Causal Inference for Computer Scientists: Challenges and Opportunities

Saber Salehkaleybar

LIACS, Leiden University

Leiden $\leftrightarrow$ Chile Research Seminar

27 July 2026 ▪ FCFM, Universidad de Chile

Plan for today (3 hours)

1. **Foundations of causal inference** *45 min*
   SCMs, interventions, counterfactuals

   Break + questions 15 min

2. **Challenges** + hands-on *45 min*
   Confounding, identifiability, biased data (Python)

   Break + questions 15 min

3. **Applications in modern AI** *45–60 min*
   Reinforcement learning, interpretable/generative AI, gene perturbation prediction

What you'll take away

A working model for asking and answering *interventional* and *counterfactual* questions.

Outline

1. Part 1 — Foundations of Causal Inference
2. Part 2 — Challenges & Hands-on
3. Part 3 — Applications in AI

Part 1 — Foundations of Causal Inference

Correlation vs. Causation

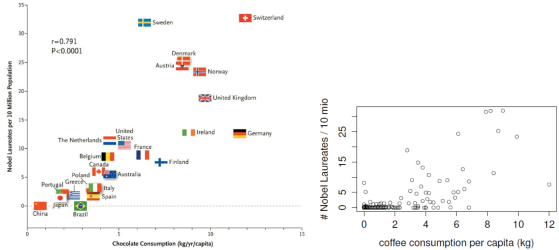
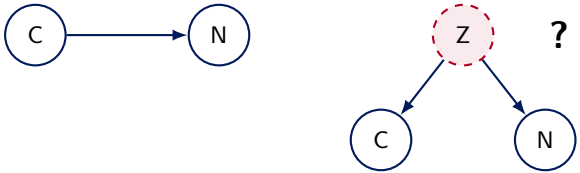


Figure 1: Nobel laureates vs. chocolate consumption per country



Does chocolate cause Nobel prizes or is there a common cause Z (e.g. wealth)?

Correlation vs Causation (in Public)

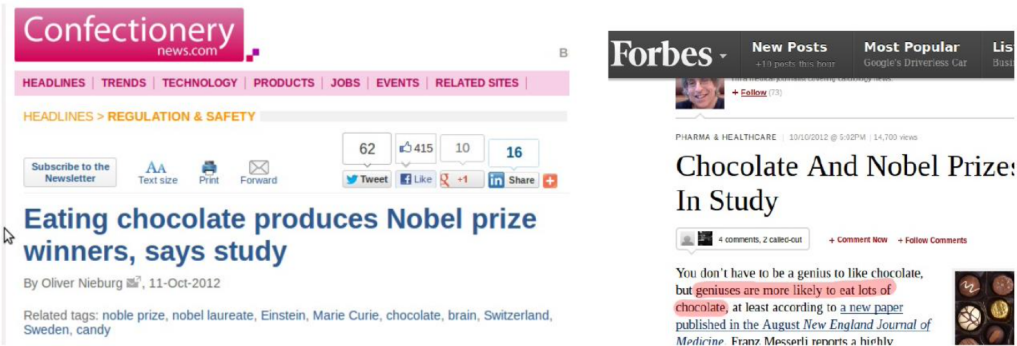
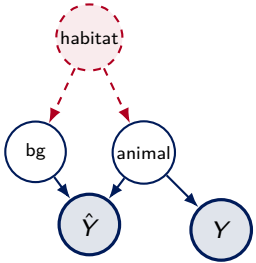


Figure 2: Two online articles drawing causal conclusions from the observed correlation.

Spurious Correlation in ML

The model labels this photo of **cows on a beach** as *"camel."*

- In training, cows appear on *green background*, camels on *sandy backgrounds*.
- The model attends to **background colour**, a shortcut correlated with the label, not the animal.
- For the cows living on the beach, the prediction breaks.



Habitat confounds background & animal; model uses bg

When Statistics May Be Misleading (Simpson's Paradox)

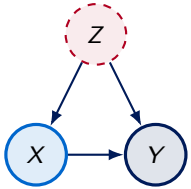
Observational study: 700 patients, 350 took a drug, 350 did not.

	Drug	**No drug**
Men	81/87 (93%)	234/270 (87%)
Women	192/263 (73%)	55/80 (69%)
Combined	273/350 (78%)	289/350 (83%)

- Drug helps *both* men and women, but appears *worse* overall.
- Should a doctor prescribe it? The answer needs the **causal story**, not summary statistics.

What Explains the Reversal? "A Causal Story"

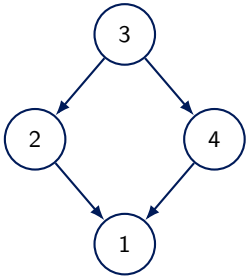
- Women recover less often *regardless* of the drug.
- Women are much more likely to take the drug (263 vs. 80).
- Gender is a **common cause** of taking the drug (X) and recovery (Y).
- Comparing drug vs. no-drug without accounting for gender mixes different populations.



