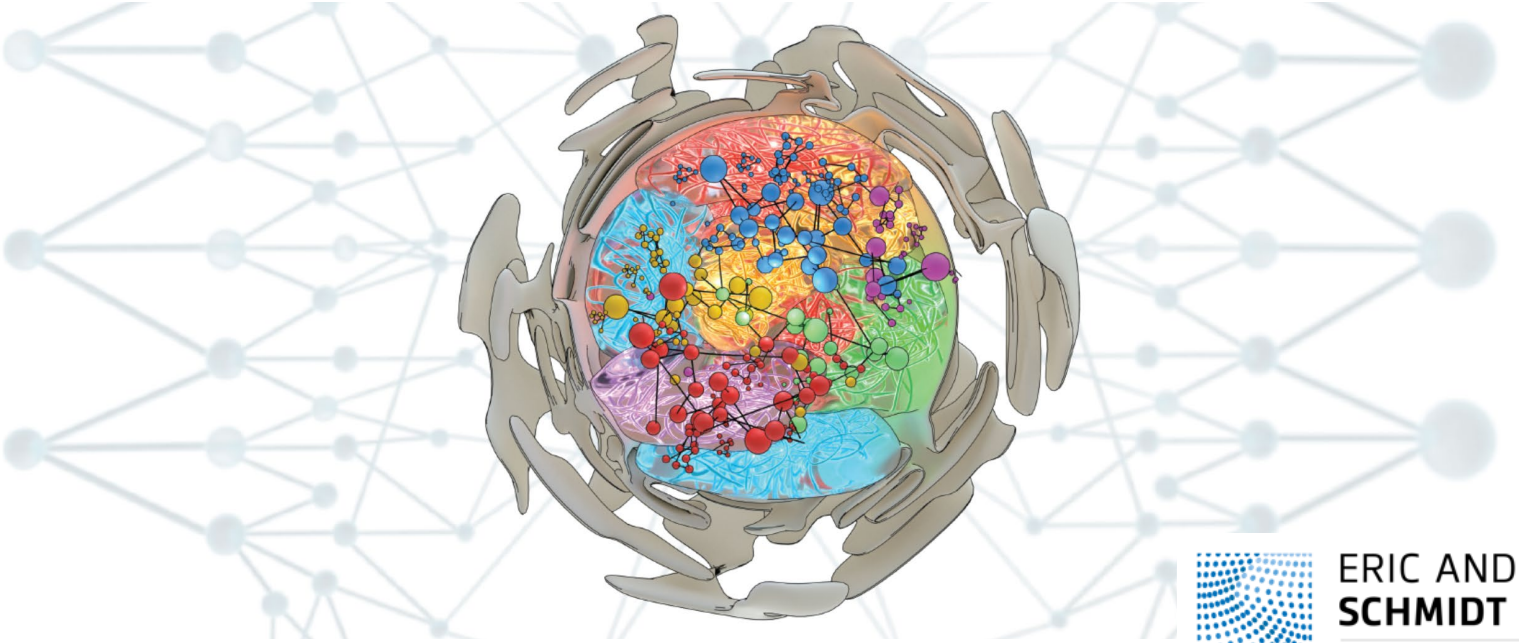
Causal Representation Learning and Optimal Intervention Design

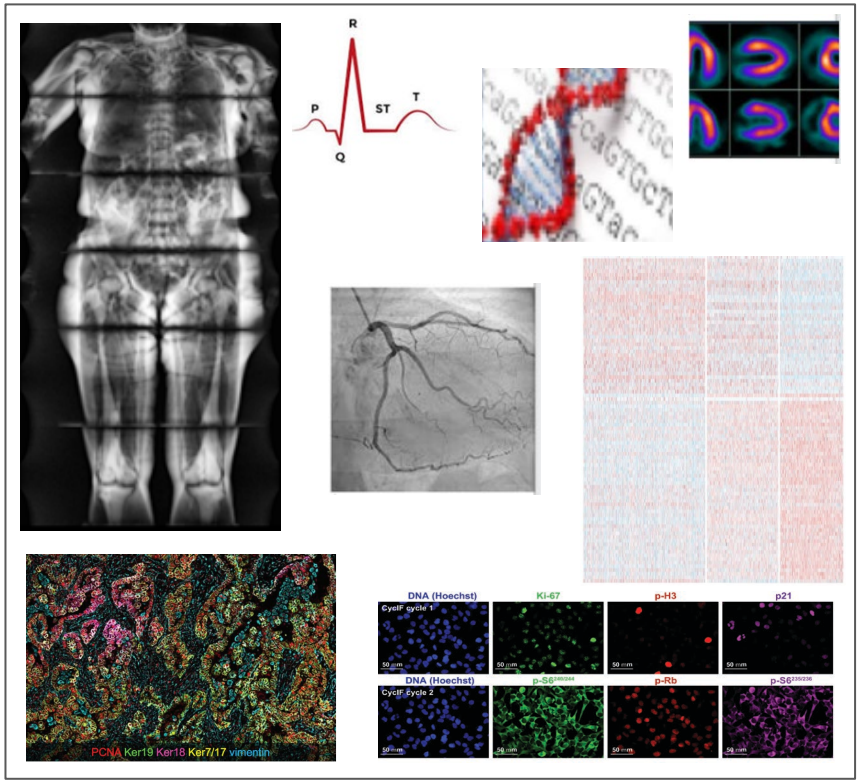
Caroline Uhler (MIT & Broad Institute)

