AIM 2: Artificial Intelligence in Medicine II

Harvard - BMIF 203 and BMI 702, Spring 2025

Lecture 9: Knowledge graph learning, Building multimodal knowledge graphs, Structure-inducing pre-training, Knowledge-based foundation models

HARVARD MEDICAL SCHOOL

Kempner INSTITUTE For the Study of Natural & Artificial Intelligence at Harvard University

BROAD INSTITUTE

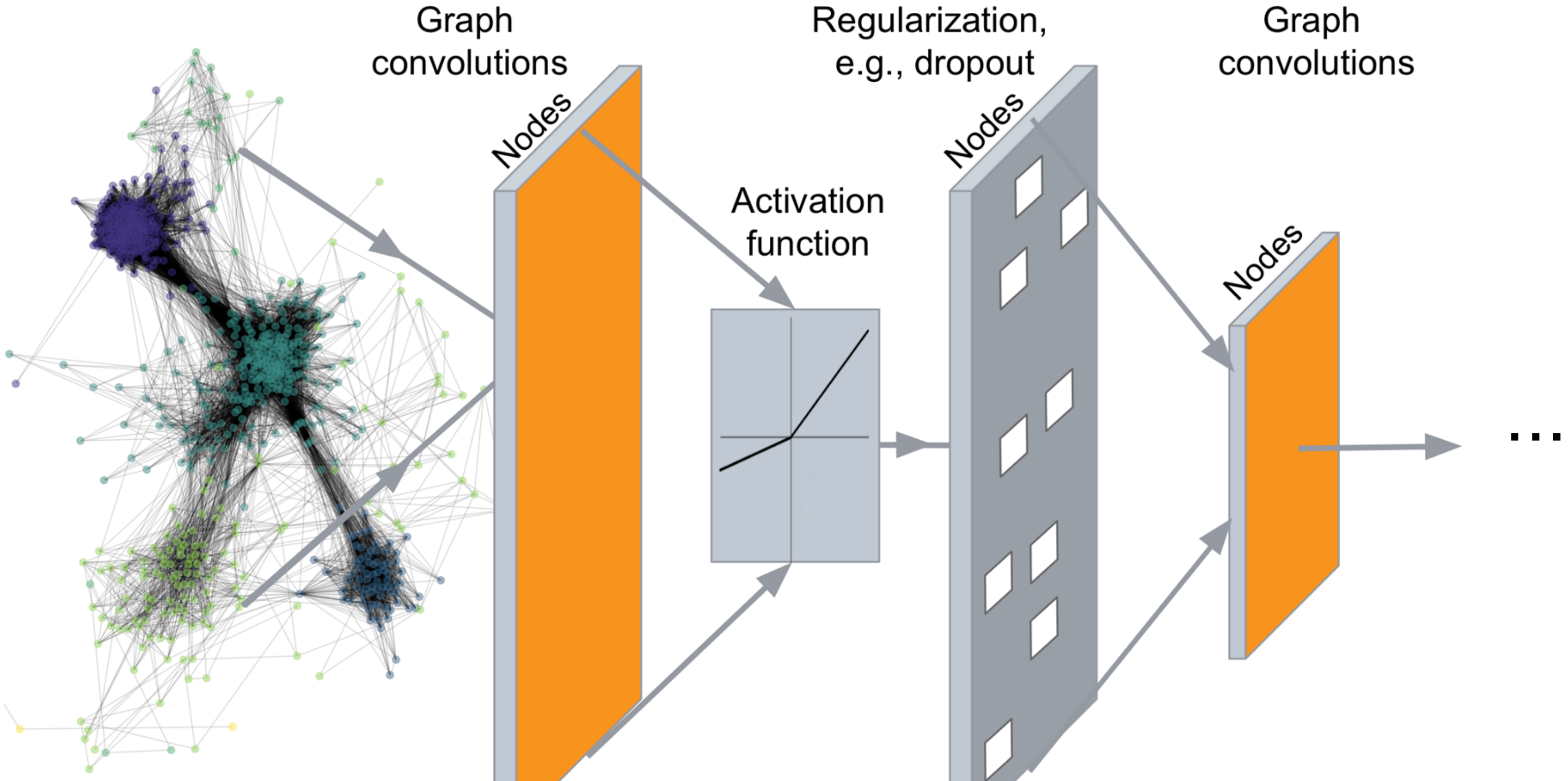
Marinka Zitnik
marinka@hms.harvard.edu

Deep graph representation learning

Recap of message passing neural network (MPNN) strategies

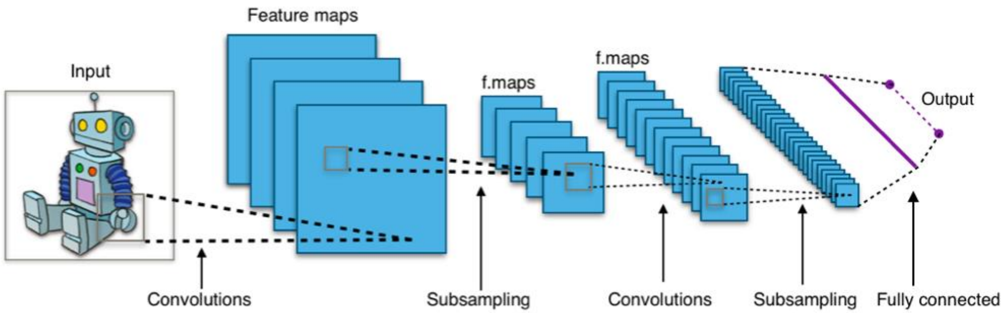
Graph neural networks

- Encoder: Multiple layers of nonlinear transformation of graph structure

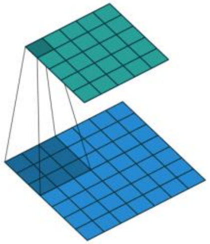


Convolutional networks

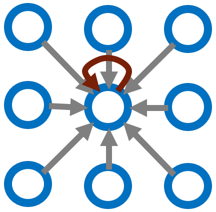
- Let’s start with convolutional networks on an image:



- Single convolutional network with a 3x3 filter:

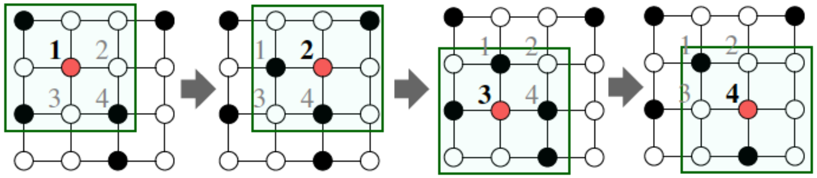


Image



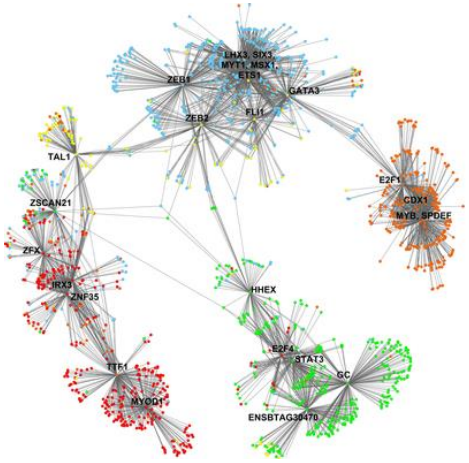
Graph

- Transform information (or messages) from the neighbors and combine them: $\sum_i W_i \, h_i$

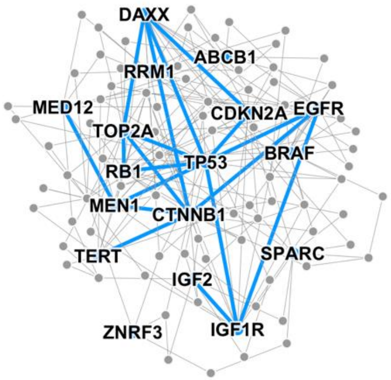


Real world graphs

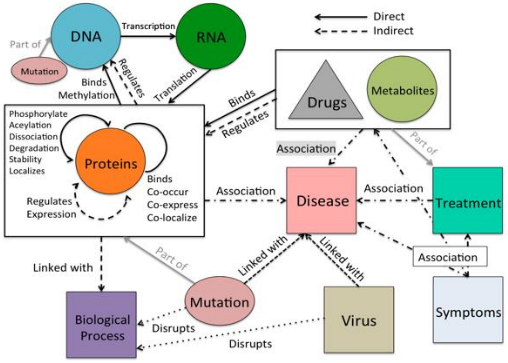
- But what if your graphs look like this?



Gene interaction network



Disease pathways

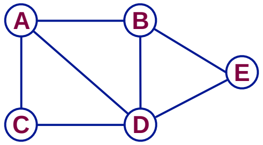


Biomedical knowledge graphs

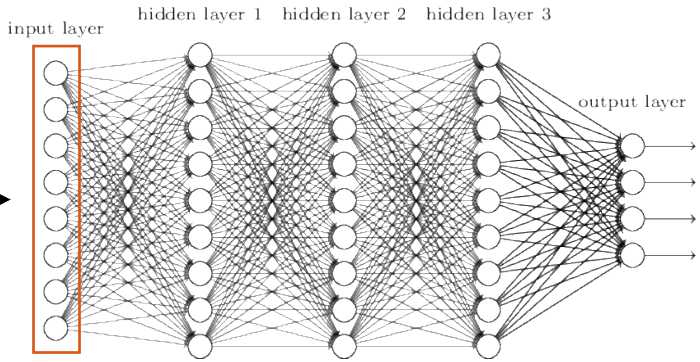
- Examples:
  - Biological or medical networks
  - Social networks
  - Information networks
  - Knowledge graphs
  - Communication networks
  - Web graphs
  - …

Naïve approach

- Join adjacency matrix and features
- Feed them into a deep neural network:



$$\begin{array}{c|ccccc|cc} & A & B & C & D & E & \multicolumn{2}{c}{Feat} \\ A & 0 & 1 & 1 & 1 & 0 & 1 & 0 \\ B & 1 & 0 & 0 & 1 & 1 & 0 & 0 \\ C & 1 & 0 & 0 & 1 & 0 & 0 & 1 \\ D & 1 & 1 & 1 & 0 & 1 & 1 & 1 \\ E & 0 & 1 & 0 & 1 & 0 & 1 & 0 \end{array}$$

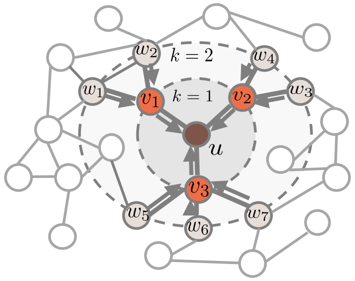


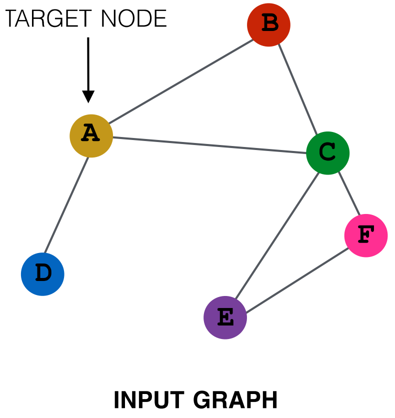
?

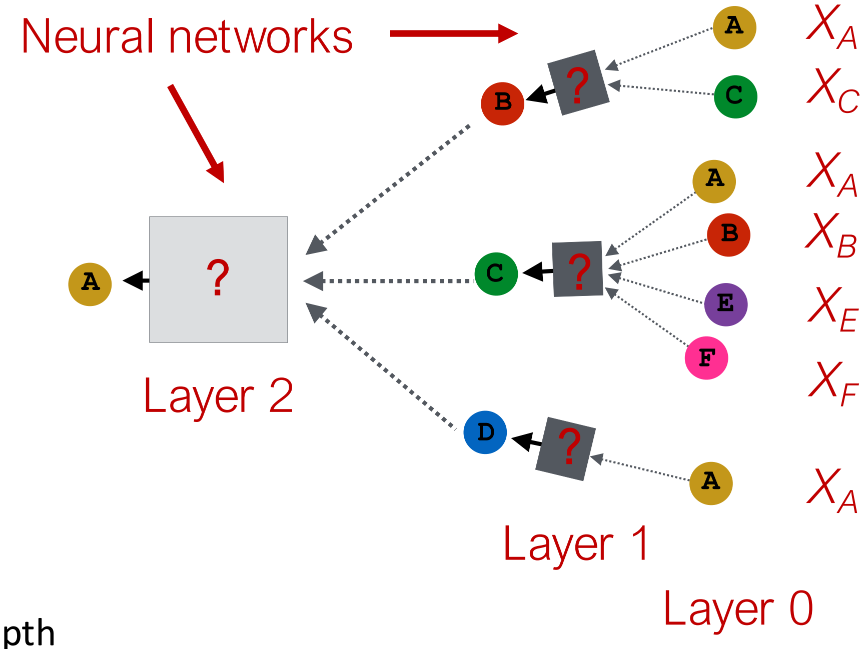
- Issues with this idea:
  - $O(N)$ parameters
  - Not applicable to graphs of different sizes
  - Not invariant to node ordering

Graph neural networks

- Intuition:
  - Each node’s neighborhood defines a computational graph
  - Generate node embeddings based on local network neighborhoods
- Neighborhood aggregation:

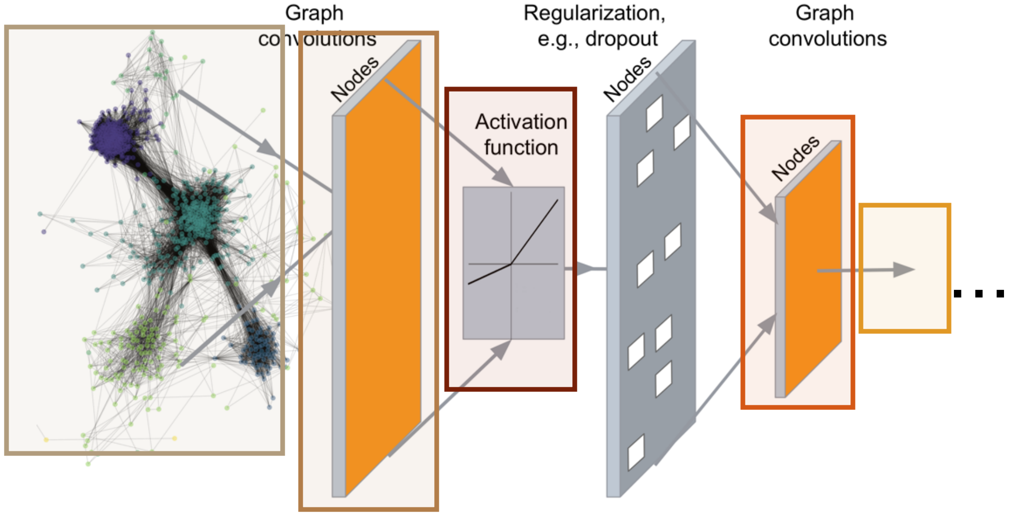






  - Model can be of arbitrary depth
    - Nodes have embeddings at each layer
    - Layer 0 embedding of node *u* is its input features X_u
- Basic neighborhood aggregation: Average information from neighbors and apply a neural network

Basic approach



Initial 0-th layer embeddings are equal to node features

$$\boxed{\mathbf{h}_v^0 = \mathbf{x}_v}$$

Previous layer embedding of v

$$\mathbf{h}_v^k = \boxed{\sigma}\left(\mathbf{W}_k \boxed{\sum_{u\in N(v)} \frac{\mathbf{h}_u^{k-1}}{|N(v)|}} + \mathbf{B}_k \boxed{\mathbf{h}_v^{k-1}}\right), \quad \forall k \in \{1, ..., K\}$$

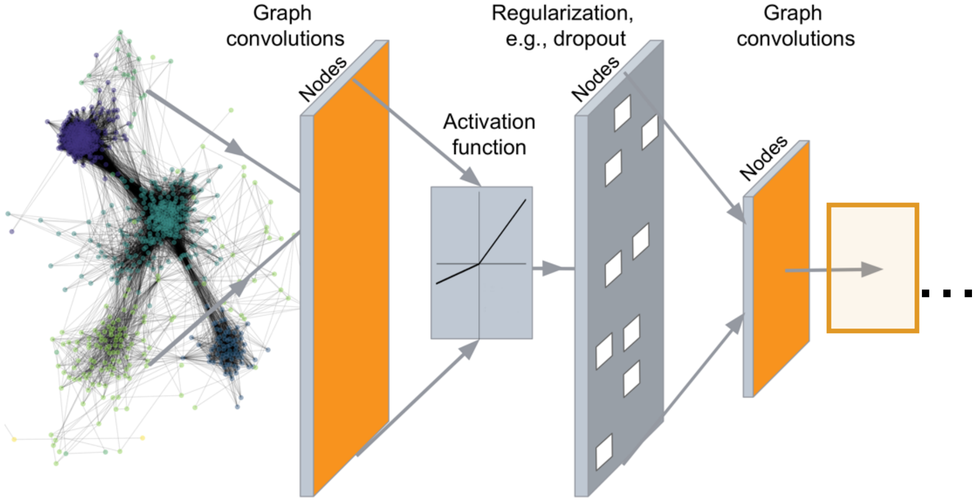
$$\boxed{\mathbf{z}_v} = \mathbf{h}_v^K$$

Average of neighbor's previous layer embeddings

Embedding after K layers of neighborhood aggregation

Non-linearity (e.g., ReLU)

Basic approach



$$\mathbf{h}_v^0 = \mathbf{x}_v$$

trainable weight matrices
(i.e., what we learn)

$$\mathbf{h}_v^k = \sigma\left(\boxed{\mathbf{W}_k} \sum_{u\in N(v)} \frac{\mathbf{h}_u^{k-1}}{|N(v)|} + \boxed{\mathbf{B}_k}\mathbf{h}_v^{k-1}\right), \quad \forall k \in \{1, ..., K\}$$

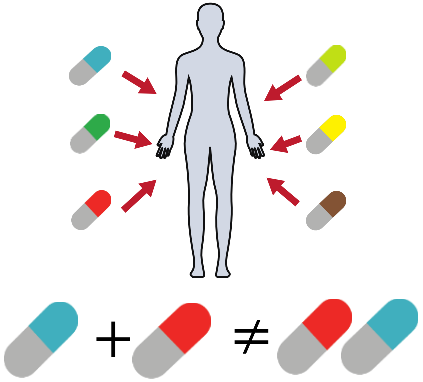
$$\boxed{\mathbf{z}_v} = \mathbf{h}_v^K$$

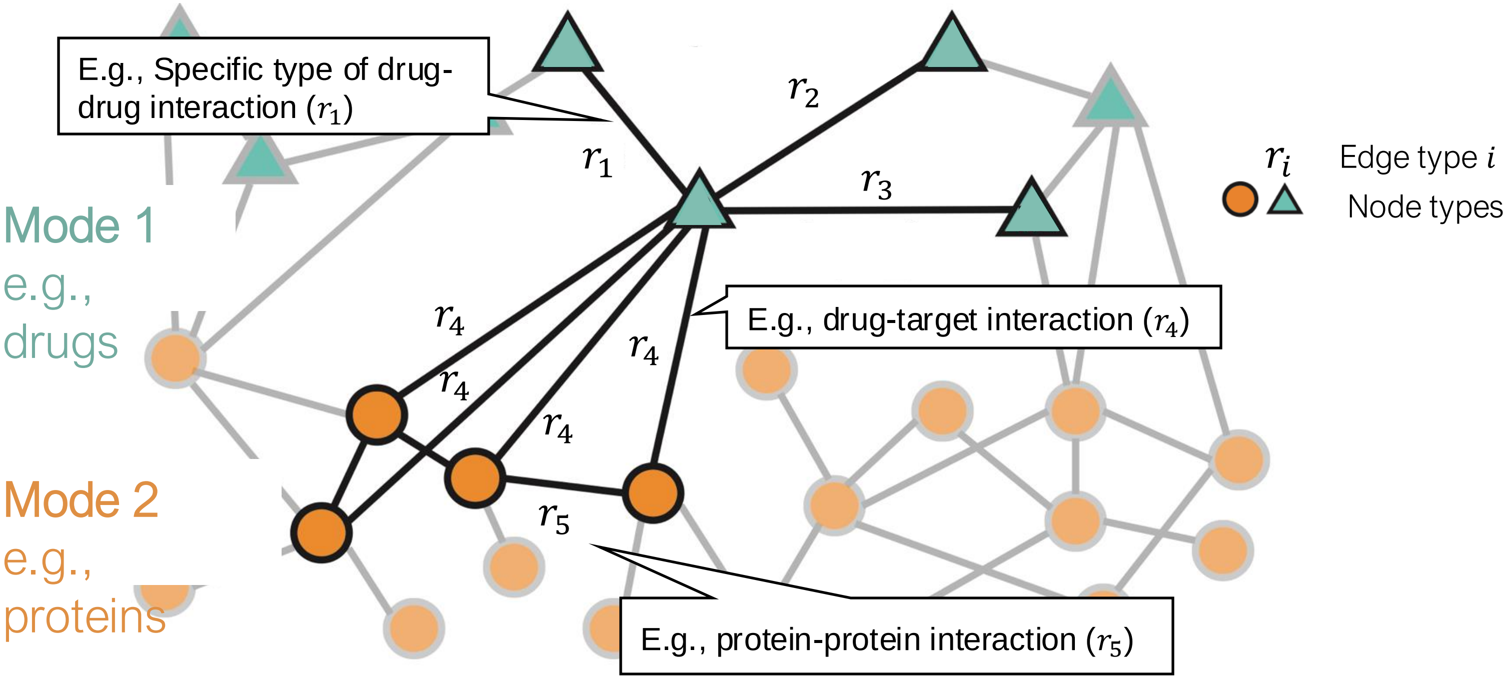
We can feed these into any loss function and run stochastic gradient descent to train the weight parameters

Polypharmacy modeling and antibiotic discovery

Application: Drug combinations

- Combinatorial explosion
  - >13 million possible combinations of 2 drugs
  - >20 billion possible combinations of 3 drugs
- Non-linear & non-additive interactions
  - Different effect than the additive effect of individual drugs
- Small subsets of patients
  - Side effects are interdependent
  - No info on drug combinations not yet used in patients




E.g., Specific type of drug-drug interaction (r_1)

E.g., drug-target interaction (r_4)

E.g., protein-protein interaction (r_5)

Mode 1
e.g.,
drugs

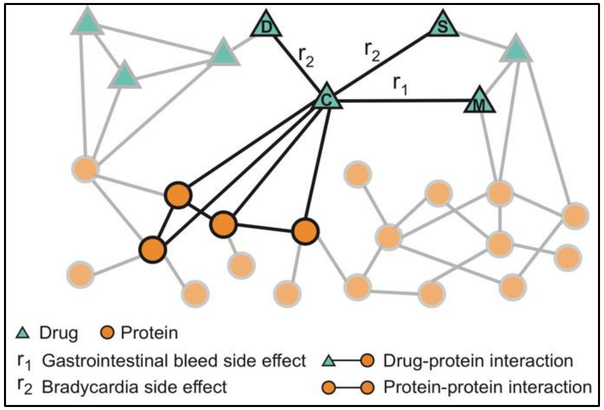
Mode 2
e.g.,
proteins

r_i Edge type i

● ▲ Node types

Modeling Polypharmacy Side Effects with Graph Convolutional Networks, Bioinformatics, 2018

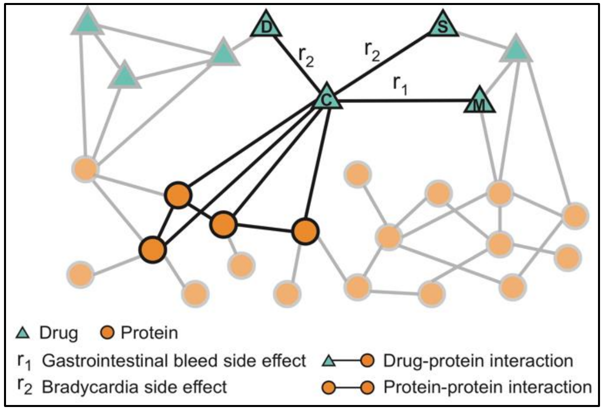
Polypharmacy dataset


D
S
C
M
r_2
r_2
r_1
Drug
Protein
r_1 Gastrointestinal bleed side effect
r_2 Bradycardia side effect
Drug-protein interaction
Protein-protein interaction

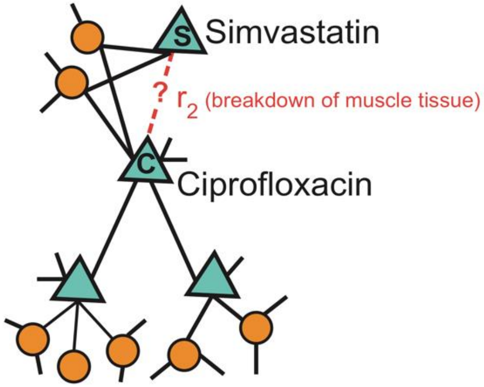
- Molecular, drug, and patient data for all US-approved drugs
  - 4,651,131 drug-drug edges: Patient data from adverse event system, tested for confounders [FDA]
  - 18,596 drug-protein edges
  - 719,402 protein-protein edges: Physical, metabolic enzyme-coupled, and signaling interactions
  - Drug and protein features: drugs' chemical structure, proteins' membership in pathways
- This is a multimodal network with over 5 million edges separated into 1,000 different edge types

Modeling Polypharmacy Side Effects with Graph Convolutional Networks, Bioinformatics, 2018

Experimental setup


D
S
C
M
r_2
r_2
r_1
Drug
Protein
r_1 Gastrointestinal bleed side effect
r_2 Bradycardia side effect
Drug-protein interaction
Protein-protein interaction

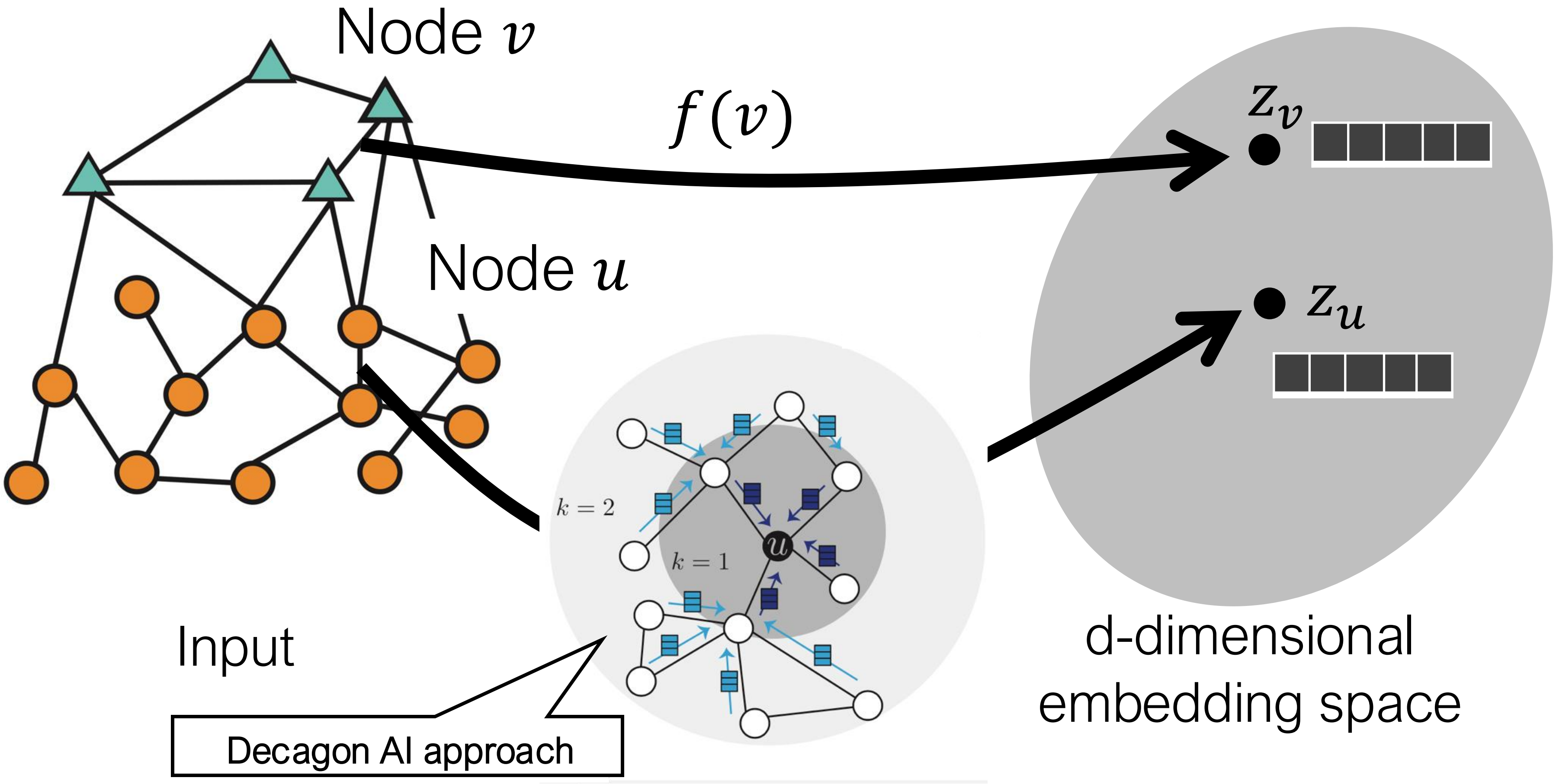
- Two main stages:
  1. Learn an embedding for every node in polypharmacy network
  2. Predict a score for every drug-drug, drug-protein, protein-protein pair in the test set based on the embeddings



Example: How likely will Simvastatin and Ciprofloxacin, when taken together, break down muscle tissue?