X: Drug, Y: Recovery, Z: Gender

Graphs: A Quick Vocabulary

- Nodes = variables; edges = direct causes ($X \to Y$).
- PA_j / CH_j: parents / children of node j.
- **Directed path**: follow the arrows; gives ancestors / descendants.
- **v-structure** (immorality): $i \to k \leftarrow j$.
- **DAG**: directed graph with no cycles.



node 3 is a root; $\mathrm{PA}_1 = \mathrm{CH}_3 = \{2, 4\}$

Structural Causal Models (Bivariate Case)

Two sets of variables: **exogenous** noises and **endogenous** variables.

Definition: SCM with graph $C \to E$

$$C := N_C, \qquad E := f_E(C, N_E), \qquad N_C \perp\!\!\!\perp N_E.$$

- C is the direct cause of E.
- N_C, N_E are the exogenous noises.
- *Independence of cause and mechanism*: $N_C \perp\!\!\!\perp N_E$.

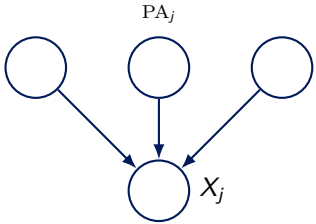
Structural Causal Models (Multivariate Case)

Definition: SCM $\mathcal{C} := (S, P_N)$

A collection of d assignments

$$X_j := f_j(\mathrm{PA}_j, N_j), \quad j = 1, \ldots, d, \qquad N_1, \ldots, N_d \text{ jointly independent.}$$

- The graph = which PA_j appear on each right-hand side.
- *Principle of Independent Mechanisms.*
- Induces the observational distribution over the variables.

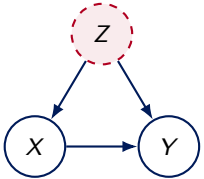


Interventions: What We Really Want

- Most studies are really about the effect of **interventions**:
  - treating illness by administering medication,
  - reducing wildfires by changing controllable factors.
- **Randomized controlled trials** are the gold standard (randomization balances other causes).
- But RCTs are often impossible: ethics, feasibility, cost, compliance.

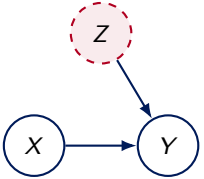
The do-Operator: Intervention as Graph Surgery

Observe $X = x$



back-door path $X \leftarrow Z \rightarrow Y$ stays open

Do $\mathrm{do}(X := x)$



incoming edges to X are removed

$$P(Y \mid \mathrm{do}(X := x)) \neq P(Y \mid X = x) \quad \text{in general.}$$

Intervening vs. Conditioning

Conditioning $P(Y \mid X = x)$

- We do *not* change the world.
- Restrict attention to cases where $X = x$.
- Causal system stays intact.

Intervention $P(Y \mid \mathrm{do}(X := x))$

- We *set* X to x, overriding its causes.
- Changes the data-generating process.
- Graph surgery: cut incoming arrows.

Interventional Distribution

Definition: Interventional distribution

Replace one (or several) assignment in $\mathcal{C}$ to get a new SCM $\tilde{\mathcal{C}}$:

$$X_k := \tilde{f}(\tilde{\mathrm{PA}}_k, \tilde{N}_k).$$

The entailed distribution is written $P_X^{\mathcal{C};\, \mathrm{do}(X_k := \cdots)}$.

- **Hard intervention**: $X_k := c$ (a constant).
- **Soft/imperfect**: replace f_k but keep some parents.

Counterfactuals: Example

While driving home last night, I came to a fork in the road, where I had to make a choice: to take the freeway ($X = 1$) or go on a surface street named James Boulevard ($X = 0$). I took James, only to find out that the traffic was touch and go. As I arrived home, an hour later, I said to myself: "Gee, I should have taken the freeway."

Example: What Does "I Should Have" Mean?

Let Y = driving time. The factual world is fixed: $X{=}0$ (James), $Y{=}60$ min.

- **Not** a question about drivers in general: $P(Y \mid \mathrm{do}(X{:=}1))$ averages over *everyone* who takes the freeway.
- **The regret is personal**: same night, same weather, same traffic pattern; only *my* choice changes. That is the counterfactual

$$Y_{\mathrm{do}(X:=1)} \mid \underbrace{X{=}0,\ Y{=}60}_{\text{what actually happened}} .$$

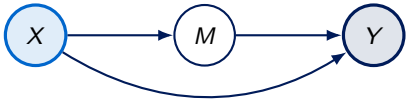
- "I should have" = I believe $Y_{\mathrm{do}(X=1)} < 60$ for *that* realization of all other factors; the same noise U that produced my actual trip.

Counterfactuals in Three Steps

1. **Abduction** — update $P(N)$ on the observed evidence.
2. **Action** — apply $\mathrm{do}(\cdot)$ to the model.
3. **Prediction** — read off the counterfactual quantity $P(Y_x \mid X', Y')$.

