Beyond ICA: Causal Disentanglement via Interventions

Chandler Squires

04/19/2023

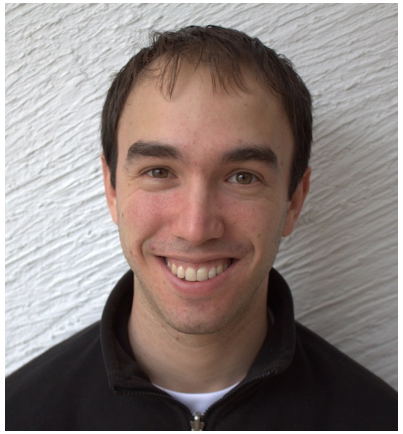
Anna Seigal

Salil Bhate

Caroline Uhler

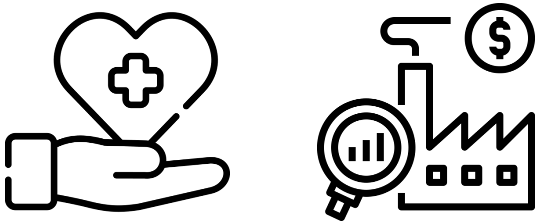
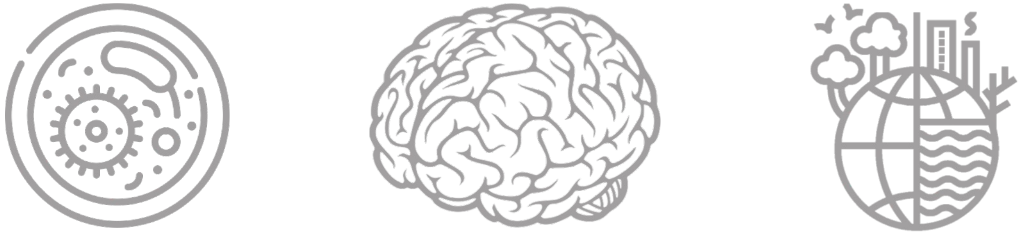
Nils Sturma

Matthias Drton



David Sontag

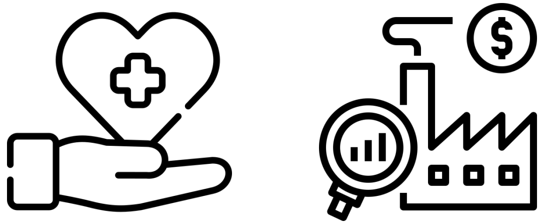
Goal: Introduce the tools of causal reasoning to new, complex domains.

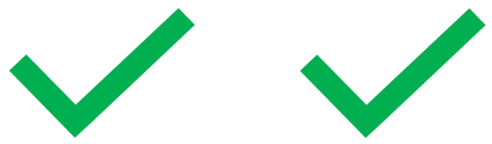
		Known causal graph?	Known causal variables?
Type 1 domains: *causally familiar*	**Causal inference** 	✓	✓
Type 2 domains: *conceptually familiar*	**Causal structure learning** 	✗	✓
Type 3 domains: *conceptually novel*	**Causal representation learning** 	✗	✗

Known causal graph?	Known causal variables?

Type 1 domains:
causally familiar

Causal inference



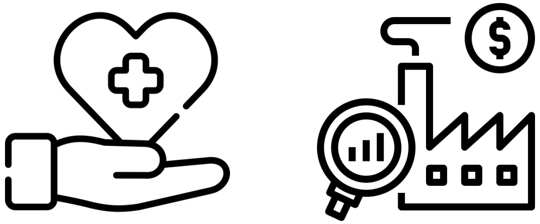


Much heavy lifting is done by humans. Causal relationships are determined from subconscious “common sense” principles or by conscious, domain-specific reasoning.

Analogous to “rule-based” systems in artificial intelligence.

	Known causal graph?	Known causal variables?
Type 1 domains: *causally familiar*	✓	✓

Causal inference



Rich and active area of research, including several topics:

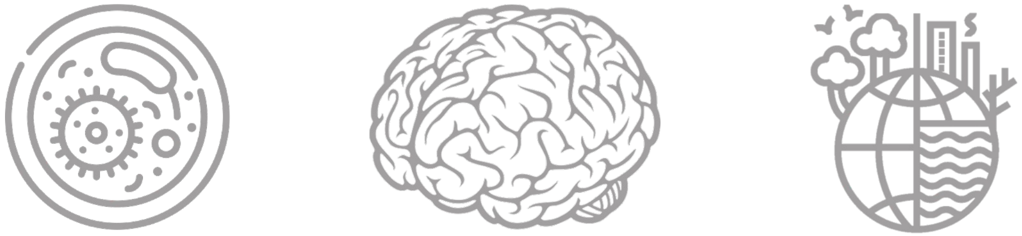
- Identifiability and transportability (Shpitser '06, Drton '16, Lee '20)
- Instrumental variable methods (Newey '03, Singh '19)
- Proxy variable methods (Miao '18, Kallus '21)
- Sensitivity analysis
- …

		Known causal graph?	Known causal variables?
Type 1 domains: *causally familiar*	**Causal inference**	✓	✓
Type 2 domains: *conceptually familiar*	**Causal structure learning**	✗	✓
Type 3 domains: *conceptually novel*	**Causal representation learning**	✗	✗

Causal inference

Causal structure learning

Causal representation learning



		Known causal graph?	Known causal variables?
Type 2 domains: *conceptually familiar*	**Causal structure learning**	✗	✓

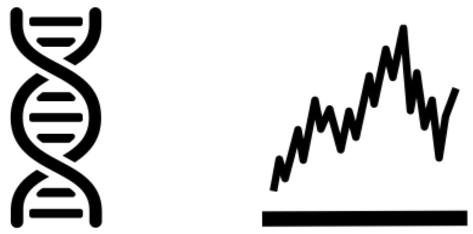
Less heavy lifting is done by humans. Practitioners pick the relevant variables, often by designing technologies to measure these variables. Machines learn the causal relationships.

Analogous to “feature engineering” in machine learning.

Known causal graph? Known causal variables?

Type 2 domains:
conceptually familiar

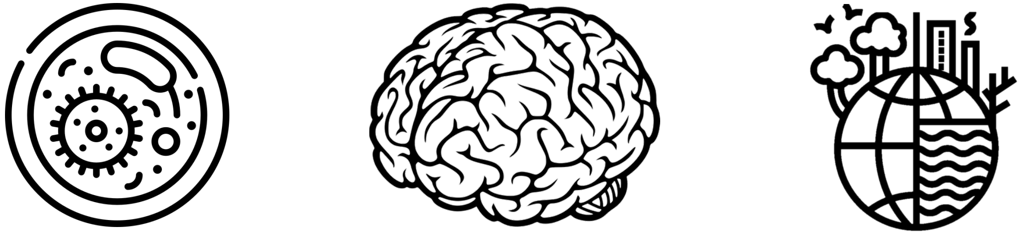
Causal structure learning

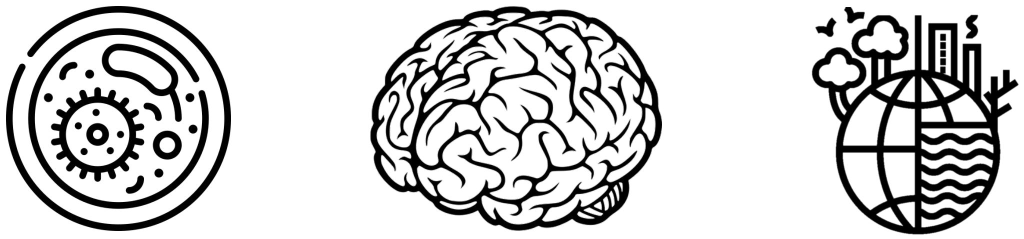


My own "home" area of research, very active area:

- Differentiable approaches (Zheng '18, Lachapelle '19, Brouillard '20)
- Bayesian methods (Friedman '03, Lorch '21, Castelletti '22)
- Interventions and multiple environments (Eaton '07, Hauser '12, Mooij '20)
- Targeted approaches (Peters '16, Wang '18)
- Experimental design (Eberhardt '05, Hyttinen '13, Agrawal '19)
- …

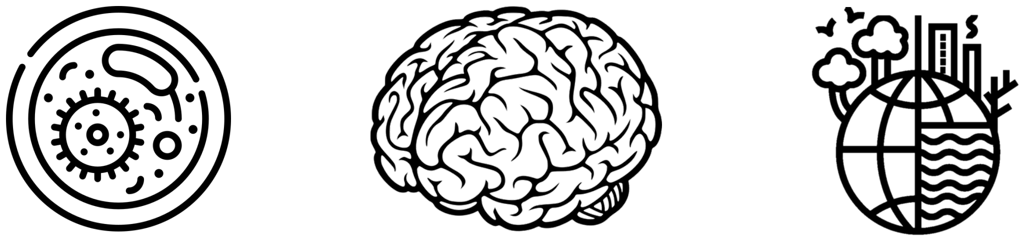
		Known causal graph?	Known causal variables?
Type 1 domains: *causally familiar*	Causal inference	✓	✓
Type 2 domains: *conceptually familiar*	Causal structure learning	✗	✓
Type 3 domains: ***conceptually novel***	**Causal representation learning**	✗	✗



		Known causal graph?	Known causal variables?
Type 3 domains: *conceptually novel*	**Causal representation learning** 	✗	✗

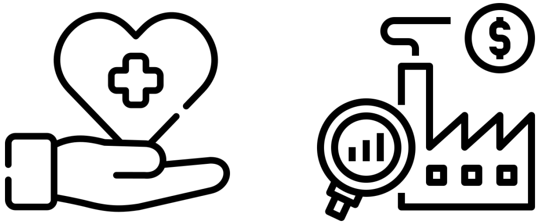
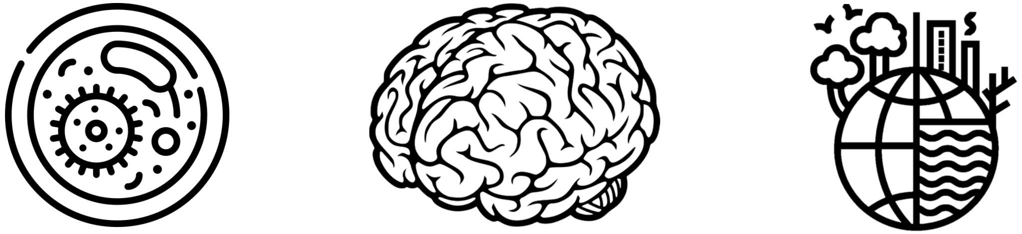
Involves high-dimensional measurements of complex systems with which humans have little or no direct experience. Thus, it is infeasible to rely on humans for any heavy lifting.

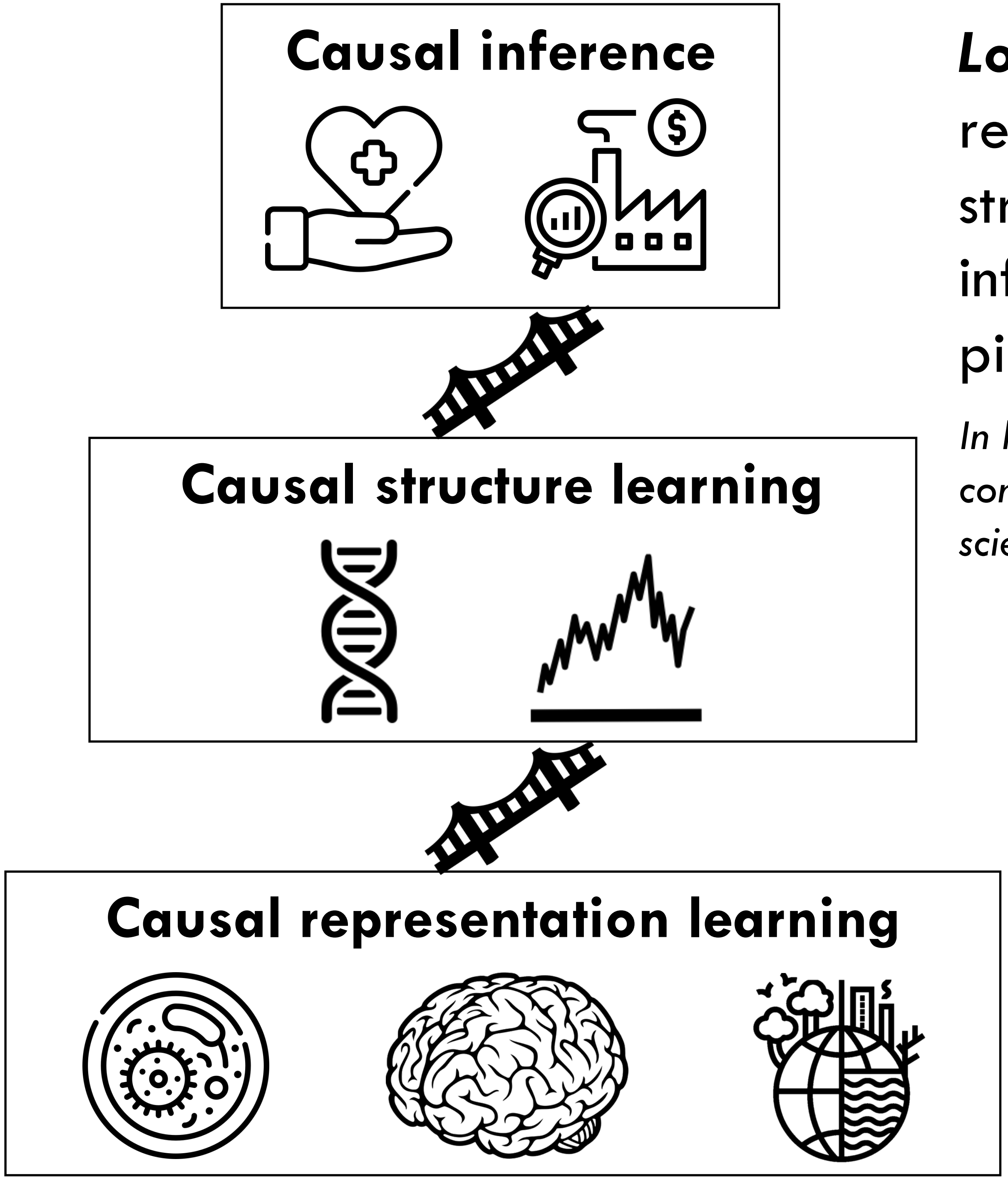
Analogous to “feature learning” in machine learning.

