

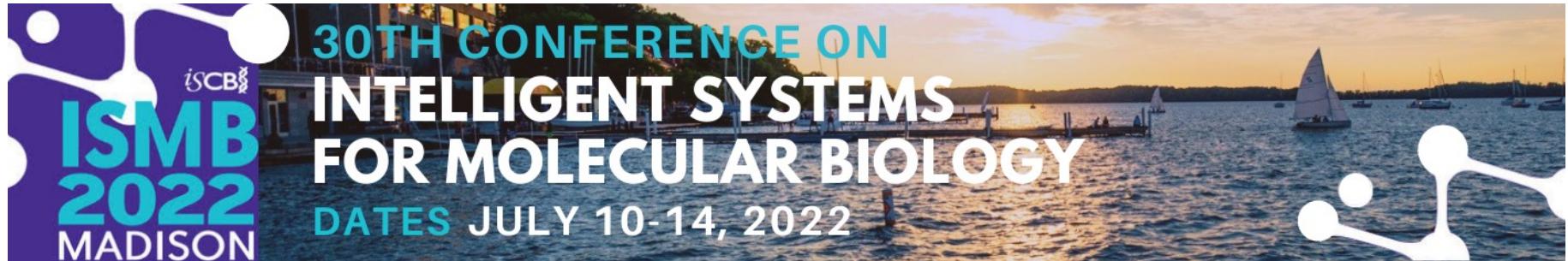
# Towards Precision Medicine with Graph Representation Learning

Michelle M. Li & Marinka Zitnik

Department of Biomedical Informatics  
Broad Institute of Harvard and MIT  
Harvard Data Science

[zitniklab.hms.harvard.edu/biomedgraphml](http://zitniklab.hms.harvard.edu/biomedgraphml)





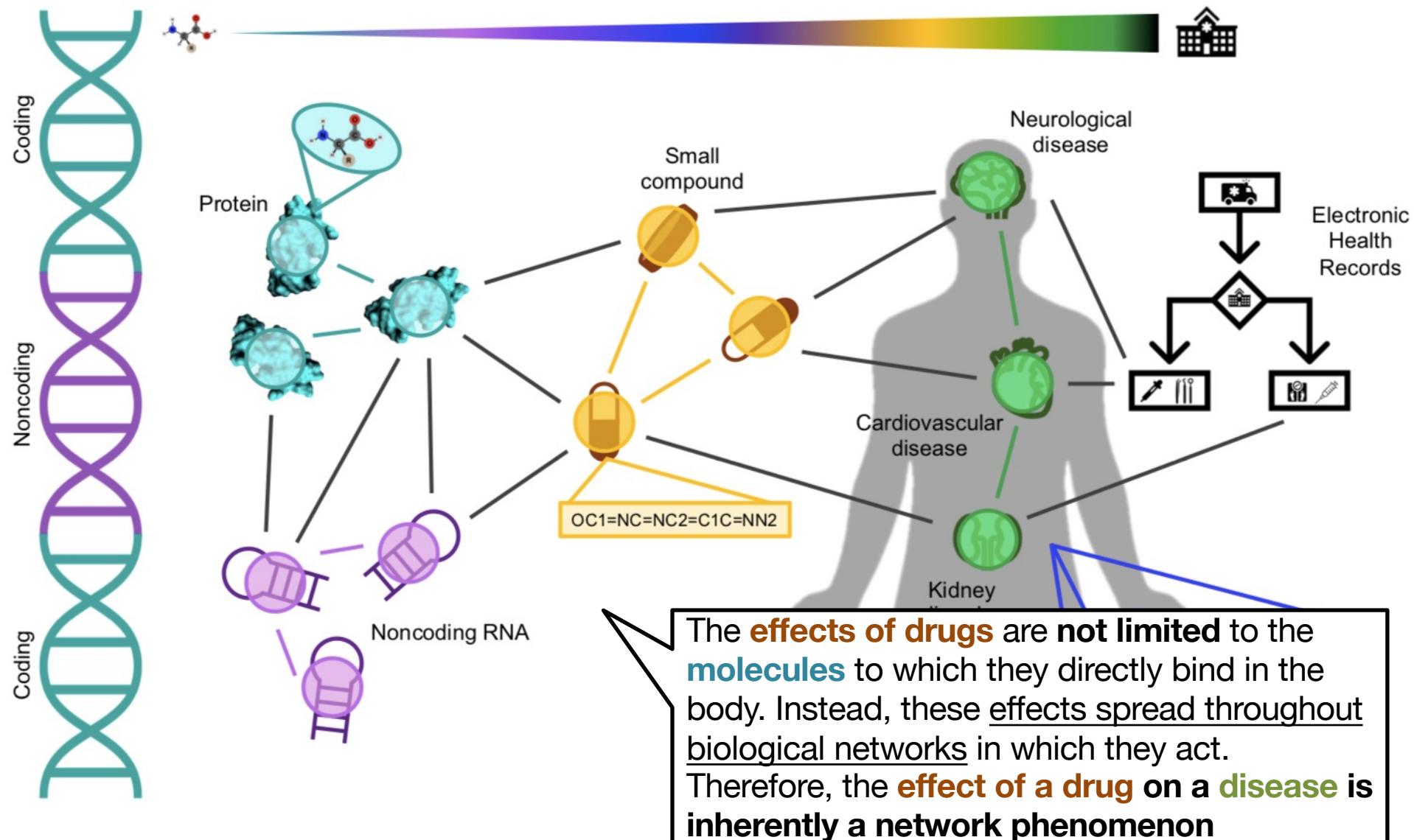
## Tutorial VT4

July 7, 2022 at 9am – 1pm CDT

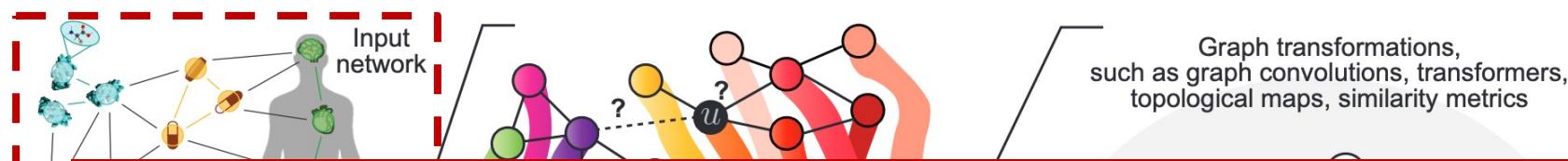


All tutorial materials are available at  
[zitniklab.hms.harvard.edu/biomedgraphml](http://zitniklab.hms.harvard.edu/biomedgraphml)

# Biology is interconnected



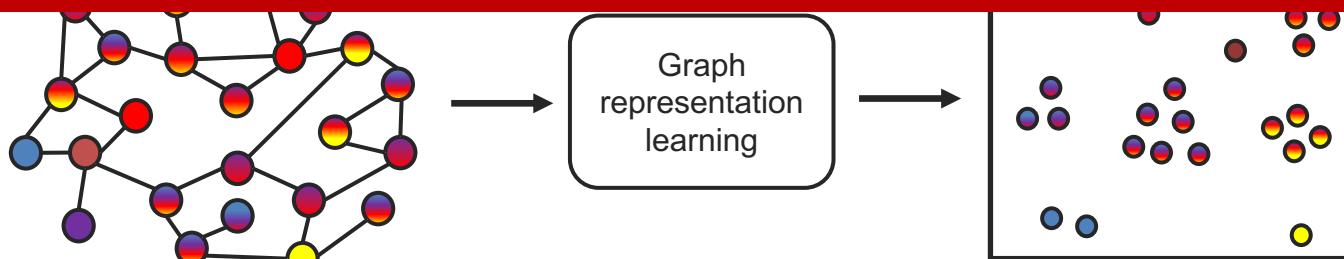
# Graph representation learning realizes key network principles for data-rich biomedicine



Cellular components associated with a specific disease (phenotype) show a tendency to cluster in the same network neighborhood



Deep graph representation learning methods are well-suited for the analysis of biological networks



# This Tutorial

- ✓ 1. Methods: Network diffusion, shallow network embeddings, graph neural networks, equivariant neural networks
- 👉 2. Applications: Fundamental biological discoveries and precision medicine
- 3. Hands-on exercises: Demos, implementation details, tools, and tips

Applications of graph representation learning on...

# DISEASES

1. Single-cell transcriptomics data
2. Spatial transcriptomics data

Applications of graph representation learning on...

# DISEASES

1. Single-cell transcriptomics data
2. Spatial transcriptomics data

## Disease State Prediction From Single-Cell Data Using Graph Attention Networks

Neal G. Ravindra\*†  
Yale University  
neal.ravindra@yale.edu

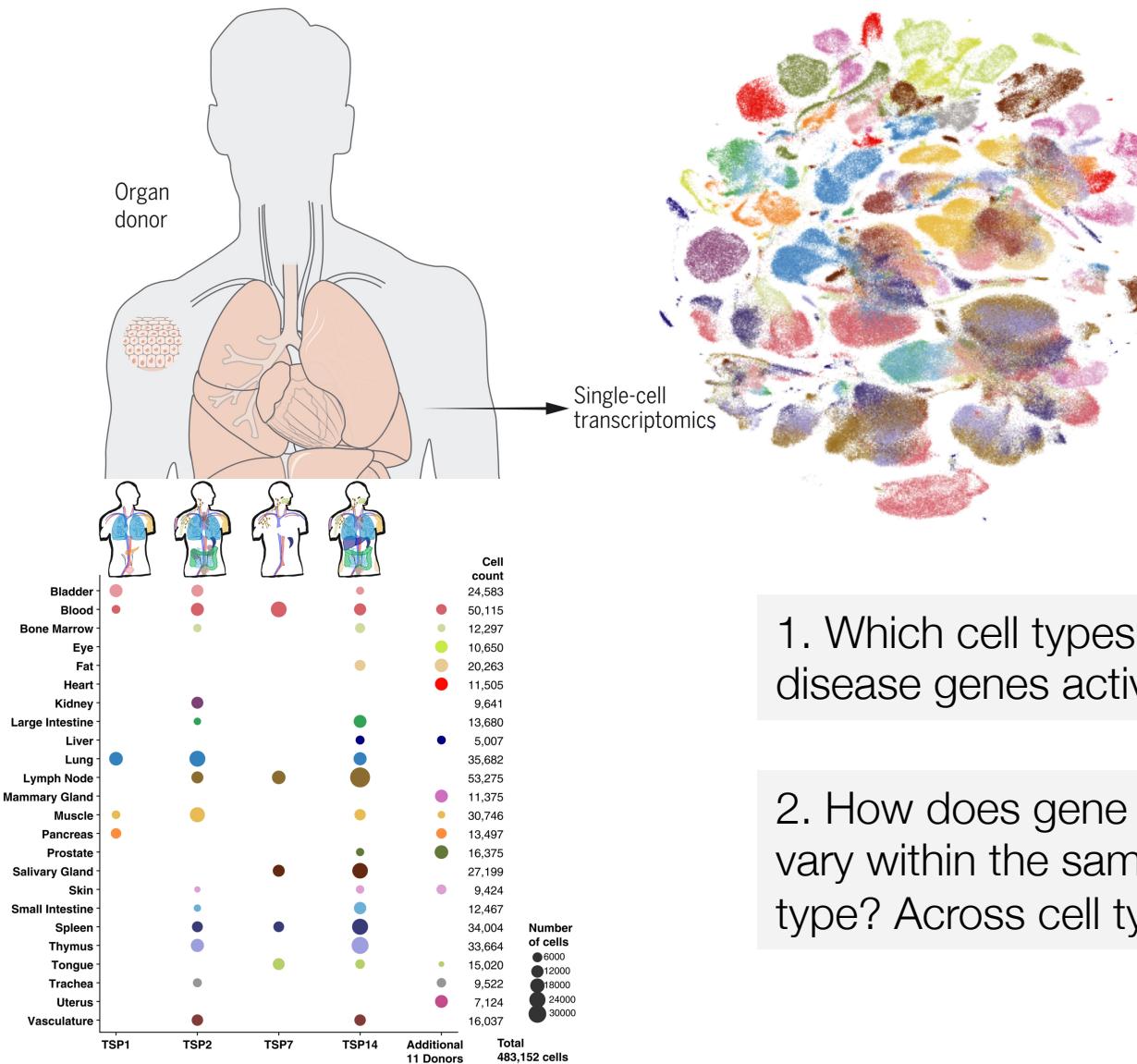
Arijit Sehanobish\*†  
Yale University  
arijit.sehanobish@yale.edu

Jenna L. Pappalardo‡  
Yale University  
jenna.pappalardo@yale.edu

David A. Hafler‡  
Yale University  
david.hafler@yale.edu

David van Dijk†  
Yale University  
david.vandijk@yale.edu

# Cross-tissue human cell atlases

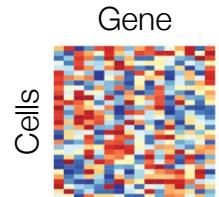


1. Which cell types are disease genes active?

2. How does gene expression vary within the same cell type? Across cell types?

# Limitations of existing methods

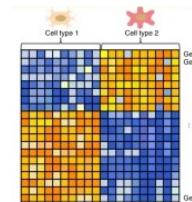
## scRNA-seq pipelines studying genetics & disease



Filtering, normalizing,  
and clustering



Differential gene expression  
(DGE) analysis



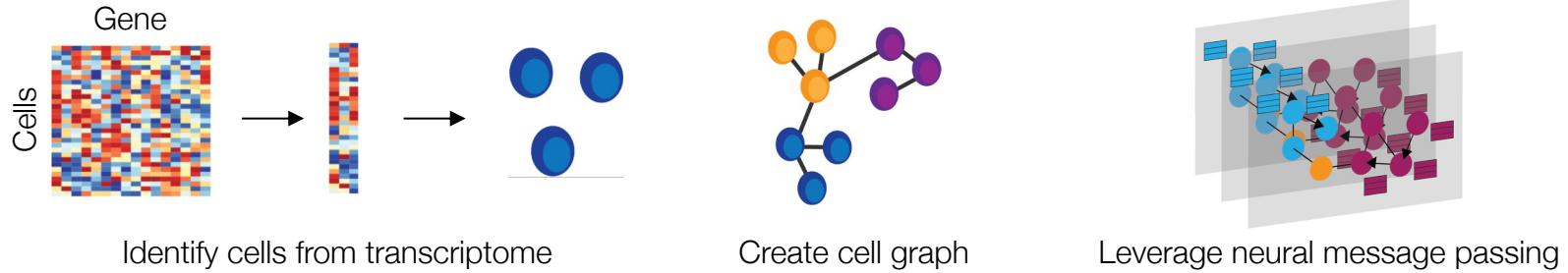
### Limitations of using differential gene expression (DGE)

- DGE is not necessarily associated with disease or cell state
- The “most differentially” expressed genes do not yield causal structure

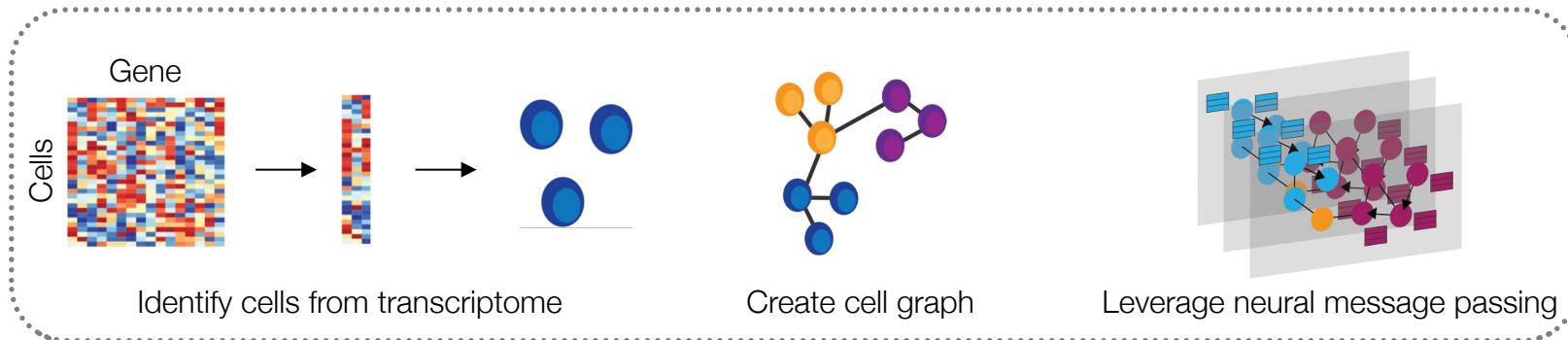
### Methodological limitations for differential gene expression (DGE)

- Most DGE methods don’t allow for interactions between features
- Pipeline error and missing trivial non-linear, causal covariance structures

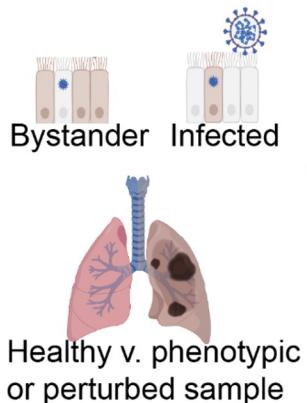
# Leveraging cell-cell interactions



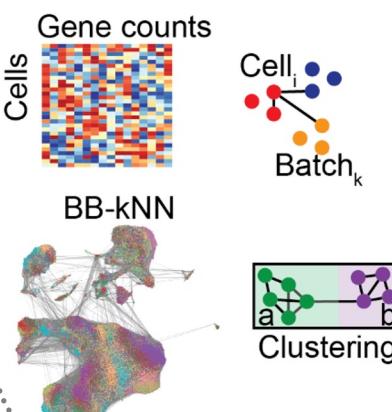
# Leveraging cell-cell interactions



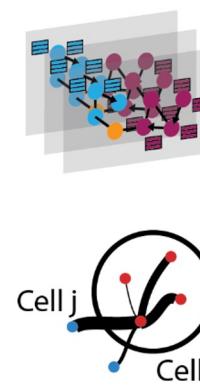
Single-cell measurements



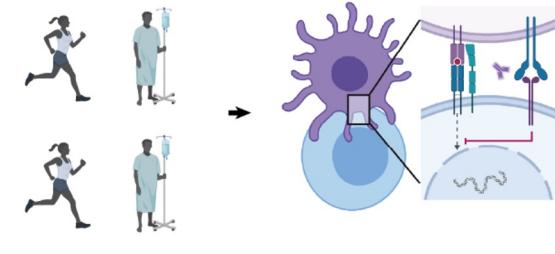
scRNA-seq pipeline



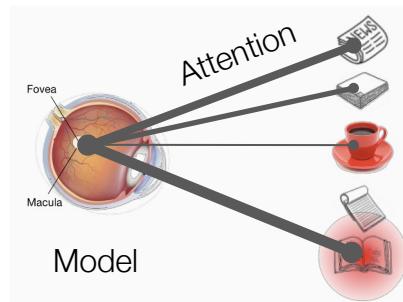
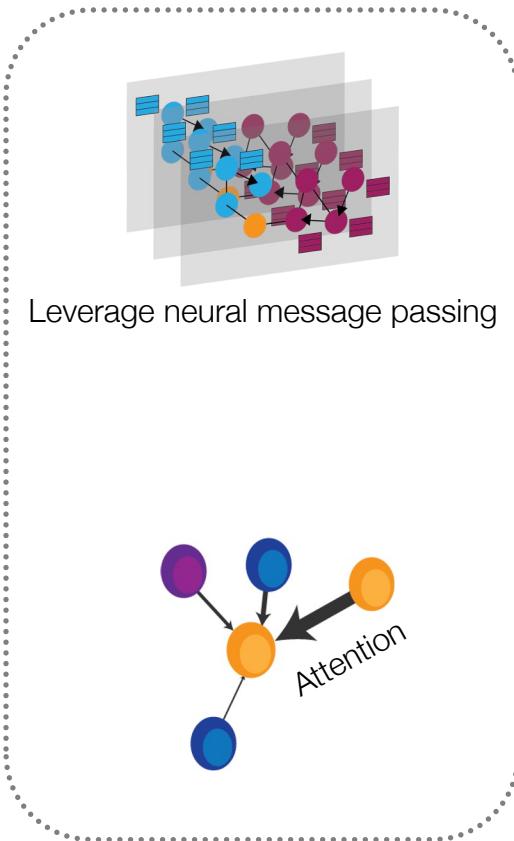
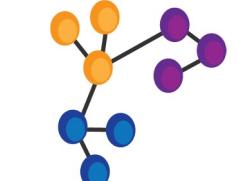
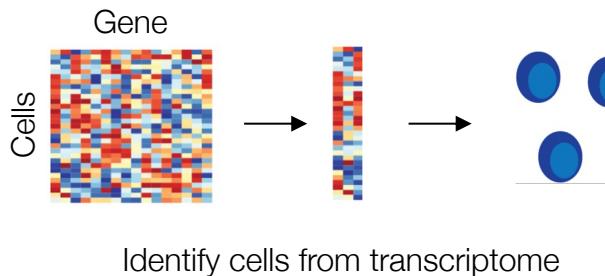
scGAT



Cell-by-cell prediction and hypothesis generation



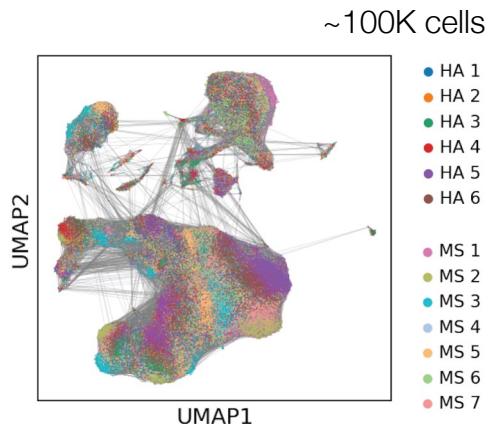
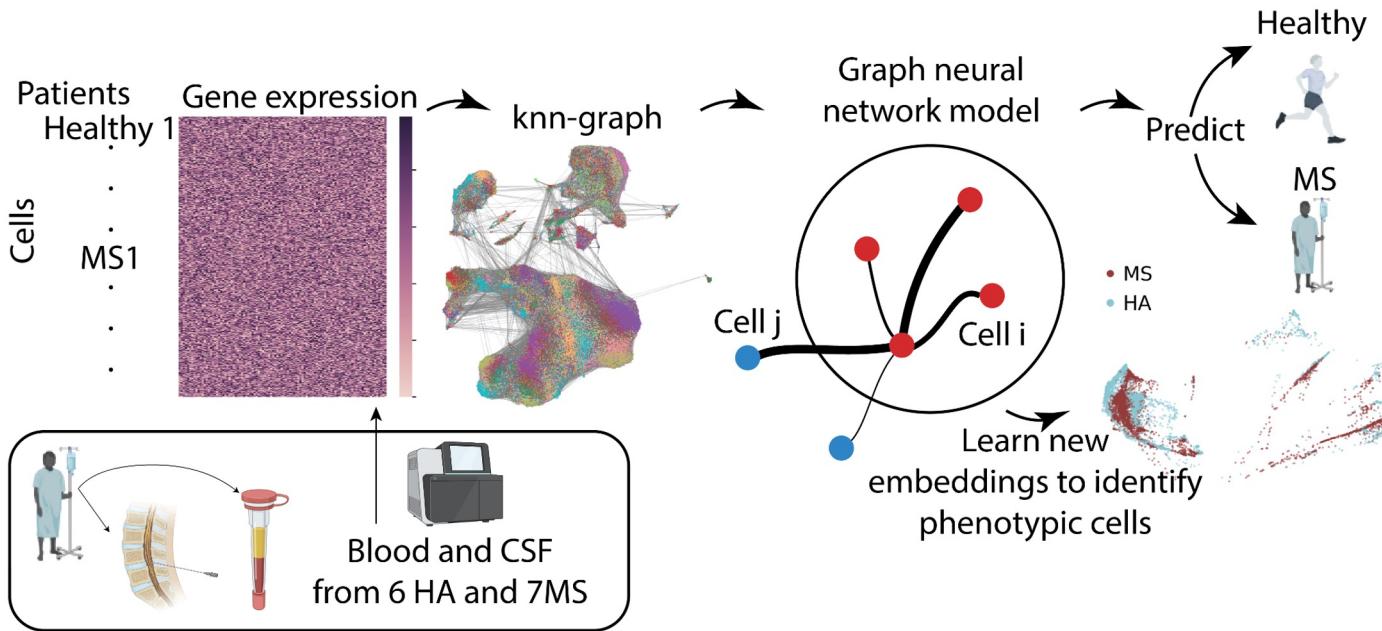
# Leveraging cell-cell interactions



Which cell-cell interactions contribute the most/least to a specific disease state?

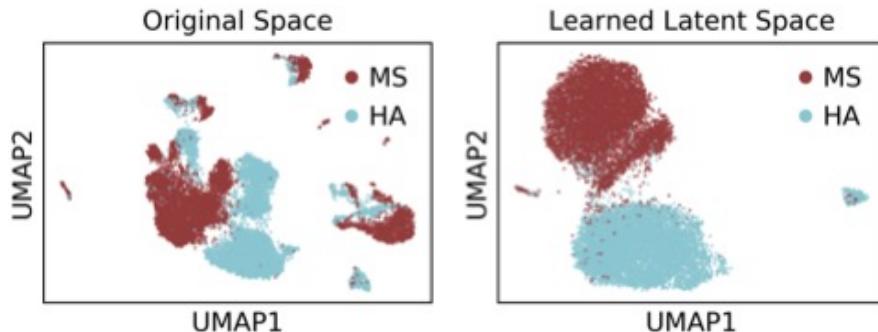
## Overview

## Patient Dataset



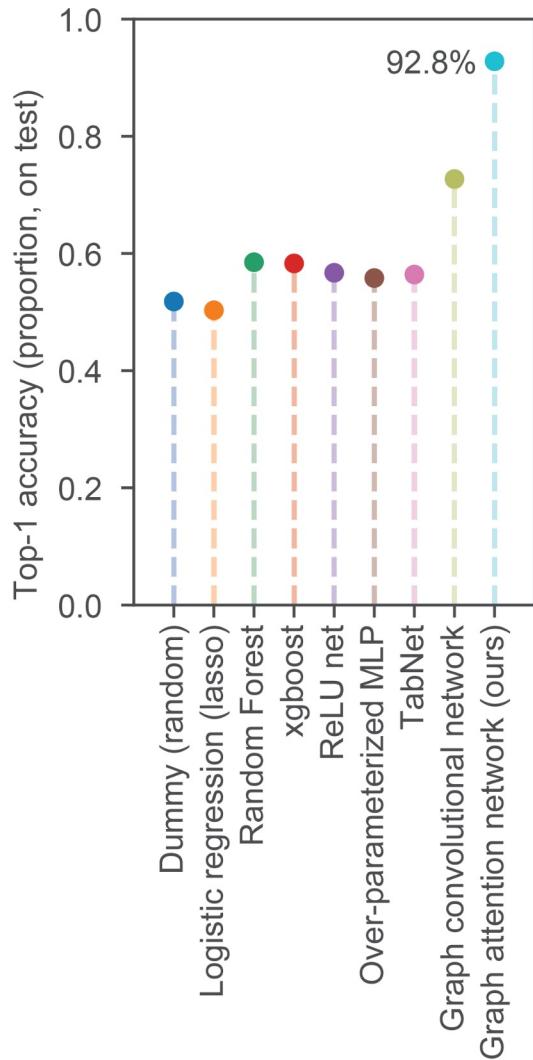
Task		Train	Dev	Test
Inductive	# Nodes	43866	9686	13033
	# Edges	332398	73552	100715
	# Features	22005	22005	22005
	# Classes	2	2	2
	# Graphs	1	1	1
Transductive	# Nodes	54000	6000	6667
	# Features	22005	22005	22005
	# Classes	2	2	2
	# Edges	5007093		

# Results

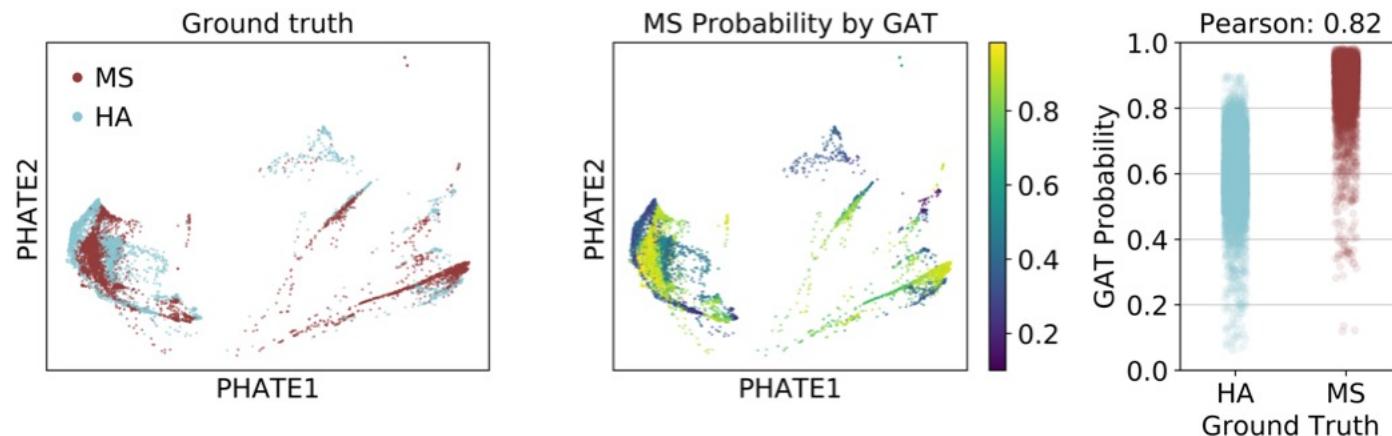


Task	Model	Accuracy
Inductive	Random	51.8
	MLP	56.7
	Random Forest	58.5
	Graph Convolutional Network	72.1
	Graph Attention Network(our)	<b>92.3 ± .7</b>
Transductive	Graph Convolutional Network	82.91
	Graph Attention Network(our)	<b>86 ± .3</b>

