Causal Representation Learning

AAAI 2025 Tutorial

Room 118C – Pennsylvania Convention Center

Philadelphia, Pennsylvania, US

08:30am-12:30pm, February 26th, 2025

Burak Varıcı

CMU

bvarici@andrew.cmu.edu

Emre Acartürk

RPI

acarte@rpi.edu

Karthikeyan Shanmugam

Google DeepMind

karthikeyanvs@google.com

Ali Tajer

RPI

tajer@ecse.rpi.edu

Some Housekeeping

slides available at:

https://www.isg-rpi.com/aaai-25

Topic	Presenter	Duration	Time
Introduction and Background	Ali Tajer	45 minutes	08:30-09:15
Foundations of Interventional CRL	Burak Varıcı	75 minutes	09:15-10:30
Coffee Break		30 minutes	10:30-11:00
Empirical Aspects	Emre Acartürk	40 minutes	11:00-11:40
Other Recent Developments	Emre Acartürk	35 minutes	11:40-12:15
Closing Thoughts	Burak Varıcı	15 minutes	12:15-12:30

Part I – Introduction and Background

08:30-09:15 (45 minutes)

- What is CRL?
- Why is it important?
- How to formalize it?
- What does causality signify?
- Necessary background in causal learning
- Theoretical questions in CRL

Ali Tajer

What is CRL?

Unpacking …

Data-generation process

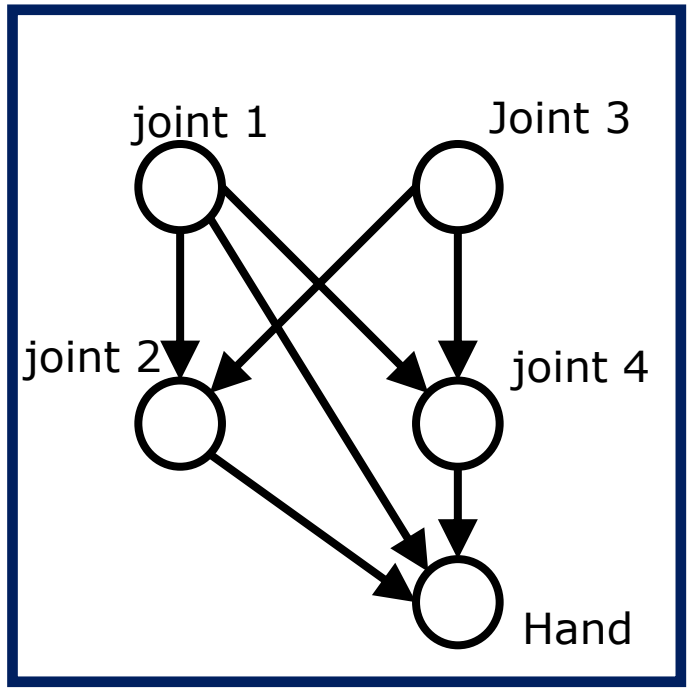
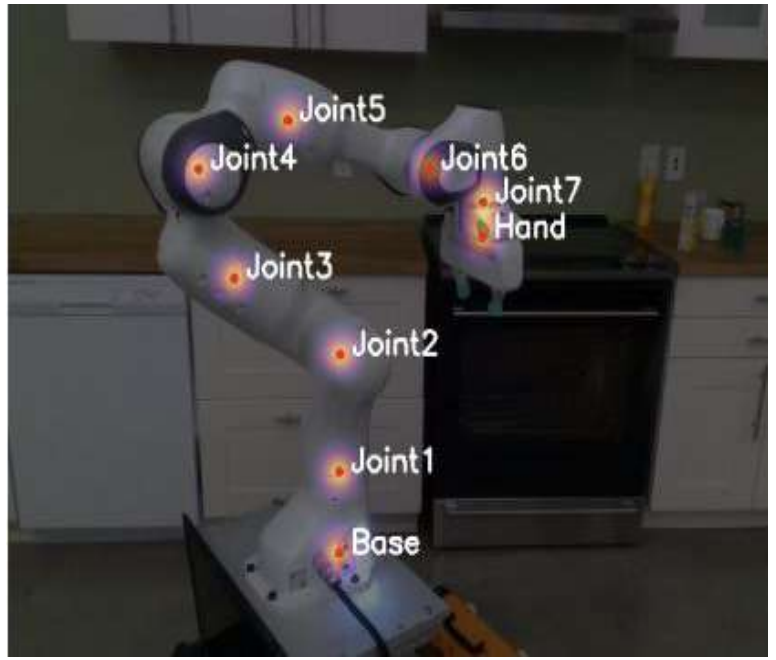


Image of the robot



Disentangle generating factors

Positions of various robot's joints

Unpacking …

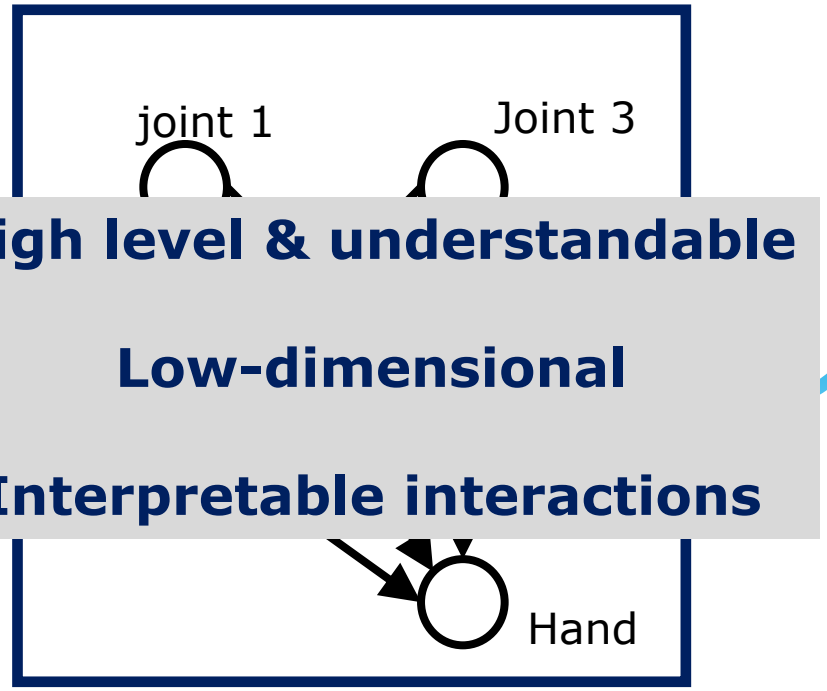
Data-generation process

Image of the robot

Low level & complex

High-dimensional

Uninterpretable



High level & understandable

Low-dimensional

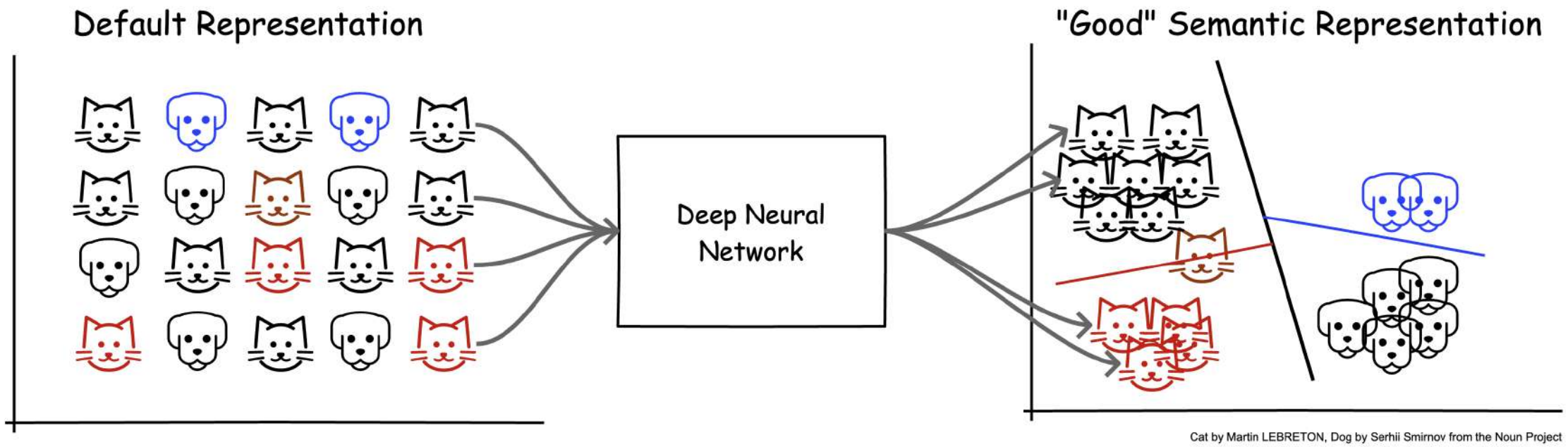
Interpretable interactions

Disentangle generating factors

Positions of various robot’s joints

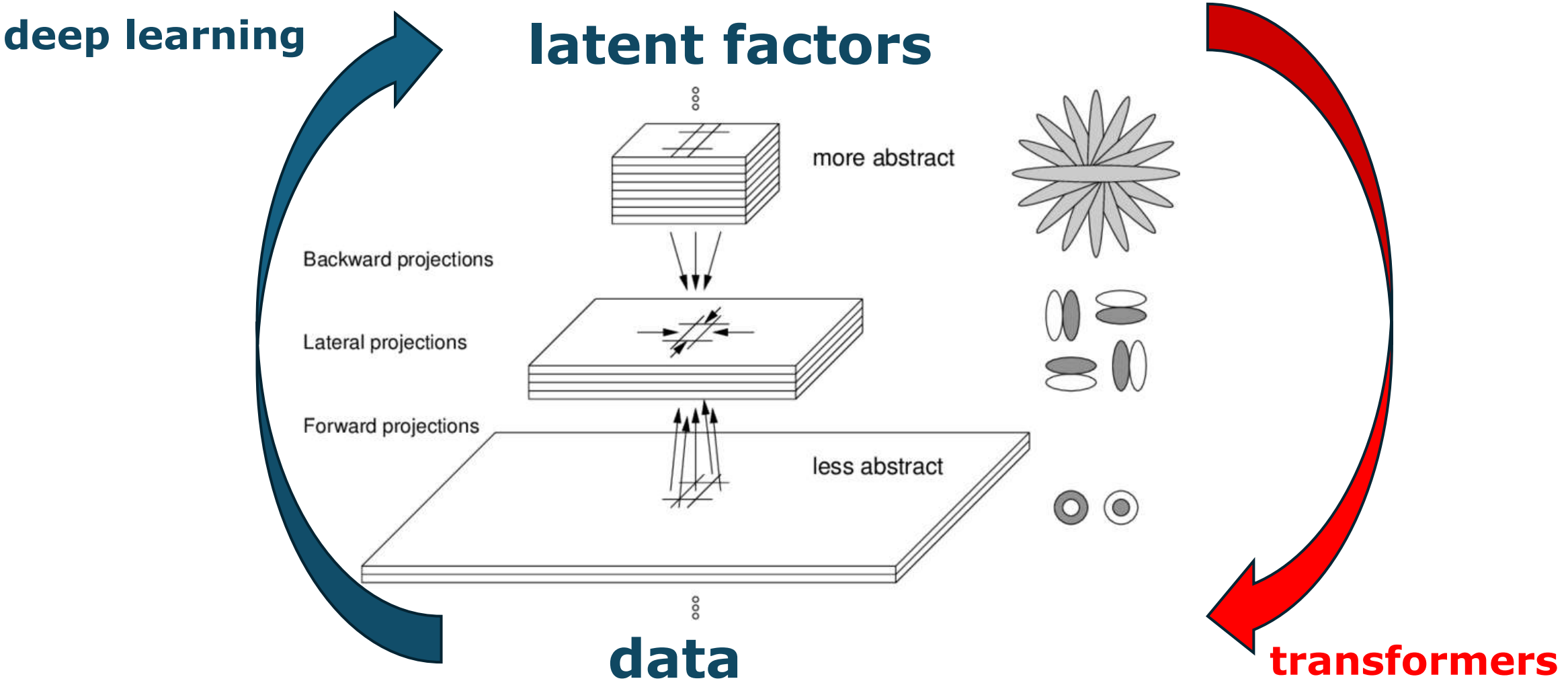
Representation Learning

data: often high-dimensional, complex, unexplainable structure



representation learning: disentangle high-level meaningful latent factors

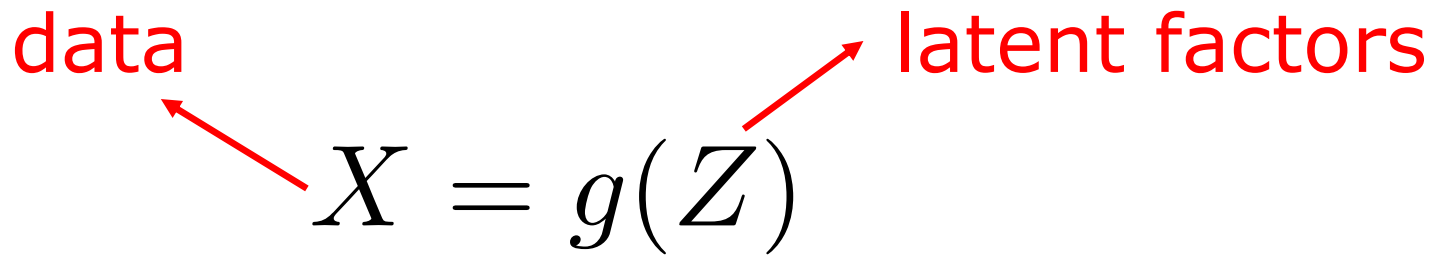
Success of Representation Learning



Formalize it!

definition:

transform raw data into useful, structured latent representations



objectives:

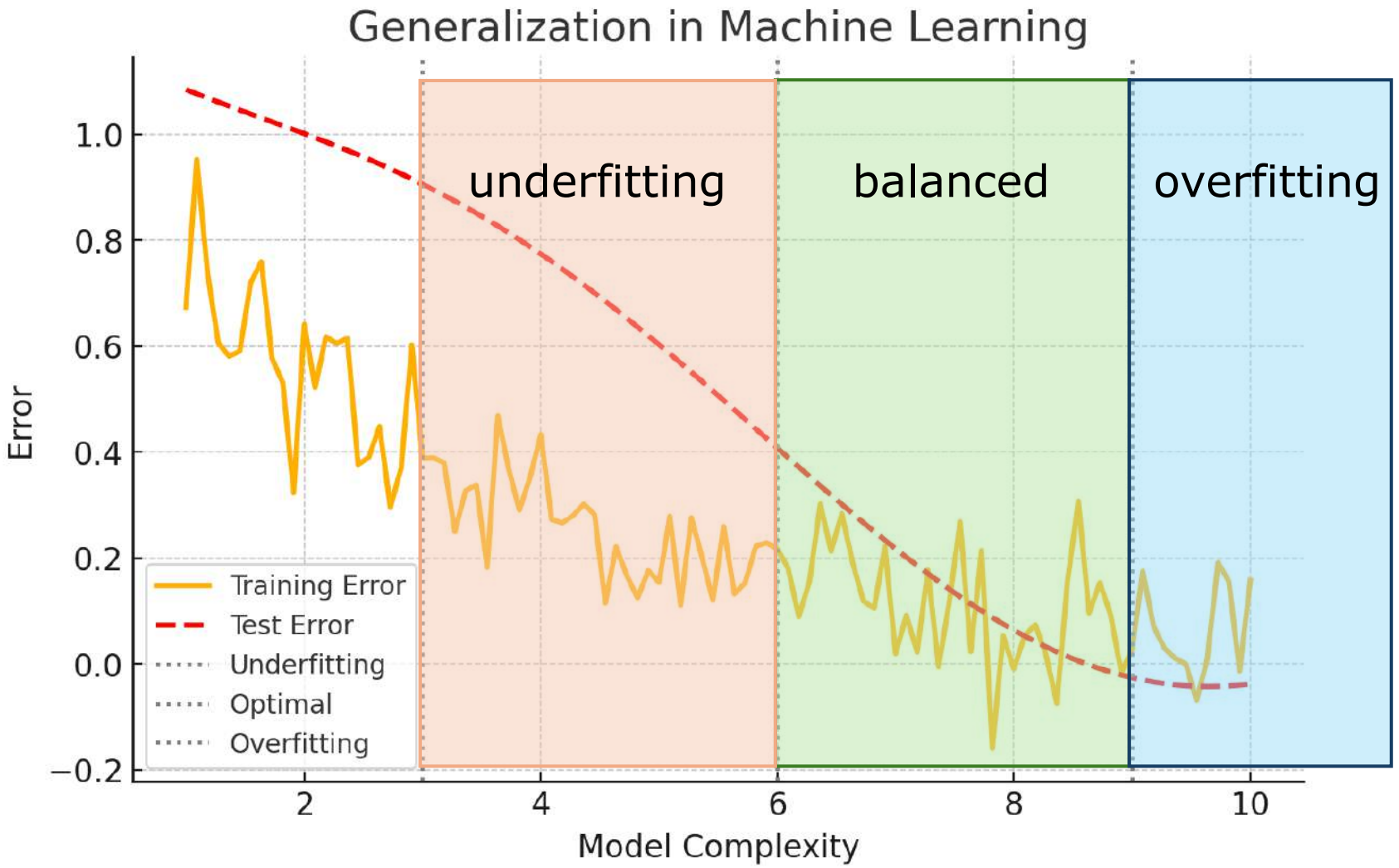
Disentanglement: Separate factors of variation in data

Invariance: robust to transformations

Interpretability: explanation of learned features.

A Key Advantage: Generalization

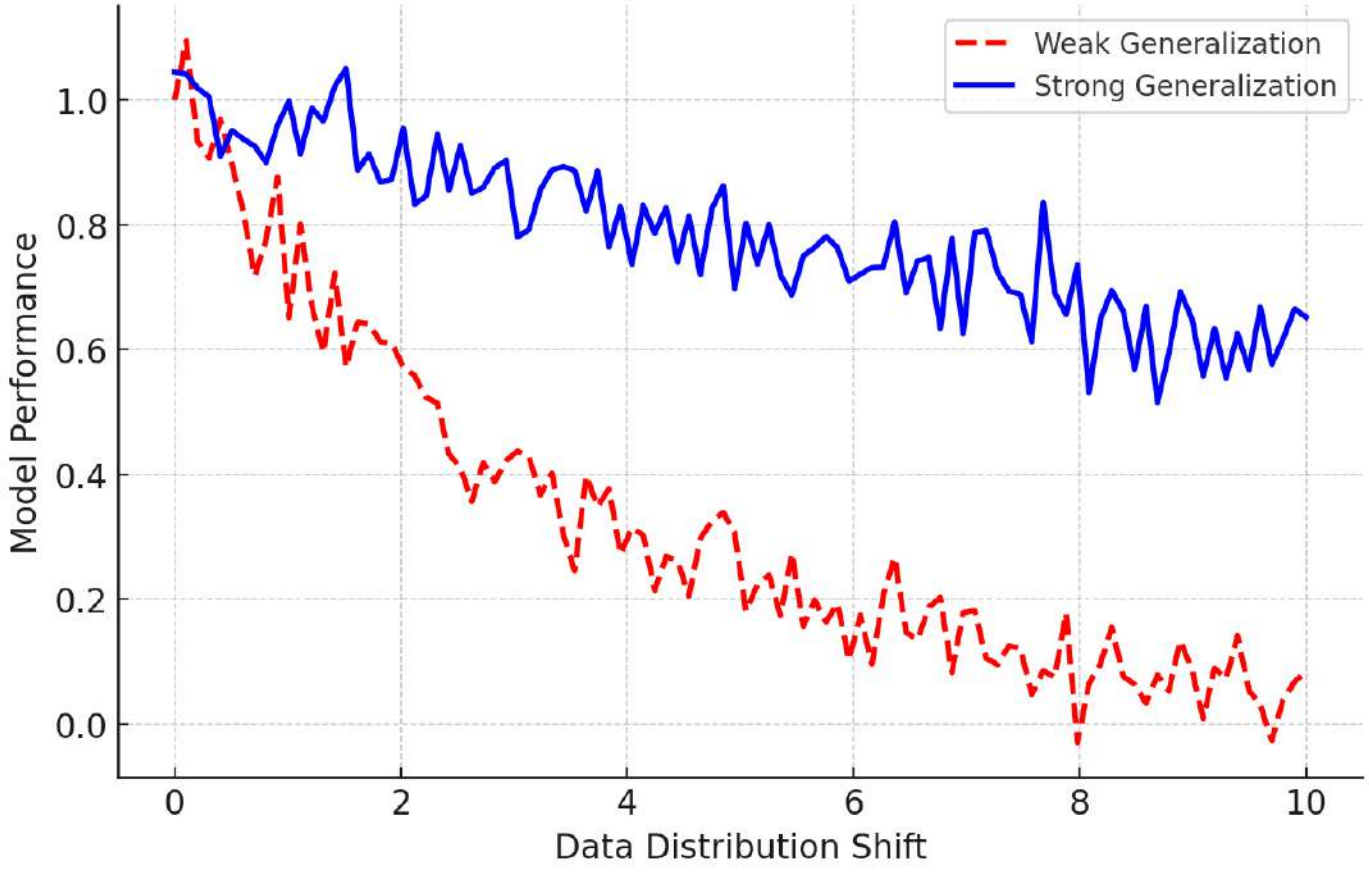
Perform well on new (unseen) data



Strong Generalization … Still a Challenge

Weak: perform well on new data from the same distribution

Strong: stable performance even under model shifts



Strong Generalization

training environment

testing environment

A Necessary Step

weak
generalization

Strong
generalization

“good enough”
representation

“ground truth”
representation

one well-defined representation, or a class of representations

A Necessary Step

weak
generalization

Strong
generalization

“good enough”
representation

“ground truth”
representation

advantages of ground truth representation

- out-of-distribution robustness
- reusable models
- counterfactual system analysis

Causal Representation Learning

Toward Causal Representation Learning

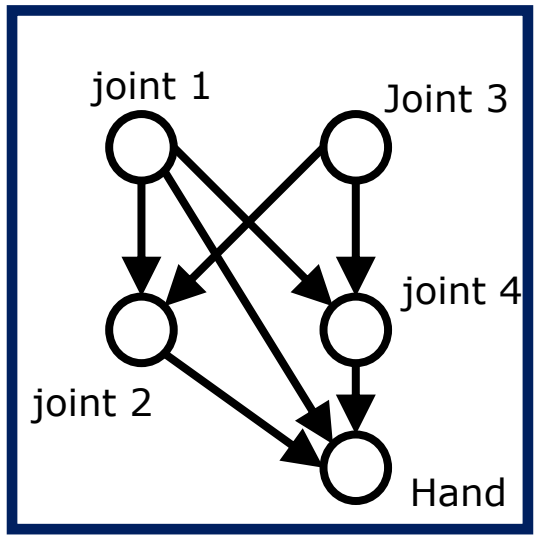
This article reviews fundamental concepts of causal inference and relates them to crucial open problems of machine learning, including transfer learning and generalization, thereby assaying how causality can contribute to modern machine learning research.

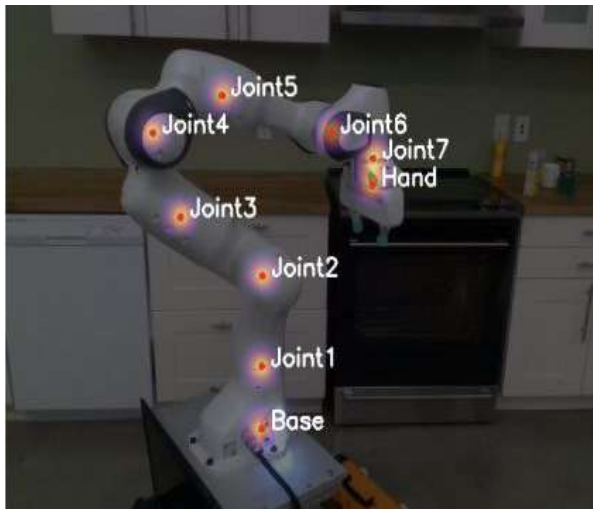
By Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio

CRL: Two Premises

1- There is a (class of) ground truth latent model(s)

2- Latent variables can have statistical interactions





Disentangle generating factors

Variable Interactions

robot's motions via a set of kinetic equations

$$\frac{dx}{dt} = v\cos(\theta)$$

$$\frac{dy}{dt} = v\sin(\theta)$$

$$\frac{d\theta}{dt} = \omega$$

$$v = \frac{v_L + v_R}{2}$$

$$\omega = \frac{v_R - v_L}{d}$$

where:

- v is the linear velocity of the robot (m/s),
- ω is the angular velocity of the robot (rad/s),
- v_L and v_R are the linear velocities of the left and right wheels (m/s),
- d is the distance between the wheels (m).

Variable Interactions

robot's motions via a set of kinetic equations

$$\frac{dx}{dt} = v\cos(\theta)$$

The gold standard to model variable interactions

(complex) sets of coupled equations

$$\frac{d\theta}{dt} = \omega$$

$$\omega = \frac{v_R - v_L}{d}$$

where:

- v is the linear velocity of the robot (m/s),
- ω is the angular velocity of the robot (rad/s),
- v_L and v_R are the linear velocities of the left and right wheels (m/s),
- d is the distance between the wheels (m).

