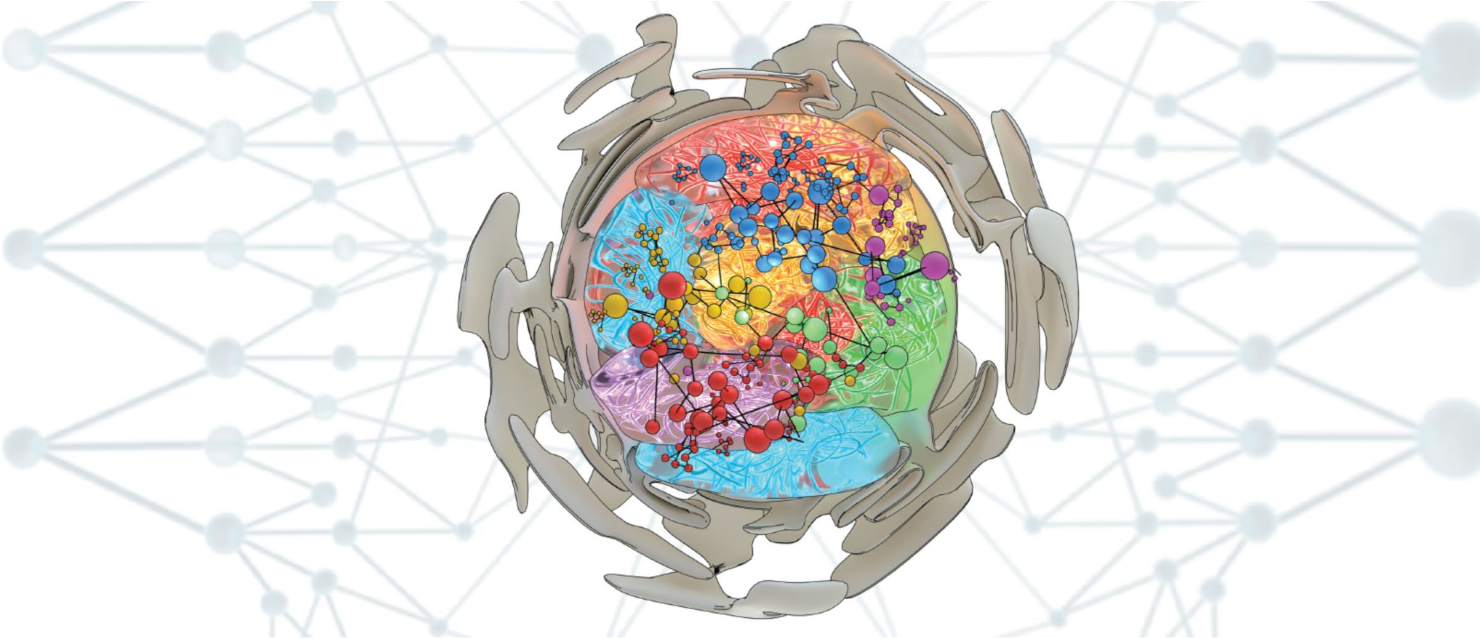
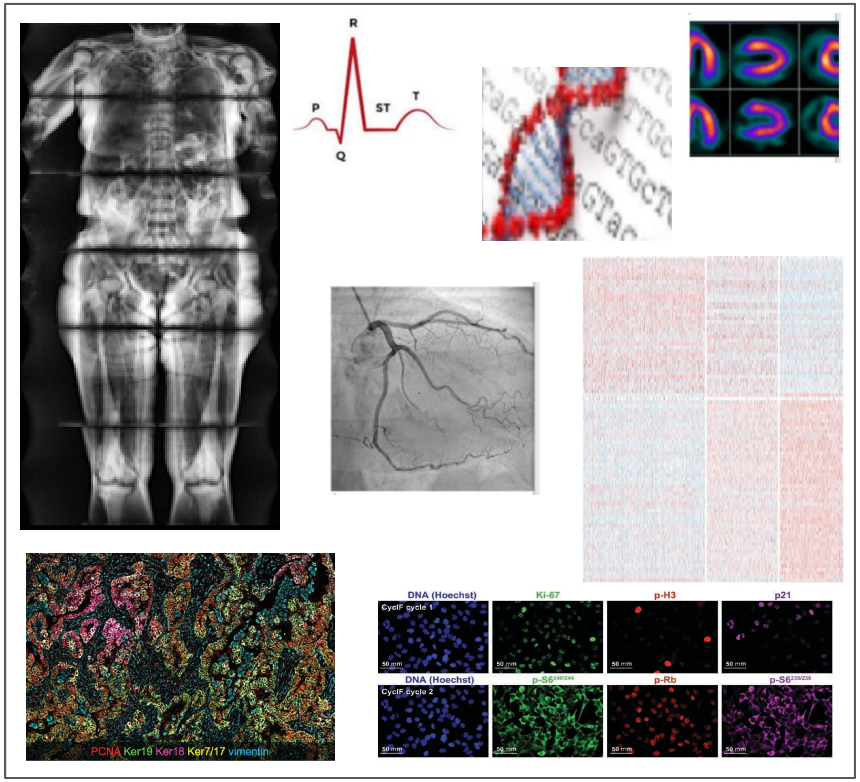
From Interventions to Causality using Over-Parameterized Neural Networks

Caroline Uhler (MIT & Broad Institute)



Need for causal representation learning

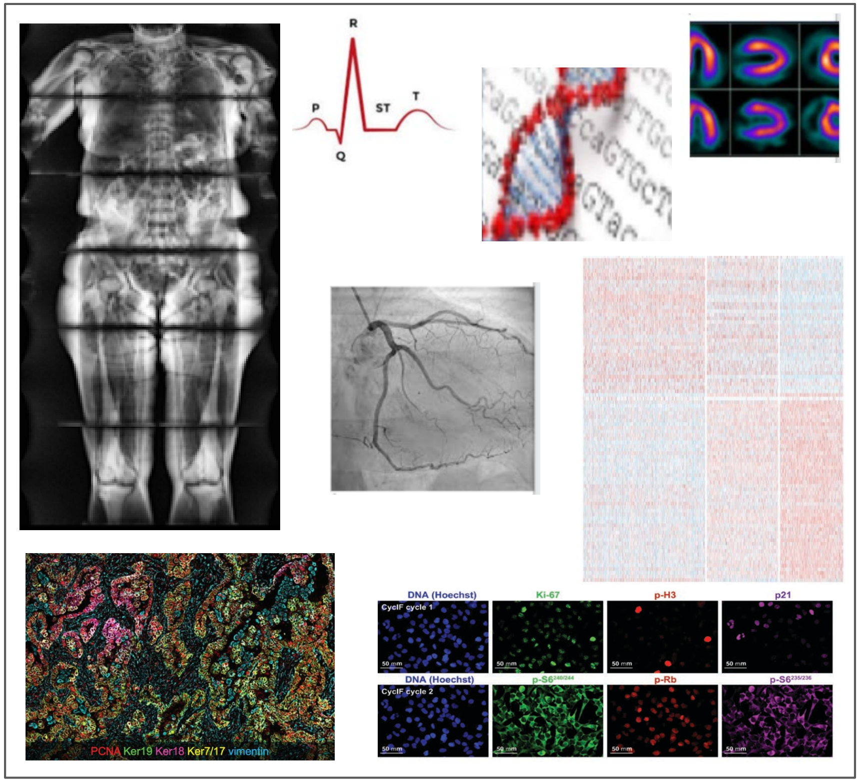
Huge amounts of unlabeled data of many different modalities



Representation learning allows integrating different modalities and extracting latent structures that capture intrinsic behavior without labeled data

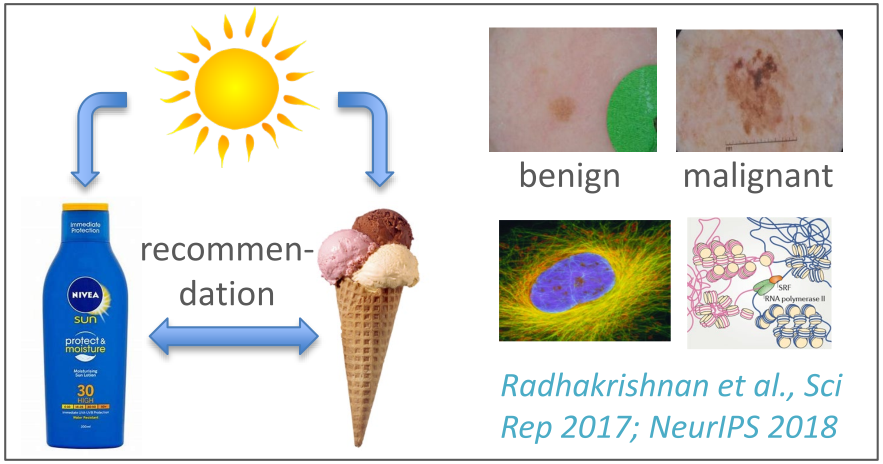
Need for causal representation learning

Huge amounts of unlabeled data of many different modalities



Representation learning allows integrating different modalities and extracting latent structures that capture intrinsic behavior without labeled data

