Machine Learning for Drug Development

Marinka Zitnik

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

marinka@hms.harvard.edu
https://zitniklab.hms.harvard.edu

HARVARD MEDICAL SCHOOL HDSI | Harvard Data Science Initiative BROAD INSTITUTE

Outline

✓Overview and introduction

Part 1: Virtual drug screening and drug repurposing ☜

Part 2: Adverse drug effects, drug-drug interactions

Part 3: Clinical trial site identification, patient recruitment

Part 4: Molecule optimization, molecular graph generation, multimodal graph-to-graph translation

Part 5: Molecular property prediction and transformers

Demos, resources, wrap-up & future directions

Method:

Subgraph Neural Networks

Alsentzer, Finlayson, Li, and Zitnik, Subgraph Neural Networks, *NeurIPS* 2020

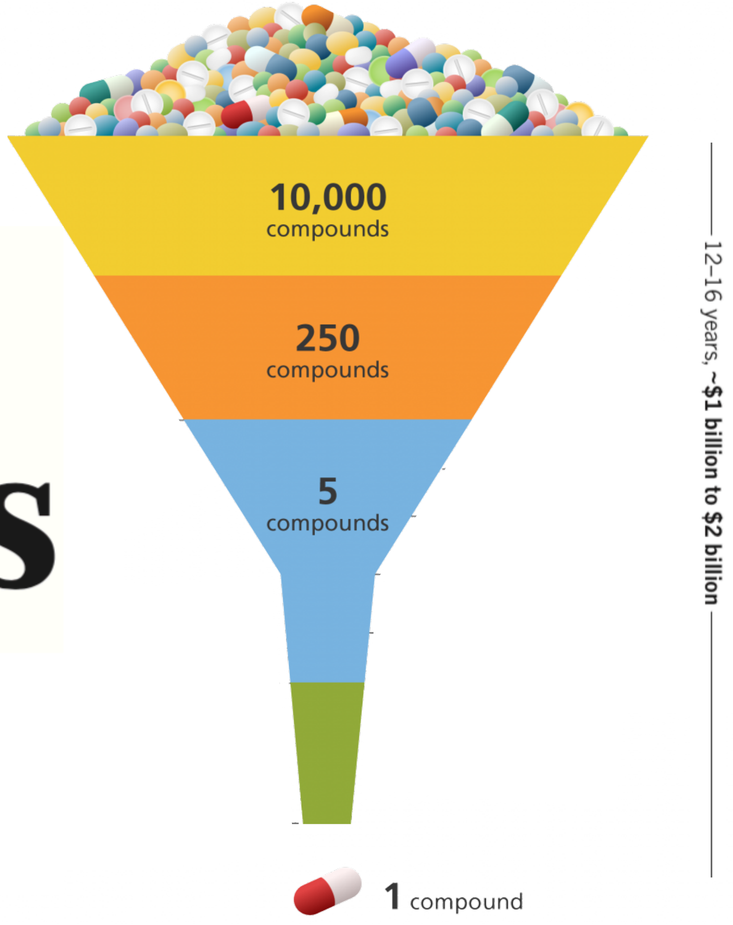
Application:

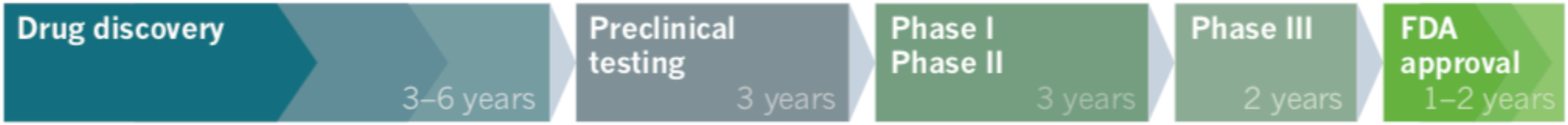
Finding Effective Drug Treatments

In submission

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.




Drug discovery 3–6 years

Preclinical testing 3 years

Phase I Phase II 3 years

Phase III 2 years

FDA approval 1–2 years

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

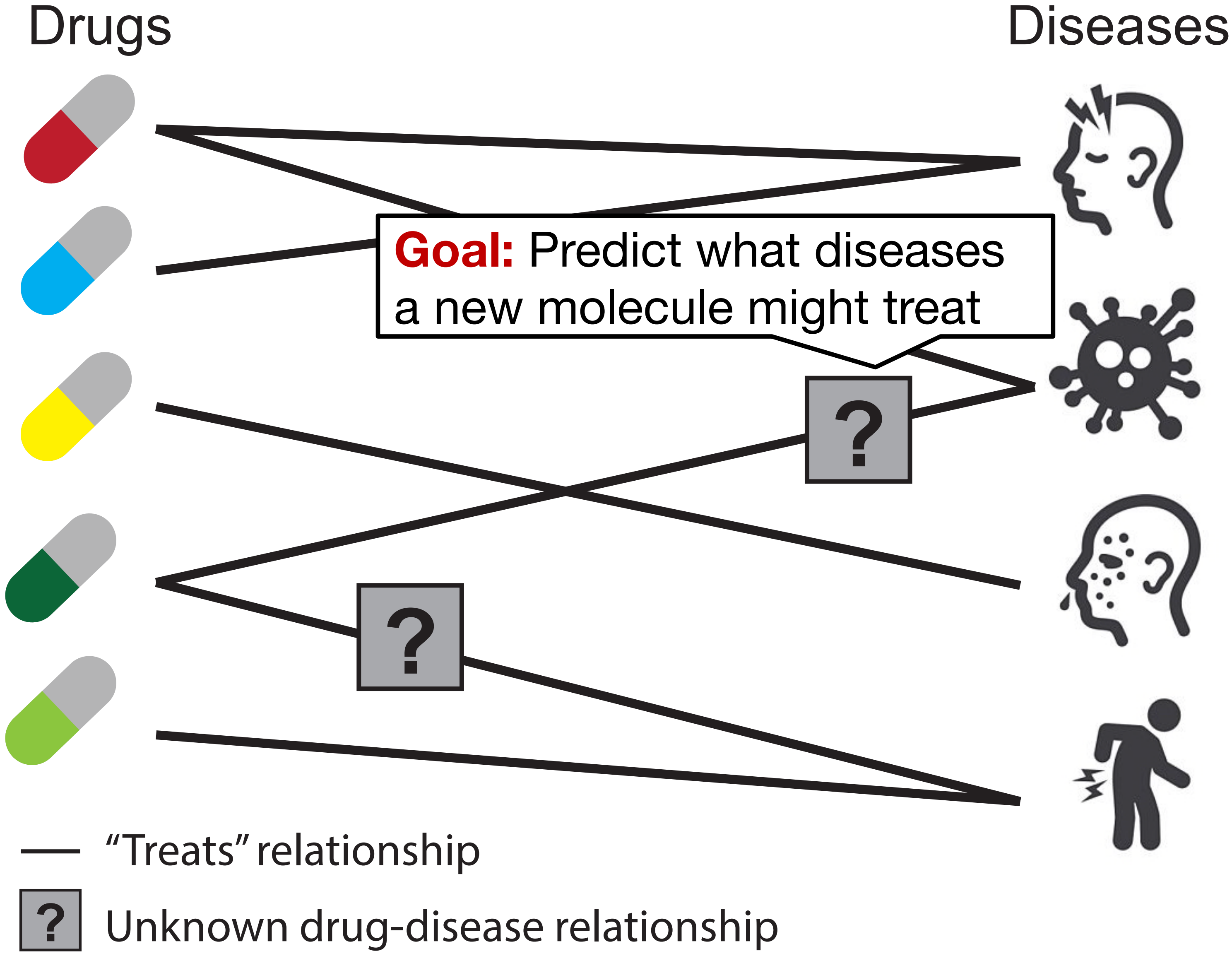
A SHORTER TIMESCALE

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

Drug repositioning

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

What drug treats what disease?