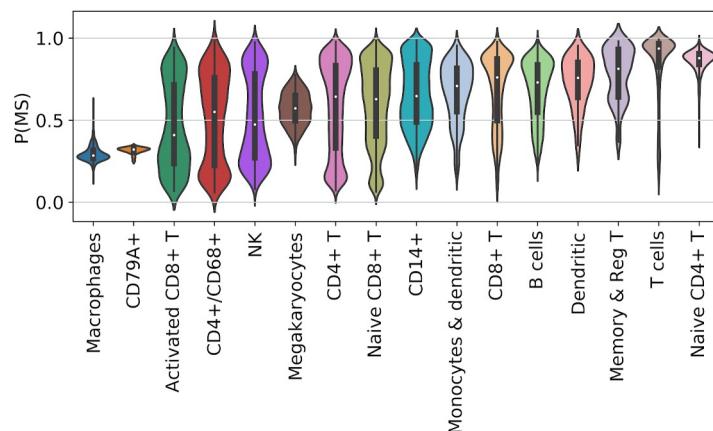
- **Transductive task:** Randomly assign 10% nodes for validation & 10% for testing
  - Keeping ratio of healthy & MS cells same as in full dataset
- **Inductive task:** Randomly choose a healthy adult & MS patient
  - Train on remaining 5 MS patients and 4 healthy adults



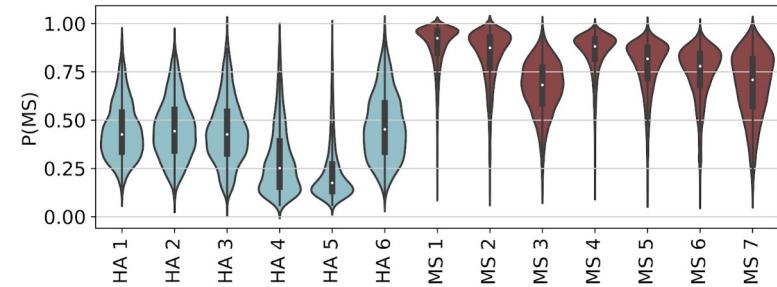
# Results



Predicted probabilities from induction task per cell



Aggregating predicted probabilities shows cell types important for predicting disease state

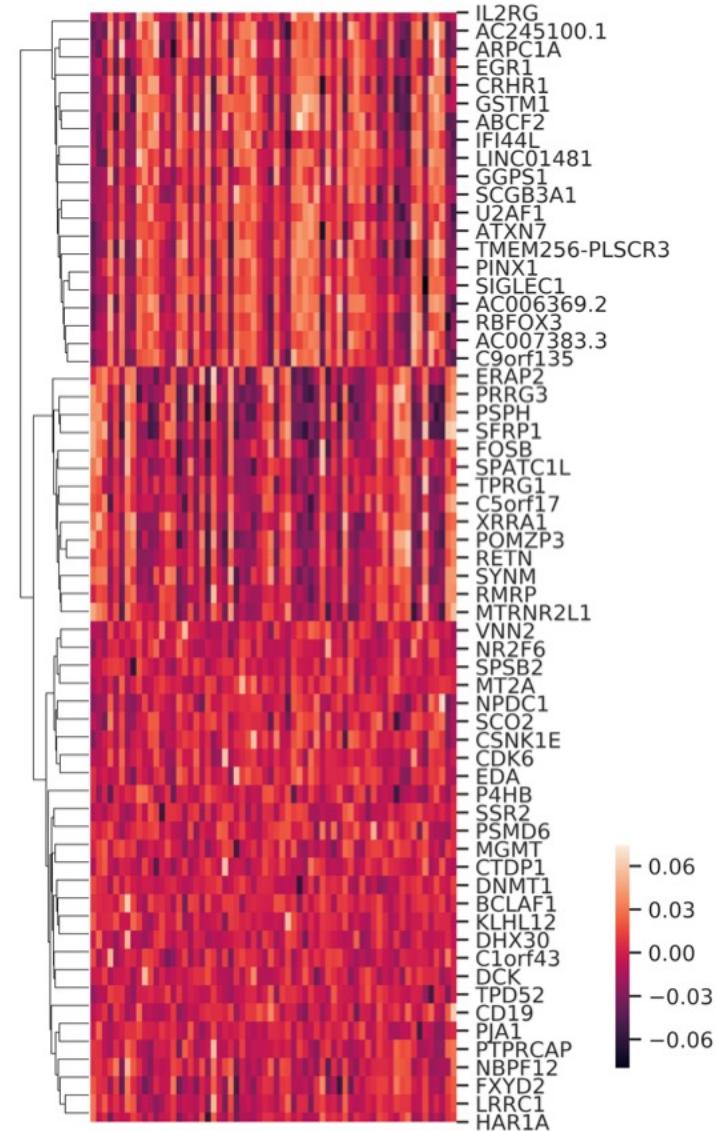


Variance of a patient's cells' probability of being in an MS state may indicate timing of flare-up

# Results

Per  $k$  head,  $g_i^k = \max_j(|w_{ij}|)$

- Interleukin-2 receptor subunit (IL2RG) among top 10 predictive features per head
- Marker for therapeutically targeted B cells (CD19) also among top features
- Top predictive features regulate hormone secretion, nerve cell development, and lipid metabolism, suggesting relevant but novel hits



# Key Takeaways

- Cell graphs enable modeling of cell-cell interactions
  - Typically not considered in standard scRNA-seq pipelines studying disease
- GAT outperforms classic models as well as related GNNs (without attention mechanism) in predicting disease state from a transcriptome
  - Aggregating predicted probabilities shows cell types important for predicting disease state
  - Variance of a patient's cells' probability of being in an MS state may indicate timing of flare-up
  - Top predictive features regulate may be candidates for therapeutic targets
- Resources
  - Paper: [dl.acm.org/doi/10.1145/3368555.3384449](https://dl.acm.org/doi/10.1145/3368555.3384449)
  - GitHub: [github.com/vandijklab/scGAT](https://github.com/vandijklab/scGAT)
  - Follow-up work on COVID-19: [arxiv.org/abs/2007.04777](https://arxiv.org/abs/2007.04777)

Applications of graph representation learning on...

# DISEASES

1. Single-cell transcriptomics data
2. Spatial transcriptomics data

Applications of graph representation learning on...

# DISEASES

1. Single-cell transcriptomics data
2. Spatial transcriptomics data

METHOD

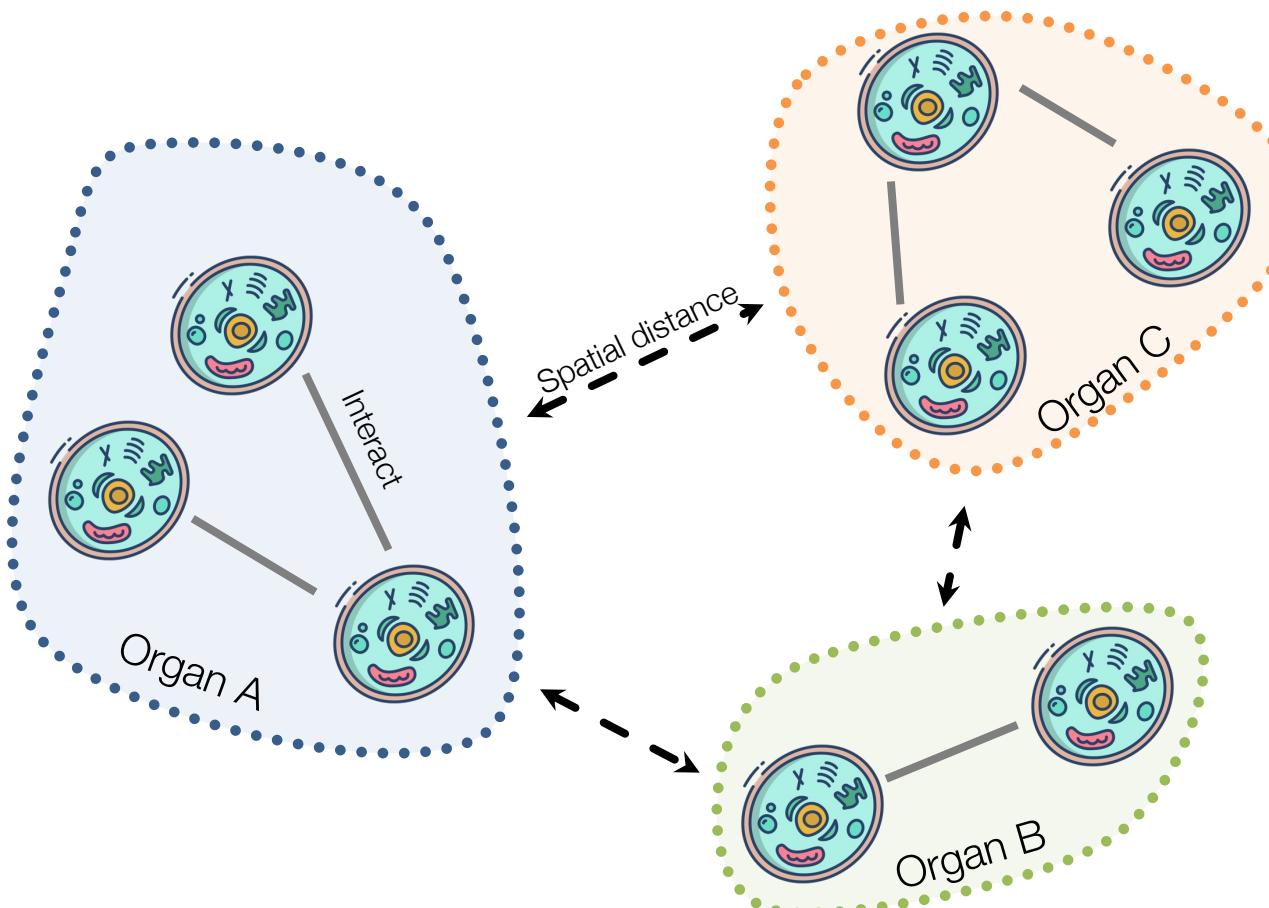
Open Access

## GCNG: graph convolutional networks for inferring gene interaction from spatial transcriptomics data



Ye Yuan<sup>1</sup> and Ziv Bar-Joseph<sup>1,2\*</sup> 

# Motivation

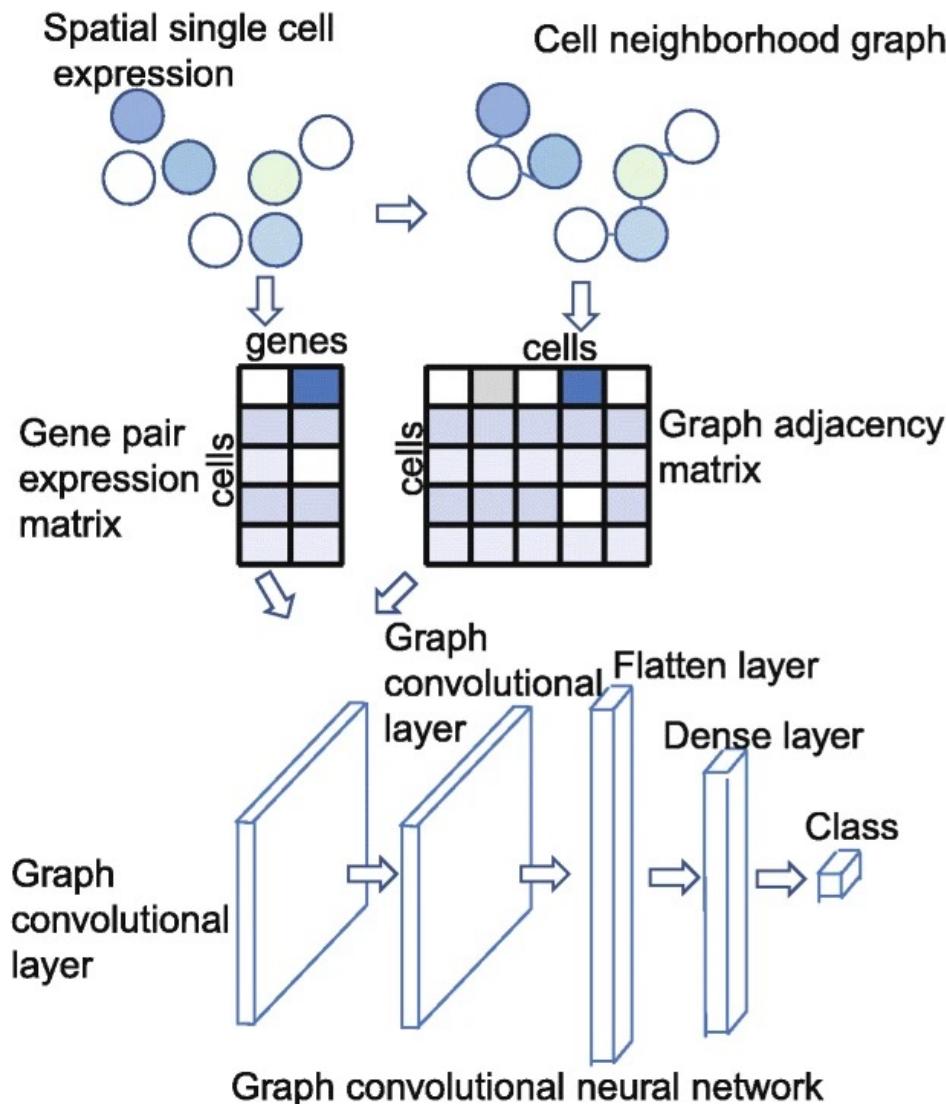


1. Leverage higher-order interactions between genes & cells

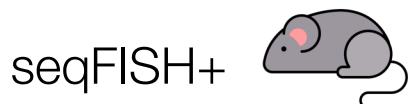
2. Utilize both gene expression & cellular relationships

3. Overcome incomplete spatial relationships

# Overview



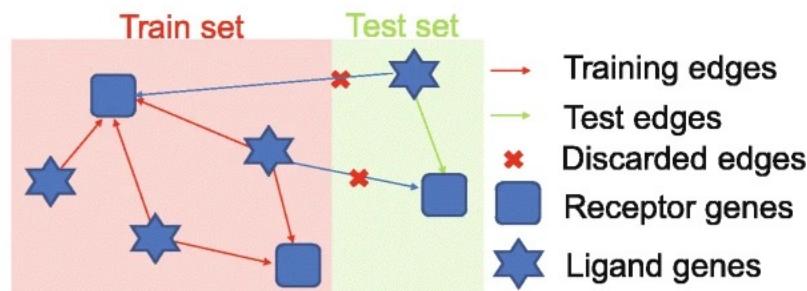
# Experimental setup



- Mouse cortex tissue
- Expression data: 10,000 genes in 913 cells
- Labeled ligand-receptor pairs: 1056 known interactions between 309 ligands and 481 receptors

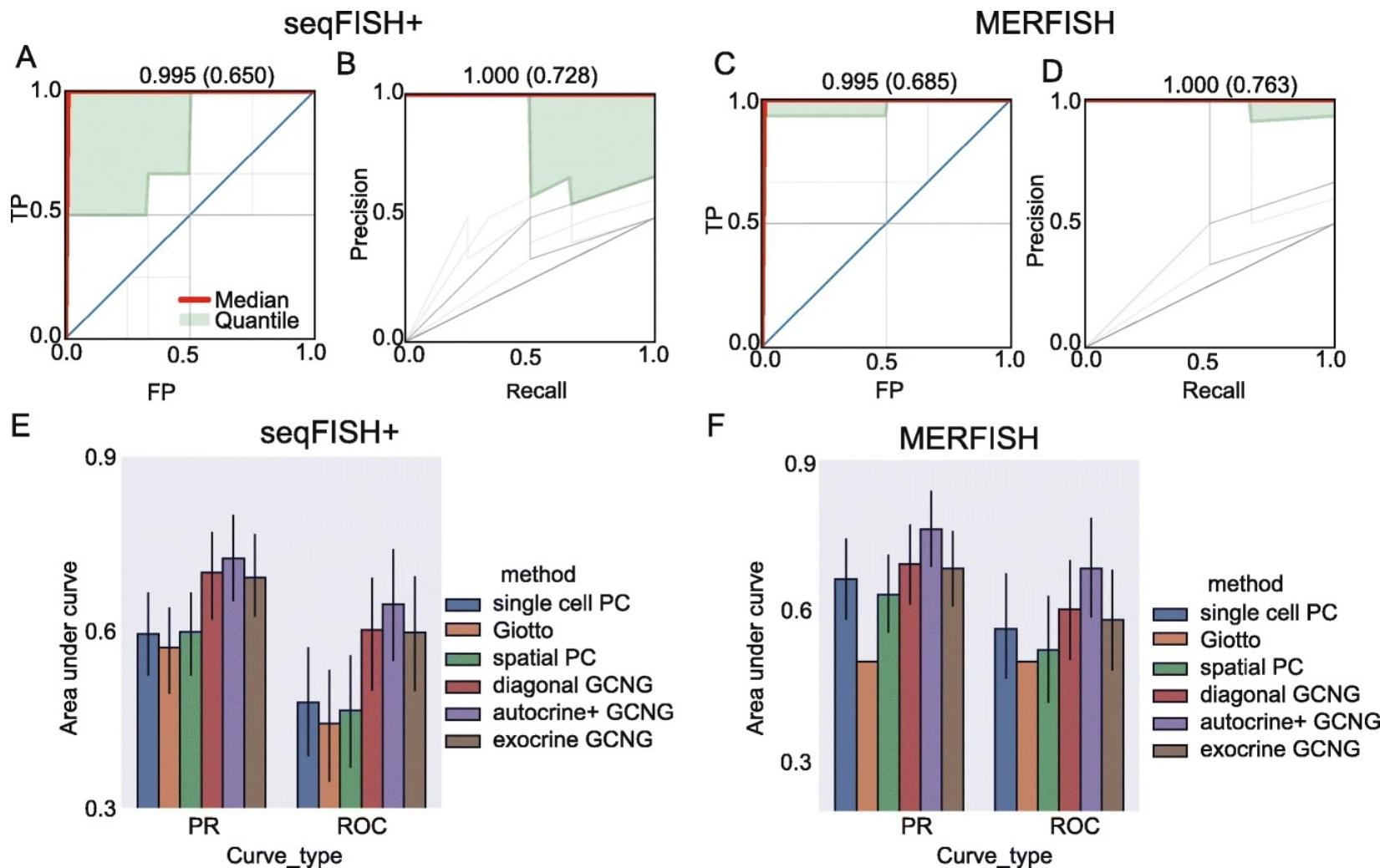


- Cells (in vitro)
- Expression data: 10,050 genes from 1368 cells
- Labeled ligand-receptor pairs: 841 known interactions between 270 ligands and 376 receptors

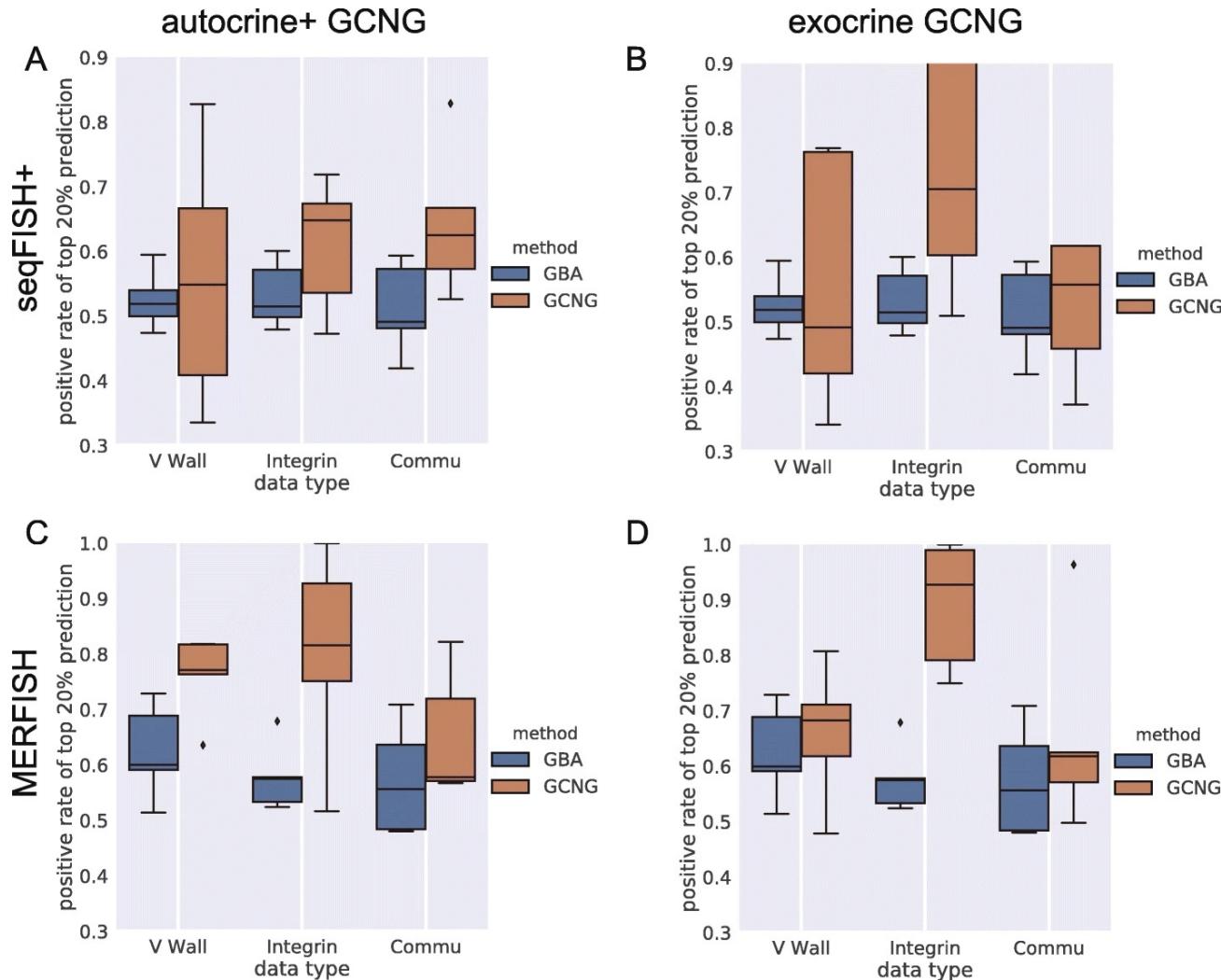


Ligand a	Interacting receptor b	1
Ligand a	Non Interacting receptor x	0
Ligand a	Interacting receptor c	1
Ligand a	Non Interacting receptor y	0

# Results



# Results



# Key Takeaways

- GCNG
  - Encodes the spatial information as a graph
  - Combines the spatial cell neighborhood graph with expression data using supervised learning
    - Unlike standard approaches, which rely on unsupervised correlation-based analysis
  - Can propose novel pairs of extracellular interacting genes
  - Outputs can be used for downstream analysis, including functional assignment
- Resources
  - Paper: [genomebiology.biomedcentral.com/articles/10.1186/s13059-020-02214-w](https://genomebiology.biomedcentral.com/articles/10.1186/s13059-020-02214-w)
  - GitHub: [github.com/xiaoyeye/GCNG](https://github.com/xiaoyeye/GCNG)
  - Relevant papers:
    - Wang et al. *Nature Communications* (2021). [scGNN is a novel graph neural network framework for single-cell RNA-seq analysis](#)
    - Ding and Regev, *Nature Communications* (2021). [Deep generative model embedding of single-cell RNA-seq profiles on hyperspheres and hyperbolic spaces](#)

# Graph RL for diseases

## Summary

- **Single cell GAT:** Model cellular interactions to learn disease state of cells while identifying (via attention mechanism) the cell types and biomarkers that contributed most to disease
- **Spatial transcriptomics:** Construct a spatial cell neighborhood graph and combine with expression data to model cellular interactions with gene- to tissue-level organization

## Poll Question

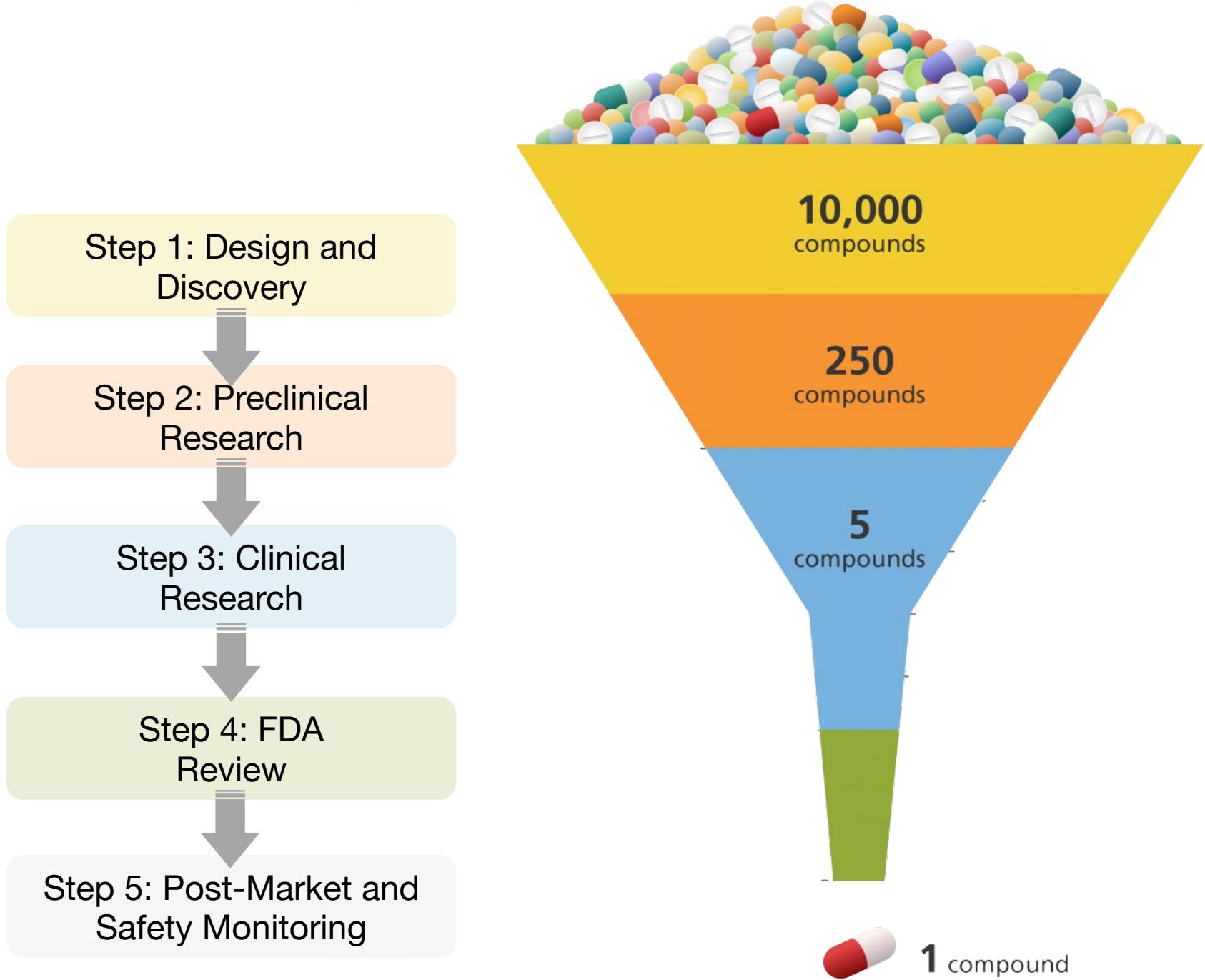
What diseases might the use of graph representation learning on single-cell/spatial transcriptomics data be the MOST or LEAST impactful for? *Fill in the blank*

## Q&A Session

Applications of graph representation learning on...

# THERAPEUTICS

1. Molecular property prediction, drug-target interaction prediction, molecular generation
2. Drug discovery
3. Drug repurposing



Applications of graph representation learning on...