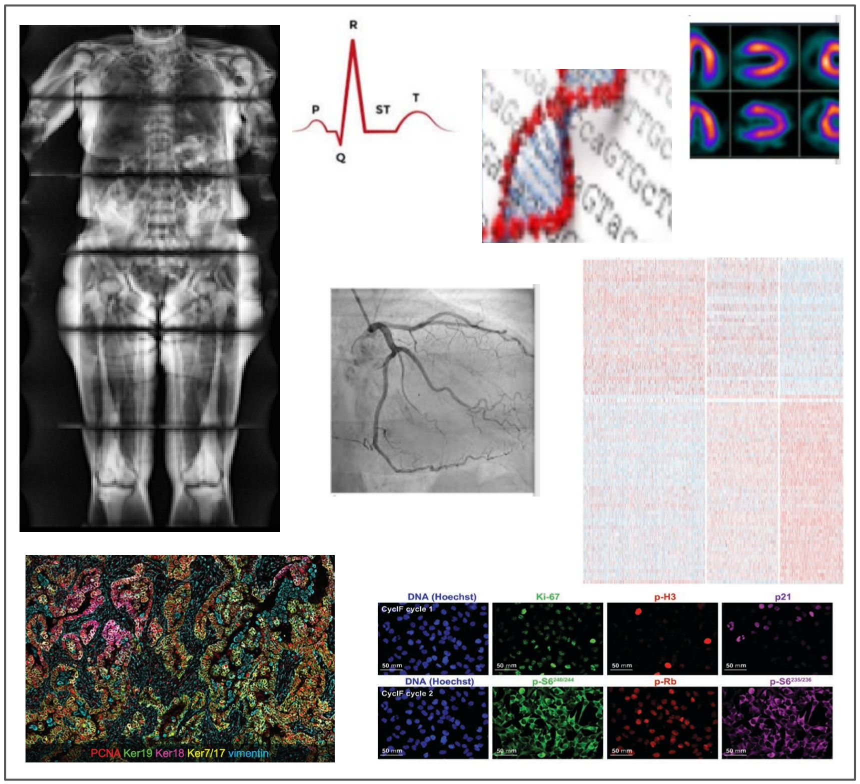
Understanding the underlying mechanisms / causal relationships is critical in biomedical sciences



Radhakrishnan et al., Sci Rep 2017; NeurIPS 2018

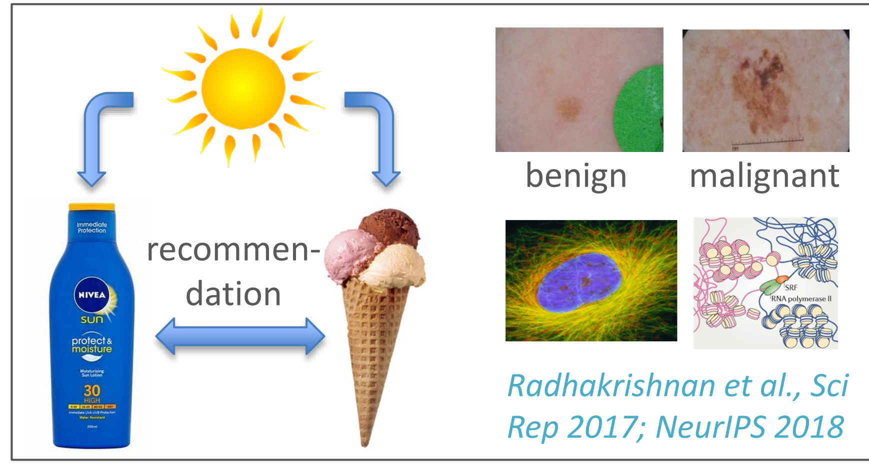
Need for causal representation learning

Huge amounts of unlabeled data of many different modalities



Representation learning allows integrating different modalities and extracting latent structures that capture intrinsic behavior without labeled data

Understanding the underlying mechanisms / causal relationships is critical in biomedical sciences



Causal Inference

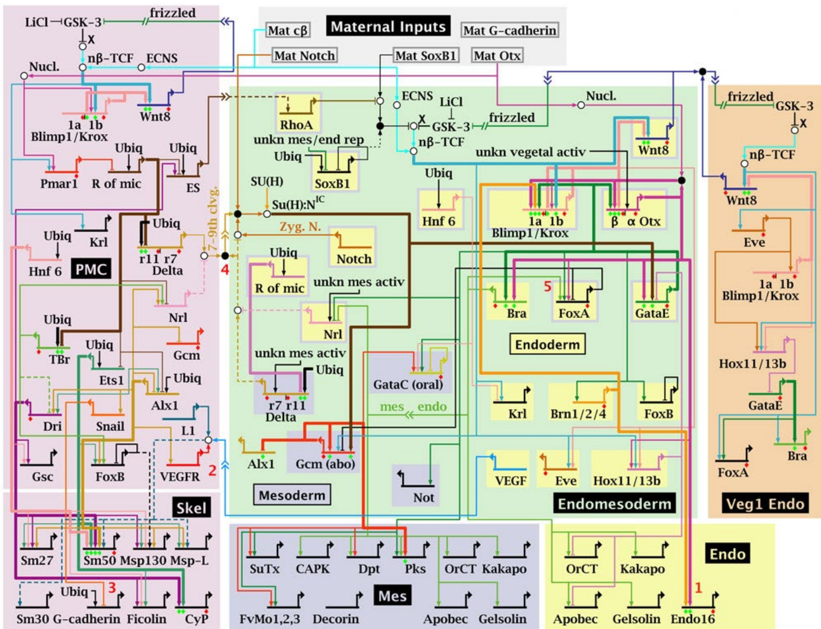
Representation learning

We need a theory of causal representation learning!

Perturbations (CRISPR, drugs, …) represent unique opportunity!

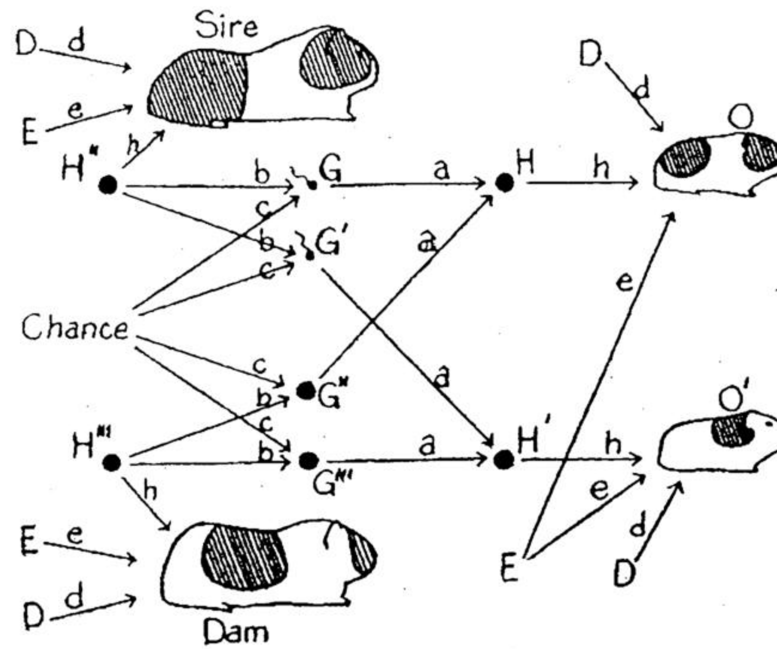
Gene regulation and structural equation models

Ex: Gene regulatory network for pregastrular endomesoderm specification in sea urchins



Eric H. Davidson, 2006

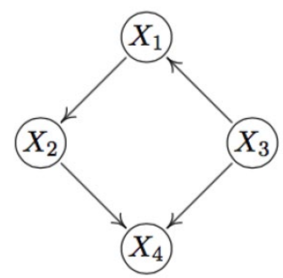
Sewell Wright developed the foundation of causal inference by studying heredity



Sewell Wright, 1920

Causal structural equation models:

- Represent causal relations by directed network
- Each node associated with random variable, stochasticity introduced by independent noise variables ϵ_i

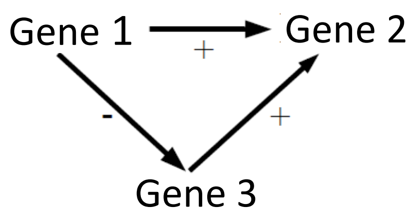


$$X_1 \leftarrow f_1(X_3, \epsilon_1)$$
$$X_2 \leftarrow f_2(X_1, \epsilon_2)$$
$$X_3 \leftarrow f_3(\epsilon_3)$$
$$X_4 \leftarrow f_4(X_2, X_3, \epsilon_4)$$

Causal structure discovery

Learning causal networks from observational data has long history starting with work by Spirtes in 1990s

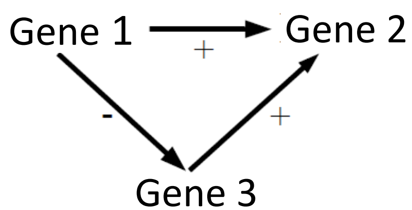
These algorithms assume faithfulness, i.e., that causal effects cannot cancel each other out

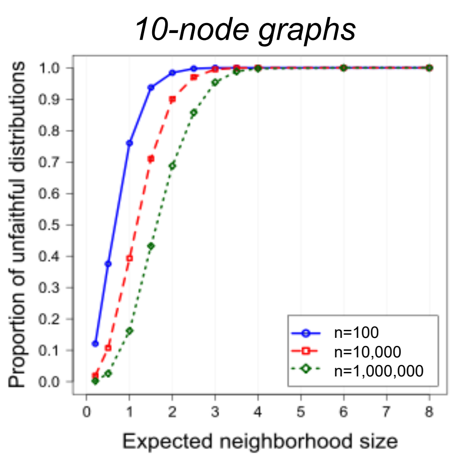


Causal structure discovery

Learning causal networks from observational data has long history starting with work by Spirtes in 1990s