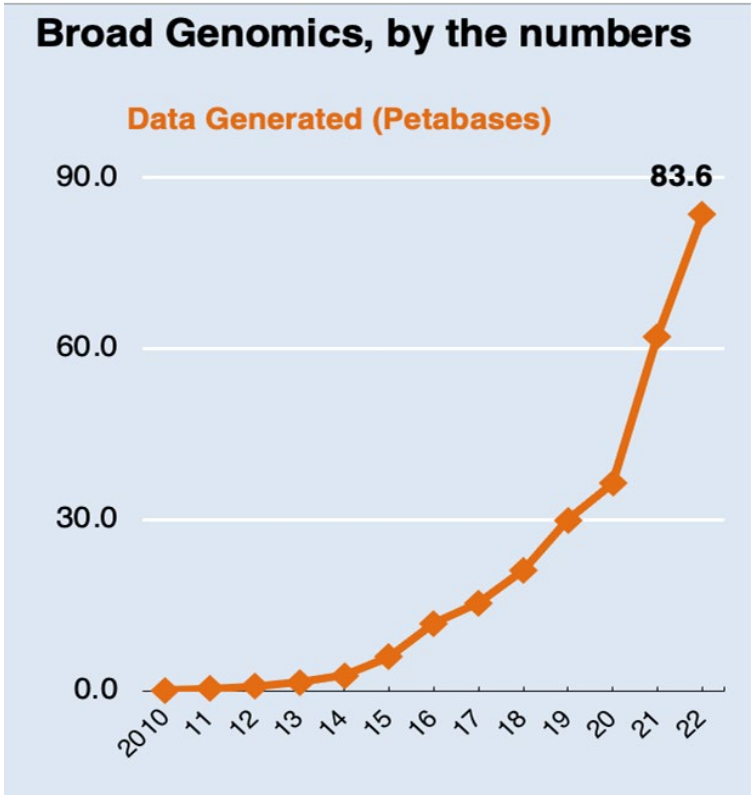


ERIC AND WENDY
SCHMIDT CENTER
AT BROAD INSTITUTE

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

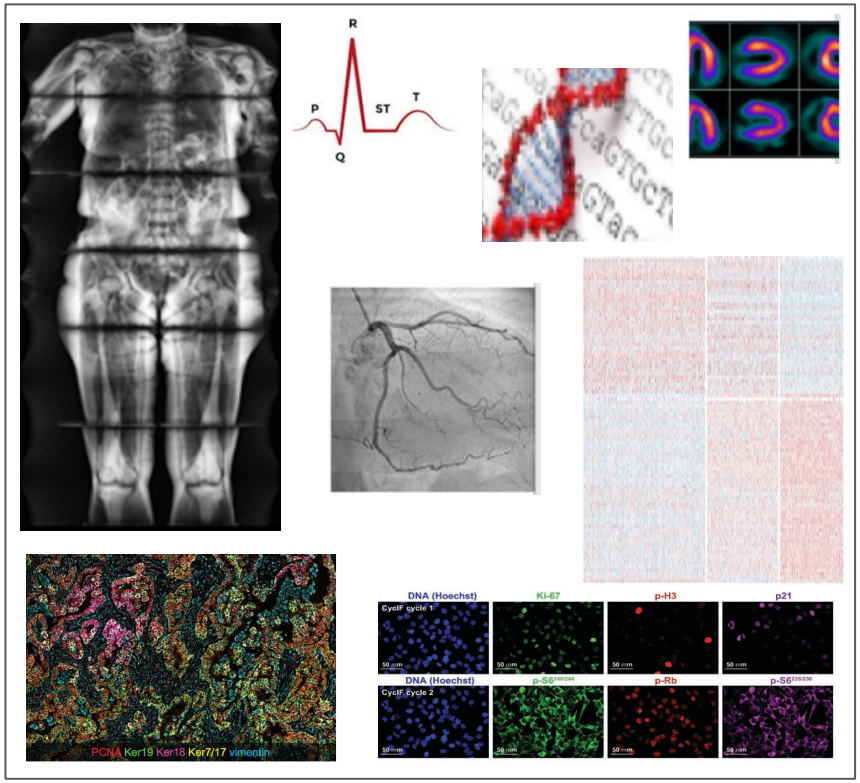
Similar to Twitter

Netflix movie corpus: 60pb

Total data currently under management at Broad: ~100pb

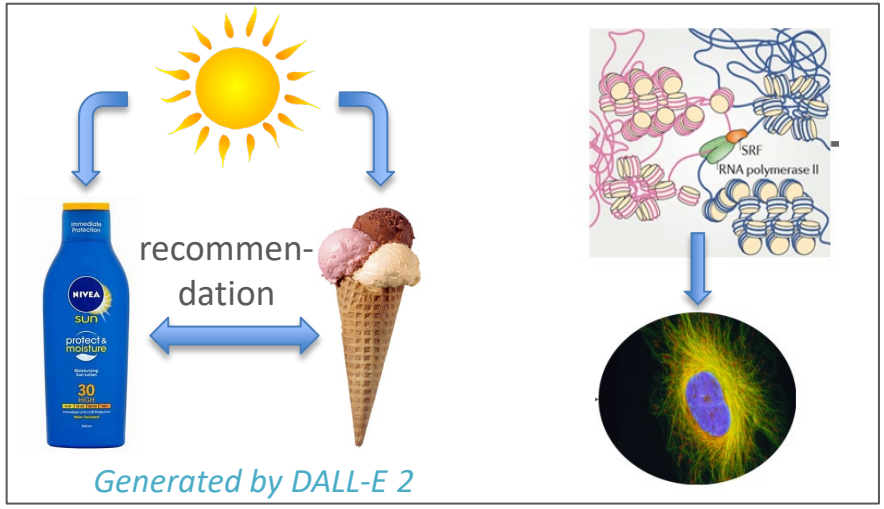
Need for causal representation learning

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



Generated by DALL-E 2

Causal Inference

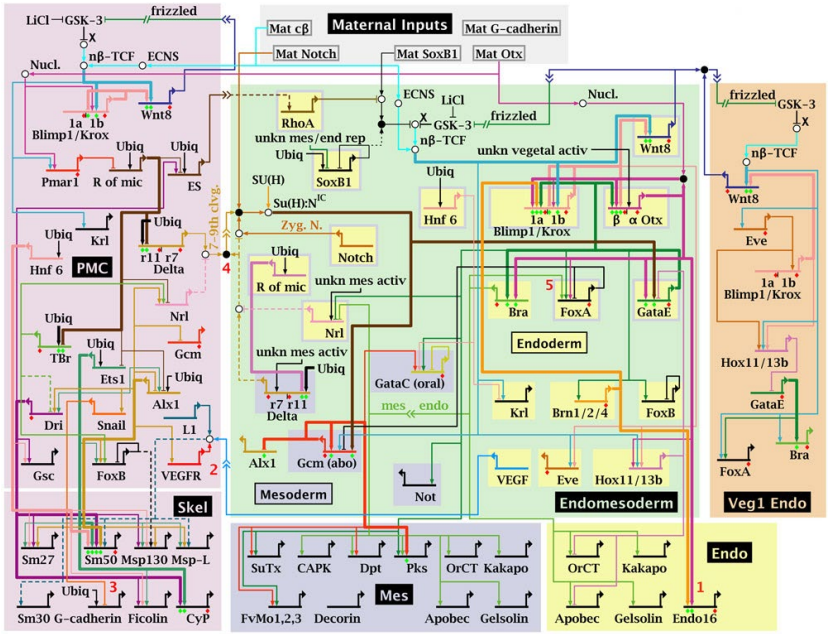
Representation learning

We need a theory of causal representation learning!

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

Gene regulation and structural equation models

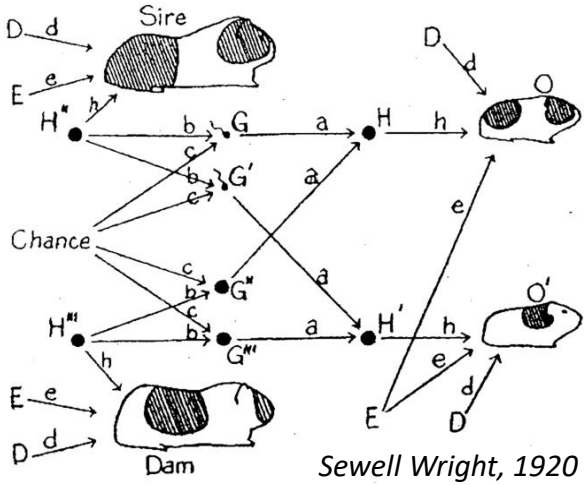
Ex: Gene regulatory network for pregastrular endomesoderm specification in sea urchins



