Identifying Confounding from Causal Mechanism Shifts

Sarah Mameche **Jilles Vreeken** **David Kaltenpoth**
CISPA Helmholtz Center for Information Security

Abstract

Causal discovery methods commonly assume that all data is independently and identically distributed (i.i.d.) and that there are no unmeasured confounding variables. In practice, neither is likely to hold, and detecting confounding in non-i.i.d. settings poses a significant challenge. Motivated by this, we explore how to discover confounders from data in multiple environments with causal mechanism shifts. We show that the mechanism changes of observed variables can reveal which variable sets are confounded. Based on this idea, we propose an empirically testable criterion based on mutual information, show under which conditions it can identify confounding, and introduce CoCo to discover confounders from data in multiple contexts. Our experiments confirm that CoCo works well on synthetic and real-world data.

1 INTRODUCTION

In most scientific fields, we aim to understand a system of interest not just in terms of statistical regularities, but in terms of its underlying causal mechanisms. A causal understanding is necessary to predict how a system will behave upon intervention, and hence relevant to guide study design, medical treatment formulation, and effective intervention targeting (Pearl, 2009a).

Yet, determining causality is notoriously challenging. While it can be established through controlled experiments, direct control is often impractical or infeasible. Consider genomics: while modern tools can activate or silence specific genes directly, these changes often introduce unforeseen consequences downstream of the targeted gene, known as off-target effects (Dominguez et al., 2016). Through observational studies, on the other hand, we only capture statistical correlations that could be spurious and thus misleading.

Most prominently, spurious correlations result from unmeasured confounding factors. A standard practice is to assume that we measure *all* relevant variables such that the system is *causally sufficient*, which assumes this problem away. In practice, however, sufficiency is neither realistic nor verifiable. In addition, spurious correlations can also appear due to distribution shifts. The common assumption that we obtain *independently and identically distributed (i.i.d)* samples does not apply in many practical applications where data originates from different contexts.

Several approaches to causal discovery relax sufficiency, but some offer ambiguous results (variables *might* be confounded) (Spirtes, 2001; Bhattacharya et al., 2021), others require strong parametric assumptions (Kaltenpoth and Vreeken, 2023a,b). Conversely, a growing literature in causal discovery accommodates non-i.i.d. data (Huang et al., 2020; Mooij et al., 2020; Perry et al., 2022) but existing work only allows for limited confounding effects (Huang et al., 2020). Motivated by this, we take a closer look at the interplay of the sufficiency and i.i.d. assumptions, with the aim of discovering latent variables from multi-context data.

We show that, surprisingly, confounding is identifiable without further assumptions on the type of data or the functional form of causal mechanisms, by considering distribution shifts of the observed variables. We explain this by example of three variables in five contexts in Fig. 1. As the graphical model G shows, X is a cause of Y and the pair is either unconfounded (left column) or confounded (right column) through a latent variable Z. In each context, interventions affect specific variables (hammers) and change their underlying generating process $P^\star$ (colored boxes).

Our main observation is that changes in an unmeasured common cause will translate into measurable changes in the observed variables. For example, consider the intervention on Z in context c_1 (purple). As this does not directly affect the causal mechanisms of

Proceedings of the 27th International Conference on Artificial Intelligence and Statistics (AISTATS) 2024, Valencia, Spain. PMLR: Volume 238.

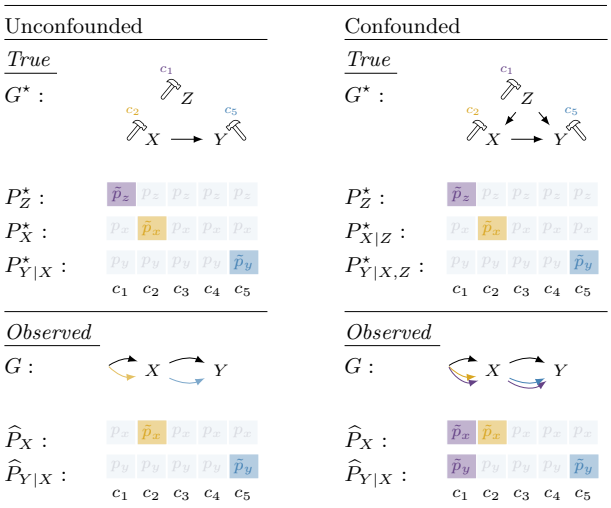


Figure 1: *Confounding introduces dependent mechanism shifts.* We consider two systems, one causal (left) and one confounded (right) in five different contexts. The true causal mechanisms $P^\star$ (top) change independently of each other, here due to targeted interventions in certain contexts (colored). If there is an unobserved confounder Z, however, we observe dependencies in the mechanism shifts of X and Y (bottom right).

the other variables, the conditionals $P^\star_{X|Z}$ and $P^\star_{Y|X,Z}$ remain the same; they change independently in other contexts (yellow, blue). However, the situation is different for the observed distributions $\widehat{P}$ (bottom). Under confounding, we observe changes for both variables in c_1 as Z is unaccounted for. Guided by this, we propose detecting confounders from correlations in causal mechanism shifts. We interpret such changes as discrete clusterings over contexts, as illustrated by the colored boxes in Fig. 1, and measure their dependency using mutual information. We show that allows to discover not only confounders, but also the variable sets that are affected by the same latent variable, with or without knowing the true causal directions.

