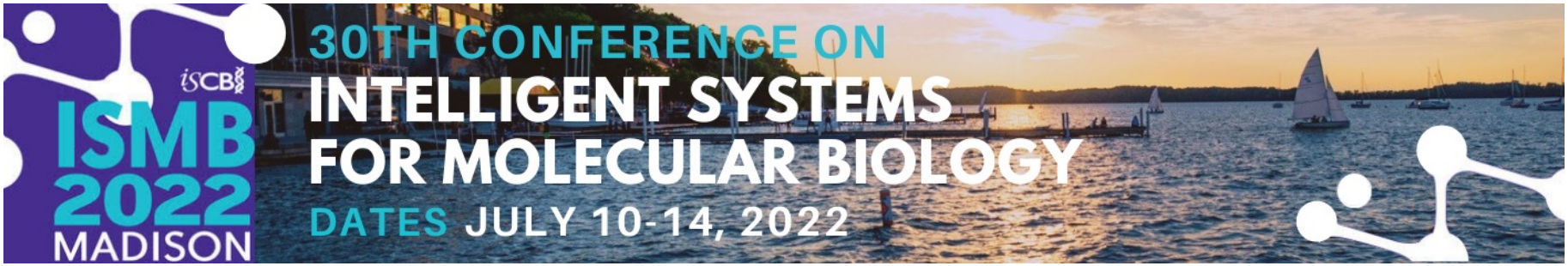
Towards Precision Medicine with Graph Representation Learning

Michelle M. Li & Marinka Zitnik

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

zitniklab.hms.harvard.edu/biomedgraphml



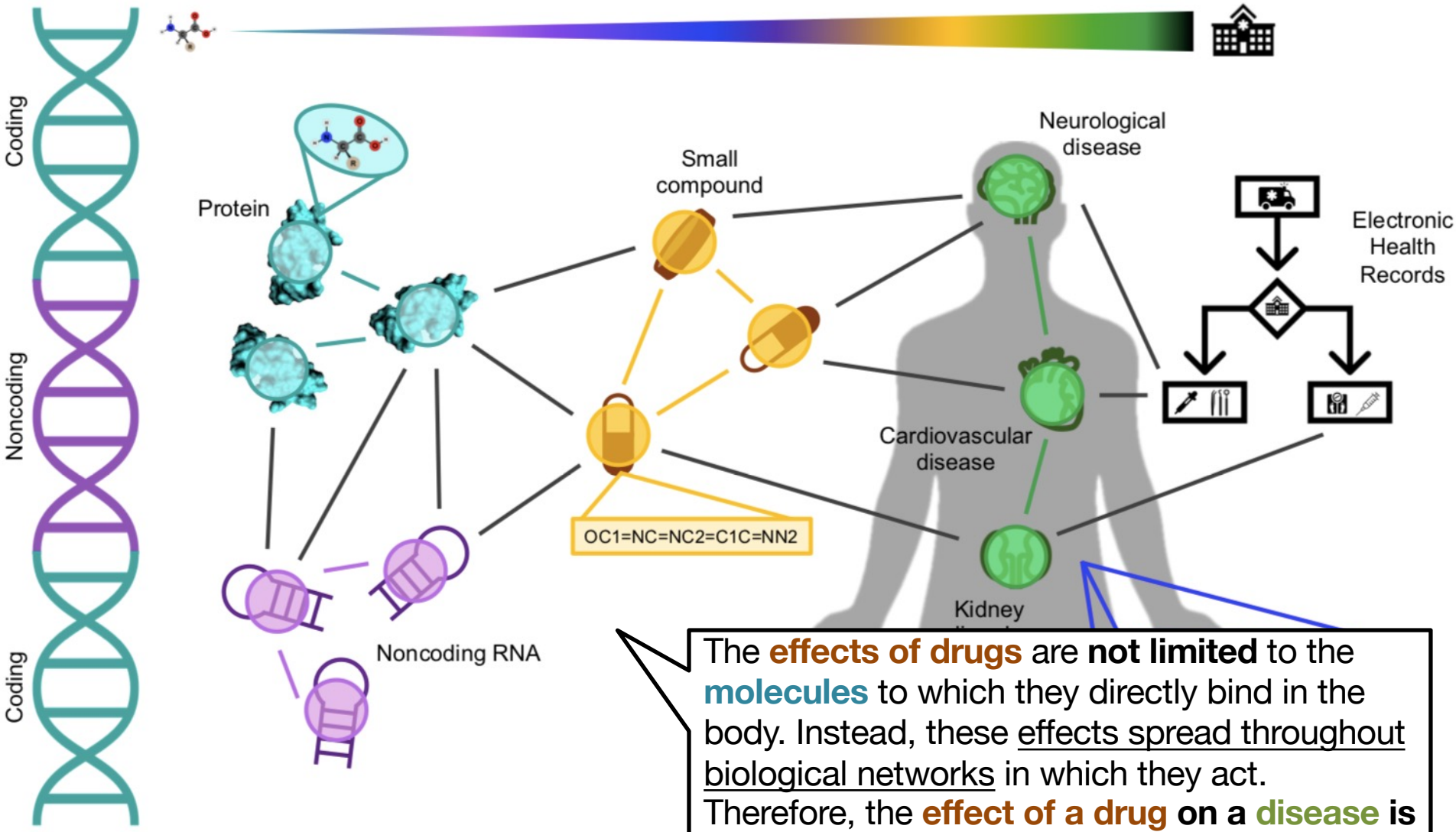
Tutorial VT4

July 7, 2022 at 9am – 1pm CDT

All tutorial materials are available at

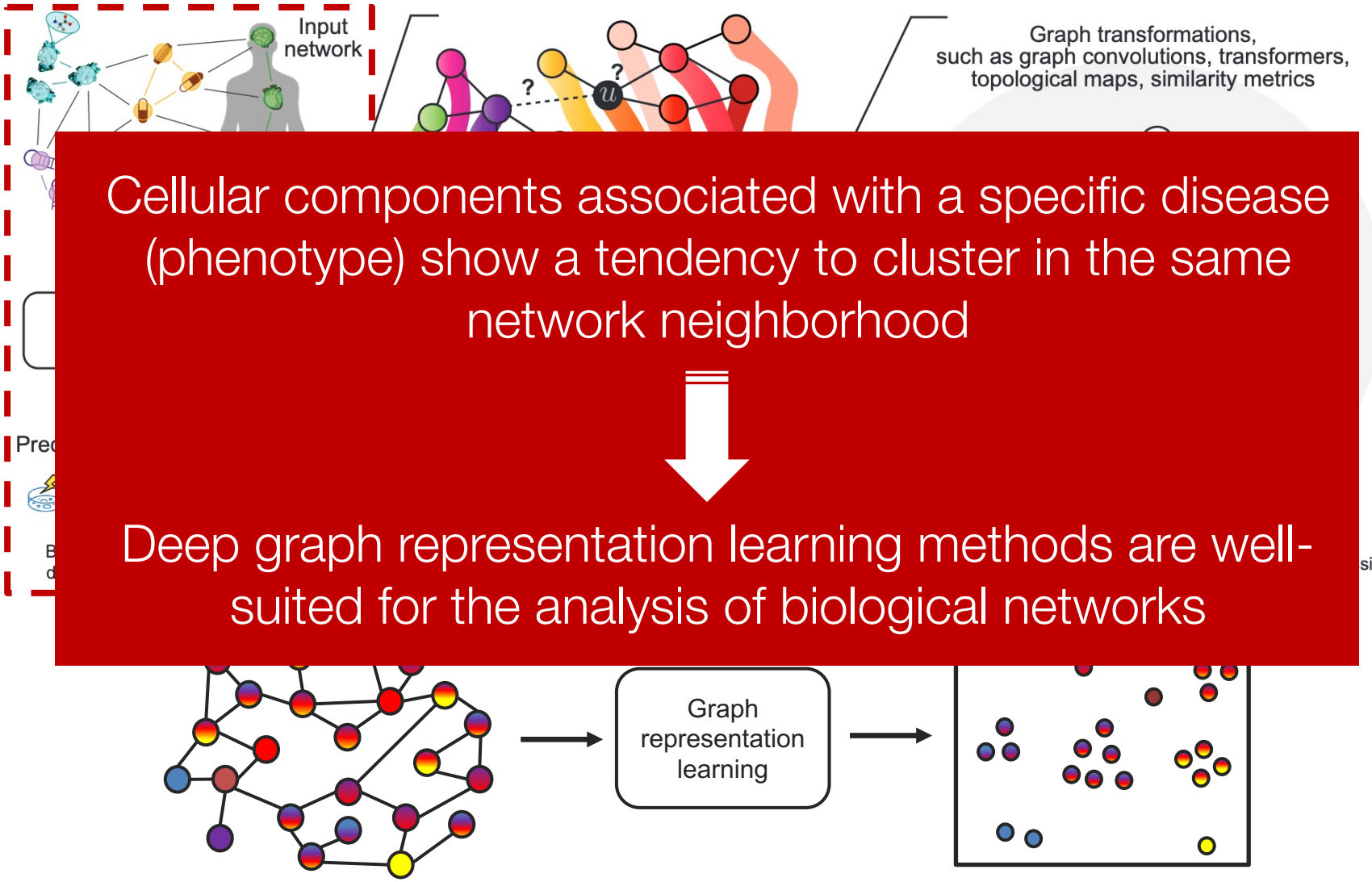
zitniklab.hms.harvard.edu/biomedgraphml

Biology is interconnected



The **effects of drugs** are **not limited** to the **molecules** to which they directly bind in the body. Instead, these effects spread throughout biological networks in which they act. Therefore, the **effect of a drug** **on a disease** **is inherently a network phenomenon**

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



Graph Representation Learning for Biomedicine, *Nature Biomedical Engineering* (in press), 2022, arXiv:2104.04883

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

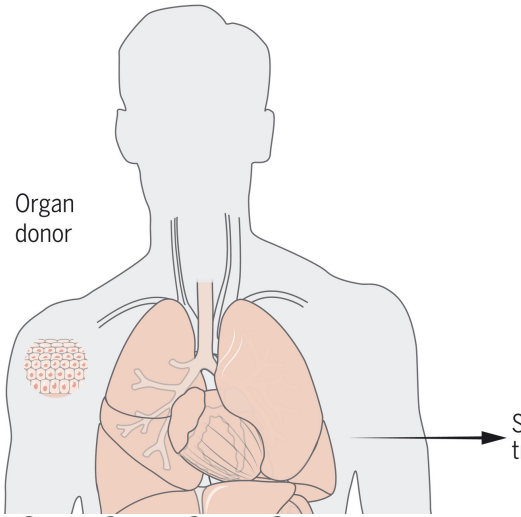
Arijit Sehanobish*† Yale University arijit.sehanobish@yale.edu

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

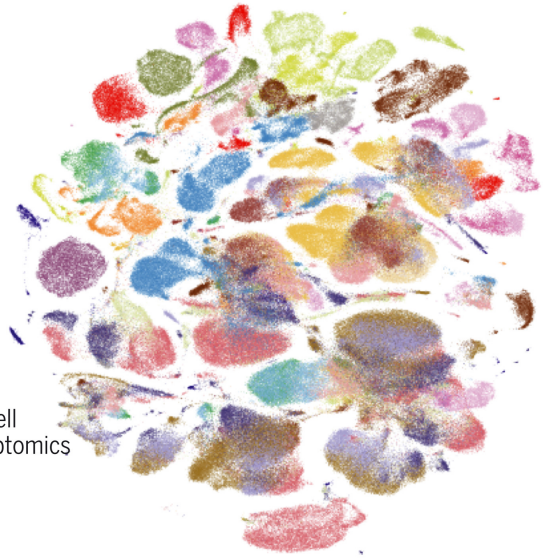
David A. Hafler‡ Yale University david.hafler@yale.edu

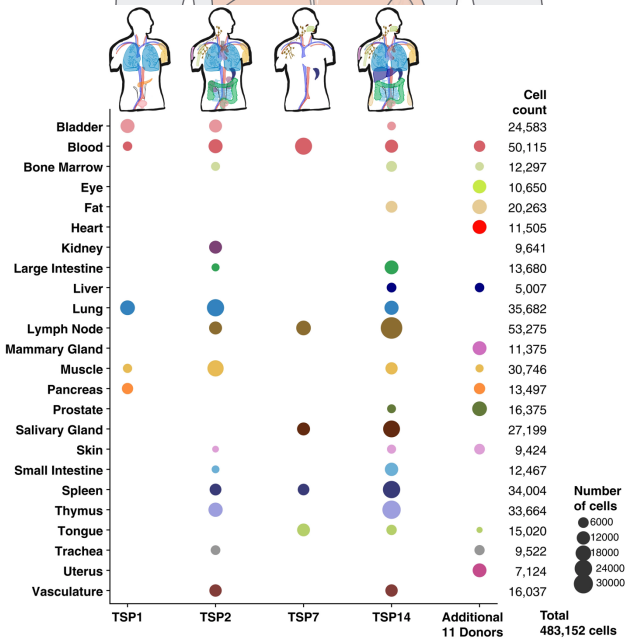
David van Dijk† Yale University david.vandijk@yale.edu

Cross-tissue human cell atlases



Single-cell
transcriptomics





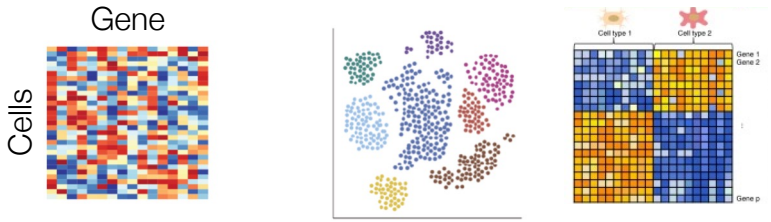
1. Which cell types are disease genes active?

2. How does a disease gene's expression vary within the same cell type? Across cell types?

The Tabula Sapiens: A multiple-organ, single-cell transcriptomic atlas of humans, *Science*, 2022.

Limitations of existing methods

scRNA-seq pipelines studying genetics & disease



Filtering, normalizing, and clustering

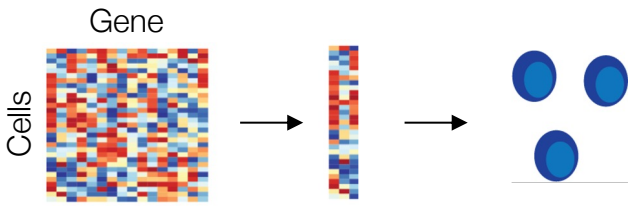
Differential gene expression (DGE) analysis

Limitations of using differential gene expression (DGE)

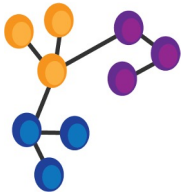
- DGE is not necessarily *associated* with disease or cell state
- The “most differentially” expressed genes do not yield causal structure
- Most DGE methods don’t allow for interactions between features

Disease State Prediction From Single-Cell Data Using Graph Attention Networks, *ACM CHIL*, 2020.

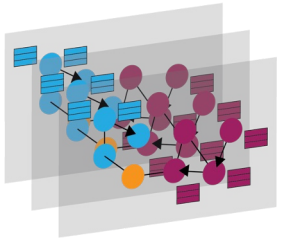
Leveraging cell-cell interactions



Identify cells from transcriptome



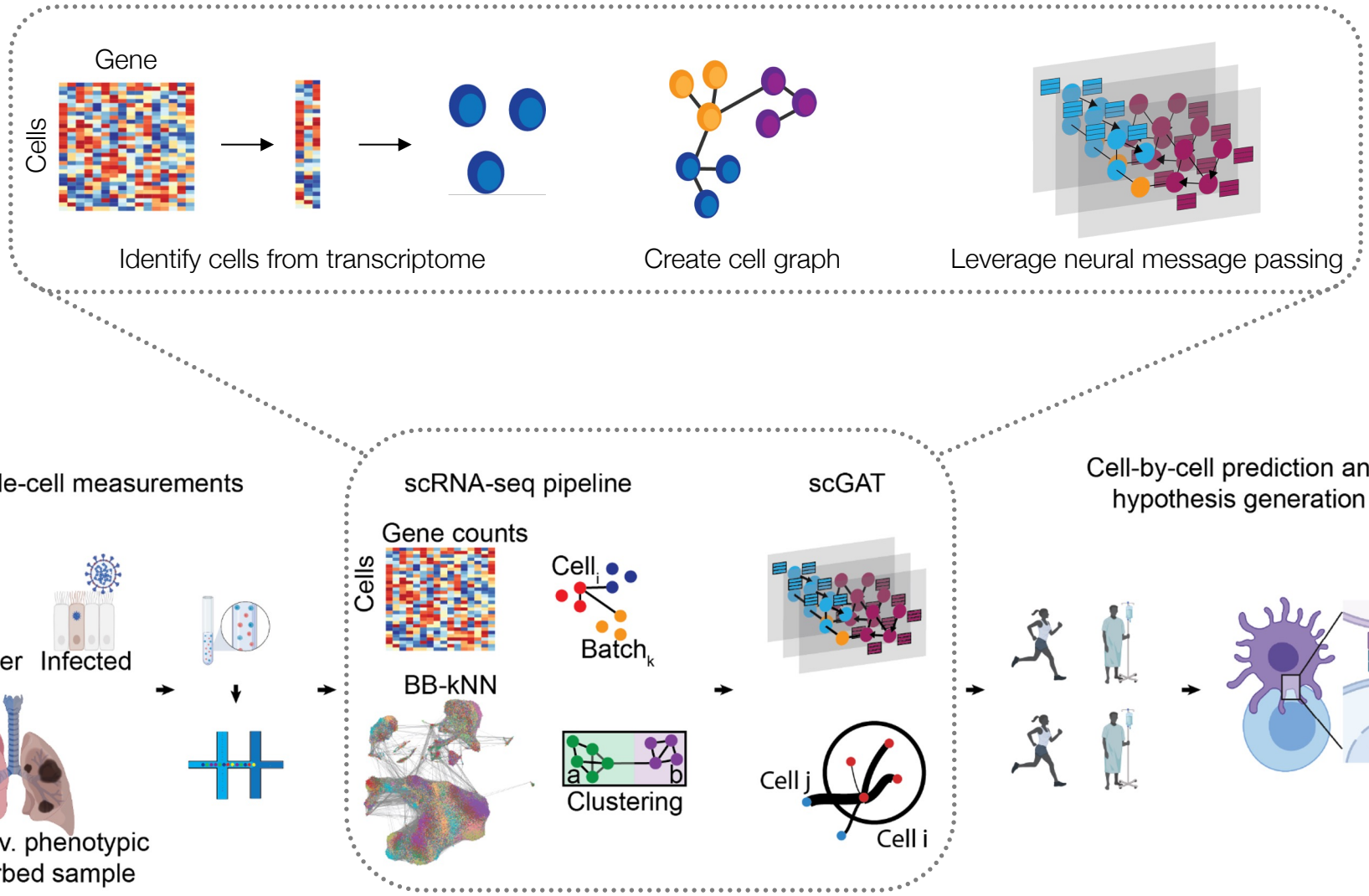
Create cell graph



Leverage neural message passing

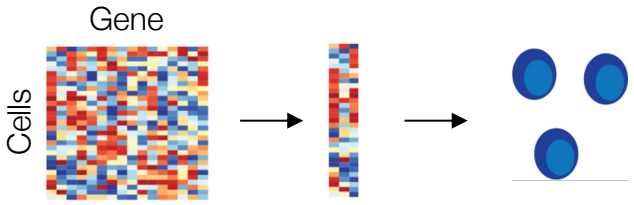
Disease State Prediction From Single-Cell Data Using Graph Attention Networks, *ACM CHIL*, 2020.

Leveraging cell-cell interactions

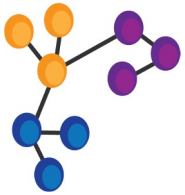


Disease State Prediction From Single-Cell Data Using Graph Attention Networks, *ACM CHIL*, 2020.

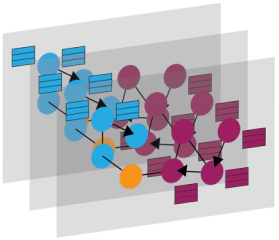
Leveraging cell-cell interactions



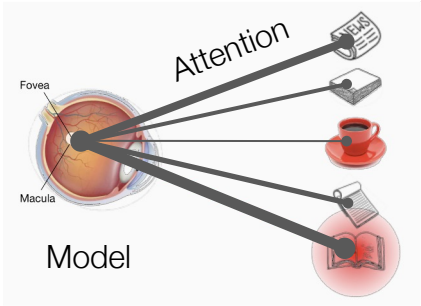
Identify cells from transcriptome

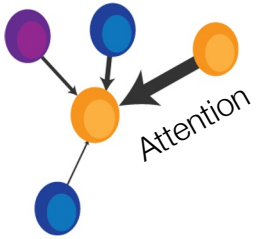


Create cell graph



Leverage neural message passing



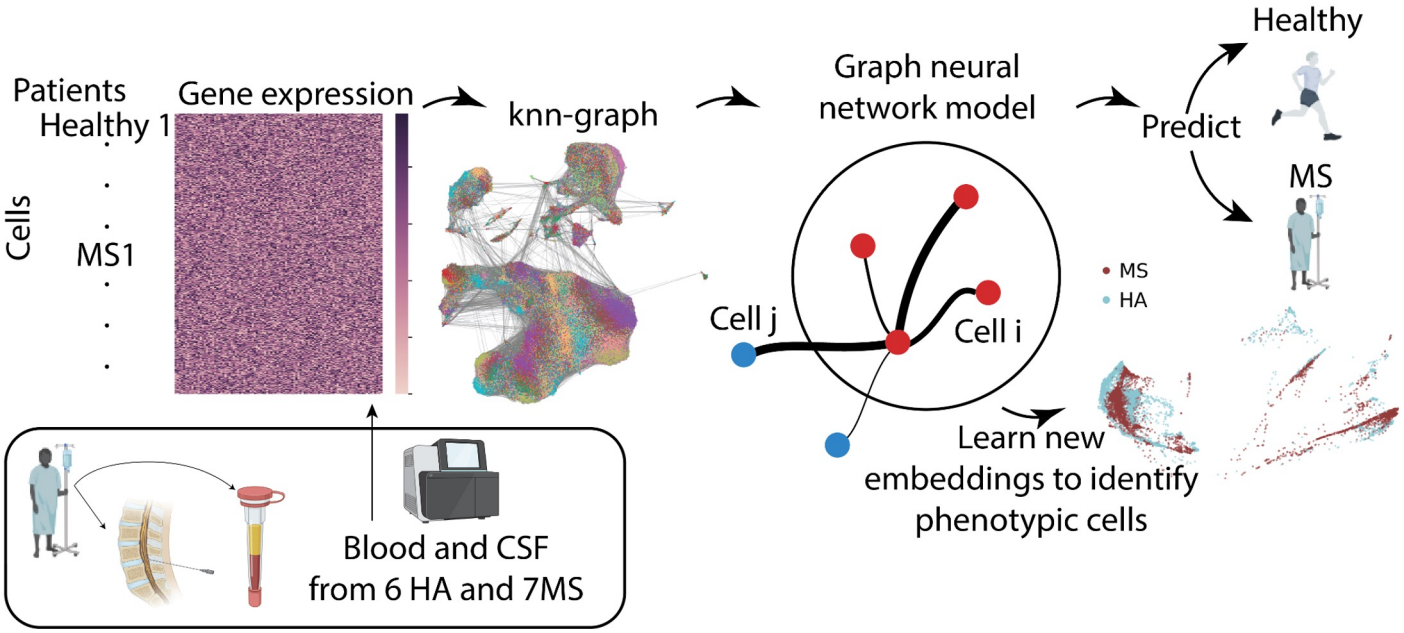


Which cell-cell interactions contribute the most/least to a specific disease state?