		Known causal graph?	Known causal variables?
Type 3 domains: *conceptually novel*	**Causal representation learning** 	✗	✗

An emerging area of research

- Learning latent DAGs from observational data (Silva ’06, Cai ’19, Kivva ‘21, Xie ‘22)
- Causal feature learning (Chalupka ’15, ‘16, ’17)
- Domain generalization (Arjovsky ’19, Rosenfeld ’20, Zhou ‘22)
- Learning latent DAGs from paired counterfactual data (Brehmer ’22, Ahuja ‘22)
- Learning latent DAGs from interventional data (Ahuja ‘22, Liu ‘22, **Squires ‘23**, Varici ‘23)
- …

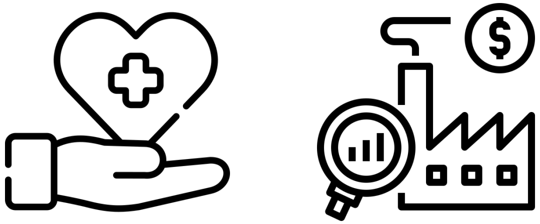
		Known causal graph?	Known causal variables?
Type 1 domains: *causally familiar*	**Causal inference** 	✓	✓
Type 2 domains: *conceptually familiar*	**Causal structure learning** 	✗	✓
Type 3 domains: *conceptually novel*	**Causal representation learning** 	✗	✗



Long-term goal: Treat representation learning, structure learning, and inference as a single pipeline.

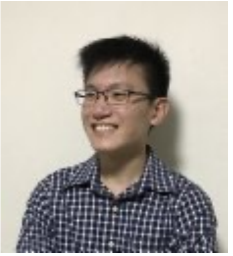
In line with Bin Yu's vision of considering the entire "data science life cycle".

Causal inference



Post-selection inference
(Chernozhukov '15, Gradu '22)

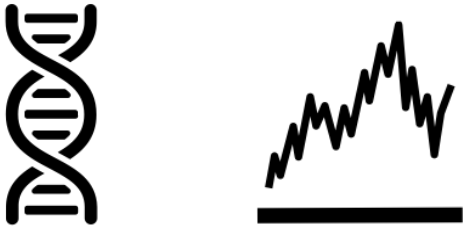
Decision-centric causal structure learning
(ongoing work)



Davin Choo

Invariant causal prediction
(Peters '16, Heinze-Deml '18)

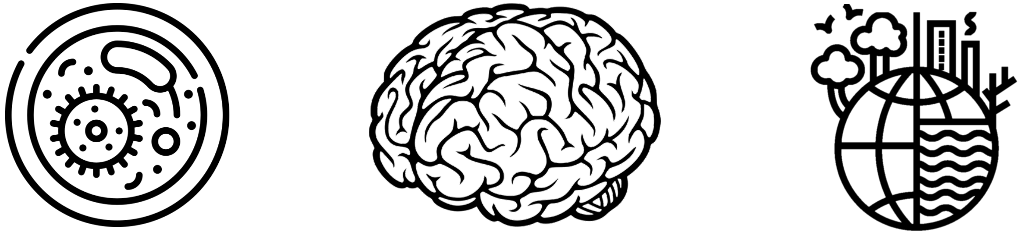
Causal structure learning



Domain generalization
(Arjovsky '19)

Future work

Causal representation learning



Vapnik's principle: When solving a problem of interest, do not solve a more general problem as an intermediate step.

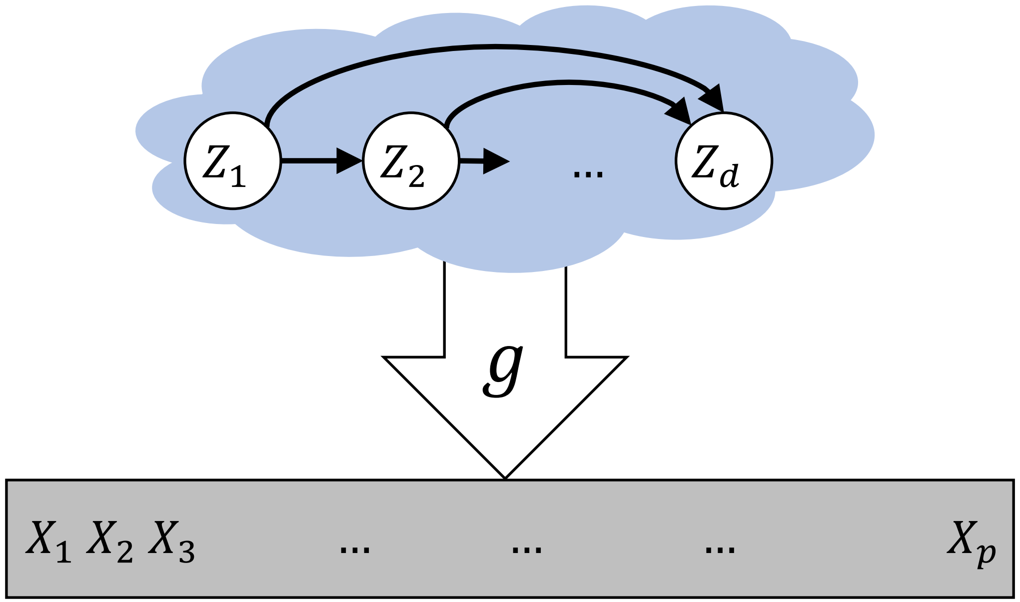
Two addenda to Vapnik's principle

- **The "hidden structure" principle.**
  - Optimally solving the problem of interest might require leveraging hidden structure that is only apparent when solving the general problem.
    - Estimating composite functions (Baraud '14)
    - Prediction-centric learning (Karzand '15, Bresler '16, Boix-Adsera '21)
    - Semi-parametric inference: need to estimate "nuisance functions".
  - *If we solve an intermediate problem, we must take that into account: avoid simply "plugging in".*

- **The "first move" principle.**
  - The general problem can be a rich source of intuition and insight.
  - Good place to develop techniques.

The first move:
Causal Disentanglement

Macro-variables



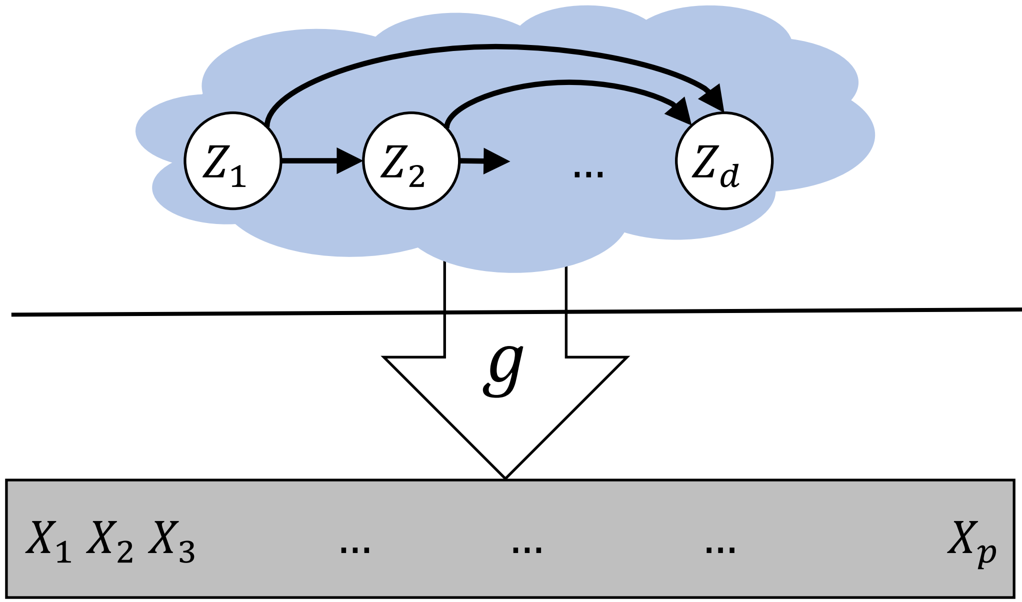
Micro-variables

View causal representation from a generative modeling perspective.

Can we infer the latent variables?

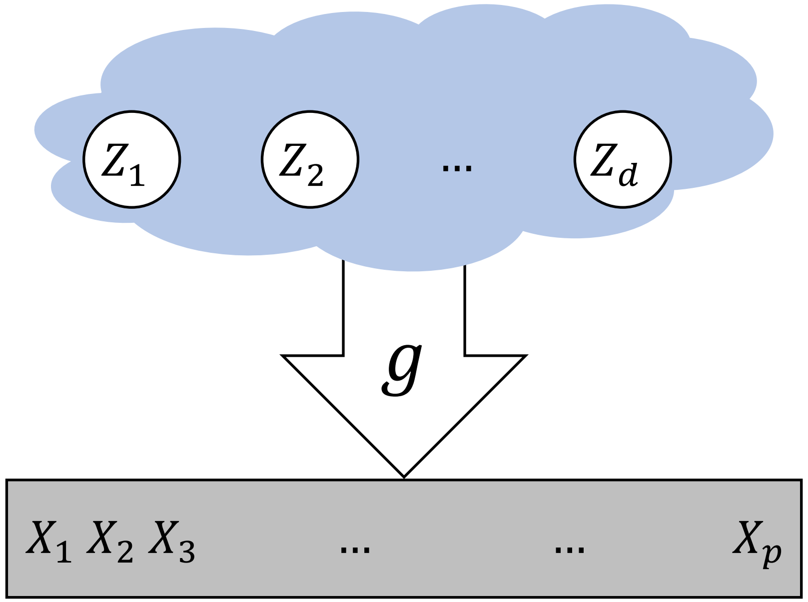
Permutation indeterminacy*: the macro-variables can always be re-labeled so that 1,2,3,…,d is a topological order.*

	Cellular Biology	Neuroscience
Macro-variables	• Protein concentrations • Cellular morphology (e.g. nucleus shape)	• Neurotransmitter concentrations • Reuptake rate
Micro-variables	• Fluorescent microscopy images • Gene expression (RNAseq)	• Neuroimaging data (fMRI) • Electrical activity (LFP)



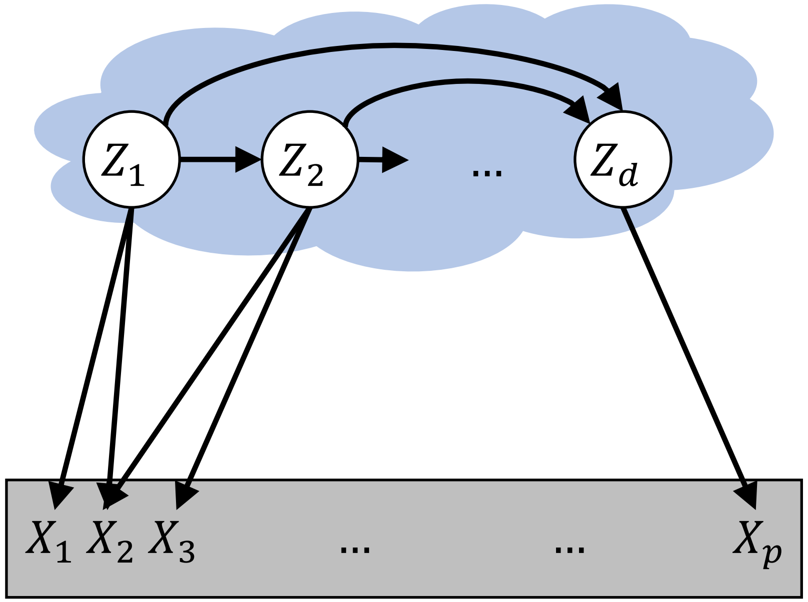
Possible approaches

Restrict latent DAG $\mathcal{G}$



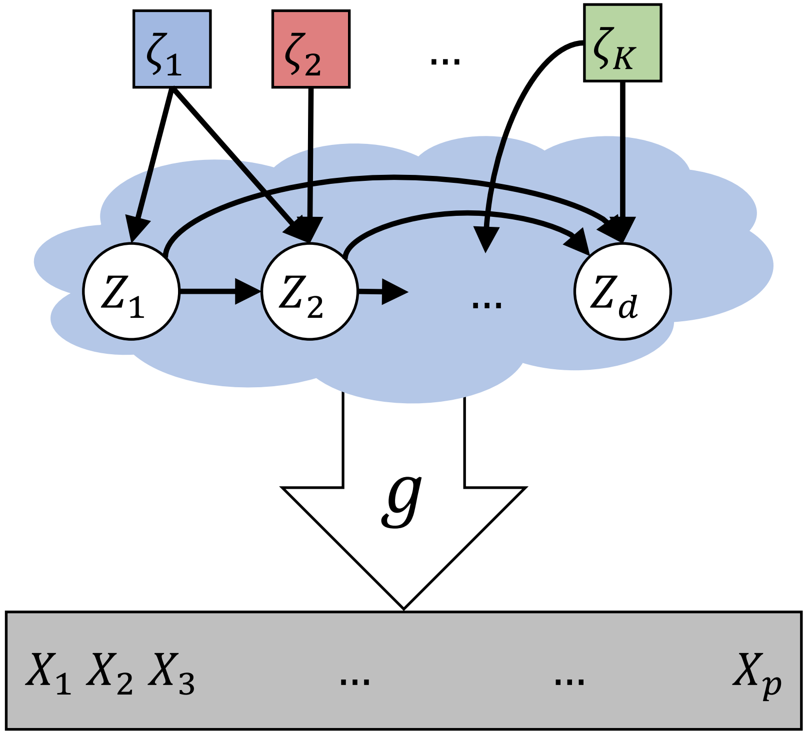
Linear ICA (Comon 1994)
Nonlinear ICA (Hyvärinen '19)

Restrict mixing function g

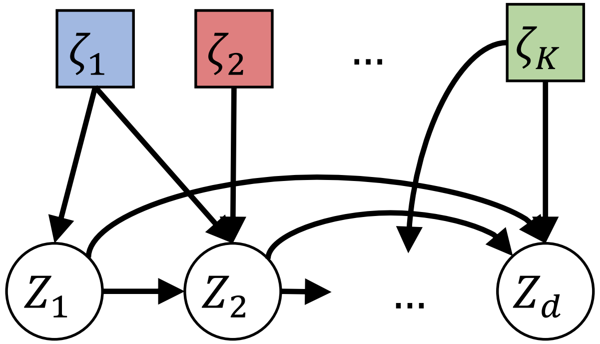


Most work on latent DAG recovery
(Silva '06, Halpern '15, Cai '19,
Kivva '21, Xie '20, Xie '22)

