Identifying Structural Biases from Causal Mechanism Shifts

Praharsh Nanavati **Jilles Vreeken** **David Kaltenpoth**
CISPA Helmholtz Center for Information Security, Saarbrücken, Germany
{praharsh.nanavati, vreeken, david.kaltenpoth}@cispa.de

Abstract

Causal discovery methods commonly assume that all data is independently and identically distributed (i.i.d.) and that there are no unmeasured variables affecting the system. In practice, these assumptions are often violated, leading to inaccurate inference. In this paper, we study how to identify hidden confounding and selection biases from causal mechanism shifts. In particular, we show that structural biases lead to dependent mechanism shifts. That is, by considering for which variables the mechanisms change given data from different environments, we can tell which variables are unbiased, which are subject to hidden confounding, and which are undergoing selection bias. We formalize this into an empirically testable criterion based on mutual information, and show under which conditions it identifies structural biases. To tell which nodes are subject to what kind of bias, we introduce the STRUBI algorithm. Experiments on synthetic and real-world data show that STRUBI works well in practice, accurately recovering affected variable sets and types of biases, outperforming the state-of-the-art by a wide margin. We provide an implementation of STRUBI at `https://github.com/niftynans/strubi`.

1 Introduction

Methods for causal inference generally require causal sufficiency and i.i.d.-ness [21, 31]. That is, to provably recover the causal ground truth, they require that all causally relevant variables are measured (sufficiency) and that the samples in the data are all independent and identically distributed (i.i.d.). While convenient in theory, neither of these are likely to hold in practice [13, 14]. Often, we do not know what the causally relevant variables are, and even if we do, we cannot always measure them. Often, we also do not know whether a system has a representative sample, or whether it is subject to selection bias. In either case, this can lead to causal models that can be arbitrarily wrong because the data will show spurious dependencies [5, 14, 32]. For example, if 'shark attacks' and 'ice cream sales' are statistically dependent, it may lead us to believe there is a causal dependency, even if they are actually conditionally independent given their *unmeasured* confounder 'season'. Similarly, if we only study the survival rates of fighter jets that returned from combat, then we get a distorted view of where more protective armor is most needed. This raises the question, can we detect whether the data at hand is unbiased, subject to hidden confounding, or subject to selection bias?

In this paper, we show that by considering non-i.i.d. data collected from different contexts, the effects of confounding and selection bias are identifiable by considering distribution shifts of the observed variables. In particular, we show that actual, *intrinsic*, mechanism shifts leave measurably different *observed* mechanism shifts between the three cases. We give an example in Fig. 1. Here, X and Y are children of observed variable W and are either unbiased (left), confounded (center), or undergoing selection bias (right) through a latent collider V. We consider data gathered in five contexts, $c_1, \ldots, c_5$. In each context, interventions affect specific variables (hammers) and change their underlying generating process (colored boxes).

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

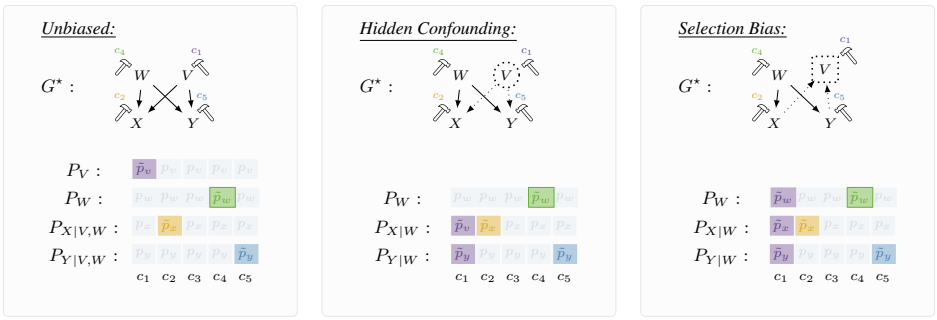


Figure 1: *Structural biases introduce dependent mechanism shifts.* We consider three systems of four variables across five different contexts. The unbiased system (left) exhibits independent mechanism shifts. Under confounding (center), an *actual* mechanism change on *hidden* confounder V in context c_1, leads to *observed* mechanism changes in variables $X|V,W$ and $Y|V,W$ for the same context. Under selection on a *hidden* collider (right), a mechanism change on that collider (V) leads to *observed* mechanism changes for *all* its ancestors, X, Y, and even W.

In the unbiased case, where we measured all relevant variables, the intervention on V in context c_1 (purple) do not affect the observed mechanisms of the other variables. The conditionals $P_{X|V,W}$ and $P_{Y|V,W}$, and the marginal P_W all remain the same, they change independently in other contexts (yellow, blue, green). Under confounding (center), however, an intrinsic mechanism change of V lead to observed changes in its children (X and Y). Under selection bias (right), however, an intrinsic mechanism shift in V leads to observed shifts in all of its upstream variables (X, Y, and W).

We formally show under which conditions we can leverage this observation to identify and classify between the three cases. To do so, we assume that mechanisms shifts are independent and non-degenerate. Importantly, unlike existing work [5, 6], our approach is non-parametric, making it generally applicable. To put our theory to practice, we propose the StruBI algorithm for structural bias identification from empirical data. StruBI operates on a given causal graph $\mathcal{G}$ and employs the KCI test by [37] to determine if mechanisms change between contexts. We use adjusted mutual information to determine whether the mechanism shifts of different variables behave independently or not, and to classify between confounding or selection, consider many of the upstream variables change together.

We empirically evaluate StruBI on a wide range of synthetic data, comparing it to five state of the art methods. The experiments show that StruBI outperforms its competitors by a wide margin, and that using an inferred causal graph gives comparable results to using the ground truth causal network. We verify that StruBI is applicable on real-world data through a case study on cell-signaling data, where it correctly classifies between confounding and selection bias, and unlike its competitors, retrieves high-quality sets of affected variables in both settings.

All code is available in the supplementary. We postpone all proofs to the appendix.

2 Related Work

Inferring causality from observational data is a fundamental challenge in statistics [21, 26, 31]. In a nutshell, it is only possible to do so under assumptions of the data generating model [21]. Standard assumptions include the causal Markov condition, causal faithfulness, causal sufficiency, and i.i.d.-ness. Under these assumptions, classical constraint-based [31, 36] and score-based [3, 23] methods can recover causal structures up to Markov equivalence.

Causal discovery under latent confounding is an active topic of research. The FCI family [4, 20, 31] can discover and mark edges that are likely confounded. Nested Markov Models (NMMs) [7, 25, 28, 29] leverage Verma constraints to identify latent factors. DCD [1] combines NMMs with differentiable constraints [38] to discover partial DAGs. Reddy et al. [24] utilize mutual information on observed distributions, and Karlsson and Krijthe [12] address latent confounding for effect estimation under a fixed graph structure. Kaltenpoth and Vreeken [9, 11] model latent confounders directly, allowing identification of which nodes *share* latent confounders.

There exists much less research on causal discovery under selection. FCI [31] is sound under selection, but Spirtes et al. [32] show that single observational studies are generally insufficient to distinguish between confounding or selection. To resolve this, Versteeg et al. [33] identify local patterns to detect causal effects under both biases. Recent work has expanded these capabilities to temporal dependencies in sequential data [39], "twin graph" frameworks for interventional settings [5], and rank-based constraints for linear Gaussian models [6]. These methods respective require it i.i.d. or sequential data, or strict parametric assumptions.

Recent research leverages non-i.i.d. data, in particular distribution shifts across different environments, to relax the assumptions necessary to identify causal DAGs. Existing methods primarily focus on handling latent confounding. JCI [18] extends existing causal discovery methods with a context variable. Perry et al. [22] focus on sparse mechanism shifts, whereas CD-NOD [8] and CoCo [14] both exploit independent mechanism changes. Our approach leverages multi-environment data to identify the presence of, and, which nodes are affected by hidden confounding resp. selection bias by detecting dependent mechanism shifts.

3 Preliminaries

In this section, we establish conditions necessary to distinguish confounding from selection bias in non-i.i.d. settings. We begin by defining context-indexed observed distributions and show how they are generated by latent causal mechanisms. We then ask the central question: what patterns in the observed context shifts differentially reveal latent confounding and respectively selection?

3.1 Problem Setting

We consider variables $V = X \cup Z \cup S$, where X, Z, and S denote the sets of observed variables, hidden confounders, and hidden selection variables, respectively. We assume that $X = (X_1 \dots, X_d)$ consists of d dimensions and the system is observed across multiple contexts $c \in \mathcal{C}$, with $n_c = |\mathcal{C}|$. For each context, the full data-generating distribution is $P^c(V) = P(V \mid C = c)$. We assume that all these distributions are *Markov* and *faithful* with respect to the same causal DAG $\mathcal{G}^* = (V, E)$, with edge $(i, j) \in E$ whenever V_i is a direct cause of V_j. We write $\mathrm{pa}_i^* = \{V_j : (j, i) \in E\}$ for the parents of V_i in $\mathcal{G}^*$, and $\mathrm{pa}_i = \mathrm{pa}_i^* \cap X$ for observed parents of variables X_i. We further write ch_i for the set of all children of V_i, anc_i for the set of all ancestors, and $\mathrm{anc}^+(U) = \mathrm{anc}(U) \cup U$.

We assume that latent confounders Z are exogenous, while latent selection variables S_ℓ are sink nodes in $\mathcal{G}^*$ with $\mathrm{pa}^*(S_\ell) \subseteq X$. The observed distribution in context c is then given by

$$P^c_{\mathrm{obs}}(X) := P^c(X \mid R(S) = 1)\,,$$

where $R = R(S) = 1$ indicates retention of a sample. Absent selection, $R \equiv 1$ and $P^c_{\mathrm{obs}}(X) = P^c(X)$. We make the standard assumption [10] that selection is *positive and nondegenerate*, i.e., $0 < P^c(R(S) = 1 \mid \mathrm{pa}(R))$ for all c, and $P^c(R(S) \mid \mathrm{pa}(R)) < 1$ for some $c \in \mathcal{C}$. This excludes selection rules where selection is so strict that there is no data left in some contexts.

We do not assume causal sufficiency over X alone. Instead, all hidden common causes of observed variables are assumed to be included in Z, and all selection-induced conditioning is represented by S.

The question is therefore: what can observed mechanisms P^c_{obs} tell us about the graph $\mathcal{G}^*$ over V?

Problem Statement *Given the set of observed distributions $\{P^c_{\mathrm{obs}}(X)\}_{c\in\mathcal{C}}$, our goal is to identify subsets of X whose observed local mechanisms shift jointly across contexts, and to distinguish, when possible, whether these shifts are better explained by latent confounding or by selection bias.*

3.2 Intrinsic and Observed Causal Mechanism Shifts

A key distinction in our setting is between *intrinsic* mechanism shifts in all variables, $V_i \in V$ and *observed* mechanism shifts after marginalizing latent confounders and conditioning on selection.

For any variable $V_i \in V$, we refer to the conditional distribution $P^c(V_i \mid \mathrm{pa}_i^*)$ as its intrinsic local mechanism in context c. This induces an intrinsic partition Π_i^* of the context set: two contexts c, c' belong to the same block when the mechanisms generating V_i are the same, $P^c(V_i \mid \mathrm{pa}_i^*) = P^{c'}(V_i \mid \mathrm{pa}_i^*)$. We denote the block to which c belongs by $\Pi_i^*(c)$ and therefore have

$\Pi_i^*(c) = \Pi_i^*(c')$ in these cases. We call a pair of contexts c, c' such that $\Pi_i^*(c) \neq \Pi_i^*(c')$ an *intrinsic shift*. We assume that these intrinsic shifts are jointly independent [14, 22].

Assumption A1 (Nondegenerate Independent Intrinsic Shifts)**.** We assume that intrinsic partitions $\{\Pi_i^*\}_i$ are jointly independent and nondegenerate, $0 < P(\Pi_i^*(C) \neq \Pi_i^*(C')) < 1$.

This assumption commonly holds in realistic settings where shifts occur due to interventions on a system [22]. For example, in biological data, data from observed distributions is much easier to obtain than interventional data, and such interventions usually target only one feature at a time.

Just as the intrinsic shifts of $P_i^c = P^c(V_i \mid \mathrm{pa}_i^*)$ induce partitions Π_i^*, the observed distributions $P_{i,\mathrm{obs}}^c = P^c(X_i \mid \mathrm{pa}_i, R = 1)$ can also shift and induce observed partition Π_i^{obs} of the contexts. In general, every intrinsic shift of X_i between c, c' also implies an observed shift between c, c'. For two contexts c, c', we define the *observed affected set*

$$A_{c,c'} := \{X_i \in X : \Pi^{\mathrm{obs}}(c) \neq \Pi^{\mathrm{obs}}(c')\}.$$

Our goal in this paper is to relate the observed partitions Π^{obs} of variables X to the intrinsic partitions Π^* of all variables V. In the next section we show that variables X_i, X_j which are members of $A_{c,c'}$ for many c, c' are prime targets for suspecting the influence of latent mechanism shifts.