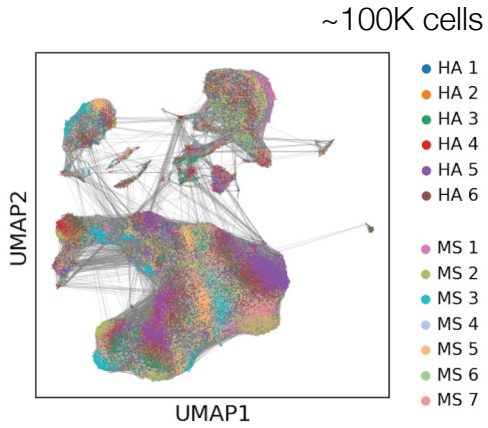
Disease State Prediction From Single-Cell Data Using Graph Attention Networks, *ACM CHIL*, 2020.

Experimental Setup

Overview



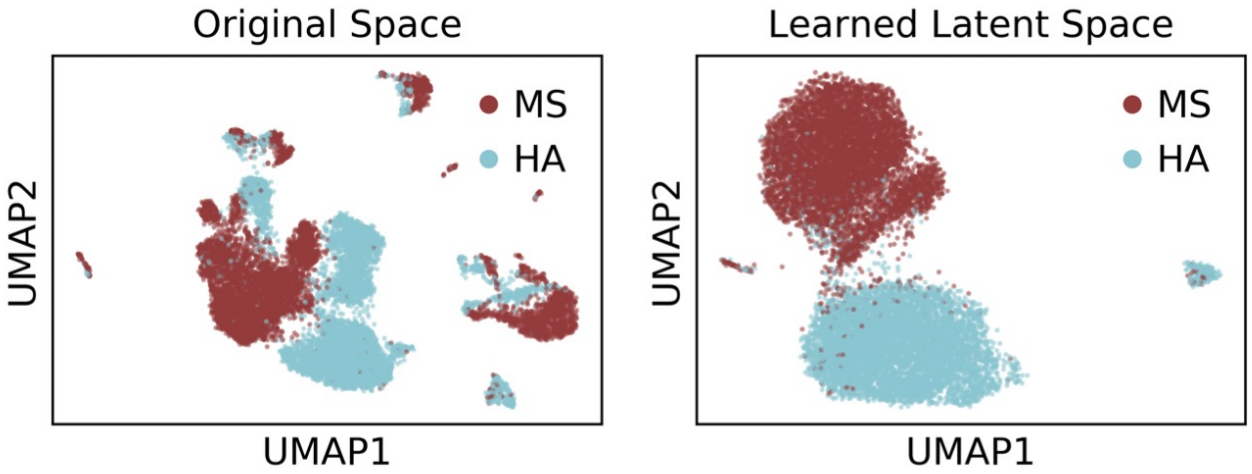
Patient Dataset



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		

Disease State Prediction From Single-Cell Data Using Graph Attention Networks, *ACM CHIL*, 2020.

Results

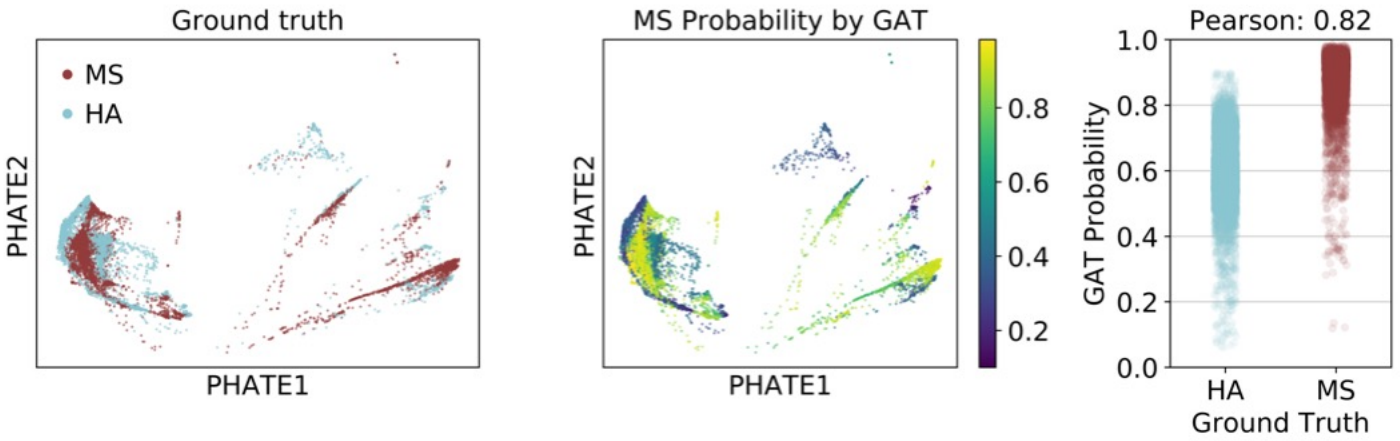


Task	Model	Accuracy
Inductive	Random	51.8
	MLP	56.7
	Random Forest	58.5
	Graph Convolutional Network	72.1
	Graph Attention Network(our)	**92.3** $\pm$ .7
Transductive	Graph Convolutional Network	82.91
	Graph Attention Network(our)	**86** $\pm$ .3

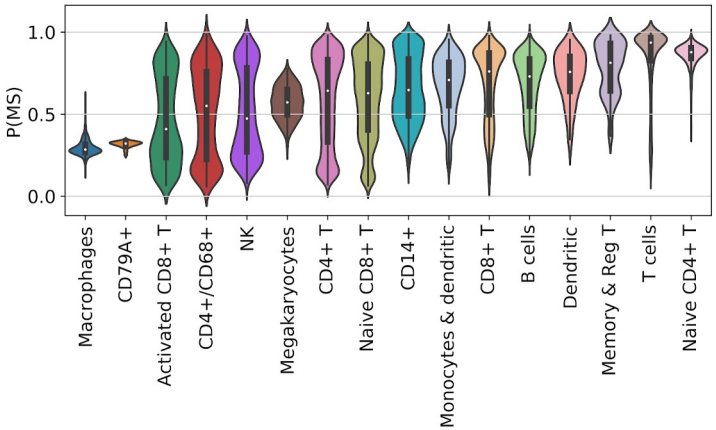
- Transductive task: Randomly assign 10% nodes for validation & 10% for testing
  - Keeping ratio of healthy & MS cells same as in full dataset
- Inductive task: Randomly choose a healthy adult & MS patient
  - Train on remaining 5 MS patients and 4 healthy adults

Disease State Prediction From Single-Cell Data Using Graph Attention Networks, *ACM CHIL*, 2020.

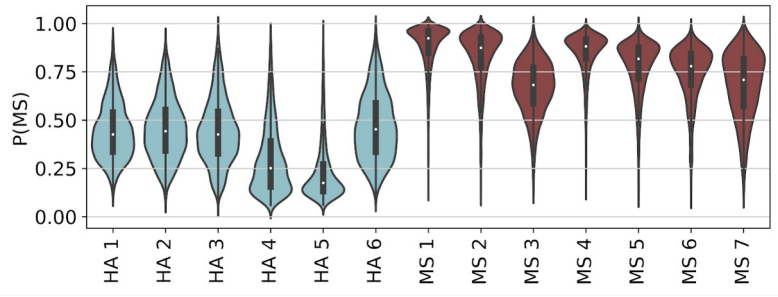
Results



Predicted probabilities from induction task per cell



Aggregating predicted probabilities shows cell types important for predicting disease state



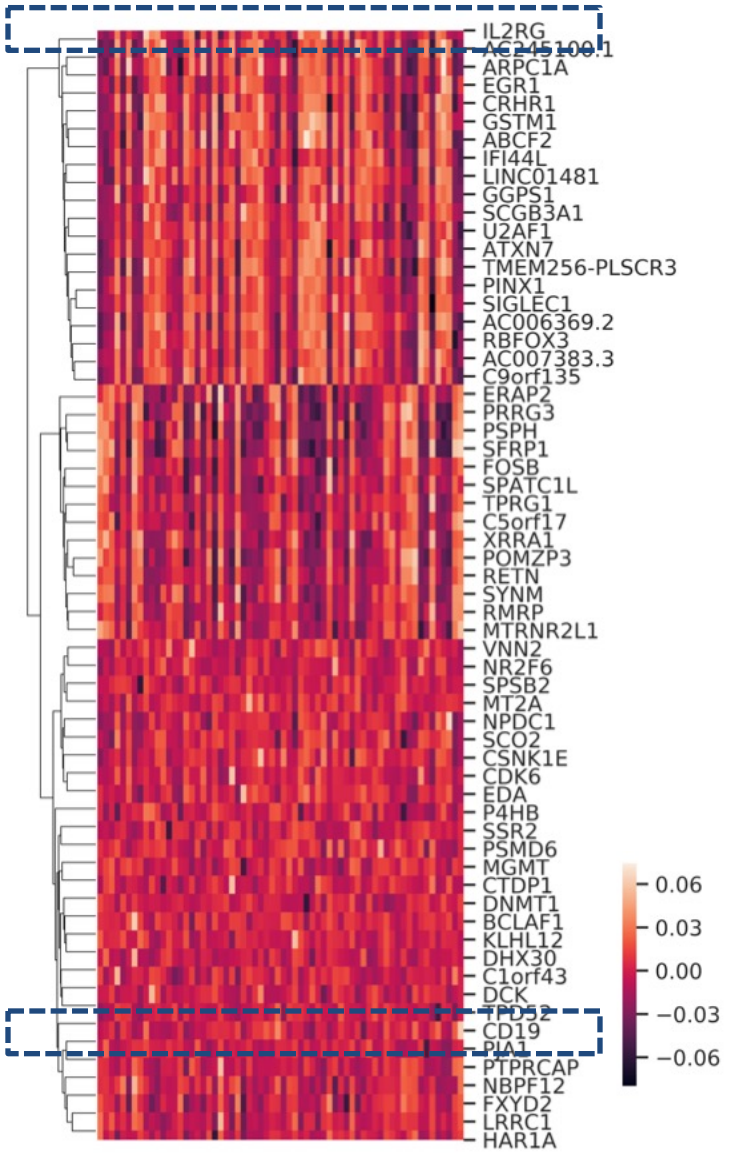
Variance of a patient's cells' probability of being in an MS state may indicate timing of flare-up

Disease State Prediction From Single-Cell Data Using Graph Attention Networks, *ACM CHIL*, 2020.

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



Disease State Prediction From Single-Cell Data Using Graph Attention Networks, *ACM CHIL*, 2020.

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
  - GitHub: github.com/vandijklab/scGAT
  - Follow-up work on COVID-19: 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

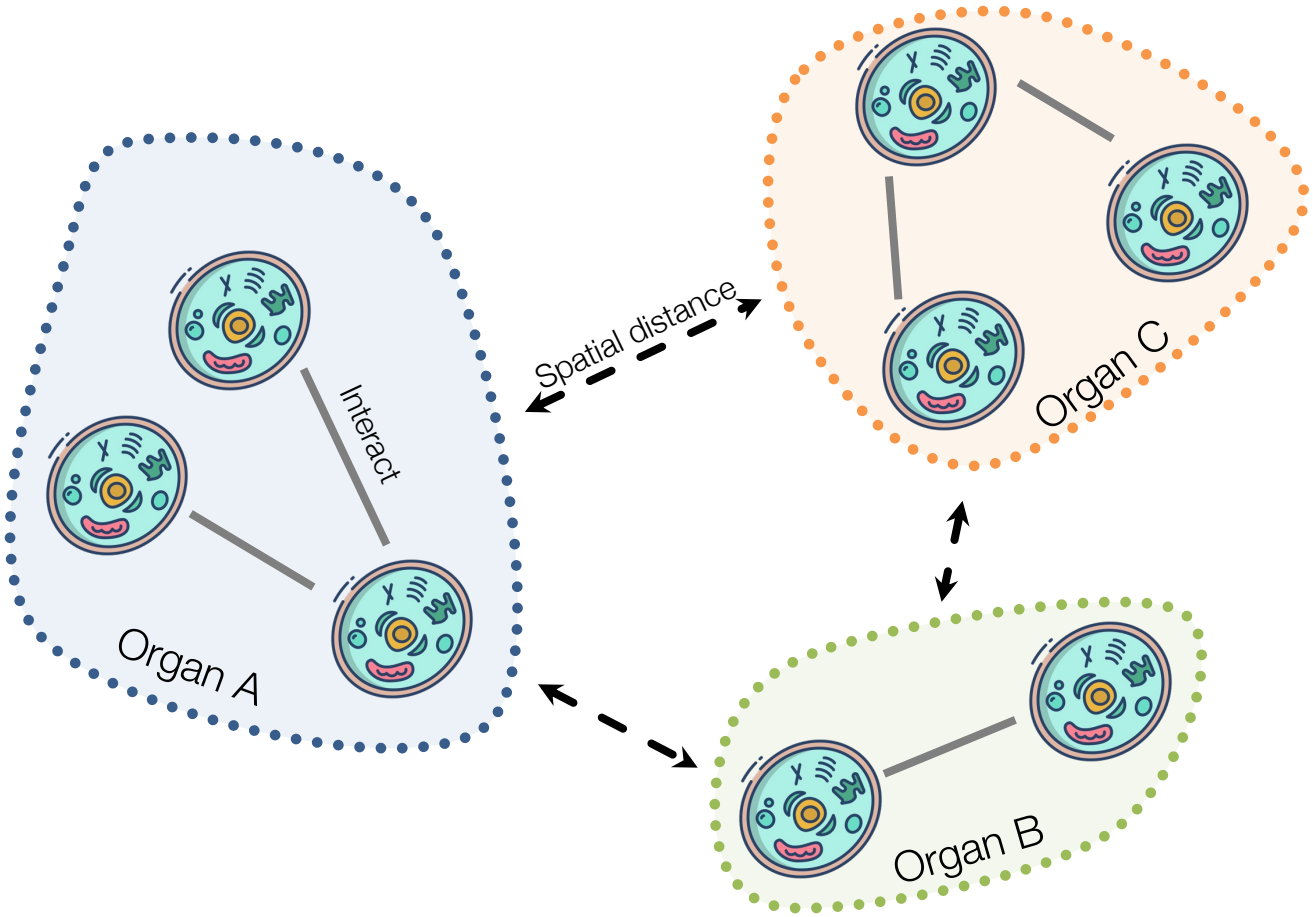
1. Single-cell transcriptomics data
2. Spatial transcriptomics data

METHOD Open Access

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

Ye Yuan[1] and Ziv Bar-Joseph[1,2*]

Motivation

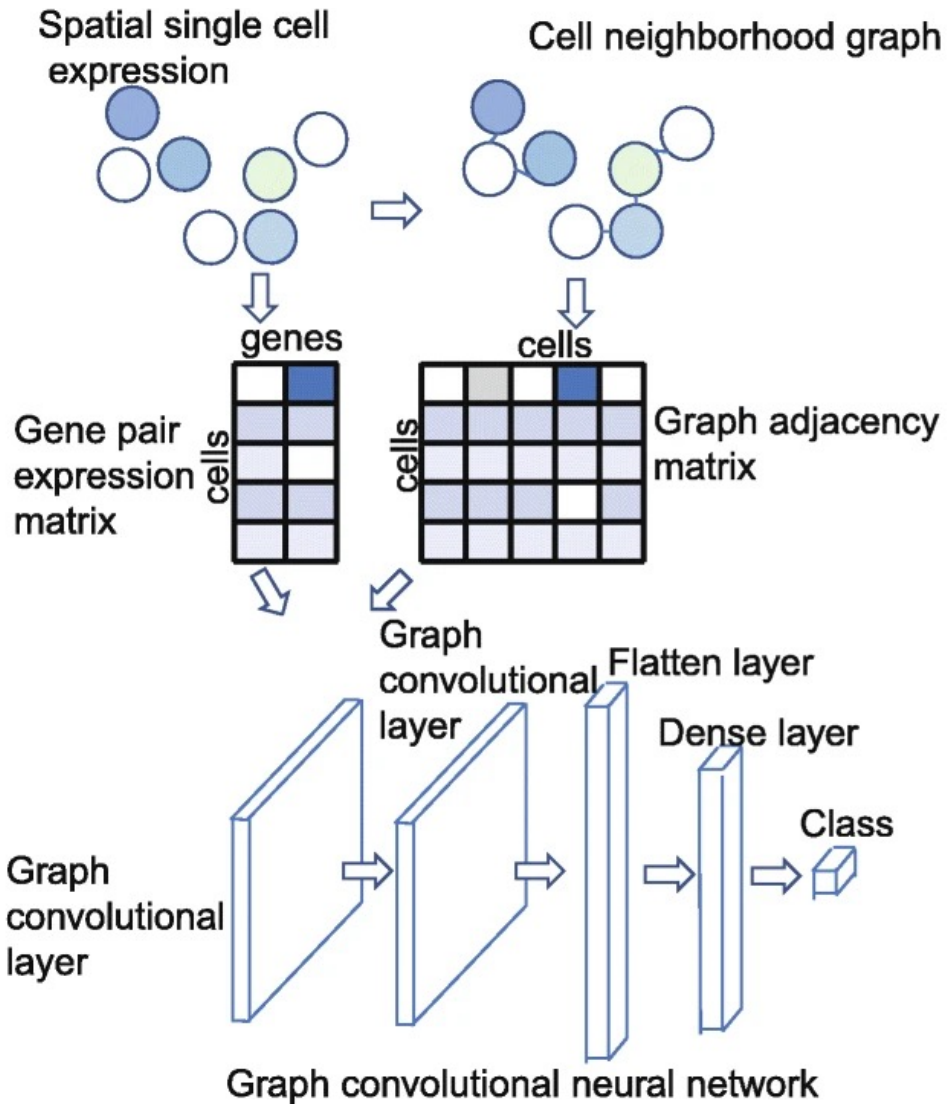


1. Leverage higher-order interactions between cells

2. Utilize both gene expression & cellular organization

3. Overcome incomplete spatial relationships

Overview

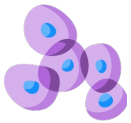


GCNG: Graph convolutional networks for inferring gene interactions from spatial transcriptomics data, *Genome Biology*, 2020.

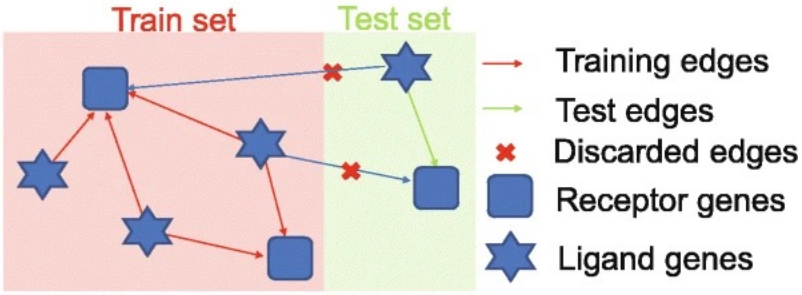
Experimental setup

SeqFISH+ 

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

MERFISH 

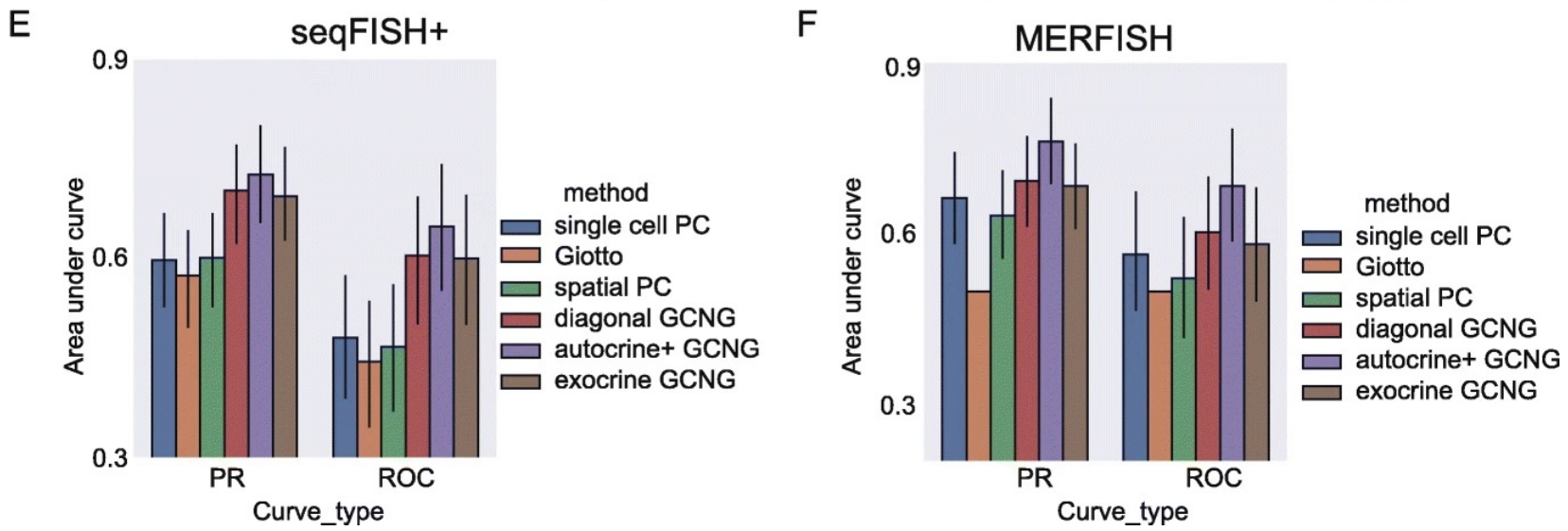
- Cells (in vitro)
- Expression data: 10,050 genes from 1368 cells
- Labeled ligand-receptor pairs: 841 known interactions between 270 ligands and 376 receptors



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

GCNG: Graph convolutional networks for inferring gene interactions from spatial transcriptomics data, *Genome Biology*, 2020.