Drugs

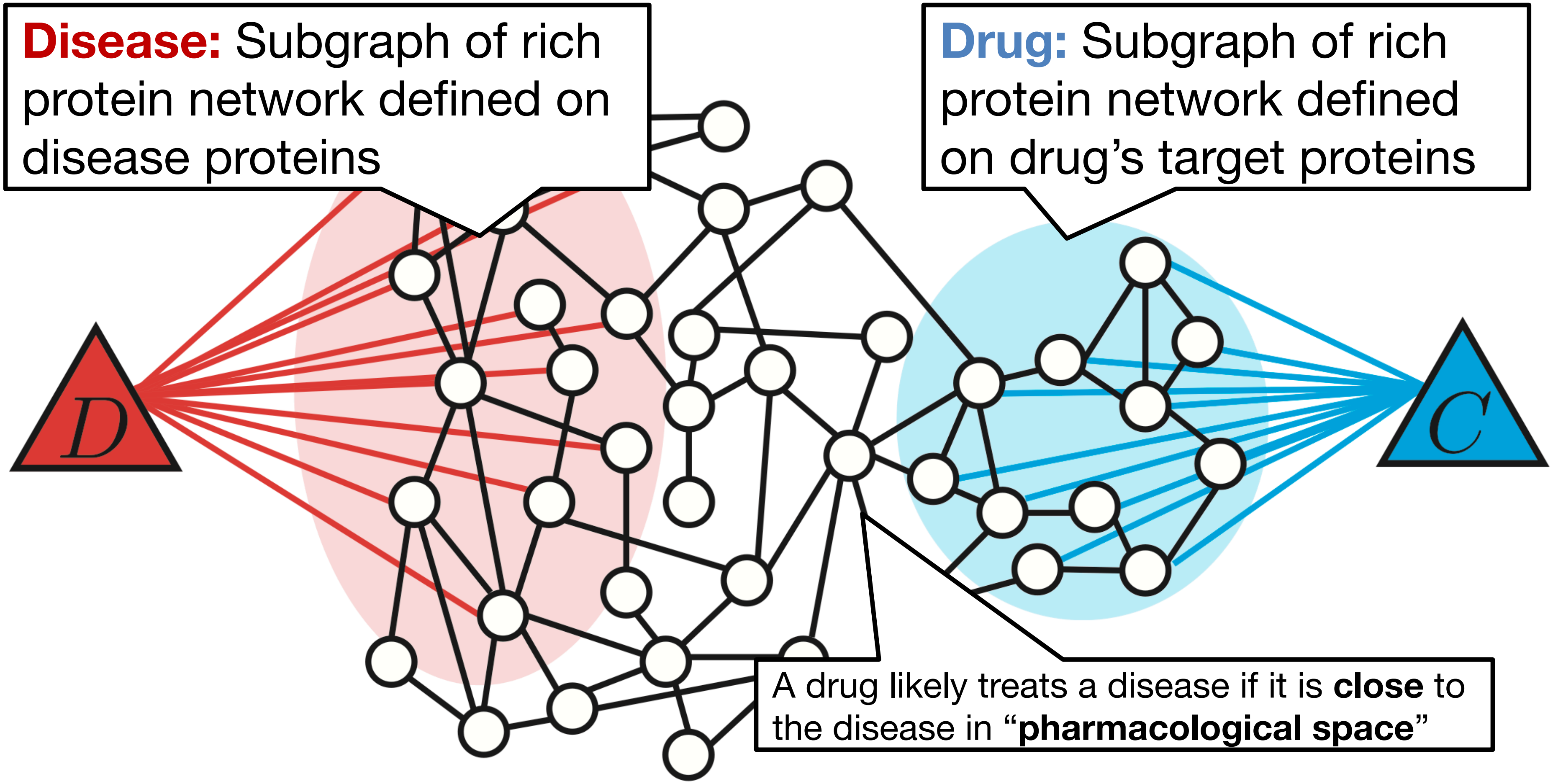
Diseases

Goal: Predict what diseases a new molecule might treat

— “Treats” relationship

? Unknown drug-disease relationship

Key Insight: Subgraphs



Disease: Subgraph of rich protein network defined on disease proteins

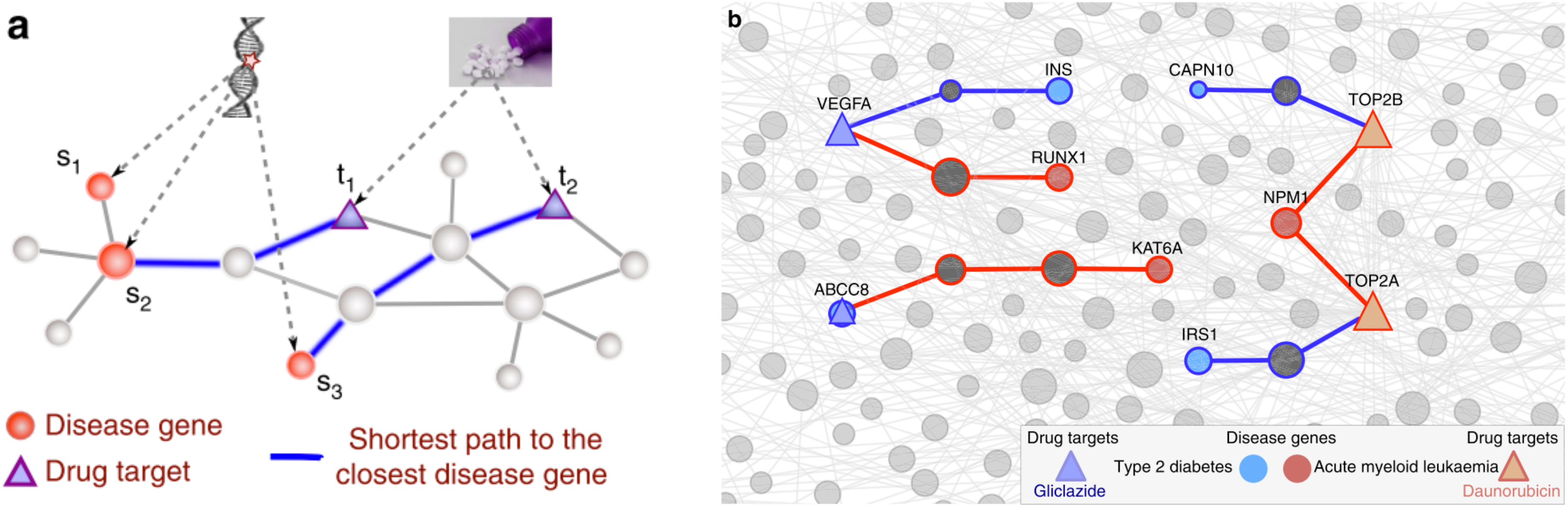
Drug: Subgraph of rich protein network defined on drug's target proteins

A drug likely treats a disease if it is **close** to the disease in "**pharmacological space**"

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

Why Subgraphs? – Part #1

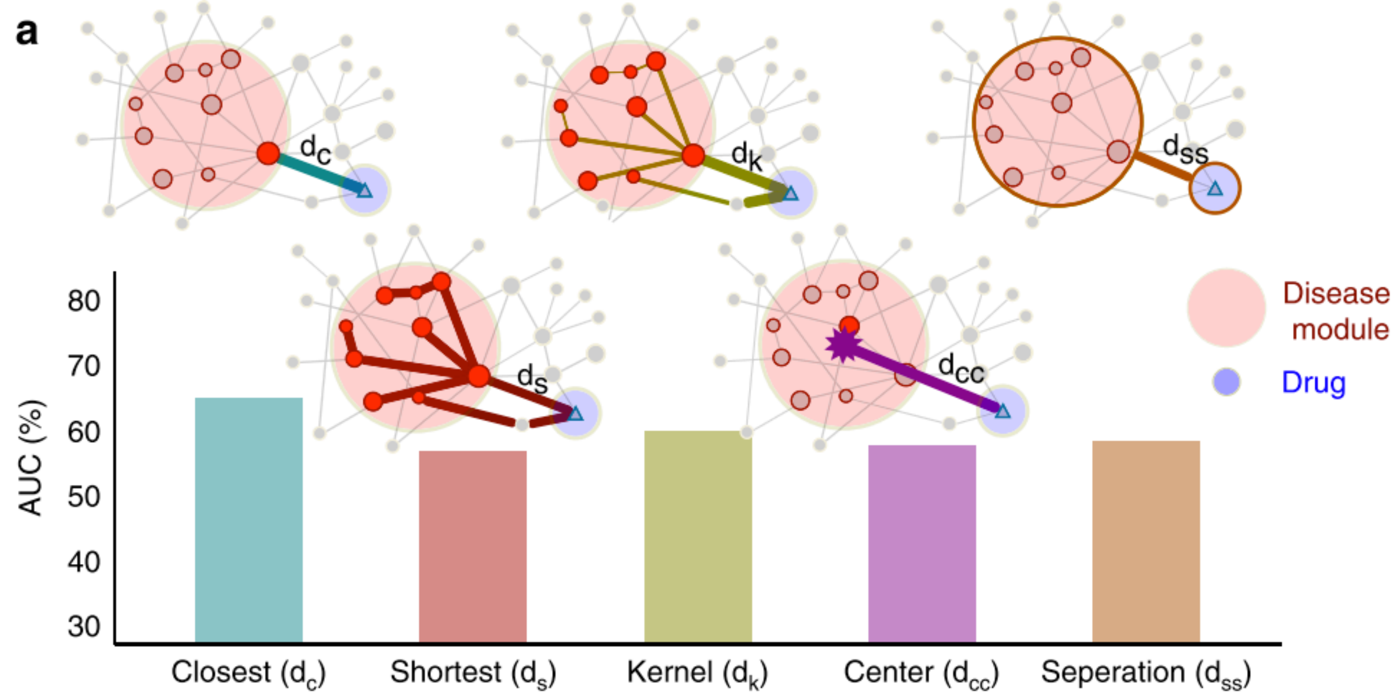
- Analysis of 238 drugs used in 78 diseases
- **Key result:** Therapeutic effect of drugs is localized in a small network neighborhood of disease genes



Guney, E., Menche, J., Vidal, M. and Barábasi, A.L., Network-based in silico drug efficacy screening. Nature Communications, 2016

Why Subgraphs? – Part #2

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



Guney, E., Menche, J., Vidal, M. and Barábasi, A.L., Network-based in silico drug efficacy screening. Nature Communications, 2016

Why Subgraphs? – Part #3

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

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

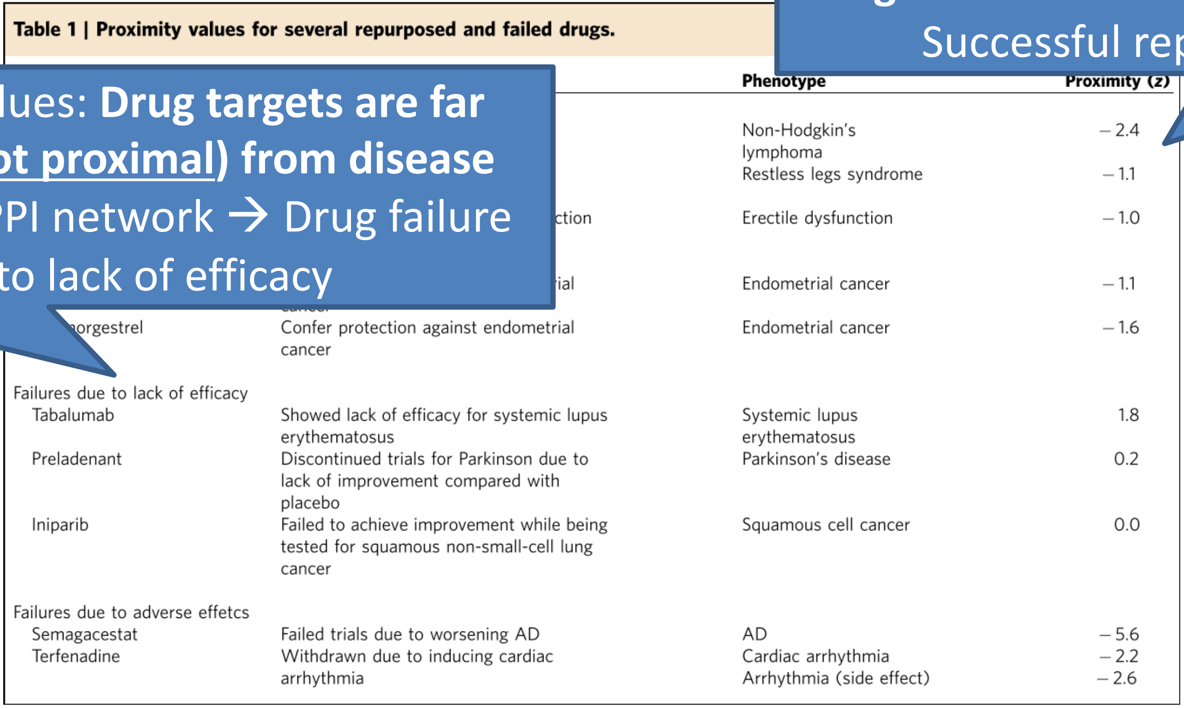


Table 1 | Proximity values for several repurposed and failed drugs.

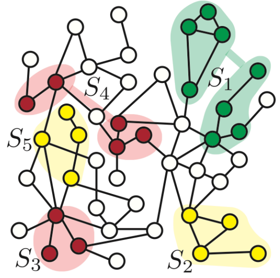
		Phenotype	Proximity (z)
		Non-Hodgkin's lymphoma	– 2.4
		Restless legs syndrome	– 1.1
	ction	Erectile dysfunction	– 1.0
	ial cancer	Endometrial cancer	– 1.1
norgestrel	Confer protection against endometrial cancer	Endometrial cancer	– 1.6
Failures due to lack of efficacy			
Tabalumab	Showed lack of efficacy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

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

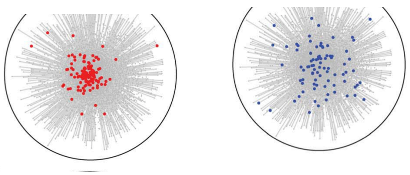
Guney, E., Menche, J., Vidal, M. and Barábasi, A.L., Network-based in silico drug efficacy screening. Nature Communications, 2016

Why are subgraphs challenging?