4 Theory

In this section we connect how observed partitions relate to the effects of latent confounding and selection by studying the network topology of nodes whose observed partitions correlate. Proofs for all results can be found in the appendix.

4.1 Difference in Shifts in Observed Variables Based on the Structural Bias

We begin with a simple example to illustrate how latent parents of a variable X and latent selection on a variable X can have different impacts on its upstream variables.

Example 4.1. Suppose that variable X is caused by a latent Z and observed W, and that Z, W, ϵ_X are jointly independent,

$$X = \alpha Z + \gamma W + \epsilon_X, \qquad \mu_c = E[Z \mid C = c]$$

If $\mu_c = 0$ and $\mu_{c'} = \delta$, then the observed conditional mean changes as

$$\mathbb{E}[X \mid W = w, C = c] = \gamma w,$$
$$\mathbb{E}[X \mid W = w, C = c'] = \alpha\delta + \gamma w.$$

Thus the observed distribution $P^c(X \mid W)$ changes across contexts because it marginalizes over the shifted latent parent Z. In contrast, the observed distribution $P^c(W)$ does not shift.

Now suppose that there is no latent confounder, but that selection acts through a descendant S of X,

$$X = \gamma W + \epsilon_X, \qquad S = aX + \epsilon_S, \qquad R = \mathbf{1}\{S > \tau_c\},$$

where $R = 1$ denotes that a sample is retained in context c. By Bayes' theorem, we get

$$p_c(x \mid w, R = 1) \propto \frac{P^c(R = 1 \mid X = x, W = w)}{P^c(R = 1 \mid W = w)} = \frac{P^c(\epsilon_S > \tau_c - ax)}{\int P^c(\epsilon_S > \tau_c - au)p(u \mid w)\, du}.$$

When $a \neq 0$, this weight depends on x, so the observed distribution $P^c(X \mid W, R = 1)$ changes when τ_c changes. Furthermore, the observed distribution $P^c(W \mid R = 1)$ also changes:

$$p_c(w \mid R = 1) \propto \frac{\int P^c(\epsilon_S > \tau_c - au)p(u \mid w)\, du\,.}{P^c(R = 1)}.$$

Hence selection makes both its parent X and its observed ancestor W shift, while a changing latent parent affects only the observed distributions of X that directly marginalize over it.

Next we show that this crucial asymmetry between latent confounding and selection holds generically for large classes of parametric systems.

Theorem 4.2 (Observed Local Signatures of Confounding and Selection)**.** *Let $\mathcal{G}^*$ be the true graph over V, with precisely one latent variable, $|Z \cup S| = 1$. Let A be given by*

$$A = \begin{cases} \mathrm{ch}_{\mathcal{G}^*}(Z), & \textit{for a latent confounder}, \\ \mathrm{anc}_{\mathcal{G}^*}(S), & \textit{for selection}\,. \end{cases}$$

Let $\{P_\theta^c : c \in C, \theta \in \Theta\}$ be a parametric family with $\Theta \subseteq \mathbb{R}^d$ open and connected, such that all joint, marginal and conditional densities are real analytic in θ on a non-empty common support. Let c and c' be two contexts in which only the latent variable's mechanism changes. Suppose that for every $X_i \in A$, there exists at least one $\theta_0 \in \Theta$ such that $P^c_{\mathrm{obs}}(X_i \mid \mathrm{pa}_{\mathcal{G}^*_X}(X_i); \theta_0) \neq P^{c'}_{\mathrm{obs}}(X_i \mid \mathrm{pa}_{\mathcal{G}^*_X}(X_i); \theta_0)$.

Then $A_{c,c'} = A$ for all θ outside a Lebesgue-null subset of Θ.

The conditions of this theorem generically hold for large classes of parametric families: exponential families, linear Gaussian SEMs, parametric additive noise models, generalized additive models, splines, neural networks with many common activation functions, normalizing flows, and many more.

In the following we therefore assume throughout that latent mechanism shifts induce the correct observed mechanism shifts.

4.2 Ancestral Coverage and Minimal Explanations

The previous theorem motivates using the graphical topology of sets of variables with correlated observed partitions. While there is no necessary topological relationship between variables whose mechanism shifts are correlated by latent confounding, in selection bias the set of variables whose partitions are correlated contains for any contained variable also all its ancestors.

Definition 4.3 (Ancestral Coverage)**.** Let $A \subseteq V$ be a subset of the nodes of the DAG $\mathcal{G} = (V, E)$. We define the *ancestral coverage* of the subset A as

$$\mathrm{cov}_{\mathcal{G}}(A) := \frac{|A|}{|\mathrm{anc}_{\mathcal{G}}(A)|}\,.$$

We call A *ancestrally closed* if $\mathrm{cov}_{\mathcal{G}_X}(A) = 1$, i.e., $A = \mathrm{anc}(A)$.

By Theorem 4.2, under independent and nondegenerate mechanism shifts, coverage $\mathrm{cov}(A_{c,c'}) \approx 1$ in the case of latent selection, but $\mathrm{cov}(A_{c,c'}) \ll 1$ in the case of latent confounding.

Proposition 4.4 (Ancestral Coverage of Confounding and Selection Effects)**.** *Let $\mathcal{G}^*$ be a directed Erdős-Rényi graph with independent edge probabilities p. Assume that the recovered set $A_{c,c'}$ is generated either by a single latent confounder or by a single latent selection variable.*

1. *If $A_{c,c'}$ is created by selection, then $\mathrm{cov}(A_{c,c'}) = 1$.*
2. *If $A_{c,c'}$ is created by a confounder, then $\mathrm{cov}(A_{c,c'}) < 1$ with probability $1 - O(e^{-pk(d-k)/2})$.*

Of course, it is always possible to construct adversarial examples in which the latent confounder Z targets an ancestrally closed set. However, such explanations are not minimal in the following sense.

Proposition 4.5 (Minimal Latent Variable Sets)**.** *Let $A_{c,c'} \subseteq X$ be nonempty and ancestrally closed in $\mathcal{G}^*$. Assume that there exists at least one parent-child pair $X_i, X_j \in A_{c,c'}$. Then among all single-latent-node explanations of the correlated set $A_{c,c'}$, the edge-minimal explanation is that of a single selector S with sink nodes of $A_{c,c'}$ as parents,* $\mathrm{pa}(S) = \{X_i \in A_{c,c'} : \mathrm{ch}(X_i) \cap X = \emptyset\}$.

We introduce the notion of minimality as it helps in explaining correlations in the mechanism shifts of the observed variables as simply as possible. We formalize this notion by defining an explanation with, (i) the minimum number of latent variables, and (ii) the minimum number of edges to and from the latent variables. For example, if a selection criteria affects a set of nodes, then in the infinite sample limit, it affects all nodes that are upstream of it. However, this can be minimally expressed by showing only the sink nodes of the affected set.

Based on these results, we next introduce a practical algorithm which leverages the notion of ancestral coverage to distinguish confounding from selection bias in the finite data setting.

Algorithm 1: Structural Bias Identification – STRUBI($\mathcal{G}$)

input : Observed data over X, C; thresholds τ_1, τ_2; observed causal graph $\mathcal{G}$.
output : Subsets of X subject to selection bias respectively confounding.

1 **foreach** variable $X_i \in X$ **do**
2 Compute $p_{c,c'} = \text{p-val}\left(P^c(X_i \mid \text{pa}_i) \neq P^{c'}(X_i \mid \text{pa}_i)\right)$ for all pairs (c, c');
3 Convert $\{p_{c,c'}\}$ to a partition Π_i (grouping contexts by mechanism invariance);
4 Construct an affinity matrix $\mathcal{M}$ where $\mathcal{M}_{ij} = \text{AMI}(\Pi_i, \Pi_j)$;
5 Find the set of connected components $\mathcal{K} = \{K_1, \ldots, K_m\}$ in the graph induced by $\mathcal{M}$;
6 Filter connected components of $\mathcal{K}$ such that $X_S = \{K_i \in \mathcal{K} : |K_i| > 1\}$;
7 **for** each subset $S \in X_S$ **do**
8 $\bar{m}(S) \leftarrow \text{mean}(\{\text{AMI}(\Pi_i, \Pi_j) \mid i, j \in X_S, i < j\})$;
9 **if** $\bar{m}(S) < \tau_1$ **then**
10 Label S as **Unbiased**;
11 **else**
12 **if** $\text{cov}_{\mathcal{G}}(S) \geq \tau_2$ **then**
13 Label S as **Selection Bias**;
14 **else**
15 Label S as **Confounding**;
16 **return** Set of subsets X_S with assigned labels;

5 The STRUBI Algorithm

In this section, we introduce the STRUBI algorithm and give the pseudocode in Algorithm 1. The algorithm takes as input the observed data over contexts, the observed causal graph $\mathcal{G}$, and thresholds τ_1 and τ_2. τ_1 is a threshold for the mean of the pairwise Adjusted Mutual Information (AMI) values $\bar{m}(\cdot)$ for each detected subset, S to detect the presence of a latent variable. We show how we calculate AMI values in Appx. B. τ_2 is a threshold for $\text{cov}_{\mathcal{G}}$ to distinguish selection bias from hidden confounding. We arrive at the thresholds 0.1 and 0.6 for τ_1 and τ_2 respectively, as shown in Appx. Sec. C.2.

Detecting Causal Mechanism Shifts We perform a conditional independence test to detect mechanism changes, for each causal mechanism of an observed variable X_i and each pair of environments (line 2) as done by Mameche et al. [14], resulting in the following p-values,

$$p_{c,c'} = \text{p-val}\left(P^c(X_i \mid \text{pa}_i) \neq P^{c'}(X_i \mid \text{pa}_i)\right).$$

We perform this independence test using the KCI test by Zhang et al. [37] for all variables that have parents in $\mathcal{G}$. For marginal tests, we perform the Kolmogorov-smirnov test [15]. Based on $p_{c,c'}$, we obtain partitions, $\left\{\Pi_i^{\text{obs}}\right\}_{i \in \{1,\ldots,|V|\}}$, as defined in Section 3, that describe which observed variables have significant overlaps in their causal mechanism shifts (l. 3).

Detecting presence of latent variables We take inspiration from Mameche et al. [14] and construct an affinity matrix $\mathcal{M}$ using AMI scores between observed partitions as edge weights (l. 4). We extract non-trivial connected components from the resulting weighted graph by considering subsets with $\bar{m}(S) > \tau_1$, to identify node subsets jointly influenced by the same latent variable (l. 5-6). However, we use binary indicators of distribution shifts to build our partitions as opposed to Mameche et al. [14] who cluster p-values to derive partitions and then detect the presence of a latent variable by doing spectral clustering on the affinity matrix obtained from the pairwise AMI scores instead.

Identifying Structural Biases We then distinguish selection bias from confounding using STRUBI. To do this, we check for each subset if $\text{cov}_{\mathcal{G}} < \tau_2$ and predict confounding if true, else selection bias (l. 13-15).

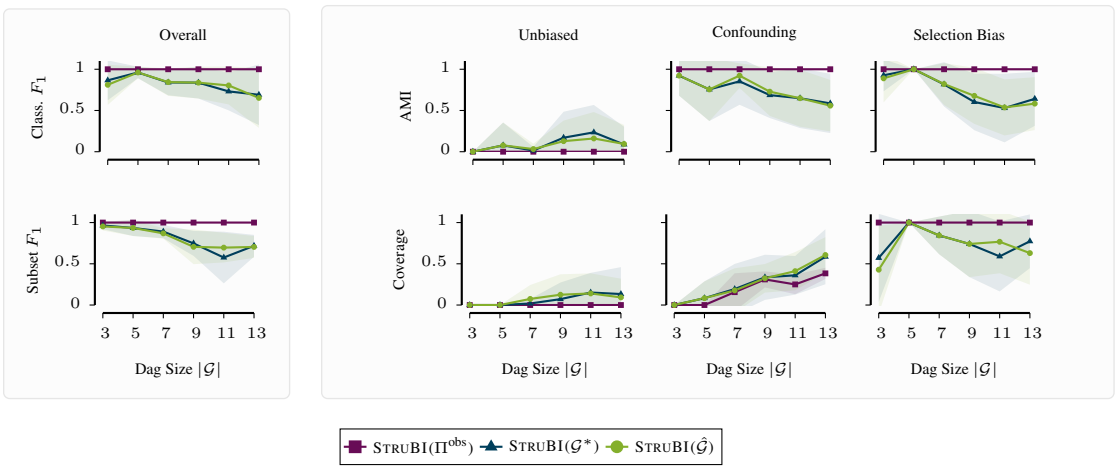


Figure 2: *Structural Bias Detection with* StruBI. Given are $F1$-scores for bias classification and subset recovery (left) and the mean pairwise AMI and Coverage values for the unbiased, confounding, and the selection bias cases (right). The values for the AMI and Coverage are low for unbiased subsets, low and high respectively for confounded cases, and high for both for the selection cases.