# THERAPEUTICS

1. Molecular property prediction, drug-target interaction prediction, molecular generation
  2. Drug discovery
  3. Drug repurposing
- 

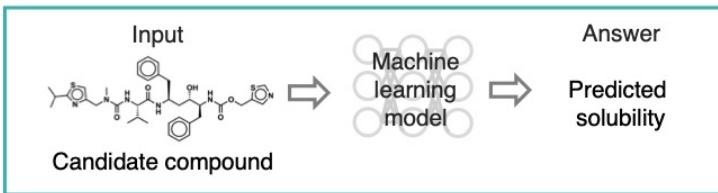
## **Therapeutics Data Commons: Machine Learning Datasets and Tasks for Drug Discovery and Development**

---

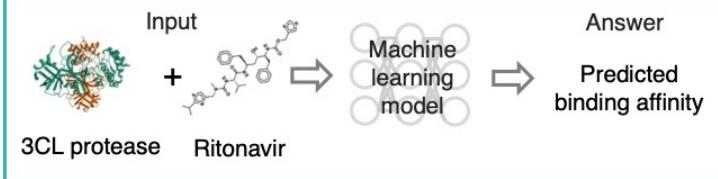
**Kexin Huang<sup>1,\*</sup>, Tianfan Fu<sup>2,\*</sup>, Wenhao Gao<sup>3,\*</sup>, Yue Zhao<sup>4</sup>, Yusuf Roohani<sup>5</sup>, Jure Leskovec<sup>5</sup>, Connor W. Coley<sup>3</sup>, Cao Xiao<sup>6</sup>, Jimeng Sun<sup>7</sup>, Marinka Zitnik<sup>1</sup>**  
<sup>1</sup>Harvard <sup>2</sup>Georgia Tech <sup>3</sup>MIT <sup>4</sup>CMU <sup>5</sup>Stanford <sup>6</sup>Amplitude <sup>7</sup>UIUC  
contact@tdcommons.ai

# Compelling applications of graph RL

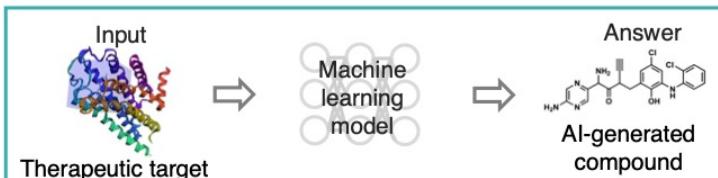
I want to know the solubility of a compound of interest.



I want to know the binding affinity of Ritonavir to 3CL protease.



I want to generate a highly potent compound that effectively binds a therapeutic target.



I want to iteratively design a small molecule drug, informed by ADMET and binding affinity predictions.



AMINO ACID SEQUENCE

SGFRKMAFPFGKVKVSCMVQVTGCTTTLNGLMDDVVYCPHRVICTSEDMLNPNEYEDLLIRKSNNENFLVQAGNVQLRIVIGRSNQNVCVLKLVDTANPKTFKKYKVFIQPGQTFSVLAICYGSFVGYCMMRPNFTIKGSGCGSVGFNIDYEDCFCSFCYMHMELPFGVRAGTDLDEGNPFGPFVDRDQTQAAGDTDTITVNLVLMLYAAVINGDRWFLNRFTTTLINDFNLVAMKVNTEPLTQDHVDILGPLSAQTGIAVLDMCASLKELLQ

MOLECULE

AFFINITY PREDICTION MODEL TYPE: MPNN-CNN

ADMET PREDICTION MODEL TYPE: MPNN

BINDING AFFINITY (KD): 749.91 nM

CANONICAL SMILES: O=[NH]c1=O[nH]c1([C#H])2([C#H])((C#H)([C#H](F)C)O)COP(=O)([N][C#H])C(OCC(C)=O)C=OCCCCC3

BINDING AFFINITY (KD): 749.91 nM

PREDICTED ADMET PROPERTY

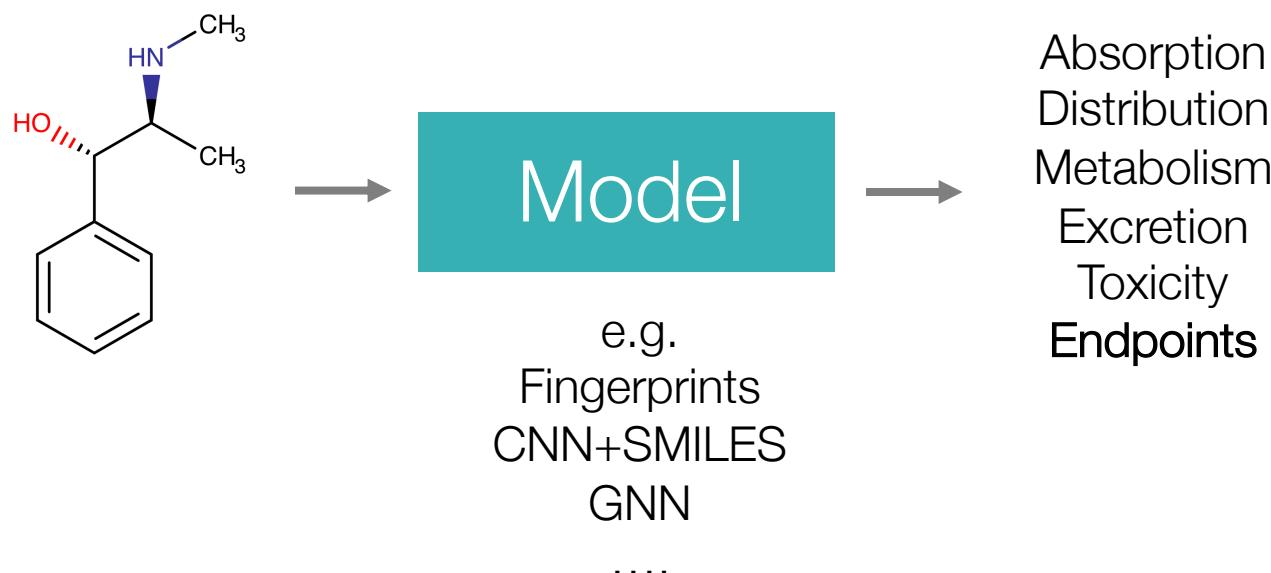
Property	Value
Solubility	-2.89 log mol/L
Lipophilicity	1.21 (log-ratio)
(Absorption) Caco-2	-5.39 cm/s
(Absorption) HIA	2.13 %
(Absorption) BBB	67.58 %
(Distribution) PBR	74.56 %
(Metabolism) CYP2D6	57.17 %
(Metabolism) CYP3A4	26.57 %
(Metabolism) CYP2A6	9.32 %
(Metabolism) CYP1A2	1.15 %
(Metabolism) CYP2E1	10.25 %
(Excretion) Biotransformation	1.63 %
(Excretion) Half life	1.24 h
(Excretion) Clearance	8.28 h
Clinical Toxicity	8.08 mL/min/kg
	28.47 %

Latency: 0.3s

SCREENSHOT FLAG

Powered by Gradio

# ADMET property prediction



# Datasets

22 datasets with ADMET endpoints

## Absorption

Caco2 (Cell Permeability)  
HIA (Intestinal Absorption)  
Pgp (P-glycoprotein)  
Bioavailability  
Lipophilicity  
Solubility

## Excretion

Half Life  
Clearance (Hepatocyte)  
Clearance (Microsome)

## Distribution

BBB (Blood-Brain Barrier)  
PPBR (Plasma Protein Binding)  
VDss (Volume of Distribution)

## Toxicity

LD50 (Acute Toxicity)  
hERG blocker  
Ames Mutagenicity  
Drug Induced Liver Injury

## Metabolism

CYP2C9/2D6/3A4 Inhibition  
CYP2C9/2D6/3A4 Substrate



Paper  
Supplementary



Public  
Database



# Results: ADMET prediction (1/3)

Raw Feature Type		Expert-Curated Methods		SMILES	Molecular Graph-Based Methods (state-of-the-Art in ML)					
Dataset	Metric	Morgan [31]	RDKit2D [24]	CNN [18]	NeuralFP [7]	GCN [23]	AttentiveFP [43]	AttrMasking [16]	ContextPred [16]	
	# Params.	1477K	633K	227K	480K	192K	301K	2067K	2067K	
TDC.Caco2 (↓)	MAE	0.908±0.060	<b>0.393±0.024</b>	0.446±0.036	0.530±0.102	0.599±0.104	<u>0.401±0.032</u>	0.546±0.052	0.502±0.036	
TDC.HIA (↑)	AUROC	0.807±0.072	<b>0.972±0.008</b>	0.869±0.026	0.943±0.014	0.936±0.024	<u>0.974±0.007</u>	<b>0.978±0.006</b>	0.975±0.004	
TDC.Pgp (↑)	AUROC	0.880±0.006	0.918±0.007	0.908±0.012	0.902±0.020	0.895±0.021	0.892±0.012	<b>0.929±0.006</b>	0.923±0.005	
TDC.Bioav (↑)	AUROC	0.581±0.086	<b>0.672±0.021</b>	0.613±0.013	0.632±0.036	0.566±0.115	0.632±0.039	0.577±0.087	0.671±0.026	
TDC.Lipo (↓)	MAE	0.701±0.009	<b>0.574±0.017</b>	0.743±0.020	0.563±0.023	<u>0.541±0.011</u>	0.572±0.007	0.547±0.024	<b>0.535±0.012</b>	
TDC.AqSol (↓)	MAE	1.203±0.019	<b>0.827±0.047</b>	1.023±0.023	0.947±0.016	0.907±0.020	<b>0.776±0.008</b>	1.026±0.020	1.040±0.045	
TDC.BBB (↑)	AUROC	0.823±0.015	<b>0.889±0.016</b>	0.781±0.030	<u>0.836±0.009</u>	0.842±0.016	0.855±0.011	<u>0.892±0.012</u>	<b>0.897±0.004</b>	
TDC.PPBR (↓)	MAE	12.848±0.362	<b>9.994±0.319</b>	11.106±0.358	<b>9.292±0.384</b>	10.194±0.373	<u>9.373±0.335</u>	<u>10.075±0.202</u>	9.445±0.224	
TDC.VD (↑)	Spearman	0.493±0.011	<b>0.561±0.025</b>	0.226±0.114	0.258±0.162	0.457±0.050	<u>0.241±0.145</u>	<u>0.559±0.019</u>	0.485±0.092	
TDC.CYP2D6-I (↑)	AUPRC	0.587±0.011	<b>0.616±0.007</b>	0.544±0.053	0.627±0.009	0.616±0.020	0.646±0.014	<u>0.721±0.009</u>	<b>0.739±0.005</b>	
TDC.CYP3A4-I (↑)	AUPRC	0.827±0.009	<b>0.829±0.007</b>	0.821±0.003	0.849±0.004	0.840±0.010	0.851±0.006	<u>0.902±0.002</u>	<b>0.904±0.002</b>	
TDC.CYP2C9-I (↑)	AUPRC	0.715±0.004	<b>0.742±0.006</b>	0.713±0.006	0.739±0.010	0.735±0.004	0.749±0.004	<u>0.829±0.003</u>	<b>0.839±0.003</b>	
TDC.CYP2D6-S (↑)	AUPRC	0.671±0.066	<b>0.677±0.047</b>	0.485±0.037	0.572±0.062	0.617±0.039	0.574±0.030	<u>0.704±0.028</u>	<b>0.736±0.024</b>	
TDC.CYP3A4-S (↑)	AUROC	0.633±0.013	<b>0.639±0.012</b>	<b>0.662±0.031</b>	0.578±0.020	0.590±0.023	0.576±0.025	<u>0.582±0.021</u>	0.609±0.025	
TDC.CYP2C9-S (↑)	AUPRC	0.380±0.015	<b>0.360±0.040</b>	0.367±0.059	0.359±0.059	0.344±0.051	0.375±0.032	<u>0.381±0.045</u>	<b>0.392±0.026</b>	
TDC.Half_Life (↑)	Spearman	<b>0.329±0.083</b>	<b>0.184±0.111</b>	0.038±0.138	0.177±0.165	<b>0.239±0.100</b>	0.085±0.068	0.151±0.068	0.129±0.114	
TDC.CL-Micro (↑)	Spearman	0.492±0.020	<b>0.386±0.014</b>	0.252±0.116	0.529±0.015	<b>0.532±0.033</b>	0.365±0.055	0.585±0.034	0.578±0.007	
TDC.CL-Hepa (↑)	Spearman	0.272±0.068	<b>0.382±0.007</b>	0.235±0.021	0.401±0.037	<b>0.366±0.063</b>	0.289±0.022	0.413±0.028	<b>0.439±0.026</b>	
TDC.hERG (↑)	AUROC	0.736±0.023	<b>0.841±0.020</b>	0.754±0.037	0.722±0.034	0.738±0.038	<u>0.825±0.007</u>	0.778±0.046	0.756±0.023	
TDC.AMES (↑)	AUROC	0.794±0.008	<b>0.823±0.011</b>	0.776±0.015	0.823±0.006	0.818±0.010	<u>0.814±0.008</u>	<b>0.842±0.008</b>	0.837±0.009	
TDC.DILI (↑)	AUROC	0.832±0.021	<b>0.875±0.019</b>	0.792±0.016	0.851±0.026	0.859±0.033	<u>0.886±0.015</u>	<b>0.919±0.008</b>	0.861±0.018	
TDC.LD50 (↓)	MAE	0.649±0.019	<b>0.678±0.003</b>	0.675±0.011	0.667±0.020	0.649±0.026	<u>0.678±0.012</u>	<b>0.685±0.025</b>	0.669±0.030	

- Finding 1: No single method has the best performance across all scenarios

# Results: ADMET prediction (2/3)

Raw Feature Type		Expert-Curated Methods		SMILES	Molecular Graph-Based Methods (state-of-the-Art in ML)					
Dataset	Metric	Morgan [31]	RDKit2D [24]	CNN [18]	NeuralFP [7]	GCN [23]	AttentiveFP [43]	AttrMasking [16]	ContextPred [16]	
	# Params.	1477K	633K	227K	480K	192K	301K	2067K	2067K	
<b>TDC.Caco2</b> (↓)	MAE	0.908±0.060	<b>0.393±0.024</b>	0.446±0.036	0.530±0.102	0.599±0.104	<u>0.401±0.032</u>	0.546±0.052	0.502±0.036	
<b>TDC.HIA</b> (↑)	AUROC	0.807±0.072	<b>0.972±0.008</b>	0.869±0.026	0.943±0.014	0.936±0.024	<u>0.974±0.007</u>	<b>0.978±0.006</b>	0.975±0.004	
<b>TDC.Pgp</b> (↑)	AUROC	0.880±0.006	0.918±0.007	0.908±0.012	0.902±0.020	0.895±0.021	0.892±0.012	<b>0.929±0.006</b>	0.923±0.005	
<b>TDC.Bioav</b> (↑)	AUROC	0.581±0.086	<b>0.672±0.021</b>	0.613±0.013	0.632±0.036	0.566±0.115	0.632±0.039	<u>0.577±0.087</u>	0.671±0.026	
<b>TDC.Lipo</b> (↓)	MAE	0.701±0.009	<u>0.574±0.017</u>	0.743±0.020	0.563±0.023	<u>0.541±0.011</u>	0.572±0.007	0.547±0.024	<b>0.535±0.012</b>	
<b>TDC.AqSol</b> (↓)	MAE	1.203±0.019	<b>0.827±0.047</b>	1.023±0.023	0.947±0.016	0.907±0.020	<b>0.776±0.008</b>	1.026±0.020	1.040±0.045	
<b>TDC.BBB</b> (↑)	AUROC	0.823±0.015	<b>0.889±0.016</b>	0.781±0.030	0.836±0.009	0.842±0.016	0.855±0.011	<u>0.892±0.012</u>	<b>0.897±0.004</b>	
<b>TDC.PPBR</b> (↓)	MAE	12.848±0.362	<u>9.994±0.319</u>	11.106±0.358	<b>9.292±0.384</b>	10.194±0.373	<u>9.373±0.335</u>	<u>10.075±0.202</u>	9.445±0.224	
<b>TDC.VD</b> (↑)	Spearman	0.493±0.011	<b>0.561±0.025</b>	0.226±0.114	0.258±0.162	0.457±0.050	<u>0.241±0.145</u>	<u>0.559±0.019</u>	0.485±0.092	
<b>TDC.CYP2D6-I</b> (↑)	AUPRC	0.587±0.011	<u>0.616±0.007</u>	0.544±0.053	0.627±0.009	0.616±0.020	0.646±0.014	<u>0.721±0.009</u>	<b>0.739±0.005</b>	
<b>TDC.CYP3A4-I</b> (↑)	AUPRC	0.827±0.009	<u>0.829±0.007</u>	0.821±0.003	0.849±0.004	0.840±0.010	0.851±0.006	<u>0.902±0.002</u>	<b>0.904±0.002</b>	
<b>TDC.CYP2C9-I</b> (↑)	AUPRC	0.715±0.004	<u>0.742±0.006</u>	0.713±0.006	0.739±0.010	0.735±0.004	0.749±0.004	<u>0.829±0.003</u>	<b>0.839±0.003</b>	
<b>TDC.CYP2D6-S</b> (↑)	AUPRC	0.671±0.066	<u>0.677±0.047</u>	0.485±0.037	0.572±0.062	0.617±0.039	0.574±0.030	<u>0.704±0.028</u>	<b>0.736±0.024</b>	
<b>TDC.CYP3A4-S</b> (↑)	AUROC	0.633±0.013	<u>0.639±0.012</u>	<b>0.662±0.031</b>	0.578±0.020	0.590±0.023	0.576±0.025	<u>0.582±0.021</u>	0.609±0.025	
<b>TDC.CYP2C9-S</b> (↑)	AUPRC	0.380±0.015	<u>0.360±0.040</u>	0.367±0.059	0.359±0.059	0.344±0.051	0.375±0.032	<u>0.381±0.045</u>	<b>0.392±0.026</b>	
<b>TDC.Half_Life</b> (↑)	Spearman	<b>0.329±0.083</b>	0.184±0.111	0.038±0.138	0.177±0.165	0.239±0.100	0.085±0.068	0.151±0.068	0.129±0.114	
<b>TDC.CL-Micro</b> (↑)	Spearman	0.492±0.020	<b>0.586±0.014</b>	0.252±0.116	0.529±0.015	<u>0.532±0.033</u>	0.365±0.055	<u>0.585±0.034</u>	<u>0.578±0.007</u>	
<b>TDC.CL-Hepa</b> (↑)	Spearman	0.272±0.068	<u>0.382±0.007</u>	0.235±0.021	0.401±0.037	0.366±0.063	0.289±0.022	<u>0.413±0.028</u>	<b>0.439±0.026</b>	
<b>TDC.hERG</b> (↑)	AUROC	0.736±0.023	<b>0.841±0.020</b>	0.754±0.037	0.722±0.034	0.738±0.038	<u>0.825±0.007</u>	<u>0.778±0.046</u>	0.756±0.023	
<b>TDC.AMES</b> (↑)	AUROC	0.794±0.008	<u>0.823±0.011</u>	0.776±0.015	0.823±0.006	0.818±0.010	<u>0.814±0.008</u>	<b>0.842±0.008</b>	0.837±0.009	
<b>TDC.DILI</b> (↑)	AUROC	0.832±0.021	<u>0.875±0.019</u>	0.792±0.016	0.851±0.026	0.859±0.033	<u>0.886±0.015</u>	<b>0.919±0.008</b>	0.861±0.018	
<b>TDC.LD50</b> (↓)	MAE	0.649±0.019	<u>0.678±0.003</u>	0.675±0.011	0.667±0.020	0.649±0.026	<u>0.678±0.012</u>	<b>0.685±0.025</b>	0.669±0.030	

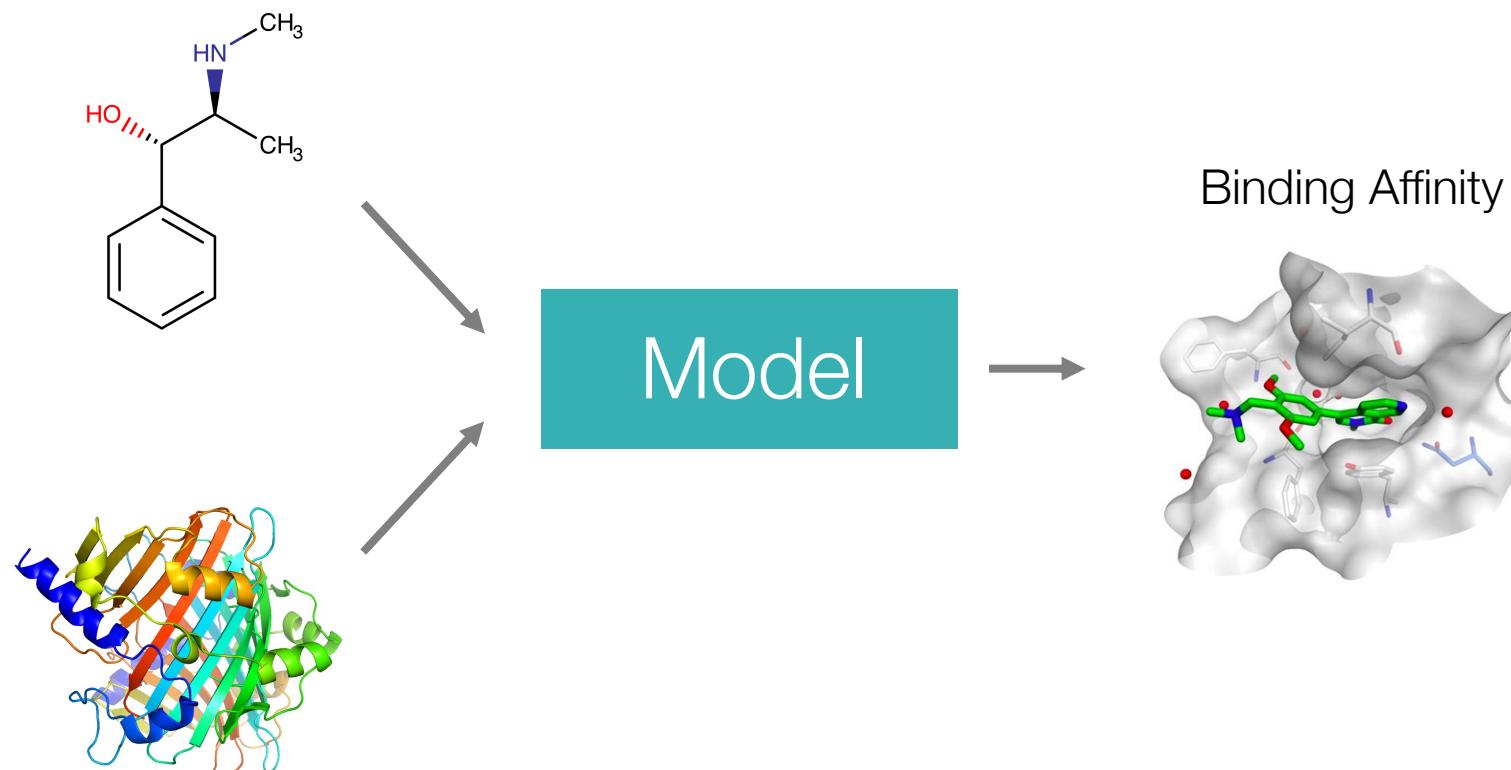
- Finding 2: Expert-curated methods, such as Morgan's fingerprints can outperform graph RL methods on some endpoints

# Results: ADMET prediction (3/3)

Raw Feature Type		Expert-Curated Methods		SMILES	Molecular Graph-Based Methods (state-of-the-Art in ML)				
Dataset	Metric	Morgan [31]	RDKit2D [24]	CNN [18]	NeuralFP [7]	GCN [23]	AttentiveFP [43]	AttrMasking [16]	ContextPred [16]
	# Params.	1477K	633K	227K	480K	192K	301K	2067K	2067K
<b>TDC.Caco2</b> (↓)	MAE	0.908±0.060	<b>0.393±0.024</b>	0.446±0.036	0.530±0.102	0.599±0.104	<u>0.401±0.032</u>	0.546±0.052	0.502±0.036
<b>TDC.HIA</b> (↑)	AUROC	0.807±0.072	0.972±0.008	0.869±0.026	0.943±0.014	0.936±0.024	<u>0.974±0.007</u>	<b>0.978±0.006</b>	0.975±0.004
<b>TDC.Pgp</b> (↑)	AUROC	0.880±0.006	0.918±0.007	0.908±0.012	0.902±0.020	0.895±0.021	0.892±0.012	<b>0.929±0.006</b>	0.923±0.005
<b>TDC.Bioav</b> (↑)	AUROC	0.581±0.086	<b>0.672±0.021</b>	0.613±0.013	0.632±0.036	0.566±0.115	0.632±0.039	0.577±0.087	0.671±0.026
<b>TDC.Lipo</b> (↓)	MAE	0.701±0.009	0.574±0.017	0.743±0.020	0.563±0.023	<u>0.541±0.011</u>	0.572±0.007	0.547±0.024	<b>0.535±0.012</b>
<b>TDC.AqSol</b> (↓)	MAE	1.203±0.019	<u>0.827±0.047</u>	1.023±0.023	0.947±0.016	0.907±0.020	<b>0.776±0.008</b>	1.026±0.020	1.040±0.045
<b>TDC.BBB</b> (↑)	AUROC	0.823±0.015	<b>0.889±0.016</b>	0.781±0.030	0.836±0.009	0.842±0.016	0.855±0.011	<u>0.892±0.012</u>	<b>0.897±0.004</b>
<b>TDC.PPBR</b> (↓)	MAE	12.848±0.362	9.994±0.319	11.106±0.358	<b>9.292±0.384</b>	10.194±0.373	<u>9.373±0.335</u>	<u>10.075±0.202</u>	9.445±0.224
<b>TDC.VD</b> (↑)	Spearman	0.493±0.011	<b>0.561±0.025</b>	0.226±0.114	0.258±0.162	0.457±0.050	<u>0.241±0.145</u>	<u>0.559±0.019</u>	0.485±0.092
<b>TDC.CYP2D6-I</b> (↑)	AUPRC	0.587±0.011	0.616±0.007	0.544±0.053	0.627±0.009	0.616±0.020	0.646±0.014	<u>0.721±0.009</u>	<b>0.739±0.005</b>
<b>TDC.CYP3A4-I</b> (↑)	AUPRC	0.827±0.009	0.829±0.007	0.821±0.003	0.849±0.004	0.840±0.010	0.851±0.006	<u>0.902±0.002</u>	<b>0.904±0.002</b>
<b>TDC.CYP2C9-I</b> (↑)	AUPRC	0.715±0.004	0.742±0.006	0.713±0.006	0.739±0.010	0.735±0.004	0.749±0.004	<u>0.829±0.003</u>	<b>0.839±0.003</b>
<b>TDC.CYP2D6-S</b> (↑)	AUPRC	0.671±0.066	0.677±0.047	0.485±0.037	0.572±0.062	0.617±0.039	0.574±0.030	<u>0.704±0.028</u>	<b>0.736±0.024</b>
<b>TDC.CYP3A4-S</b> (↑)	AUROC	0.633±0.013	<u>0.639±0.012</u>	<b>0.662±0.031</b>	0.578±0.020	0.590±0.023	0.576±0.025	<u>0.582±0.021</u>	0.609±0.025
<b>TDC.CYP2C9-S</b> (↑)	AUPRC	0.380±0.015	0.360±0.040	0.367±0.059	0.359±0.059	0.344±0.051	0.375±0.032	<u>0.381±0.045</u>	<b>0.392±0.026</b>
<b>TDC.Half_Life</b> (↑)	Spearman	<b>0.329±0.083</b>	0.184±0.111	0.038±0.138	0.177±0.165	0.239±0.100	0.085±0.068	0.151±0.068	0.129±0.114
<b>TDC.CL-Micro</b> (↑)	Spearman	0.492±0.020	<b>0.586±0.014</b>	0.252±0.116	0.529±0.015	<u>0.532±0.033</u>	0.365±0.055	<u>0.585±0.034</u>	0.578±0.007
<b>TDC.CL-Hepa</b> (↑)	Spearman	0.272±0.068	0.382±0.007	0.235±0.021	0.401±0.037	0.366±0.063	0.289±0.022	<u>0.413±0.028</u>	<b>0.439±0.026</b>
<b>TDC.hERG</b> (↑)	AUROC	0.736±0.023	<b>0.841±0.020</b>	0.754±0.037	0.722±0.034	0.738±0.038	<u>0.825±0.007</u>	0.778±0.046	0.756±0.023
<b>TDC.AMES</b> (↑)	AUROC	0.794±0.008	0.823±0.011	0.776±0.015	0.823±0.006	0.818±0.010	<u>0.814±0.008</u>	<b>0.842±0.008</b>	<u>0.837±0.009</u>
<b>TDC.DILI</b> (↑)	AUROC	0.832±0.021	0.875±0.019	0.792±0.016	0.851±0.026	0.859±0.033	<u>0.886±0.015</u>	<b>0.919±0.008</b>	0.861±0.018
<b>TDC.LD50</b> (↓)	MAE	0.649±0.019	<u>0.678±0.003</u>	0.675±0.011	0.667±0.020	0.649±0.026	<u>0.678±0.012</u>	<b>0.685±0.025</b>	0.669±0.030

- Finding 3: Pre-training can be helpful. Pre-trained graph RL models yield strongest predictors overall

# Drug-target interaction prediction

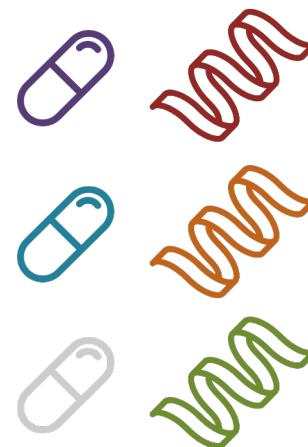
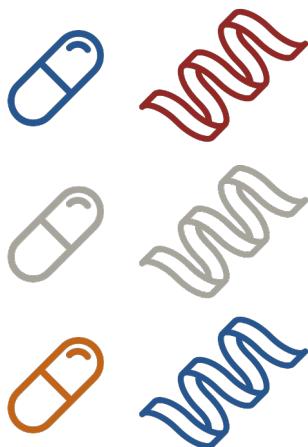


# Setup: Distribution shifts and generalization

DTI datasets are typically split into train/validation/test sets in a random manner. Identifying drug targets in the real-world, however, requires generalization to novel drugs and proteins.