Worked Example: Counterfactual in a 3-Variable ANM

$$X := N_X$$
$$M := 2X + N_M$$
$$Y := M + X + N_Y$$



Observed: $X'=1,\ M'=3,\ Y'=5$.

Ask: *had $X=0$, what would Y be?*

1. **Abduction**:

$$N_X = 1, \quad N_M = M' - 2X' = 1, \quad N_Y = Y' - M' - X' = 1.$$

2. **Action** — apply $\mathrm{do}(X = 0)$, *keep* N_M, N_Y:

$$X := 0.$$

3. **Prediction** — recompute downstream:

$$M_{X=0} = 2 \cdot 0 + N_M = 1,$$

$$Y_{X=0} = M_{X=0} + 0 + N_Y = 1 + 0 + 1 = \mathbf{2}.$$

The Three-Layer Causal Hierarchy

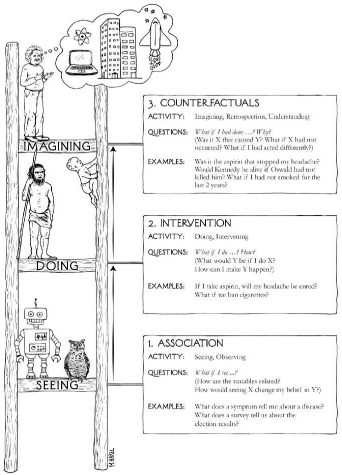


Figure 3: Ladder of Causation

1. **Association** $P(Y \mid X)$ *seeing*
2. **Intervention** $P(Y \mid \text{do}(X))$ *doing*
3. **Counterfactual** $P(Y_x \mid X', Y')$ *imagining*

- Most ML lives on Layer 1 (association).
- This tutorial: Layers 2 & 3.

Some Terminology

Causal effect identification

Given the graph, estimate the effect of an intervention:

$$P(Y \mid \text{do}(X := x)) = ?$$

Causal discovery

Learn the graph itself from data:

$$P_{\mathbf{X}} \longrightarrow \hat{G}$$

Two frameworks

Structural Causal Models: networks of variables, a CS flavor.
Potential Outcomes: social science / medicine.
We focus on SCMs.

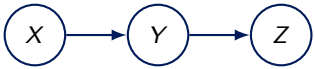
Break

15 minutes ▪ questions welcome

Part 2 — Challenges & Hands-on

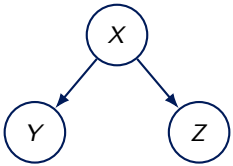
Three Building Blocks: Chain, Fork, Collider

Chain



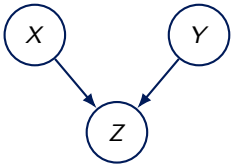
$X \perp\!\!\!\perp Z \mid Y$

Fork



$Y \perp\!\!\!\perp Z \mid X$

Collider



$X \perp\!\!\!\perp Y$, but *not* given Z

Collider intuition ($Z = X + Y$, $X \perp\!\!\!\perp Y$)

Knowing $X = 3$ tells nothing about Y. But if $Z = 10$ *and* $X = 3$, then $Y = 7$: conditioning on a collider **creates** dependence.

d-Separation

Definition: d-separation

A path is **blocked** by a set S if it has a node i_k with either:

(i) $i_k \in S$ and it is a chain $(\rightarrow i_k \rightarrow)$ or fork $(\leftarrow i_k \rightarrow)$; or

(ii) i_k is a collider $(\rightarrow i_k \leftarrow)$ and neither i_k nor its descendants are in S.

$A \perp\!\!\!\perp_G B \mid S$ if *every* path between A and B is blocked by S.

Markov property

If P_X is induced by an SCM with graph G, then $A \perp\!\!\!\perp_G B \mid S \Rightarrow A \perp\!\!\!\perp B \mid S$: graph structure entails (conditional) independence in the distribution.

Markov Equivalence Class

- Different DAGs can imply the *same* independencies.

Theorem: Graphical criterion

Two DAGs are Markov equivalent iff they have the same **skeleton** and the same **v-structures**.

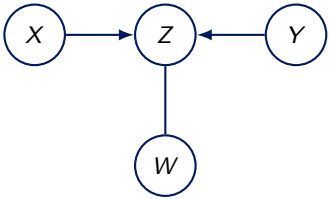
- A Markov equivalence class (MEC) is summarized by a **CPDAG**: edges shared by all members are directed; others undirected.
- From observational data alone, we can generally recover only the CPDAG.

$X - Y - Z$: three DAGs, one CPDAG

Finding the MEC: the IC Algorithm

Inductive Causation (Verma & Pearl, 1991) recovers the CPDAG of the Markov equivalence class from conditional-independence tests alone.