Modeling Polypharmacy Side Effects with Graph Convolutional Networks, Bioinformatics, 2018

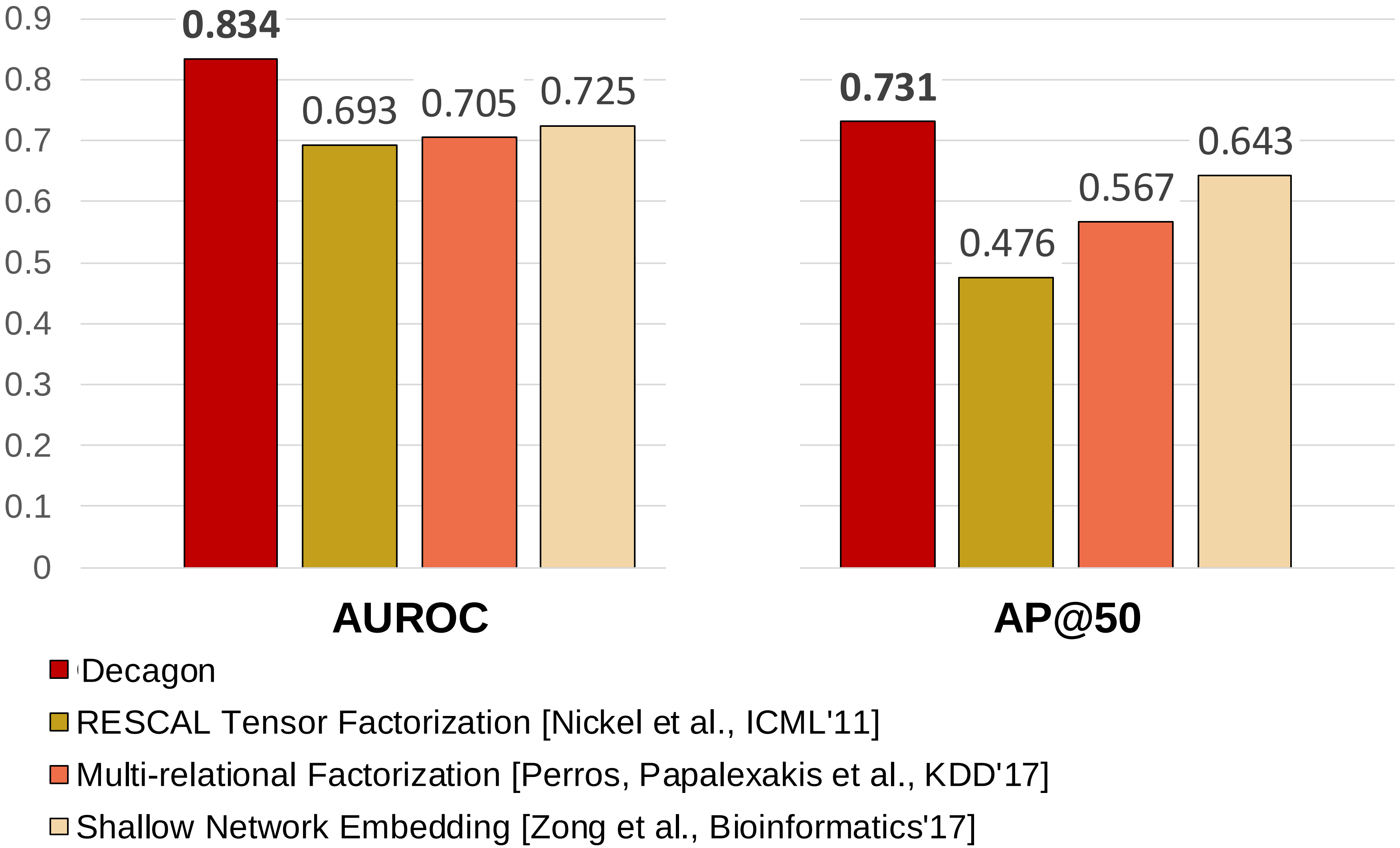
Approach: Graph Neural Network



Map nodes to d-dimensional embeddings such that nodes with similar network neighborhoods are embedded close together

Modeling Polypharmacy Side Effects with Graph Convolutional Networks, Bioinformatics, 2018

Results: Polypharmacy side effects

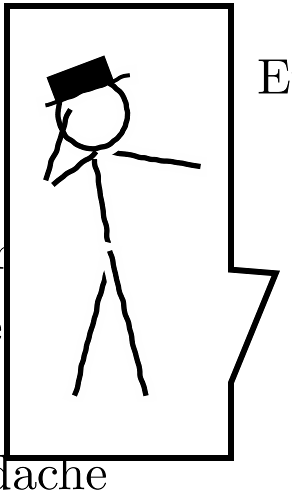


Modeling Polypharmacy Side Effects with Graph Convolutional Networks, Bioinformatics, 2018

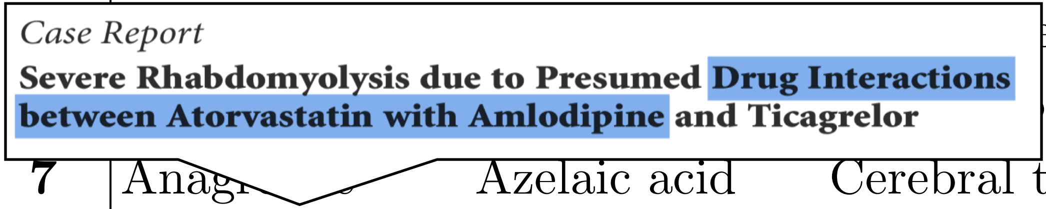
Results: Polypharmacy side effects

Approach:

1) Train deep model on data generated **prior to 2012**
2) How many **predictions** have been **confirmed after 2012**?



Rank	Drug	Drug	Side effect	Evidence found
1	Pyrimethamine	Aliskiren	Sarcoma	
2	Tigecycline	Bimatoprost	Autonomic n	
3	Telangiectases	Omeprazole	Dacarbazine	
4	Tolcapone	Pyrimethamine	Blood brain	
5			eadache	
6			ular acidosis	
7	Anagr	Azelaic acid	Cerebral thrombosis	
8	Atorvastatin	Amlodipine	Muscle inflammation	
9	Aliskiren	Tioconazole	Breast inflammation	
10	Estradiol	Nadolol	Endometriosis	



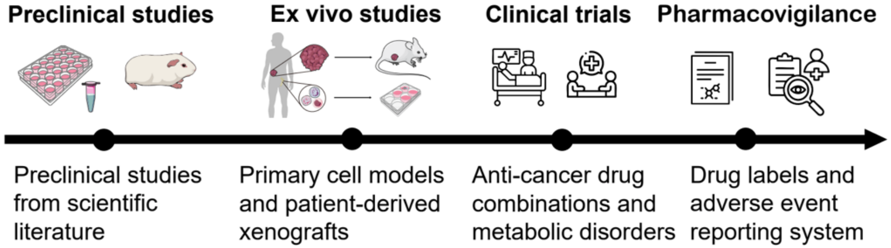
Case Report

Severe Rhabdomyolysis due to Presumed Drug Interactions between Atorvastatin with Amlodipine and Ticagrelor

Modeling Polypharmacy Side Effects with Graph Convolutional Networks, Bioinformatics, 2018

Multimodal AI predicts clinical outcomes of drug combinations from preclinical data

- **Personalized oncology therapy:** Predicts leukemia drug combination responses using patient genomics and xenograft models
- **Drug safety & transporter interactions:** Identifies organ-specific toxicities and transporter-based risks for early drug development
- **Oncology drug combinations & polypharmacy:** Assesses PARP inhibitor safety, differentiating approved vs. investigational regimens
- **Metabolic disease insights:** Ranked Resmetirom among the safest candidates for MASH, supporting FDA approval

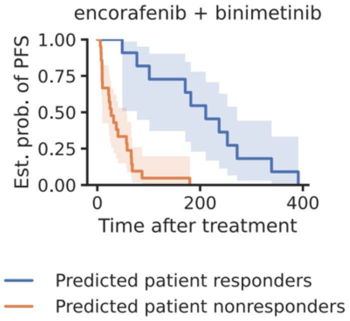


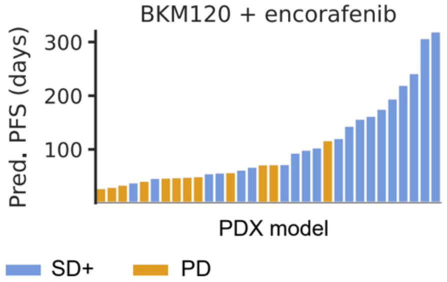
Preclinical studies — Preclinical studies from scientific literature

Ex vivo studies — Primary cell models and patient-derived xenografts

Clinical trials — Anti-cancer drug combinations and metabolic disorders

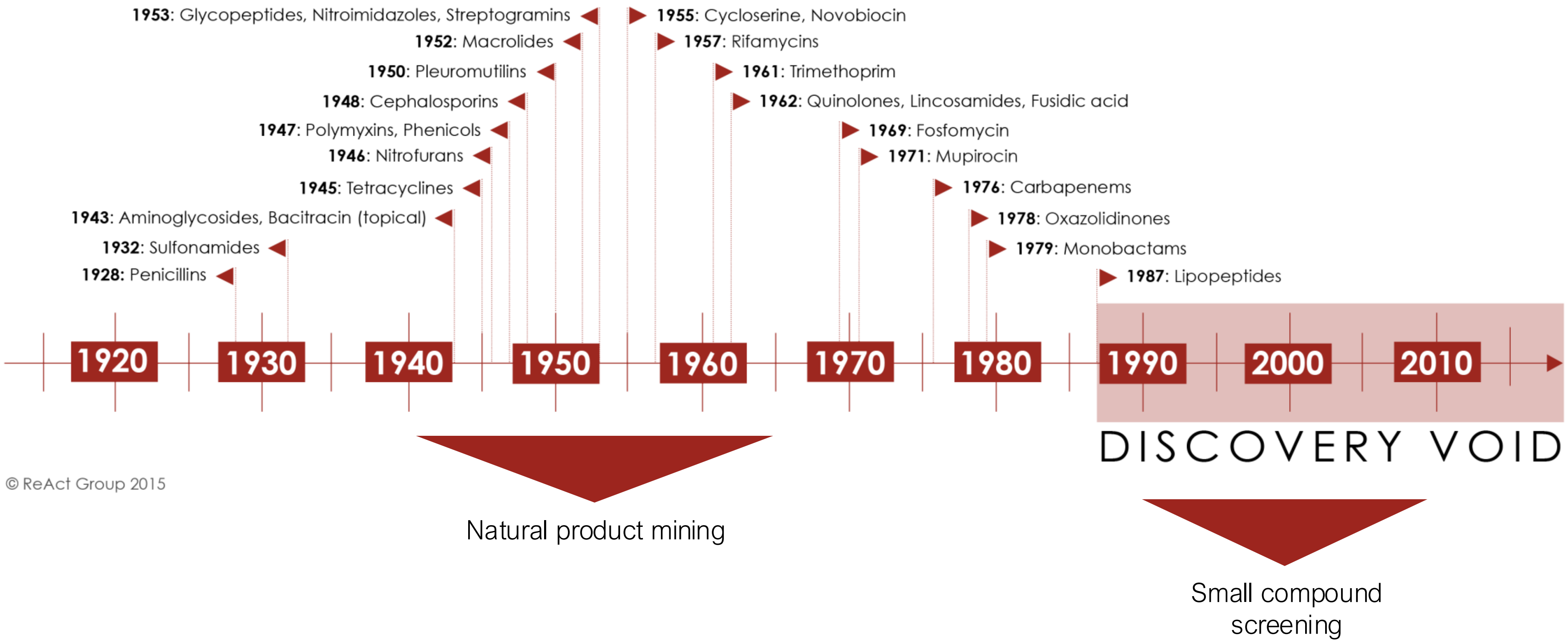
Pharmacovigilance — Drug labels and adverse event reporting system





Multimodal AI predicts clinical outcomes of drug combinations from preclinical data, arXiv 2025

Application: Antibiotic discovery



1928: Penicillins

1932: Sulfonamides

1943: Aminoglycosides, Bacitracin (topical)

1945: Tetracyclines

1946: Nitrofurans

1947: Polymyxins, Phenicols

1948: Cephalosporins

1950: Pleuromutilins

1952: Macrolides

1953: Glycopeptides, Nitroimidazoles, Streptogramins

1955: Cycloserine, Novobiocin

1957: Rifamycins

1961: Trimethoprim

1962: Quinolones, Lincosamides, Fusidic acid

1969: Fosfomycin

1971: Mupirocin

1976: Carbapenems

1978: Oxazolidinones

1979: Monobactams

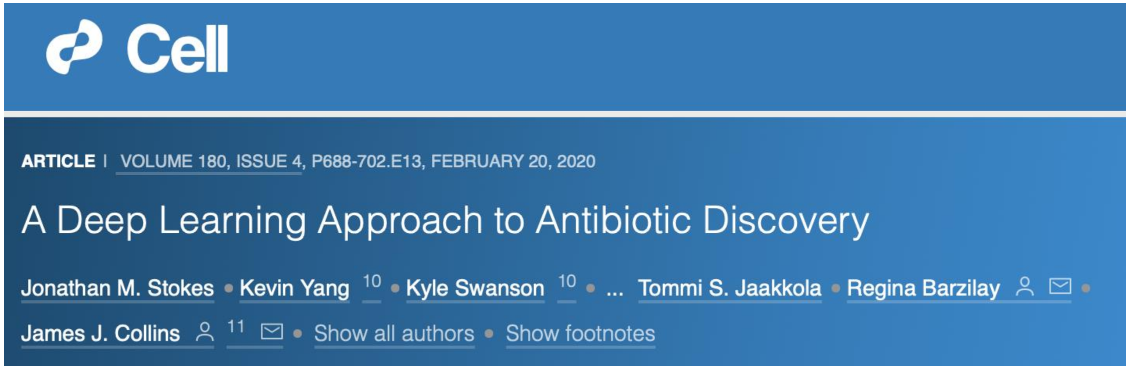
1987: Lipopeptides

DISCOVERY VOID

© ReAct Group 2015

Natural product mining

Small compound screening

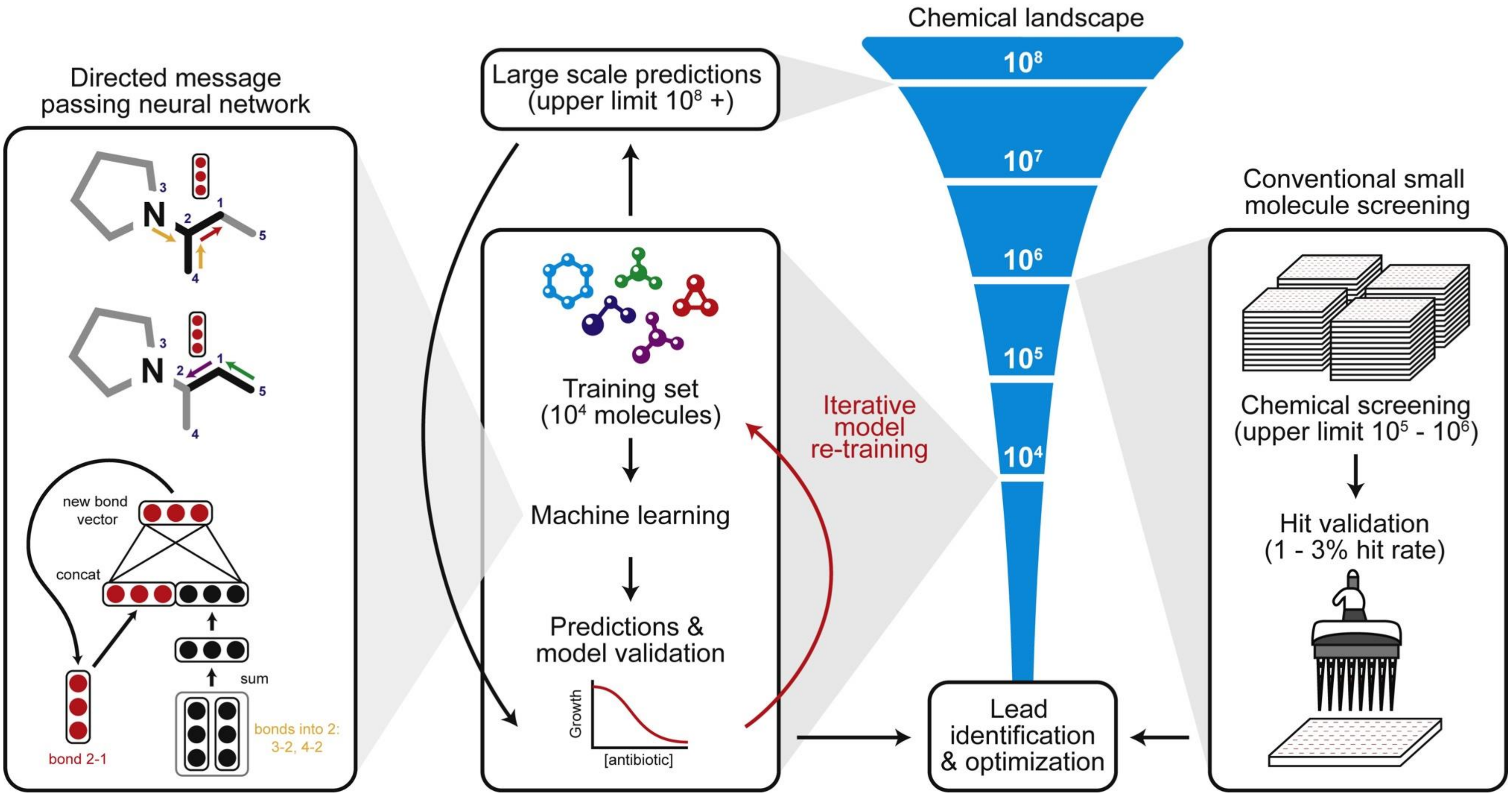


ARTICLE | VOLUME 180, ISSUE 4, P688-702.E13, FEBRUARY 20, 2020

A Deep Learning Approach to Antibiotic Discovery

Jonathan M. Stokes • Kevin Yang [10] • Kyle Swanson [10] • ... Tommi S. Jaakkola • Regina Barzilay • James J. Collins [11] • Show all authors • Show footnotes

GNNs to learn molecular structure

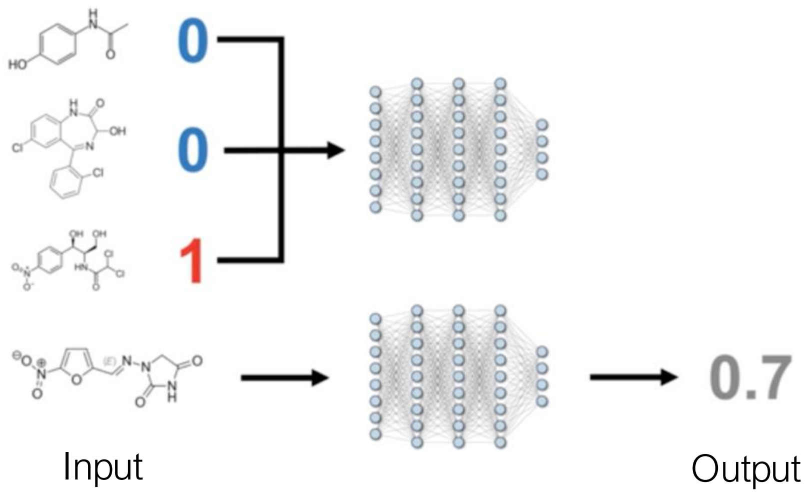

Directed message passing neural network
3
N
2
1
5
4
3
N
2
1
5
4
new bond vector
concat
sum
bond 2-1
bonds into 2: 3-2, 4-2
Large scale predictions (upper limit 10⁸ +)
Training set (10⁴ molecules)
Machine learning
Predictions & model validation
Growth
[antibiotic]
Iterative model re-training
Chemical landscape
10⁸
10⁷
10⁶
10⁵
10⁴
Lead identification & optimization
Conventional small molecule screening
Chemical screening (upper limit 10⁵ - 10⁶)
Hit validation (1 - 3% hit rate)

Directed message passing neural network model iteratively (1) learns representations of molecules and (2) optimizes the representations for predicting growth inhibition

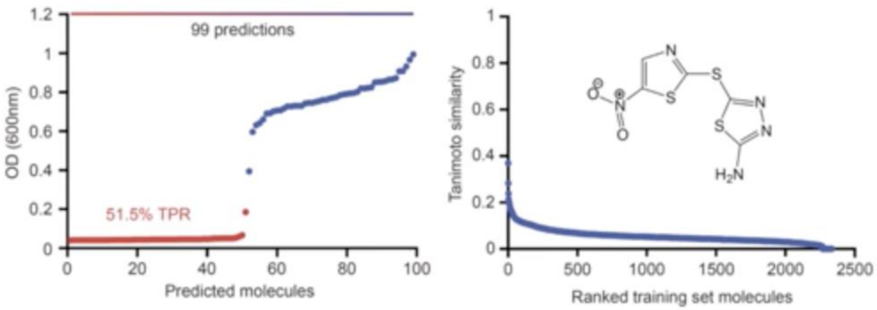
A Deep Learning Approach to Antibiotic Discovery, *Cell*, 2020.

