Identifiability of Causal Models

and Applications to Perturb-seq data

Chandler Squires

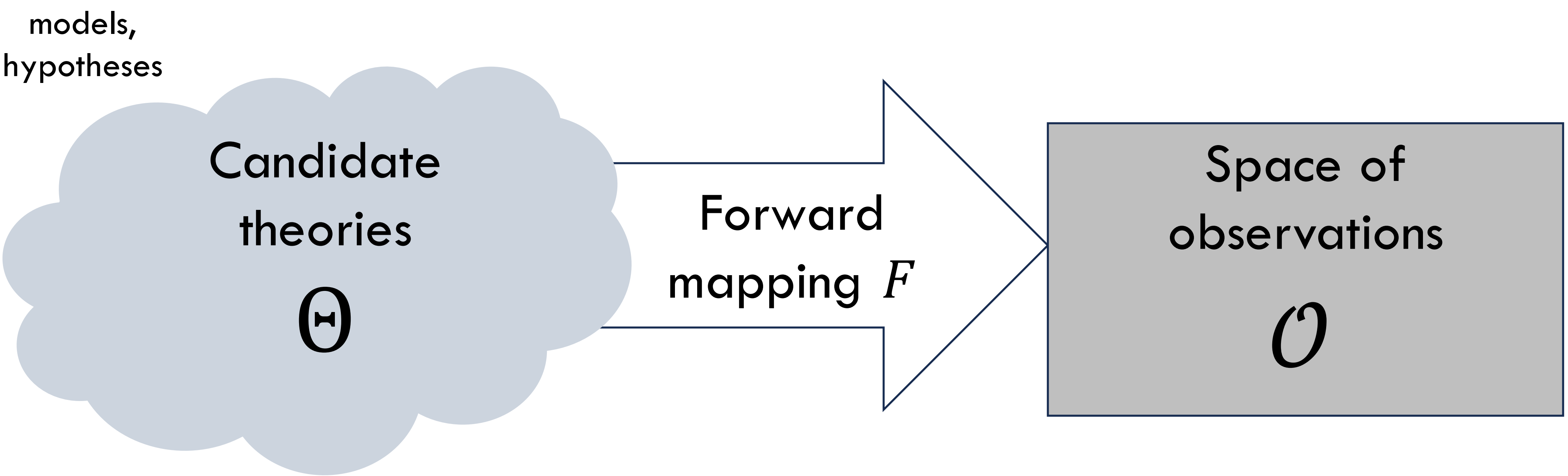
08/03/2023

Outline

- *Background:* Identifiability
- *Background:* Structural causal models
- Causal structure learning
    - Learning from unknown-target interventions
    - Learning in the presence of unobserved variables
    - Experimental design for causal structure learning
- Causal disentanglement

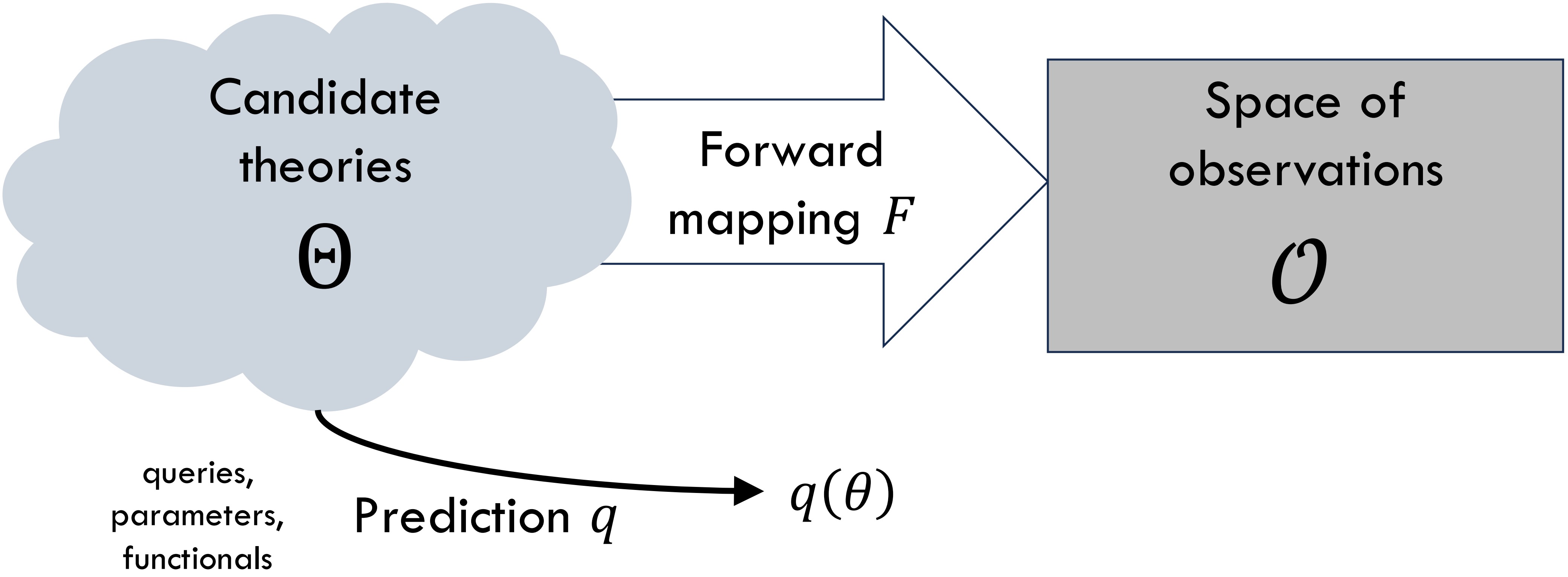
Outline

- *Background:* Identifiability
- *Background:* Structural causal models
- Causal structure learning
  - Learning from unknown-target interventions
  - Learning in the presence of unobserved variables
  - Experimental design for causal structure learning
- Causal disentanglement



Model identifiability: Does a unique theory fits the observed data?

What can be estimated? Identifiability, estimability, causal inference and ill-posed inverse problems. Maclaren and Nicholson (2020).
Asymptotic methods in statistical theory. Le Cam (1986).



Parameter identifiability: Do all theories which fit the observed data make the same predictions?

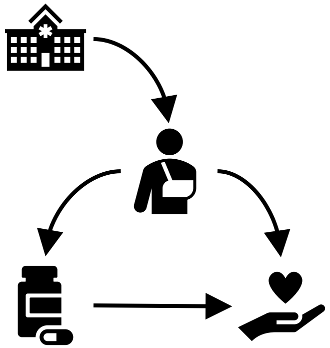
What can be estimated? Identifiability, estimability, causal inference and ill-posed inverse problems. Maclaren and Nicholson (2020).
Asymptotic methods in statistical theory. Le Cam (1986).

Why should we care about identifiability?

Identifiability provides a rigorous license for **extrapolation** (aka **generalization**).

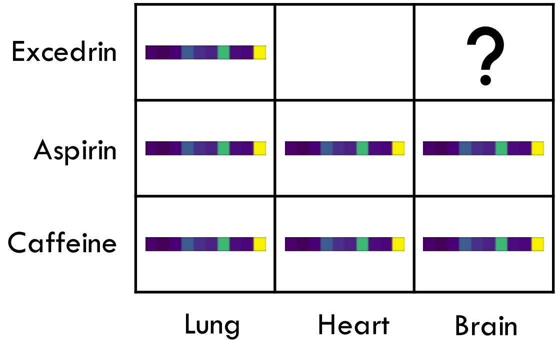
Model robustness

Generalizing to distribution shifts requires identifying non-spurious features from spurious ones [1].



Healthcare

Extrapolating to novel patient populations requires identifying conditional treatment effects [2].

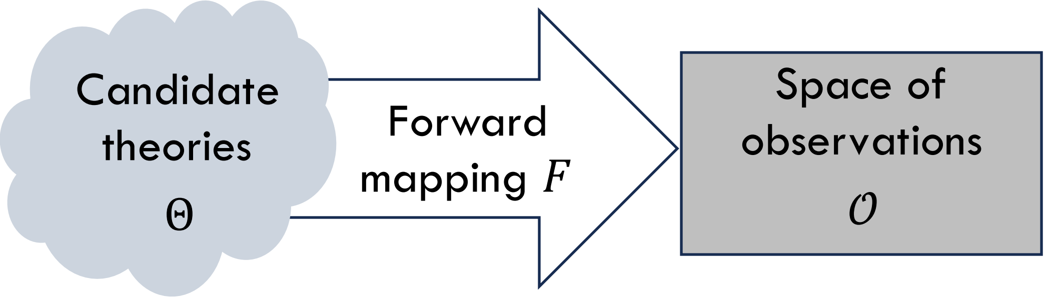


Biology

Predicting the effect of a drug on a new cell type requires identifying how drugs and cells interact [3].

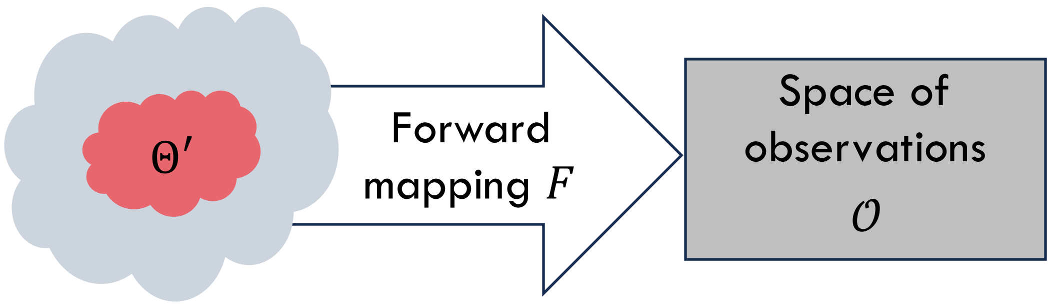
[1] ***Invariant risk minimization.*** Arjovsky, Bottou, Gulrajani, Lopez-Paz (2020).
[2] ***Counterfactual off-policy evaluation with Gumbel-max structural causal models.*** Oberst and Sontag (2019).
[3] ***Causal Imputation via Synthetic Interventions.*** Squires*, Shen*, Agarwal, Shah, Uhler (CLeaR 2022).

What if a model (or parameter) is not identifiable?



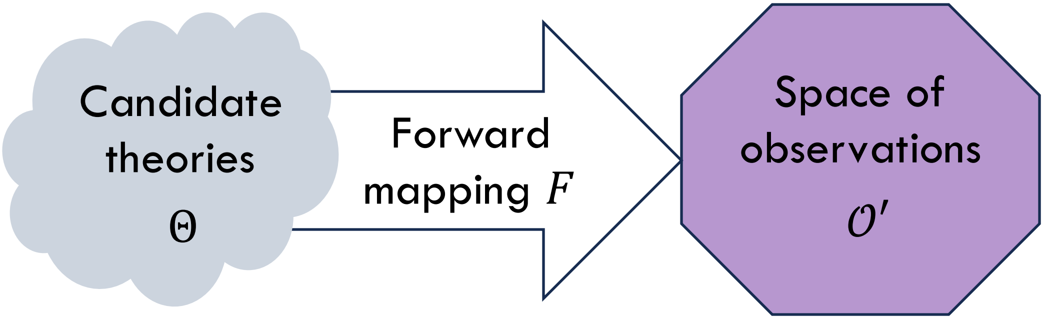
Restriction approach:

consider a smaller space of theories.



Expansion approach:

consider a richer form of observations.



Structural Causal Models

Let $\mathcal{G}$ be a directed acyclic graph (DAG).
The following equations define a ***structural causal model (SCM)*** $\mathcal{M}$ with causal DAG $\mathcal{G}$.

$$
\begin{aligned}
Z_1 &= f_1\big(pa_{\mathcal{G}}(Z_1), \varepsilon_1\big) & \qquad \varepsilon_1 &\sim \mathbb{P}_{\varepsilon_1} \\
Z_2 &= f_2\big(pa_{\mathcal{G}}(Z_2), \varepsilon_2\big) & \qquad \varepsilon_2 &\sim \mathbb{P}_{\varepsilon_2} \\
&\vdots & &\vdots \\
Z_d &= f_d\big(pa_{\mathcal{G}}(Z_d), \varepsilon_d\big) & \qquad \varepsilon_d &\sim \mathbb{P}_{\varepsilon_d}
\end{aligned}
$$

Structural Causal Models

Let $\mathcal{G}$ be a directed acyclic graph (DAG). The following equations define a ***structural causal model (SCM)*** $\mathcal{M}$ with causal DAG $\mathcal{G}$.

$$Z_1 = f_1\left(pa_{\mathcal{G}}(Z_1), \varepsilon_1\right) \qquad \varepsilon_1 \sim \mathbb{P}_{\varepsilon_1}$$

$$Z_2 = f_2\left(pa_{\mathcal{G}}(Z_2), \varepsilon_2\right) \qquad \varepsilon_2 \sim \mathbb{P}_{\varepsilon_2}$$

$$\vdots \qquad \qquad \vdots$$

$$Z_d = f_d\left(pa_{\mathcal{G}}(Z_d), \varepsilon_d\right) \qquad \varepsilon_d \sim \mathbb{P}_{\varepsilon_d}$$

Causal mechanisms
(aka "structural equations")

Denote the set of SCMs with causal DAG $\mathcal{G}$ as $\mathfrak{M}(\mathcal{G})$.

Structural Causal Models

Let $\mathcal{G}$ be a directed acyclic graph (DAG). The following equations define a ***structural causal model (SCM)*** $\mathcal{M}$ with causal DAG $\mathcal{G}$.

$$Z_1 = f_1\big(pa_{\mathcal{G}}(Z_1), \varepsilon_1\big) \qquad \varepsilon_1 \sim \mathbb{P}_{\varepsilon_1}$$

$$Z_2 = f_2\big(pa_{\mathcal{G}}(Z_2), \varepsilon_2\big) \qquad \varepsilon_2 \sim \mathbb{P}_{\varepsilon_2}$$

$$\vdots \qquad\qquad\qquad \vdots$$

$$Z_d = f_d\big(pa_{\mathcal{G}}(Z_d), \varepsilon_d\big) \qquad \varepsilon_d \sim \mathbb{P}_{\varepsilon_d}$$

Causal mechanisms
(aka "structural equations")

Exogenous noise variables
(assumed mutually independent)

Structural Causal Models

Let $\mathcal{G}$ be a directed acyclic graph (DAG). The following equations define a ***structural causal model (SCM)*** $\mathcal{M}$ with causal DAG $\mathcal{G}$.

$$
\begin{aligned}
Z_1 &= f_1\big(pa_{\mathcal{G}}(Z_1), \varepsilon_1\big) & \qquad \varepsilon_1 &\sim \mathbb{P}_{\varepsilon_1} \\
Z_2 &= f_2\big(pa_{\mathcal{G}}(Z_2), \varepsilon_2\big) & \qquad \varepsilon_2 &\sim \mathbb{P}_{\varepsilon_2} \\
&\vdots & &\vdots \\
Z_d &= f_d\big(pa_{\mathcal{G}}(Z_d), \varepsilon_d\big) & \qquad \varepsilon_d &\sim \mathbb{P}_{\varepsilon_d}
\end{aligned}
$$

Causal mechanisms
(aka "structural equations")

Exogenous noise variables
(assumed mutually independent)

Denote the set of SCMs with causal DAG $\mathcal{G}$ as $\mathfrak{M}(\mathcal{G})$.

Interventions

An intervention I consists of:

- a set $T(I)$ of ***intervention targets*** and
- an indexed set $\{f_i^I\}_{i \in T(I)}$ of ***interventional mechanisms***[1].

Control

$$
\begin{aligned}
Z_1 &= f_1(\varepsilon_1) \\
Z_2 &= f_2(Z_1, \varepsilon_2) \\
&\vdots \\
Z_d &= f_d(Z_1, Z_2, \dots, \varepsilon_d)
\end{aligned}
$$

$$
\begin{aligned}
Z_1 &= f_1(\varepsilon_1) \\
Z_2 &= f_2^I(Z_1, \varepsilon_2) \\
&\vdots \\
Z_d &= f_d(Z_1, Z_2, \dots, \varepsilon_d)
\end{aligned}
$$

An SCM $\mathcal{M}$ and an intervention I define a new **interventional SCM** $\mathcal{M}_I$.

[1]Changes to $\mathbb{P}_{\varepsilon_i}$ can generally be absorbed into a change in f_i^I.

Control

$$Z_1 = f_1(\varepsilon_1)$$
$$Z_2 = f_2(Z_1, \varepsilon_2)$$
$$\vdots$$
$$Z_d = f_d(Z_1, Z_2, \ldots, \varepsilon_d)$$

$$Z_1 = f_1(\varepsilon_1)$$
$$Z_2 = f_2^I(Z_1, \varepsilon_2)$$
$$\vdots$$
$$Z_d = f_d(Z_1, Z_2, \ldots, \varepsilon_d)$$

"do" intervention

aka *"point"*

$$Z_2 = \hat{z}_2$$

Sets the target to a ***constant value.***

Perfect intervention

aka *"hard"*

$$Z_2 = f_2^I(\varepsilon_2)$$

Removes dependence of the target on its parents.