1. **Skeleton.** For each pair (X_i, X_j), join them by an undirected edge *unless* some set S_{ij} makes $X_i \perp\!\!\!\perp X_j \mid S_{ij}$. Record that S_{ij}.
2. **v-structures.** For every path $X_i - X_k - X_j$ with X_i, X_j non-adjacent, orient $X_i \to X_k \leftarrow X_j$ *iff* $X_k \notin S_{ij}$.
3. **Propagate.** Apply Meek's rules: orient remaining edges wherever the alternative would create a new v-structure or a cycle.



$X \perp\!\!\!\perp Y,\ Z \notin S_{XY} \Rightarrow$ v-structure

$X \to Z \leftarrow Y$

Output: a **CPDAG**: directed edges shared by all DAGs in the MEC, undirected where the direction is unidentifiable. (PC is its practical, order-efficient variant.)

Causal Discovery: Learning the Graph

- **No assumptions** $\Rightarrow$ recover only the MEC (CPDAG).
  - Constraint-based: IC, PC (conditional independence tests)
  - Score-based: GES
- **Functional assumptions** $\Rightarrow$ full DAG identifiable:
  - **LiNGAM**: linear, non-Gaussian noise (Shimizu et al. 2006)
  - **ANM**: additive noise, nonlinear f_j (Peters et al. 2014)

In Part 2

So far every variable was observed. What if some are **hidden**?

Causal Discovery in the Bivariate Case: ANM

With only $P(X, Y)$, both $X \to Y$ and $X \leftarrow Y$ are Markov equivalent; independence tests *cannot* orient a single edge. Restricting the mechanism breaks the tie.

Definition: Additive Noise Model (ANM)

$Y := f(X) + N_Y, \quad$ with $N_Y \perp\!\!\!\perp X$ and f nonlinear.

Identifiability (Hoyer et al. 2009)

For generic (f, P_X, P_N) there is *no* backward ANM $X = g(Y) + N_X$ with $N_X \perp\!\!\!\perp Y$.
Exception: f linear *and* both noises Gaussian.

Algorithm for a pair (X, Y):

1. Regress Y on $X \Rightarrow \hat{f}$; residual $R_{\to} = Y - \hat{f}(X)$.
2. Test $R_{\to} \perp\!\!\!\perp X$ (HSIC); score $s_{\to}$.
3. Regress X on $Y \Rightarrow \hat{g}$; $R_{\leftarrow} = X - \hat{g}(Y)$; test $R_{\leftarrow} \perp\!\!\!\perp Y$; score $s_{\leftarrow}$.
4. $s_{\to} \gg s_{\leftarrow} \Rightarrow X \to Y$.

Identifying Effects: Valid Adjustment Set

Suppose that the causal graph is given (from domain expert or the output of causal discovery)

Definition: Valid adjustment set

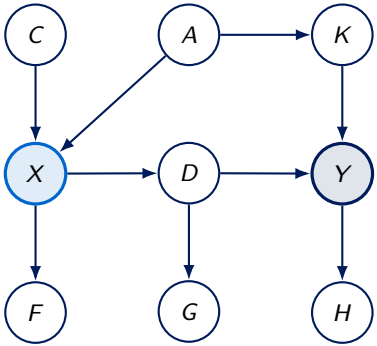
Z is valid for (X, Y) if

$$P(Y \mid \text{do}(X := x)) = \sum_z P(Y \mid X = x, Z = z)\, P(Z = z).$$

Back-door criterion

Z contains no descendant of X and blocks every back-door path from X to Y.

Example: Back-door Adjustment on a Graph



Goal: $P(Y \mid \mathrm{do}(X = x))$. Back-door path

$$X \leftarrow C, \quad X \leftarrow A \rightarrow K \rightarrow Y$$

must be blocked *without* conditioning on a descendant of X.

Valid adjustment sets

$Z = \{K\}$ or $Z = \{C, A\}$.

- D, F, G, H are **descendants of** X; conditioning on them violates the first back-door condition.
- $\{K\}$ blocks $X \leftarrow A \rightarrow K \rightarrow Y$.

Example: Kidney Stones (Adjustment in Action)

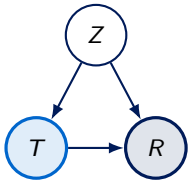
Two treatments a, b; Z = stone size.
$P(Z{=}\text{small}){=}0.51$, $P(Z{=}\text{large}){=}0.49$.

	Overall	Small	Large
a	78%	93%	73%
b	83%	87%	69%

Naive (marginal): $b\,(0.83) > a\,(0.78)$.
Adjusted via back-door on Z:

$$P^{\text{do}(a)}(R{=}1) = 0.93(0.51) + 0.73(0.49)$$
$$= 0.474 + 0.358 = \mathbf{0.83}$$
$$P^{\text{do}(b)}(R{=}1) = 0.87(0.51) + 0.69(0.49)$$
$$= 0.444 + 0.338 = \mathbf{0.78}$$

$$\text{ACE} = 0.83 - 0.78 = +0.05 \;\Rightarrow\; a \text{ wins.}$$

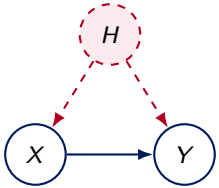


$Z \to T$: doctors gave a to harder (large-stone) cases.