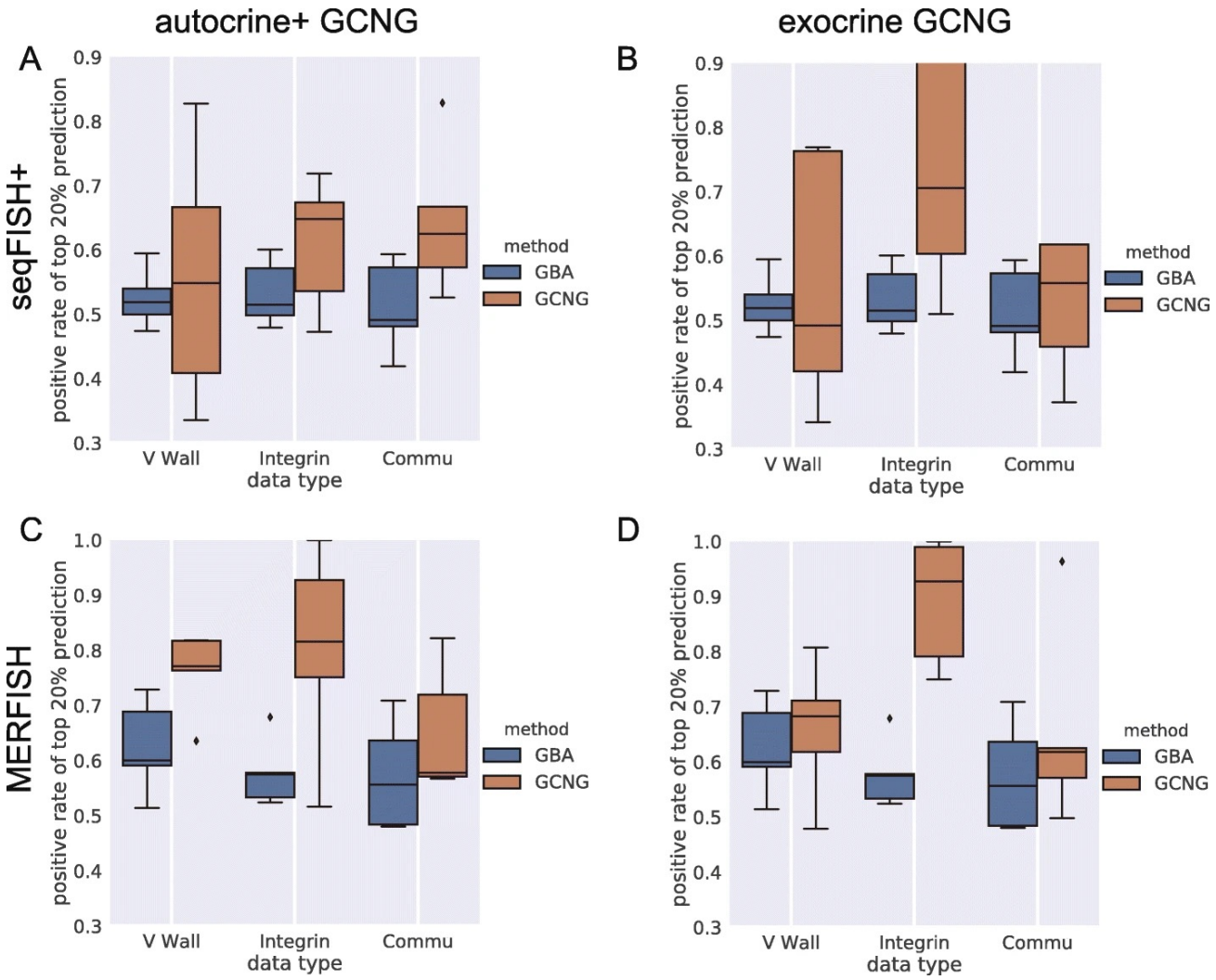
Results: Inferring L-R interaction



- Single cell Pearson correlation (PC): Pearson correlation between the expression of ligands and receptors within each cell
- Giotto: Calculate similarity score for all pairs of genes in all pairs of neighboring cell types → Rank pairs based on average score
- Spatial PC: PC between ligand and receptors in neighboring cells
- Diagonal GCNG: Only uses a diagonal matrix to represent the graph → Only autocrine interactions are possible
- Exocrine GCNG: Only exocrine interaction between cells are allowed
- Autocrine+ GCNG: Both autocrine and exocrine interactions

GCNG: Graph convolutional networks for inferring gene interactions from spatial transcriptomics data, *Genome Biology*, 2020.

Results: Functional prediction



GCNG: Graph convolutional networks for inferring gene interactions from spatial transcriptomics data, *Genome Biology*, 2020.

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
  - GitHub: 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

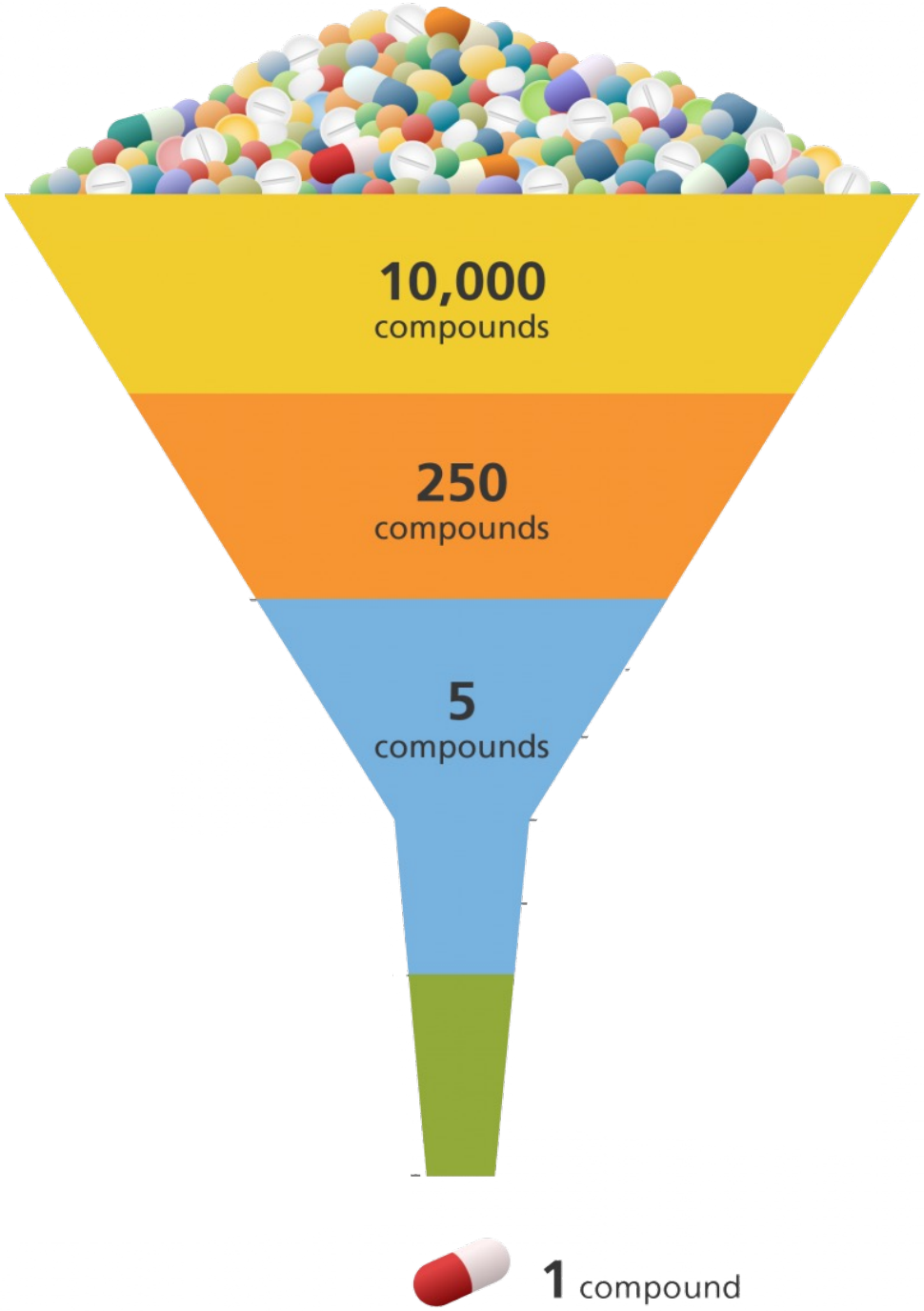
Step 1: Design and Discovery

Step 2: Preclinical Research

Step 3: Clinical Research

Step 4: FDA Review

Step 5: Post-Market and Safety Monitoring



Applications of graph representation learning on…

THERAPEUTICS

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

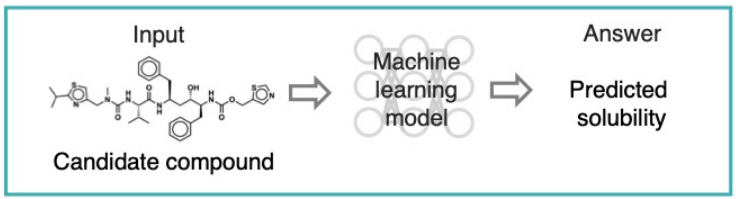
Therapeutics Data Commons: Machine Learning Datasets and Tasks for Drug Discovery and Development

---

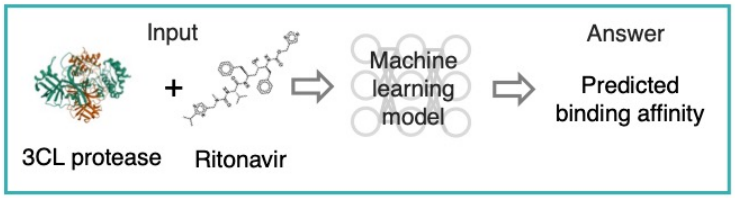
Kexin Huang[1]*, Tianfan Fu[2]*, Wenhao Gao[3]*, Yue Zhao[4], Yusuf Roohani[5], Jure Leskovec[5], Connor W. Coley[3], Cao Xiao[6], Jimeng Sun[7], Marinka Zitnik[1]
[1]Harvard [2]Georgia Tech [3]MIT [4]CMU [5]Stanford [6]Amplitude [7]UIUC
contact@tdcommons.ai

Compelling applications of graph RL

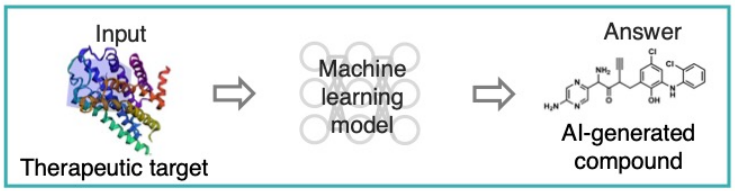
I want to know the solubility of a compound of interest.



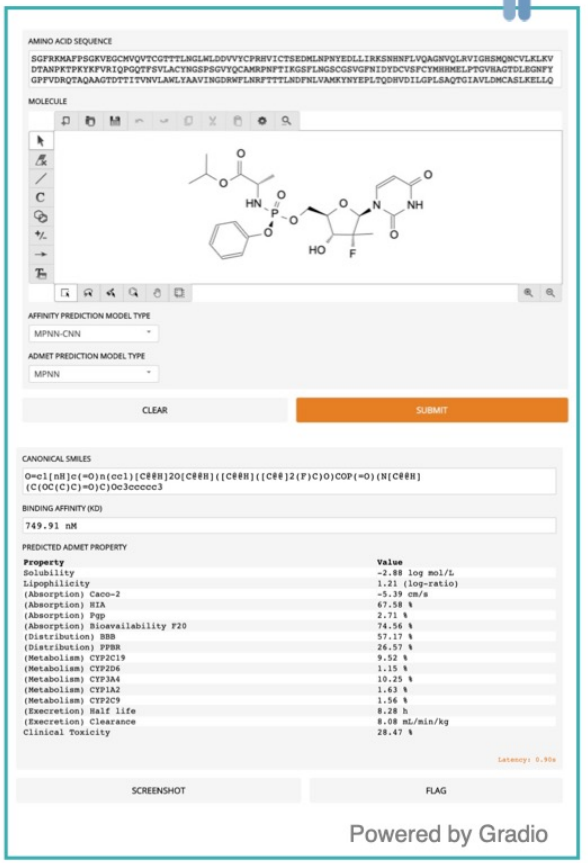
I want to know the binding affinity of Ritonavir to 3CL protease.



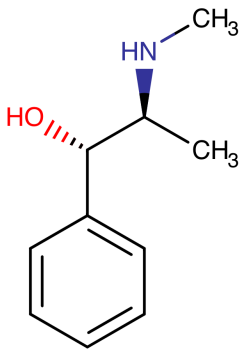
I want to generate a highly potent compound that effectively binds a therapeutic target.



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



ADMET property prediction



Model

e.g.
Fingerprints
CNN+SMILES
GNN
....

Absorption
Distribution
Metabolism
Excretion
Toxicity
Endpoints

Datasets

Paper
Supplementary

Public
Database

22 datasets with ADMET endpoints

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

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

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

Excretion
Half Life
Clearance (Hepatocyte)
Clearance (Microsome)

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

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	**0.393±0.024**	0.446±0.036	0.530±0.102	0.599±0.104	0.401±0.032	0.546±0.052	0.502±0.036
TDC.HIA (↑)	AUROC	0.807±0.072	0.972±0.008	0.869±0.026	0.943±0.014	0.936±0.024	0.974±0.007	**0.978±0.006**	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	**0.929±0.006**	0.923±0.005
TDC.Bioav (↑)	AUROC	0.581±0.086	**0.672±0.021**	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	0.574±0.017	0.743±0.020	0.563±0.023	0.541±0.011	0.572±0.007	0.547±0.024	**0.535±0.012**
TDC.AqSol (↓)	MAE	1.203±0.019	0.827±0.047	1.023±0.023	0.947±0.016	0.907±0.020	**0.776±0.008**	1.026±0.020	1.040±0.045
TDC.BBB (↑)	AUROC	0.823±0.015	0.889±0.016	0.781±0.030	0.836±0.009	0.842±0.016	0.855±0.011	0.892±0.012	**0.897±0.004**
TDC.PPBR (↓)	MAE	12.848±0.362	9.994±0.319	11.106±0.358	**9.292±0.384**	10.194±0.373	9.373±0.335	10.075±0.202	9.445±0.224
TDC.VD (↑)	Spearman	0.493±0.011	**0.561±0.025**	0.226±0.114	0.258±0.162	0.457±0.050	0.241±0.145	0.559±0.019	0.485±0.092
TDC.CYP2D6-I (↑)	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	0.721±0.009	**0.739±0.005**
TDC.CYP3A4-I (↑)	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	0.902±0.002	**0.904±0.002**
TDC.CYP2C9-I (↑)	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	0.829±0.003	**0.839±0.003**
TDC.CYP2D6-S (↑)	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	0.704±0.028	**0.736±0.024**
TDC.CYP3A4-S (↑)	AUROC	0.633±0.013	0.639±0.012	**0.662±0.031**	0.578±0.020	0.590±0.023	0.576±0.025	0.582±0.021	0.609±0.025
TDC.CYP2C9-S (↑)	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	0.381±0.045	**0.392±0.026**
TDC.Half_Life (↑)	Spearman	**0.329±0.083**	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
TDC.CL-Micro (↑)	Spearman	0.492±0.020	**0.586±0.014**	0.252±0.116	0.529±0.015	0.532±0.033	0.365±0.055	0.585±0.034	0.578±0.007
TDC.CL-Hepa (↑)	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	0.413±0.028	**0.439±0.026**
TDC.hERG (↑)	AUROC	0.736±0.023	**0.841±0.020**	0.754±0.037	0.722±0.034	0.738±0.038	0.825±0.007	0.778±0.046	0.756±0.023
TDC.AMES (↑)	AUROC	0.794±0.008	0.823±0.011	0.776±0.015	0.823±0.006	0.818±0.010	0.814±0.008	**0.842±0.008**	0.837±0.009
TDC.DILI (↑)	AUROC	0.832±0.021	0.875±0.019	0.792±0.016	0.851±0.026	0.859±0.033	0.886±0.015	**0.919±0.008**	0.861±0.018
TDC.LD50 (↓)	MAE	0.649±0.019	0.678±0.003	0.675±0.011	0.667±0.020	0.649±0.026	0.678±0.012	**0.685±0.025**	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
TDC.Caco2 (↓)	MAE	0.908±0.060	**0.393±0.024**	0.446±0.036	0.530±0.102	0.599±0.104	0.401±0.032	0.546±0.052	0.502±0.036
TDC.HIA (↑)	AUROC	0.807±0.072	0.972±0.008	0.869±0.026	0.943±0.014	0.936±0.024	0.974±0.007	**0.978±0.006**	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	**0.929±0.006**	0.923±0.005
TDC.Bioav (↑)	AUROC	0.581±0.086	**0.672±0.021**	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	0.574±0.017	0.743±0.020	0.563±0.023	0.541±0.011	0.572±0.007	0.547±0.024	**0.535±0.012**
TDC.AqSol (↓)	MAE	1.203±0.019	0.827±0.047	1.023±0.023	0.947±0.016	0.907±0.020	**0.776±0.008**	1.026±0.020	1.040±0.045
TDC.BBB (↑)	AUROC	0.823±0.015	0.889±0.016	0.781±0.030	0.836±0.009	0.842±0.016	0.855±0.011	0.892±0.012	**0.897±0.004**
TDC.PPBR (↓)	MAE	12.848±0.362	9.994±0.319	11.106±0.358	**9.292±0.384**	10.194±0.373	9.373±0.335	10.075±0.202	9.445±0.224
TDC.VD (↑)	Spearman	0.493±0.011	**0.561±0.025**	0.226±0.114	0.258±0.162	0.457±0.050	0.241±0.145	0.559±0.019	0.485±0.092
TDC.CYP2D6-I (↑)	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	0.721±0.009	**0.739±0.005**
TDC.CYP3A4-I (↑)	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	0.902±0.002	**0.904±0.002**
TDC.CYP2C9-I (↑)	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	0.829±0.003	**0.839±0.003**
TDC.CYP2D6-S (↑)	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	0.704±0.028	**0.736±0.024**
TDC.CYP3A4-S (↑)	AUROC	0.633±0.013	0.639±0.012	**0.662±0.031**	0.578±0.020	0.590±0.023	0.576±0.025	0.582±0.021	0.609±0.025
TDC.CYP2C9-S (↑)	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	0.381±0.045	**0.392±0.026**
TDC.Half_Life (↑)	Spearman	**0.329±0.083**	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
TDC.CL-Micro (↑)	Spearman	0.492±0.020	**0.586±0.014**	0.252±0.116	0.529±0.015	0.532±0.033	0.365±0.055	0.585±0.034	0.578±0.007
TDC.CL-Hepa (↑)	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	0.413±0.028	**0.439±0.026**
TDC.hERG (↑)	AUROC	0.736±0.023	**0.841±0.020**	0.754±0.037	0.722±0.034	0.738±0.038	0.825±0.007	0.778±0.046	0.756±0.023
TDC.AMES (↑)	AUROC	0.794±0.008	0.823±0.011	0.776±0.015	0.823±0.006	0.818±0.010	0.814±0.008	**0.842±0.008**	0.837±0.009
TDC.DILI (↑)	AUROC	0.832±0.021	0.875±0.019	0.792±0.016	0.851±0.026	0.859±0.033	0.886±0.015	**0.919±0.008**	0.861±0.018
TDC.LD50 (↓)	MAE	0.649±0.019	0.678±0.003	0.675±0.011	0.667±0.020	0.649±0.026	0.678±0.012	**0.685±0.025**	0.669±0.030

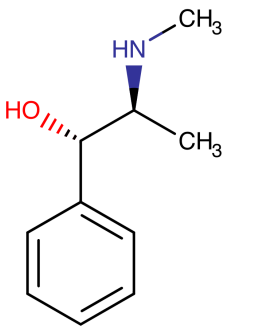
- **Finding 2**: Expert-curated methods, such as Morgan's fingerprints can outperform graph RL methods on some endpoints

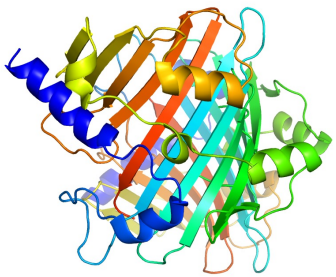
Results: ADMET prediction (3/3)

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
TDC.Caco2 (↓)	MAE	0.908±0.060	**0.393±0.024**	0.446±0.036	0.530±0.102	0.599±0.104	0.401±0.032	0.546±0.052	0.502±0.036
TDC.HIA (↑)	AUROC	0.807±0.072	0.972±0.008	0.869±0.026	0.943±0.014	0.936±0.024	0.974±0.007	**0.978±0.006**	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	**0.929±0.006**	0.923±0.005
TDC.Bioav (↑)	AUROC	0.581±0.086	**0.672±0.021**	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	0.574±0.017	0.743±0.020	0.563±0.023	0.541±0.011	0.572±0.007	0.547±0.024	**0.535±0.012**
TDC.AqSol (↓)	MAE	1.203±0.019	0.827±0.047	1.023±0.023	0.947±0.016	0.907±0.020	**0.776±0.008**	1.026±0.020	1.040±0.045
TDC.BBB (↑)	AUROC	0.823±0.015	0.889±0.016	0.781±0.030	0.836±0.009	0.842±0.016	0.855±0.011	0.892±0.012	**0.897±0.004**
TDC.PPBR (↓)	MAE	12.848±0.362	9.994±0.319	11.106±0.358	**9.292±0.384**	10.194±0.373	9.373±0.335	10.075±0.202	9.445±0.224
TDC.VD (↑)	Spearman	0.493±0.011	**0.561±0.025**	0.226±0.114	0.258±0.162	0.457±0.050	0.241±0.145	0.559±0.019	0.485±0.092
TDC.CYP2D6-I (↑)	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	0.721±0.009	**0.739±0.005**
TDC.CYP3A4-I (↑)	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	0.902±0.002	**0.904±0.002**
TDC.CYP2C9-I (↑)	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	0.829±0.003	**0.839±0.003**
TDC.CYP2D6-S (↑)	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	0.704±0.028	**0.736±0.024**
TDC.CYP3A4-S (↑)	AUROC	0.633±0.013	0.639±0.012	**0.662±0.031**	0.578±0.020	0.590±0.023	0.576±0.025	0.582±0.021	0.609±0.025
TDC.CYP2C9-S (↑)	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	0.381±0.045	**0.392±0.026**
TDC.Half_Life (↑)	Spearman	**0.329±0.083**	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
TDC.CL-Micro (↑)	Spearman	0.492±0.020	**0.586±0.014**	0.252±0.116	0.529±0.015	0.532±0.033	0.365±0.055	0.585±0.034	0.578±0.007
TDC.CL-Hepa (↑)	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	0.413±0.028	**0.439±0.026**
TDC.hERG (↑)	AUROC	0.736±0.023	**0.841±0.020**	0.754±0.037	0.722±0.034	0.738±0.038	0.825±0.007	0.778±0.046	0.756±0.023
TDC.AMES (↑)	AUROC	0.794±0.008	0.823±0.011	0.776±0.015	0.823±0.006	0.818±0.010	0.814±0.008	**0.842±0.008**	0.837±0.009
TDC.DILI (↑)	AUROC	0.832±0.021	0.875±0.019	0.792±0.016	0.851±0.026	0.859±0.033	0.886±0.015	**0.919±0.008**	0.861±0.018
TDC.LD50 (↓)	MAE	0.649±0.019	0.678±0.003	0.675±0.011	0.667±0.020	0.649±0.026	0.678±0.012	**0.685±0.025**	0.669±0.030

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

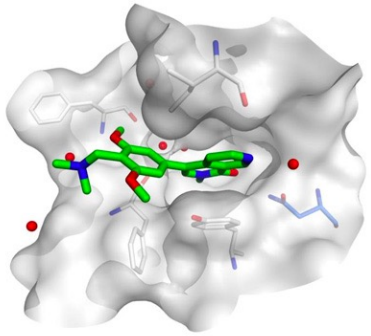
Drug-target interaction prediction





Model

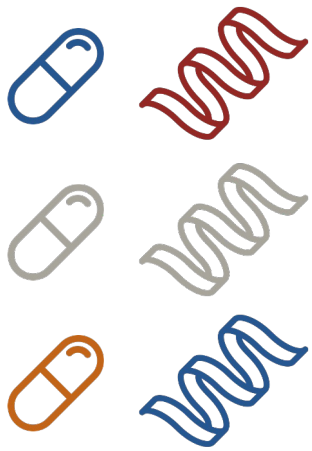
Binding Affinity

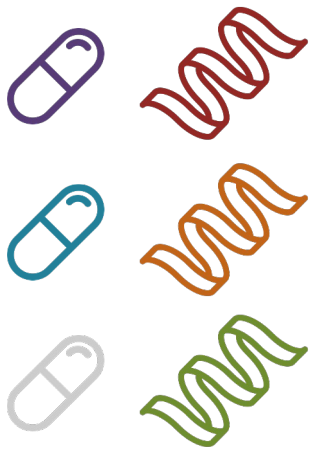


Setup: Distribution shifts and generalization

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!