A domain generalization problem!



Train-Valid: DTIs Patented in 2013-18

Test: DTIs Patented in 2019-21

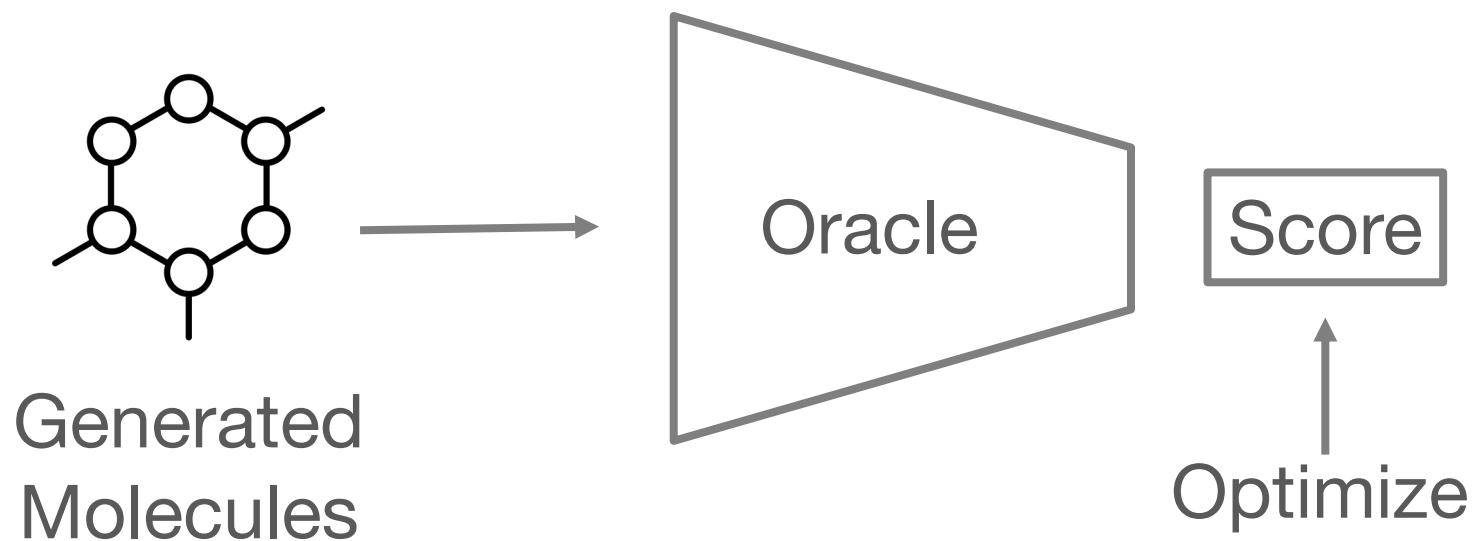
# Results



**ERM (Empirical Risk Minimization)** is a standard training strategy where errors across all domains are minimized. State-of-the-art domain generalization methods: **MMD (Maximum Mean Discrepancy)** optimizes similarities between predicted and observed values using maximum mean discrepancy score across domains. **CORAL (Correlation Alignment)** matches the mean and covariance of features across domains. **IRM (Invariant Risk Minimization)** optimizes features using a cross-domain optimized linear classifier. **GroupDRO (distributionally robust neural networks for group shifts)** optimizes ERM and adjusts weights of domains with larger errors. **MTL (marginal transfer learning)** concatenates original features with an augmented vector of marginal feature distributions. **ANDMask** masks gradients that have inconsistent signs in the corresponding weights across domains

- **Finding 1:** OOD (Out-of-distribution) performance drops from 33.9%-43.6%.
- **Finding 2:** Standard supervised models have similar performance as state-of-the-art domain generalization methods.

# Molecule generation



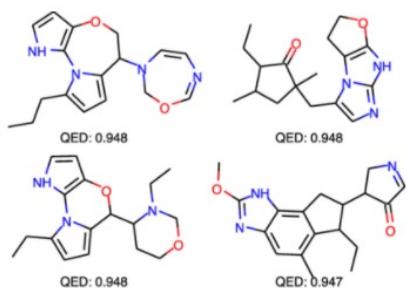
# Setup: High-capacity oracles (1/2)

Real-world oracles (e.g., bioassays and experimental validation of predictions) are expensive and resource-intensive



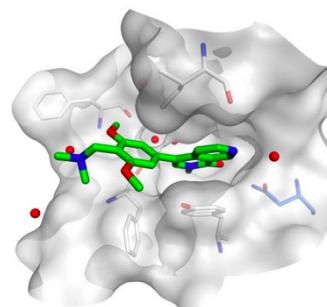
Molecule generation given a small budget,  
i.e., limited number of oracle calls!

Previous oracle



**Milliseconds in RDKit**  
SOTA methods call millions of times!

Docking oracle



vs

**Minutes in Vina**  
Restricted to thousands of calls only!

# Setup: High-capacity oracles (2/2)

Optimizing for a single target property is not sufficient. It does not generate molecules with many drug-like properties



We need effective indicators of performance of these methods in real-world scenarios

**Established performance metrics:**

Top100/Top10/Top1 docking scores, Diversity, Novelty

**Additional performance metrics:**

Synthesizability with Molecule.One\*

% Pass filters (PAINS/SureChEMBL/Glaxo)

# Results: Docking molecule generation (1/3)

Method Category			Domain-Specific Methods		State-of-the-Art Methods in ML			
Metric	Best-in-data	# Calls	Screening	Graph-GA [20]	LSTM [34]	GCPN [45]	MolDQN [46]	MARS [42]
# Params.	-	-	0	0	3149K	18K	2694K	153K
Top100 (↓)	-12.080		-9.693±0.019	<b>-11.224±0.484</b>	-9.971±0.115	-9.053±0.080	-6.738±0.042	-8.224±0.196
Top10 (↓)	-12.590		-10.777±0.189	<b>-12.400±0.782</b>	-11.163±0.141	-11.027±0.273	-7.506±0.085	-9.843±0.068
Top1 (↓)	-12.800		-11.500±0.432	<b>-13.233±0.713</b>	-11.967±0.205	-12.033±0.618	-7.800±0.042	-11.100±0.141
Diversity (↑)	0.864	1000	0.873±0.003	0.815±0.046	0.871±0.004	<b>0.913±0.001</b>	0.904±0.001	0.871±0.004
Novelty (↑)	-		-	1.000±0.000	1.000±0.000	1.000±0.000	1.000±0.000	1.000±0.000
%Pass (↑)	0.780		0.757±0.026	<b>0.777±0.096</b>	0.777±0.026	0.170±0.022	0.033±0.005	0.563±0.052
Top1 Pass (↓)	-11.700		-9.167±0.047	<b>-10.600±0.374</b>	-9.367±0.094	-8.167±0.047	-6.450±0.085	-7.367±0.205
m1 (↓)	5.100		<u>5.527±0.780</u>	7.695±0.909	<b>4.818±0.541</b>	10.000±0.000	10.000±0.000	6.037±0.137
Top100 (↓)	-12.080	5000	-10.542±0.035	<b>-14.811±0.413</b>	-13.017±0.385	-10.045±0.226	-8.236±0.089	-9.509±0.035
Top10 (↓)	-12.590		-11.483±0.056	<b>-15.930±0.336</b>	-14.030±0.421	-11.483±0.581	-9.348±0.188	-10.693±0.172
Top1 (↓)	-12.800		-12.100±0.356	<b>-16.533±0.309</b>	-14.533±0.525	-12.300±0.993	-9.990±0.194	-11.433±0.450
Diversity (↑)	0.864		0.872±0.003	0.626±0.092	0.740±0.056	<b>0.922±0.002</b>	0.893±0.005	0.873±0.002
Novelty (↑)	-		-	1.000±0.000	1.000±0.000	1.000±0.000	1.000±0.000	1.000±0.000
%Pass (↑)	0.780		<b>0.683±0.073</b>	0.393±0.308	0.257±0.103	0.167±0.045	0.023±0.012	<u>0.527±0.087</u>
Top1 Pass (↓)	-11.700		-10.100±0.000	<b>-14.267±0.450</b>	-12.533±0.403	-9.367±0.170	-7.980±0.112	-9.000±0.082
m1 (↓)	5.100		<b>5.610±0.805</b>	9.669±0.468	<u>5.826±1.908</u>	10.000±0.000	10.000±0.000	7.073±0.798

- **Finding 1:** Models perform poorly in challenging yet realistic setting (i.e., they do not beat best-in-data reference when they are given 1,000 # calls)

# Results: Docking molecule generation (2/3)

Method Category		Domain-Specific Methods			State-of-the-Art Methods in ML			
Metric	Best-in-data	# Calls	Screening	Graph-GA [20]	LSTM [34]	GCPN [45]	MolDQN [46]	MARS [42]
# Params.	-	-	0	0	3149K	18K	2694K	153K
Top100 (↓)	-12.080	1000	-9.693±0.019	<b>-11.224±0.484</b>	<u>-9.971±0.115</u>	<u>-9.053±0.080</u>	<u>-6.738±0.042</u>	<u>-8.224±0.196</u>
Top10 (↓)	-12.590		-10.777±0.189	<b>-12.400±0.782</b>	<u>-11.163±0.141</u>	<u>-11.027±0.273</u>	<u>-7.506±0.085</u>	<u>-9.843±0.068</u>
Top1 (↓)	-12.800		-11.500±0.432	<b>-13.233±0.713</b>	<u>-11.967±0.205</u>	<u>-12.033±0.618</u>	<u>-7.800±0.042</u>	<u>-11.100±0.141</u>
Diversity (↑)	0.864		0.873±0.003	0.815±0.046	0.871±0.004	<b>0.913±0.001</b>	<u>0.904±0.001</u>	<u>0.871±0.004</u>
Novelty (↑)	-		-	1.000±0.000	1.000±0.000	1.000±0.000	1.000±0.000	1.000±0.000
%Pass (↑)	0.780		0.757±0.026	<b>0.777±0.096</b>	<u>0.777±0.026</u>	0.170±0.022	0.033±0.005	0.563±0.052
Top1 Pass (↓)	-11.700		-9.167±0.047	<b>-10.600±0.374</b>	<u>-9.367±0.094</u>	<u>-8.167±0.047</u>	<u>-6.450±0.085</u>	<u>-7.367±0.205</u>
m1 (↓)	5.100		5.527±0.780	7.695±0.909	<b>4.818±0.541</b>	10.000±0.000	10.000±0.000	6.037±0.137
Top100 (↓)	-12.080	5000	-10.542±0.035	<b>-14.811±0.413</b>	<u>-13.017±0.385</u>	<u>-10.045±0.226</u>	<u>-8.236±0.089</u>	<u>-9.509±0.035</u>
Top10 (↓)	-12.590		-11.483±0.056	<b>-15.930±0.336</b>	<u>-14.030±0.421</u>	<u>-11.483±0.581</u>	<u>-9.348±0.188</u>	<u>-10.693±0.172</u>
Top1 (↓)	-12.800		-12.100±0.356	<b>-16.533±0.309</b>	<u>-14.533±0.525</u>	<u>-12.300±0.993</u>	<u>-9.990±0.194</u>	<u>-11.433±0.450</u>
Diversity (↑)	0.864		0.872±0.003	0.626±0.092	0.740±0.056	<b>0.922±0.002</b>	<u>0.893±0.005</u>	<u>0.873±0.002</u>
Novelty (↑)	-		-	1.000±0.000	1.000±0.000	1.000±0.000	1.000±0.000	1.000±0.000
%Pass (↑)	0.780		<b>0.683±0.073</b>	0.393±0.308	0.257±0.103	0.167±0.045	0.023±0.012	<u>0.527±0.087</u>
Top1 Pass (↓)	-11.700		-10.100±0.000	<b>-14.267±0.450</b>	<u>-12.533±0.403</u>	<u>-9.367±0.170</u>	<u>-7.980±0.112</u>	<u>-9.000±0.082</u>
m1 (↓)	5.100		<b>5.610±0.805</b>	9.669±0.468	<u>5.826±1.908</u>	10.000±0.000	10.000±0.000	7.073±0.798

- Finding 2: Graph-GA method with 0 learnable parameters performs the best. SOTA ML methods report excellent results when resources are unlimited

# Results: Docking molecule generation (3/3)

Method Category		Domain-Specific Methods		State-of-the-Art Methods in ML				
Metric	Best-in-data	# Calls	Screening	Graph-GA [20]	LSTM [34]	GCPN [45]	MolDQN [46]	MARS [42]
# Params.	-	-	0	0	3149K	18K	2694K	153K
Top100 (↓)	-12.080	1000	-9.693±0.019	<b>-11.224±0.484</b>	<u>-9.971±0.115</u>	<u>-9.053±0.080</u>	<u>-6.738±0.042</u>	<u>-8.224±0.196</u>
Top10 (↓)	-12.590		-10.777±0.189	<b>-12.400±0.782</b>	<u>-11.163±0.141</u>	<u>-11.027±0.273</u>	<u>-7.506±0.085</u>	<u>-9.843±0.068</u>
Top1 (↓)	-12.800		-11.500±0.432	<b>-13.233±0.713</b>	<u>-11.967±0.205</u>	<u>-12.033±0.618</u>	<u>-7.800±0.042</u>	<u>-11.100±0.141</u>
Diversity (↑)	0.864		0.873±0.003	0.815±0.046	0.871±0.004	<b>0.913±0.001</b>	<u>0.904±0.001</u>	<u>0.871±0.004</u>
Novelty (↑)	-		-	1.000±0.000	1.000±0.000	1.000±0.000	1.000±0.000	1.000±0.000
%Pass (↑)	0.780		0.757±0.026	<b>0.777±0.096</b>	<u>0.777±0.026</u>	0.170±0.022	0.033±0.005	<b>0.563±0.052</b>
Top1 Pass (↓)	-11.700		-9.167±0.047	<b>-10.600±0.374</b>	<u>-9.367±0.094</u>	<u>-8.167±0.047</u>	<u>-6.450±0.085</u>	<u>-7.367±0.205</u>
m1 (↓)	5.100		5.527±0.780	<u>7.695±0.909</u>	<b>4.818±0.541</b>	10.000±0.000	10.000±0.000	<u>6.037±0.137</u>
Top100 (↓)	-12.080	5000	-10.542±0.035	<b>-14.811±0.413</b>	<u>-13.017±0.385</u>	<u>-10.045±0.226</u>	<u>-8.236±0.089</u>	<u>-9.509±0.035</u>
Top10 (↓)	-12.590		-11.483±0.056	<b>-15.930±0.336</b>	<u>-14.030±0.421</u>	<u>-11.483±0.581</u>	<u>-9.348±0.188</u>	<u>-10.693±0.172</u>
Top1 (↓)	-12.800		-12.100±0.356	<b>-16.533±0.309</b>	<u>-14.533±0.525</u>	<u>-12.300±0.993</u>	<u>-9.990±0.194</u>	<u>-11.433±0.450</u>
Diversity (↑)	0.864		0.872±0.003	0.626±0.092	0.740±0.056	<b>0.922±0.002</b>	<u>0.893±0.005</u>	<u>0.873±0.002</u>
Novelty (↑)	-		-	1.000±0.000	1.000±0.000	1.000±0.000	1.000±0.000	1.000±0.000
%Pass (↑)	0.780		<b>0.683±0.073</b>	<u>0.393±0.308</u>	0.257±0.103	0.167±0.045	0.023±0.012	<b>0.527±0.087</b>
Top1 Pass (↓)	-11.700		-10.100±0.000	<b>-14.267±0.450</b>	<u>-12.533±0.403</u>	<u>-9.367±0.170</u>	<u>-7.980±0.112</u>	<u>-9.000±0.082</u>
m1 (↓)	5.100		<b>5.610±0.805</b>	<u>9.669±0.468</u>	<u>5.826±1.908</u>	10.000±0.000	10.000±0.000	<u>7.073±0.798</u>

- Finding 3: The greater the number of calls, the worse the quality of generated molecules (drug-likeness)

# Machine learning foundation for therapeutics



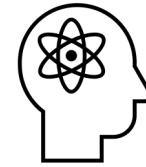
Domain  
scientists

Identify meaningful  
tasks and datasets

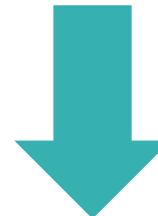


THERAPEUTICS  
DATA COMMONS

Design  
AI/ML methods



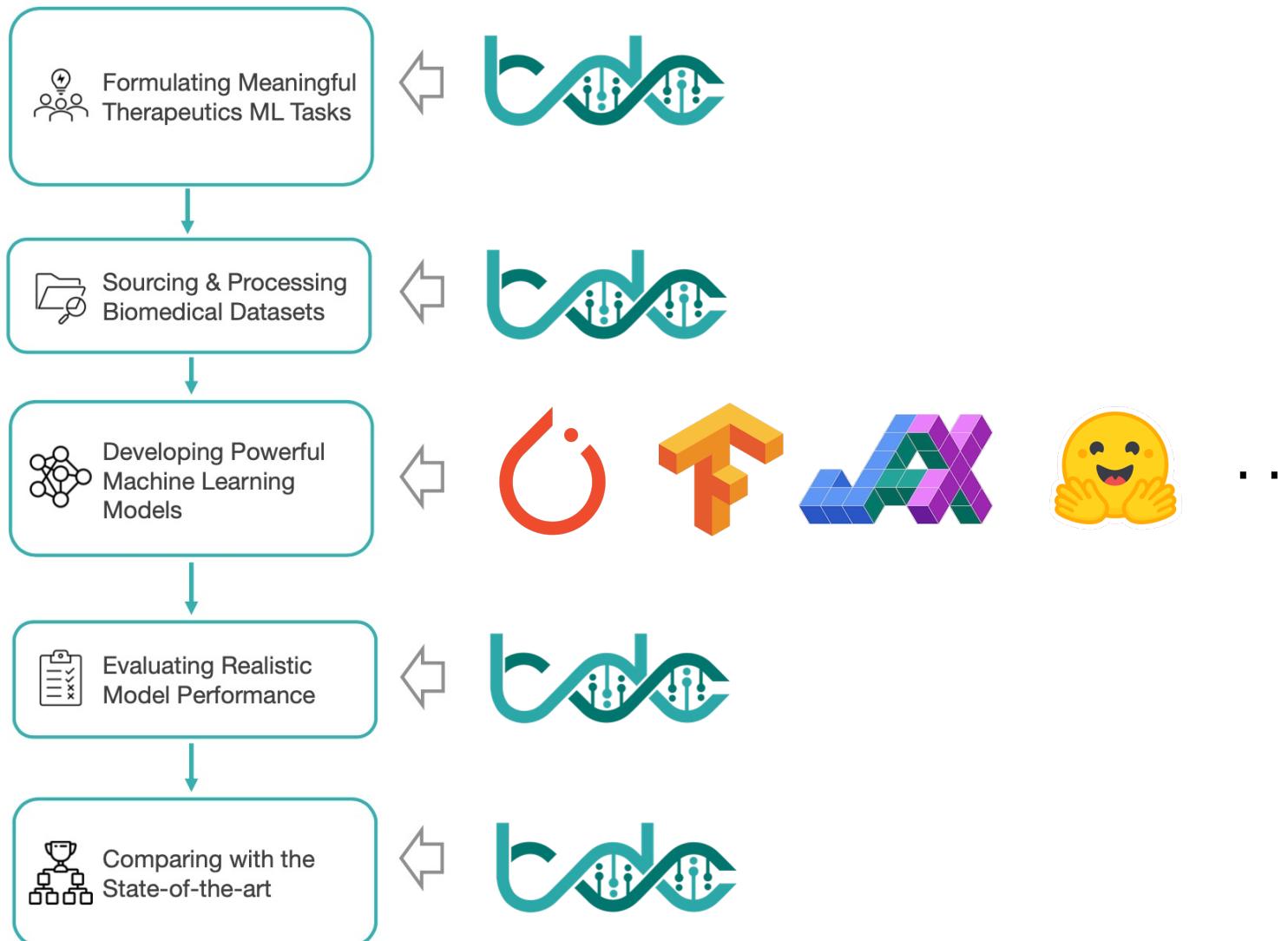
AI/ML  
scientists



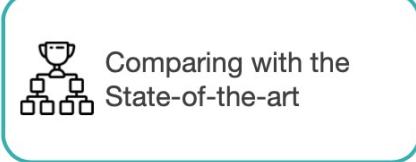
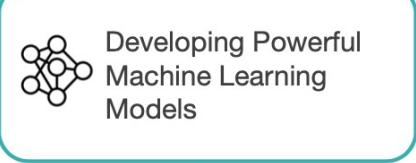
Facilitate algorithmic and scientific advance  
in therapeutics

TDC supports the development of novel ML theory and methods, with a strong bent towards developing the mathematical foundations of which ML algorithms are most suitable for drug discovery applications and why

# Lifecycle of therapeutics ML



## Lifecycle of Therapeutics Machine Learning



### 22 Therapeutics ML Tasks



### 66 Therapeutics ML-ready Datasets

Drug_ID	Drug	Y
CHEMBL15932	COc1cccc2[nH]ncc12	2.10
CHEMBL1527751	Oc1ncncc2scc(-c3ccsc3)c12	2.25

### TDC Data Functions

- 4 Realistic TDC Data Splits Functions
- 17 TDC Molecule Generation Oracles
- 11 TDC Data Processing Helpers

### 23 TDC Evaluator Functions

- |                               |                            |
|-------------------------------|----------------------------|
| Regression: <b>6 Metrics</b>  | Binary: <b>8 Metrics</b>   |
| Multi-class: <b>3 Metrics</b> | Molecule: <b>6 Metrics</b> |

### TDC Leaderboards

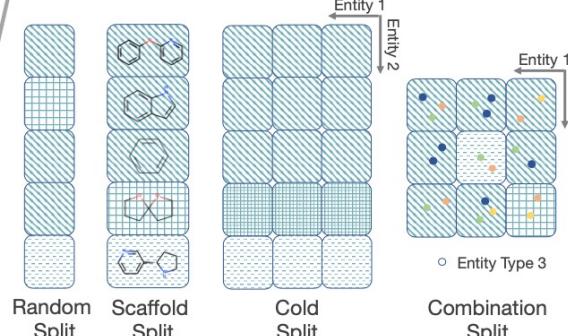
#### 22 ADMET Group Benchmarks

#### 5 Drug Combination Group Benchmarks

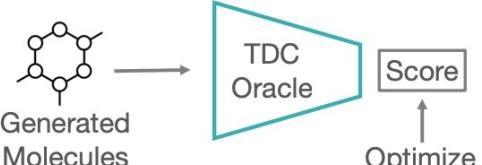
#### 4 Docking Score Molecule Generation Benchmarks

### TDC Data Splits Functions

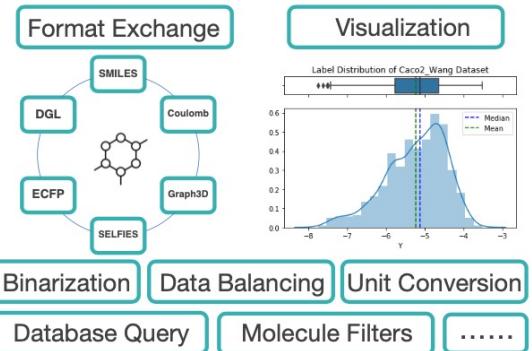
Train Valid Test



### TDC Molecule Generation Oracles



### TDC Data Processing Helpers



# Key Takeaways

- TDC provides an artificial intelligence foundation for therapeutic science
  - **Python package:** Tools, libraries, leaderboards, and resources, including data functions, strategies for systematic model evaluation, meaningful data splits, data processors, and molecule generation oracles
  - **AI-ready datasets** cover a range of therapeutic modalities, including small molecules, biologics, antibodies, peptides, miRNAs, and gene therapies
  - **Solvable AI tasks** cover all stages of drug discovery:
    - **Target discovery:** Tasks to identify candidate therapeutic targets
    - **Activity modeling:** Tasks to screen and generate individual or combinatorial candidates with high binding activity
    - **Efficacy and safety:** Optimize signatures indicative of safety and efficacy
    - **Manufacturing:** Tasks on the manufacturing and synthesis of therapeutics
- Resources
  - Website: <https://tdcommons.ai>
  - Paper: <https://arxiv.org/abs/2102.09548>
  - GitHub: <https://github.com/mims-harvard/TDC>

Applications of graph representation learning on...

# THERAPEUTICS

1. Molecular property prediction, drug-target interaction prediction, molecular generation
2. Drug discovery
3. Drug repurposing

Applications of graph representation learning on...

# THERAPEUTICS

1. Molecular property prediction, drug-target interaction prediction, molecular generation
2. **Drug discovery**
3. Drug repurposing

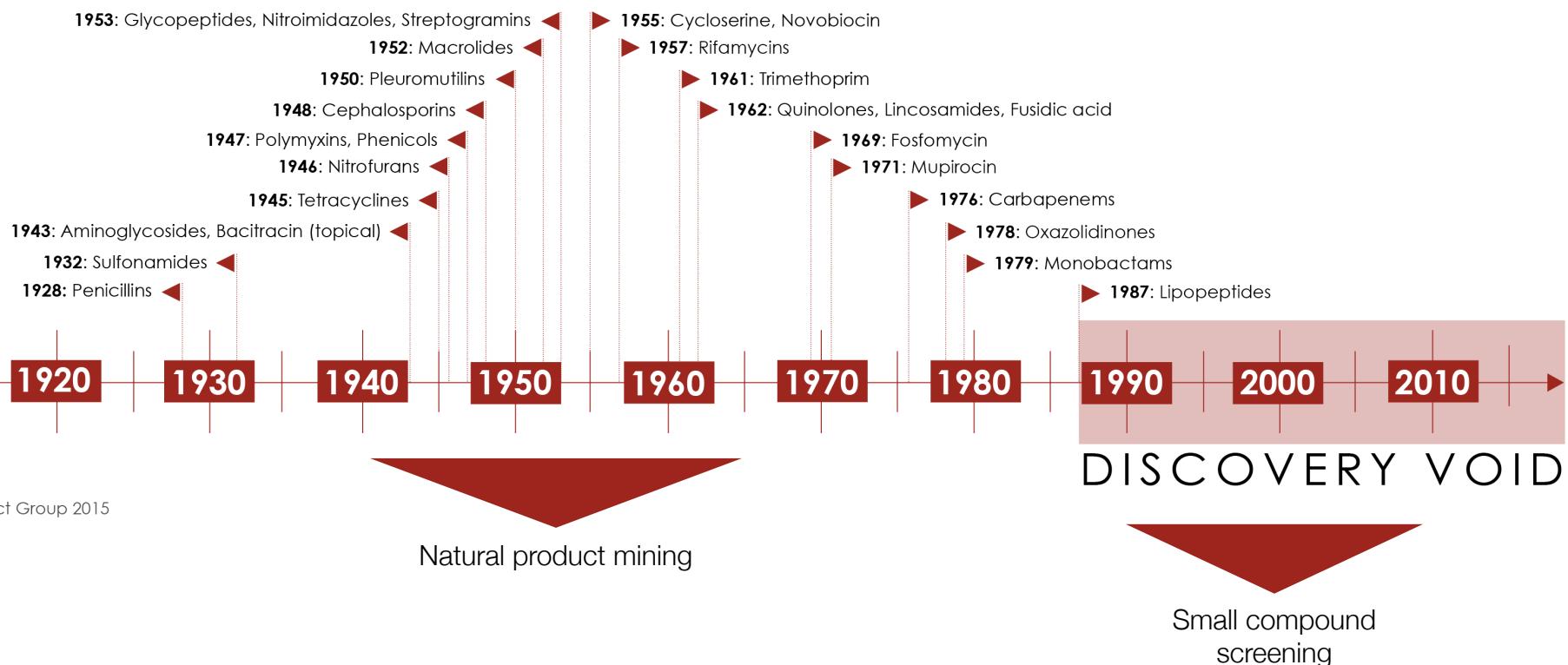
The image shows a screenshot of a journal article from the Cell journal. The header features the Cell logo and the word "Cell". Below the header, the text "ARTICLE | VOLUME 180, ISSUE 4, P688-702.E13, FEBRUARY 20, 2020" is displayed. The main title of the article is "A Deep Learning Approach to Antibiotic Discovery". Below the title, the author list includes Jonathan M. Stokes, Kevin Yang<sup>10</sup>, Kyle Swanson<sup>10</sup>, Tommi S. Jaakkola, Regina Barzilay, James J. Collins<sup>11</sup>, and links to "Show all authors" and "Show footnotes".