Experimental setup

Training Dataset
(Human Medicines and Natural Products)



Empirical Validation
(Broad Repurposing Hub)



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

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

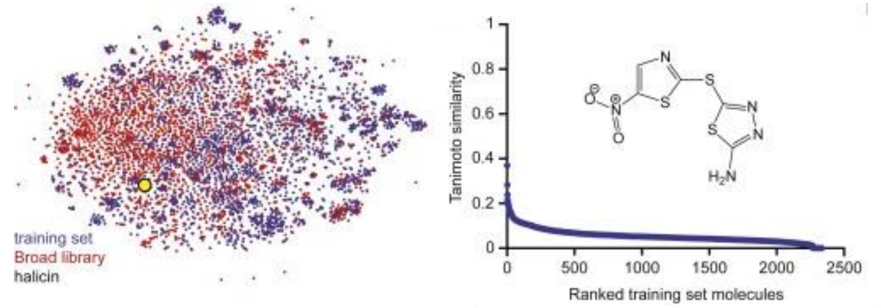
Task: Test top 99 predictions & prioritize based on similarity to known antibiotics or predicted toxicity

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

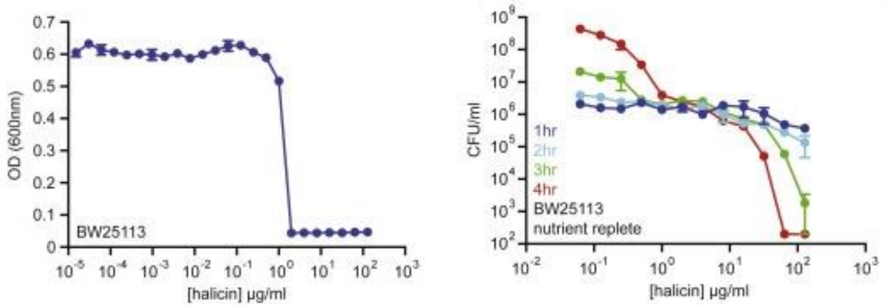
Results

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

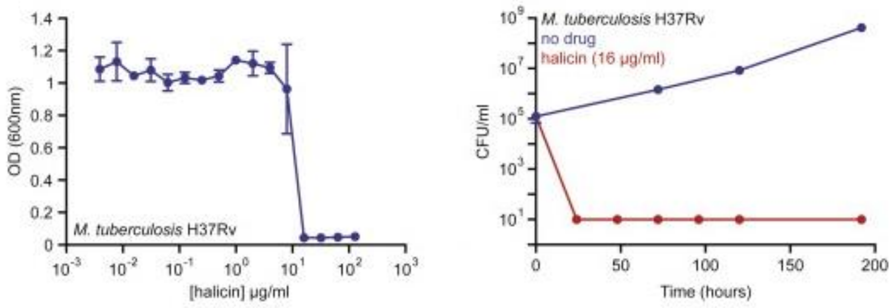
Halicin predicted to be antibacterial



Halicin against *E. coli*



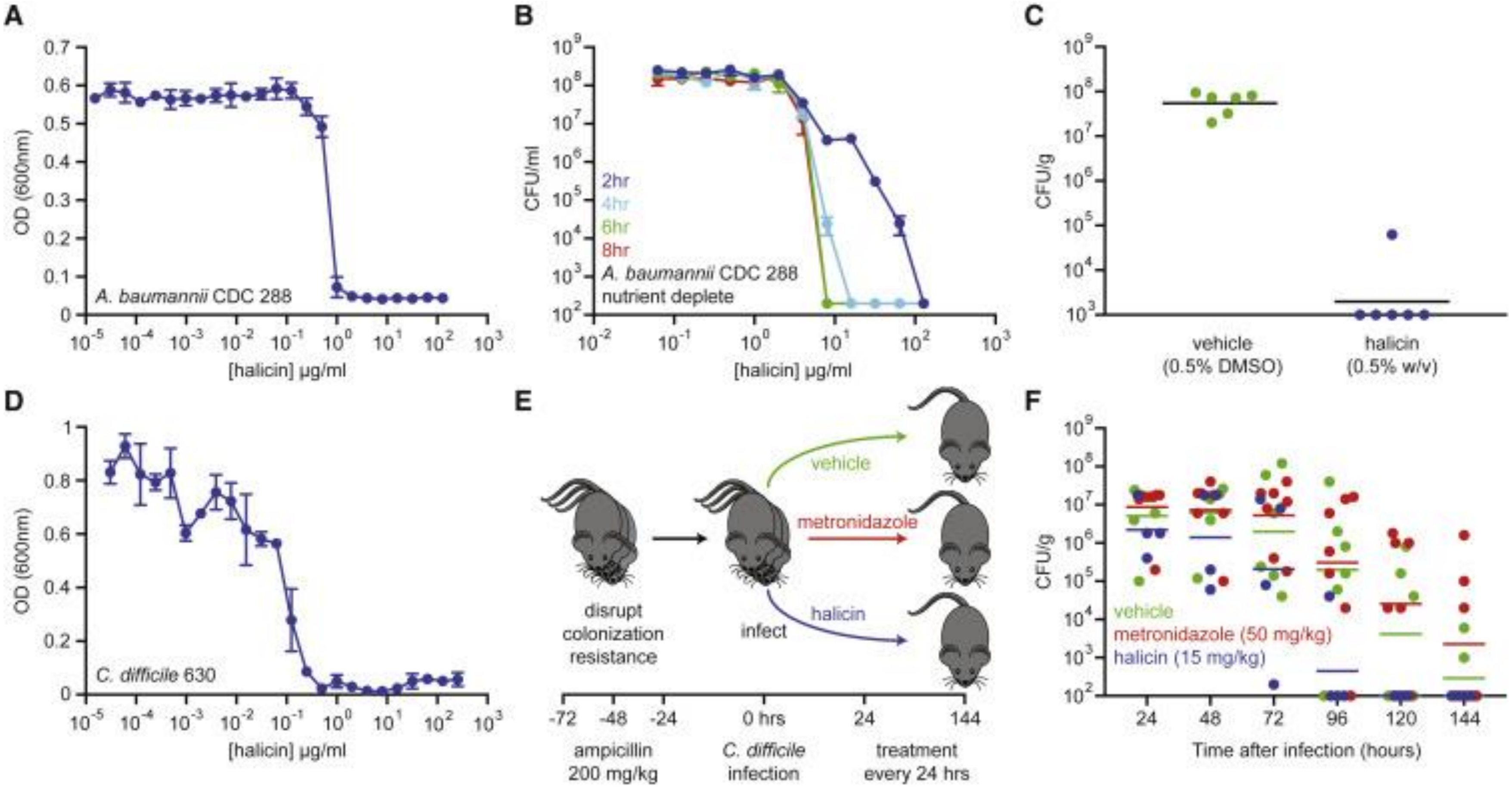
Halicin against *M. tuberculosis*



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

Results

Halicin's efficacy in murine models of infection



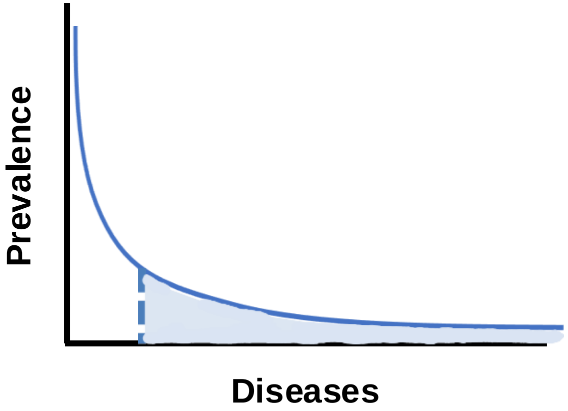
Validated against ~6K molecules to identify halicin, a novel candidate antibiotic

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

Rare disease diagnosis

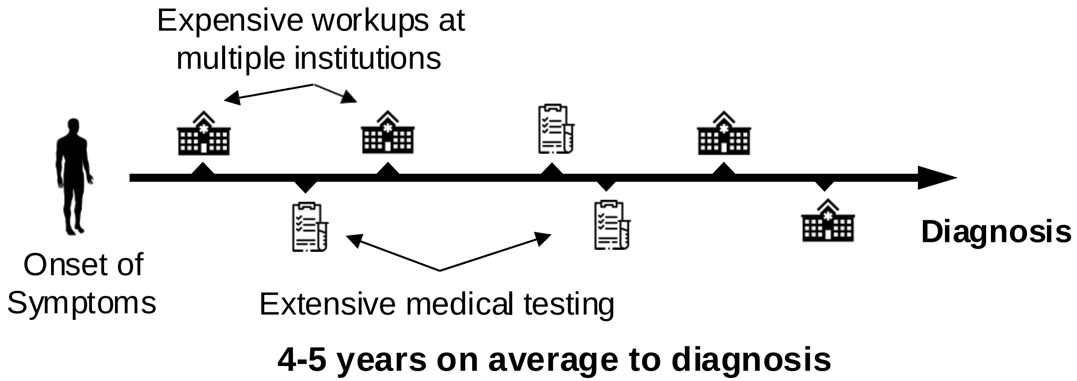
Rare disease diagnosis

- Rare diseases affect between 300-400 million or 1 in 20 people worldwide, yet each disease affects no more than 50 per 100,000 individuals
- Diagnosis is challenging due to the heterogeneity of clinical presentations and small patient populations



Rare disease diagnosis

- Many patients suffering from rare diseases are undiagnosed. It currently takes 4-5 years on average for patients to receive a diagnosis.


Expensive workups at multiple institutions
Onset of Symptoms
Extensive medical testing
Diagnosis
4-5 years on average to diagnosis

Can AI help shorten diagnostic odysseys for rare disease patients?

Haendel et al. How many rare diseases are there? *Nature Review Drug Discovery* (2020).
Wakap et al. Estimating cumulative point prevalence of rare diseases: analysis of the Orphanet database. *EJHG* (2020).

Diagnostic odysseys

- Over 7,000 rare diseases, each affects < 200,000 patients in the US
  - Most diseases are phenotypically heterogeneous
  - Front-line clinicians might lack disease experience, resulting in expensive clinical workups for patients across multiple years
  - Diagnosis often requires a specialist, sub-specialist, or multi-disciplinary referrals

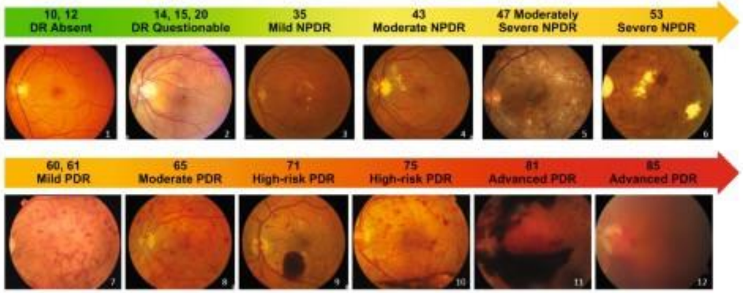
- On average, the long search for a **rare disease diagnosis takes 5 to 7 years, 4 up to 8 physicians, and 2 to 3 misdiagnoses**

- Diagnostic delay is so pervasive that it leads to problems for patients:
  - Undergoing **redundant testing and procedures**
  - Substantial delay in obtaining disease-appropriate management and **inappropriate therapies**
  - **Irreversible disease progression**—time window for intervention can be missed leading to disease progression

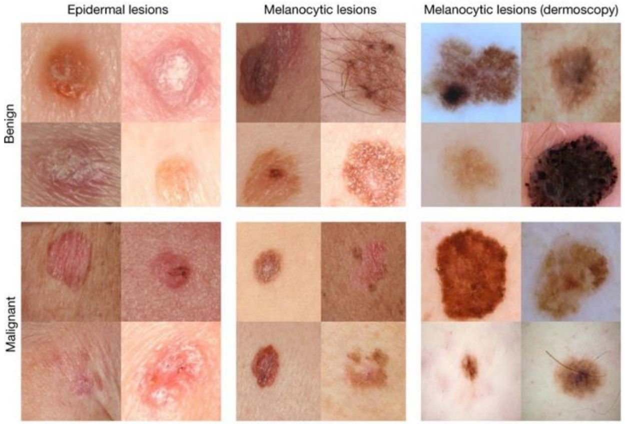
Can AI help shorten diagnostic odysseys for rare disease patients?

AI models for disease diagnosis

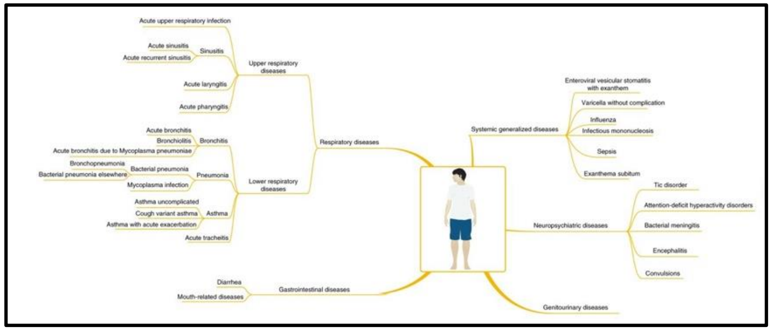
Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopahty in Retinal Fundus Photographs (*JAMA*)



Dermatologist-level Classification of Skin Cancer (*Nature*)

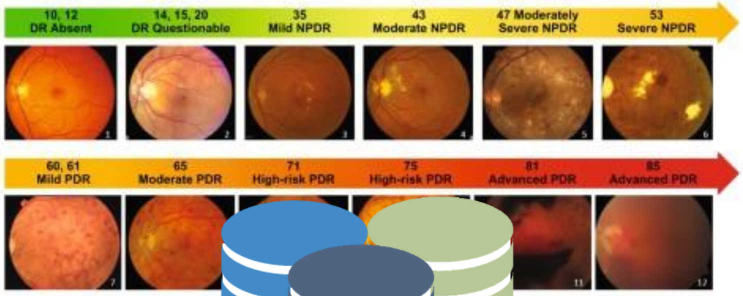


Evaluation and Accurate Diagnoses of Pediatric Diseases Using AI (*Nature Medicine*)


Acute upper respiratory infection
Acute sinusitis
Acute recurrent sinusitis
Sinusitis
Upper respiratory diseases
Acute laryngitis
Acute pharyngitis
Acute bronchitis
Bronchiolitis
Bronchitis
Acute bronchitis due to Mycoplasma pneumoniae
Bronchopneumonia
Bacterial pneumonia
Bacterial pneumonia elsewhere
Pneumonia
Mycoplasma infection
Lower respiratory diseases
Respiratory diseases
Asthma uncomplicated
Cough variant asthma
Asthma
Asthma with acute exacerbation
Acute tracheitis
Diarrhea
Mouth-related diseases
Gastrointestinal diseases
Systemic generalized diseases
Enteroviral vesicular stomatitis with exanthem
Varicella without complication
Influenza
Infectious mononucleosis
Sepsis
Exanthema subitum
Neuropsychiatric diseases
Tic disorder
Attention-deficit hyperactivity disorders
Bacterial meningitis
Encephalitis
Convulsions
Genitourinary diseases

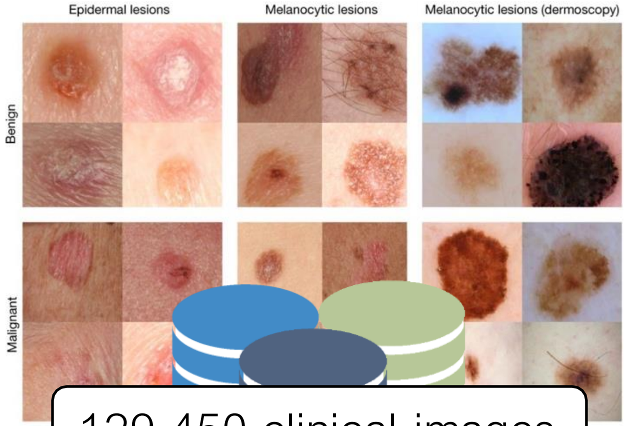
AI models for disease diagnosis

Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopahty in Retinal Fundus Photographs (*JAMA*)



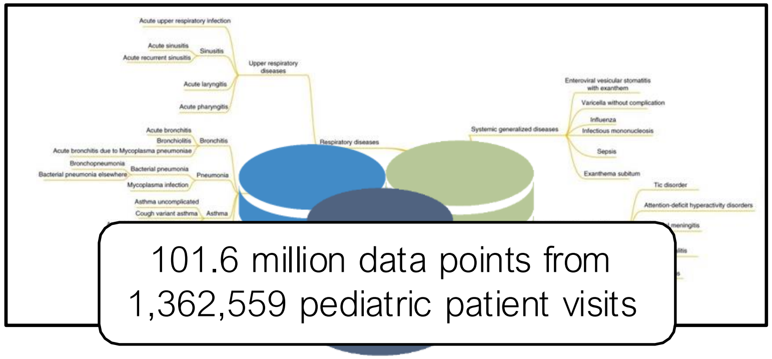
128,175 retinal images

Dermatologist-level Classification of Skin Cancer (*Nature*)



129,450 clinical images

Evaluation and Accurate Diagnoses of Pediatric Diseases Using AI (*Nature Medicine*)


Acute upper respiratory infection
Acute sinusitis
Acute recurrent sinusitis
Sinusitis
Upper respiratory diseases
Acute laryngitis
Acute pharyngitis
Acute bronchitis
Bronchiolitis
Bronchitis
Acute bronchitis due to Mycoplasma pneumoniae
Bronchopneumonia
Bacterial pneumonia
Bacterial pneumonia elsewhere
Pneumonia
Mycoplasma infection
Asthma uncomplicated
Cough variant asthma
Asthma
Respiratory diseases
Systemic generalized diseases
Enteroviral vesicular stomatitis with exanthem
Varicella without complication
Influenza
Infectious mononucleosis
Sepsis
Exanthema subitum
Tic disorder
Attention-deficit hyperactivity disorders
meningitis

101.6 million data points from
1,362,559 pediatric patient visits

Rare disease diagnosis is hard!

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

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

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

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

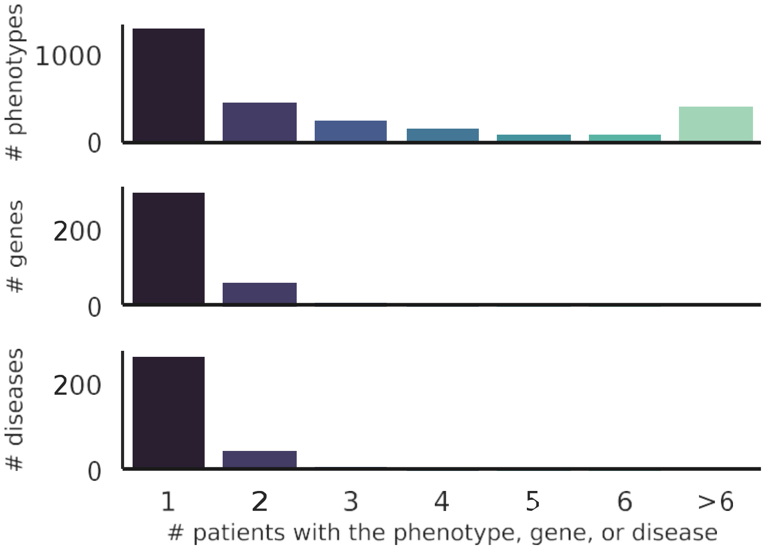
Rare disease diagnosis is hard!

1. Need to extrapolate beyond training distribution to never-before-seen genetic conditions
2. Approaches must be able to learn from limited data given the lack of large annotated datasets of patients with rare genetic diseases & low prevalence of each disease

Low overlap of phenotypes, causal genes, and diseases across patients

Undiagnosed Diseases Network

Of 465 diagnosed patients in the UDN, there are 378 unique causal genes and 299 unique diseases.



Rare disease diagnosis is hard!

1. Need to extrapolate beyond training distribution to never-before-seen genetic conditions
2. Approaches must be able to learn from limited data given the lack of large annotated datasets of patients with rare genetic diseases & low prevalence of each disease

Undiagnosed Diseases Network

Of 465 diagnosed patients in the UDN, there are 378 unique causal genes and 299 unique diseases.

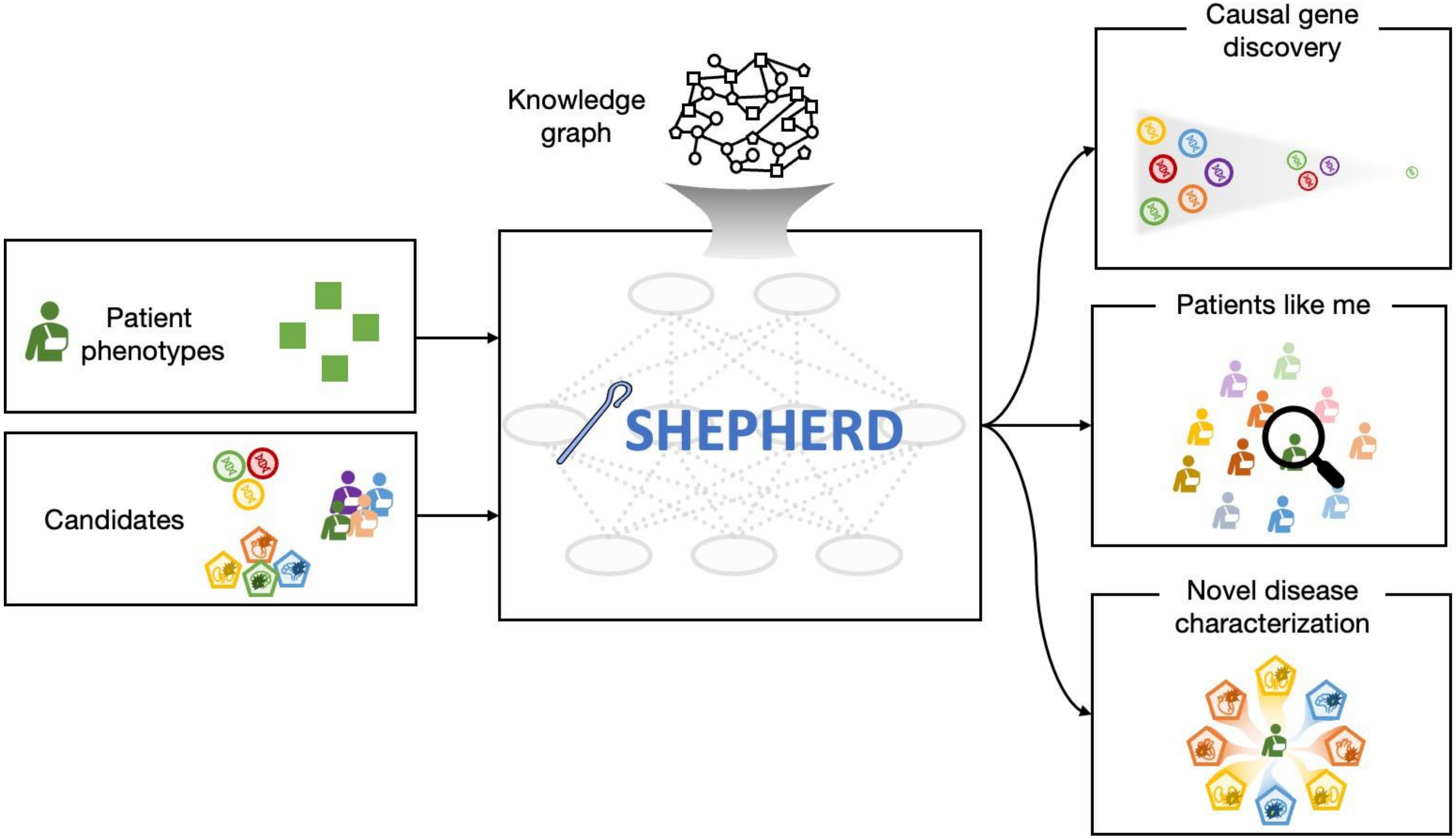
Phenotypic heterogeneity

% phenotypic overlap in patients with the same diseases
67% +/- 43%

Novel / atypical conditions

% patient phenotypes with known association to causal gene
28% +/- 21%

SHEPHERD: KG-based AI for rare disease diagnosis

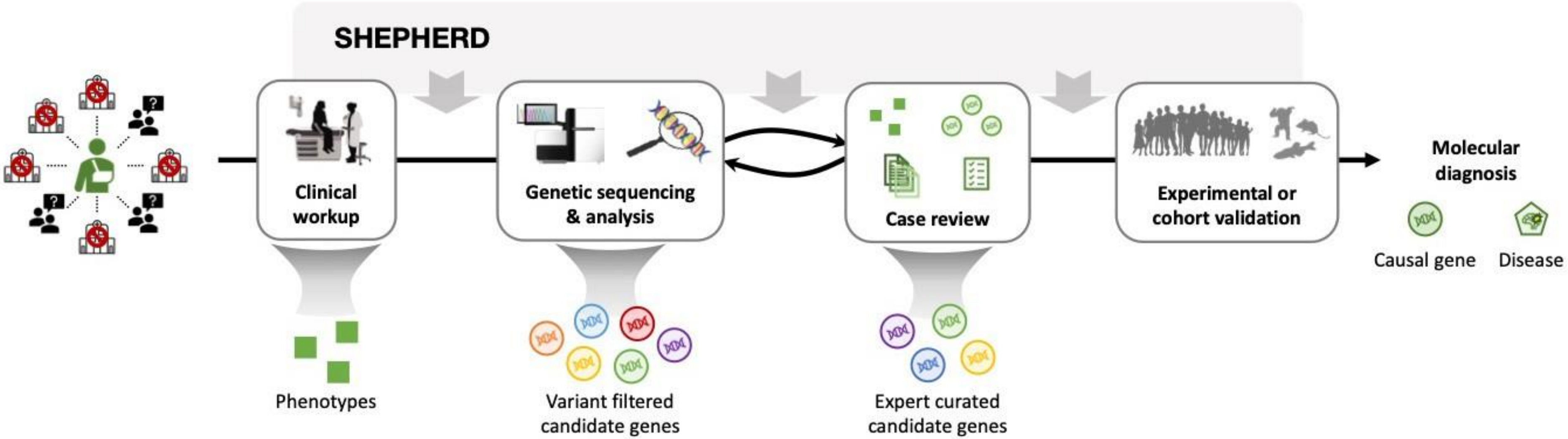


Given a patient’s set of **phenotypes** and a list of **candidate genes**, which gene(s) are most likely to explain the patient's presenting symptoms?

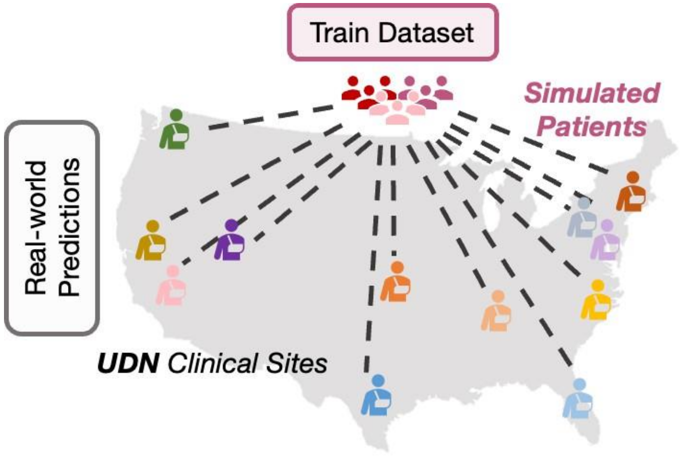
Given a patient’s set of **phenotypes** and a **cohort of rare disease patients**, can we identify other patients with similar genetic conditions?