Learning from interventions



Squires '23
Liu '22, Ahuja '22, Varici '23



Control			…
$Z_1 = f_1(\varepsilon_1)$	$Z_1 = f_1'(\varepsilon_1)$	$Z_1 = f_1(\varepsilon_1)$	
$Z_2 = f_2(Z_1, \varepsilon_2)$	$Z_2 = f_2'(Z_1, \varepsilon_2)$	$Z_2 = f_2''(Z_1, \varepsilon_2)$	
$\vdots$	$\vdots$	$\vdots$	
$Z_d = f_d(Z_1, Z_2, \dots, \varepsilon_d)$	$Z_d = f_d(Z_1, Z_2, \dots, \varepsilon_d)$	$Z_d = f_d(Z_1, Z_2, \dots, \varepsilon_d)$	

Do-intervention — Replaces mechanism with a constant — $Z_2 = \hat{z}_2$

Perfect intervention — Removes dependence of parents — $Z_2 = f_2'(\varepsilon_2)$

Soft intervention (mechanism shift) — Changes mechanism to any function — $Z_2 = f_2'(Z_1, \varepsilon_2)$

↓ More general

Wishlist

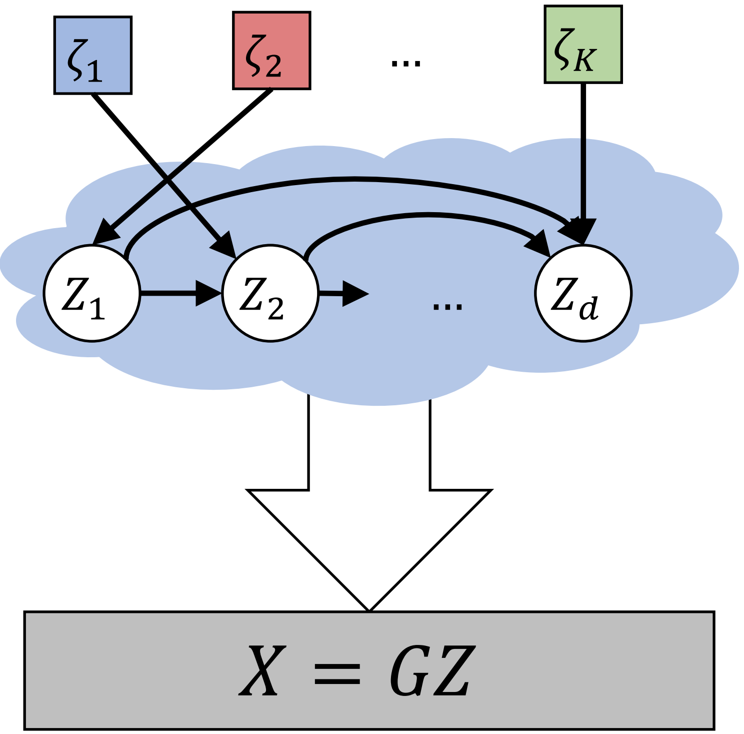
- Identifiability theory
  - Given any set of interventions, what indeterminacies remain (similar to Markov equivalence)?
- Algorithms
  - Score-based (e.g., penalized maximum likelihood)
    - Exact search
    - Greedy search
    - Gradient-based search
- Statistical and computational theory
  - Minimax rates
  - Rate-optimal algorithms

Linear Causal Disentanglement via Interventions

Chandler Squires, Anna Seigal, Salil Bhate, Caroline Uhler

Limitations

1. Single-node interventions
2. Linear mixing
3. Linear latent causal model



$G \in \mathbb{R}^{p \times d}$ with full column rank

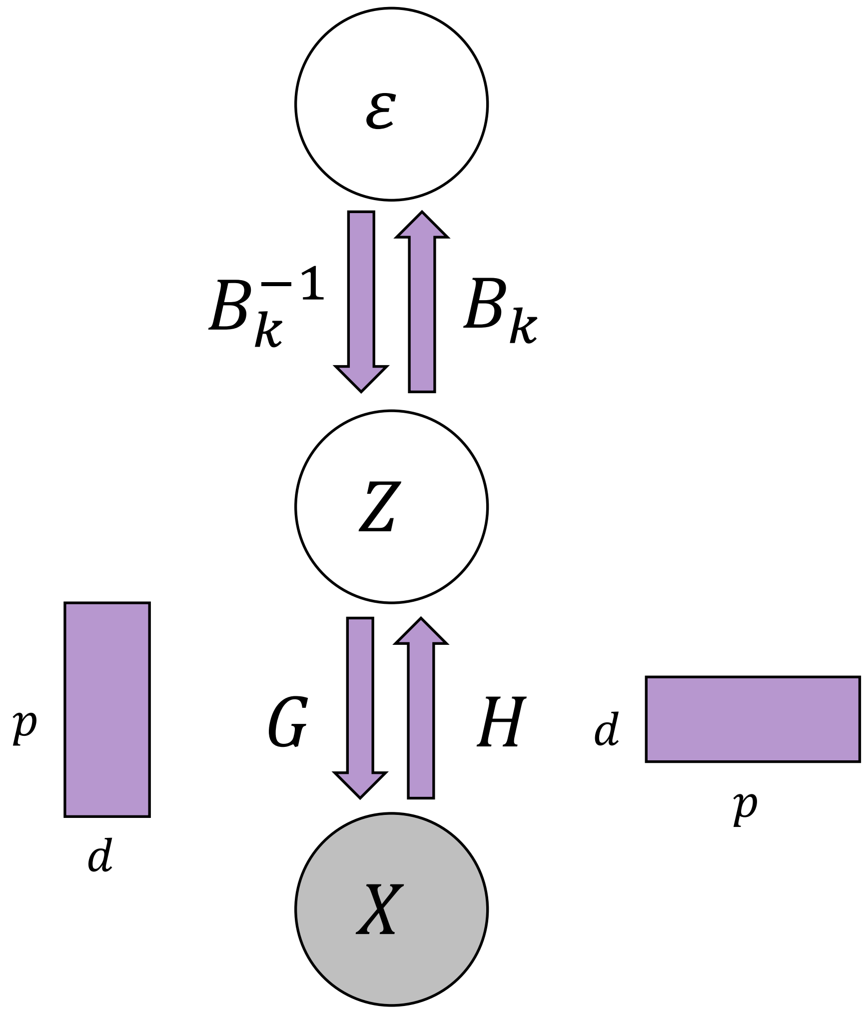
Control			…
$Z_1 = \sigma_1 \varepsilon_1$	$Z_1 = \sigma_1' \varepsilon_1$	$Z_1 = \sigma_1 \varepsilon_1$	
$Z_2 = A_{12} Z_1 + \sigma_2 \varepsilon_2$	$Z_2 = A_{12} Z_1 + \sigma_2 \varepsilon_2$	$Z_2 = A_{12}' Z_1 + \sigma_2' \varepsilon_2$	
$\vdots$	$\vdots$	$\vdots$	
$Z_d = A_{1d} Z_1 + A_{2d} Z_2 + \cdots + \sigma_d \varepsilon_d$	$Z_d = A_{1d} Z_1 + A_{2d} Z_2 + \cdots + \sigma_d \varepsilon_d$	$Z_d = A_{1d} Z_1 + A_{2d} Z_2 + \cdots + \sigma_d \varepsilon_d$	

Compact version:

In context k, $Z = A_k Z + \Omega_k^{1/2} \varepsilon.$

Equivalently,

$Z = B_k^{-1} \varepsilon$ for $B_k = \Omega_k^{-1/2} (I - A_k).$ ← Upper triangular



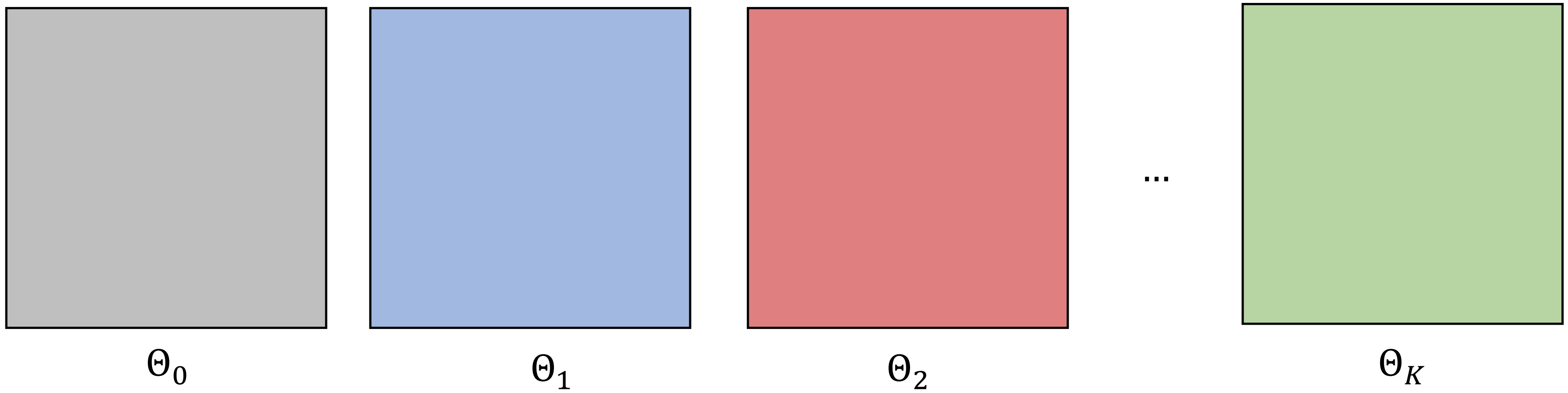
$H = G^{\dagger}$

$$\mathrm{Cov}(\varepsilon)^{-1} = I_d$$

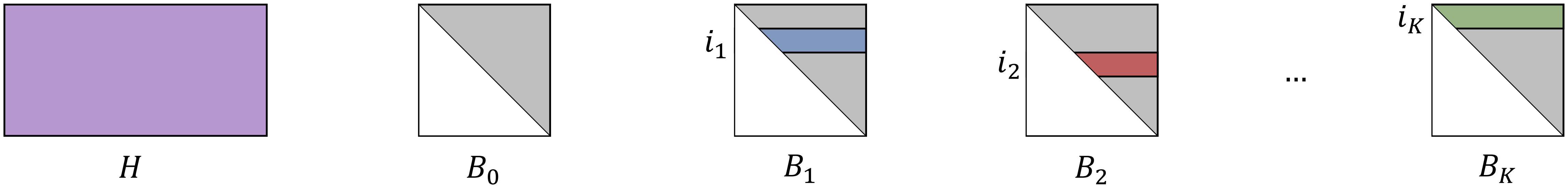
$$\mathrm{Cov}_k(Z)^{-1} = B_k^{\top} B_k$$

$$\Theta_k \coloneqq \mathrm{Cov}_k(X)^{-1} = H^{\top} B_k^{\top} B_k H$$

Input:



Output:



such that $\Theta_k = H^\top B_k^\top B_k H$ for all k.

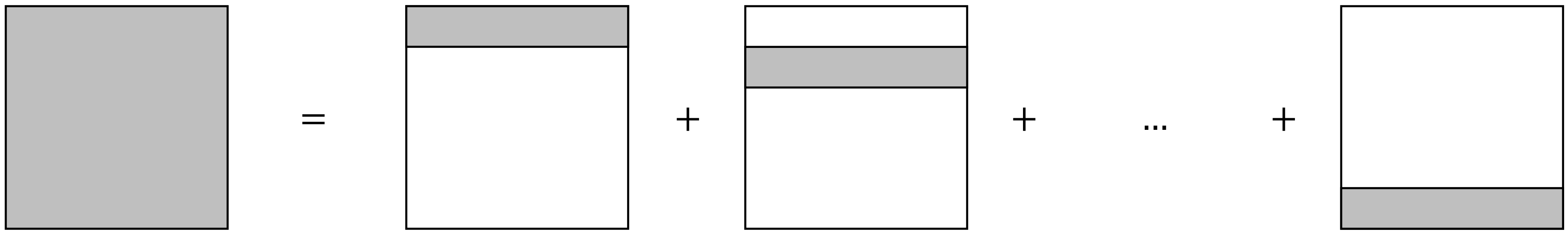
Theorem (perfect interventions): one intervention per latent node is **sufficient**, and in the worst-case, **necessary**, to recover $H = G^{\dagger}$ and $B_0, B_1, \ldots, B_K$.

Note: “Recovery” is only up to an indeterminacy that comes from re-labeling nodes.

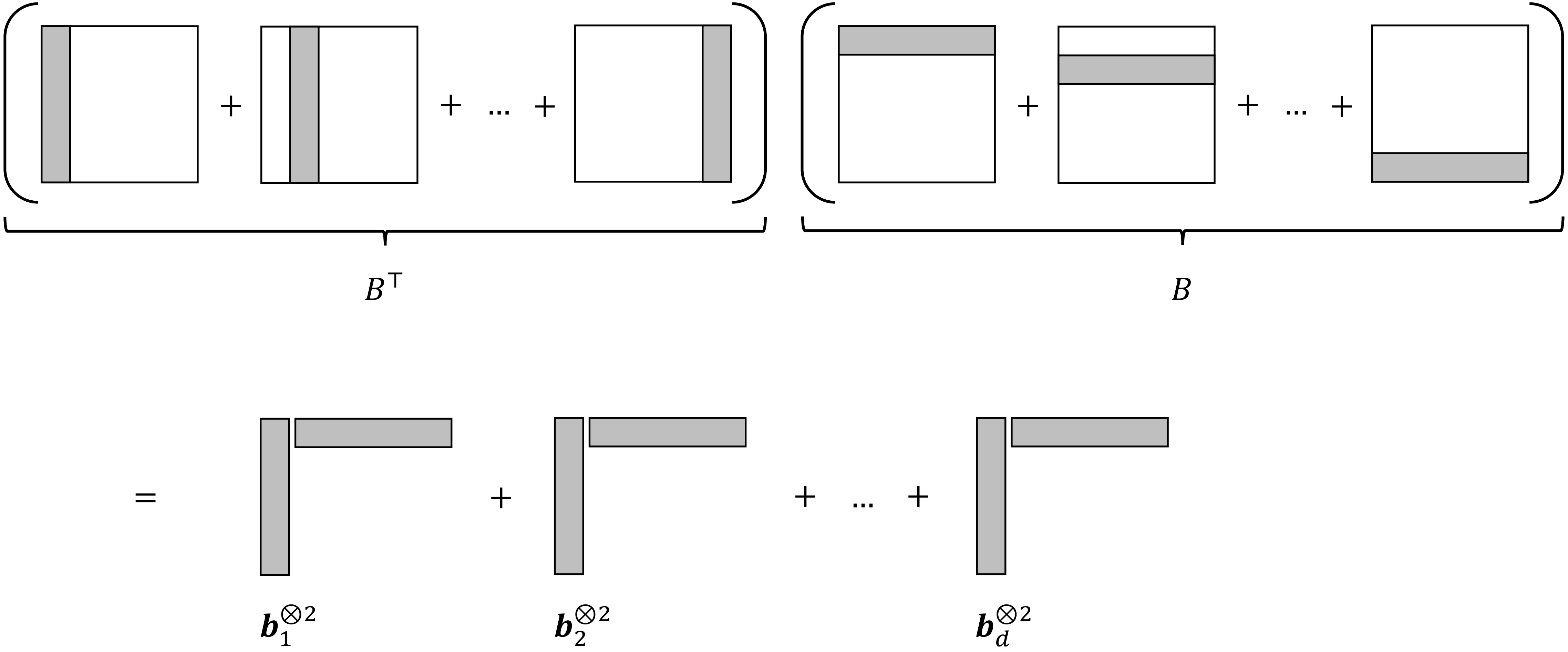
Theorem (soft interventions): one intervention per latent node is **sufficient**, and in the worst-case, **necessary**, to recover $\mathcal{G}$ up to transitive closure.