Conditioning *keeps* that bias; the do-adjustment reweights by $P(Z)$ and **flips** the ranking.

Where Causal Inference Gets Hard

- **Latent confounding** — hidden common causes.
- **Biased observational data** — selection, missingness.
- **Feedback loops**



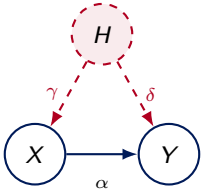
hidden H: back-door adjustment fails

Challenge 1 — Latent Confounding

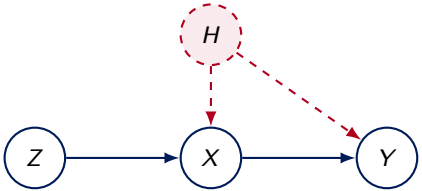
- With hidden H, regressing Y on X gives

$$\frac{\mathrm{Cov}(X,Y)}{\mathrm{Var}(X)} = \alpha + \underbrace{\frac{\delta\,\mathrm{Var}(H)}{\mathrm{Var}(X)}}_{\neq 0 \text{ bias}}$$

- The naive estimate is biased away from the true effect α.



Instrumental Variables



Z is an instrument if

(a) $Z \perp\!\!\!\perp H$

(b) $Z \not\perp\!\!\!\perp X$

(c) Z affects Y only through X

Two-stage estimate (linear)

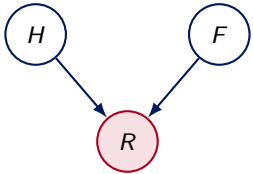
(1) regress X on $Z \Rightarrow \hat{\beta}$; (2) regress Y on $\hat{\beta}Z \Rightarrow \hat{\alpha}$. *Example:* proximity to college as an instrument for education $\rightarrow$ earnings (confounder: ability).

Challenge 2 — Biased Data

- **Selection bias** = conditioning on a collider (Berkson's paradox).

Berkson toy model

$H, F \sim \mathrm{Ber}(0.5)$ independent, $R = \min(H, F) \oplus N_R$ where $N_R \sim \mathrm{Ber}(0.1)$. Then $H \not\perp\!\!\!\perp F \mid R{=}0$: selecting on R induces a spurious correlation.



$H \perp F$ marginally; conditioning on the collider R (selection), they become dependent.

Hands-on

Three short Python experiments

Hands-on Setup (Python)

```python
# pip install dowhy causal-learn numpy pandas networkx
import numpy as np, pandas as pd, networkx as nx
from dowhy import CausalModel
```

Notebook: `https://github.com/sabersalehk/Causal-Inference-Seminar`

Ex. 1 ANM causal discovery (bivariate) on the Tübingen cause-effect pairs
Ex. 2 Causal discovery with DoWhy
Ex. 3 Finding a valid adjustment set with DoWhy

scan for the notebook

Exercise 1 — ANM Causal Discovery (Bivariate)

```
from causallearn.search.\
  FCMBased.ANM.ANM import ANM
# x, y : one cause-effect pair
anm = ANM()
p_fwd, p_bwd = anm.cause_or_effect(
    x.reshape(-1,1),
    y.reshape(-1,1))
# larger residual-indep p-value
# => that direction is preferred
print(p_fwd, p_bwd)
```

Data: `webdav.tuebingen.mpg.de/cause-effect/` (108 pairs).

Idea (ANM): fit $Y = f(X) + N$; the true direction leaves residuals *independent* of the input.

- Load a pair from the Tübingen cause-effect dataset.
- Regress both directions; test residual independence (HSIC).
- Predict the causal direction; check against ground truth.

Exercise 2 — Causal Discovery with DoWhy

```
from dowhy import CausalModel
import networkx as nx
# 1) discover a graph (PC via
#    causal-learn), export to DOT
from causallearn.search.\
  ConstraintBased.PC import pc
cg = pc(df.values)
# 2) hand the learned graph to DoWhy
model = CausalModel(df, "X", "Y",
    graph=learned_dot_string)
model.view_model()
```

Task

- Simulate multivariate data from a known DAG.
- Recover the CPDAG, convert it to a DoWhy graph.
- Compare recovered vs. true graph (SHD); note undirected edges (MEC).

Discuss

Why can only the CPDAG be identified?

Exercise 3 — Valid Adjustment Set with DoWhy

```
model = CausalModel(df, "X", "Y",
                    graph=dag_dot)
est = model.identify_effect()
print(est.get_backdoor_variables())
res = model.estimate_effect(est,
  method_name=
    "backdoor.linear_regression")
print(res.value)
```

Task

- Build the kidney-stone / confounded DAG from Part 1.
- Let DoWhy *identify* the back-door adjustment set.
- Estimate the ATE; compare adjusting vs. not adjusting for Z.