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

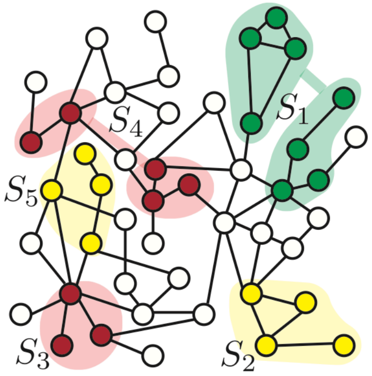


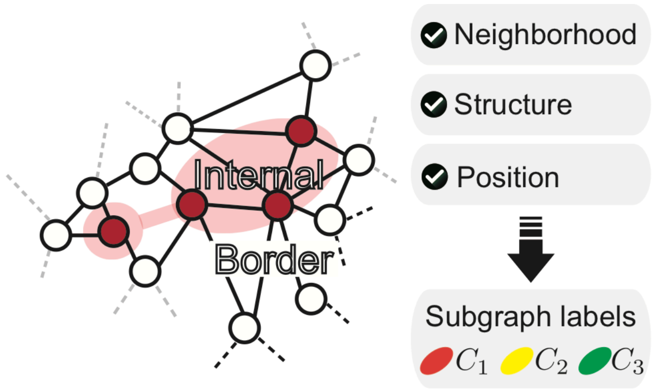
- Rich connectivity patterns, both internally a externally through interactions with the rest of G:
  - How to inject this information into a GNN?
- Subgraphs can be:
  - Localized and reside in our region of the graph
  - Distributed across multiple local neighborhoods



Problem Formulation

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





✓ Neighborhood

✓ Structure

✓ Position

Subgraph labels

🟥 C_1 🟨 C_2 🟩 C_3

SubGNN: Overview

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

Emily Alsentzer

Sam Finlayson

Michelle Li

Alsentzer, Finlayson, Li, and Zitnik, Subgraph Neural Networks, *NeurIPS* 2020

Part 1: Neural Message Passing

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

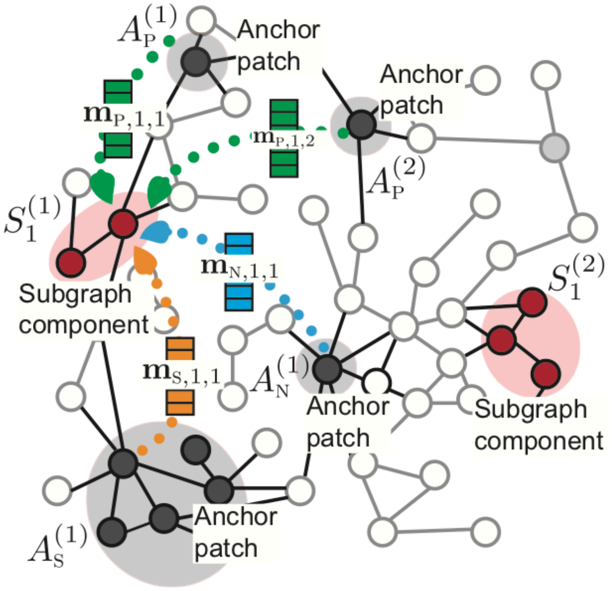
similarity function between a subgraph component and an anchor patch

$$\text{MSG}_X = \boxed{\gamma_X}\left(S^{(C)}, A_X\right) \cdot p_X$$

$$\mathbf{a}_{X,c} = \text{AGG}_M\left(\left\{\text{MSG}_X(S^{(C)}, A_X, p_X), \forall A_X \in \mathcal{A}_X\right\}\right),$$

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

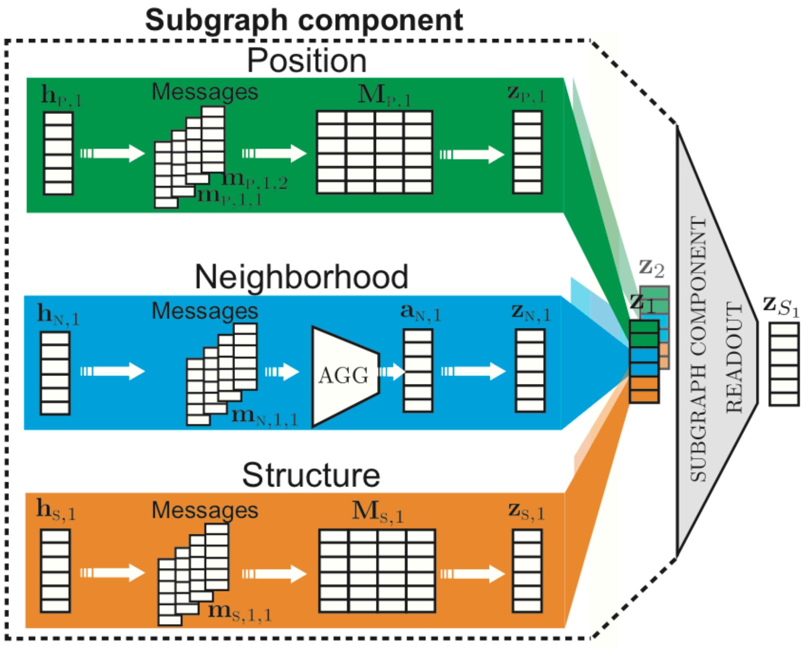
property-specific representation of a subgraph component; passed to the next layer



Part 2: Property-aware Routing

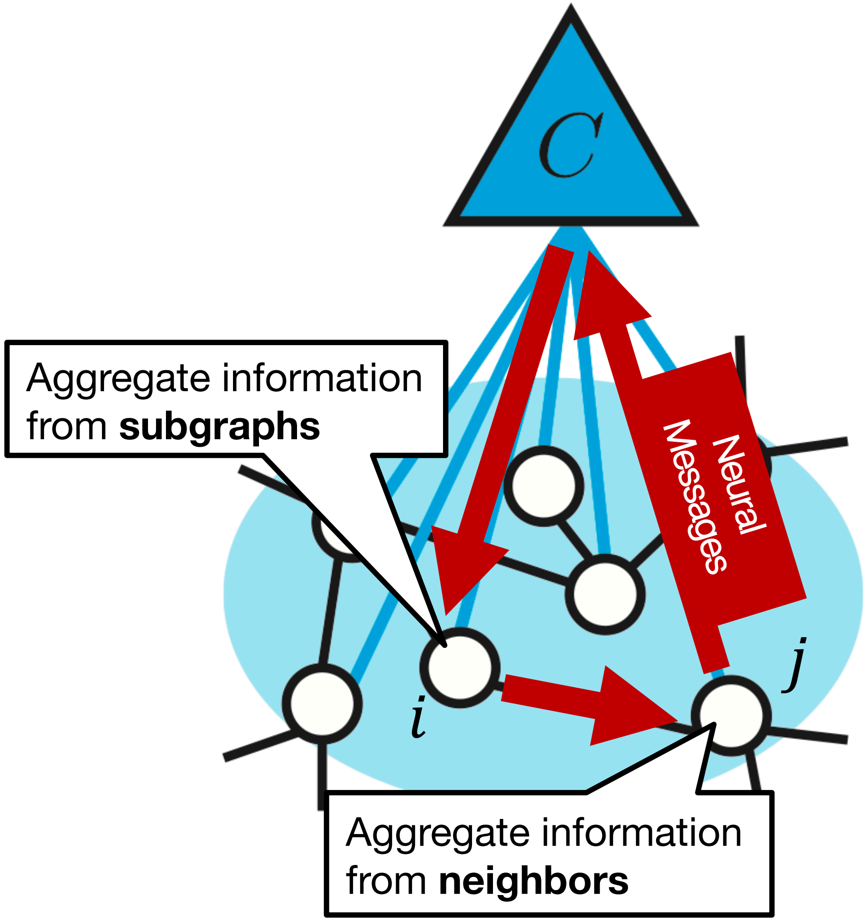
- SubGNN specifies three channels, each designed to capture a distinct subgraph property
  - Position, neighborhood, and structure

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

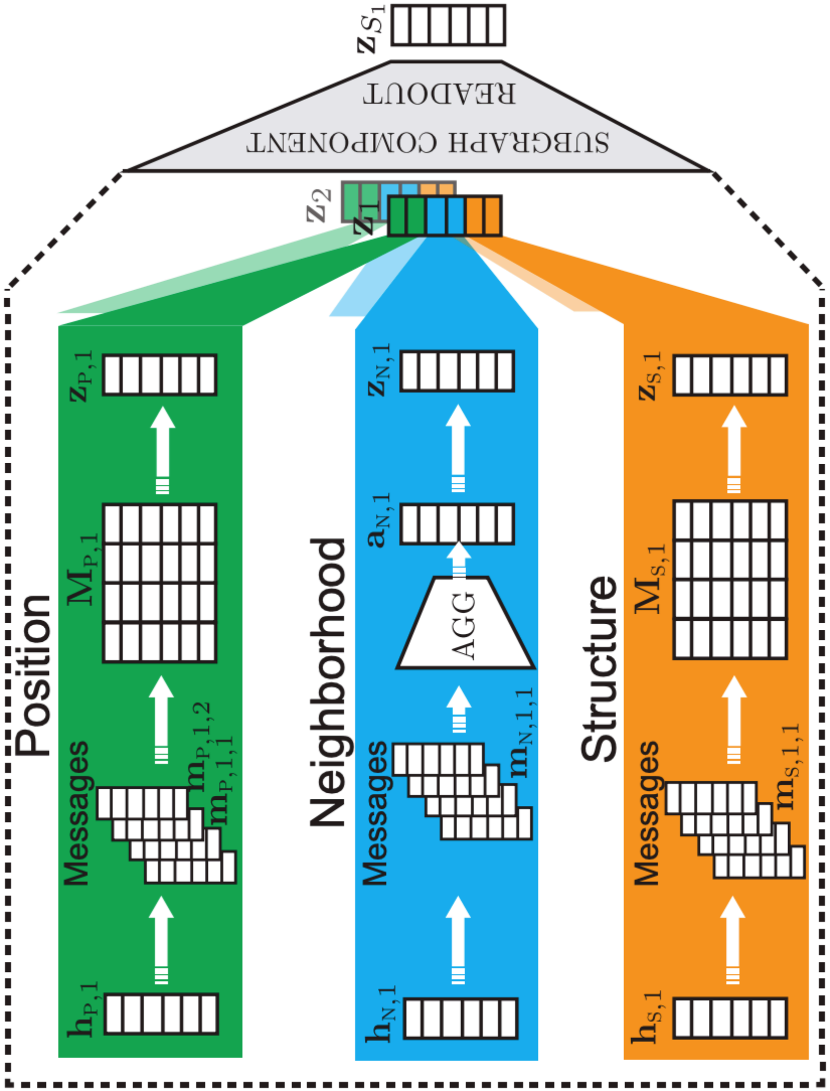


Channel outputs $\boldsymbol{z}_x$ are concatenated to produce **a final subgraph representation $\boldsymbol{z}_S$**

SubGNN: Overview


Aggregate information from **subgraphs**

Aggregate information from **neighbors**



Setup: Drug Repurposing dataset

- Protein-protein interaction network culled from 15 knowledge databases with 19K nodes, 350K edges

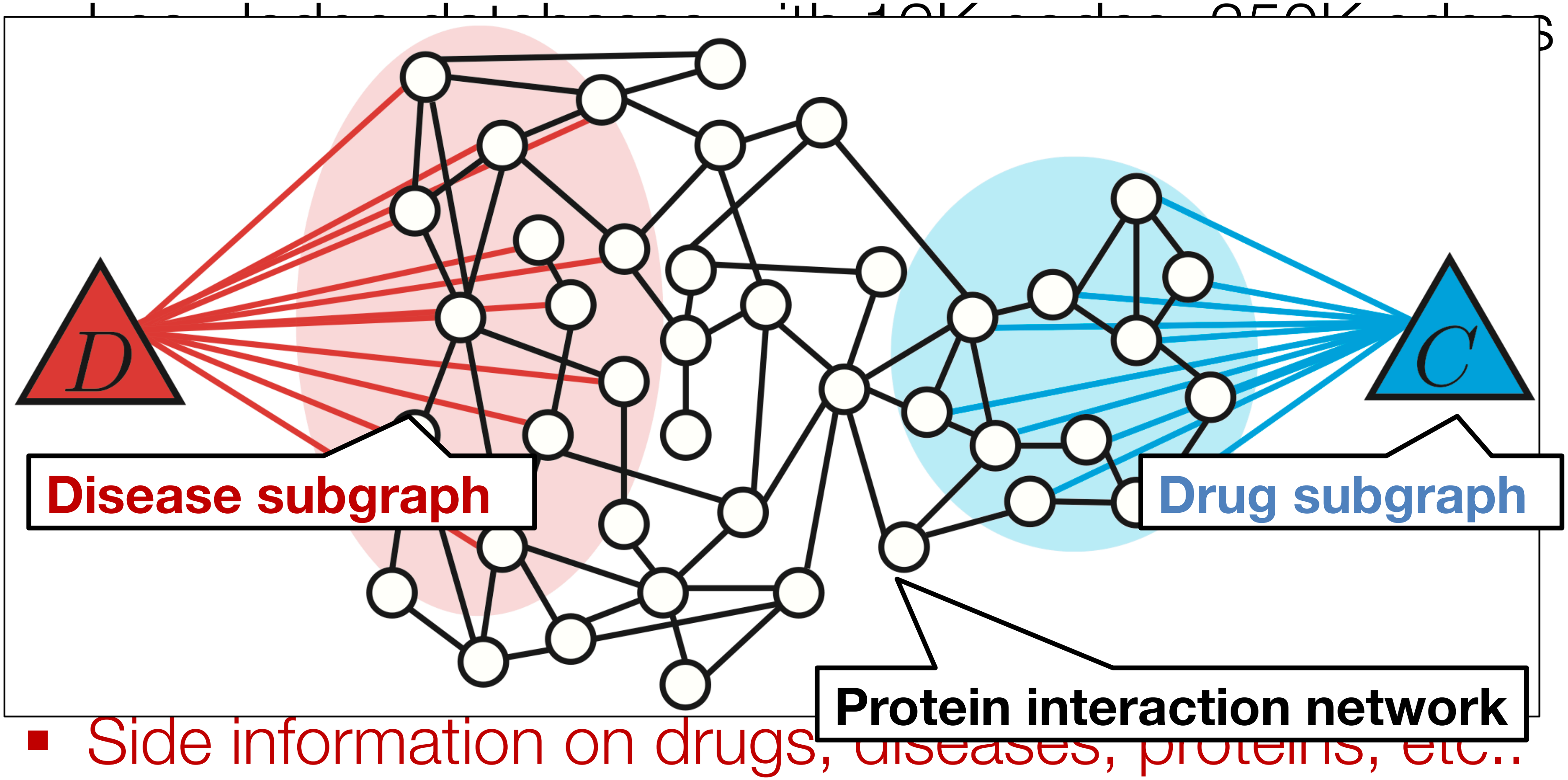
- Drug-protein and disease-protein links:
  - DrugBank, OMIM, DisGeNET, STITCH DB and others
  - 20K drug-protein links, 560K disease-protein links

