Prescient
Design
A Genentech Accelerator

Combinatorial prediction of therapeutic targets using a causally inspired GNN

Application of Deep Learning on Graphs ETH

Guadalupe Gonzalez
Senior ML Scientist, Frontier Research

Genentech
A Member of the Roche Group

Collaborators

Marinka Zitnik

Michael Bronstein

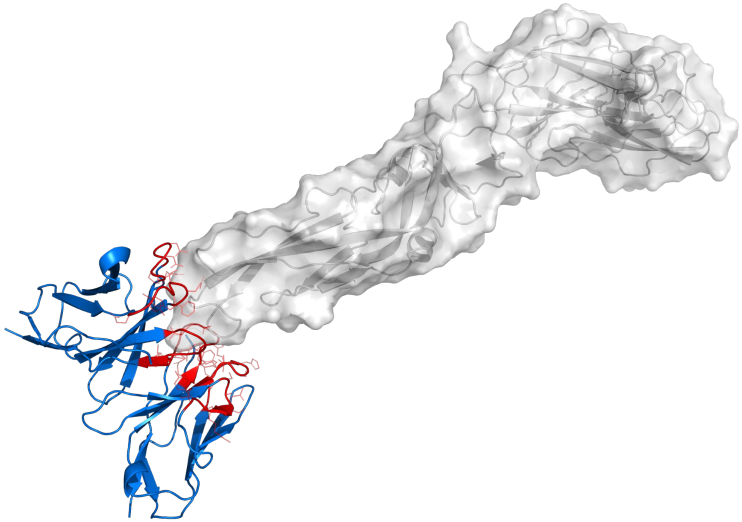
Kirill Veselkov

Isuru Herath
Domen Mohorcic

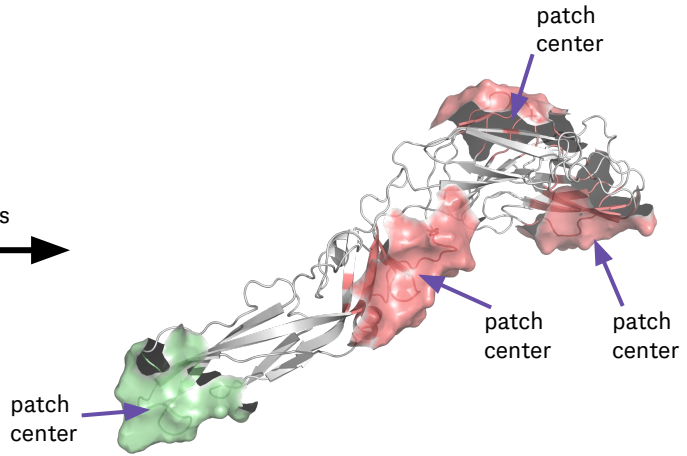
Introduction

A familiar use case | Epiphany Affinity

Prescient Design
A Genentech Accelerator



Extract antigen patches



(1) do they bind?

(2) is this the epitope?

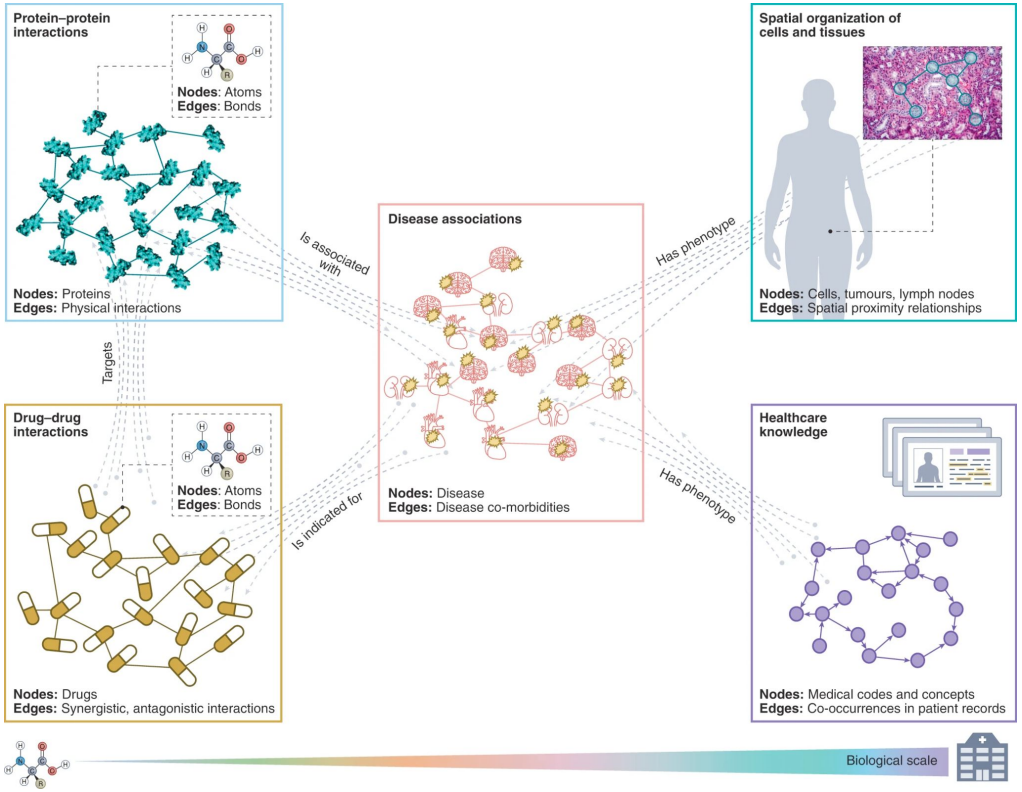
(3) is this the paratope?

Images courtesy of Jan Ludwiczak

Genentech
A Member of the Roche Group

Zooming out | Graph representations are everywhere in biology

Prescient Design
A Genentech Accelerator



Li et al. 2022

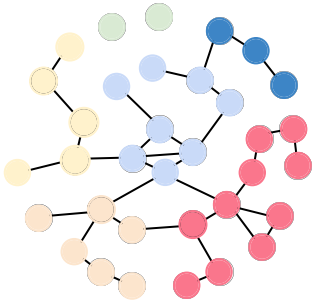
Genentech
A Member of the Roche Group

Current GNN paradigm | Message passing

Prescient Design
A Genentech Accelerator

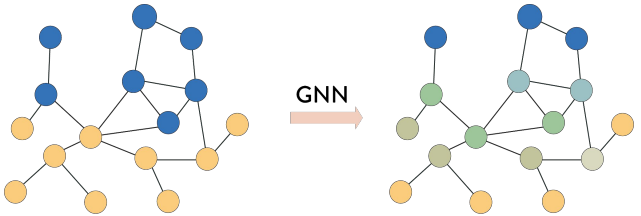
Biological principles that deep learning on graphs naturally includes:

Local hypothesis: interacting entities are more similar than non-interacting entities



PPI

Homophily principle: aggregates messages from neighbors to compute enhanced representations under the assumption that neighbors are similar



Li et al. 2022

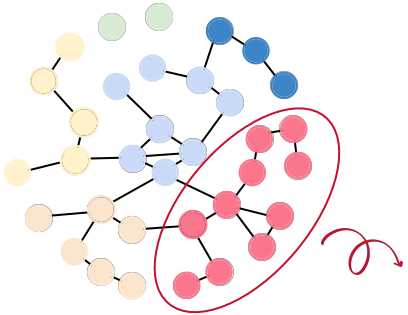
Genentech
A Member of the Roche Group

Current GNN paradigm | Message passing

Prescient Design
A Genentech Accelerator

Biological principles that deep learning on graphs naturally includes:

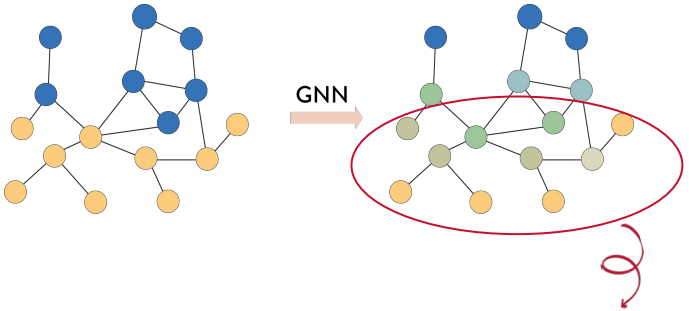
Disease module hypothesis: cellular components tend to cluster in the same network neighborhood