Note: “Recovery” is only up to an indeterminacy that comes from re-labeling nodes.

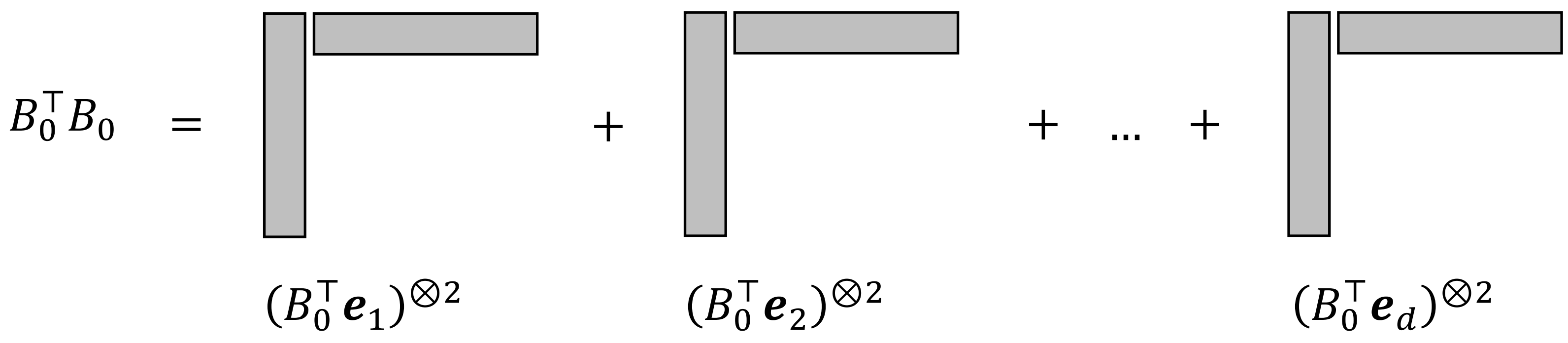
Proof of sufficiency (perfect interventions)

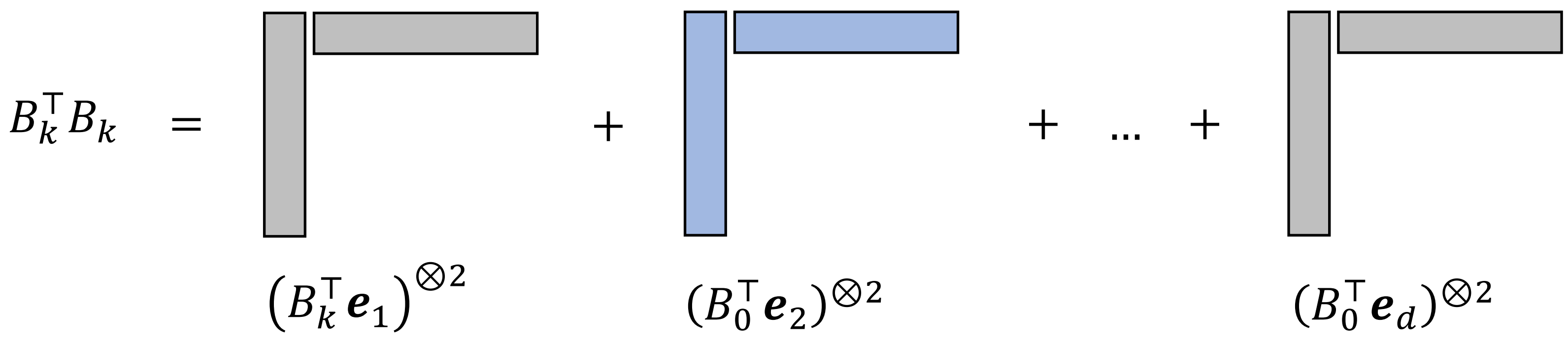


$$B = \boldsymbol{e}_1 \boldsymbol{b}_1^\top + \boldsymbol{e}_1 \boldsymbol{b}_1^\top + \ldots + \boldsymbol{e}_d \boldsymbol{b}_d^\top$$



$$\boldsymbol{v}^{\otimes 2} = \boldsymbol{v}\boldsymbol{v}^\top$$





$$\Rightarrow B_k^\top B_k - B_0^\top B_0 = \left(B_k^\top \boldsymbol{e}_{i_k}\right)^{\otimes 2} - \left(B_0^\top \boldsymbol{e}_{i_k}\right)^{\otimes 2}$$

$$\Rightarrow \Theta_k - \Theta_0 = \left(H^\top B_k^\top \boldsymbol{e}_{i_k}\right)^{\otimes 2} - \left(H^\top B_0^\top \boldsymbol{e}_{i_k}\right)^{\otimes 2}$$

Key identity:

$$\Theta_k - \Theta_0 = \left(H^\top B_k^\top \boldsymbol{e}_{i_k}\right)^{\otimes 2} - \left(H^\top B_0^\top \boldsymbol{e}_{i_k}\right)^{\otimes 2}$$

$$H^\top B_k^\top \boldsymbol{e}_{i_k} = \sum_{i \in \overline{pa}(i_k)} (B_k)_{i_k,i} \boldsymbol{h}_i$$

H

Thus, $\mathrm{rowspan}(\Theta_k - \Theta_0) \subseteq \langle \boldsymbol{h}_i : i \in \overline{pa}(i_k) \rangle$

$\Rightarrow \Theta_k - \Theta_0$ is rank one if i_k is a source node.

In fact, $\Theta_k - \Theta_0$ is rank two if i_k is not a source node.

Essential idea of the algorithm:

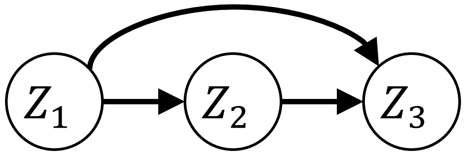
1. Use rank test to find source nodes.
2. Recover corresponding row of H up to scale.
3. "Get rid of" source nodes and repeat.

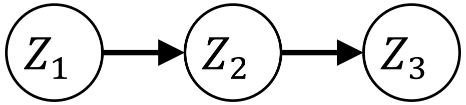
"Getting rid of" nodes:

- Form a vector space V from the already-recovered rows of H.
- Project $\Theta_k - \Theta_0$ onto the orthogonal complement of V.
- Subtleties involved in recovering a row of H instead of an orthogonal basis for H.

Other remarks on theoretical results

- **Worst-case necessity:** If we are missing an intervention on a sink node (a node with no children), we can't recover the corresponding row of H.
- **Soft interventions:** We can only recover the graph up to transitive closure, for example, we can't tell apart the two graphs below.

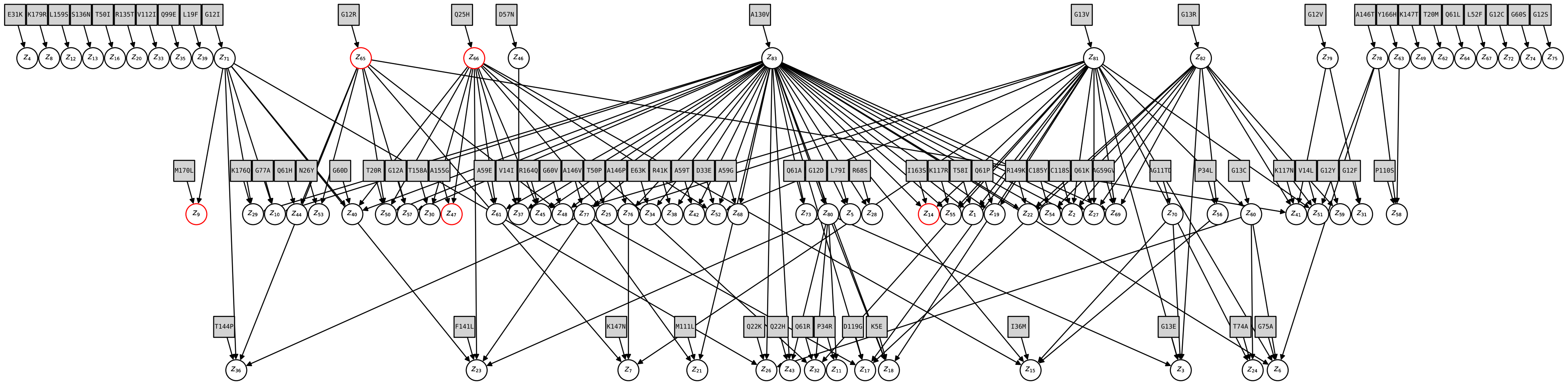




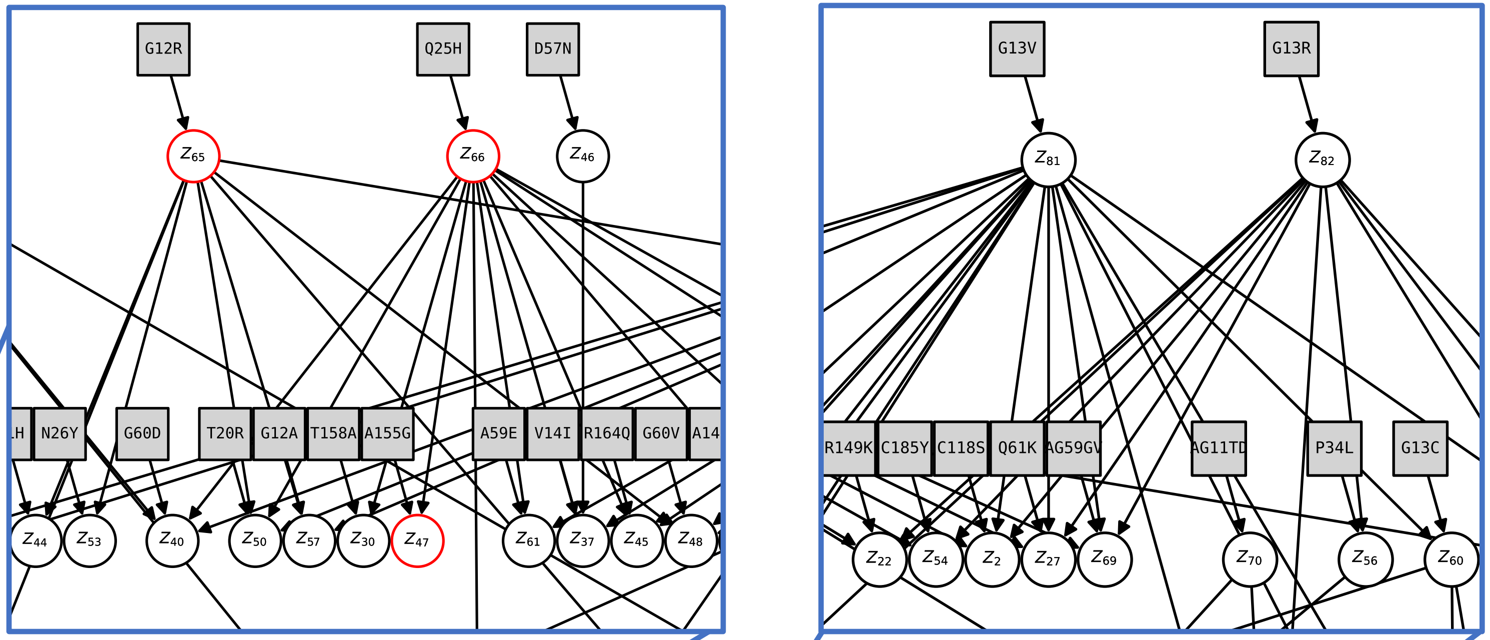
A hypothetical workflow

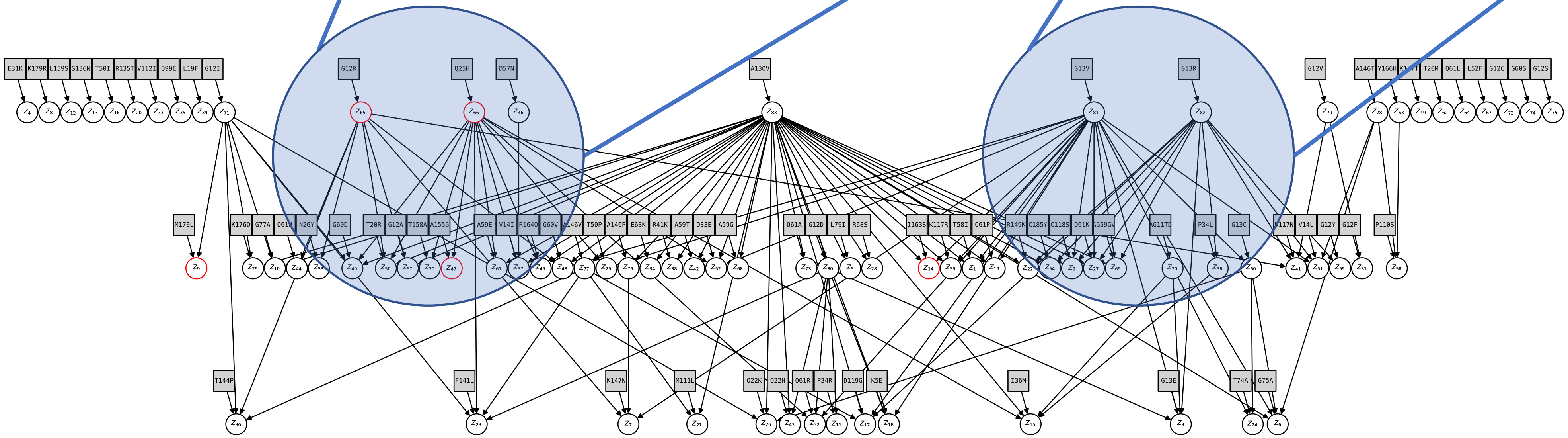
Biological application:

- Single-cell RNA sequencing of 90,000 lung cancer cells
- Contexts: $K = 83$ mutations of the KRAS oncogene
- Used $p = 83$ most variable genes as observed X variables.