- Medical indications and contra-indications:
  - DrugBank, MEDI-HPS, DailyMed, Drug Central, RepoDB
  - 6K drug-disease indications

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

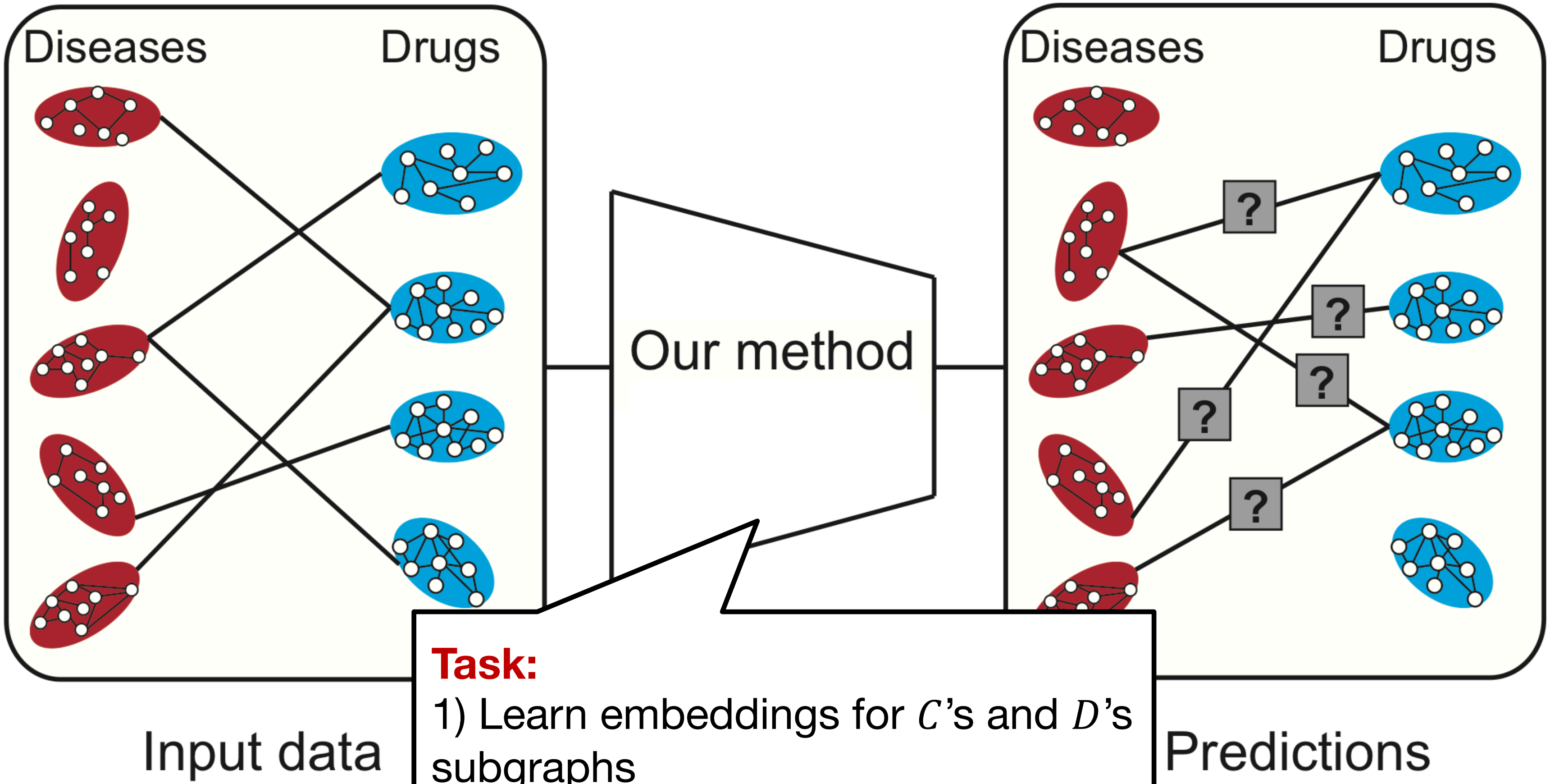
Setup: Drug repurposing dataset

- Protein-protein interaction network culled from 15 knowledge databases with 19K nodes, 350K edges



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

Predict links between drug and disease subgraphs

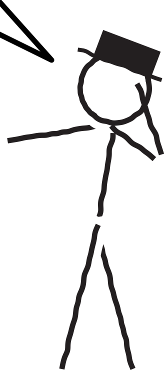


Task:
1) Learn embeddings for C's and D's subgraphs
2) Predict whether C should be **indicated** or **contra-indicated** for D

Results: Drug Repurposing

Stanford MEDICINE | SPARK Translational Research Program
From Bench to Bedside

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



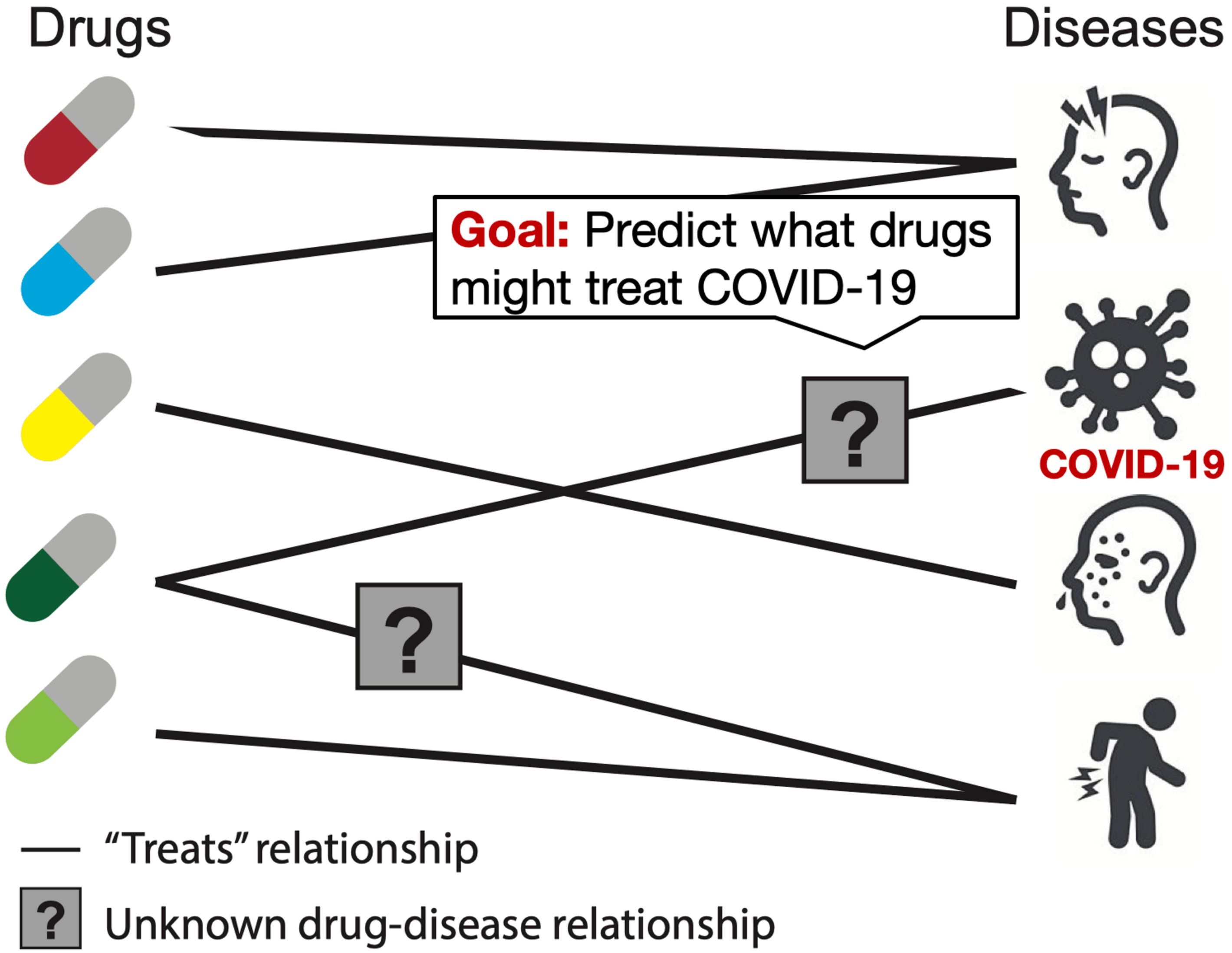
Drug	Disease	
N-acetyl-cysteine	cystic fibrosis	
Xamoterol	neurodegenerat	
Plerixafor	cancer	
Sodium selenite	cancer	Rank: 36/5000
Ebselen	C difficile	Rank: 10/5000
Itraconazole	cancer	Rank: 26/5000
Bestatin	lymphedema	Rank: 11/5000
Bestatin	pulmonary arterial hypertension	Rank: 16/5000
Ketaprofen	lymphedema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

Drug Repurposing for Emerging Pathogens

Paper:

Deisy Morselli Gysi, Ítalo Do Valle, Marinka Zitnik, Asher Ameli, Xiao Gan, et al. Network Medicine Framework for Identifying Drug Repurposing Opportunities for COVID-19, *arXiv:2004.07229*