Soft intervention

aka *"imperfect"* or *"mechanism shift"*

$$Z_2 = f_2^I(Z_1, \varepsilon_2)$$

Arbitrarily changes the mechanism of the target.

Least general ⟷ Most general

Interventions

Given a tuple of intervention targets $\mathcal{I} = (\emptyset, T_1, T_2, \dots, T_K)$, we can define $\mathfrak{M}_{\mathcal{I}}(\mathcal{G})$ as all tuples $(\mathcal{M}_0, \mathcal{M}_1, \mathcal{M}_2, \dots, \mathcal{M}_K)$ of SCMs where $\mathcal{M}_k \in \mathfrak{M}(\mathcal{G})$ for all k and $\mathcal{M}_k$ differs from $\mathcal{M}_0$ only in the mechanisms $\{f_i\}_{i \in T_k}$.

For $\mathcal{I} = (\emptyset, \{2\}, \{1, d\}, \dots, \{1\})$, a typical element looks like:

$$\begin{aligned} Z_1 &= f_1(\varepsilon_1) \\ Z_2 &= f_2(Z_1, \varepsilon_2) \\ &\vdots \\ Z_d &= f_d(Z_1, Z_2, \dots, \varepsilon_d) \end{aligned}$$

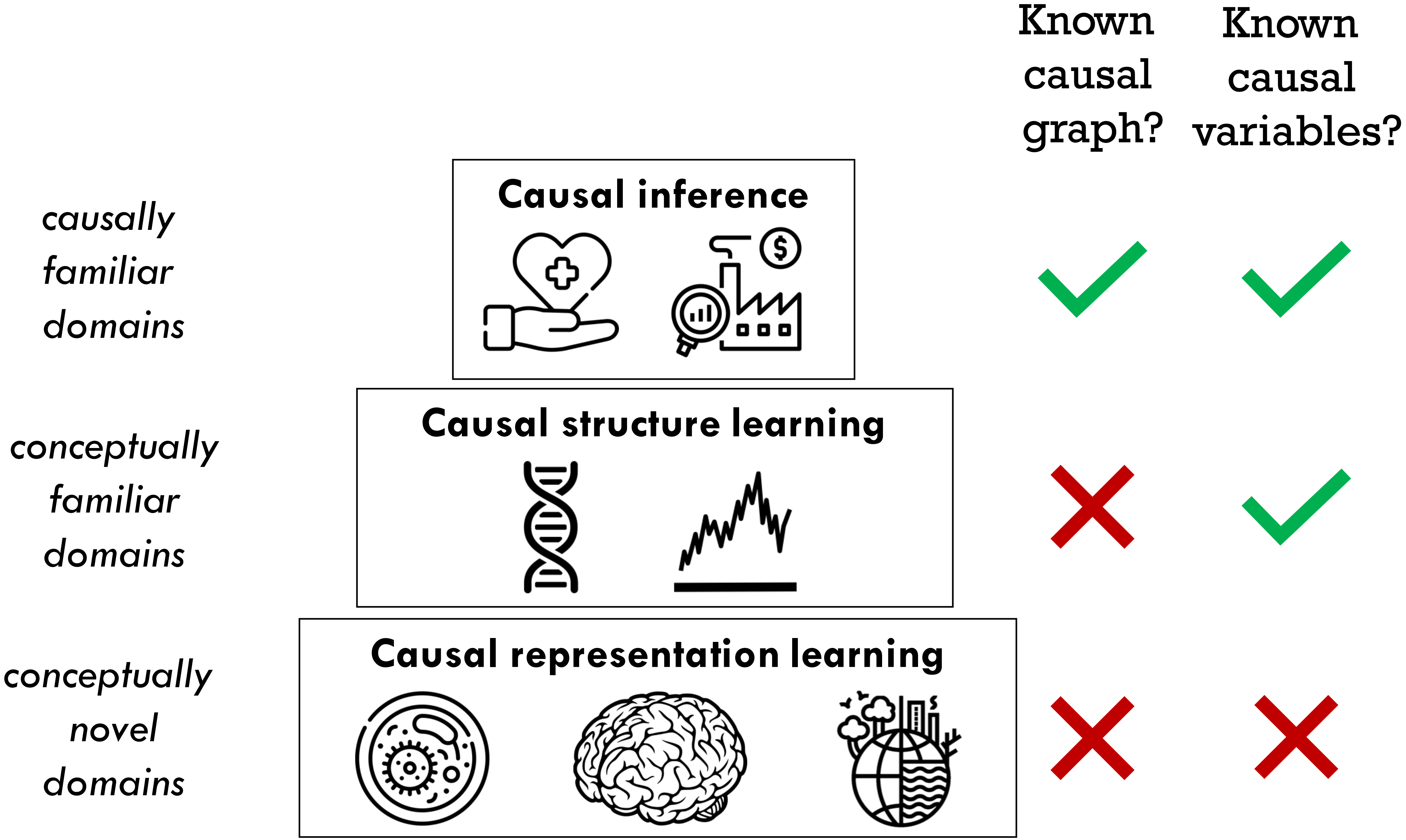
$$\begin{aligned} Z_1 &= f_1(\varepsilon_1) \\ Z_2 &= f_2^{I_1}(Z_1, \varepsilon_2) \\ &\vdots \\ Z_d &= f_d(Z_1, Z_2, \dots, \varepsilon_d) \end{aligned}$$

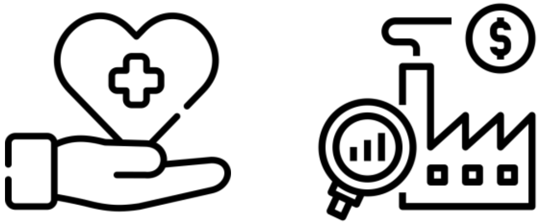
$$\begin{aligned} Z_1 &= f_1^{I_2}(\varepsilon_1) \\ Z_2 &= f_2(Z_1, \varepsilon_2) \\ &\vdots \\ Z_d &= f_d^{I_2}(Z_1, Z_2, \dots, \varepsilon_d) \end{aligned}$$

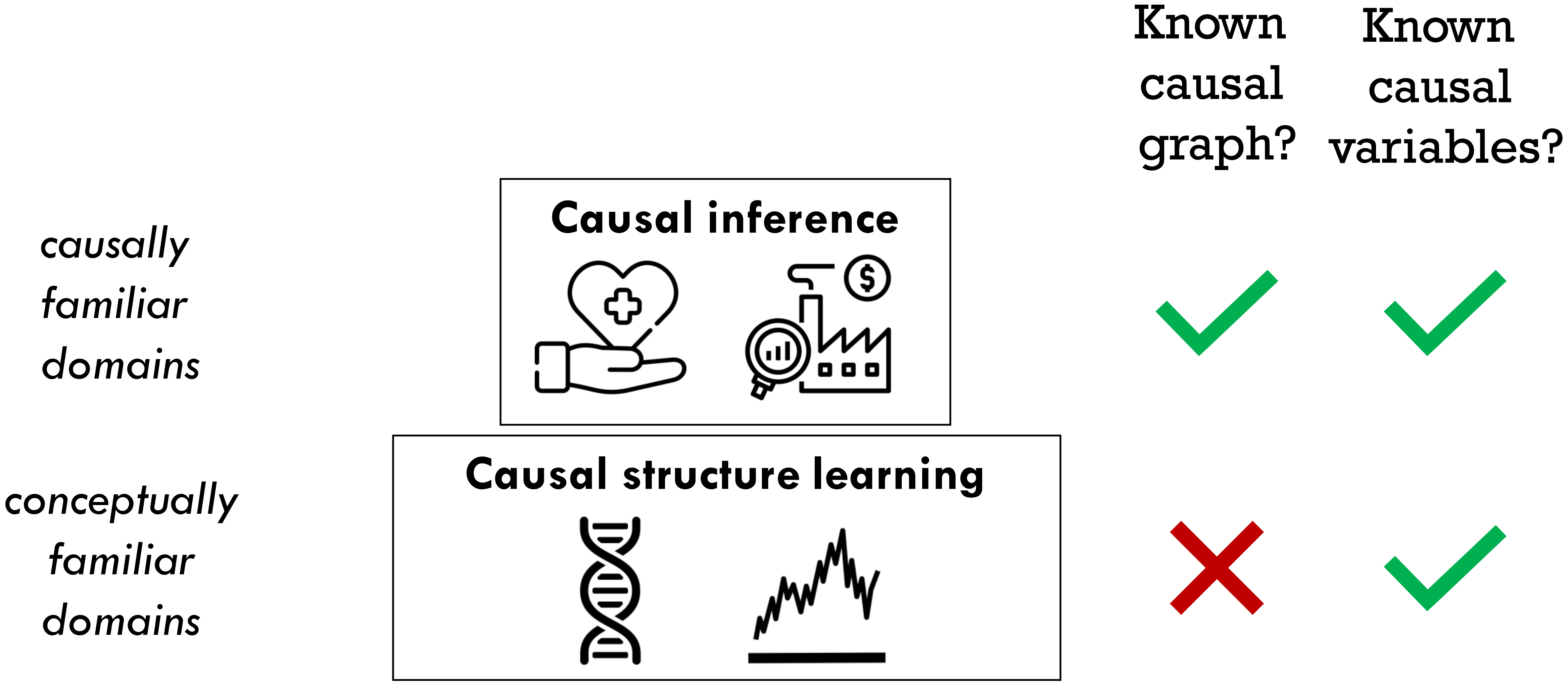
$\dots$

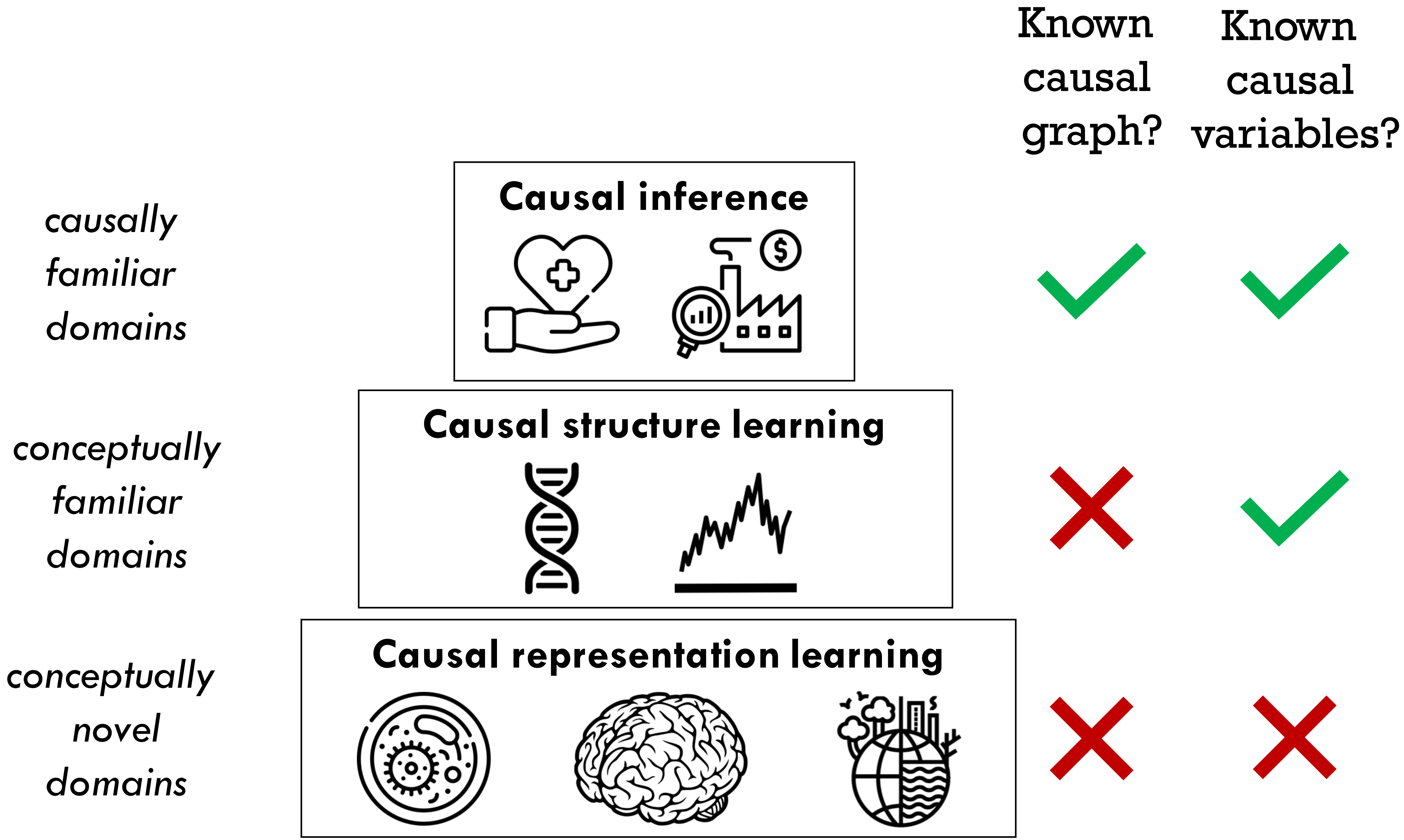
$$\begin{aligned} Z_1 &= f_1^{I_K}(\varepsilon_1) \\ Z_2 &= f_2(Z_1, \varepsilon_2) \\ &\vdots \\ Z_d &= f_d^{I_K}(Z_1, Z_2, \dots, \varepsilon_d) \end{aligned}$$

Identifiability in causality


Known causal graph?
Known causal variables?
causally familiar domains
Causal inference
conceptually familiar domains
Causal structure learning
conceptually novel domains
Causal representation learning

		Known causal graph?	Known causal variables?
causally familiar domains	**Causal inference** 	✓	✓

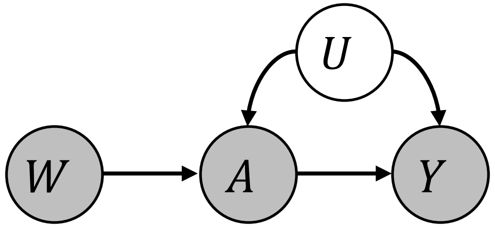

Known causal graph?
Known causal variables?
causally familiar domains
Causal inference
conceptually familiar domains
Causal structure learning



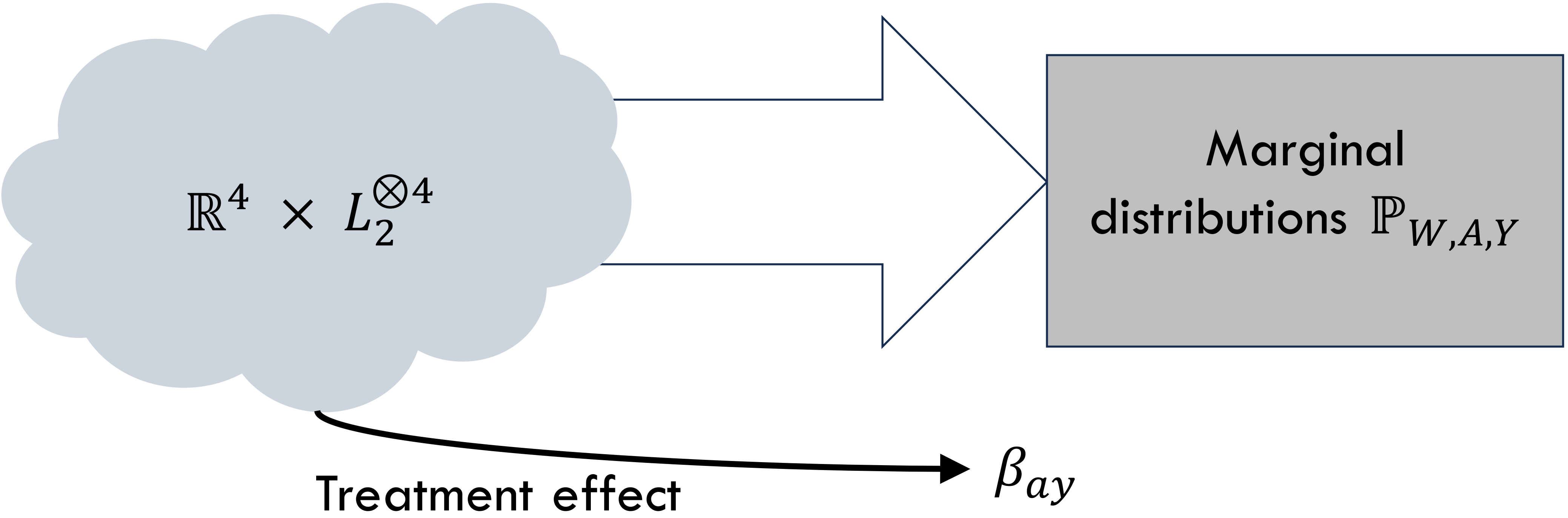
Known causal graph?
Known causal variables?
causally familiar domains
Causal inference
conceptually familiar domains
Causal structure learning
conceptually novel domains
Causal representation learning