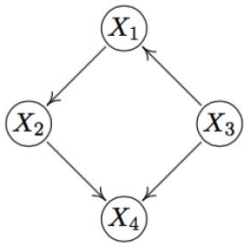
Eric H. Davidson, 2006

Sewell Wright developed the foundation of causal inference by studying heredity

causal structural equation models



Sewell Wright, 1920



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

Causal relations given by directed network; each node is associated with a random variable; stochasticity introduced by independent noise variables ϵ_i

Defining interventional distributions:

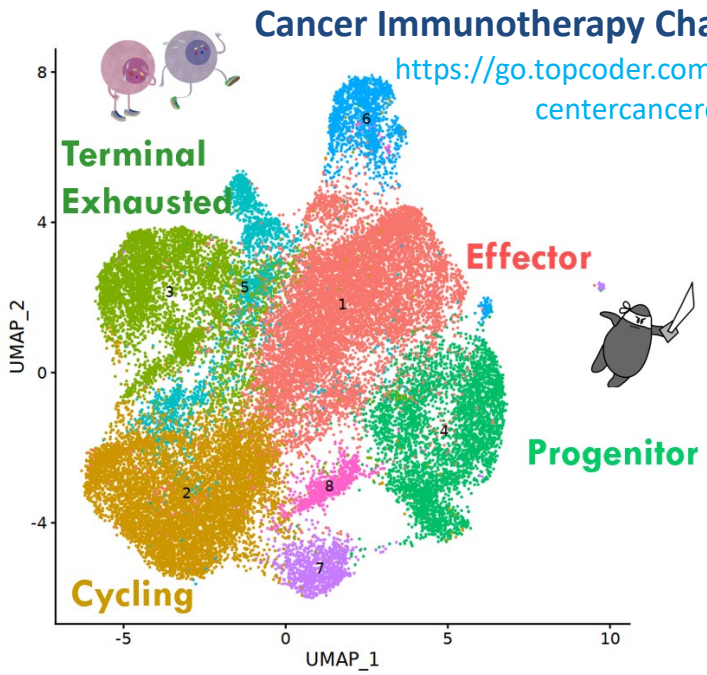
- Intervention on X_2: $\quad \mathrm{do}(X_2 = c)$
- $p(X_3 \mid \mathrm{do}(X_4 = c)) = p(x_3) \neq p(x_3 \mid x_4)$, but $p(X_4 \mid \mathrm{do}(X_3 = c)) = p(x_4 \mid x_3) \neq p(x_4)$

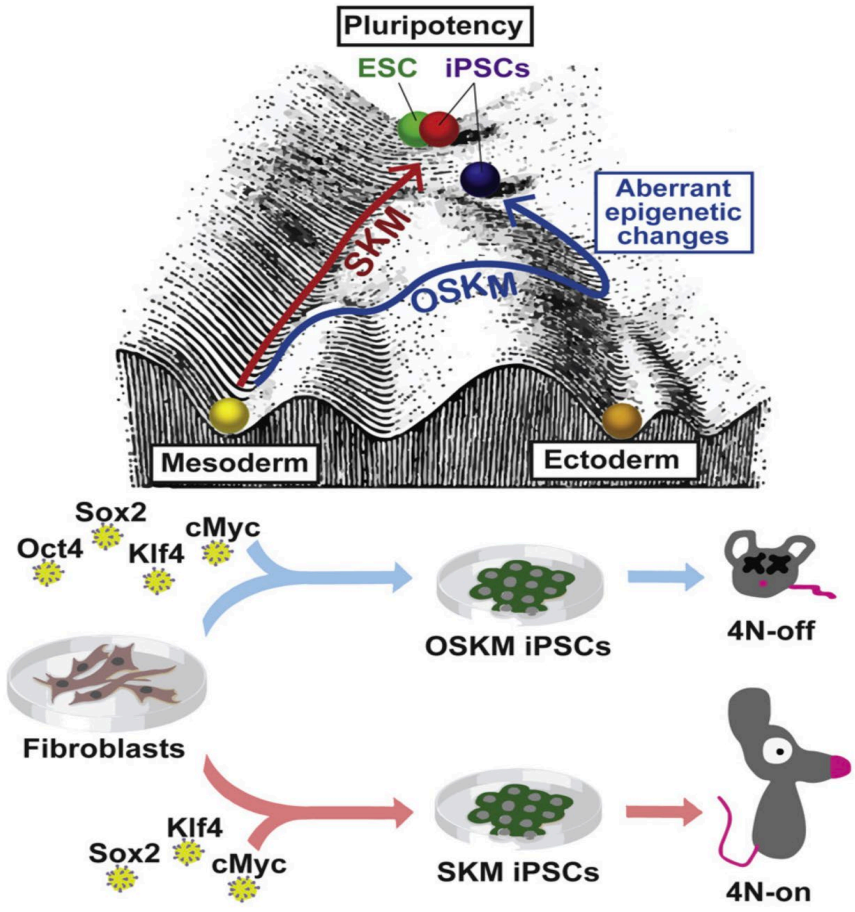
Intervention defines probabilistic operation that is different from conditioning and marginalization