Emerging Pathogens



Goal: Predict what drugs might treat COVID-19

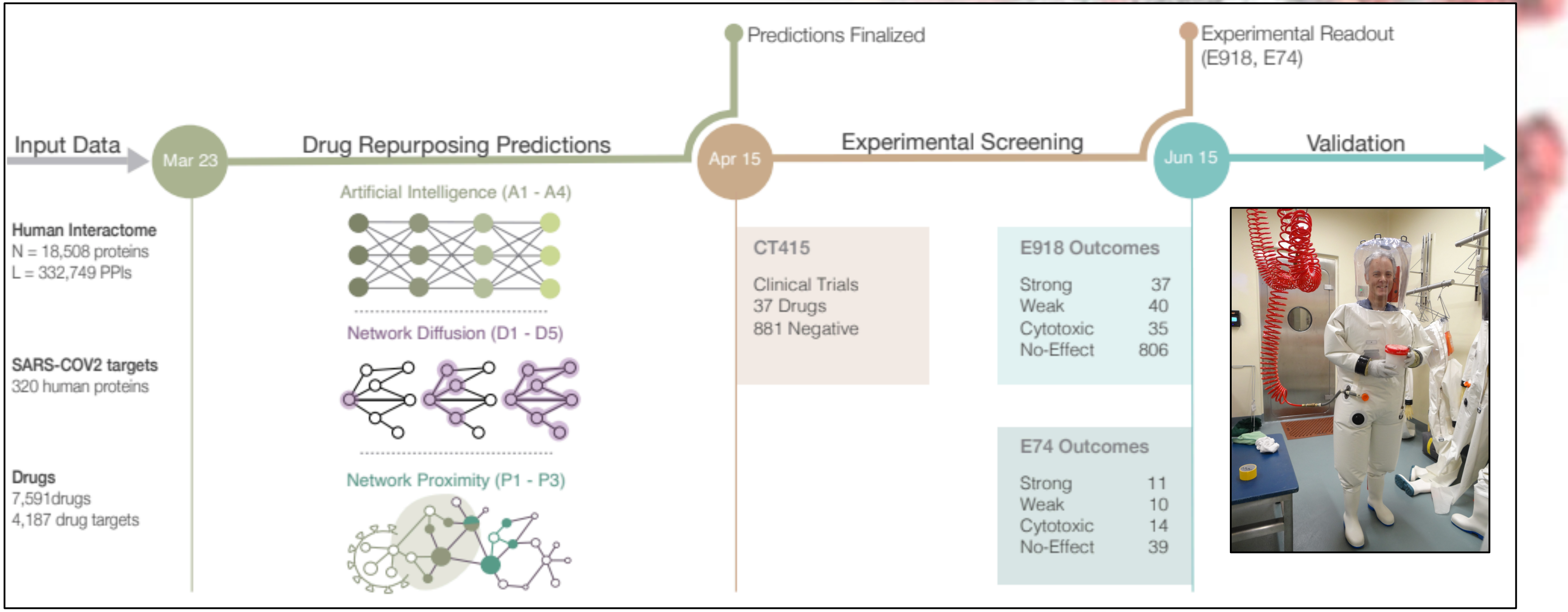
— "Treats" relationship

? Unknown drug-disease relationship

Never-Before-Seen Disease

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

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



Predictions Finalized
Experimental Readout
(E918, E74)
Input Data
Mar 23
Drug Repurposing Predictions
Apr 15
Experimental Screening
Jun 15
Validation
Artificial Intelligence (A1 - A4)
Network Diffusion (D1 - D5)
Network Proximity (P1 - P3)

Human Interactome
N = 18,508 proteins
L = 332,749 PPIs

SARS-COV2 targets
320 human proteins

Drugs
7,591drugs
4,187 drug targets

CT415

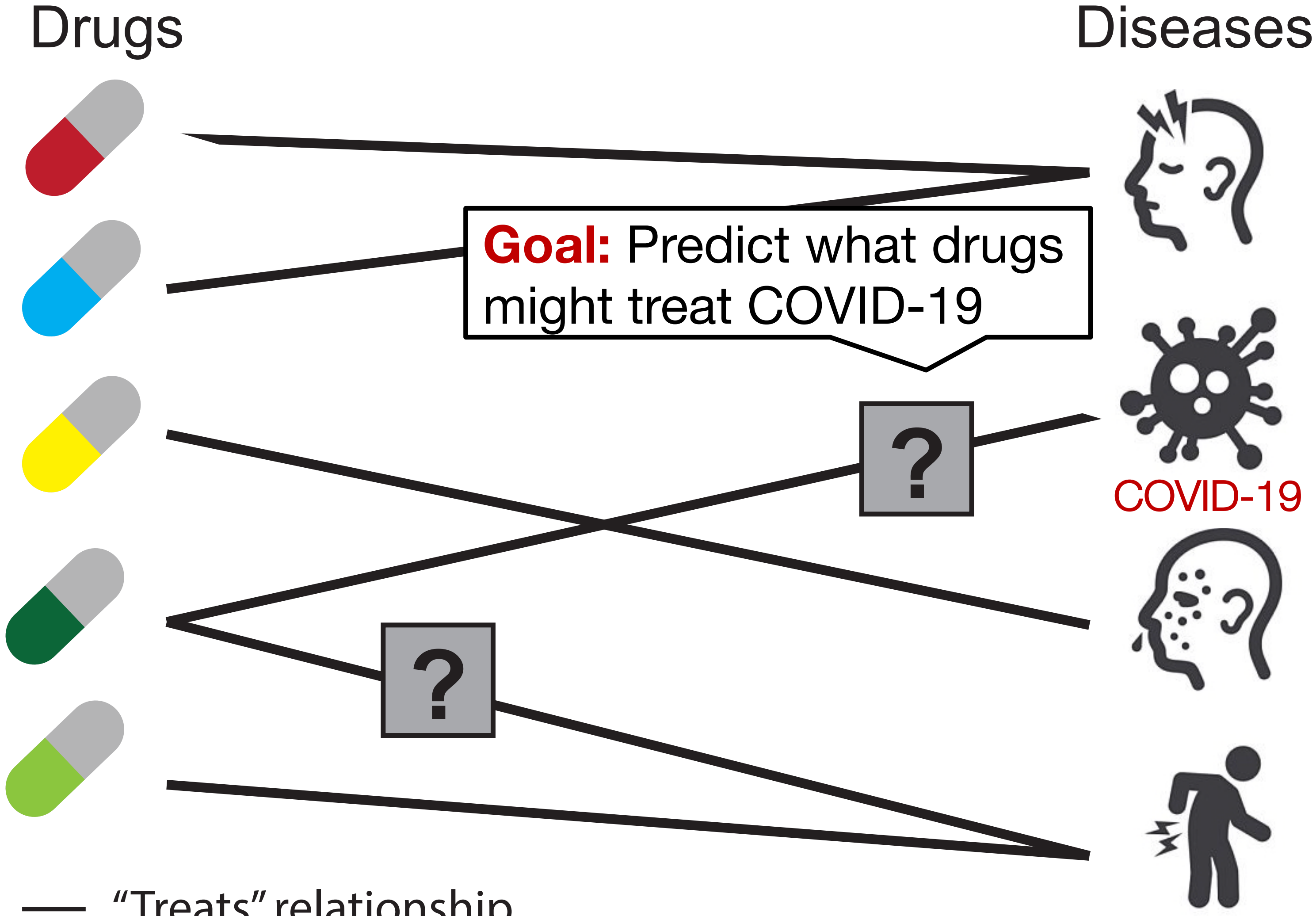
Clinical Trials
37 Drugs
881 Negative

E918 Outcomes	
Strong	37
Weak	40
Cytotoxic	35
No-Effect	806

E74 Outcomes	
Strong	11
Weak	10
Cytotoxic	14
No-Effect	39

Network Medicine Framework for Identifying Drug Repurposing Opportunities for Covid-19, *arXiv:2004.07229*