ML is Insufficient

coupled equations: precisely trace how changing θ causes changes in x

ML algorithms: estimate the statical correction of θ and x

+ data-driven
- cannot trace cause-effect relationships

$$\frac{dx}{dt} = v\cos(\theta)$$

$$\frac{dy}{dt} = v\sin(\theta)$$

$$\frac{d\theta}{dt} = \omega$$

+ traces cause-effect relationships
- critically **model-based**
- models inaccessible/unavailable for latent spaces)

ML is Insufficient

coupled equations: precisely trace how changing θ causes changes in x

ML algorithms: estimate the statical correction of θ and x

causal inference lies in between:

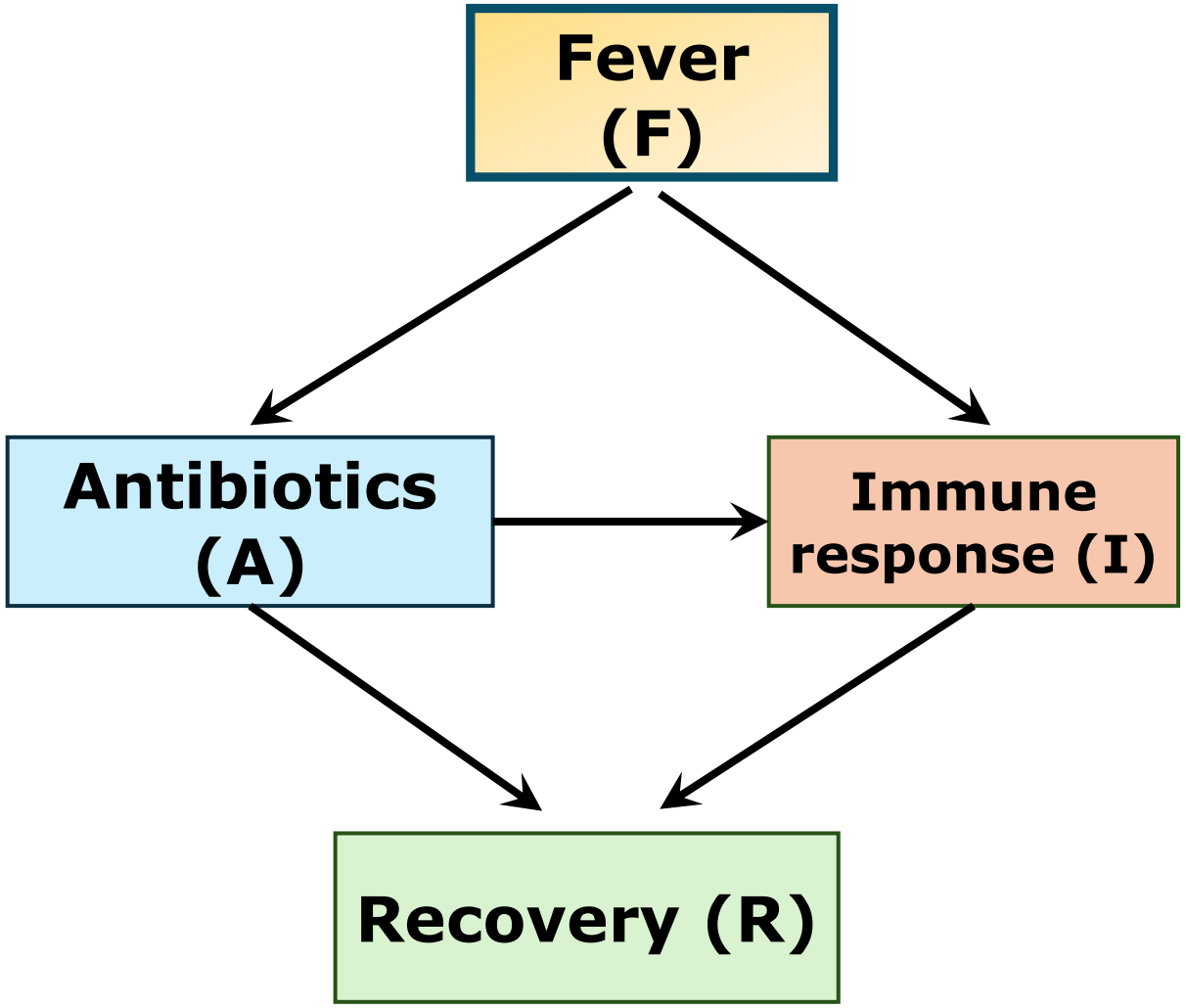
Data-driven + captures causal effects

$$\frac{dx}{dt}$$

$$\frac{d\theta}{dt} = \omega$$

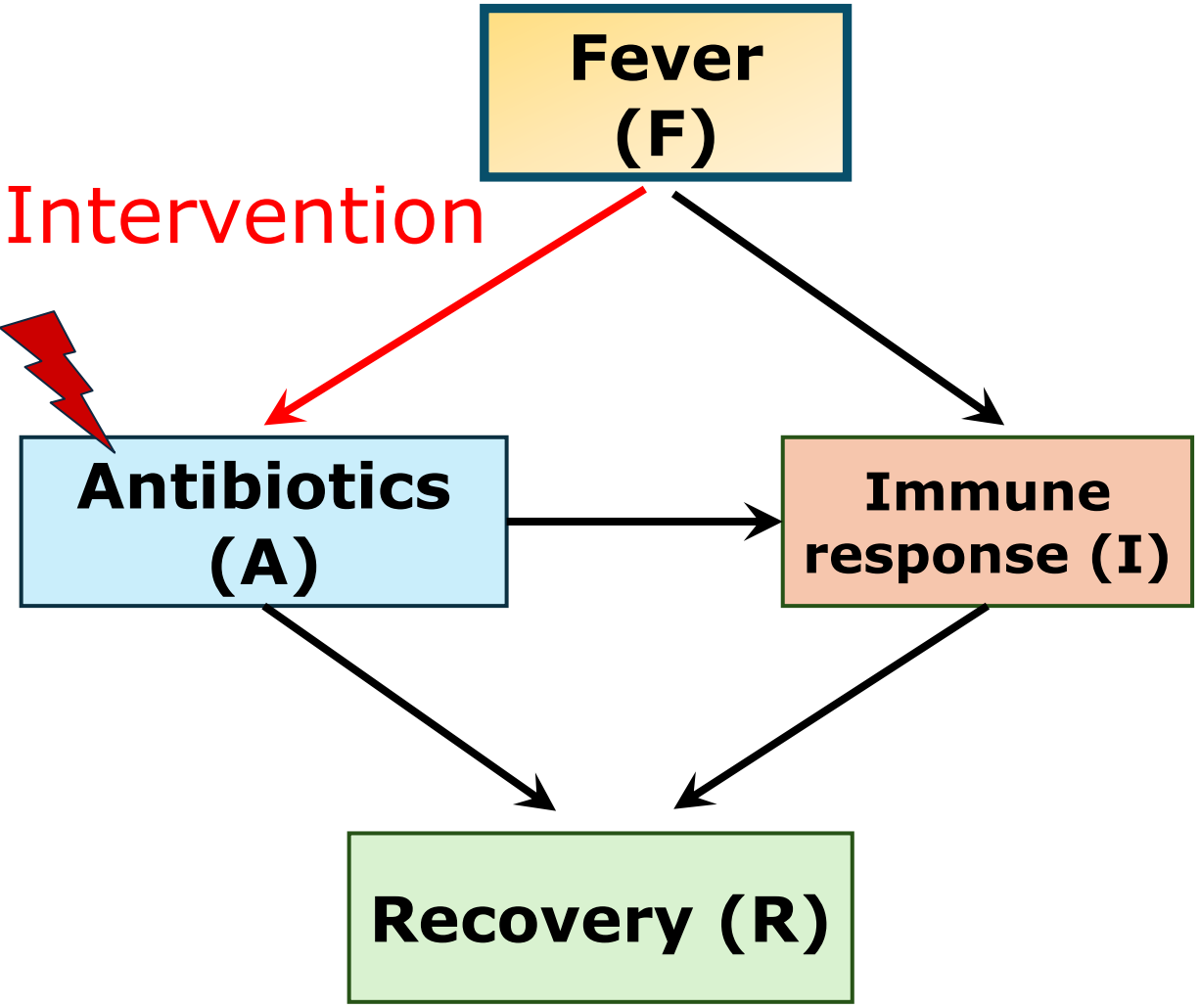
+ traces cause-effect relationships
- critically **model-based**
- models inaccessible/unavailable for latent spaces)

Cause – Effect Relationships



Antibiotics Prescription		
	$P(A = 1 \mid F = 1)$	$P(A = 1 \mid F = 0)$
Observational	0.80	0.20

Cause – Effect Relationships



Antibiotics Prescription		
	$P(A=1 \mid F=1)$	$P(A=1 \mid F=0)$
Observational	0.80	0.20

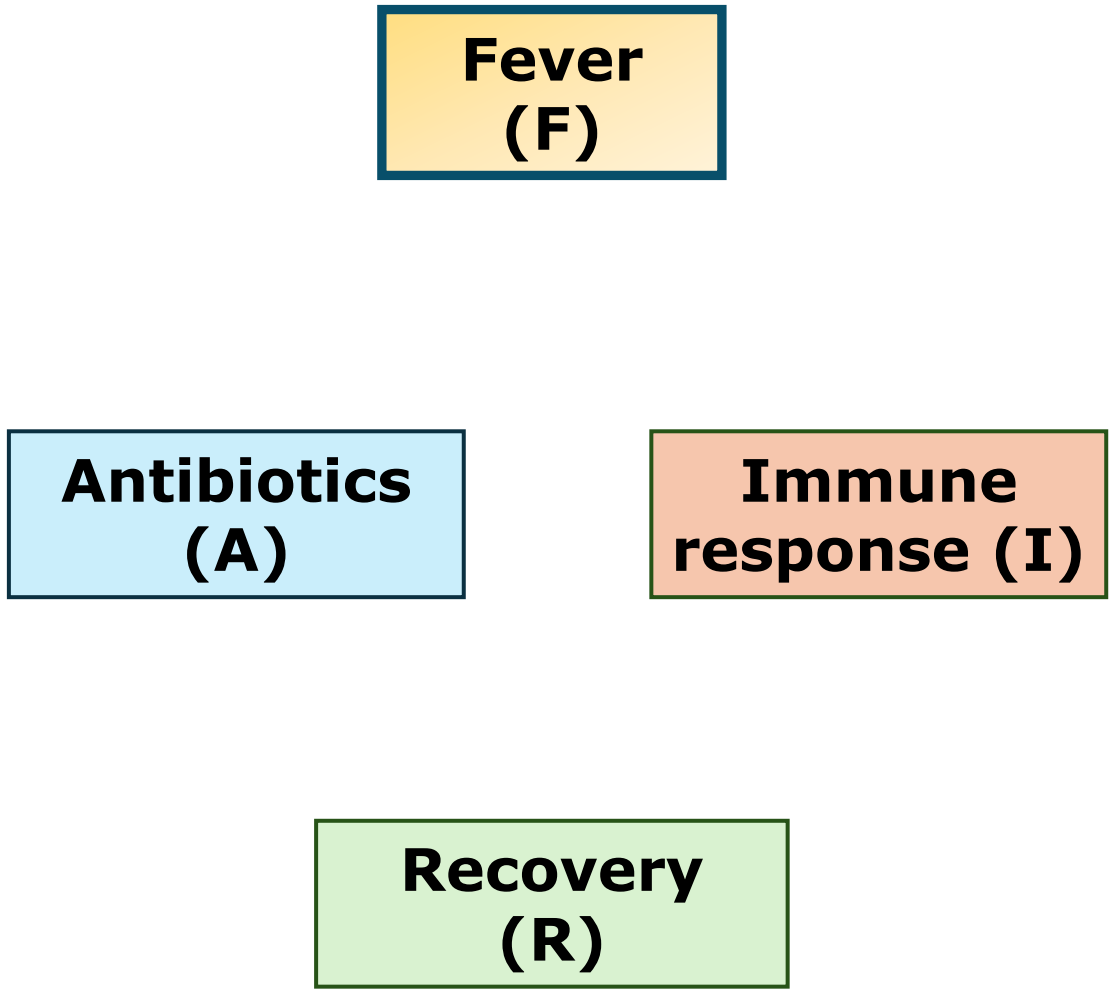
do intervention

$$P(A=1)=1$$

Stochastic Intervention		
	$P(A=1 \mid F=1)$	$P(A=1 \mid F=0)$
Interventional	0.95	0.50

Learning Cause – Effect Relationships

- use observation and interventional data
- determine the topology (edges)
- cause-effect direction (edge orientation)
- predict causal effects



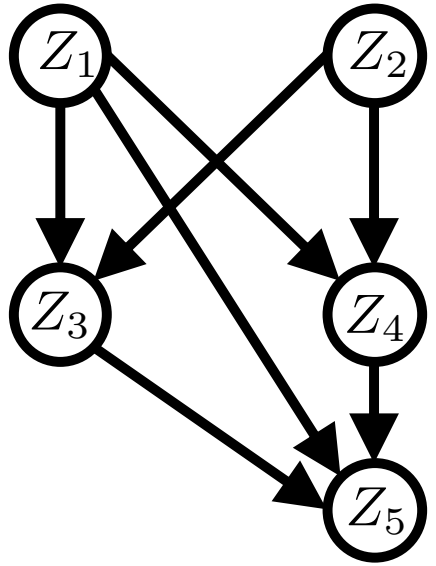
Structural Causal Model

- RVs: $\mathbf{Z} = \{Z_1, \ldots, Z_n\}$
- exogenous noise variables $\epsilon = (\epsilon_1, \ldots, \epsilon_n)$
- **probability model:** $\mathbb{P}_\epsilon$ a probability measure on ϵ
- **structural functions**

$$Z_i = f_i(\mathbf{Z}_{\mathrm{pa}(i)}, \epsilon_i) \quad \Rightarrow \quad F = \{f_i : i \in [n]\}$$

- directed acyclic graph (DAG): $\mathcal{G}$

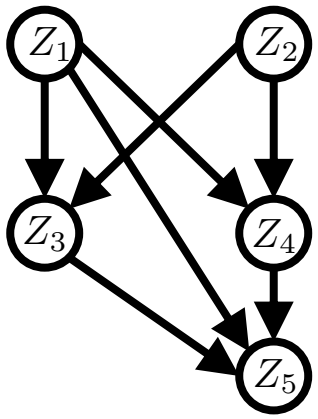
$$p(Z) = \prod_{i=1}^{n} p(Z_i \mid Z_{\mathrm{pa}(i)})$$



a structural causal model is specified by $(\mathbf{Z},\ \epsilon,\ \mathbb{P}_\epsilon,\ F)$

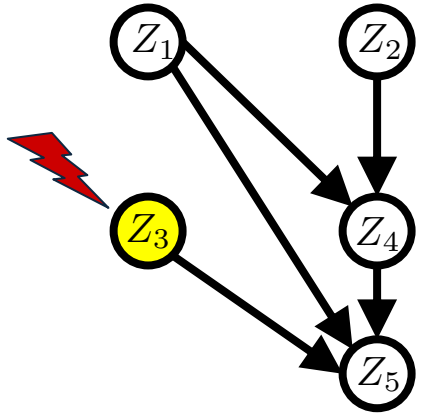
Interventions: Deliberate Manipulations

observational



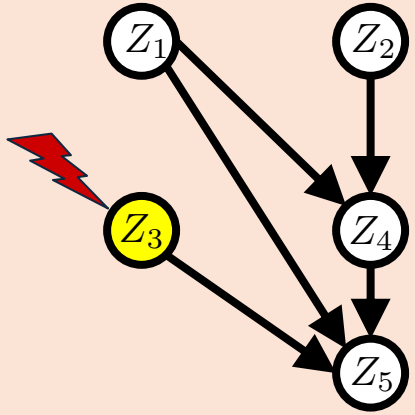
$$p(Z_3|Z_1, Z_2)$$

do



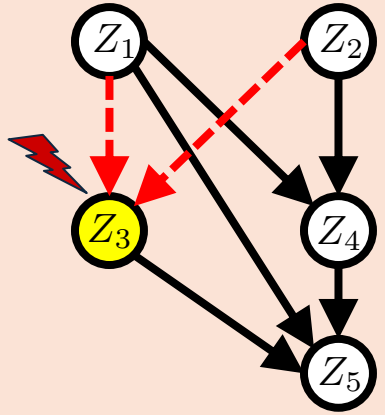
$$\begin{cases} 1 & \text{for } Z_3 = z^* \\ 0 & \text{for } Z_3 \neq z^* \end{cases}$$

hard (perfect)



$$q(Z_3)$$

soft (imperfect)



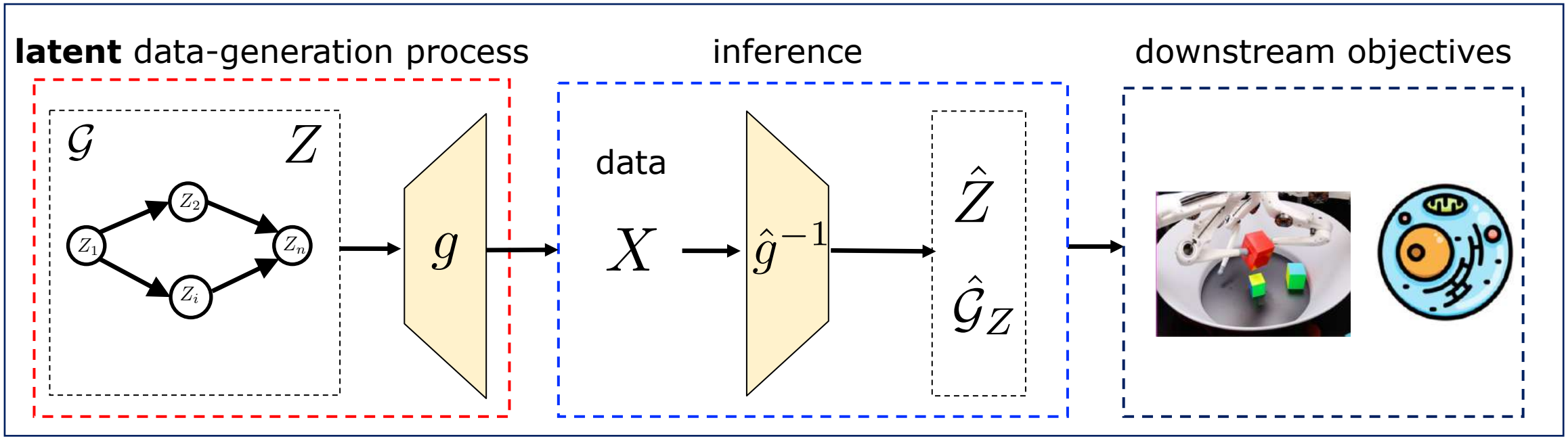
$$q(Z_3|Z_1, Z_2)$$

intervention on Z_i: $p_i(z_i \mid \mathrm{pa}(z_i)) \rightarrow q_i(z_i \mid \mathrm{pa}(z_i))$

functionally: $Z_i = f_i(\mathrm{pa}(Z_i), \epsilon_i) \rightarrow \tilde{f}_i(\mathrm{pa}(Z_i), \tilde{\epsilon}_i)$

Formalizing CRL

Causal Representation Learning

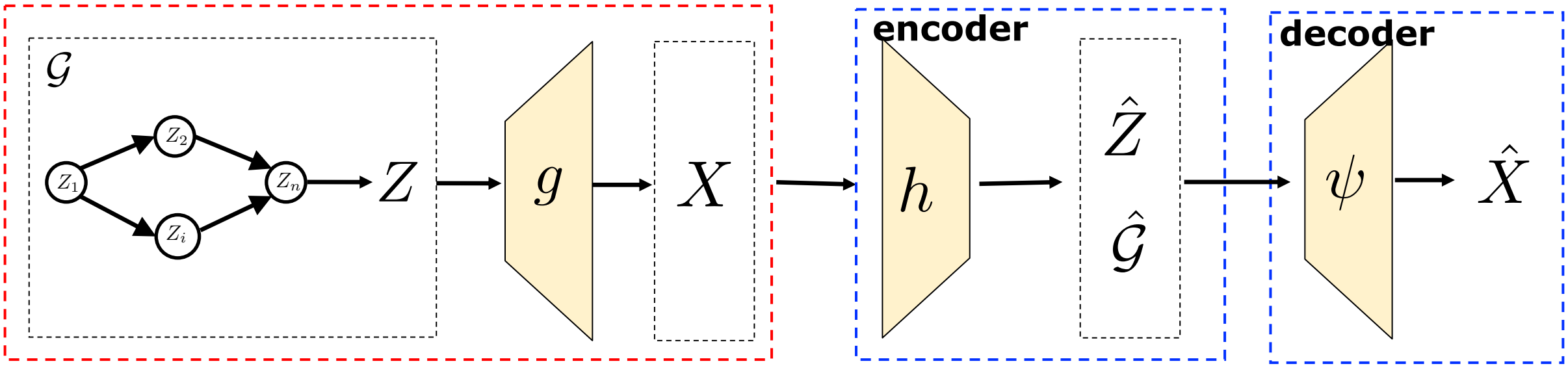


CRL objective

Use data X to learn ground truth representation:

variables Z and their inherent interaction structure $\mathcal{G}$

Common Strategy



Find encoder-decoder pair (h, ψ) that allows perfect recovery:

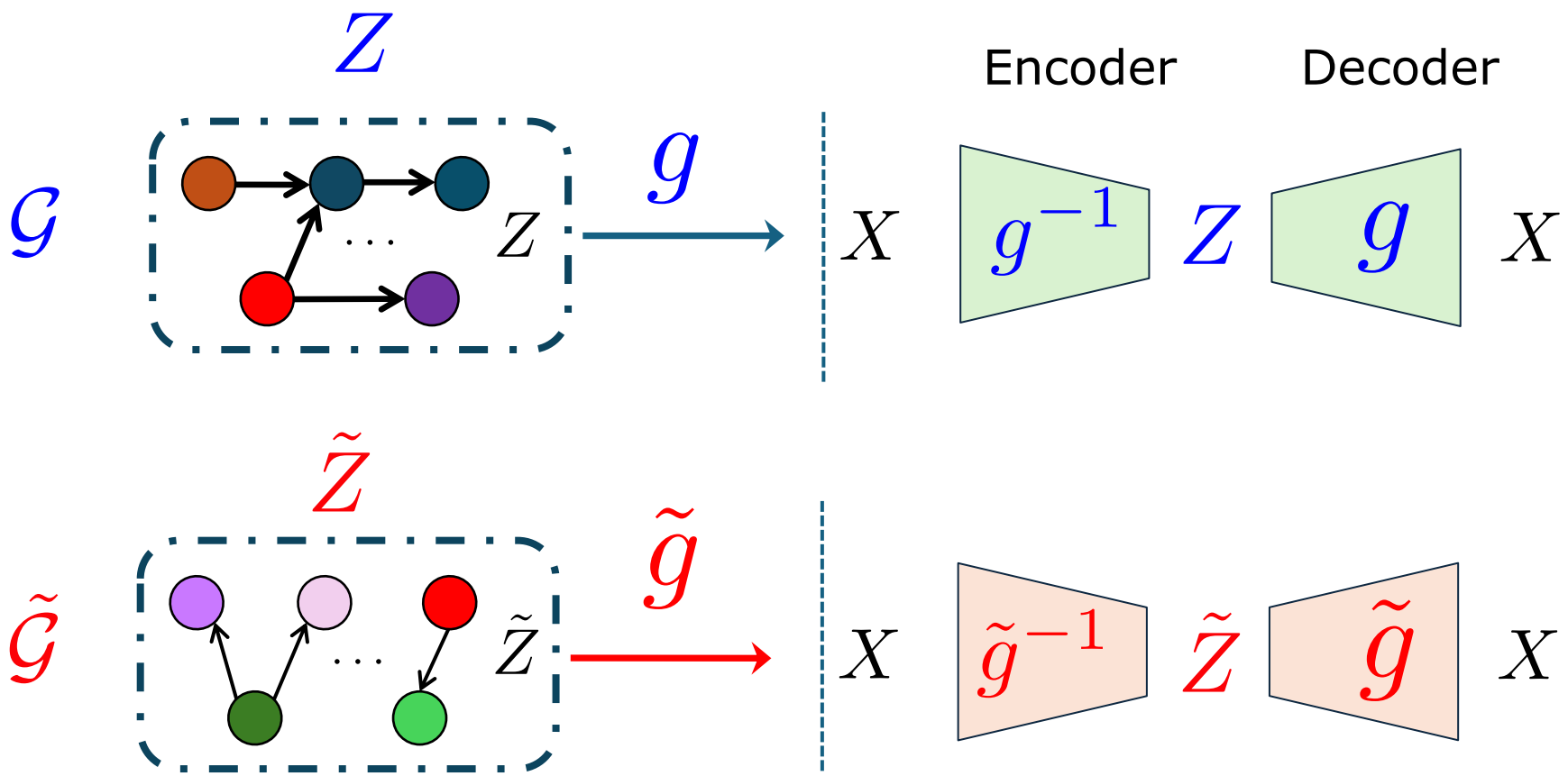
- **variables:** $\hat{X} = X$
- **graph:** pdf of $\hat{Z}$ factorizes wrt a DAG $\mathcal{G}$

Key tool: leverage invariance properties
Goal: unique solution (up to a specified equiv. class)

Why is CRL Challenging?