Given a patient’s set of **phenotypes** and a set of **known diseases**, can we characterize the clinical presentation based on our current knowledge of rare diseases?

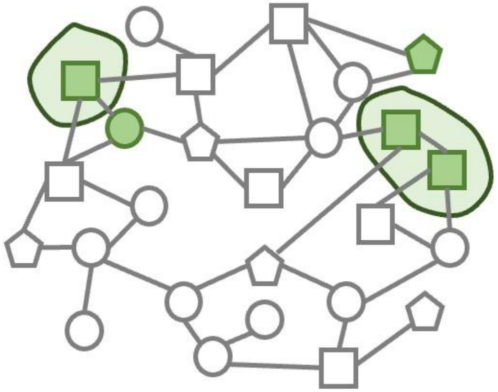
AI for hard-to-diagnose diseases


SHEPHERD
Clinical workup
Genetic sequencing & analysis
Case review
Experimental or cohort validation
Molecular diagnosis
Causal gene
Disease
Phenotypes
Variant filtered candidate genes
Expert curated candidate genes

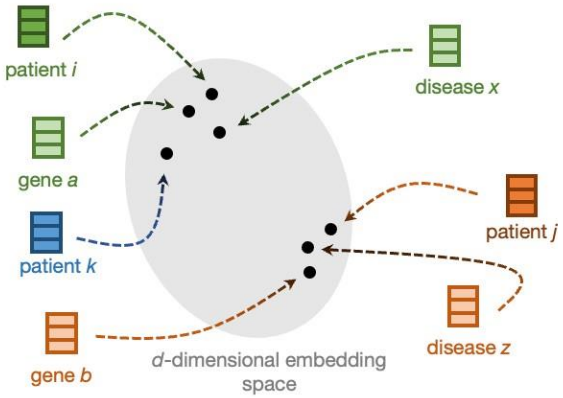
Key features of SHEPHERD



Train on simulated patients, evaluate on UDN patients



Model patients as subgraphs in knowledge graph



Perform label-efficient model training

Training data: Simulated patients

42,680 simulated patients across 2,134 diseases in Orphanet

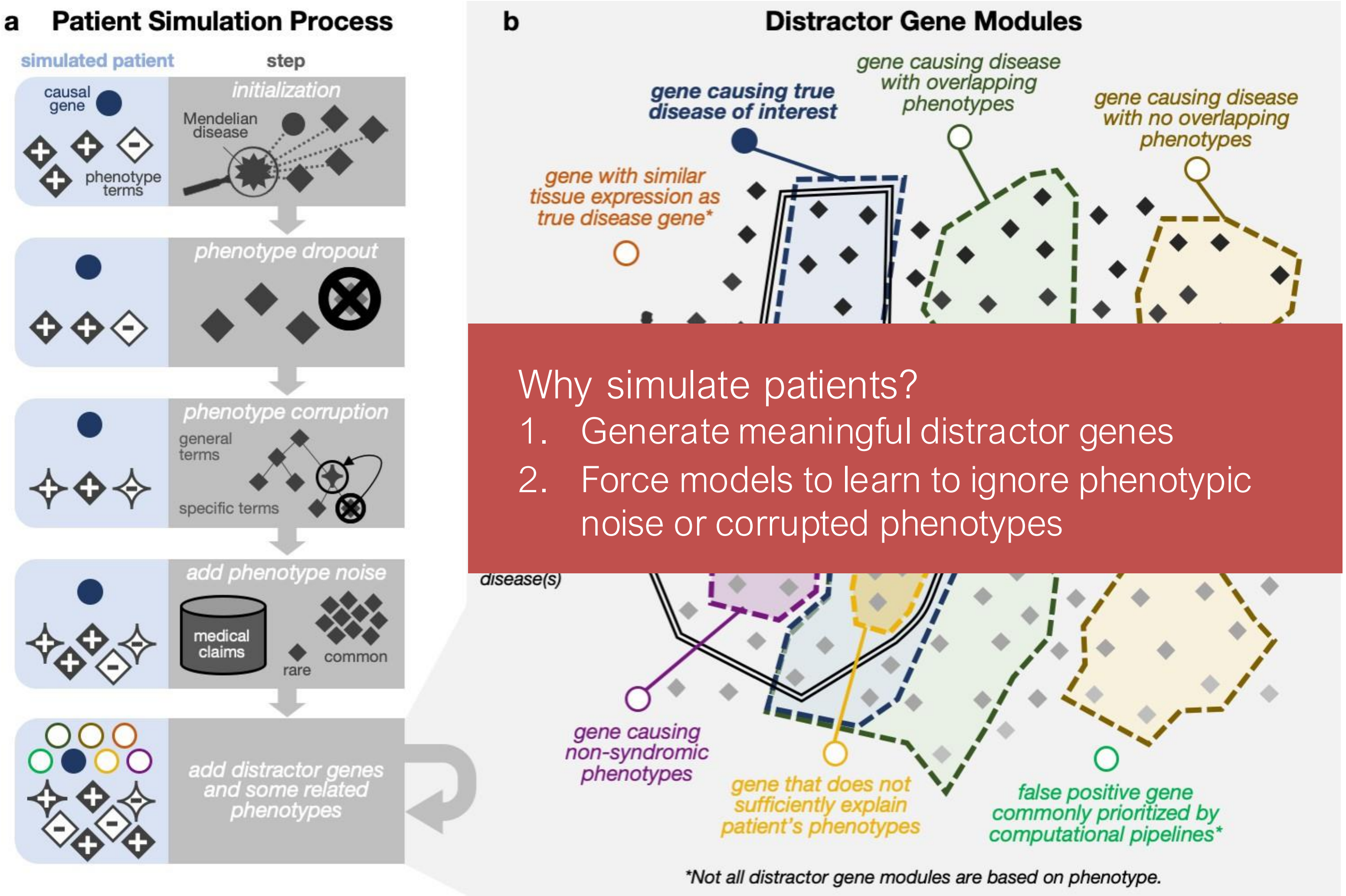
Train set
(N = 36,224)

Validation set
(N = 6,400)

Disease-split training and validation to select for generalizable models

Alsentzer et al. Simulation of undiagnosed patients with novel genetic conditions. *Nature Communications* 2024

Simulation process


a
Patient Simulation Process
simulated patient
step
causal gene
phenotype terms
initialization
Mendelian disease
phenotype dropout
phenotype corruption
general terms
specific terms
add phenotype noise
medical claims
rare
common
add distractor genes and some related phenotypes
b
Distractor Gene Modules
gene causing true disease of interest
gene causing disease with overlapping phenotypes
gene causing disease with no overlapping phenotypes
gene with similar tissue expression as true disease gene*
disease(s)
gene causing non-syndromic phenotypes
gene that does not sufficiently explain patient's phenotypes
false positive gene commonly prioritized by computational pipelines*
*Not all distractor gene modules are based on phenotype.

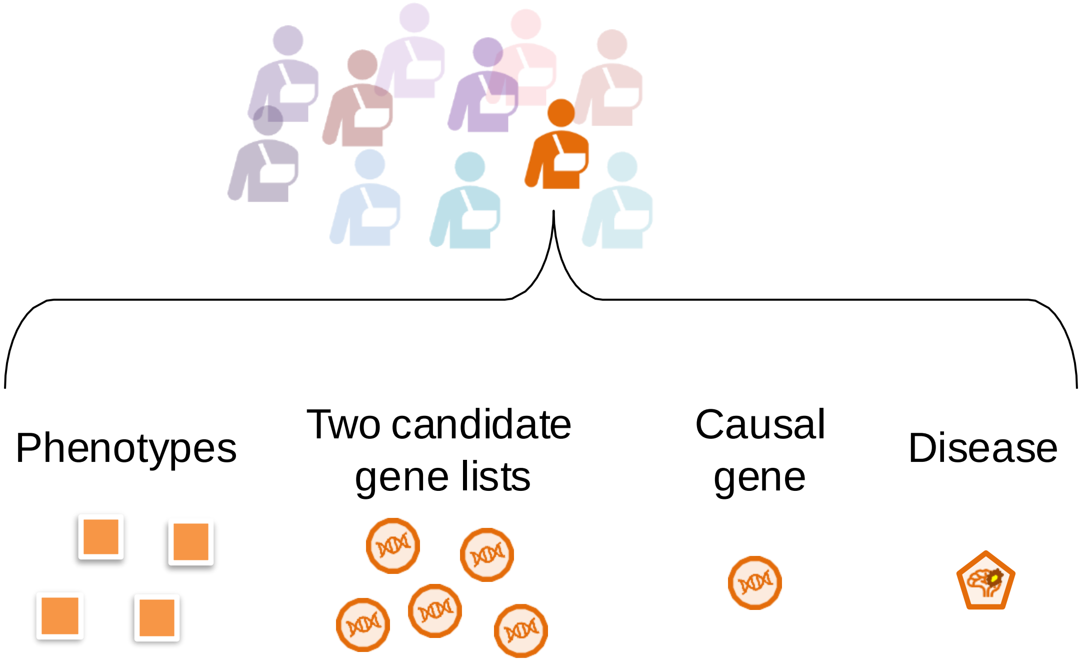
Why simulate patients?

1. Generate meaningful distractor genes
2. Force models to learn to ignore phenotypic noise or corrupted phenotypes

Alsentzer et al. Simulation of undiagnosed patients with novel genetic conditions. *Nature Communications* 2024

Undiagnosed Disease Network (UDN) cohort

465 patients who have received a molecular diagnosis



Phenotypes

Two candidate gene lists

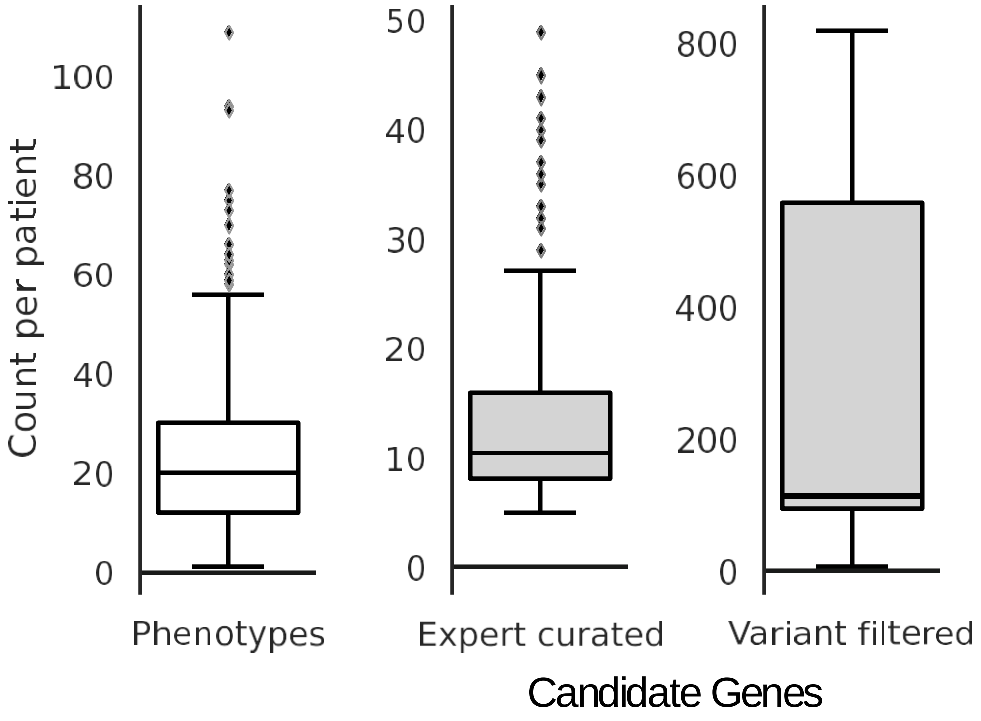
Causal gene

Disease

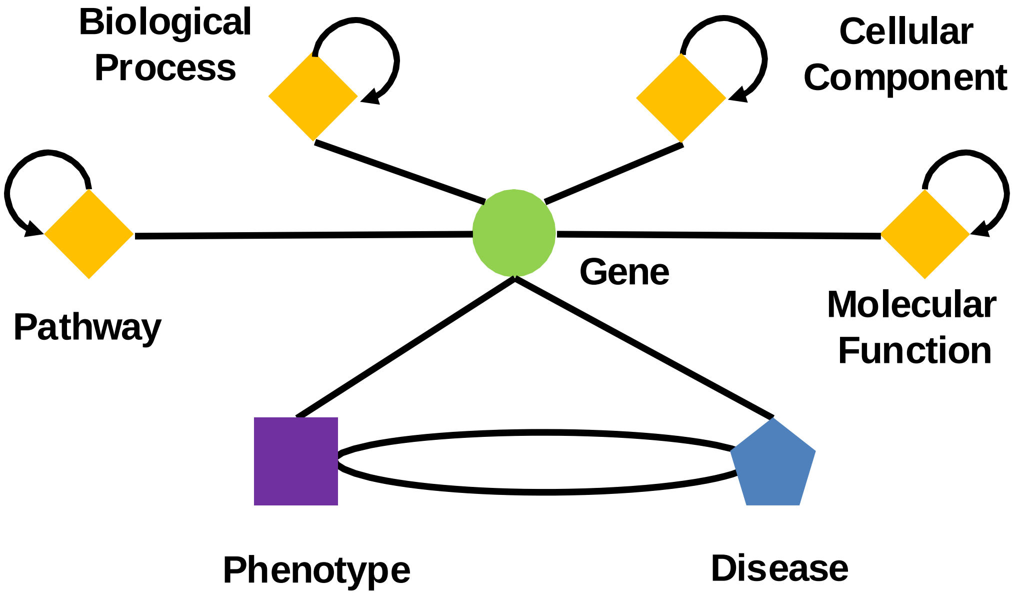
Undiagnosed Disease Network (UDN) cohort

465 patients who have received a molecular diagnosis

Number of phenotypes and candidate genes per patient



Rare disease knowledge graph (KG)



KG	# Types	Count
Nodes	7	100,272
Edges	15	2,092,690

Protein-Protein Interaction — STRING, TRANSFAC 2.0

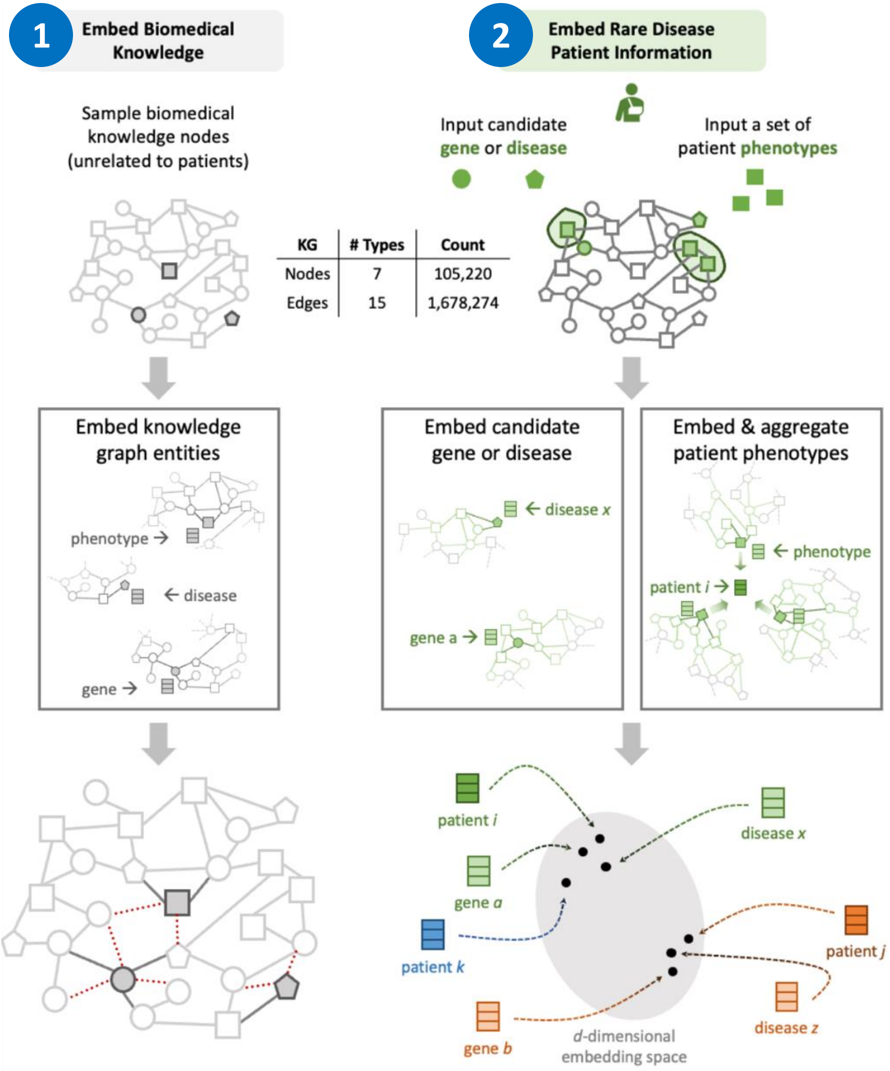
Pathway Membership — reactome

Functional Similarity — GENEONTOLOGY Unifying Biology

Phenotypic Similarity — human phenotype ontology, orphanet, OMIM Human Genetics Knowledge for the World, DisGeNET

KG Modified from zitniklab.hms.harvard.edu/projects/PrimeKG/

Knowledge graph learning


1
Embed Biomedical Knowledge
2
Embed Rare Disease Patient Information
Sample biomedical knowledge nodes (unrelated to patients)
Input candidate gene or disease
Input a set of patient phenotypes
Embed knowledge graph entities
phenotype →
← disease
gene →
Embed candidate gene or disease
← disease x
gene a →
Embed & aggregate patient phenotypes
← phenotype
patient i →
patient i
disease x
gene a
patient k
patient j
disease z
gene b
d-dimensional embedding space

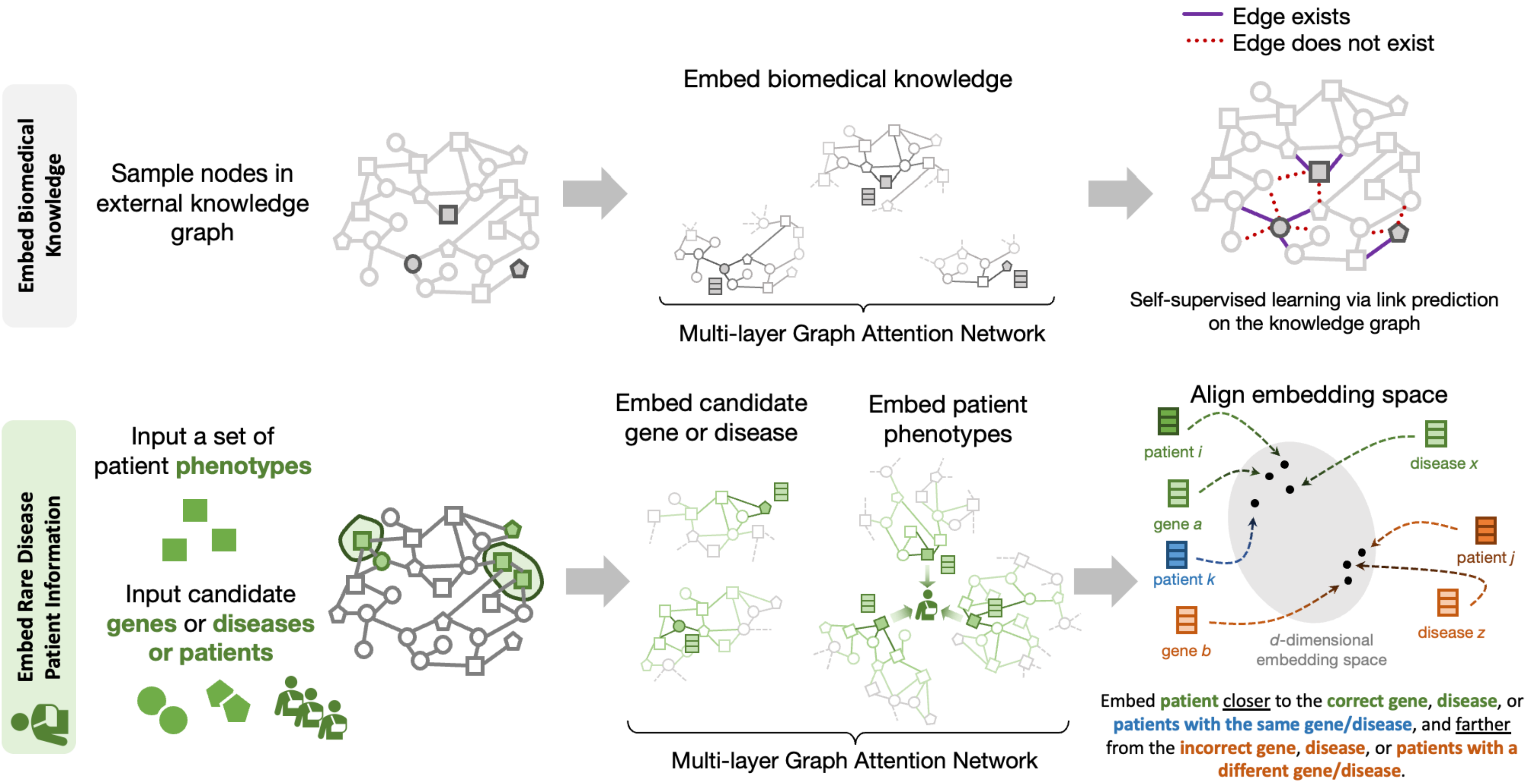
KG	# Types	Count
Nodes	7	105,220
Edges	15	1,678,274

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

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

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

SHEPHERD’s model


Edge exists
Edge does not exist
Embed Biomedical Knowledge
Sample nodes in external knowledge graph
Embed biomedical knowledge
Multi-layer Graph Attention Network
Self-supervised learning via link prediction on the knowledge graph
Embed Rare Disease Patient Information
Input a set of patient phenotypes
Input candidate genes or diseases or patients
Embed candidate gene or disease
Embed patient phenotypes
Multi-layer Graph Attention Network
Align embedding space
patient *i*
disease *x*
gene *a*
patient *j*
patient *k*
gene *b*
disease *z*
d-dimensional embedding space

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

Experimental setup

https://undiagnosed.hms.harvard.edu

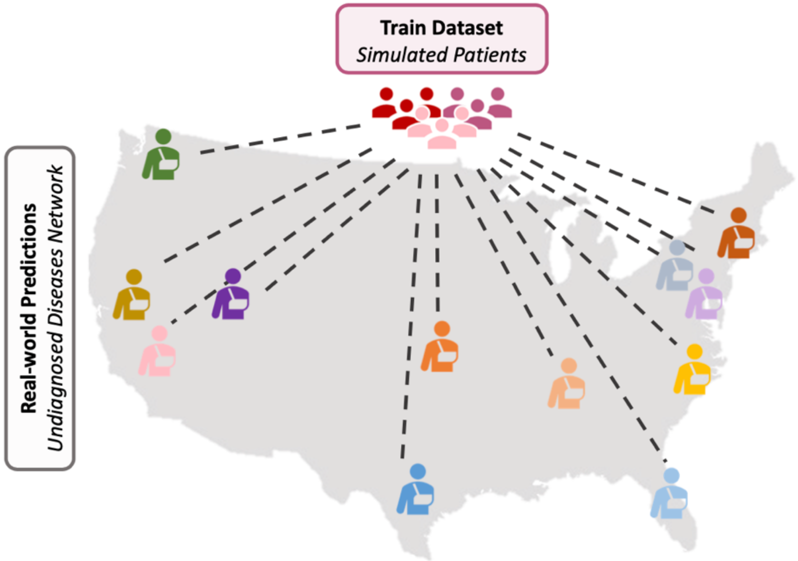
Undiagnosed Diseases Network

SHEPHERD's model training:

- 42K synthetic patients

SHEPHERD's model evaluation

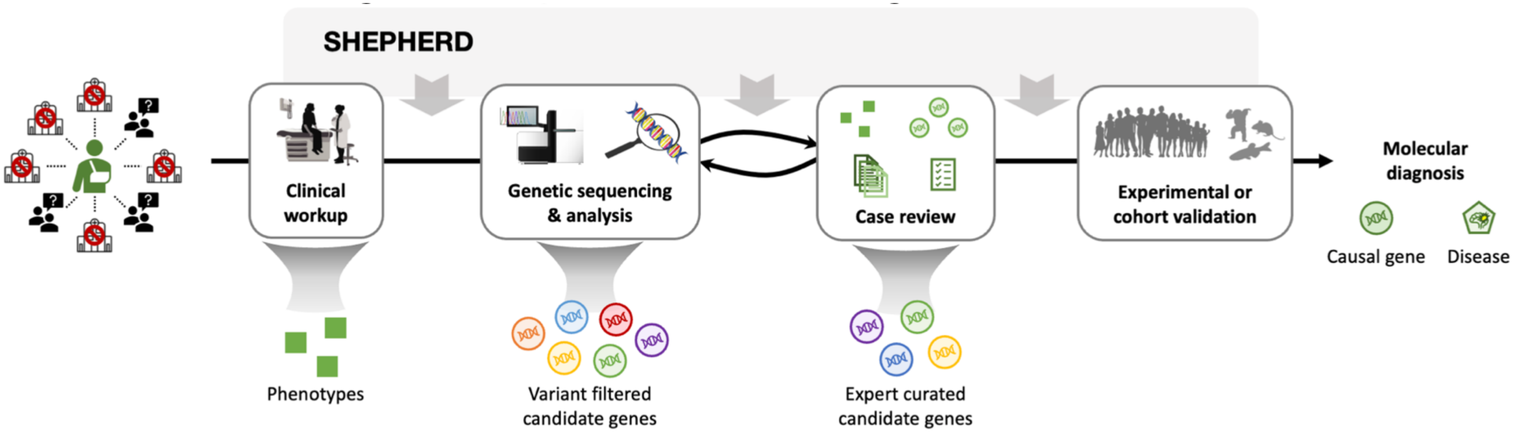
- **UDN patient cohort:** 465 rare disease patients with labeled diagnoses, spanning 299 diseases
  - 79% of genes and 83% of diseases are represented in only a single patient
- **MyGene2 patient cohort:** 146 rare disease patients, spanning 55 diseases



