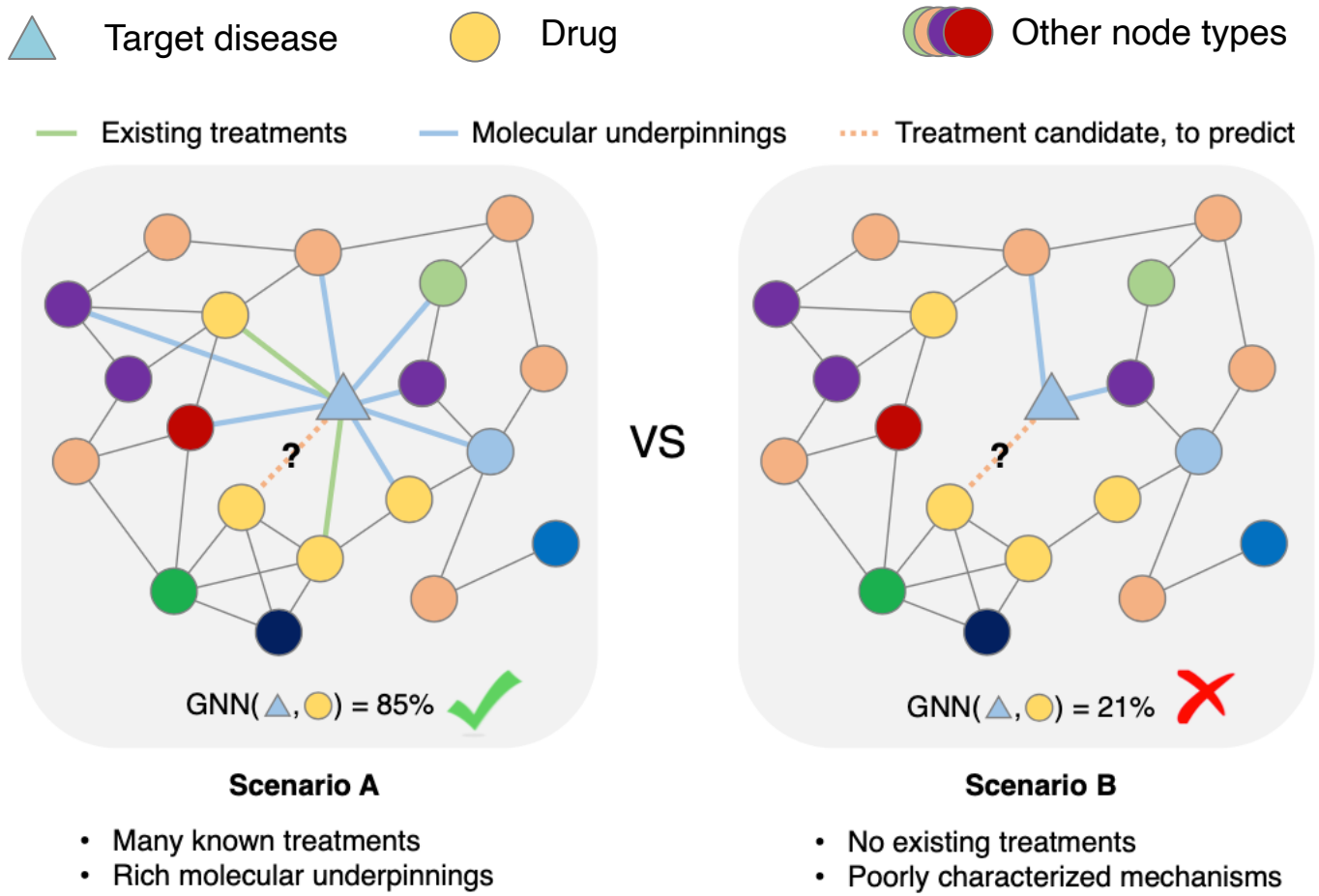
- Many known treatments
- Rich molecular underpinnings

In this setting, existing methods perform well



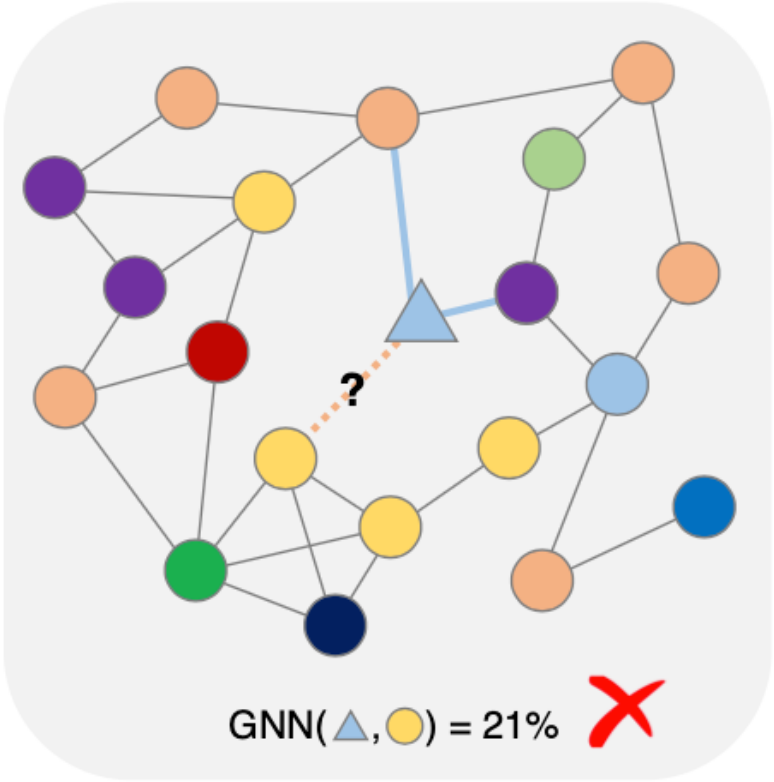


How about other settings?



No treatments = No links between disease and any drug nodes
Poorly characterized mechanisms = Sparse local neighborhoods

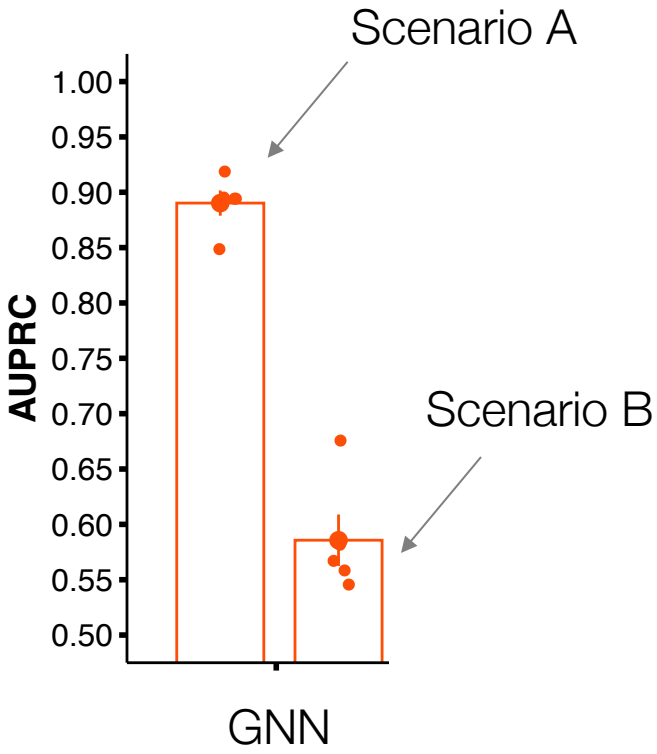
Performance in other settings



Scenario B

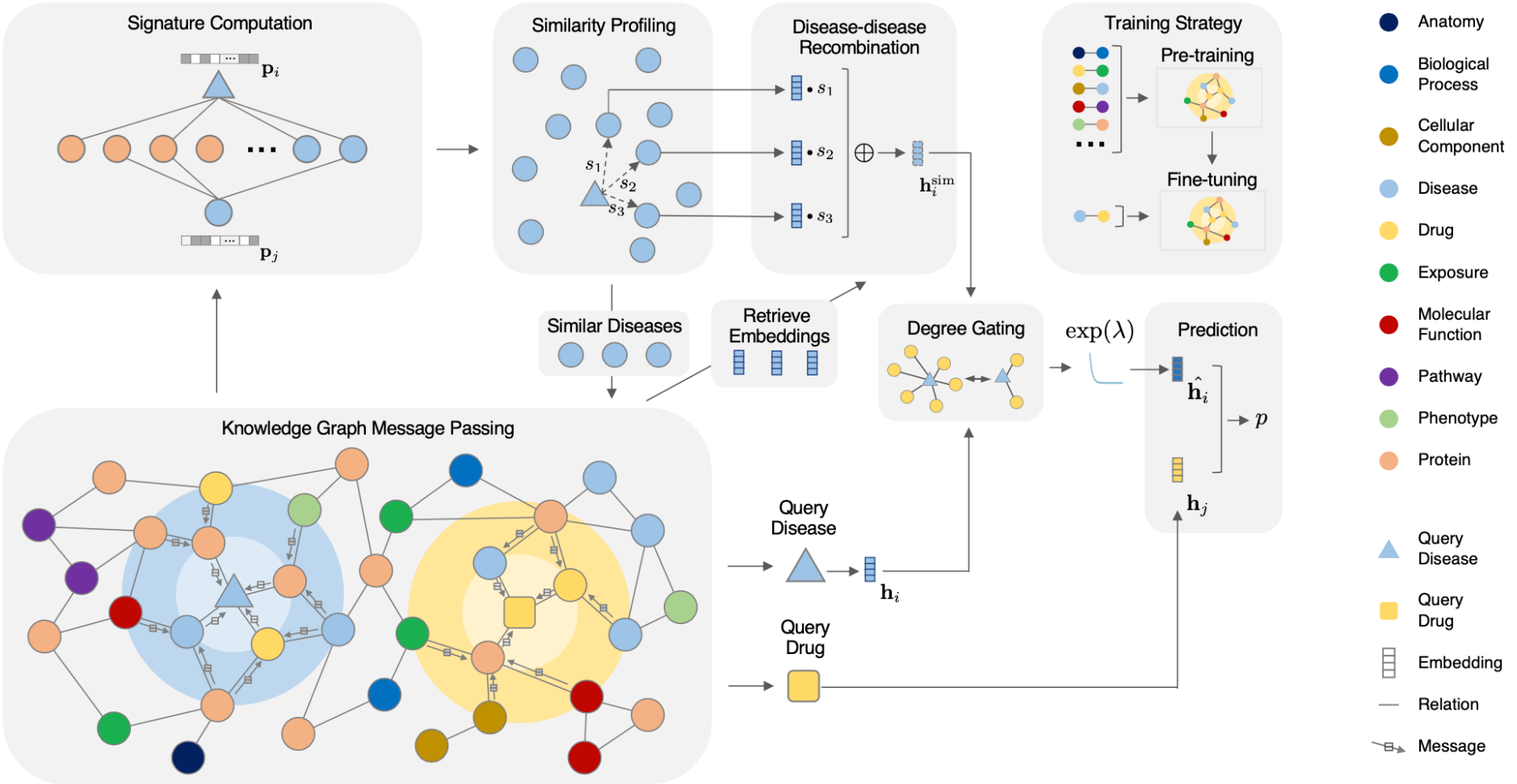
- No existing treatments
- Poorly characterized mechanisms
- Challenging to predict

Disease embeddings are less meaningful because so many relationships are unknown



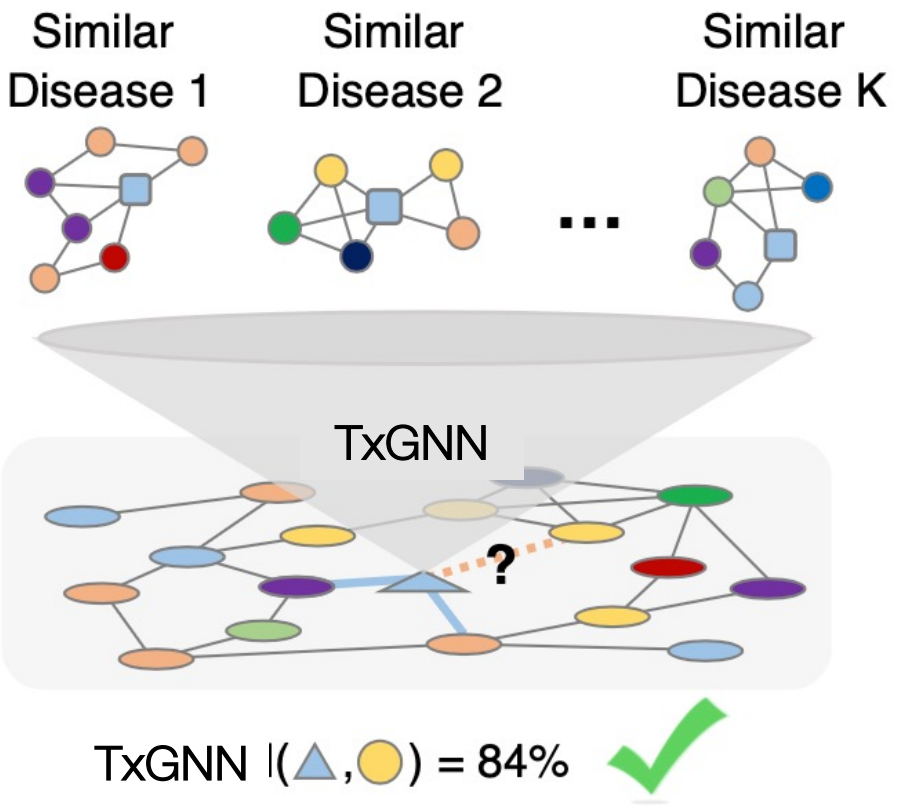
Need better disease embeddings -- Is there an inductive bias (biological rationale) that can be incorporated into the ML model?