These algorithms assume faithfulness, i.e., that causal effects cannot cancel each other out





Problem: Faithfulness violations are frequent when sample size n isn't infinite

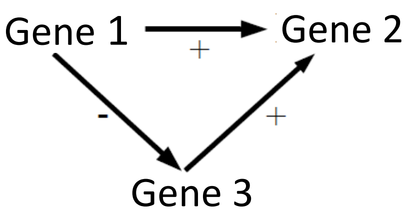
Learning network on 100 nodes requires >>10^100 samples

Uhler et al., Ann. Statist., 2013; Raskutti & Uhler, Stat, 2018

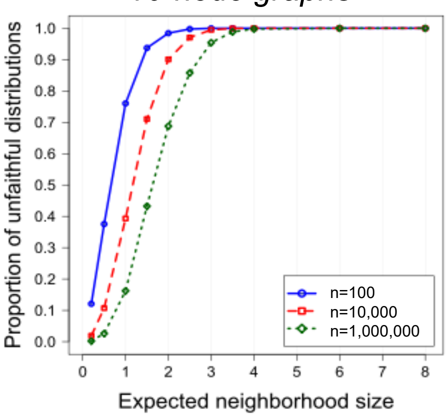
Causal structure discovery

Learning causal networks from observational data has long history starting with work by Spirtes in 1990s

These algorithms assume faithfulness, i.e., that causal effects cannot cancel each other out



10-node graphs

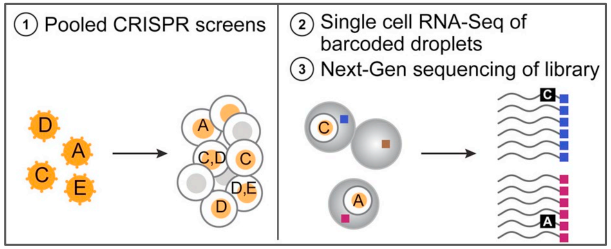


Problem: Faithfulness violations are frequent when sample size n isn't infinite

Learning network on 100 nodes requires >>10^100 samples

Uhler et al., Ann. Statist., 2013; Raskutti & Uhler, Stat, 2018

High-throughput perturbational data is available!



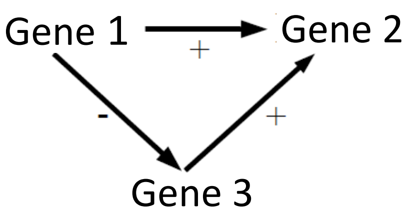
Perturb-seq

Dixit et al, Cell, 2016

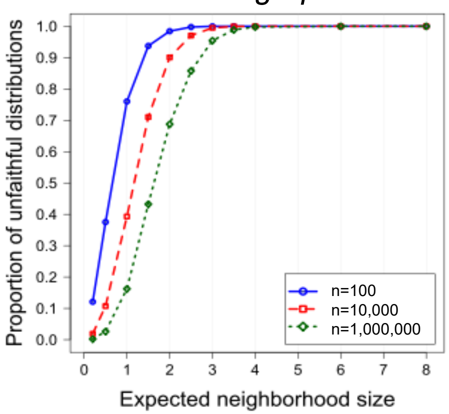
Causal structure discovery

Learning causal networks from observational data has long history starting with work by *Peter Spirtes* in 1990s

These algorithms assume faithfulness, i.e., that causal effects cannot cancel each other out



10-node graphs

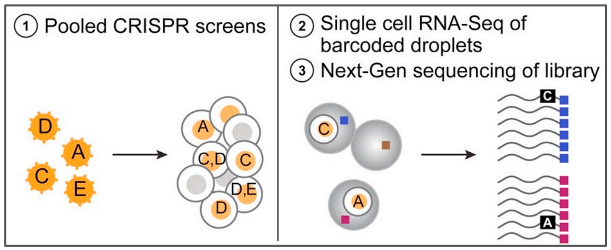


Problem: Faithfulness violations are frequent when sample size n isn't infinite

Learning network on 100 nodes requires >>10^100 samples

Uhler et al., Ann. Statist., 2013; Raskutti & Uhler, Stat, 2018

High-throughput perturbational data is available!

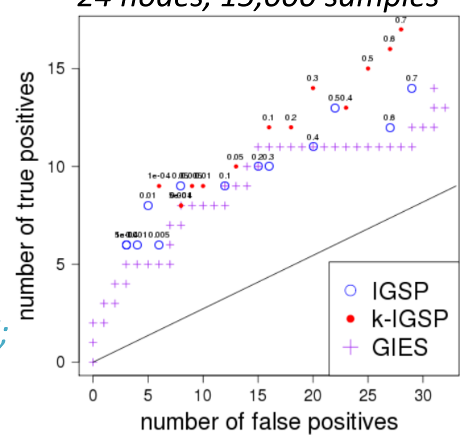


Perturb-seq

Dixit et al, Cell, 2016

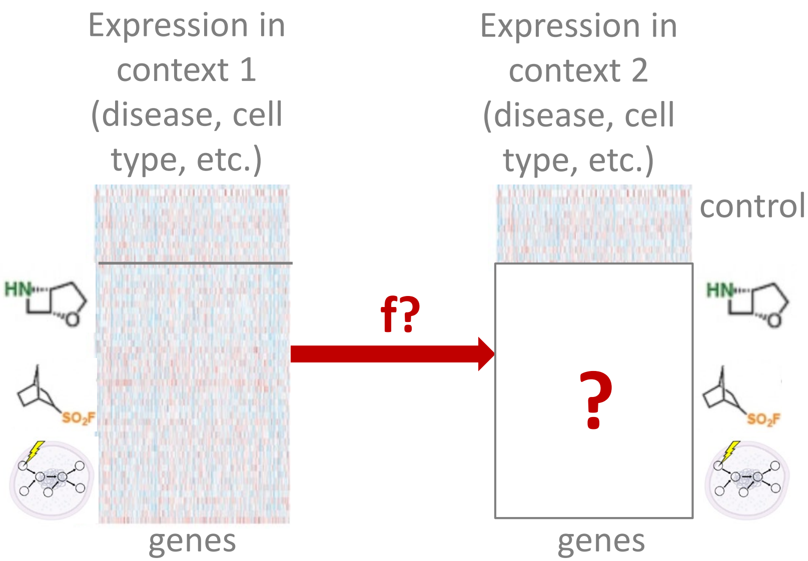
- Building on *Frederick Eberhardt*'s formalism, we developed first provably consistent algorithm for inferring causal network from observational & interventional data
- Scales to graphs with 1000s of nodes