Discuss

What if a needed confounder is unobserved?

Break

15 minutes ▪ questions welcome

Part 3 — Applications in AI

Why Causality for AI?

- Robustness under distribution shift.
- Credit assignment and transfer in **reinforcement learning**.
- **Mechanistic interpretability** of generative models.

This part

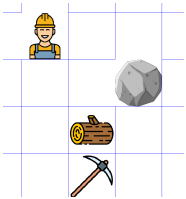
Three threads: (a) causality $\times$ RL, (b) causality $\times$ gene perturbation prediction, (c) causality $\times$ interpretable generative AI.

Thread 1: Causality in Reinforcement Learning

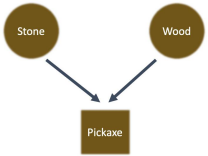
- MDPs as SCMs; actions as interventions do(a).
- Hierarchical RL, causal credit assignment.

Hierarchical RL (HRL)

Long-horizon, sparse-reward tasks are decomposed into a hierarchy of subgoals. The challenge: *discover* the subgoal structure and *explore* it efficiently.



The mini-craft game



Subgoal structure of mini-craft game

Figure 4: Mini-craft: subgoals stone + wood $\rightarrow$ pickaxe (subgoal structure)

HRC: the algorithm (one iteration of the loop)

Three sets of subgoals: **CS** controllable (learned to reach), **IS** intervention (currently held ON), **CCS** reachable (newly discovered).

1. **Select** a subgoal g_{sel} from CS and move it into IS *(which one = the key question)*
2. **Intervention**: play, forcing every subgoal in IS to stay achieved; log trajectories D_I.
3. **Causal discovery (SSD)**: from D_I, estimate the subgoal graph $\hat{\mathcal{G}}$.
4. **Reachable set** CCS = subgoals whose parents are all now in IS.
5. **Train** policies for CCS, then move them into CS.

Repeat until the final goal g_n enters IS.

Step by step in mini-craft (S, W $\rightarrow$ P)

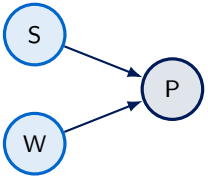
(a) Start: CS = IS = $\varnothing$. Agent knows the resources exist, not how they relate.

(b) Pre-train roots: learn to fetch S and W $\Rightarrow$ both controllable.

(c) Put S in IS (intervene), play, run SSD $\Rightarrow$ P *not* yet a child (P is AND, needs W too).

(d) Add W to IS as well. Now SSD sees P appear whenever S *and* W are held $\Rightarrow$ edges $S \rightarrow P$, $W \rightarrow P$ discovered.

(e) Train P; final goal reached. Hierarchy $\{S, W\} \rightarrow P$ recovered.



final recovered subgoal structure

blue=controllable (CS), red=intervened (IS), black=goal.

Results

Theoretical — training cost

Targeted (HRC_h) vs. random (HRC_b) subgoal selection:

	Tree $G(n, b)$	**Semi-ER** $G(n, p)$
HRC_h (ours)	$O(\log^2 n\, b)$	$O(n^{\frac{4}{3}+\frac{2}{3}c} \log n)$
HRC_b (random)	$\Omega(n^2 b)$	$\Omega(n^2)$

Discovery accuracy (SHD, lower is better):

Method	SHD	Missing	Extra
SSD (ours)	**12.3**	6.0	**6.3**
SDI (Ke et al.)	19.8	4.2	15.6

Targeted exploration is *provably* cheaper, and SSD adds fewer spurious edges.

Experimental — 2D-Minecraft

- Long-horizon, reward only at the final goal.
- All HRC variants reach the goal with *fewer system probes* than CDHRL, HAC, HER, OHRL, and vanilla PPO.

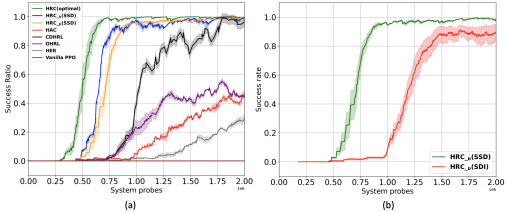


Figure 5: Success ratio vs. system probes: HRC variants above baselines

Khorasani, Salehkaleybar, Kiyavash, Grossglauser. ICML 2025. arXiv:2507.04373

Thread 2: Causality for Perturbation Prediction

The obstacle

Full time trajectories are often *unavailable*.
Single-cell RNA sequencing destroys the cell it measures. There is no path to observe.

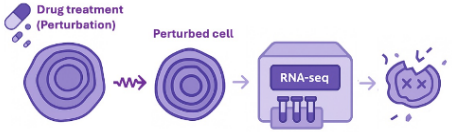


Figure 6: Conceptual illustration of sampling under drug treatment

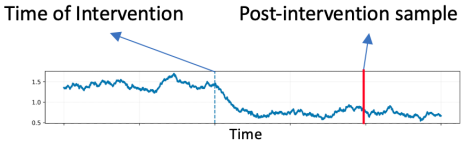


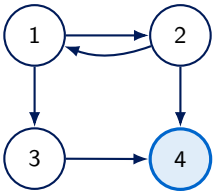
Figure 7: Illustration of post-intervention sample in stochastic dynamical systems.