Approach: TxGNN



Zero-Shot Prediction of Therapeutic Use with Geometric Deep Learning and Clinician Centered Design, medRxiv, 2023

TxGNN: Transfer learning across diseases

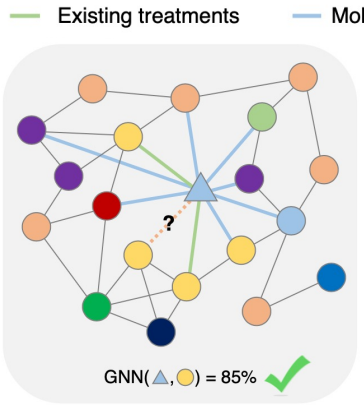


(1) identify similar diseases

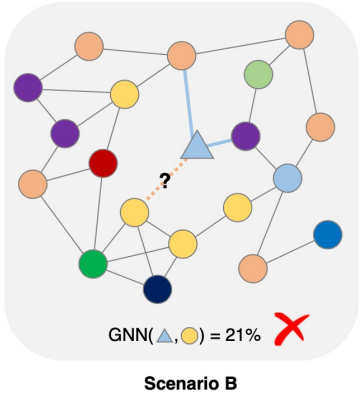
(2) leverage disease similarities

Results: Therapeutic use prediction

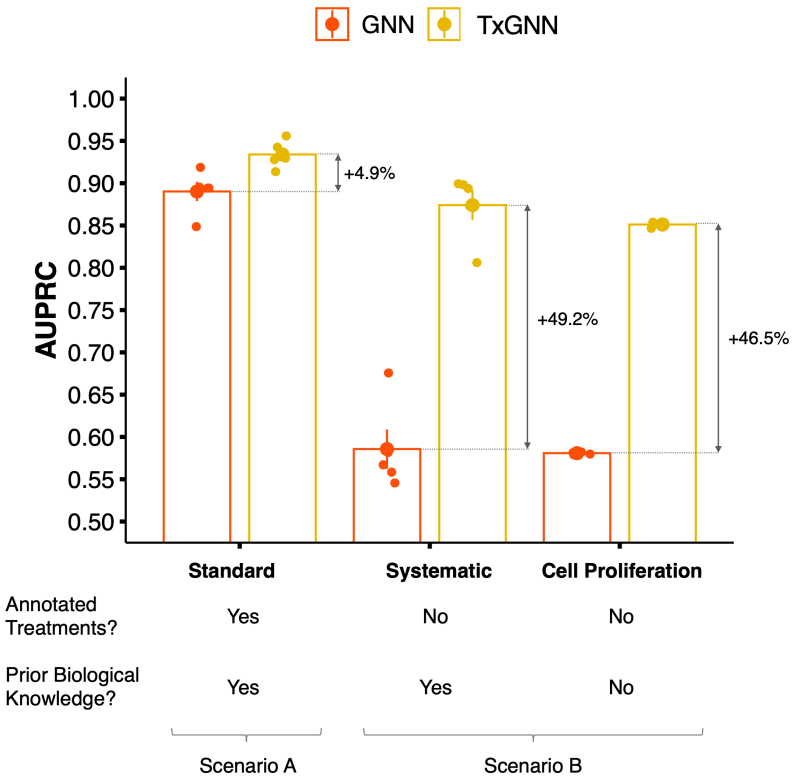
- Once trained, TXGNN can perform zero-shot inference on new diseases without additional parameters or fine-tuning on ground truth labels



- Many known treatments
- Known molecular understanding
- "Easy" to predict

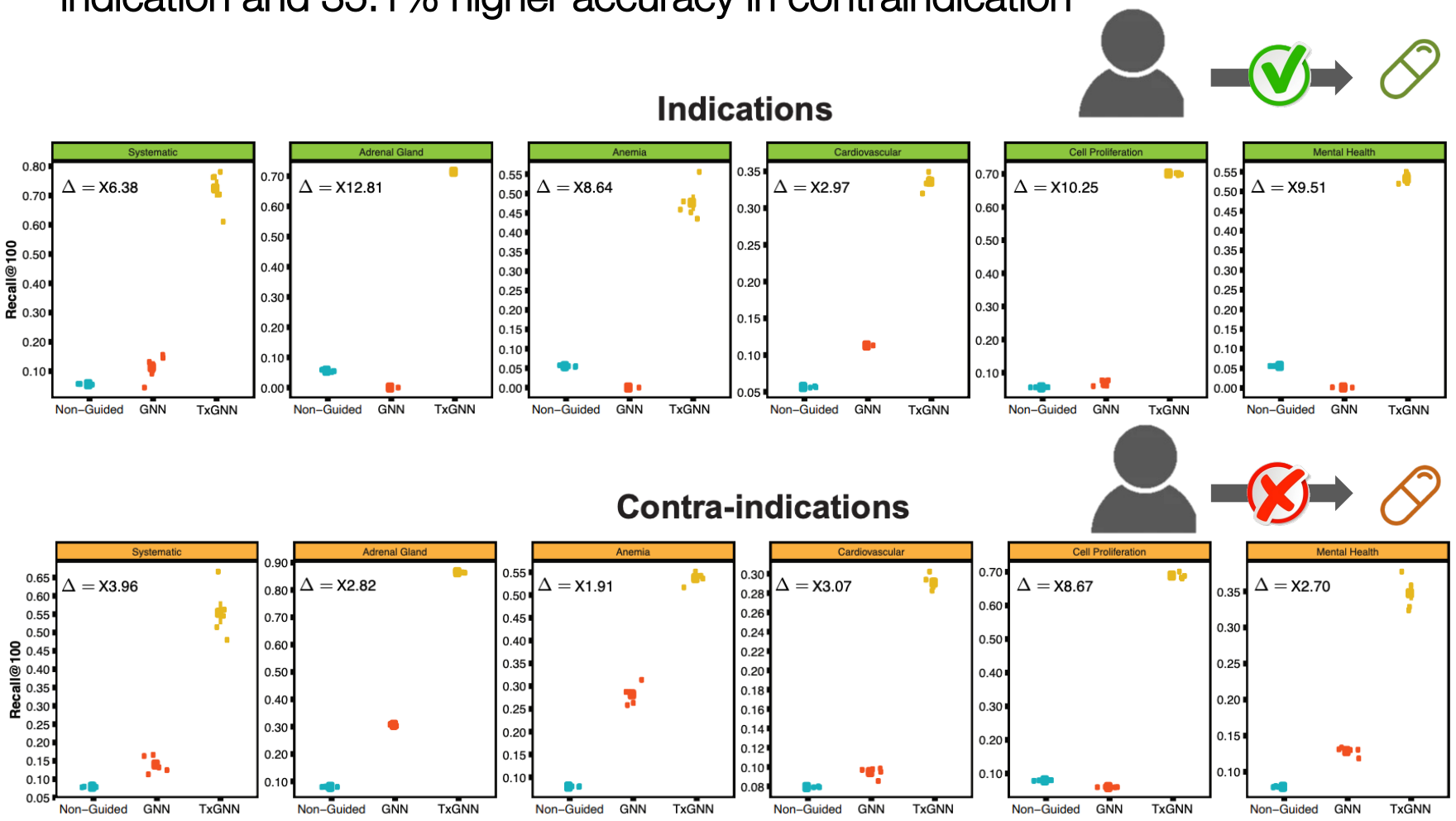


- No known treatments
- Poor molecular understanding
- "Hard" to predict



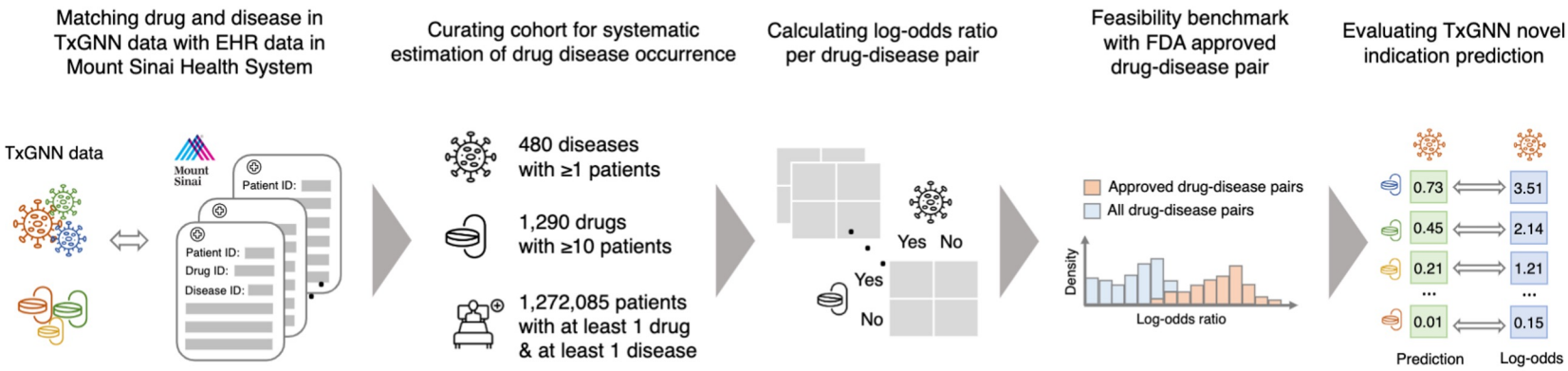
Results: Therapeutic use prediction

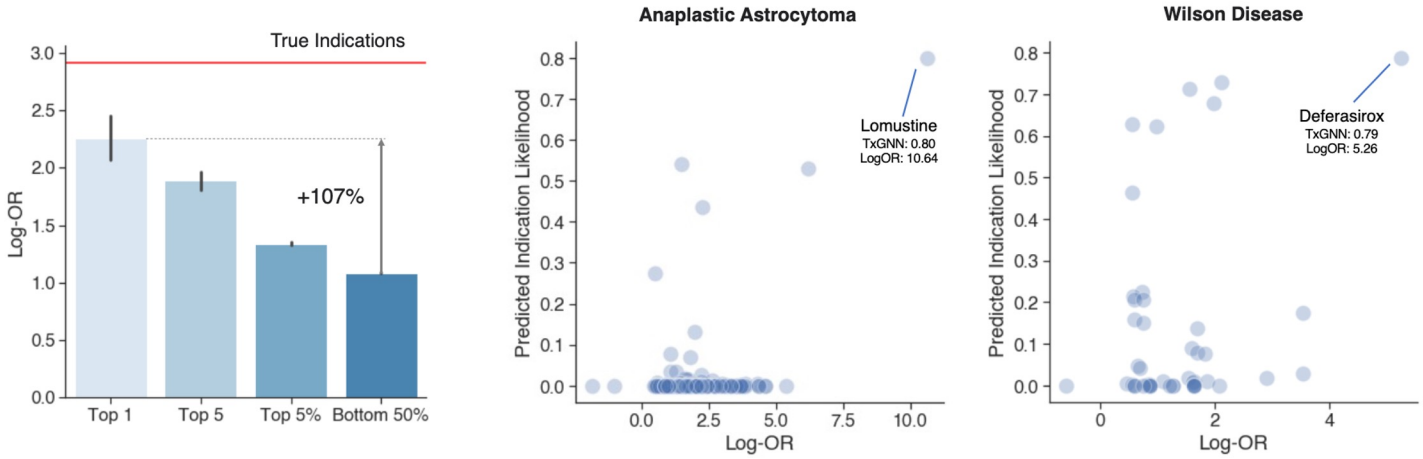
- TxGNN improves over existing methods, with up to 49.2% higher accuracy in indication and 35.1% higher accuracy in contraindication



Results: Therapeutic use prediction

- TxGNN's novel predictions are consistent with off-label prescription decisions made by clinicians in a large healthcare system





Zero-Shot Prediction of Therapeutic Use with Geometric Deep Learning and Clinician Centered Design, medRxiv, 2023

Results: Therapeutic use prediction

- TxGNN can also predict therapeutic use for recent FDA approvals

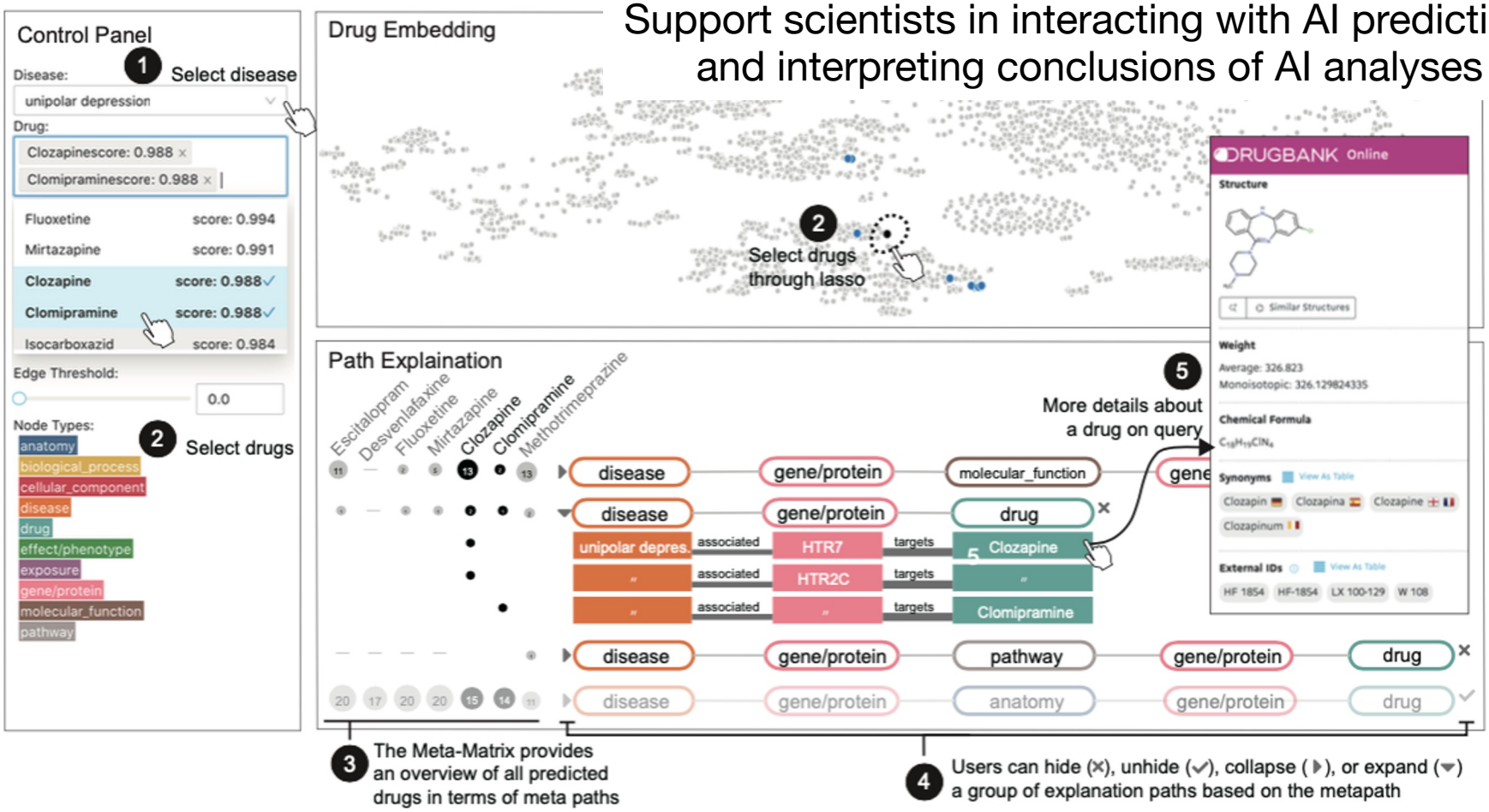
Drug name	Ingredient	Disease	Approval date	Company	FDA Number	Orphan	Prediction	Percentile
Welireg	Belzutifan	von Hippel-Lindau disease	08/13/2021	Merck	NDA215383	Yes	0.720	4.11%
Livtencity	Maribavir	Cytomegalovirus infection	11/23/2021	Takeda	NDA215596	Yes	0.033	66.37%
Tezspire	Tezepelumab-Ekko	Asthma	12/17/2021	Astrazeneca	BLA761224	No	0.233	32.41%
Leqvio	Inclisiran Sodium	Familial hypercholesterolemia	12/22/2021	Novartis	NDA214012	No	0.301	19.32%
Adbry	Tralokinumab	Atopic dermatitis	12/27/2021	Leo Pharma	BLA761180	No	0.040	50.37%
Vabysmo	Faricimab-Svoa	Macular degeneration	01/28/2022	Genentech	BLA761235	No	0.938	2.25%
Vonjo	Pacritinib Citrate	Myelofibrosis	02/28/2022	Cti Biopharma	NDA208712	Yes	0.011	63.14%
Ztalmy	Ganaxolone	CDKL5 disorder	03/18/2022	Marinus	NDA215904	Yes	0.335	18.73%
Mounjaro	Tirzepatide	Type 2 diabetes mellitus	05/13/2022	Eli Lilly	NDA215866	No	0.286	12.50%
Vtama	Tapinarof	Psoriasis	05/23/2022	Dermavant	NDA215272	No	0.261	32.70%

Zero-Shot Prediction of Therapeutic Use with Geometric Deep Learning and Clinician Centered Design, medRxiv, 2023

AI-clinician collaboration

"Will clozapine treat unipolar depression? What is the disease treatment mechanism?"