Instrumental variable regression

causally familiar



$$
\begin{aligned}
U &= \varepsilon_u & \varepsilon_u &\sim \mathbb{P}_{\varepsilon_u}\\
W &= \varepsilon_w & \varepsilon_w &\sim \mathbb{P}_{\varepsilon_w}\\
A &= \beta_{ua} U + \beta_{wa} W + \varepsilon_a & \varepsilon_a &\sim \mathbb{P}_{\varepsilon_a}\\
Y &= \beta_{uy} U + \beta_{ay} A + \varepsilon_y & \varepsilon_y &\sim \mathbb{P}_{\varepsilon_y}
\end{aligned}
$$



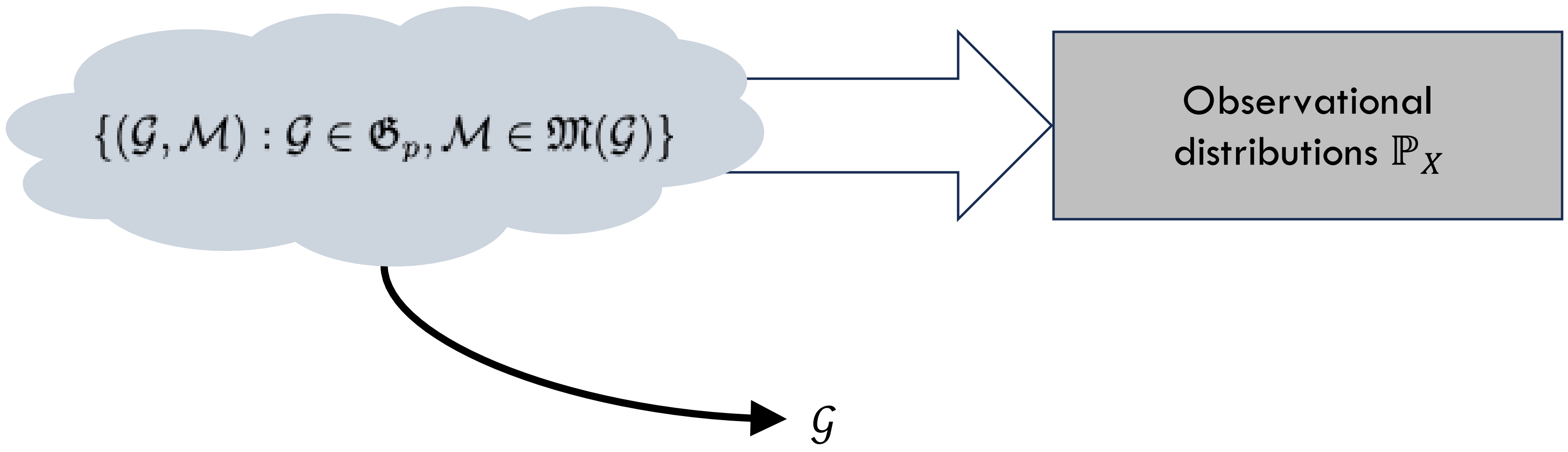
Outline

- *Background:* Identifiability
- *Background:* Structural causal models
- Causal structure learning
  - Learning from unknown-target interventions
  - Learning in the presence of unobserved variables
  - Experimental design for causal structure learning
- Causal disentanglement

Causal structure learning (observational)

conceptually familiar

$$\mathfrak{G}_p := \text{the set of all DAGs on } p \text{ nodes}$$

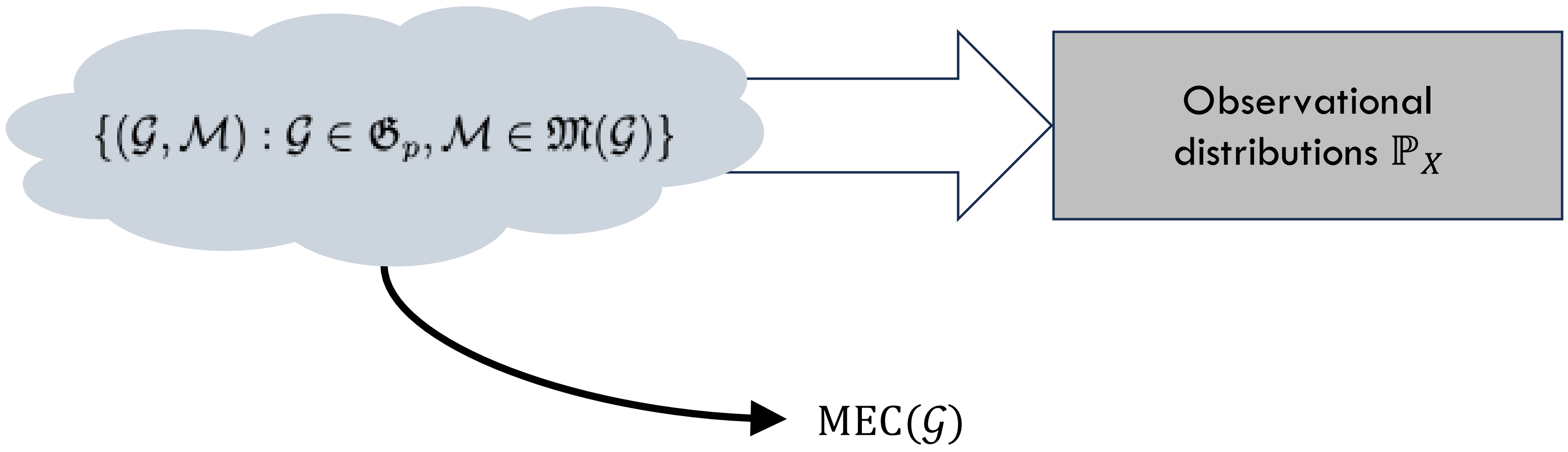


Causation, prediction, and search. Spirtes, Glymour, Scheines (1993).
Optimal structure identification with greedy search. Chickering (2002).

Causal structure learning (observational)

conceptually familiar

Without further restrictions on $\mathbb{P}_X$, $\mathcal{G}$ is only identifiable up to Markov equivalence.



Causation, prediction, and search. Spirtes, Glymour, Scheines (1993).
Optimal structure identification with greedy search. Chickering (2002).

Causal structure learning (interventional)

conceptually familiar

Suppose known intervention targets $\mathcal{I} = (\emptyset, T_1, \dots, T_K)$.

$$\{(\mathcal{G}, \mathcal{M}_0, \mathcal{M}_1, \ldots, \mathcal{M}_K) : \mathcal{G} \in \mathfrak{G}_p, \; (\mathcal{M}_0, \mathcal{M}_1, \ldots, \mathcal{M}_K) \in \mathfrak{M}_{\mathcal{I}}(\mathcal{G})\}$$

Tuples of distributions

$(\mathbb{P}_X, \mathbb{P}_X^{I_1}, \dots, \mathbb{P}_X^{I_K})$

$\mathcal{I} - \mathrm{MEC}(\mathcal{G})$

Characterization and greedy learning of interventional Markov equivalence classes of directed acyclic graphs. Hauser and Bühlmann (2012).
Permutation-based causal inference algorithms with interventions. Wang, Solus, Wang, Uhler (2017).
Characterizing and learning equivalence classes of causal DAGs under interventions. Yang, Katcoff, Uhler (2018).

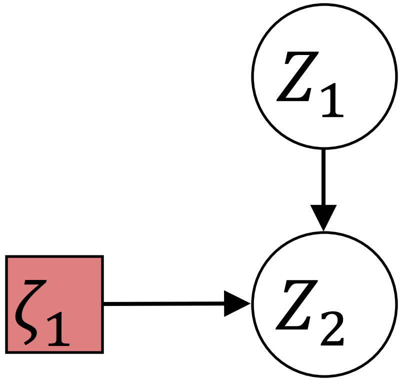
Outline

- *Background:* Identifiability
- *Background:* Structural causal models
- Causal structure learning
  - Learning from unknown-target interventions
  - Learning in the presence of unobserved variables
  - Experimental design for causal structure learning
- Causal disentanglement

We often don't know all targets of an intervention. For example, CRISPR is well-known to have *off-target effects [1]*.

[1] ***Unbiased detection of off-target cleavage by CRISPR-Cas9 and TALENs using integrase-defective lentiviral vectors.*** Wang, Wang, Wu et al. (2015).

We often don't know all targets of an intervention. For example, CRISPR is well-known to have *off-target effects [1]*.

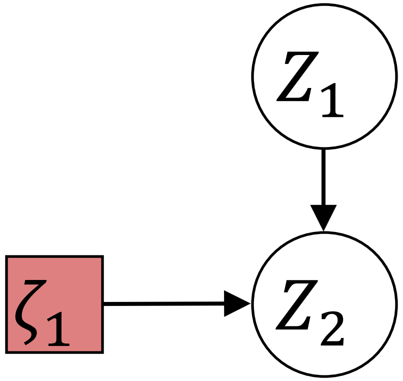


$$\zeta_1 = \mathbb{1}_{\text{sample from intervention } k}$$
$$Z_1 = f_1(\varepsilon_1)$$
$$Z_2 = f_2(Z_1, \varepsilon_2) \cdot (1 - \zeta_1) + f_2^I(Z_1, \varepsilon_2) \cdot \zeta_1$$

Solution: Treat each intervention as a variable. Its children correspond to its targets [2, 3].

[1] Unbiased detection of off-target cleavage by CRISPR-Cas9 and TALENs using integrase-defective lentiviral vectors. Wang, Wang, Wu et al. (2015).
[2] Joint causal inference from multiple contexts. Mooij, Magliacane, Claassen (2020).
[3] Causal discovery from soft intervention with unknown targets. Jaber, Kocaoglu, Shanmugam, Bareinboim (2020).

Conditional independence statements involving interventional variables correspond to **conditional invariances.**



$$\zeta_1 = \mathbb{1}_{\text{sample from intervention } k}$$
$$Z_1 = f_1(\varepsilon_1)$$
$$Z_2 = f_2(Z_1, \varepsilon_2) \cdot (1 - \zeta_1) + f_2^I(Z_1, \varepsilon_2) \cdot \zeta_1$$

$$\mathbb{P}(Z_1 \mid \zeta_1 = 0) = \mathbb{P}(Z_1 \mid \zeta_1 = 1)$$
$$\Leftrightarrow$$
$$\mathbb{P}(Z_1) = \mathbb{P}^{I_1}(Z_1)$$

[1] Joint causal inference from multiple contexts. Mooij, Magliacane, Claassen (2020).

Suggests that we can re-purpose existing methods for causal structure learning…

… but we need to handle intervention variables specially.

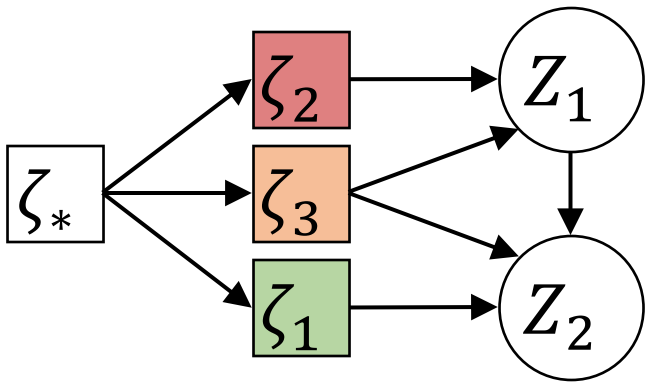
Background knowledge about intervention variables

Exogeneity [1]

They have no “system variables” as children.

Generic context [1]

They are fully connected (since $\zeta_k = 1$ implies $\zeta_{k'} = 0$ for any other k')



[1] Joint causal inference from multiple contexts. Mooij, Magliacane, Claassen (2020).

Permutation-based causal structure learning with unknown interventions targets.

Squires, Wang, Uhler (UAI 2020)

Yuhao
Wang

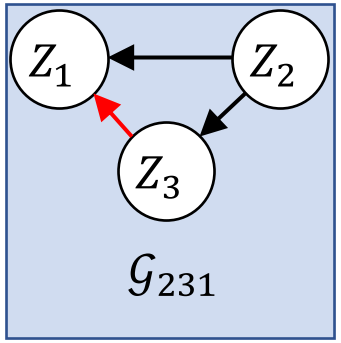
Caroline
Uhler

Unknown-Target Interventional Greedy Sparsest Permutation (UT-IGSP)

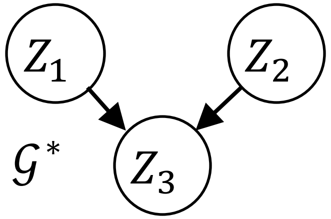
Re-purpose the Greedy Sparsest Permutation (GSP) algorithm [1].

Idea of GSP: search over the space of permutations (node orderings) instead of the space DAGs.

Each permutation can be associated with a **minimal I-MAP**: the sparsest graph consistent with the permutation which fits the data.



Covered edge:
$Z_i \rightarrow Z_j$ such that
$pa(Z_i) = pa(Z_j) \setminus \{Z_i\}$



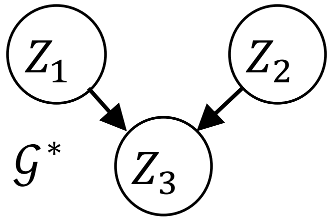
[1] *Consistency guarantees for greedy permutation-based causal inference algorithms.* Solus, Wang, Uhler (2021).

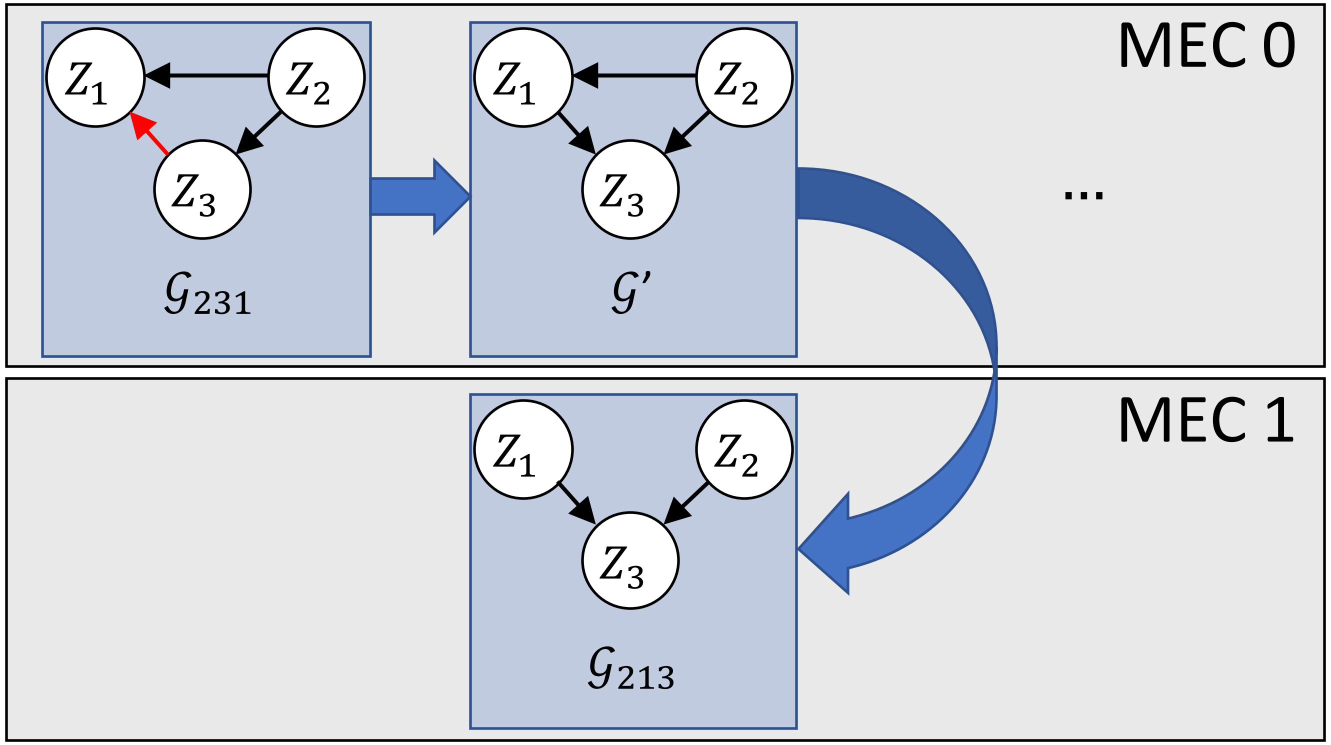
Unknown-Target Interventional Greedy Sparsest Permutation (UT-IGSP)

Re-purpose the Greedy Sparsest Permutation (GSP) algorithm [1].

Idea of GSP: search over the space of permutations (node orderings) instead of the space DAGs.

Each permutation can be associated with a **minimal I-MAP**: the sparsest graph consistent with the permutation which fits the data.