Judea Pearl, 1995

Motivation: Cell state engineering

- **Engineering cell states:** rejuvenation, regenerative & personalized medicine
- Achievable e.g. through: combinations of transcription factors (humans ~2000)





https://www.sciencedirect.com/science/article/pii/S1934590919304230

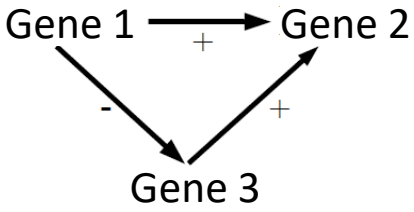
Space of possible perturbations and relevant contexts is combinatorial & continuous!

Causal structure discovery

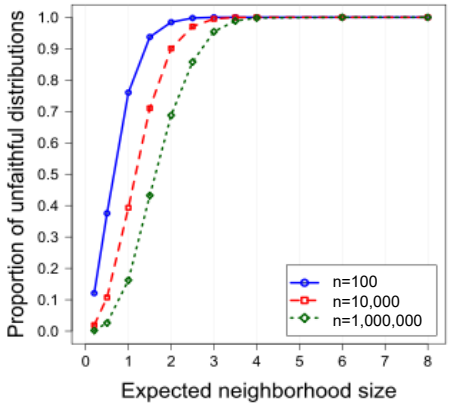
UAI contributions: Malinsky, Rios, Moffa, Kuipers, Kiyavash, Choo, Shiragur, Claassen, Mooij, Koivisto, Evans, Cussens, Richardson, Cooper, Shimizu, Meek,…

Foundations for learning causal networks from observational data were developed at CMU by **Spirtes, Glymour & Scheines** 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

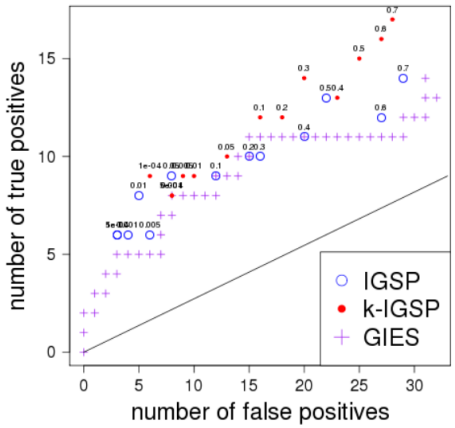
Uhler et al., Ann. Statist., 2013; Raskutti & Uhler, Stat, 2018

- **Problem:** number of conditional independence tests
- Developed greedy sparest permutation algorithm that is consistent under strictly weaker conditions [Wang, Solus, Yang & U., '17]
- Building on **Eberhardt**'s formalism, we extended this to first provably consistent algorithm for inferring causal network from observational & interventional data
- Computationally scales to graphs with 1000s of nodes, but not performance-wise

Yang, Katcoff & U., '18, Squires, Wang & U., '20,

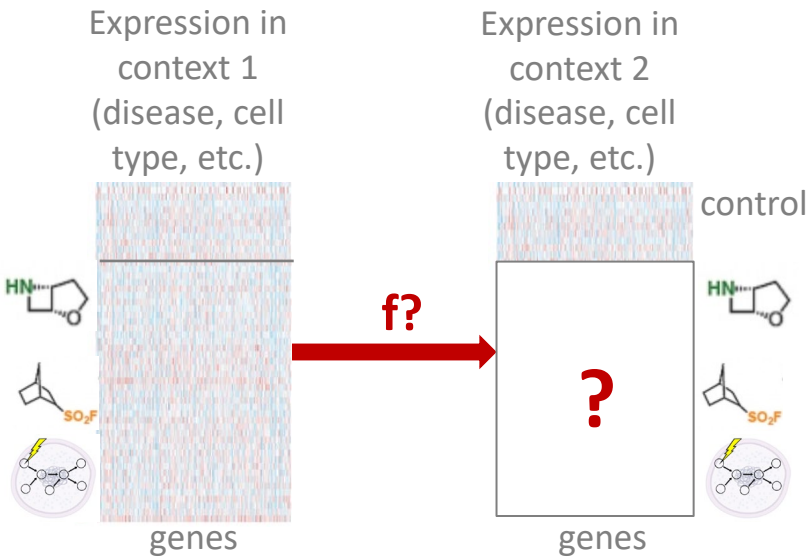
Recent review: Squires & U., Causal structure learning: a combinatorial perspective, FoCM 2022

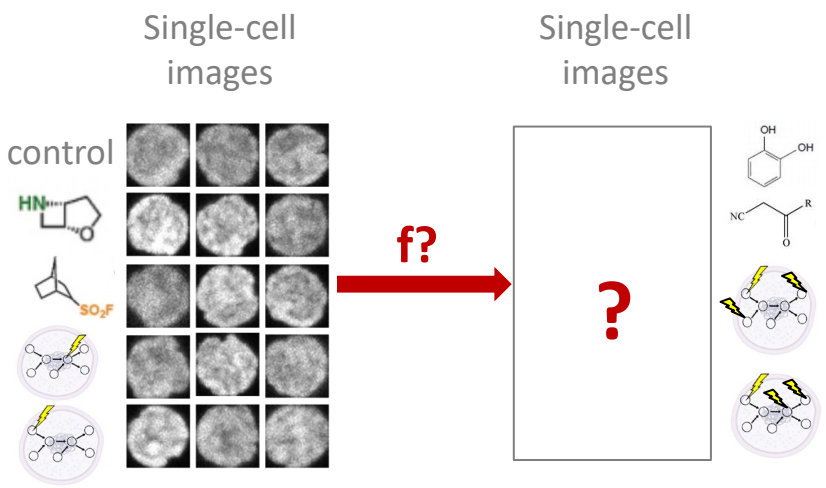
24 nodes, 15,000 samples



Causal transport and multi-modality

How to predict the effect of unseen interventions/perturbations?