Support scientists in interacting with AI predictions and interpreting conclusions of AI analyses

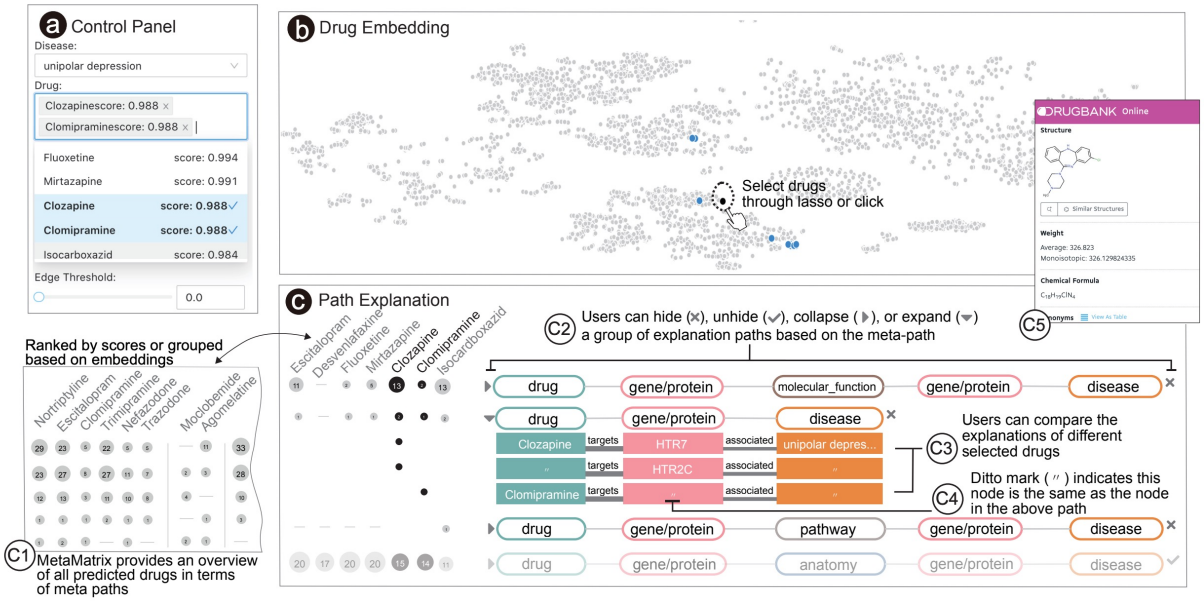


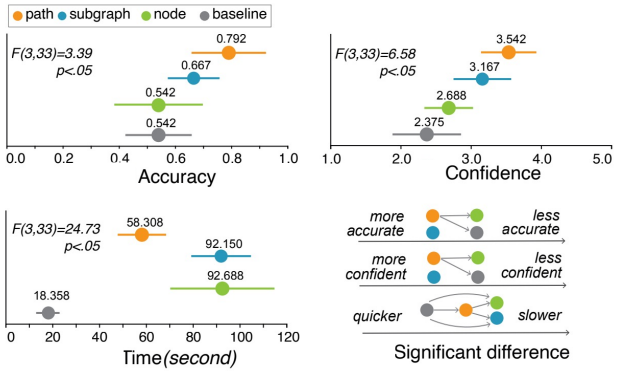
Probing GNN Explainers: A Rigorous Theoretical and Empirical Analysis of GNN Explanation Methods, *AISTATS* 2022

Extending the Nested Model for User-Centric XAI: A Design Study on GNN-based Drug Repurposing, *IEEE VIS* 2022 (**Best Paper Award**)

Identification of Disease Treatment Mechanisms through the Multiscale Interactome, Nature Communications 2021

Clinician-centered AI design

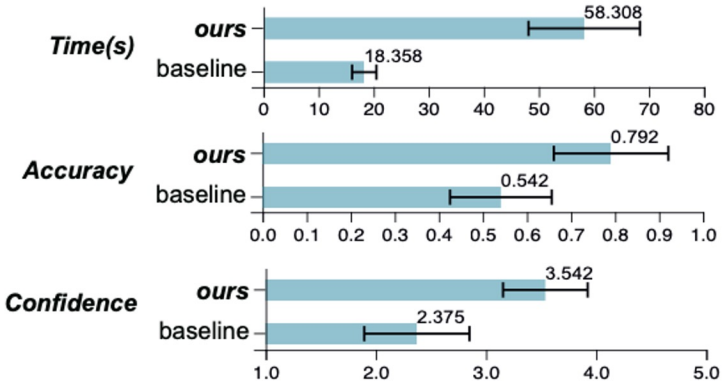




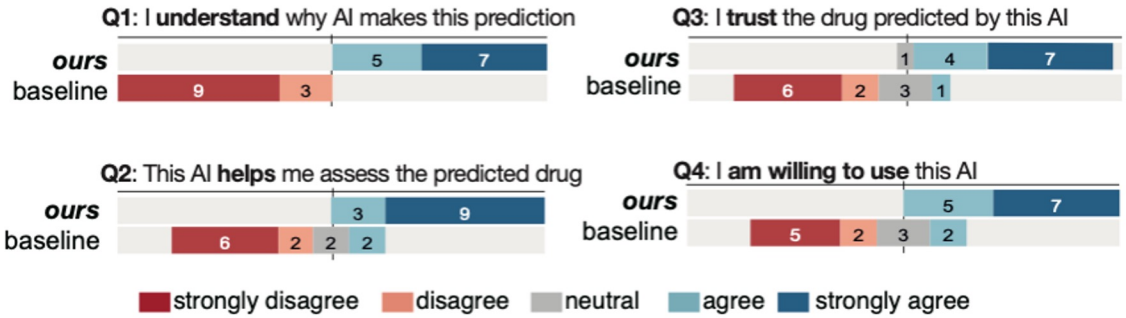
Zero-shot prediction of therapeutic use with geometric deep learning and clinician centered design, medRxiv, 2023
Probing GNN Explainers: A Rigorous Theoretical and Empirical Analysis of GNN Explanation Methods, *AISTATS* 2022
Extending the Nested Model for User-Centric XAI: A Design Study on GNN-based Drug Repurposing, *IEEE VIS* 2022 (**Best Paper Award**)
Identification of Disease Treatment Mechanisms through the Multiscale Interactome, Nature Communications 2021

Usability study with end users

Compared to a no-explanation baseline in terms of **user answer accuracy**, **exploration time**, **user confidence**, and **user agreement** across a spectrum of usability questions



Error bars indicate the 95% confidence intervals

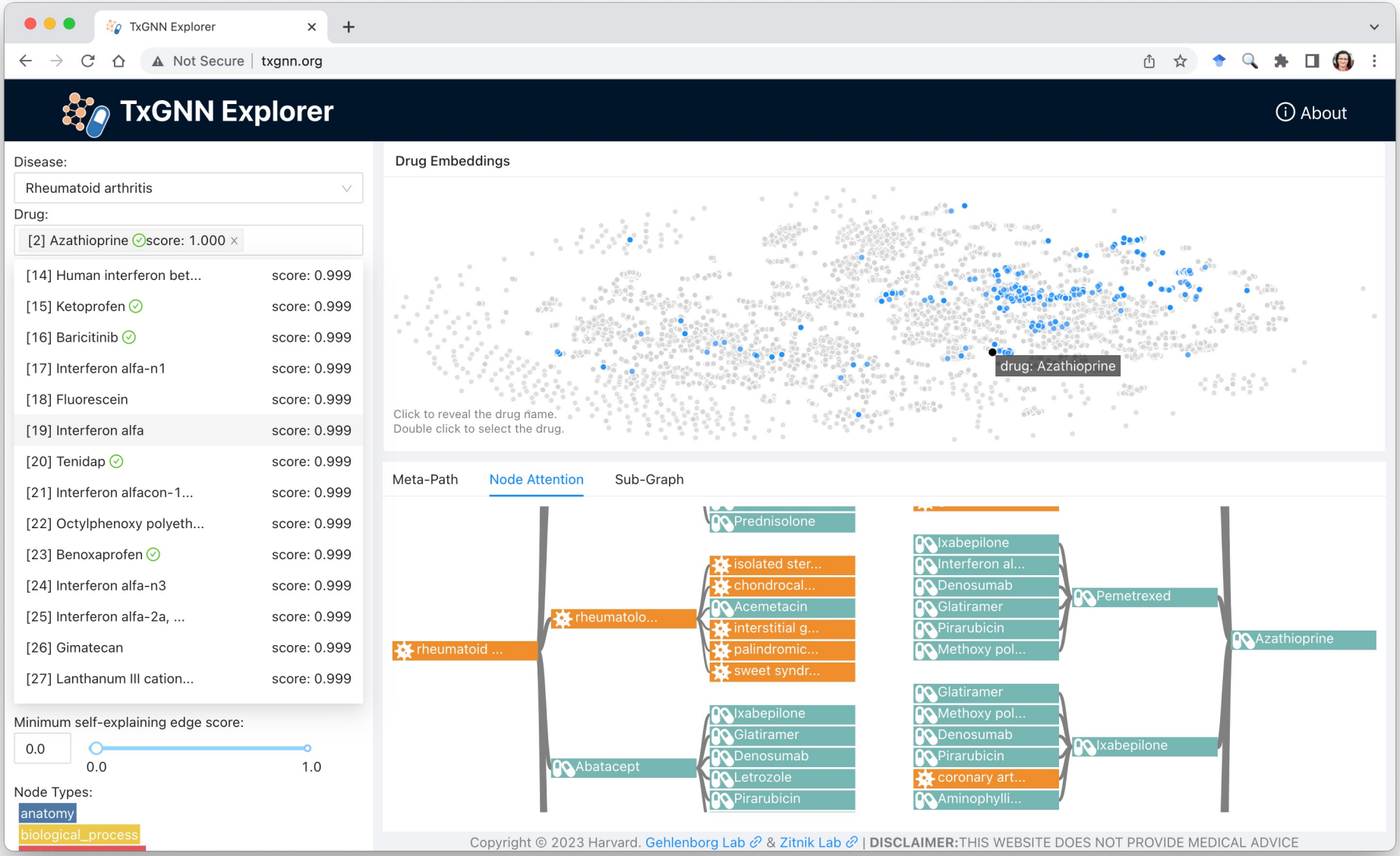


Agree scores are placed to the right, disagree to the left

Mount Sinai

Zero-Shot Prediction of Therapeutic Use with Geometric Deep Learning and Clinician Centered Design, medRxiv, 2023
Extending the Nested Model for User-Centric XAI: A Design Study on GNN-based Drug Repurposing, *IEEE VIS* 2022 (**Best Paper Award**)

http://txgnn.org



L13 Quick Check

https://forms.gle/B5PBaa2DCTLZpEqh8

BMI 702: Biomedical Artificial Intelligence

Foundations of Biomedical Informatics II, Spring 2023

Quick check quiz for lecture 13: Label-efficient learning, few-shot learning, biomarker discovery, indication and contra-indication inference, drug repurposing, adverse event prediction.

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

Not shared

* Indicates required question

First and last name *

Your answer

Harvard email address *

Your answer

Go to http://txgnn.org and examine predictions for **rheumatoid arthritis**. Our evaluation will focus on disease-modifying antirheumatic drugs (DMARDs), which is a class of drugs indicated for the treatment of several inflammatory arthritides, including rheumatoid arthritis, as well as for the management of other connective tissue diseases and some cancers. Answer the following four questions. *

1) What is the predicted rank of **sulfasalazine**, a common conventional DMARD?

2) What is the predicted rank of **methotrexate**, another common DMARD?

3) Give two examples of reasoning paths (meta-paths) used by the algorithm to relate **rheumatoid arthritis** with **sulfasalazine.** Comment the results.

4) Give two examples of reasoning paths (meta-paths) used by the algorithm to relate **rheumatoid arthritis** with **methotrexate.** Comment the results. *Examine meta-paths that use this template: Disease-Drug-Gene/Protein-Drug.*

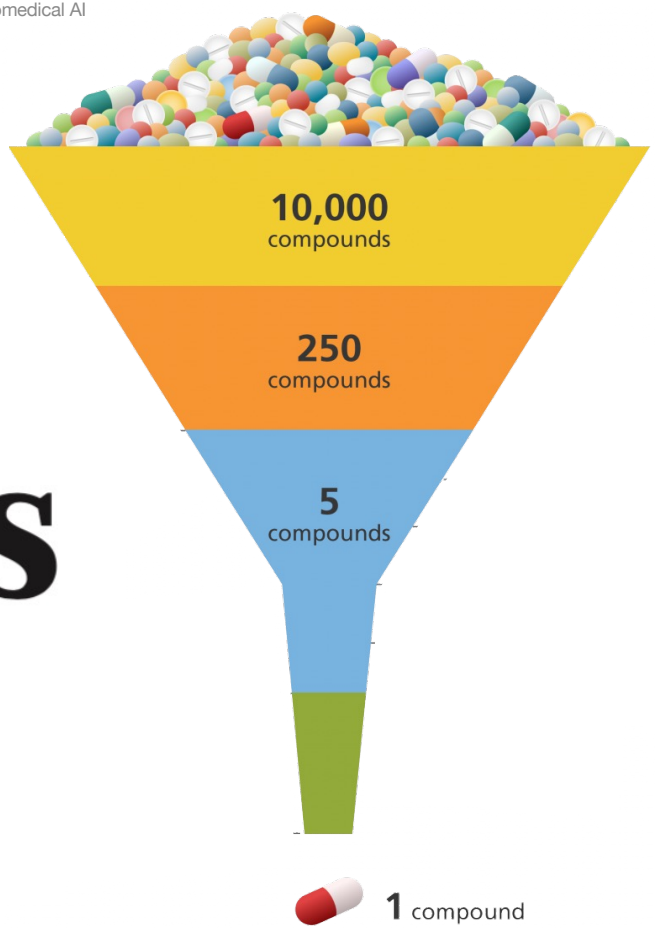
Your answer

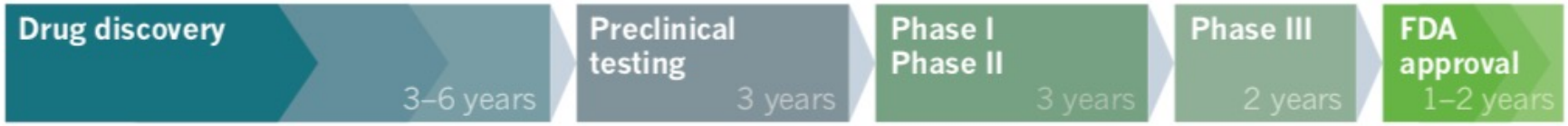
Outline for today's class

- High-throughput genetic and chemical perturbations
- Therapeutic use prediction, indication and contra-indication inference
- Drug repurposing

New tricks for old drugs

Faced with skyrocketing costs for developing new drugs, researchers are looking at ways to repurpose older ones — and even some that failed in initial trials.





12–16 years, **~$1 billion to $2 billion**

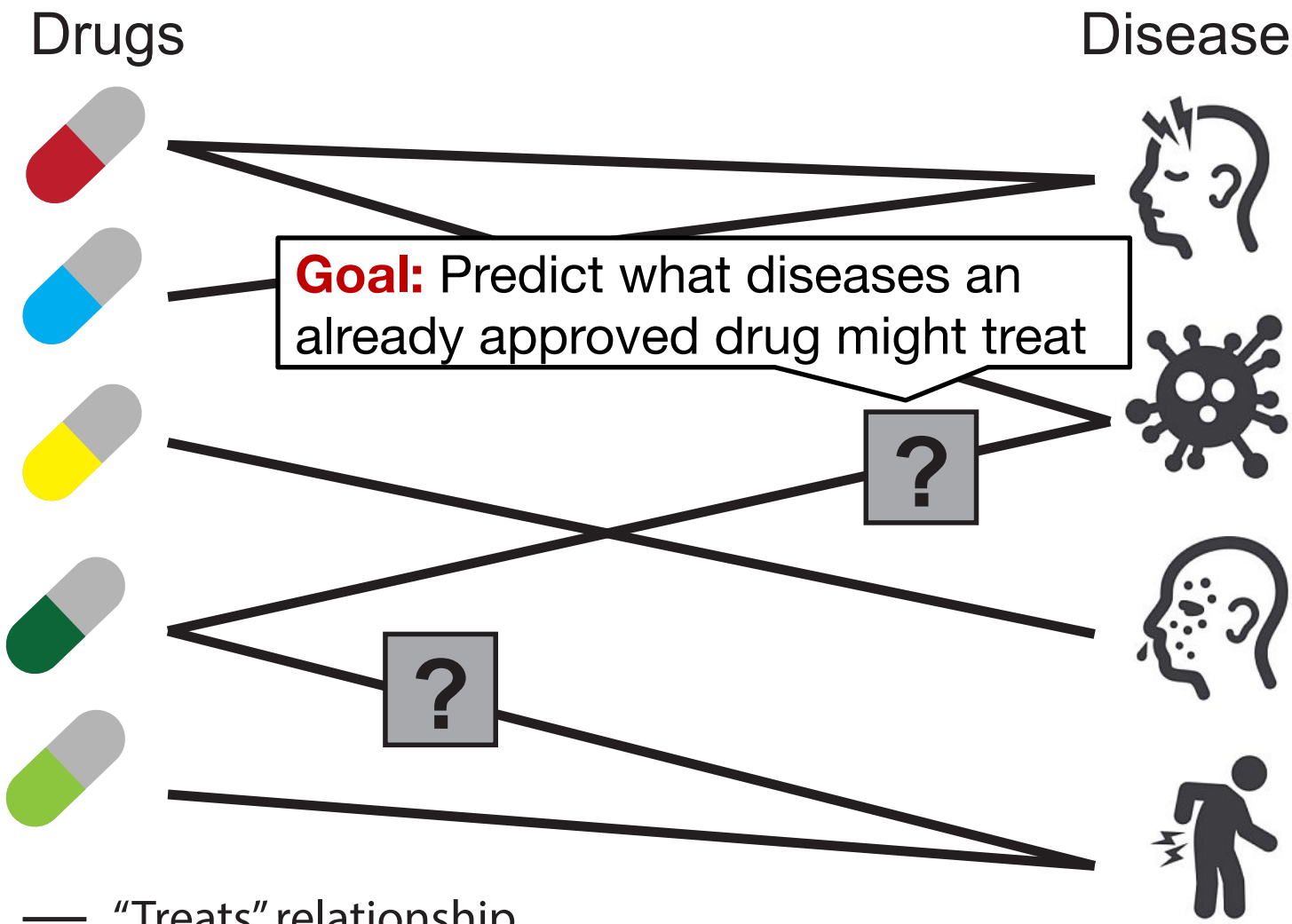
A SHORTER TIMESCALE

Because most repositioned drugs have already passed the early phases of development and clinical testing, they can potentially win approval in less than half the time and at one-quarter of the cost.