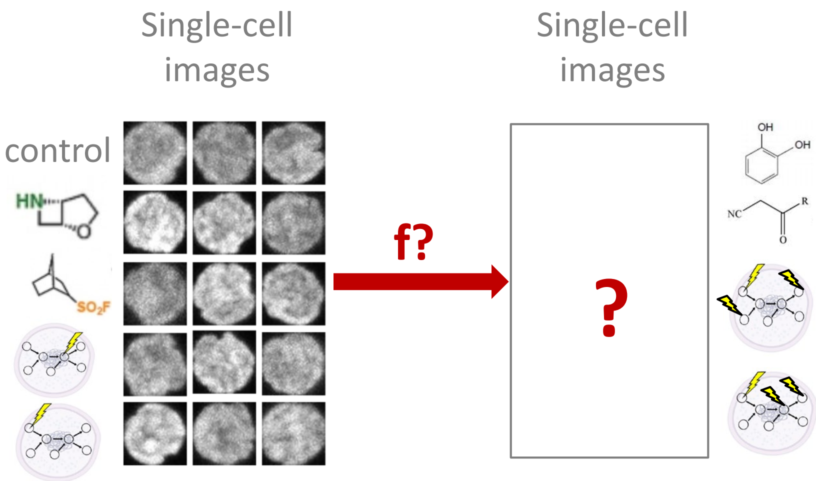
24 nodes, 15,000 samples

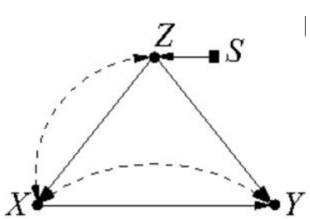


Wang et al., NeurIPS 2017; Yang et al., ICML 2018; Squires, et al. UAI 2020

Causal imputation problems in single-cell biology



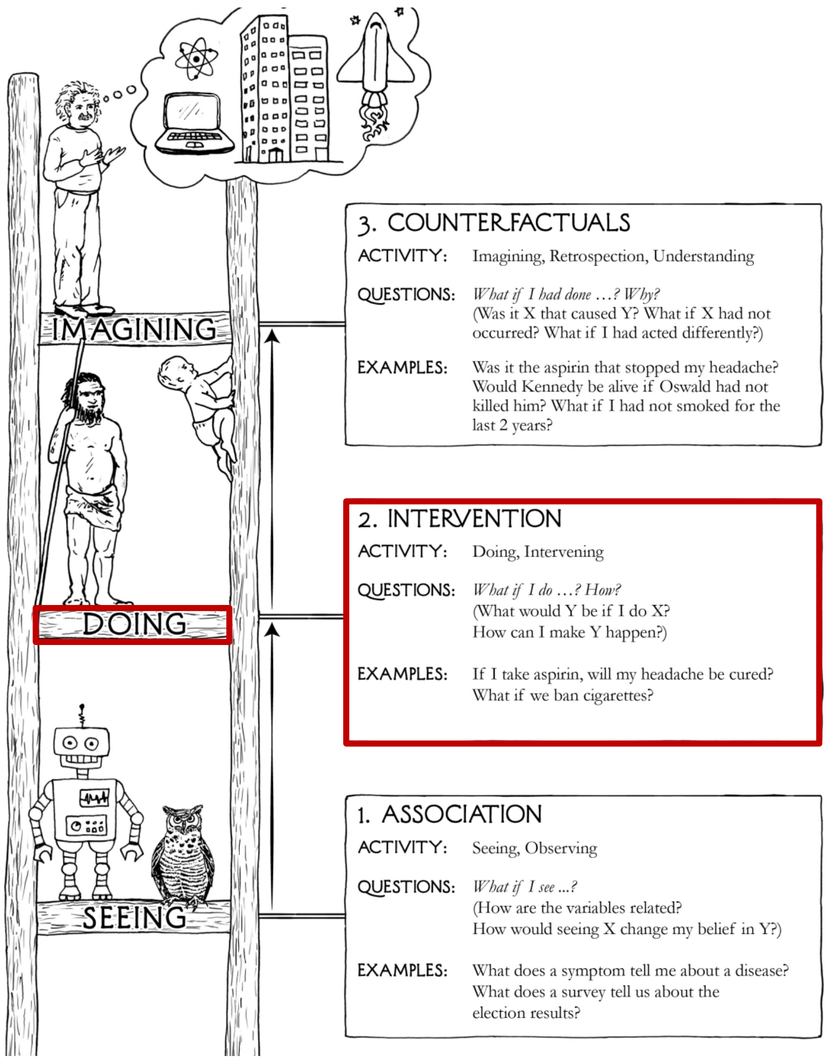




Causal transportability: Bareinboim, Pearl
Synthetic control / interventions: Abadie, Agarwal, Shah, Shen

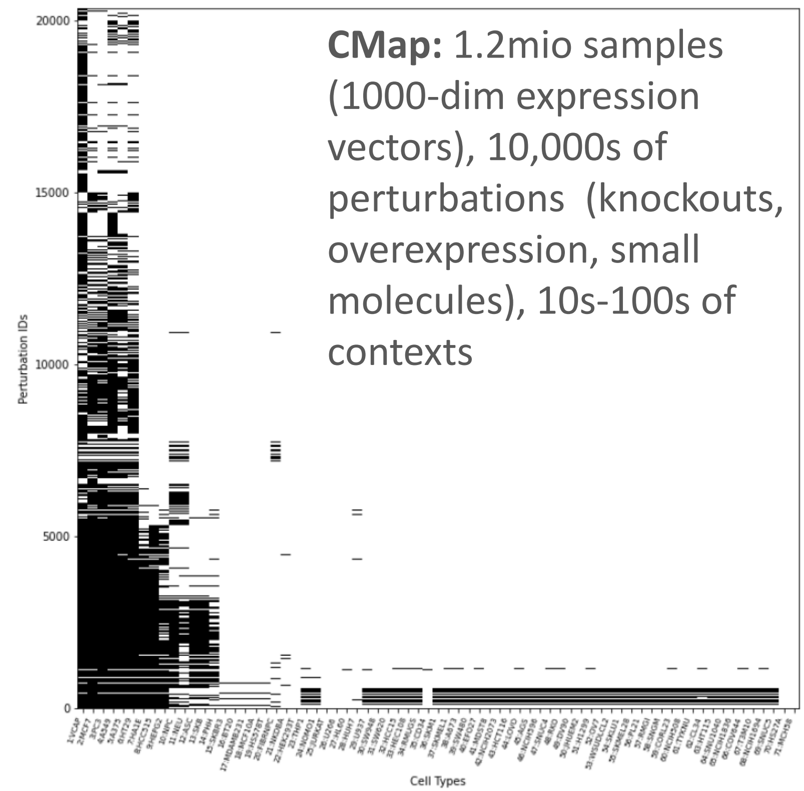
Idea 1: Causal inference by predicting interventions

Judea Pearl's causal hierarchy



J. Pearl, The Book of Why, 2018

Causal imputation

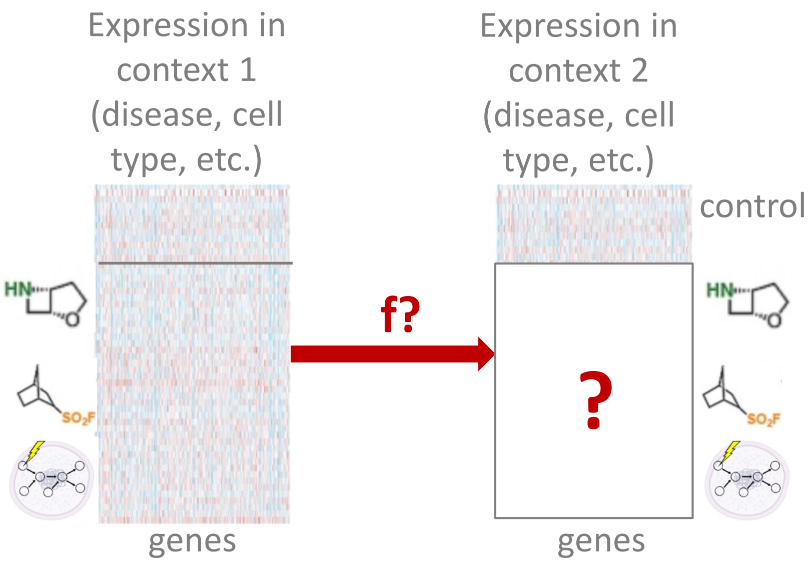


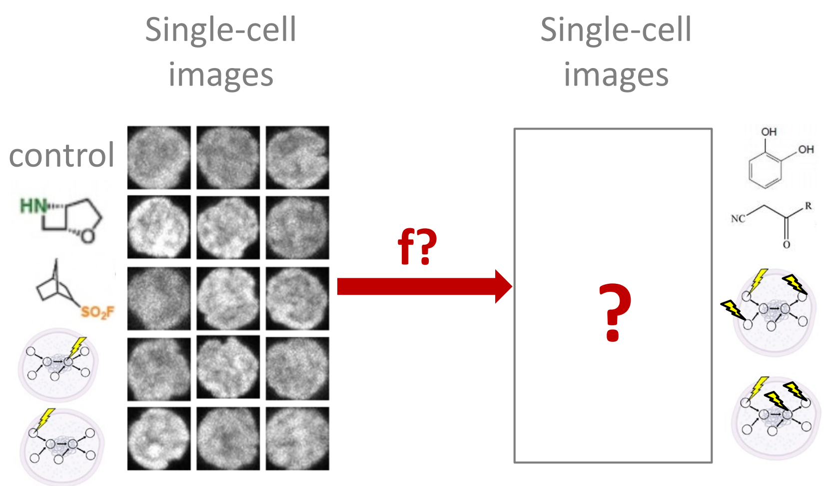
CMap: 1.2mio samples (1000-dim expression vectors), 10,000s of perturbations (knockouts, overexpression, small molecules), 10s-100s of contexts

We are great at solving prediction problems:

Availability of interventional data allows us to turn causal questions into (causal) prediction problems!

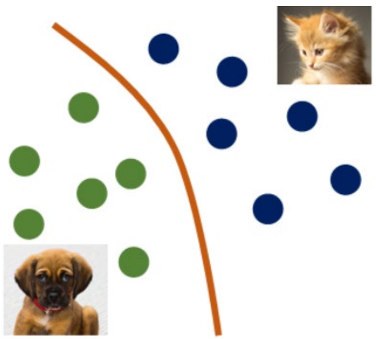
Causal imputation problems in single-cell biology



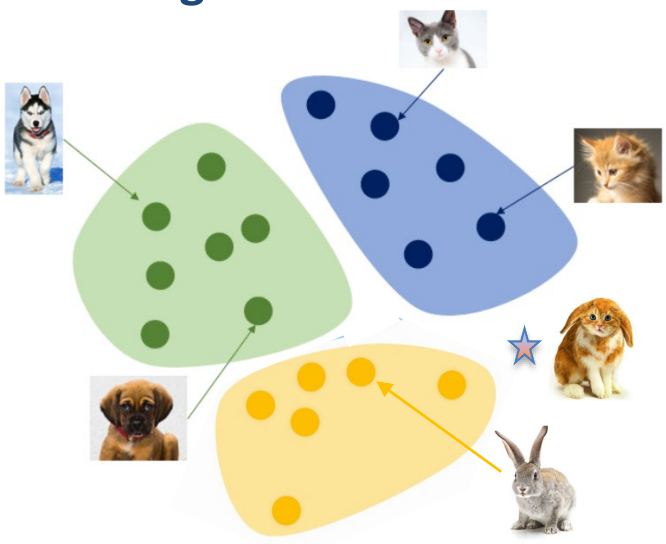


Over-parameterized neural networks

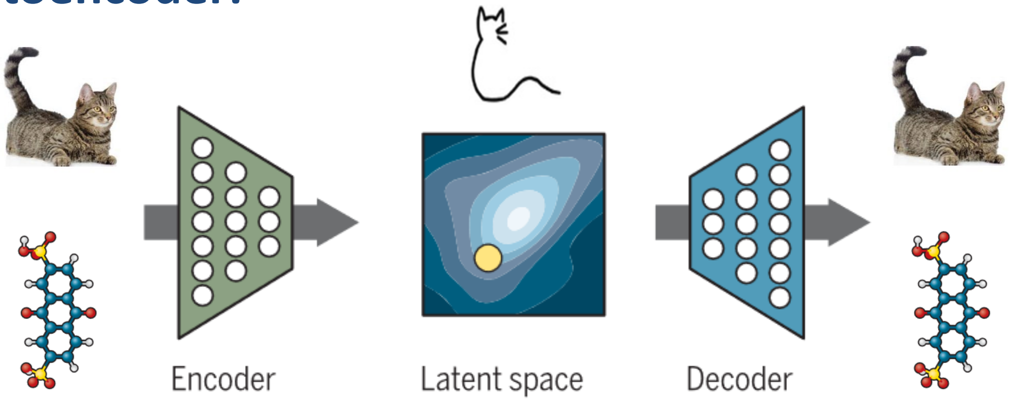
Discriminative modeling:



Generative modeling:



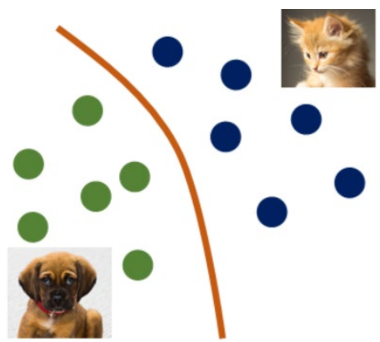
Autoencoder:



Sanchez-Lengeling *et al.*, *Science* **361**, 360–365 (2018)

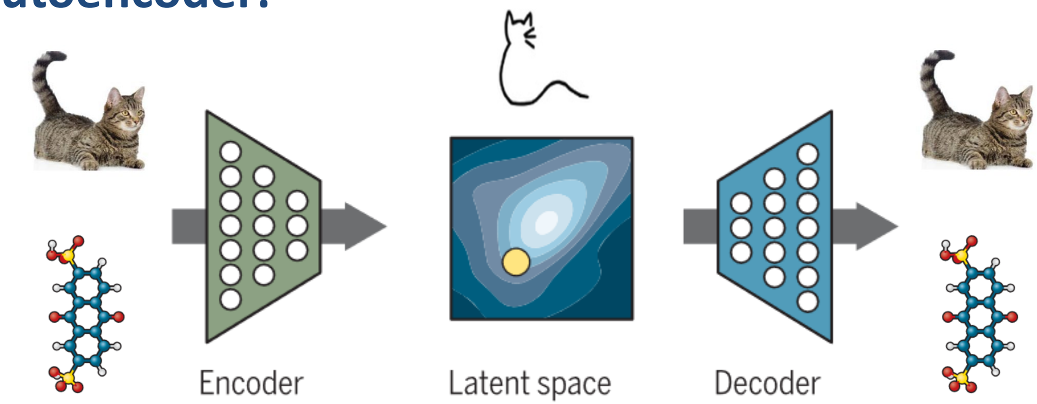
Over-parameterized neural networks

Discriminative modeling:



Generative modeling:

Autoencoder:



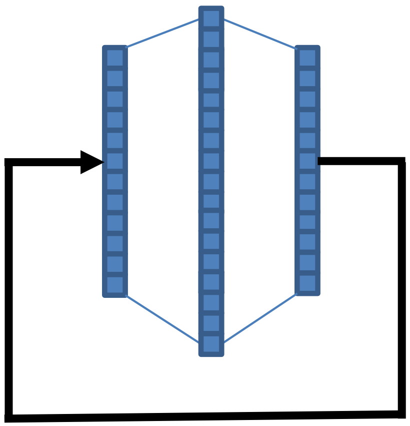
Sanchez-Lengeling *et al.*, *Science* **361**, 360–365 (2018)

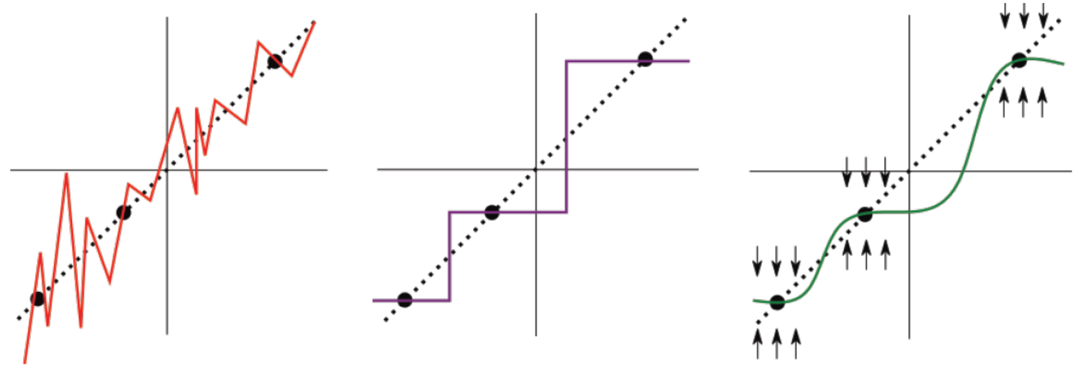
Over-parameterized neural networks:

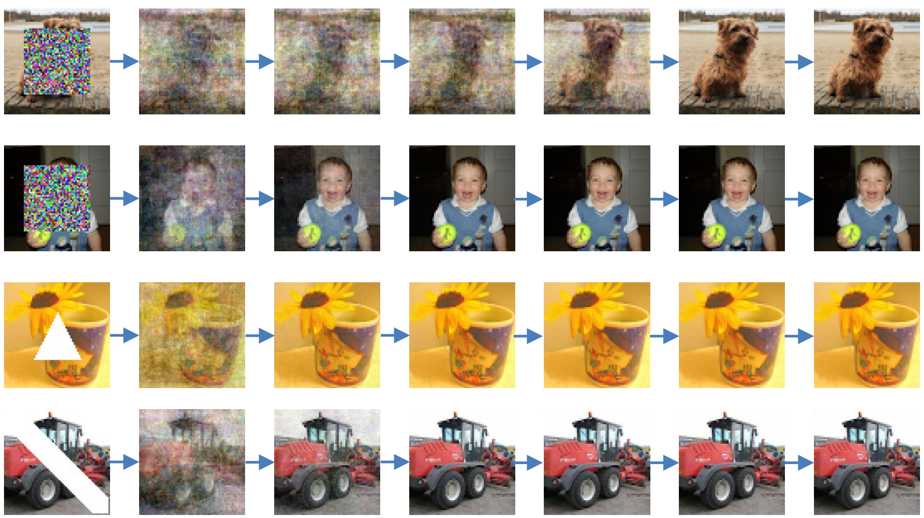
- Deep neural networks can generalize while interpolating the training data *Belkin et al., PNAS 2019*
- Infinitely wide neural networks converge to the neural tangent kernel *Jacot et al., NeurIPS 2018*
  6.S088 Modern Machine Learning: Simple Methods that Work (mit.edu)
- Neural tangent kernel with specific activation function is Bayes optimal for classification