Disease-associated proteins

PPI

Community detection: predict communities based on node embedding similarities which are higher for connected nodes → it will typically predict nodes closer in the network as belonging to the same community



Community detection influenced by network proximity

Li et al. 2022

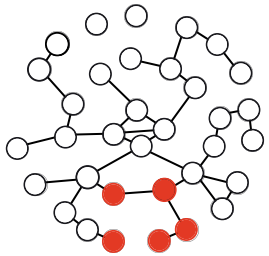
Genentech
A Member of the Roche Group

Current GNN paradigm | Message passing

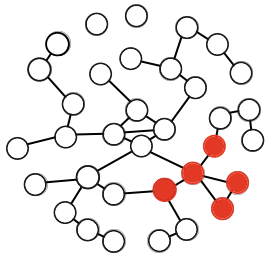
Prescient Design
A Genentech Accelerator

Biological principles that deep learning on graphs naturally includes:

Shared-component hypothesis: diseases driven by perturbations of the same components (or close in the network) are phenotypically similar



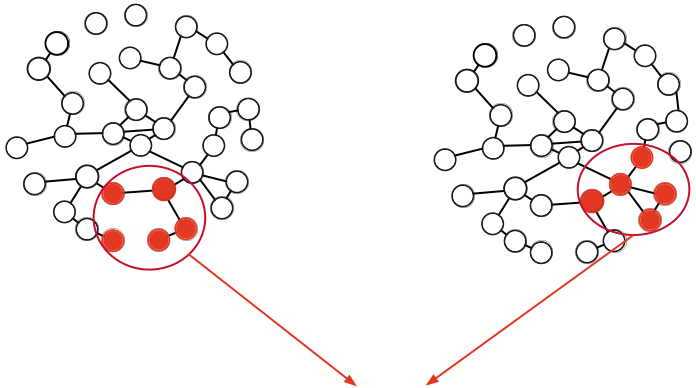
Disease A



Disease B

PPI

Subgraph representations: similar representations for overlapping or similar subpgrahs



Similar representations

Li et al. 2022

A brief history of therapeutics discovery:

Phenotype-driven vs target-driven approaches

Phenotype-driven vs target-driven drug discovery

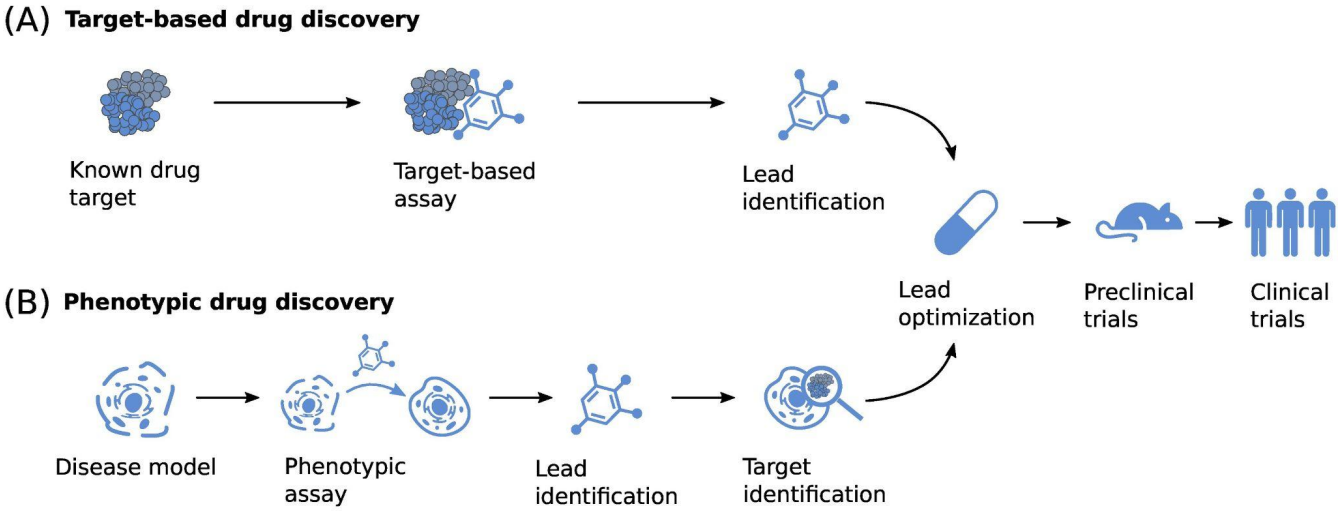


Image source = Krentzel, D., Shorte, S.L. and Zimmer, C. (2023) 'Deep learning in image-based phenotypic drug discovery', *Trends in Cell Biology*, 33(7), pp. 538–554. doi:10.1016/j.tcb.2022.11.011.

Phenotype-driven vs target-driven drug discovery

Prescient Design
A Genentech Accelerator

Majority of the first-in-class drugs approved by the US FDA between 1999-2008 were discovered empirically without a drug target hypothesis

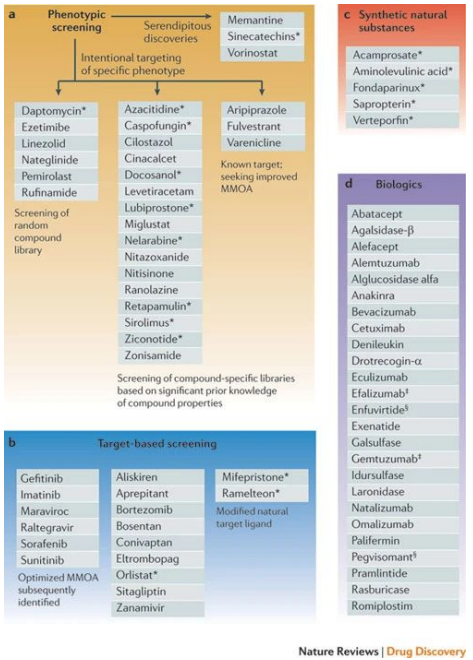


Image source = Swinney, D., Anthony, J. How were new medicines discovered?. *Nat Rev Drug Discov* 10, 507–519 (2011). https://doi.org/10.1038/nrd3480