6 Experiments

In this section, we empirically evaluate StruBI on synthetic and real-world data. We focus on classifying between unbiasedness, confounding, and selection, and on how well the affected sets of nodes are recovered.

StruBI. We consider three main versions of StruBI. First, we consider StruBI(Π^{obs}), which is given the *true* partitions Π^{obs} of the observed variables. Second, we consider StruBI($\mathcal{G}^*$), which is given the *true* observed causal structure ${\mathcal{G}^*}_X$. Third, we evaluate StruBI($\hat{\mathcal{G}}$), which is given a *learned* causal graph $\hat{\mathcal{G}}$ obtained using Topic [35].

In addition, to assess the robustness of StruBI against graph mis-specification, we also consider StruBI($\tilde{\mathcal{G}}$), which operates on a perturbed graph $\tilde{\mathcal{G}}$ generated by randomly flipping edges of the ground truth DAG. We provide details and the results for this variant in Appx. C.6.

Baselines We compare StruBI against five strong baselines. FCI [30] can perform causal discovery in the presence of hidden confounding and selection bias. We consider FCI applied on data pooled across all contexts (FCI-P), as well as JCI-FCI [18], which extends FCI to multi-context data. We also compare against CoCo [14] as a method that can determine hidden confounding from mechanism changes. We include Latent-Selection [6] which focuses on latent variable causal discovery under selection bias. We apply it both on pooled data (LS-P) as well as independently on each context and aggregating the results (LS-C). For more details, see Appx. C.15.

Synthetic Data To evaluate with known ground truth, we consider synthetic data. We adopt the experimental setup of Huang et al. [8] and Mameche et al. [14]. We generate synthetic data using an Erdős-Rényi model over $m \in \{3, 5, 7, 9, 11, 13\}$ observed variables. The number of contexts m_c is scaled as $2m$. We sample from the following structural equation model (SEM)

$$X_i^{(c)} = \sum_{j \in \text{pa}(X_i)} w_{ij} \cdot f\left(X_j^{(c)}\right) + \mu_i^{(c)} + N_i^{(c)} \ ,$$

where the edge weights w_{ij} are sampled from $\mathcal{U}(0.7, 1.3)$ for observed interactions and $\mathcal{U}(1.7, 2.3)$ for interactions involving latent variables. The functional forms f are drawn uniformly at random from $\{x, x^2, x^3, \tanh, \text{sinc}\}$ and remain invariant across contexts. Mechanism shifts are introduced by re-sampling the bias term $\mu_i^{(c)}$ from a jittered discrete set $\{0, 5, 10\} \pm \mathcal{U}(-0.5, 0.5)$. The exogenous noise $N_i^{(c)}$ is drawn with equal probability from either a Gaussian distribution $\mathcal{N}(0, \sigma^2)$ or a Uniform distribution $\mathcal{U}(-\sigma\sqrt{3}, \sigma\sqrt{3})$ with $\sigma = 0.05$.

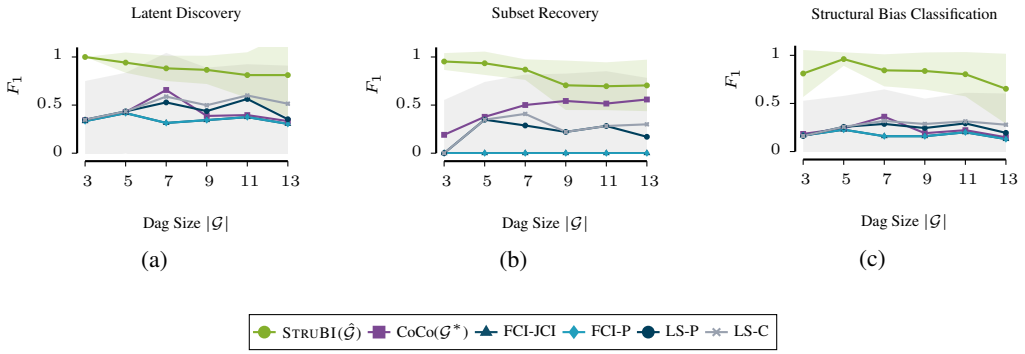


Figure 3: Macro-F_1 scores for (a) latent discovery, (b) subset recovery, and (c) structural bias classification. The shaded gray region illustrates the aggregate performance with maximum error bars across all baseline methods for the sake of readability. We include all error bars in Appx. Figure 11.

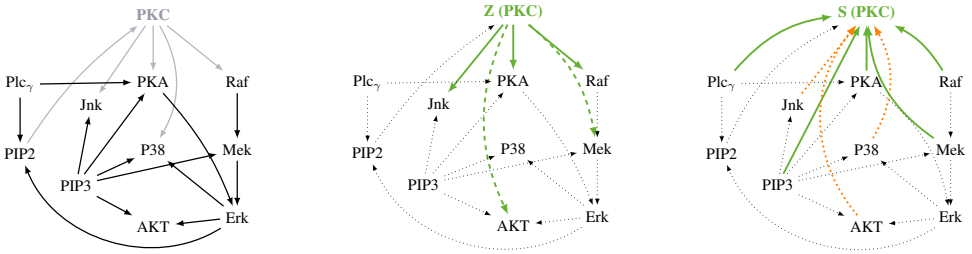
Each latent confounder Z_i is modeled as a source node with edges to a random subset of at least two variables from a pool of nodes in the graph avoiding explicit ancestral closure. Selection bias is introduced via hidden colliders S_i, whose parents are a randomly selected subset of observed variables. The final dataset is filtered by retaining only samples where S_i exceeds the s^{th} percentile of its marginal distribution. We set $s = 85$ for our synthetic data, creating a strong selection bias. All data is normalized globally to zero mean and unit variance.

6.1 Detecting Structural Biases

First, we are interested in understanding whether and how well StruBI can detect structural biases. To this end, we generate synthetic data as above, and consider StruBI(Π^{obs}), StruBI($\mathcal{G}^*$), and StruBI($\tilde{\mathcal{G}}$). We give the results in Fig. 2. On the left-hand side, we report the $F1$ scores for classifying the type of structural bias (i.e. none, confounding, selection) resp. for determining which nodes are affected. StruBI(Π^{obs}) shows our theory works; knowing the actual mechanism shifts, StruBI performs perfectly. More interestingly, StruBI given an inferred graph $\hat{\mathcal{G}}$ performs essentially as well as when given the true graph $\mathcal{G}^*$. The right-hand panel shows the AMI and coverage values. We see that classification performance goes hand in hand with how well partitions are recovered (top row), as well as that when more nodes share a confounder, it becomes harder to classify it from the case of selection with many ancestral nodes.

Next, we evaluate how well StruBI performs compared to the baselines. To this end we consider StruBI given graph $\hat{\mathcal{G}}$ inferred using Topic [35], JCI-FCI and FCI-P given the causal skeletons, CoCo given the ground truth graph $\mathcal{G}^*$, and LS-P and LS-C. We are interested in (i) how well can each identify the presence of latent variables, (ii) identify the affected variables, and (iii) how well can they identify the type of structural bias. We run each method on the same data as above, and present the results in Fig. 3. We see that for each of the three tasks, StruBI outperforms all baselines by a large margin. LS-P and LS-C do not do very well, probably because of their parametric and structural assumptions and model misspecification. CoCo is the best performing competitor, and the only one that improves in subset recovery as the DAG size increases. StruBI also outperforms the baselines in terms of the accuracy, precision, and recall. We give those results in Appx. Fig. 12.

We present various additional results, including several ablation studies, and a runtime analysis, in the Appx C. Specifically, we show that StruBI performs well even when many assumptions are not met in practice. For example, StruBI works well even with correlated mechanism shifts, class imbalances, and with out of training sample splits. In order to further validate its effectiveness, we also include results where the observed graph is inferred using CD-NOD [8], which we use as another off-the-shelf causal discovery method. StruBI also accommodates simultaneous confounding and selection bias for the identifiable case where the selection criteria is upstream of the confounders.



(a) Unbiased Consensus (b) Confounding case (c) Selection Bias case

Figure 4: STRUBI correctly recovers hidden confounding resp. selection bias on the Sachs et al. [27] data. (a) Unbiased consensus graph. (b) STRUBI correctly determines hidden confounding and recovers three directly and two indirectly confounded nodes when PKC is withheld. (c) STRUBI correctly determines selection bias and recovers six out of eight true ancestors when withholding and selecting on PKC. Green lines are correct (solid) resp. on a causal path (dashed), while orange lines are false positives (dotted).

6.2 Case Study – Cell Signaling

To see how well STRUBI performs in a real-world settings, we consider the (*infamous*) Sachs et al. [27] flow cytometry dataset. The data contains samples of eleven protein and phospholipid components in human immune cells that were studied under different molecular interventions. We show the consensus causal graph for the Sachs dataset in Fig. 4a. When we run STRUBI, we find dependencies in the mechanism shifts of proteins Akt and Mek, PKA and PKC, and Plcg and P38. This aligns with documented signaling crosstalk and shared regulatory architectures [2, 17, 19]. We give more details in Section D.

For the main experiment, we focus on the PKC node, as it possesses a sufficient number of descendants and a complex upstream topology beyond its immediate parents. Following the design of Mameche et al. [13], we induce hidden confounding by dropping variable PKC from the dataset, and selection bias by first filtering the data to retain only samples where PKC exceeds the 80th percentile of its marginal distribution, and then dropping it. We consider STRUBI and CoCo given the consensus causal network. The other baselines do not accept a causal graph as input, so we let those infer it themselves. We run each method on the unbiased data, data where we induce confounding, and on data where we induce selection. We present the results of STRUBI in Fig. 4 and of the baselines in Appx. Fig. 15.

STRUBI correctly classifies confounding and selection bias in their respective cases. In the confounded setting, STRUBI correctly infers three confounded edges (Jnk, PKA, Raf), reports two edges to causal descendants (Mek, AKT) that according to *another* consensus graph by Meinshausen et al. [16] are in fact correct, and misses only one edge (P38). In the case of selection bias, it correctly retrieves six out of eight ancestors, additionally recovering edges from Jnk, P38, and AKT that are correct compared to the graph by Meinshausen et al. [16] that includes cycles are at least partly correct. Of the competitors, CoCo performs well on the confounded setting, but fails to recover accurate subsets in the case of selection bias. JCI-FCI recovers in both cases only two indirectly related nodes (Erk and Akt). LS-C in both cases incorrectly returns that all nodes in the graph are affected.

STRUBI is able to recover meaningful subsets affected by latent variables even though many assumptions are violated, because coverage depends on nodes detectably affected by mechanism shifts. For instance, in the case of selection bias, if the affected nodes are part of a cycle, the coverage would depend on the number of nodes affected by the selection criteria, the strength of the selection bias, and how far upstream the shifts propagate. Barring cases where a majority of nodes in the entire feedback loop are directly affected or in the case of noisy distributions, the coverage acts as a good measure to separate confounding and selection. This works in the case of feedback loops if and only if the fraction of upstream nodes subsumed in the set of affected nodes diverge in the two cases. In cases where shifts do not strongly propagate upstream, i.e., in the case of a weaker selection criteria, weak mechanism shifts, as well as in cases with many latent variables, the coverage values may very

well overlap. In theory, the joint distribution of the observed variables is in general defined in terms of an equilibrium solution, so that a confounder on a subset of the strongly connected component (SCC) also leads to mechanism shifts in all other variables inside the SCC.

In summary, STRUBI performs well, even on real-world data on which our assumptions are unlikely to hold, as AMI and coverage are sound metrics to identify and classify the type of structural biases at play. It correctly determines unbiasedness, confounding, and selection bias, and recovers many of the truly affected nodes. We provide further details, along with a detailed discussion and additional results in Appx. D.

7 Discussion & Conclusion

We studied the problem of identifying structural biases from observational data. In particular, we showed under which conditions it is possible to infer the presence of latent confounding or selection bias on latent colliders, given data from multiple contexts by detecting dependent mechanism changes. To tell which nodes are subject to what kind of bias in practice, we proposed the STRUBI algorithm. Through an extensive set of experiments, including a case study on cell-signalling data, we showed that STRUBI performs well in practice, outperforming the state of the art in detecting confounding, selection bias, and classifying between them, even when our assumptions do not strictly hold.