G12 and G13 positions of KRAS: key functional residues that are known causal drivers of cancer.





Ongoing work

Extension to multi-node interventions

Álvaro Ribot

Cathy Cai

Extension to non-linear mixing

Jiaqi Zhang

Unpaired Multi-Domain Causal Representation Learning

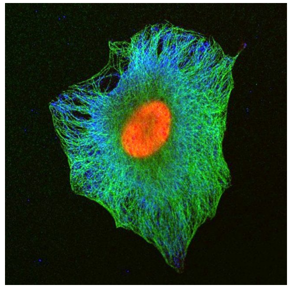
Nils Sturma, **Chandler Squires**, Matthias Drton, Caroline Uhler

Multiple modalities

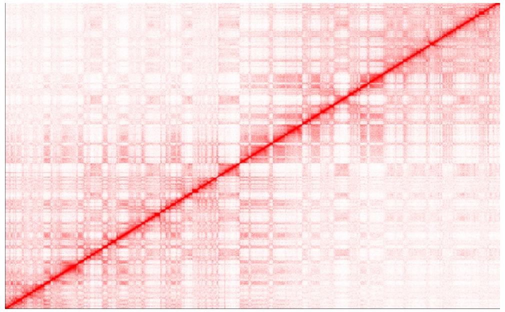
- Humans process the world through **sight, sound, smell, touch, taste…**
- Each input modality gives information about different, possibly overlapping, aspects of the world.
- Taken together, multiple modalities provide a richer picture than any single modality can provide on its own.

Multiple modalities in biology

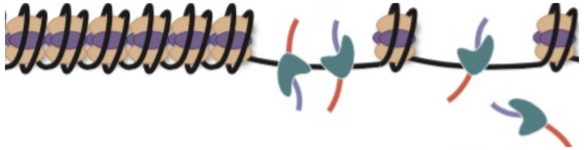
Gene expression



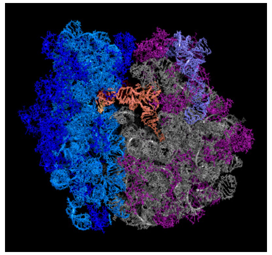
Fluorescent imaging



Chromosome organization

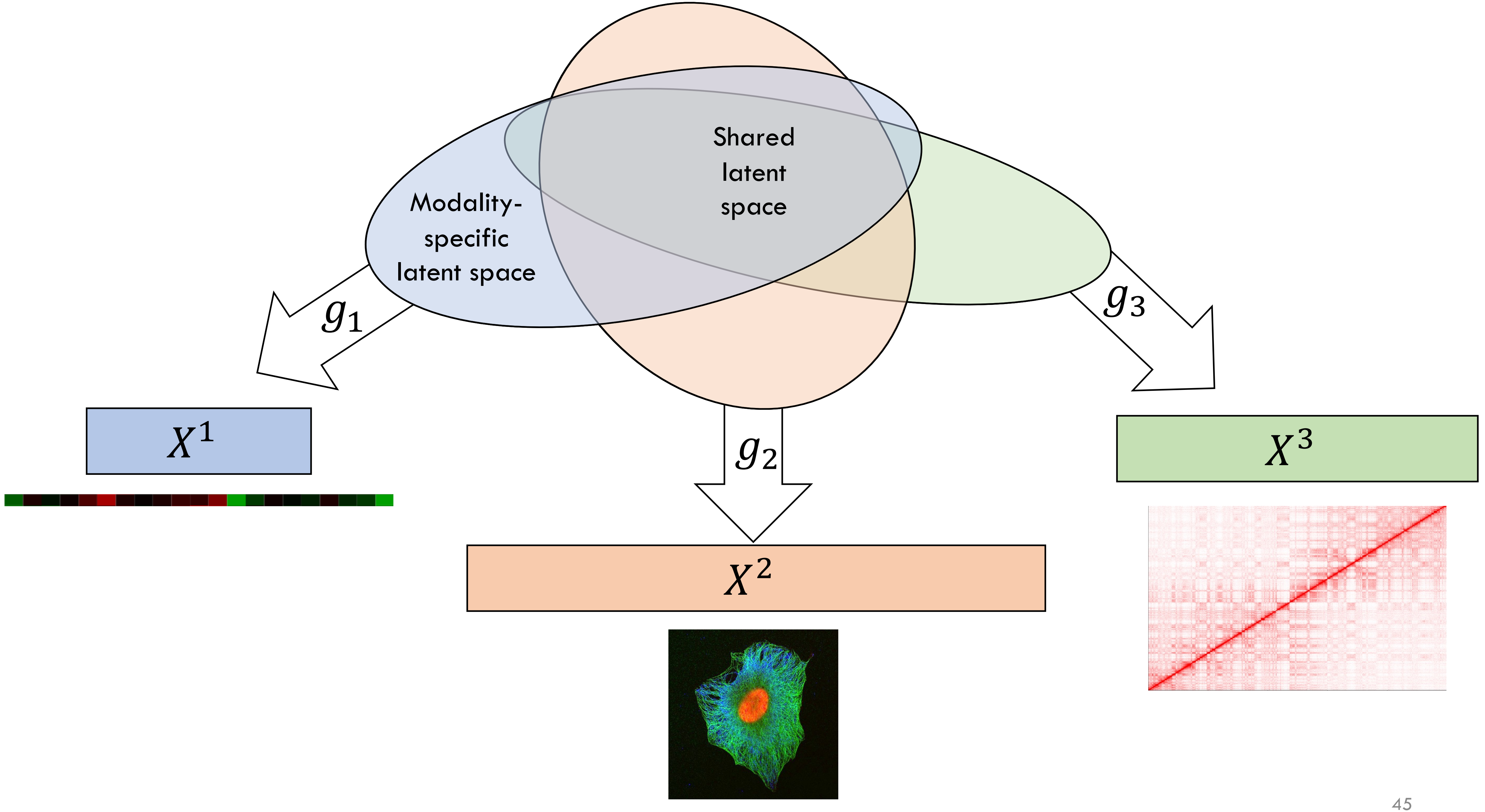


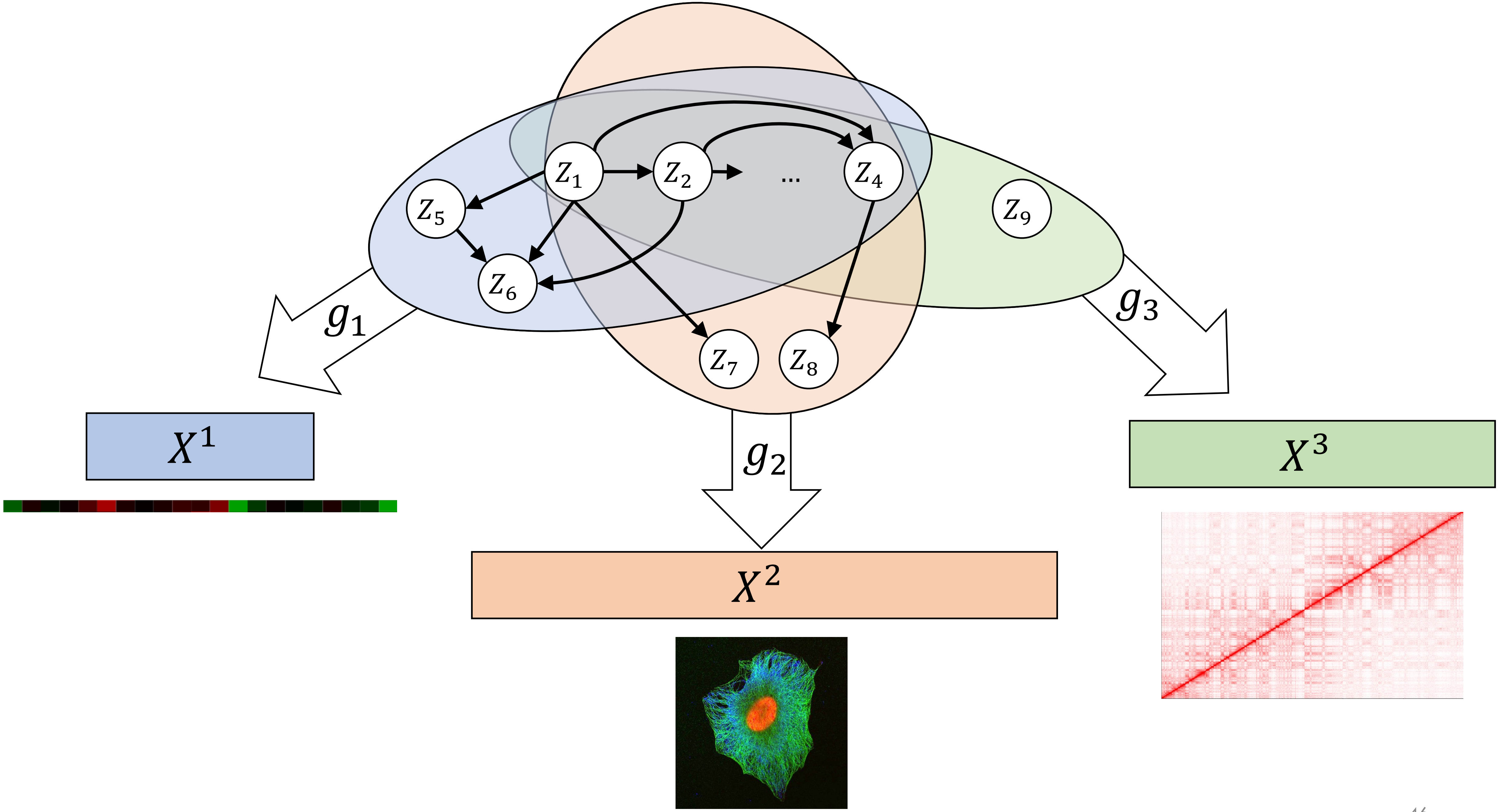
Chromatin accessibility



Protein expression

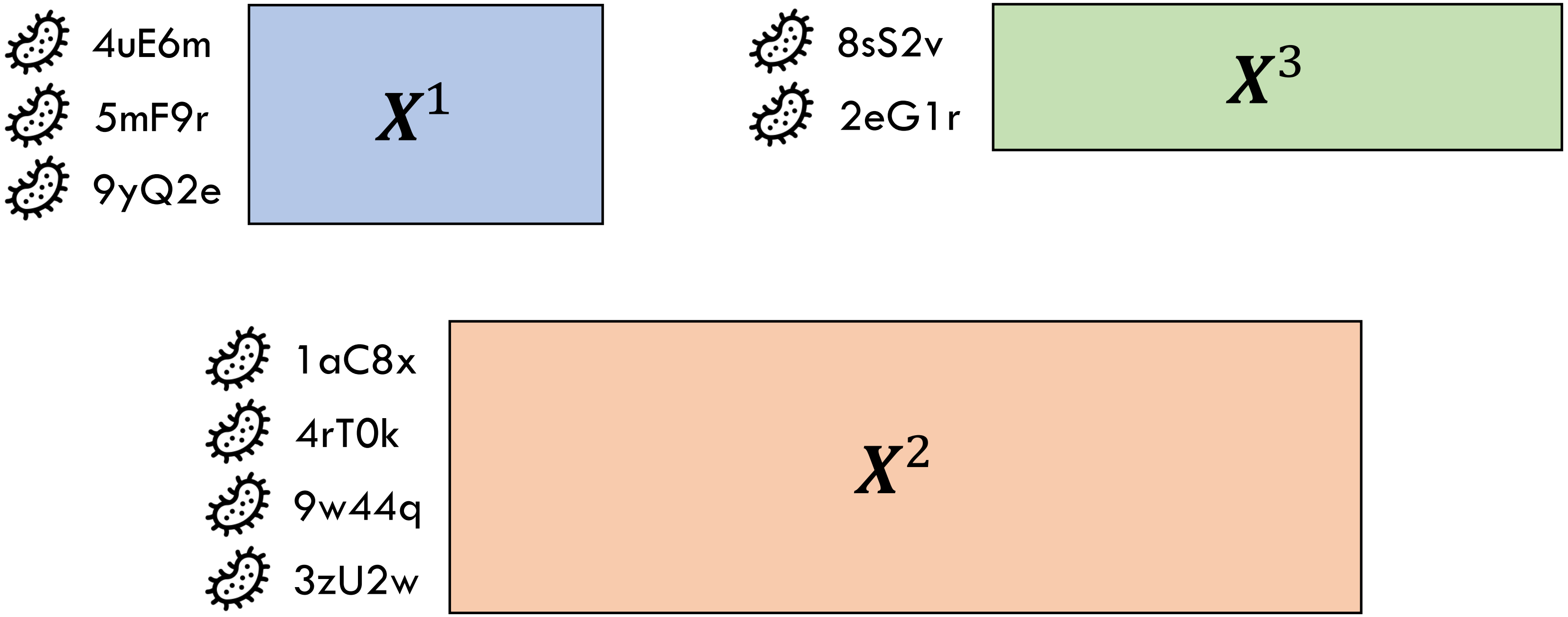
...