Genentech
A Member of the Roche Group

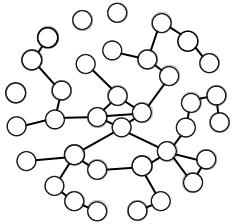
Problem formulation

Prescient Design
A Genentech Accelerator

Problem formulation

healthy

PPI network



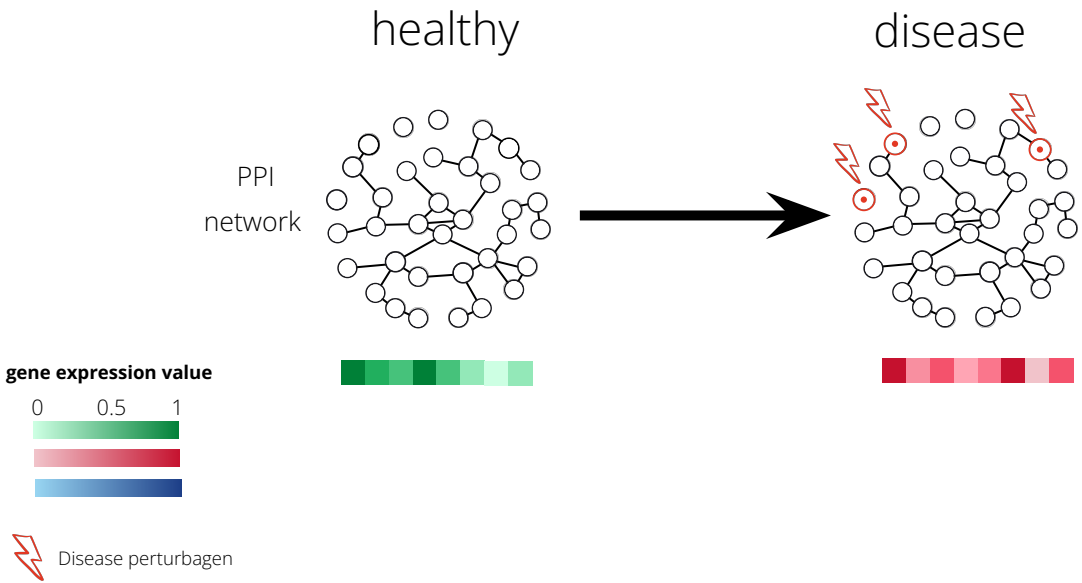
gene expression value

0 0.5 1

Genentech
A Member of the Roche Group

Prescient Design
A Genentech Accelerator

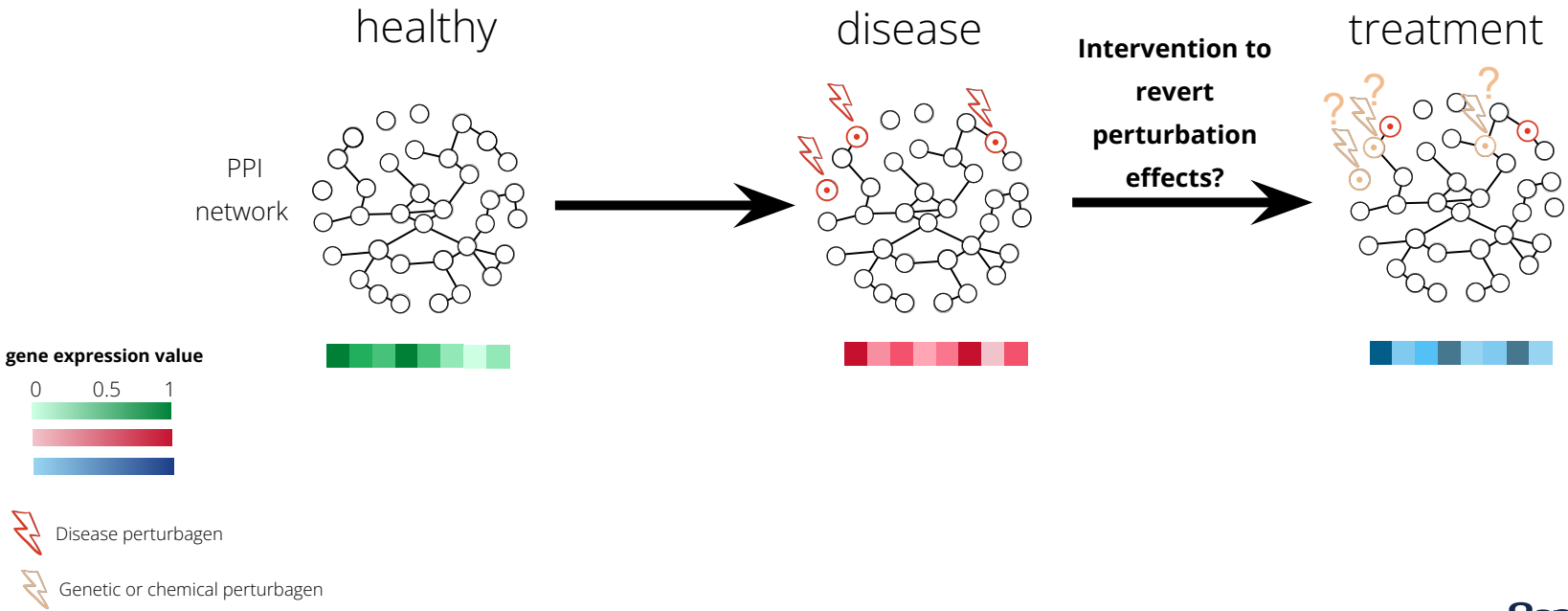
Problem formulation



Genentech
A Member of the Roche Group

Prescient Design
A Genentech Accelerator

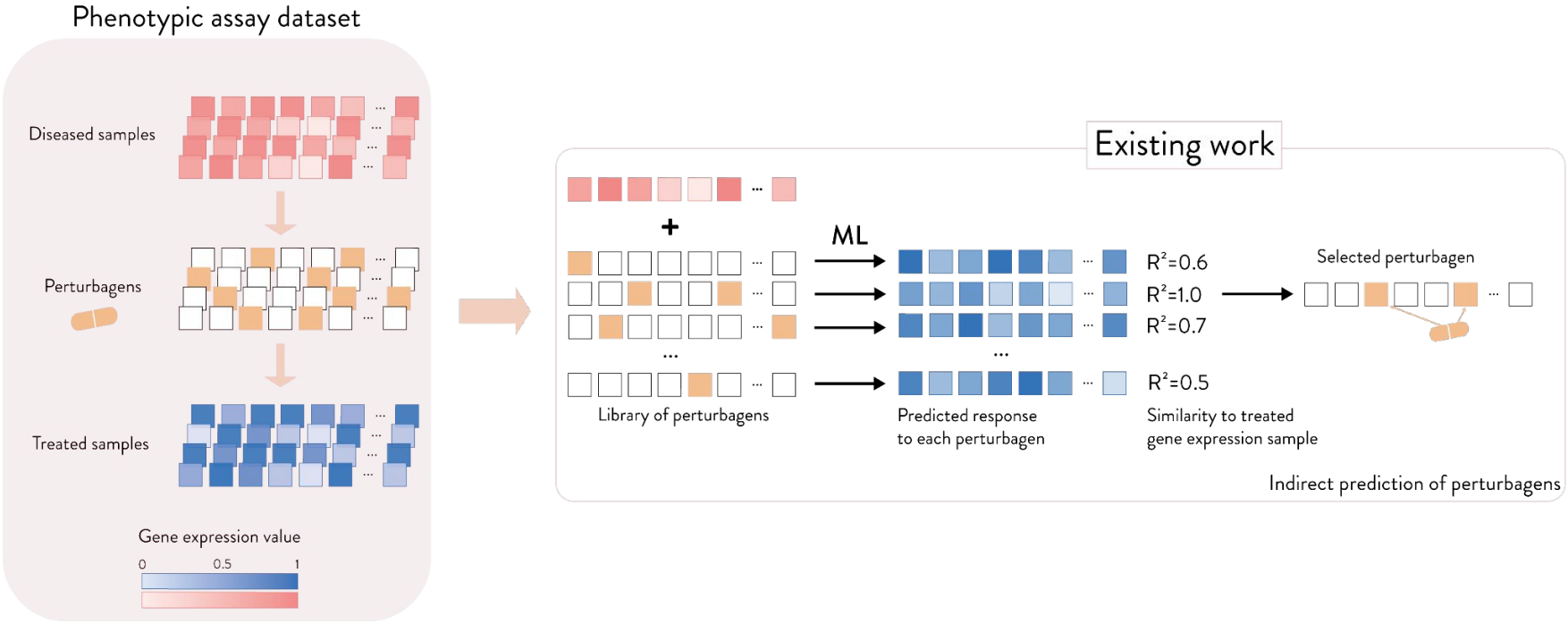
Problem formulation



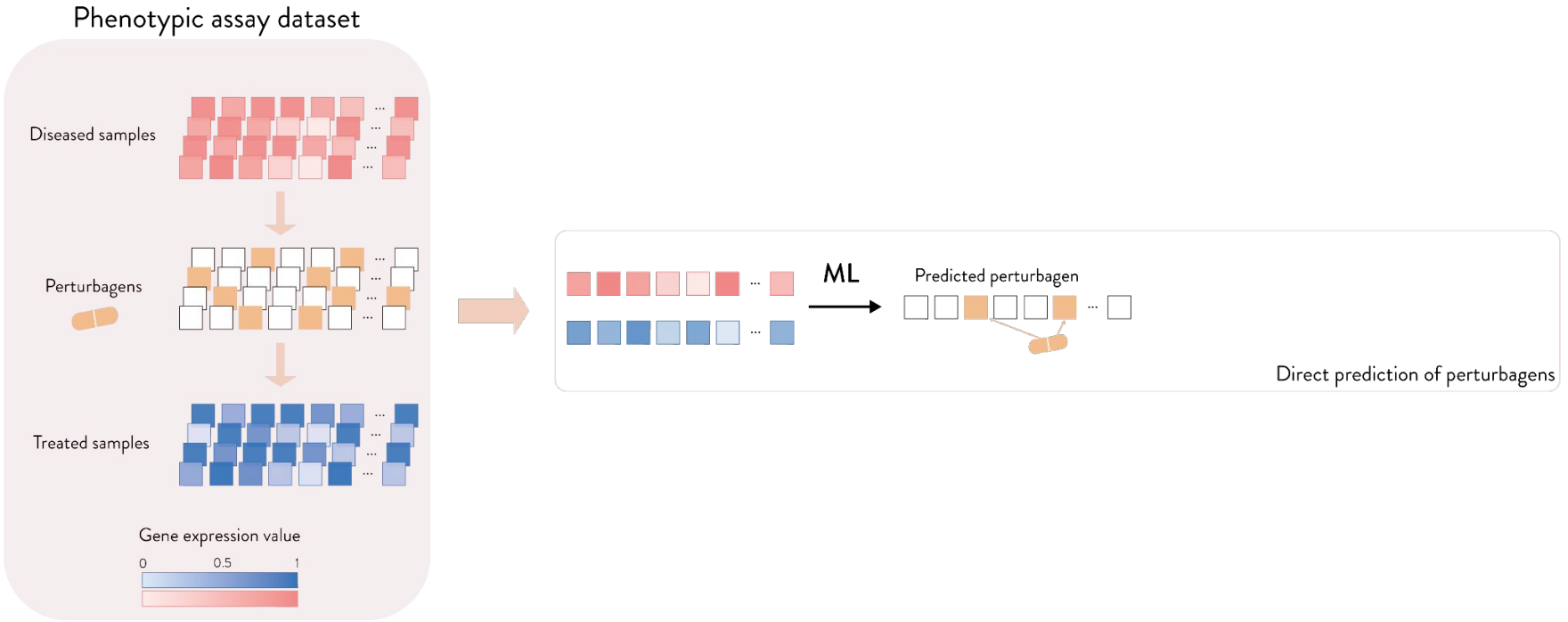
Genentech
A Member of the Roche Group

Existing work

Existing work



Existing work



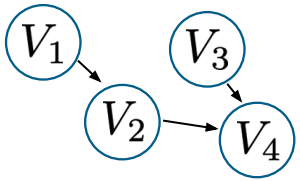
Combinatorial prediction of therapeutic targets using a causally inspired neural network

Causal inference primer

Structural **C**ausal **M**odels

Consists of a **causal graph** and **structural equations**

causal model: mathematical abstraction that quantitatively describes the causal relations between variables.



exogenous variables

U

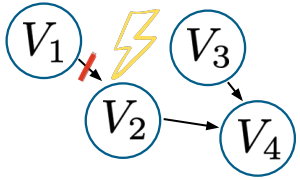
quantify causal effects shown in causal graph

$$\begin{cases} V_1 = f(U_{V_1}) \\ V_2 = f(U_{V_2}, V_1) \\ V_3 = f(U_{V_3}) \\ V_4 = f(U_{V_4}, V_2, V_3) \end{cases}$$

intervention

$do(V_2) = v_2$

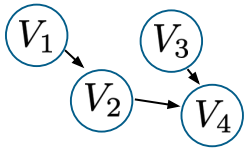
"do" operator



$$\begin{aligned} V_1 &= f(U_{V_1}) \\ V_2 &= v_2 \\ V_3 &= f(U_{V_3}) \\ V_4 &= f(U_{V_4}, V_2, V_3) \end{aligned}$$

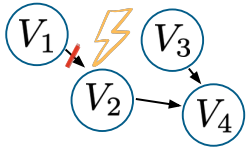