ARTICLE | VOLUME 180, ISSUE 4, P688-702.E13, FEBRUARY 20, 2020

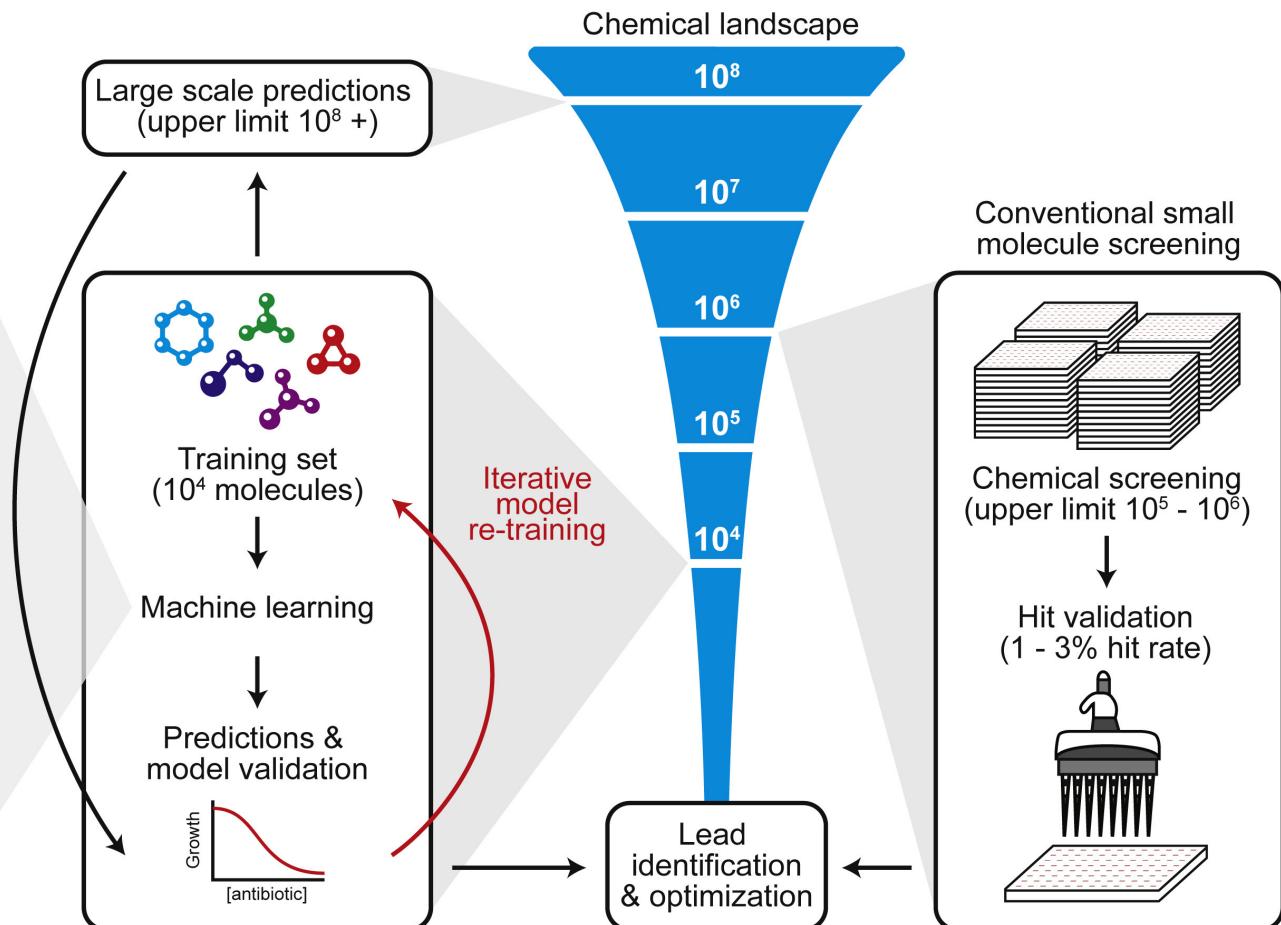
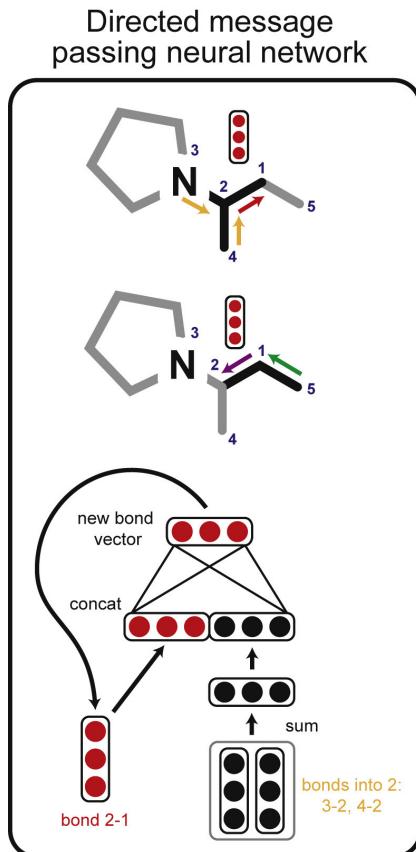
A Deep Learning Approach to Antibiotic Discovery

Jonathan M. Stokes • Kevin Yang<sup>10</sup> • Kyle Swanson<sup>10</sup> • ... Tommi S. Jaakkola • Regina Barzilay • James J. Collins<sup>11</sup> • Show all authors • Show footnotes

# Antibiotic discovery timeline

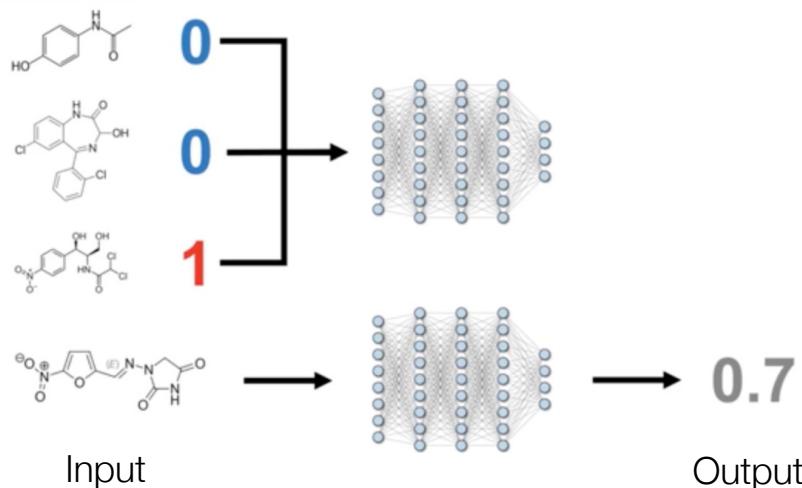


# GNN to learn molecular structure



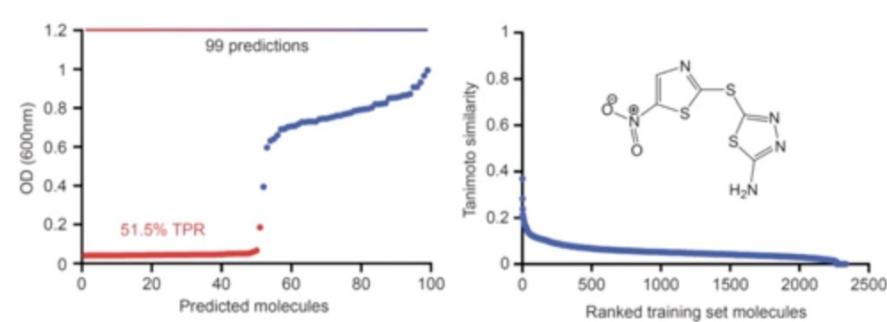
# Experimental setup

**Training Dataset**  
(Human Medicines and Natural Products)



**Data:** 2,335 molecules (human medicines and natural products) screened for growth inhibition

**Empirical Validation**  
(Broad Repurposing Hub)



**Data:** 6,111 molecules (at various stages of investigation for human diseases) in Broad Repurposing Hub

**Task:** Test top 99 predictions & prioritize based on similarity to known antibiotics or predicted toxicity

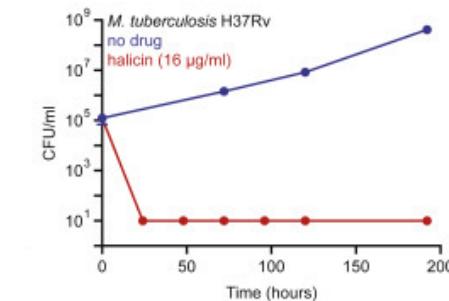
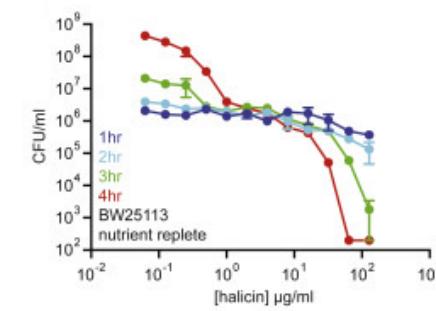
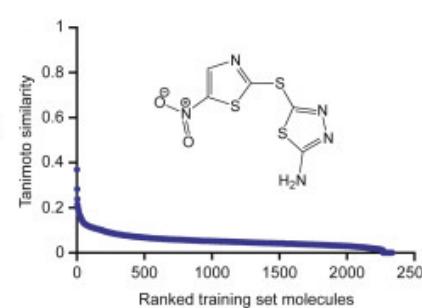
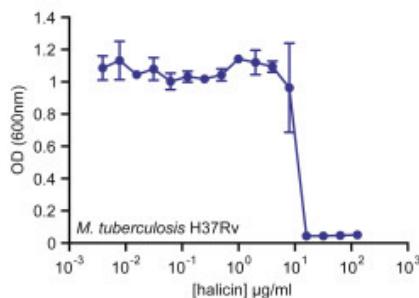
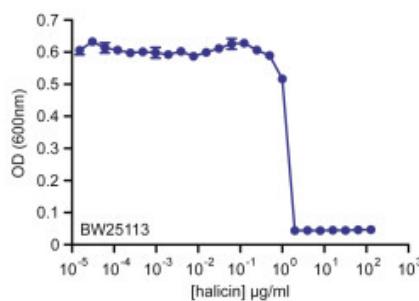
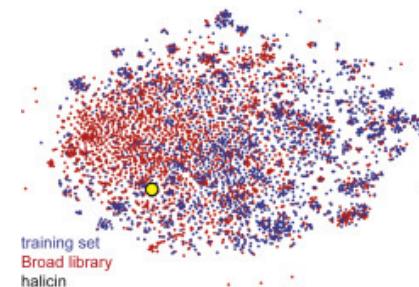
# Results

**Halicin** was developed to be an anti-diabetic drug, but the development was discontinued due to poor results in testing.

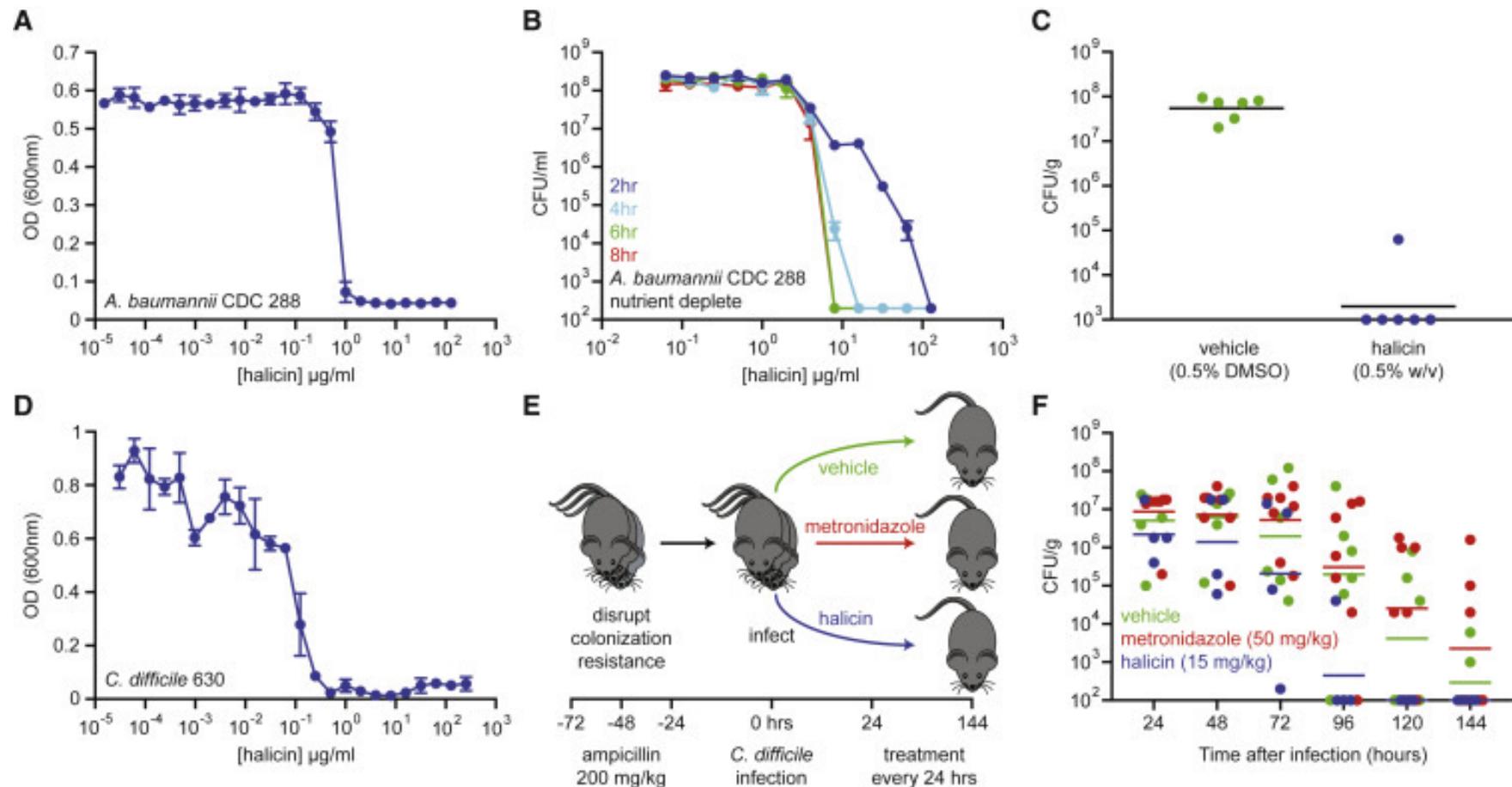
Halicin predicted to be antibacterial

Halicin against *E. coli*

Halicin against *M. tuberculosis*



# Results



# Key Takeaways

- Directed message passing neural network model iteratively (1) learns representations of molecules and (2) optimizes the representations for predicting growth inhibition
- Validated against ~6K molecules in the Broad Repurposing Hub to identify candidate antibiotics
- Halicin, initially developed to be an anti-diabetic drug (but discontinued due to poor results in testing), is identified and verified through experiments as a promising antibiotic
- Resources
  - Paper: [doi.org/10.1016/j.cell.2020.01.021](https://doi.org/10.1016/j.cell.2020.01.021)
  - Chemprop resources:
    - Paper: [doi.org/10.1021/acs.jcim.9b00237](https://doi.org/10.1021/acs.jcim.9b00237)
    - GitHub: [github.com/chemprop/chemprop](https://github.com/chemprop/chemprop)

Applications of graph representation learning on...

# THERAPEUTICS

1. Molecular property prediction, drug-target interaction prediction, molecular generation
2. Drug discovery
3. Drug repurposing

Applications of graph representation learning on...

# THERAPEUTICS

1. Molecular property prediction, drug-target interaction prediction, molecular generation
2. Drug discovery
3. **Drug repurposing**

## Network medicine framework for identifying drug-repurposing opportunities for COVID-19

Deisy Morselli Gysi<sup>a,b,c,1</sup>, Ítalo do Valle<sup>a,b,1</sup>, Marinka Zitnik<sup>d,e,1</sup>, Asher Ameli<sup>b,f,1</sup>, Xiao Gan<sup>a,b,c,1</sup>, Onur Varol<sup>a,b,g,1</sup>, Susan Dina Ghiassian<sup>f,2</sup>, J. J. Patten<sup>h</sup>, Robert A. Davey<sup>h</sup>, Joseph Loscalzo<sup>i</sup>, and Albert-László Barabási<sup>a,b,j,2</sup>

---

**GNNEExplainer: Generating Explanations for Graph Neural Networks**

---

<https://doi.org/10.1038/s41467-021-21770-8>

OPEN

## Identification of disease treatment mechanisms through the multiscale interactome

Camilo Ruiz<sup>1,2</sup>, Marinka Zitnik<sup>3</sup> & Jure Leskovec<sup>1,4</sup>

Rex Ying<sup>†</sup> Dylan Bourgeois<sup>†,‡</sup> Jiaxuan You<sup>†</sup> Marinka Zitnik<sup>†</sup> Jure Leskovec<sup>†</sup>

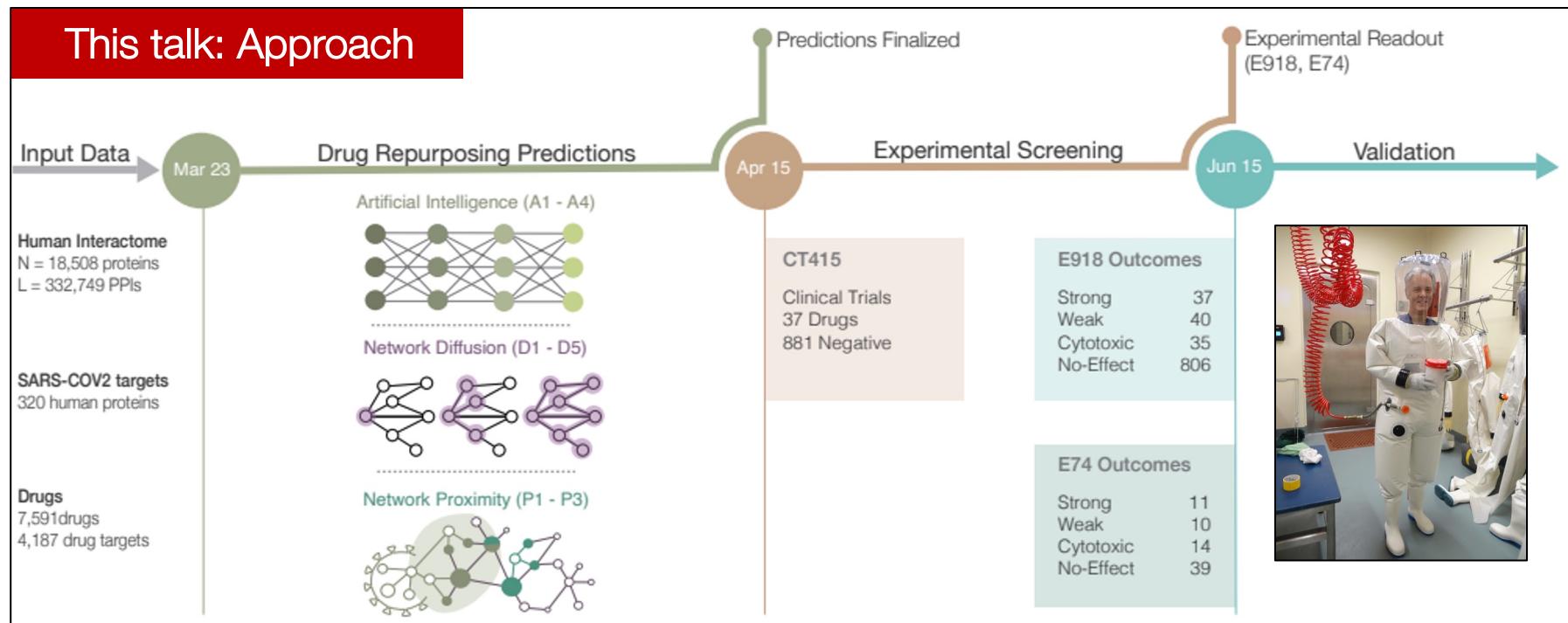
<sup>†</sup>Department of Computer Science, Stanford University

<sup>‡</sup>Robust.AI

{rexying, dtsbourg, jiaxuan, marinka, jure}@cs.stanford.edu

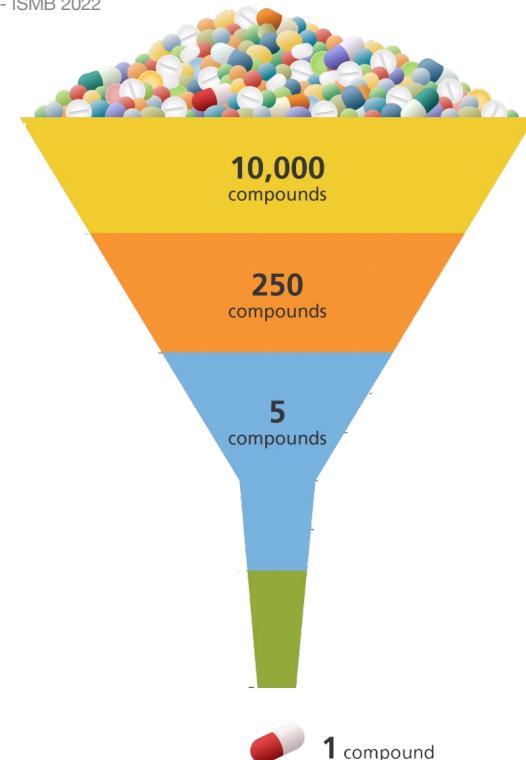
# Rapid therapeutic innovation

- Traditional, iterative development, experimental & clinical testing, and approval of new drugs sometimes not feasible
  - Certain therapeutic areas, public health emergencies
- Challenge:** How to compress years of work into months or even weeks through AI, automation, and new data resources?



# 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.*



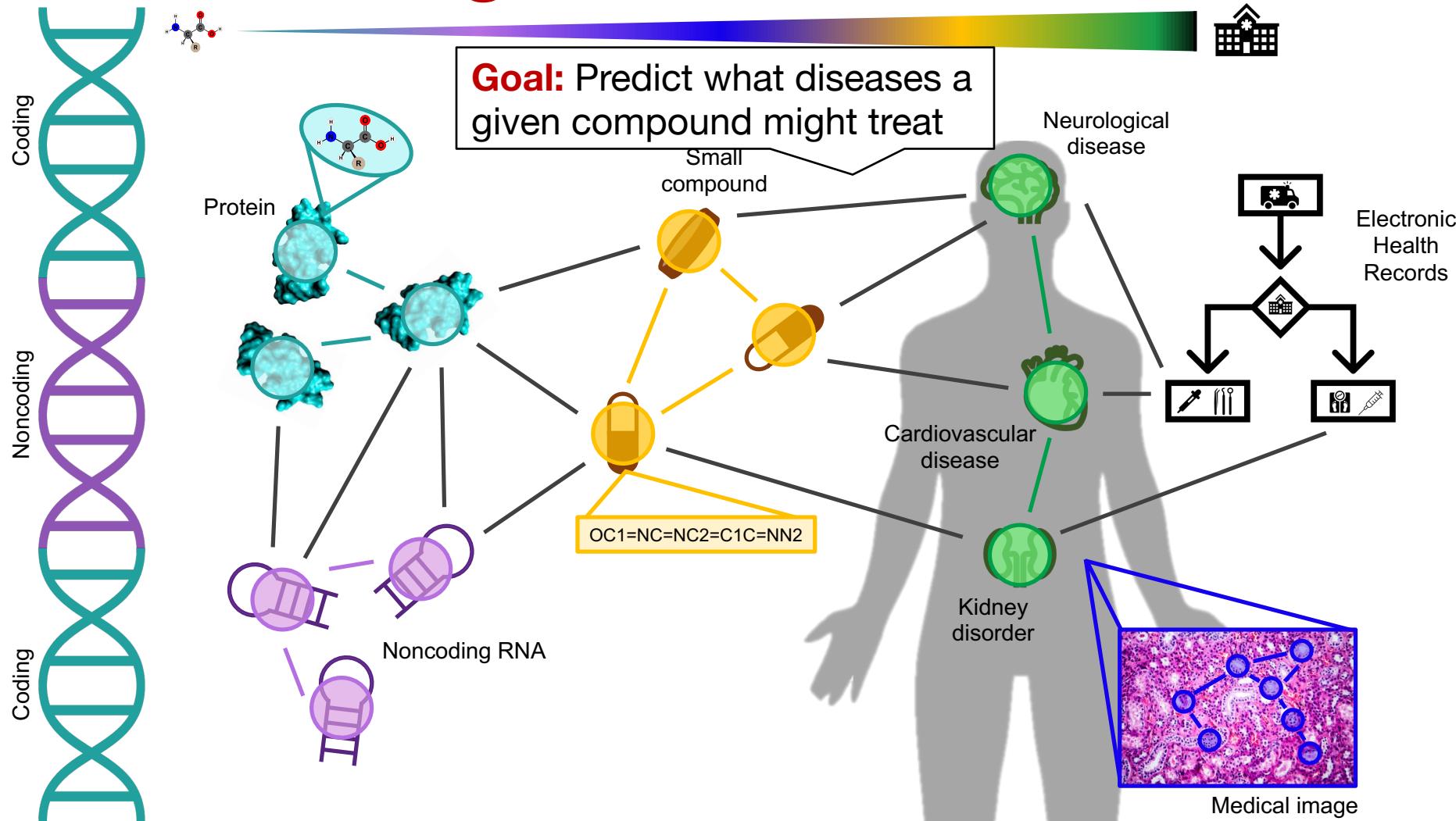
## A SHORTER TIMESCALE

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.