Shared latent space
Modality-specific latent space
g_1
g_2
g_3
X^1
X^2
X^3

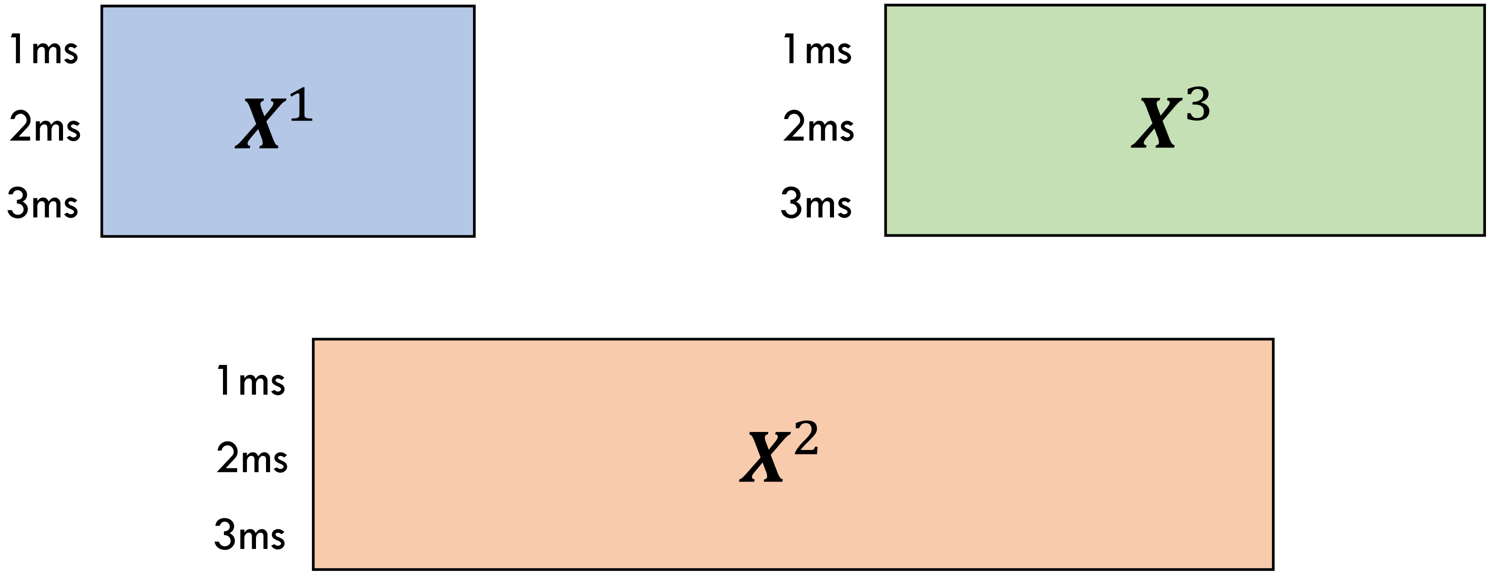

Z_1
Z_2
...
Z_4
Z_5
Z_6
Z_7
Z_8
Z_9
g_1
g_2
g_3
X^1
X^2
X^3

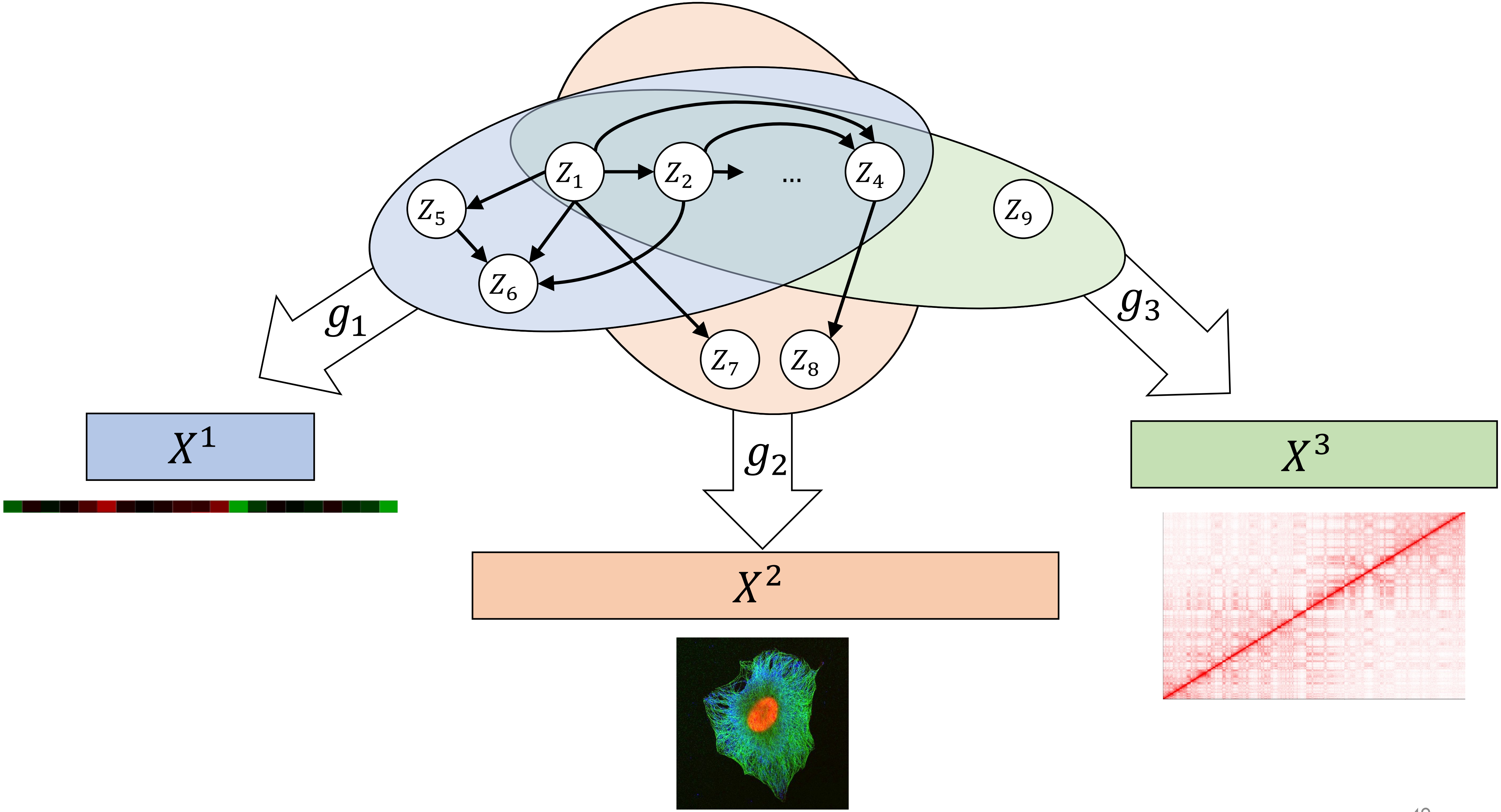
Technological limitation:

- Most experimental technologies (RNA sequencing, microscopic imaging, and chromatin conformation capture) destroy the cell in the process of measurement.
- Thus, we never observe samples from the joint distribution P_X over $(X^1, X^2, \ldots, X^m)$, but only from the marginals $P_{X^1}, P_{X^2}, \ldots, P_{X^m}$.

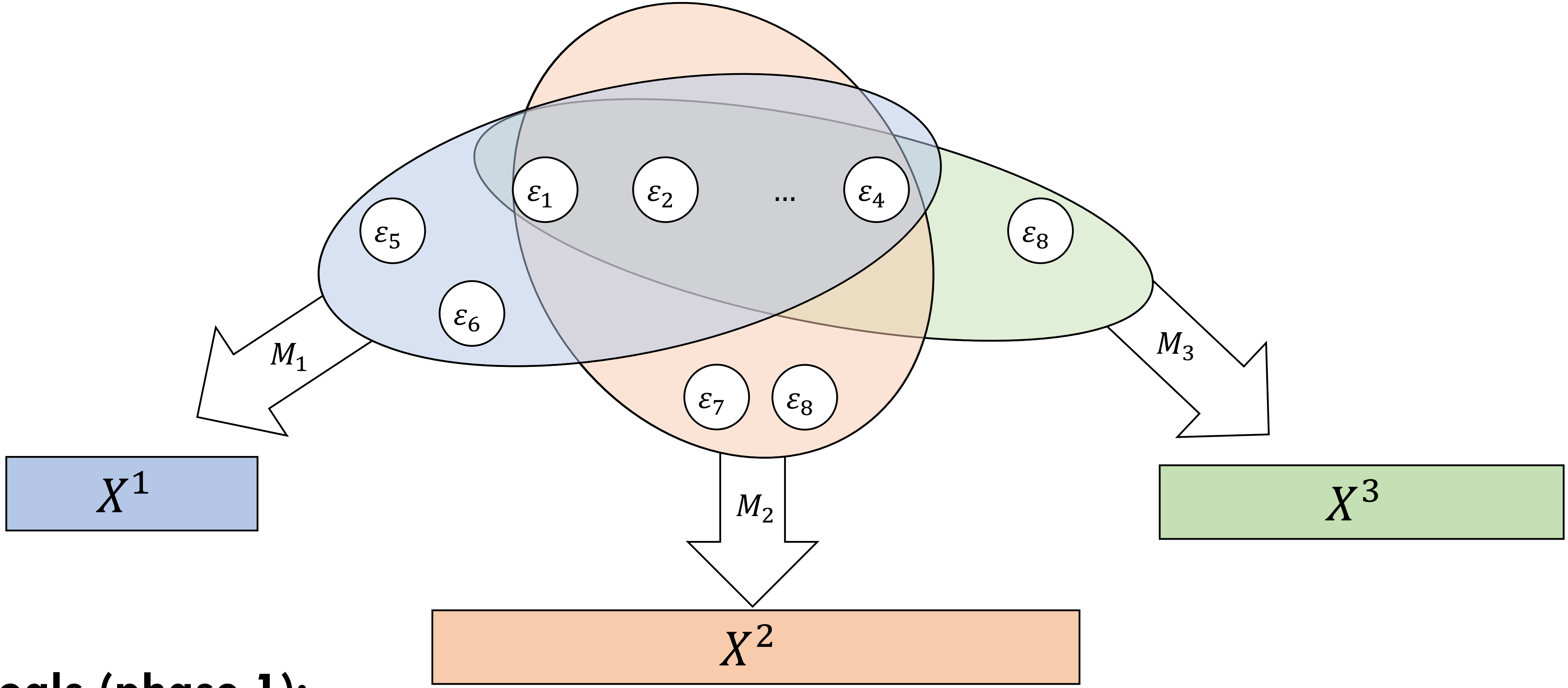


- This prevents the use of prior group ICA / multiset canonical correlation analysis methods (Calhoun 2001, Nielsen 2002, Beckman 2005, Richard 2021, …)
- These methods assume access to **paired** data, e.g., different subjects in an fMRI experiment have corresponding time points or voxels.





Z_1
Z_2
...
Z_4
Z_5
Z_6
Z_7
Z_8
Z_9
g_1
g_2
g_3
X^1
X^2
X^3

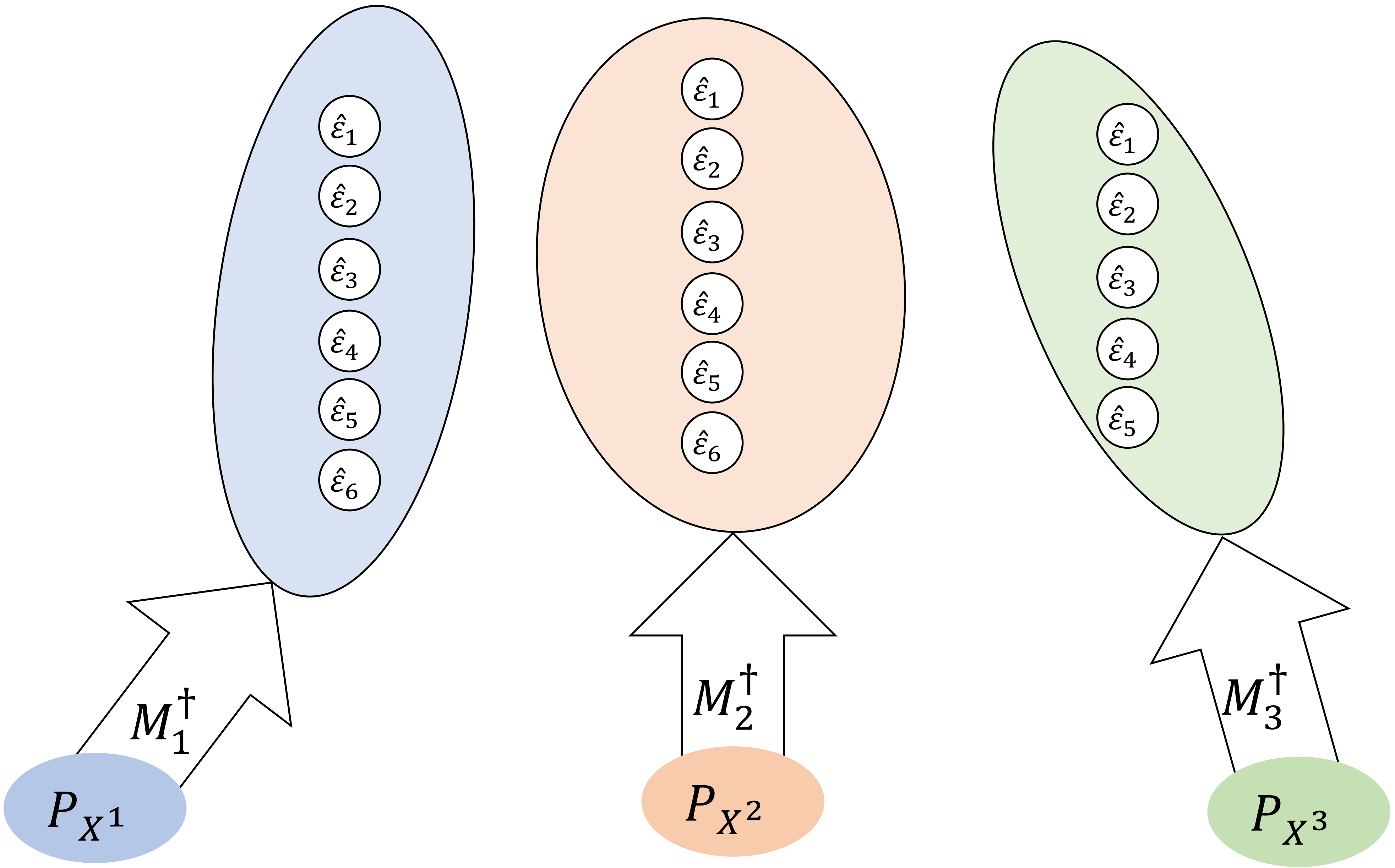


Goals (phase 1):