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

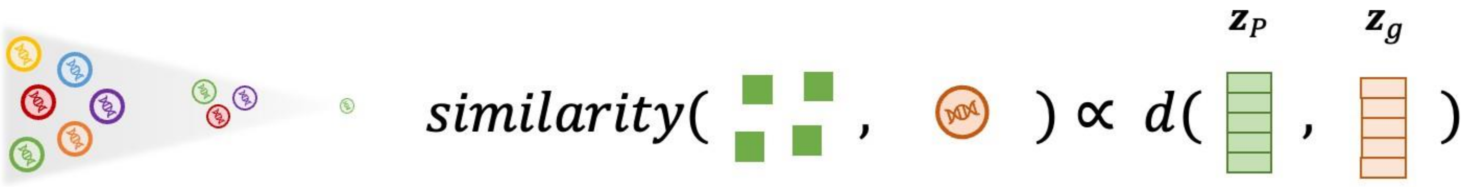
Diagnostic tasks

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

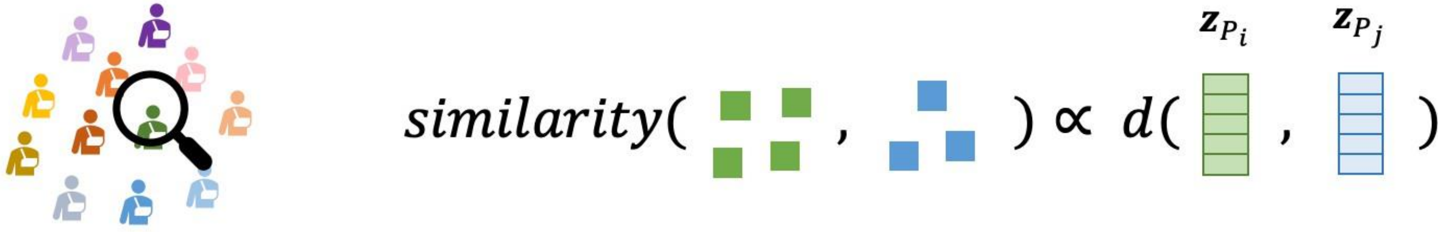

SHEPHERD
Clinical workup
Genetic sequencing & analysis
Case review
Experimental or cohort validation
Molecular diagnosis
Causal gene
Disease
Phenotypes
Variant filtered candidate genes
Expert curated candidate genes

Diagnostic tasks

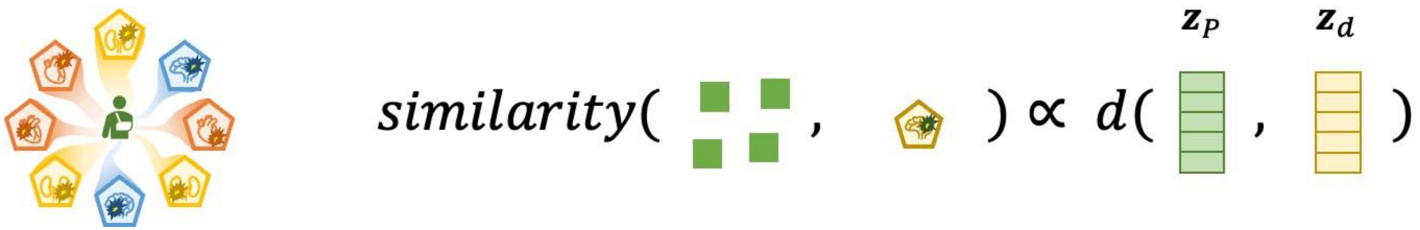
Causal Gene Discovery



Patients-like-me



Novel disease characterization



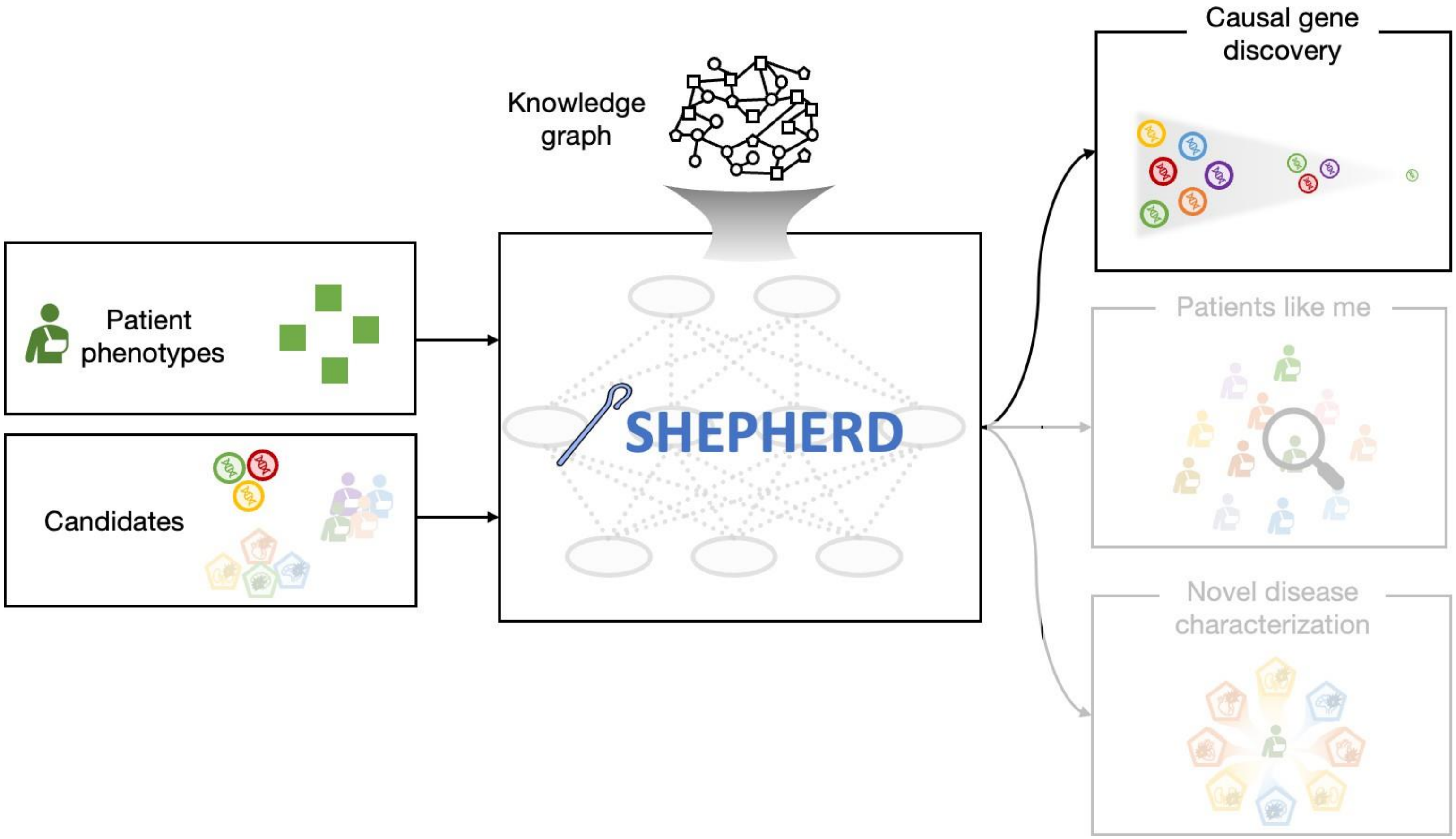
Legend

■ Patient phenotypes P

Gene g

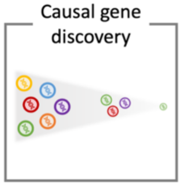
Disease d

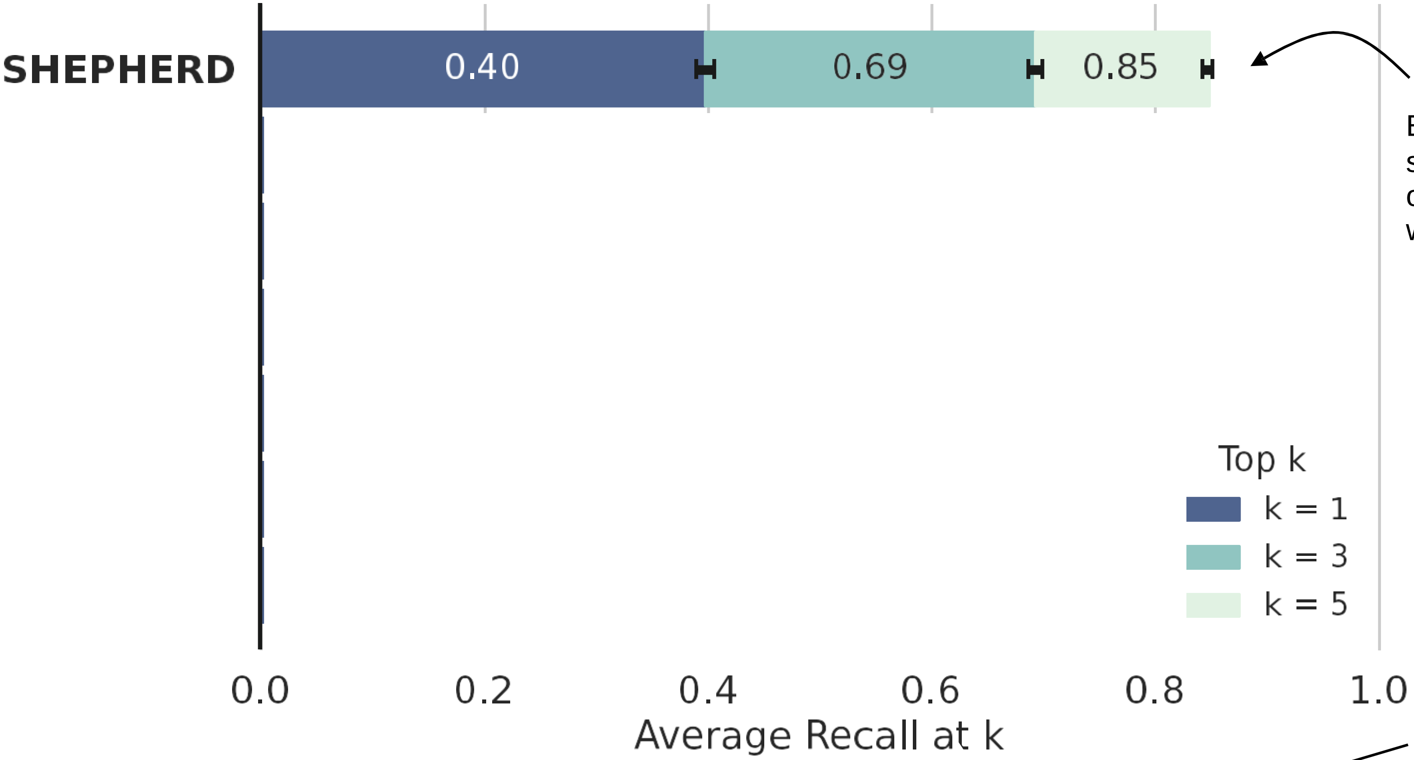
Causal gene discovery: Results



Given a patient's set of **phenotypes** and a list of **candidate genes**, which gene(s) are most likely to explain the patient's presenting symptoms?

Causal gene discovery: Results





Error bars denote standard deviation over models trained with 5 random seeds

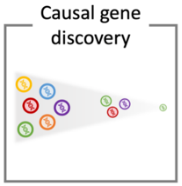
# of causal genes retrieved in top k ranked genes on average

* LR = logistic regression

† Jagadeesh et al. Phrank measures phenotype sets similarity to greatly improve Mendelian diagnostic disease prioritization. Genetics in Medicine.

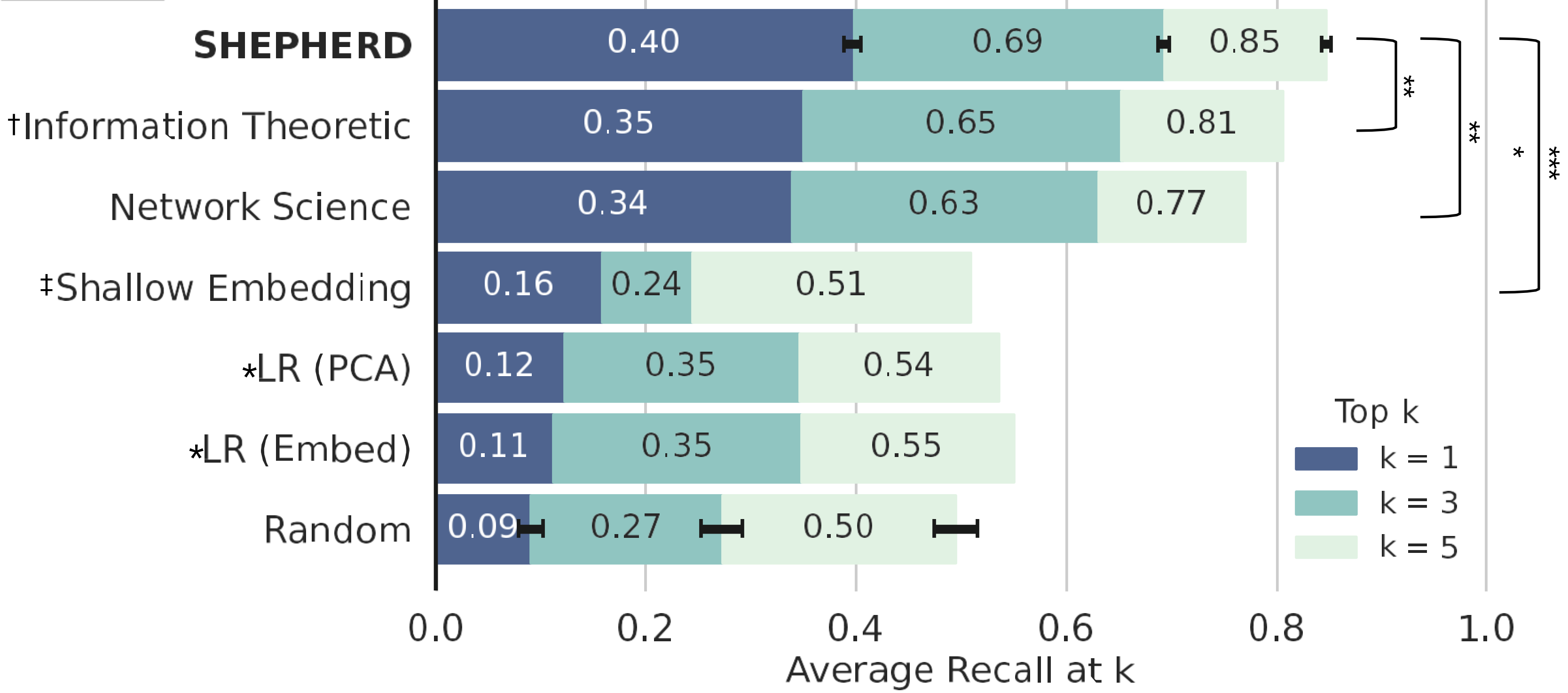
‡ Peng et al. CADA: phenotype-driven gene prioritization based on a case-enriched knowledge graph. NAR Genom Bioinform.

Causal gene discovery: Results



** p-value < 0.005

**** p-value < 0.00005

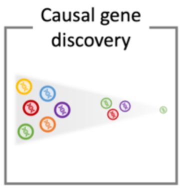


⋆ LR = logistic regression

[†] Jagadeesh et al. Phrank measures phenotype sets similarity to greatly improve Mendelian diagnostic disease prioritization. Genetics in Medicine.

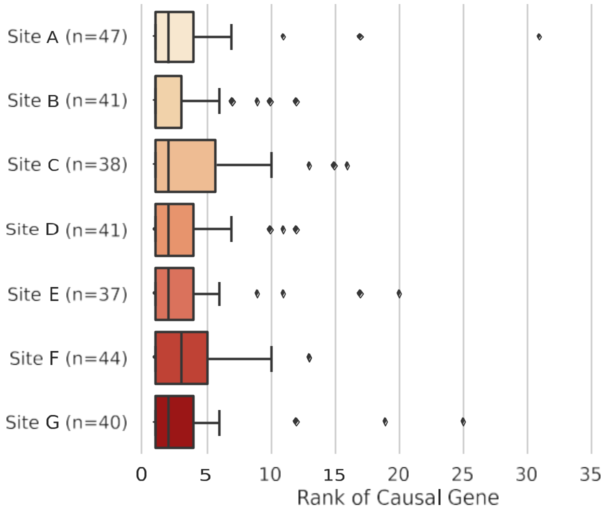
[‡] Peng et al. CADA: phenotype-driven gene prioritization based on a case-enriched knowledge graph. NAR Genom Bioinform.

Causal gene discovery: Results

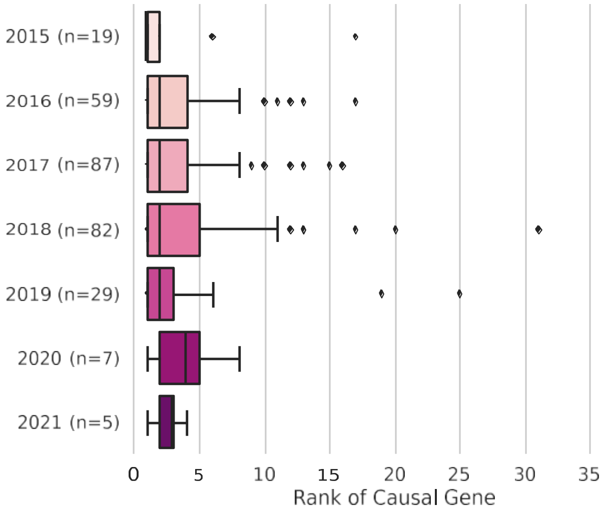


SHEPHERD generalizes across…

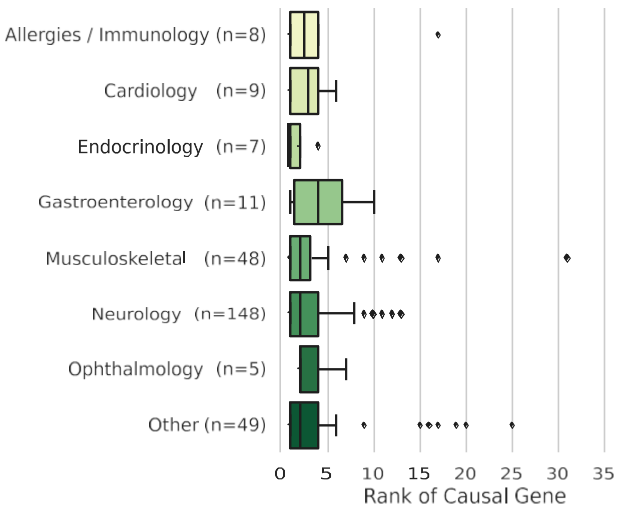
Performance by Clinical Site



Performance by Evaluation Year



Performance by Primary Symptoms



Atypical disease presentation

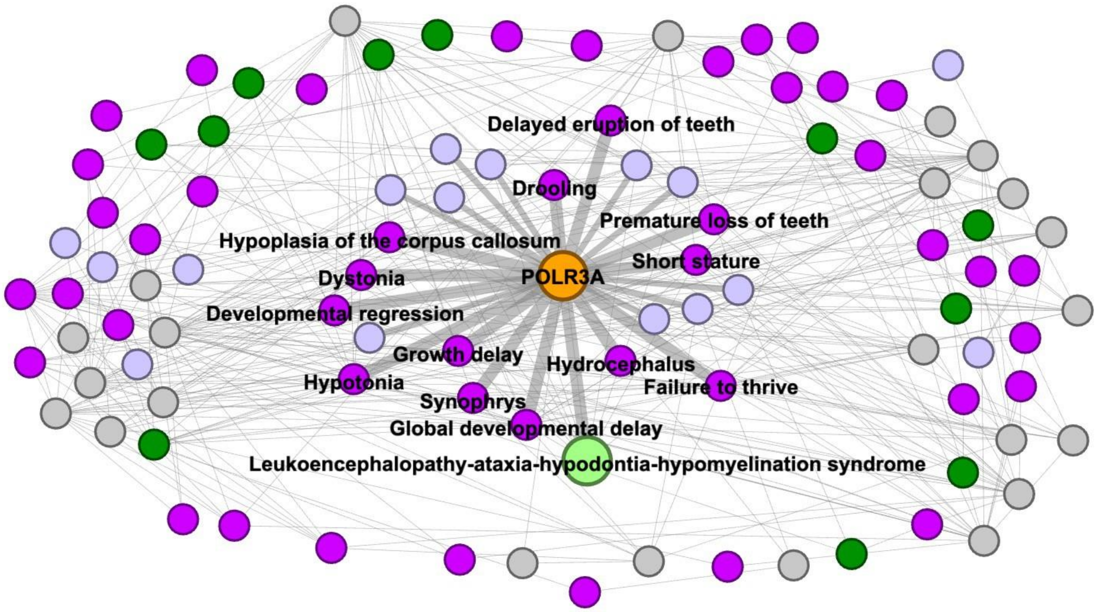
Patient: UDN-1
Admitted: 2016 **Diagnosed**: 2019
Causal gene: *POLR3A*
Disease: POLR3-Related Leukodystrophy
Atypical Phenotypes: – lack of tear production, premature adrenarche, laryngeal cleft, hearing loss, and high blood pressure

Legend
- ● Patient phenotype
- ● Causal gene
- ● Disease
- ● Related phenotype
- ● Non-causal gene
- ● Related disease

Only **28.3%** of the patient's 46 phenotypes are directly connected to *POLR3A*

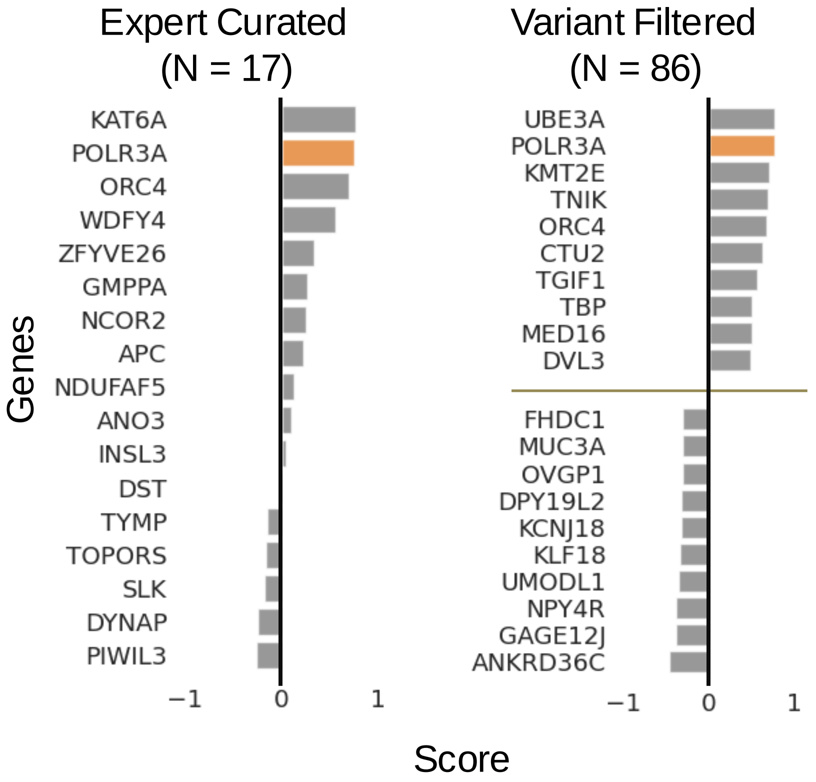
94% of the 205 phenotypes directly connected to *POLR3A* are <u>not</u> associated with the patient

Subset of Rare Disease Knowledge Graph



Delayed eruption of teeth
Drooling
Premature loss of teeth
Hypoplasia of the corpus callosum
Short stature
Dystonia
POLR3A
Developmental regression
Growth delay
Hydrocephalus
Hypotonia
Failure to thrive
Synophrys
Global developmental delay
Leukoencephalopathy-ataxia-hypodontia-hypomyelination syndrome

Atypical disease presentation



		Attention	Phenotype (N = 46)
Top 10 phenotypes		0.037	Short stature
		0.034	Failure to thrive
		0.033	Central hypotonia
		0.032	Microcephaly
		0.032	Prominent eyelashes
		0.032	Respiratory insufficiency
		0.031	Gastrostomy tube feeding in infancy
		0.031	Chronic lung disease
		0.028	Ventriculomegaly
		0.027	Growth delay
		…	…
Bottom 10 phenotypes	★	0.014	Alacrima
		0.014	Premature loss of primary teeth
	★	0.013	Moderate sensorineural hearing impairment
		0.013	Pancreatitis
		0.012	Abnormal sternum morphology
		0.011	T2 hypointense basal ganglia
		0.011	Febrile seizure (within the age range of 3 months to 6 years)
		0.009	Chronic pancreatitis
	★	0.006	Laryngeal cleft
		0.0003	T2 hypointense brainstem