Contributions To summarize, we show how confounding can be identified from violations of the independent mechanism shift principle (Janzing and Schölkopf, 2010), a special form of the independent causal mechanisms assumption (Peters et al., 2017). We propose an empirical estimator based on mutual information and give an analysis of both bivariate and multivariate settings, and show that they remain robust when the true causal directions among observed variables are unknown. To apply our theoretical results, we propose the CoCo algorithm for discovering confounding in different contexts. Empirically, we show that our approach effectively discovers latent variables and gives insights into cell signalling.

2 PROBLEM AND ASSUMPTIONS

In this section, we present our problem setting and state our assumptions.

2.1 Problem Setting

We consider a system of observed variables X and unobserved variables Z, collectively called $V = X \cup Z$. The values of X, Z may be continuous, categorical, or mixed. We assume that the system is observed in multiple contexts, represented by a categorical variable C taking values $c \in \mathcal{C}$, and denote their number $n_c = |\mathcal{C}|$. We allow the distribution $P^c(V) = P(V \mid C = c)$ to depend on c, as we describe in the following section.

For any fixed $c \in \mathcal{C}$, we assume the causal relationships between variables V to be described by a Directed Acyclic Graph (DAG) $\mathcal{G}^* = (V, E)$ with edges $(i, j) \in E$ when V_i is a cause of V_j. We write pa_i^* for the set of direct causes of V_i in $\mathcal{G}^*$. W.l.o.g. the indices of X_i and V_i are assumed to be ordered such that, whenever clear from the context, we can write pa_i^* to denote the parents of X_i. We assume the latent Z to be jointly independent and exogenous to X—there exist no edges $X_i \to Z_j$—so that for all Z_j, $\mathrm{pa}^*_{Z_j} = \emptyset$.

We assume causal sufficiency over $X \cup Z \cup C$, i.e., all common parents of two or more observed variables are themselves included among X, Z, C. We can now state our problem informally as follows.

Problem Statement. Given data over the observed variables X in contexts C, we want to determine which observed variables in X are jointly confounded.

2.2 Causal Mechanism Shifts

We now describe the data-generating process across multiple contexts. While we assume that the same causal structure applies in all contexts, in many applications such as gene editing experiments (Dominguez et al., 2016), a system is subject to interventions or other causal mechanism changes. That is, the generating process $P^c(V_i \mid \mathrm{pa}_i^*)$ of each variable V_i may be different across contexts. Nevertheless, as interventions typically affect only a small number of causal mechanisms at a time, the causal mechanism governing a specific V_i will generally be the same for most $c \in \mathcal{C}$ and differ only in few. To represent this, for every variable V_i, we partition the contexts so that within each set, the causal mechanism remains constant. That is, for each V_i we have a partition $\Pi_i^* = \left\{\pi_i^1, \ldots, \pi_i^{k_i}\right\}$ of $\mathcal{C} = \pi_i^1 \cup \ldots \cup \pi_i^{k_i}$ into disjoint π_i^j such that $P^c(V_i \mid \mathrm{pa}_i^*) = P^{c'}(V_i \mid \mathrm{pa}_i^*)$ for c, c' in the same $\pi \in \Pi_i^*$. We refer to this part π as $\Pi_i^*(c)$ and call the corresponding mechanism $P^\pi(V_i \mid \mathrm{pa}_i^*)$.

We allow all partitions Π_i^* to be distinct. More precisely, we regard partition Π_i^* of the contexts $\mathcal{C}$ as a random variable, and assume some joint distribution $P(\Pi^*)$ over all partitions Π_i^* of V_i. We hence assume the distribution of the observed V to be as follows.

Assumption 1 (Markov Property under Mechanism Changes) *The distribution $P(V)$ is given by*

$$\begin{aligned} P(V) &= \int P^C(V) dP(C) \\ &= \int \prod_i P^{\Pi_i^*(C)}(V_i \mid \mathrm{pa}_i^*) dP(C) \\ &= \int \prod_i P^{\Pi_i^*}(V_i \mid \mathrm{pa}_i^*) dP(\Pi^*) \,. \end{aligned}$$

In other words, variables V are assumed to be *conditionally exchangeable*, so that the same graph $\mathcal{G}^*$ applies in every context $c \in \mathcal{C}$ (Guo et al., 2022). Importantly, the distribution $P(V)$ does not depend on $P(C)$, but only on the mechanism generating the Π_i^*.

For an overview of our problem setting, we refer to Fig. 1 showing a fixed causal DAG in all contexts (G), as well as interventions which we represent as changes in the functional cause-effect dependencies (arrows). Each variable is associated with a partition Π_i^* showing such changes (colored boxes). We next move to properties of causal mechanism shifts that will be relevant for confounder identification.

2.3 Independent Mechanism Shifts

A common principle in causal discovery is the independence of causal mechanisms (Janzing and Schölkopf, 2010). It states that the distribution $P^c(V_i \mid \mathrm{pa}_i^*)$ is uninformative of $P^{c'}(V_j \mid pa_j)$ when either $i \neq j$ or $c \neq c'$. A change in the mechanism of V_i therefore provides no information about changes in the mechanism of V_j, and partitions Π_i^* and Π_j^* are independent.