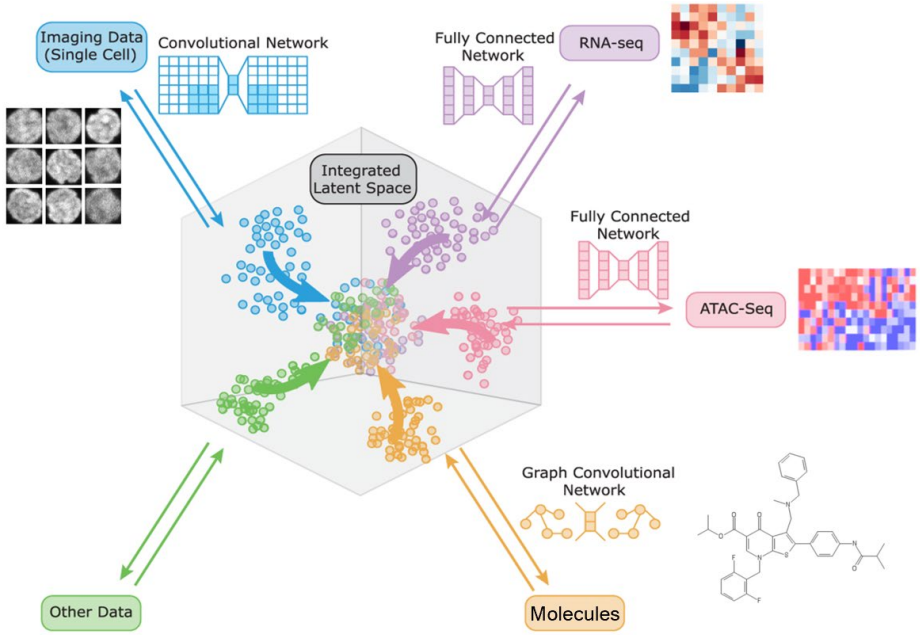
Transport to new contexts

Transport to new perturbations

How to think of causal variables in images?
Can multi-modality help?

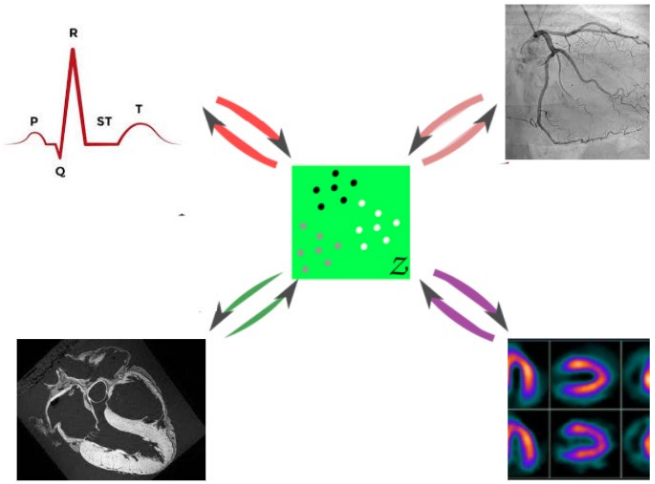
Multi-modal autoencoders for learning causal features

Multi-modal integration / translation



Yang et al., ICML Workshop 2019 & Nature Communications, 2021

Multi-modal integration for genetic association studies



Radhakrishnan et al.,
Nature Communications 2023

Multi-modal 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

Related work:

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

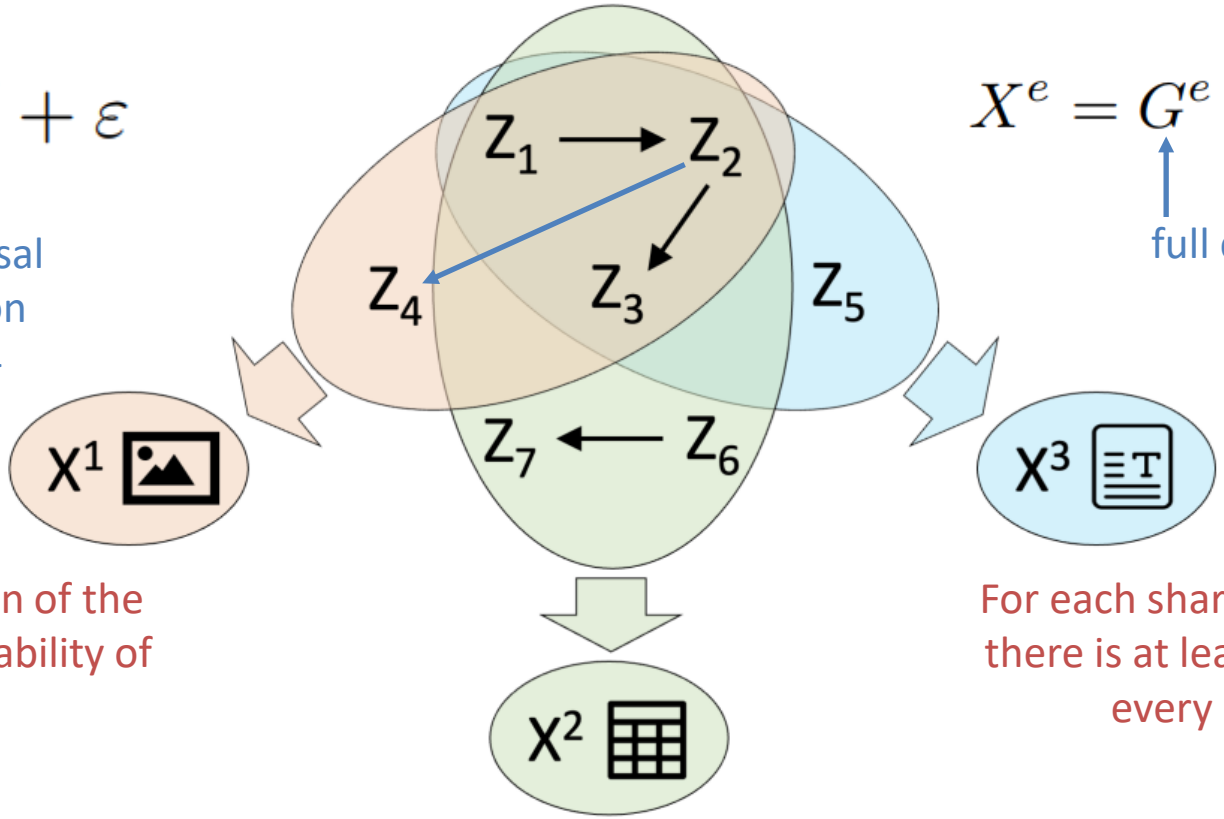
Learning latent causal graph from multi-modal data