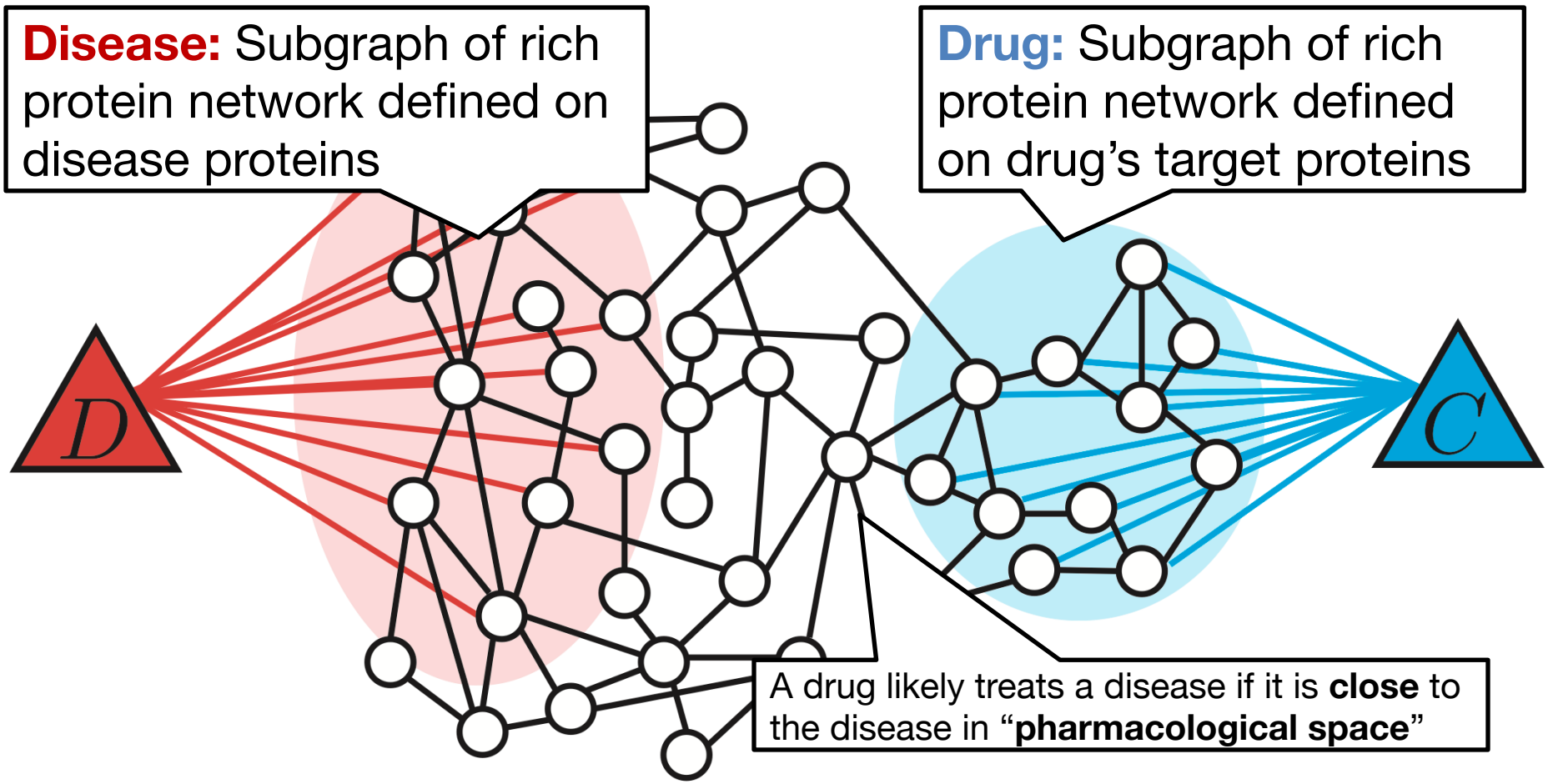
~6 years, **~$300 million**

12–16 years, **~$1 billion to $2 billion**

What drug treats what disease?



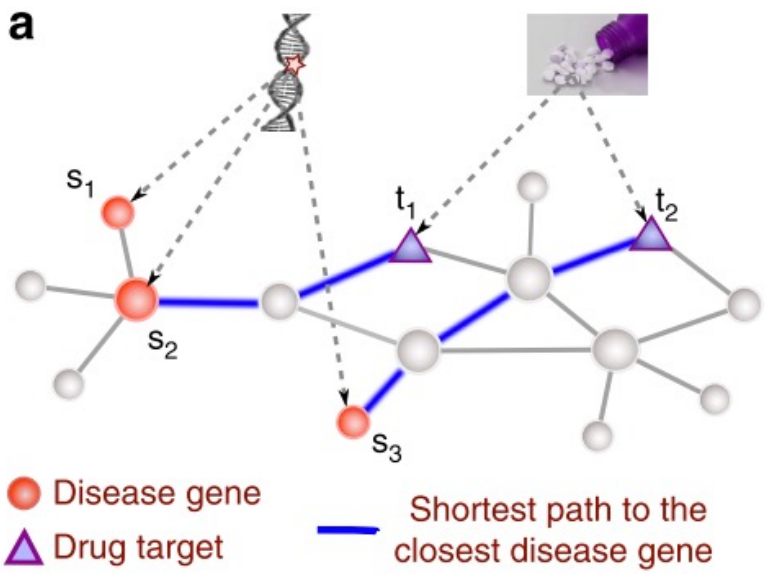
Key insight: subgraphs

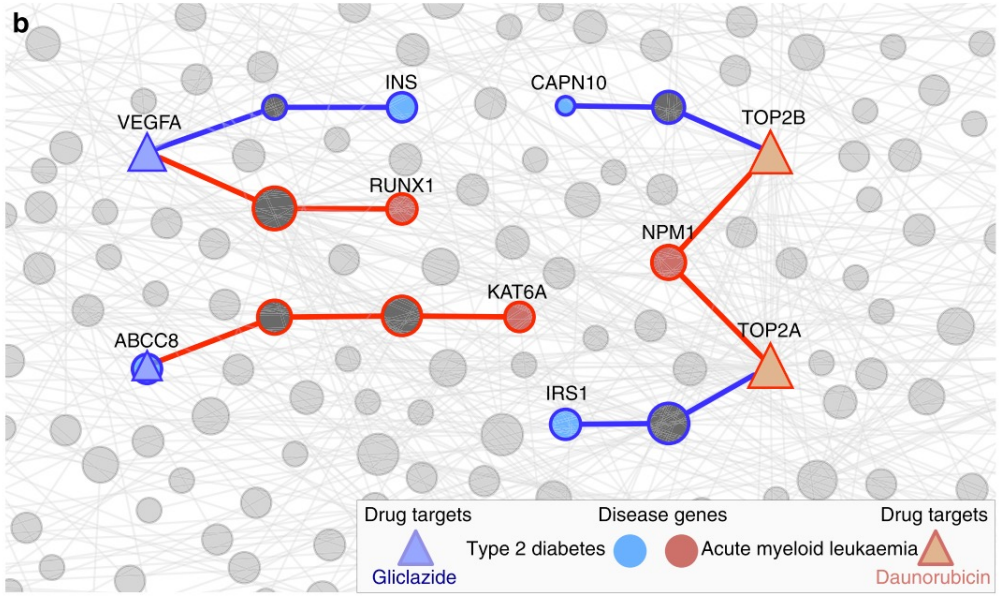


Idea: Use the paradigm of embeddings to operationalize the concept of closeness in pharmacological space

Why subgraphs? – Part #1

- Analysis of 238 drugs used in 78 diseases
- **Key result:** Therapeutic effect of drugs is localized in a small network neighborhood of disease genes

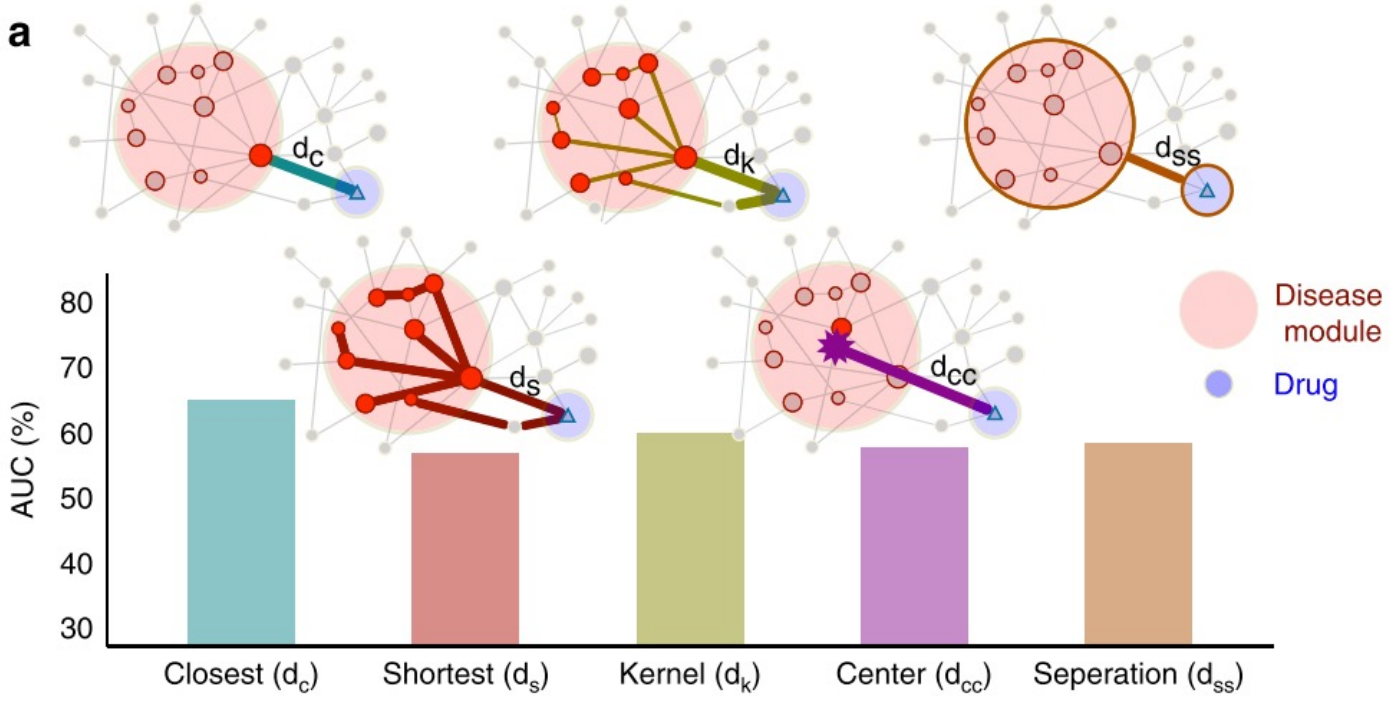




Guney, E., Network-based in silico drug efficacy screening. Nature Communications, 2016

Why subgraphs? – Part #2

- Analysis of 238 drugs used in 78 diseases
- **Key result:** Therapeutic effect of drugs is localized in a small network neighborhood of disease genes



Guney, E., Network-based in silico drug efficacy screening. Nature Communications, 2016

Why subgraphs? – Part #3

- Analysis of 238 drugs used in 78 diseases
- **Key result:** Therapeutic effect
  in a small network neighborho

Negative *z*-values: **Drug targets are close (i.e., proximal) to disease genes** in the PPI network → Successful repurposing

Positive *z*-values: **Drug targets are far away (i.e., not proximal) from disease genes** in the PPI network → Drug failure due to lack of efficacy

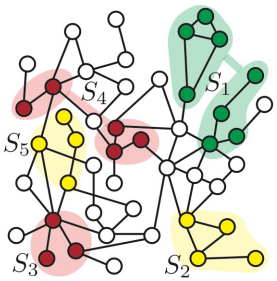
Table 1 | Proximity values for several repurposed and failed drugs.

		Phenotype	Proximity (z)
		Non-Hodgkin's lymphoma	−2.4
		Restless legs syndrome	−1.1
	ction	Erectile dysfunction	−1.0
	ial	Endometrial cancer	−1.1
	norgestrel	Confer protection against endometrial cancer	Endometrial cancer	−1.6
Failures due to lack of efficacy			
Tabalumab	Showed lack of efficacy for systemic lupus erythematosus	Systemic lupus erythematosus	1.8
Preladenant	Discontinued trials for Parkinson due to lack of improvement compared with placebo	Parkinson's disease	0.2
Iniparib	Failed to achieve improvement while being tested for squamous non-small-cell lung cancer	Squamous cell cancer	0.0
Failures due to adverse effetcs			
Semagacestat	Failed trials due to worsening AD	AD	−5.6
Terfenadine	Withdrawn due to inducing cardiac arrhythmia	Cardiac arrhythmia	−2.2
		Arrhythmia (side effect)	−2.6

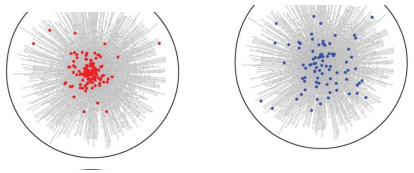
Guney, E., Network-based in silico drug efficacy screening. Nature Communications, 2016

Why are subgraphs (drug and disease modules) challenging?

- Need to predict over structures of varying size:
  - How to represent subgraphs that are not k-hop neighborhoods?