Enforcing reconstruction? Required but not enough!

$$\tilde{g}(\tilde{Z}) = g(Z)$$



Enforcing Reconstruction is not Enough – Linear

- **Special case**: **linear** transformation and **independent** latents (**linear ICA**)

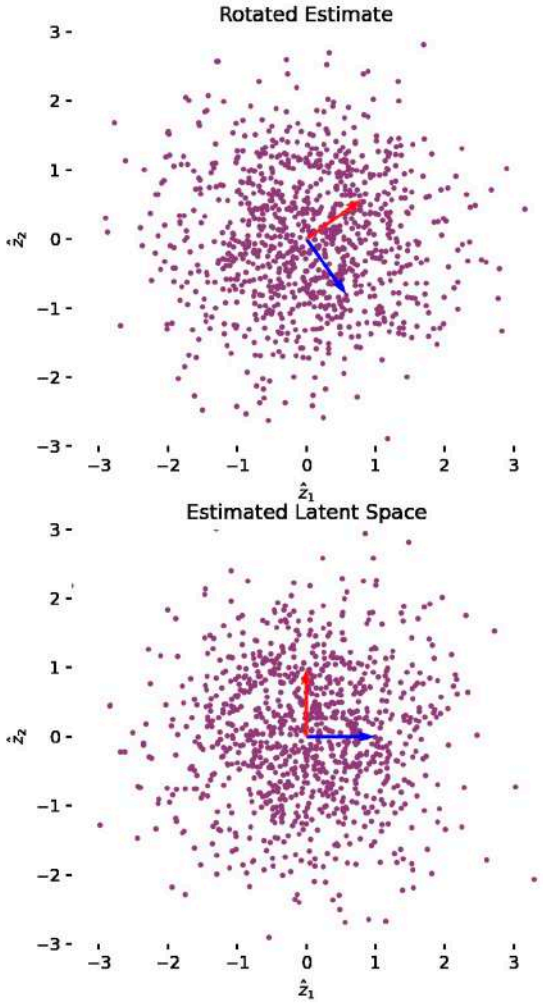
$$X = \mathbf{G} \cdot Z \ , \qquad p(Z) = \prod_i p(Z_i)$$

- Is linear ICA solution set unique? **no** – e.g., Gaussians are rotation invariant

$$\{(\hat{Z}, \hat{\mathbf{G}}) \ : \ X = \hat{\mathbf{G}} \cdot \hat{Z} \ \text{ and } \ \hat{Z}_i \perp\!\!\!\perp \hat{Z}_j \ \ \forall i, j\}$$

- What can be guaranteed?
  If at most one Z_i is Gaussian: ID up to permutation & scaling

$$\hat{Z} = \mathbf{P}_\sigma \cdot \mathbf{D} \cdot Z$$



Enforcing Reconstruction is not Enough – Linear

- **Special case**: **linear** transformation and **independent** latents (**linear ICA**)

$$X = \mathbf{G} \cdot Z \ , \qquad p(Z) = \prod_i p(Z_i)$$

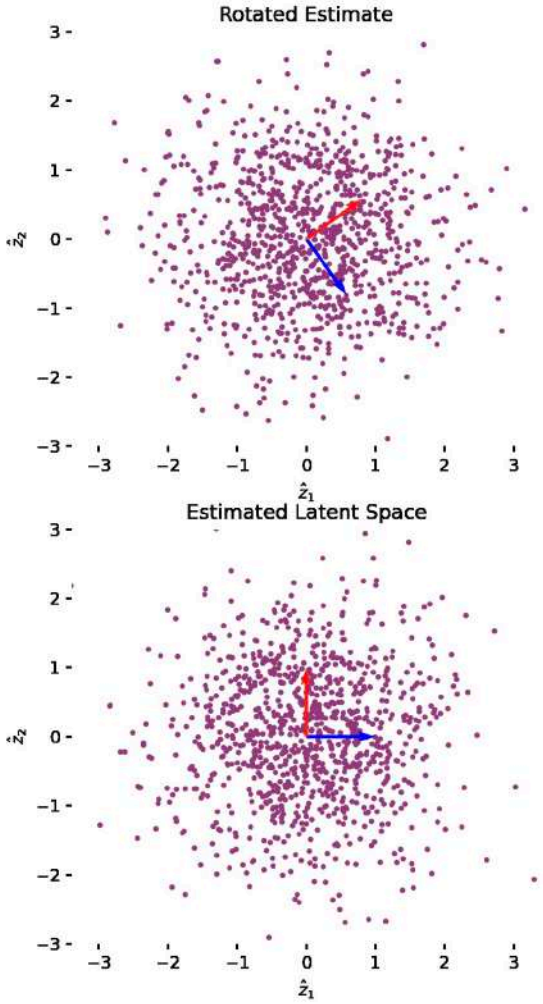
- Is linear ICA solution set unique? **no** – e.g., Gaussians are rotation invariant

$$\{(\hat{Z}, \hat{\mathbf{G}}) \ : \ X = \hat{\mathbf{G}} \cdot \hat{Z} \ \text{ and } \ \hat{Z}_i \perp\!\!\!\perp \hat{Z}_j \ \ \forall i, j\}$$

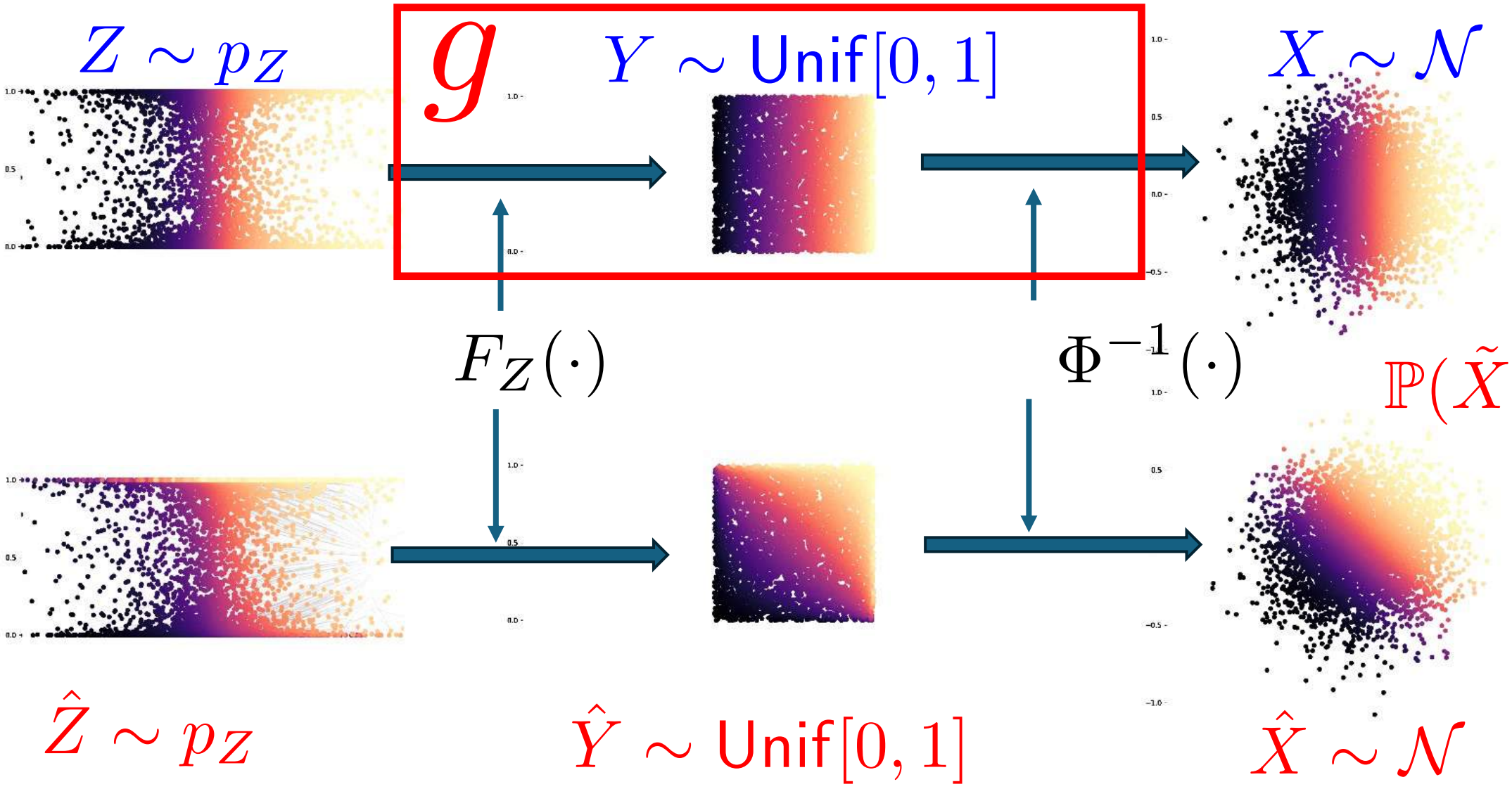
some ambiguity is clearly unavoidable

permutation ≡ relabeling

scaling ≡ units of measurement

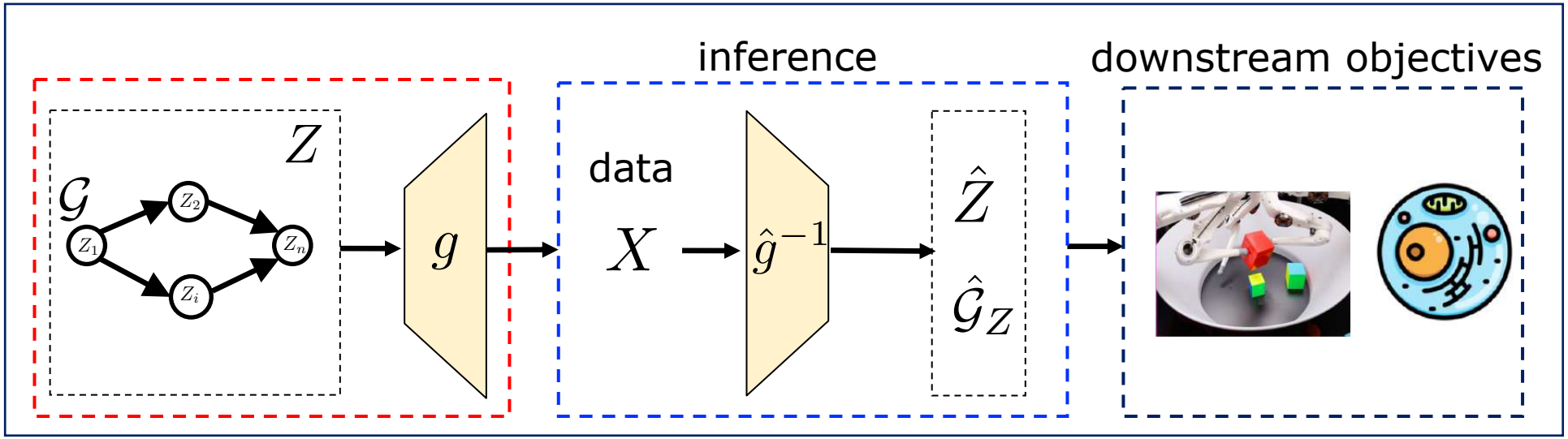


Enforcing Reconstruction is not Enough – Nonlinear



Slide credit: Jason Hartford & Dhanya Shridhar

CRL Objectives



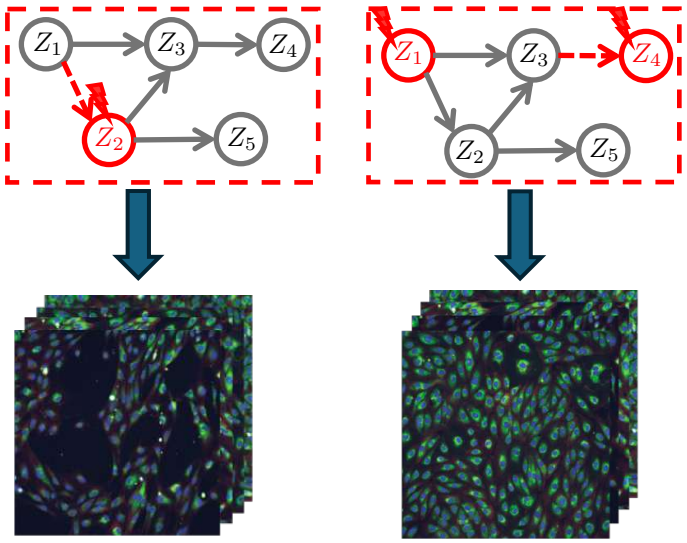
two symbiotic questions:

- **identifiability (algorithm-agnostic)**.
  - determine whether a unique model (up to tolerable ambiguities) can explain the data.
  - determine the necessary and sufficient conditions under which $\mathcal{G}$ and Z can be recovered.

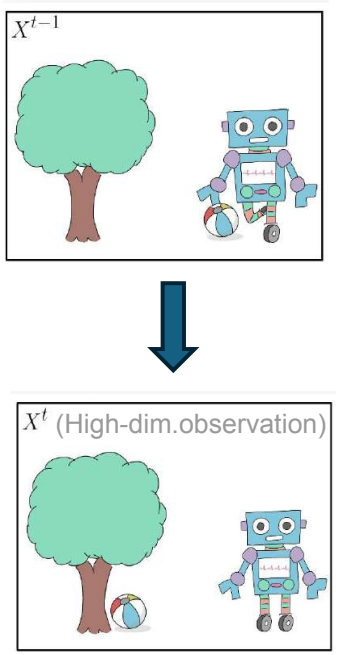
- **achievability (algorithmic)**.
  - it complements identifiability by designing inference algorithms that estimate $\mathcal{G}$ and Z
  - it hinges on the information available about the data, transformation, and the latent model.

CRL Taxonomy (non-exhaustive)

Interventional



Temporal Data

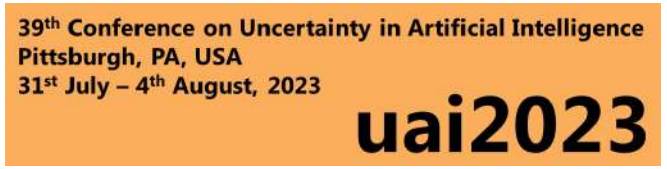


Multiview

Different ways of inducing statistical diversity

Growing resources: workshops, tutorials, & surveys

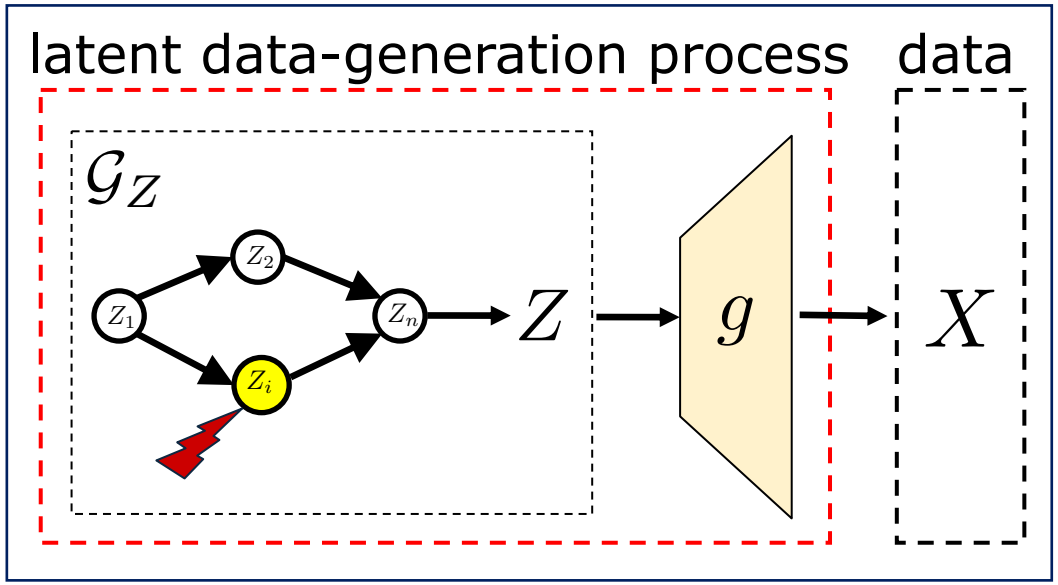
- **UAI** 2022 Workshop, **UAI** 2023 Tutorial
- **NeurIPS** 2023 Workshop, **NeurIPS** 2024 Workshop
- A Survey on Causal Generative Modeling (TMLR 2024)



Our focus: developments in the last ~2 years

Emphasis: interventional aspects

Interventional CRL



Intervention models
(do, hard, soft)

transformation g models
(parametric and non-parametric)

causal models
(parametric and non-parametric)

Part II - Foundations of Interventional CRL

09:15-10:30 (75 minutes)

- Why interventional CRL?
- Deterministic interventions
- Stochastic interventions and score functions
- Score-based CRL
  - General transforms
  - Linear transforms
  - Multi-node interventions

Burak Varıcı

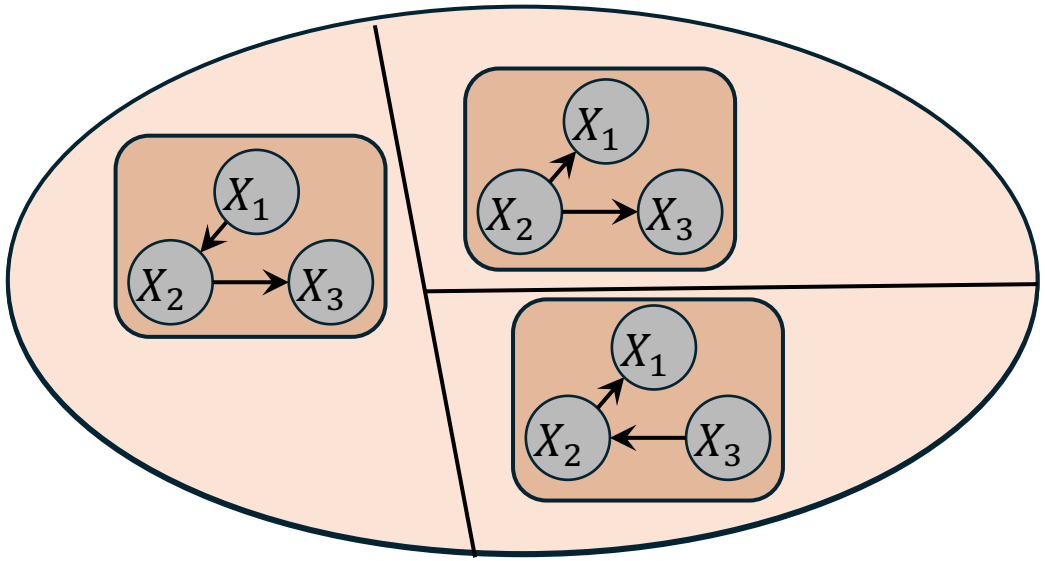
Carnegie Mellon University

Interventional CRL

Identifiability & Algorithm Design

Leveraging Interventions

- Natural choice for standard causal discovery/inference
- Each intervention introduces new constraints
- We can learn from them (**density ratio**, contrastive learning etc.)



All three graph can generate same distribution

+ intervention on X_2

All three are distinguishable!

Figure credit: Chandler Squires

Why Interventional CRL?

Distribution level information: almost* unsupervised

$$z \sim p_Z \ , \quad \tilde{z} \sim \tilde{p}_X$$
$$x \sim p_X \ , \quad \tilde{x} \sim \tilde{p}_X$$

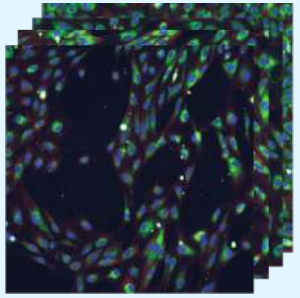
distribution level

$$z \sim p_Z \quad , \quad \tilde{z} \leftarrow \text{interv}(z)$$

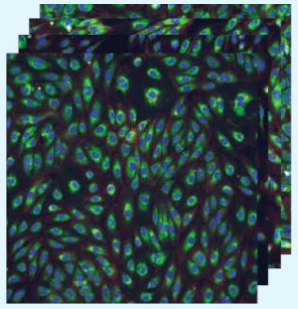
sample level

Active observations

Intervene and observe the system after some fixed time



$x \sim p_X$



$\tilde{x} \sim \tilde{p}_X$

Passive observations

- Given multiple datasets / domains
- Sparse distributions shifts ≡ latent interventions

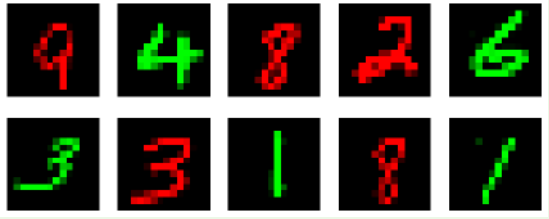
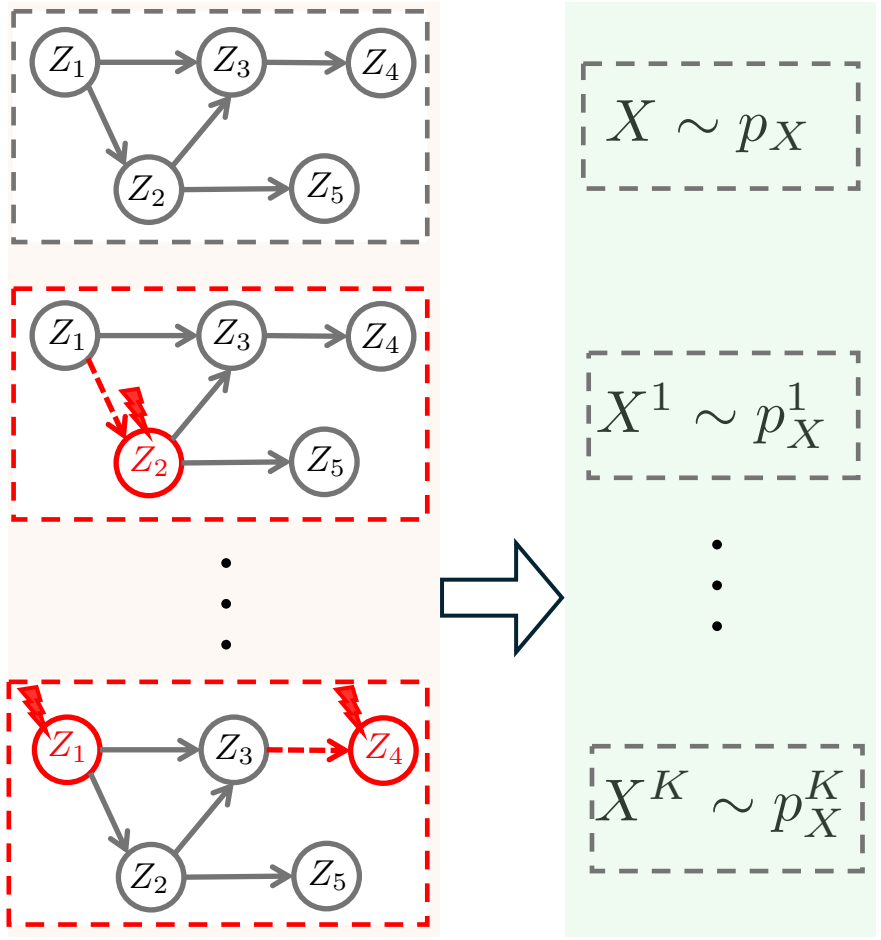


Figure credit: Jason Hartford

Formulation



An interventional environment $\mathcal{E}^m$:

- Intervention I^m acting on targets $T(I^m) \subset [n]$
- For $i \in T(I^m)$: $p_i(z_i \mid \mathrm{pa}(z_i)) \to q_i^m(z_i \mid \mathrm{pa}(z_i))$

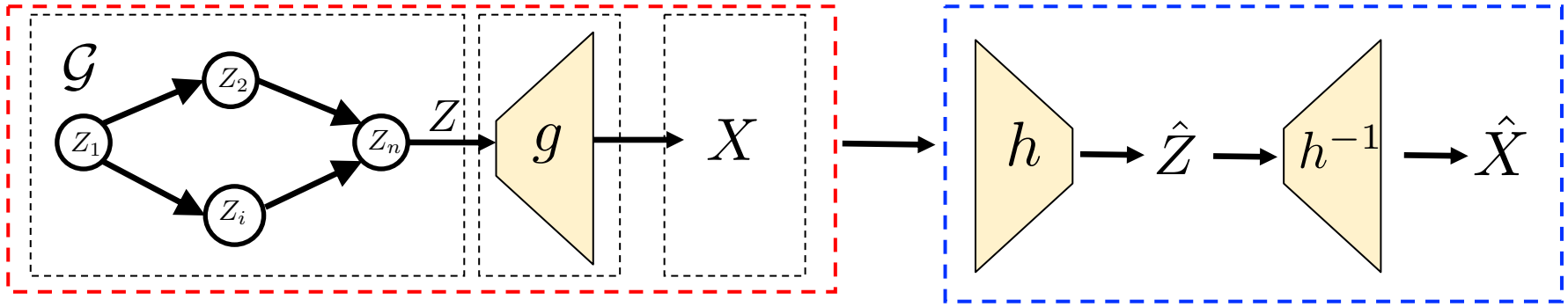
$$p_Z^m(z) = \prod_{i \in T(I^m)} q_i^m(z_i \mid \mathrm{pa}(z_i)) \prod_{i \notin T(I^m)} p_i(z_i \mid \mathrm{pa}(z_i))$$

- Shared transformation g: $X^m = g(Z^m)$ for all m
- **Observe samples from** P_X^m

We'll have have 1 or 2 int. mechanisms, simplify

$q_i := q_i^m$ and $I^m := T(I^m) = \{m\}$

Warm-up: Reconstruction Constraint



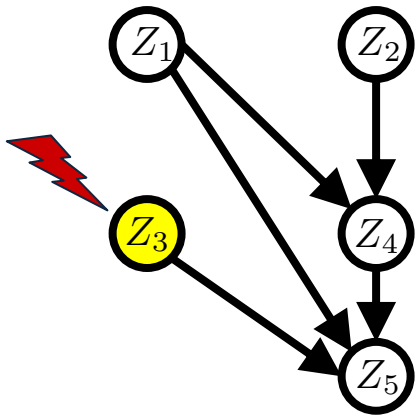
Suppose **polynomial** transform

- Use only *observational* data
- Constraint: perfect reconstruction + decoder is polynomial

Affine transform: $\hat{Z} = AZ + c$

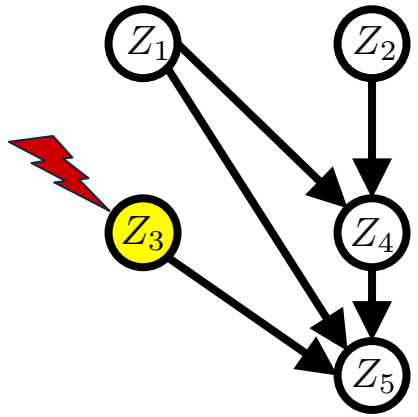
Discussion Order

do



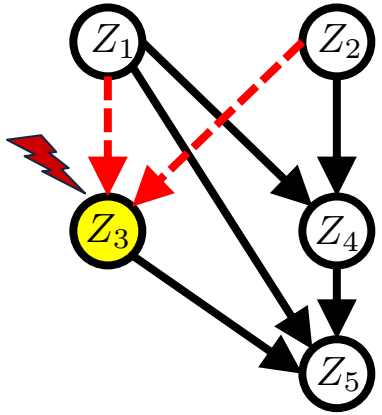
$$\begin{cases} 1 & \text{for } Z_3 = z^* \\ 0 & \text{for } Z_3 \neq z^* \end{cases}$$

hard (perfect)



$$q(Z_3)$$

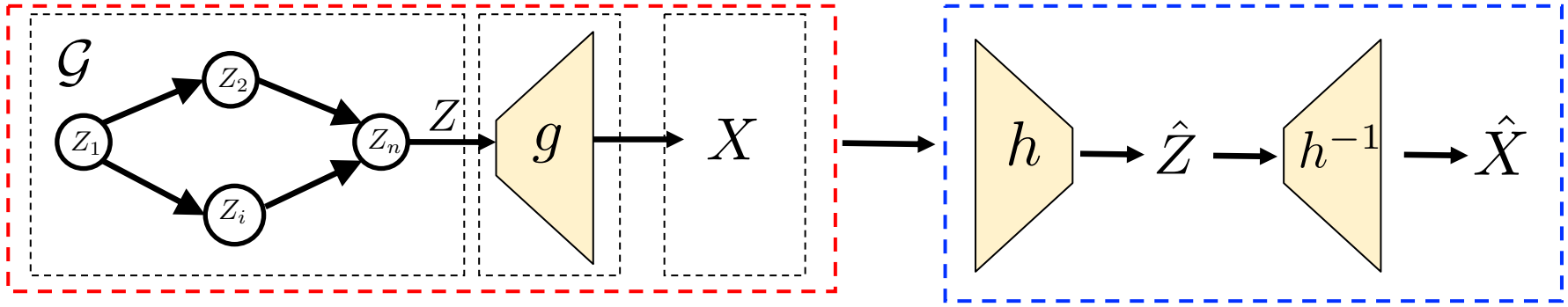
soft (imperfect)



$$q(Z_3|Z_1, Z_2)$$

stronger → weaker (more real)

do interventions



Polynomial g. Let do intervention $Z_k = z_k^*$ for fixed z_k^*

- Constraint: $\hat{Z}_k := [h(X)]_k = z_k^*$ for all $X \in \mathcal{X}^k$

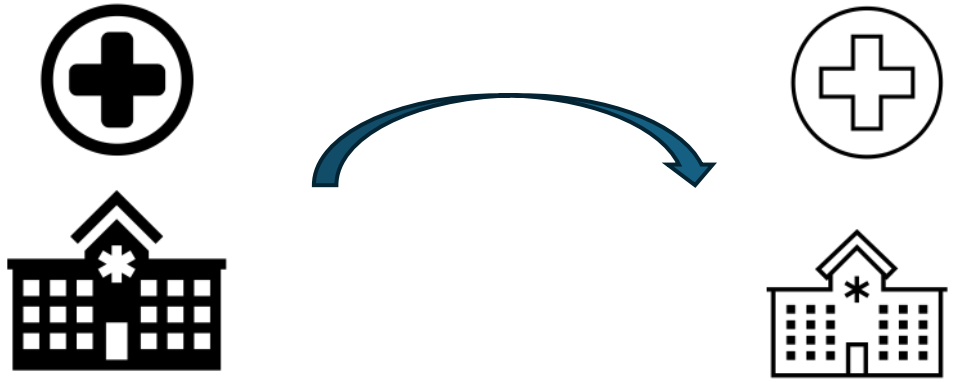
Theorem. observational data + one **do** intervention/node ensure

1. Graph recovery is **perfect**: $\hat{\mathcal{G}} = \mathcal{G}$
2. latent variables are recovered **up to scaling**: $\hat{Z}_k = c_k \cdot Z_k + b$

Stochastic Interventions

deterministic intervention is not always possible…

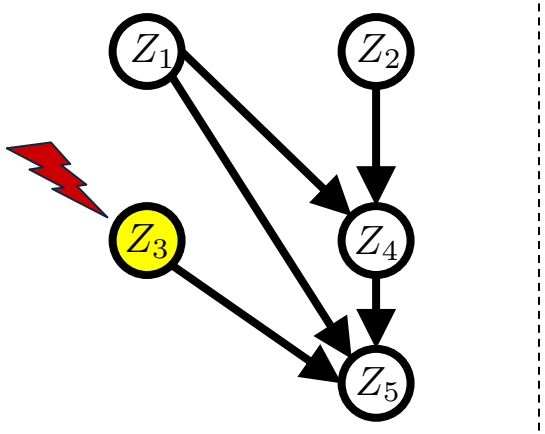
causal mechanisms are still probabilistic



$$p_i(z_i \mid \mathrm{pa}(z_i)) \rightarrow q_i(z_i \mid \mathrm{pa}(z_i))$$

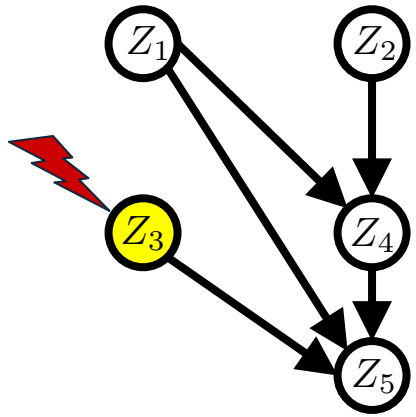
Discussion Order

do



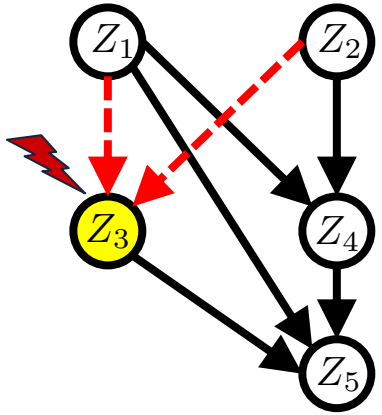
$$\begin{cases} 1 & \text{for } Z_3 = z^* \\ 0 & \text{for } Z_3 \neq z^* \end{cases}$$

hard (perfect)



$$q(Z_3)$$

soft (imperfect)

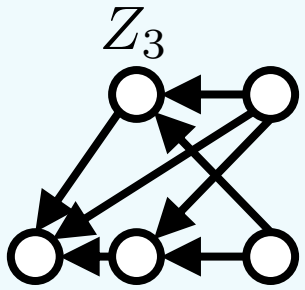


$$q(Z_3|Z_1, Z_2)$$

stronger → weaker (more real)

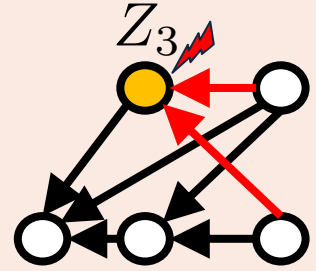
How do stochastic interventions help?