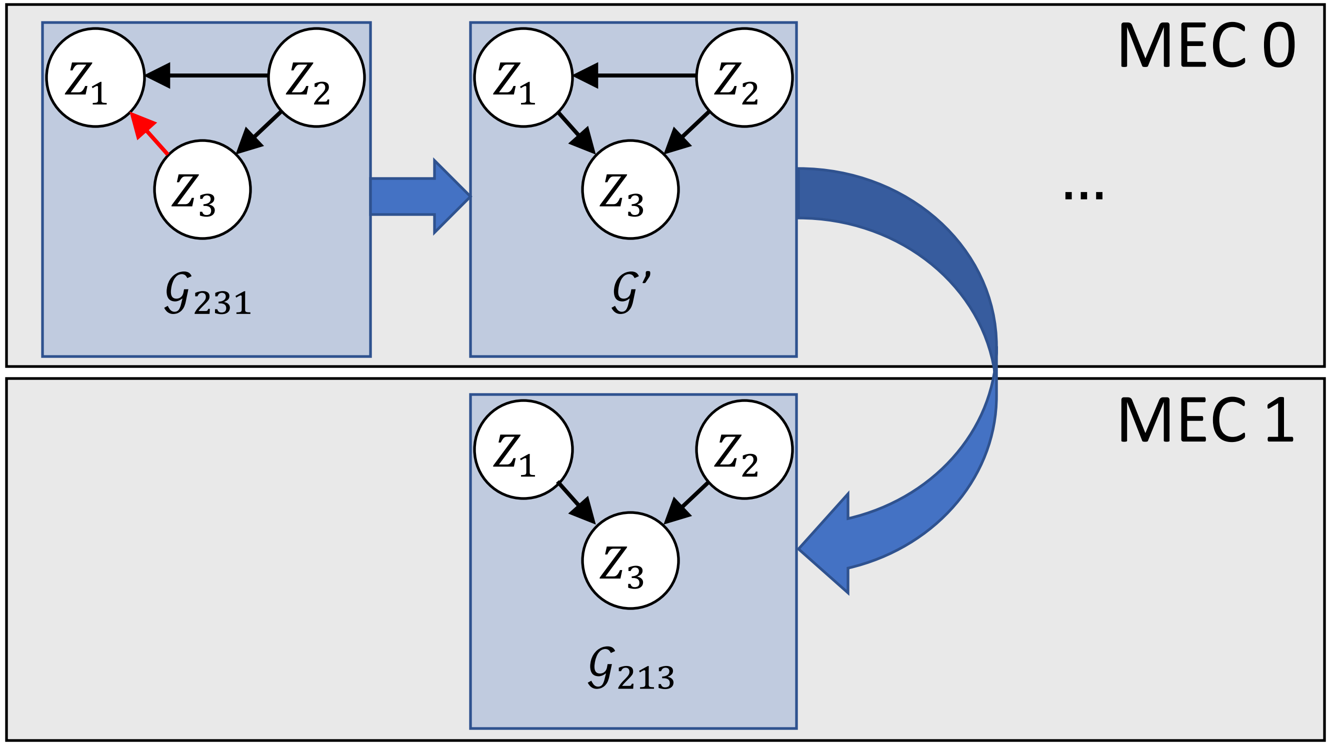


[1] Consistency guarantees for greedy permutation-based causal inference algorithms. Solus, Wang, Uhler (2021).

Unknown-Target Interventional Greedy Sparsest Permutation (UT-IGSP)

Consistency of greedy search: no local minima [1]

Let $\mathcal{G}$ be any I-MAP of $\mathcal{G}^*$. Then there exists a sequence of covered edge reversals and edge deletions from $\mathcal{G}$ to $\mathcal{G}^*$. We call this a **Chickering sequence.**



[1] Optimal structure identification with greedy search. Chickering (2002).

Unknown-Target Interventional Greedy Sparsest Permutation (UT-IGSP)

Interventional Chickering sequences

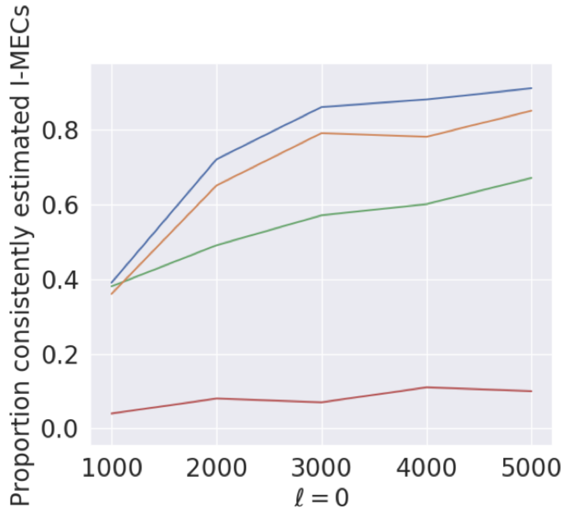
Theorem: If we start with all interventions satisfying exogeneity, then there are Chickering sequences which never violate exogeneity.

Unknown-Target Interventional Greedy Sparsest Permutation (UT-IGSP)

Interventional Chickering sequences

Theorem: If we start with all interventions satisfying exogeneity, then there are Chickering sequences which never violate exogeneity.

Code:
github.com/uhlerlab/causaldag



(b) Proportion of correctly estimated $\mathcal{I}$-MECs

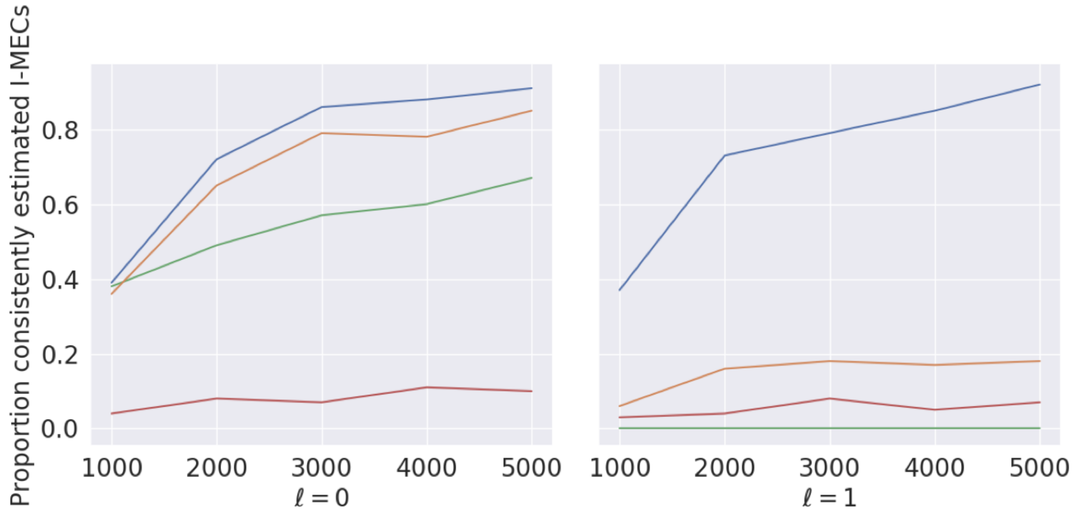
ℓ: number of off-target effects

Unknown-Target Interventional Greedy Sparsest Permutation (UT-IGSP)

Interventional Chickering sequences

Theorem: If we start with all interventions satisfying exogeneity, then there are Chickering sequences which never violate exogeneity.

Code:
github.com/uhlerlab/causaldag



(b) Proportion of correctly estimated $\mathcal{I}$-MECs

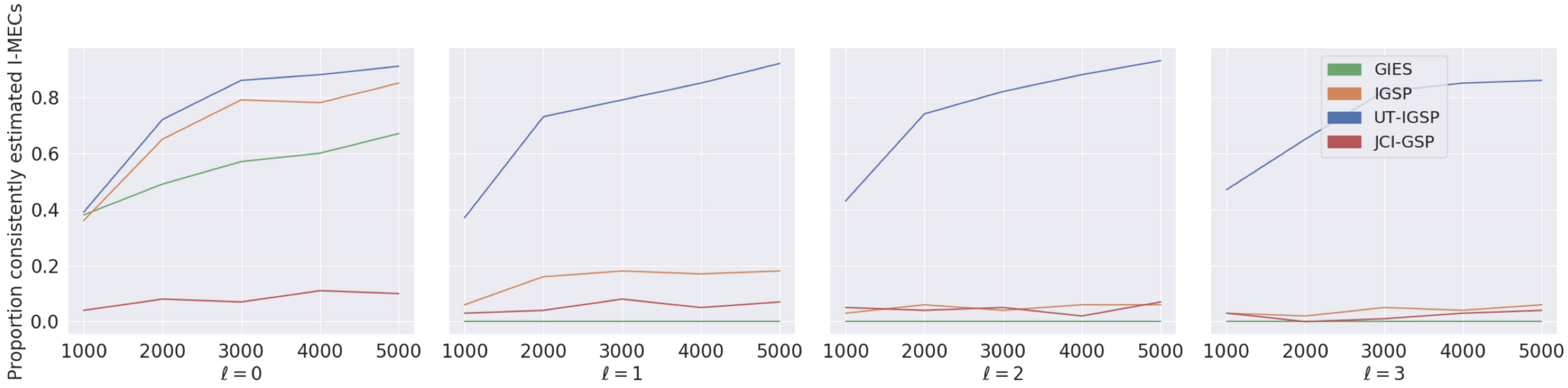
ℓ: number of off-target effects

Unknown-Target Interventional Greedy Sparsest Permutation (UT-IGSP)

Interventional Chickering sequences

Theorem: If we start with all interventions satisfying exogeneity, then there are Chickering sequences which never violate exogeneity.

Code:
github.com/uhlerlab/causaldag



(b) Proportion of correctly estimated $\mathcal{I}$-MECs

ℓ: number of off-target effects

Outline

- *Background:* Identifiability
- *Background:* Structural causal models
- Causal structure learning
  - Learning from unknown-target interventions
  - Learning in the presence of unobserved variables
  - Experimental design for causal structure learning
- Causal disentanglement

We often don't know have measurements of all relevant causal variables. There are many approaches to dealing with this.

We often don't know have measurements of all relevant causal variables. There are many approaches to dealing with this.

Change from DAG models to mixed graphs.

FCI [1], GSPo [2], DCD-UC [3]

[1] ***An anytime algorithm for causal inference.*** Spirtes (2001).

[2] ***Ordering-based causal structure learning in the presence of latent variables.*** Bernstein*, Saeed*, Squires*, Uhler (UAI 2020).

[3] ***Differentiable causal discovery under unmeasured confounding.*** Bhattacharya, Nagarajan, Malinsky, Shpitser (2021).

We often don't know have measurements of all relevant causal variables. There are many approaches to dealing with this.

Change from DAG models to mixed graphs.

FCI [1], GSPo [2], DCD-UC [3]

"Deconfound" by trying to learn latent variables.

LrPS + GES [4], DeCAMfounder [5]

[1] *An anytime algorithm for causal inference.* Spirtes (2001).

[2] *Ordering-based causal structure learning in the presence of latent variables.* Bernstein*, Saeed*, Squires*, Uhler (UAI 2020).

[3] *Differentiable causal discovery under unmeasured confounding.* Bhattacharya, Nagarajan, Malinsky, Shpitser (2021).

[4] *Robust causal structure learning with some hidden variables.* Frot, Nandy, Matthius (2017).

[5] *The DeCAMfounder: non-linear causal discovery in the presence of hidden variables.* Agrawal, Squires, Prasad, Uhler (JRSS-B 2023).

We often don’t know have measurements of all relevant causal variables. There are many approaches to dealing with this.

Change from DAG models to mixed graphs.

FCI [1], GSPo [2], DCD-UC [3]

“Deconfound” by trying to learn latent variables.

LrPS + GES [4], DeCAMfounder [5]

Use “low-rank” structures induced by latent variables

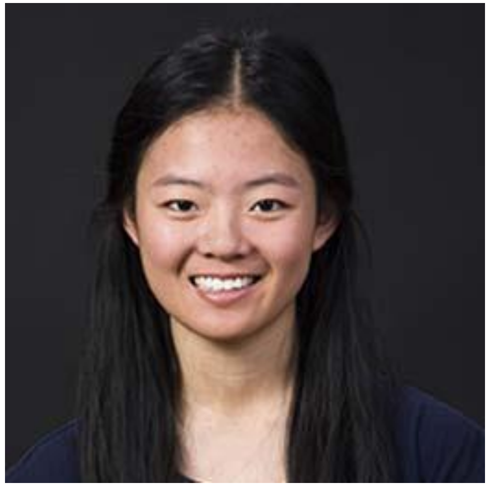
Adams [6], DCD-FG [7]

Our approach

[1] *An anytime algorithm for causal inference.* Spirtes (2001).
[2] *Ordering-based causal structure learning in the presence of latent variables.* Bernstein*, Saeed*, Squires*, Uhler (UAI 2020).
[3] *Differentiable causal discovery under unmeasured confounding.* Bhattacharya, Nagarajan, Malinsky, Shpitser (2021).
[4] *Robust causal structure learning with some hidden variables.* Frot, Nandy, Matthius (2017).
[5] *The DeCAMfounder: non-linear causal discovery in the presence of hidden variables.* Agrawal, Squires, Prasad, Uhler (JRSS-B 2023).
[6] *Identification of partially observed linear causal models: Graphical conditions for the non-gaussian and heterogeneous cases.* Adams, Hansen, Zhang (2021
[7] *Large-Scale Differentiable Causal Discovery of Factor Graphs.* Lopez, Hütter, Pritchard, Regez (2022).

Causal structure learning between clusters of nodes induced by latent factors.

Squires*, Yun*, Nichani, Agrawal, Uhler (CLeaR 2022).



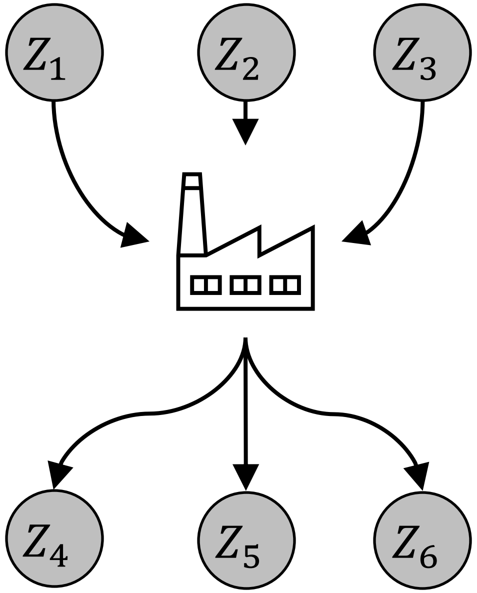
Annie
Yun

Eshaan
Nichani

Raj
Agrawal

Caroline
Uhler

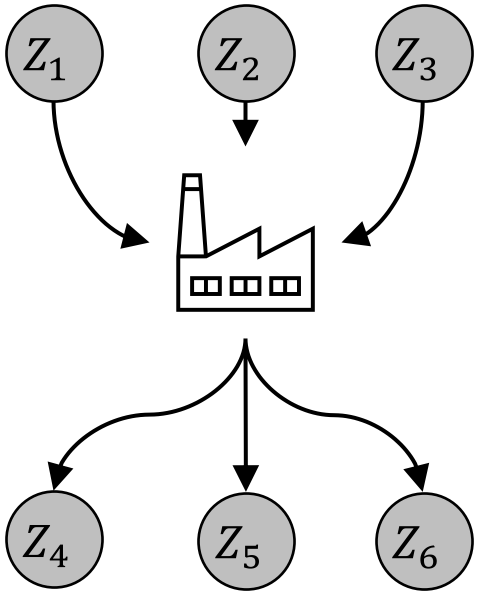
We use structure inspired by “transcription factories” [1]: locations in the nucleus where several transcription factors come together to regulate genes.



[1] ***Transcription factors and 3d genome conformation in cell-fate decisions.*** Stadhouders, Filian, Graf (2019).

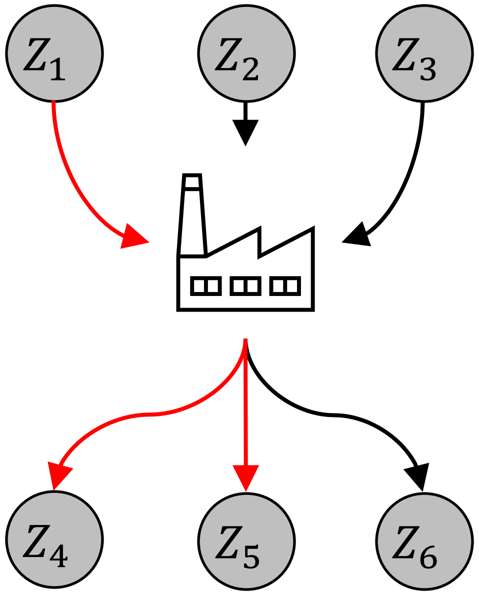
Treks

A trek between Z_i and Z_j is a tuple of directed paths with the same source and with Z_i and Z_j as their sinks.



Treks

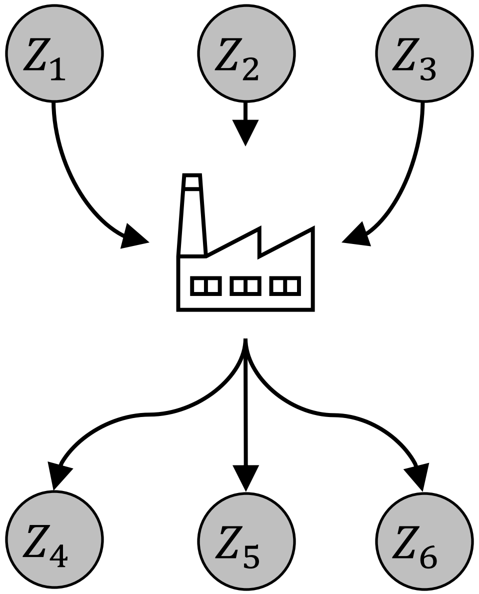
A trek between Z_i and Z_j is a tuple of directed paths with the same source and with Z_i and Z_j as their sinks.