Results: Patients-like-me

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

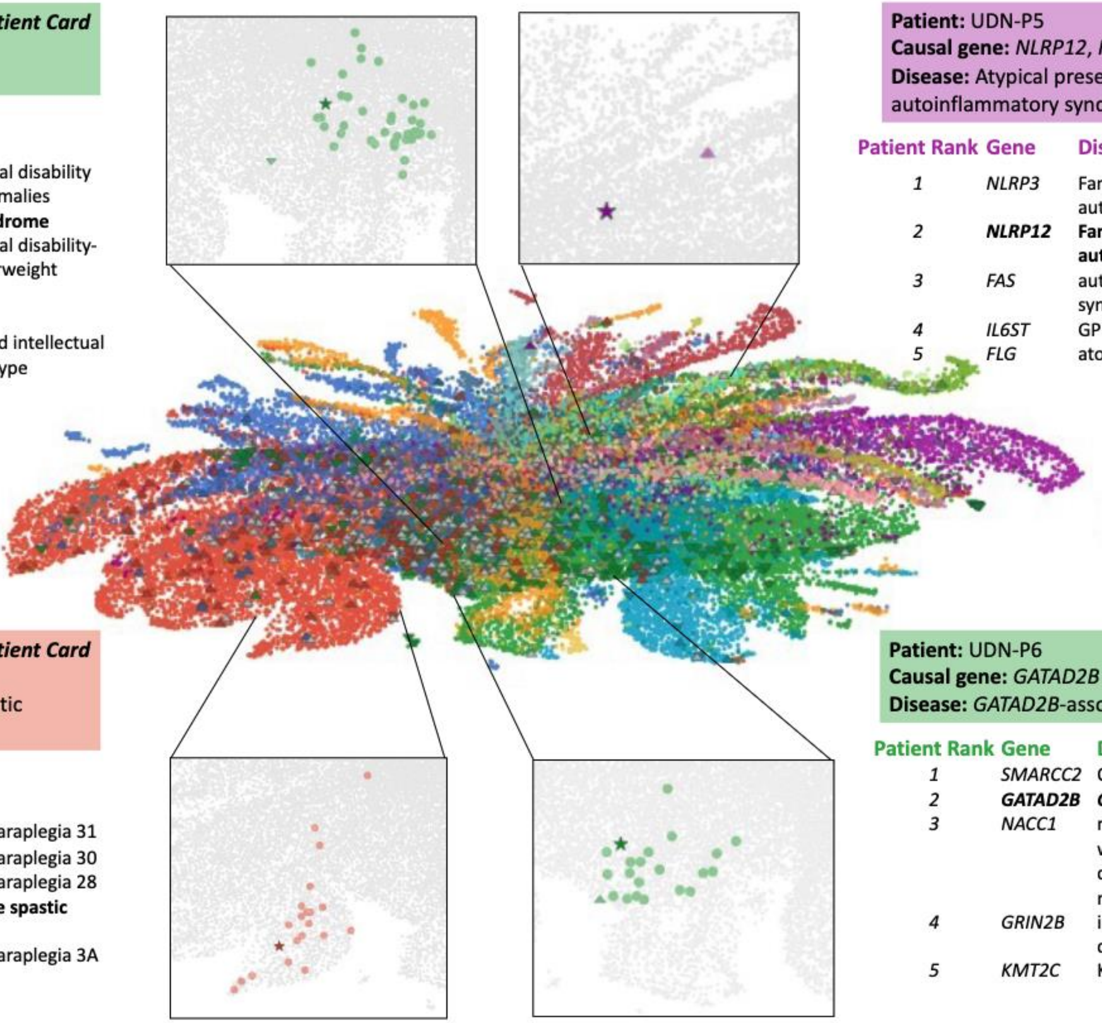
a

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

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

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

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



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

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

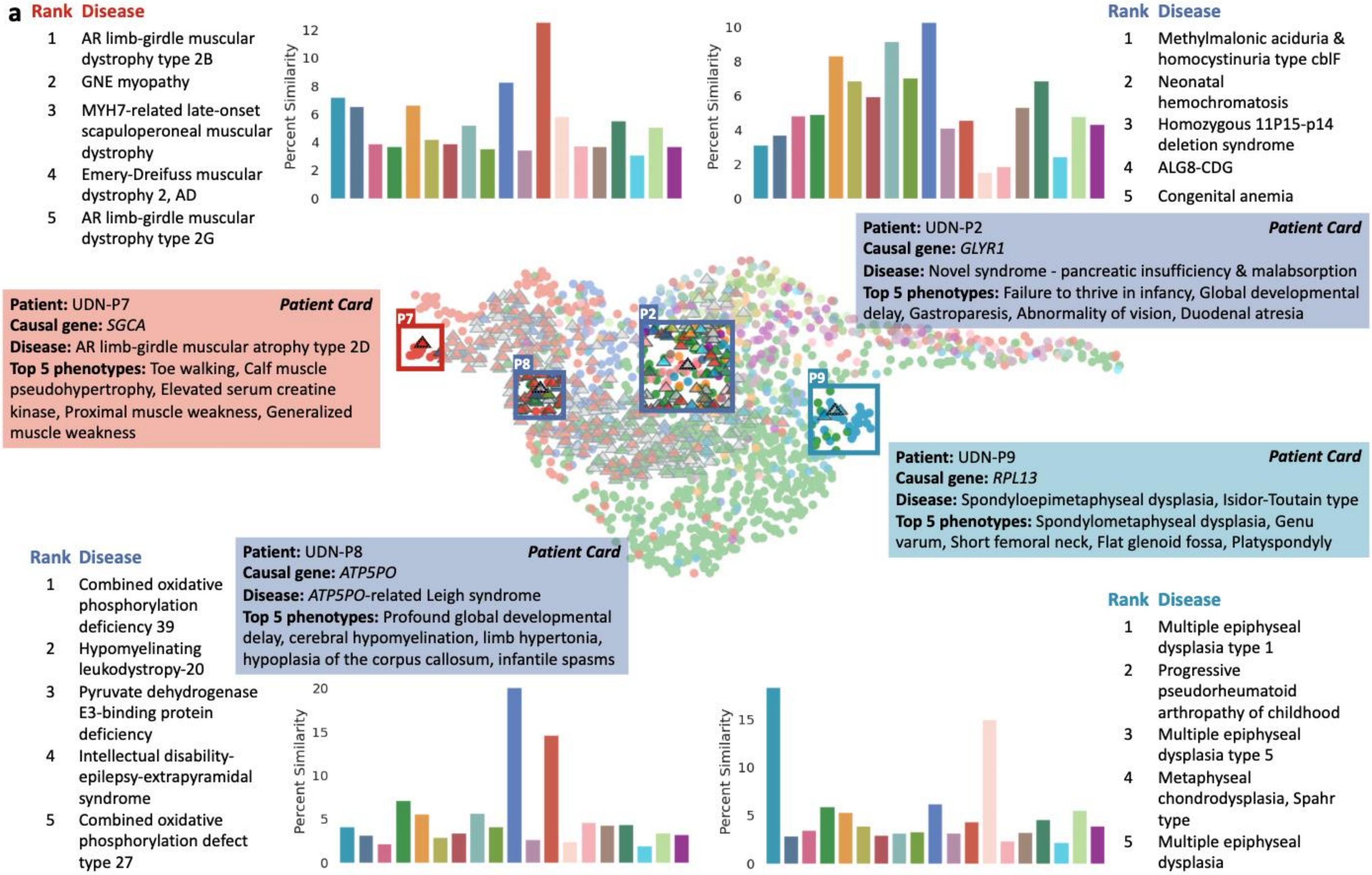
Patient: UDN-P6 ***Patient Card***
Causal gene: *GATAD2B*
Disease: *GATAD2B*-associated syndrome

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

Deep learning for diagnosing patients with rare genetic diseases, *medRxiv,* 2022

Novel disease characterization

Results: New disease naming



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

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

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

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

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

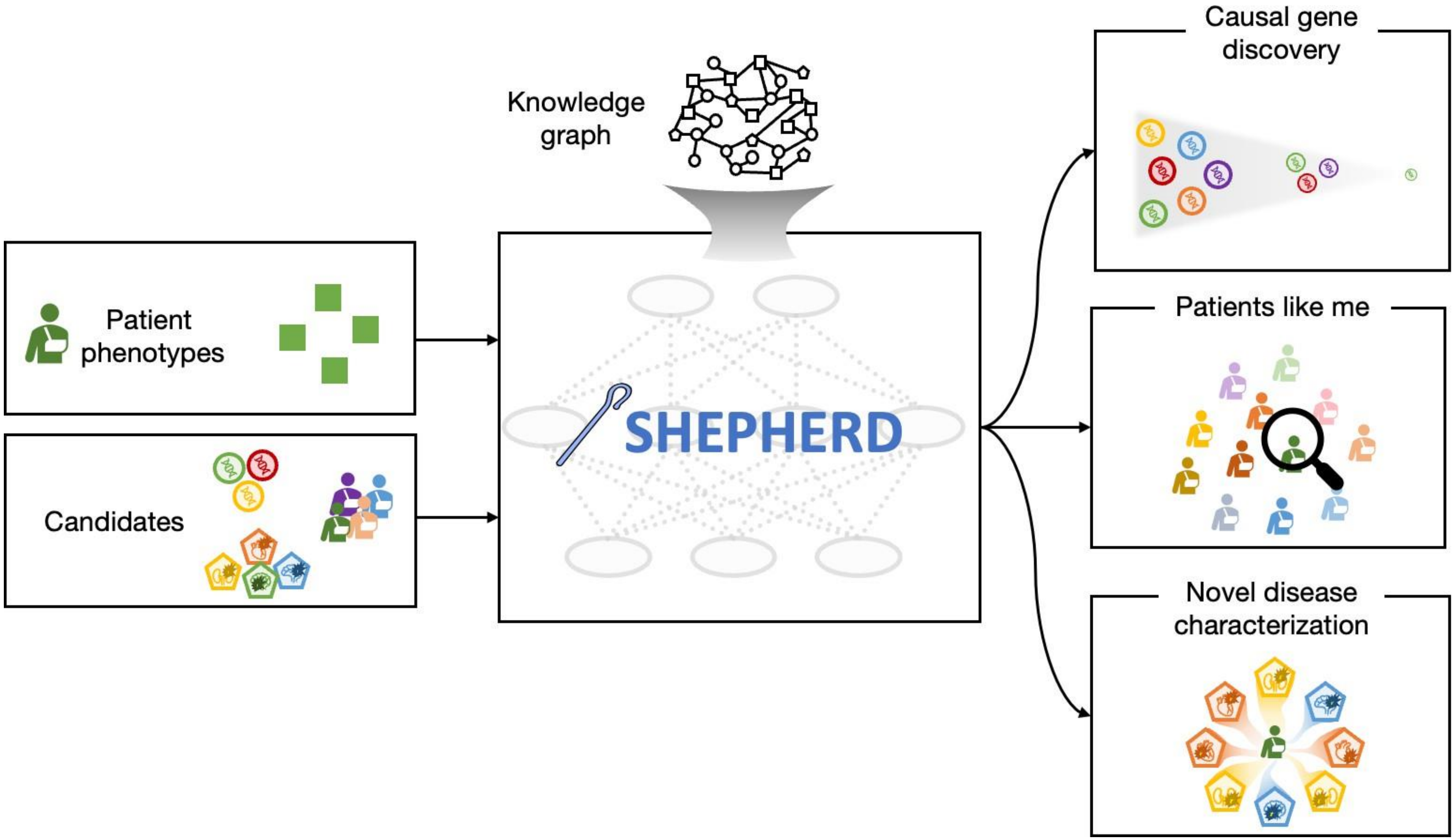
Rank	Disease
1	Combined oxidative phosphorylation deficency 39
2	Hypomyelinating leukodystropy-20
3	Pyruvate dehydrogenase E3-binding protein deficiency
4	Intellectual disability-epilepsy-extrapyramidal syndrome
5	Combined oxidative phosphorylation defect type 27

Patient: UDN-P8 ***Patient Card***
Causal gene: *ATP5PO*
Disease: *ATP5PO*-related Leigh syndrome
Top 5 phenotypes: Profound global developmental delay, cerebral hypomyelination, limb hypertonia, hypoplasia of the corpus callosum, infantile spasms

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

Deep learning for diagnosing patients with rare genetic diseases, *medRxiv*, 2022

SHEPHERD: KG-based AI for rare disease diagnosis



Given a patient's set of **phenotypes** and a list of **candidate genes**, which gene(s) are most likely to explain the patient's presenting symptoms?

Given a patient's set of **phenotypes** and a **cohort of rare disease patients**, can we identify other patients with similar genetic conditions?

Given a patient's set of **phenotypes** and a set of **known diseases**, can we characterize the clinical presentation based on our current knowledge of rare diseases?

Take-away messages

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

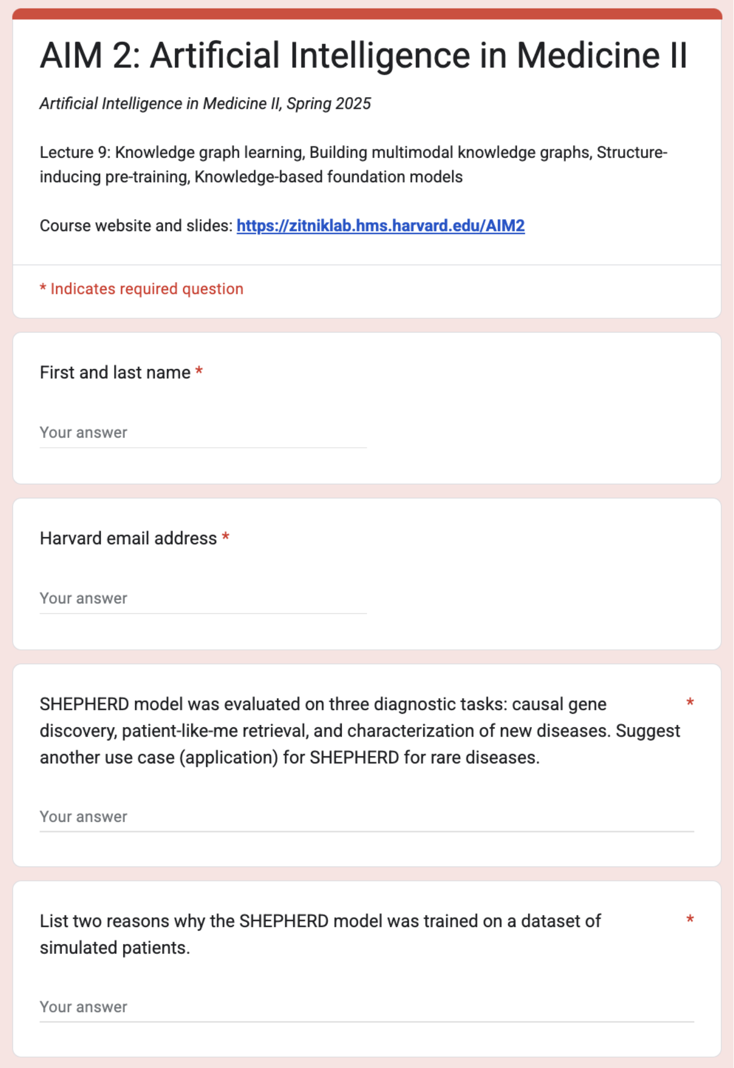
First deep learning approach for individualized diagnosis of rare genetic diseases

Graph learning approach is not only helpful but necessary

Deep learning for diagnosing patients with rare genetic diseases, *medRxiv*, 2022

Quick check

https://forms.gle/AfRT7pdXGa7MoJxJA



AIM 2: Artificial Intelligence in Medicine II

Artificial Intelligence in Medicine II, Spring 2025

Lecture 9: Knowledge graph learning, Building multimodal knowledge graphs, Structure-inducing pre-training, Knowledge-based foundation models

Course website and slides: https://zitniklab.hms.harvard.edu/AIM2

* Indicates required question

First and last name *

Your answer

Harvard email address *

Your answer

SHEPHERD model was evaluated on three diagnostic tasks: causal gene discovery, patient-like-me retrieval, and characterization of new diseases. Suggest another use case (application) for SHEPHERD for rare diseases. *

Your answer

List two reasons why the SHEPHERD model was trained on a dataset of simulated patients. *

Your answer

Towards foundation models for knowledge graphs

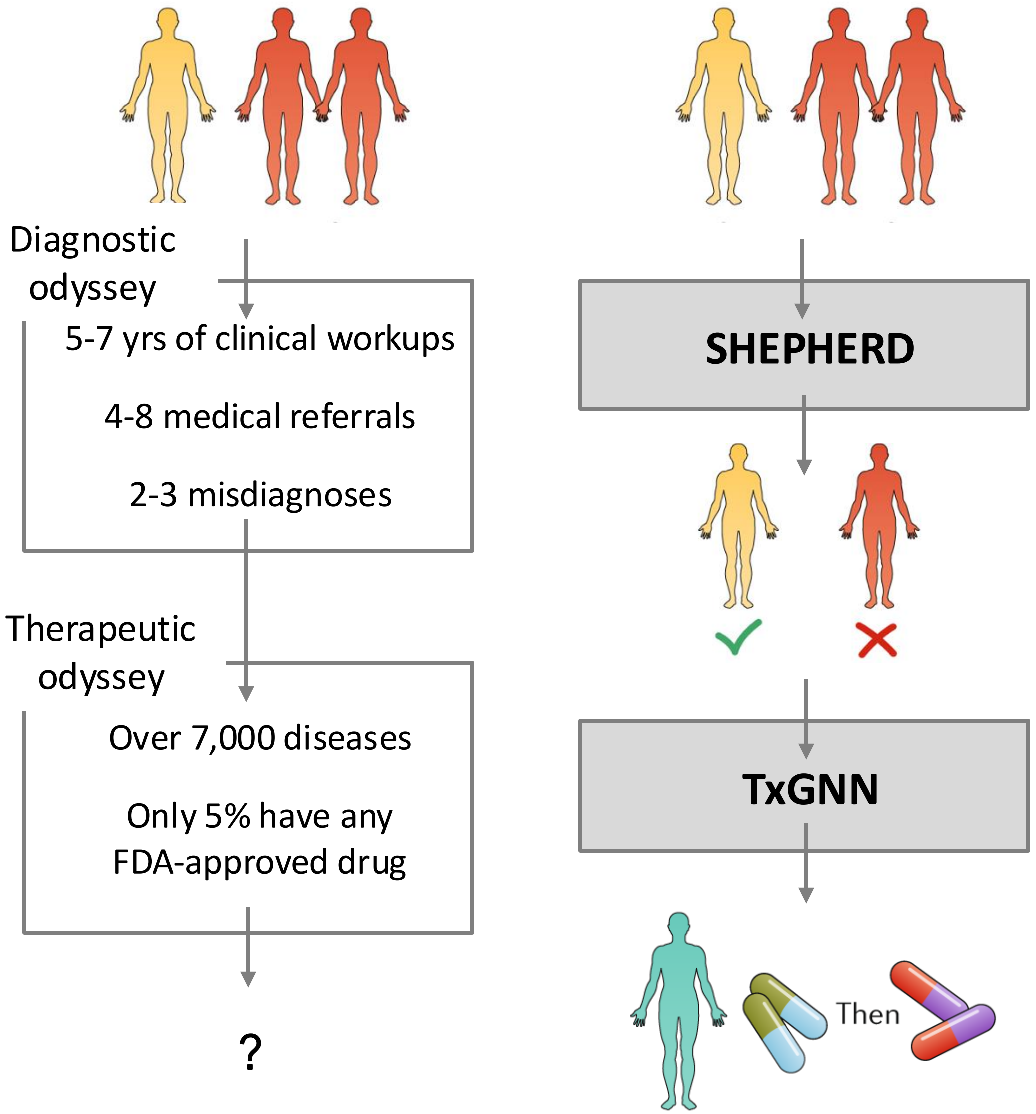
Future with AI: From mysteries to therapies

Knowledge graph models for diagnosing rare disease patients

SHEPHERD: Deep learning for diagnosing patients with rare genetic diseases, medRxiv 2025

Knowledge graph models for universal drug repurposing

TxGNN: A foundation model for clinician-centered drug repurposing, *Nature Medicine* 2024



Diagnostic odyssey

5-7 yrs of clinical workups

4-8 medical referrals

2-3 misdiagnoses

Therapeutic odyssey

Over 7,000 diseases

Only 5% have any FDA-approved drug

?

SHEPHERD

✓ ✗

TxGNN

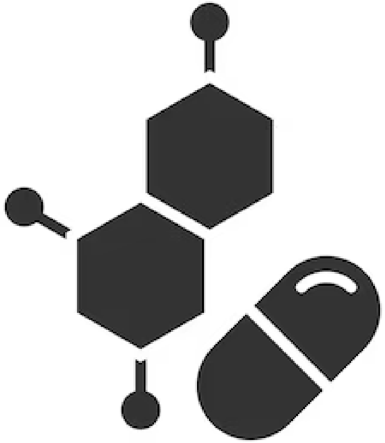
Then

Precision medicine (treatments)

Measure phenotype and mechanisms

Design therapeutic agents or select optimal perturbations

+



→

Provide each patient with the right drug, at the right dose, at the right time

Clinical phenotypes and diseases

17,000	Diseases
7,000	Rare diseases
5-7%	Rare diseases with treatments
No	Treatment options for many disease subtypes

Medicines and drugs

40-50	New molecules per year
30%	Drugs are issued at least one post-approval new indication
Many	Drugs have accrued over 10 drug indications over the years

Matching drugs to clinical outcomes across thousands of diseases

Drug repurposing as an effective drug development strategy for many diseases

1 No effective treatments for rare and even many complex diseases:

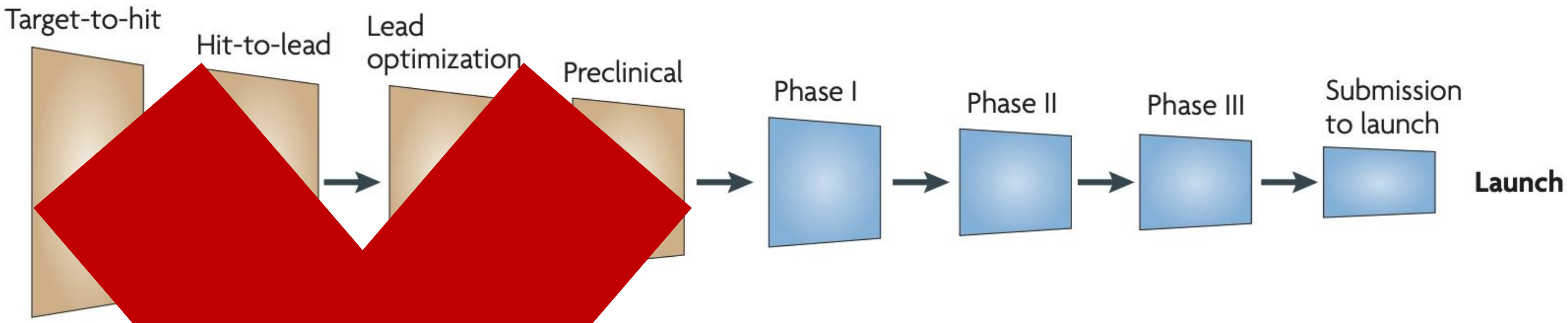
- Over 7,000 rare diseases affect 300-400 million people worldwide. Only 5% of rare diseases have FDA-approved drugs
- Even for diseases with approved treatments, new drugs can offer alternative options that cause fewer side effects and replace drugs that are ineffective for patient subpopulations

2 Faster translation to the clinic and lower development costs

- 30% of drugs approved were issued at least one post-approval new indication. Many drugs have accrued over 10 indications over years
- Most repurposed drugs are the results of serendipity (luck is not a strategy!)

Zero-shot drug repurposing with geometric deep learning and clinician centered design, *Nature Medicine* (in press) 2024

Phases of drug discovery from initial stage (target-to-hit) to final stage (launch)



	Target-to-hit	Hit-to-lead	Lead optimization	Preclinical	Phase I	Phase II	Phase III	Submission to launch	Launch
p(TS)	80%	[illegible]	[illegible]	69%	54%	34%	70%	91%	
***WIP* needed for 1 launch**	24.3	[illegible]	[illegible]	12.4	8.6	4.6	1.6	1.1	1
Cost per WIP per Phase	$1	[illegible]	[illegible]	$5	$15	$40	$150	$40	
Cycle time (years)	1.0	[illegible]	[illegible]	1.0	1.5	2.5	2.5	1.5	
Cost per launch (out of pocket)	$2[illegible]	[illegible]	[illegible]	[illegible]2	$128	$185	$235	$44	$873
% Total cost per NME	[illegible]	[illegible]	[illegible]	[illegible]	15%	21%	27%	5%	
Cost of capital	11[illegible]								
Cost per launch (capitalized)	$94	[illegible]166	$414	$150	$273	$319	$314	$48	[illegible]

▨ Discovery ▨ Development

p(TS) – probability of successful transition from one stage to the next; NME – new molecular entity; WIP – work in process

All-disease model for drug repurposing

Biomedical data span multiple scales and multiple data modalities

Transcriptomics

Physical contacts

Molecular pathways and patient subtypes

Treatment information

TxGNN:
All-disease drug repurposing model

Once trained, models are adapted to an array of tasks, with no or minimal training

What patient populations will respond to treatment?

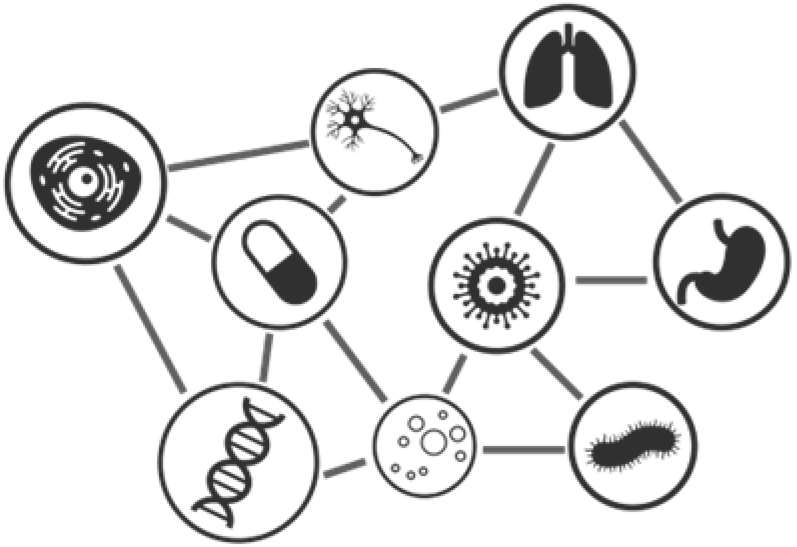
What candidate therapeutics will have an acceptable safety profile for patients with metastatic melanoma?

What small-molecule compounds will inhibit a kinase?