### Drug repositioning

~6 years, ~\$300 million

# What drug treats what disease?



Graph Representation Learning in Biomedicine, *Nature Biomedical Engineering*, 2021 (in press), arXiv:2104.04883

Machine Learning for Integrating Data in Biology and Medicine: Principles, Practice, and Opportunities, *Information Fusion* 2019

Representation Learning for Networks in Biology and Medicine: Advancements, Challenges, and Opportunities, 2021, arXiv:2104.04883

# COVID-19 disease module

Viral-Human  
Protein-Protein Interaction



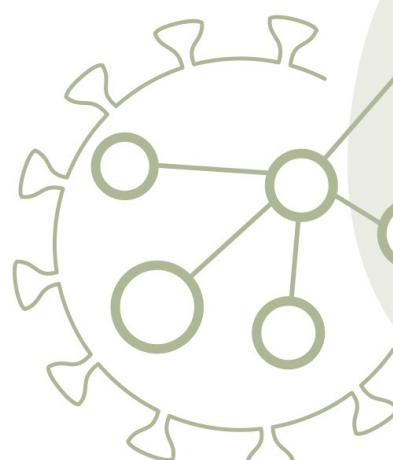
Human-Human  
Protein-Protein Interaction



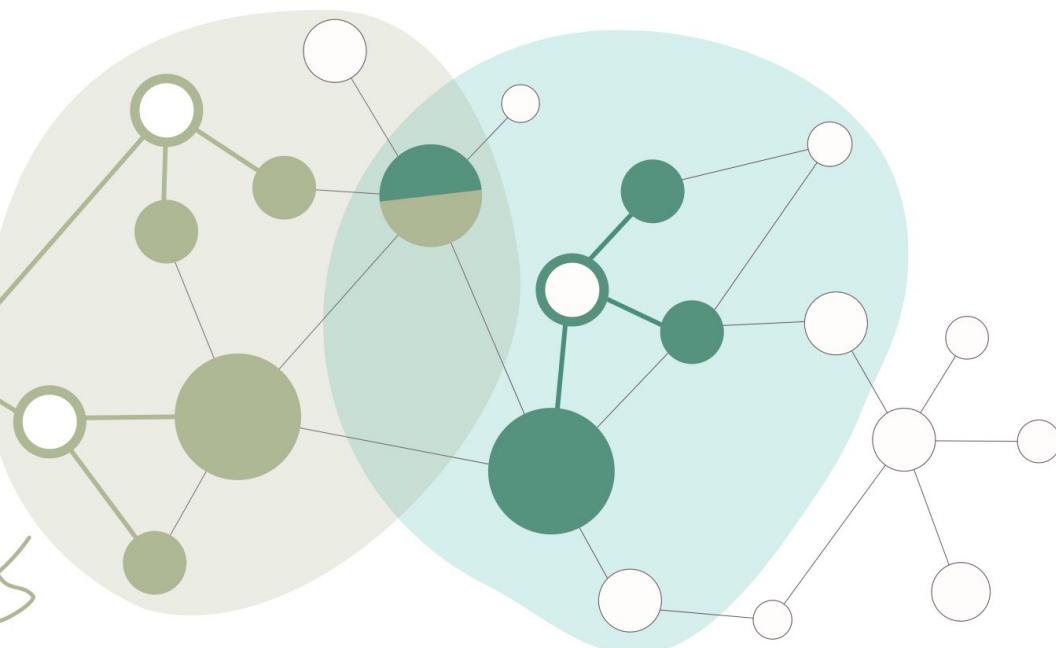
Drug-Human  
Protein-Protein Interaction



Viral Interactome



Human Interactome



Viral Disease Module

Drug Disease Module

**Viral 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<sub>62</sub>, 2021

# Dataset and experimental setup

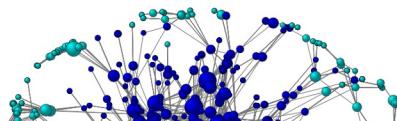
- COVID-19 repurposing knowledge graph:
  - Human protein-protein interaction graph
  - All U.S. approved drugs and proteins they bind to
  - All common diseases and proteins they cause them
  - COVID-19 disease and proteins causing the disease
  - All approved treatments for common diseases
- Goal: Given common diseases and treatments for them, **identify candidate treatments for COVID-19 in a zero-shot manner**

# Why is this task challenging?

**Challenge:** Generalizing to new phenomena is hard:

- Prevailing methods require abundant label information
- However, labeled examples are scarce
- Examples: Novel drugs in development, emerging pathogens, rare diseases, hard-to-diagnose patients

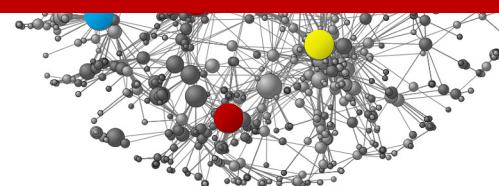
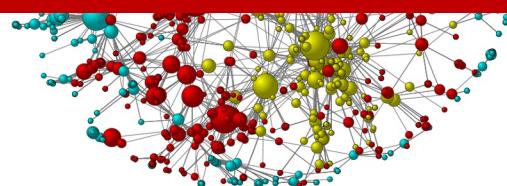
**What prevailing  
methods assume**



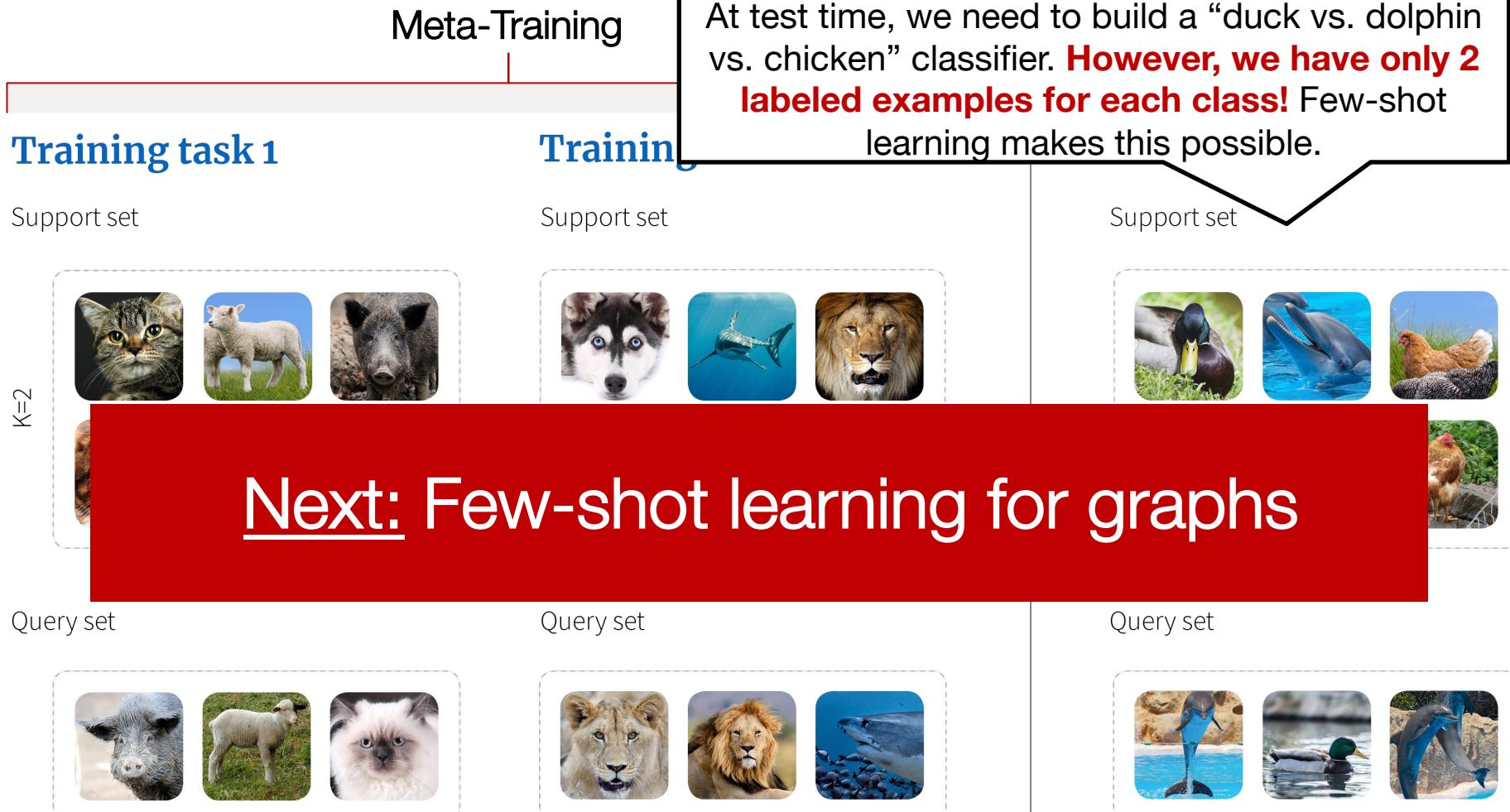
**What happens in  
real world**



Today: Few-shot learning for graphs



# Background: Few-shot learning



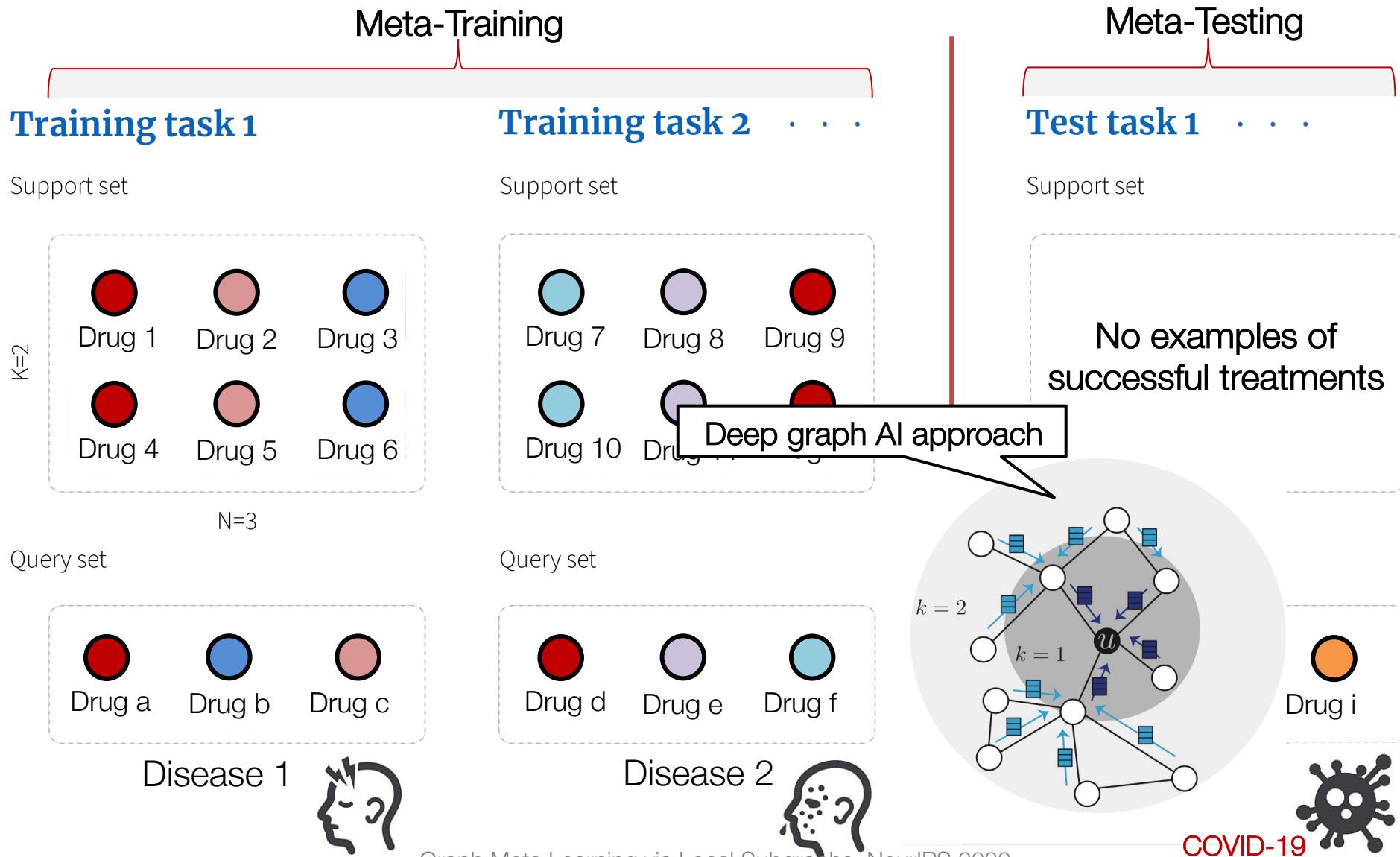
Next: Few-shot learning for graphs

An example of 2-shot 3-way image classification

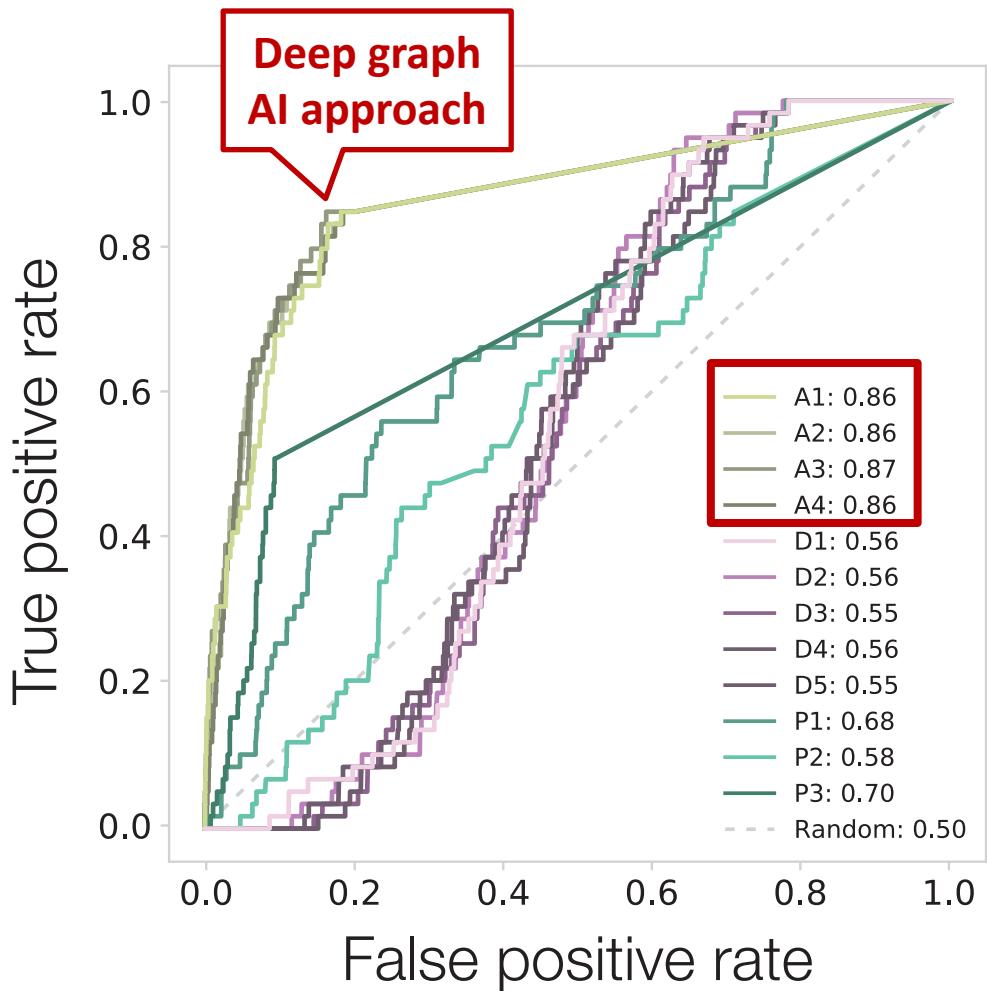
**Few-shot learning:** Instantiation of meta learning in the field of supervised learning

**K-shot N-class classification:** K labeled examples for each of N classes

# Few-shot learning for drugs



# Results: COVID-19 repurposing



We test each method's ability to recover drugs currently in clinical trials for COVID-19 (67 drugs from ClinicalTrials.gov)

---

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

# Results: Experimental screening



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
27	Dehorseranol

Predicted 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 & human cells:

- We screened in human cells the top-ranked drugs, obtaining a 62% success rate, in contrast to the 0.8% hit rate of nonguided screenings
- This is an order of magnitude higher hit rate among top 100 drugs than alternative approach

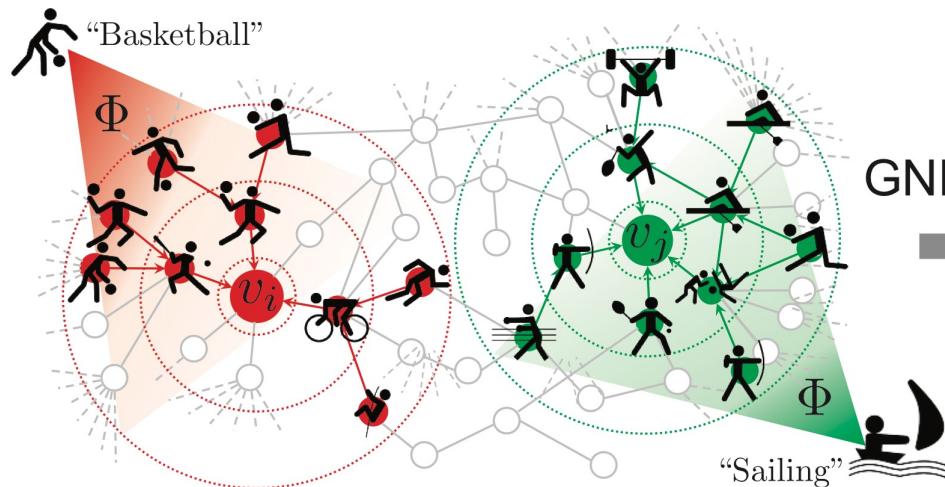
# Explaining machine predictions

Key idea:

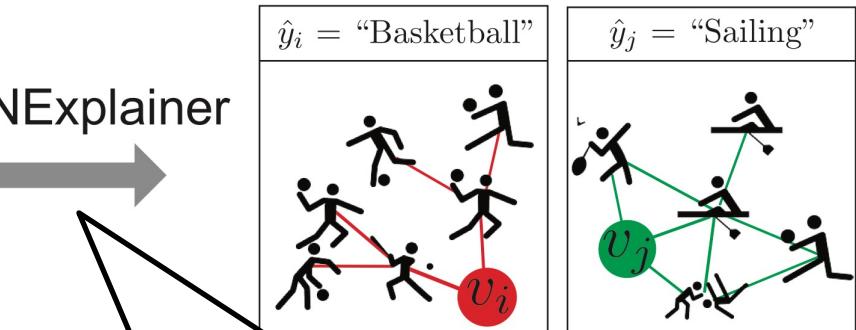
- Summarize where in the data the model “looks” for evidence for its prediction
- Find a small subgraph **most influential** for the prediction



GNN model training and predictions



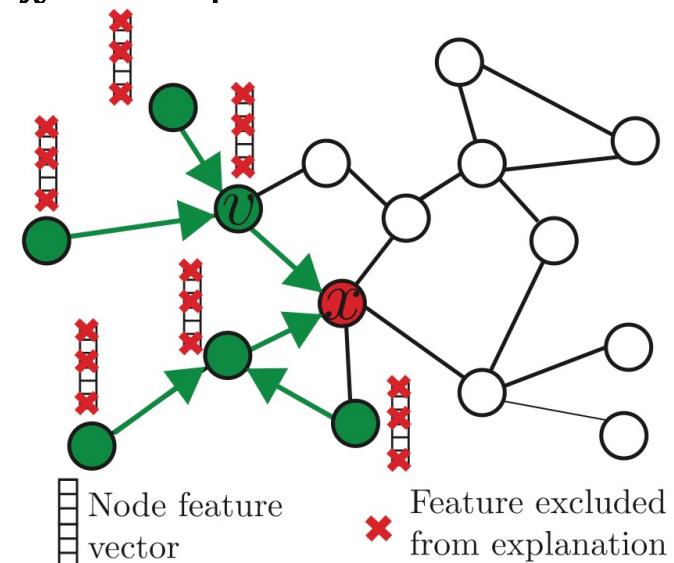
Explaining GNN’s predictions



Approach to generate explanations  
for graph neural networks based  
on **counterfactual reasoning**

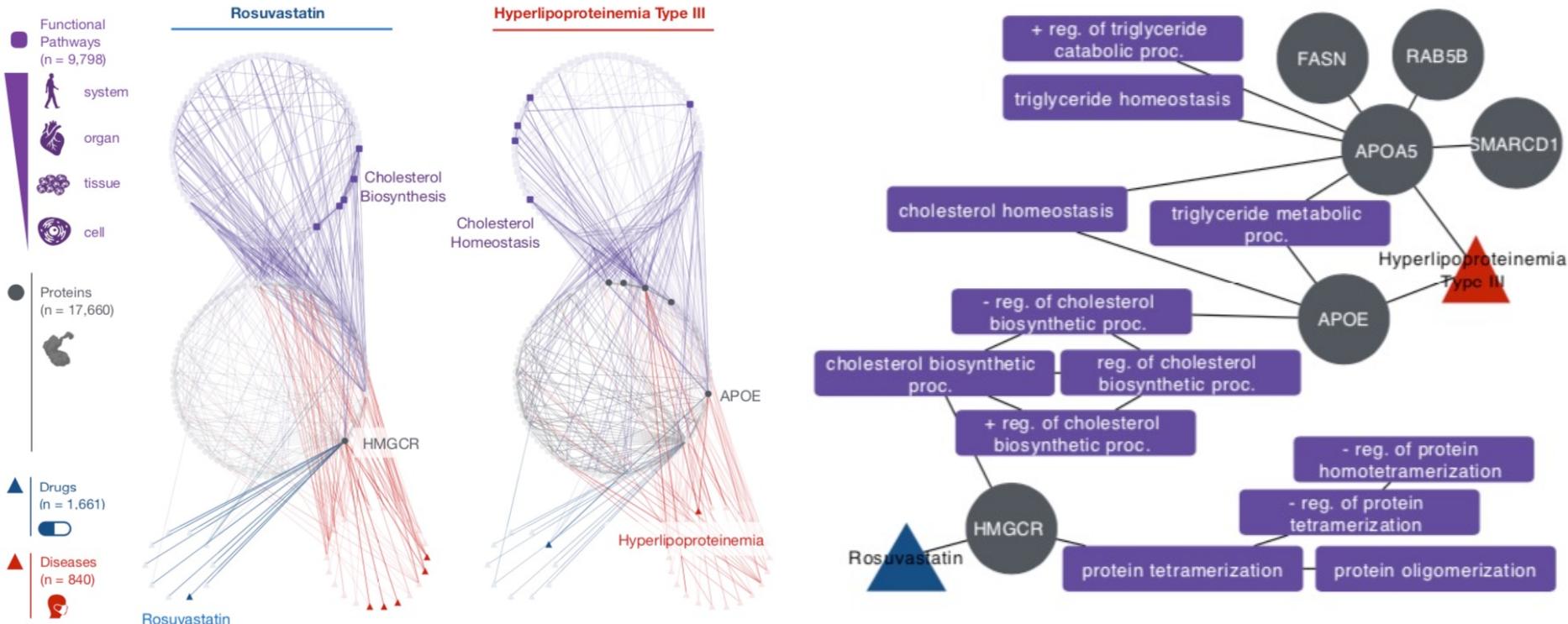
# GNNExplainer: key idea

- **Input:** Given prediction  $f(x)$  for node/link  $x$
- **Output:** Explanation, a small subgraph  $M_x$  together with a small subset of node features:
  - $M_x$  is most influential for prediction  $f(x)$
- **Approach:** Optimize mask  $M_x$  in a post-hoc manner
  - **Intuition:** If removing  $v$  from the graph strongly decreases the probability of prediction  $\Rightarrow v$  is a good counterfactual explanation for the prediction



# Example of explanations

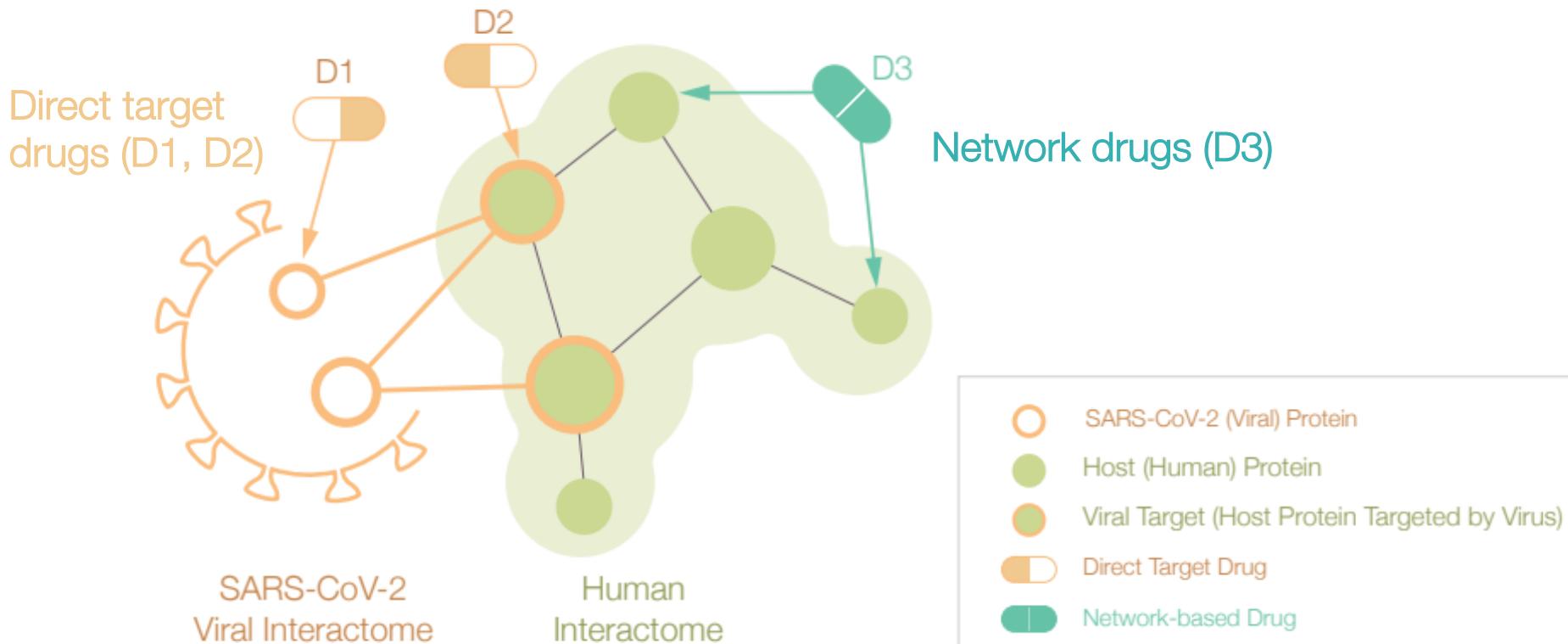
"Will rosuvastatin treat **hyperlipidemia**? What is the disease treatment mechanism?"



# Predictions → 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



# Key Takeaways

- Approach to identify **repurposable drugs for future pathogens** and **neglected diseases underserved by the costs** and extended timeline of de novo drug development
- Algorithms we deployed algorithms relying on artificial intelligence, network diffusion, and network proximity:
  - No single predictive algorithm offers consistently reliable outcomes across all datasets and metrics
  - **Multimodal approach fused predictions of all algorithms**, finding that a consensus among different predictive methods and consistently exceeding performance of the best individual algorithm
  - **Top-ranked drugs** screened in human cells yield a 62% success rate in contrast to the 0.8% hit rate of **nonguided screenings**
- Resources
  - Paper: <https://www.pnas.org/doi/full/10.1073/pnas.2025581118>
  - Webinar: [https://www.youtube.com/watch?v=jS8\\_WViNj4](https://www.youtube.com/watch?v=jS8_WViNj4)
  - GitHub:
    - COVID-19 repurposing: <https://github.com/Barabasi-Lab/COVID-19>
    - Multimodal fusion: <https://github.com/mims-harvard/crank>

# Graph RL for therapeutics

## Summary

- **TDC:** Open-science initiative with AI-ready datasets, AI tasks, and benchmarks for therapeutic science
- **Deep learning for antibiotic discovery:** Generative methods can examine several orders of magnitude larger chemical spaces than standard chemical libraries and generate compounds with desired drug-like properties
- **COVID-19 drug repurposing:** When designing new drugs from scratch is not feasible, repurposing offers an enticing alternative. Few-shot methods can identify promising therapeutic opportunities for diseases with few treatment options

## Poll Question

What is your dream AI/ML-ready dataset and AI/ML task for therapeutics? *Fill in the blank*

## Q&A Session

Applications of graph representation learning on...

# PRECISION MEDICINE

1. Histopathology images of tissue biopsies
2. Patient electronic health records

Applications of graph representation learning on...

# PRECISION MEDICINE

1. Histopathology images of tissue biopsies
2. Patient electronic health records



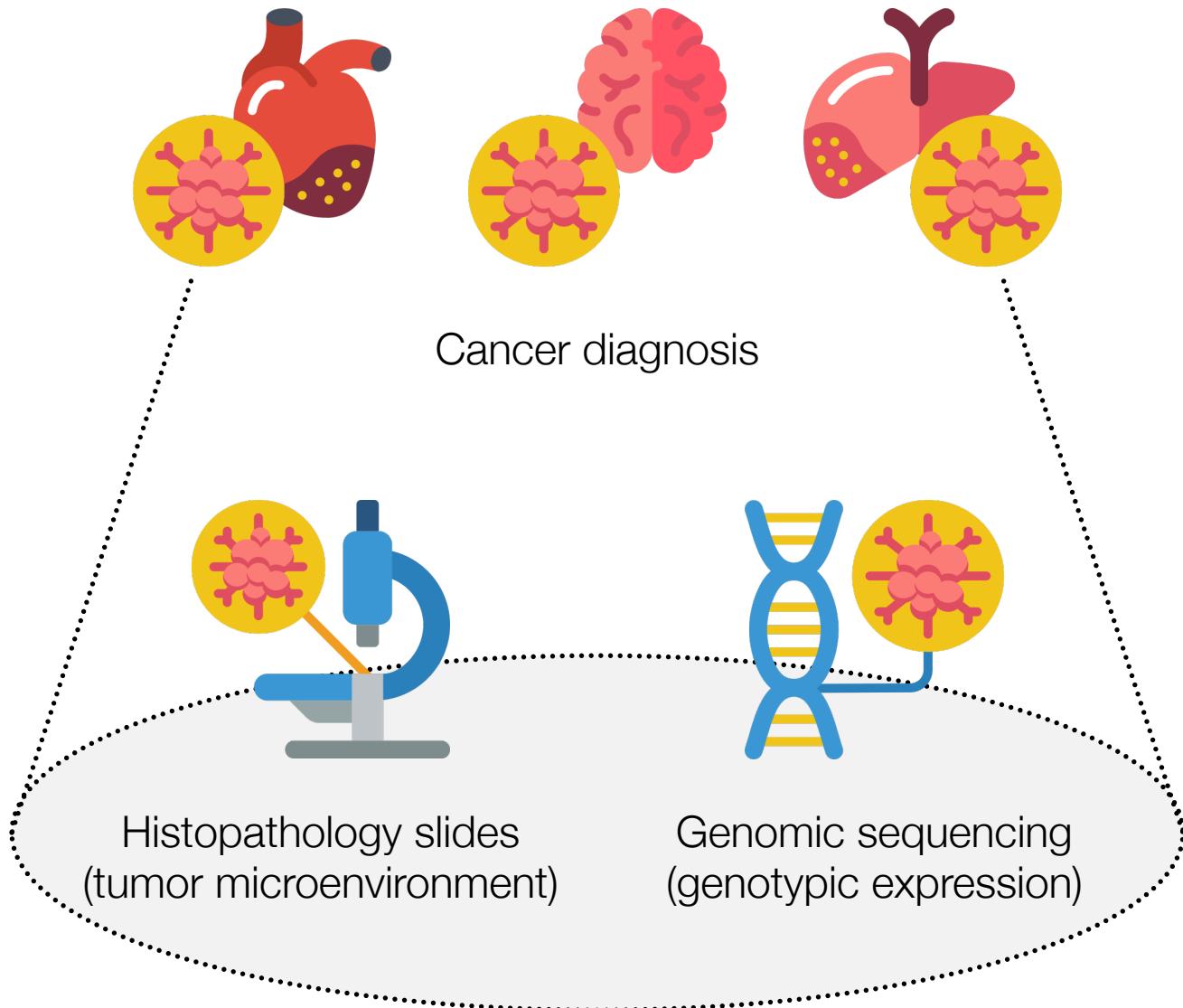
IEEE TRANSACTIONS ON MEDICAL IMAGING, VOL. 41, NO. 4, APRIL 2022