causal Markov condition

Causal inference primer



$$V_1 = U_{V_1}$$
$$V_2 = AND(U_{V_2}, V_1)$$
$$V_3 = U_{V_3}$$
$$V_4 = AND(U_{V_4}, V_2, V_3)$$

$U_{V_2} = U_{V_4} = 1$



$$V_1 = U_{V_1}$$
$$V_2 = v_2$$
$$V_3 = U_{V_3}$$
$$V_4 = AND(U_{V_4}, V_2, V_3)$$

$U_{V_2} = U_{V_4} = 1$

V_1	V_2	V_3	V_4
1	1	0	0
1	1	1	1
0	0	1	0
1	1	1	1
0	0	0	0
1	1	0	1

$P(V_4|V_2, V_3)$

observational distribution
$P^o(V_1, V_2, V_3, V_4)$

=

?

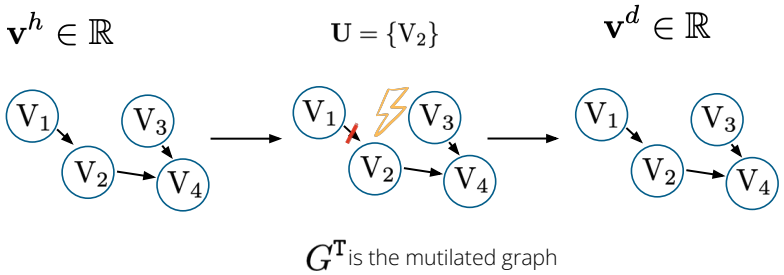
V_1	V_2	V_3	V_4
0	1	0	0
1	1	1	1
1	0	1	0
0	1	1	1
1	0	0	0
0	1	0	0

$P(V_4|do(V_2), V_3)$

interventional distribution
$P^i(V_1, V_2, V_3, V_4)$

Problem formulation

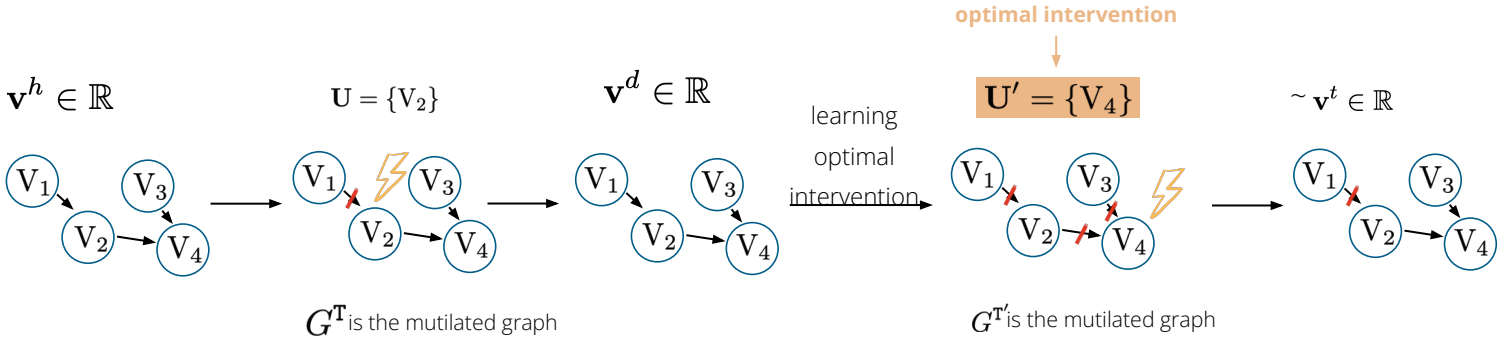
Let $\mathtt{M} =< \mathbf{E}, \mathbf{V}, \mathcal{F}, P(\mathbf{E}) >$ be an SCM associated with causal graph G. Dataset: $\mathcal{T} = \{\mathtt{T}_1, \ldots, \mathtt{T}_m\}$



$\mathtt{T} =< \mathbf{v}^h, \mathbf{U}, \mathbf{v}^d >$ **disease intervention data**

Problem formulation

Let $\mathtt{M} =< \mathbf{E}, \mathbf{V}, \mathcal{F}, P(\mathbf{E}) >$ be an SCM associated with causal graph G. Dataset: $\mathcal{T} = \{\mathtt{T}_1, \ldots, \mathtt{T}_m\}$



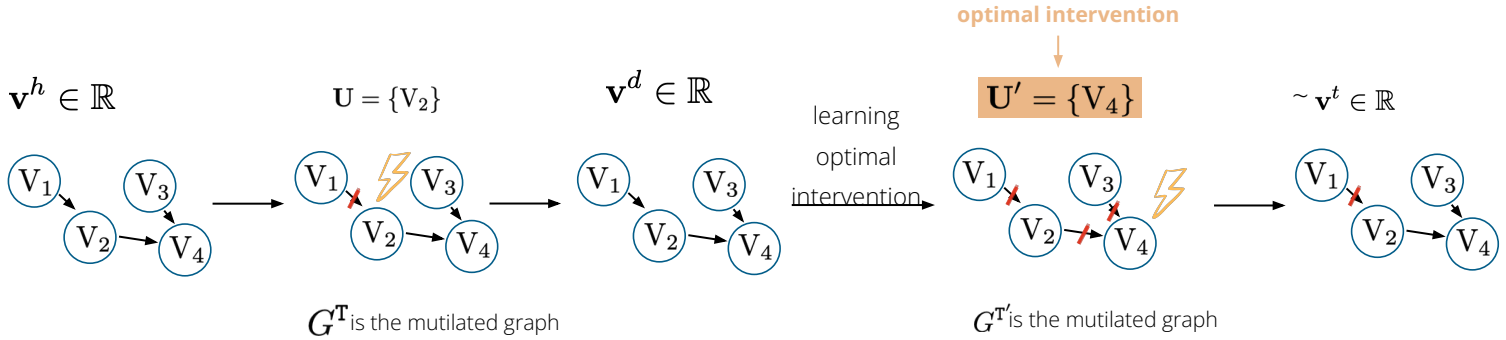
$\mathtt{T} =< \mathbf{v}^h, \mathbf{U}, \mathbf{v}^d >$

disease intervention data

combinatorial prediction

Problem formulation

Let $\mathtt{M} =< \mathbf{E}, \mathbf{V}, \mathcal{F}, P(\mathbf{E}) >$ be an SCM associated with causal graph G. Dataset: $\mathcal{T} = \{\mathtt{T}_1, \ldots, \mathtt{T}_m\}$