Assumption 2 (Independent Mechanism Shifts) *We assume that mechanisms $P^c(V_i \mid \mathrm{pa}_i^*)$ change independently but identically distributed across environments. That is, we assume that*

$$P(\Pi^*) = \prod_{V_i} P(\Pi_i^*) \,.$$

Mere independence of mechanism shifts is not a significant constraint. The mechanisms of V_i and V_j both differing across all (or no) environments would trivially satisfy this condition, but reveal no information about the core causal mechanisms. We therefore additionally assume that mechanism shifts are sparse, so that mechanisms remain the same across most environments (Guo et al., 2022; Schölkopf et al., 2021).

Assumption 3 (Sparse Mechanism Shifts) *Let C and C' be two i.i.d. samples from the same distribution $P(C)$. We assume that the probability of mechanism changes between two contexts is*

$$p = P(\Pi^*(C) \neq \Pi^*(C')) < 0.5 \,.$$

With this we assume that mechanism shifts occur infrequently, implying that causal functions persist across the majority of environments. This assumption is valid in many study settings where specific targets are interventions in a small number of contexts, and has been adopted in the causal discovery literature (Perry et al., 2022; Mameche et al., 2023).

Conversely, we assume that two contexts c, c' are assigned to different sets of the partition Π^*, then the corresponding causal mechanisms indeed change.

Assumption 4 (Π-faithfulness) *Let Π_i^* be the partition of V_i. Then for any two environments c, c',*

$$\Pi^*(c) \neq \Pi^*(c') \longrightarrow P^c(V_i \mid \mathrm{pa}_i^*) \neq P^{c'}(V_i \mid \mathrm{pa}_i^*) \,.$$

This faithfulness condition ensures that our partitions precisely capture the changes in causal functions. Next, we show how these assumptions, which we assume to hold when variables in V are measured, are violated when some latent factors Z are not observed.

3 IDENTIFYING CONFOUNDING FROM MECHANISM SHIFTS

We begin with an analysis of the effect of latent confounding on the partitions of the causal mechanisms, then propose a score for determining whether a given pair of variables is confounded, and conclude by giving consistency guarantees for our score. We include proofs of all results in the supplementary material.

3.1 Confounding Introduces Dependent Mechanism Shifts

All assumptions we made in the previous section address the *true* partitions Π_i^* of the *true* causal mechanisms over the *true* causal parents pa_i^*. When not all variables are observed, the situation changes. To see this, let us consider the following linear example,

$$\begin{aligned} Z &\sim N(0, \sigma_z^2(c)) \\ X &= \alpha Z + \epsilon_x \\ Y &= \beta X + \gamma Z + \epsilon_y \,, \end{aligned}$$

where the only source of mechanism shifts is the non-constant variance $\sigma_z^2(c)$ of the unobserved Z.

Then, regressing the variable Y on X, we obtain

$$X \sim N(0, \sigma_x^2 + \alpha^2 \sigma_z^2(c))$$
$$\widehat{\beta}_{Y|X} = \frac{\text{cov}(X,Y)}{\text{var}(X)} = \frac{\beta\sigma_x^2 + \alpha\gamma\sigma_z^2(c)}{(\sigma_x^2 + \alpha^2\sigma_z^2(c))}, \qquad (1)$$

so that in general clearly both the distributions $P^c(X)$ and $P^c(Y \mid X)$ change as $\sigma_z^2(c)$ changes.

Note however that in exceptional circumstances the above does not apply. If the parameters are chosen as $\beta = 1$ and $\alpha = \gamma$ in Eq. (1), then $\widehat{\beta}_{Y|X}$ will not change despite a change in σ_z^2. This kind of fine-tuning of the parameters likely happens only in adversarial cases. When we assume that parameters are sampled from a continuous probability distribution, the probability of obtaining a set of parameters where a change in the mechanism of the confounder Z does not translate into a change in the mechanisms affecting X and Y is zero. We therefore make the following general assumption.

Assumption 5 (Shift Faithfulness) *Let Z be an unobserved common parent of all variables in the subset $X_S \subset X$. Then each mechanism change in Z between two contexts c, c' entails a mechanism change for each $X_i \in X_S$ between the same contexts c, c'.*

Note that we do not strictly need *all* mechanism shifts of Z to be reflected in X, Y, but only that some (non-zero) fraction is captured. However, for ease of exposition we make the above assumption. Hence, changes in the causal mechanism of Z lead to correlations between the observed partitions Π_i of variables affected by Z. We therefore now turn to the question of how to measure these correlations.

3.2 Measuring Dependence of Mechanism Shifts via Mutual Information

To measure whether the mechanism changes of variables are dependent, we consider the Mutual Information (MI) between partitions.

For two partitions Π_1, Π_2 of the set of contexts $\mathcal{C}$ into I, J sets, we consider the contingency table $\mathcal{M}$,

$$\mathcal{M} = \begin{pmatrix} n_{11} & n_{12} & \dots & n_{1J} \\ n_{21} & n_{22} & \dots & n_{2J} \\ \vdots & \vdots & \ddots & \vdots \\ n_{I1} & n_{I2} & \dots & n_{IJ} \end{pmatrix}$$

where n_{ij} measures the number of contexts in $\pi_1^i \cap \pi_2^j$, and with row margins $u_i = |\pi_1^i|$ and column margins $v_j = |\pi_2^j|$ counting the size of partition elements.