$p_3(z_3 \mid z_{\mathrm{pa}(3)})$



$$p(Z) = p_3(z_3 \mid \mathrm{pa}(z_3)) \prod_{i \neq 3} p_i(z_i \mid \mathrm{pa}(z_i))$$

$q_3(z_3 \mid z_{\mathrm{pa}(3)})$



$$p^3(Z) = q_3(z_3 \mid \mathrm{pa}(z_3)) \prod_{i \neq 3} p_i(z_i \mid \mathrm{pa}(z_i))$$

$$\log p(z) - \log p^3(z) = \log \frac{p_3}{q_3}(z_3 \mid \mathrm{pa}(z_3))$$

function of only z_3 and $\mathrm{pa}(z_3)$

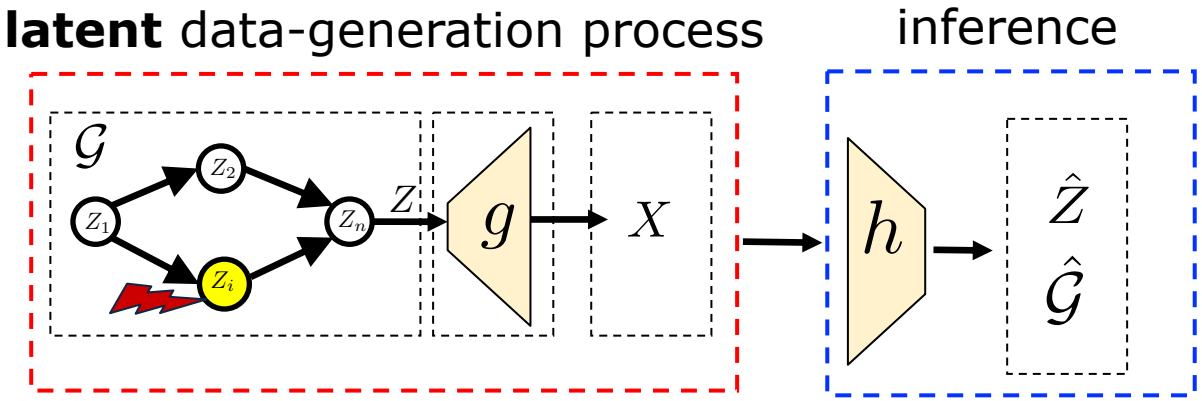
Score functions: $\underbrace{\nabla_z \log p(Z)}_{s(Z)} - \underbrace{\nabla_z \log p^3(Z)}_{s^3(Z)} = \begin{bmatrix} x \\ 0 \\ x \\ 0 \\ 0 \\ x \\ 0 \end{bmatrix}$

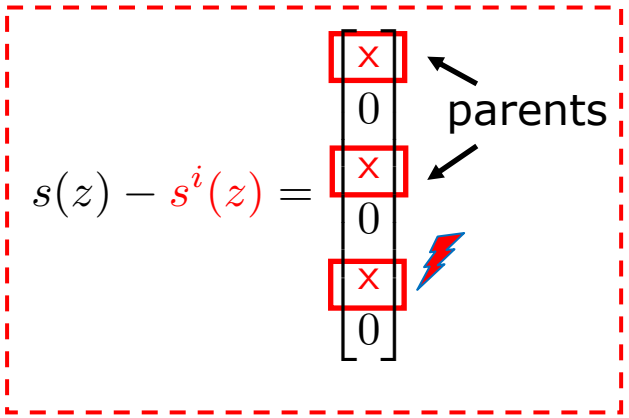
parent coordinates

intervened node

Learning Signal: Score Difference Sparsity

$s(z) - s^i(z) = \nabla_z p_i(z_i \mid \mathrm{pa}(z_i)) - \nabla_z q_i(z_i \mid \mathrm{pa}(z_i))$: function of only z_i and $\mathrm{pa}(z_i)$



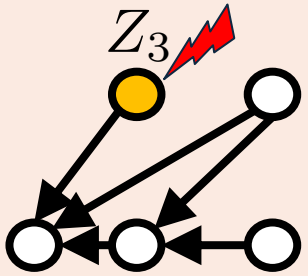


1. non-zero coordinates of score difference = **DAG structure**!
2. *(informal)* estimated score differences *cannot* be sparser than the true score differences

incorrect encoder, i.e., $h(g(Z)) \neq Z \quad \rightarrow \quad s(\hat{z}) - s^i(\hat{z})$ **not** a function of only z_i & $\mathrm{pa}(z_i)$

Even Sparser Score Difference

$q_3(z_3)$

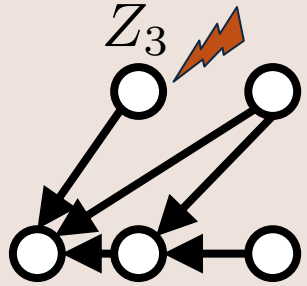


$$p^3(Z) = q_3(z_3) \prod_{i \neq 3} p_i(z_i \mid \mathrm{pa}(z_i))$$

Two hard interventions on the same node

$q_i(z_i)$ & $\tilde{q}_i(z_i)$

$\tilde{q}_3(z_3)$



$$\tilde{p}^3(Z) = \tilde{q}_3(z_3) \prod_{i \neq 3} p_i(z_i \mid \mathrm{pa}(z_i))$$

$$\log p^i(z) - \log \tilde{p}^i(z) = \log q_i(z_i) - \log \tilde{q}_i(z_i)$$ function of only z_i

Score functions: $$s^i(Z) - \tilde{s}^i(Z) = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ \mathrm{x} \\ 0 \end{bmatrix} \leftarrow$$ intervened node

For single-node interventions, reverse direction is also true!

Observational and Interventional environment

Hard or (soft + additive noise model): $\mathcal{E}^0$ and $\mathcal{E}^m$ with the target i

$$\mathbb{E}\left[\left|s(Z) - s^m(Z)\right|_k\right] \neq 0 \quad \Longleftrightarrow \quad k \in \mathrm{pa}(i)$$

$$s(z) - s^m(z) = \begin{bmatrix} \times \\ 0 \\ \times \\ 0 \\ \times \\ 0 \end{bmatrix}$$

parents

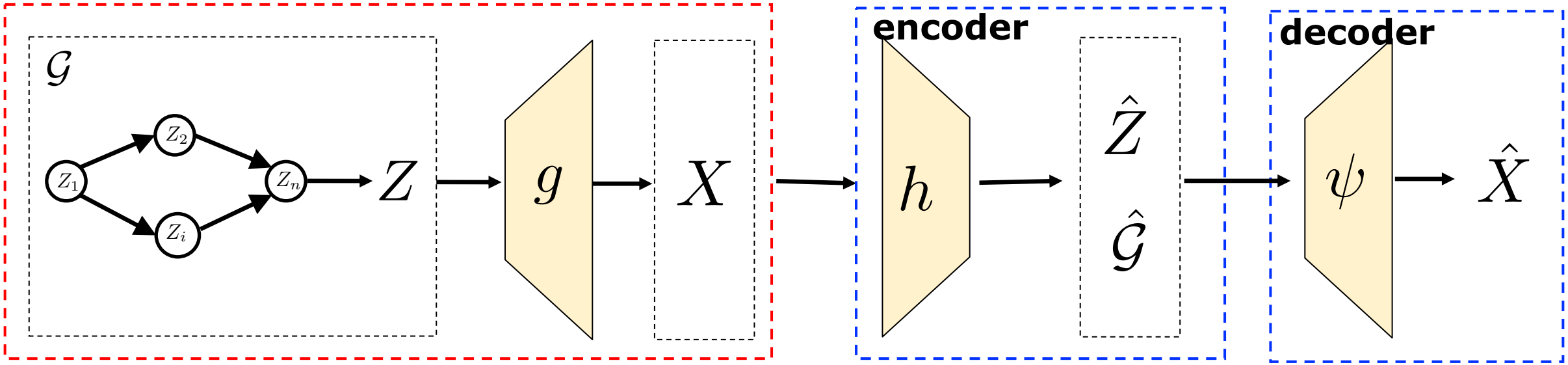
Two interventional environments

Coupled hard: $\mathcal{E}^m$ and $\tilde{\mathcal{E}}^m$ with the same target i

$$\mathbb{E}\left[\left|\tilde{s}(Z) - \tilde{s}^m(Z)\right|_k\right] \neq 0 \quad \Longleftrightarrow \quad k = i$$

$$s^m(z) - \tilde{s}^m(z) = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ \times \\ 0 \end{bmatrix}$$

Common Strategy



Find encoder-decoder pair (h, ψ) that allows perfect recovery:

- **variables:** $\hat{X} = X$
- **graph:** pdf of $\hat{Z}$ factorizes wrt a DAG $\mathcal{G}$

score functions $s(\hat{Z})$ have similar sparsity structures as $s(Z)$

Score-based CRL Guarantees

CRL guarantees rely on an array of model/intervention assumptions

- **Latent score functions:** How to compute them?
- **Causal model:** parametric (e.g., linear) versus nonparametric SCMs
- **Transformation:** parametric (e.g., linear) versus nonparametric
- **Intervention size & type:** How many hard & soft int. per node?
- **Intervention targets:** What happens when targets are unknown?
- **Finite-sample:** What are the sample complexity guarantees?

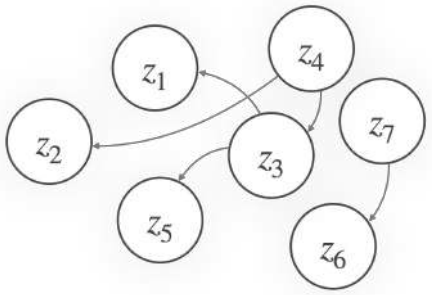
Score Difference Transform

Compute **latent** score differences using **observed** score differences

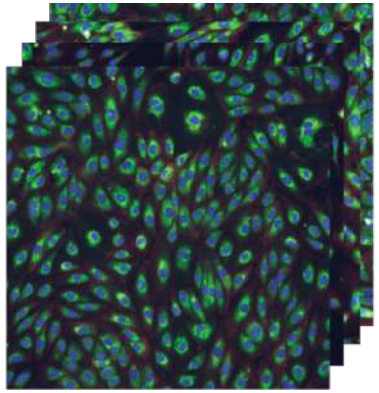
$$X = g(Z) \qquad p_X(x) = p_Z(z) \times \left|\det\left(J_g(z)^\top J_g(z)\right)\right|^{-\frac{1}{2}}$$

$$\implies \quad s_Z(z) = \left[J_g(z)\right]^\top \cdot s_X(x) + \nabla_z \log \left|\det\left(\left[J_g(z)\right]^\top \cdot J_g(z)\right)\right|^{1/2}$$

$$s_Z(z) - s_Z^m(z) = \left[J_g(z)\right]^\top \cdot \left(s_X(x) - s_X^m(x)\right)$$



$$Z \xrightarrow[\text{true dec.}]{g} X \xrightarrow[\text{cand. enc.}]{h} \hat{Z} \xrightarrow[\text{cand. dec.}]{h^{-1}} X$$

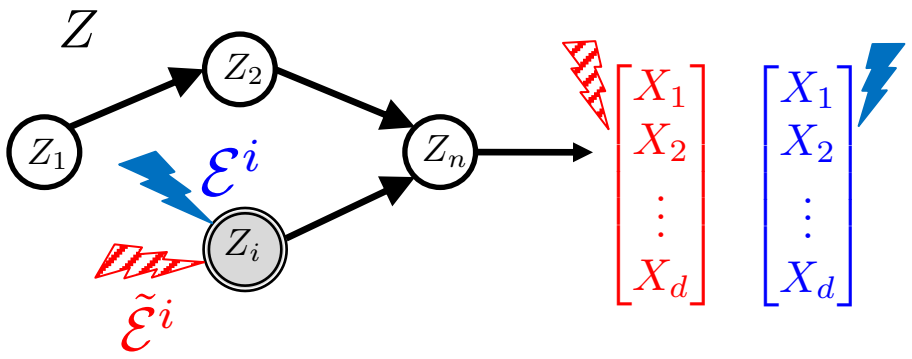


$$s_{\hat{Z}}(\hat{z}) - s_{\hat{Z}}^m(\hat{z}) = \left[J_{h^{-1}}(x)\right]^\top \cdot \left(s(x) - s^m(x)\right)$$

General CRL: Two Hard Interventions

$$X = g(Z)$$

- **two** hard interventions on Z_i : $p_i(z_i \mid \mathrm{pa}(z_i)) \rightarrow \mathcal{E}^i : q_i(z_i)$ and $\tilde{\mathcal{E}}^i : \tilde{q}_i(z_i)$
- mild regulatory assumption: $\frac{\partial}{\partial z_i} \frac{q_i(z_i)}{\tilde{q}_i(z_i)} \neq 0$ almost everywhere
- Sparse score difference: $\mathbb{1}\left(\mathbb{E}\left[s^i(Z) - \tilde{s}^i(Z) \right] \right) = \mathbf{e}_i$

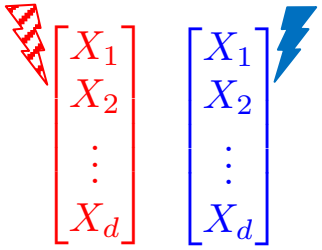


$$s^i(z) - \tilde{s}^i(z) = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ \mathrm{x} \\ 0 \end{bmatrix}$$

General CRL: Node-level Identifiability

if given **only**

two environments



Solve for the encoder

$$h_i^* = \arg\min_{h \in \mathcal{H}} \left\| \mathbb{E}\left[\left| s^i(\hat{z}) - \tilde{s}^i(\hat{z}) \right| \right] - \mathbf{e}_i \right\|^2$$

node-level ID guarantee $\hat{Z}_i = [h_i^*(X)]_i = \phi_i(Z_i)$

ϕ_i: diffeomorphism (bijection, differentiable)

$$s^i(\hat{z}) - \tilde{s}^i(\hat{z}) = J_\phi^{-\top}(z) \cdot \left(s^i(z) - \tilde{s}^i(z) \right) \qquad \text{where} \quad \hat{Z} = \phi(Z)$$

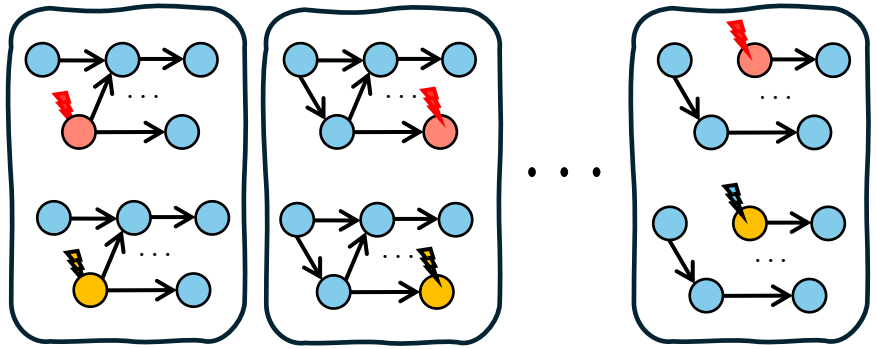
how it works?

$$\begin{bmatrix} 0 \\ 0 \\ 0 \\ x \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} \frac{\partial \phi_1}{\partial Z_1} & \frac{\partial \phi_1}{\partial Z_2} & \cdots & \frac{\partial \phi_1}{\partial Z_n} \\ \frac{\partial \phi_2}{\partial Z_1} & \frac{\partial \phi_2}{\partial Z_2} & \cdots & \frac{\partial \phi_2}{\partial Z_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial \phi_m}{\partial Z_1} & \frac{\partial \phi_m}{\partial Z_2} & \cdots & \frac{\partial \phi_m}{\partial Z_n} \end{bmatrix} \cdot \begin{bmatrix} 0 \\ x \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

$$\frac{\partial \hat{Z}_i}{\partial Z_j} \neq 0 \iff i = j$$

General CRL: Complete Identifiability

if given two interventions for **all** nodes



$s_X^1 - \tilde{s}_X^1 \quad \cdots \quad s_X^n - \tilde{s}_X^n$

(observational – interventional score): non-zero at coordinates i and parents of i

Solve for the encoder

$$h^* = \arg\min_{h \in \mathcal{H}} \sum_{i=1}^{n} \left\| \mathbb{E}\left[\left| s^i(\hat{z}) - \tilde{s}^i(\hat{z}) \right| \right] - \mathbf{e}_i \right\|^2$$

complete ID guarantee $\hat{Z}_i = \phi_i(Z_i)$ for all $i \in [n]$

Exact graph recovery

$$\mathbb{E}\left[\left| s(\hat{z}) - s^i(\hat{z}) \right| \right]_k \neq 0 \iff k \in \mathrm{pa}(i) \cup i$$

General CRL Summary

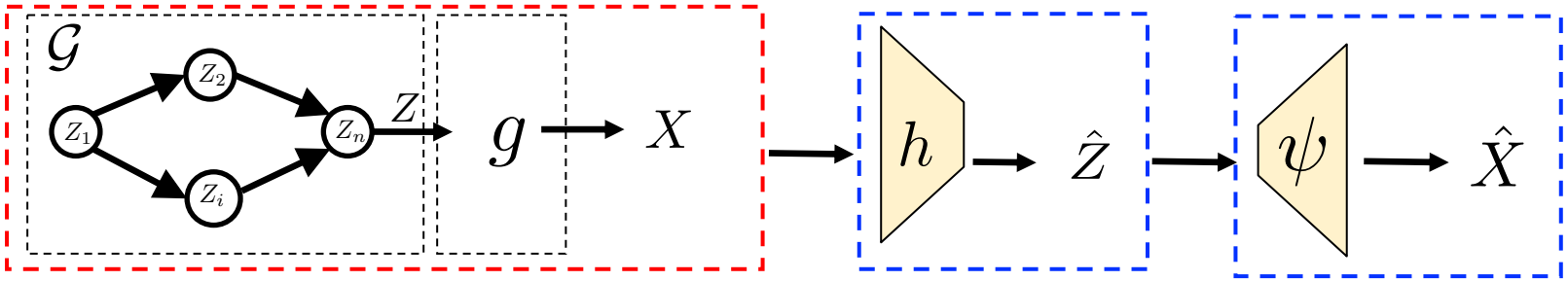
Theorem. observational data + **two hard interventions/node** ensure

1. Graph recovery is **perfect**: $\hat{\mathcal{G}} = \mathcal{G}$
2. Variable recovery is up to **component-wise transforms***, for every intervened node:

$$\hat{Z}_i = \phi_i(Z_i)$$

- Variable recovery*: information-theoretic ID limit (von Kügelgen et al. NeurIPS 2023)
- Observational data: can be relaxed under mild assumptions

General CRL: Implementation



- Parameterize encoder/decoder, learn via gradient descent. Scalable!
- Latent scores via observed score differences $s^i(\hat{z}) - \tilde{s}^i(\hat{z}) = [J_\psi(\hat{z})]^\top \cdot (s^i(x) - \tilde{s}^i(x))$
- No further restrictions (e.g., faithfulness)

$$h^*, \psi^* = \arg\min_{h,\psi} \sum_{i=1}^{n} \left\| \mathbb{E}\left[\left| s^i(\hat{z}) - \tilde{s}^i(\hat{z}) \right| \right] - \mathbf{e}_i \right\|^2 + \|(\psi \circ h)(X) - X\|^2$$

(encoder, decoder)

reconstruction loss for invertibility

Score Estimation Landscape

detailed discussed soon (empirical studies)

General

Sliced score-matching
Denoising score matching

Parametric

Gaussian

Non-parametric

Kernel-based
RKHS-based

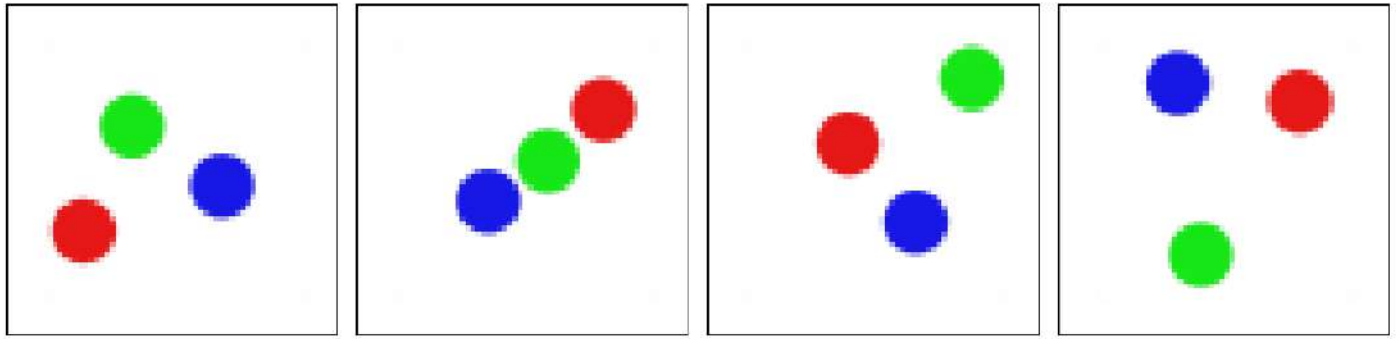
Y. Song, S. Garg, J. Shi and S. Ermon. *"Sliced score matching: A scalable approach to density and score estimation"*. UAI'20
P. Vincent. *"A connection between score matching and denoising autoencoders"*. Neural computation, 2011.
A. Wibisono, Y. Wu, and K.Y. Yang. *"Optimal score estimation via empirical Bayes smoothing"*. CoLT'24
Y. Zhou, J. Shi, and J. Zhu. *"Nonparametric score estimators"*. ICML'20

Simulations: Image Dataset

- **Image rendering**: 3 balls with different colors (dimension: 64x64x3)
- Latent variables: coordinates of ball centers
- Linear Gaussian SEMs over 6 nodes (Erdős–Rényi random graphs)
- 5 runs, 10^4 samples per node in each run
- Learn score differences via density ratio estimation

$$Z_1 = [0.0172, 0.1845, 0.2701, 0.7289, 0.7369, 0.4203]^\top$$
$$Z_2 = [0.8973, 0.8146, 0.6091, 0.5495, 0.2992, 0.3394]^\top$$
$$Z_3 = [0.3443, 0.6400, 0.9879, 0.9780, 0.6948, 0.2323]^\top$$
$$Z_4 = [0.8713, 0.8564, 0.3557, 0.0617, 0.2374, 0.9535]^\top$$

g



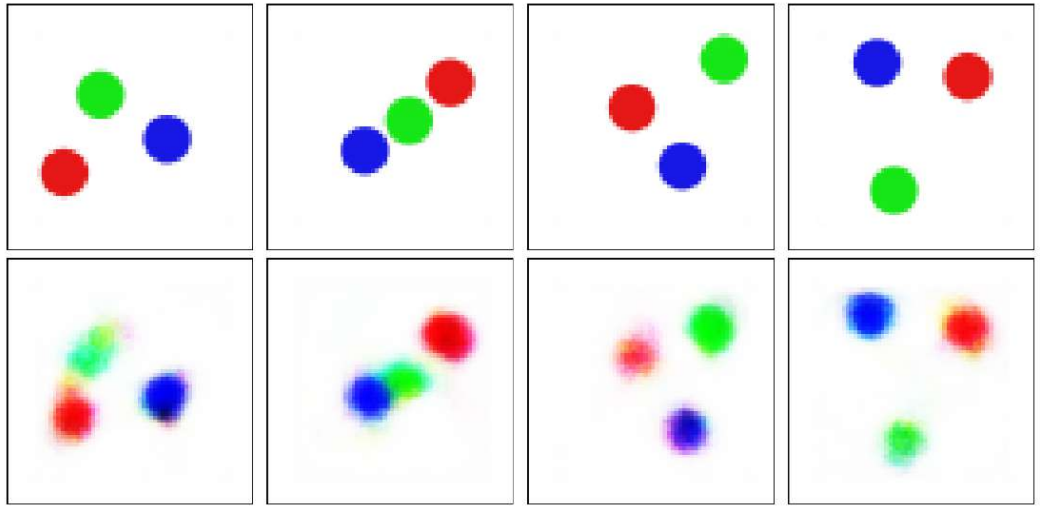
g is unknown and non-parametric

Simulations: CRL and Reconstruction

- Performance varies from near-perfect to OK (but all visually distinguishable)
- Variable recovery metric: mean correlation coefficient **(MCC):**
- Linear correlations between estimated and true variables (ideally 1)
- Similar experiments: (Ahuja'23, do interv.) and (Buchholz'23, contrastive learning)

Ground truth

Reconstruction



$$\mathrm{MCC}(Z, \hat{Z}) \triangleq \frac{1}{n} \sum_{i \in [n]} \mathrm{corr}(Z_i, \hat{Z}_i)$$

Mean MCC : 0.76 ± 0.17

General CRL – One intervention/node?