$\mathtt{T} =< \mathbf{v}^h, \mathbf{U}, \mathbf{v}^d >$

disease intervention data

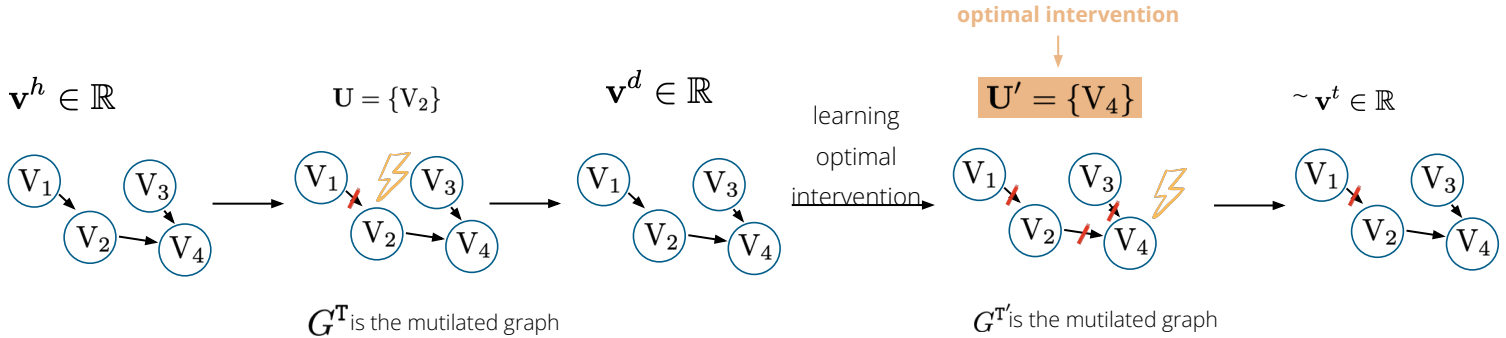
combinatorial prediction

$$argmax_{\mathbf{U}'} P^{G^{\mathtt{T}}}(\mathbf{V} = \mathbf{v}^t | do(\mathbf{U}'))$$

"do" operator

Problem formulation

Let $\mathtt{M} =< \mathbf{E}, \mathbf{V}, \mathcal{F}, P(\mathbf{E}) >$ be an SCM associated with causal graph G. Dataset: $\mathcal{T} = \{\mathtt{T}_1, \ldots, \mathtt{T}_m\}$



$\mathtt{T} =< \mathbf{v}^h, \mathbf{U}, \mathbf{v}^d >$ **disease intervention data**

combinatorial prediction

which in the absence of unobserved confounders is:

$$argmax_{\mathbf{U}'} P^{G^{\mathtt{T}}}(\mathbf{V} = \mathbf{v}^t | do(\mathbf{U}'))$$

"do" operator

$$argmax_{\mathbf{U}'} P^{G^{\mathtt{T}'}}(\mathbf{V} = \mathbf{v}^t | \mathbf{U}')$$

Problem formulation

Prescient Design
A Genentech Accelerator

causal inference

The goal is to find the perturbation set $\mathbf{U}'$ with the highest likelihood of achieving treated state $\mathbf{v}^t \in \mathbb{R}$

$$argmax_{\mathbf{U}'} P^{G^{\intercal}}(\mathbf{V} = \mathbf{v}^t | do(\mathbf{U}'))$$

"do" operator

which under markovianity is:

our goal $$argmax_{\mathbf{U}'} P^{G^{\intercal'}}(\mathbf{V} = \mathbf{v}^t | \mathbf{U}')$$

sets of random variables

Genentech
A Member of the Roche Group

Problem formulation

Prescient Design
A Genentech Accelerator

causal inference

The goal is to find the perturbation set $\mathbf{U}'$ with the highest likelihood of achieving treated state $\mathbf{v}^t \in \mathbb{R}$

$$argmax_{\mathbf{U}'} P^{G^{\mathbf{T}}}(\mathbf{V} = \mathbf{v}^t | do(\mathbf{U}'))$$

"do" operator

which under markovianity is:

our goal $$argmax_{\mathbf{U}'} P^{G^{\mathbf{T}'}}(\mathbf{V} = \mathbf{v}^t | \mathbf{U}')$$

sets of random variables

graph representation learning

$$f : G^{\mathbf{T}'}, \mathrm{x}^d, \mathrm{x}^t \rightarrow argmax_{\mathcal{U}'} P^{G^{\mathbf{T}'}}(\mathrm{x} = \mathrm{x}^t | \mathrm{x}^d, \mathcal{U}')$$

vectors, sets

Intuitively, we need to search the space of all interventions and score how good they are

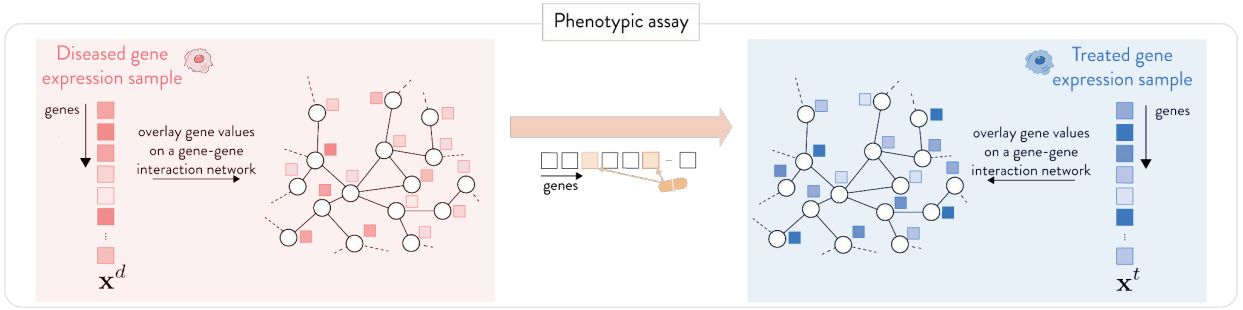
- intractable

What we propose:

- a smart way to search the space of interventions (**intervention discovery module**)
- a way to score each intervention (**response prediction module**)