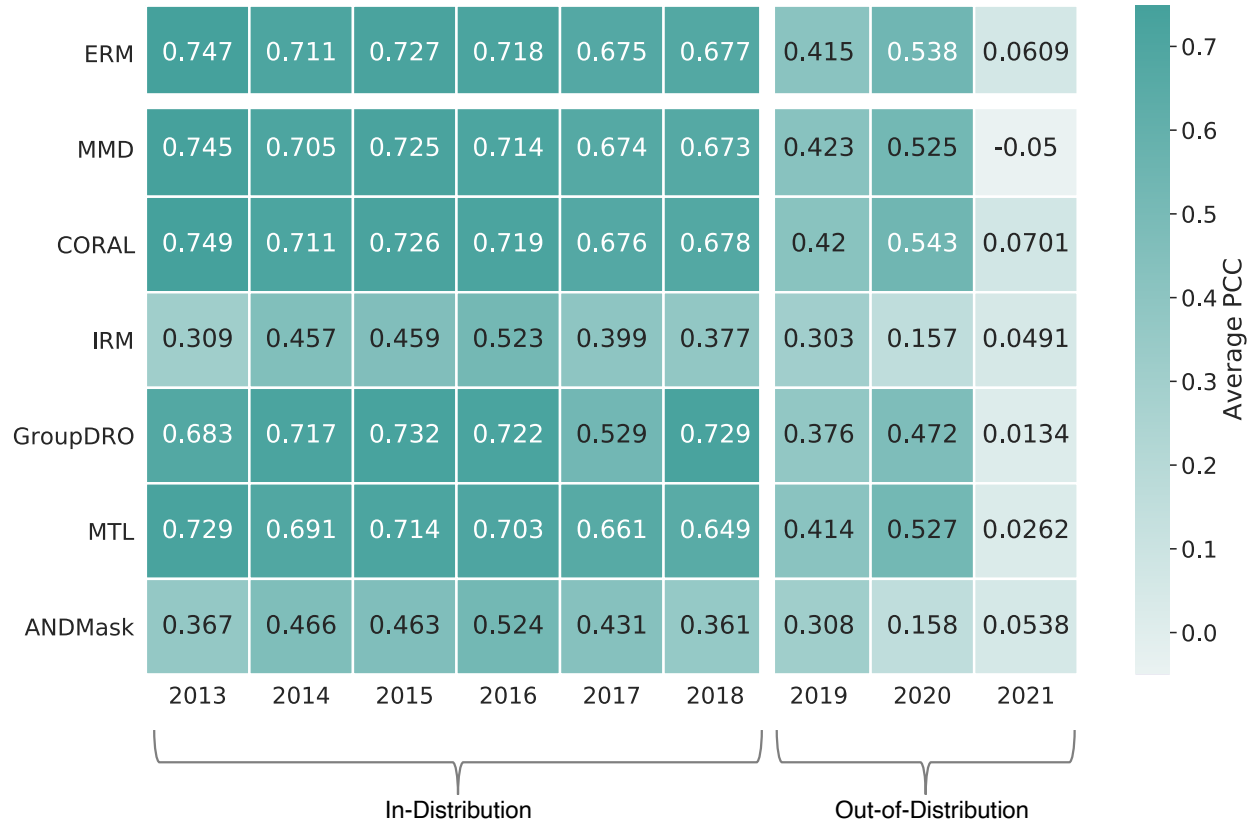




Train-Valid: DTIs Patented in **2013-18**

Test: DTIs Patented in **2019-21**

Results

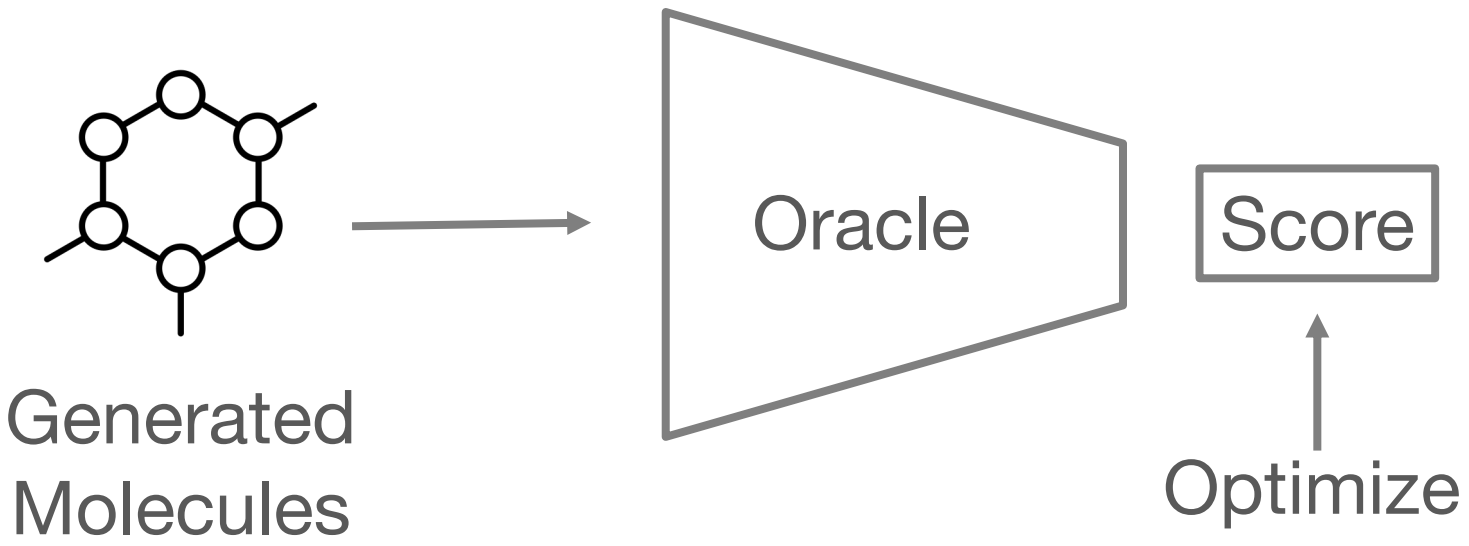


ERM (Empirical Risk Minimization) is a standard training strategy where errors across all domains are minimized.

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

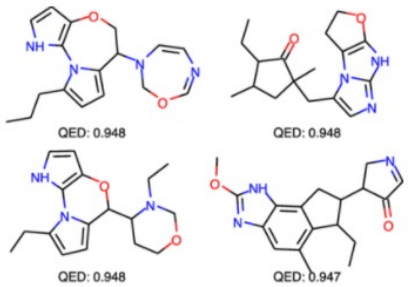


Setup: High-capacity oracles (1/2)

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!

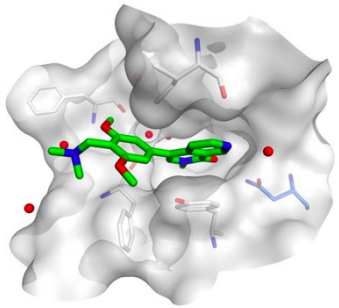
Previous oracle



Milliseconds in RDKit
SOTA methods call millions of times!

vs

Docking oracle



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)

*https://molecule.one

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	1000	-9.693±0.019	**-11.224±0.484**	-9.971±0.115	-9.053±0.080	-6.738±0.042	-8.224±0.196
Top10 (↓)	-12.590		-10.777±0.189	**-12.400±0.782**	-11.163±0.141	-11.027±0.273	-7.506±0.085	-9.843±0.068
Top1 (↓)	-12.800		-11.500±0.432	**-13.233±0.713**	-11.967±0.205	-12.033±0.618	-7.800±0.042	-11.100±0.141
Diversity (↑)	0.864		0.873±0.003	0.815±0.046	0.871±0.004	**0.913±0.001**	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	**0.777±0.096**	0.777±0.026	0.170±0.022	0.033±0.005	0.563±0.052
Top1 Pass (↓)	-11.700		-9.167±0.047	**-10.600±0.374**	-9.367±0.094	-8.167±0.047	-6.450±0.085	-7.367±0.205
m1 (↓)	5.100		5.527±0.780	7.695±0.909	**4.818±0.541**	10.000±0.000	10.000±0.000	6.037±0.137
Top100 (↓)	-12.080	5000	-10.542±0.035	**-14.811±0.413**	-13.017±0.385	-10.045±0.226	-8.236±0.089	-9.509±0.035
Top10 (↓)	-12.590		-11.483±0.056	**-15.930±0.336**	-14.030±0.421	-11.483±0.581	-9.348±0.188	-10.693±0.172
Top1 (↓)	-12.800		-12.100±0.356	**-16.533±0.309**	-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	**0.922±0.002**	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		**0.683±0.073**	0.393±0.308	0.257±0.103	0.167±0.045	0.023±0.012	0.527±0.087
Top1 Pass (↓)	-11.700		-10.100±0.000	**-14.267±0.450**	-12.533±0.403	-9.367±0.170	-7.980±0.112	-9.000±0.082
m1 (↓)	5.100		**5.610±0.805**	9.669±0.468	5.826±1.908	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	**-11.224±0.484**	-9.971±0.115	-9.053±0.080	-6.738±0.042	-8.224±0.196
Top10 (↓)	-12.590		-10.777±0.189	**-12.400±0.782**	-11.163±0.141	-11.027±0.273	-7.506±0.085	-9.843±0.068
Top1 (↓)	-12.800		-11.500±0.432	**-13.233±0.713**	-11.967±0.205	-12.033±0.618	-7.800±0.042	-11.100±0.141
Diversity (↑)	0.864		0.873±0.003	0.815±0.046	0.871±0.004	**0.913±0.001**	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	**0.777±0.096**	0.777±0.026	0.170±0.022	0.033±0.005	0.563±0.052
Top1 Pass (↓)	-11.700		-9.167±0.047	**-10.600±0.374**	-9.367±0.094	-8.167±0.047	-6.450±0.085	-7.367±0.205
m1 (↓)	5.100		5.527±0.780	7.695±0.909	**4.818±0.541**	10.000±0.000	10.000±0.000	6.037±0.137
Top100 (↓)	-12.080	5000	-10.542±0.035	**-14.811±0.413**	-13.017±0.385	-10.045±0.226	-8.236±0.089	-9.509±0.035
Top10 (↓)	-12.590		-11.483±0.056	**-15.930±0.336**	-14.030±0.421	-11.483±0.581	-9.348±0.188	-10.693±0.172
Top1 (↓)	-12.800		-12.100±0.356	**-16.533±0.309**	-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	**0.922±0.002**	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		**0.683±0.073**	0.393±0.308	0.257±0.103	0.167±0.045	0.023±0.012	0.527±0.087
Top1 Pass (↓)	-11.700		-10.100±0.000	**-14.267±0.450**	-12.533±0.403	-9.367±0.170	-7.980±0.112	-9.000±0.082
m1 (↓)	5.100		**5.610±0.805**	9.669±0.468	5.826±1.908	10.000±0.000	10.000±0.000	7.073±0.798

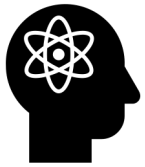
- **Finding 2**: Graph-GA method with 0 learnable parameters performs the best. SOTA ML methods report excellent results when resources are unlimited

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	**-11.224±0.484**	-9.971±0.115	-9.053±0.080	-6.738±0.042	-8.224±0.196
Top10 (↓)	-12.590		-10.777±0.189	**-12.400±0.782**	-11.163±0.141	-11.027±0.273	-7.506±0.085	-9.843±0.068
Top1 (↓)	-12.800		-11.500±0.432	**-13.233±0.713**	-11.967±0.205	-12.033±0.618	-7.800±0.042	-11.100±0.141
Diversity (↑)	0.864		0.873±0.003	0.815±0.046	0.871±0.004	**0.913±0.001**	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	**0.777±0.096**	0.777±0.026	0.170±0.022	0.033±0.005	0.563±0.052
Top1 Pass (↓)	-11.700		-9.167±0.047	**-10.600±0.374**	-9.367±0.094	-8.167±0.047	-6.450±0.085	-7.367±0.205
m1 (↓)	5.100		5.527±0.780	7.695±0.909	**4.818±0.541**	10.000±0.000	10.000±0.000	6.037±0.137
Top100 (↓)	-12.080	5000	-10.542±0.035	**-14.811±0.413**	-13.017±0.385	-10.045±0.226	-8.236±0.089	-9.509±0.035
Top10 (↓)	-12.590		-11.483±0.056	**-15.930±0.336**	-14.030±0.421	-11.483±0.581	-9.348±0.188	-10.693±0.172
Top1 (↓)	-12.800		-12.100±0.356	**-16.533±0.309**	-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	**0.922±0.002**	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		**0.683±0.073**	0.393±0.308	0.257±0.103	0.167±0.045	0.023±0.012	0.527±0.087
Top1 Pass (↓)	-11.700		-10.100±0.000	**-14.267±0.450**	-12.533±0.403	-9.367±0.170	-7.980±0.112	-9.000±0.082
m1 (↓)	5.100		**5.610±0.805**	9.669±0.468	5.826±1.908	10.000±0.000	10.000±0.000	7.073±0.798

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

Machine learning foundation for therapeutics



Identify meaningful tasks and datasets

Design AI/ML methods

Domain scientists

THERAPEUTICS DATA COMMONS

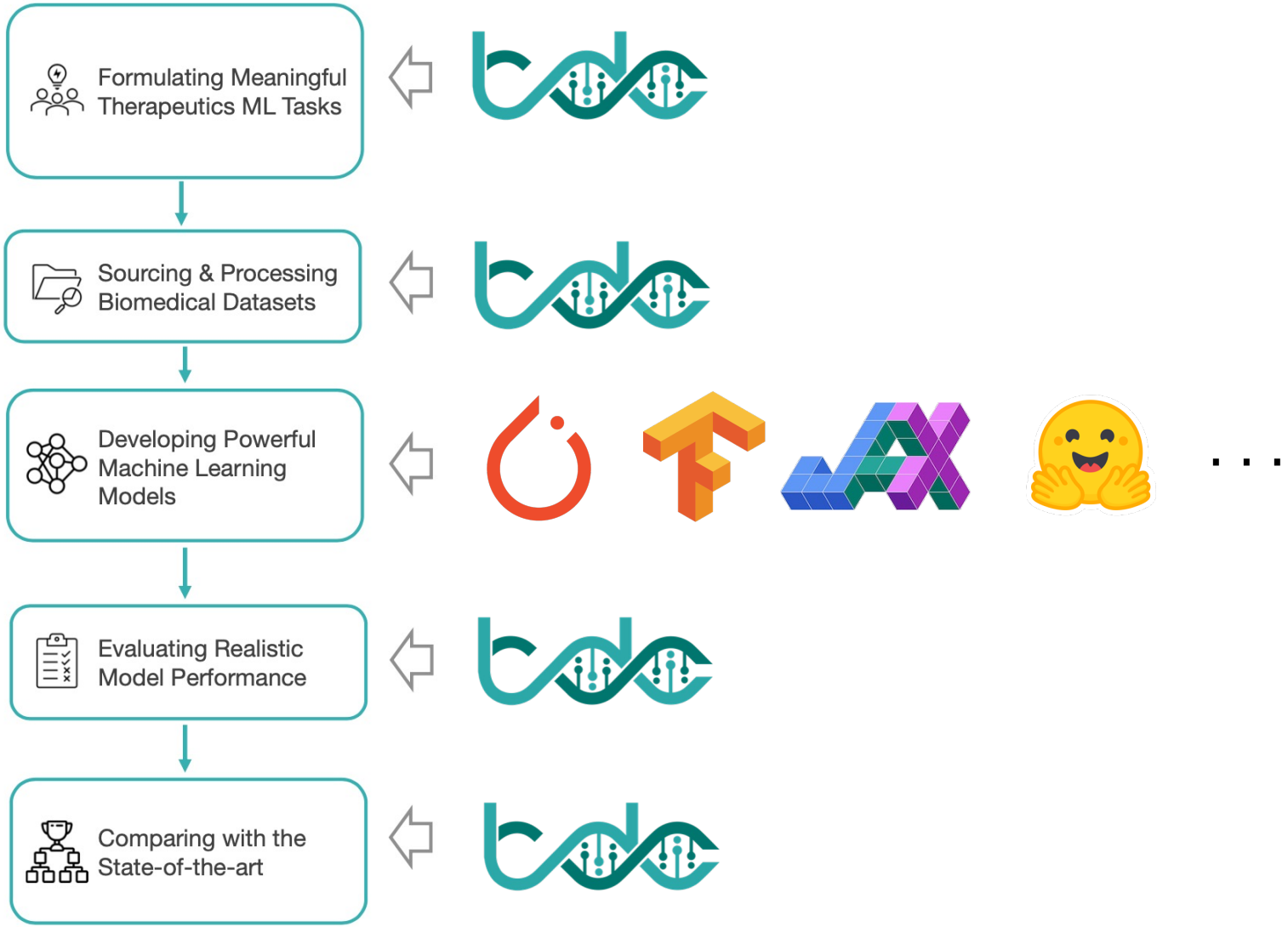
AI/ML scientists

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

Therapeutics Data Commons: Machine Learning Datasets and Tasks for Therapeutics, NeurIPS, 2021
Machine Learning Foundation for Drug Discovery and Development, *Nature Chemical Biology,* (in press), 2022

Lifecycle of therapeutics ML



Therapeutics Data Commons: Machine Learning Datasets and Tasks for Therapeutics, NeurIPS, 2021
Machine Learning Foundation for Drug Discovery and Development, *Nature Chemical Biology,* (in press), 2022

https://tdcommons.ai

Lifecycle of Therapeutics Machine Learning

- Formulating Meaningful Therapeutics ML Tasks
- Sourcing & Processing Biomedical Datasets
- Developing Powerful Machine Learning Models
- Evaluating Realistic Model Performance
- Comparing with the State-of-the-art

THERAPEUTICS DATA COMMONS

22 Therapeutics ML Tasks

- Small-Molecule: **15 Tasks**
- Macromolecule: **8 Tasks**
- Cell & Gene Therapy: **2 Tasks**
- Target Discovery: **5 Tasks**
- Activity Modeling: **13 Tasks**
- Efficacy & Safety: **6 Tasks**
- Manufacturing: **4 Tasks**

66 Therapeutics ML-ready Datasets

15,919,332 Data Points

Drug_ID	Drug	Y
CHEMBL15932	COc1cccc2[nH]ncc12	2.10
CHEMBL1527751	Oc1ncnc2scc(-c3ccsc3)c12	2.25

TDC Data Functions

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

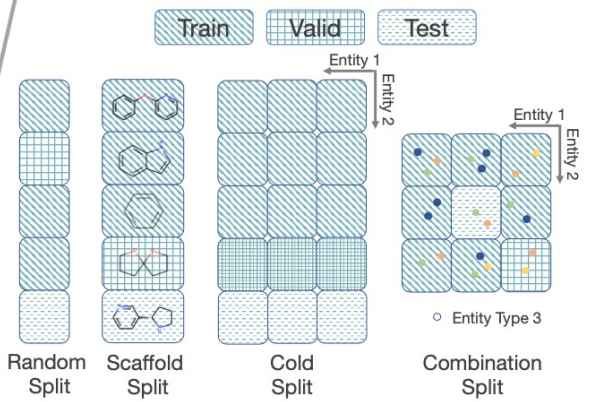
23 TDC Evaluator Functions

- Regression: **6 Metrics**
- Binary: **8 Metrics**
- Multi-class: **3 Metrics**
- Molecule: **6 Metrics**

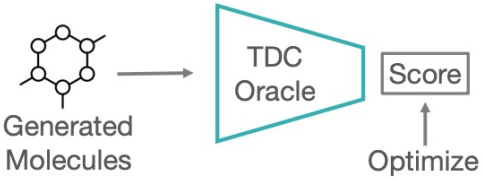
TDC Leaderboards

- **22** ADMET Group Benchmarks
- **5** Drug Combination Group Benchmarks
- **4** Docking Score Molecule Generation Benchmarks

TDC Data Splits Functions



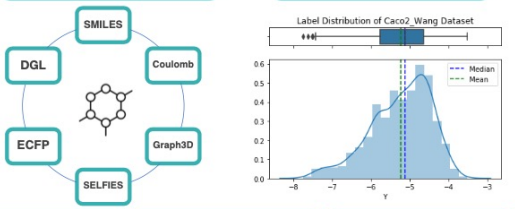
TDC Molecule Generation Oracles



Docking | ASKCOS | JNK3 | GSKB | DRD2 | M1 | QED | .....