757

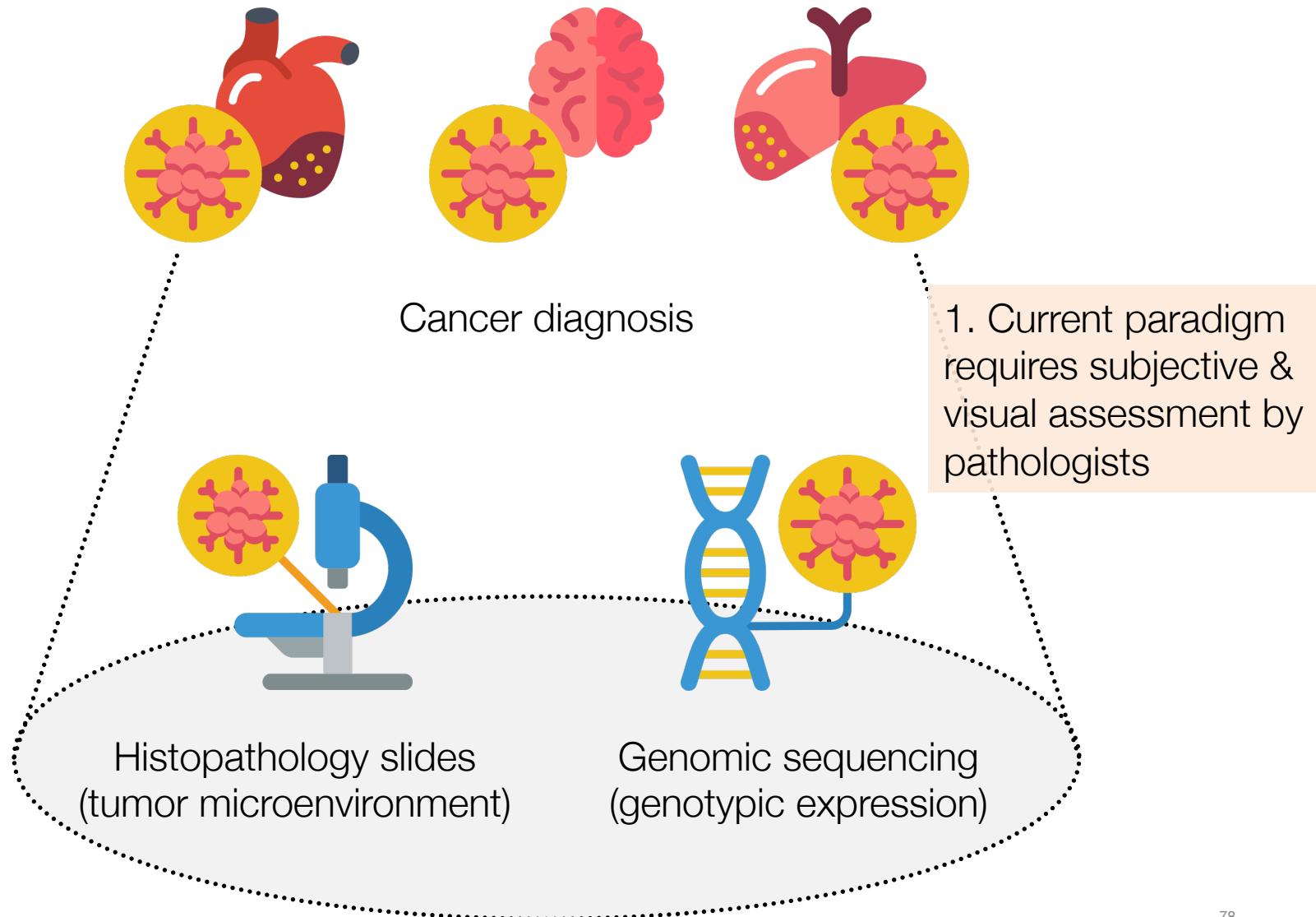
## Pathomic Fusion: An Integrated Framework for Fusing Histopathology and Genomic Features for Cancer Diagnosis and Prognosis

Richard J. Chen, Ming Y. Lu, Jingwen Wang, Drew F. K. Williamson, Scott J. Rodig, Neal I. Lindeman, and Faisal Mahmood<sup>ID</sup>, Member, IEEE

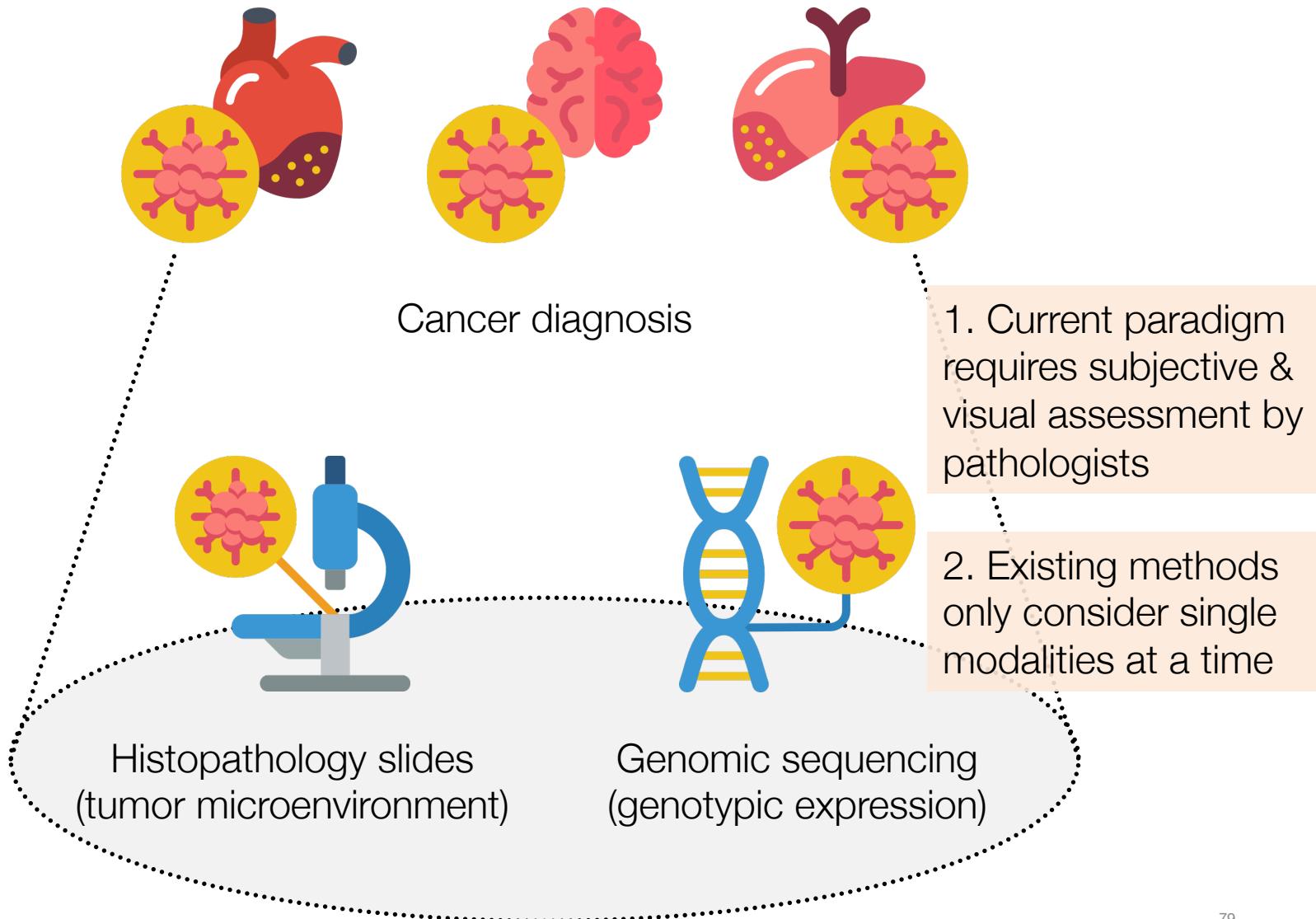
# Motivation



# Motivation



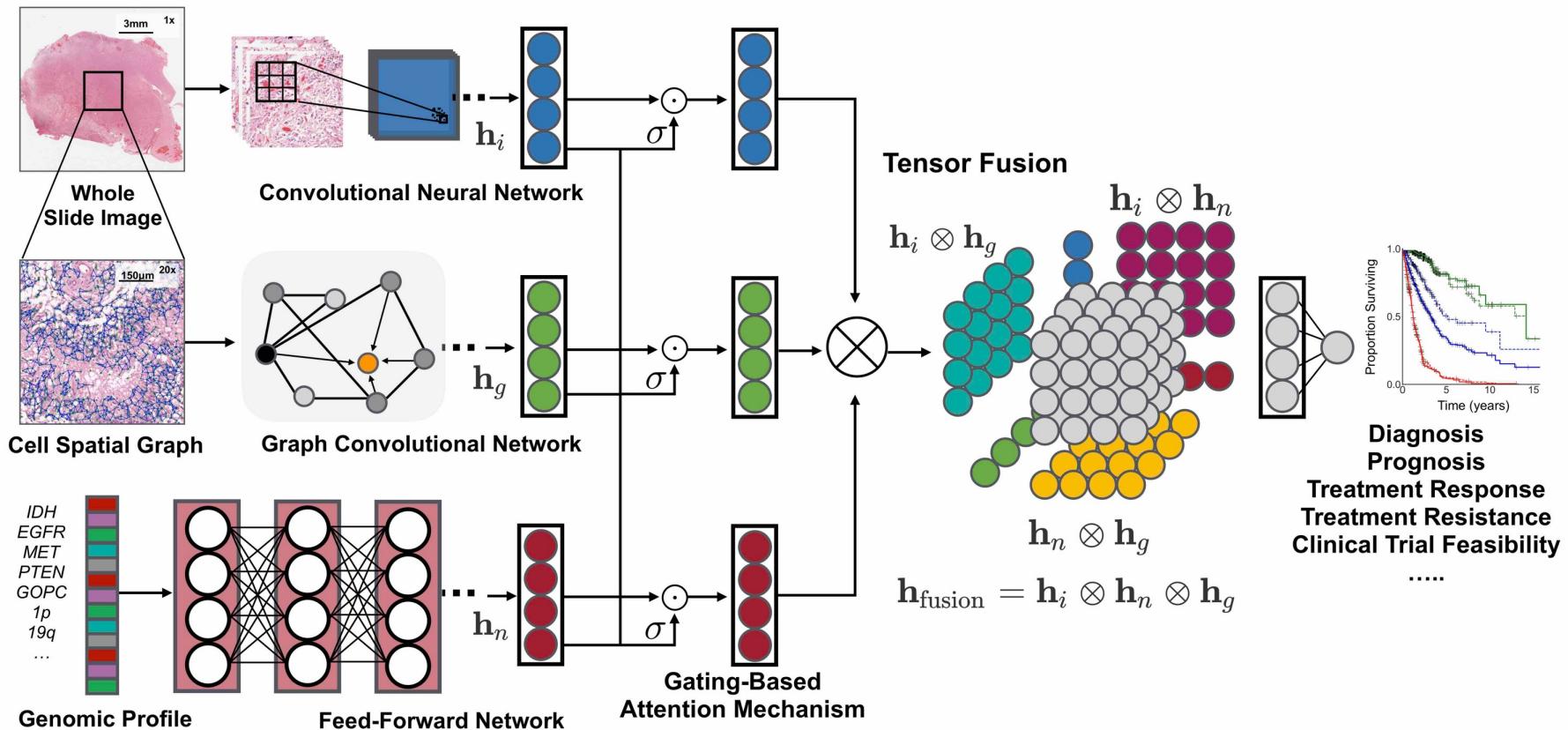
# Motivation



# Challenges

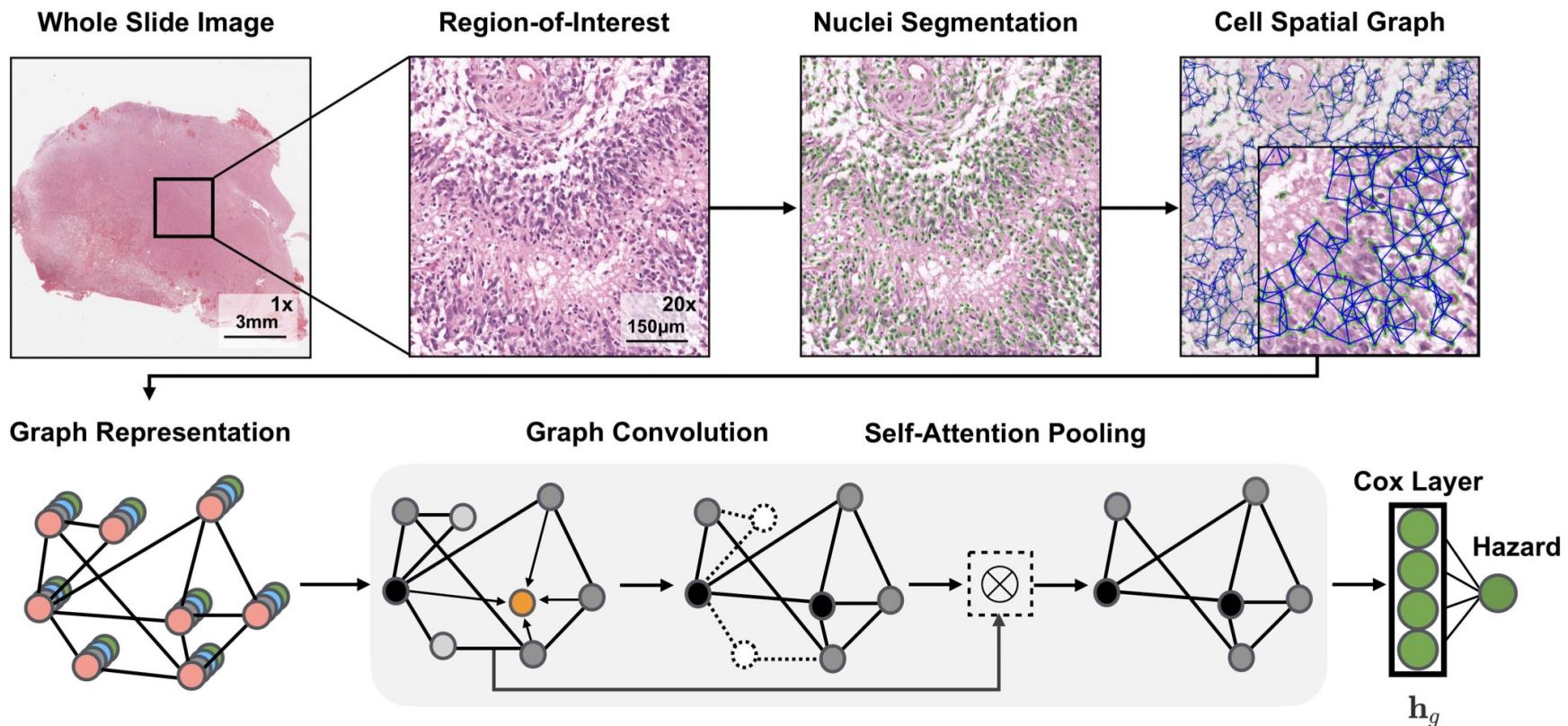
- **Goal:** Create an objective, interpretable, and scalable framework to diagnose patients
  - **Challenge:** Current standard of care relies on subjective and qualitative approaches
- **Goal:** Learn meaningful representations of tissue biopsies from histopathology images
  - **Challenge:** Existing methods typically only use CNNs, which do not capture the underlying structure in tissues and cells
- **Goal:** Integrate multi-modal data (e.g., tissue biopsies, genotypic information) to diagnose patients' cancers
  - **Challenge:** Existing methods only focus on one data modality at a time

# Overview of Pathomic Fusion

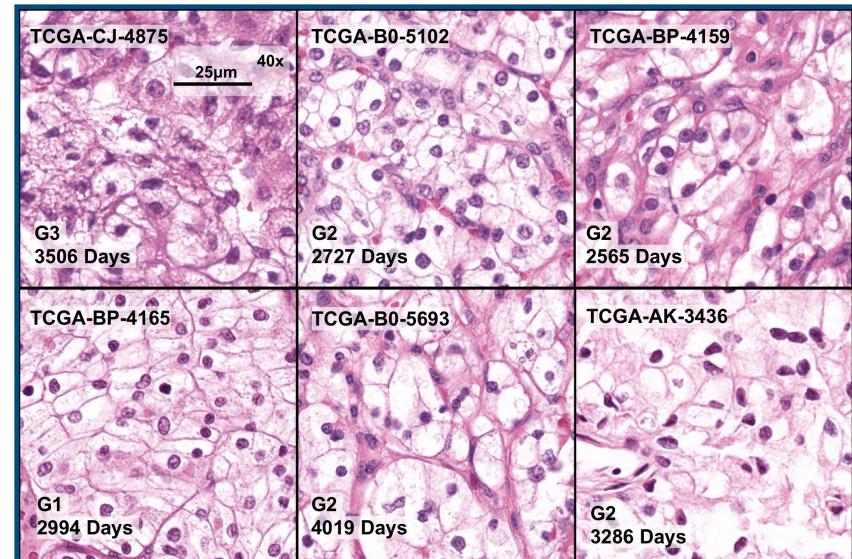
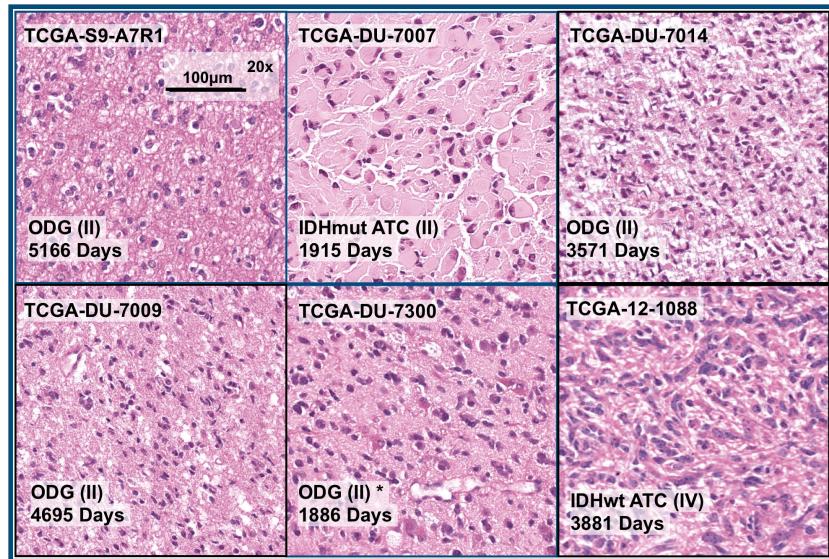


Pathomic Fusion: An Integrated Framework for Fusing Histopathology and Genomic Features for Cancer Diagnosis and Prognosis, *IEEE Transactions on Medical Imaging*, 2020.

# GCN for whole slide images



# Data



## Data

- 470 paired samples
- 20 x 1024 x 1024 Histology ROIs (1-3 per patient)
- 1 Mutation, 79 CNV, 240 RNA-Seq

## Experiments

- Compare to WHO Grade + Subtype
- 15-Fold CV

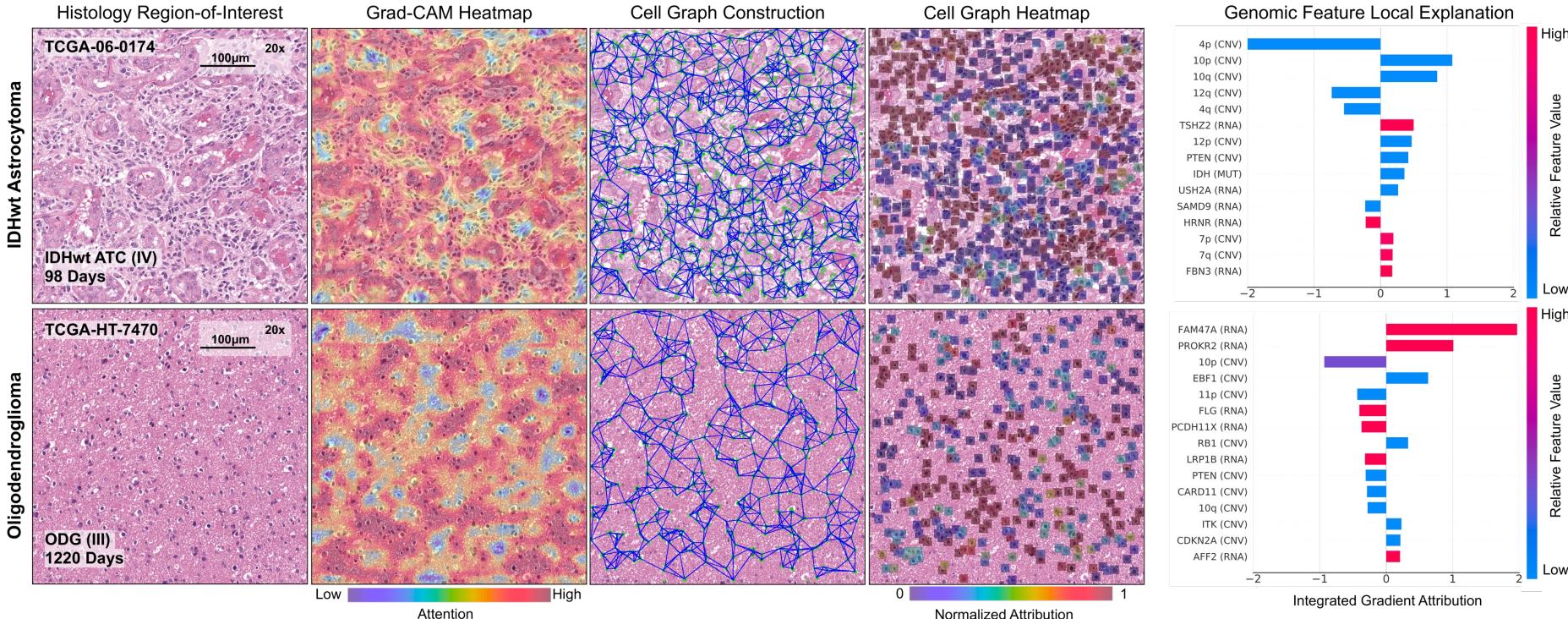
## Data

- 417 paired samples
- 40 x 512 x 512 Histology ROIs (3 per patient)
- 117 CNV, 240 RNA-Seq

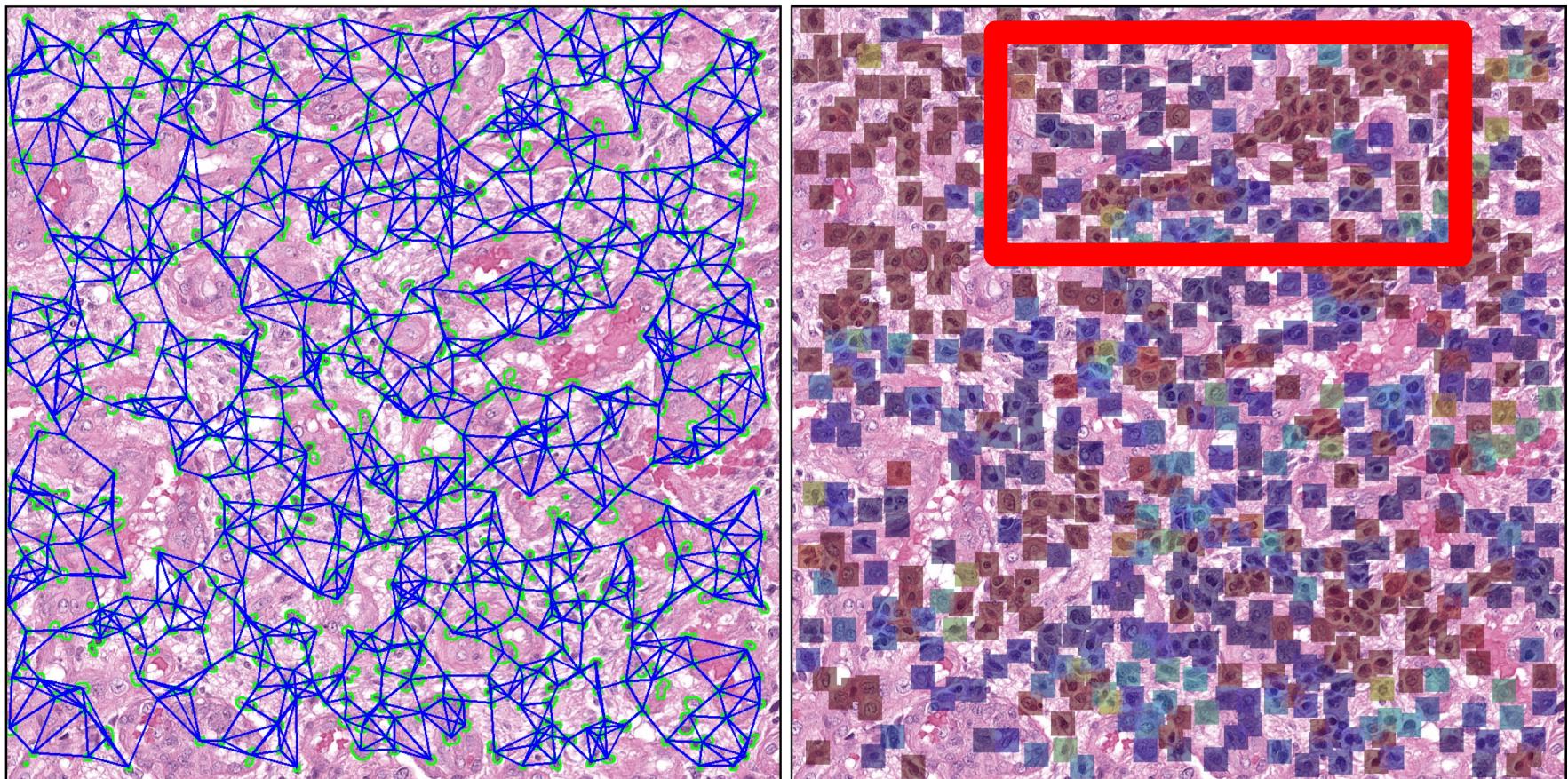
## Experiments

- Compare to Fuhrman Grade
- 15-Fold CV

# Results

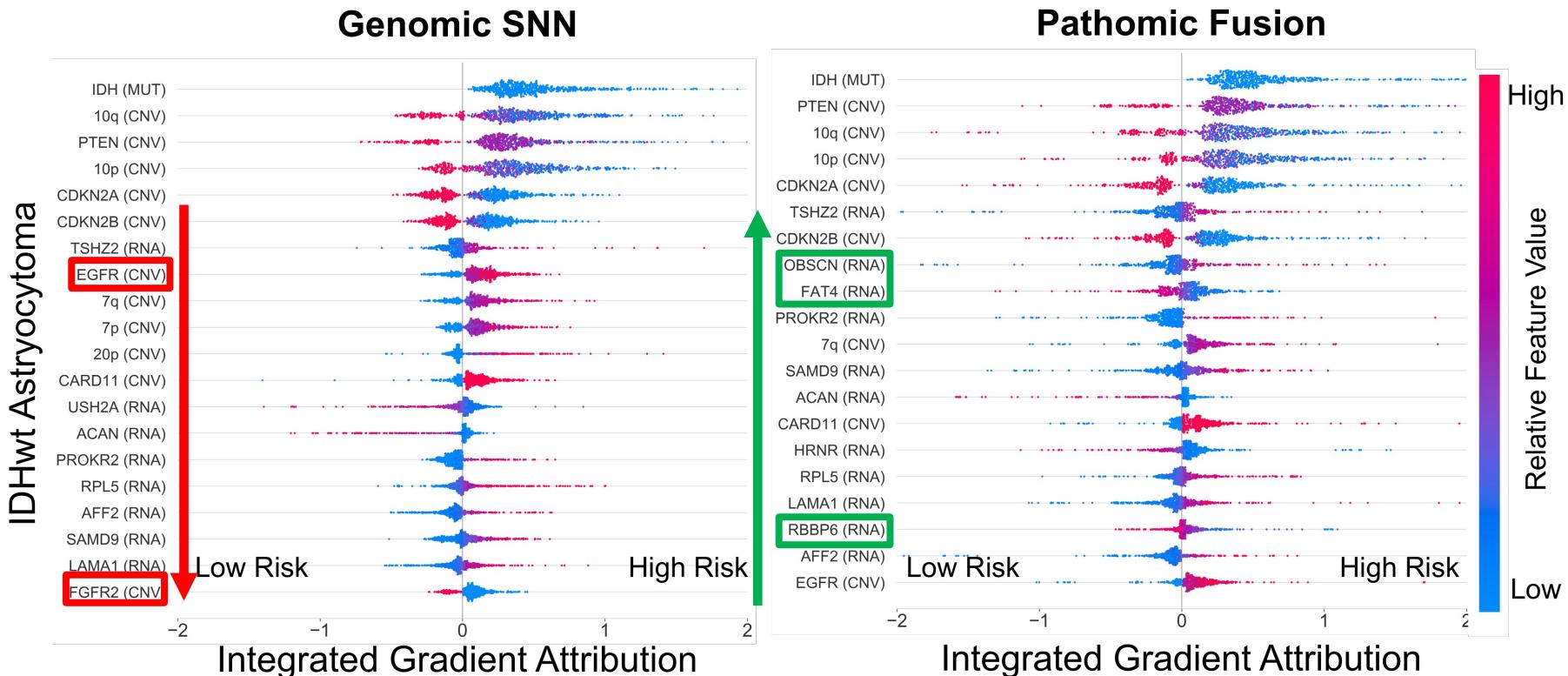


# Results



Pathomic Fusion: An Integrated Framework for Fusing Histopathology and Genomic Features for Cancer Diagnosis and Prognosis, *IEEE Transactions on Medical Imaging*, 2020.