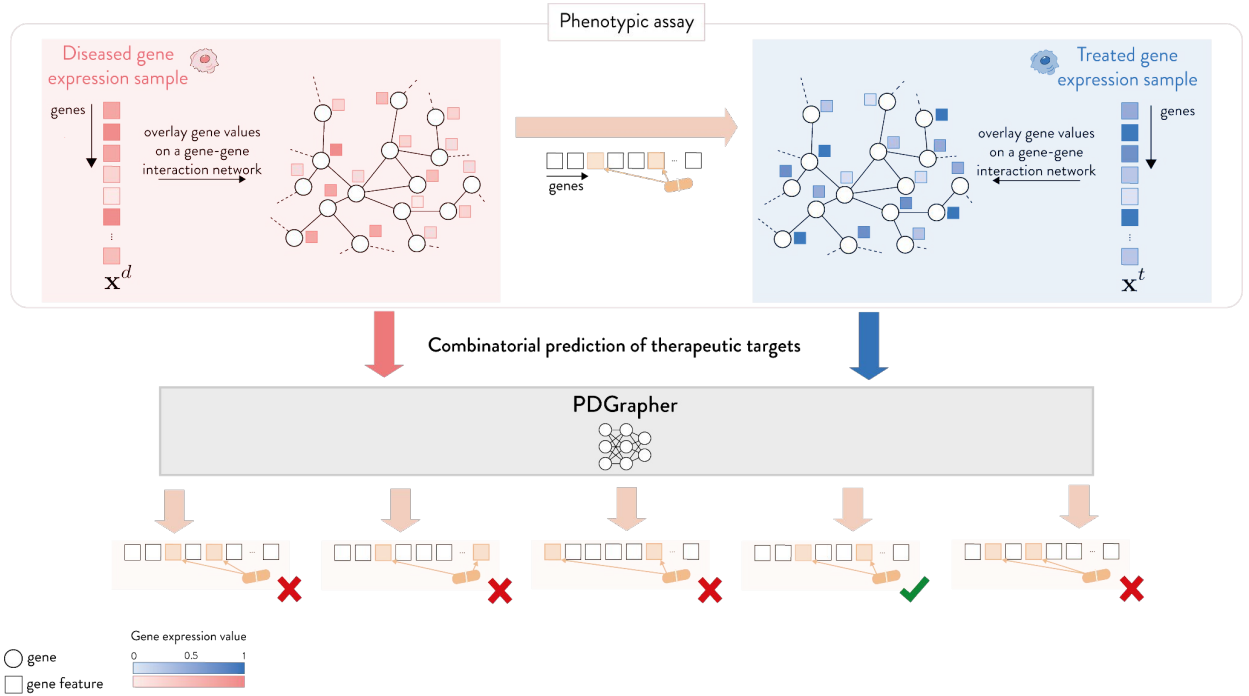
Genentech
A Member of the Roche Group

Our solution: PDGrapher



Prescient Design
A Genentech Accelerator

Our solution: PDGrapher

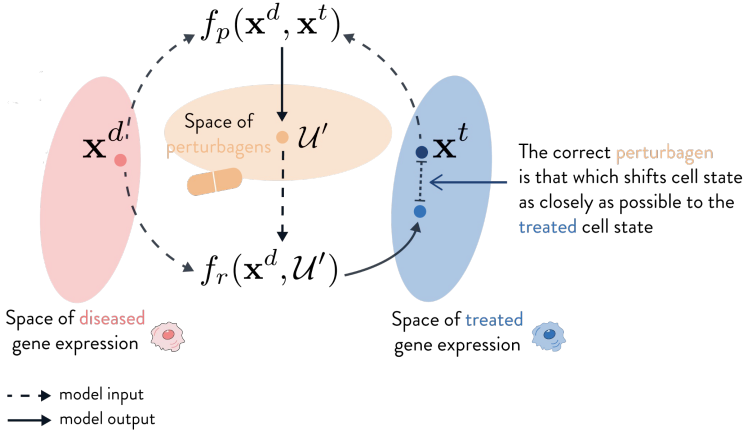


Genentech
A Member of the Roche Group

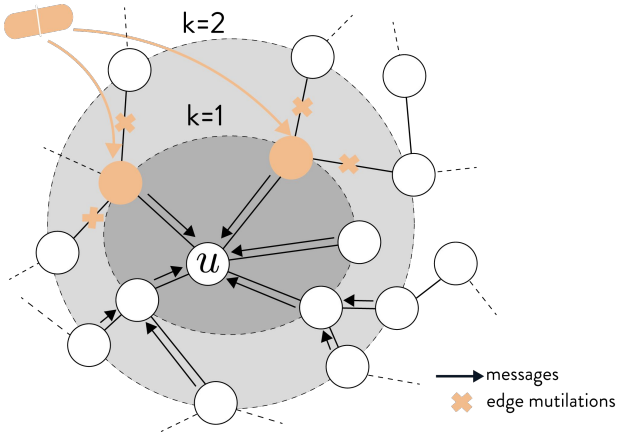
Prescient Design
A Genentech Accelerator

Our solution: PDGrapher

Optimization



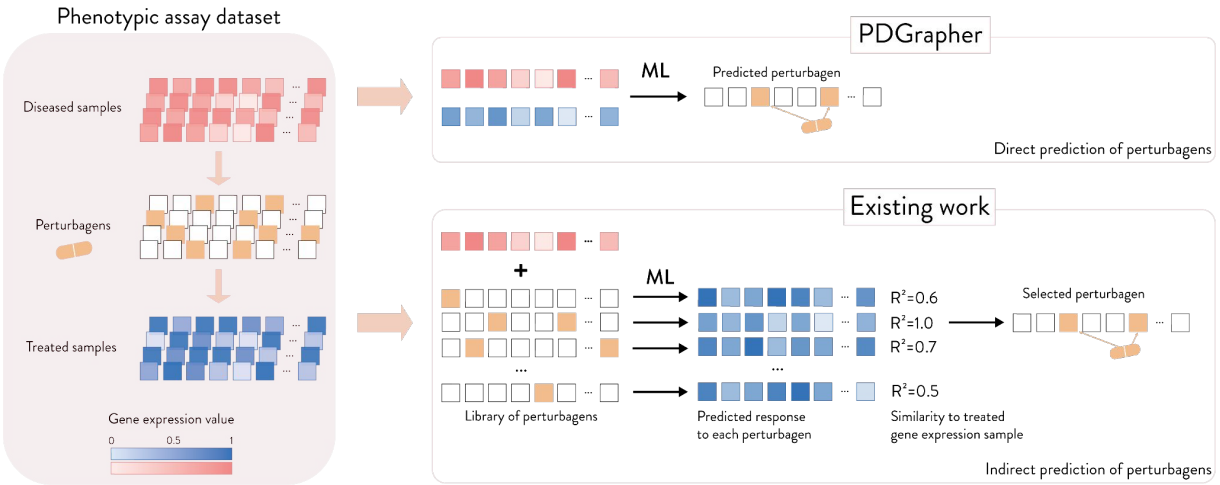
Mutilations in message passing



Genentech
A Member of the Roche Group

Prescient Design
A Genentech Accelerator

Our solution: PDGrapher



PDGrapher

1. Models perturbations explicitly
2. Performs direct prediction of perturbagens

Genentech
A Member of the Roche Group

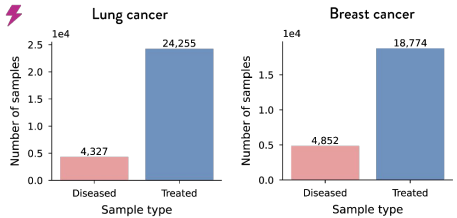
Prescient Design
A Genentech Accelerator

Datasets

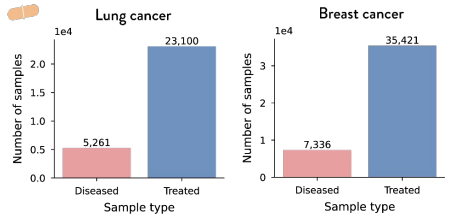
NIH LINCS PROGRAM
LIBRARY OF INTEGRATED NETWORK-BASED CELLULAR SIGNATURES

L1000 dataset

single-gene knockouts



compounds



The Human Reference Protein Interactome Mapping Project

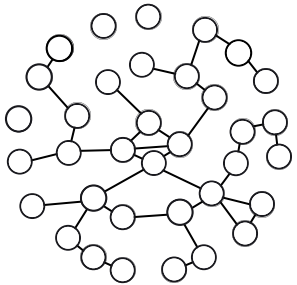
BioGRID 4.4

DISEASE NETWORKS

Uncovering disease-disease relationships through the incomplete interactome

Jörg Menche, Amitabh Sharma, Maksim Kitsak, Susan Dina Ghiassian, Marc Vidal, Joseph Loscalzo, Albert-László Barabási*

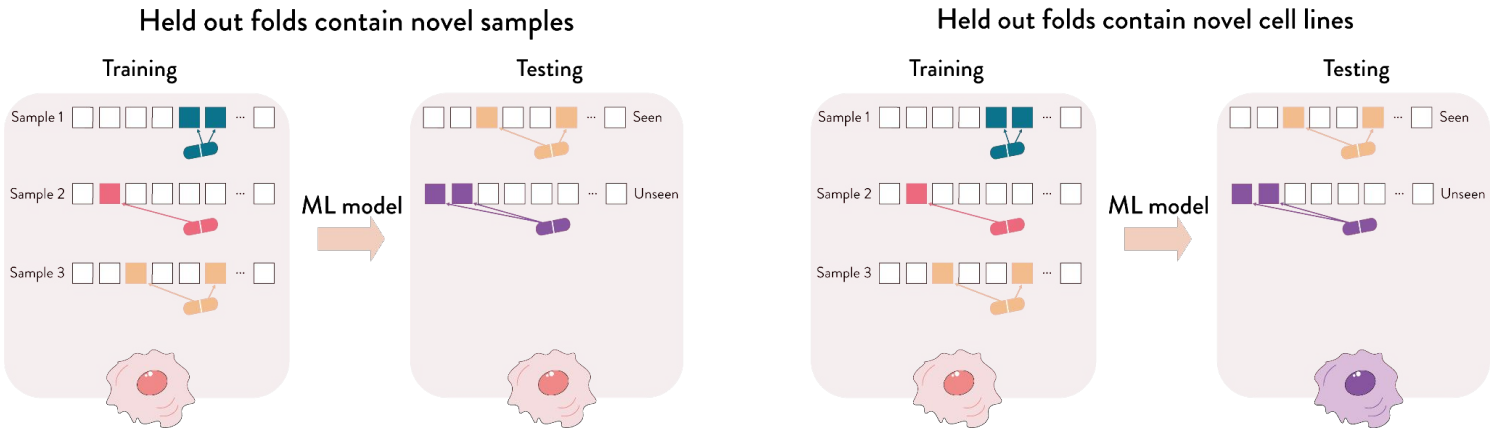
Protein-Protein Interaction network



10,716 nodes
149,270 edges

Genentech
A Member of the Roche Group

Training settings



Baselines

- **Random:** Randomly ranks list of genes
- **Cancer targets:** First M genes are targets of existing cancer drugs, N-M genes are sorted randomly
- **Cancer genes:** First M genes are cancer-associated genes, N-M genes are sorted randomly
- **scGen:** Indirect prediction of perturbagen.