Trek between Z_4 and Z_5 with paths $Z_1 \rightarrow F \rightarrow Z_4$ and $Z_1 \rightarrow F \rightarrow Z_5$

Treks

A trek between Z_i and Z_j is a tuple of directed paths with the same source and with Z_i and Z_j as their sinks.



Trek separation [1]

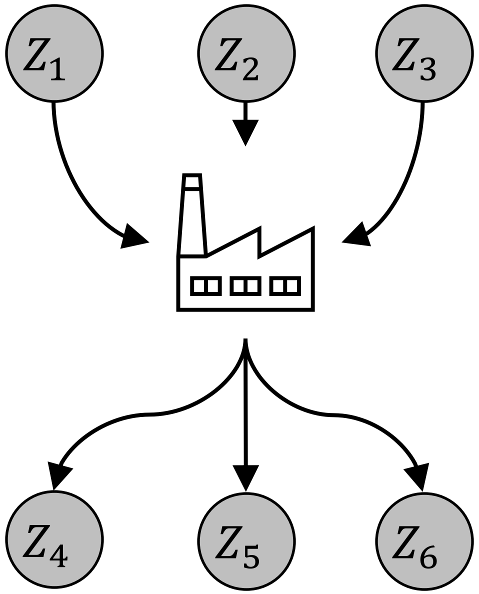
(specialized version)

A node F trek-separates (t-separates) sets $\boldsymbol{A}$ and $\boldsymbol{B}$ if all treks from $Z_i \in \boldsymbol{A}$ to $Z_j \in \boldsymbol{B}$ pass through F.

*[1] **Trek separation in Gaussian graphical models.*** Sullivant, Talaska, Draisma (2010).

Treks

A trek between Z_i and Z_j is a tuple of directed paths with the same source and with Z_i and Z_j as their sinks.



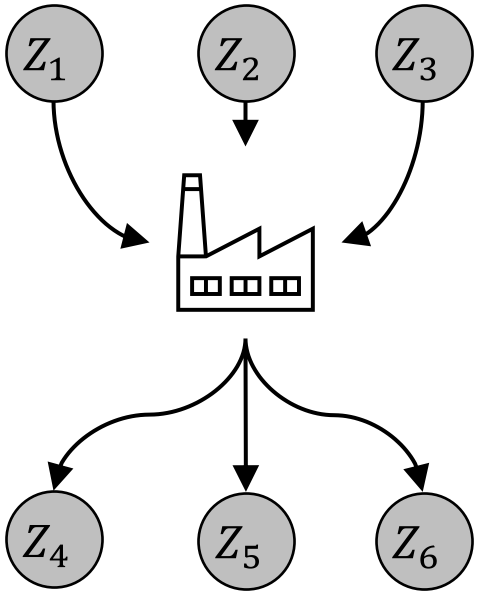
F t-separates
$\boldsymbol{A} = \{Z_4, Z_5\}$
from
$\boldsymbol{B} = \{Z_6\}$

Trek separation [1]

(specialized version)

A node F trek-separates (t-separates) sets $\boldsymbol{A}$ and $\boldsymbol{B}$ if all treks from $Z_i \in \boldsymbol{A}$ to $Z_j \in \boldsymbol{B}$ pass through F.

[1] Trek separation in Gaussian graphical models. Sullivant, Talaska, Draisma (2010).



Treks

A trek between Z_i and Z_j is a tuple of directed paths with the same source and with Z_i and Z_j as their sinks.

Trek separation [1]

(specialized version)

A node F trek-separates (t-separates) sets $\boldsymbol{A}$ and $\boldsymbol{B}$ if all treks from $Z_i \in \boldsymbol{A}$ to $Z_j \in \boldsymbol{B}$ pass through F.

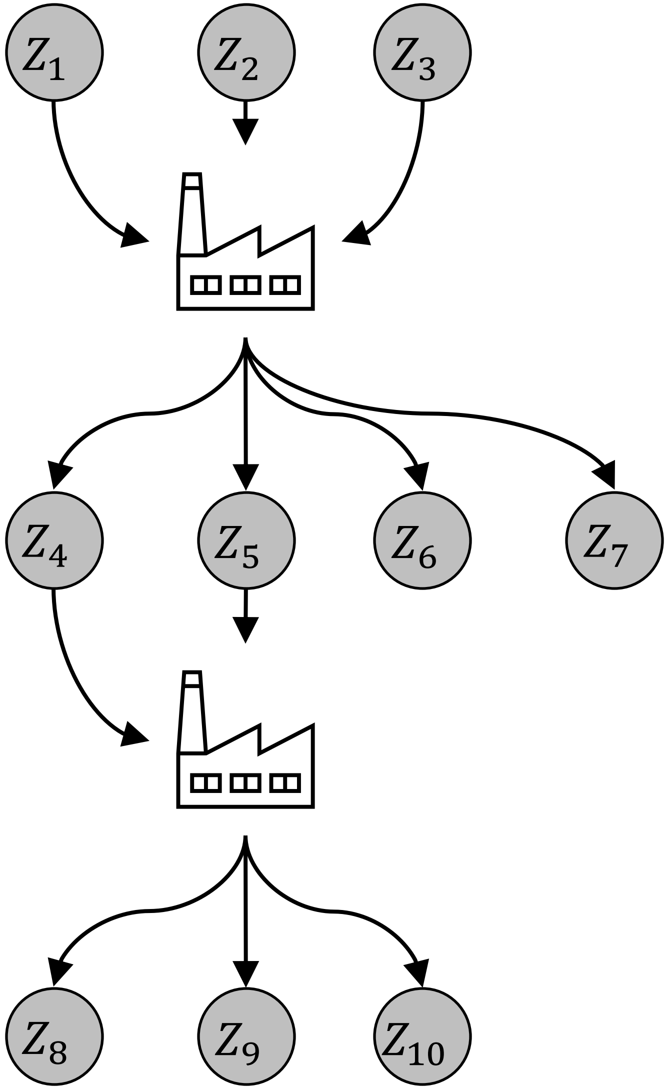
Trek separation theorem [1]

(specialized version)

If a node F t-separates $\boldsymbol{A}$ and $\boldsymbol{B}$, then $\mathrm{rank}(\Sigma_{\boldsymbol{A},\boldsymbol{B}}) = 1$.

If no node t-separates $\boldsymbol{A}$ and $\boldsymbol{B}$, then $\mathrm{rank}(\Sigma_{\boldsymbol{A},\boldsymbol{B}}) = 2$, generically.

[1] Trek separation in Gaussian graphical models. Sullivant, Talaska, Draisma (2010).



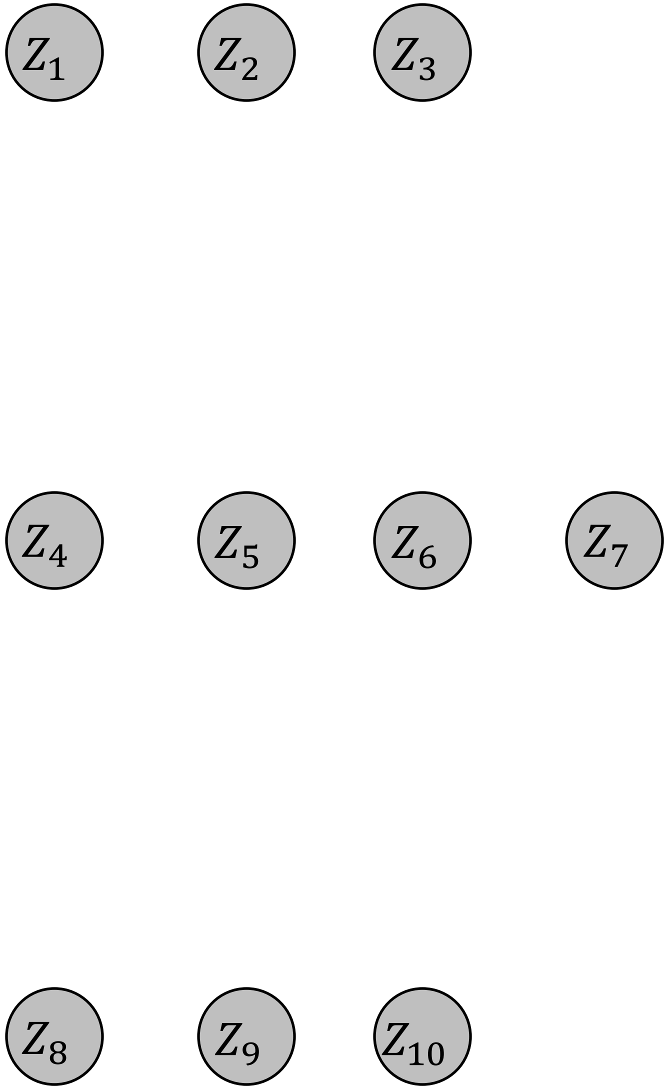
Latent Factor Causal Models (LFCMs)

(a) **Unique cluster:** Each observed variable has a single latent parent.
(b) **Bipartite:** No latent → latent or observed → observed edges.
(c) **Triple child:** Each latent variable has ≥ 3 children.
(d) **Double parent:** Latents "connected" by at least two children.

Z_i and Z_j are in the bottom-most cluster

$\Leftrightarrow$

$\boldsymbol{A} = \{Z_i, Z_j\}$ is t-separated from $\boldsymbol{B} = [d] \setminus \boldsymbol{A}$ by a single node



Z_i and Z_j are in the bottom-most cluster

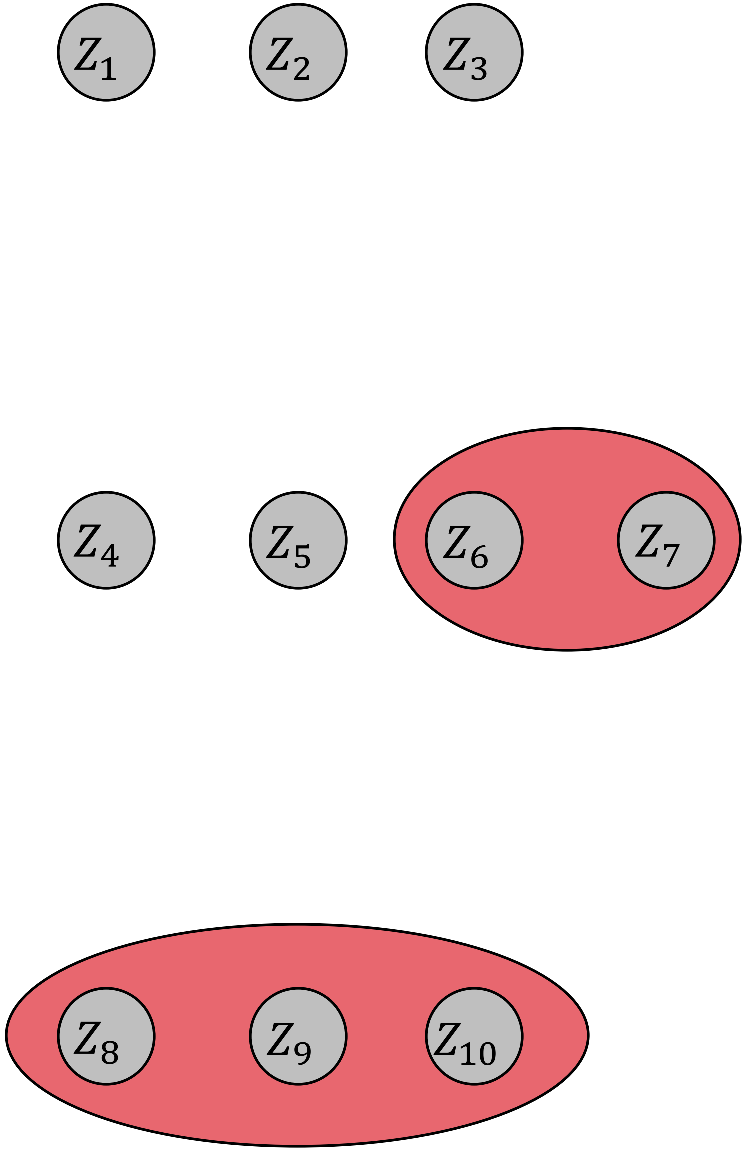
$\Leftrightarrow$

$\boldsymbol{A} = \{Z_i, Z_j\}$ is t-separated from $\boldsymbol{B} = [d] \setminus \boldsymbol{A}$ by a single node

Consequence: We can identify nodes in the same bottom-most cluster by testing whether certain sub-matrices are rank one.

EstimateLFCM algorithm:

(1) Iteratively find bottom-most clusters.



Z_i and Z_j are in the bottom-most cluster

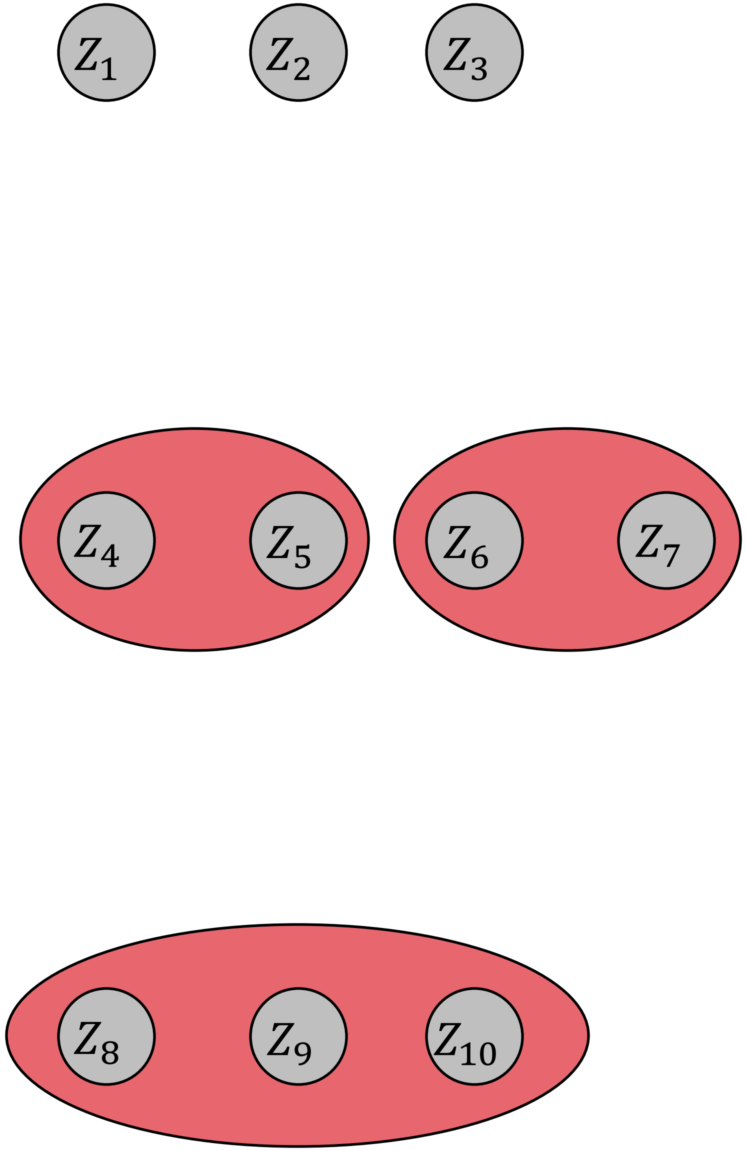
$\Leftrightarrow$

$\boldsymbol{A} = \{Z_i, Z_j\}$ is t-separated from $\boldsymbol{B} = [d] \setminus \boldsymbol{A}$ by a single node

Consequence: We can identify nodes in the same bottom-most cluster by testing whether certain sub-matrices are rank one.

EstimateLFCM algorithm:

(1) Iteratively find bottom-most clusters.