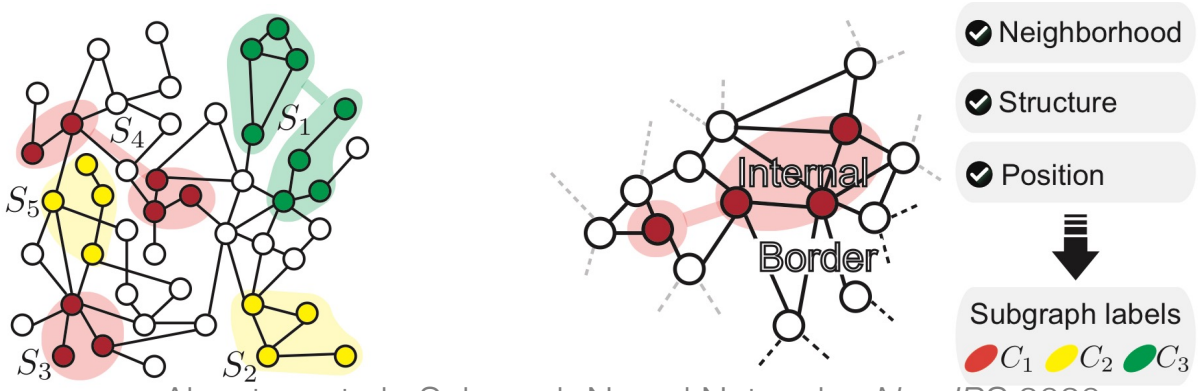


- Rich connectivity patterns, both internally a externally through interactions with the rest of G:
  - How to inject this information into a GNN?
- Subgraphs can be:
  - Localized and reside in our region of the graph
  - Distributed across multiple local neighborhoods



Problem formulation

- **Goal:** Learn subgraph embeddings such that the likelihood of preserving subgraph topology is maximized in the embedding space
  - S_i and S_j with similar subgraph topology should be embedded close together in the embedding space
- SubGNN: Representation learning framework for all key properties of subgraph topology



Alsentzer et al., Subgraph Neural Networks, *NeurIPS* 2020

SubGNN: Overview

- SubGNN: Representation learning framework for all key properties of subgraph topology
- Two key parts:
  - **Part 1:** Hierarchical propagation of information in G:
    - Propagate messages from anchor patches to subgraphs
    - Aggregate messages into a final subgraph embedding
  - **Part 2:** Routing of messages through 3 channels, each capturing a distinct property of subgraph topology: position, neighborhood, and structure channels

Alsentzer et al., Subgraph Neural Networks, *NeurIPS* 2020

Part 1: Neural message passing

- Property x-specific messages m_x are propagated from anchor patch A_x^q to subgraph component S_i^c
- Anchor patches are helper subgraphs randomly sampled from G; patches A_P, A_N, and A_S for position, neighborhood and structure

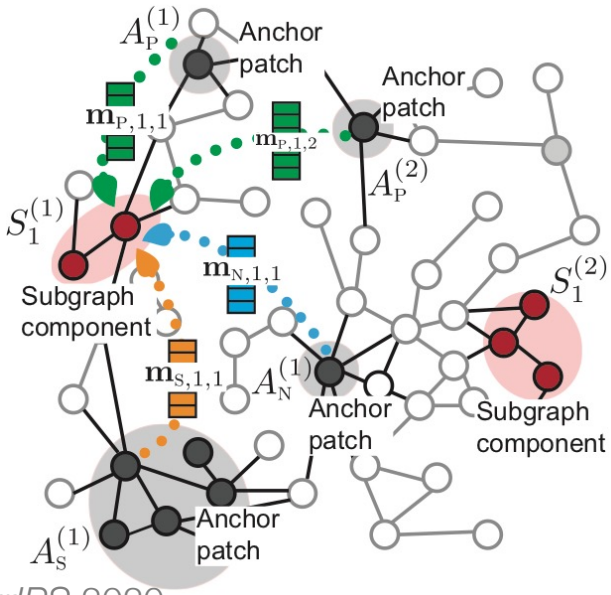
similarity function between a subgraph component and an anchor patch

$$\mathbf{MSG}_{\mathrm{X}} = \boxed{\gamma_{\mathrm{X}}}\left(S^{(C)}, A_{\mathrm{X}}\right) \cdot p_{\mathrm{X}}$$

$$\mathbf{a}_{\mathrm{X},c} = \mathbf{AGG}_M\left(\left\{\mathbf{MSG}_{\mathrm{X}}(S^{(C)}, A_{\mathrm{X}}, p_{\mathrm{X}}), \forall A_{\mathrm{X}} \in \mathcal{A}_{\mathrm{X}}\right\}\right),$$

$$\boxed{\mathbf{h}_{\mathrm{X},c}^{(l)}} = \sigma\left(\mathbf{W}_h \cdot [\mathbf{a}_{\mathrm{X},c}; \mathbf{h}_{\mathrm{X},c}^{(l-1)}]\right),$$

property-specific representation of a subgraph component; passed to the next layer

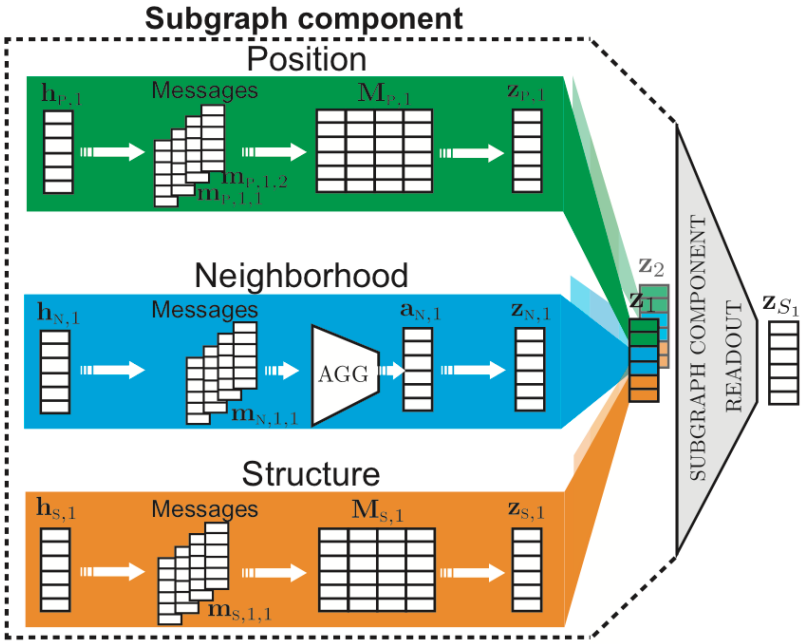


Alsentzer et al., Subgraph Neural Networks, *NeurIPS* 2020

Part 2: Property-aware routing

- SubGNN specifies three channels, each designed to capture a distinct subgraph property
  - Position, neighborhood, and structure

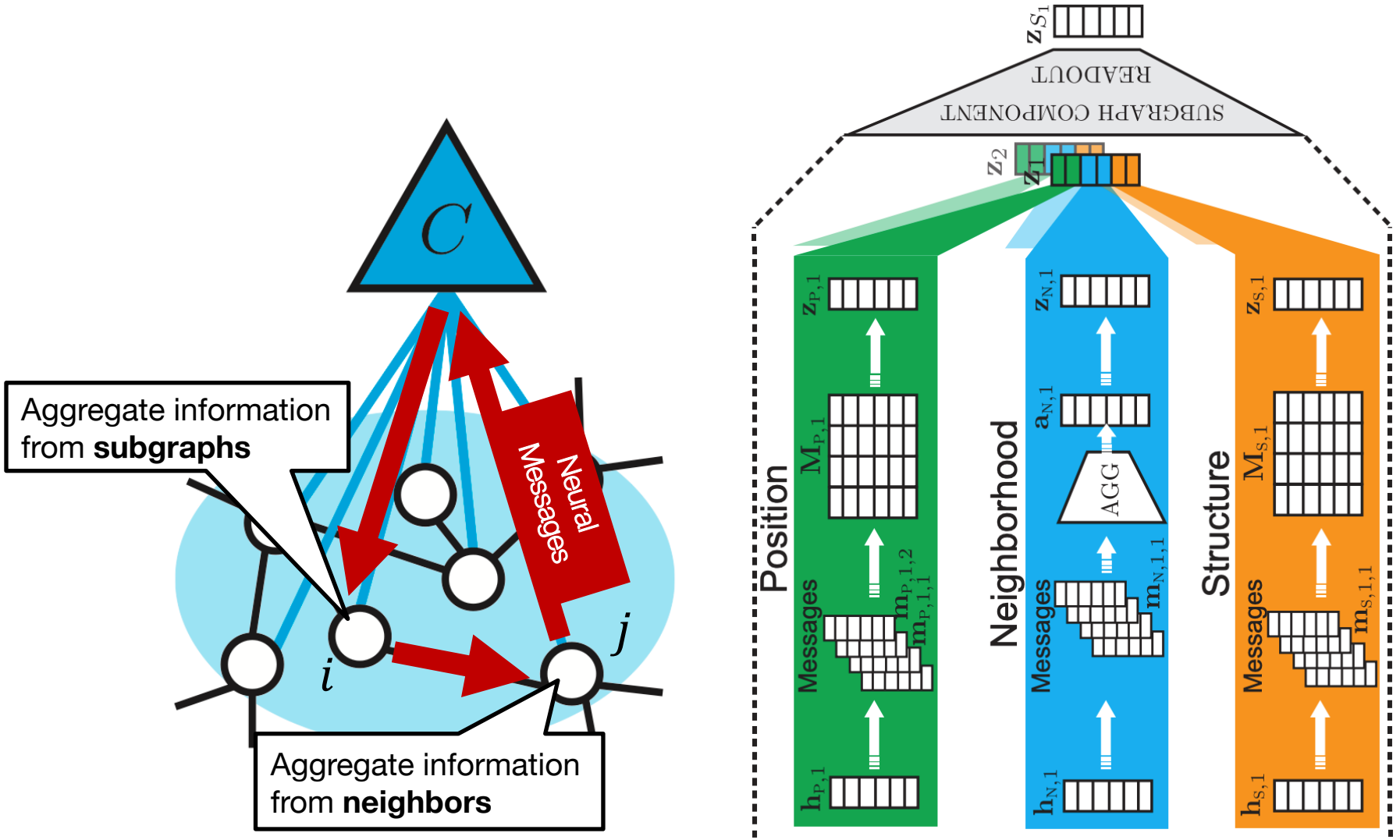
- Channel x has three key elements:
  - Similarity function γ_x to weight messages sent between anchor patches and subgraph components
  - Sampling function φ_x to generate anchor patches
  - Anchor patch encoder ψ_x



Channel outputs $\boldsymbol{z}_x$ are concatenated to produce a final subgraph representation $\boldsymbol{z}_S$

Alsentzer et al., Subgraph Neural Networks, *NeurIPS* 2020

SubGNN: Overview



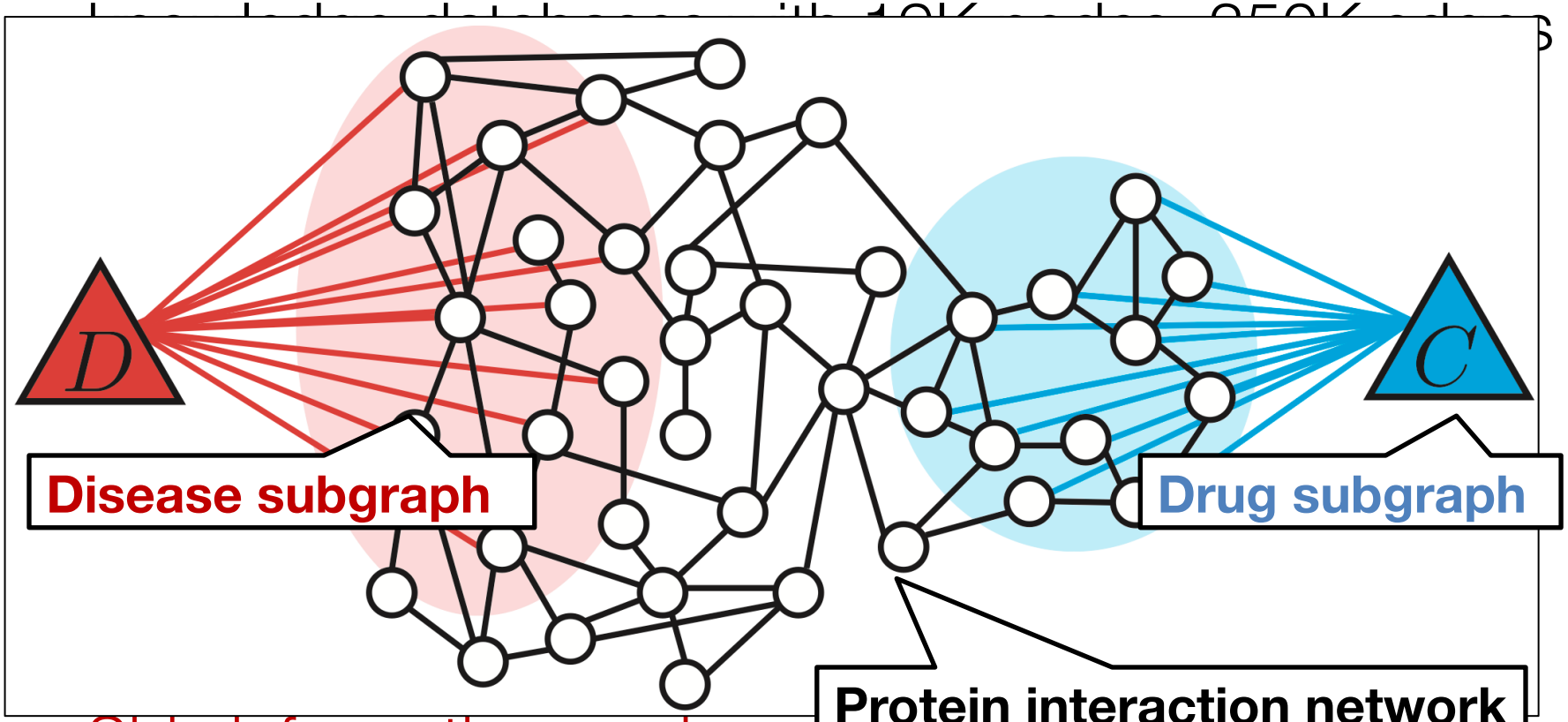
Alsentzer et al., Subgraph Neural Networks, *NeurIPS* 2020

Setup: Drug repurposing dataset

- Protein-protein interaction network culled from 15 knowledge databases with 19K nodes, 350K edges
- Drug-protein and disease-protein links:
  - DrugBank, OMIM, DisGeNET, STITCH DB and others
  - 20K drug-protein links, 560K disease-protein links
- Medical indications and contra-indications:
  - DrugBank, MEDI-HPS, DailyMed, Drug Central, RepoDB
  - 6K drug-disease indications
- Side information on drugs, diseases, proteins, etc.:
  - Molecular pathways, disease symptoms, side effects

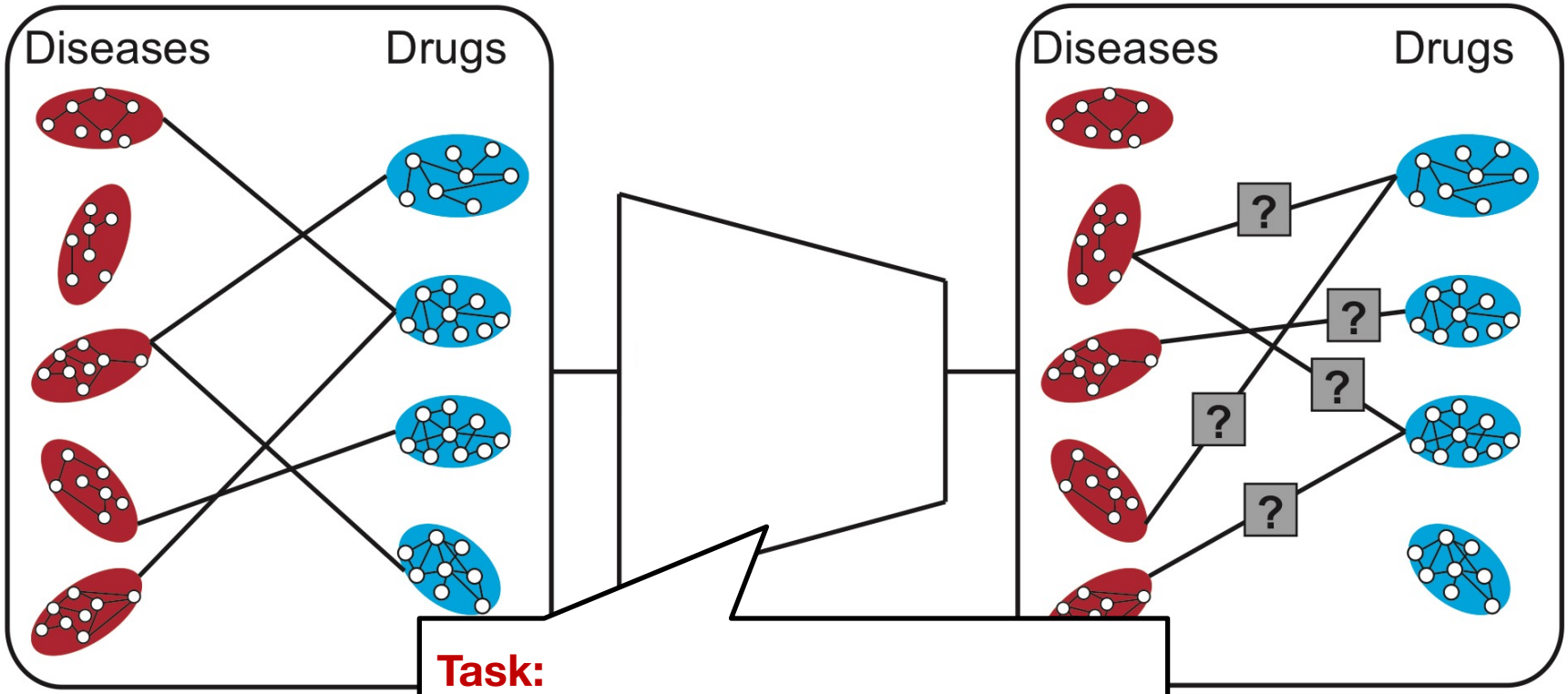
Setup: Drug repurposing dataset

- Protein-protein interaction network culled from 15



- Side information on drugs, diseases, proteins, etc..
  - Molecular pathways, disease symptoms, side effects