MS92, Wed 2.45-3pm: Radhakrishnan, Belkin & Uhler, arXiv:2204.14126

Inductive bias of over-parameterized autoencoders



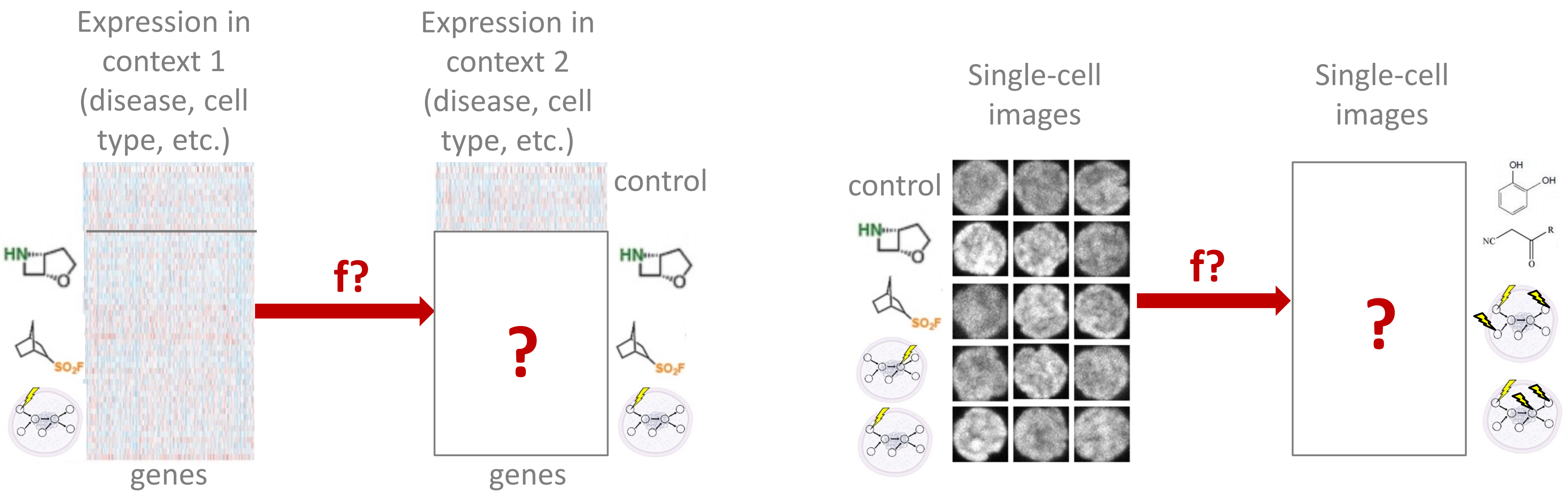




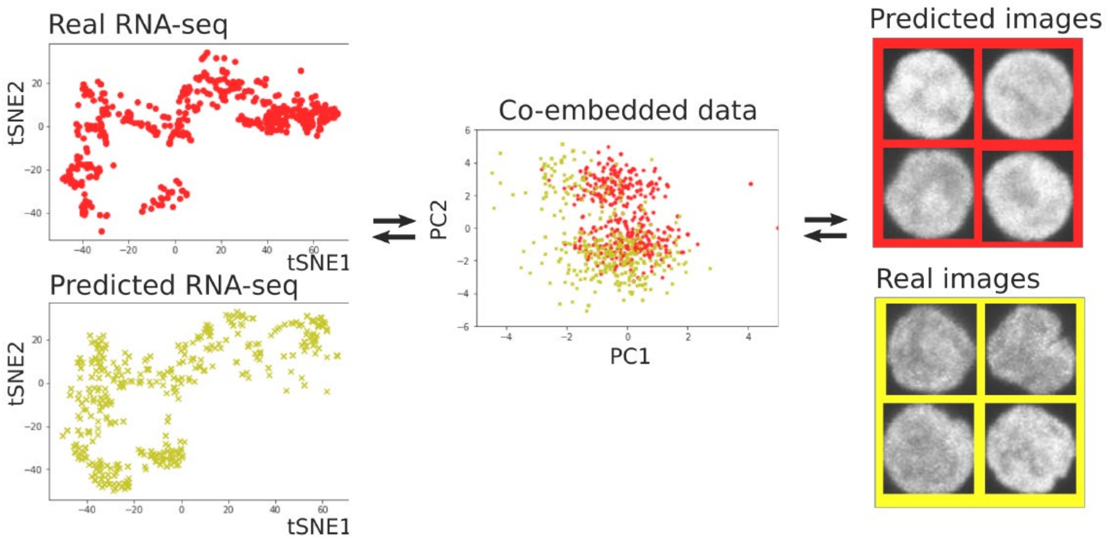
- Deep over-parameterized CNNs can interpolate training data even with random labels (*Arpit et al. ICML 17; Zhang et al. ICLR 17*)
- There are many ways to interpolate training data
- Over-parameterized autoencoders learn maps that are contractive at training examples

Radhakrishnan, Belkin & Uhler, PNAS 2020

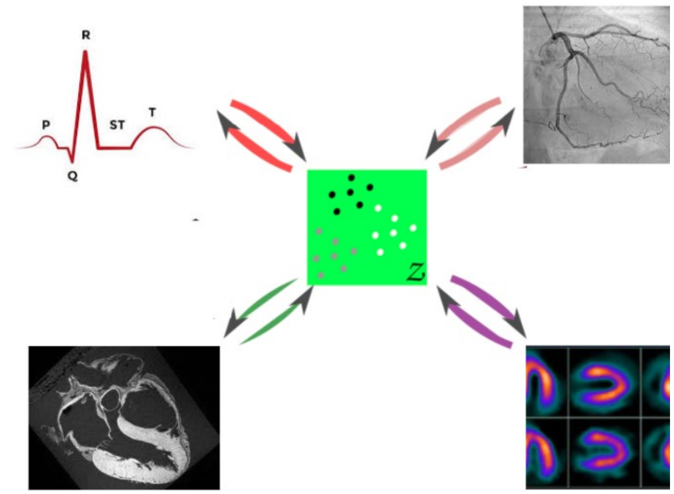
Causal imputation problems in single-cell biology



Idea 2: Multi-modal learning to discover causal feature



Yang et al., Nat Comm 2021



Radhakrishnan et al., bioRxiv

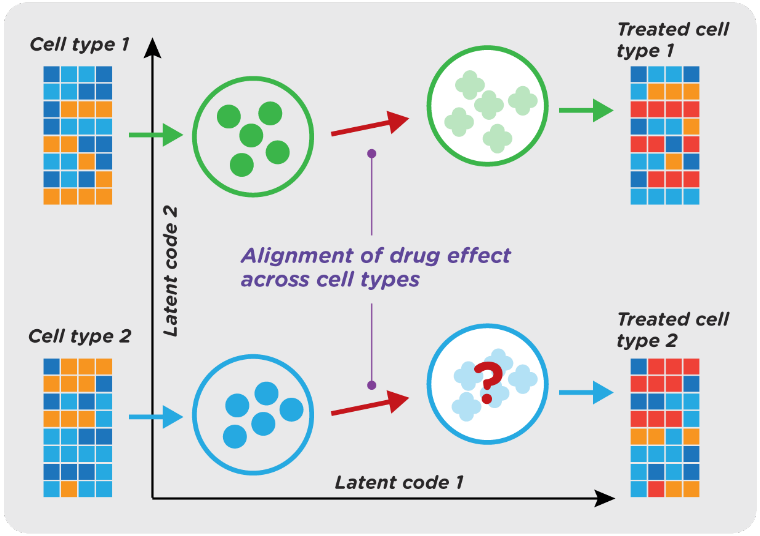
Representation learning as a tool for causal feature discovery
by learning integrated latent spaces:
Causal features should be invariant to modality in which they are measured!

Invariant prediction for causal inference: Peters, Buehlmann, Meinshausen
Invariant risk minimization: Arjovsky, Bottou, Gulrajani, Lopez-Paz

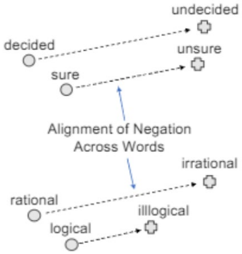
Causal feature learning: Chalupka, Perona, Eberhardt
Disentanglement: Schoelkopf, Bengio,…

Idea 3: Over-parameterization to align causal effects

Latent spaces that align causal effects

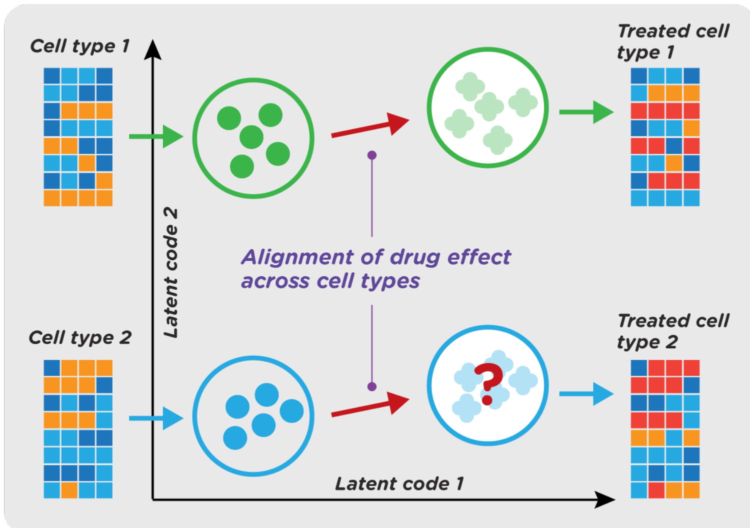


Word2Vec analogy

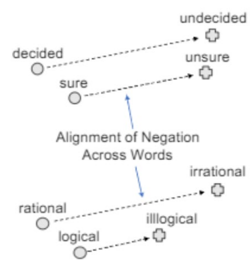


Idea 3: Over-parameterization to align causal effects

Latent spaces that align causal effects

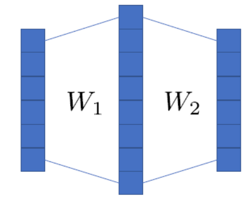


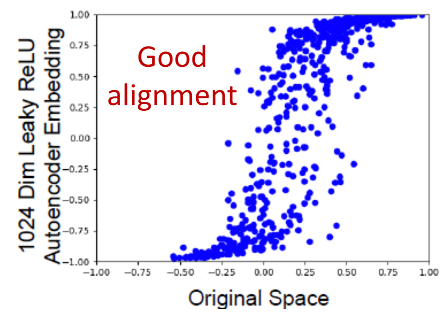
Word2Vec analogy

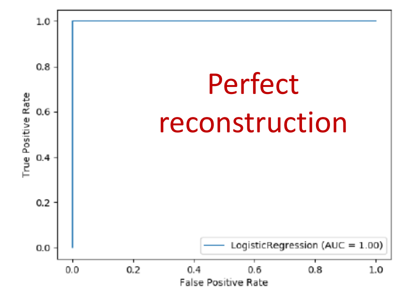


Over-parameterized neural nets

- Over-parameterized autoencoders: interpolate, generalize, and align drug signatures across cell types!