# Results



# Key Takeaways

- Pathomic Fusion is
  - Objective and multimodal
  - Interpretable
  - Adaptable to any type or combination of modalities
  - Locally and globally interpretable
  - Reproducible and publicly available
- Resources
  - Paper: [ieeexplore.ieee.org/document/9186053](https://ieeexplore.ieee.org/document/9186053)
  - GitHub: [github.com/mahmoodlab/PathomicFusion](https://github.com/mahmoodlab/PathomicFusion)
  - Talk: [youtube.com/watch?v=TrjGEUVX5YE](https://youtube.com/watch?v=TrjGEUVX5YE)
  - Synthetic dataset: [doi.org/10.1038/s41551-021-00751-8](https://doi.org/10.1038/s41551-021-00751-8)

Applications of graph representation learning on...

# PRECISION MEDICINE

1. Histopathology images of tissue biopsies
2. Patient electronic health records

Applications of graph representation learning on...

# PRECISION MEDICINE

1. Histopathology images of tissue biopsies
2. Patient electronic health records

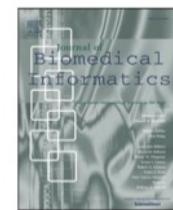
*Journal of Biomedical Informatics* 127 (2022) 104000



Contents lists available at [ScienceDirect](#)

**Journal of Biomedical Informatics**

journal homepage: [www.elsevier.com/locate/yjbin](http://www.elsevier.com/locate/yjbin)



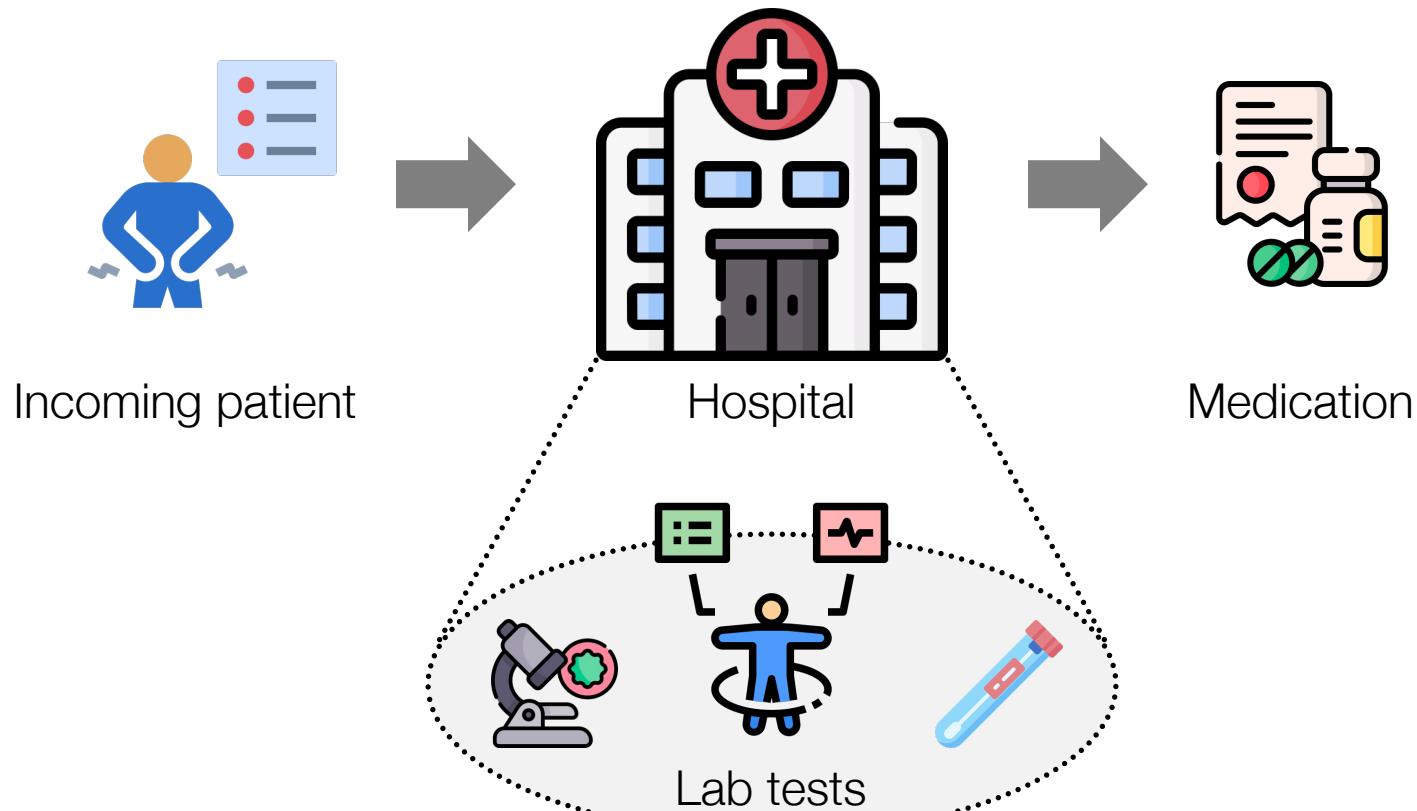
**MedGCN: Medication recommendation and lab test imputation via graph convolutional networks**

Chengsheng Mao, Liang Yao, Yuan Luo \*

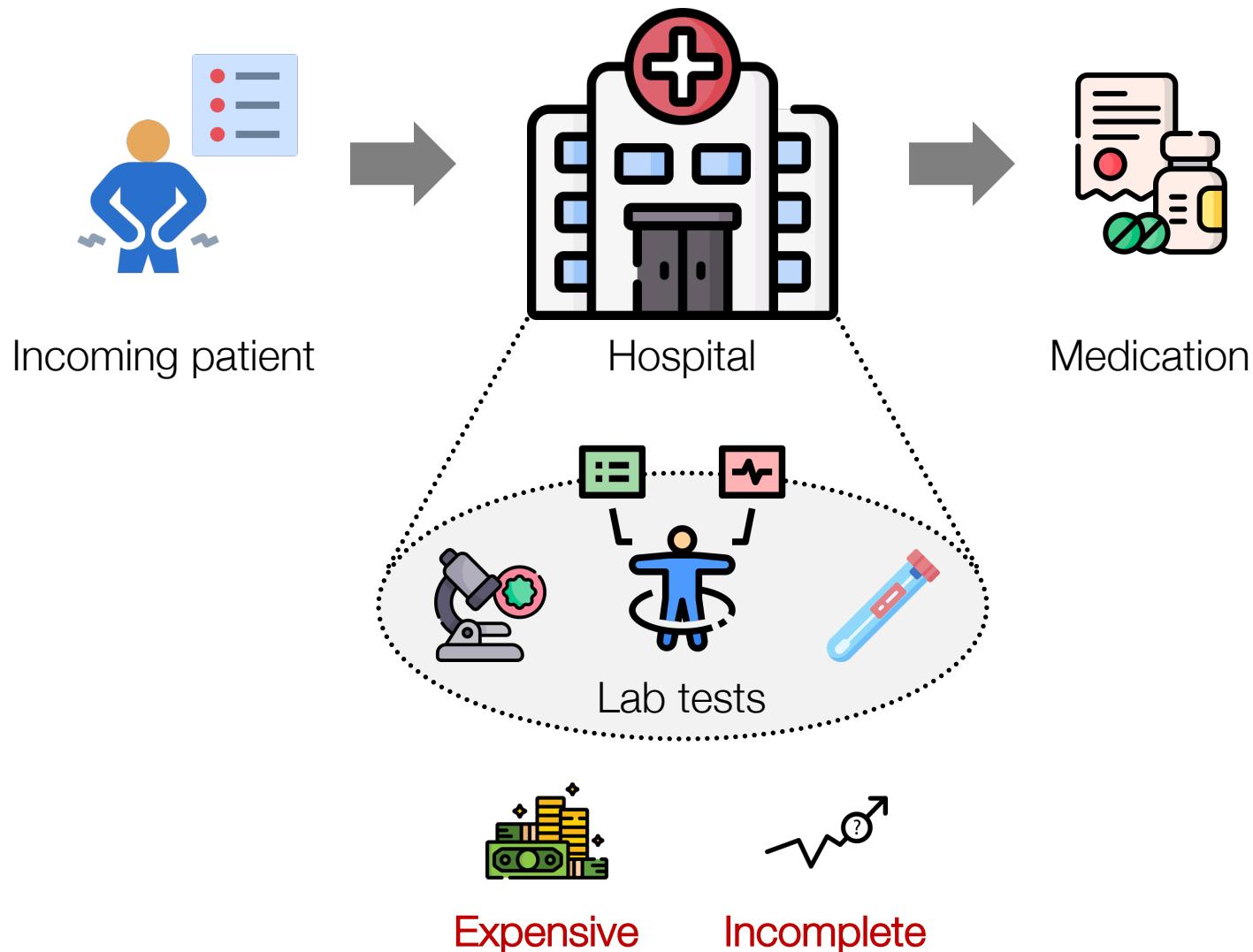
*Department of Preventive Medicine, Feinberg School of Medicine, Northwestern University, Chicago, IL, USA*



# Motivation



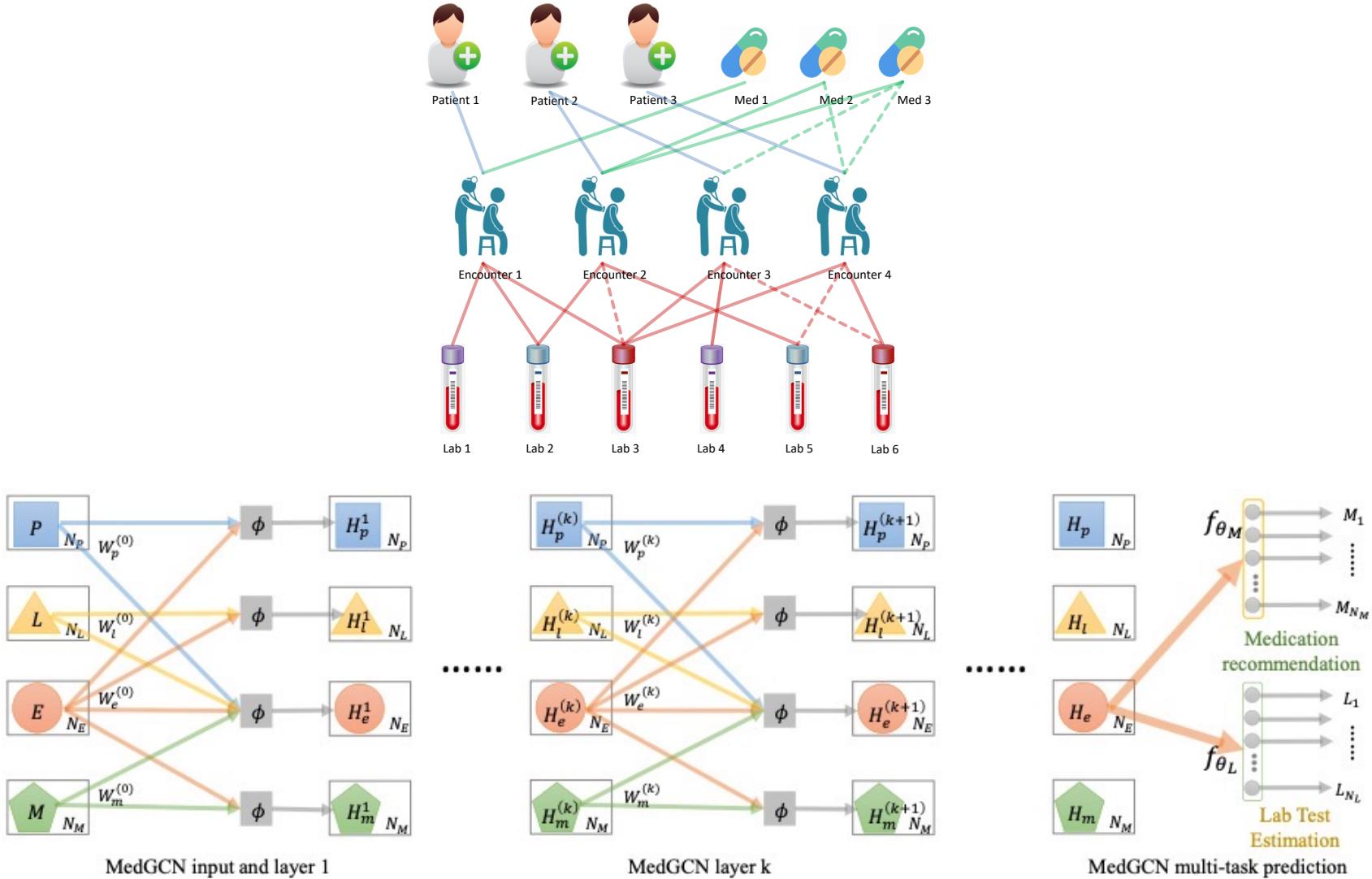
# Motivation



# Challenges

- **Goal:** Capture the complex relationships between patients, encounters, labs, and medications
  - **Challenge:** Existing methods typically represent a single entity (e.g., cannot model heterogeneous networks)
- **Goal:** Impute missing lab tests' values
  - **Challenge:** Prior work assume certain characteristics of the missing values (e.g., measured at the same time point, has temporal correlation)
- **Goal:** Recommending medications without prior knowledge of patient diagnoses
  - **Challenge:** Existing methods typically use diagnosis codes as input, and are thus reliant on physicians' domain expertise.

# Overview of MedGCN



# MedGCN Message Propagation

	P1	P2	P3		M1	M2	M3		L1	L2	L3	L4	L5	L6		L1	L2	L3	L4	L5	L6
E1	1	0	0	E1	1	0	0	E1	.3	0	.1	0	0	0	E1	1	1	1	0	0	0
E2	0	1	0	E2	0	1	1	E2	0	.1	0	0	.5	0	E2	0	1	0	0	1	0
E3	0	1	0	E3	0	0	0	E3	0	0	.1	.8	0	0	E3	0	0	1	1	0	0
E4	0	0	1	E4	0	0	0	E4	0	0	.0	0	0	.6	E4	0	0	1	0	0	1
$A_{E \times P}$			$A_{E \times M}$			$A_{E \times L}$			$M_{E \times L}$												

$$H_e^{(k+1)} = \phi\left(A_{E \times P} \cdot H_p^{(k)} \cdot W_p^{(k)} + A_{E \times L} \cdot H_l^{(k)} \cdot W_l^{(k)} + A_{E \times M} \cdot H_m^{(k)} \cdot W_m^{(k)} + H_e^{(k)} \cdot W_e^{(k)}\right)$$

$$H_p^{(k+1)} = \phi\left(A_{P \times E} \cdot H_e^{(k)} \cdot W_e^{(k)} + H_p^{(k)} \cdot W_p^{(k)}\right)$$

$$H_l^{(k+1)} = \phi\left(A_{L \times E} \cdot H_e^{(k)} \cdot W_e^{(k)} + H_l^{(k)} \cdot W_l^{(k)}\right)$$

$$H_m^{(k+1)} = \phi\left(A_{m \times E} \cdot H_e^{(k)} \cdot W_e^{(k)} + H_m^{(k)} \cdot W_m^{(k)}\right)$$

# Datasets

## NMEDW

#E: 1260; #P: 865; #L: 197, #M: 57

Matrix	Size	Edges	Sparsity	Values
$A_{E \times P}$	$1260 \times 865$	1260	99.88%	binary: 0, 1
$A_{E \times L}$	$1260 \times 197$	43806	82.35%	continuous: 0–1
$A_{E \times M}$	$1260 \times 57$	2475	96.55%	binary: 0, 1

## MIMIC-III

#E: 18190; #P: 15153; #L: 219, #M: 117

Matrix	Size	Edges	Sparsity	Values
$A_{E \times P}$	$18190 \times 15153$	18190	99.99%	binary: 0, 1
$A_{E \times L}$	$18190 \times 219$	1029964	68.96%	continuous: 0–1
$A_{E \times M}$	$18190 \times 117$	23395	98.68%	binary: 0, 1

# Results

NMEDW

MIMIC-III

## Medication Recommendation

Methods	LRAP	MAP@2
MedGCN (ours)	<b>.7588±.0028</b>	<b>.7558±.0035</b>
MedGCN-ind (ours)	.7491±.0067*	.7558±.0073
MedGCN-Med (ours)	.7477±.0032*	.7457±.0046*
MLP	.7331±.0126*	.6965±.0113*
GBDT	.7120±.0018*	.6864±.0023*
RF	.6872±.0072*	.7055±.0068*
LR	.5325*	.4133*
SVM	.4324*	.3353*
CC	.6276±.0116*	.6182±.0159*

Methods	LRAP	MAP@2
MedGCN (ours)	<b>.8349±.0008</b>	.8069±.0022
MedGCN-ind (ours)	.8345±.0007	<b>.8070±.0029</b>
MedGCN-Med (ours)	.8346±.0005	.8061±.0020
MLP	.8325±.0003*	.8030±.0030*
GBDT	.5793±.0001*	.5019±.0002*
RF	.8215±.0007*	.8030±.0011*
LR	.3367*	.1839*
SVM	.6642*	.6146*
CC	.7660±.0005*	.7153±.0003*

## Lab Test Imputation

Methods	MSE
MedGCN (ours)	<b>.0229±.0025</b>
MedGCN-ind (ours)	.0264±.0034*
MedGCN-Lab (ours)	.0254±.0003*
MICE	.0474±.0010*
MGCNN	.0369±.0009*
GCMC	.0426±.0025*
GCMC+FEAT	.0359±.0030*

Methods	MSE
MedGCN(ours)	<b>.0140±.0002</b>
MedGCN-ind(ours)	.0143±.0002*
MedGCN-Lab(ours)	.0143±.0001*
MICE	.0146±.0001*
MGCNN	.0413±.0048*
GCMC	.0296±.0004*
GCMC+FEAT	.0290±.0001*

# Key Takeaways

- MedGCN
  - Incorporates complex associations between multiple medical entities (e.g., patients, labs, encounters, medications)
  - Extends general GCN model to heterogeneous graphs and missing feature values for medical settings
  - Learn multiple tasks via cross regularization
  - Is inductive to efficiently generate representations for new data
- Resources
  - Paper: [doi.org/10.1016/j.jbi.2022.104000](https://doi.org/10.1016/j.jbi.2022.104000)
  - GitHub: [github.com/mocherson/MedGCN](https://github.com/mocherson/MedGCN)

# Why are precision medicine applications so challenging?

- Methods presented so far optimize for accuracy
- Accuracy alone is no longer enough
- Life or death decisions
  - Need **robust** algorithms
  - Ensure that models behave **responsibly**
  - Ensure that models are **trustworthy**
  - **Checks and balances built** into ML deployment
- Other criteria are important too:
  - Explainable predictions and interpretable models
  - Privacy-preserving, causal, and robust predictions

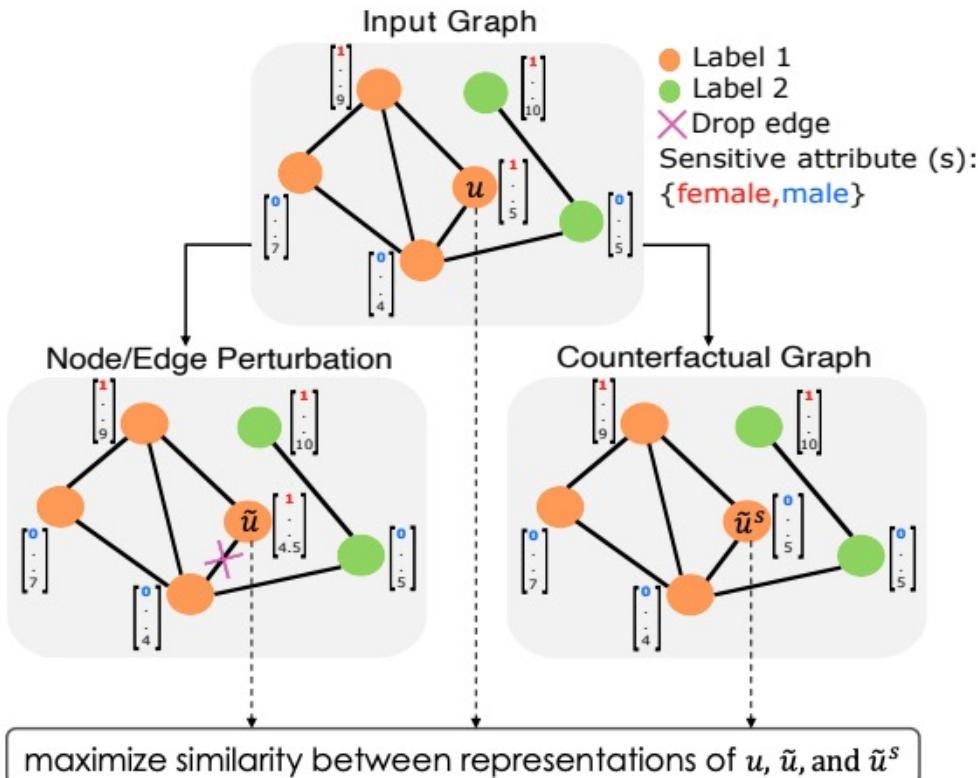


High-stakes decisions

# Towards fair & stable GNNs (1/3)

- As the representations output by GNNs are considered for real-world implementation, it is important that **representations are fair and stable**
- NIFTY (uNIfying Fairness and stabiliTY) is a novel framework:
  - It can be **used with any GNN** to learn fair and stable representations
  - It develops:
    - an **objective function that simultaneously** accounts for fairness and stability
    - a **layer-wise weight normalization using the Lipschitz constant** to enhance neural message passing in GNNs
  - **Theoretical proved** that NIFTY promotes counterfactual fairness and stability in the resulting representations

# Towards fair & stable GNNs (2/3)



- NIFTY learn **node representations that are both fair and stable**
  - Invariant to sensitive attribute value
  - Invariant to perturbations of the graph structure and non-sensitive attributes
- NIFTY's objective function jointly optimizes for fairness and stability:
  - Maximize similarity between:
    - Representations of original nodes
    - Representation of nodes in augmented graph
  - Augmented graph is generated by:
    - Slightly perturbing original node attributes and edges
    - Considering counterfactuals of the original nodes where the value of the sensitive attribute is modified

# Towards fair & stable GNNs (3/3)

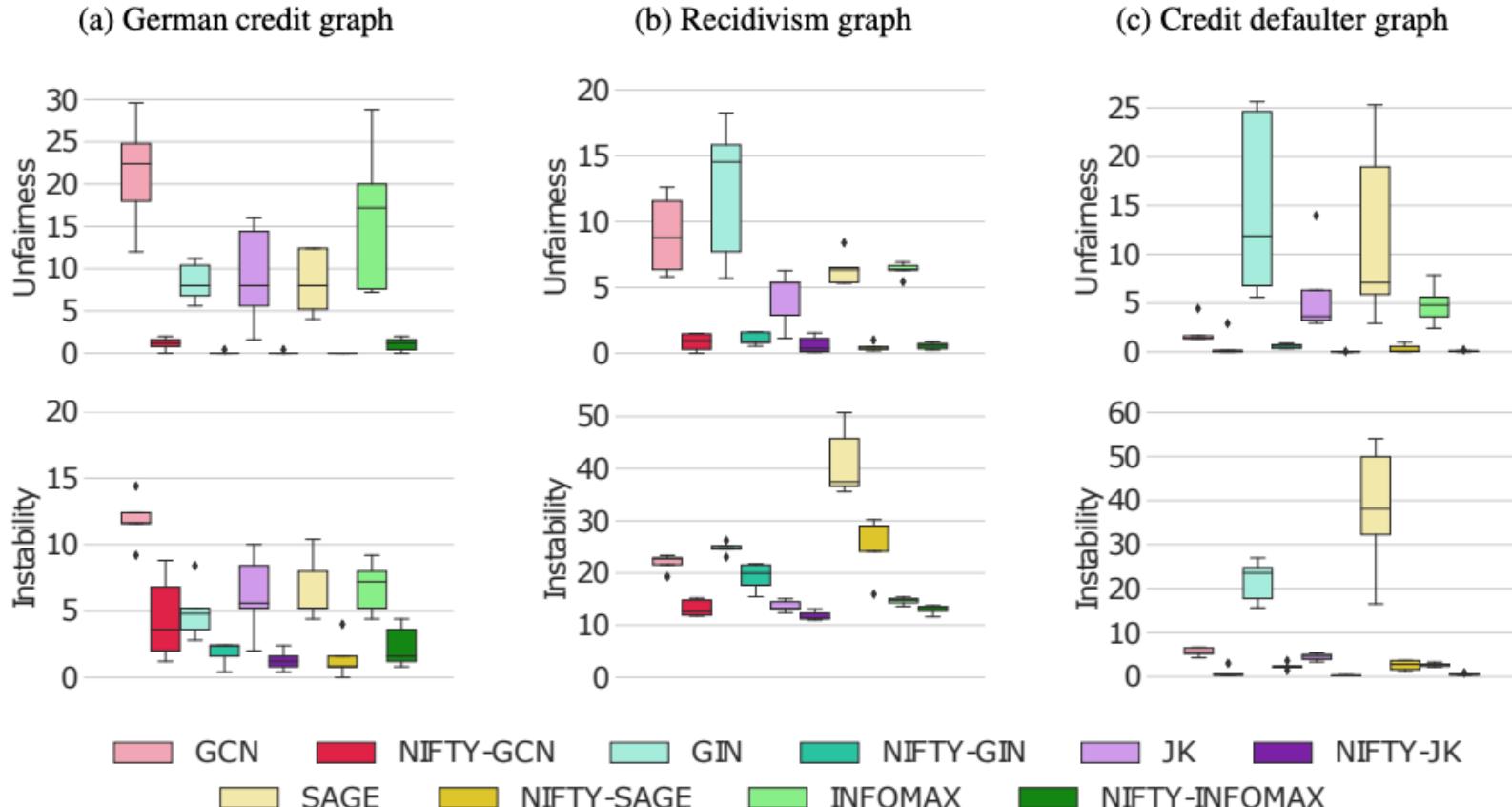


Figure 2: Unfairness (top) and instability (bottom) error rates for five GNNs and their NIFTY counterparts. NIFTY-enhanced GNNs give fairer and more stable predictions than their unmodified counterparts across all three datasets and five GNNs.

# Graph RL for precision medicine

## Summary

- **Pathomic Fusion:** Applies a graph convolutional network to represent & integrate histopathology slides with genomic features for patient cancer diagnosis
- **MedGCN:** Simultaneously represents the complexity of relationships between patients, encounters, labs, and medications while imputing missing lab tests' values to recommend medications for patients

## Poll Question

What other applications in precision medicine require (or *should* require) ethical considerations? *Fill in the blank*

## Q&A Session

# This Tutorial

- ✓ 1. Methods: Network diffusion, shallow network embeddings, graph neural networks, equivariant neural networks
- ✓ 2. Applications: Fundamental biological discoveries and precision medicine
- 3. Hands-on exercises: Demos, implementation details, tools, and tips

# Resources

- Books & survey papers
  - William Hamilton, *Graph Representation Learning* ([morganclaypool.com/doi/abs/10.2200/S01045ED1V01Y202009AIM046](http://morganclaypool.com/doi/abs/10.2200/S01045ED1V01Y202009AIM046))
  - Li et al., Graph Representation Learning for Biomedicine ([arxiv.org/abs/2104.04883](https://arxiv.org/abs/2104.04883))
- Keynotes & seminars
  - Michael Bronstein, “Geometric Deep Learning: The Erlangen Programme of ML” (ICLR 2021 keynote) ([youtube.com/watch?v=w6Pw4MOzMu0](https://youtube.com/watch?v=w6Pw4MOzMu0))
  - Broad Institute Models, Inference & Algorithms: Actionable machine learning for drug discovery; Primer on graph representation learning ([youtube.com/watch?v=9YpTYdru0Rg](https://youtube.com/watch?v=9YpTYdru0Rg))
  - Stanford University (CS224W Lecture): Graph neural networks in computational biology ([youtube.com/watch?v=\\_hy9AgZXhbQ](https://youtube.com/watch?v=_hy9AgZXhbQ))
  - AI Cures Drug Discovery Conference ([youtube.com/watch?v=wNXSkISMTw8](https://youtube.com/watch?v=wNXSkISMTw8))
- Conferences & summer schools
  - London Geometry and Machine Learning Summer School ([logml.ai](https://logml.ai))
  - Learning on Graphs Conference ([logconference.github.io](https://logconference.github.io))

# Resources

- Software & packages
  - PyTorch Geometric
  - NetworkX
  - Stanford Network Analysis Platform (SNAP)
- Tutorials & code bases
  - Pytorch Geometric Colab Notebooks ([pytorch-geometric.readthedocs.io/en/latest/notes/colabs.html](https://pytorch-geometric.readthedocs.io/en/latest/notes/colabs.html))
  - Zitnik Lab Graph ML Tutorials ([github.com/mims-harvard/graphml-tutorials](https://github.com/mims-harvard/graphml-tutorials))
  - Stanford University's CS224 ([web.stanford.edu/class/cs224w](https://web.stanford.edu/class/cs224w))
- Datasets
  - Precision Medicine Oriented Knowledge Graph (PrimeKG) ([zitniklab.hms.harvard.edu/projects/PrimeKG](https://zitniklab.hms.harvard.edu/projects/PrimeKG))
  - Therapeutic Data Commons (TDC) ([tdcommons.ai](https://tdcommons.ai))
  - BioSNAP ([snap.stanford.edu/biodata/](https://snap.stanford.edu/biodata/))
  - Open Graph Benchmark (OGB) ([ogb.stanford.edu](https://ogb.stanford.edu))