$$Z = AZ + \varepsilon$$

linear additive causal structural equation model with non-Gaussian noise

$$X^e = G^e \cdot \begin{pmatrix} Z_{\mathcal{L}} \\ Z_{I_e} \end{pmatrix}$$

full column rank



Allows application of the results on identifiability of linear ICA

For each shared latent node there is at least one child in every domain

Theorem: The number of shared latent nodes and the joint domain distribution is identifiable. If there are no edges between the shared and domain-specific latent components and each shared latent node has at least 2 pure children, then also the shared latent graph is identifiable.

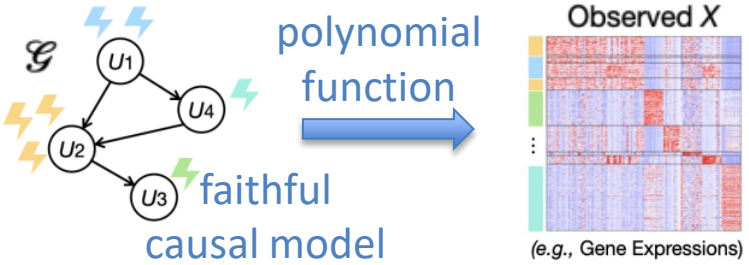
Sturma, Squires, Drton & Uhler, arXiv:2302.00993

(building heavily on Kun Zhang's recent work)

Learning latent causal graph from interventional data

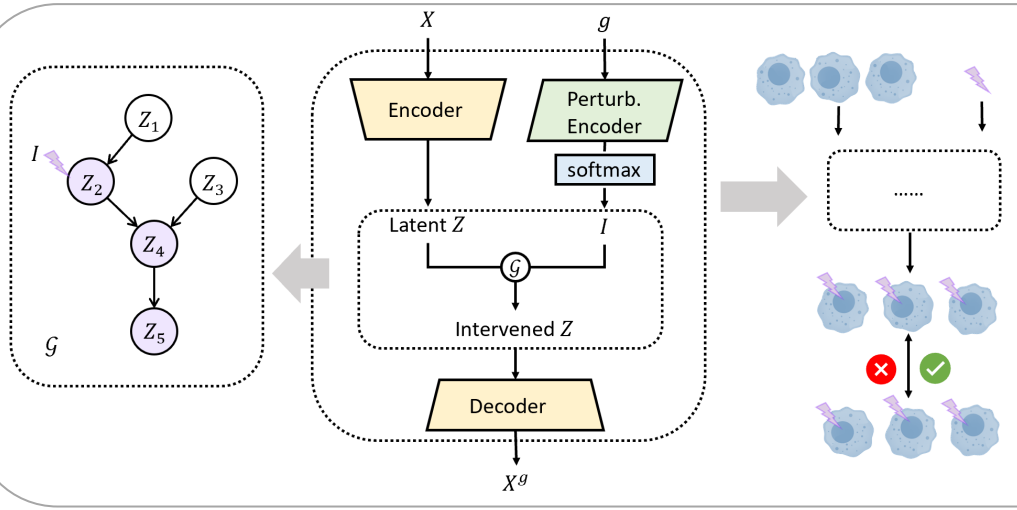
Given observational and perturbational data X, X^g:

- learn a generative model for $\mathbb{P}(X^g|X, g)$,
- the grouping of targeted variables $I = \{g\}$,
- and the causal graph between I.



Theorem: If interventional data from at least one intervention per latent node is available, then the latent interventional targets and the causal structure between the latent variables are identifiable (up to permutation), in theory as well as algorithmically using our discrepancy-VAE.

Zhang, Squires, Greenewald, Srivastava, Shanmugam & Uhler, arXiv:2307.06250



Discrepancy-based VAE Loss:

$$\mathbb{E}_g\{\mathbb{E}_X[\ \mathbb{E}_{Z|X} \log p_\theta(X|Z) - D_{KL}\left(q_\phi(Z|X) \parallel \mathcal{N}(0, I)\right) - \alpha MMD\left(\hat{X}^g, X^g\right)\ ]\}$$

Reconstruction — Prior: Standard VAE loss (ELBO)