Z_i and Z_j are in the bottom-most cluster

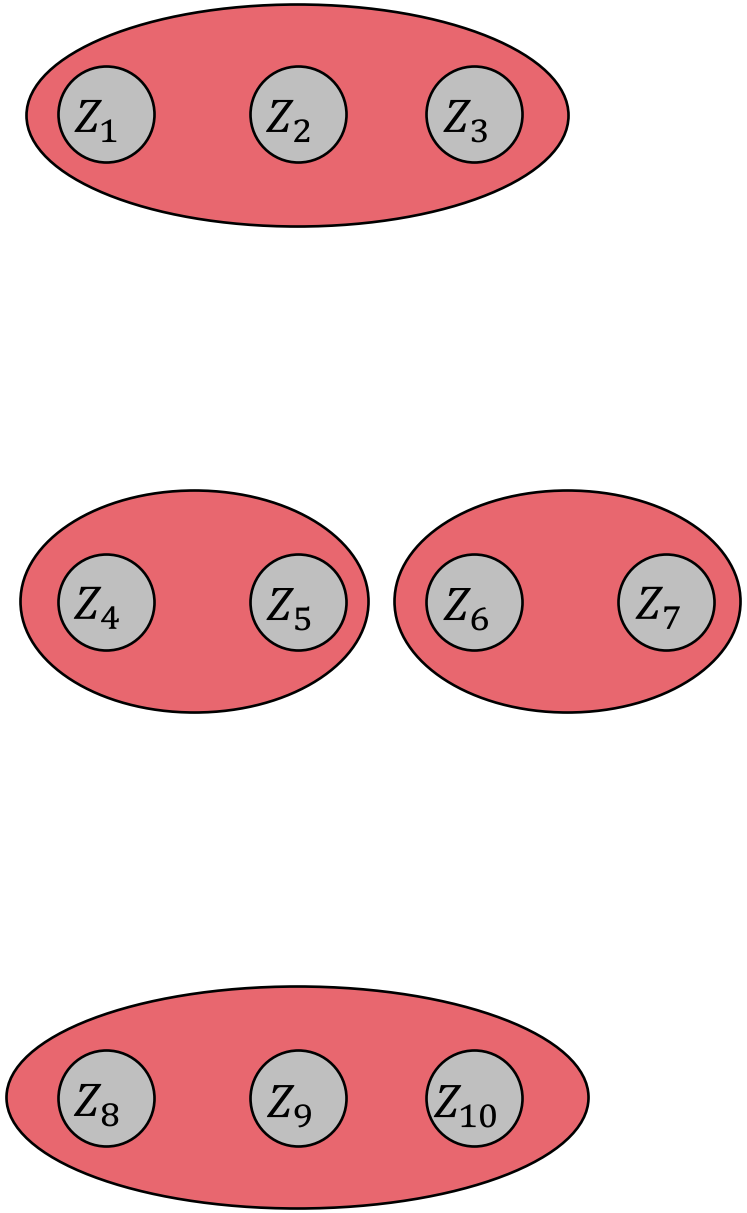
$\Leftrightarrow$

$\boldsymbol{A} = \{Z_i, Z_j\}$ is t-separated from $\boldsymbol{B} = [d] \setminus \boldsymbol{A}$ by a single node

Consequence: We can identify nodes in the same bottom-most cluster by testing whether certain sub-matrices are rank one.

EstimateLFCM algorithm:

(1) Iteratively find bottom-most clusters.



Z_i and Z_j are in the bottom-most cluster

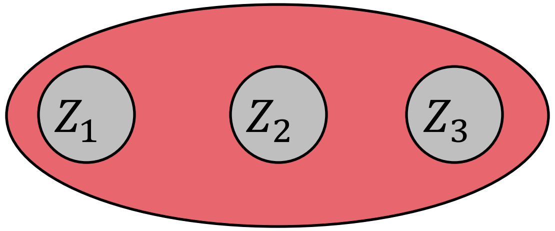
$\Leftrightarrow$

$\boldsymbol{A} = \{Z_i, Z_j\}$ is t-separated from $\boldsymbol{B} = [d] \setminus \boldsymbol{A}$ by a single node

Consequence: We can identify nodes in the same bottom-most cluster by testing whether certain sub-matrices are rank one.

EstimateLFCM algorithm:

(1) Iteratively find bottom-most clusters.

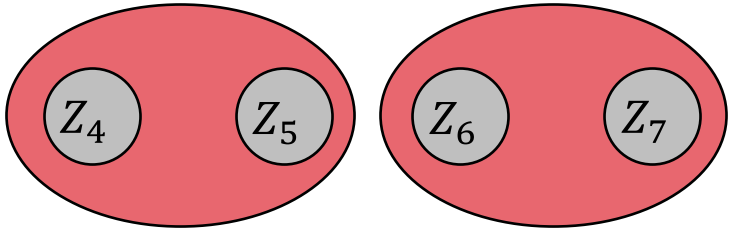


Z_i and Z_j are in the bottom-most cluster

$\Leftrightarrow$

$\boldsymbol{A} = \{Z_i, Z_j\}$ is t-separated from $\boldsymbol{B} = [d] \setminus \boldsymbol{A}$ by a single node

Consequence: We can identify nodes in the same bottom-most cluster by testing whether certain sub-matrices are rank one.

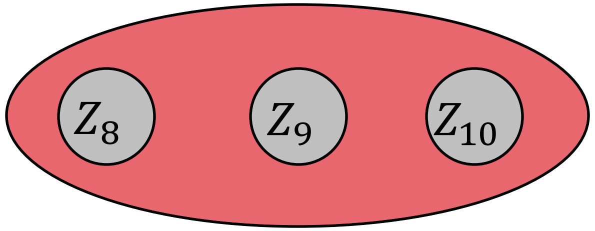


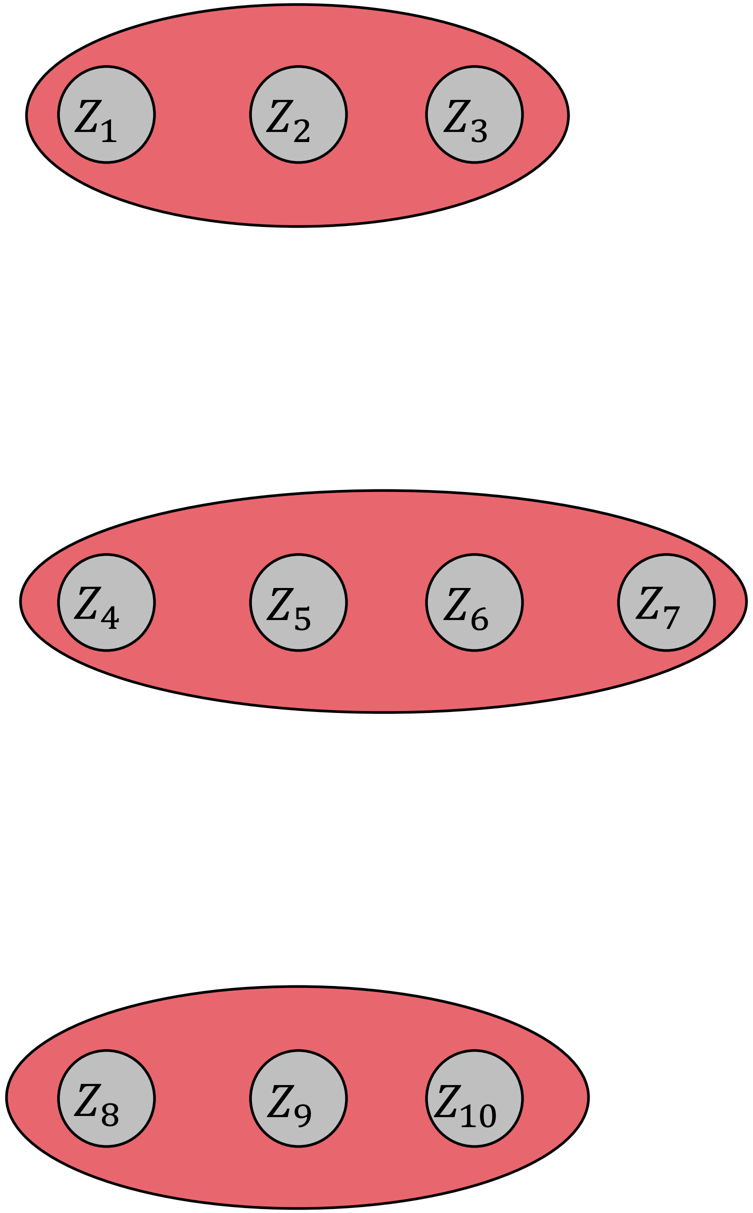
EstimateLFCM algorithm:

(1) Iteratively find bottom-most clusters.

(2) “Merging” handles split-up clusters.

$\{Z_4, Z_5\}$ are split up from $\{Z_6, Z_7\}$ in the first step.





Z_i and Z_j are in the bottom-most cluster

$\Leftrightarrow$

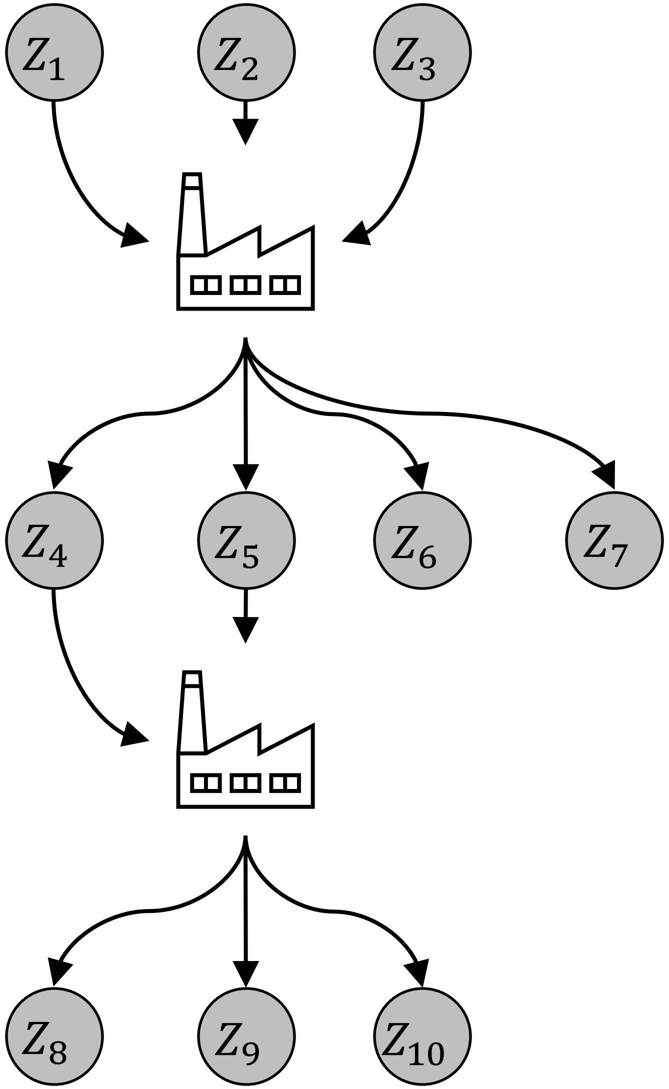
$\boldsymbol{A} = \{Z_i, Z_j\}$ is t-separated from $\boldsymbol{B} = [d] \setminus \boldsymbol{A}$ by a single node

Consequence: We can identify nodes in the same bottom-most cluster by testing whether certain sub-matrices are rank one.

EstimateLFCM algorithm:

(1) Iteratively find bottom-most clusters.

(2) "Merging" handles split-up clusters.



Z_i and Z_j are in the bottom-most cluster

$\Leftrightarrow$

$\boldsymbol{A} = \{Z_i, Z_j\}$ is t-separated from $\boldsymbol{B} = [d] \setminus \boldsymbol{A}$ by a single node

Consequence: We can identify nodes in the same bottom-most cluster by testing whether certain sub-matrices are rank one.

EstimateLFCM algorithm:

(1) Iteratively find bottom-most clusters.

(2) "Merging" handles split-up clusters.

(3) Add edges between clusters.

Outline

- *Background:* Identifiability
- *Background:* Structural causal models
- Causal structure learning
  - Learning from unknown-target interventions
  - Learning in the presence of unobserved variables
  - Experimental design for causal structure learning
- Causal disentanglement

Experimental design: how to learn a causal graph in as few interventions as possible?

[1] ***n-1 experiments suffice to determine the causal relations among n variables.*** Eberhardt, Glymour, Scheines (2006).
[2] ***Two optimal strategies for active learning of causal models from interventional data.*** Hauser and Bühlmann (2014).
[3] ***ABCD-Strategy: Budgeted experimental design for targeted causal structure discovery.*** Agrawal, Squires, Yang, Shanmugam, Uhler (NeurIPS 2019).
[4] ***Matching a desired causal state via shift interventions.*** Zhang, Squires, Uhler (NeurIPS 2021).
[5] ***Active learning for optimal intervention design in causal models.*** Zhang, Cammarata, Squires, Sapsis, Uhler (NMI 2023, to appear).

Active structure learning of causal DAGs via directed clique trees.

Squires, Magliacane, Greenewald, Katz, Kocaoglu, Shanmugam (NeurIPS 2020).

Sara Magliacane

Kristjan Greenewald

Dmitry Katz

Murat Kocaoglu

Karthikeyan Shanmugam

Worst-case lower bounds

Previous results [1,2] lower bound how many interventions are required in the *worst-case* over all DAGs in an equivalence class.

[1] ***Two optimal strategies for active learning of causal models from interventional data.*** Hauser and Bühlmann (2014).
[2] ***Learning causal graphs with small interventions***. Shanmugam, Kocaoglu, Dimakis, Vishwanath (2015).

Worst-case lower bounds

Previous results [1,2] lower bound how many interventions are required in the *worst*-case over all DAGs in an equivalence class.

Not necessarily a good measure of difficulty

Some graphs could be much harder to learn than others.

Worst-case lower bounds

Previous results [1,2] lower bound how many interventions are required in the *worst*-*case* over all DAGs in an equivalence class.

Not necessarily a good measure of difficulty

Some graphs could be much harder to learn than others.

Instance-wise difficulty

We define a **verifying intervention set (VIS)** for a DAG $\mathcal{G}$: a set $\mathcal{I}$ of intervention targets which makes the DAG identifiable. Difficulty can be measured by $m(\mathcal{G})$, the size of the smallest VIS.

Worst-case lower bounds

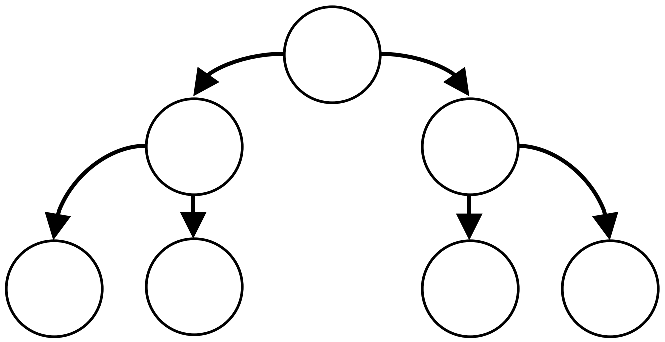
Previous results [1,2] lower bound how many interventions are required in the *worst-case* over all DAGs in an equivalence class.

Not necessarily a good measure of difficulty

Some graphs could be much harder to learn than others.

Instance-wise difficulty

We define a **verifying intervention set (VIS)** for a DAG $\mathcal{G}$: a set $\mathcal{I}$ of intervention targets which makes the DAG identifiable. Difficulty can be measured by $m(\mathcal{G})$, the size of the smallest VIS.



$m(\mathcal{G})$ for a tree = 1

Optimal experimental design for tree on d nodes: $O(\log d)$ interventions [1].

[1] ***Sample-efficient active learning of causal trees.*** Greenewald, Katz, Shanmugam, Magliacane, Kocaoglu, Boix-Adsera, Bresler (2019)

Worst-case lower bounds

Previous results [1,2] lower bound how many interventions are required in the *worst*-*case* over all DAGs in an equivalence class.

Not necessarily a good measure of difficulty

Some graphs could be much harder to learn than others.

Instance-wise difficulty

We define a **verifying intervention set (VIS)** for a DAG $\mathcal{G}$: a set $\mathcal{I}$ of intervention targets which makes the DAG identifiable. Difficulty can be measured by $m(\mathcal{G})$, the size of the smallest VIS.

Interpretation of $m(\mathcal{G})$

"Teacher" knows $\mathcal{G}$ and wants "Student" to learn it. What's the most efficient way to teach?

Characterization of $m(\mathcal{G})$ via directed clique trees (DCTs)

We extend the notion of clique trees (aka "junction trees" or "tree decompositions") of undirected chordal graphs to DAGs.

This lets us uniformly lower bound $m(\mathcal{G})$ over an equivalence class.

DCT Policy

We develop an intervention policy which is within a $\log(d)$ factor of $m(\mathcal{G})$, for any $\mathcal{G}$ in a restricted class[1].

[1]The class restriction was removed in [1].

[1] Verification and search algorithms for causal DAGs. Choo, Shiragur, Bhattacharyya (2022).