TDC Data Processing Helpers

Format Exchange | Visualization

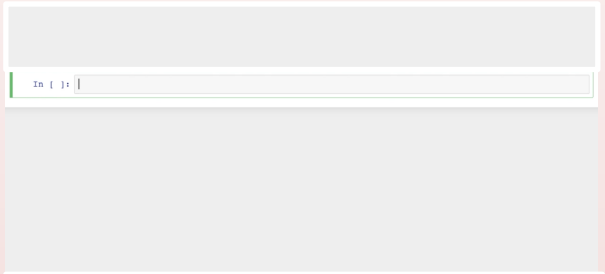


Binarization | Data Balancing | Unit Conversion | Database Query | Molecule Filters | .....

@ProjectTDC

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…

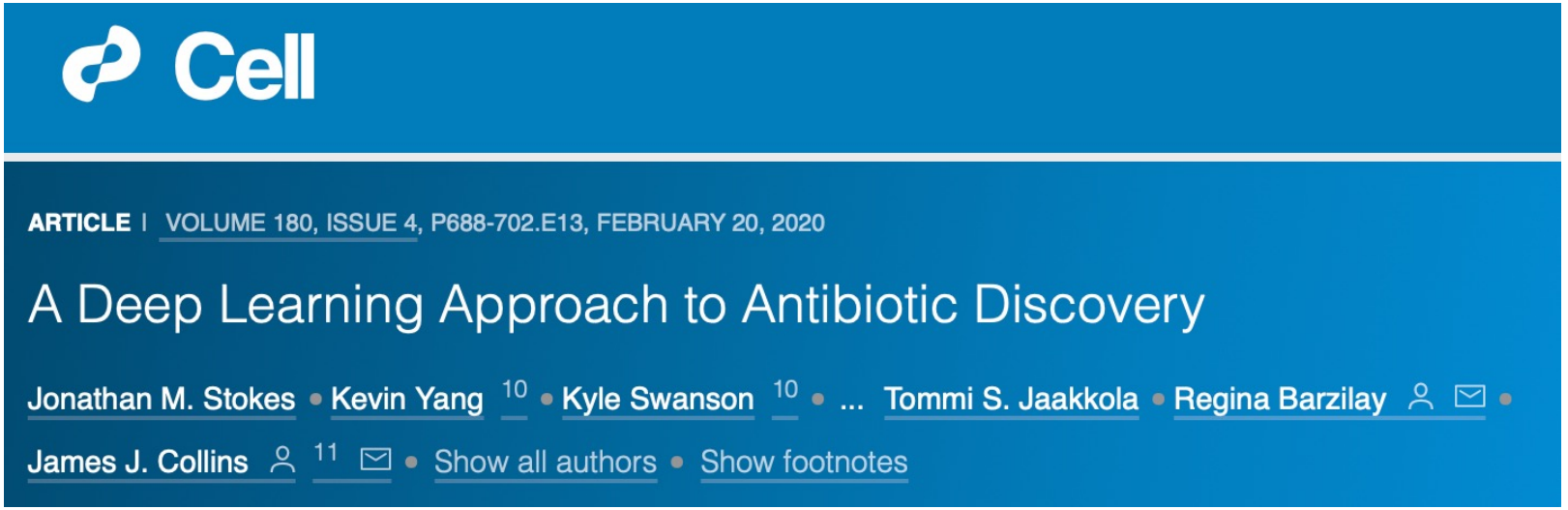
THERAPEUTICS

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

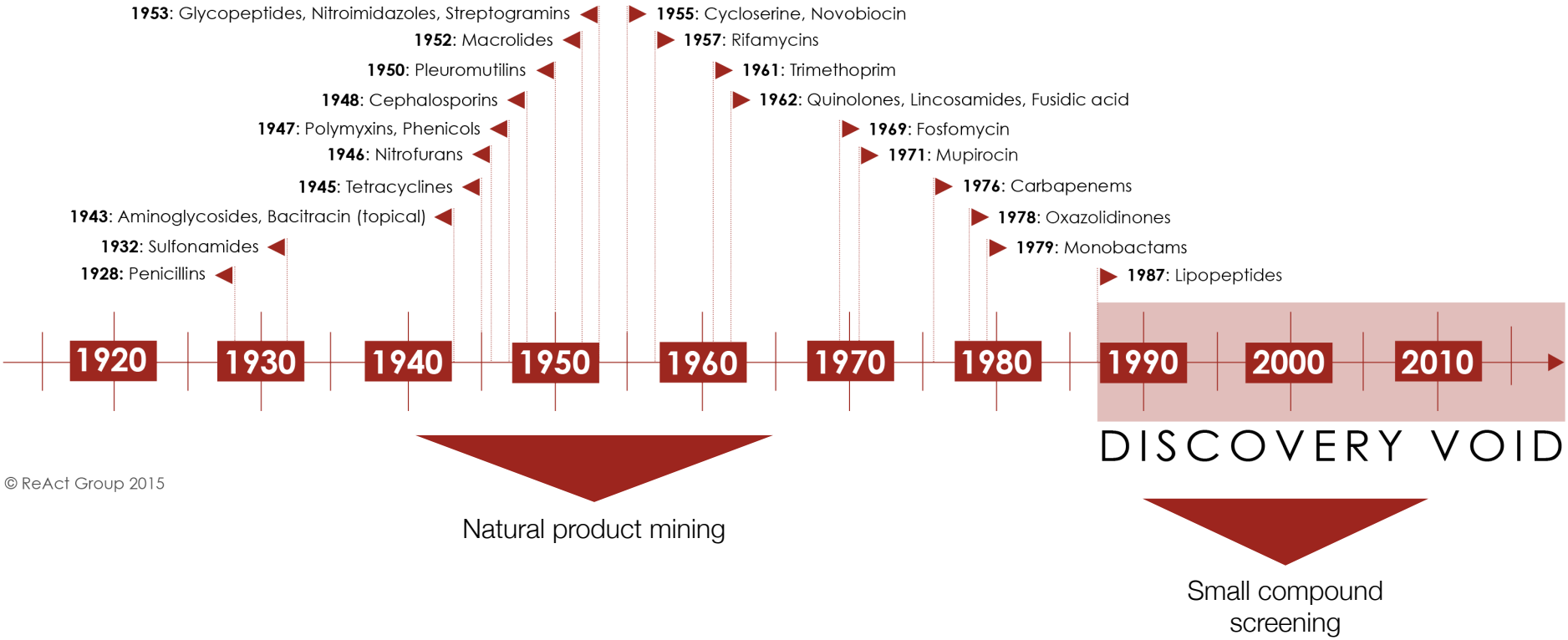
Applications of graph representation learning on…

THERAPEUTICS

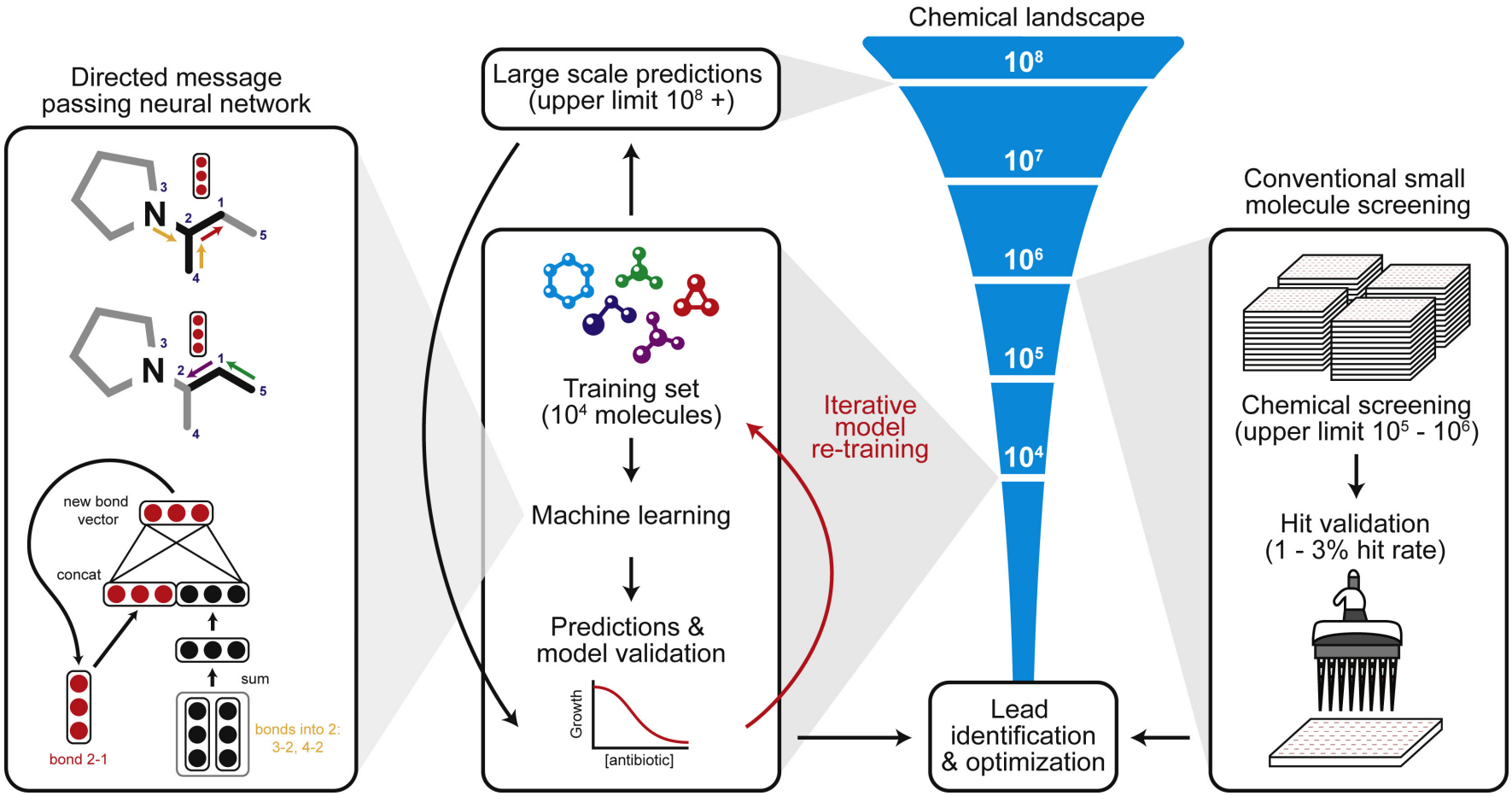
1. Molecular property prediction, drug-target interaction prediction, molecular generation
2. Drug discovery
3. Drug repurposing



Antibiotic discovery timeline



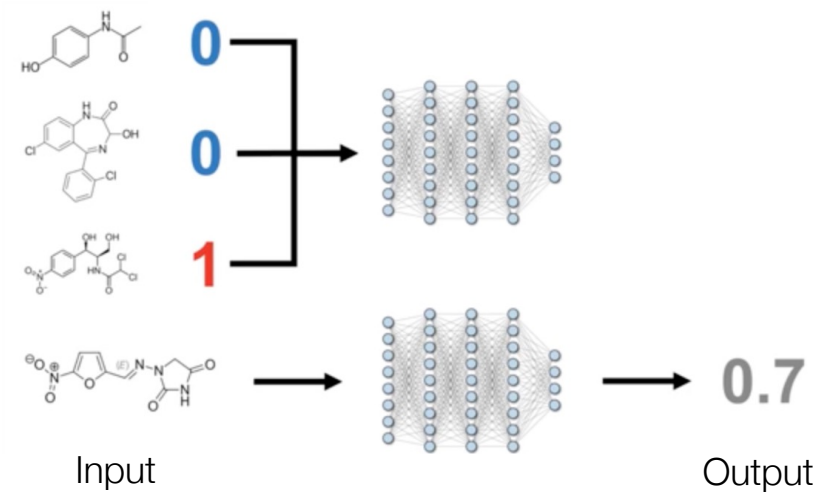
GNN to learn molecular structure



A Deep Learning Approach to Antibiotic Discovery, *Cell*, 2020.

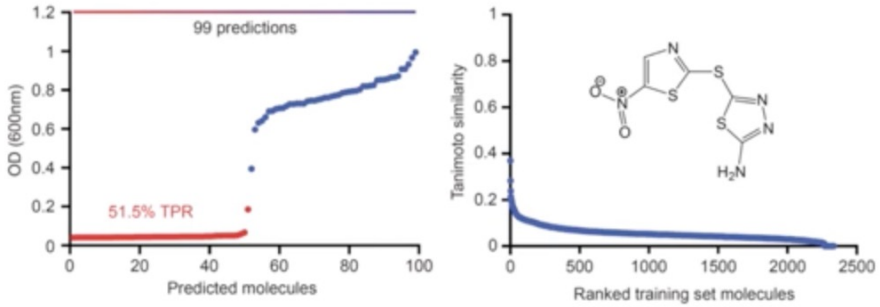
Experimental setup

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

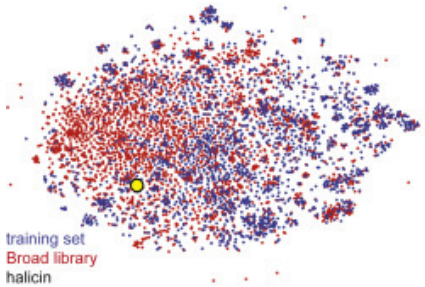
Task: Test top 99 predictions & prioritize based on similarity to known antibiotics or predicted toxicity

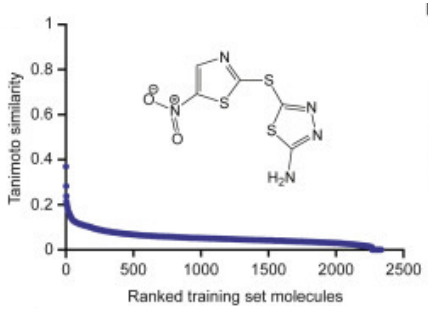
A Deep Learning Approach to Antibiotic Discovery, *Cell*, 2020.

Results

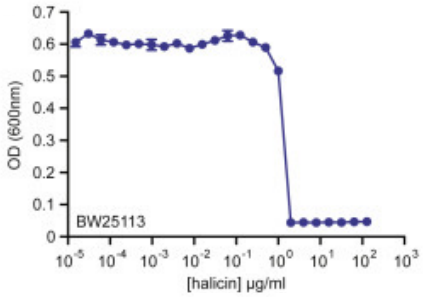
Halicin was developed to be an anti-diabetic drug, but the development was discontinued due to poor results in testing.

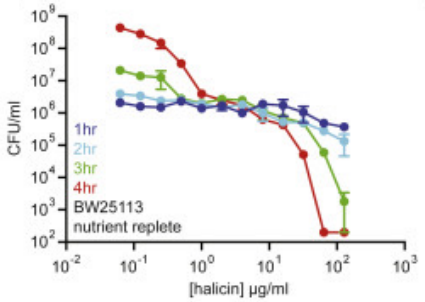
Halicin predicted to be antibacterial



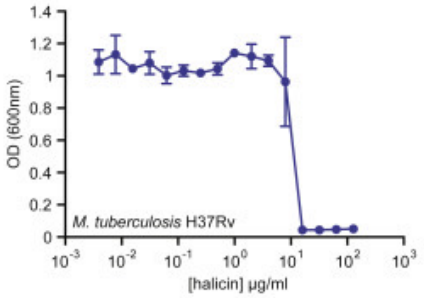


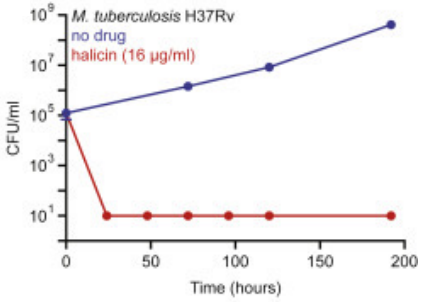
Halicin against *E. coli*





Halicin against *M. tuberculosis*

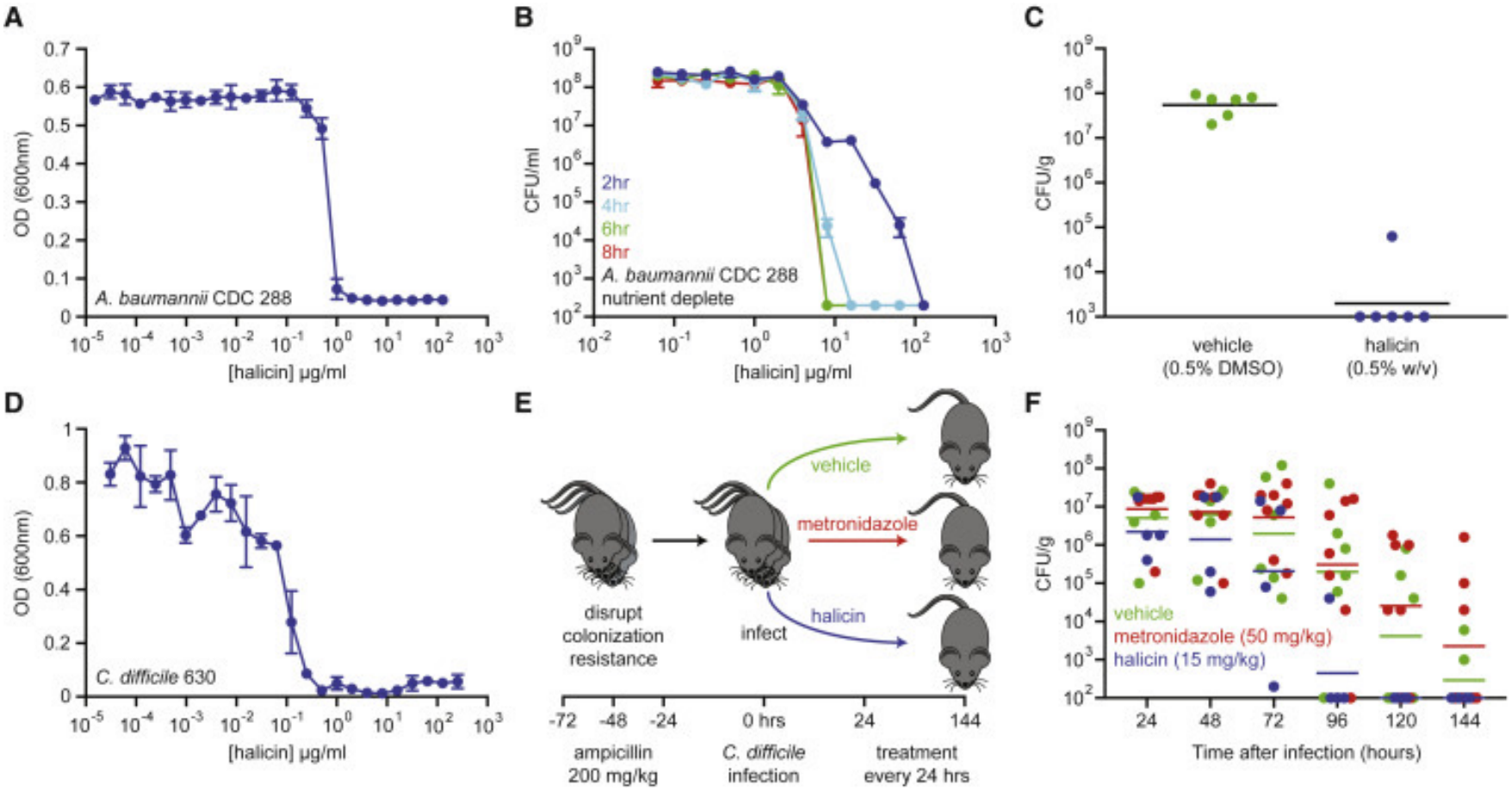




A Deep Learning Approach to Antibiotic Discovery, *Cell*, 2020.

Results

Halicin's efficacy in murine models of infection



A Deep Learning Approach to Antibiotic Discovery, *Cell*, 2020.