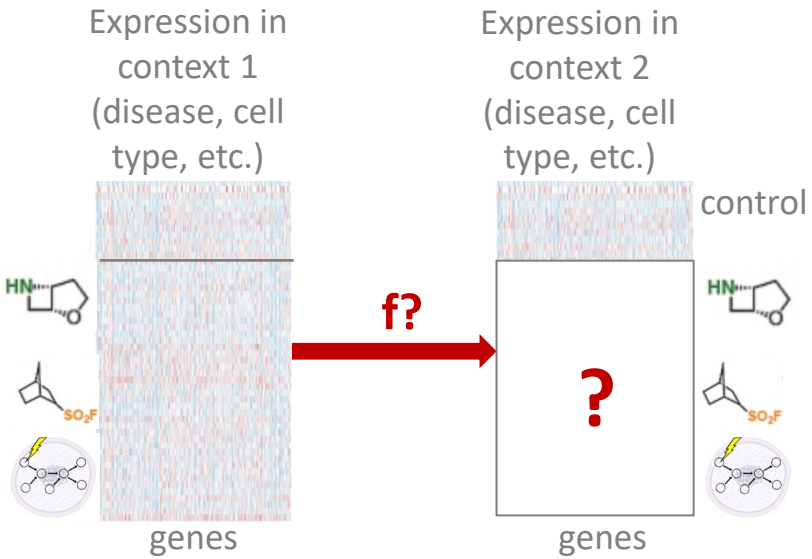
Distributional loss: Max Mean Discrepancy

Related: Identifiability results under hard interventions in linear model *Squires, Seigal, Bhate & Uhler, ICML 2023*

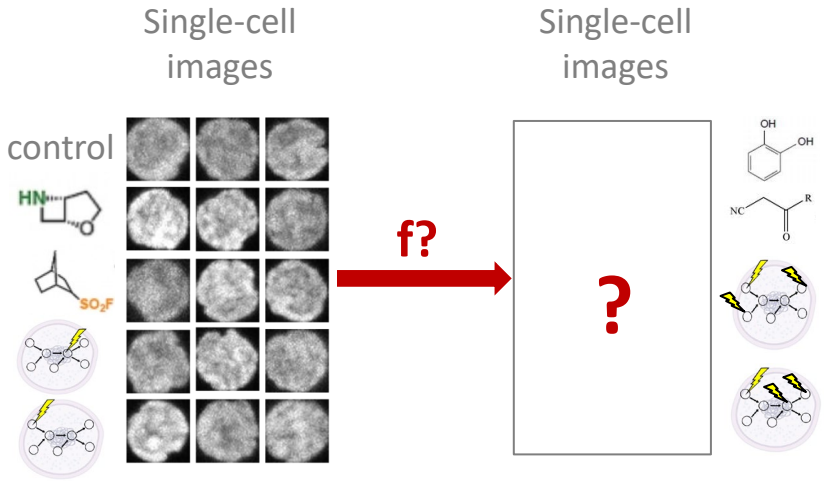
Causal variable learning: Kun Zhang, Eberhardt, Sridhar, Hartford,... **Disentanglement:** Schoelkopf, Bengio,...

Causal transport: blackbox or causal model?

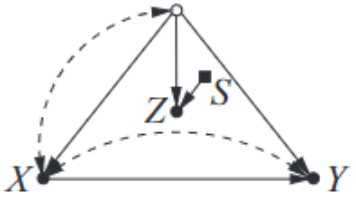
Transport to new contexts



Transport to new perturbations

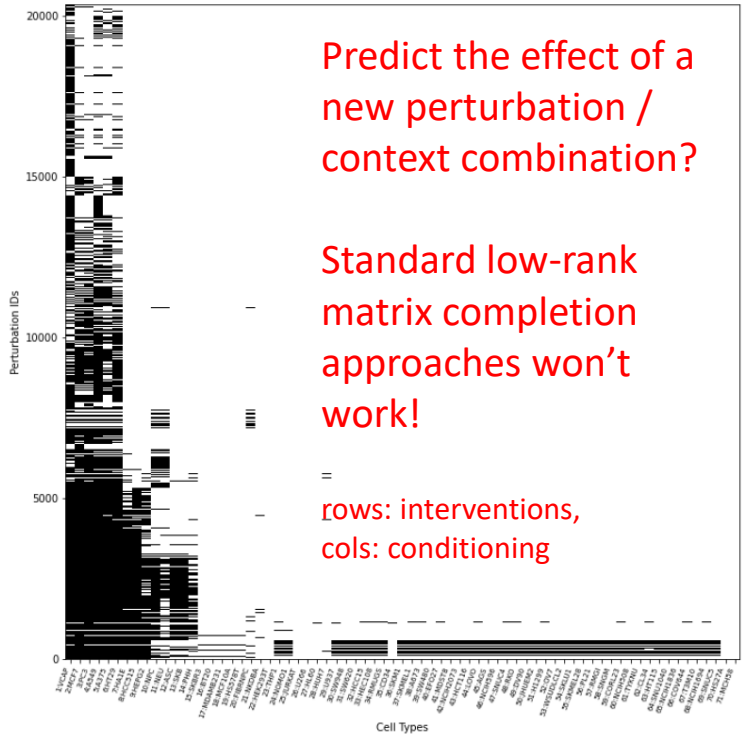


Given the causal graph, then necessary and sufficient conditions for causal transportability (i.e. transport across contexts) are known [Bareinboim & Pearl, NeurIPS 2014, PNAS 2016, etc.]



Black-box versus causal model: Transport problem to new genetic perturbations seems the hardest problem because of missing prior?

Causal matrix completion using neural tangent kernel



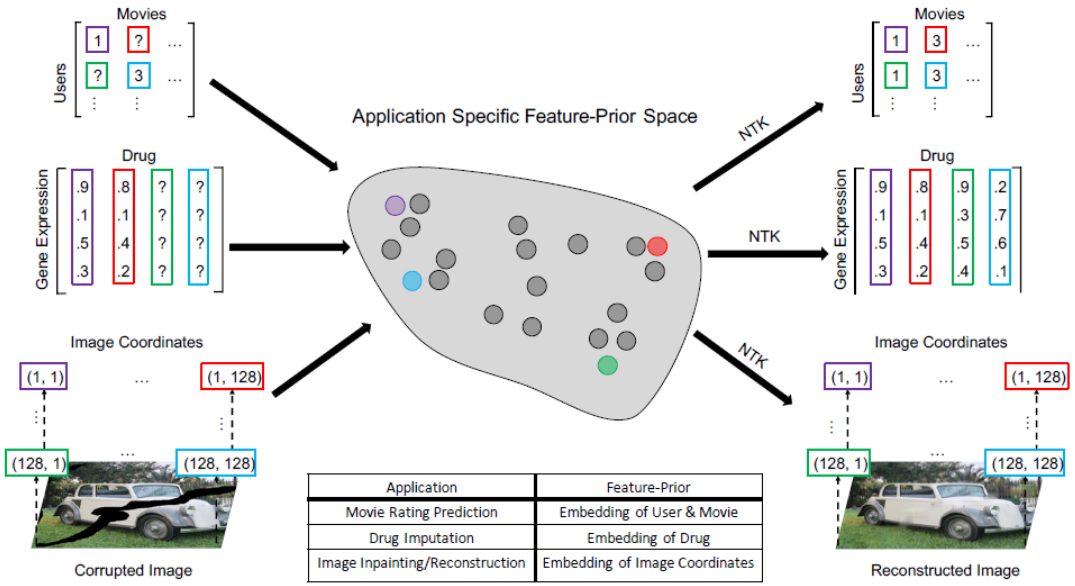
Predict the effect of a new perturbation / context combination?