If the partitions describe causal mechanism shifts of two different variables X_i, X_j, then a latent confounder affecting both X_i, X_j leads to correlations between these partitions. To measure these, we consider the mutual information between Π_1 and Π_2. The marginal entropy of Π_1 and joint entropy of Π_1, Π_2 are

$$H(\Pi_1) = -\sum_i \frac{u_i}{N} \log \frac{u_i}{N},$$
$$H(\Pi_1, \Pi_2) = -\sum_{ij} \frac{n_{ij}}{N^2} \log \frac{n_{ij}}{N^2},$$

with $H(\Pi_2)$ similar, and their mutual information is

$$\begin{aligned} \text{I}(\Pi_1, \Pi_2) &= H(\Pi_1) + H(\Pi_2) - H(\Pi_1, \Pi_2) \\ &= \sum_{ij} \frac{n_{ij}}{N} \log \frac{n_{ij} N}{u_i v_j}. \end{aligned}$$

In general, given data from only a finite number of contexts, this plug-in estimate of the MI between partitions will be positively biased (Vinh et al., 2009). We can correct for this by standardizing our score using the expectation for two independent partitions.

Expected MI under Independent Shifts We first consider two independent partitions Π_1', Π_2' with contingency table $\mathcal{M}$ with column sums u and row sums v. To define their mutual information, the hypergeometric model of randomness has been adopted in the literature (Vinh et al., 2009, 2010). That is, given the marginal counts u, v, the joint counts are assumed to follow a hypergeometric distribution $N_{ij} \sim \mathcal{H}(u, v, N)$, with probability mass function

$$\mathcal{P}(n_{ij} \mid u, v, N) = \frac{\binom{n_{ij}}{v_j}\binom{N - v_j}{u_i - n_{ij}}}{\binom{N}{u_i}}.$$

The expected mutual information between the independent partitions is then computed as

$$\begin{aligned} \mathbb{E}\left[\text{I}(\Pi_1', \Pi_2')\right] &= \sum_{\mathcal{M}} \text{I}(\mathcal{M}) \mathcal{P}(\mathcal{M}) \\ &= \sum_{ij} \sum_{n_{ij}} \text{I}(n_{ij}) \mathcal{P}(n_{ij} \mid u, v, N) \qquad (2) \end{aligned}$$

where $\text{I}(n_{ij}) = \frac{n_{ij}}{N} \log \frac{n_{ij} N}{u_i v_j}$ and the inner sum runs over $n_{ij} \in [\max\{0, u_i + v_j - N\}, \min\{u_i, v_j\}]$. By replacing the term $\text{I}(n_{ij})$ by $\text{I}(n_{ij})^2$, one can similarly compute the second moment, and thus the variance

$$\text{Var}(\text{I}(\Pi_1', \Pi_2')) = \mathbb{E}\left[\text{I}(\Pi_1', \Pi_2')^2\right] - \mathbb{E}\left[\text{I}(\Pi_1', \Pi_2')\right]^2.$$

With this, we can compute the standardized score of our observed mutual information $I(\Pi_1, \Pi_2)$

$$t = \frac{I(\Pi_1, \Pi_2) - \mathbb{E}\left[\text{I}(\Pi_1', \Pi_2')\right]}{\sqrt{\text{Var}(\text{I}(\Pi_1', \Pi_2'))}},$$

whose properties we study in the next section.

3.3 Identifying Confounded Variable Pairs

In the following, we make sure that the MI-based score is sensible in our setting.

We begin by considering the bivariate case and show that when the causal direction between a pair X, Y is known, we indeed obtain the correct results in the limit of large numbers of contexts.

Lemma 3.1 (Significance and Power) Let X, Y be unconfounded and $X \to Y$. Let Π_X, Π_Y be the corresponding partitions.

$$\lim_{n_c \to \infty} \mathcal{P}(t > q_{1-\alpha}) \to \alpha \ ,$$

where $q_{1-\alpha}$ is the $1-\alpha$-quantile of standard normal distribution. Conversely, if X, Y are confounded by Z, then for any $\alpha > 0$ we obtain a power of 1 in the limit,

$$\lim_{n_c \to \infty} \mathcal{P}(t > q_{1-\alpha}) \to 1 \ .$$

This result tells us that with data from enough environments, we are guaranteed to discover which pairs of variables are confounded. Next, we move on to the problem of recovering sets of jointly confounded nodes.

3.4 Beyond Confounded Pairs

To determine whether a set of variables shares a joint confounder, we extend our score beyond pairs of variables. A natural extension of mutual information for a set of partitions is total correlation (Watanabe, 1960),

$$\begin{aligned} T(\Pi_1, \dots, \Pi_s) &= \sum_i H(\Pi_i) - H(\Pi_1, \dots, \Pi_s) \\ &= \sum_i I(\Pi_i, \Pi_{>i} \mid \Pi_{<i}) \end{aligned}$$

where $\Pi_{<i} = \{\Pi_1, \dots, \Pi_{i-1}\}$ and similarly for $\Pi_{>i}$. It is straightforward to correct this score analog as done above for pair-wise MI terms. As both corrected and uncorrected scores are asymptotically equivalent, we will consider T as is in our theoretical analysis.

First we discuss how to use this score for detecting joint confounding. To this end, consider three variables X_1, X_2, X_3. By Assumption 5 and Lemma 3.1, we know these can only be jointly confounded iff all X_i, X_j are pair-wise confounded. It could of course be that rather than jointly confounded, there are three disjoint confounders Z_{12}, Z_{13}, Z_{23} affecting each of the individual pairs. We can distinguish these two cases? Only if all three variables share the same latent confounder Z, knowing about the partition of one variable explains away some of the correlation between the other two partitions, so that we have $I(\Pi_i, \Pi_j \mid \Pi_k) < I(\Pi_i, \Pi_j)$ for any permutation of the variables.