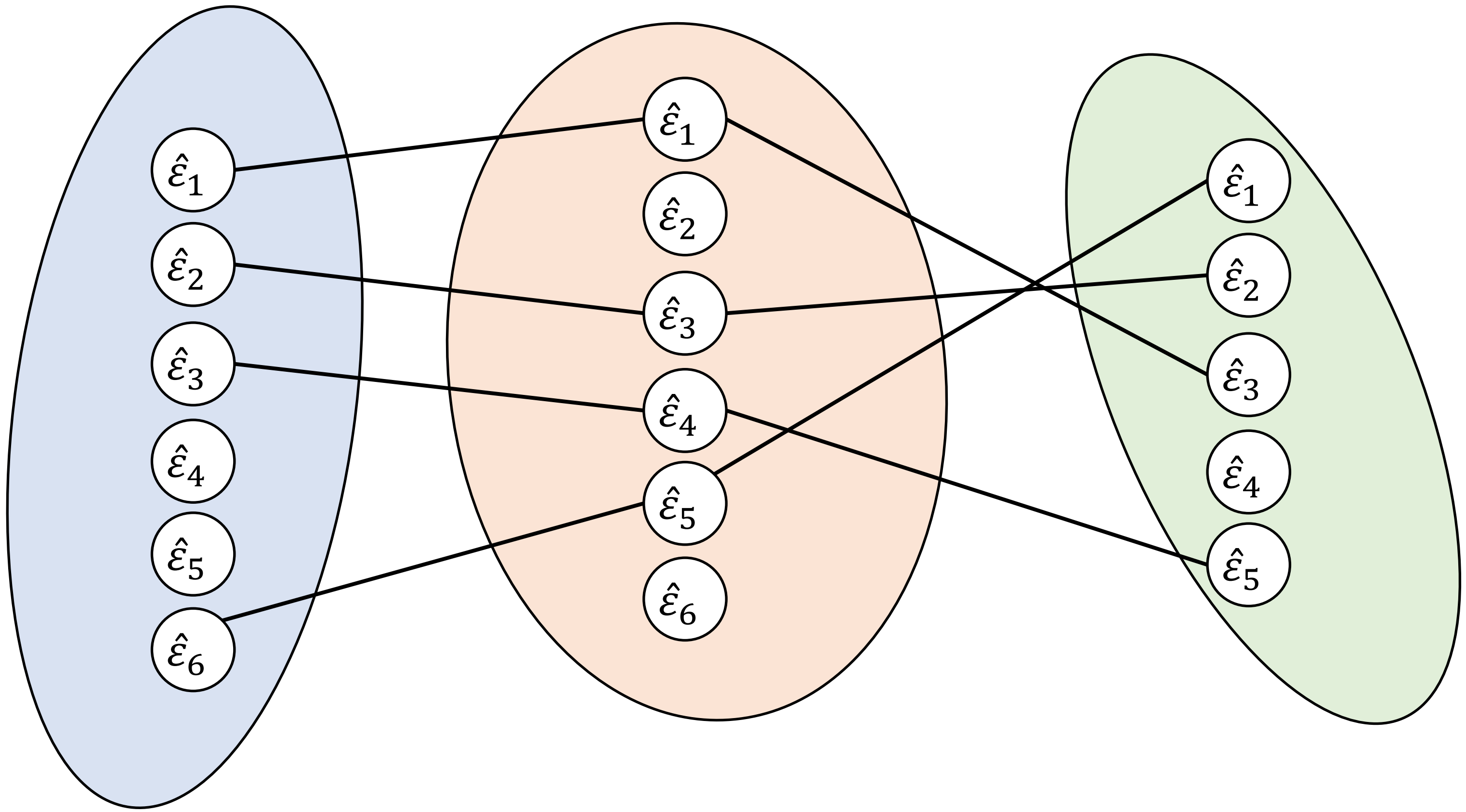
- Recover P, the distribution over the exogenous variables $\boldsymbol{\varepsilon} = \varepsilon_1, \varepsilon_2, \ldots, \varepsilon_8$.
- Recover the joint mixing matrix $M: \boldsymbol{\varepsilon} \mapsto (X_1, X_2, X_3)$.

The pushforward distribution $M\#P$ is the joint distribution P_X.

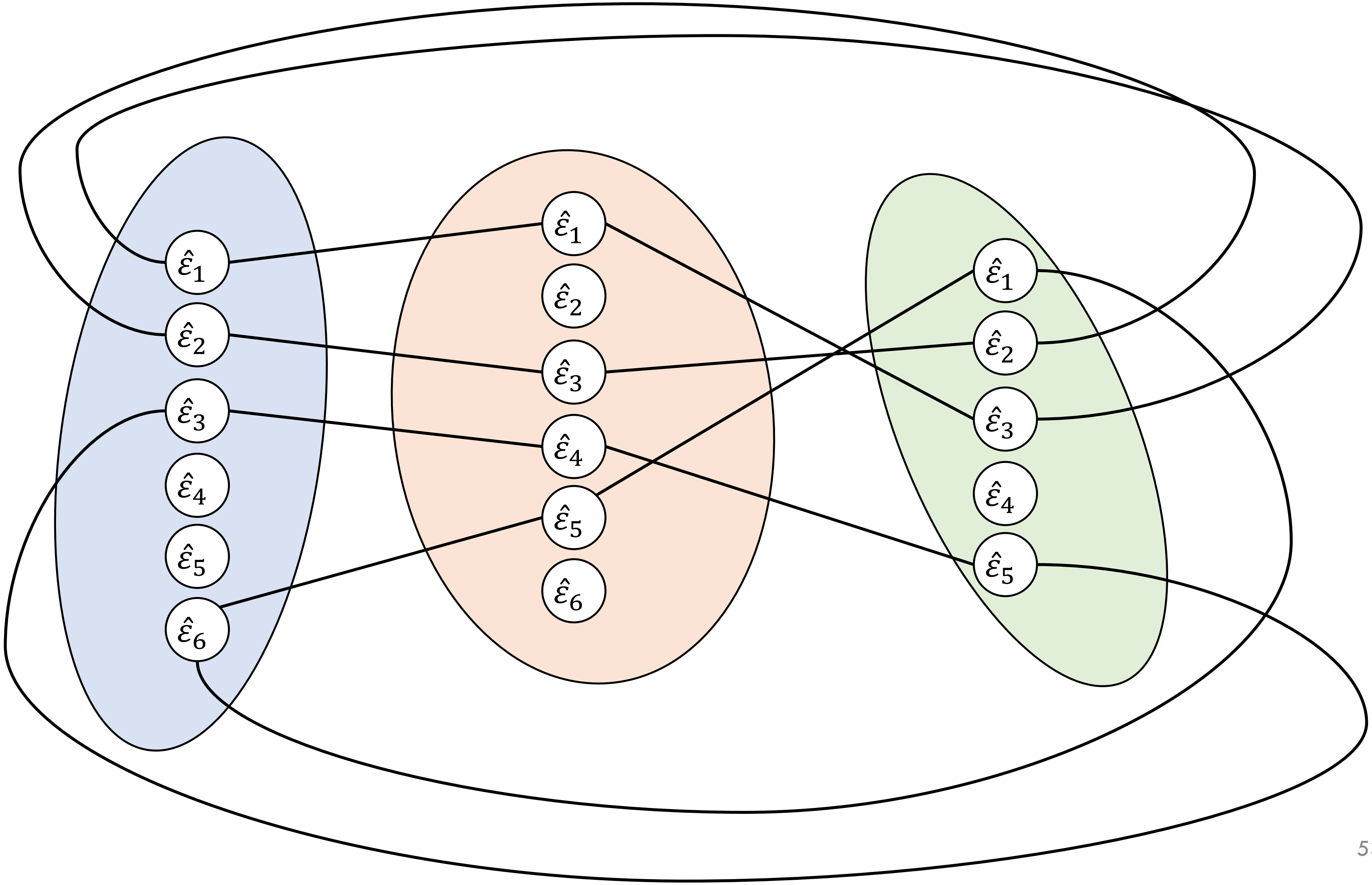
Step 1: Perform linear ICA separately in each domain.



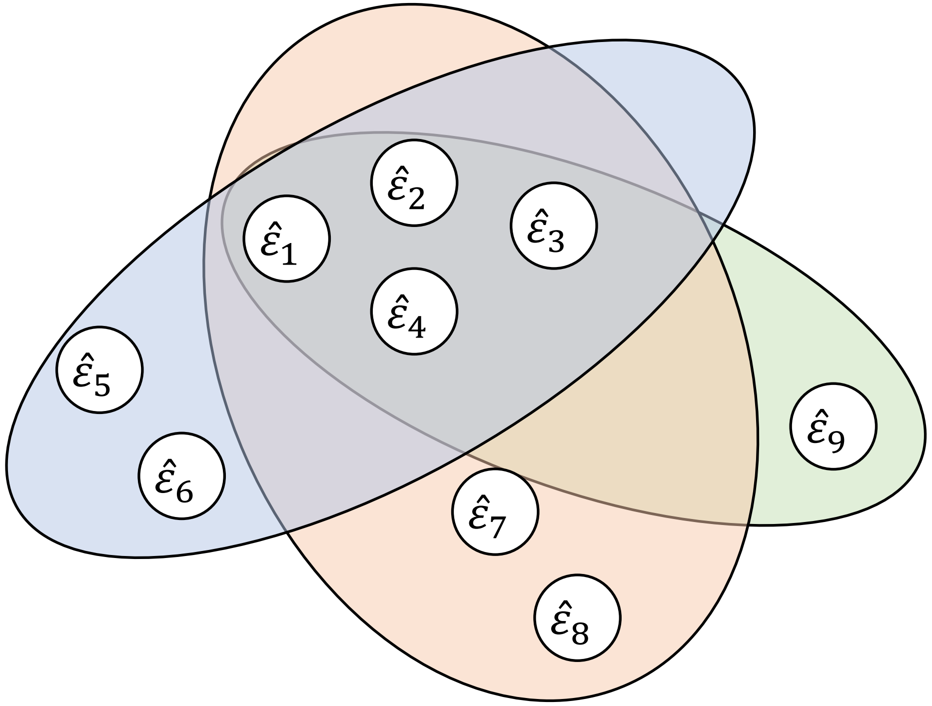
Step 2: Match latent distributions between domains based on Kolmogorov-Smirnov testing.



Step 2: Match latent distributions between domains based on Kolmogorov-Smirnov testing.



Step 3: Merge latent spaces



Assumptions:

(C1) Exogenous variables have unit variance (w.l.o.g.), are non-symmetric, and have distinct distributions up to sign (i.e., $d(P_i, P_j) > 0, d(P_i, -P_j) > 0$ for all $i \neq j$).

(C2) The latent SCM and the mixing functions are linear, i.e. $X^e = G^e Z$ for each domain $e \in [m]$. The stacked mixing matrix $G = [G^1; G^2; \ldots; G^m]$ is full column rank.

Theorem: Suppose access to $m \geq 2$ domains. Under **(C1)** and **(C2)**, P and M are recoverable.

Note: "Recovery" is only up to an indeterminacy that comes from re-labeling nodes.

Additional results:

1. **Matching gets better with more modalities.** Each added modality (assuming enough samples) gives another estimate of the distributions of the shared latent variables. Enforcing transitivity between matches gives better power for a fixed false discovery rate.
2. **Latent graph recovery.** After recovering, we can use standard techniques involving restrictions on g to recover the latent graph.

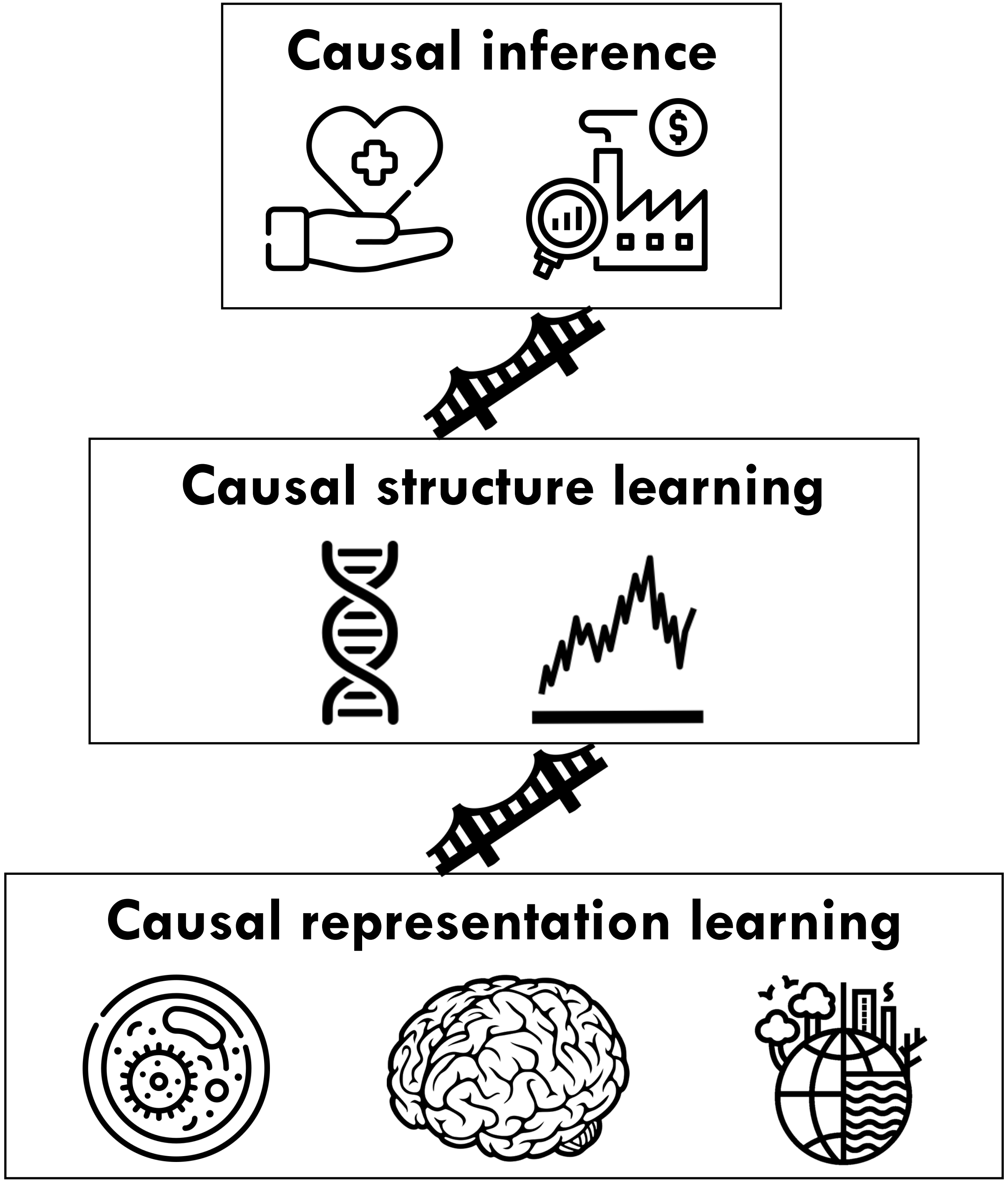
Future Work

- Nonlinear setting
  - Would provide identifiability theory for several existing approaches (e.g., Yang '21)
- Combining interventions and multiple modalities

Type 1 domains:
causally familiar

Type 2 domains:
conceptually familiar

Type 3 domains:
conceptually novel



Selected References

Agrawal, R., Squires, C., Yang, K., Shanmugam, K., & Uhler, C. (2019). Abcd-strategy: Budgeted experimental design for targeted causal structure discovery.