Never-before-seen disease

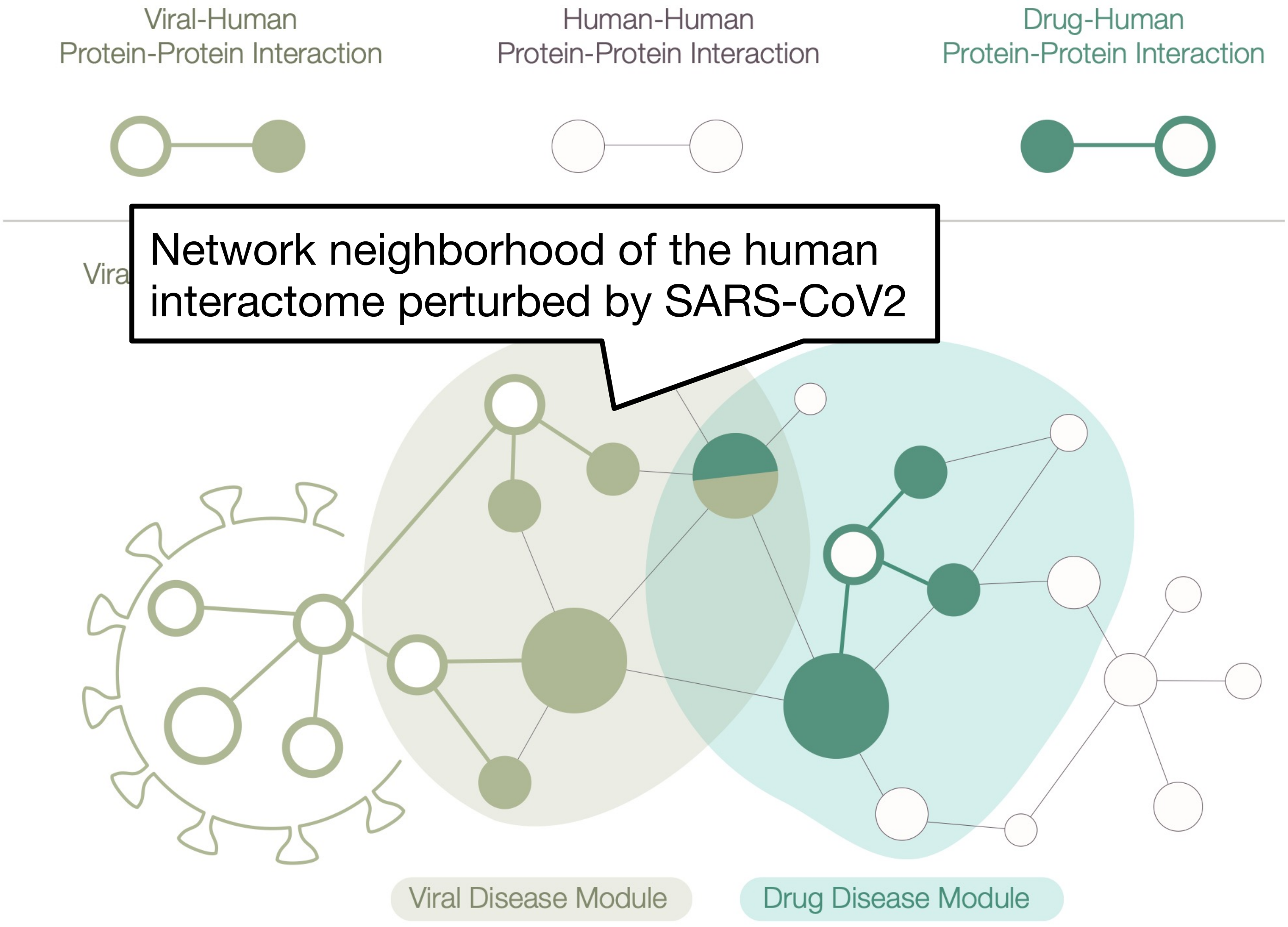


Goal: Predict what drugs might treat COVID-19

— "Treats" relationship

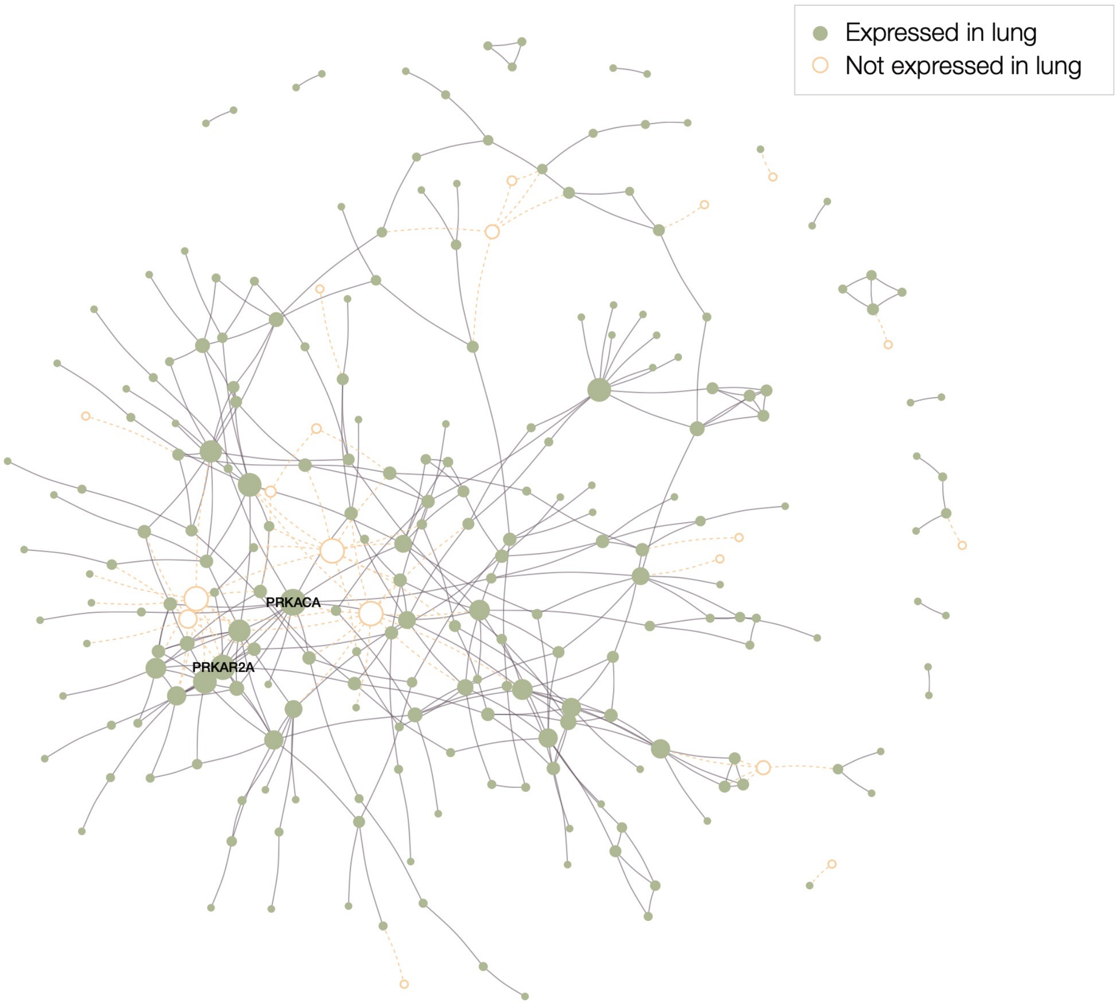
? Unknown drug-disease relationship

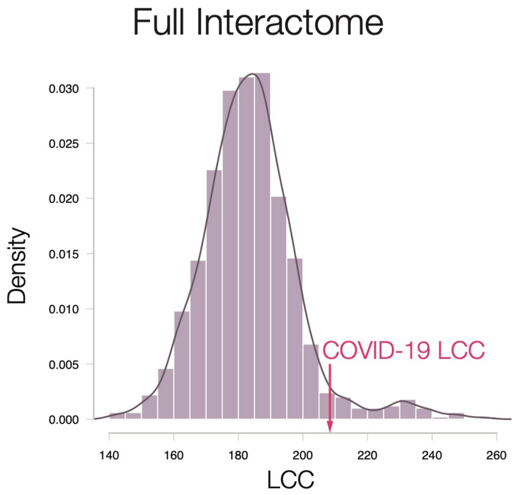
How to represent COVID-19? Map SARS-CoV2 targets to the human interactome

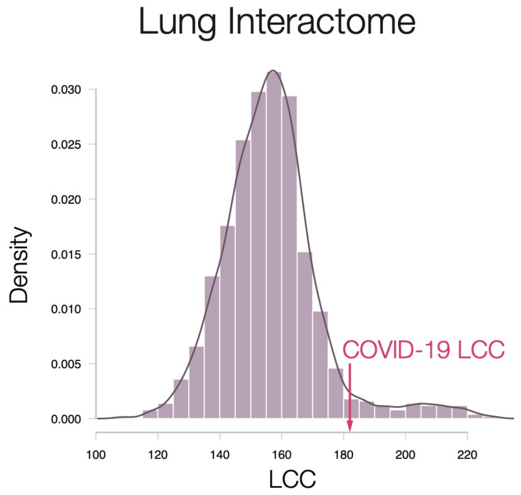

Viral-Human
Protein-Protein Interaction
Human-Human
Protein-Protein Interaction
Drug-Human
Protein-Protein Interaction
Vira
Viral Disease Module
Drug Disease Module

Network neighborhood of the human interactome perturbed by SARS-CoV2

COVID-19 Subgraph

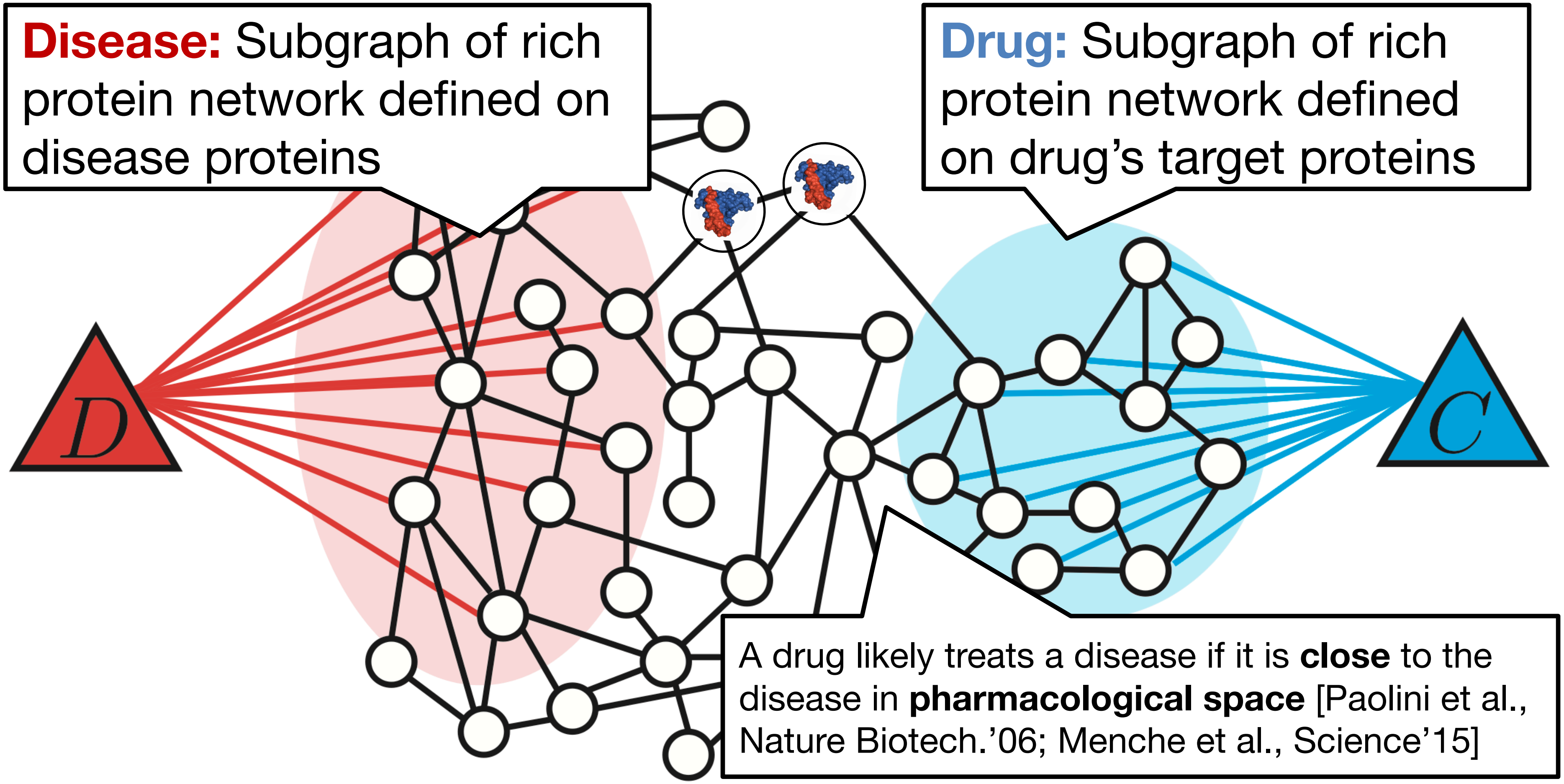



Full Interactome


Lung Interactome

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

Key Insight: Subgraphs



Disease: Subgraph of rich protein network defined on disease proteins

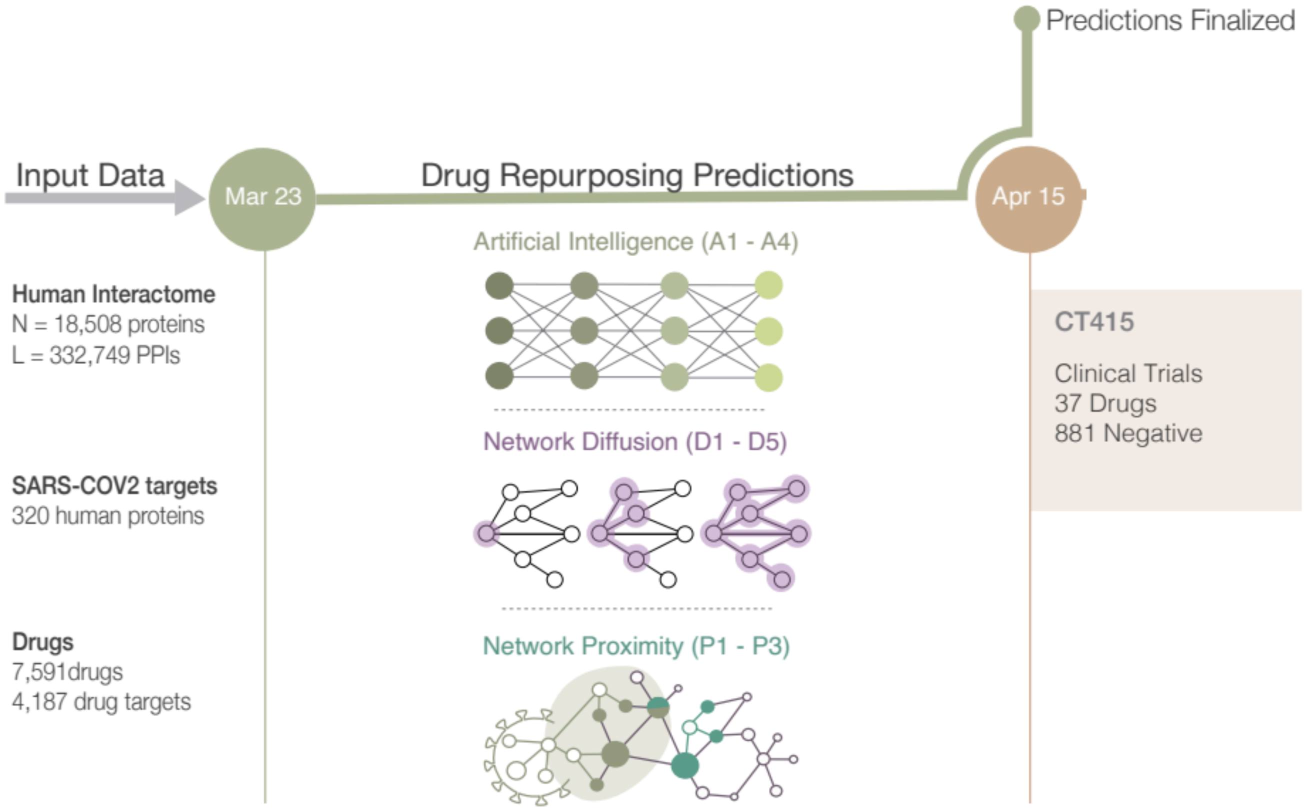
Drug: Subgraph of rich protein network defined on drug’s target proteins