Note that in general, for a set of size s to permit such an equivalent explanation in the first place, we would need to add a total of $\binom{s}{2}$ confounders with $s(s-1)$ outgoing edges to obtain the same structure of pairwise confounding. While this may *plausibly* occur for small sets of variables that appear to be pair-wise correlated, we assume the true graph $\mathcal{G}^*$ to be causally minimal in the following sense.

Assumption 6 (Confounder Minimality) *For every subset X_S of at least $|S| \geq 4$ variables, there are at most $2\,|S|$ edges incoming into X_S from latent confounders Z_j with at least three children in X_S.*

This minimality assumption ensures that variables that appear to be jointly confounded are indeed confounded; put differently, when a small number of latent variables suffice to explain the observed correlations, there should indeed exist only few confounders. With this, we can guarantee that the identification of joint confounding is possible from the total correlation T.

Theorem 3.1 Let X_S be a set of variables such that all $X_i, X_j \in X_S$ are pair-wise confounded. Then X_S is jointly confounded if and only if for each triple $X_i, X_j, X_k \in S$ we have

$$\begin{aligned} \lim_{n_c \to \infty} \mathcal{P}(T(\Pi_i, \Pi_j, \Pi_k) < I(\Pi_i, \Pi_j) + I(\Pi_j., \Pi_k)) \\ = \begin{cases} 1, & X_1, X_j, X_k \text{ jointly confounded} \\ 0, & \text{otherwise} \, . \end{cases} \end{aligned}$$

With this, we can recover how many latent confounders Z_j there are, and sets of jointly confounded nodes by each Z_j are uniquely identifiable by our score. Due to both the large number of tests that the above objective involves, as well as issues due to false negatives for the pair-wise test, we propose a more efficient and robust method in Section 4.

As we assumed causal directions among all variables to be known up to this point, the remaining question is what happens if this is not the case.

3.5 (Spurious) Spurious Correlations

We now address the case where the true causal structure is unknown and we estimate partitions in the presence of misdirected edges. We want to ensure we can use the MI-based score consistently.

First, let us return to the case of two variables X, Y such that in the true graph $\mathcal{G}^*$, the causal direction $X \to Y$ applies. What would happen if we instead considered the partitions obtained by considering the graph $\mathcal{G}$ differing from $\mathcal{G}^*$ by inverting this edge to $Y \to X$? To compare the resulting partitions, we write Π_X, Π_Y for the partitions of causal mechanisms in $\mathcal{G}$, and similarly Π^* for $\mathcal{G}^*$ and Π' for a graph $\mathcal{G}'$.

It turns out that, with high probability, the misdirected edge will introduce additional correlations between the inferred partitions Π_X, Π_Y. Intuitively, this is because distribution shifts in $P^c(Y)$ now need to come with matching mechanism shifts of $P^c(X \mid Y)$ to ensure that $P^c(X)$ does *not* change (Huang et al., 2020). This leads to the following result.

Proposition 3.1 (Consistency for Pairs of Variables) If a variable pair X, Y is confounded by a variable Z, then there exists some $\rho > 0$ such that

$$\mathcal{P}(\mathrm{I}(\Pi_X^*, \Pi_Y^*) < \mathrm{I}(\Pi_X, \Pi_Y)) = 1 - O(e^{-\rho n_c}) .$$

When X, Y are part of a larger graph, the situation becomes more involved. Based purely on dependencies between the observed variables X, we can at best recover the Markov equivalence class (MEC, Pearl (2009a)). However, due to the effects of latent confounders, the MEC over X will contain large numbers of spurious edges (Elidan et al., 2000; Kaltenpoth and Vreeken, 2023a). We therefore show that so long as the number of latent confounders affecting and spurious siblings of a given target X_i are not too large, then we can still recover the correct parents of the target.

Proposition 3.2 (Recovering Parents) Let X_i be a target variable and let $\mathcal{G}$ and $\mathcal{G}'$ be two graphs in the MEC of the marginal distribution $P^c(X)$. Assume that only one of the two graphs correctly recovers the parents of X_i, $\mathrm{pa}_i = \mathrm{pa}_i^*$ and $\mathrm{pa}_i' \neq \mathrm{pa}_i^*$, and further assume that the number of latent confounders affecting X_i plus spurious siblings is bounded by $\frac{\log(0.5)}{\log(1-p)}$. Then

$$\begin{aligned}\mathcal{P}(\mathrm{I}(\Pi_i, \{\Pi_j : j \in \mathrm{pa}_i\}) < \mathrm{I}(\Pi_i', \{\Pi_j' : j \in \mathrm{pa}_i'\})) \\ = 1 - O(e^{-\rho n_c})\end{aligned}$$

Summing up over all variables gives us the following consistency of the entire causal ordering over X.

Theorem 3.2 (Consistency) Let $\mathcal{G}^*$ be the true graph over V and let $\mathcal{G}_x^*$ be the induced graph on X, and assume that for all X_i the number of latent confounders plus spurious siblings is bounded by $\frac{\log(0.5)}{\log(1-p)}$. Then with high probability, $\mathcal{G}_x^*$ and its partitions $\Pi_1^*, \ldots, \Pi_k^*$ are the unique minimum of total correlation,

$$\mathcal{P}(\arg\min_{\mathcal{G}} T(\Pi_1, \ldots, \Pi_m) = \{\mathcal{G}_x^*\}) = 1 - O(e^{-\rho n_c}) .$$

With these theoretical guarantees in hand, we now move on to provide an effective algorithm for discovering which variables are indeed confounded.