Predict links between drug and disease subgraphs

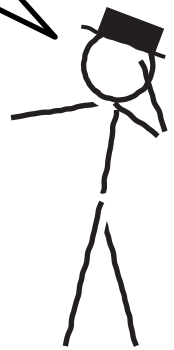


Task:
1) Learn embeddings for C's and D's subgraphs
2) Predict whether C should be **indicated** or **contra-indicated** for D

Results: Drug repurposing

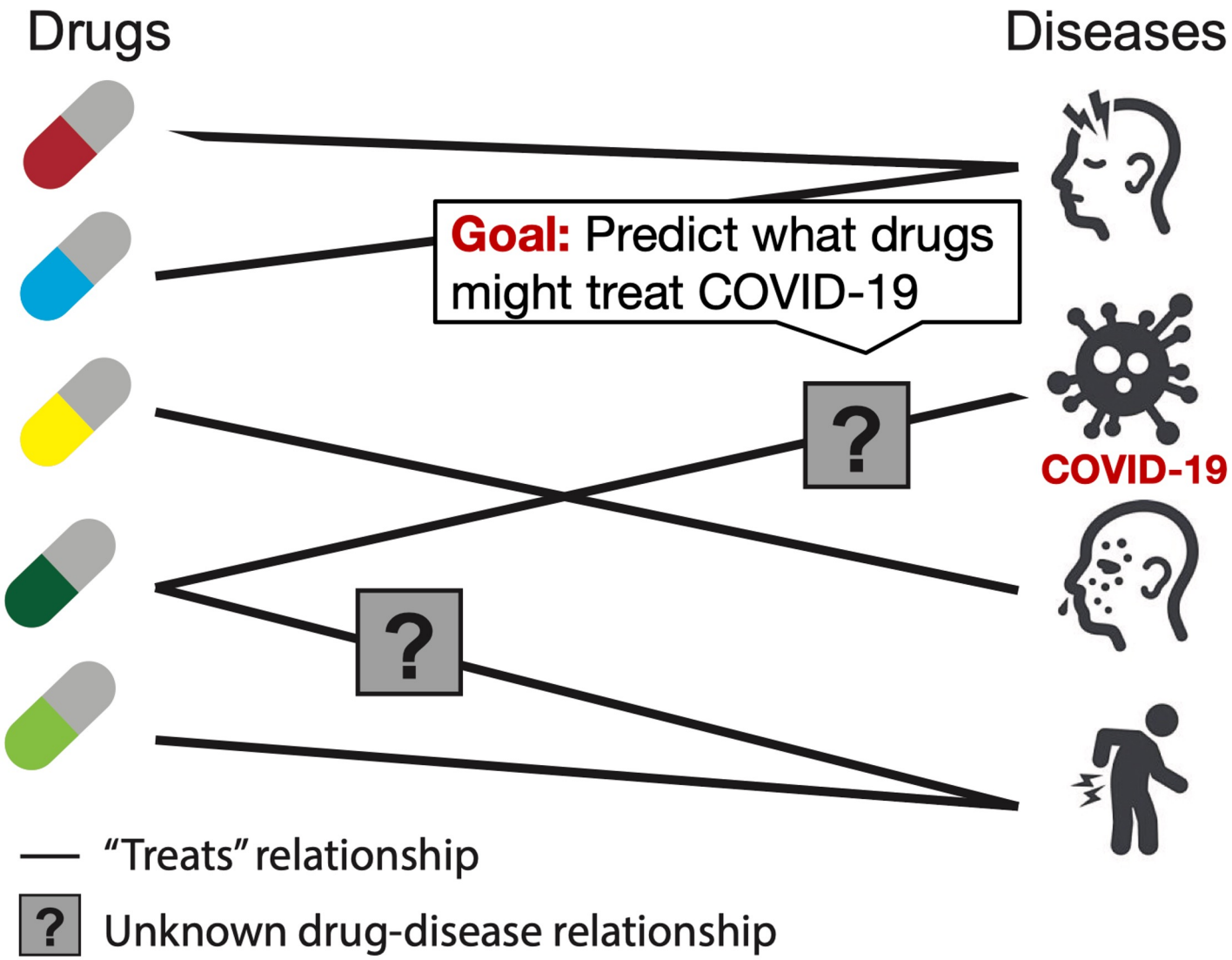
Stanford MEDICINE | SPARK Translational Research Program
From Bench to Bedside

Task: Predict if an existing drug can be repurposed for a new disease



Drug	Disease	
N-acetyl-cysteine	cystic fibrosis	
Xamoterol	neurodegenerat	
Plerixafor	cancer	
Sodium selenite	cancer	Rank: 36/5000
Ebselen	C difficile	Rank: 10/5000
Itraconazole	cancer	Rank: 26/5000
Bestatin	lymphedema	Rank: 11/5000
Bestatin	pulmonary arterial hypertension	Rank: 16/5000
Ketaprofen	lymphedema	Rank: 28/5000
Sildenafil	lymphatic malformation	Rank: 26/5000
Tacrolimus	pulmonary arterial hypertension	Rank: 46/5000
Benzamil	psoriasis	Rank: 114/5000
Carvedilol	Chagas' disease	Rank: 9/5000
Benserazide	BRCA1 cancer	Rank: 41/5000
Pioglitazone	interstitial cystitis	Rank: 13/5000
Sirolimus	dystrophic epidermolysis bullosa	Rank: 46/5000

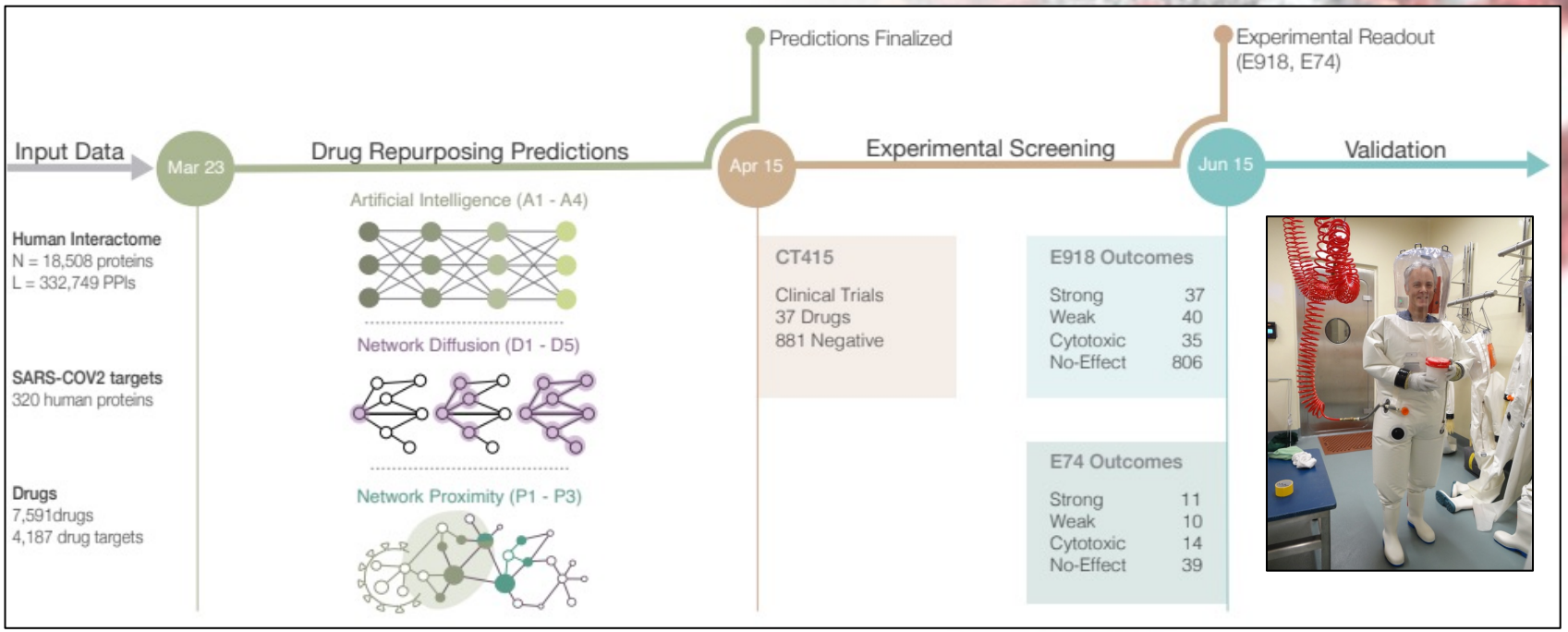
Emerging pathogens



Emerging pathogens

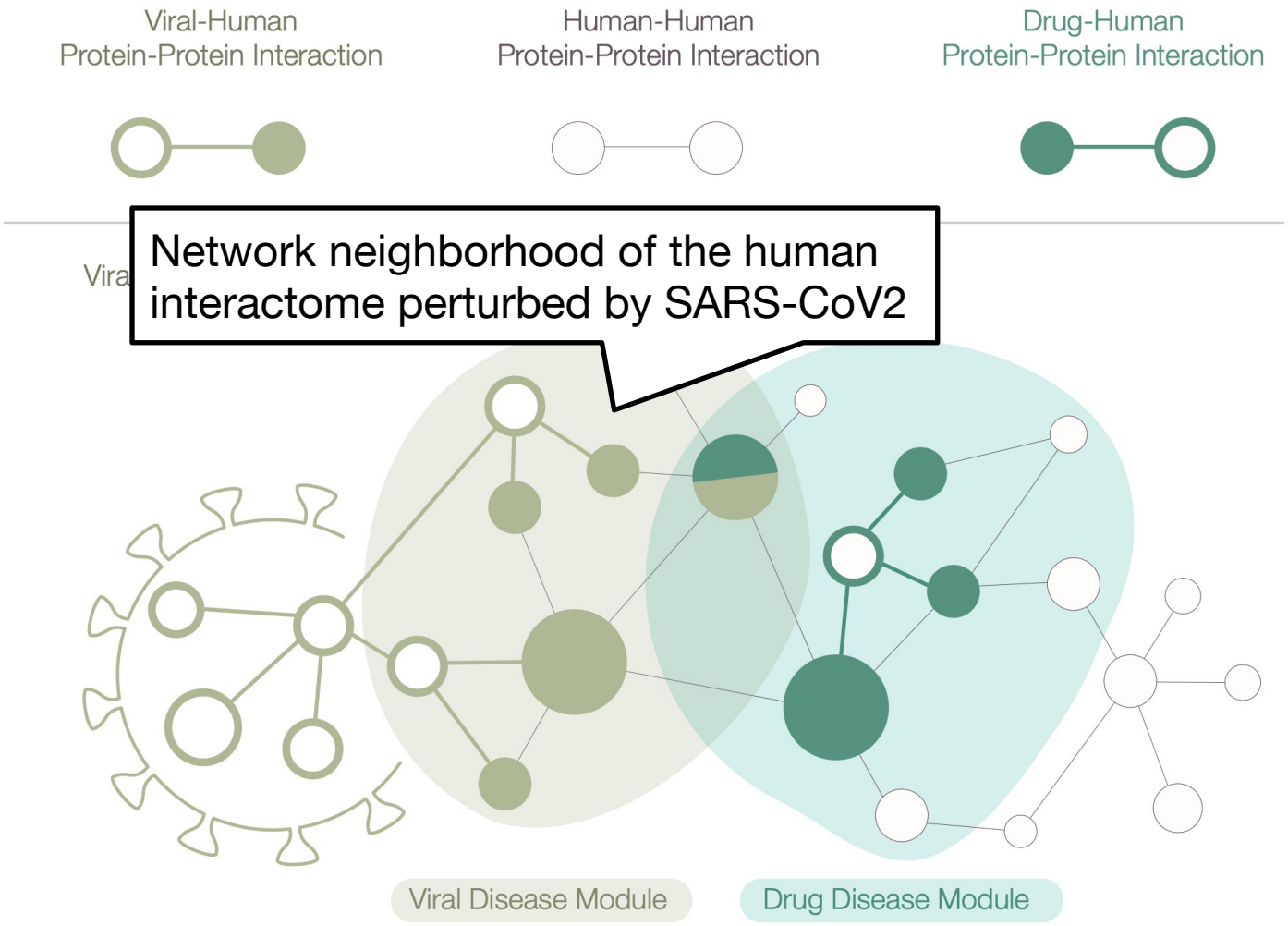
The traditional approach of iterative development, experimental testing, clinical validation, and approval of new drugs are not feasible

A more realistic strategy relies on drug repurposing, requiring us to identify clinically approved drugs that have a therapeutic effect in COVID-19 patients



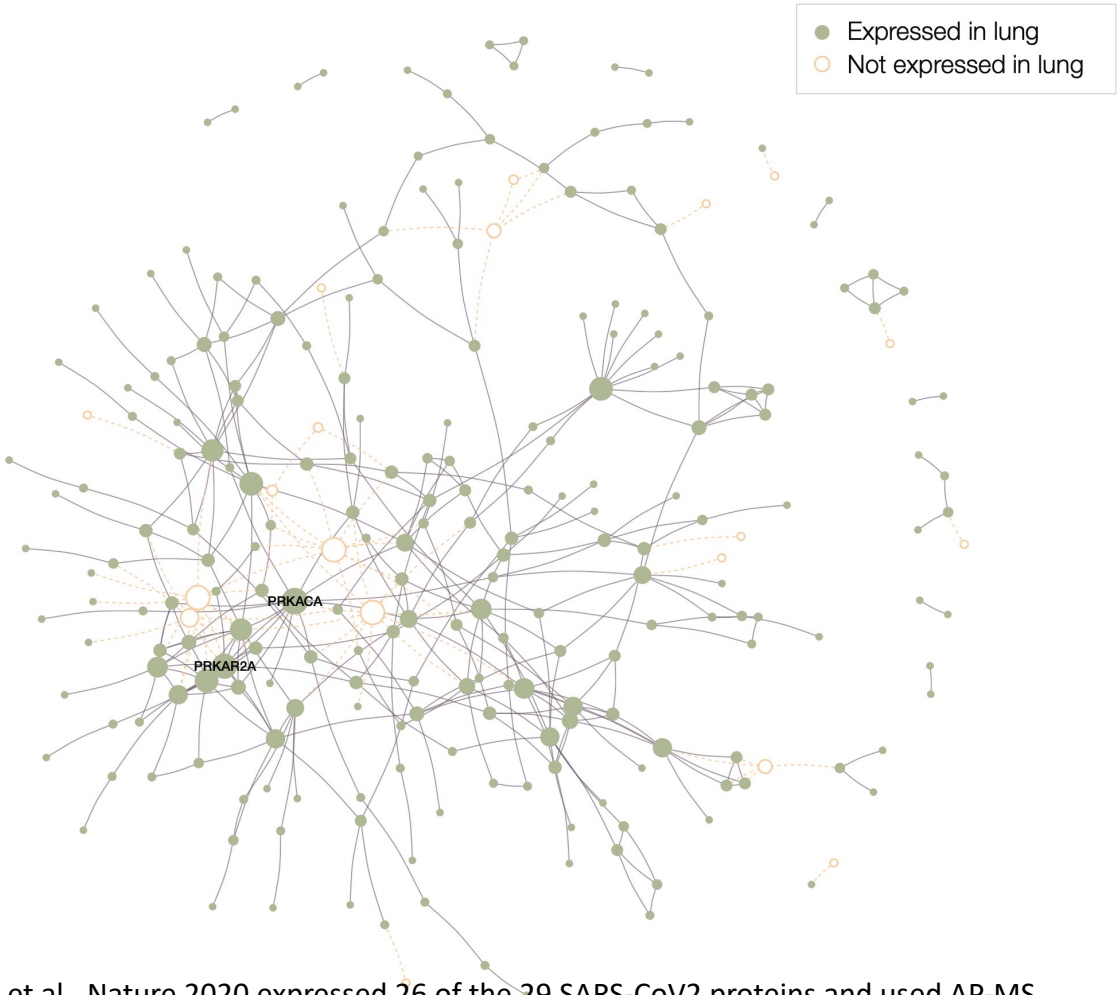
Network Medicine Framework for Identifying Drug Repurposing Opportunities for Covid-19, *PNAS* 2021