Limitations As with all approaches in causal inference, we too need to make assumptions on how the world works. While we relax the assumptions of sufficiency and i.i.d.-ness, we do require the Markov condition and faithfulness. This means that for systems with cyclic dependencies or in cases where two competing causal pathways exactly cancel each other, our guarantees do not hold. We see this as an important direction for future work. A key assumption in our work is that mechanism changes are independent and nondegenerate. While generally reasonable, this will not always hold in practice, and it will be interesting to generalize our theory given (limited) knowledge about intervention targets, or by including recent results on identifying intervention targets.

A practical limitation is that STRUBI relies on a conditional independence test, and therewith requires sufficiently many i.i.d. samples per context, and similarly, to determine independence of mechanism shifts, requires data from sufficiently many contexts. The availability of contexts therefore affects performance. We show that a higher context-to-dimension ratio yields better results, as a larger number of distinct contexts provides more interventional shifts, which improves identifiability. At lower ratios STRUBI remains sound, but it can only identify the base distribution and the mechanism shifts that the available contexts reveal. Evaluating all pairs of contexts (c, c') has a complexity of $\mathcal{O}(|C|^2)$, where $|C|$ is the number of distinct contexts. In most datasets $|C|$ is small to moderate, so this is unlikely to become a bottleneck.

Another practical consideration is that we need access to a causal graph built from domain knowledge (such as protein signalling, process mining pipelines, climate change data, etc.) or learnt using an off-the-shelf causal discovery algorithm. However, how well this would work depends on the extent to which the biases as present. Although STRUBI works well if the graph is not highly mis-specified, it would be interesting to explore removing the need for a graph to begin with. We could improve STRUBI's guarantees if we extend our theory and method to allow for probabilistic outputs to determine which structural bias is at play, and to amortize the process across different datasets. Another interesting idea would be to extend STRUBI to time-series data, where contexts can possibly be replaced with regimes.

Acknowledgements

Praharsh Nanavati is a Ph.D. student at the Saarbrücken Graduate School for Computer Science. He is funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – GRK 2853/1 "Neuroexplicit Models of Language, Vision, and Action" - project number 471607914.

References

[1] Rohit Bhattacharya, Tushar Nagarajan, Daniel Malinsky, and Ilya Shpitser. Differentiable causal discovery under unmeasured confounding. In *Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 2314–2322. PMLR, 2021.

[2] Begoña Canovas and Angel R Nebreda. Diversity and versatility of p38 kinase signalling in health and disease. *Nature reviews Molecular cell biology*, 22(5):346–366, 2021.

[3] David Maxwell Chickering. Learning equivalence classes of bayesian-network structures. *Journal of Machine Learning Research*, 2:445–498, 2002.

[4] Diego Colombo, Marloes H. Maathuis, Markus Kalisch, and Thomas S. Richardson. Learning high-dimensional directed acyclic graphs with latent and selection variables. *The Annals of Statistics*, 40(1):294–321, 2012.

[5] Haoyue Dai, Ignavier Ng, Jianle Sun, Zeyu Tang, Gongxu Luo, Xinshuai Dong, Peter Spirtes, and Kun Zhang. When selection meets intervention: Additional complexities in causal discovery. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2025.

[6] Haoyue Dai, Yiwen Qiu, Ignavier Ng, Xinshuai Dong, Peter Spirtes, and Kun Zhang. Latent variable causal discovery under selection bias. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2025.

[7] Robin J Evans and Thomas S Richardson. Smooth, identifiable supermodels of discrete dag models with latent variables. *Bernoulli*, 25(2):848–876, 2019.

[8] Biwei Huang, Kun Zhang, Jiji Zhang, Joseph Ramsey, Ruben Sanchez-Romero, Clark Glymour, and Bernhard Schölkopf. Causal discovery from heterogeneous/nonstationary data. *Journal of Machine Learning Research*, 21(89):1–53, 2020.

[9] David Kaltenpoth and Jilles Vreeken. Causal Discovery with Hidden Confounders using the Algorithmic Markov Condition. In *Proceedings of the International Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 1016–1026. PMLR, 2023.

[10] David Kaltenpoth and Jilles Vreeken. Identifying selection bias from observational data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 8177–8185, 2023.

[11] David Kaltenpoth and Jilles Vreeken. Nonlinear Causal Discovery with Latent Confounders. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 15639–15654. PMLR, 2023.

[12] Rickard Karlsson and Jesse Krijthe. Detecting hidden confounding in observational data using multiple environments. In *Proceedings of the Annual Conference on Neural Information Processing Systems (NeurIPS)*, pages 44280–44309. Curran, 2023.

[13] Sarah Mameche, David Kaltenpoth, and Jilles Vreeken. Learning causal mechanisms under independent changes. In *Proceedings of the Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2023.

[14] Sarah Mameche, Jilles Vreeken, and David Kaltenpoth. Identifying confounding from causal mechanism shifts. In *Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 4897–4905. PMLR, 2024.

[15] Frank J Massey Jr. The kolmogorov-smirnov test for goodness of fit. *Journal of the American Statistical Association*, 46(253):68–78, 1951.

[16] Nicolai Meinshausen, Alain Hauser, Joris M Mooij, Jonas Peters, Philip Versteeg, and Peter Bühlmann. Methods for causal inference from gene perturbation experiments and validation. *Proceedings of the National Academy of Sciences*, 113(27):7361–7368, 2016.

[17] Michelle C Mendoza, E Emrah Er, and John Blenis. The ras-erk and pi3k-mtor pathways: cross-talk and compensation. *Trends in biochemical sciences*, 36(6):320–328, 2011.

[18] Joris M Mooij, Sara Magliacane, and Tom Claassen. Joint causal inference from multiple contexts. *Journal of Machine Learning Research*, 21(99):1–108, 2020.

[19] Yasutomi Nishizuka. Protein kinase c and lipid signaling for sustained cellular responses. *The FASEB journal*, 9(7):484–496, 1995.

[20] Juan Miguel Ogarrio, Peter Spirtes, and Joe Ramsey. A hybrid causal search algorithm for latent variable models. In *Proceedings of the International Conference on Probabilistic Graphical Models (PGM)*, pages 368–379, 2016.

[21] Judea Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, 2nd edition, 2009.

[22] Ronan Perry, Julius Von Kügelgen, and Bernhard Schölkopf. Causal discovery in heterogeneous environments under the sparse mechanism shift hypothesis. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Proceedings of the Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2022.

[23] Joseph Ramsey, Madelyn Glymour, Ruben Sanchez-Romero, and Clark Glymour. A million variables and more: The Fast Greedy Equivalence Search algorithm for learning high-dimensional graphical causal models, with an application to functional magnetic resonance images. *Int J of Data Sci Anal.*, 3:121–129, 2017.

[24] Abbavaram Gowtham Reddy, Saloni Dash, Amit Sharma, and Vineeth N. Balasubramanian. Counterfactual generation under confounding. In *NeurIPS 2022 Workshop on Causality for Real-world Impact*, 2022.

[25] Thomas S Richardson, Robin J Evans, James M Robins, and Ilya Shpitser. Nested markov properties for acyclic directed mixed graphs. *The Annals of Statistics*, 51(1):334–361, 2023.

[26] Donald B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66:688–701, 1974.

[27] Karen Sachs, Omar Perez, Dana Pe'er, Douglas A Lauffenburger, and Garry P Nolan. Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, 308(5721): 523–529, 2005.

[28] Ilya Shpitser, Robin J Evans, Thomas S Richardson, and James M Robins. Introduction to nested markov models. *Behaviormetrika*, 41(1):3–39, 2014.

[29] Ilya Shpitser, Robin J Evans, and Thomas S Richardson. Acyclic linear SEMs obey the nested Markov property. In *Proceedings of the International Conference on Uncertainty in Artificial Intelligence (UAI)*. NIH Public Access, 2018.

[30] Peter Spirtes, Christopher Meek, and Thomas Richardson. An algorithm for causal inference in the presence of latent variables and selection bias. *Computation, causation, and discovery*, 21: 1–252, 1999.

[31] Peter Spirtes, Clark N Glymour, and Richard Scheines. *Causation, prediction, and search*. MIT press, 2000.

[32] Peter L Spirtes, Christopher Meek, and Thomas S Richardson. Causal inference in the presence of latent variables and selection bias. In *Proceedings of the International Conference on Uncertainty in Artificial Intelligence (UAI)*, 1995.

[33] Philip Versteeg, Joris Mooij, and Cheng Zhang. Local constraint-based causal discovery under selection bias. In *Proceedings of the Conference on Causal Learning and Reasoning (CLeaR)*, pages 840–860, 2022.

[34] Nguyen Xuan Vinh, Julien Epps, and James Bailey. Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *Journal of Machine Learning Research*, 11:2837–2854, 2010.

[35] Sascha Xu, Sarah Mameche, and Jilles Vreeken. Information-theoretic causal discovery in topological order. In *Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2025.

[36] Jiji Zhang. On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias. *Artificial Intelligence*, 172:1873–1896, 2008.

[37] Kun Zhang, Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. Kernel-based conditional independence test and application in causal discovery. In *Proceedings of the International Conference on Uncertainty in Artificial Intelligence (UAI)*, page 804–813. AUAI Press, 2011.

[38] Xun Zheng, Bryon Aragam, Pradeep K Ravikumar, and Eric P Xing. Dags with no tears: Continuous optimization for structure learning. *Proceedings of the Annual Conference on Neural Information Processing Systems (NeurIPS)*, 31, 2018.

[39] Yujia Zheng, Zeyu Tang, Yiwen Qiu, Bernhard Schölkopf, and Kun Zhang. Detecting and identifying selection structure in sequential data. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2024.

A Proofs

Throughout this section, let G_X denote the observed subgraph induced by X, and write $\mathrm{anc}^+_{G_X}(A) = \mathrm{anc}_{G_X}(A) \cup A$ for observed ancestors including the set itself. For fixed contexts c, c', define the observed local module

$$Q_i^c(x_i \mid a; \theta) := P^c_{\mathrm{obs}}\big(x_i \mid \mathrm{pa}_{G_X}(X_i) = a; \theta\big).$$

We use the standard analytic zero-set fact that a real analytic function on an open connected parameter set is either identically zero or has a Lebesgue-null zero set.

Proof of Theorem 4.2

Theorem 4.2 (Observed Local Signatures of Confounding and Selection)**.** *Let $\mathcal{G}^*$ be the true graph over V, with precisely one latent variable, $|Z \cup S| = 1$. Let A be given by*

$$A = \begin{cases} \mathrm{ch}_{\mathcal{G}^*}(Z), & \textit{for a latent confounder}, \\ \mathrm{anc}_{\mathcal{G}^*}(S), & \textit{for selection}\,. \end{cases}$$

Let $\{P^c_\theta : c \in C, \theta \in \Theta\}$ be a parametric family with $\Theta \subseteq \mathbb{R}^d$ open and connected, such that all joint, marginal and conditional densities are real analytic in θ on a non-empty common support. Let c and c' be two contexts in which only the latent variable's mechanism changes. Suppose that for every $X_i \in A$, there exists at least one $\theta_0 \in \Theta$ such that $P^c_{\mathrm{obs}}(X_i \mid \mathrm{pa}_{\mathcal{G}^*_X}(X_i); \theta_0) \neq P^{c'}_{\mathrm{obs}}(X_i \mid \mathrm{pa}_{\mathcal{G}^*_X}(X_i); \theta_0)$.

Then $A_{c,c'} = A$ for all θ outside a Lebesgue-null subset of Θ.

Proof. Fix the two contexts c, c'. We prove two statements: nodes outside the set A have invariant observed local modules for all parameter values, while nodes in A have changed observed local modules except on a null parameter set.

Latent confounding. There is no selection, so $P^c_{\mathrm{obs}} = P^c$, and the only changed intrinsic mechanism is the exogenous law of the single latent confounder Z. If $X_i \notin \mathrm{ch}_{G^*}(Z)$, then Z is not a parent of X_i. Since all other parents of X_i are observed and all observed mechanisms are unchanged, the Markov factorization gives

$$Q_i^c(\cdot \mid \mathrm{pa}_{G_X}(X_i); \theta) = Q_i^{c'}(\cdot \mid \mathrm{pa}_{G_X}(X_i); \theta)$$

for every $\theta \in \Theta$. Equivalently, any downstream effect of the shifted Z reaches X_i through observed parents and is blocked by conditioning on them.