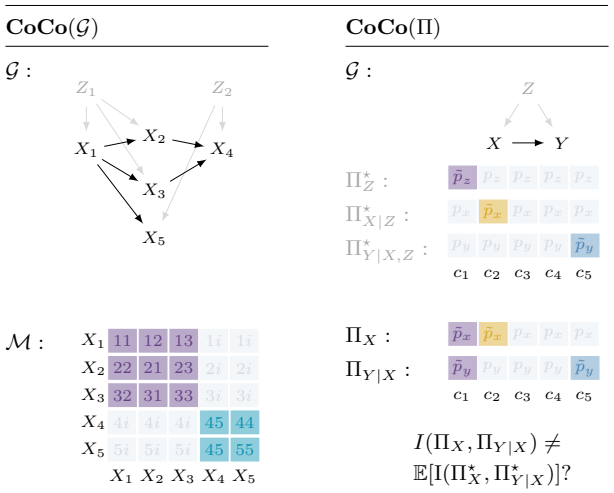


Figure 2: *Components of* CoCo. In a DAG $\mathcal{G}$ with unobserved confounders Z_1, Z_2 (top left), we consider each pair of nodes (top right) discover their partitions, and test them for dependency using MI (bottom right). We obtain an affinity matrix showing which nodes are affected by the same confounder (bottom left).

4 DISCOVERING CONFOUNDERS IN DIFFERENT CONTEXTS

Based on our theory, we develop the CoCo algorithm for discovering **Co**nfounders in different **Co**ntexts.

Determining Causal Mechanism Shifts There exist approaches to discovering causal mechanisms and their changes in multiple contexts. Since it agrees well with our shift testing approach, we build upon the MSS estimator developed by Perry et al. (2022) which starts from the correct MEC and directs edges in a way that induces the fewest conditional shifts.

For each causal mechanism of a target variable X_i and each pair of environments, we perform a conditional independence test to detect mechanism changes, resulting in the following p-values,

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

We hereby use the Kernel Conditional Independence test (KCI, Zhang et al. (2011)) for all practical purposes, but other instantiations are possible (Park et al., 2021). In case a variable has no parents in $\mathcal{G}$, the above reduces to testing the marginal distributions $P(X_i)$ for equality for which we use the Maximum Mean Discrepancy (MMD, Gretton et al. (2012)).

As the pair-wise p-values between pairs of contexts are correlated and hence do not allow a well-defined measure of dependency, we convert them to a partition to use our MI-based measure. We obtain a clustering

naively from the pair-wise tests by including c_i, c_j in the same group iff the pair-wise testing does not indicate that $\Pi(c_i) \neq \Pi(c_j)$. Hence, if there are disagreements between the correlated tests, we resolve these in favor of more mechanism changes, although other options are possible depending on the sensitivity of the test. In the bivariate example shown in Fig. 1, for instance, we obtain partitions Π_X and $\Pi_{Y|X}$ corresponding to the shown changes in $\widehat{P}_X$ and $\widehat{P}_{Y|X}$, which we test for independence as described in the following.

Discovering Confounding Variables Next, for every pair of variables X_i and X_j, we determine whether it is confounded by using a one-tailed Z-test, resulting in the p-values

$$p_{ij} = \Phi^{-1}\left(\frac{I(\Pi_i, \Pi_j) - \mathbb{E}\left[\mathrm{I}(\Pi_i^\star, \Pi_j^\star)\right]}{\sqrt{\mathrm{Var}(\mathrm{I}(\Pi_1^\star, \Pi_2^\star))}}\right) ,$$

where Φ is the cumulative density function of the standard normal distribution.

In the second stage, we aim to discover those subsets of variables that are affected by the same latent variable. While our theoretical analysis suggests considering the total correlation over variable subsets, performing such a test for every given subset $X_S \subseteq X$ is both infeasible and faces us with a multiple testing problem involving enormous numbers of tests. We therefore infer confounders directly from pairwise tests.

If our tests for discovering causal mechanism shifts and confounding were perfect, variables subject to the same confounder would form distinct clusters with high pairwise MI, and these clusters could be used as direct estimates of confounded variable sets. In practice, however, we will find some variable pairs to incorrectly be judged (un)confounded. To mitigate this issue, we perform a clustering on the pairwise MI terms.

More precisely, we consider the affinity matrix $\mathcal{M}$ with entries $\mathcal{M}_{ij} = I(\Pi_i, \Pi_j)$, using MI as pairwise similarity, and use spectral clustering (Donath and Hoffman, 1972) to discover strongly connected components in this matrix. As a result, we obtain multiple subsets X_{S_j} that are likely subject to the same confounder.

CoCo To summarize, we present the pseudocode for CoCo as Alg. 1, and an illustration in Fig. 2. In the first phase, we test all pairs of contexts for mechanism shifts (l. 1–4), and repeat this for each variable to obtain its partition. In the second phase, we test all pairs of variables for confounding (l. 5–6). Last, we cluster the variables into subsets that are affected by the same confounder (l. 7–8).

Regarding the complexity of our method, shift testing is in $\mathcal{O}(|\mathcal{G}|\cdot|\mathcal{C}|^2)$, testing for confoundedness in $\mathcal{O}(|\mathcal{G}|^2)$, plus spectral clustering in $\mathcal{O}(|\mathcal{G}|^3)$.