Question: Is it possible to solve CRL with **one hard int/node?**

Intuition: If node *i* has no parents, one intervention suffices (recall node-level ID)

Answer: It is possible, but under more assumptions …

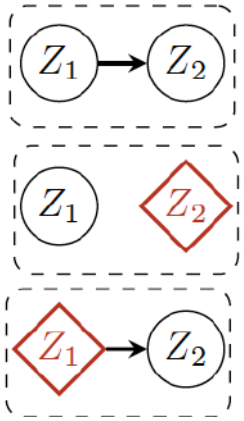
von Kügelgen et al. (2023): **n=2** nodes & nonlinear *genericity* assumptions (effect of node 1 on 2)

Wendong et al. (2023): known latent graph

Yao et al. (2025): known topological order

need to have

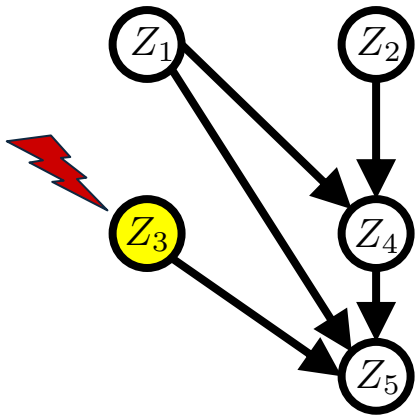
$$s(z) - s^i(z) = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ \times \\ 0 \end{bmatrix}$$



J. von Kügelgen et al., *Nonparametric Identifiability of Causal Representations from Unknown Interventions,* NeurIPS'2023
L. Wendong et al., *Causal Component Analysis,* NeurIPS'2023
D. Yao, D. Rancati, R. Cadei, M. Fumero, F. Locatello, *Unifying CRL with the Invariance Principle,* ICLR'2025

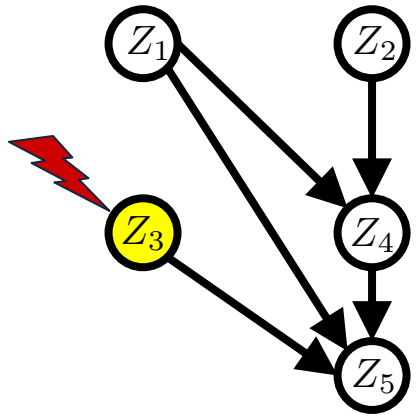
Discussion Order

do



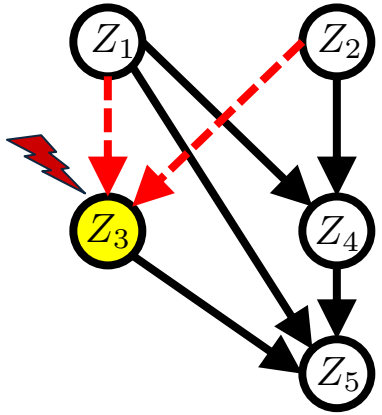
$$\begin{cases} 1 & \text{for } Z_3 = z^* \\ 0 & \text{for } Z_3 \neq z^* \end{cases}$$

hard (perfect)



$$q(Z_3)$$

soft (imperfect)



$$q(Z_3|Z_1, Z_2)$$

stronger → weaker (more real)

Soft Interventions

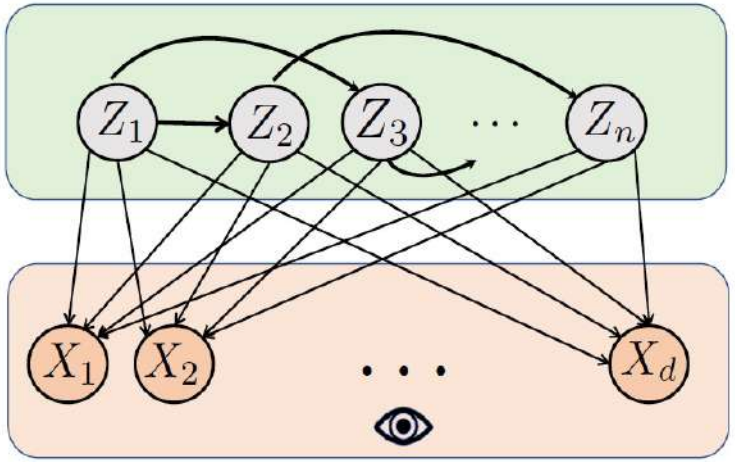
Linear Causal Representation Learning

linear transformation from latent to observable space

$$X = \mathbf{G} \cdot Z$$

restriction to linear transforms enables extensions for:

- ✔ **soft interventions**
- ✔ fewer interventional mechanisms (one per node)
- ✔ unknown multi-node interventions
- ✔ finite-sample guarantees



What Happens to Score Functions?

in general: $s(x) - s^m(x) = \left[J_{g^{-1}}(x)\right]^\top \left(s(z) - s^m(z)\right)$

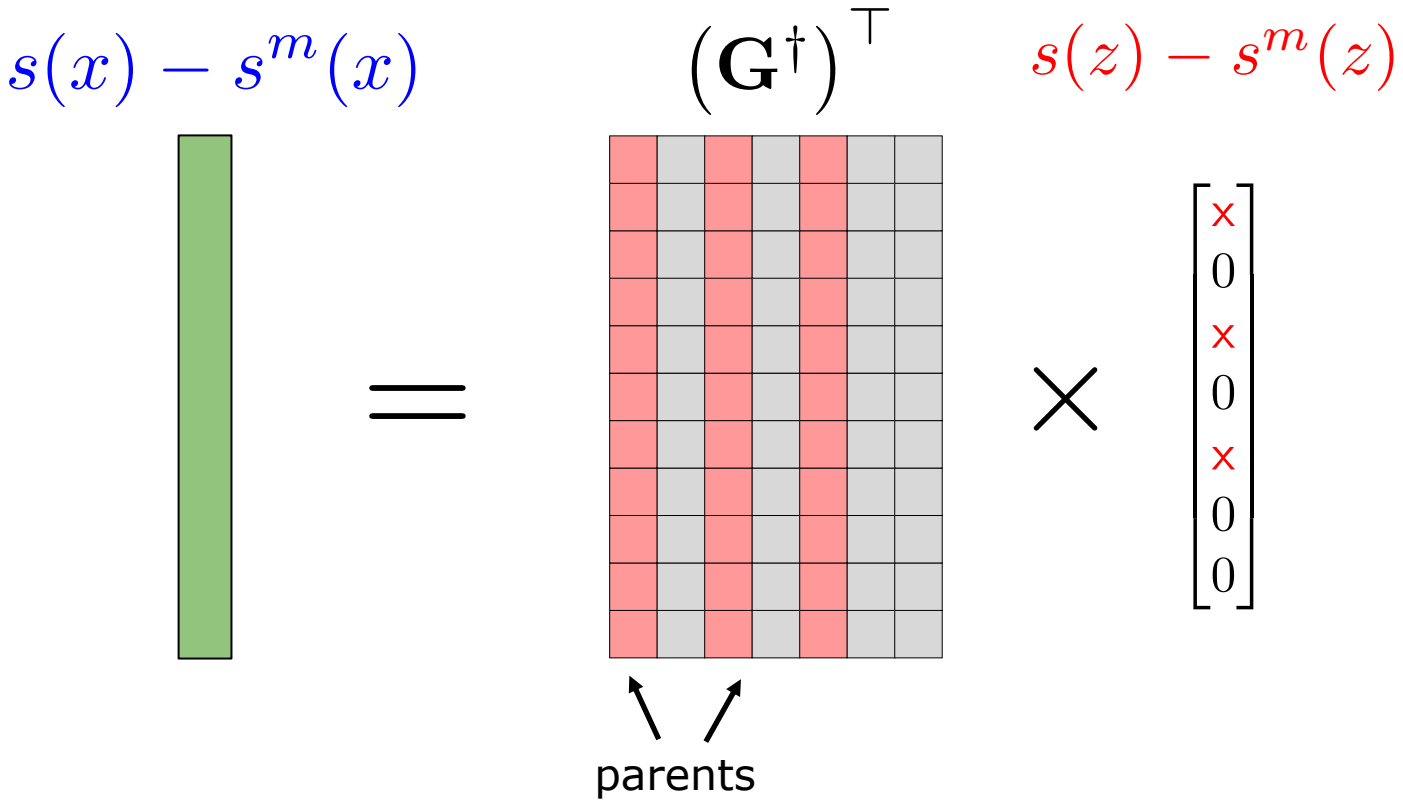
$$Z = \mathbf{G}^\dagger \cdot X \implies J_{g^{-1}}(x) = \mathbf{G}^\dagger \rightarrow \quad s(x) - s^m(x) = \left(\mathbf{G}^\dagger\right)^\top \cdot \left(s(z) - s^m(z)\right)$$

- Score differences are linearly related
- Observed score differences secretly captures the inverse transform!
- Define correlation matrices for score differences

$$\mathbf{R}_X^i := \mathrm{Cor}\left[s(x) - s^i(x)\right]$$
$$\mathbf{R}_Z^i := \mathrm{Cor}\left[s(z) - s^i(z)\right]$$

$$\mathbf{R}_X^i = (\mathbf{G}^\dagger)^\top \cdot \mathbf{R}_Z^i \cdot \mathbf{G}^\dagger$$

Soft Interventions

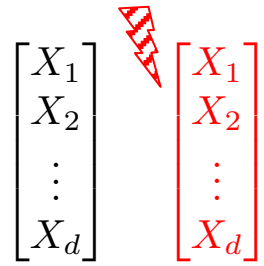


Infer about the inverse transform using observed score difference

$$\left(s(x) - s^m(x)\right) \in \text{span}\{\mathbf{G}_j^\dagger : j \in \text{pa}(i) \cup i\}$$

Linear CRL: Partial Identifiability

if given **only one** int. environment



compute correlation matrix of score difference

$$\mathbf{R}_X^i := \mathrm{Cor}\left[s(x) - s^i(x)\right]$$

estimate for the i-th row of the true decoder $\mathbf{G}^\dagger$

$$\mathbf{H}_i \leftarrow \mathbf{R}_X^i \cdot y = c_i \cdot \mathbf{G}_i^\dagger + \sum_{k \in \mathrm{pa}(i)} c_k \cdot \mathbf{G}_k^\dagger \quad , \quad y \in \mathbb{S}^{d-1}$$

node-level partial ID **soft**:

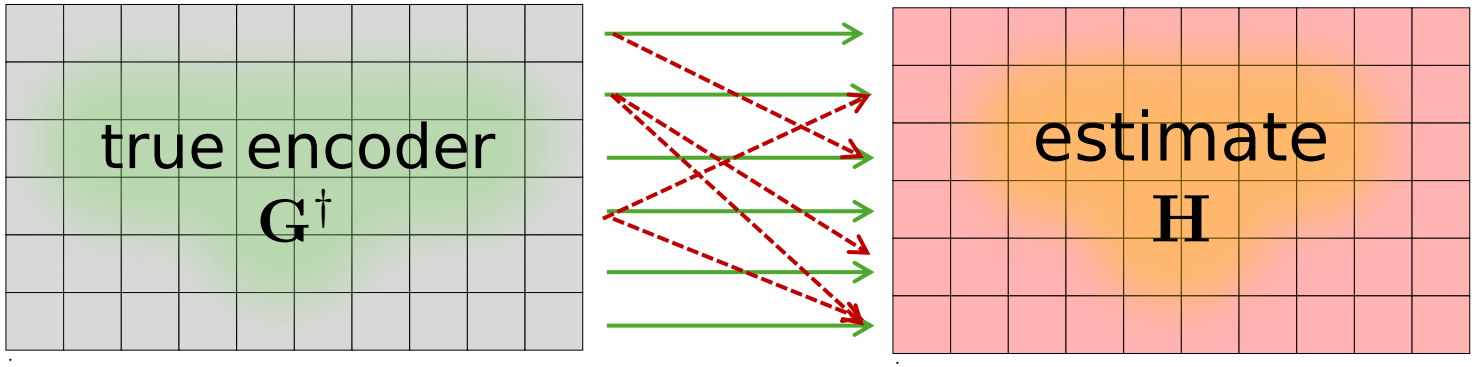
identifiability up to mixing with parents

$$\hat{Z}_i = \mathbf{H}_i \cdot X = c_i \cdot Z_i + \sum_{k \in \mathrm{pa}(i)} c_k \cdot Z_k$$

Encoder Recovery

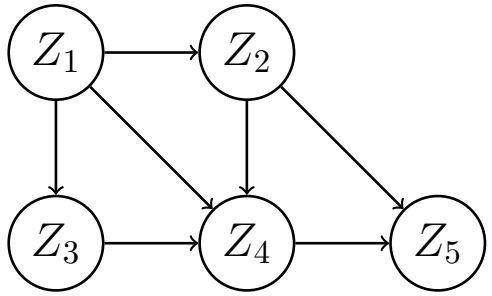
if given one soft intervention for **each** node,

$$\mathbf{H}_i \leftarrow \mathbf{R}_X^i \cdot y_i \quad , \quad \forall i \in [n],\ y_i \sim \mathbb{S}^{d-1}$$

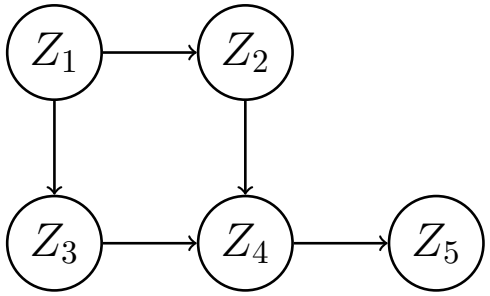


$$\hat{Z} = \mathbf{H} \cdot X = \mathbf{C} \cdot Z \qquad \text{where} \quad \mathbf{C}_{i,j} = 0 \ \ \forall j \notin \mathrm{pa}(i)$$

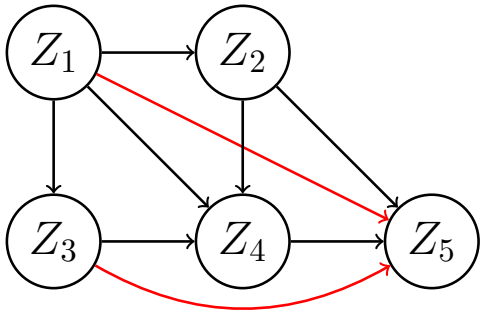
Latent Graph Recovery



original DAG



transitive reduction $\iff$



transitive closure

$$\hat{Z} = \mathbf{H} \cdot X : \qquad \mathbb{E}\left[\left|s(\hat{z}) - s^i(\hat{z})\right|\right]_k \neq 0 \implies k \in \mathrm{an}(i)$$

$$k \to i \in \mathcal{G}_{\text{trans.reduc.}} \implies \mathbb{E}\left[\left|s(\hat{z}) - s^i(\hat{z})\right|\right]_k \neq 0 \qquad (*)$$

one soft intervention/node: transitive closure/reduction of $\mathcal{G}$

(*): Under a mild assumption (satisfied for additive noise models and/or hard interventions etc.)

One Soft Intervention/Node Summary

Theorem. For linear transforms and general causal models, observational data + **one soft** intervention/node ensure

1. Graph recovery: up to transitive closure: $\hat{\mathcal{G}}_{\text{trans. clos.}} = \mathcal{G}_{\text{trans. clos.}}$
2. Latent variable recovery: up to **mixing with parents**:

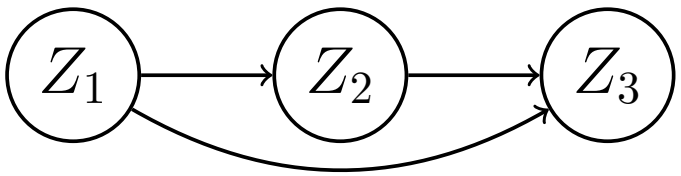
$$\hat{Z}_i = c_i \cdot Z_i + \sum_{k \in \text{pa}(i)} c_k \cdot Z_k$$

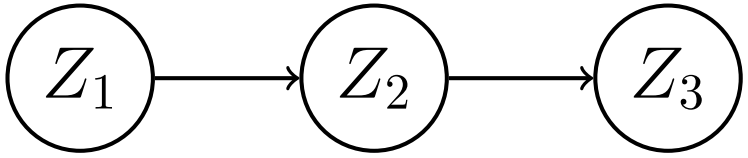
Soft Interventions – Limitations

Consider a linear causal model: $Z = \mathbf{A} \cdot Z + \epsilon$

Given one soft intervention from each node;

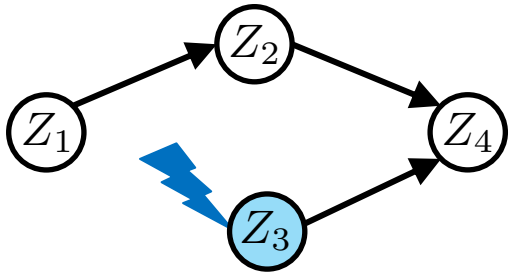
- Latent variable non-identifiability: for 2-node graph $Z_1 \to Z_2$, cannot get $\hat{Z}_2 = c \cdot Z_2$
- Latent graph non-identifiability: see 3-node graphs
- Without further assumptions, one soft int/node is not enough.





Hard Interventions – Latent Variables

- **One hard int/node:** target node becomes independent of its non-descendants.
- Additional step: use this property to resolve mixing with parents
- **Linear MMSE** estimator to update the encoder (in topological order)



$$Z_3 \perp\!\!\!\perp Z_1, Z_2$$

$$\mathbf{u} \leftarrow \mathrm{Cov}(\hat{Z}_i, \hat{\mathrm{pa}}(\hat{Z}_i)) \cdot \left[\mathrm{Cov}(\hat{\mathrm{pa}}(\hat{Z}_i))\right]^{-1}$$

$$\mathbf{H}_i \leftarrow \mathbf{H}_i - \mathbf{u} \cdot \mathbf{H}_{\hat{\mathrm{pa}}(i)}$$

identifiability up to scaling

$$\mathbf{H}_i = c_i \cdot \mathbf{G}_i^{\dagger} \quad \rightarrow \quad \hat{Z}_i = c_i \cdot Z_i$$

Linear CRL: Hard Interventions Summary

Construct $\mathcal{G}$ with: $\hat{\text{pa}}(i) := \{k \neq i \ : \mathbb{E}\left[\left|s(\hat{z}) - s^i(\hat{z})\right|\right]_k \neq 0\}$

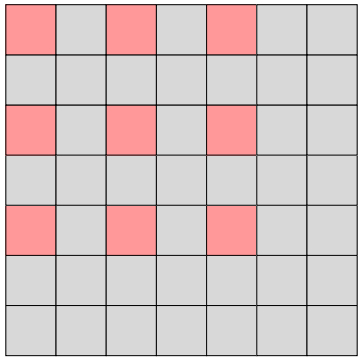
Theorem. For linear transforms and general causal models, observational data + **one hard** intervention/node ensure

1. **Perfect** graph recovery: $\hat{\mathcal{G}} = \mathcal{G}$
2. Latent variable recovery up to **scaling** $\hat{Z}_i = c_i \cdot Z_i$ for some constant c_i

Score Difference Rank

$\mathbf{R}_Z^i$: has non-zero $|\mathrm{pa}(i)+1| \times |\mathrm{pa}(i)+1|$ submatrix $\implies \mathrm{rank}(\mathbf{R}_Z^i) \leq |\mathrm{pa}(i)|+1$

- Linear causal models are *rank-deficient* [Squires et al. 2023]: $\mathrm{rank}(\mathbf{R}_Z^i) \leq 2$
- *Sufficiently* nonlinear models [Varıcı et al. 2023]: $\mathrm{rank}(\mathbf{R}_Z^i) = |\mathrm{pa}(i)|+1$
  - ✔ (full-rank) 2-layer MLP with $Z_i = \nu^\top \sigma(\mathbf{W} \cdot \mathrm{pa}(Z_i)) + \epsilon_i$
  - ✔ (full-rank) Quadratic: $Z_i = \sqrt{\mathrm{pa}(Z_i)^\top \cdot \mathbf{Q} \cdot \mathrm{pa}(Z_i)} + \epsilon_i$



$$\mathbf{R}_Z^i := \mathrm{Cor}\left[s(z) - s^i(z)\right]$$

Can we leverage this to improve soft intervention results?

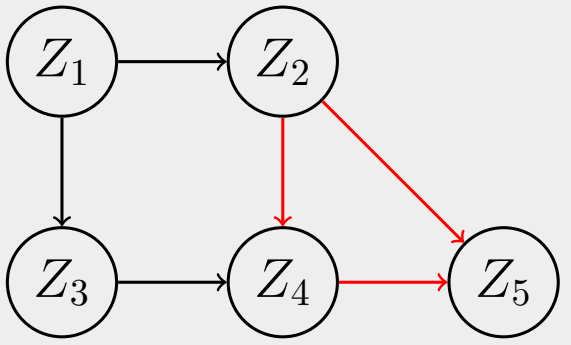
C.Squires, A.Seigal, S.Bhate, C. Uhler *"Linear Causal Disentanglement via Interventions"*. ICML'23
B.Varıcı, E.Acartürk, K.Shanmugam, A.Kumar, A.Tajer *"Score-based CRL from Interventions"*. 2023

Soft Interventions – Sufficiently Nonlinear SCMs

Theorem. observational data + one soft intervention/node ensure

1. Graph recovery is **perfect**
2. latent variables are recovered up to linear mixing with **surrounding parents**
3. Also, $\hat{Z}$ is Markov with respect to $\mathcal{G}$

- $Z_k \in \mathrm{pa}(Z_i)$ is a surrounding parent if $\mathrm{ch}(Z_i) \subset \mathrm{ch}(Z_k)$
- Result is tight (Jin & Syrgkanis, NeurIPS'24)



$\hat{Z}_1 : Z_1$

$\hat{Z}_2 : Z_2$

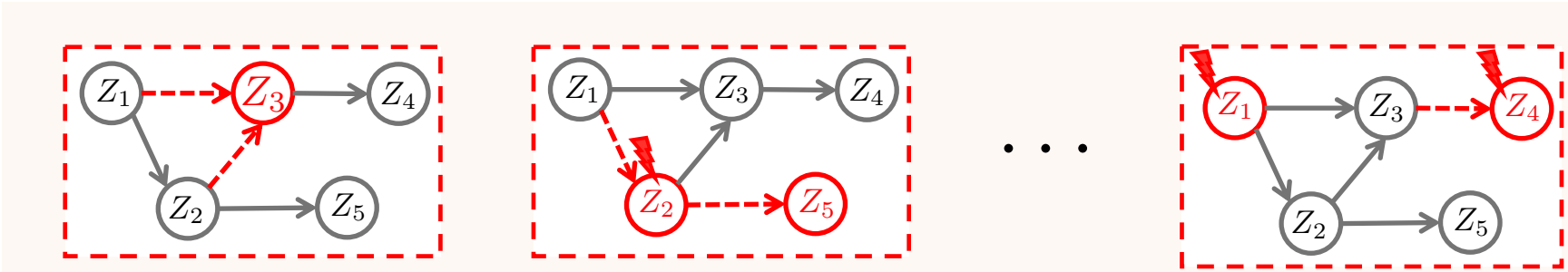
$\hat{Z}_3 : Z_3$

$\hat{Z}_4 : Z_4, Z_2$

$\hat{Z}_5 : Z_4, Z_2, Z_4$

surrounding parents

Linear CRL – Unknown Multi-node Interventions

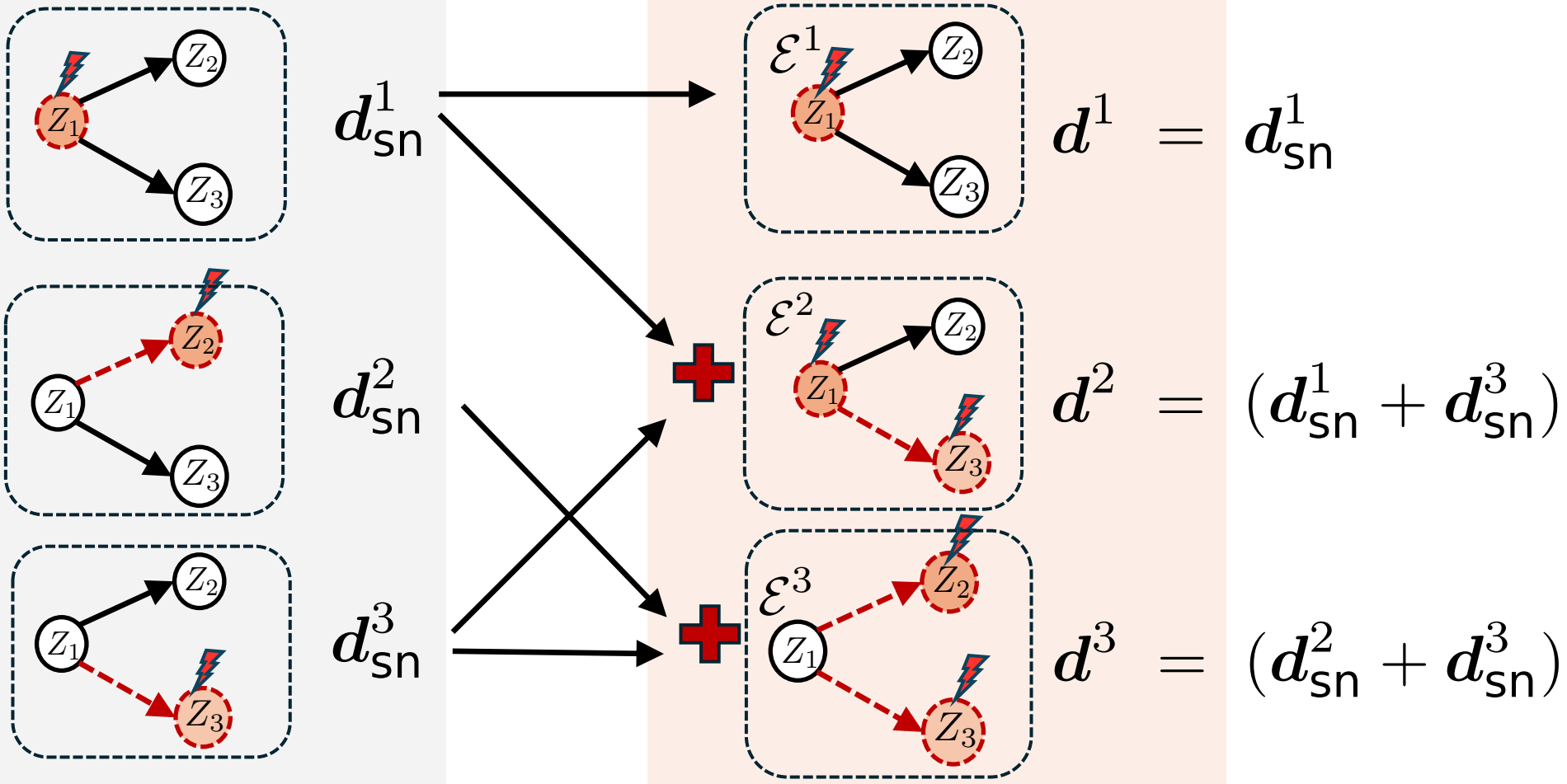


- M interventional environments $\{\mathcal{E}^1, \ldots, \mathcal{E}^M\}$
- environment $\mathcal{E}^M$ impacts an **unknown** set of nodes I^m

$$p^m(z) = \prod_{i \in I^m} q_i(z_i \mid \mathrm{pa}(z_i)) \prod_{i \notin I^m} p_i(z_i \mid \mathrm{pa}(z_i))$$

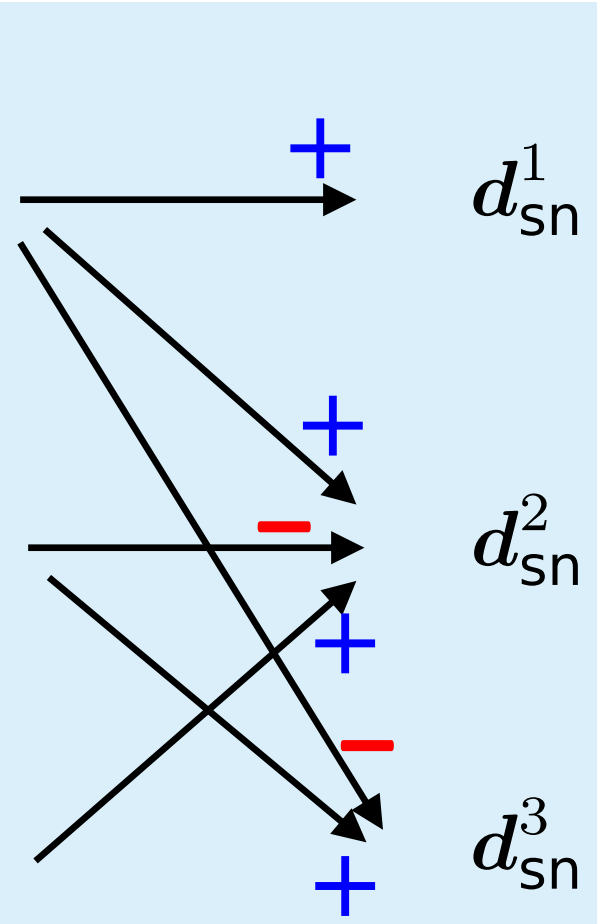
- **Challenge:** score differences are not sparse anymore
- **Question:** under what conditions do the single-node intervention guarantees hold?