Thread 2: Causality for Perturbation Prediction

Many systems (gene networks, neuroscience) can be modeled by SDEs. A linear SDE is:

$$dx = (-\Lambda x + b)\, dt + \sigma\, dW_t.$$

- The drift matrix Λ *is* the causal graph: edge $j \to i \iff \Lambda_{ij} \neq 0$.
- In biology we rarely see *trajectories*; only **steady-state "snapshots"** (mean μ, covariance Σ).
- Interventions = gene *knockdowns*: zero out row i's off-diagonals (cut incoming edges), keep self-regulation.



$\{1, 2\}$ is one SCC; knockdown on node 4 (blue)

Question

From steady-state (μ, Σ) plus a few interventions, can we *recover* Λ, b, D?

Main result: One Intervention per Component is Enough

Identifiability

One intervention per strongly-connected component (SCC) of the drift graph suffices to recover *all* OU parameters (generically, up to a single global scale) provided the SCC condensation graph is connected with one root and mild spectral/rank non-degeneracy holds.

Why up to scale? The stationary moment equations

$$\mu = \Lambda^{-1} b, \qquad \Lambda\Sigma + \Sigma\Lambda^{\top} = D$$

are invariant under $(\Lambda, b, D) \mapsto (c\Lambda, cb, cD)$; an intrinsic ambiguity of steady-state data.

Structure Learning After a single-node intervention, the means that *change* are exactly the **downstream** SCCs:

$$\mu_k^{(i)} \neq \mu_k \iff k \in \mathrm{Desc}(C).$$

$\Rightarrow$ recovers the SCC decomposition *and* a topological order over SCCs.

Proposed Estimator: a Steady-State Moment Loss

Recover $\Theta = [\mathrm{vec}(\Lambda); b; d]$ by jointly fitting the stationary mean and covariance equations across observational *and* interventional data:

$$\min_{\Lambda,b,D} \underbrace{\left\|\Lambda\mu - b\right\|^2 + \left\|\Lambda\Sigma + \Sigma\Lambda^\top - D\right\|_F^2}_{\text{observational residuals}} + \sum_{i\in I} \underbrace{\left\|\tilde{\Lambda}^{(i)}\mu^{(i)} - b\right\|^2 + \left\|\tilde{\Lambda}^{(i)}\Sigma^{(i)} + \Sigma^{(i)}\tilde{\Lambda}^{(i)\top} - D\right\|_F^2}_{\text{interventional residuals}} + \lambda \left\|\Lambda\right\|_1.$$

- $\tilde{\Lambda}^{(i)}$ = drift with row i's off-diagonals zeroed (the knockdown).
- Convex least-squares in the moment equations; ℓ_1 term encourages a sparse drift graph.
- The identifiability results guarantee the minimizer is the true Θ up to global scale.

Experimental Results

Synthetic ($n = 10$)

- Relative error of Λ, *b decreases* with the number of interventions (up to global scale); matching the theory.

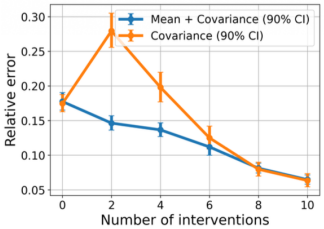


Figure 8: Relative error of Λ

Real data (three Perturb-seq datasets)

Metrics DES / PDS (higher = better); shaded row has oracle access to interventional means.

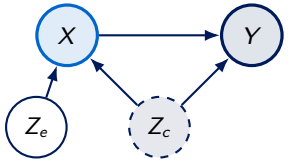
	Co-Culture		Control		IFN-γ	
Method	DES	PDS	DES	PDS	DES	PDS
Observational mean	0.33	0.57	0.33	0.57	0.35	0.67
Ours (Covariance)	0.51	0.43	**0.46**	0.43	0.42	0.43
Ours (Mean + Cov.)	0.43	**0.67**	0.33	0.63	**0.47**	**0.70**
Interventional mean	**0.52**	0.57	0.42	**0.67**	0.36	**0.70**

S. Salehkaleybar. One Intervention per Component is Enough. ICML 2026 (Spotlight).

Causal Probing

Correlational probing trains a classifier to read a property off hidden states. It shows what is *encoded*, not whether the model *uses* it.

- **Causal probing** instead *intervenes* on a hidden state $h_\ell(x)$, changing a property Z_c from value z to z', and checks whether the next-token prediction Y changes accordingly; the do-operator inside the LLM.
- Example: "*The key [MASK] on the table.*" Base model prefers `is` (Z_c=SG). Flip the encoded number to PL. Does it now predict `are`?
- Two properties: **causal** Z_c (drives Y) and **nuisance** Z_e (context, should not).