Outline

- *Background:* Identifiability
- *Background:* Structural causal models
- Causal structure learning
  - Learning from unknown-target interventions
  - Learning in the presence unobserved variables
  - Experimental design for causal structure learning
- Causal disentanglement

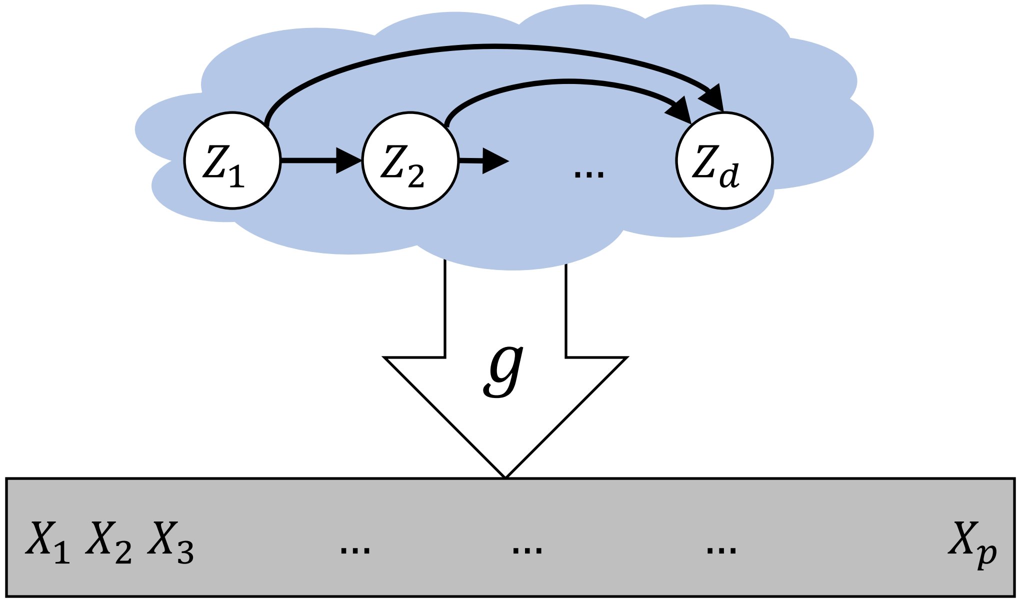
Causal Disentanglement Models

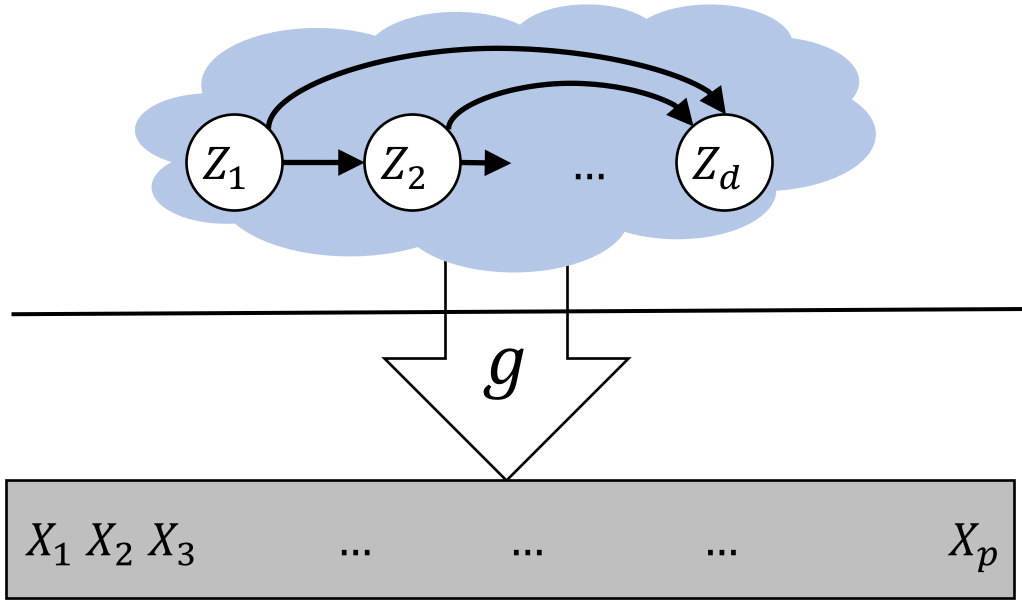
Conceptually novel

(Unobserved)
Macro-variables

(Unobserved)
Mixing function

(Observed)
Micro-variables

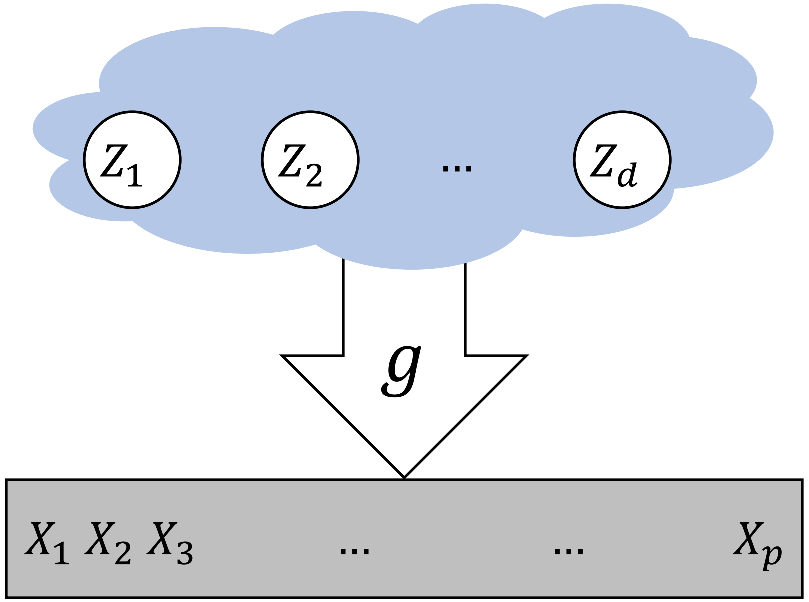




	Cellular Biology	Neuroscience
	• Protein concentrations • Cellular morphology (e.g. nucleus shape)	• Neurotransmitter concentrations • Reuptake rate
	• Fluorescent microscopy images • Gene expression (RNAseq)	• Neuroimaging data (fMRI) • Electrical activity (LFP)

Approaches to identifiability of causal disentanglement

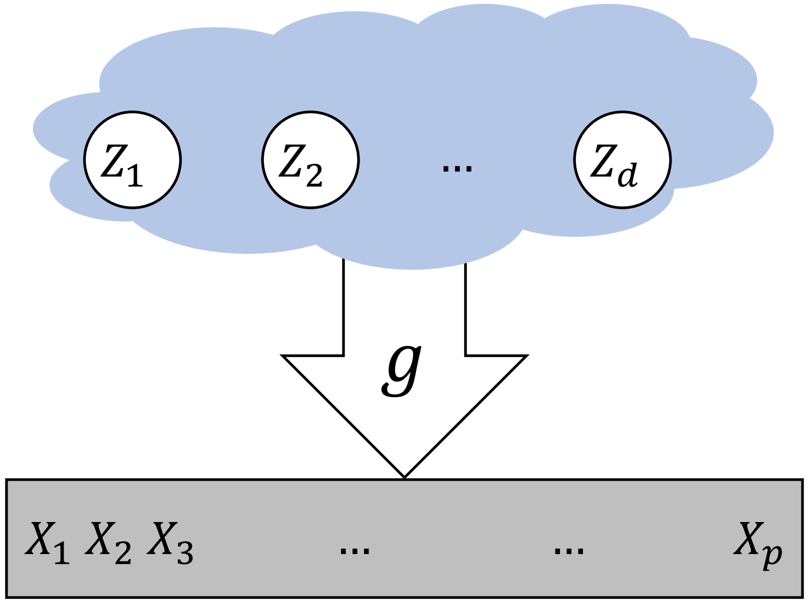
Restrict latent DAG $\mathcal{G}$



[1], [2]

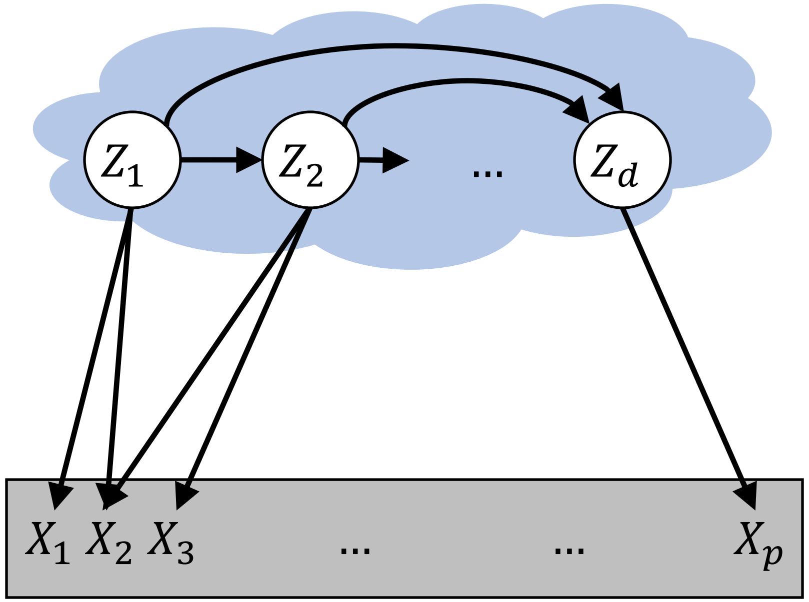
[1] ***Independent component analysis: a new concept?*** Comon (1994).
[2] ***Nonlinear ICA using auxiliary variables and generalized contrastive learning.*** Hyvärinen, Sasaki, Turner (2019).

Restrict latent DAG $\mathcal{G}$



[1], [2]

Restrict mixing function g



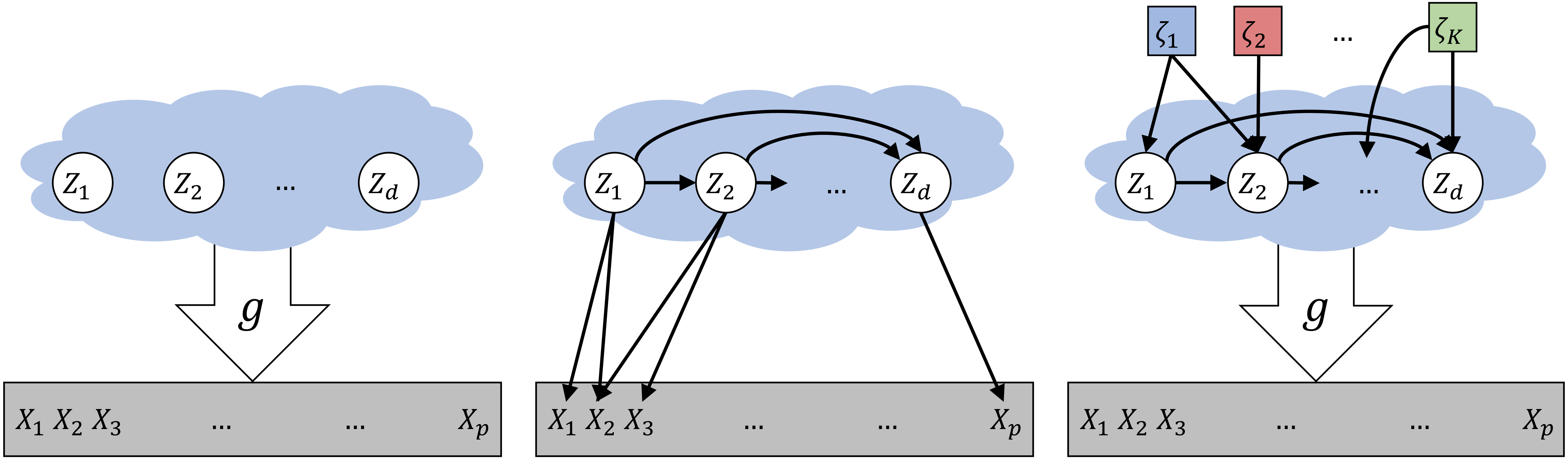
[3], [4], [5], [6], [7]

[1] ***Independent component analysis: a new concept?*** Comon (1994).
[2] ***Nonlinear ICA using auxiliary variables and generalized contrastive learning.*** Hyvärinen, Sasaki, Turner (2019).
[3] ***Learning the structure of linear latent variable models.*** Silva, Scheines, Glymour, Spirtes, Chickering (2006).
[4] ***Anchored discrete factor analysis.*** Halpern, Hong, Sontag (2015).
[5] ***Triad constraints for learning causal structure of latent variables.*** Cai, Xie, Glymour, Hao, Zhang (2019).
[6] ***Learning latent causal graphs via mixture oracles.*** Kivva, Rajendran, Ravikumar, Aragam (2019).
[7] ***Identification of linear non-Gaussian latent hierarchical structure.*** Xie, Huang, Chen, He, Geng, Zhang (2022).

Restrict latent DAG $\mathcal{G}$

Restrict mixing function g

Incorporate multiple contexts



[1], [2]

[3], [4], [5], [6], [7]

Papers on the next slide

[1] ***Independent component analysis: a new concept?*** Comon (1994).
[2] ***Nonlinear ICA using auxiliary variables and generalized contrastive learning.*** Hyvärinen, Sasaki, Turner (2019).
[3] ***Learning the structure of linear latent variable models.*** Silva, Scheines, Glymour, Spirtes, Chickering (2006).
[4] ***Anchored discrete factor analysis.*** Halpern, Hong, Sontag (2015).
[5] ***Triad constraints for learning causal structure of latent variables.*** Cai, Xie, Glymour, Hao, Zhang (2019).
[6] ***Learning latent causal graphs via mixture oracles.*** Kivva, Rajendran, Ravikumar, Aragam (2019).
[7] ***Identification of linear non-Gaussian latent hierarchical structure.*** Xie, Huang, Chen, He, Geng, Zhang (2022).

	Latent model	Mixing function	Intervention type	Intervention size
Squires et al. (2023)	Linear	Linear*	Perfect†	1
Ahuja et al. (2023)	Non-parametric	Linear*	Do	1
Buchholz et al. (2023)	Linear	Non-parametric	Perfect†	1
Varici et al. (2023)	Non-parametric	Linear*	Perfect†	1
von Kügelgen et al. (2023)	Non-parametric	Non-parametric	Perfect††	1
Zhang et al. (2023)	Non-parametric	Linear*	Soft	1

*Can be extended to mixing by a "full-rank" polynomial.
†Also show that soft interventions give a weaker form of identifiability.
††Only for $d = 2$, extension to arbitrary d with additional data.

[1] ***Interventional causal representation learning.*** Ahuja, Mahajan, Bengio (2023).
[2] ***Score-based causal representation learning with interventions.*** Varici, Acatürk, Shanmugam Kumar, Tajer (2023).
[3] ***Learning linear causal representations from interventions under general nonlinear mixing.*** Buchholz, Rajendran, Rosenfeld, Aragam, Schölkopf, Ravikumar (2023).
[4] ***Nonparametric identifiability of causal representation from unknown interventions.*** von Kügelgen, Besserve, Liang, Gresele, Kekić, Bareinboim, Blei, Schölkopf (2023).
[5] ***Identifiability Guarantees for Causal Disentanglement from Soft Interventions.*** Zhang, Squires, Greenewald, Srivastava, Shanmugam, Uhler (2023, under submission).

	Latent model	Mixing function	Intervention type	Intervention size
Squires et al. (2023)	**Linear**	**Linear***	**Perfect†**	**1**
Ahuja et al. (2023)	Non-parametric	Linear*	Do	1
Buchholz et al. (2023)	Linear	Non-parametric	Perfect†	1
Varici et al. (2023)	Non-parametric	Linear*	Perfect†	1
von Kügelgen et al. (2023)	Non-parametric	Non-parametric	Perfect††	1
Zhang et al. (2023)	Non-parametric	Linear*	Soft	1

*Can be extended to mixing by a "full-rank" polynomial.
†Also show that soft interventions give a weaker form of identifiability.
††Only for $d = 2$, extension to arbitrary d with additional data.

*[1] **Interventional causal representation learning.*** Ahuja, Mahajan, Bengio (2023).
*[2] **Score-based causal representation learning with interventions.*** Varici, Acatürk, Shanmugam Kumar, Tajer (2023).
*[3] **Learning linear causal representations from interventions under general nonlinear mixing.*** Buchholz, Rajendran, Rosenfeld, Aragam, Schölkopf, Ravikumar (2023).
*[4] **Nonparametric identifiability of causal representation from unknown interventions.*** von Kügelgen, Besserve, Liang, Gresele, Kekić, Bareinboim, Blei, Schölkopf (2023).
*[5] **Identifiability Guarantees for Causal Disentanglement from Soft Interventions.*** Zhang, Squires, Greenewald, Srivastava, Shanmugam, Uhler (2023, under submission).

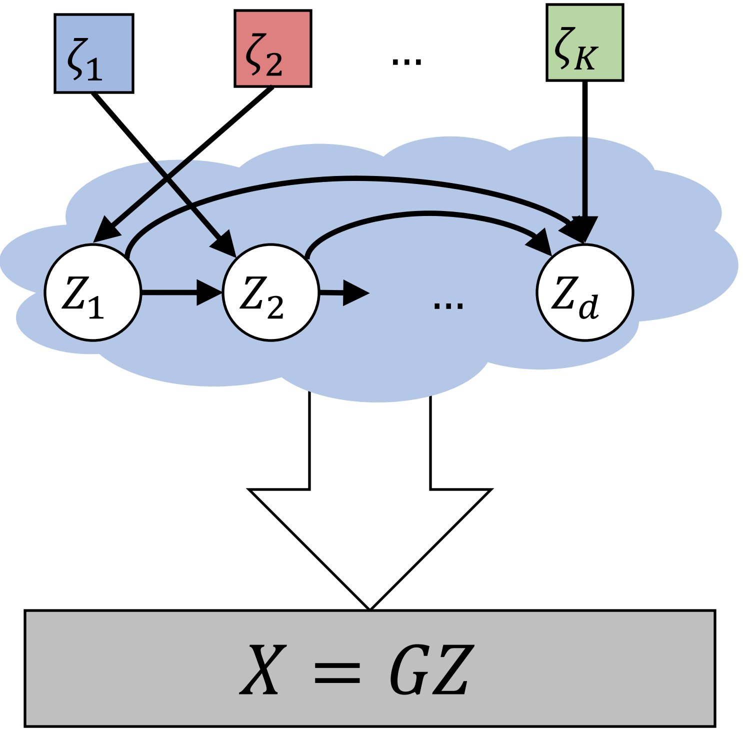
Linear causal disentanglement via interventions.