Idea: Aggregate Score Differences



Single-node interventions and score differences

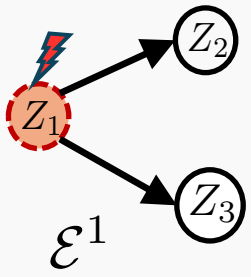
Multi-node interventions and score differences

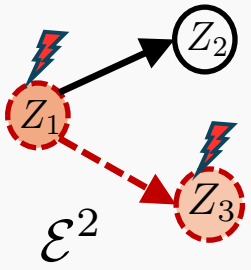


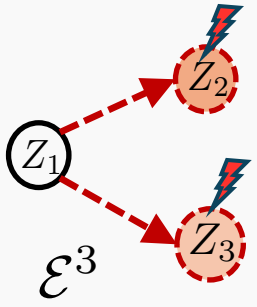
Create single-node score differences

Diverse unknown multi-node interventions

Requirement*: full-rank intervention matrix $\mathbf{D} \in \{0,1\}^{n \times M}$ with $\mathbf{D}_{i,m} = \mathbb{I}(i \in I^m)$



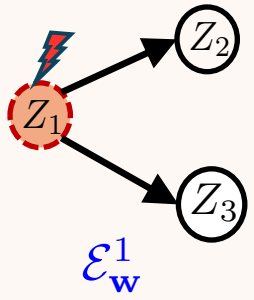


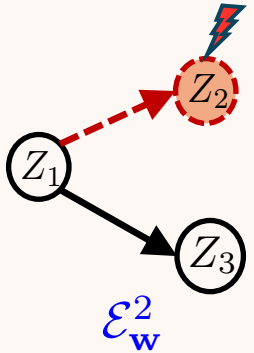


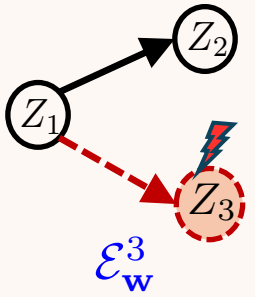
$$\mathbf{D} = \begin{array}{c} \begin{matrix} \mathcal{E}^1 & \mathcal{E}^2 & \mathcal{E}^3 \end{matrix} \\ \begin{bmatrix} 1 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 1 \end{bmatrix} \end{array} \begin{matrix} -Z_1 \\ -Z_2 \\ -Z_3 \end{matrix}$$

$$\begin{bmatrix} 1 & 1 & -1 \\ 0 & -1 & 1 \\ 0 & 1 & 0 \end{bmatrix} \times \mathbf{D} =$$

mixing matrix: $\mathbf{W}$







*Intervention regularity: effect of an intervention is distinct on the scores associated with different nodes.

Multi-node Intervention Guarantees

- Goal: find such a mixing matrix (with integer entries) to obtain sparser interventions.
- Strategy: use subspaces of aggregated score functions $\dim\Big(\text{proj.image}\big(\sum_{i=0}^{n} w_i \cdot s_X^i\big)\Big) = 1$

Theorem. observational data + diverse unknown multi-node interv.

1. **soft**: identifiability up to ancestors

$$\hat{Z}_i = c_i \cdot Z_i + \sum_{k \in \text{an}(i)} c_k \cdot Z_k \ , \qquad \hat{\mathcal{G}}_{\text{trans.clos.}} = \mathcal{G}_{\text{trans.clos.}}$$

2. **hard**: perfect identifiability

$$\hat{Z}_i = c_i \cdot Z_i \ , \qquad \hat{\mathcal{G}} = \mathcal{G}$$

Part III – Empirical Aspects & Recent Developments

11:00-12:15 (75 minutes)

- Score estimation
- Empirical Results for Interventional CRL
- Sample-Complexity Analysis
- Other recent developments
  - Multiview
  - Temporal
  - Mechanism Sparsity
  - and more …

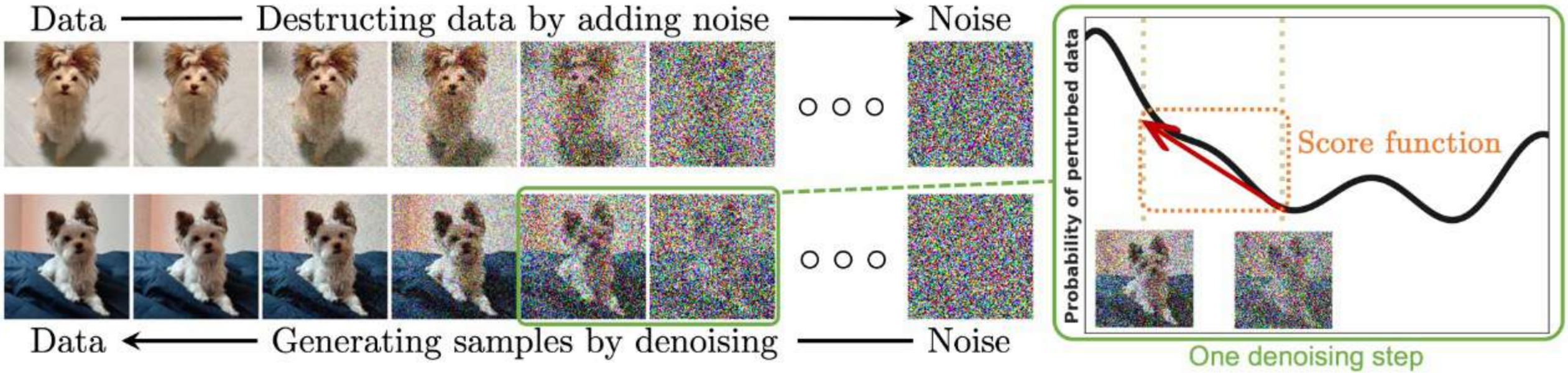
Emre Acartürk

Rensselaer Polytechnic Institute

Implementation and Experiments

Score Estimation

growing interest due to applications in diffusion models



directly uses score network

Figure credit: Yang et al. *"Diffusion Models: A Comprehensive Survey of Methods and Applications"*. 2024.

Score Estimation Landscape

General

Sliced score-matching
Denoising score matching

Parametric

Gaussian

Non-parametric

Kernel-based
RKHS-based

Y. Song, S. Garg, J. Shi and S. Ermon. *"Sliced score matching: A scalable approach to density and score estimation"*. UAI'20
P. Vincent. *"A connection between score matching and denoising autoencoders"*. Neural computation, 2011.
A. Wibisono, Y. Wu, and K.Y. Yang. *"Optimal score estimation via empirical Bayes smoothing"*. CoLT'24
Y. Zhou, J. Shi, and J. Zhu. *"Nonparametric score estimators"*. ICML'20

Score Estimation: Gaussian

$$Z \sim \mathcal{N}_n(\mu, \Sigma) \qquad\qquad s(z) = \nabla_z \log p(z) = -\Sigma^{-1} \cdot (z - \mu)$$

- Score function depends only on mean and (inverse) covariance
- How to estimate ? **Parameter estimation**
- Denote estimates of mean and covariance by $\hat{\mu}$ and $\hat{\Sigma}$.

Score function estimate is given by $\hat{s}(z) = -\hat{\Sigma}^{-1} \cdot (z - \hat{\mu})$.

Score Estimation: Sliced Score Matching

idea: match a score network $s_\theta(x)$ to true score function $s_X(x)$ in ℓ_2 norm

$$\underset{\theta}{\text{minimize}} \ \frac{1}{2}\,\mathbb{E}_{p_X}\left[\|\nabla_x \log p_X(x) - s_\theta(x)\|_2^2\right] \qquad \text{infeasible}$$

$$\underset{\theta}{\text{minimize}}\ \mathbb{E}_{p_X}\mathbb{E}_{\mathcal{N}_n}\left[v^\top \nabla_x s_\theta(x) v + \frac{1}{2}\|s_\theta(x)\|_2^2\right] \qquad \text{solvable}$$

properties:

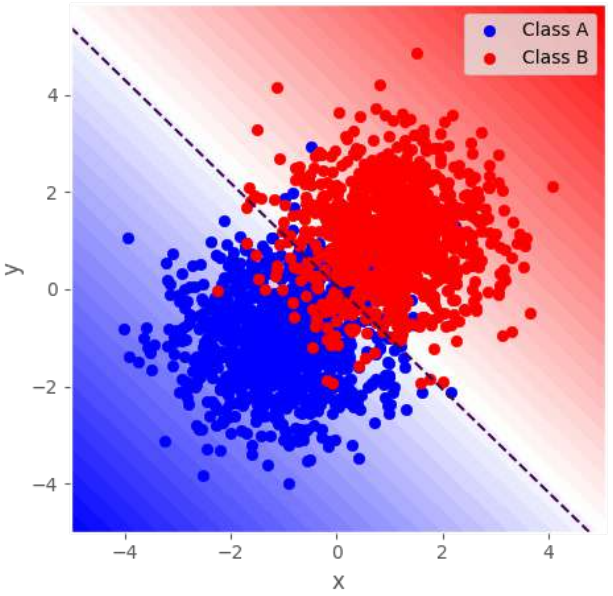
- consistent estimation
- differentiable loss function $\Rightarrow$ amenable to efficient optimization
- scalable (only matrix-vector multiplications)

*for various choices of p_v including standard Gaussian

Score (Difference) Estimation: Classification-based

Log density ratio (LDR) gives score difference

$$s(x) - s^i(x) = \nabla_x \log p(x) - \nabla_x \log p^i(x) = \nabla_x \log \frac{p(x)}{p^i(x)}$$



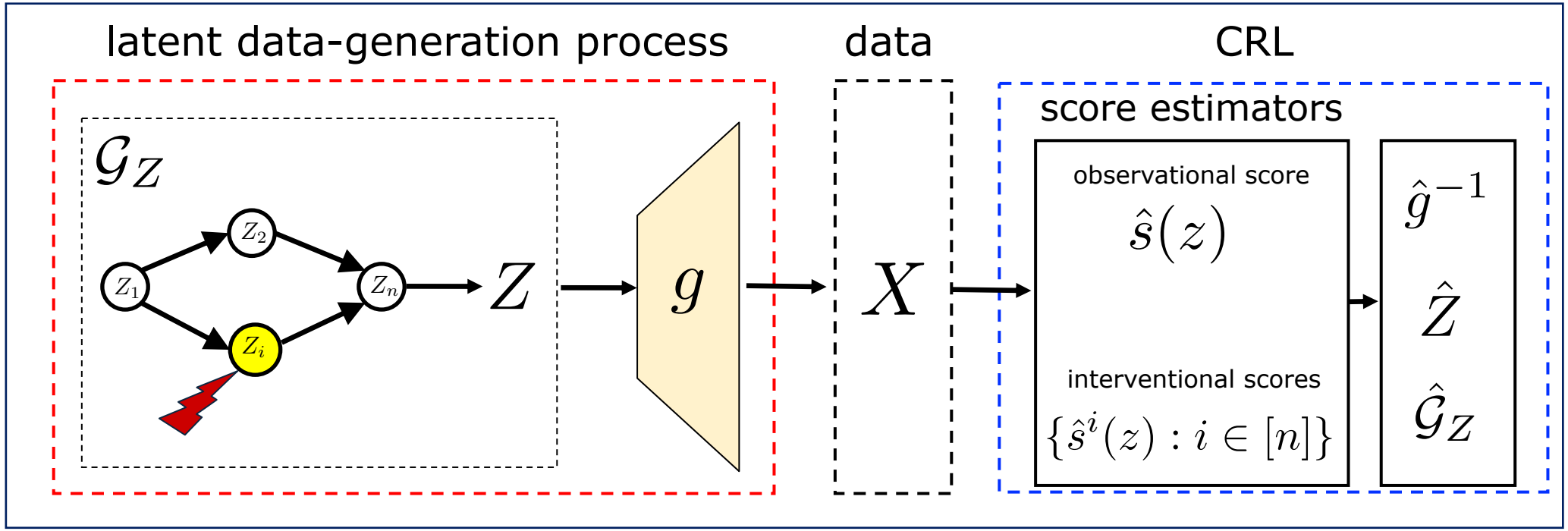
How to get LDR: Binary classification

$$\underset{\theta}{\text{minimize}} \; -\sum_t y_t \log(h_\theta(x_t)) + (1 - y_t) \log(1 - h_\theta(x_t)) \,.$$

$$\text{LDR estimate} = \log\left(\nu \cdot \frac{h_{\theta*}(x)}{1 - h_{\theta*}(x)}\right)$$

ν is the class prior $\frac{\sum_t y_t}{\sum_t (1-y_t)}$

Score Estimation Used Modularly



Empirical Results - Metrics

Data generation:

- Graph: Erdős–Rényi random graphs
- SCMs: Linear, quadratic, MLPs

Graph recovery: use Structural Hamming Distance (SHD):

count # of (edge additions/removals /flips)

Variable recovery: mean correlation coefficient (ideally 1)

$$\text{MCC}(Z, \hat{Z}) \triangleq \frac{1}{n} \sum_{i \in [n]} \text{corr}(Z_i, \hat{Z}_i)$$

Simulations – Linear SEMs

- **SEM:** Linear Gaussian
- **Intervention:** One/node (soft and hard)
- **Score estimation**: Parametric for Gaussian

# of samples	Soft SHD	Soft MCC	Hard SHD	Hard MCC
5,000	0.59 ± 0.11	0.98	0.17 ± 0.04	1.00
10,000	0.36 ± 0.08	0.98	0.09 ± 0.03	1.00
50,000	$\mathbf{0.28} \pm 0.06$	$\mathbf{0.98}$	$\mathbf{0.03} \pm 0.01$	$\mathbf{1.00}$

Latent dimension: n=5, observed dimension: d=100

Simulations – Nonlinear SEMs

- **SEM:** Quadratic Gaussian $Z_i = \sqrt{\mathrm{pa}(Z_i)^\top \cdot \mathbf{Q}_i \cdot \mathrm{pa}(Z_i)} + \epsilon_i$
- **Intervention:** One/node (soft and hard)
- **Score estimation**: Sliced score matching

	Score oracle	
Intervention	**SHD**	**MCC**
Soft	**0.11**	**0.93**
Hard	**0.03**	**1.00**

	Score Estimation	
Intervention	**SHD**	**MCC**
Soft	**2.79**	**0.60**
Hard	**2.62**	**0.93**

Latent dimension: n=5, observed dimension: d=100

Simulations – Discussion

Recovering the variables is easier than the graph

- Variables: Almost no hyperparameters
- Graph: More delicate …

perfect scores $\quad \mathbb{E}\left[\left|s(Z) - s^i(Z)\right|_k\right] = 0 \quad \Longleftrightarrow \quad k \notin \overline{\mathrm{pa}}(i)$

noisy scores $\quad \mathbb{E}\left[\left|\hat{s}(\hat{Z}) - \hat{s}^i(\hat{Z})\right|_k\right] \leq \epsilon \quad \Longleftrightarrow \quad k \notin \widehat{\overline{\mathrm{pa}}}(i)$

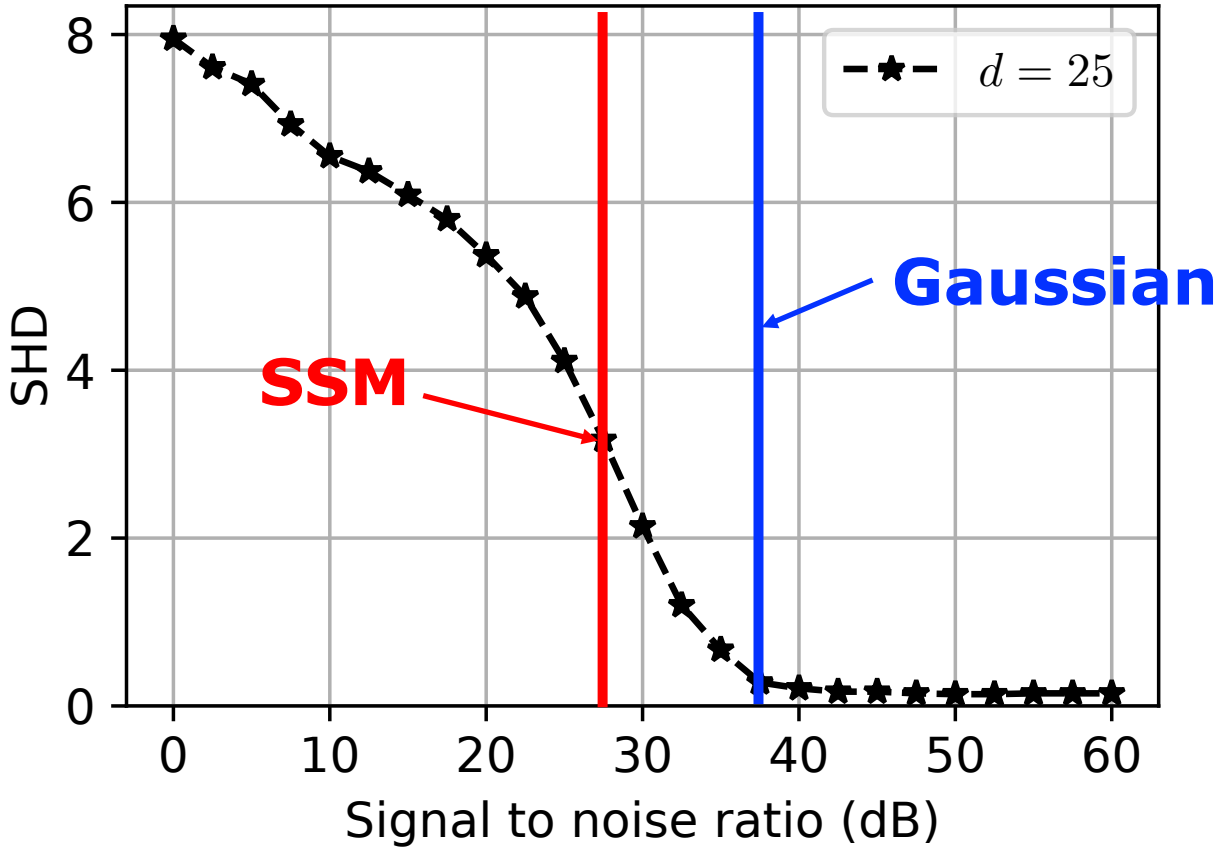
Noisy Estimates

How do noisy estimates degrade performance?

Manually add noise to true score functions (n=5 & d=25)

$$\hat{s}(x;\sigma^2) = s(x)\cdot\left(1+\Xi\right)$$

where $\Xi \sim \mathcal{N}(0, \sigma^2 \cdot \mathbf{I}_{d\times d})$.



Finite-sample Regime

Important observation:

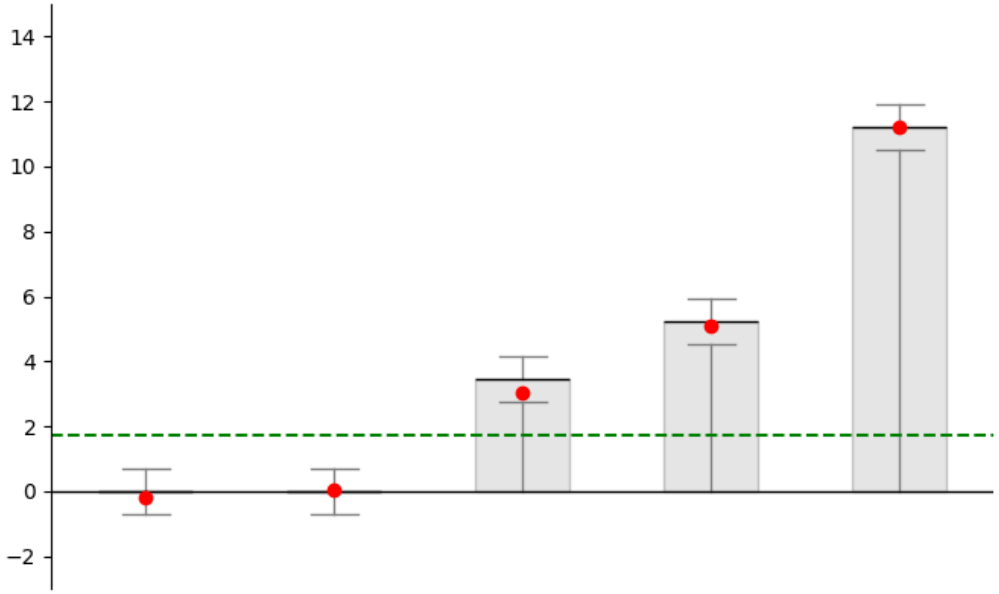
when estimate noise is low enough it doesn't hurt graph recovery

Why?

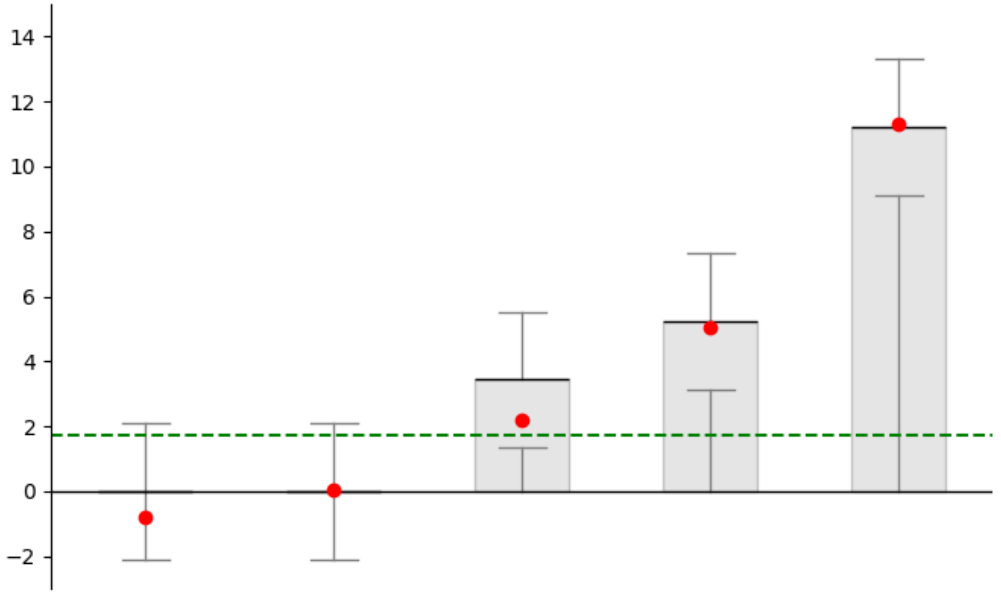
1. Our algorithm relies on estimating covariance matrix rank/eigenspace

2. rank/eigenspace estimates are stable (robust to noise)

Eigenvalue Stability



small score estimate noise



large score estimate noise

Finite-sample Recovery Metrics

Infinite sample identifiability:

holds w/ prob $1-\delta$

1. Graph recovery: up to transitive closure: $\hat{\mathcal{G}}_{\text{trans. clos.}} = \mathcal{G}_{\text{trans. clos.}}$
2. Latent variable recovery: up to mixing with parents:

$$\hat{Z}_i = c_i \cdot Z_i + \sum_{k \in \mathrm{pa}(i)} c_k \cdot Z_k$$

Finite sample identifiability: for given (ϵ, δ),

Sample Complexity of Linear CRL

Theorem. Adopting the RKHS-based score estimator, using

$$N \geq C \cdot \left(\frac{1}{\min\{\epsilon, c\}} \right)^4 \cdot \left(\log \frac{8n}{\delta} \right)^4$$

samples per environment ensures (ϵ, δ)–PAC identifiability.

Interventional CRL Recap

Score-based CRL framework:

- accommodates various **transformations** (parametric and non-parametric)
- accommodates various **causal models** (parametric and non-parametric)
- accommodates various **data** settings, including multi-node interventions
- Unifies identifiability analysis and algorithm design

Improvement in practice:

- More accurate score estimation
- More efficient/scalable score estimation

Other Interventional CRL Methods

- Interventional CRL via contrastive learning
- Discrepancy VAE
- CRL with known graphs

S. Buchholz et al. *"Learning linear causal representations from interventions under general nonlinear mixing".* NeurIPS'23
J. Zhang et al. *"Identifiability guarantees for causal disentanglement from soft interventions".* NeurIPS'23
L. Wendong et al. *"Causal Component Analysis".* NeurIPS'23
A. Li et al. *"Disentangled Representation Learning in Non-Markovian Causal Systems"*. NeurIPS'24

Interventional CRL via Contrastive Learning

Setting: $X = g(Z)$ where $Z \sim$ linear Gaussian SEM

Enforcing the latent variables $\hat{Z}$ to be Gaussian ensures $\hat{Z} = \mathbf{A}Z + \mathbf{b}$.

- **ID Result:** Using 1 hard int./node, latent variables recovery up to **scaling**
- How to learn a "**general**" encoder?