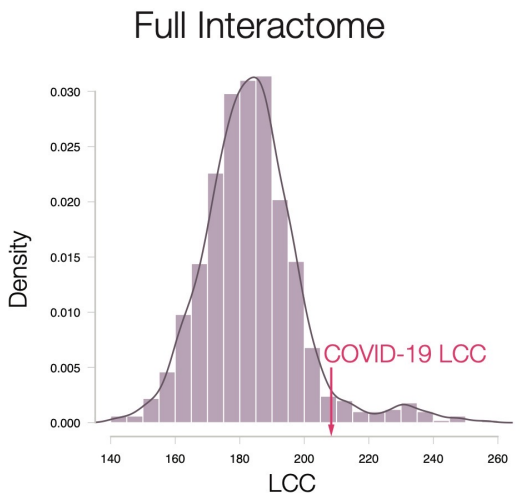
How to represent COVID-19? Map SARS-CoV2 targets to the human interactome

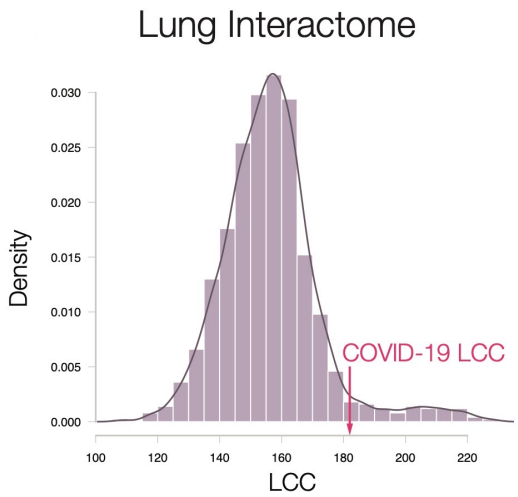


Network Medicine Framework for Identifying Drug Repurposing Opportunities for Covid-19, *PNAS* 2021

COVID-19 disease module







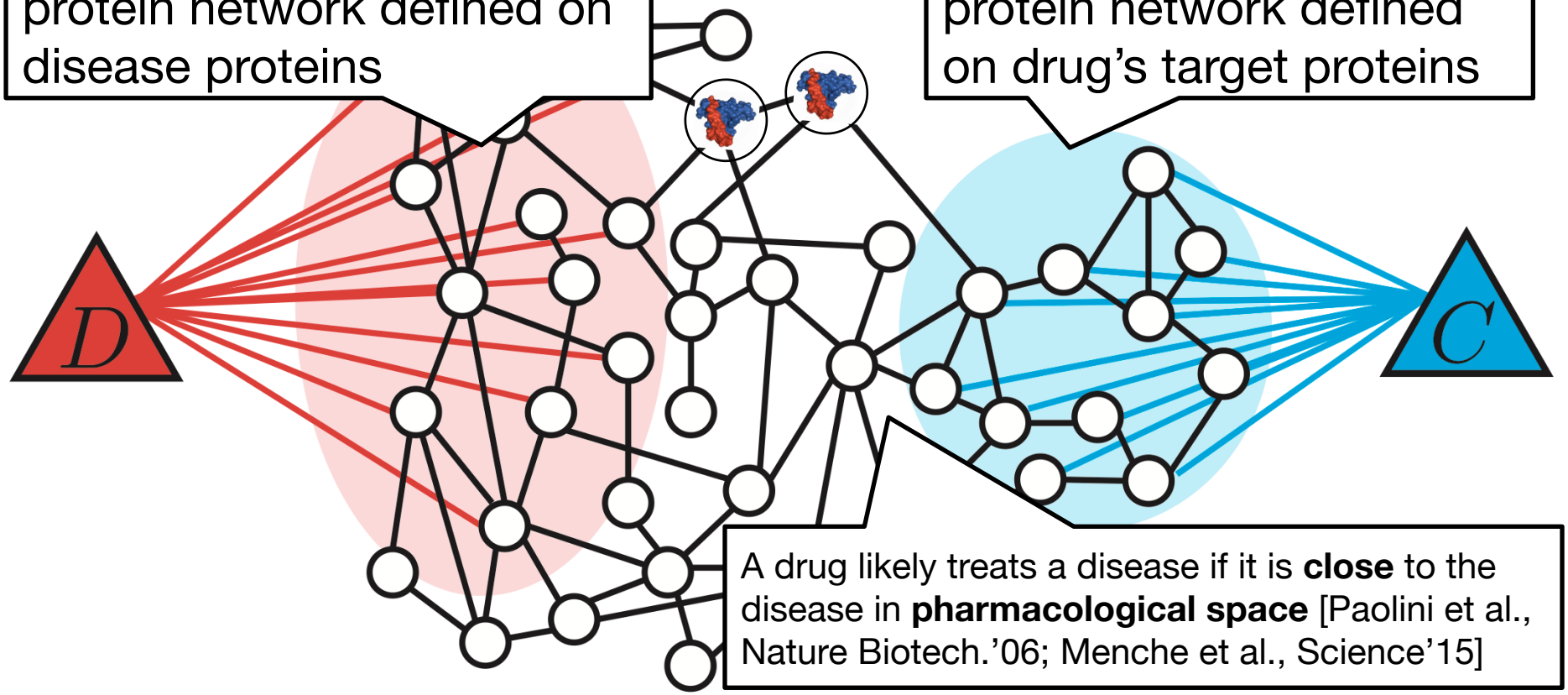
Gordon et al., Nature 2020 expressed 26 of the 29 SARS-CoV2 proteins and used AP-MS to identify 332 human proteins to which viral proteins bind

Network Medicine Framework for Identifying Drug Repurposing Opportunities for Covid-19, *PNAS* 2021

Key Insight: subgraphs

Disease: Subgraph of rich protein network defined on disease proteins

Drug: Subgraph of rich protein network defined on drug's target proteins



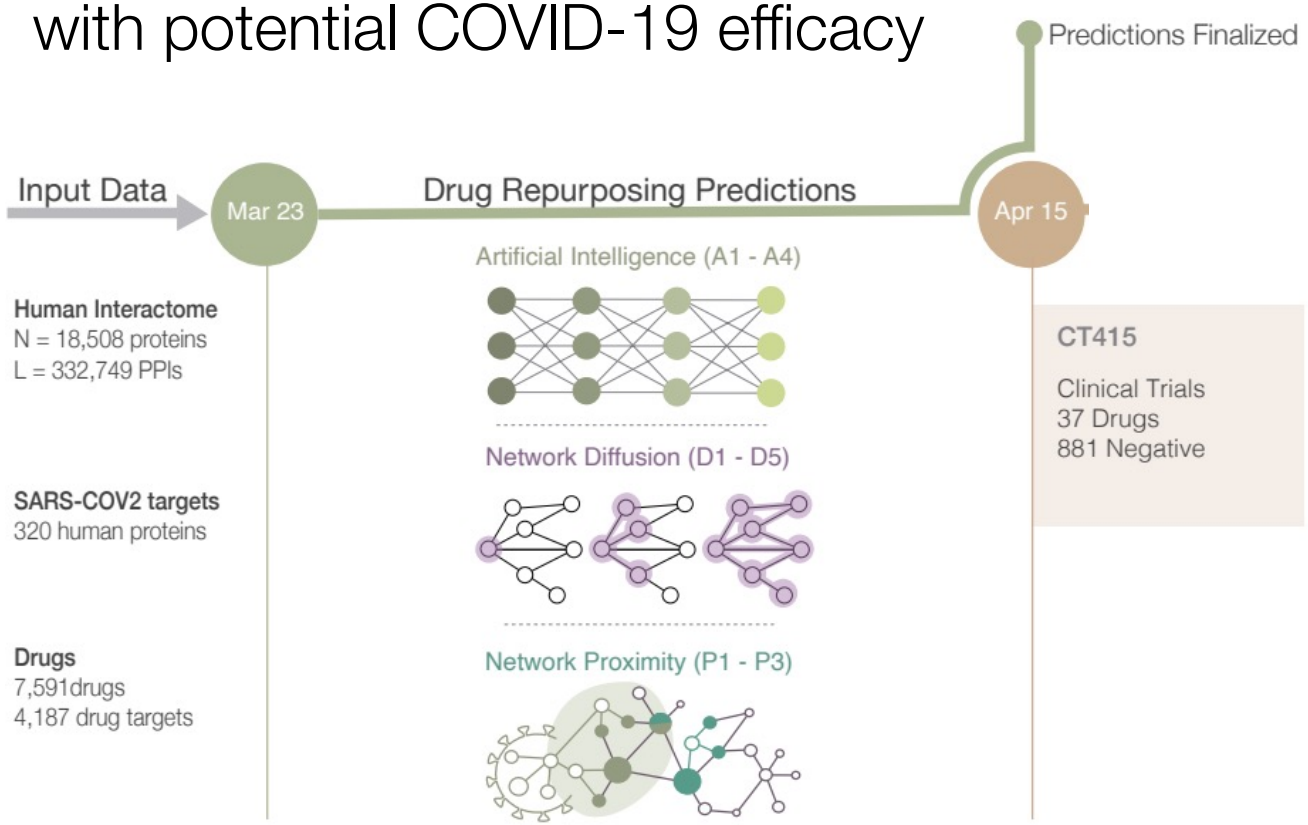
A drug likely treats a disease if it is **close** to the disease in **pharmacological space** [Paolini et al., Nature Biotech.'06; Menche et al., Science'15]

Idea: Use the paradigm of embeddings to operationalize the concept of closeness in pharmacological space

Network Medicine Framework for Identifying Drug Repurposing Opportunities for Covid-19, *PNAS* 2021

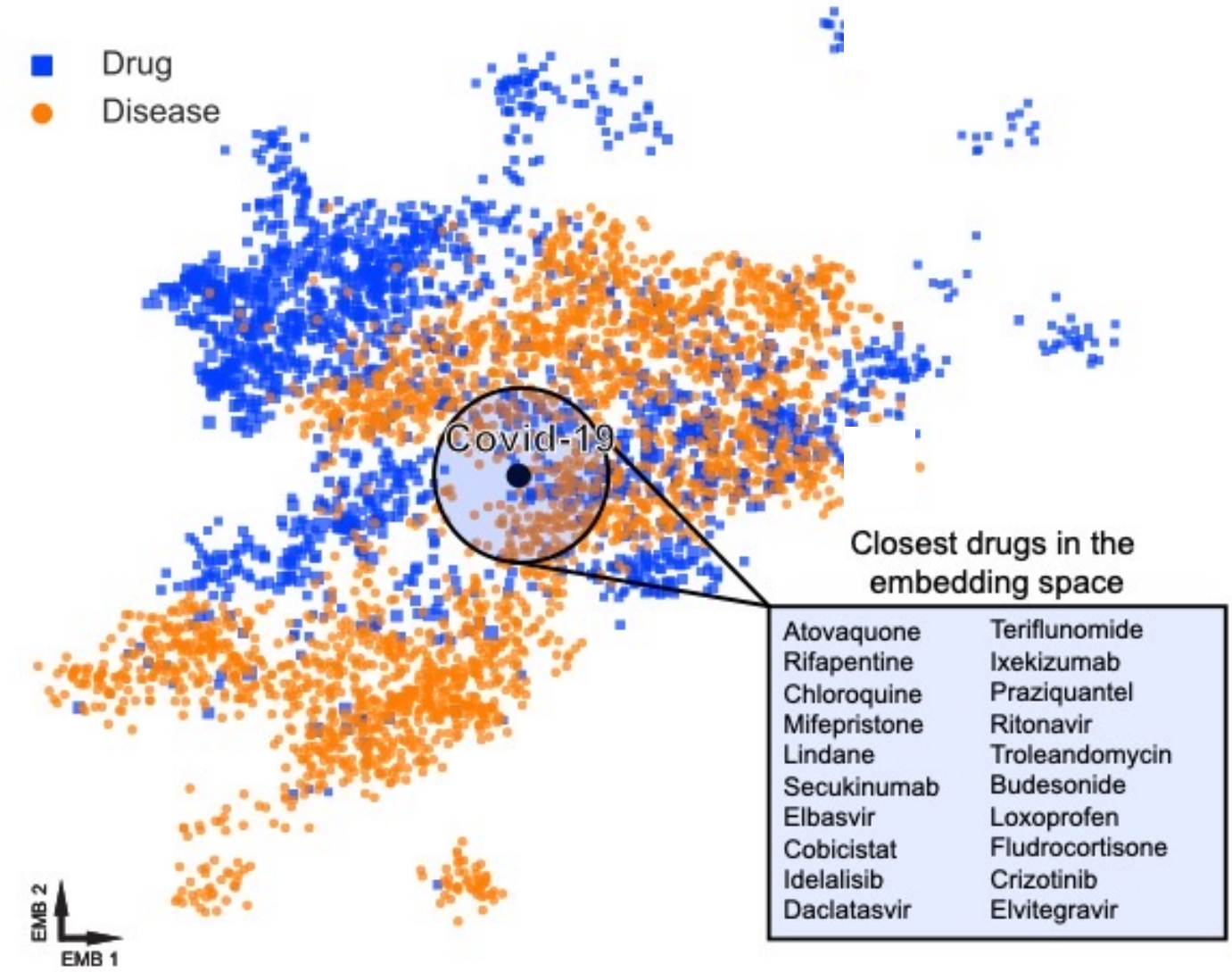
Computational setup

- Proxy for ground-truth information:
  - Monitor drugs under clinical trials
  - Capture the medical community's assessment of drugs with potential COVID-19 efficacy



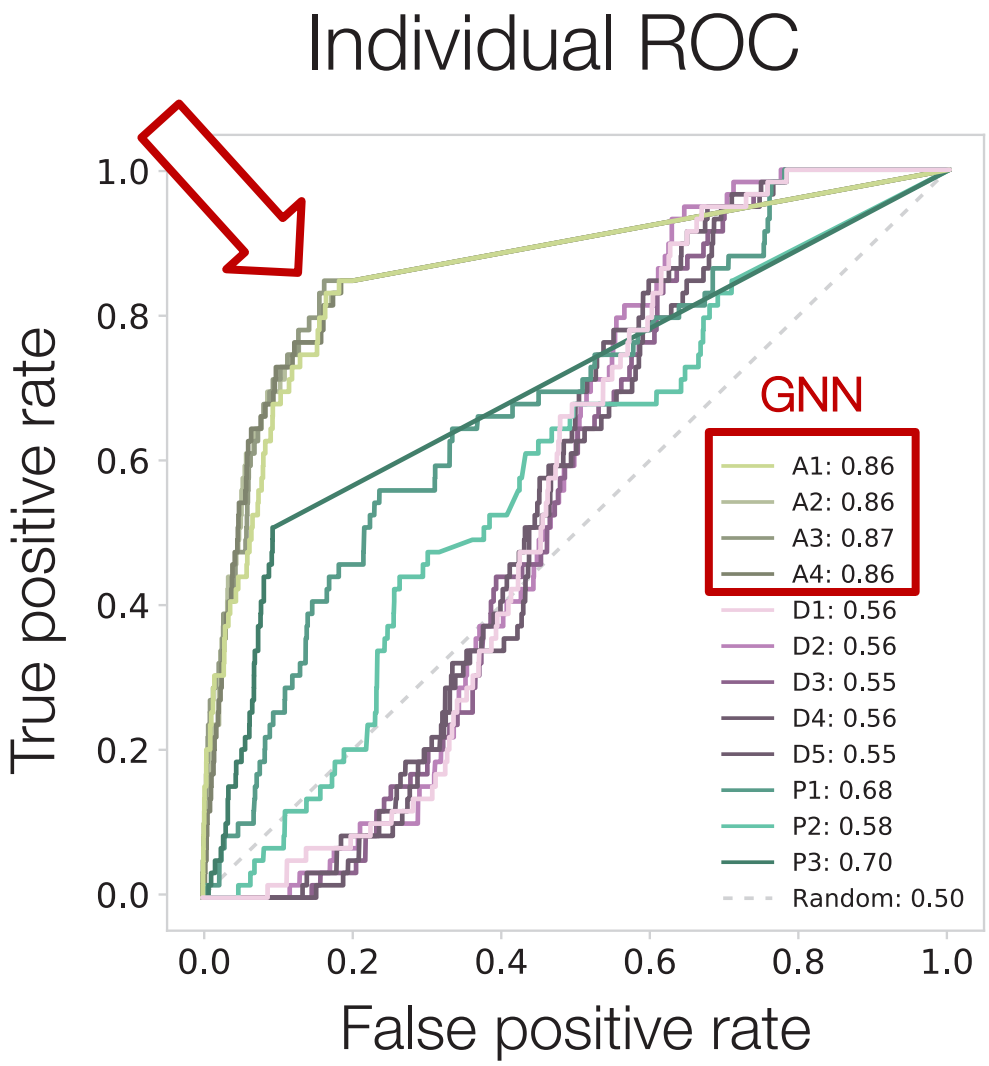
Network Medicine Framework for Identifying Drug Repurposing Opportunities for Covid-19, *PNAS* 2021