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

=

rows: interventions,
cols: conditioning

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

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

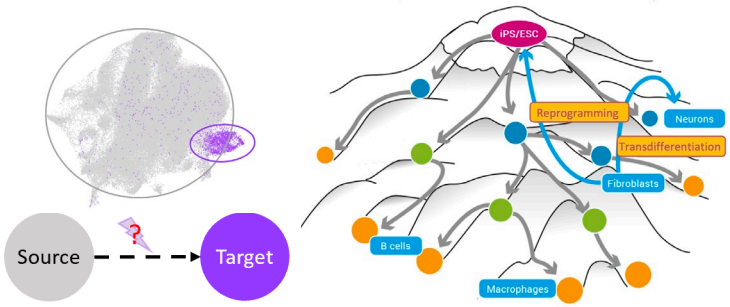
Related work: Synthetic controls / interventions: Shah, Agarwal, Abadie

Where we are headed: from prediction to control

If we are able to predict the effect of an unseen intervention, we should be able to optimize interventions to induce a particular cell state transition

Active learning of interventions:

For cell state engineering:



Algorithm iteratively updates causal model belief using samples acquired so far from different interventions, and selects next intervention that is most informative and will move the distribution to the desired state using causally informed acquisition function:

$$\min_{g} \sum_{g'} \omega(g') \cdot Var(\| m_\theta(g') - m_{target} \|^2 \mid \mathcal{D} \vee g)$$

All candidate perturbations

Distance estimate *Augmented dataset*

Uncertainty of estimating goodness of a candidate

Our acquisition function is theoretically sound (information-theoretic bounds, provably recovers optimal intervention) and computationally efficient (closed-form evaluation and fast gradient-based optimization for linear Gaussian SEM and the problem of mean matching)

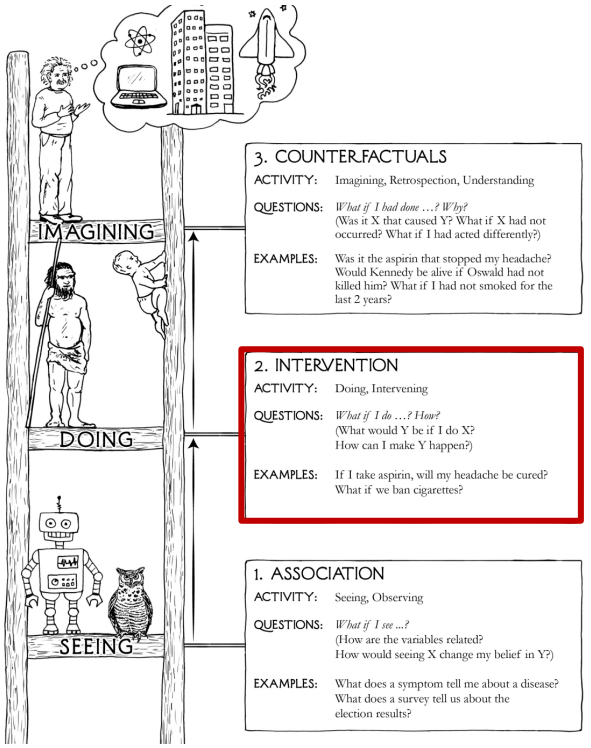
Zhang, Squires et al., to appear in Nature Machine Intelligence 2023

Related work: Active learning/Bayesian opt/Bandits: Gulchin, Aglietti, Bareinboim; also: Zhang, Squires & U., NeurIPS 2021

Biomedical sciences are uniquely suited not only to being one of the greatest beneficiaries of research in causality/ML but also one of the greatest sources of inspiration for it.

- While in biology we have access to large-scale multi-modal and interventional datasets, the underlying causal model incl. causal variables is generally unknown
- Optimally making use of large-scale multi-modal and interventional datasets requires a theoretical and algorithmic framework for causal representation learning
- Many open problems regarding how to optimally combine approaches from representation learning with causality
- Concentrated on Pearl's level 2 (predicting the effect of unseen interventions): instead of combining methods from level 1 and level 3, do we need a completely new theoretical and algorithmic framework?

Pearl's causal hierarchy



J. Pearl, The Book of Why, 2018

Acknowledgments

I mainly presented work led by 3 of my PhD students:

Jiaqi Zhang

Chandler Squires

Adit Radhakrishnan

PhD students:

- Louis Cammarata
- Daniel Paysan
- Adityanarayanan Radhakrishnan
- Hannah Schlueter
- Chandler Squires
- Yitong Tseo
- Chenyu Wang
- Jiaqi Zhang
- Xinyi Zhang

with contributions from my whole group

Postdocs:

- Matthew Amodio
- Salil Bhate
- Wengong Jin
- Kiran Shiragur

MSc/undergraduate students

- Alvaro Ribot Barrado
- Cathy Cai
- Emily Liu
- Nten Nyiam
- Ishika Shah

Funding:

ERIC AND WENDY SCHMIDT CENTER AT BROAD INSTITUTE

DEPARTMENT OF THE NAVY ONR Science & Technology

NIH

SIMONS FOUNDATION

IBM

NSF