S. Buchholz, G. Rajendran, E. Rosenfeld, B. Aragam, B. Schölkopf, and P. Ravikumar. *"Learning linear causal representations from interventions under general nonlinear mixing".* NeurIPS'23

Interventional CRL via Contrastive Learning

Contrastive learning: distinguish observational and interventional samples

$$\sum_{i\in\mathcal{I}} \mathcal{L}^{(i)}_{\text{cross-entropy}} + \tau_1(\operatorname{tr}\exp(W\circ W) - n) + \tau_2\|W\|_1$$

linear SEM parameters

enforce linear DAGness

promote sparsity

Experiments: Good performance on both synthetic and image (ball) data (>0.8 MCC)

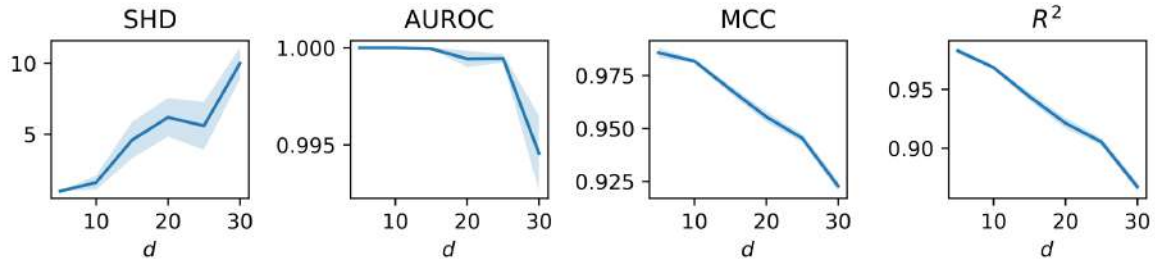
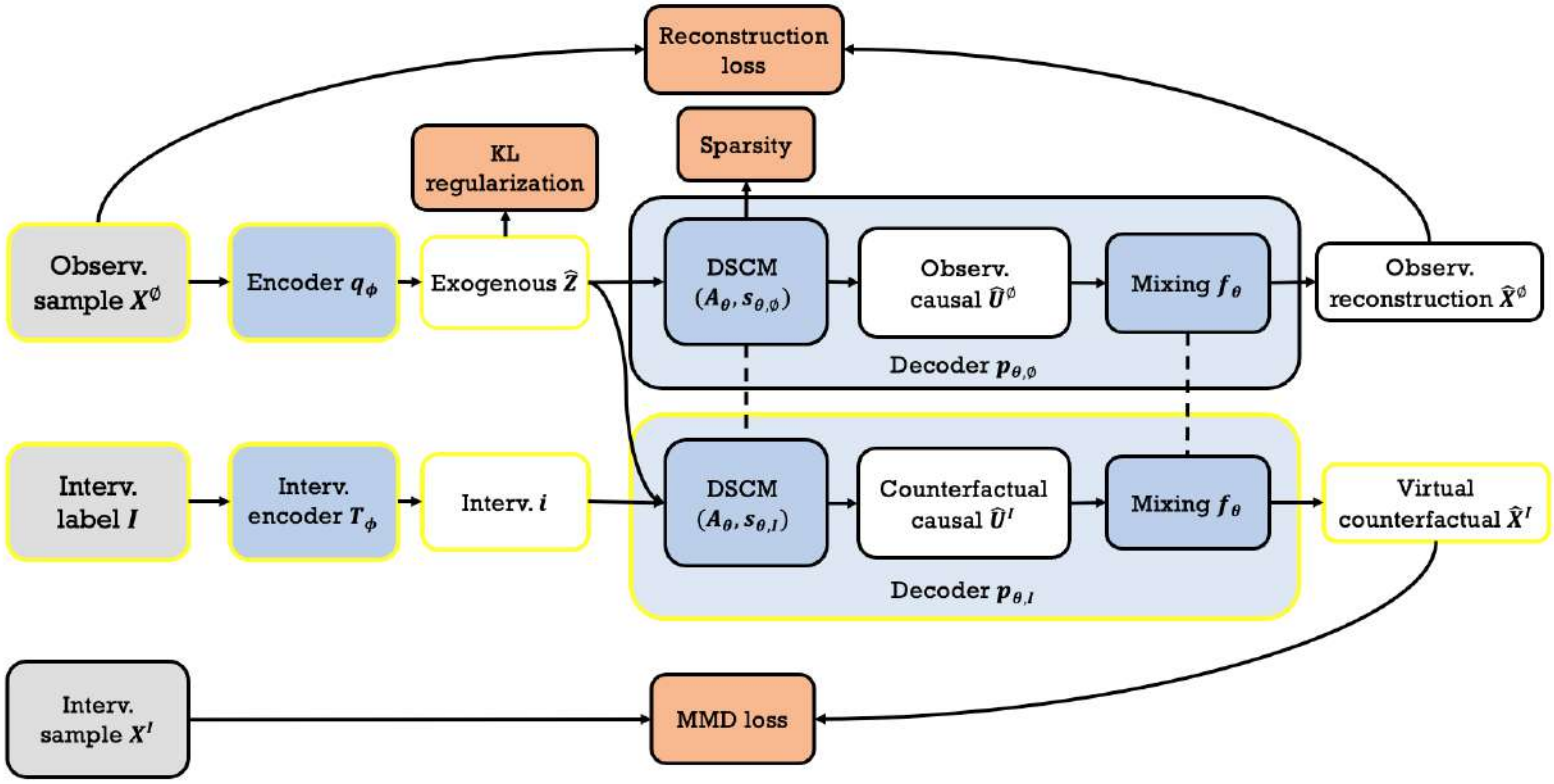


Figure 2: Dependence of performance metrics for $ER(d, 2)$ graphs with $d' = 100$ and nonlinear mixing f on the dimension d.

Synthetic data results

Sample from image dataset

DiscrepancyVAE for Soft Interventions



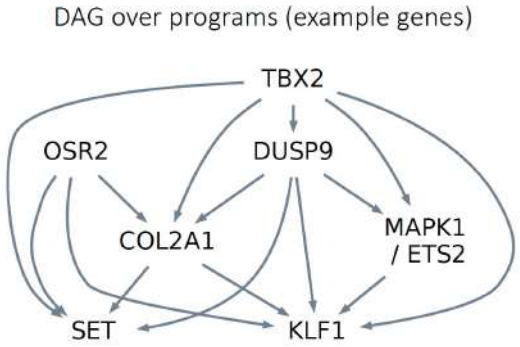
DiscrepancyVAE architecture
Learn transform, SCM and intervention effects altogether

- **Setting:**
  - Linear* transforms
  - nonlinear SCMs
  - 1 soft int./node
- **Result:** ID up to ancestors

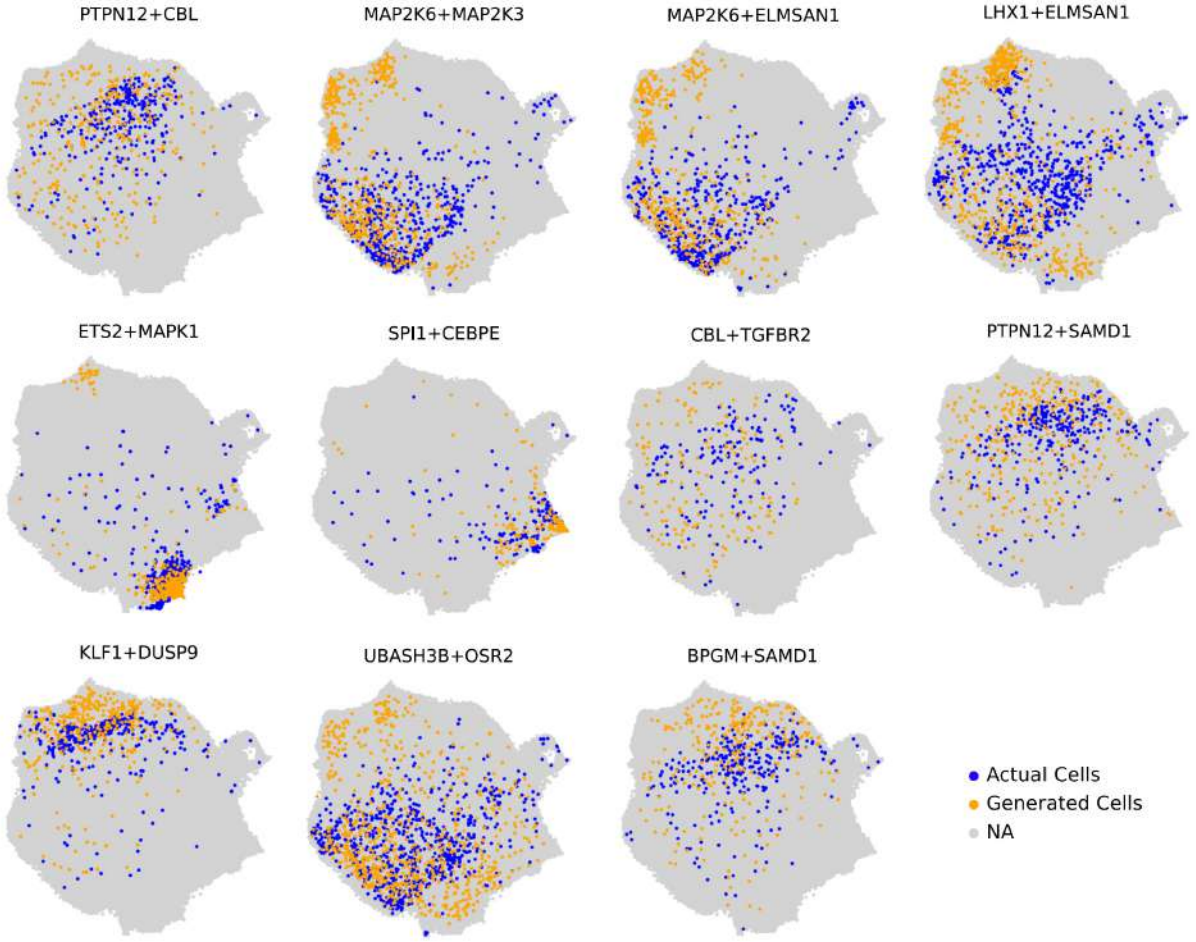
DiscrepancyVAE for Soft Interventions

Experiments:

- Graph estimation and intervention extrapolation on perturb-seq dataset [Norman et al. 2019]



Learned causal graph



Intervention extrapolation

CRL with a Known Graph

If we assume to the graph, things get easier

Theorem. General transform + **one** intervention per node

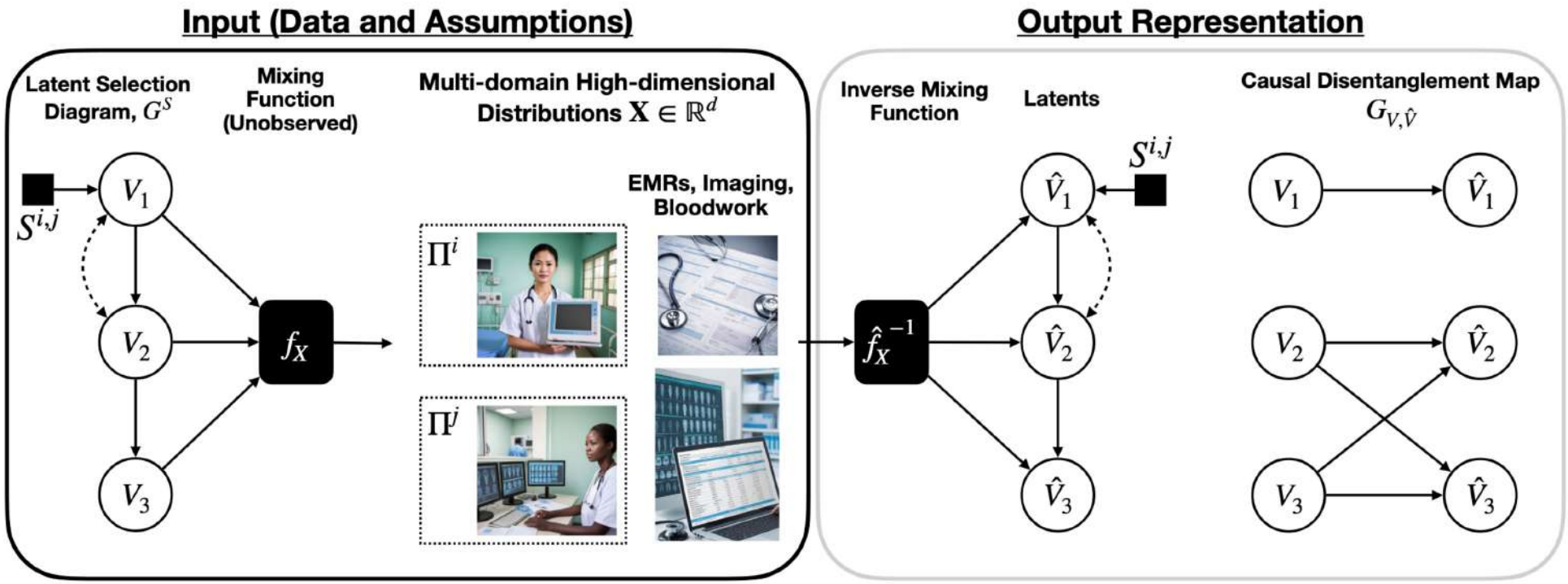
- Soft interventions: ID up to ancestors
- Hard interventions: ID up to elementwise transforms

Algorithm: Maximum likelihood estimation (e.g., via normalizing flows)

CRL with a Known Graph

This idea works even with semi-Markovian graphs (can have **bidirected** edges)

- **Causal Disentanglement Map:** Learn what *can* be **disentangled** from *what*



… but doesn't specify **how** to recover the variables

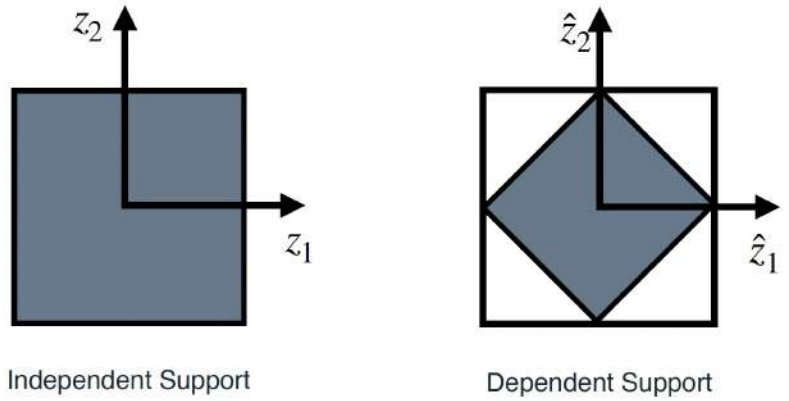
Other Recent Developments

Independent Supports

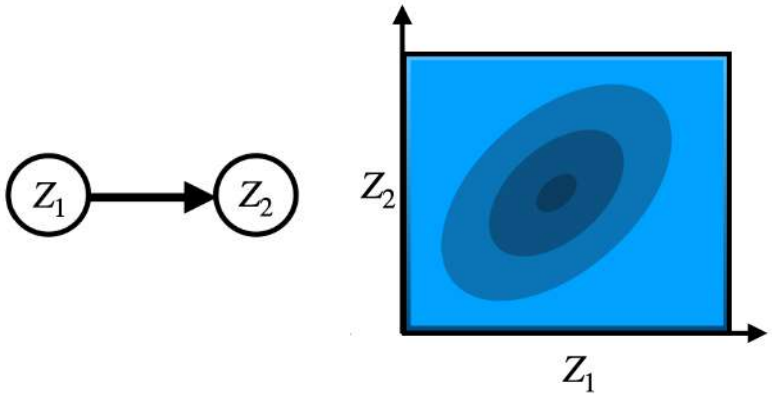
- Assume polynomial g and **independent (bounded) supports:** $\mathcal{Z}^k_{k,m} = \mathcal{Z}^k_k \times \mathcal{Z}^k_m$
- Constraint: $\hat{Z}_k$ and $\hat{Z}_m$ have independent supports.

Theorem. If soft interventions on Z_k and Z_m have independent supports

- block-affine identifiability: $\hat{Z}_k = a^\top \cdot Z + b, \quad$ where $a_m = 0$



Geometric intuition



Statistical independence is not necessary

Multi-domain CRL via Weak Invariances

- Goal: **Identify stable variables in multi-domain data**
  - Stability examples: Marginal distribution invariance, support invariance
- Method: Enforce known invariances

Domain 1 Domain 2

$$p^{(1)}(X_S) = p^{(2)}(X_S)$$

Let S be the stable set. Block-identifiability of stable variables:

$$\hat{Z}_S = f(Z_S)$$

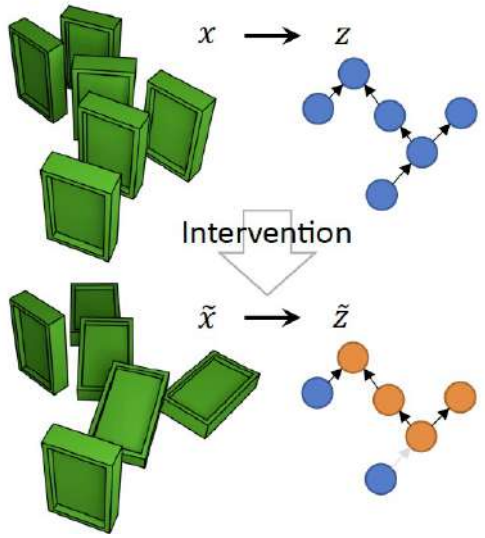
Counterfactual CRL

- Randomly sample from exhaustive, atomic intervention sets

$$\epsilon \sim p_{\mathcal{E}_i}\,, \qquad I \sim p_{\mathcal{I}}\,, \qquad \forall i \in I\,,\ \tilde{\epsilon} \sim p_{\tilde{\mathcal{E}}_i}\,, \qquad \forall i \notin I\,,\ \tilde{\epsilon}_i = \epsilon_i\,,$$

$$\begin{aligned} z &= s(\epsilon)\,, & x &= g(z)\,, \\ \tilde{z} &= \tilde{s}_I(\tilde{\epsilon})\,, & \tilde{x} &= g(\tilde{z})\,. \end{aligned}$$

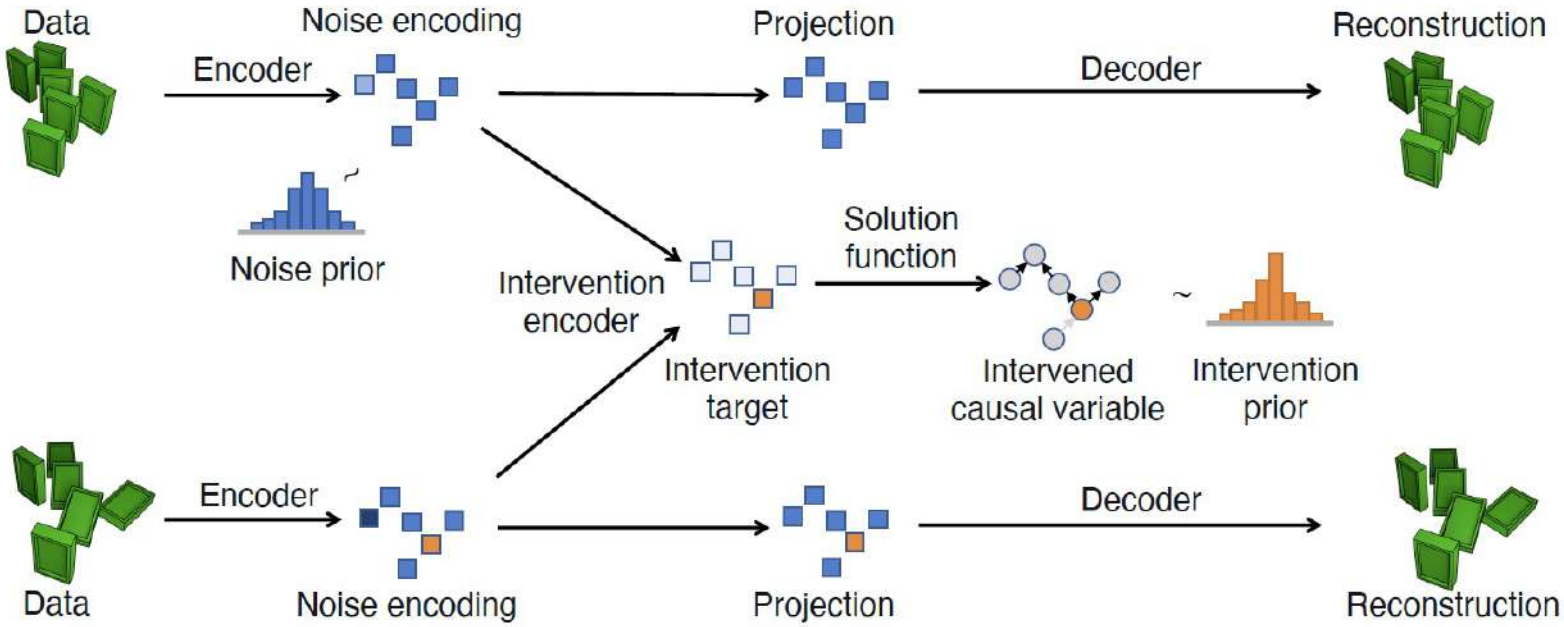
Same sample - two views

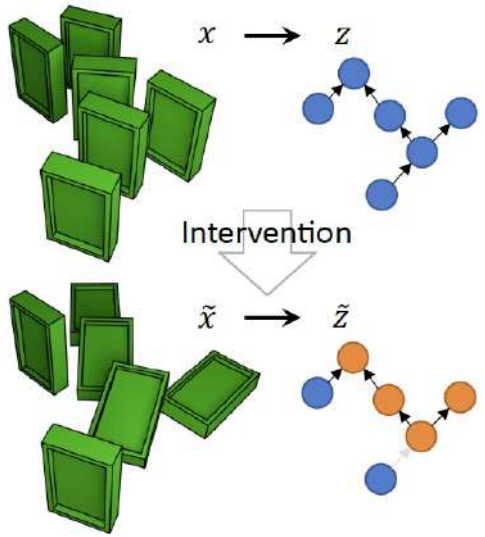


- The causal model is **identifiable**: If an estimated latent SCM produces the same distribution $p(x, \tilde{x})$, then it achieves perfect identifiability

Counterfactual CRL

- Learn a VAE to recover noise and estimate intervention target
- Enforce exogenous noise invariance on learned noise





Multiview CRL with Partial Observability

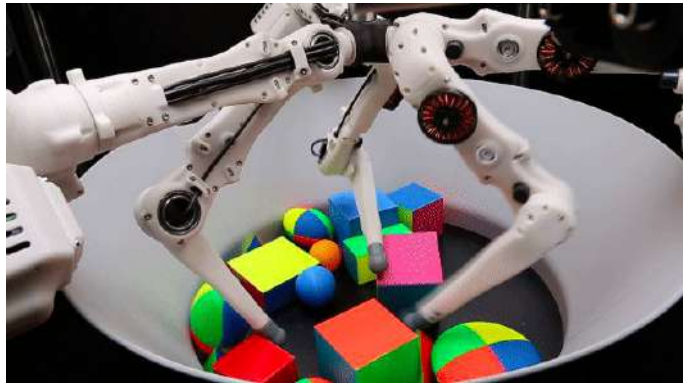
- **Motivation**: multiple-views of the *"same sample"*

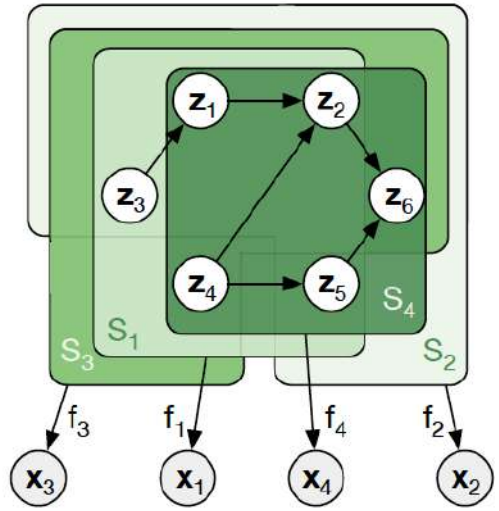
variables & mixing are not fully shared (multiple-cameras, camera+robot sensors)

$$X^k = f_k(Z_{S_k}) \qquad \text{s.t. } S_k \subset [n], \quad \forall k \in [K]$$

Focus: identifying causal variables (rather than the causal relations)

ID target: disentangling the **shared** variables from the rest.





Multiview CRL with Partial Observability

Method: enforce **sample-level invariance** with contrastive learning

$$\min_{h_1,\ldots,h_k} \sum_{j,k} \underbrace{\mathbf{E}\left[\|h_j(X^j) - h_k(X^k)\|\right]_2}_{\text{Content alignment}} - \underbrace{\sum_k \text{entropy}(h_k(X^k))}_{\text{entropy regularization}}$$

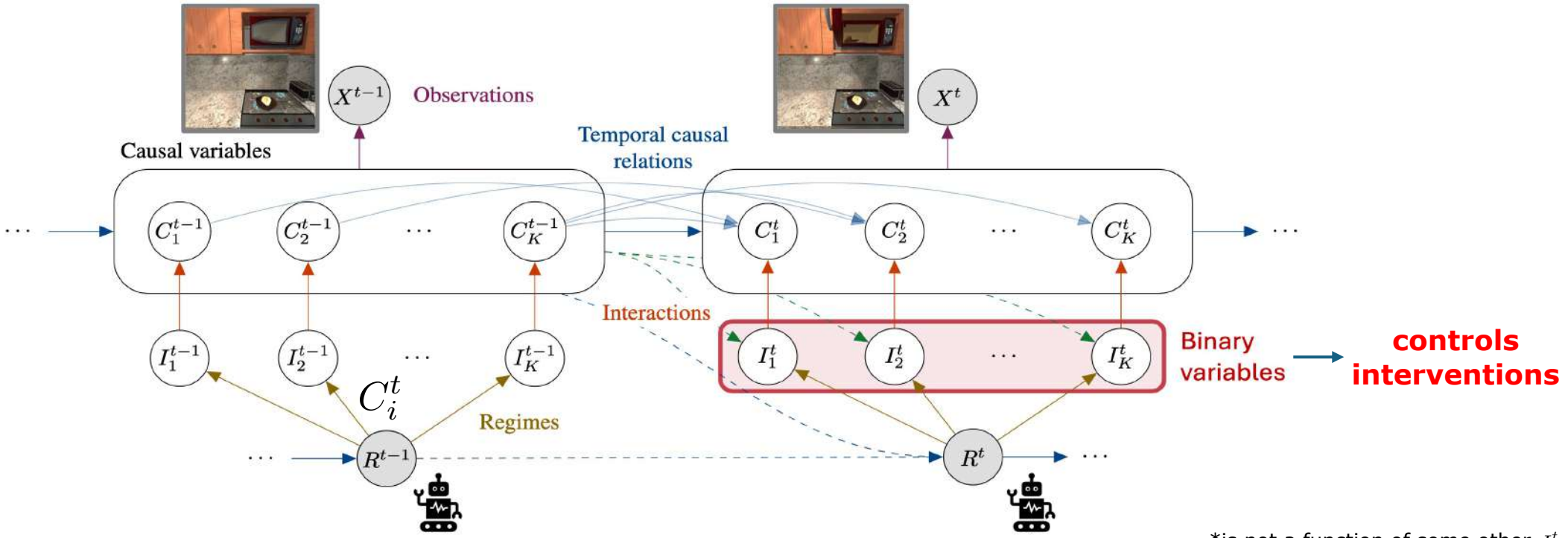
Let $A = \bigcap_k S_k$. Block-identifiability of shared variables:

$$\hat{Z}_A = f(Z_A)$$

*With *"information-sharing"* between different views, loss function can be modified.