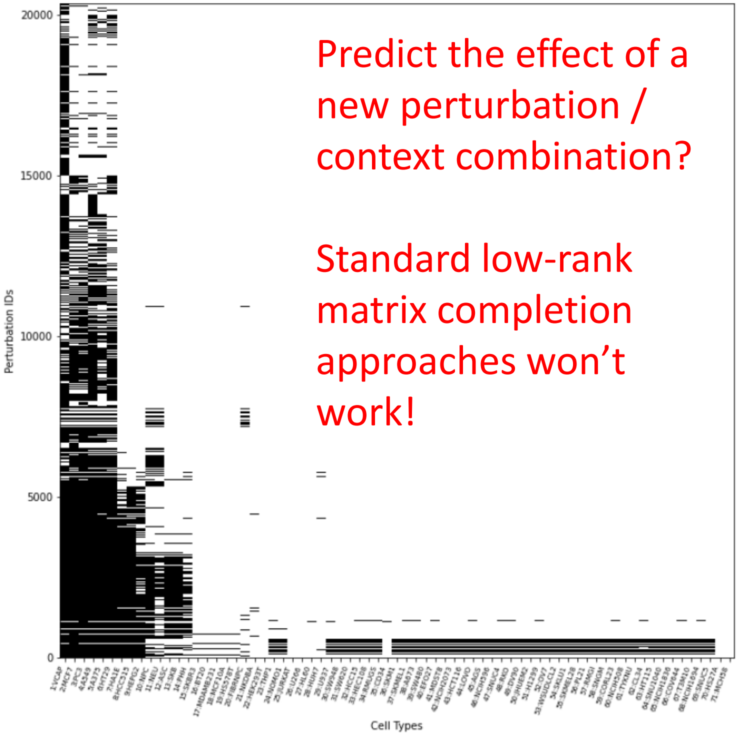


Belyaeva et al., Nat Comm 2021

Over-parameterized autoencoders provide more "space" to align causal effects:

We need to study their inductive biases and the interplay with causality!

Idea 4: Matrix completion using neural tangent kernel

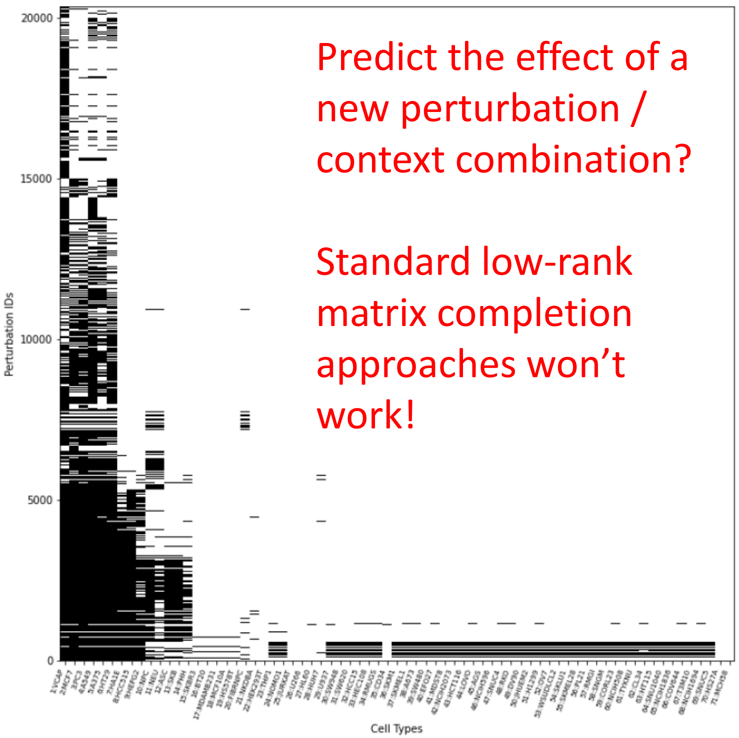


Predict the effect of a new perturbation / context combination?

Standard low-rank matrix completion approaches won't work!

Radhakrishnan et al., PNAS 2022

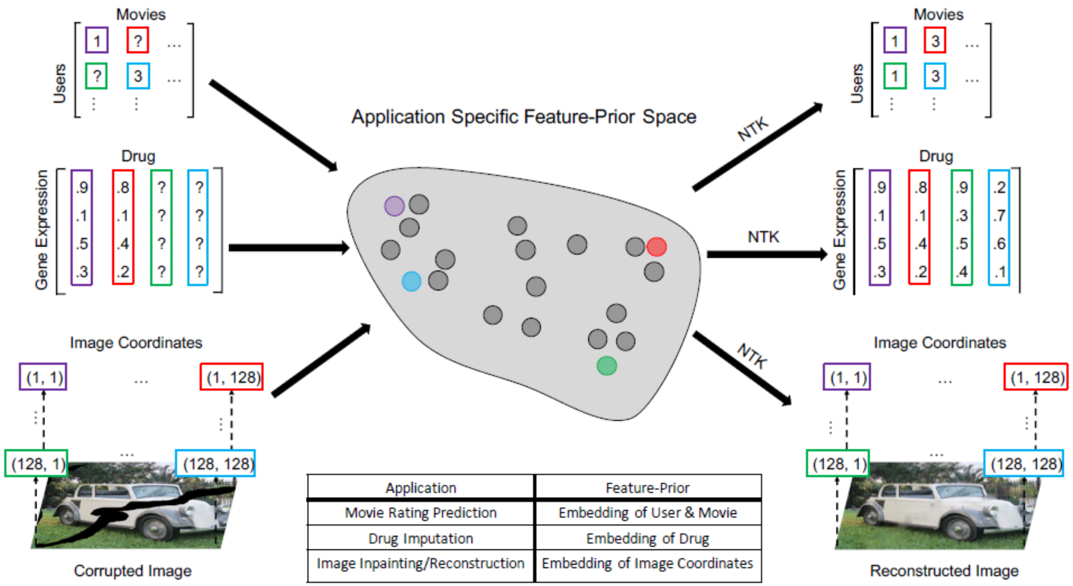
Idea 4: Matrix completion using neural tangent kernel



Predict the effect of a new perturbation / context combination?

Standard low-rank matrix completion approaches won't work!

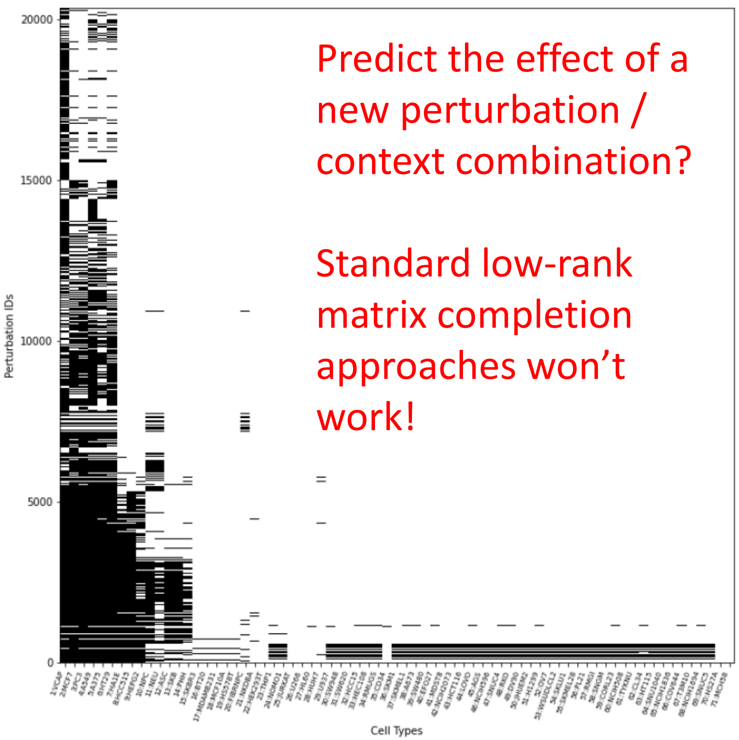
We built an NTK framework for matrix completion that can make use of feature priors on rows and columns



Application	Feature-Prior
Movie Rating Prediction	Embedding of User & Movie
Drug Imputation	Embedding of Drug
Image Inpainting/Reconstruction	Embedding of Image Coordinates

Radhakrishnan et al., PNAS 2022

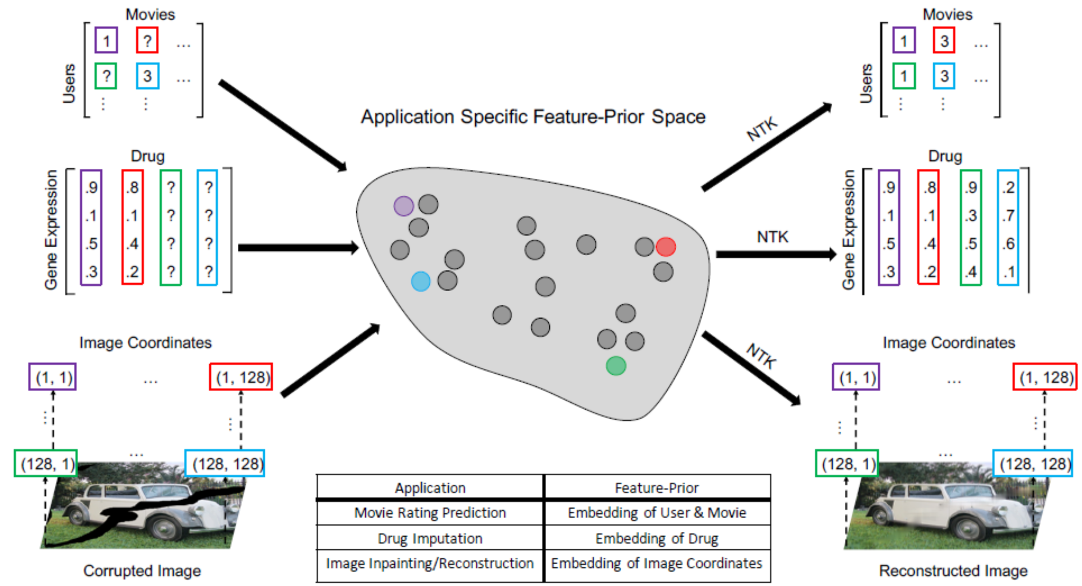
Idea 4: Matrix completion using neural tangent kernel



Predict the effect of a new perturbation / context combination?