A foundation model for clinician-centered drug repurposing, *Nature Medicine* 2024

All-disease model for drug repurposing

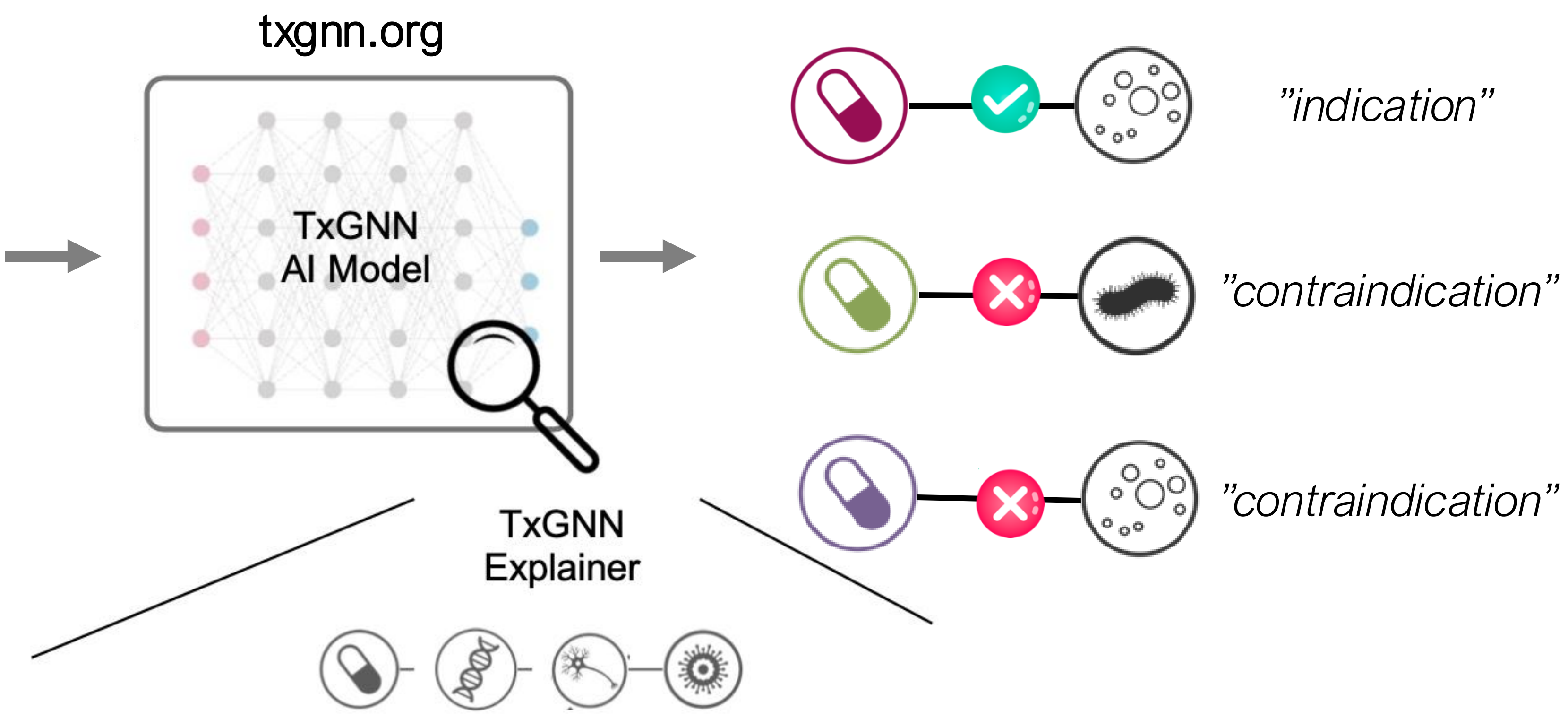
Multimodal knowledge graph of 17,080 disease phenotypes



Semi-automatic KG rebuild when new datasets become available

Building a knowledge graph to enable precision medicine, Scientific Data 2023

Process therapeutic tasks and predict candidate indications and contraindications

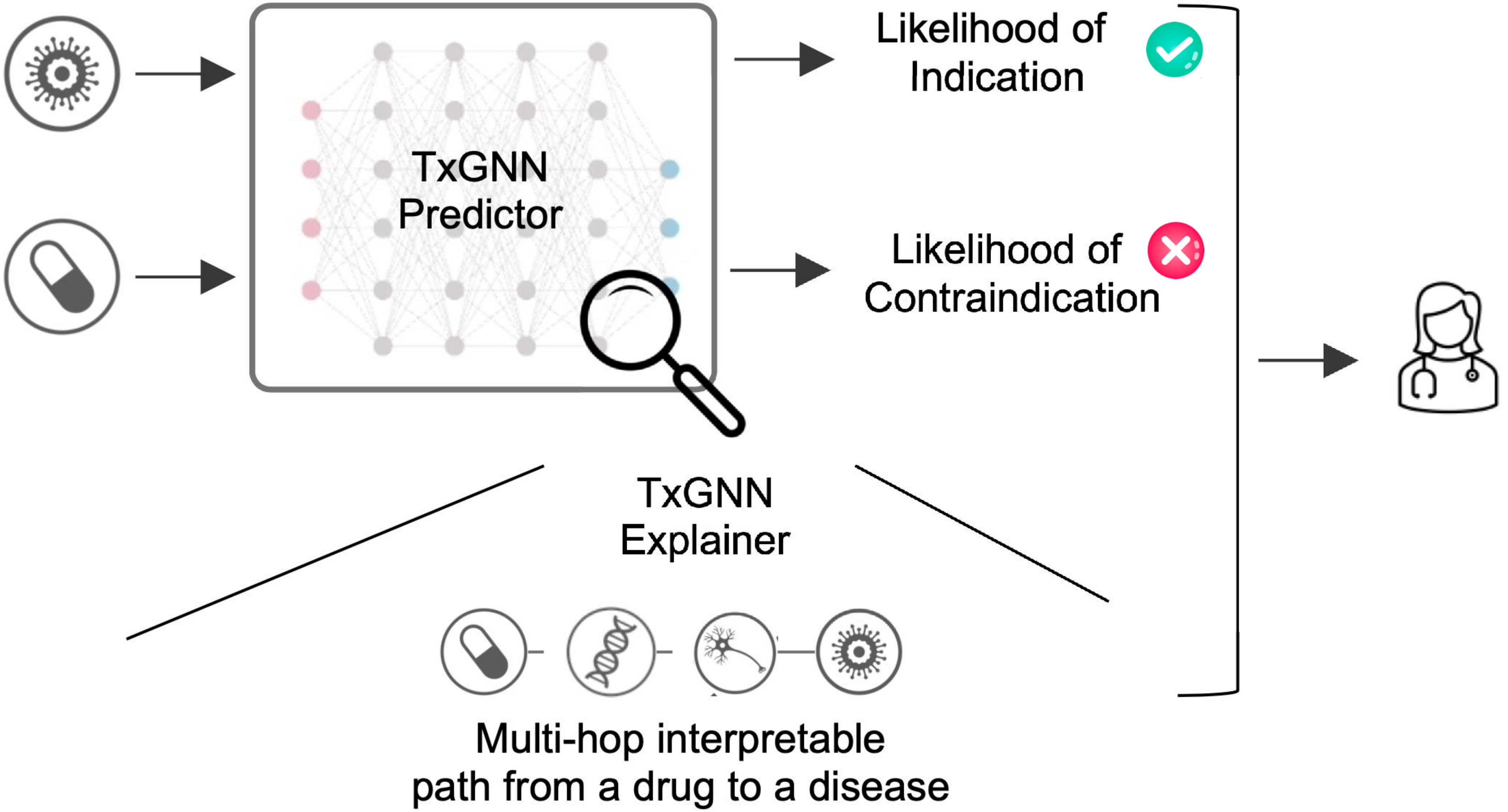


txgnn.org

Mechanistic path from drug to disease

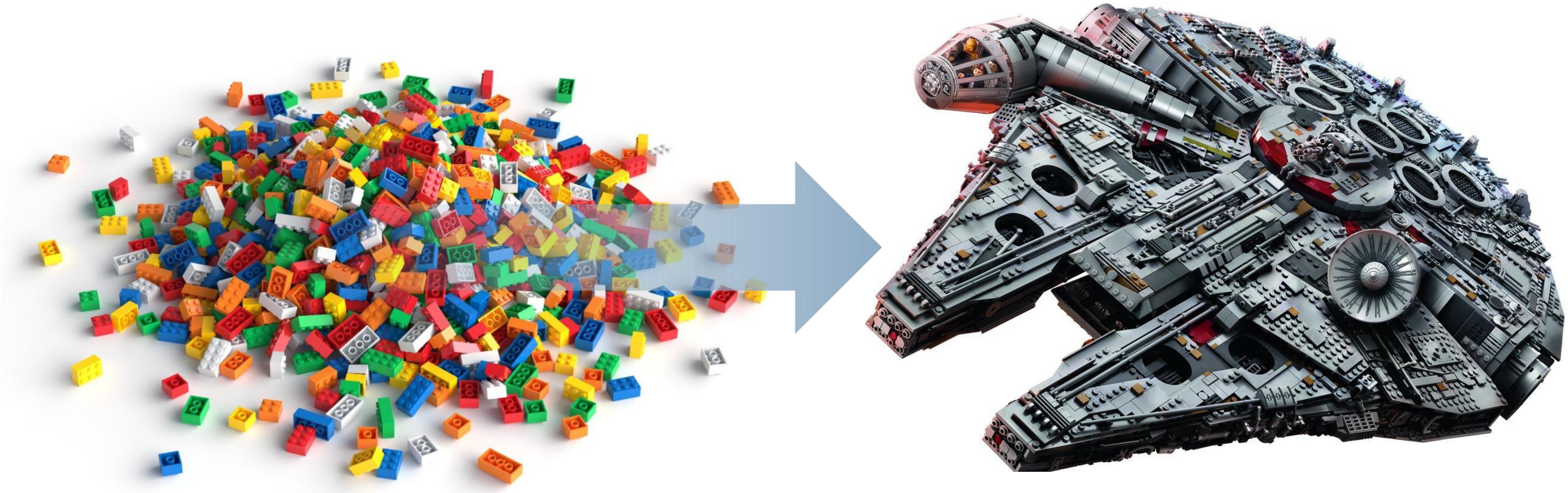
Structure-inducing pre-training, *Nature Machine Intelligence* 2023; Multimodal learning with graphs, *Nature Machine Intelligence* 2023; Graph Representation Learning in Biomedicine and Healthcare, *Nature Biomedical Engineering* 2022; Multimodal Learning with Graphs, *Nature Machine Intelligence* 2023; A foundation model for clinician-centered drug repurposing, *Nature Medicine* 2024

All-disease model for drug repurposing



A foundation model for clinician-centered drug repurposing, *Nature Medicine* 2024

Building knowledge graphs: Medical data are multimodal and scattered across databases



VAST UNORGANIZED KNOWLEDGE

CURATED KNOWLEDGE GRAPH

Building knowledge graphs: Medical data are multimodal and scattered across databases

ChEMBL

includes 1.6M assays
covering 2.4M compounds

Bgee

includes 31,467 bulk and
single-cell RNA-seq libraries

STRING

includes 20B interactions
between 59.3M proteins

GENEONTOLOGY
Unifying Biology

includes 6M gene annotations
derived from 150K publications

reactome

includes 2,711 pathways
manually curated by PhDs

DRUGBANK

includes 17K FDA-approved
and experimental drugs

NIH National Library of Medicine
National Center for Biotechnology Information

includes annotations for 192K
human genetic elements

SIDER

includes 139K adverse
reactions for marketed drugs

human phenotype ontology

includes 13K phenotypes and
156K disease annotations

Building knowledge graphs: Medical data are multimodal and scattered across databases

ChEMBL

includes 1.6M assays
covering 2.4M compounds

ChEMBL evidence integration pipeline

repeat for all 36 databases

1. Include only drugs approved for marketing by the FDA or clinical candidates.
2. Machine learning analysis to evaluate clinical trials that ended earlier than scheduled.
3. Score all drug-disease edges by clinical precedence (*i.e.*, Phase 0, I, II, III, IV).
4. Down-weight scores for trials that stop early due to negative outcomes or safety concerns.
5. Construct typed edges in knowledge graph based on strength of ChEMBL evidence.

includes 6M gene annotations

National Library of Medicine
National Center for Biotechnology Information

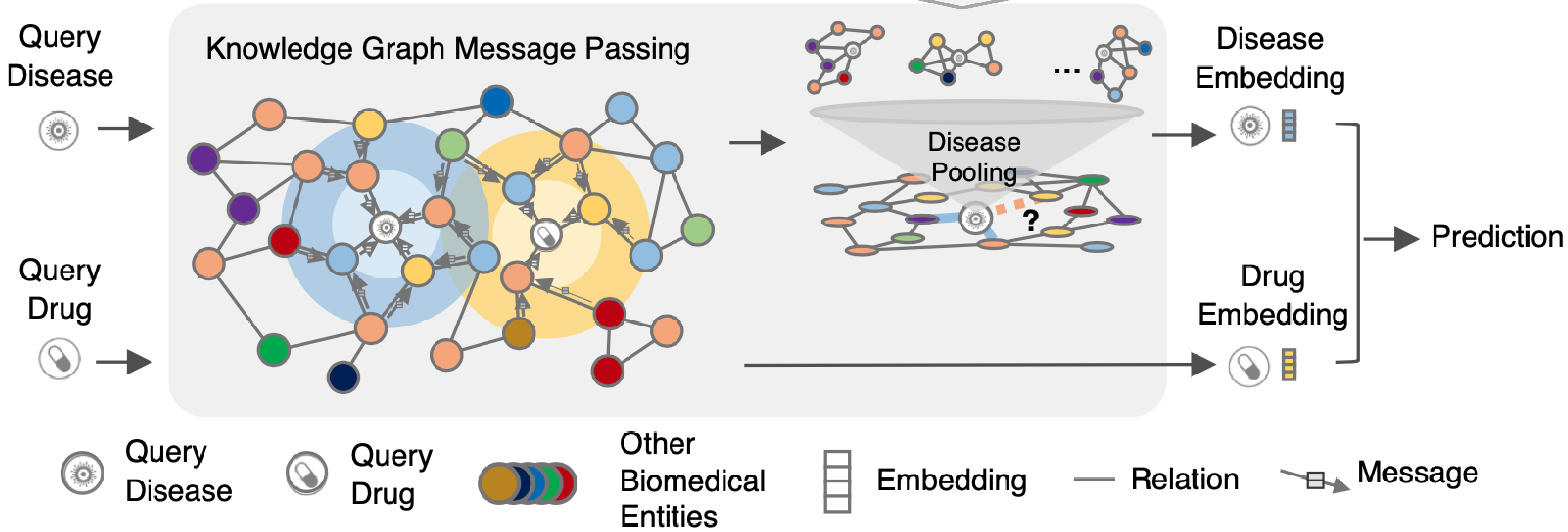
includes annotations for 192K
human genetic elements

manually curated by PhDs

and experimental drugs

Knowledge graph based TxGNN model enables transfer learning across 17,080 disease phenotypes

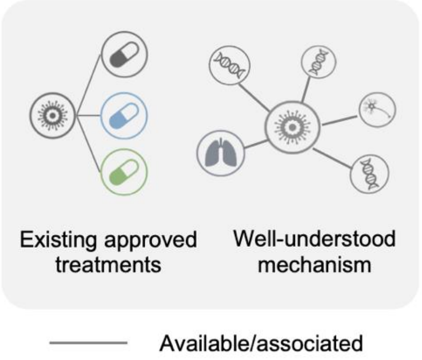
Disease pooling is a module that identifies diseases similar to a query disease and transfers information from related diseases to the query disease



A foundation model for clinician-centered drug repurposing, *Nature Medicine* 2024

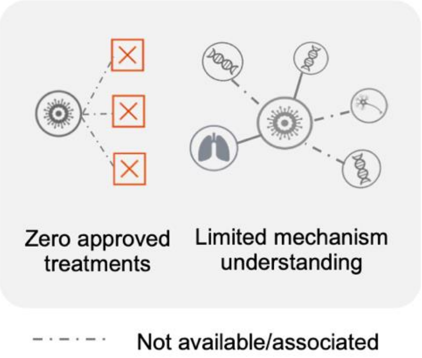
TxGNN identifies candidate drugs for diseases with no treatment options

Once trained, TxGNN can perform zero-shot prediction on new diseases without additional parameters or fine-tuning on labeled data



Scenario A: Current state-of-the-art

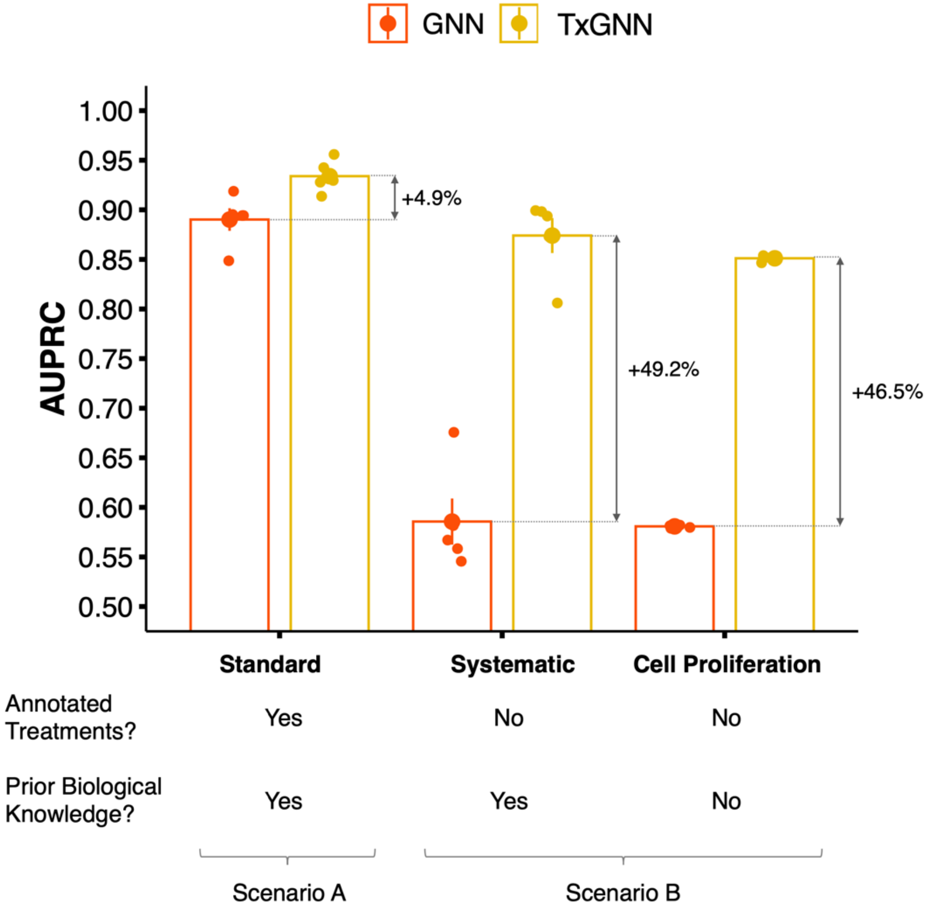
- Disease with existing treatments
- Easier to predict



Scenario B: Zero-shot prediction

- Diseases with no existing treatments
- Much harder to predict

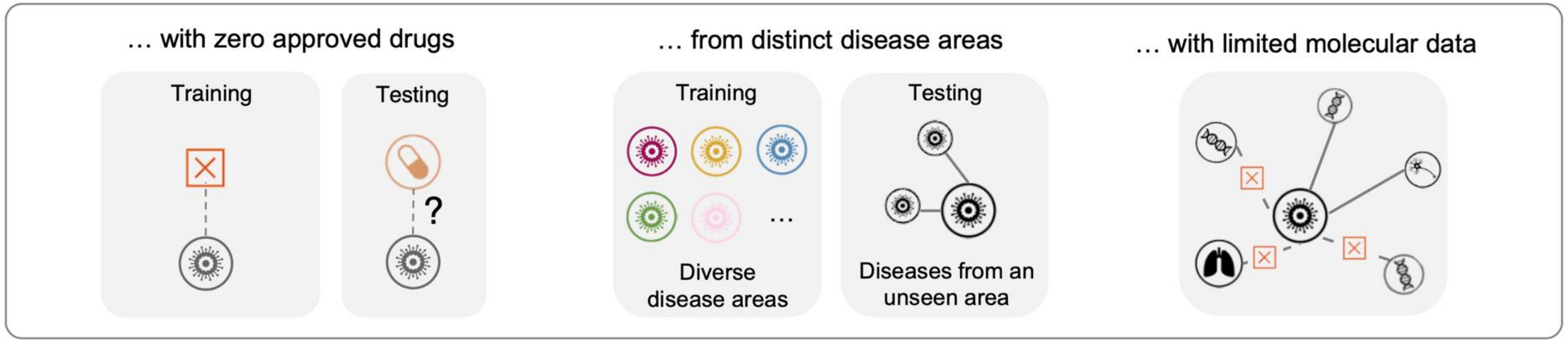
7,000+ rare diseases affect 300-400M globally; only 5% have FDA-approved drugs. New drugs can offer better, side-effect reduced options for specific patients



A foundation model for clinician-centered drug repurposing, *Nature Medicine* 2024

Benchmarking TxGNN on challenging dataset splits across disease areas

Held-out folds contain diseases …



… with zero approved drugs

Training

Testing

… from distinct disease areas

Training

Diverse disease areas

Testing

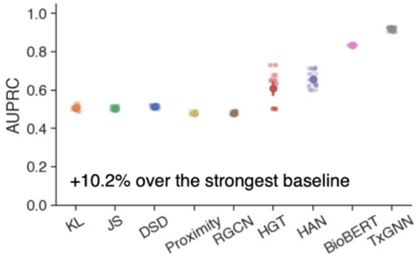
Diseases from an unseen area

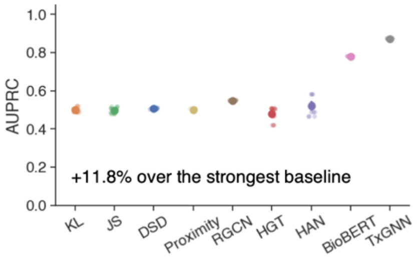
… with limited molecular data

Held-out folds contain **cancer diseases**

Diseases in this area include:

- Leydig cell tumor
- Neurofibroma
- Acute myeloid leukemia

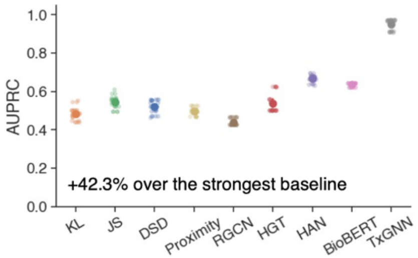


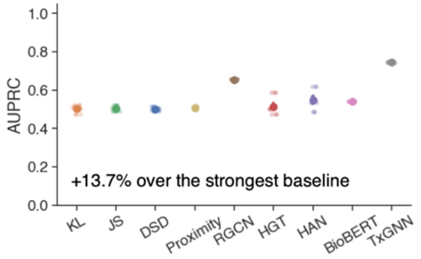


Held-out folds contain **anemia-related diseases**

Diseases in this area include:

- Thalassemia
- Aplastic anemia
- Hemoglobin C disease

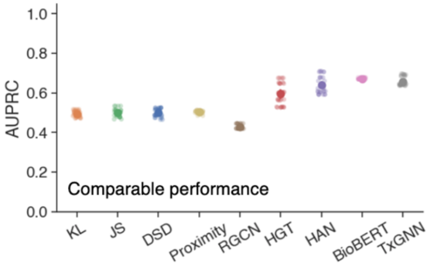


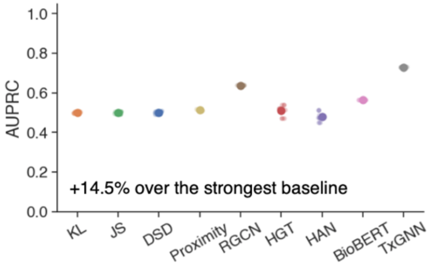


Held-out folds contain **cardiovascular diseases**

Diseases in this area include:

- Mitral valve stenosis
- Congestive heart failure
- Long QT syndrome



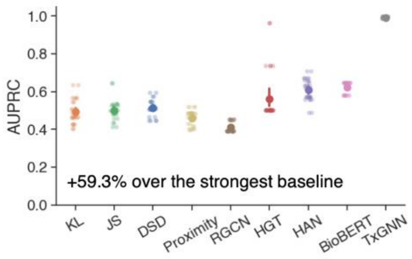


Held-out folds contain **adrenal gland diseases**

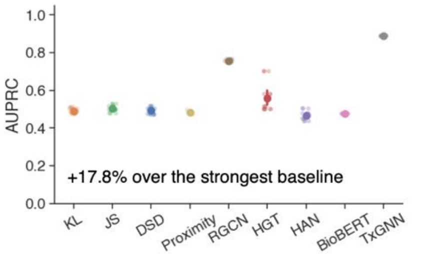
Diseases in this area include:

- Hyperaldosteronism
- Addison disease
- Ectopic cushing syndrome

Zero-shot indication prediction



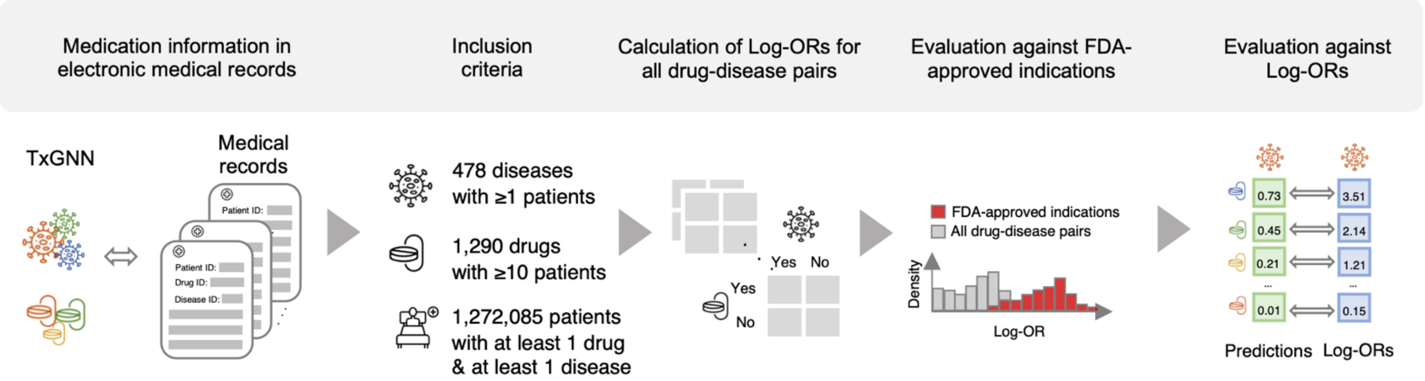
Zero-shot contraindication prediction



A foundation model for clinician-centered drug repurposing, *Nature Medicine* 2024