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
  - Chemprop resources:
    - Paper: doi.org/10.1021/acs.jcim.9b00237
    - GitHub: 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[a,b,c,1], Ítalo do Valle[a,b,1], Marinka Zitnik[d,e,1], Asher Ameli[b,f,1], Xiao Gan[a,b,c,1], Onur Varol[a,b,g], Susan Dina Ghiassian[f], J. J. Patten[h], Robert A. Davey[h], Joseph Loscalzo[i], and Albert-László Barabási[a,b,j,2]

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

Identification of disease treatment mechanisms through the multiscale interactome

Camilo Ruiz[1,2], Marinka Zitnik[3] & Jure Leskovec[1,4]✉

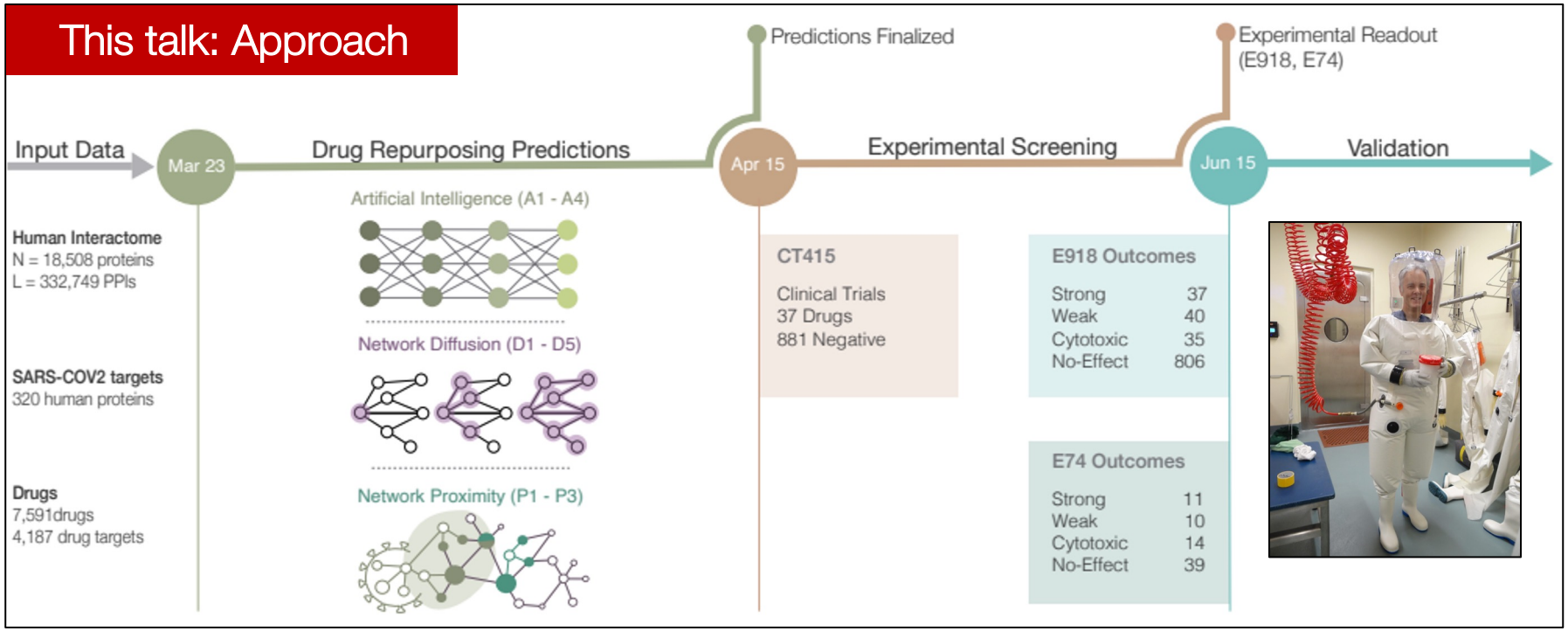
GNNExplainer: Generating Explanations for Graph Neural Networks

Rex Ying[†] Dylan Bourgeois[†,‡] Jiaxuan You[†] Marinka Zitnik[†] Jure Leskovec[†]

[†]Department of Computer Science, Stanford University
[‡]Robust.AI
{rexying, dtsbourg, jiaxuan, marinka, jure}@cs.stanford.edu

Rapid therapeutic innovation

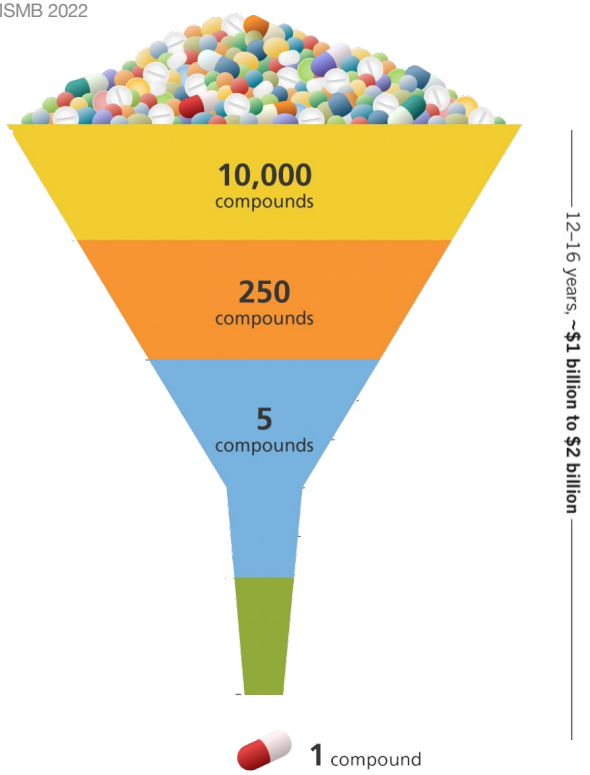
- Traditional, iterative development, experimental & clinical testing, and approval of **new drugs** sometimes not feasible
  - Certain therapeutic areas, public health emergencies
- **Challenge:** How to compress years of work into months or even weeks through AI, automation, and new data resources?



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

New tricks for old drugs

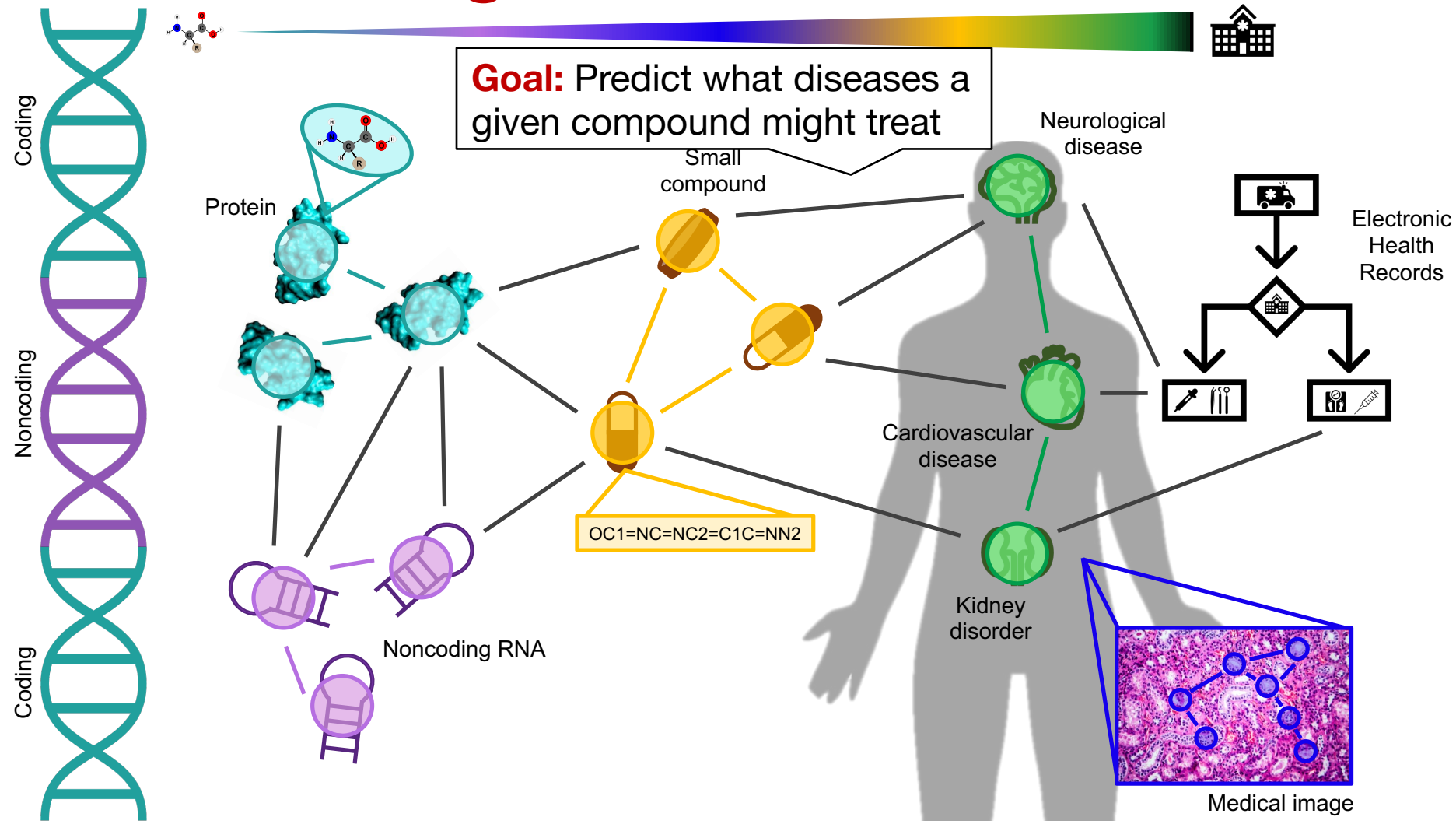
Faced with skyrocketing costs for developing new drugs, researchers are looking at ways to repurpose older ones — and even some that failed in initial trials.



A SHORTER TIMESCALE

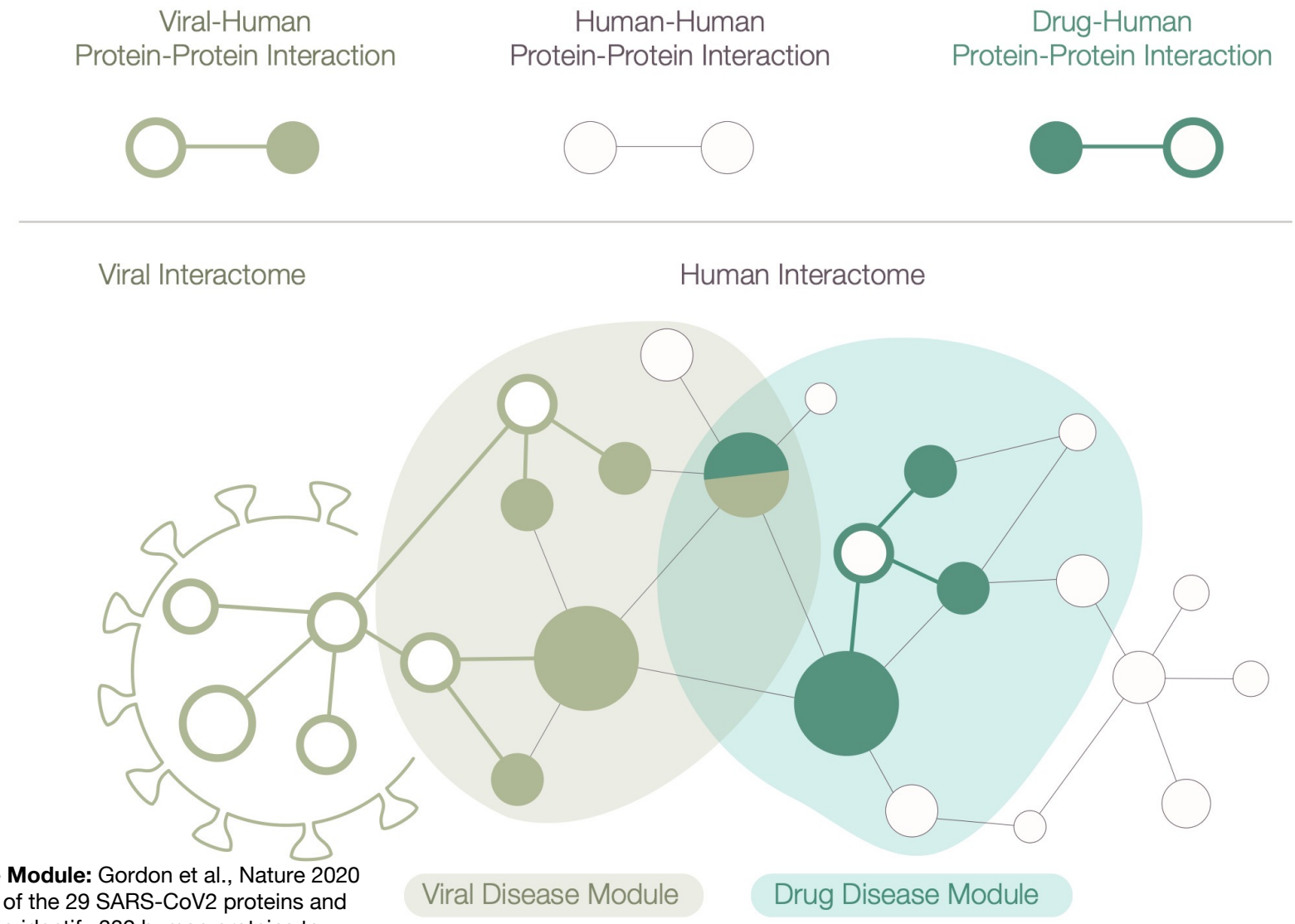
Because most repositioned drugs have already passed the early phases of development and clinical testing, they can potentially win approval in less than half the time and at one-quarter of the cost.

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 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

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

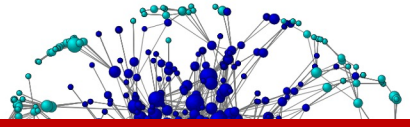
Network Medicine Framework for Identifying Drug Repurposing Opportunities for Covid-19, *PNAS*, 2021

Why is this task challenging?

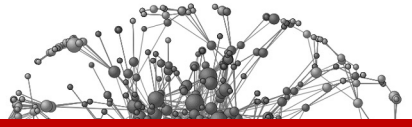
Challenge: Generalizing to new phenomena is hard:

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

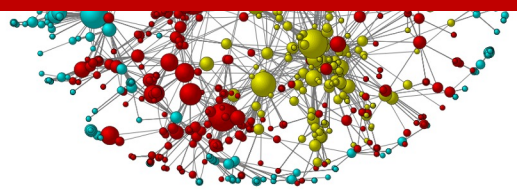
What prevailing methods assume

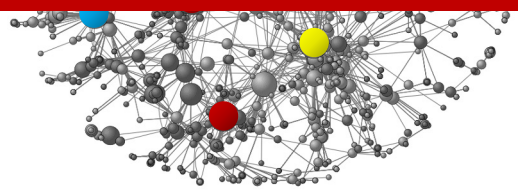


What happens in real world



Today: Few-shot learning for graphs





Background: Few-shot learning

Meta-Training

At test time, we need to build a “duck vs. dolphin vs. chicken” classifier. **However, we have only 2 labeled examples for each class!** Few-shot learning makes this possible.

Training task 1

Support set

K=2

Query set

Training

Support set

Query set

Support set

Query set

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

Meta-Training

Meta-Testing

Training task 1

Support set

K=2

Drug 1 Drug 2 Drug 3

Drug 4 Drug 5 Drug 6

N=3

Query set

Drug a Drug b Drug c

Disease 1

Training task 2 . . .

Support set

Drug 7 Drug 8 Drug 9

Drug 10 Dru

Query set

Drug d Drug e Drug f

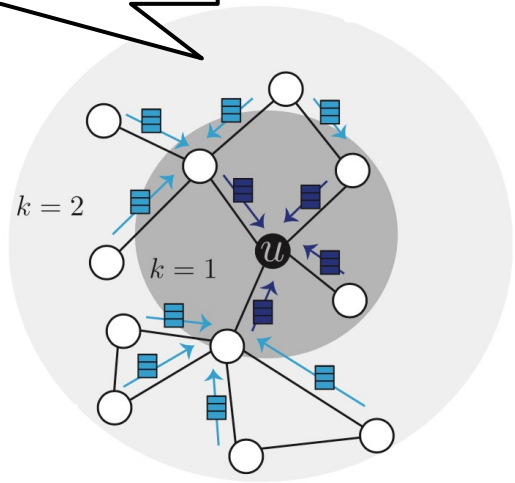
Disease 2

Test task 1 . . .

Support set

No examples of successful treatments

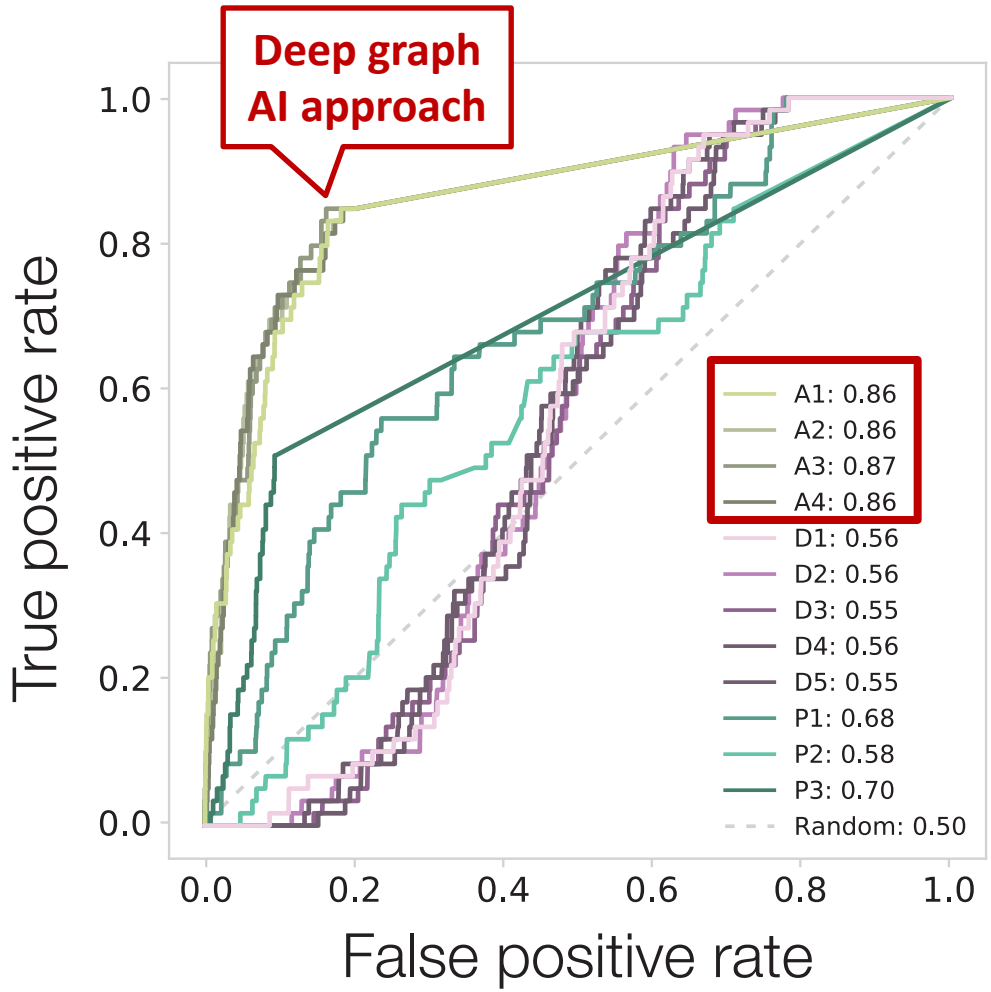
Deep graph AI approach



Drug i

COVID-19

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

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

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

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**

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

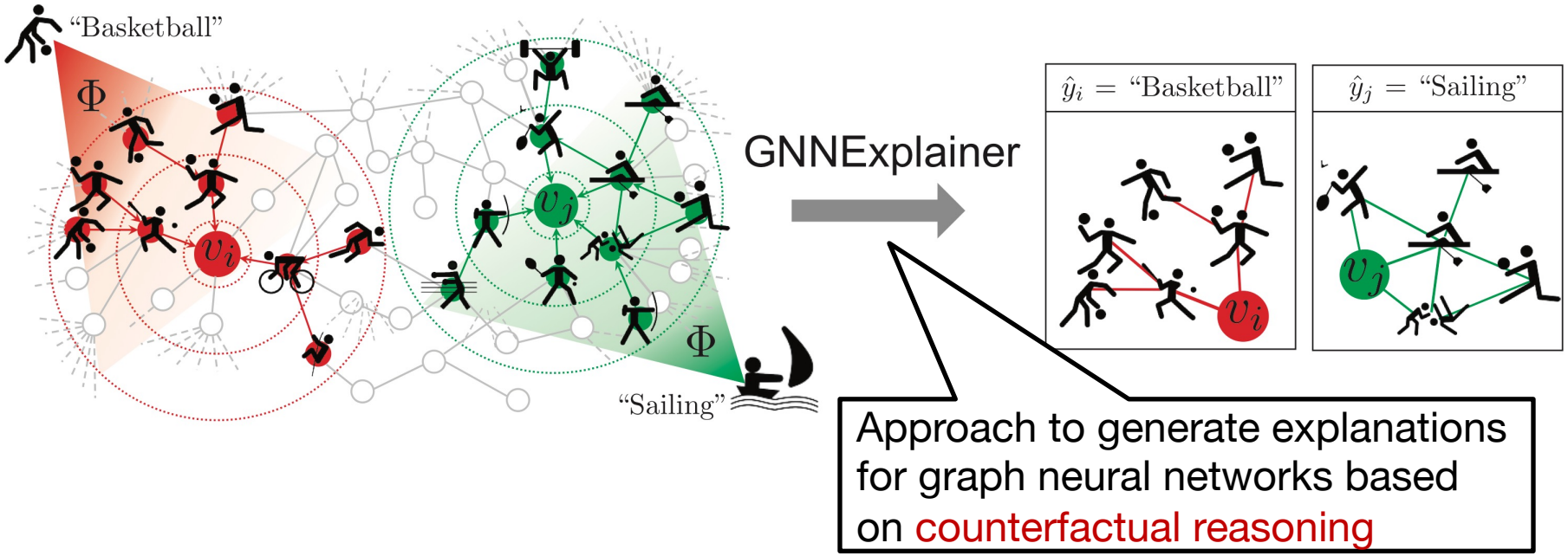
Explaining machine predictions

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

Explaning GNN's predictions

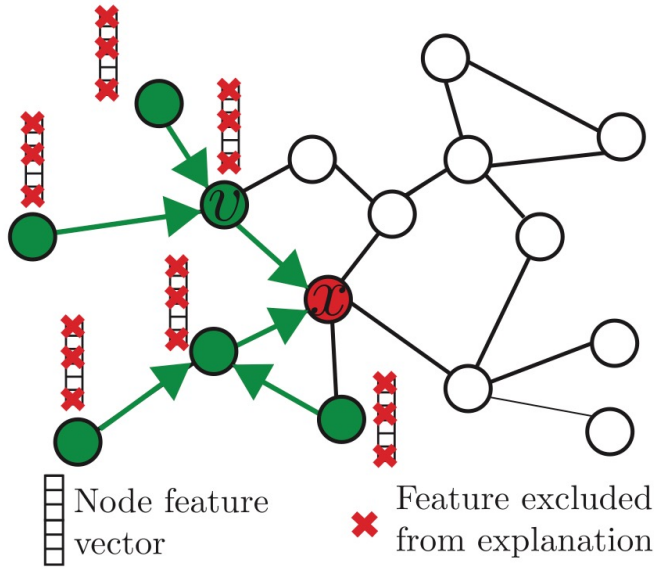


Approach to generate explanations for graph neural networks based on counterfactual reasoning

GNNExplainer: Generating Explanations for Graph Neural Networks, *NeurIPS* 2019

GNNExplainer: key idea

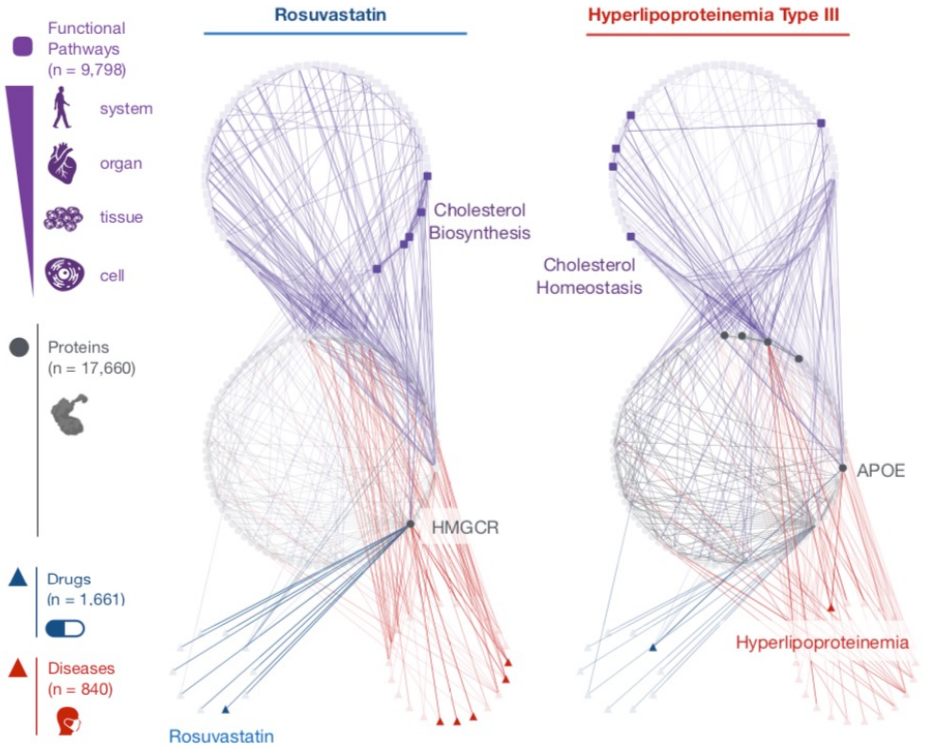
- Input: Given prediction $f(x)$ for node/link x
- Output: Explanation, a small subgraph M_x together with a small subset of node features:
  - M_x is most influential for prediction $f(x)$
- Approach: Optimize mask M_x in a post-hoc manner
  - Intuition: If removing v from the graph strongly decreases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