If $X_i \in \mathrm{ch}_{G^*}(Z)$, then the latent parent is marginalized out of the observed local module. For parent values a with positive probability,

$$Q_i^c(x_i \mid a; \theta) = \int p_\theta(x_i \mid a, z)\, dP^c_\theta(z \mid a),$$

with the same conditional kernel $p_\theta(x_i \mid a, z)$ in both contexts and a context-dependent mixing law for Z. By assumption, for this X_i there is some $\theta_0 \in \Theta$ at which the two observed local modules differ. Hence, for some (x_i, a) on the common support,

$$f_i(\theta) := Q_i^c(x_i \mid a; \theta) - Q_i^{c'}(x_i \mid a; \theta)$$

is a real analytic function that is not identically zero. Its zero set is therefore Lebesgue null. Thus X_i is in $A_{c,c'}$ for all θ outside a null set.

Selection. Now suppose the single latent variable is a selector S. Then there is no latent confounder, all parents of observed variables are observed, and only the selection mechanism changes. If $X_i \notin \mathrm{anc}_{G^*}(S)$, then S is a non-descendant of X_i. By the local Markov property, $X_i \perp S \mid \mathrm{pa}_{G_X}(X_i)$, and therefore also $X_i \perp R(S) \mid \mathrm{pa}_{G_X}(X_i)$. Hence

$$Q_i^c(\cdot \mid \mathrm{pa}_{G_X}(X_i); \theta) = P^c(\cdot \mid \mathrm{pa}_{G_X}(X_i); \theta) = P^{c'}(\cdot \mid \mathrm{pa}_{G_X}(X_i); \theta) = Q_i^{c'}(\cdot \mid \mathrm{pa}_{G_X}(X_i); \theta),$$

where the middle equality uses that no observed intrinsic mechanism changes.

If $X_i \in \mathrm{anc}_{G^*}(S)$, Bayes' rule gives, for any observed parent value a with positive probability,

$$Q_i^c(x_i \mid a;\theta) = P^c(x_i \mid a, R=1;\theta) = \frac{P^c(R=1 \mid x_i, a;\theta)P^c(x_i \mid a;\theta)}{P^c(R=1 \mid a;\theta)}.$$

The factor $P^c(x_i \mid a;\theta)$ is unchanged across c, c', whereas the selection likelihood changes with the mechanism of S. The theorem assumes that this reweighting is nontrivial for at least one θ_0. As above, analyticity then implies that equality of Q_i^c and $Q_i^{c'}$ holds only on a Lebesgue-null subset of Θ.

There are finitely many observed variables. Taking the union of the null exceptional sets over all $X_i \in A$ is still null, while all $X_i \notin A$ are invariant for every θ. Therefore $A_{c,c'} = A$ for all θ outside a Lebesgue-null subset of Θ. □

Proof of Proposition 4.4

Proposition 4.4 (Ancestral Coverage of Confounding and Selection Effects)**.** *Let $\mathcal{G}^*$ be a directed Erdős-Rényi graph with independent edge probabilities p. Assume that the recovered set $A_{c,c'}$ is generated either by a single latent confounder or by a single latent selection variable.*

1. *If $A_{c,c'}$ is created by selection, then* $\mathrm{cov}(A_{c,c'}) = 1$.
2. *If $A_{c,c'}$ is created by a confounder, then* $\mathrm{cov}(A_{c,c'}) < 1$ *with probability* $1 - O(e^{-pk(d-k)/2})$.

Proof. Let $A = A_{c,c'}$, $k = |A|$, and $d = |X|$. We write

$$\mathrm{cov}_{G_X}(A) = \frac{|A|}{|\,\mathrm{anc}^+_{G_X}(A)|}.$$

If A is created by selection, Theorem 4.2 gives $A = \mathrm{anc}_{G^*}(S) \cap X$, the observed ancestors of the selector. Any observed ancestor of a node in A is again an observed ancestor of S, so $\mathrm{anc}^+_{G_X}(A) = A$ and hence $\mathrm{cov}_{G_X}(A) = 1$.

If A is created by a latent confounder, Theorem 4.2 gives $A = \mathrm{ch}_{G^*}(Z)$. For A to be ancestrally closed, no observed node outside A may be an observed ancestor of any node in A; in particular, there can be no directed edge from $X \setminus A$ into A. In the standard directed Erdős–Rényi DAG model, each cross pair contributes such an outside-to-inside edge with probability $p/2$. Therefore

$$\mathbb{P}\big(A \text{ has no outside parent} \mid |A| = k\big) \le (1 - p/2)^{k(d-k)} \le \exp\{-pk(d-k)/2\}.$$

On the complementary event, $\mathrm{anc}^+_{G_X}(A)$ strictly contains A, so $\mathrm{cov}_{G_X}(A) < 1$. Thus

$$\mathbb{P}\big(\mathrm{cov}_{G_X}(A) < 1 \mid |A| = k\big) \ge 1 - \exp\{-pk(d-k)/2\} = 1 - O(e^{-pk(d-k)/2}),$$

which proves the claimed bound. □

Proof of Proposition 4.5

Proposition 4.5 (Minimal Latent Variable Sets)**.** *Let $A_{c,c'} \subseteq X$ be nonempty and ancestrally closed in $\mathcal{G}^*$. Assume that there exists at least one parent-child pair $X_i, X_j \in A_{c,c'}$. Then among all single-latent-node explanations of the correlated set $A_{c,c'}$, the edge-minimal explanation is that of a single selector S with sink nodes of $A_{c,c'}$ as parents,* $\mathrm{pa}(S) = \{X_i \in A_{c,c'} : \mathrm{ch}(X_i) \cap X = \emptyset\}$.

Proof. Let $A = A_{c,c'}$. The sink nodes of A are the maximal nodes of the subgraph induced by A,

$$M := \{X_i \in A : \mathrm{ch}_{G_X}(X_i) \cap A = \emptyset\}.$$

Because A is ancestrally closed, every node in A is an ancestor of some node in M, and every observed ancestor of a node in M lies in A. Hence

$$A = \mathrm{anc}^+_{G_X}(M).$$

A single selector S with $\mathrm{pa}(S) = M$ therefore generates exactly the observed shifted set A by Theorem 4.2, using $|M|$ edges into the latent node.

No single-selector explanation can use fewer such edges. Indeed, suppose another selector parent set B also satisfies $\mathrm{anc}^+_{G_X}(B) = A$. For any $m \in M$, since $m \in A = \mathrm{anc}^+_{G_X}(B)$, either $m \in B$ or m is an ancestor of some $b \in B$. In the latter case $b \in A$ is a descendant of m; maximality of m in A forces $b = m$. Thus $m \in B$, so $M \subseteq B$.

A single-confounder explanation is local by Theorem 4.2: to generate exactly A, the latent confounder must be a parent of every node in A, and hence uses $|A|$ latent edges. Since A contains at least one parent-child pair, not every node of A is maximal, so $|M| < |A|$. Therefore the selector with parent set M is edge-minimal among all single-latent-node explanations. □

Coverage bound used for thresholding. If a confounded set A has size k and is not ancestrally closed, then $\mathrm{anc}^+_{G_X}(A)$ contains at least one node outside A. Consequently,

$$\mathrm{cov}_{G_X}(A) = \frac{|A|}{|\,\mathrm{anc}^+_{G_X}(A)|} \leq \frac{k}{k+1}.$$

This gives a simple finite-size thresholding heuristic for separating low-coverage confounding patterns from ancestrally closed selection patterns.

B Calculating Adjusted Mutual Information (AMI)

We show here, how we compute the AMI score for two partitions, Π_i^{obs} and Π_j^{obs}. Following [14], each variable V_i induces a partition Π_i^{obs} of the context set $\mathcal{C}$, where two contexts lie in the same block if the conditional mechanism of V_i is invariant across them. To quantify whether two variables exhibit coordinated mechanism shifts, we compare their induced partitions.

Let $\Pi_i^{\text{obs}} = \{\pi_i^1, \ldots, \pi_i^I\}$ and $\Pi_j^{\text{obs}} = \{\pi_j^1, \ldots, \pi_j^J\}$ be two partitions of $\mathcal{C}$. Their overlap is summarized by the contingency table $\mathbf{N} = (n_{ab})$, where

$$n_{ab} = \left|\pi_i^a \cap \pi_j^b\right|, \qquad u_a = \sum_b n_{ab}, \qquad v_b = \sum_a n_{ab}, \qquad N = |\mathcal{C}|.$$

As in [14], the mutual information between the two partitions is

$$\mathrm{MI}(\Pi_i^{\text{obs}}, \Pi_j^{\text{obs}}) = \sum_{a=1}^{I}\sum_{b=1}^{J} \frac{n_{ab}}{N} \log \frac{n_{ab}N}{u_a v_b},$$

with entropies

$$H(\Pi_i^{\text{obs}}) = -\sum_{a=1}^{I} \frac{u_a}{N} \log \frac{u_a}{N}, \qquad H(\Pi_j^{\text{obs}}) = -\sum_{b=1}^{J} \frac{v_b}{N} \log \frac{v_b}{N}.$$

Large values indicate that knowing the shift pattern of one variable is informative about the shift pattern of the other, which is precisely the signal we expect when both are jointly affected by the same confounder or collider.

With finitely many contexts, however, raw mutual information is positively biased even for unrelated partitions. We therefore use adjusted mutual information (AMI) [34], which subtracts the expected mutual information under random partitions and normalizes the result:

$$\mathrm{AMI}(\Pi_i^{\text{obs}}, \Pi_j^{\text{obs}}) = \frac{\mathrm{MI}(\Pi_i^{\text{obs}}, \Pi_j^{\text{obs}}) - \mathbb{E}[\mathrm{MI}(\Pi_i^{\text{obs}}, \Pi_j^{\text{obs}})]}{\frac{1}{2}\left(H(\Pi_i^{\text{obs}}) + H(\Pi_j^{\text{obs}})\right) - \mathbb{E}[\mathrm{MI}(\Pi_i^{\text{obs}}, \Pi_j^{\text{obs}})]}.$$

This yields $\mathrm{AMI} = 1$ for identical partitions and an expected value of 0 for independent random partitions, making it a more reliable finite-sample affinity score than raw mutual information. The expectation $\mathbb{E}[\mathrm{MI}(\Pi_i^{\text{obs}}, \Pi_j^{\text{obs}})]$ is computed according to the hypergeometric distribution.

As a simple example, let $\mathcal{C} = \{c_1, c_2, c_3, c_4\}$ and consider

$$\Pi_i^{\text{obs}} = \{\{c_1, c_2\}, \{c_3, c_4\}\}, \qquad \Pi_j^{\text{obs}} = \{\{c_1, c_2\}, \{c_3, c_4\}\}.$$

Then the contingency table is

$$\mathbf{N} = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix},$$

so the two partitions agree exactly and $\mathrm{AMI}(\Pi_i^{\text{obs}}, \Pi_j^{\text{obs}}) = 1$. In contrast, if $\Pi_j^{\text{obs}} = \{\{c_1, c_3\}, \{c_2, c_4\}\}$, the overlap is spread evenly across the contingency table, the mutual information vanishes, and the adjusted score is correspondingly near 0.

In our algorithm, the pairwise AMI values define the affinity matrix $\mathcal{M}$, whose connected components are candidate subsets of variables that shift together.

C Additional Results on Synthetic Data

Table 1: Experimental Parameters for Synthetic Data Generation

Parameter	Description
n	The number of samples (observations) available within each context.
m	The total number of observed variables in the causal Directed Acyclic Graph (DAG).
r	The ratio of mechanism shifts assigned to observed nodes relative to latent nodes.
mcf	Scaling factor for context availability, where the number of contexts $m_c = m \times$ *mcf*.
δ	The shift density; the ratio of contexts containing a shift to the total number of contexts (baseline $\delta = 0.48$).

Synthetic Data Generation Process We generate 25 instances of ground truth DAGs across for each node size, $n \in \{3, 5, 7, 9, 11, 13\}$. For our main experiments, we set the total number of contexts m_c to be $2n$, with 500 samples per context. Within these contexts, we allow $0.48 * m_c$ contexts (to ensure sparsity) to appear to be shifted in order to ensure sparse mechanism shifts. We set the ratio of the number of shifts per each observed node to the number of shifts per each latent node to be 0.25.

To ensure exclusive mechanism shifts while maintaining a controlled graph density, we define the shift budget allocation for latent and observed nodes as follows. The total shift budget B_{total} is defined as:

$$B_{total} = \lfloor m_c \cdot \delta \rfloor$$

Given a set of latent nodes S_l (comprising confounders and colliders) and a set of observed nodes X, we define the effective denominator D using the observed shift ratio r:

$$D = |S_l| + (r \cdot |X|)$$

The number of shifts assigned to each latent node (V) and each observed node (X_i) is then calculated as:

$$V = \max\left(1, \left\lfloor \frac{B_{total}}{D} \right\rfloor\right)$$

$$X_i = \lfloor r \cdot V \rfloor$$

To rigorously evaluate the proposed method, we conduct a series of controlled ablations across several key dimensions. These experiments are designed to test the robustness of our information-theoretic signatures under varying structural and environmental constraints.