Evaluating new drug repurposing predictions

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



Medication information in electronic medical records

Inclusion criteria

Calculation of Log-ORs for all drug-disease pairs

Evaluation against FDA-approved indications

Evaluation against Log-ORs

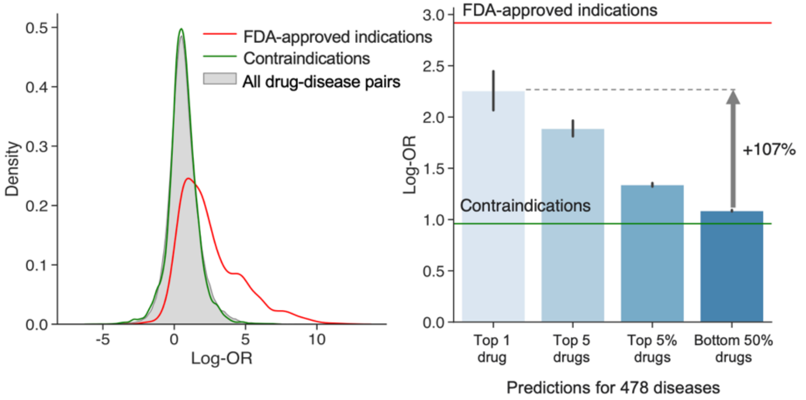
TxGNN

Medical records

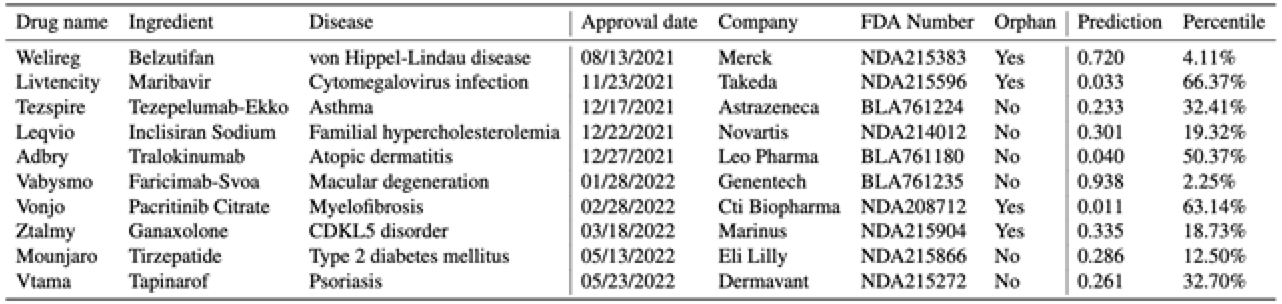
478 diseases with ≥1 patients

1,290 drugs with ≥10 patients

1,272,085 patients with at least 1 drug & at least 1 disease

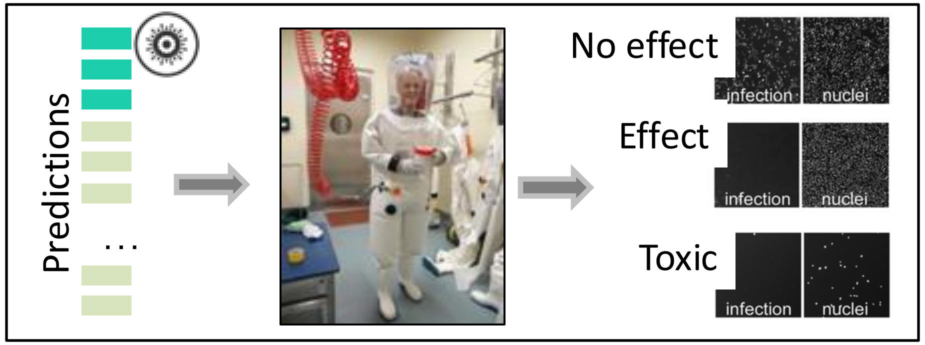


- TxGNN predicts therapeutic use for recent FDA approvals and informs laboratory testing



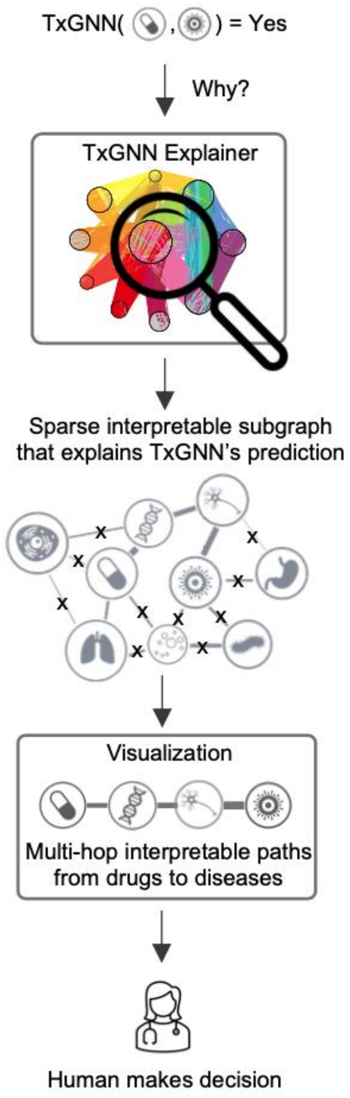
Drug name	Ingredient	Disease	Approval date	Company	FDA Number	Orphan	Prediction	Percentile
Welireg	Belzutifan	von Hippel-Lindau disease	08/13/2021	Merck	NDA215383	Yes	0.720	4.11%
Livtencity	Maribavir	Cytomegalovirus infection	11/23/2021	Takeda	NDA215596	Yes	0.033	66.37%
Tezspire	Tezepelumab-Ekko	Asthma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

PNAS'21



A foundation model for clinician-centered drug repurposing, *Nature Medicine* 2024

Clinician-centered design: txgnn.org


TxGNN(,) = Yes

Why?

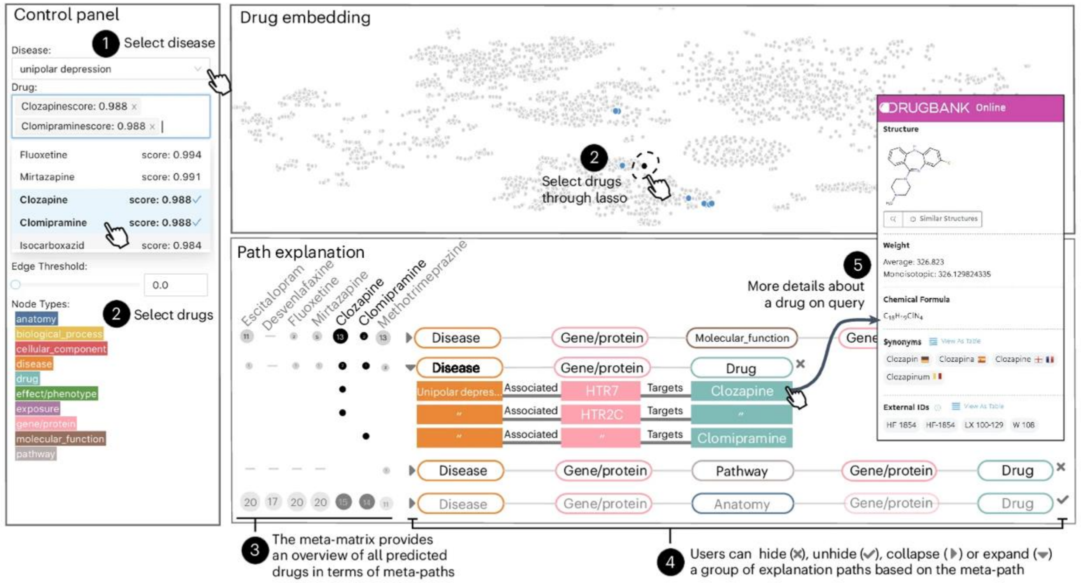
TxGNN Explainer

Sparse interpretable subgraph that explains TxGNN's prediction

Visualization

Multi-hop interpretable paths from drugs to diseases

Human makes decision


Control panel

1 Select disease

Disease: unipolar depression

Drug: Clozapinescore: 0.988 × Clomipraminescore: 0.988 ×

Fluoxetine	score: 0.994
Mirtazapine	score: 0.991
Clozapine	score: 0.988 ✓
Clomipramine	score: 0.988 ✓
Isocarboxazid	score: 0.984

Edge Threshold: 0.0

Node Types: anatomy, biological_process, cellular_component, disease, drug, effect/phenotype, exposure, gene/protein, molecular_function, pathway

2 Select drugs

Drug embedding

2 Select drugs through lasso

Path explanation

5 More details about a drug on query

3 The meta-matrix provides an overview of all predicted drugs in terms of meta-paths

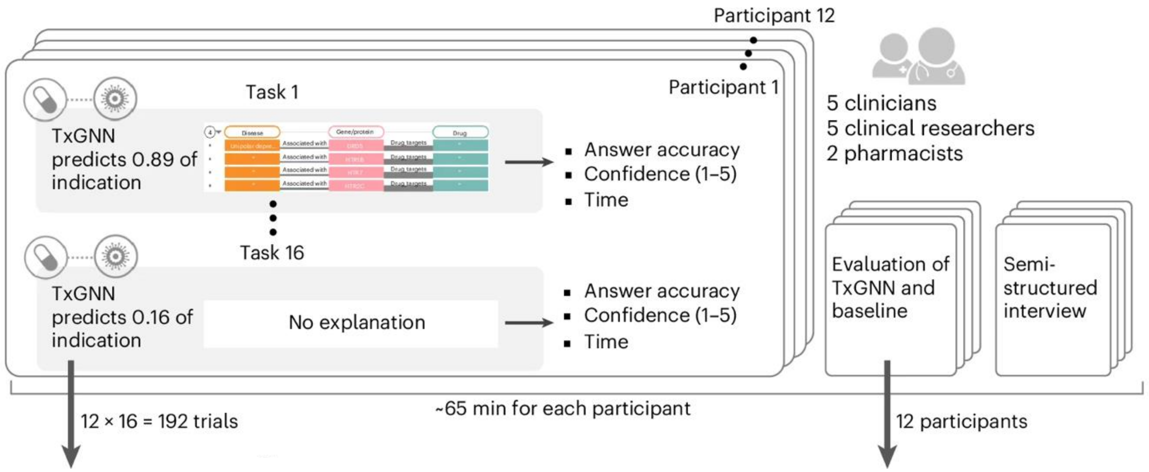
4 Users can hide (×), unhide (✓), collapse (▶) or expand (▼) a group of explanation paths based on the meta-path

Escitalopram
Desvenlafaxine
Fluoxetine
Mirtazapine
Clozapine
Clomipramine
Methotrimeprazine
Disease — Gene/protein — Molecular_function — Gene
Disease — Gene/protein — Drug
Unipolar depres... Associated HTR7 Targets Clozapine
" Associated HTR2C Targets "
" Associated " Targets Clomipramine
Disease — Gene/protein — Pathway — Gene/protein — Drug
Disease — Gene/protein — Anatomy — Gene/protein — Drug
20 17 20 20 15 14 11
DRUGBANK Online
Structure
Similar Structures
Weight
Average: 326.823
Monoisotopic: 326.129824335
Chemical Formula
C18H19ClN4
Synonyms View As Table
Clozapin Clozapina Clozapine
Clozapinum
External IDs View As Table
HF 1854 HF-1854 LX 100-129 W 108

Panels of clinicians, clinical researchers and pharmacists test usability of TxGNN:

- Scientific and medical consensus
- User confidence and trust
- User agreement
- Time used for exploring predictions

Path-based explanations perform significantly better than **node-based explanations** and **subgraph-based explanations** across three usability metrics: accuracy, confidence, time


Participant 12

Participant 1

Task 1

TxGNN predicts 0.89 of indication

- Answer accuracy
- Confidence (1–5)
- Time

Task 16

TxGNN predicts 0.16 of indication

No explanation

- Answer accuracy
- Confidence (1–5)
- Time

5 clinicians
5 clinical researchers
2 pharmacists

Evaluation of TxGNN and baseline

Semi-structured interview

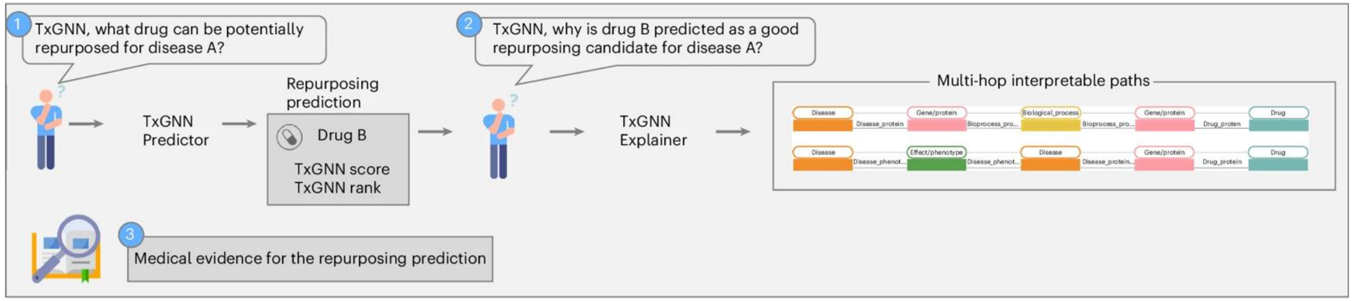
12 × 16 = 192 trials

~65 min for each participant

12 participants

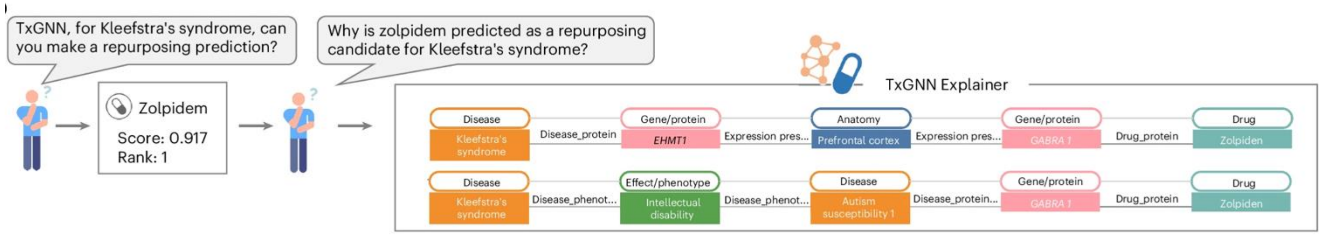
A foundation model for clinician-centered drug repurposing, *Nature Medicine* 2024

Clinician-centered design: txgnn.org


1 TxGNN, what drug can be potentially repurposed for disease A?

2 TxGNN, why is drug B predicted as a good repurposing candidate for disease A?

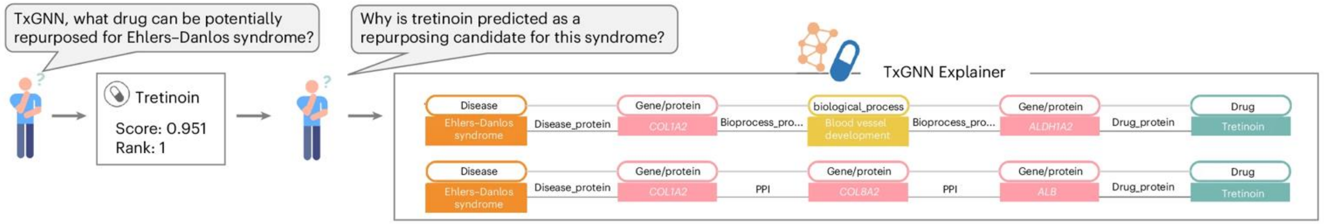
3 Medical evidence for the repurposing prediction


TxGNN, for Kleefstra's syndrome, can you make a repurposing prediction?

Why is zolpidem predicted as a repurposing candidate for Kleefstra's syndrome?

Medical reasoning

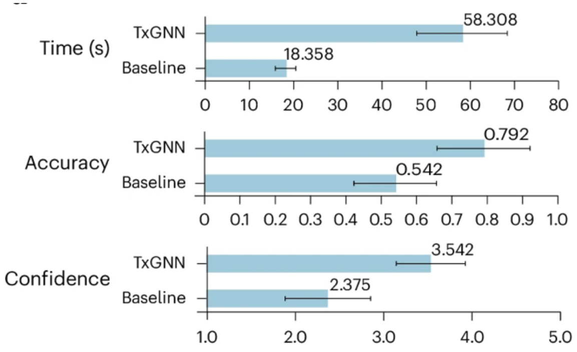
Kleefstra's syndrome is a rare genetic disorder caused by mutations in the *EHMT1* gene and marked by intellectual disability, delayed speech and autism. Zolpidem, a sedative primarily used for treating insomnia, has shown surprising neurostimulating effects in various medical case studies of neurodevelopmental disorders. This paradoxical activity of zolpidem can lead to temporary improvements in speech, motor skills and alertness, offering a potential therapeutic avenue for this syndrome.

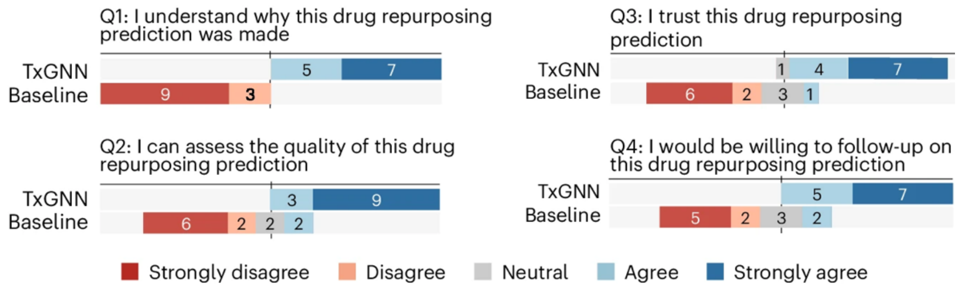

TxGNN, what drug can be potentially repurposed for Ehlers–Danlos syndrome?

Why is tretinoin predicted as a repurposing candidate for this syndrome?

Medical reasoning

Ehlers–Danlos syndrome is a rare connective tissue disorder caused by mutations in collagen-coding genes (*COL1A1*/*COL1A2*) that lead to poor wound healing and abnormal scars. Tretinoin, a Vitamin A derivative, carried by albumin (*ALB*) and acting on *ALDH1A2*, may help improve these symptoms by promoting collagen production in the skin. In ClinVar, Ehlers–Danlos subtypes are linked to *ALB* mutations associated with *ALDH1AI*.



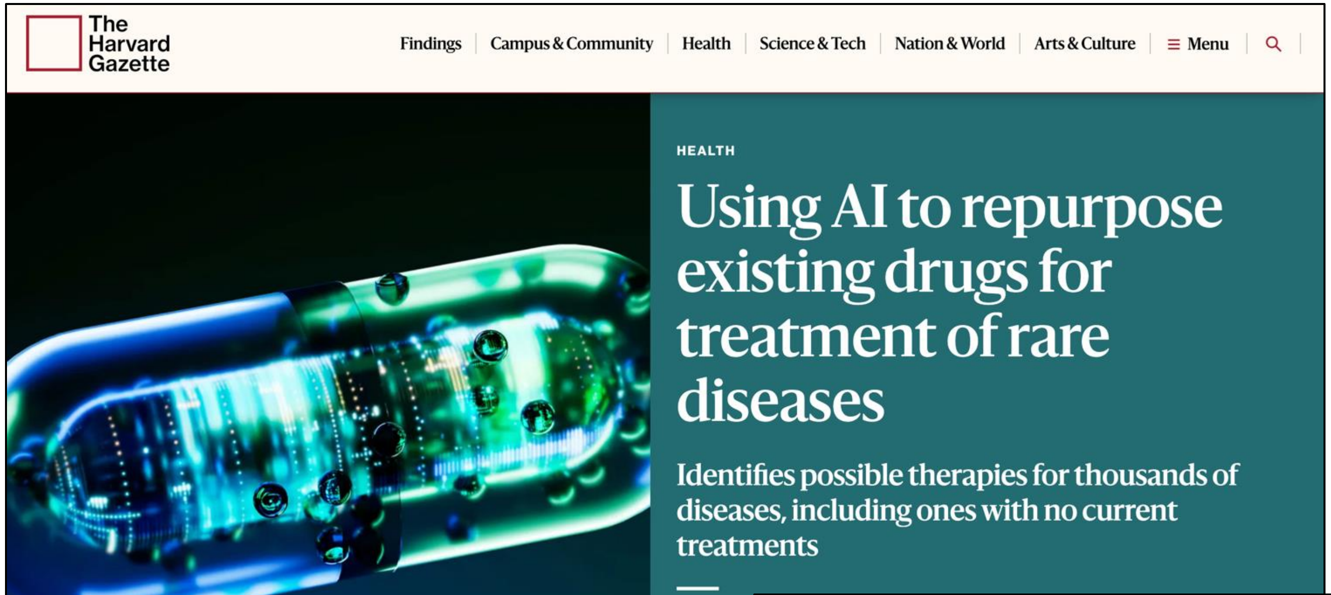


Mount Sinai Hospital

- Better accuracy (+46%) and confidence (+49%) when explanations provided
- Support scientists in interacting with TxGNN and interpreting TxGNN predictions

A foundation model for clinician-centered drug repurposing, *Nature Medicine* 2024

Open models, open datasets, and evaluations



The Harvard Gazette
Findings | Campus & Community | Health | Science & Tech | Nation & World | Arts & Culture | ≡ Menu

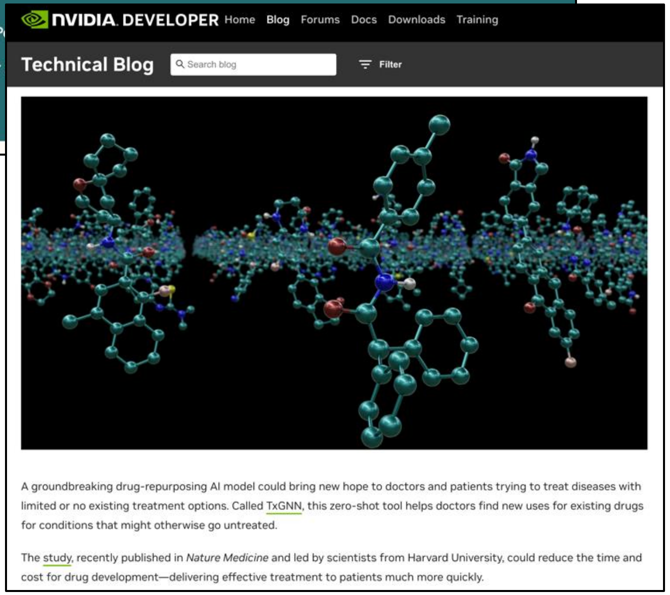
HEALTH

Using AI to repurpose existing drugs for treatment of rare diseases

Identifies possible therapies for thousands of diseases, including ones with no current treatments

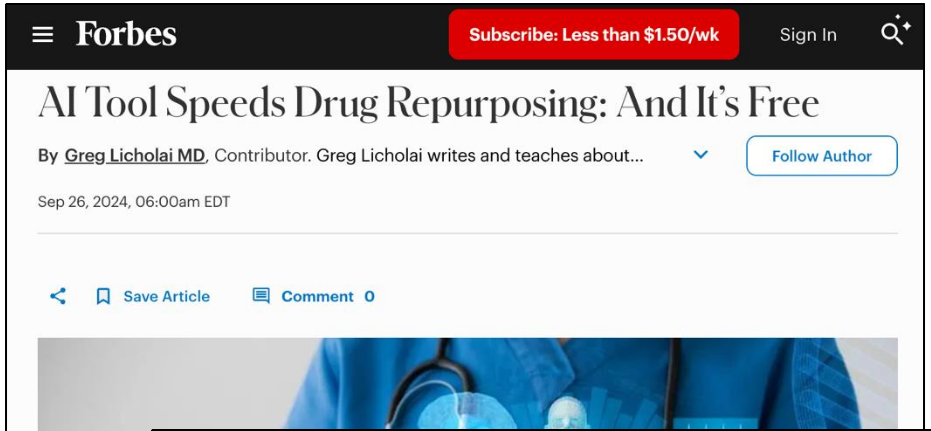
Ekaterina P

September



A groundbreaking drug-repurposing AI model could bring new hope to doctors and patients trying to treat diseases with limited or no existing treatment options. Called TxGNN, this zero-shot tool helps doctors find new uses for existing drugs for conditions that might otherwise go untreated.

The study, recently published in *Nature Medicine* and led by scientists from Harvard University, could reduce the time and cost for drug development—delivering effective treatment to patients much more quickly.



AI Tool Speeds Drug Repurposing: And It's Free

By Greg Licholai MD, Contributor. Greg Licholai writes and teaches about...

Sep 26, 2024, 06:00am EDT

AI for drug repur

Innovative a
the myriad c
expensive, w

PRESS RELEASE

With TxGNN, Kempner Researchers Introduce an AI "Dr. House" to Find Treatments for Rare Diseases

By Yohan J. John, Ph.D. | September 30, 2024

Kempner scientists are using powerful AI technology to identify potential drug-disease pairings that could help advance treatment for rare diseases.

- Real-world implementation
- Clinical collaborations for 20+ diseases, including neurology, cancer, and rare diseases