Algorithm 1: CoCo($\mathcal{G}$)

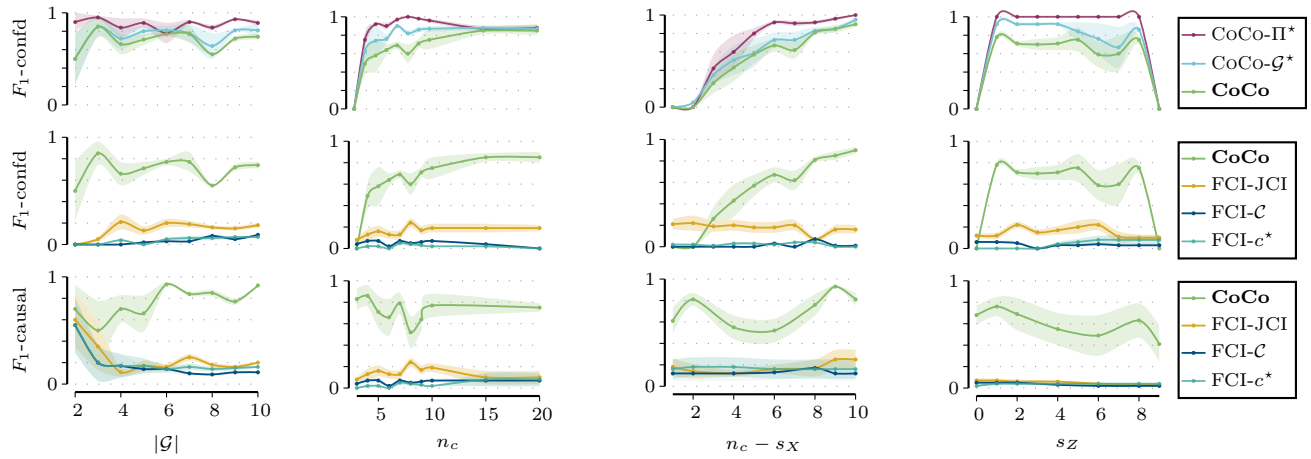
input : Data over X, C; causal DAG $\mathcal{G}$.
output: Subsets of X that are jointly confounded by a latent variable Z_j.

1 **foreach** variable X_i **do**
2 **foreach** pair of contexts c, c' **do**
3 $p_{c,c'} = p\text{-val}\left(P^c(X_i \mid \mathrm{pa}_i) \neq P^{c'}(X_i \mid \mathrm{pa}_i)\right)$.
4 Convert $\{p_{c,c'}\}$ to a partition Π_i
5 **foreach** pair of variables X_i, X_j **do**
6 $p_{ij} = p\text{-val}(I(\Pi_i, \Pi_j) \neq \mathbb{E}\left[\mathrm{I}(\Pi_i^\star, \Pi_j^\star)\right])$.
7 Construct an affinity matrix $\mathcal{M}$
8 Discover subsets X_S of X that are connected components in $\mathcal{M}$, using spectral clustering
9 **return** Subsets X_S

5 RELATED WORK

As arguably one of the most important problems in statistical inference, causal inference has attracted a lot of recent research attention (Rubin, 1974; Spirtes et al., 2000; Pearl, 2009a). Unfortunately, the existence of confounders, selection bias and other statistical problems make it impossible to infer causality from observational data without further assumptions (Pearl, 2009b). Given purely observational data, classical constraint-based (Spirtes et al., 2000, 1999; Zhang, 2008) and score-based (Chickering, 2002; Scanagatta et al., 2015; Ramsey et al., 2017) reconstruct causal graphs up to Markov equivalence, assuming sufficiency as well as i.i.d. data.

When causal sufficiency does not hold, a number of algorithms such as the FCI family (Spirtes et al., 2000; Colombo et al., 2012; Ogarrio et al., 2016), and convex optimization-based approaches (Chandrasekaran et al., 2010) can find confounding to a limited extent. Specifically, Nested Markov Models (NMMs) (Shpitser et al., 2014, 2018; Richardson et al., 2017; Evans and Richardson, 2019) allow identifiability of causal models with latent factors by using (pair-wise) Verna constraints. The recent approach DCD by Bhattacharya et al. (2021) combines NMMs with the differentiable constraint by Zheng et al. (2018) to discover a partially directed causal network and likely confounded nodes. In contrast, Kaltenpoth and Vreeken (2023a,b) explicitly model the latent confounders Z by exploiting patterns of the observed causal graph structure and violations of causal mechanism independence. Reddy et al. (2022) propose a related but different mutual



(a) DAG size (b) Contexts (c) Invariant Contexts (d) Latent Shifts

Figure 3: *Detecting Confounding with* CoCo. We evaluate CoCo on discovering confounding in DAGs $\mathcal{G}$ over multiple contexts. We compare (top) CoCo with MSS and the KCI test (green) to oracle versions that start from the true partitions $\Pi^\star$ (purple) resp. the fully directed DAG $\mathcal{G}^\star$ (blue). We compare (middle, bottom) against JCI-FCI (yellow), FCI ($\mathcal{C}$) on pooled data, and FCI ($c^\star$) per context (blue). We report F1 scores computed over each pair of nodes, evaluating whether it is confounded (top, middle), resp. causal (bottom).

information estimator, which applies directly to the distributions of observed variables as opposed to partitions induced by causal mechanism shifts, as well as assumes that no direct causal edges are present. Karlsson and Krijthe (2023) also address violations in exchangeability under latent confounding, but focus on causal effect estimation under a fixed graph structure.