C.1 Parameter Definitions

The synthetic data generation and subsequent evaluations are governed by the parameters detailed in Table 1. For all ablations, we vary the hyperparameter in question over the defined range, and fix the other parameters in their default values, i.e., $n = 500$, $m = 5$, our main oracle is $\textsc{StruBI}(\mathcal{G}^*)$, $r = 0.25$, $mcf = 3.0$, and $\delta = 0.48$.

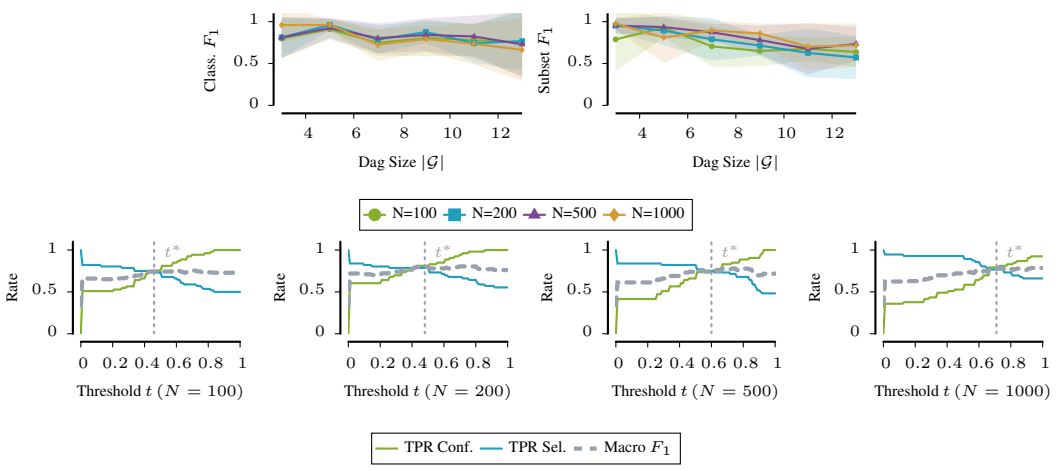


Figure 5: Ablation results for Sample Complexity across DAG sizes. The average execution runtimes are 1.57s, 2.91s, 16.79s, and 124.23s for $N = 100, 200, 500$, and 1000, respectively. We choose $n = 500$ for our main experiments as it provides a good balance of performance and runtime and matches base requirements for the baselines. The structural bias classification and subset recovery $F1$ scores (upper) are similar.

C.2 Ablation on Sample Complexity (n)

We evaluate the data efficiency and asymptotic stability of the AMI and coverage signatures by varying $n \in \{100, 200, 500, 1000\}$. This ablation is performed across a range of DAG sizes ($m \in \{3, \ldots, 15\}$) to determine the optimal number of samples required per context for reliable results. As we see in Fig. 5, we get reliable results with lower number of samples. However, we stick to 500 samples as the variance is low and runtime is not high.

The final classification is governed by a hierarchical search for optimal thresholds based on AMI and Coverage. A threshold of 0.1 on the AMI values suffices for distinguishing the presence of a latent variable. It is important to note that the thresholds (lower) obtained for $n = 100, 200$ are lower, $0.46, 0.48$ respectively, for smaller sample sizes and higher for $n = 500, 1000$ ($0.60, 0.71$ respectively) which is expected as the estimates of mutual information are less reliable with fewer samples. Since we have $n = 500$, we have $t^* = 0.6$. Subsets with coverage $< t^*$ are classified as *Confounded* as they are more prone to having localized mechanism influence, while those exceeding t^* suggest *Selection Bias* due to widespread influence across upstream parents.

C.3 Ablation on the Structural Shift Ratio (r)

This experiment investigates the impact of structural bias by varying the frequency of shifts in observed versus latent variables using $r \in \{0, 0.25, 0.5, 1.0, 1.5\}$. We assess the method's capability to isolate latent selection effects when observed nodes exhibit varying levels of mechanism volatility for graphs of size 5. As we can see in in Fig. 6, as r increases, the performance of the method degrades as the observed nodes become more volatile and the signatures become less exclusive to latent shifts. [1]

C.4 Ablation on the Shift Density (δ)

We test the robustness of STRUBI against sparsity. We ablate over shift densities $\delta \in \{0.2, 0.48, 0.6, 0.8\}$, which denote the fraction of contexts in which any causal mechanism shifts occur, across the entire set of variables, V. Higher values of δ imply noisy mechanism shifts. We

[1] An interesting note here is that, since we choose a low value for r, it subdues the noise arising from backdoors in the underlying causal graph. If the ratio is higher, then we may misclassify confounding as selection bias in many cases. This is because, due to the presence of latent variables, while looking at conditional distributions, an observed parent that is conditioned on might open up a latent v-path, leading to unforeseen correlations. This would lead to noise as well. Hence, having a low r ratio also helps us in reducing the chance of seeing spurious effects via backdoors.

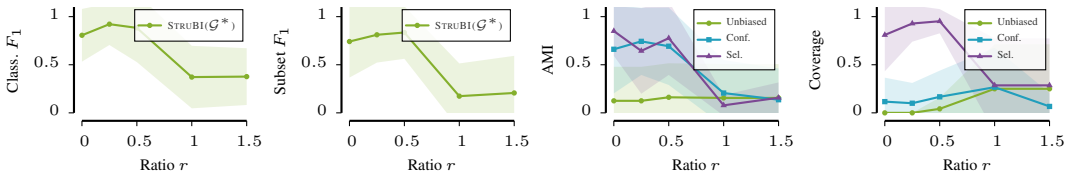


Figure 6: Ablation over values of ratio r. As r increases, performance metrics decline, indicating higher sensitivity to observed node noise.

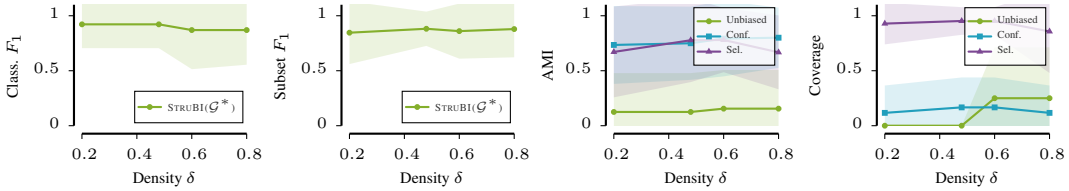


Figure 7: Ablation over the fraction of contexts in which variables undergo causal mechanism shifts. StruBI($\mathcal{G}^*$) works well even when the assumption is relaxed.

show in Fig. 7 that StruBI works well even if we don't have sparse shifts. However, we need independent and nondegenerate shifts in order to separate the effects of a latent variable from the unbiased cases.

C.5 Ablation on number of Contexts (m_c)

We perform StruBI's effectiveness across the number of given contexts. We vary *mcf* $\in \{0.5, 1.0, 2.0, 3.0\}$ which is the ratio between the number of observed nodes in the graph, and the number of contexts m_c. More the number of contexts, less the number of noisy mechanism shifts detected. As we see in Fig. 8, higher the ratio, better the performance.

C.6 StruBI's performance on incorrect graphs

To simulate structural errors in causal discovery (e.g., from a search algorithm), we introduce noise into the DAG via edge-flipping. We evaluate the *Perturbed* oracle at $m = 5$ with flip fractions, $f \in \{0.1, 0.25, 0.4, 0.5\}$ in Fig. 9. This measures the degree of structural degradation the algorithm can tolerate before the identification of latent types becomes unreliable.

C.7 StruBI and its variants

To rigorously evaluate the framework components, we define three oracles:

- StruBI*(Π^{obs}):* This variant receives the partitions and then performs classification.
- StruBI*($\mathcal{G}^*$):* This variant operates on the true underlying causal graph.
- StruBI*($\hat{\mathcal{G}}$):* This variant operates on a learned causal graph using Topic.

We showcase the performance on structural bias classification of these variants according to different metrics in Fig. 10. The runtimes for each oracle across different DAG sizes are given in Table 2.

C.8 Analysis of runtimes

All experiments were conducted on an Apple M3 Pro (14-core CPU, 36GB RAM), utilizing 10 CPU cores to parallelize all context evaluations via Python's multiprocessing. Our main oracle, StruBI($\mathcal{G}^*$) exhibits a mean runtime of approximately 34s for 13-node graphs. The StruBI(Π) oracle takes less than 1s across all sizes. The StruBI($\hat{\mathcal{G}}$) variant, where the underlying causal graph is estimated using Topic, closely tracks the StruBI($\mathcal{G}^*$)'s performance. As for the baselines, the

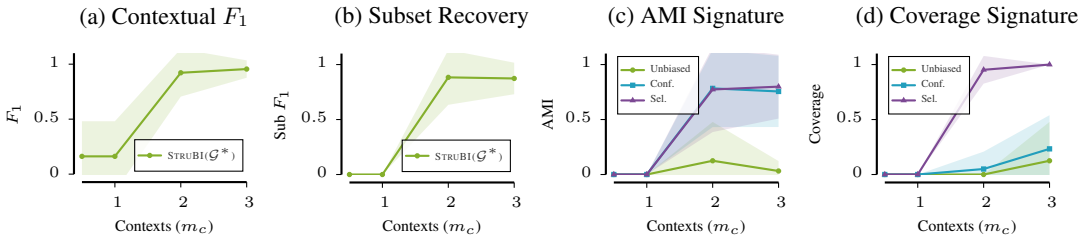


Figure 8: Ablation over the ratio of the number of observed variables with the total number of contexts. The X-axis denotes the factor multiplied to the DAG size to obtain the contexts.

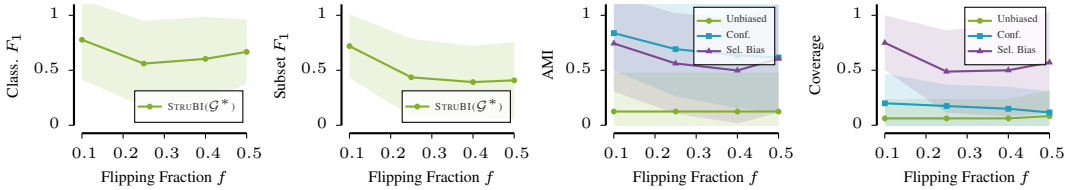


Figure 9: Ablation of performance mettrics across flipping fractions. Since the performance drop is not huge, we can say that StruBI can accommodate graph mis-specification.

Algorithm 2: StruBI Suite: Discovery and Classification

Sub-procedure A: StruBI(Π^{obs})
input : Partitions Π^{obs} for every observed variable
output : Subsets of X suffering from selection bias or confounding.

Construct an affinity matrix $\mathcal{M}$ where $\mathcal{M}_{ij} = \text{AMI}(\Pi_i^{\text{obs}}, \Pi_j^{\text{obs}})$;
Identify subsets $X_S \subseteq X$ as connected components in $\mathcal{M}$;
if $X_S = \emptyset$ **then**
  return Unbiased
else
  foreach identified subset X_S **do**
    Calculate coverage $\text{cov}(X_S) \geq \tau$? *Selection Bias* : *Confounding*;
  return Subsets X_S with labels;

Sub-procedure B: StruBI($\mathcal{G}$)
input : Observed data over X, C; observed causal graph $\mathcal{G}$.
foreach variable $X_i \in X$ **do**
  Compute $p_{c,c'}$ and convert to partition Π_i^{obs};
return StruBI(Π);

Sub-procedure C: StruBI($\hat{\mathcal{G}}$)
input : Observed data over X, C
$\hat{\mathcal{G}}$ = Topic (X, C);
return StruBI($\hat{\mathcal{G}}$);

latent selection variants (LS-Pooled, LS-Contexts) and FCI variants are very fast, with sub-second execution times even for the larger graphs, CoCo scales much more aggressively, and is slower than StruBI($\mathcal{G}^*$). The runtimes for each oracle are shown in Table 2 and the baselines in Fig. 14 in Section C of the Appendix.