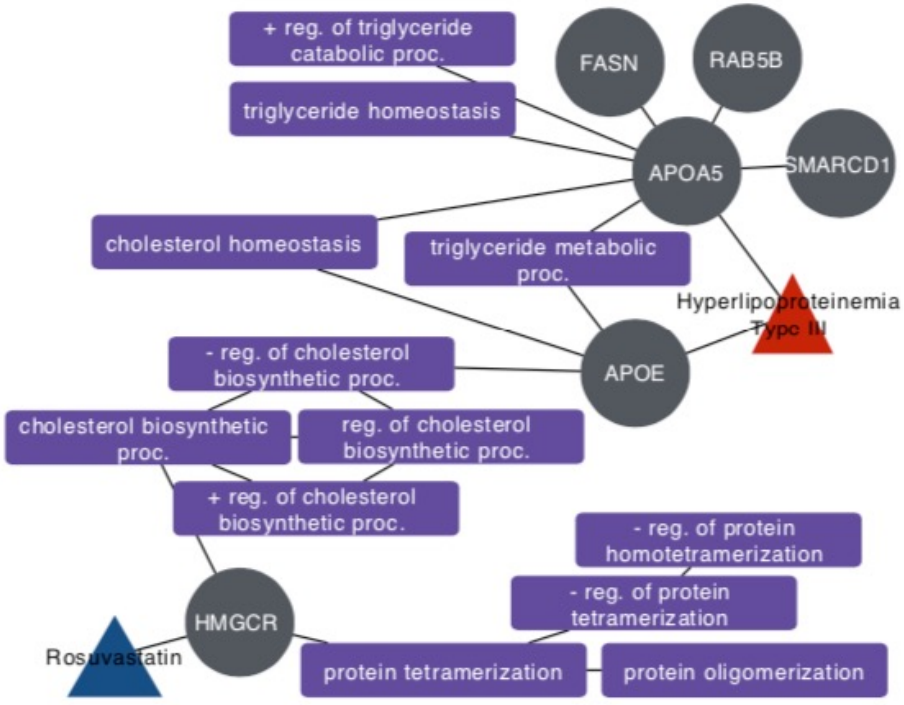


GNN Explainer: Generating Explanations for Graph Neural Networks, *NeurIPS* 2019

Example of explanations

”Will rosuvastatin treat hyperlipidemia? What is the disease treatment mechanism?”

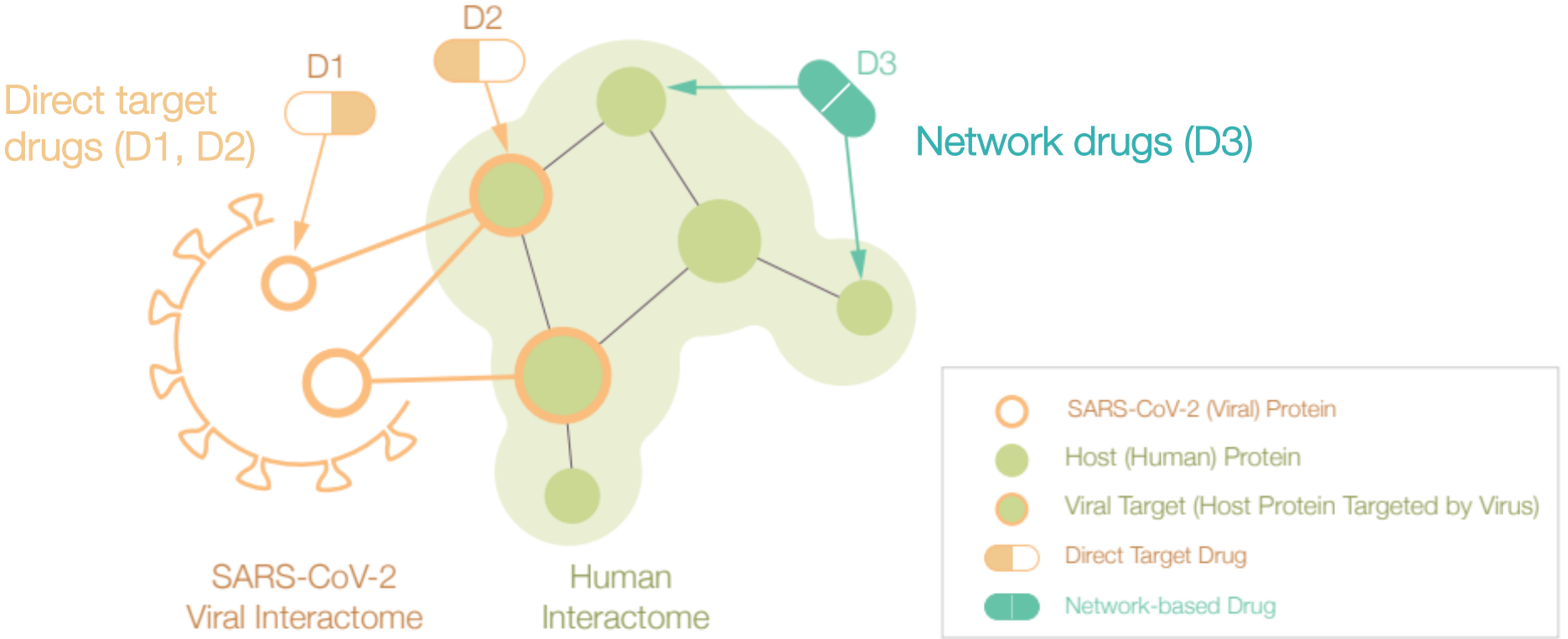




GNNExplainer: Generating Explanations for Graph Neural Networks, NeurIPS 2019
Discovery of Disease Treatment Mechanisms through the Multiscale Interactome, Nature Communications, 2021

Predictions ➡  ➡ Network drugs

- 76/77 drugs that successfully reduced viral infection <u>do not bind</u> 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
  - 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

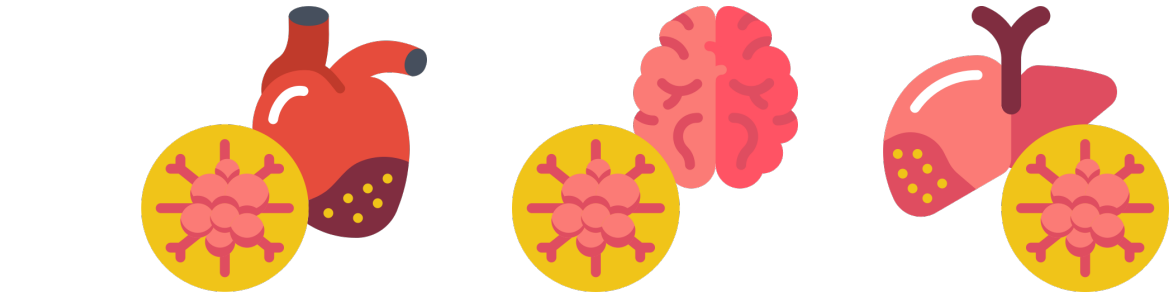
1. Histopathology images of tissue biopsies
2. Patient electronic health records

IEEE TRANSACTIONS ON MEDICAL IMAGING, VOL. 41, NO. 4, APRIL 2022 757

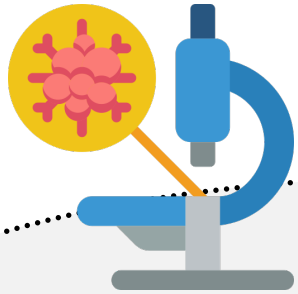
Pathomic Fusion: An Integrated Framework for Fusing Histopathology and Genomic Features for Cancer Diagnosis and Prognosis

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

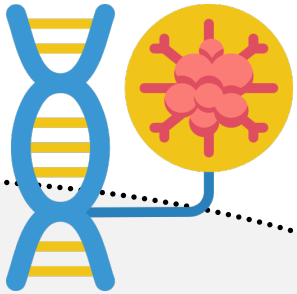
Motivation



Cancer diagnosis



Histopathology slides
(tumor microenvironment)

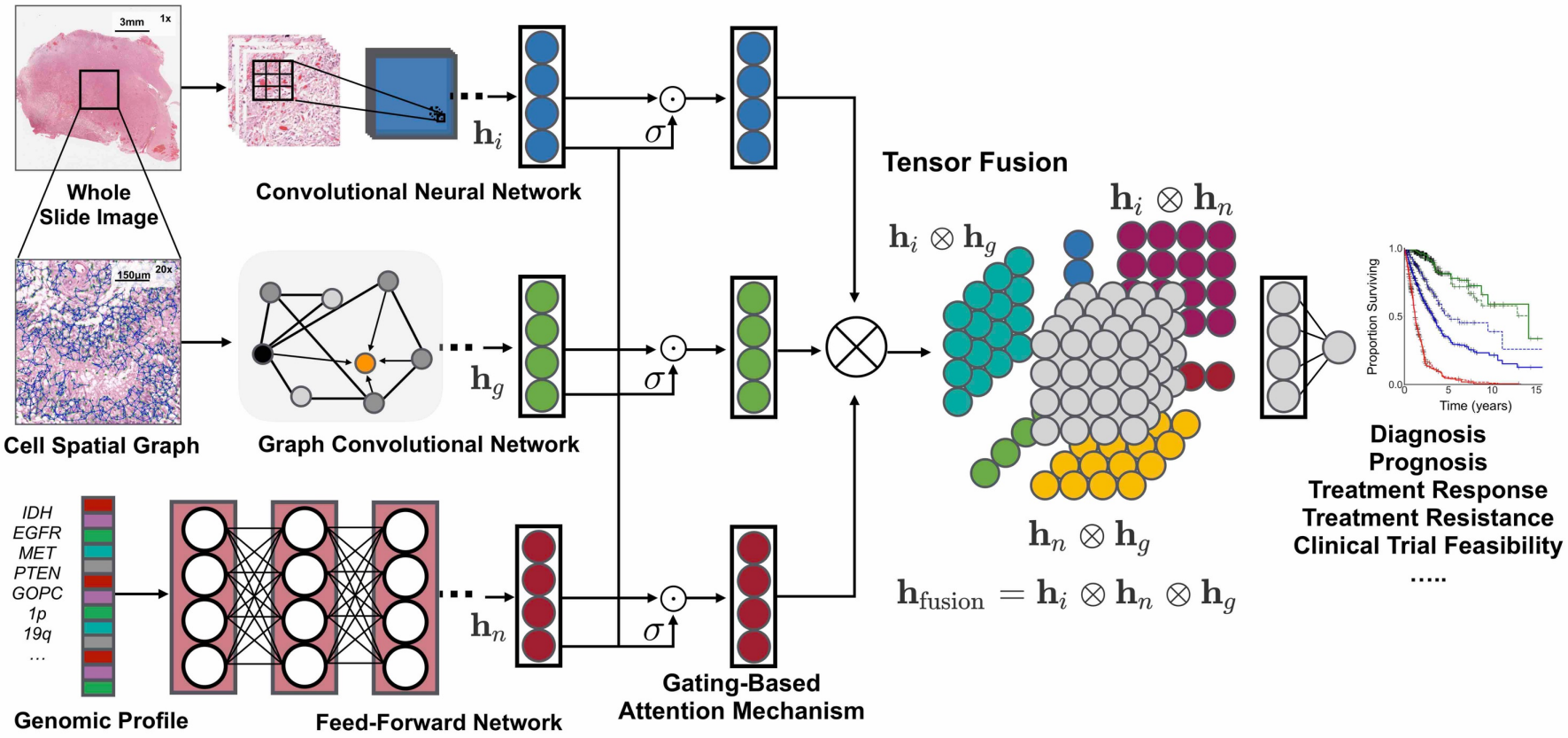


Genomic sequencing
(genotypic expression)

1. Current paradigm requires subjective & visual assessment by pathologists

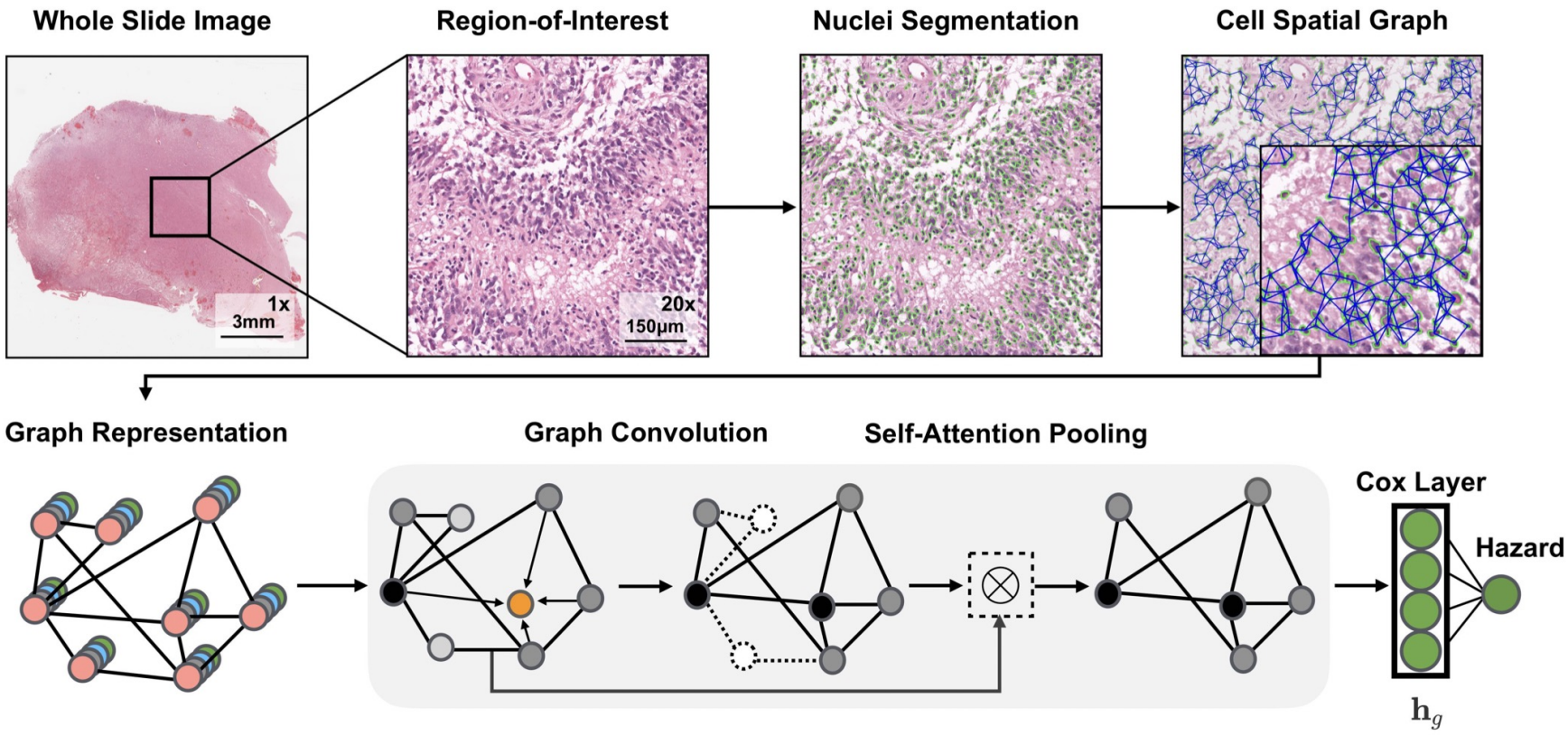
2. Existing methods only consider single modalities at a time

Overview of Pathomic Fusion



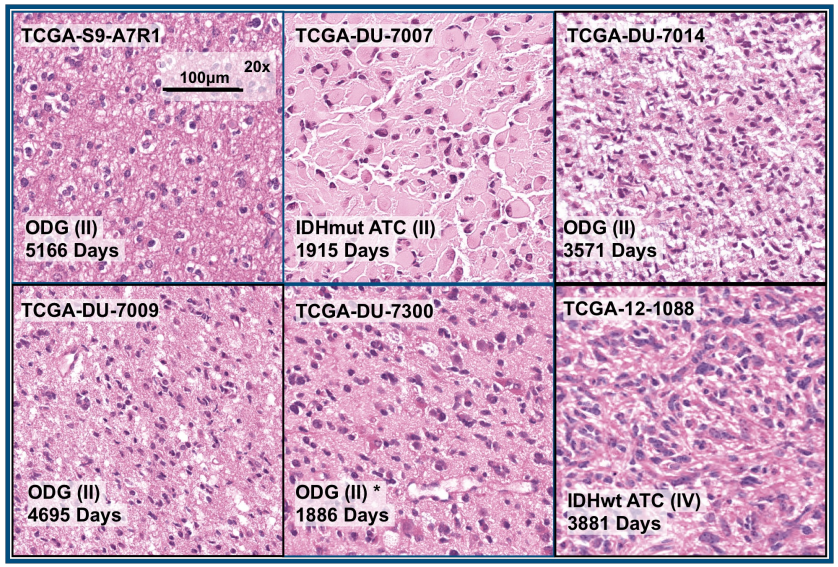
Pathomic Fusion: An Integrated Framework for Fusing Histopathology and Genomic Features for Cancer Diagnosis and Prognosis, *IEEE Transactions on Medical Imaging*, 2020.

GCN for whole slide images



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

Experimental setup

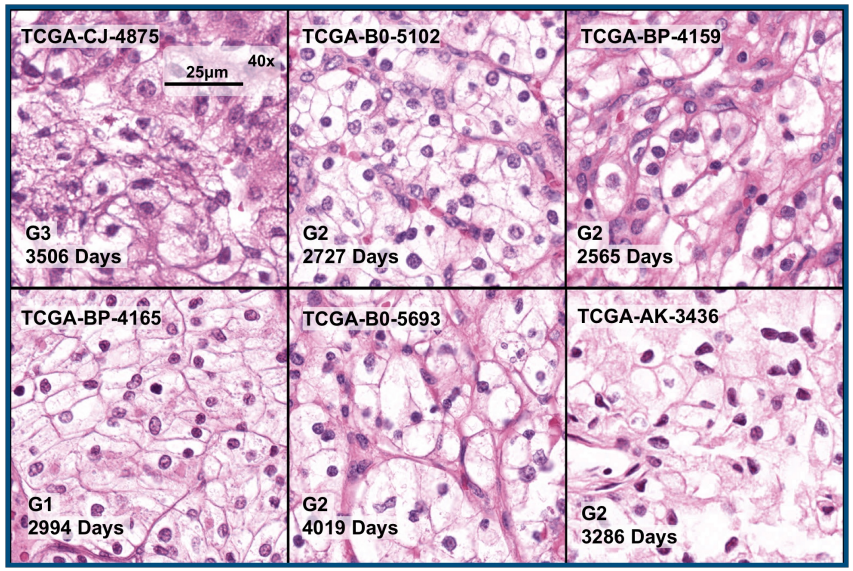


Data

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

Experiments

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



Data

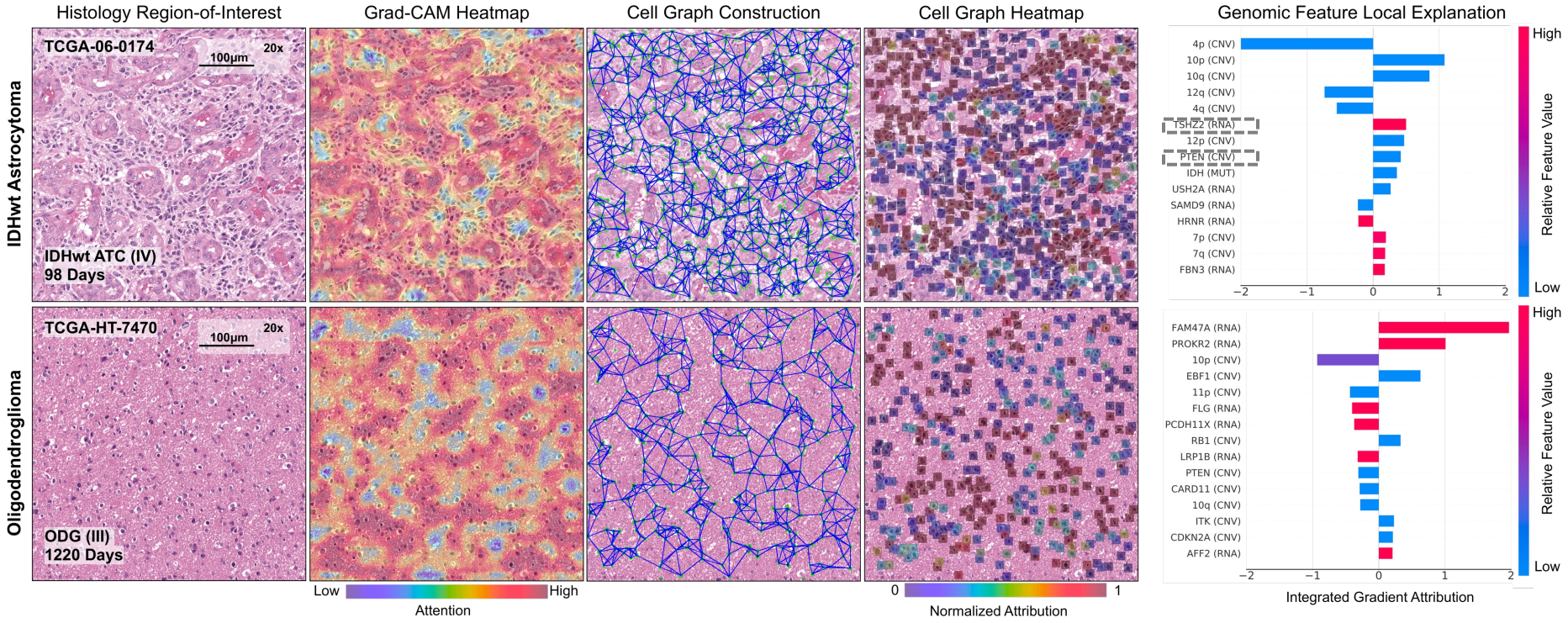
- 417 paired samples
- 40 x 512 x 512 Histology regions of interest (ROIs) (3 per patient)
- 117 CNV, 240 RNA-Seq