Standard low-rank matrix completion approaches won't work!

We built an NTK framework for matrix completion that can make use of feature priors on rows and columns



CMap
(Full Dataset)

Evaluation Metric*	Mean Over Cell Type (Naïve Baseline)	FaLRTC (Liu et al. 2013)	DNPP (Hodos et al. 2018)	NTK (Ours)
Pearson r	0.374 ± 0.0004	0.545 ± 0.0003	0.556 ± 0.0003	**0.572 ± 0.0002**
Mean R^2	0.134 ± 10^(-5)	0.286 ± 0.0003	0.296 ± 0.0004	**0.320 ± 0.0002**
Mean Cosine Similarity	0.371 ± 10^(-5)	0.536 ± 0.0004	0.541 ± 0.0004	**0.554 ± 0.0002**

CMap
(Sparse Regime)

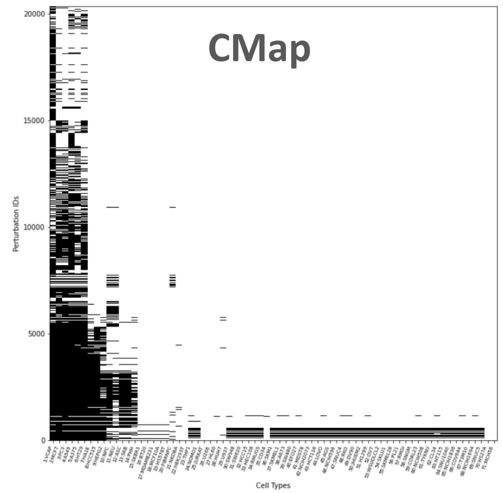
Evaluation Metric*	Mean Over Cell Type (Naïve Baseline)	FaLRTC (Liu et al. 2013)	DNPP (Hodos et al. 2018)	NTK (Ours)
Pearson r	0.450	0.544	0.538	**0.573**
Mean R^2	0.197	0.285	0.278	**0.324**
Mean Cosine Similarity	0.448	0.536	0.532	**0.565**

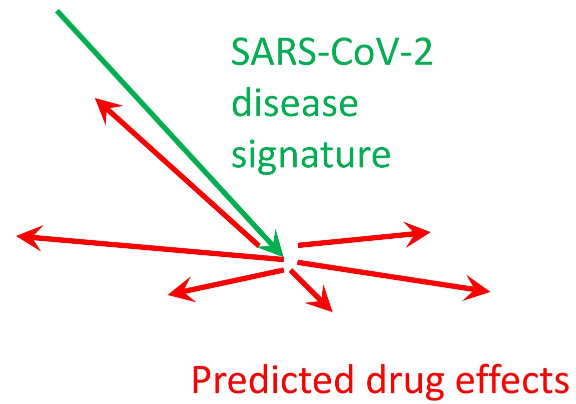
*Higher is better, with a maximum of 1.

Radhakrishnan et al., PNAS 2022

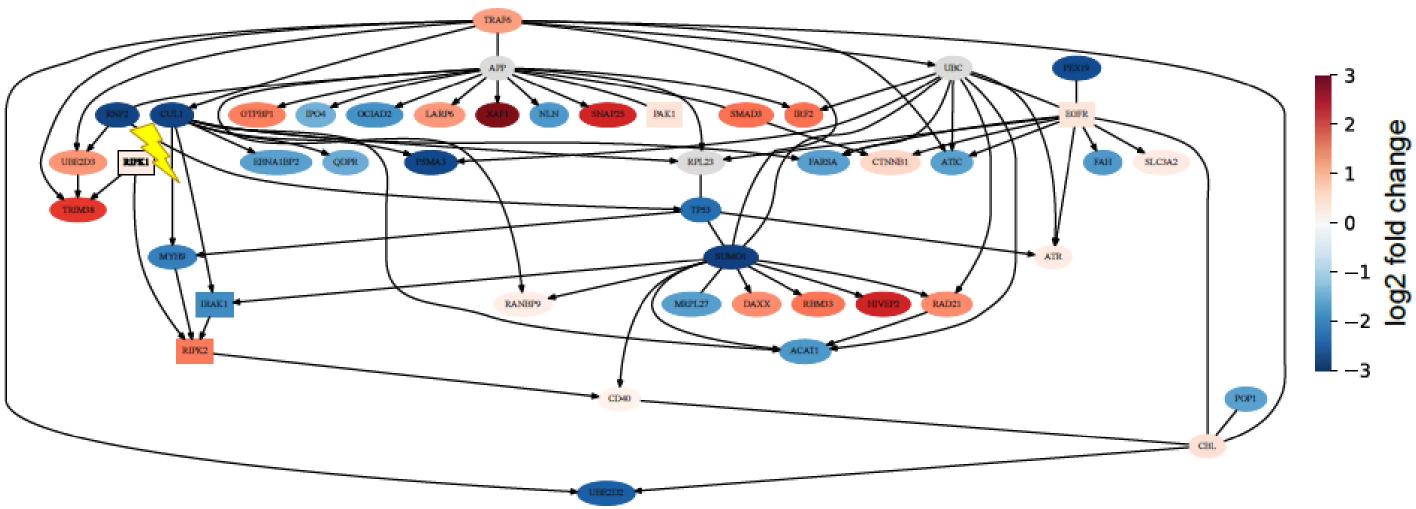
Target identification in the context of COVID-19

Transport effect of perturbations from CMap to SARS-CoV-2 infected A549 cells and find drug that is most anticorrelated with disease signature



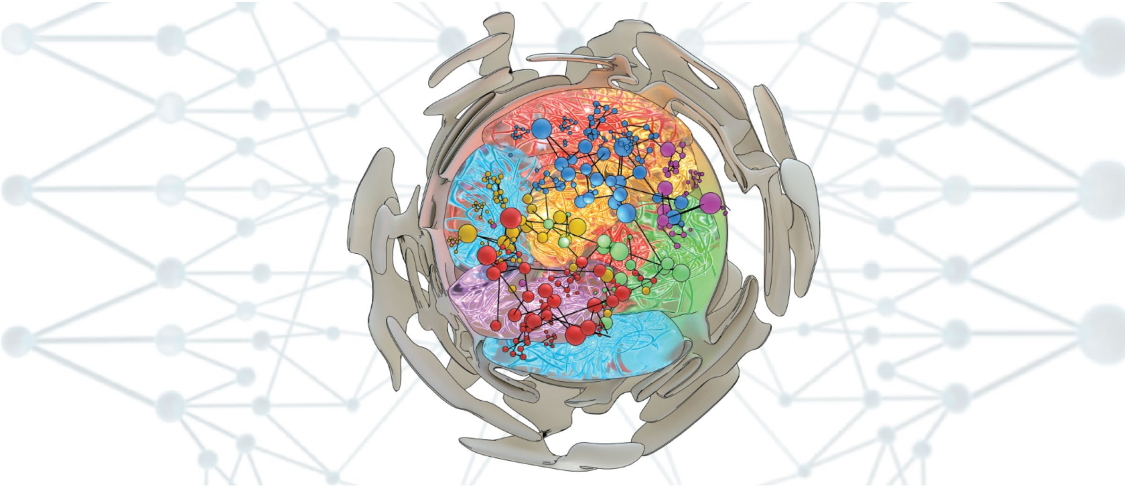


Causal structure discovery to identify putative causal drug targets:



Belyaeva et al., Nat Comm, 2021

Summary and outlook



- Developed a theoretical and algorithmic framework for integrating and translating between observational and interventional data
- Autoencoders are not only extremely useful for data integration and translation, but also for studying the theoretical properties of neural networks
- Over-parameterization leads to remarkable self-regularization properties and computational gains
- If we are able to predict the effect of unseen perturbations, we can build active framework for optimal intervention design to induce desired distribution shift

Zhang et al., NeurIPS 2021 & arXiv:2209.04744

Acknowledgments

PhD students:

- Louis Cammarata
- Parmida Davarmanesh
- Daniel Paysan
- **Adityanarayanan Radhakrishnan**
- Chandler Squires
- Yitong Tseo
- Karren Yang
- Jiaqi Zhang
- Xinyi Zhang

MSc/undergraduate students

- Cathy Cai
- Emily Liu
- Max Ruiz Luyten
- Nten Nyiam
- Ishika Shah

Postdocs:

- Salil Bhate
- Wengong Jin
- Neriman Tokcan

Collaborators:

Misha Belkin (UCSD)

GV Shivashankar (ETH Zurich)

Funding:

ERIC AND WENDY SCHMIDT CENTER AT BROAD INSTITUTE

DEPARTMENT OF THE NAVY ONR Science & Technology

MIT

SIMONS FOUNDATION

IBM

NSF