Results - Chemical-PPI, Random splitting

Prediction for a single perturbagen:

Gene 1
Gene 2
Gene 3
Gene 4
Gene 5
...
Gene 10,000

We want ground truth targets to be at the top

Prescient Design
A Genentech Accelerator

Results - Chemical-PPI, Random splitting

Prediction for a single perturbagen:

Gene 1
Gene 2
Gene 3
Gene 4
Gene 5
...
Gene 10,000

If ground truth target is predicted at the top → ranking = 1

Results - Chemical-PPI, Random splitting

Prediction for a single perturbagen:

Gene 1

Gene 2

Gene 3

Gene 4

Gene 5

...

Gene 10,000

If ground truth target is predicted at the bottom → ranking = 0

Prescient Design
A Genentech Accelerator

Results - Chemical-PPI, Random splitting

Prediction for a single perturbagen:

Gene 1 → If ground truth target is predicted at the top → ranking = 1

Gene 2

Gene 3

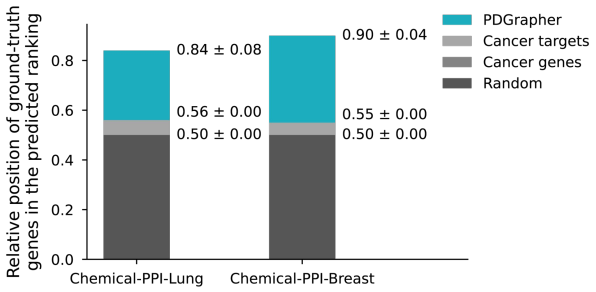
Gene 4

Gene 5

...

Gene 10,000 → If ground truth target is predicted at the bottom → ranking = 0

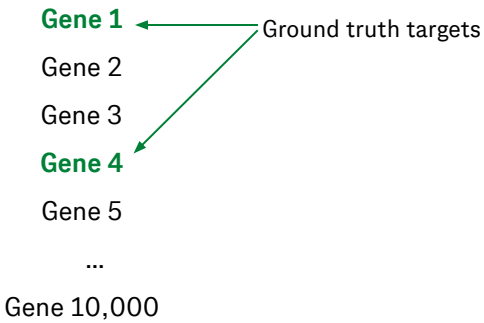
Ranking of ground truth targets



Genentech
A Member of the Roche Group

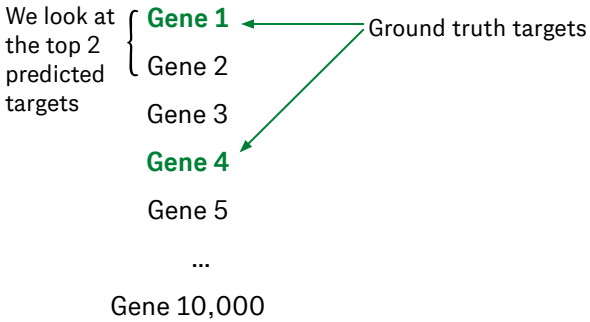
Results - Chemical-PPI, Random splitting

Prediction for a single perturbagen **with 2 targets**:



Results - Chemical-PPI, Random splitting

Prediction for a single perturbagen **with 2 targets**:



Prescient Design
A Genentech Accelerator

Results - Chemical-PPI, Random splitting

Prediction for a single perturbagen **with 2 targets**:

We look at the top 2 predicted targets

- **Gene 1**
- Gene 2
- Gene 3
- **Gene 4**
- Gene 5
- ...
- Gene 10,000

If at least one of the top K predicted genes is a target → partially accurate prediction

Genentech
A Member of the Roche Group

Results - Chemical-PPI, Random splitting

Prediction for a single perturbagen **with 2 targets**:

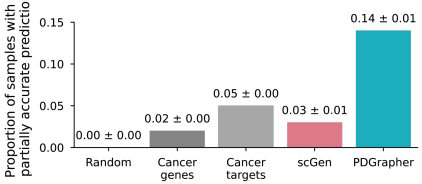
We look at the top 2 predicted targets

- **Gene 1**
- Gene 2
- Gene 3
- **Gene 4**
- Gene 5
- ...
- Gene 10,000

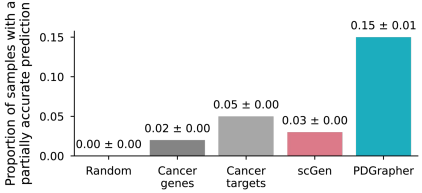
If at least one of the top K predicted genes is a target → partially accurate prediction

Proportion of samples with a partially accurate prediction

Chemical-PPI-Lung



Chemical-PPI-Breast



Prescient Design
A Genentech Accelerator

Results - Chemical-PPI, Random splitting

Prediction for a single perturbagen **with 2 targets**:

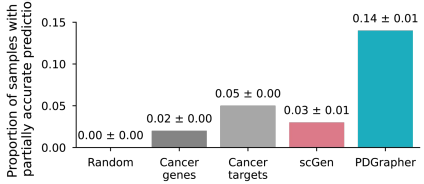
We look at the top 2 predicted targets

- **Gene 1**
- Gene 2
- Gene 3
- **Gene 4**
- Gene 5
- ...
- Gene 10,000

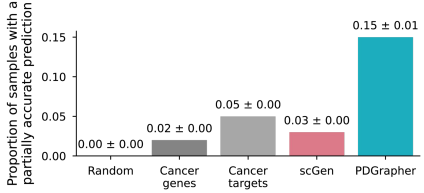
If at least one of the top K predicted genes is a target → partially accurate prediction

Proportion of samples with a partially accurate prediction

Chemical-PPI-Lung

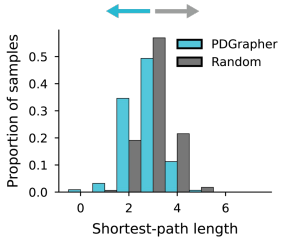


Chemical-PPI-Breast

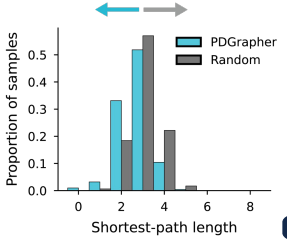


Shortest-path length between ground-truth targets and predicted targets

Chemical-PPI-Lung



Chemical-PPI-Breast



Genentech
A Member of the Roche Group

Prescient Design
A Genentech Accelerator

Results - Chemical-PPI, Random splitting

Prediction for a single perturbagen **with 2 targets**:

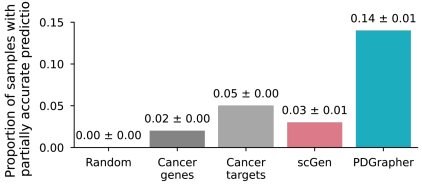
We look at the top 2 predicted targets

- **Gene 1**
- Gene 2
- Gene 3
- **Gene 4**
- Gene 5
- ...
- Gene 10,000

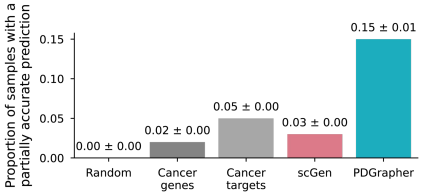
If at least one of the top K predicted genes is a target → partially accurate prediction

Proportion of samples with a partially accurate prediction

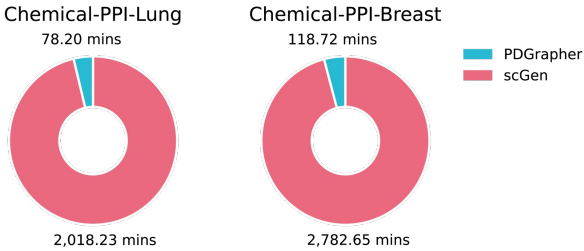
Chemical-PPI-Lung



Chemical-PPI-Breast



Training time of a single model



Our model needs 15x less time to train!