Ahuja, K., Wang, Y., Mahajan, D., & Bengio, Y. (2022) Interventional Causal Representation Learning.

Ahuja, K., Hartford, J. S., & Bengio, Y. (2022). Weakly supervised representation learning with sparse perturbations.

Arjovsky, M., Bottou, L., Gulrajani, I., & Lopez-Paz, D. (2019). Invariant risk minimization.

Baraud, Y., & Birgé, L. (2014). Estimating composite functions by model selection.

Beckmann, C. F., & Smith, S. M. (2005). Tensorial extensions of independent component analysis for multisubject FMRI analysis.

Boix-Adsera, E., Bresler, G., & Koehler, F. (2022, February). Chow-Liu++: Optimal prediction-centric learning of tree Ising models.

Brehmer, J., De Haan, P., Lippe, P., & Cohen, T. S. (2022). Weakly supervised causal representation learning.

Cai, R., Xie, F., Glymour, C., Hao, Z., & Zhang, K. (2019). Triad constraints for learning causal structure of latent variables.

Calhoun, V. D., Adali, T., Pearlson, G. D., & Pekar, J. J. (2001). A method for making group inferences from functional MRI data using independent component analysis.

Chalupka, K., Eberhardt, F., & Perona, P. (2017). Causal feature learning: an overview.

Comon, P. (1994). Independent component analysis, a new concept?

Eaton, D., & Murphy, K. (2007, March). Exact Bayesian structure learning from uncertain interventions.

Gradu, P., Zrnic, T., Wang, Y., & Jordan, M. I. (2022). Valid Inference after Causal Discovery.

Heinze-Deml, C., Peters, J., & Meinshausen, N. (2018). Invariant causal prediction for nonlinear models.

Hyvarinen, A., Sasaki, H., & Turner, R. (2019). Nonlinear ICA using auxiliary variables and generalized contrastive learning.

Kivva, B., Rajendran, G., Ravikumar, P., & Aragam, B. (2021). Learning latent causal graphs via mixture oracles.

Mooij, J. M., Magliacane, S., & Claassen, T. (2020). Joint causal inference from multiple contexts.

Richard, H., Ablin, P., Thirion, B., Gramfort, A., & Hyvarinen, A. (2021). Shared Independent Component Analysis for Multi-Subject Neuroimaging.

Rosenfeld, E., Ravikumar, P., & Risteski, A. (2020). The risks of invariant risk minimization.

Peters, J., Bühlmann, P., & Meinshausen, N. (2016). Causal inference by using invariant prediction: identification and confidence intervals.

Silva, R., Scheines, R., Glymour, C., Spirtes, P., & Chickering, D. M. (2006). Learning the Structure of Linear Latent Variable Models.

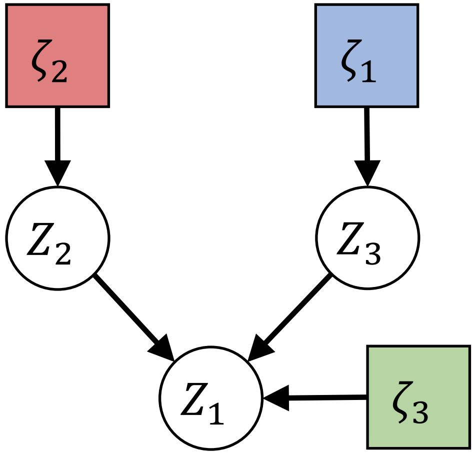
Wang, Y., Squires, C., Belyaeva, A., & Uhler, C. (2018). Direct estimation of differences in causal graphs.

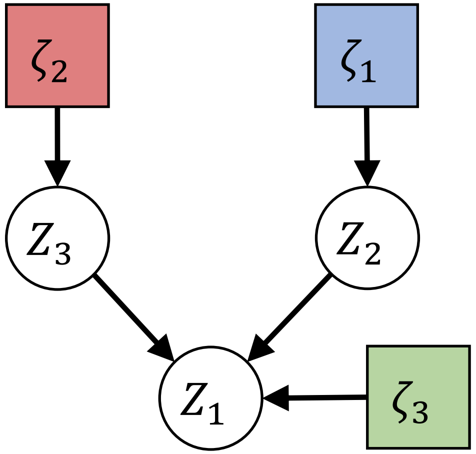
Zheng, X., Aragam, B., Ravikumar, P. K., & Xing, E. P. (2018). Dags with no tears: Continuous optimization for structure learning.

Extra Slides

Recovery up to re-labeling

$\sigma_1 = (1, 2, 3)$





$\sigma_2 = (1, 3, 2)$

$S(\mathcal{G})$: permutations consistent with $\mathcal{G}$
($\sigma(j) > \sigma(i)$ for all edges $j \to i$)

Permutation matrix: $(P_\sigma)_{ij} = \mathbb{1}_{i=\sigma(j)}$

$B_k^\sigma = P_\sigma B_k P_\sigma^\top \qquad \mathrm{H}^\sigma = P_\sigma H$

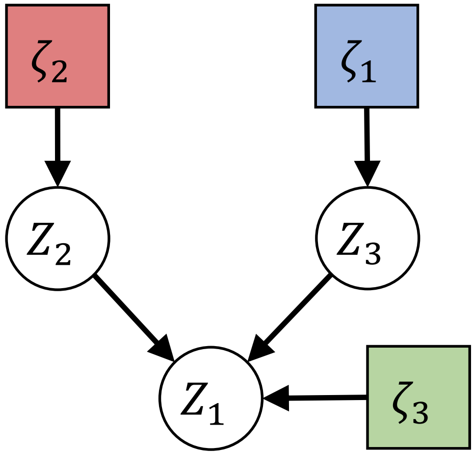
$$B_0 = \begin{bmatrix} (B_0)_{11} & (B_0)_{12} & (B_0)_{13} \\ 0 & (B_0)_{22} & 0 \\ 0 & 0 & (B_0)_{33} \end{bmatrix}$$

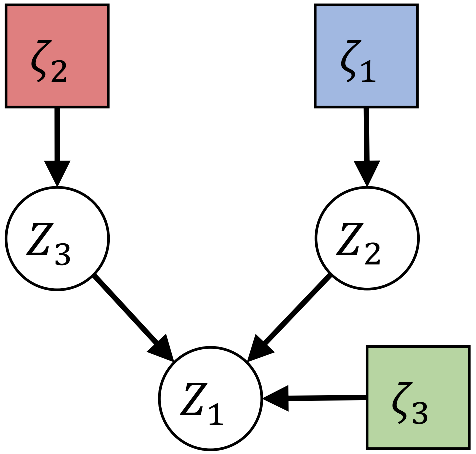
$$B_1 = \begin{bmatrix} (B_0)_{11} & (B_0)_{12} & (B_0)_{13} \\ 0 & (B_0)_{22} & 0 \\ 0 & 0 & (B_1)_{33} \end{bmatrix}$$

$$B_2 = \begin{bmatrix} (B_0)_{11} & (B_0)_{12} & (B_0)_{13} \\ 0 & (B_2)_{22} & 0 \\ 0 & 0 & (B_0)_{33} \end{bmatrix}$$

$$B_3 = \begin{bmatrix} (B_3)_{11} & (B_3)_{12} & (B_3)_{13} \\ 0 & (B_0)_{22} & 0 \\ 0 & 0 & (B_0)_{33} \end{bmatrix}$$

$$H = \begin{bmatrix} H_{11} & H_{12} & \dots & H_{1p} \\ H_{21} & H_{22} & \dots & H_{2p} \\ H_{31} & H_{32} & \dots & H_{3p} \end{bmatrix}$$





$$B_0^{\sigma_2} = \begin{bmatrix} (B_0)_{11} & (B_0)_{13} & (B_0)_{12} \\ 0 & (B_0)_{33} & 0 \\ 0 & 0 & (B_0)_{22} \end{bmatrix}$$

$$B_1^{\sigma_2} = \begin{bmatrix} (B_0)_{11} & (B_0)_{13} & (B_0)_{12} \\ 0 & (B_1)_{33} & 0 \\ 0 & 0 & (B_0)_{22} \end{bmatrix}$$

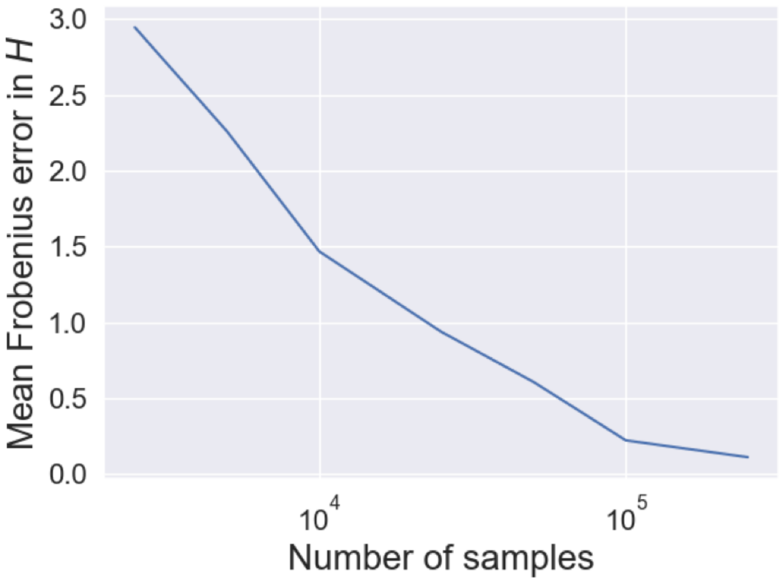
$$B_2^{\sigma_2} = \begin{bmatrix} (B_0)_{11} & (B_0)_{13} & (B_0)_{12} \\ 0 & (B_0)_{33} & 0 \\ 0 & 0 & (B_2)_{22} \end{bmatrix}$$

$$B_3^{\sigma_2} = \begin{bmatrix} (B_3)_{11} & (B_3)_{13} & (B_3)_{12} \\ 0 & (B_0)_{33} & 0 \\ 0 & 0 & (B_0)_{22} \end{bmatrix}$$

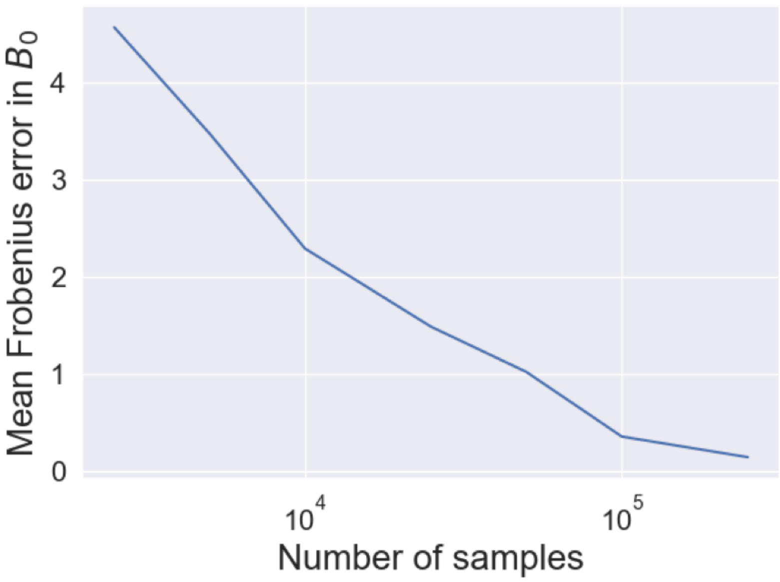
$$H^{\sigma_2} = \begin{bmatrix} H_{11} & H_{12} & \dots & H_{1p} \\ H_{31} & H_{32} & \dots & H_{3p} \\ H_{21} & H_{22} & \dots & H_{2p} \end{bmatrix}$$

Synthetic data results

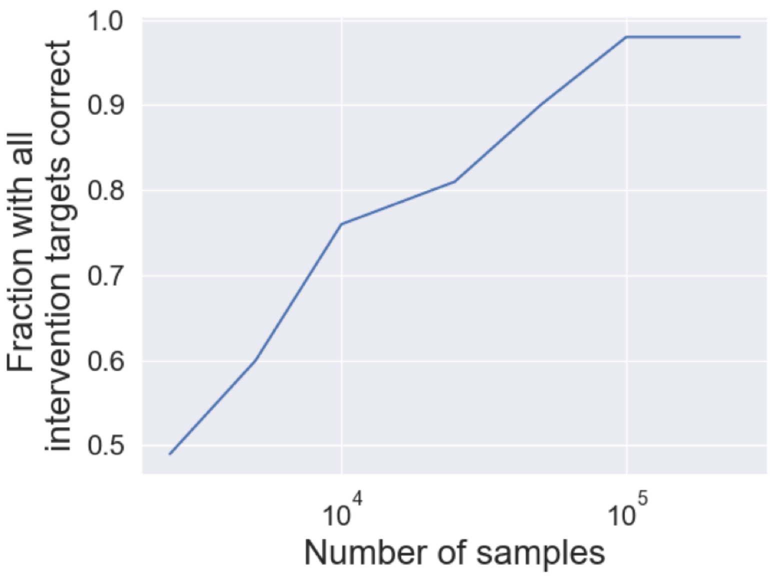
$d = 5$ latent variables
$p = 10$ observed variables
500 random models, Erdős-Rényi structure with density 0.75



(a) Error in estimating H



(b) Error in estimating B_0



(c) Intervention targets