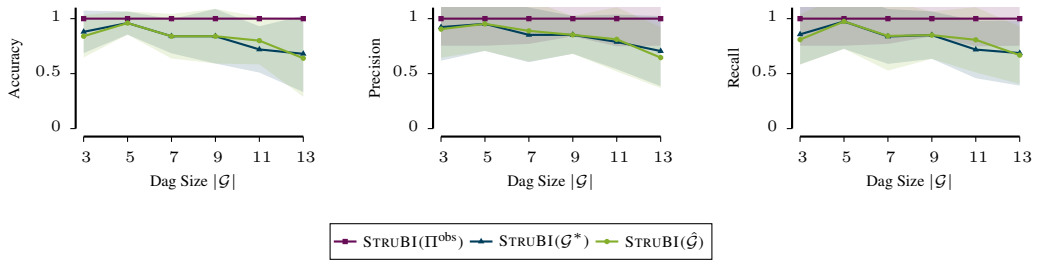


Figure 10: The accuracy, precision, and recall for StruBI's oracles for the task of structural bias classification

Table 2: Runtimes across variants and graph sizes (n).

Oracle	n	**Runtime (s)**	**Oracle**	n	**Runtime (s)**
full	3	0.19 ± 0.01	perturbed	3	0.38 ± 0.02
	5	1.40 ± 0.31		5	2.89 ± 0.55
	7	4.51 ± 0.72		7	9.24 ± 1.02
	9	10.79 ± 0.82		9	21.46 ± 1.49
	11	19.92 ± 1.37		11	40.70 ± 2.82
	13	34.47 ± 2.56		13	68.37 ± 4.48
TOPIC	3	0.27 ± 0.40	partitions	3	0.38 ± 0.78
	5	1.39 ± 0.34		5	0.13 ± 0.13
	7	4.35 ± 0.70		7	0.21 ± 0.15
	9	10.96 ± 0.73		9	0.31 ± 0.18
	11	19.39 ± 1.47		11	0.37 ± 0.24
	13	35.13 ± 2.77		13	0.51 ± 0.24

C.9 Graph Density Ablation

In our base experiments, we perform our experiments on Erdos-Renyi graphs with graph density $p = 0.5$. Since the coverage gap is very high between the confounded and selection cases, this threshold can be chosen flexibly. We obtain a suitable threshold, τ_2, for the coverage values to distinguish confounding from selection bias by analyzing the TPR rates w.r.t. τ_2. However, if we increase the graph density, we see a rise in the coverage of the confounding cases due to (1) natural coverage increase, and (2) spurious mechanism shifts propagating. We look at the average coverage values for selection and confounding via an ablation on p for 500 samples in table 3.

We see that for very low values of p, i.e., where there is sparsity, coverage becomes a weaker metric as upstream ancestry reduces. We show in table 4 for high p, especially in large graphs, the shift effects spillover and noise increases. As we see with these results, coverage becomes less effective in larger graphs with high edge densities due to noise and natural coverage overlaps, and also in small sparse graphs as shifts cannot propagate.

C.10 Dependent and Overlapping Shifts

Assuming that the mechanisms do not appear to be correlated due to the presence of a latent variable, distinguishing if the shifts appear due to a confounder, selection criteria, or from their intrinsic shifts becomes difficult. If intrinsic shifts of the variables themselves are correlated, then it is very easy to mistake it for the presence of a latent variable. However, usually mechanisms appear to be correlated due to the presence of confounders and selection criteria. StruBI might infer an extra latent variable due to the correlated shifts, and also not retrieve the entire affected node set. However, if the correlations are not very high, then StruBI can recover a fraction of the affected subset of nodes while determining the type of bias at play. It is most likely that interventions appear to be correlated due to the presence of a confounder or a selection criteria, in practice. We perform an ablation on the overlap in the intrinsic shifts of all variables in table 5. Here, the Overlap Ratio determines the pair-wise overlap of shared shifts. We define a pairwise overlap ratio, α and define it as shown in table 6.

Table 3: Average coverage for colliders and confounders across varying graph sizes and densities.

Graph Size	No. of Samples	Graph density (p)	Avg Coverage (Collider)	Avg Coverage (Confounder)
5	500	0.1	0.167 ± 0.408	0.167 ± 0.408
5	500	0.3	0.833 ± 0.408	0.000 ± 0.000
5	500	0.5	1.000 ± 0.000	0.083 ± 0.204
5	500	0.7	0.944 ± 0.136	0.167 ± 0.258
5	500	0.9	0.944 ± 0.136	0.181 ± 0.214
9	500	0.1	0.500 ± 0.527	0.229 ± 0.407
9	500	0.3	0.883 ± 0.315	0.256 ± 0.323
9	500	0.5	0.840 ± 0.312	0.384 ± 0.211
9	500	0.7	0.969 ± 0.065	0.499 ± 0.195
9	500	0.9	0.968 ± 0.071	0.479 ± 0.123
13	500	0.1	0.677 ± 0.451	0.200 ± 0.283
13	500	0.3	0.745 ± 0.262	0.722 ± 0.137
13	500	0.5	0.860 ± 0.191	0.855 ± 0.077
13	500	0.7	0.748 ± 0.315	0.913 ± 0.005
13	500	0.9	0.735 ± 0.200	0.833 ± 0.118

Table 4: Results across varying graph sizes and densities (p). '*' denotes no overlaps in the coverage metric, and 'X' denotes large overlaps (>0.4). For cases with nonzero overlaps in coverage, we chose values in between the overlapping regions.

Graph Size	$p = 0.1$	$p = 0.3$	$p = 0.5$	$p = 0.7$	$p = 0.9$
5	X	$[0.000, 0.425]^*$	$[0.287, 1.000]^*$	$[0.425, 0.808]^*$	$[0.394, 0.808]^*$
9	X	$[0.569, 0.579]$	$[0.529, 0.595]^*$	$[0.694, 0.904]^*$	$[0.602, 0.897]^*$
13	$[0.226, 0.483]$	$[0.483, 0.858]$	$[0.669, 0.932]$	X	X

The accuracy, precision, recall, F1 score, drop as the ratio increases. The increase in the mean Structural Hamming Distance (SHD) is quite steep, i.e., subset recovery is rapidly worsening with an increase in α. However, the drop in bias classification is not as rapid. This implies that StruBI can tell apart if the system is affected by confounding or selection bias, but is not retrieving the correct set of nodes suffering from this bias, as the shifts in the observed variables are now overlapping. As we see, if the graph is too sparse, then shifts don't propagate, and if it is too dense, then latent detection is overridden, as structurally, the coverage values come closer.

C.11 Out of Sample Generalization

In our experiments, we learn $\hat{\mathcal{G}}$ and use StruBI ($\hat{\mathcal{G}}$) on the same samples. This is because we wish to make use of whatever samples we can in order to learn the causal structure, and also conduct independence tests. However, using a separate sample split would increase statistical confidence in our method. We show here, an ablation with stratification with 250 samples per context, and see that the performance stays more or less consistent. We show our results in Tables 7 and 8.

C.12 Simultaneous confounding and selection bias

Detecting the presence of a confounder as well as a selection criteria is challenging due to issues with identifiability. For instance, if the selection criteria occurs downstream of confounding, then the effects of confounding would be subsumed by the selection node. It is also likely that unless the graph is big and the affected subsets are significantly far apart, the presence of such latent variables would give rise to backdoors. Doing this robustly would be a very important future direction. We provide an ablation over larger graphs, i.e., sizes 9, 11, and 13, and report the classification results in Table 9.

Table 5: Performance metrics across varying node sizes and overlap ratios (α).

Node Size	Overlap Ratio (α)	Accuracy	Precision	Recall	F1	Mean SHD
5	0.00	1.0000	1.0000	1.0000	1.0000	0.55
5	0.33	0.8800	0.8978	0.8800	0.8793	0.84
5	0.66	0.9200	0.9378	0.9200	0.9193	0.68
5	1.00	0.7600	0.8317	0.7600	0.7093	1.16
9	0.00	0.9000	0.9286	0.9000	0.9046	1.65
9	0.33	0.7200	0.8400	0.7200	0.7240	2.36
9	0.66	0.8000	0.8697	0.8000	0.7977	2.20
9	1.00	0.4800	0.4933	0.4800	0.4157	2.12
13	0.00	0.9231	0.9385	0.9231	0.9219	3.62
13	0.33	0.9231	0.9385	0.9231	0.9219	4.77
13	0.66	0.9167	0.9333	0.9167	0.9153	2.92
13	1.00	0.6667	0.7738	0.6667	0.6258	4.17

Table 6: Definition of overlap ratios (α) and their corresponding shift mechanics.

Overlap ratio (α)	Description
0.0	$\text{shifts}(V_1) \cap \text{shifts}(V_2) = \emptyset$
0.5	$\lvert\text{shifts}(V_1) \cap \text{shifts}(V_2)\rvert = 0.5 \times \min(\lvert\text{shifts}(V_1)\rvert, \lvert\text{shifts}(V_2)\rvert)$
1.0	$\text{shifts}(V_1) = \text{shifts}(V_2)$

C.13 STRUBI with CDNOD

We include an ablation study where we learn the causal graph using CDNOD [1] instead of TOPIC, as it is a constraint-based causal discovery method for non-stationary and heterogeneous data. Since CDNOD is a constraint-based method, it does not scale well, and since it is a conditional independence-based algorithm, it is also slower to implement. In certain cases, it fails to detect sparse shifts. However, STRUBI can work with off-the-shelf causal discovery methods.

C.14 STRUBI with class imbalance

In our base experiments, we use equal number of samples across all environments to establish a balanced baseline dataset. In this ablation, we induce imbalance by randomly down-sampling each environment independently, retaining a uniform random fraction of data between 50As for the imbalance and high variability, this depends mainly on the performance of the independence test being used. Lastly, the availability of multiple contexts is required in theory, for the sparsity assumption to hold. However, as we show in practice with the Sachs dataset, this need not hold true strongly.

Reduced sample size primarily affects the statistical power of the pairwise KCI tests. However, as long as KCI can detect mechanism shifts fairly accurately, STRUBI performs well. Our partitioning rule accounts for varying sample sizes by thresholding each node's shift frequency adaptively, relative to the observed spread between the highest and lowest number of shifts for that specific node. In the case of severe imbalance, the chance of detecting a shift for a genuinely shifted but rare context drops. If these true shifts go undetected, it leads to an underestimation of the recovered subset, which can subsequently hamper the structural bias classification. Therefore, while our adaptive thresholding mitigates moderate imbalance, extreme rarity of a context fundamentally limits the ability to perform reliable independence testing.

C.15 Additional Details on Baselines

LS-Pooled The first baseline [5], treats the non-i.i.d. dataset as a unified, exchangeable sample by marginalizing over all context labels. In this approach, causal discovery is performed in a single pass over the aggregated data matrix, which maximizes the sample size to increase the statistical power of the underlying conditional independence and rank tests.

Table 7: In-sample performance across varying graph sizes.

Graph Size	Bias Type Accuracy	Mean SHD
3	0.60	0.52
5	0.88	3.08
7	0.56	7.56
9	0.52	15.92

Table 8: Out-of-sample performance across varying graph sizes.

Graph Size	Bias Type Accuracy	Mean SHD
3	0.64	0.56
5	0.92	3.24
7	0.68	7.48
9	0.64	15.20

LS-Contexts The second baseline [5], explicitly performs structural discovery for each context. After this, the latent structure recovered in a majority of the contexts is retained.

Additional Comparisons with Baselines We show some extra results on the baselines in this section. Fig. 11 contains the reported error bars of the baselines in comparison with STRUBI. Fig. 12 gives the accuracy, precision, and recall on the task of structural bias classification, Fig. 13 compares STRUBI with the [5] variants on their data assumptions, and Fig. 14 compares the runtime of all baselines.

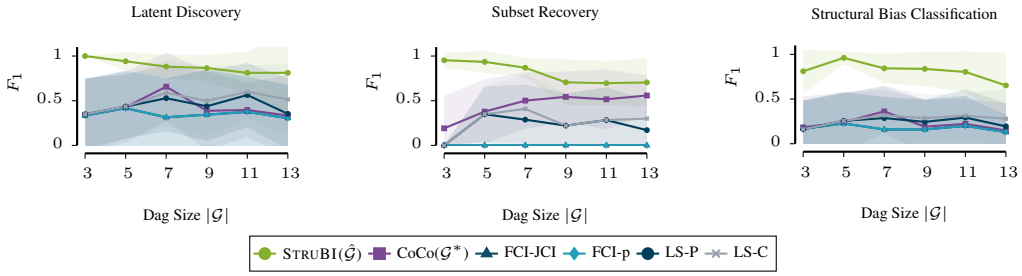


Figure 11: F1 scores on (a) latent variable discovery, (b) affected subset recovery, and (c) structural bias classification. This figure also shows the error bars, which were not included in Figure 3 for clarity.