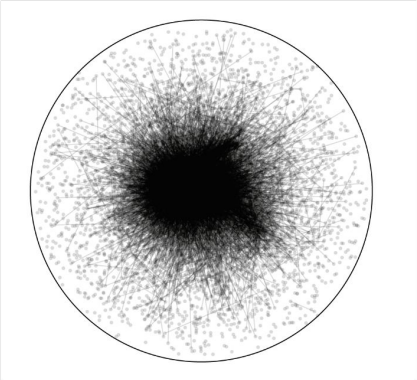
Genentech
A Member of the Roche Group

Prescient Design
A Genentech Accelerator

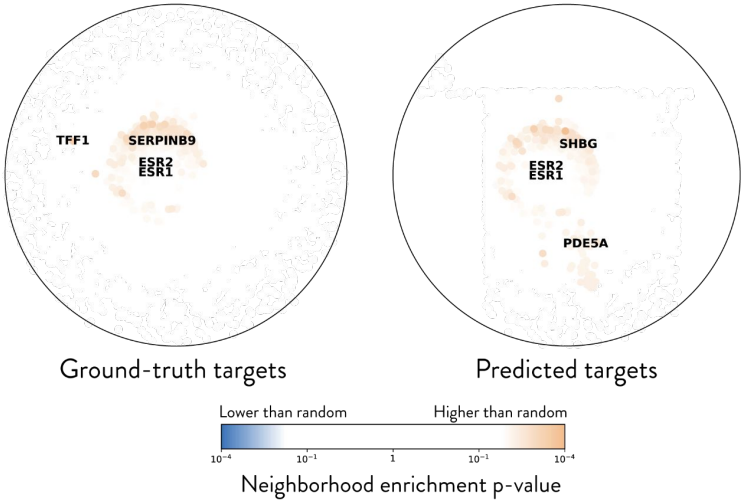
Results - Chemical-PPI, Random splitting

Raloxifene

Gene-gene interation network



Combinatorial prediction of targets: Raloxifene

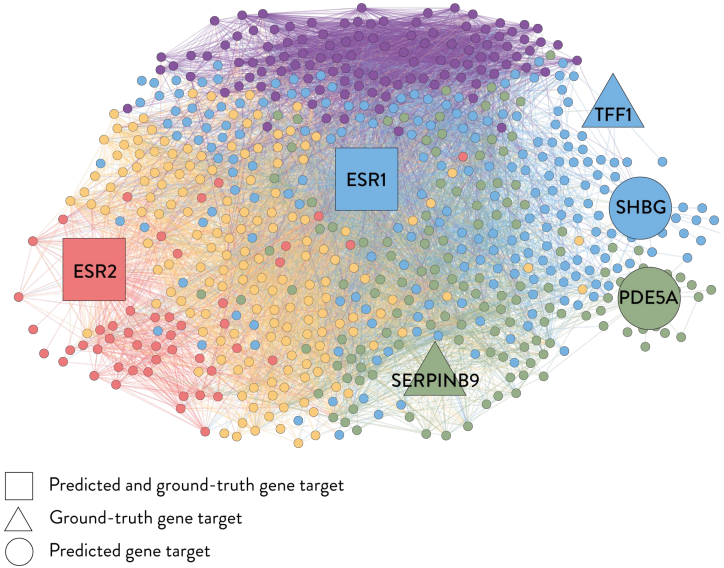


Genentech
A Member of the Roche Group

Prescient Design
A Genentech Accelerator

Results - Chemical-PPI, Random splitting

Raloxifene



Raloxifene is a second-generation selective estrogen receptor modulator. Targets: ESR1, ESR2, SERPINB9, TFF1.

Additional predictions:

- Sex Hormone Binding Globulin (SHBG)
- Phosphodiesterase 5A (PDE5A)

Genentech
A Member of the Roche Group

Takeaways

Prescient Design
A Genentech Accelerator

- Reformulate phenotype-driven lead discovery

- Propose PDGrapher for combinatorial prediction of therapeutic targets, which:
  - Outperforms mechanistic and traditional baselines
  - Performs direct prediction of perturbagens
  - Is computationally efficient
  - Illuminates mode of action of predicted perturbagens

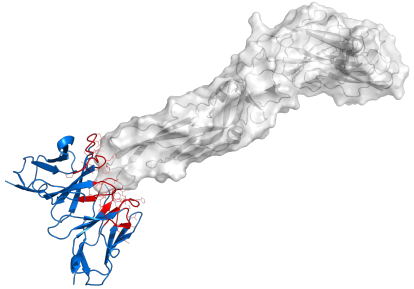
Genentech
A Member of the Roche Group

Misc

My research at Prescient

Prescient Design
A Genentech Accelerator

Epiphany affinity

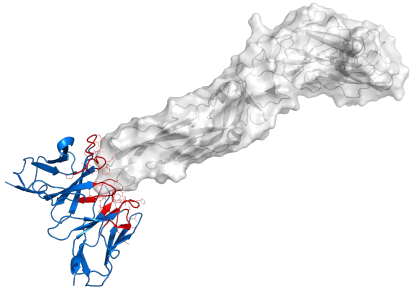


binding?

My research at Prescient

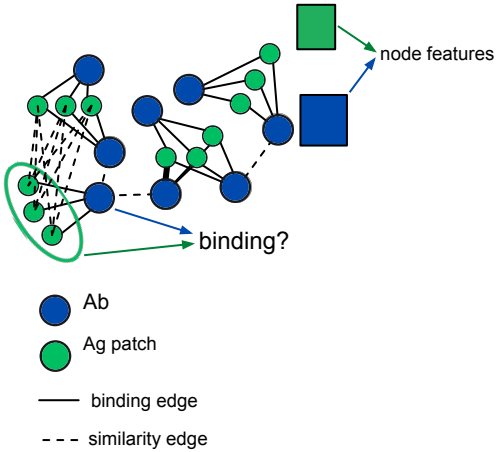
Prescient Design
A Genentech Accelerator

Epiphany affinity



binding?

KGraph



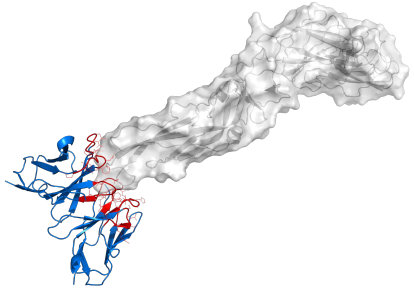
- Given an Ab node and an Ag patch node, predict link existence (binding)
- Given an Ab node and an a set of Ag patch nodes, predict link existence (binding)

Genentech
A Member of the Roche Group

My research at Prescient

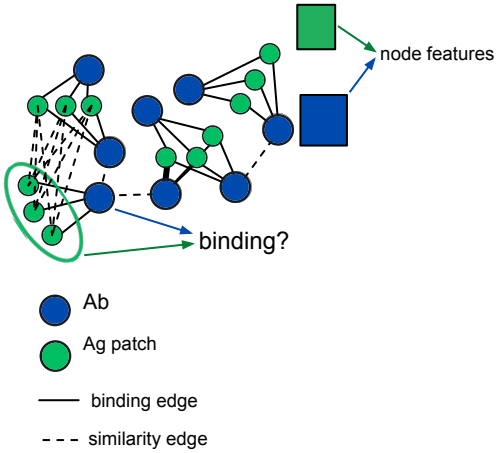
Prescient Design
A Genentech Accelerator

Epiphany affinity



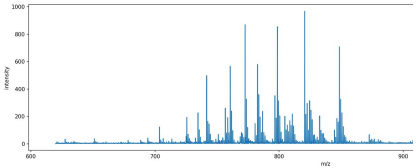
binding?

KGraph

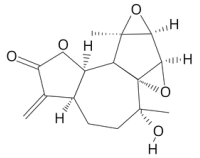


Structure elucidation

Mass spectrometry



Molecule



Genentech
A Member of the Roche Group

THE END