Squires*, Seigal*, Bhate, Uhler (ICML 2023).

Anna
Seigal

Salil
Bhate

Caroline
Uhler



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

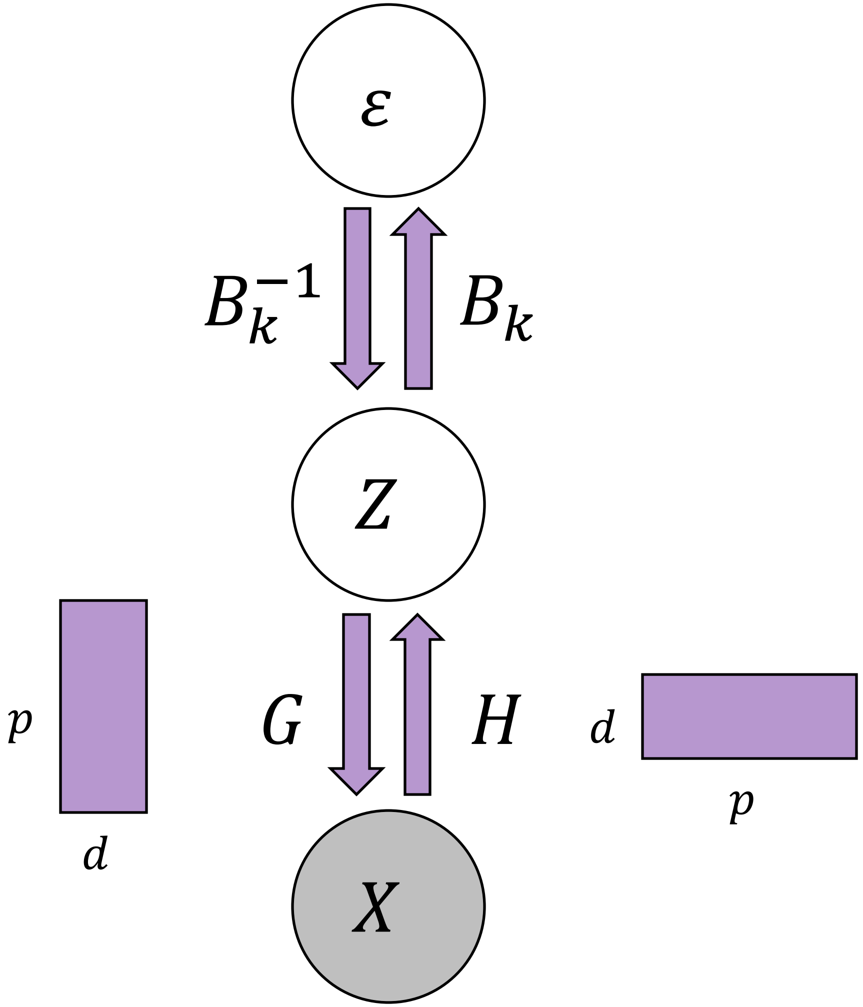
Control		…
$Z_1 = \sigma_1 \varepsilon_1$	$Z_1 = \sigma_1' \varepsilon_1$	$Z_1 = \sigma_1 \varepsilon_1$
$Z_2 = A_{12} Z_1 + \sigma_2 \varepsilon_2$	$Z_2 = A_{12} Z_1 + \sigma_2 \varepsilon_2$	$Z_2 = A_{12}' Z_1 + \sigma_2' \varepsilon_2$
$\vdots$	$\vdots$	$\vdots$
$Z_d = A_{1d} Z_1 + A_{2d} Z_2 + \cdots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

Compact version:

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

Equivalently,

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



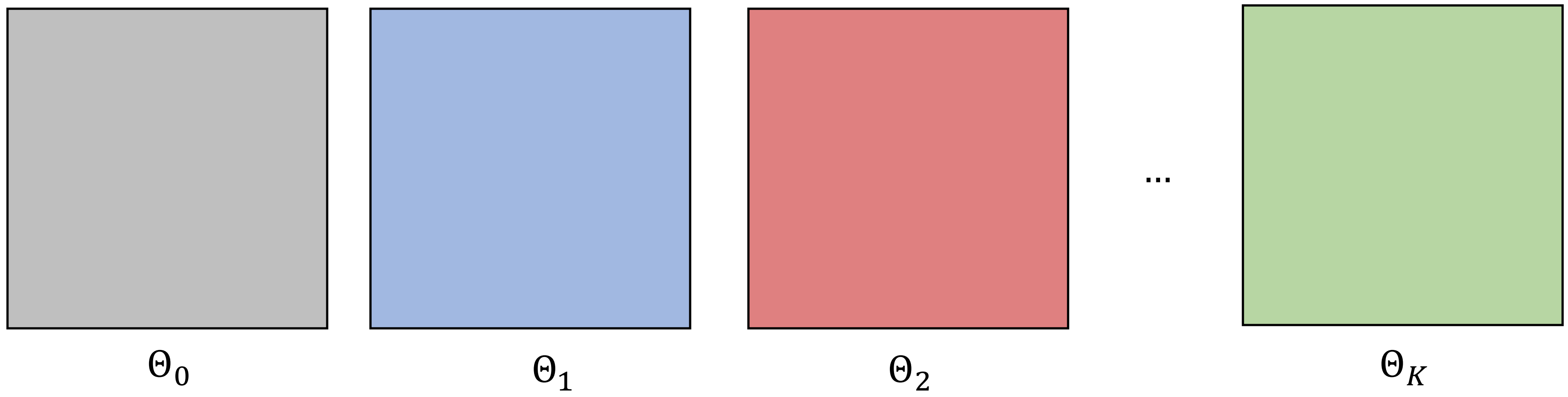
$H = G^{\dagger}$

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

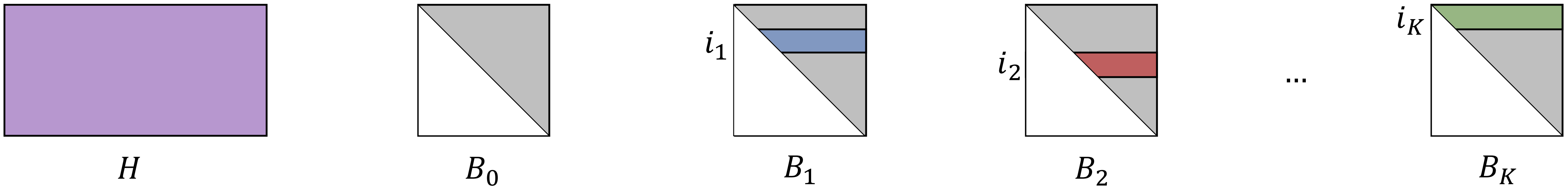
$$\mathrm{Cov}_k(Z)^{-1} = B_k^{\top} B_k$$

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

Input:



Output:



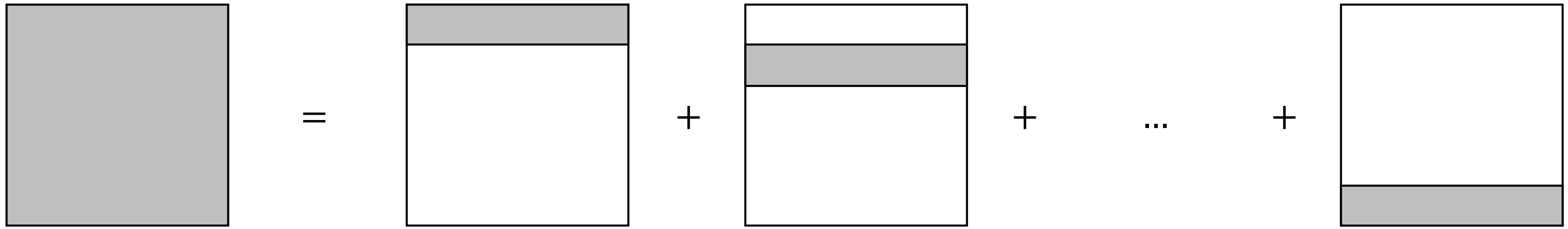
such that $\Theta_k = H^\top B_k^\top B_k H$ for all k.

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

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

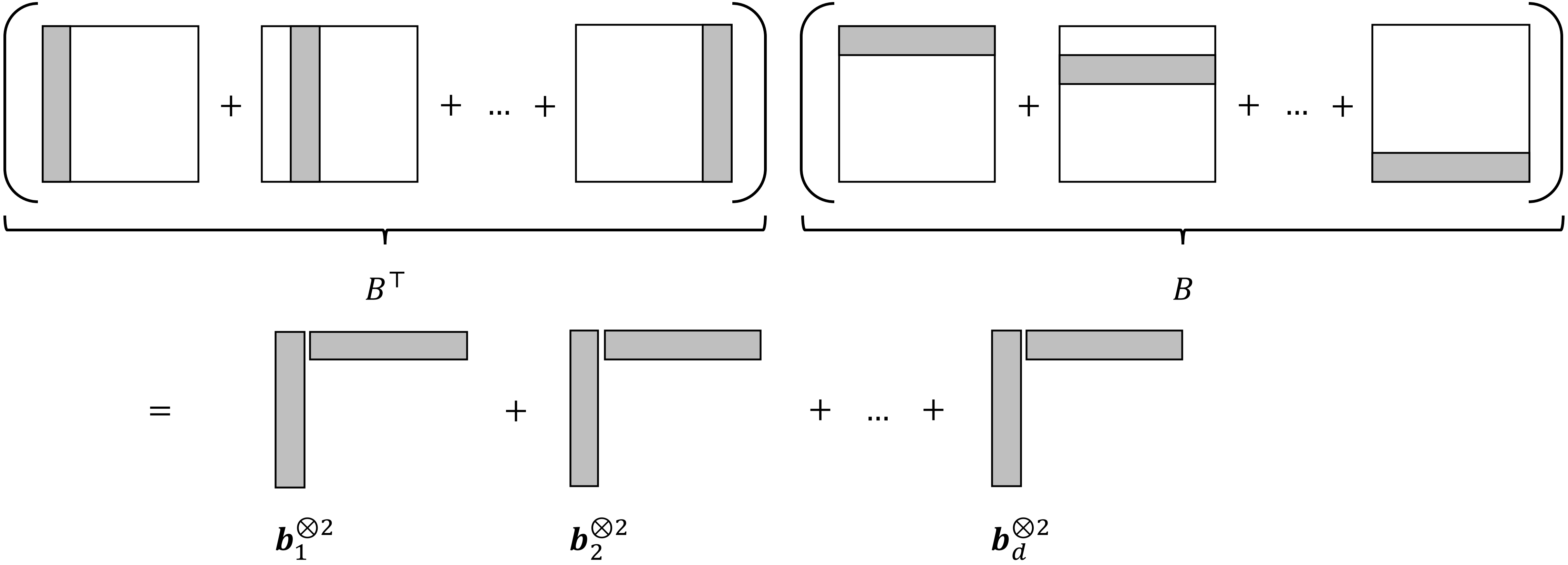
Proof of sufficiency

Row decomposition of a matrix



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

Rank-one decomposition of a matrix product

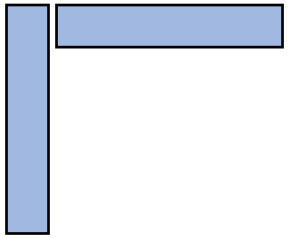


$$\boldsymbol{v}^{\otimes 2} = \boldsymbol{v}\boldsymbol{v}^\top$$

Deriving a key identity

$B_0^\top B_0 =$  $+$  $+ \ \dots \ +$ 

$(B_0^\top \boldsymbol{e}_1)^{\otimes 2}$ $\quad (B_0^\top \boldsymbol{e}_2)^{\otimes 2}$ $\quad (B_0^\top \boldsymbol{e}_d)^{\otimes 2}$

$B_k^\top B_k =$  $+$  $+ \ \dots \ +$ 

$\left(B_k^\top \boldsymbol{e}_1\right)^{\otimes 2}$ $\quad (B_0^\top \boldsymbol{e}_2)^{\otimes 2}$ $\quad (B_0^\top \boldsymbol{e}_d)^{\otimes 2}$

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

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

Identifying source nodes

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

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

H

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

$\Rightarrow \Theta_k - \Theta_0$ 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.

Algorithm

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

Getting rid of source nodes

1. Form a vector space V from the already-recovered rows of H.
2. Project $\Theta_k - \Theta_0$ onto the orthogonal complement of V.
3. 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.

Ongoing work

Identifiability Guarantees for Causal Disentanglement from Soft Interventions.

Zhang, Squires, Greenewald, Srivastava, Shanmugam, Uhler

1. Faithfulness assumptions under which causal disentanglement models are identifiable from **soft interventions**, with **non-parametric latent SCM**.
2. Algorithmic approach using VAEs and discrepancy measures.
3. Applications to extrapolating from single-gene to double-gene interventions in Perturb-seq data.

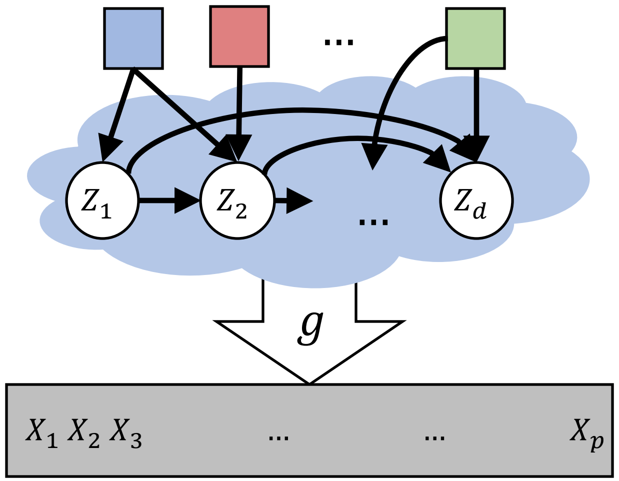
Unpaired Multi-Domain Causal Representation Learning.

Sturma, Squires, Drton, Uhler

1. Studies causal representation learning from **multiple unpaired modalities** (e.g. Perturb-seq, ATAC-seq).
2. Identifiability guarantees under restrictions on the mixing function weaker than in previous work.
3. Establishes statistical benefits of multiple modalities beyond just "more data".

Open Questions

Multi-node interventions



Statistically & computationally efficient algorithms

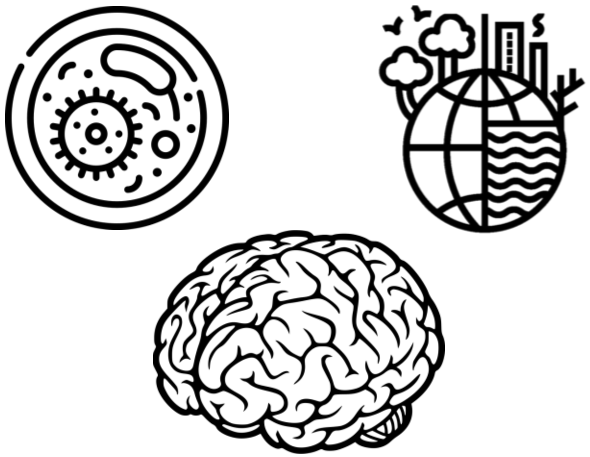
Score-based

Exact	Greedy	Gradient-based
GOBNILP (Cussens '20)	GES (Chickering '02)	NO-TEARS (Zheng '18)

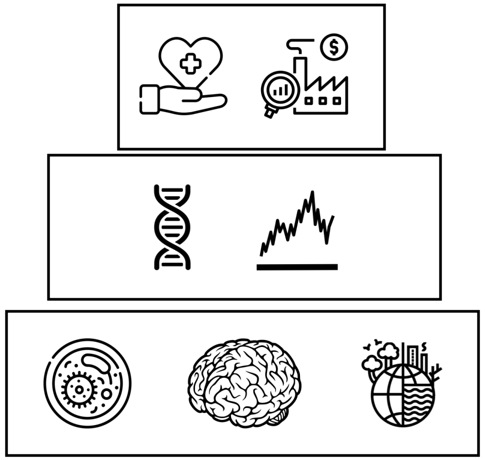
Better guarantees (in general) ⟷ More scalable (in general)

Computational-statistical trade-off

Applications



End-to-end causal reasoning



For more on causality

Causality lectures (with videos)

github.com/csquires/6.S091-causality

Causal structure learning review [1]

arxiv.org/abs/2206.01152

causaldag Python package

github.com/uhlerlab/causaldag

Simons causality bootcamp

simons.berkeley.edu/workshops/causality-boot-camp

[1] ***Causal structure learning: a combinatorial perspective.*** Squires and Uhler (JoFoCM 2022).