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

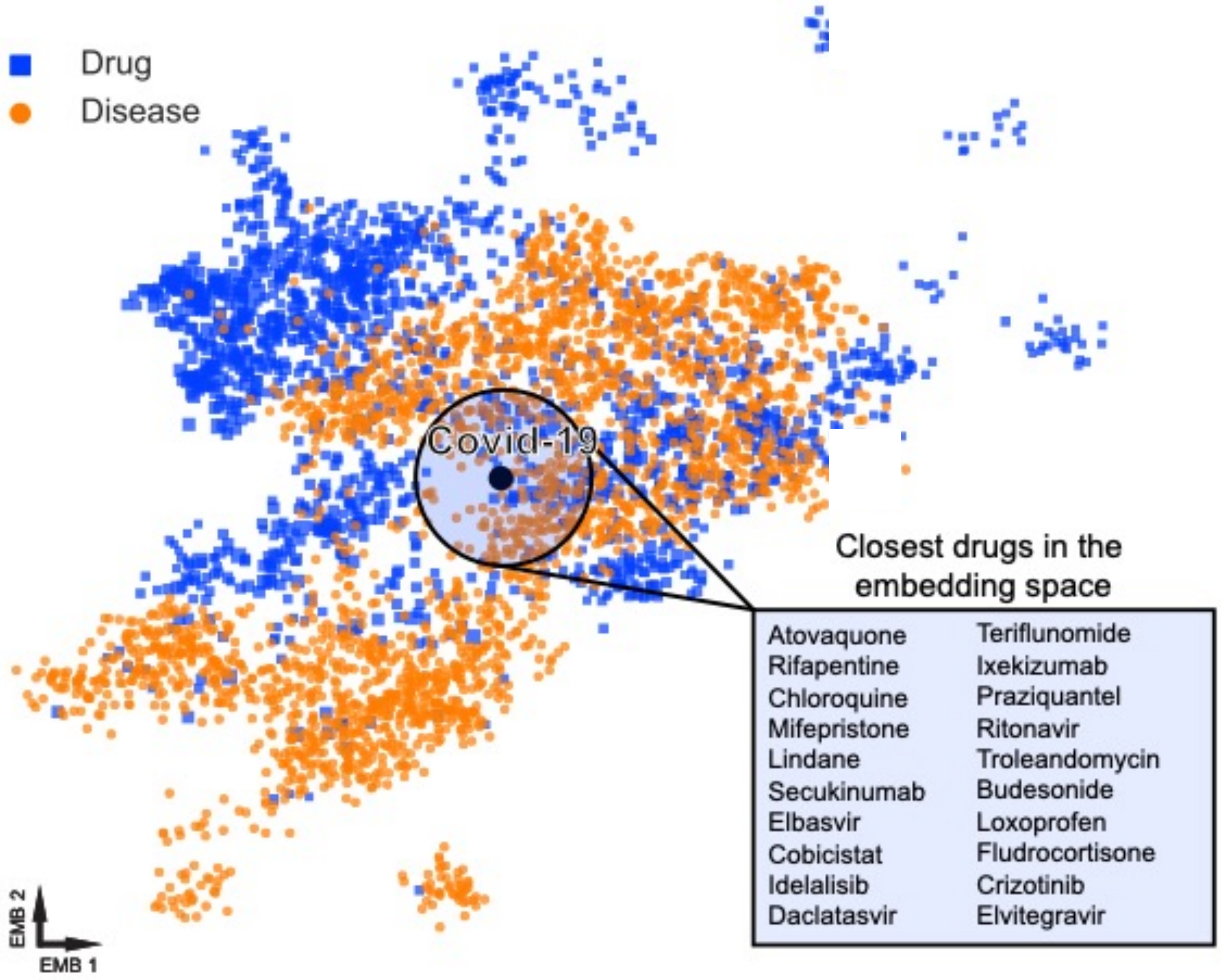
Idea: Use the paradigm of embeddings to operationalize the concept of closeness in pharmacological space

Computational Setup

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



Embedding Space

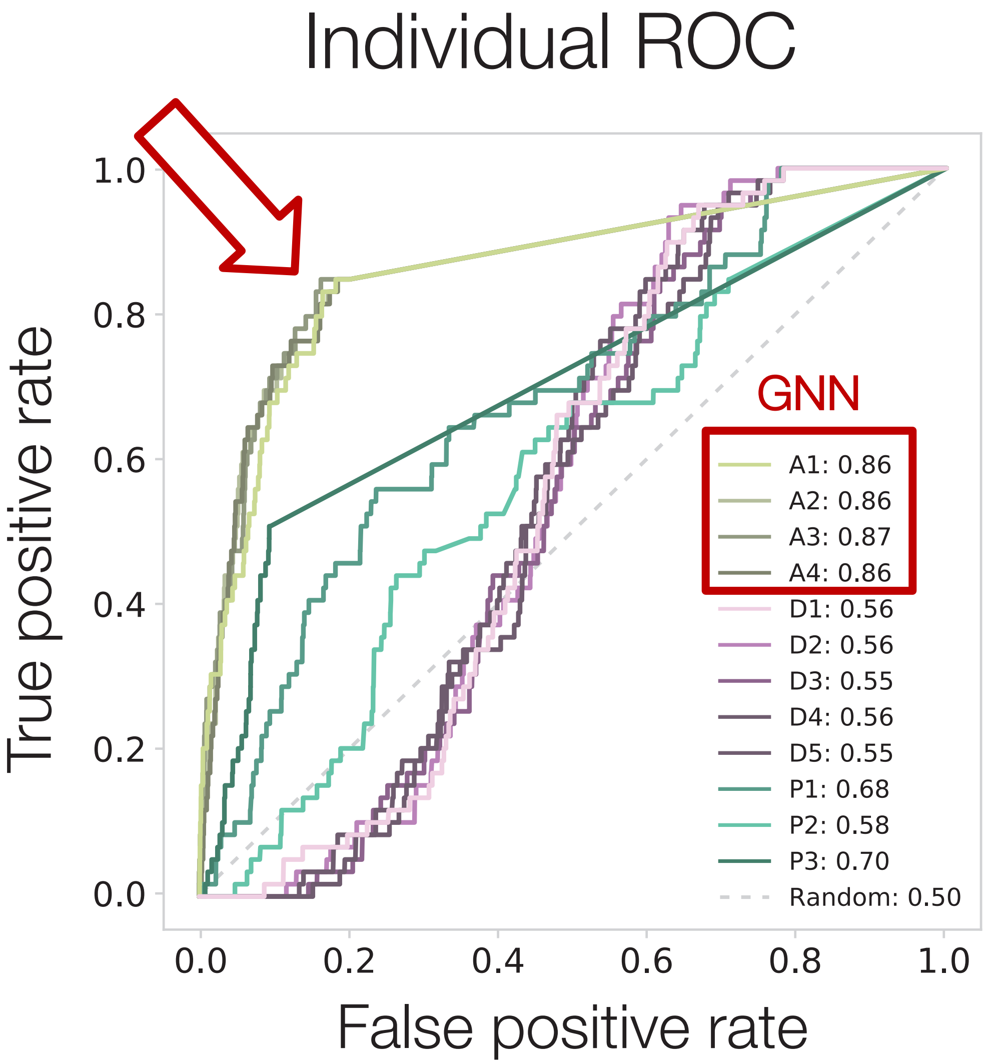


Closest drugs in the embedding space

Atovaquone
Rifapentine
Chloroquine
Mifepristone
Lindane
Secukinumab
Elbasvir
Cobicistat
Idelalisib
Daclatasvir

Teriflunomide
Ixekizumab
Praziquantel
Ritonavir
Troleandomycin
Budesonide
Loxoprofen
Fludrocortisone
Crizotinib
Elvitegravir

Results: COVID-19 Repurposing



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

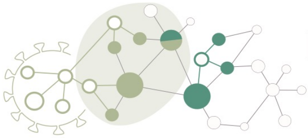
The best individual ROC curves are obtained by the GNN methods

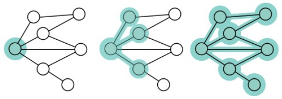
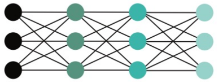
The second-best performance is provided by the proximity P3. Close behind is P1 with AUC = 0.68 and AUC = 0.58

Diffusion methods offer ROC between 0.55-0.56

Final Prediction Model – Part #1

Input Data	Methods		Outcomes

Human Interactome
N = 18,508 proteins
L = 332,749 PPIs
		Network Proximity
		3 pipelines
			Infected
			Tissues/Organs

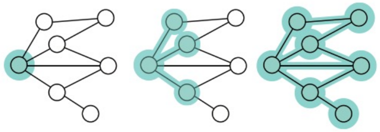
SARS-COV2 targets 320 human proteins Gordon et al, 2020		Network Diffusion 5 pipelines	Comorbidity
Drug Targets 7,591drugs 4,187 drug targets DrugBank		AI Prioritization 4 pipelines	Drug Repurposing & Validation

Final Prediction Model – Part #2

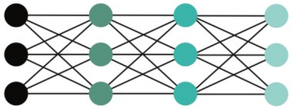
- A COVID-19 treatment can not be derived from the arsenal of therapies approved for specific diseases
- Repurposing strategies focus on drugs previously approved for other pathogens, or on drugs that target the human proteins to which viral proteins bind.
- Most approved drugs do not target directly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

Methods

Network Proximity
3 pipelines



Network Diffusion
5 pipelines



AI Prioritization
4 pipelines

Final Prediction Model – Part #3

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

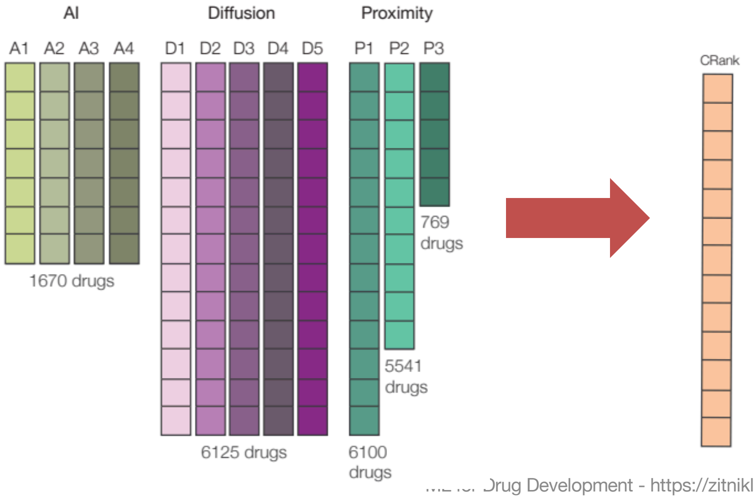
The combined performance of the AI methods is 0.87, the same as A3.