Temporal CRL from Interventions

- Latent dynamic Bayesian network, actions are observed
- If each variable has a distinct interaction pattern*, latent variables are **identifiable**



*is not a function of some other I_j^t

Temporal CRL from Interventions

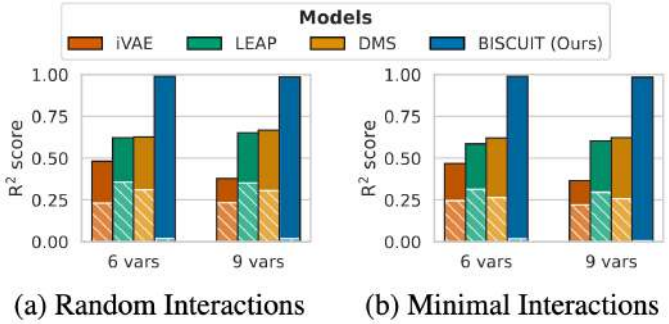
- **Model:** VAE-based*, try to learn intervention targets using MLPs
- **Idea:** Given the intervention targets, distributions should match

$$\mathcal{L}_t = \underbrace{-\mathbb{E}_{q_\phi(z^t|x^t)}[\log p_\theta(x^t|z^t)]}_{\text{Reconstruction}} + \underbrace{\mathbb{E}_{q_\phi(z^{t-1}|x^{t-1})}\left[KL\left(q_\phi(z^t|x^t)||p_\omega(z^t|z^{t-1},R^t)\right)\right]}_{\text{Prior modeling}}$$

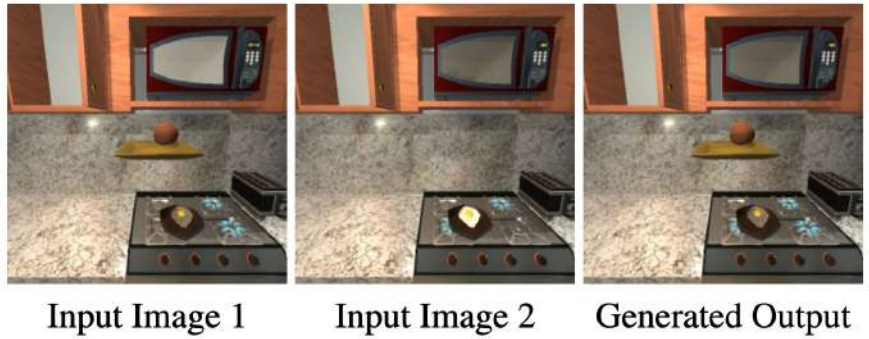
$$p_\omega(z^t|z^{t-1},R^t) = \prod_i p_\omega\left(z_i^t|z^{t-1}, f_i(R^t,z^{t-1})\right)$$

Binary function output

- **Experiments:** Image datasets
  - Voronoi [Lippe et al., 2023], Causal-World [Ahmed et al., 2020], iTHOR [Kolve et al., 2017]



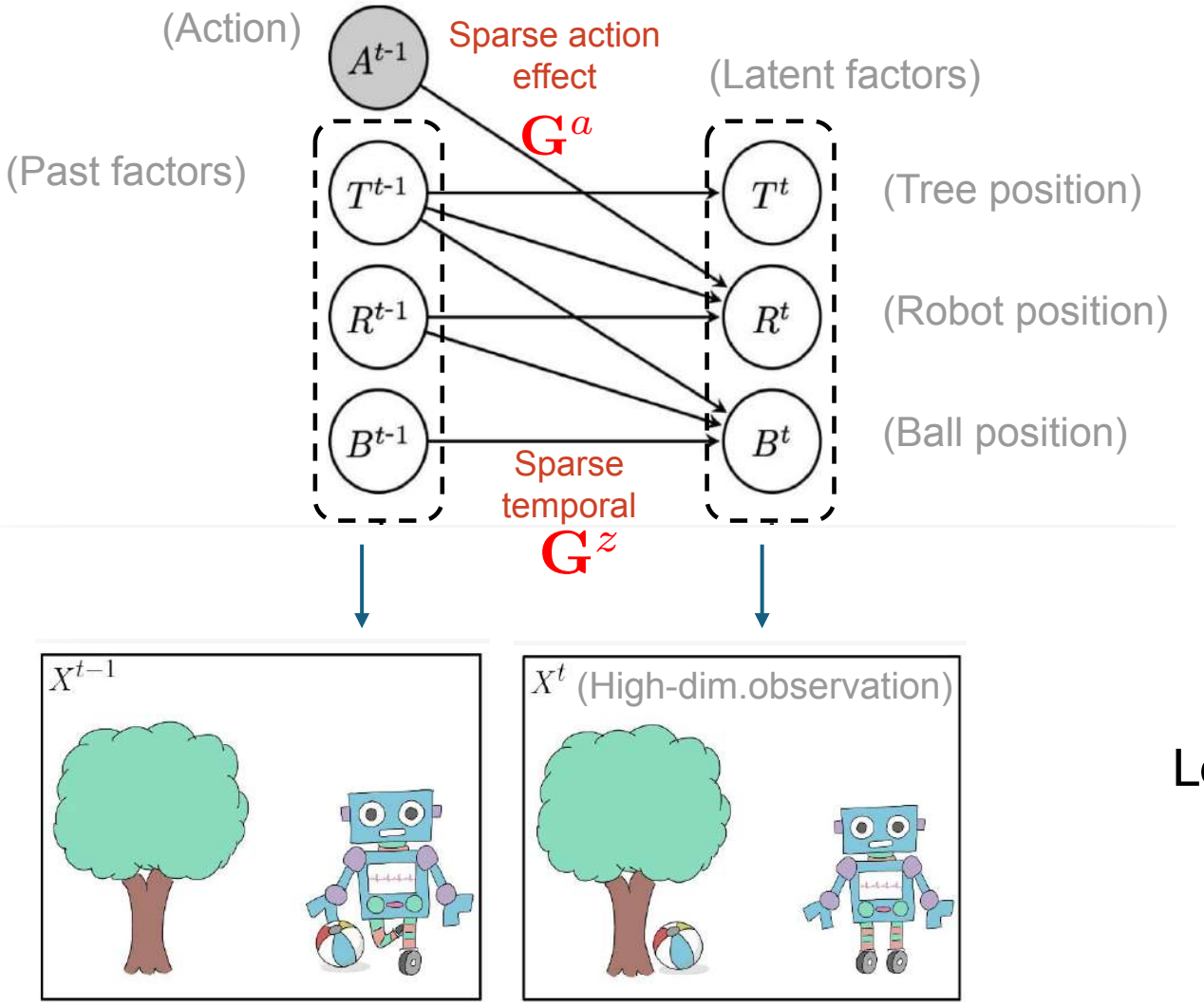
Results for Voronoi benchmark



Intervention extrapolation

*also have VAE + normalizing flow

(Partial) Disentanglement via Mechanism Sparsity



- Latent dynamic Bayesian network, no instantaneous causal relationships
- Actions are observed, conditionally factorial model

$$p(\boldsymbol{z}^t \mid \boldsymbol{z}^{<t}, \boldsymbol{a}^{<t}) = \prod_{i=1}^{d_z} p(\boldsymbol{z}_i^t \mid \boldsymbol{z}_{\mathbf{Pa}_i^z}^{<t}, \boldsymbol{a}_{\mathbf{Pa}_i^a}^{<t})$$

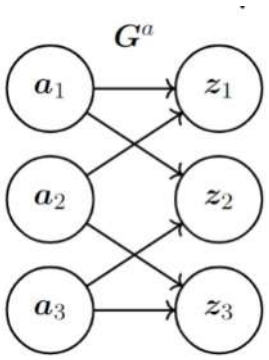
Learnable "parameters" $\theta := (g, p, \mathbf{G}^z)$

learn via VAE framework

(Partial) Disentanglement via Mechanism Sparsity

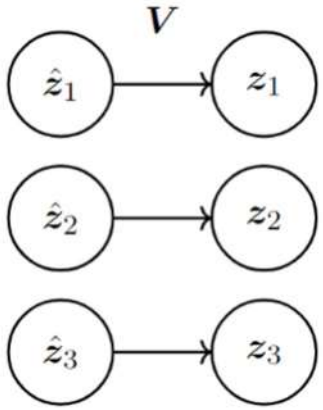
- **Sufficient influence** (informal): two variables don't have the same causes and effects.
- **Sparsity regularization**: Estimated graph is as sparse as true one
- **Importance**: Characterization of "entanglement map" (i.e., partial disentanglement) + also multi-node interventions in the temporal setting

$$\phi := g^{-1} \circ \hat{g}$$ must have sparse dependency structure



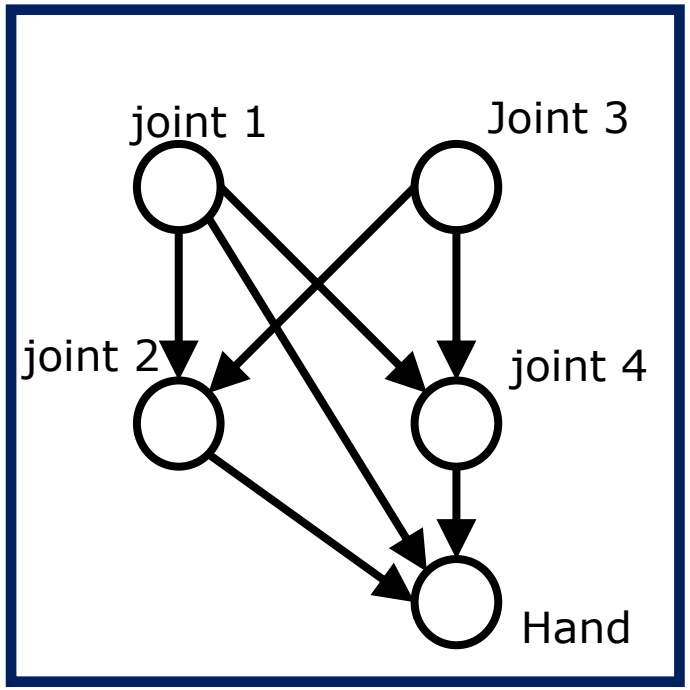
$$\boldsymbol{G}^a = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 1 \end{bmatrix}$$

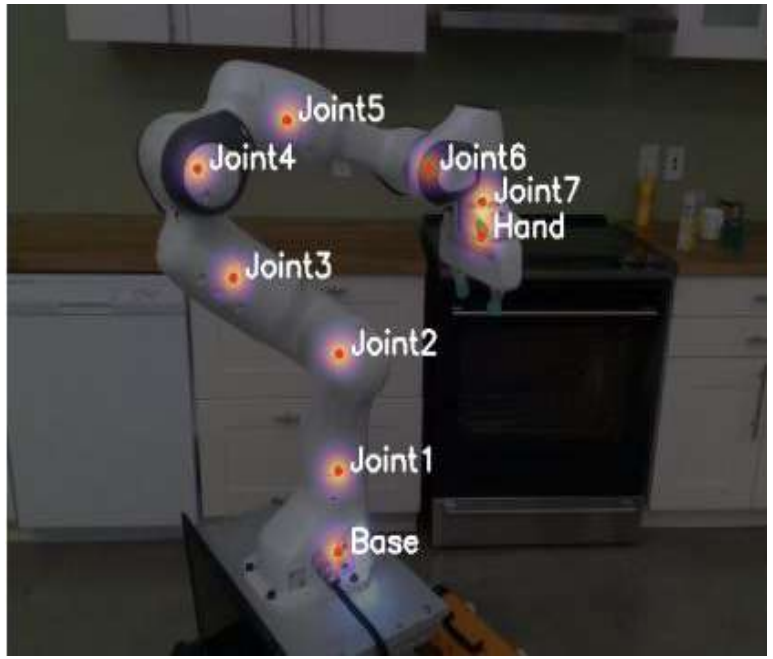
similar to *unknown multi-node interventions*



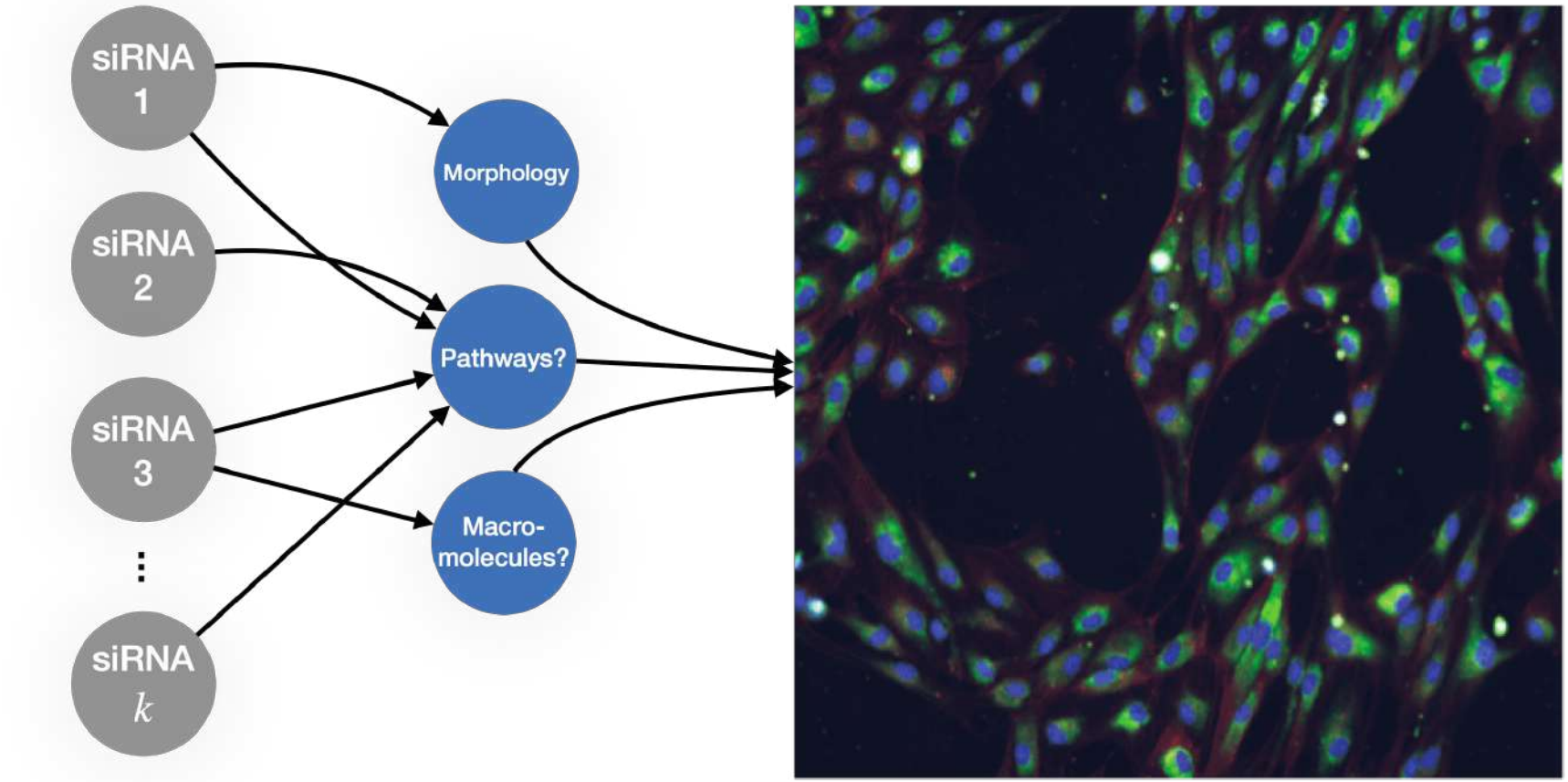
Discussion and Closing Thoughts

Time-varying Dynamical Systems





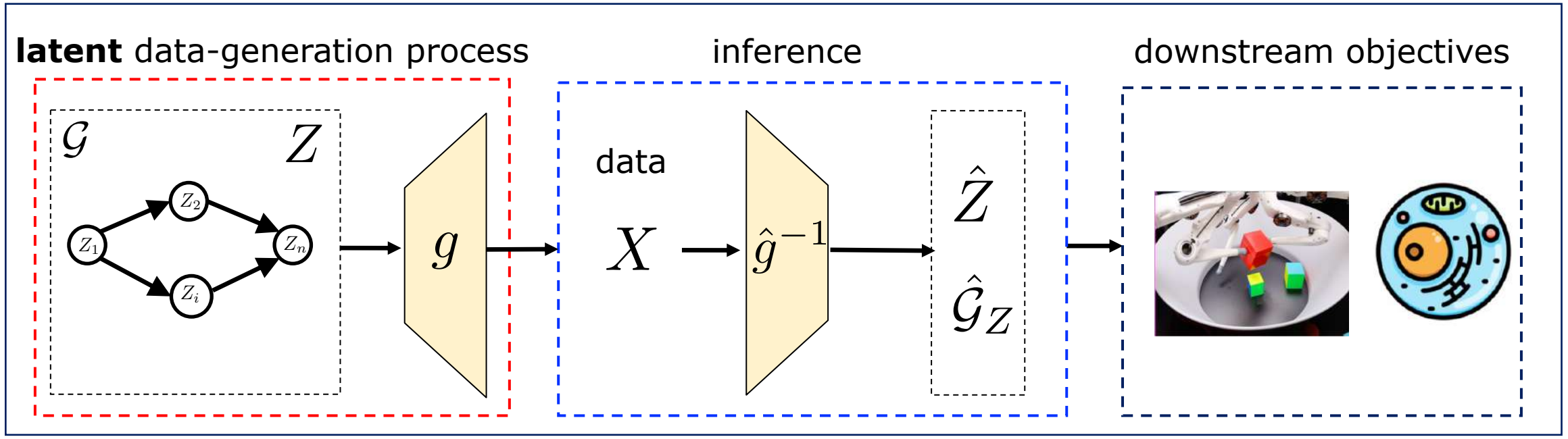
Right Level of Latent Variables



Source: https://www.rxrx.ai/rxrx1 & Jason Hartford

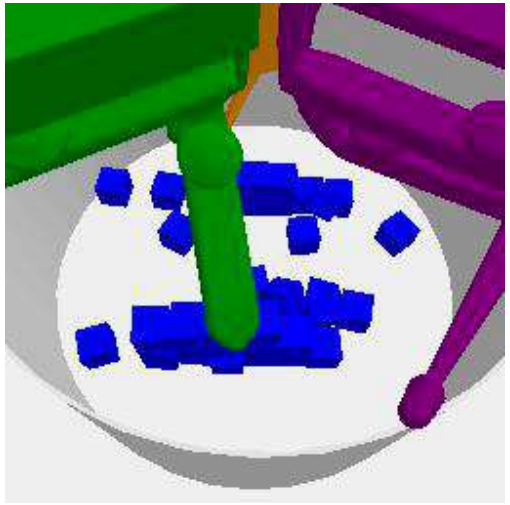
Task-oriented CRL

Is full CRL excessive?

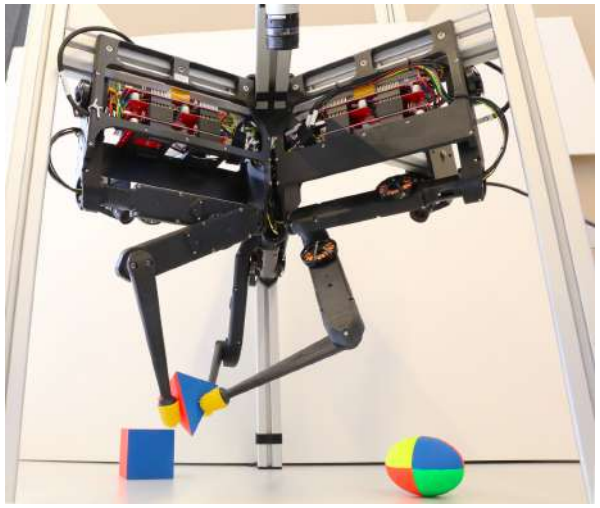


Datasets and Benchmarks

“ImageNet” for CRL?



CausalWorld



TriFinger

ISTAnt

References (at least partly) covered in this tutorial

- B. Varıcı, E. Acartürk, K. Shanmugam, and A. Tajer. "*General identifiability and achievability for causal representation learning*". AISTATS 2024.
- B. Varıcı, E. Acartürk, K. Shanmugam, A. Kumar, and A. Tajer. "*Score-based causal representation: Linear and general transformations*". arXiv: 2402.00849
- B. Varıcı, E. Acartürk, K. Shanmugam, and A. Tajer. "*Linear causal representation learning from unknown multi-node interventions*". NeurIPS 2024.
- E. Acartürk, B. Varıcı, K. Shanmugam, and A. Tajer. "*Sample complexity of interventional causal representation learning*". NeurIPS 2024
- B. Varıcı, E. Acartürk, K. Shanmugam, A. Kumar, and A. Tajer. "*Score-based causal representation learning with interventions*". arXiv: 2301.08230
- B. Schölkopf, F. Locatello, S. Bauer, N. Ke, N. Kalcbrenner, A. Goyal, Y. Bengio. *"Towards Causal Representation Learning".* IEEE 2021
- K. Ahuja, D. Mahajan, Y. Wang, and Y. Bengio. *"Interventional causal representation learning".* ICML 2023
- C. Squires, A. Seigal, S. S. Bhate, and C. Uhler. *"Linear causal disentanglement via interventions".* ICML 2023
- J. von Kügelgen, M. Besserve, W. Liang, L. Gresele, A. Kekić, E. Bareinboim, D. M. Blei, and B. Schölkopf. *"Nonparametric identifiability of causal representations from unknown interventions"*. NeurIPS 2023
- L. Wendong, A. Kekić, J. von Kügelgen, S. Buchholz, M. Besserve, L. Gresele, and B. Schölkopf. "*Causal component analysis"*. NeurIPS 2023
- S. Buchholz, G. Rajendran, E. Rosenfeld, B. Aragam, B. Schölkopf, and P. Ravikumar. *"Learning linear causal representations from interventions under general nonlinear mixing"*. NeurIPS 2023
- J. Zhang, C. Squires, K. Greenewald, A. Srivastava, K. Shanmugam, and C. Uhler. "*Identifiability guarantees for causal disentanglement from soft interventions".* NeurIPS 2023

References (at least partly) covered in this tutorial

- Y. Zhou, J. Shi, and J. Zhu. *"Nonparametric score estimators".* ICML 2020
- Y. Song, S. Garg, J. Shi, and S. Ermon. *"Sliced score matching: A scalable approach to density and score estimation".* UAI 2020
- K.Ahuja, A.Mansouri, Y.Wang. *"Multi-domain causal representation learning via weak distributional invariances".* AISTATS 2024
- P.Lippe, S.Magliacane, S.Löwe, Y.Asano, T.Cohen, E.Gavves, *"BISCUIT: Causal Representation Learning from Binary Interactions".* UAI 2023
- A. Li, Y. Pan, E. Bareinboim, *"Disentangled representation learning in non-Markovian causal systems".* NeurIPS 2024
- S. Lachapelle, P. López, Y. Sharma, K. Everett, R. Le Priol, A. Lacoste, S. Lacoste-Julien *"Nonparametric Partial Disentanglement via Mechanism Sparsity: Sparse Action, Interventions, and Sparse Temporal Dependencies".* arXiv:2401.04890
- K.Zhang, S.Xie, I.Ng, Y.Zheng. *"Causal Representation Learning from Multiple Distributions: A General Setting".* ICML 2024
- J.Brehmer, P. de Haan, P.Lippe, T.Cohen, *"Weakly Supervised Causal Representation Learning".* NeurIPS 2022
- D. Yao, D. Xu, S. Lachapelle, S. Magliacane, P. Taslakian, G. Martius, J. von Kügelgen, F. Locatello. *"Multi-view Causal Representation Learning with Partial Observability".* ICLR 2024
- H. Morioka and Hyvärinen. *"Causal Representation Learning Made Identifiable by Grouping of Observational Variables"*. ICML 2024

Additional references (non-exhaustive)

- I. Khemakhem, D. Kingma, R. Monti, A. Hyvärinen. *"Variational autoencoders and nonlinear ICA: A unifying framework".* AISTATS 2020
- A. Komanduri, X. Wu, Y. Wu, F. Chen. *"From identifiable causal representations to controllable counterfactual generation: A survey on causal generative modeling".* TMLR 2024.
- Y. Jiang and B. Aragam. "*Learning nonparametric latent causal graphs with unknown interventions*". NeurIPS 2023
- J. Jin and V. Syrgkanis. "*Learning Causal Representations from General Environments: Identifiability and Intrinsic Ambiguity*." NeurIPS 2024
- S. Bing, U. Ninad, J. Wahl, and J. Runge. "*Identifying linearly-mixed causal representation learning from multi-node interventions*", CLeaR 2024.
- D. Yao, D. Rancati, R. Cadei, M. Fumero, F. Locatello "*Unifying Causal Representation Learning with the Invariance Principle".* ICLR 2025
- D. Yao, C. Muller, F. Locatello. *"Marrying causal representation learning with dynamical systems for science".* NeurIPS 2024.
- P. Lippe, S. Magliacane, S. Löwe, Y. Asano, T. Cohen, E. Gavves, "*Causal representation learning for instantaneous and temporal effects in interactive systems*". ICLR 2023
- J. von Kügelgen, L. Gresele, Y. Sharma, W. Brendel, B. Schölkopf, M. Besserve, F. Locatello, *"Self supervised learning provably isolates content from style".* NeurIPS 2021
- B. Kivva, G. Rajendran, P. Ravikumar, B. Aragam. *"Identifiability of deep generative models without auxiliary information".* NeurIPS 2021
- S. Lachapelle, D. Mahajan, I. Mitliagkas, S. Lacoste-Julien, *"Additive Decoders for Latent Variables Identification and Cartesian-Product Extrapolation".* NeurIPS 2023

.. and many more

Additional references (non-exhaustive)

- S. Lachapelle, D. Mahajan, I. Mitliagkas, S. Lacoste-Julien, *"Additive Decoders for Latent Variables Identification and Cartesian-Product Extrapolation".* NeurIPS 2023
- F. Locatello, S. Bauer, M. Lucic, G. Raetsch, S. Gelly, B. Schölkopf, O. Bachem. *"Challenging Common Assumptions in the Unsupervised Learning of Disentangled Representations".* ICML 2019
- Y. Liu, Z. Zhang, D. Gong, M. Gong, B. Huang, A. van den Hengel, Q. Shi, *"Identifiable latent polynomial causal models through the lens of change".* ICLR'2024
- Q. Xi, B. Bloem-Reddy, *"Indeterminacy in generative models: Characterization and strong identifiability".* AISTATS 2023
- A. Hyvärinen, H. Moriola. *"Unsupervised feature extraction by time-constrastive learning and nonlinear ICA".* NeurIPS 2016
- K. Ahuja, J. Hartford, Y. Bengio. *"Properties from mechanisms: An equivariance perspective on identifiable representation learning".* ICLR 2022
- K. Ahuja, J. Hartford, Y. Bengio. *"Weakly supervised representation learning with sparse perturbations".* NeurIPS 2022
- X. Shen, F. Liu, H. Dong, Q. Lian, Z. Chen, T. Zhang. *"Weakly supervised disentagled generative causal representation learning".* JMLR 2022
- G. Chen, Y. Shen, Z. Chen, X. Song, Y. Sun, W. Yao, X. Liu, K. Zhang. *"CaRiNG: Learning temporal causal representation under non-invertible generation process".* ICML 2024
- R. Welch, J. Zhang, C. Uhler. *"Identifiability Guarantees for Causal Disentanglement from Purely Observational Data".* NeurIPS 2024

.. and many more