Embedding space



Network Medicine Framework for Identifying Drug Repurposing Opportunities for Covid-19, *PNAS* 2021

Results: COVID-19 Repurposing



We test each pipeline's ability to recover drugs currently in clinical trials for COVID-19

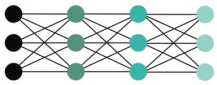
The best individual ROC curves are obtained by the GNN methods

The second-best performance is provided by the proximity P3. Close behind is P1 with AUC = 0.68 and AUC = 0.58

Diffusion methods offer ROC between 0.55-0.56

Network Medicine Framework for Identifying Drug Repurposing Opportunities for Covid-19, *PNAS* 2021

Final Prediction Model – Part #1

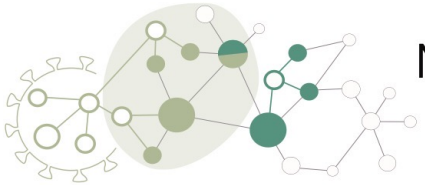
Input Data	Methods	Outcomes
Human Interactome N = 18,508 proteins L = 332,749 PPIs	 Network Proximity 3 pipelines	Infected Tissues/Organs
SARS-COV2 targets 320 human proteins Gordon et al, 2020	 Network Diffusion 5 pipelines	Comorbidity
Drug Targets 7,591drugs 4,187 drug targets DrugBank	 AI Prioritization 4 pipelines	Drug Repurposing & Validation

Network Medicine Framework for Identifying Drug Repurposing Opportunities for Covid-19, *PNAS* 2021

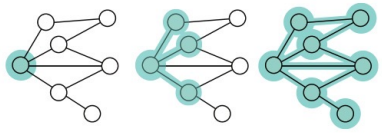
Final Prediction Model – Part #2

- A COVID-19 treatment can not be derived from the arsenal of therapies approved for specific diseases
- Repurposing strategies focus on drugs previously approved for other pathogens, or on drugs that target the human proteins to which viral proteins bind.
- Most approved drugs do not target directly disease proteins but bind to proteins in their network vicinity
- [Yildirim, Nature Biotech. 2007]
- Identify drug candidates that have the potential to perturb the network vicinity of the COVID-19 disease module.
- Implement 3 Network Repurposing Methods.

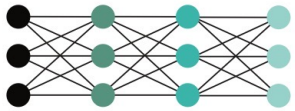
Methods



Network Proximity
3 pipelines



Network Diffusion
5 pipelines



AI Prioritization
4 pipelines

Network Medicine Framework for Identifying Drug Repurposing Opportunities for Covid-19, *PNAS* 2021

Final Prediction Model – Part #3

Rank Aggregation Algorithm: Maximize the number of pairwise agreements between the final ranking and each input ranking.

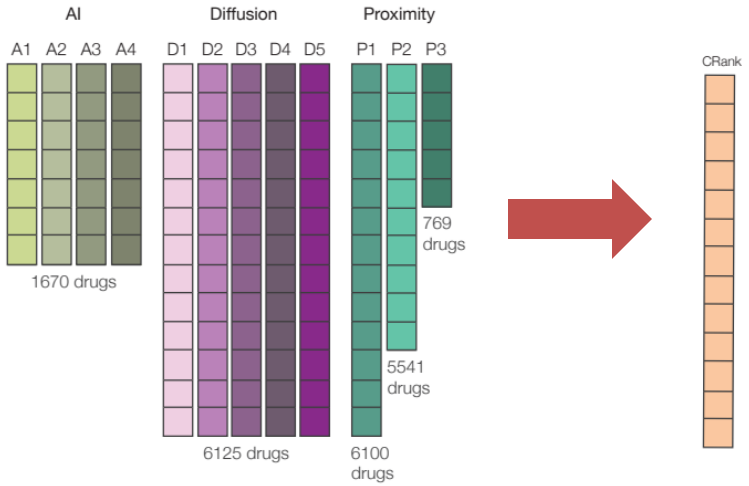
The combined performance of the AI methods is 0.87, the same as A3.

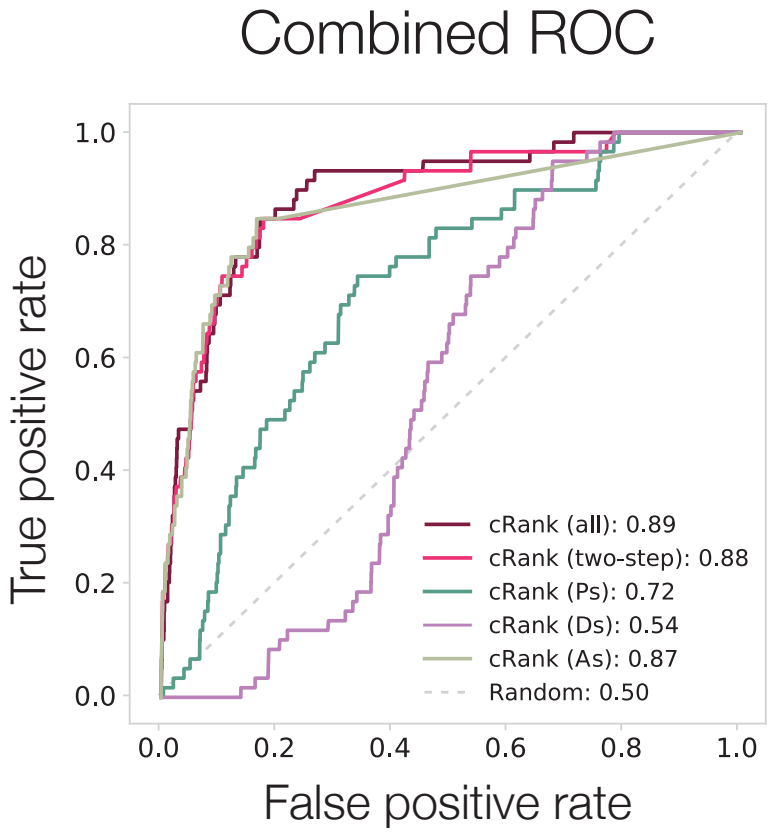
Improvement for proximity pipelines: 0.70 → 0.72.

Combined diffusion pipelines have lower performance (0.54 vs 0.56, for D1, D2, and D4).

Combining all 12 pipelines, gives AUROC=0.89, the highest of any individual or combination-based pipelines,

Individual pipelines offer complementary information harnessed by the combined ranking.





Predicted Drug Candidates

◯ # of Clinical trials from ClinicalTrials.gov

Joseph Loscalzo

86 drugs selected from the top 10% of the rank list.

Respiratory drugs (e.g., theophylline, montelukast).

Cardiovascular systems (e.g., verapamil, atorvastatin).

Antibiotics used to treat viral (e.g., ribavirin, lopinavir), parasitic (e.g., hydroxychloroquine, ivermectin, praziquantel), bacterial (e.g., rifaximin, sulfanilamide), mycotic (e.g., fluconazole), and mycobacterial (e.g., isoniazid) infections.

Immunomodulating/anti-inflammatory drugs (e.g., interferon-β, auranofin, montelukast, colchicine)

Anti-proteasomal drugs (e.g., bortezomib, carfilzomib)

Less obvious choices: aminoglutethimide, melatonin, levothyroxine, calcitriol, selegiline, deferoxamine, mitoxantrone, metformin, nintedanib, cinacalcet, and sildenafil.

	Drug	C-rank
(20)	Ritonavir	1
	Isoniazid	2
	Troleandomycin	3
	Cilostazol	4
(76)	Chloroquine	5
	Rifabutin	6
	Flutamide	7
(2)	Dexamethasone	8
	Rifaximin	9
	Azelastine	10
	Folic Acid	16
	Rabeprazole	27
	Methotrexate	32
	Digoxin	33
	Theophylline	34
	Fluconazole	41
	Aminoglutethimide	42
(67)	Hydroxychloroquine	44
	Methimazole	47
(1)	Ribavirin	49
(1)	Omeprazole	50
	Bortezomib	53
	Leflunomide	54
	Dimethylfumarate	55
(4)	Colchicine	57
	Quercetin	63
	Mebendazole	67

	Drug	C-rank
	Mesalazine	69
	Pentamidine	92
	Verapamil	98
	Melatonin	109
	Griseofulvin	112
	Auranofin	118
(1)	Atovaquone	124
	Montelukast	131
	Romidepsin	138
(1)	Cobicistat	141
(17)	Lopinavir	146
	Pomalidomide	155
	Sulfinpyrazone	157
(1)	Levamisole	161
	Calcitriol	164
(1)	Interferon-β-1a	173
	Praziquantel	176
(1)	Ascorbic acid	195
	Fluvastatin	199
(1)	Interferon-β-1b	203
	Selegiline	206
(1)	Deferoxamine	227
	Ivermectin	235
(1)	Atorvastatin	243
	Mitoxantrone	250
	Glyburide	259
(2)	Thalidomide	262

	Drug	C-rank
	Sulfanilamide	265
	Hydralazine	269
	Gemfibrozil	281
(4)	Ruxolitinib	284
	Propranolol	297
	Carbamazepine	301
	Doxorubicin	309
	Levothyroxine	329
	Dactinomycin	335
	Tenofivir	338
	Tadalafil	339
	Doxazosin	367
	Rosiglitazone	397
	Aminolevulinic acid	398
	Nitroglycerin	418
	Metformin	457
(1)	Nintedanib	466
	Allopurinol	471
	Ponatinib	491
(1)	Sildenafil	493
	Dapagliflozin	504
	Nitroprusside	515
	Cinacalcet	553
	Mexiletine	559
	Sitagliptin	706
	Carfilzomib	765
(1)	Azithromycin	786

Network Medicine Framework for Identifying Drug Repurposing Opportunities for Covid-19, *PNAS* 2021

Experimental validation of predictions

National Emerging Infectious Diseases Laboratories (NEIDL)

CRank	Drug Name
1	Ritonavir
2	Isoniazid
3	Troleandomycin
4	Cilostazol
5	Chloroquine
6	Rifabutin
7	Flutamide
8	Dexamethasone
9	Rifaximin
10	Azelastine
11	Crizotinib

17	Celecoxib
18	Betamethasone
19	Prednisolone
20	Mifepristone
21	Budesonide
22	Prednisone
23	Oxiconazole
24	Megestrol acetate
25	Idelalisib
26	Econazole

Ranked lists of drugs

New algorithms:
Prioritizing Network Communities, *Nature Communications* 2018
Subgraph Neural Networks, *NeurIPS* 2020
Graph Meta Learning via Local Subgraphs, *NeurIPS* 2020

Results: 918 compounds screened for their efficacy against SARS-CoV-2 in VeroE6 cells:

- **37 had a strong effect** being active over a broad range of concentrations
- **40 had a weak effect** on the virus
- An **order of magnitude higher hit rate** among top 100 drugs than **prior work**

Network Medicine Framework for Identifying Drug Repurposing Opportunities for Covid-19, *PNAS* 2021

Results: Network drugs

- 76/77 drugs that successfully reduced viral infection **do not bind** proteins targeted by SARS-CoV-2:
  - These drugs rely on **network-based actions** that cannot be identified by docking-based strategies

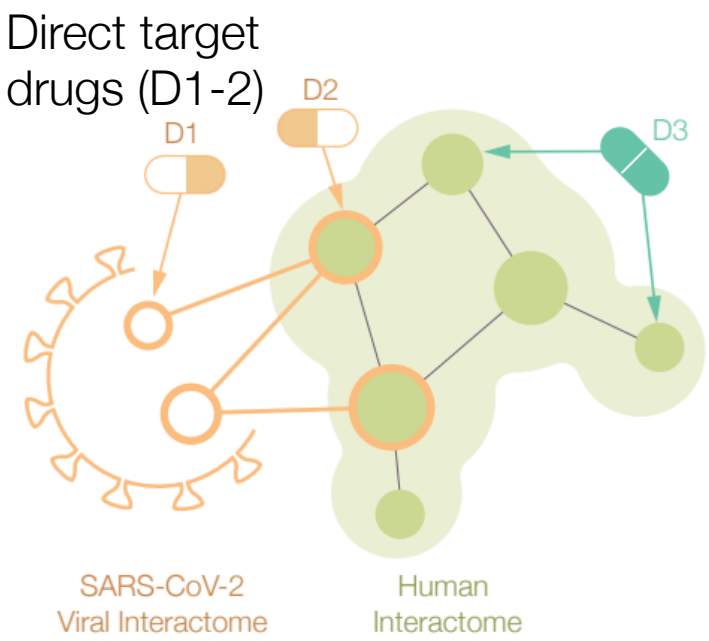
Strong
Weak

CRank	Drug Name
5	Chloroquine
6	Rifabutin
9	Rifaximin
10	Azelastine
16	Folic acid
32	Methotrexate
33	Digoxin
44	Hydroxychloroquine
50	Omeprazole
113	Clobetasol propionate
118	Auranofin
120	Vinblastine
199	Fluvastatin
210	Clomifene
233	Ibuprofen
235	Ivermectin
243	Atorvastatin
253	Pralatrexate
263	Cobimetinib
269	Hydralazine
297	Propranolol
317	Osimertinib
348	Vincristine
367	Doxazosin
397	Rosiglitazone
398	Aminolevulinic acid

CRank	Drug Name
423	Pitavastatin
431	Tenoxicam
438	Quinidine
456	Sertraline
460	Ingenol mebutate
463	Norelgestromin
493	Sildenafil
499	Eliglustat
518	Ulipristal
553	Cinacalcet
556	Perphenazine
558	Idarubicin
564	Perhexiline
569	Amiodarone
577	Duloxetine
585	Toremifene
586	Afatinib
601	Amitriptyline
626	Meclizine
635	Valsartan
651	Eletriptan
673	Sotalol
678	Thioridazine
695	Chlorcyclizine
707	Omacetaxine mepesuccinate
721	Candesartan

CRank	Drug Name
742	Mianserin
755	Clofazimine
767	Chlorpromazine
772	Imipramine
830	Promazine
900	L-Alanine
917	Moxifloxacin
933	Tasimelteon
995	Vandetanib
1000	Azilsartan medoxomil
1020	Frovatriptan
1034	Zolmitriptan
1035	Procarbazine
1093	Asenapine
1107	Dyclonine
1140.5	Clemastine
1194	Prochlorperazine
1222	Miglustat
1224	Prenylamine
1276	Dalfampridine
1314	Cinchocaine
1355	Methotrimeprazine
1396	Methylthioninium
1403	Metixene
1443	Trifluoperazine

58/77 drugs with positive experimental outcome are among top 750 ranked drugs



Network drugs (D3)