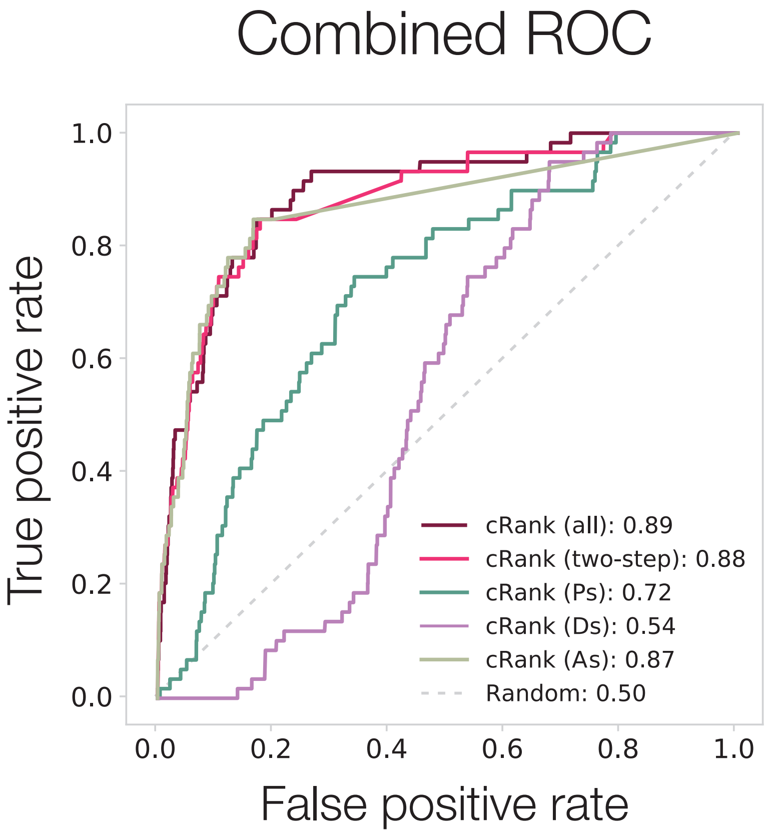
Improvement for proximity pipelines: 0.70 → 0.72.

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

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

Individual pipelines offer complementary information harnessed by the combined ranking.





Predicted Drug Candidates

○ # of Clinical trials from ClinicalTrials.gov

Joseph Loscalzo

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

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

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

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

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

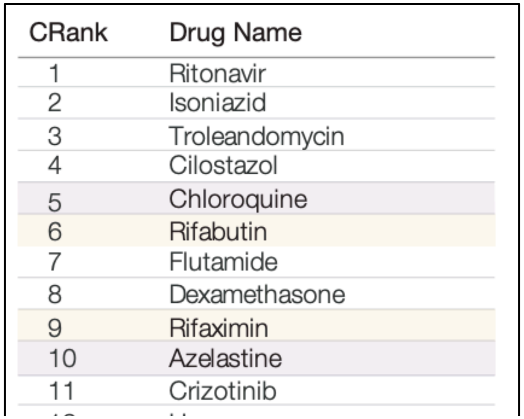
Anti-proteasomal drugs (e.g., bortezomib, carfilzomib)

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

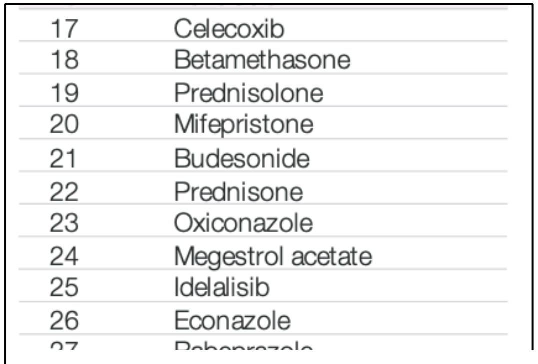
	Drug	C-rank
20	Ritonavir	1
	Isoniazid	2
	Troleandomycin	3
	Cilostazol	4
76	Chloroquine	5
	Rifabutin	6
	Flutamide	7
2	Dexamethasone	8
	Rifaximin	9
	Azelastine	10
	Folic Acid	16
	Rabeprazole	27
	Methotrexate	32
	Digoxin	33
	Theophylline	34
	Fluconazole	41
	Aminoglutethimide	42
67	Hydroxychloroquine	44
	Methimazole	47
1	Ribavirin	49
1	Omeprazole	50
	Bortezomib	53
	Leflunomide	54
	Dimethylfumarate	55
4	Colchicine	57
	Quercetin	63
	Mebendazole	67
	Mesalazine	69
	Pentamidine	92
	Verapamil	98
	Melatonin	109
	Griseofulvin	112
	Auranofin	118
1	Atovaquone	124
	Montelukast	131
	Romidepsin	138
1	Cobicistat	141
17	Lopinavir	146
	Pomalidomide	155
	Sulfinpyrazone	157
1	Levamisole	161
	Calcitriol	164
1	Interferon-β-1a	173
	Praziquantel	176
1	Ascorbic acid	195
	Fluvastatin	199
1	Interferon-β-1b	203
	Selegiline	206
1	Deferoxamine	227
	Ivermectin	235
1	Atorvastatin	243
	Mitoxantrone	250
	Glyburide	259
2	Thalidomide	262
	Sulfanilamide	265
	Hydralazine	269
	Gemfibrozil	281
4	Ruxolitinib	284
	Propranolol	297
	Carbamazepine	301
	Doxorubicin	309
	Levothyroxine	329
	Dactinomycin	335
	Tenofivir	338
	Tadalafil	339
	Doxazosin	367
	Rosiglitazone	397
	Aminolevulinic acid	398
	Nitroglycerin	418
	Metformin	457
1	Nintedanib	466
	Allopurinol	471
	Ponatinib	491
1	Sildenafil	493
	Dapagliflozin	504
	Nitroprusside	515
	Cinacalcet	553
	Mexiletine	559
	Sitagliptin	706
	Carfilzomib	765
1	Azithromycin	786

Experimental Validation of Predictions

National Emerging Infectious Diseases Laboratories (NEIDL)



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



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

Ranked lists of drugs

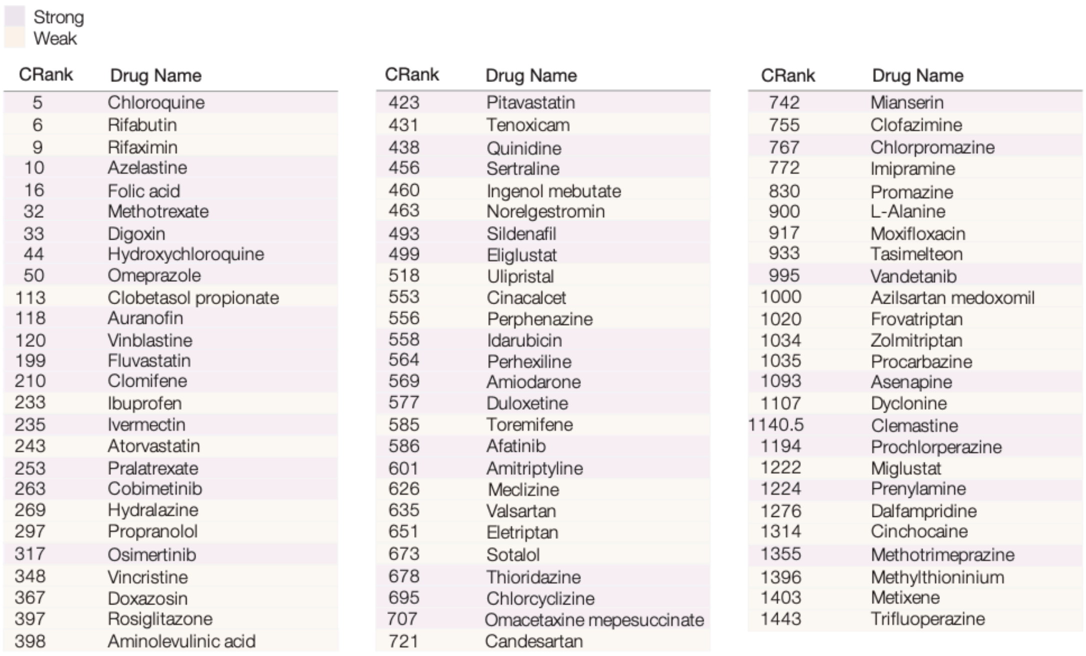
New algorithms:
Prioritizing Network Communities, *Nature Communications* 2018
Subgraph Neural Networks, *NeurIPS* 2020
Graph Meta Learning via Local Subgraphs, *NeurIPS* 2020

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

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

Results: Network Drugs

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

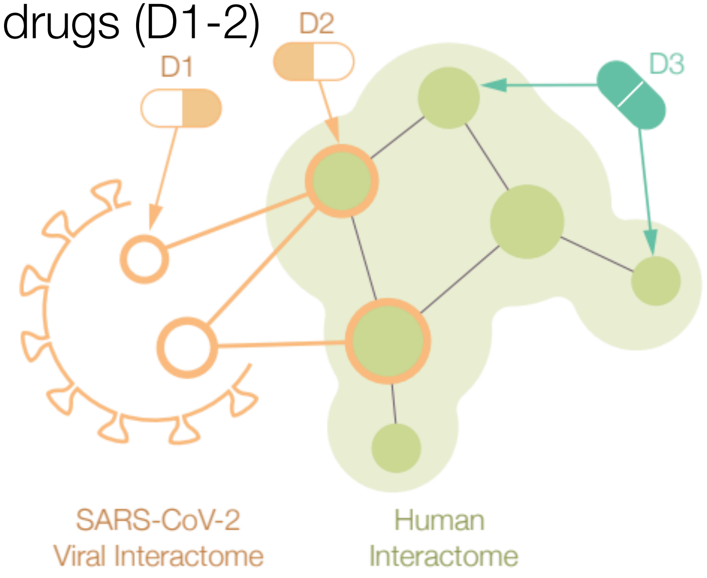


▨ Strong
□ Weak

CRank	Drug Name	CRank	Drug Name	CRank	Drug Name
5	Chloroquine	423	Pitavastatin	742	Mianserin
6	Rifabutin	431	Tenoxicam	755	Clofazimine
9	Rifaximin	438	Quinidine	767	Chlorpromazine
10	Azelastine	456	Sertraline	772	Imipramine
16	Folic acid	460	Ingenol mebutate	830	Promazine
32	Methotrexate	463	Norelgestromin	900	L-Alanine
33	Digoxin	493	Sildenafil	917	Moxifloxacin
44	Hydroxychloroquine	499	Eliglustat	933	Tasimelteon
50	Omeprazole	518	Ulipristal	995	Vandetanib
113	Clobetasol propionate	553	Cinacalcet	1000	Azilsartan medoxomil
118	Auranofin	556	Perphenazine	1020	Frovatriptan
120	Vinblastine	558	Idarubicin	1034	Zolmitriptan
199	Fluvastatin	564	Perhexiline	1035	Procarbazine
210	Clomifene	569	Amiodarone	1093	Asenapine
233	Ibuprofen	577	Duloxetine	1107	Dyclonine
235	Ivermectin	585	Toremifene	1140.5	Clemastine
243	Atorvastatin	586	Afatinib	1194	Prochlorperazine
253	Pralatrexate	601	Amitriptyline	1222	Miglustat
263	Cobimetinib	626	Meclizine	1224	Prenylamine
269	Hydralazine	635	Valsartan	1276	Dalfampridine
297	Propranolol	651	Eletriptan	1314	Cinchocaine
317	Osimertinib	673	Sotalol	1355	Methotrimeprazine
348	Vincristine	678	Thioridazine	1396	Methylthioninium
367	Doxazosin	695	Chlorcyclizine	1403	Metixene
397	Rosiglitazone	707	Omacetaxine mepesuccinate	1443	Trifluoperazine
398	Aminolevulinic acid	721	Candesartan		

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

Direct target drugs (D1-2)



Network drugs (D3)

Outline

✓ Overview and introduction

✓ Part 1: Virtual drug screening and drug repurposing

Part 2: Adverse drug effects, drug-drug interactions

Part 3: Clinical trial site identification, patient recruitment

Part 4: Molecule optimization, molecular graph generation, multimodal graph-to-graph translation

Part 5: Molecular property prediction and transformers

Demos, resources, wrap-up & future directions