C.16 Extended Discussion on Baselines

Supplying the true DAG to baseline methods such as the LS (Dai et al., 2025) and FCI variants is infeasible due to the algorithmic architecture and strict structural assumptions inherent to these approaches. We provide the true graph to CoCo as it supports taking the entire graph as input. Our experiments suggest that STRUBI ($\hat{\mathcal{G}}$) outperforms CoCo($\mathcal{G}^*$). The FCI variants are designed to infer a Partial Ancestral Graph (PAG) directly from empirical conditional independence tests. Their capacity to identify latent confounding is intrinsically tied to this structure-learning phase (e.g., through the orientation of bi-directed edges). Injecting a fully specified DAG as background knowledge circumvents this core mechanism, stripping the algorithm of its ability to detect latents and often leading to degenerate outputs if the fixed graph conflicts with empirical data.

The LS variants also cannot accept a DAG as input. These methods strictly require a "one-factor model" structure, where each latent variable must have at least two purely observed proxy measurements with no other parents. Furthermore, they rely on linear Gaussian data generation and linear

Table 9: As we expect, the subset recovery in these cases is poor, due to poor detection of mechanism shifts, due to presence of latent backdoors, upstream selection bias propagation, and natural increase in coverage due to an increase in graph size.

N	Precision	Recall	F1	Subset Recovery
9	97.4%	74.0%	77.1%	32.0%
11	97.7%	84.0%	80.6%	42.0%
13	95.6%	86.0%	83.4%	42.0%
Overall	96.8%	81.3%	80.5%	38.7%

Table 10: Accuracy and Macro F1 scores across varying node sizes.

Node Size	No. of Samples	Accuracy	Macro F1
3	500	0.8571	0.7937
5	500	0.8182	0.7778
7	500	0.7273	0.7026
9	500	0.7143	0.7074

selection mechanisms. Unless an input graph strictly conforms to this narrow measurement structure and parametric regime, it cannot be processed by these baselines. Unlike their method, we do not make any functional assumptions in order to identify selection bias. To address this and ensure a fair comparison, we use TOPIC in order to first learn a graph and then evaluate our method on it, including an ablation on mis-specified graphs. We selected TOPIC as our primary external discovery tool because its approach is well-suited for capturing non-linear dependencies without requiring strict parametric assumptions, works for multi-context data, and is faster to implement and outperforms CAM, SCORE, RESIT, and Notears.

We would also like to point out, that STRUBI cannot work with a MEC or from scratch, as the presence of selection bias correlates the shifts of its upstream variables, making their shifts symmetrical in nature, making it difficult for criteria like the Bayesian Information Criterion, or Minimium Description Length to help in compression, and hence, the orientation of the edges. This can be possible with more structural or functional assumptions and would be an interesting direction for future work.

Supplying the true graph as background knowledge is technically available to the FCI variants, but does not improve their results. Both identify latent confounding through an orientation phase that relies on the separating sets discovered during an adaptive, CI-test-driven search, and fixing the skeleton in advance removes this evidence, often producing degenerate orientations where the imposed graph conflicts with the empirical independencies.

The LS variants, on the other hand, admit no comparable mechanism, since their one factor measurement model requires each latent to have at least two purely observed, parent-free proxies, together with linear-Gaussian generation and linear selection (which is a very strong assumption), and has no analogous slot for a general DAG. We accordingly run all baselines from scratch, with the exception of CoCo, which we supply with the true graph, as it is explicitly designed to accept one. Hence, for a fair comparison with STRUBI, we now have results with two off-the-shelf-causal discovery algorithms. STRUBI's main contribution comes after this phase. It is to distinguish selection bias and confounding, via the Adjusted Mutual Information as well as the Coverage. To be able to do this, without any parametric assumptions, is our main contribution. This is a difficult problem as the presence of selection bias correlates the shifts of its upstream variables, making their shifts symmetrical in nature, making it difficult to learn causal structure.

D Additional Results on the Cell-Signaling Data

The Sachs et al. [27] dataset is particularly interesting as it contains multiple contexts with known interventions. However, it is also very challenging as the system is complex and has many cyclic dependencies [16] and hence these interventions might affect multiple nodes in overlapping contexts,

Table 11: STRUBI can work with different number of samples per context as well. However, the performance drops slightly, compared to scenarios where each context has the same number of samples.

Node Size	Imbalance Type	AMI Threshold	Coverage Threshold	Accuracy	Subset Recovery F1
5	mild_imbalance	0.01	0.67	0.7200	0.7244
5	extreme_imbalance	0.45	0.61	0.8000	0.7963
9	mild_imbalance	0.01	0.84	0.7600	0.7449
9	extreme_imbalance	0.01	0.84	0.7200	0.6898
13	mild_imbalance	0.01	0.64	0.7778	0.8056
13	extreme_imbalance	0.01	0.82	0.7941	0.7965

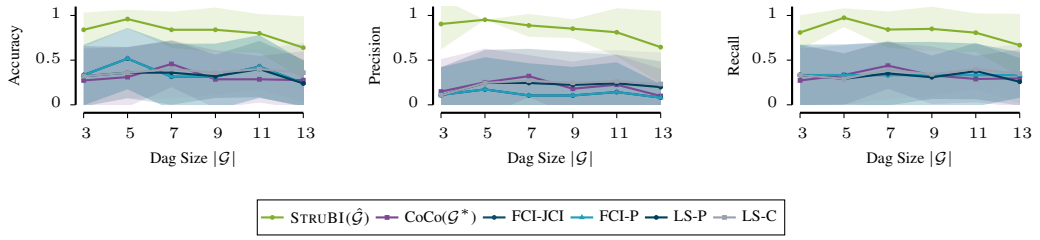


Figure 12: STRUBI outperforms the baselines for structural bias classification in terms of accuracy, precision, and recall as well.

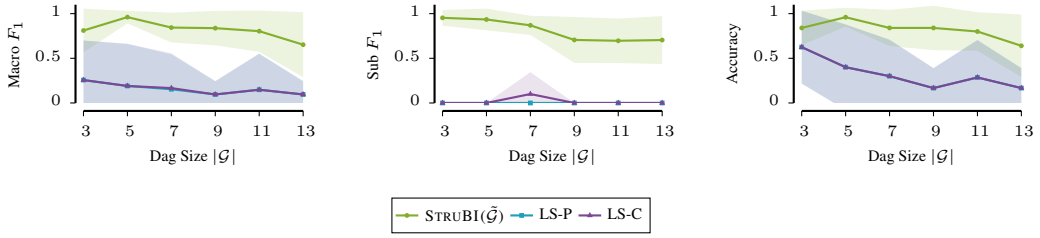


Figure 13: Our TOPIC oracle outperforms *LS-Pooled* and *LS-Contexts* on linear data with Gaussian and uniform noise sampled with equal probability. These are the primary paramteric assumptions by [6].

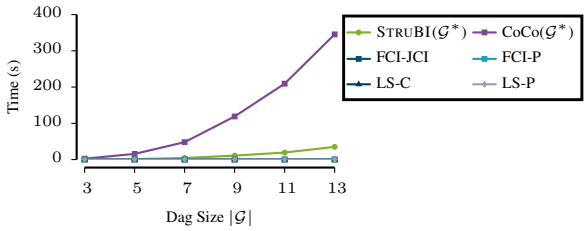


Figure 14: Runtime Analysis for all baselines across different graph sizes.

leading to noisy partitions. In the unbiased setting, STRUBI finds overlaps in causal mechanisms of Akt-Mek, Plcg-P38, and PKA-PKC. This can be due to biological crosstalk.

For instance, the dependencies detected between Akt-Mek may be attributed to the extensive pathway crosstalk between the Ras-MAPK and PI3K-mTORC1 cascades [17]. Similarly, the identified dependencies between PKA-PKC likely reflect shared regulation via phospholipase C(PLC) signaling; such simultaneous activation pathways can render conditioning on PLCγ alone insufficient for isolation [19]. Finally, the overlapping shifts in Plcg-P38 may reflect parallel activation by common upstream MAP3Ks in response to shared extracellular growth or stress signals [2].

E Discussion on the Sachs Dataset [27]

It is important to note, that in noisy datasets, if the mechanism shifts are not detected, then coverage does not remain robust in the presence of cycles as the fraction of upstream nodes subsumed in the case of confounding and selection might not diverge. Furthermore, unlike PKC, which was chosen as our primary case study due to its strong functional dependencies across environments, Erk is highly robust to mechanism shifts thereby not producing strong enough shifts in its children or ancestors. Even under direct interventions targeting PKC, Erk's mechanisms remain entirely stable. This lack of variation deprives the algorithm of the necessary cross-environmental signal needed to overcome the dataset noise and cyclical structures. CoCo was also unable to detect its effects Erk serve as potential other confounders, but CoCo does not detect their effects on its children, for the confounding case.

The mechanism shifts we obtain are given in Table 12 and the retrieved latent structures by the other baselines in Fig. 15.

Node	Causal Mechanism Shifts
Akt	**C0** (T cell activation), **C2** (AKT inhibitor)
Erk	**C3** (G06976), **C8** (beta2cAMP)
Jnk	No shifts
Mek	**C0** (activation), **C2** (AKT inhibitor), **C4** (MEK1/2), **C6** (LY294002)
P38	**C1** (ICAM-2/LFA-1), **C7** (PMA/PIP2 inhibition)
PIP2	**C3** (G06976), **C5** (U0126)
PIP3	**C0, C1, C2, C3, C4**, **C7**
PKA	**C3** (AKT inhibition), **C4** (MEK inhibition)
PKC	**C4** (Psitectorigenin)
Plcg	**C1** (LFA-1), **C7** (PIP2 inhibition), **C8** (beta2cAMP)
Raf	**C0** (activation), **C4** (MEK inhibition), **C7** (PMA), **C8** (beta2cAMP)

Table 12: Detailed inventory of causal mechanism shifts identified across the signaling network nodes.

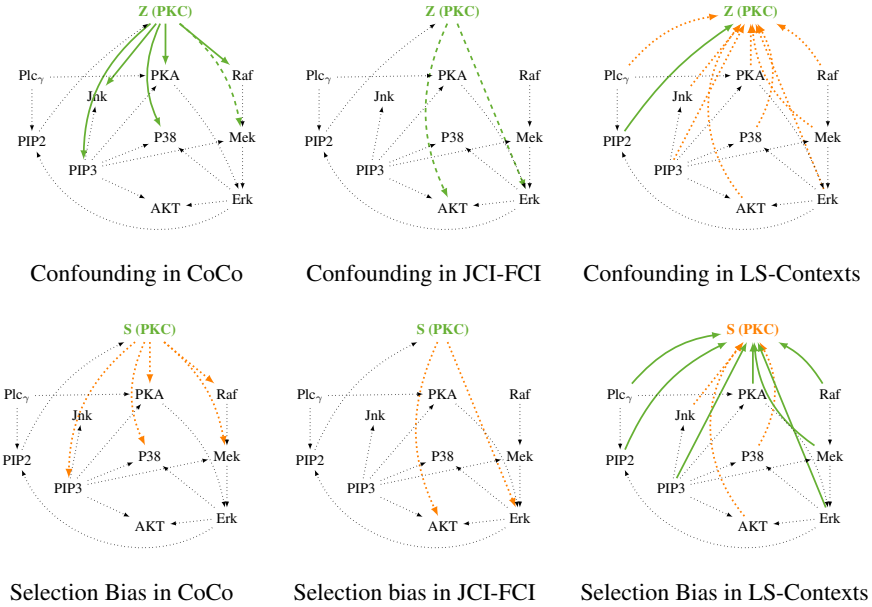


Figure 15: Retrieved latent structures across baselines.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Yes, we propose, support, and evaluate the claims made in the abstract and introduction in the remaining sections of the paper.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Yes, the limitations of the work performed can be found in Experiments, Discussions and Conclusion, and the Appendix.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: Yes, the paper contains all sets of assumptions stated clearly in the preliminaries and proofs for each theoretical result in the Appendix.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: All implementation details are provided in the Experiments section and the Appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: Yes, the entire code is shared, with implementation details specified clearly in a supplementary zip file.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Yes, all implementation details are provided in Experiments, and the Appendix, along with hyperparameters.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Yes, the paper reports error bars suitable and correctly along with error bars in the Experiments section and the Appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Yes, the paper provides a comprehensive summary of the compute resources and analysis of runtimes in the Appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Yes, the authors follow the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Yes, the paper discusses safety-critical applications of the proposed method and gives examples of real-world settings which can gain from our method.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: [N/A]

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: Yes, original owners of assets are appropriately and carefully cited.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: Yes, the new assets introduced in the paper are well documented, and the documentation is provided.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: [N/A]

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: [N/A]

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.

- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: [N/A]

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.