X: prompt $x_{1:T}$, Y: next token, Z_c: causal property, Z_e: nuisance

Gap

Prior methods train a *probe* per property to get the intervention direction; costly and possibly misaligned with the model's own geometry.

HDMI: Probe-Free Margin Intervention

Idea: steer the hidden state using the model's *own* output head (no probe). Given a source token σ and target token τ at `[MASK]`:

1. **Margin objective:** Raise target logit, lower source:

$$\mathcal{L}(x) = \phi(x)_\tau - \phi(x)_\sigma.$$

2. **Gradient** (final layer L, closed form):

$$\nabla_{h_L} \mathcal{L} = W_U^\top (e_\tau - e_\sigma).$$

3. **Ascend** K steps: $h^{(k+1)} = h^{(k)} + \alpha \nabla_h \mathcal{L}$.
4. Substitute $\tilde{h}_\ell$ back at that layer/position only.

Multi-token: sum over acceptable targets T^+ / sources T^-.

Why the margin?

A *target-only* may raise the *source* logit and *shrink* the margin. Demoting the source is essential.

Results

Evaluation: two metrics

- **Completeness**: did the target property Z_c actually flip? (validation probe on Z_c)
- **Selectivity**: were nuisance properties Z_e left intact? (validation probe on Z_e)
- **Reliability** = harmonic mean of the two.

Baselines compared

INLP, AlterRep, FGSM/PGD gradient-based probes; all probe-dependent.

Outcome

- HDMI attains the **highest reliability** (completeness–selectivity harmonic mean) across both models and tasks.
- Probe-free ⇒ no per-property training; gradient is one matrix–vector product.

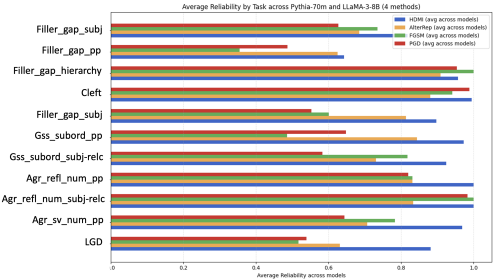


Figure 9: Reliability by tasks

Takeaways

1. Causal questions live above prediction: intervention & counterfactual.
2. Graphs + SCMs make assumptions explicit and identification checkable.
3. Hard problems (latent confounding, bias, identifiability) are active research.
4. Causality is increasingly central to robust, interpretable AI and RL.

Any Questions?

References I

- J. Pearl, M. Glymour, N. P. Jewell. *Causal Inference in Statistics: A Primer*. Wiley, 2016.
- J. Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge Univ. Press, 2nd ed., 2009.
- J. Peters, D. Janzing, B. Schölkopf. *Elements of Causal Inference*. MIT Press, 2017.
- M. A. Hernán, J. M. Robins. *Causal Inference: What If*. Chapman & Hall/CRC, 2020.
- D. B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Educational Psychology*, 1974.
- T. Verma, J. Pearl. Equivalence and synthesis of causal models. *UAI*, 1991. (IC algorithm)
- P. Spirtes, C. Glymour, R. Scheines. *Causation, Prediction, and Search*. MIT Press, 2000. (PC/FCI)
- S. Shimizu, P. O. Hoyer, A. Hyvärinen, A. Kerminen. A linear non-Gaussian acyclic model for causal discovery. *JMLR*, 2006.
- P. O. Hoyer, D. Janzing, J. M. Mooij, J. Peters, B. Schölkopf. Nonlinear causal discovery with additive noise models. *NeurIPS*, 2009.

References II

J. M. Mooij, J. Peters, D. Janzing, J. Zscheischler, B. Schölkopf. Distinguishing cause from effect using observational data: methods and benchmarks. *JMLR*, 2016. (Tübingen pairs)

J. Berkson. Limitations of the application of fourfold table analysis to hospital data. *Biometrics Bulletin*, 1946.

N. Cartwright. *Nature's Capacities and Their Measurement*. Oxford Univ. Press, 1989. ("No causes in, no causes out")

A. Sharma, E. Kiciman. DoWhy: an end-to-end library for causal inference. *arXiv:2011.04216*, 2020.

S. Khorasani, S. Salehkaleybar, N. Kiyavash, M. Grossglauser. Hierarchical RL with Targeted Causal Interventions. *ICML*, 2025. arXiv:2507.04373.

K. Meng, D. Bau, A. Andonian, Y. Belinkov. Locating and Editing Factual Associations in GPT. *NeurIPS*, 2022. arXiv:2202.05262.

S. Khorasani, S. Salehkaleybar, N. Kiyavash, M. Grossglauser. Inference-Time Causal Probing in LLMs. *arXiv:2605.07631*, 2026.

S. Salehkaleybar. One Intervention per Component is Enough: Towards Identifiability in Linear Stochastic Dynamics from Steady State. *ICML*, 2026.