Experiments

- Compare to Fuhrman Grade
- 15-Fold CV

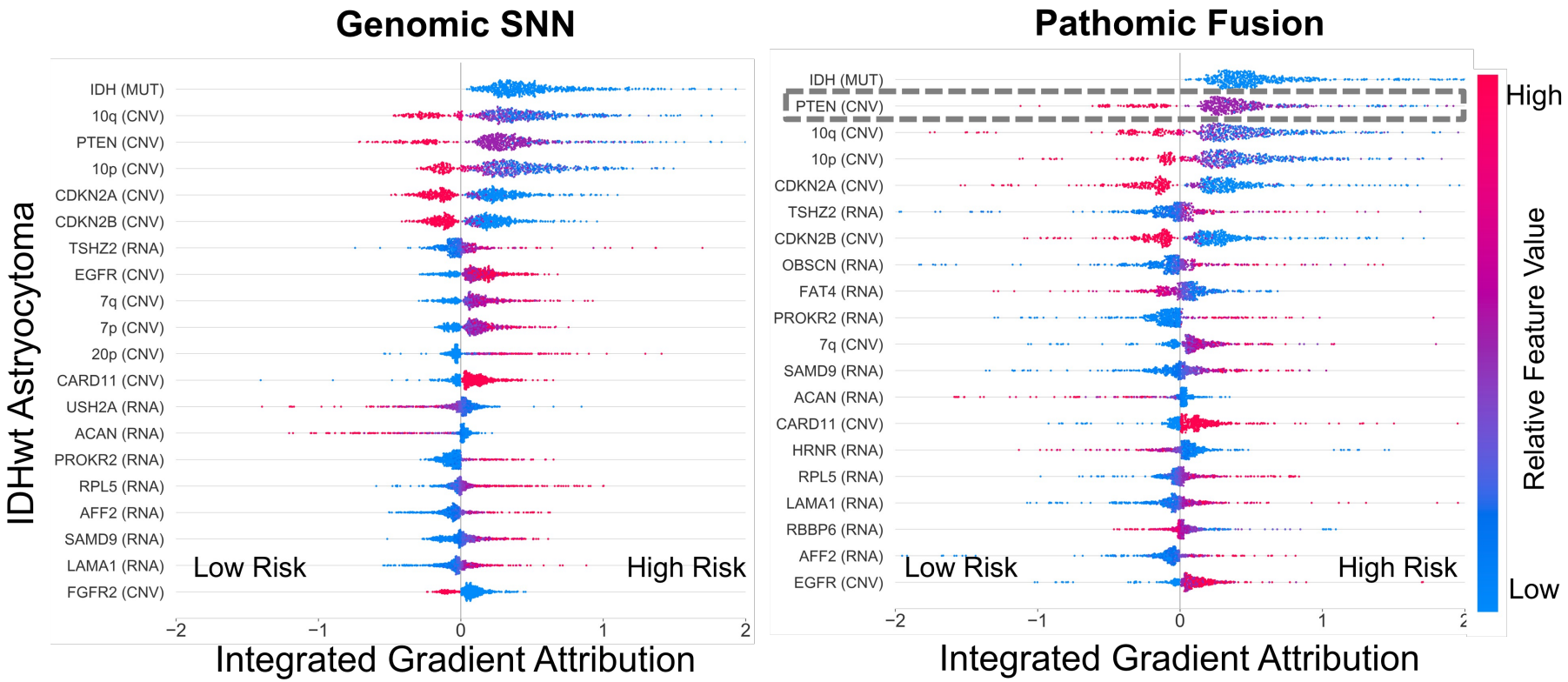
Pathomic Fusion: An Integrated Framework for Fusing Histopathology and Genomic Features for Cancer Diagnosis and Prognosis, *IEEE Transactions on Medical Imaging*, 2020.

Results



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

Results



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

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
  - GitHub: github.com/mahmoodlab/PathomicFusion
  - Talk: youtube.com/watch?v=TrjGEUVX5YE
  - Synthetic dataset: 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

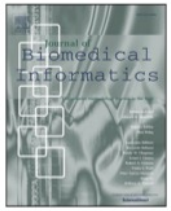
1. Histopathology images of tissue biopsies
2. Patient electronic health records

Journal of Biomedical Informatics 127 (2022) 104000

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

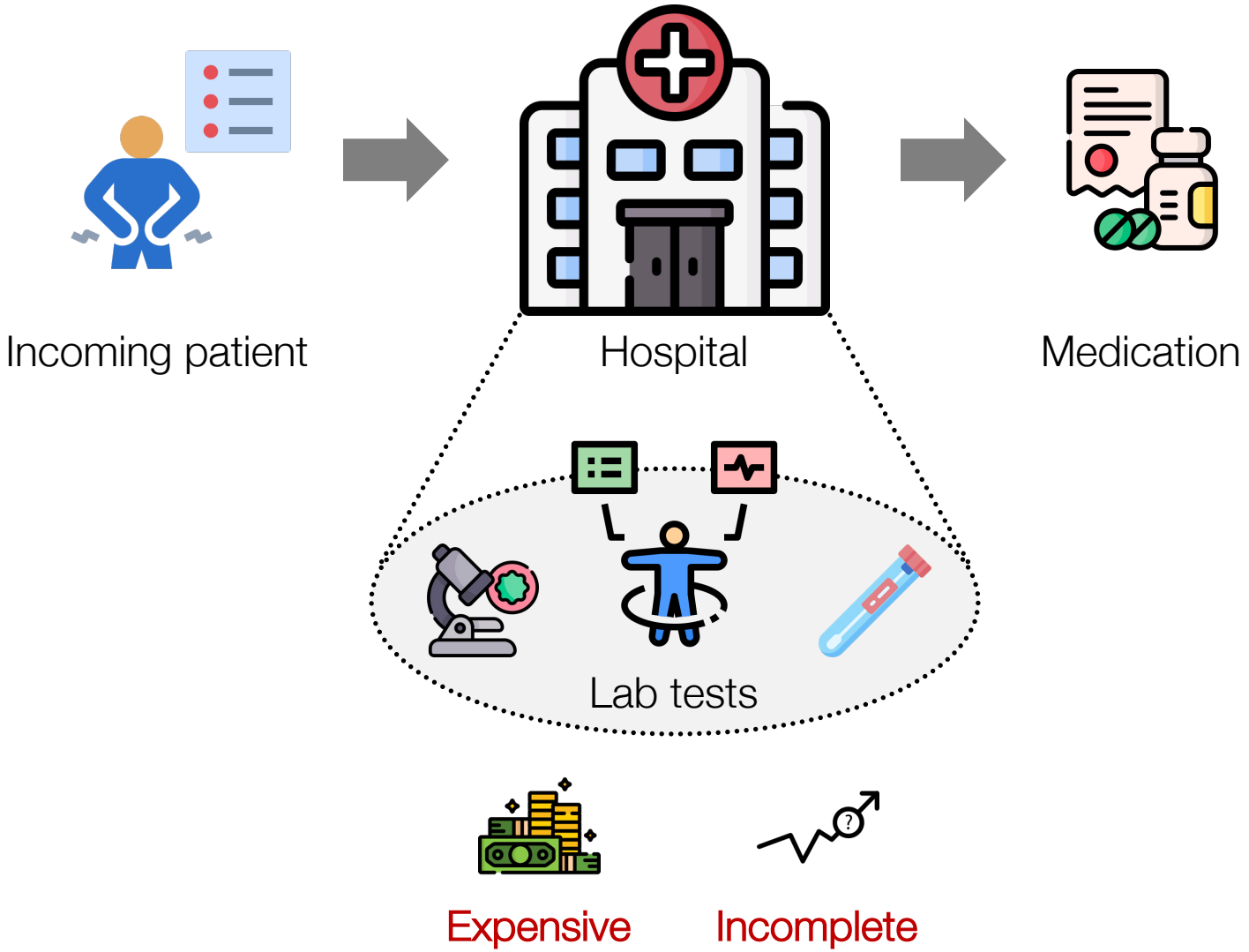


MedGCN: Medication recommendation and lab test imputation via graph convolutional networks

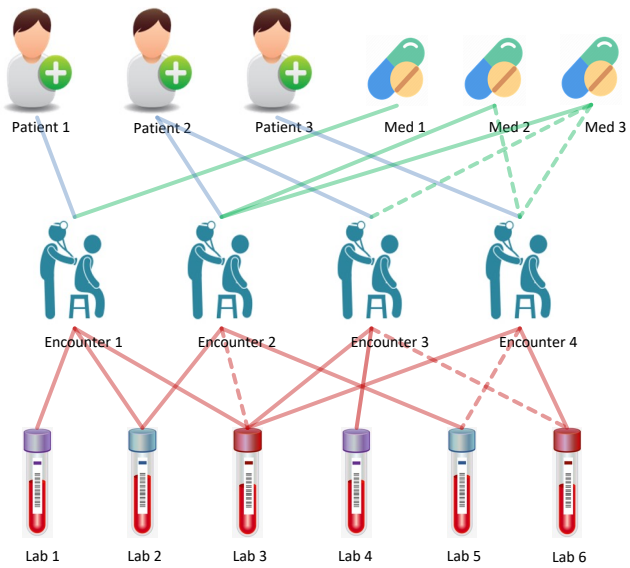
Chengsheng Mao, Liang Yao, Yuan Luo *

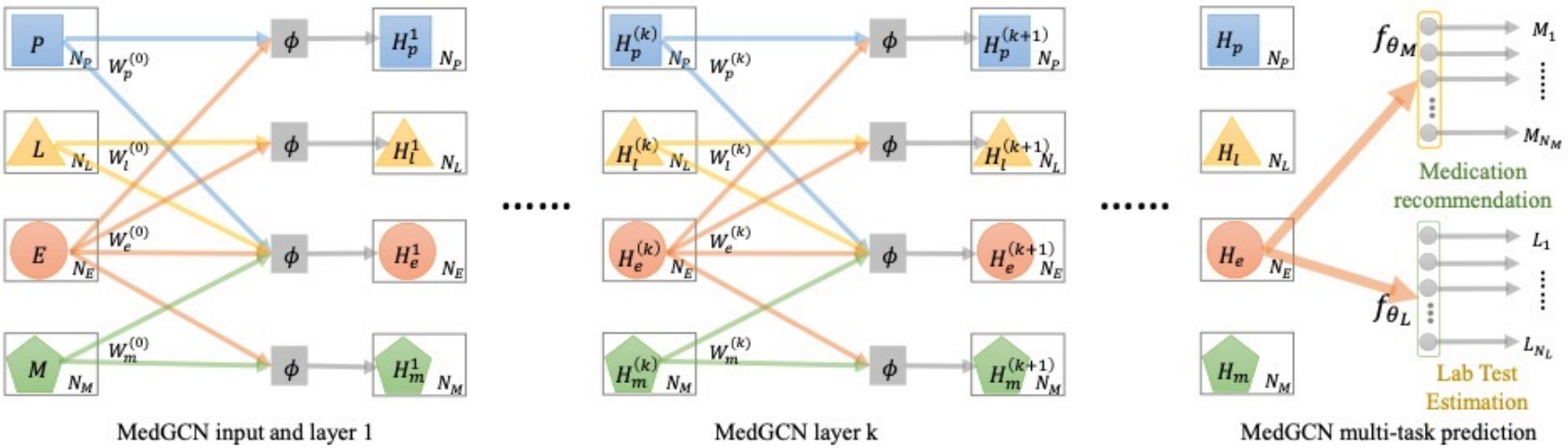
Department of Preventive Medicine, Feinberg School of Medicine, Northwestern University, Chicago, IL, USA

Motivation


Incoming patient
Hospital
Medication
Lab tests
Expensive
Incomplete

Overview of MedGCN





MedGCN: Medication recommendation and lab test imputation via graph convolutional networks, *JBI*, 2022.

MedGCN Message Propagation

	P1	P2	P3
E1	1	0	0
E2	0	1	0
E3	0	1	0
E4	0	0	1

$A_{E\times P}$

	M1	M2	M3
E1	1	0	0
E2	0	1	1
E3	0	0	0
E4	0	0	0

$A_{E\times M}$

	L1	L2	L3	L4	L5	L6
E1	.3	0	.1	0	0	0
E2	0	.1	0	0	.5	0
E3	0	0	.1	.8	0	0
E4	0	0	.0	0	0	.6

$A_{E\times L}$

	L1	L2	L3	L4	L5	L6
E1	1	1	1	0	0	0
E2	0	1	0	0	1	0
E3	0	0	1	1	0	0
E4	0	0	1	0	0	1

$M_{E\times L}$

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

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

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

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

MedGCN: Medication recommendation and lab test imputation via graph convolutional networks, *JBI*, 2022.

Experimental setup

NMEDW

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

Matrix	Size	Edges	Sparsity	Values
$A_{E\times P}$	1260 × 865	1260	99.88%	binary: 0, 1
$A_{E\times L}$	1260 × 197	43806	82.35%	continuous: 0–1
$A_{E\times M}$	1260 × 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 × 15153	18190	99.99%	binary: 0, 1
$A_{E\times L}$	18190 × 219	1029964	68.96%	continuous: 0–1
$A_{E\times M}$	18190 × 117	23395	98.68%	binary: 0, 1

MedGCN: Medication recommendation and lab test imputation via graph convolutional networks, *JBI*, 2022.

Results

NMEDW

Medication Recommendation

Methods	LRAP	MAP@2
MedGCN (ours)	**.7588±.0028**	**.7558±.0035**
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*

Lab Test Imputation

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

MIMIC-III

Medication Recommendation

Methods	LRAP	MAP@2
MedGCN (ours)	**.8349±.0008**	.8069±.0022
MedGCN-ind (ours)	.8345±.0007	**.8070±.0029**
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)	**.0140±.0002**
MedGCN-ind(ours)	.0143±.0002*
MedGCN-Lab(ours)	.0143±.0001*
MICE	.0146±.0001*
MGCNN	.0413±.0048*
GCMC	.0296±.0004*
GCMC+FEAT	.0290±.0001*

MedGCN: Medication recommendation and lab test imputation via graph convolutional networks, *JBI*, 2022.

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
  - GitHub: 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

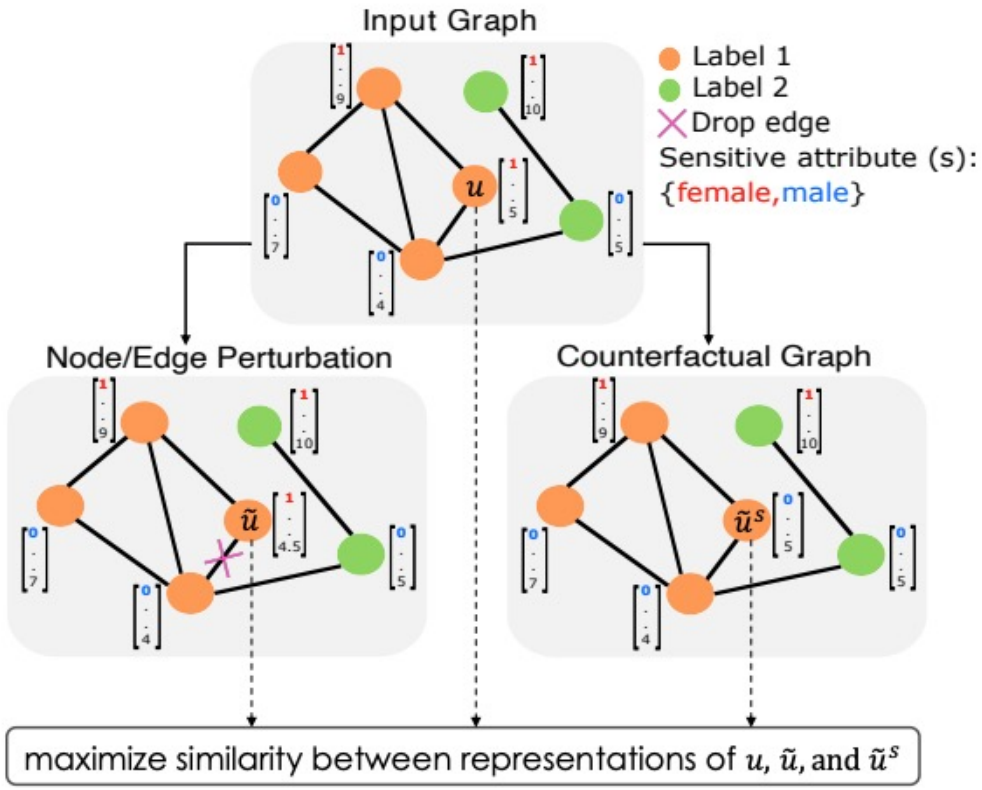
Towards fair & stable GNNs (1/3)

- 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 a Unified Framework for Fair and Stable Graph Representation Learning, UAI 2021

Code and datasets: https://zitniklab.hms.harvard.edu/projects/NIFTY

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

Code and datasets: https://zitniklab.hms.harvard.edu/projects/NIFTY

Towards fair & stable GNNs (3/3)

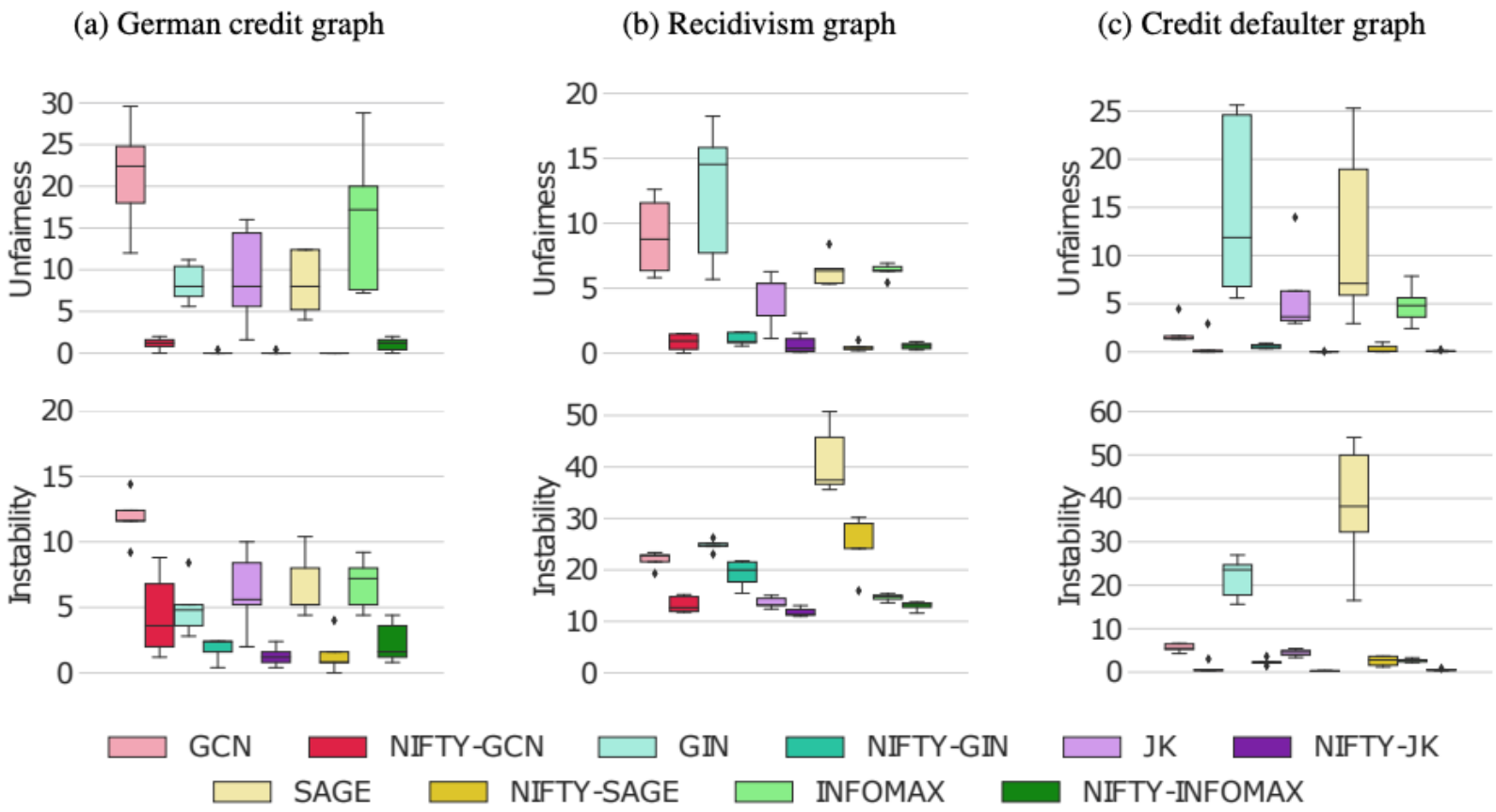


Figure 2: Unfairness (top) and instability (bottom) error rates for five GNNs and their NIFTY counterparts. NIFTY-enhanced GNNs give fairer and more stable predictions than their unmodified counterparts across all three datasets and five GNNs.

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)
  - Li et al., Graph Representation Learning for Biomedicine (arxiv.org/abs/2104.04883)

- Keynotes & seminars
  - Michael Bronstein, "Geometric Deep Learning: The Erlangen Programme of ML" (ICLR 2021 keynote) (youtube.com/watch?v=w6Pw4MOzMuo)
  - Broad Institute Models, Inference & Algorithms: Actionable machine learning for drug discovery; Primer on graph representation learning (youtube.com/watch?v=9YpTYdru0Rg)
  - Stanford University (CS224W Lecture): Graph neural networks in computational biology (youtube.com/watch?v=_hy9AgZXhbQ)
  - AI Cures Drug Discovery Conference (youtube.com/watch?v=wNXSkISMTw8)

- Conferences & summer schools
  - London Geometry and Machine Learning Summer School (logml.ai)
  - Learning on Graphs Conference (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)
  - Zitnik Lab Graph ML Tutorials (github.com/mims-harvard/graphml-tutorials)
  - Stanford University's CS224 (web.stanford.edu/class/cs224w)

- **Datasets**
  - Precision Medicine Oriented Knowledge Graph (PrimeKG) (zitniklab.hms.harvard.edu/projects/PrimeKG)
  - Therapeutic Data Commons (TDC) (tdcommons.ai)
  - BioSNAP (snap.stanford.edu/biodata/)
  - Open Graph Benchmark (OGB) (ogb.stanford.edu)