There also exists a growing literature on relaxing the i.i.d. assumption in causal discovery, showing one can obtain stronger identifiability results of the underlying causal graph under distribution shifts (Huang et al., 2020; Mooij et al., 2020). Recent approaches leverage the independent changes (Mameche et al., 2023) and sparse shift principles to discover fully directed causal DAGs from multiple environments, such as the Mechanism Shift Score (MSS, Perry et al. (2022)).

The aforementioned approaches consider an exogenous context variable which can be viewed as a special form of confounding (Huang et al., 2020). However, in practice, not all confounding can be fully explained by the effects of the environment. For example, when confounding effects are genetic, then while differences in the values of the confounder can be partially explained by membership of a subpopulation, the variance within any subpopulation is still large, that is, there may still be a confounder within each context. Most related to our method is the Joint Causal Inference (JCI) framework (Mooij et al., 2020) when instantiated with a discovery algorithm that does not require sufficiency, such as FCI (Spirtes et al., 2000).

6 EXPERIMENTS

To conclude, we evaluate CoCo empirically on both synthetic and real-world data.

CoCo and Oracles To separate the effects of discovering latent variables, mechanism changes, and causal directions, we include different oracle versions of CoCo. To study our confounding test in isolation, we consider an oracle for the true partitions, named CoCo-$\Pi^\star$. We combine it with mechanism shift testing in CoCo-$\mathcal{G}^*$, which takes as background knowledge the causal structure $\mathcal{G}^*$. Finally, we combine our approach with MSS (Perry et al., 2022) using the kernelized conditional independence test (Zhang et al., 2011) to discover a fully directed DAG $\mathcal{G}$. As MSS starts from a Markov Equivalence class, we provide all methods with the correct MEC as a starting point.

Baselines Our main competitor is JCI (Mooij et al., 2020) instantiated with the FCI algorithm (Spirtes et al., 2000), referred to as JCI-FCI. It applies FCI to an augmented causal model including the context variable and appropriate edge constraints (Mooij et al., 2020), and returns for each variable pair whether it is causal, confounded, potentially confounded, or none of the above. We also apply FCI to the pooled data from all contexts, FCI($\mathcal{C}$), and to the data of each context individually, reporting the best such result, FCI($c^\star$).

Synthetic Data Following Huang et al. (2020), we generate data from an Erdös-Rényi model as follows,

$$X_i^{(c)} = \sum_{i \in \mathrm{pa}_i^*} \omega_{ij}^{(c)} f_{ij}(X_i^{(c)}) + \sigma_j^{(c)} N_j^{(c)} \ , \qquad (3)$$

with weights $\omega_{ij}^{(c)} \sim \mathcal{U}(0.5, 2.5)$, either uniform or Gaussian noise with equal probability, and functions f sampled uniformly at random from $\{x^2, x^3, \tanh, \mathrm{sinc}\}$. For each mechanism change, we re-sample from Eq. 3. Finally, each confounder Z_j is a source node that has edges to a random subset of at least two variables.

6.1 Detecting Confounding with CoCo

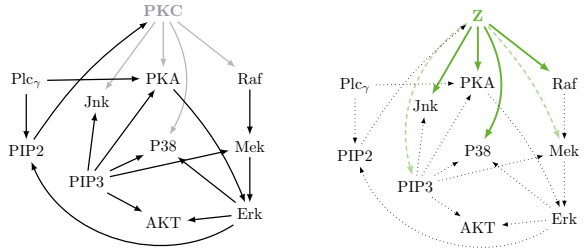
In our main experiment, we evaluate whether the methods discover confounding in a multi-context DAG $\mathcal{G}$. As the FCI variants can only determine potential confounding for node pairs, we evaluate confounding decisions for each pair of nodes in $\mathcal{G}$, using F1-scores. We show the results depending on different parameters, including the number of contexts (n_c), number of observed (n_X) and latent (n_Z) variables, and the number of observed (s_X) and latent (s_Z) mechanism shifts. We start from the parameters ($n_c = 10, n_X = 10, n_Z = 1, s_X = 1, s_Z = 2$).

We show our results in Fig. 3. As we expect from our theoretical analysis, the results for CoCo improve with more contexts (Fig. 3b) and especially with the number of *invariant* contexts, since these allow us to detect joint shifts caused by Z. Conversely, when mechanism shifts are *dense* ($n_c - s_X = 1$) or ($s_Z = n_c - 1$) or there are *no shifts*, we cannot discover any dependencies as expected. Aside from these cases, CoCo (green) clearly outperforms JCI-FCI (yellow) by a large margin in discovering confounders, while JCI-FCI in turn has a slight advantage over FCI on best-scoring single-context resp. pooled data (blue).

The gap between the oracle versions and full CoCo remains small in practice, suggesting that confounding detection scales to unknown causal directions and supporting our results in Section 3.5. To conclude, we also show how many causal edges are correctly directed in Fig. 3 (bottom). As expected for MSS, we do well under sparse shifts and with more contexts while the FCI variants generally only discover few causal edges.

6.2 Real-world Cell Signalling Data

We end with a case study on the flow cytometry dataset by Sachs et al. (2005). It contains samples of eleven protein and phospholipid components in human immune cells that were studied under different molecular interventions. To study confounding effects, we start from the consensus causal network in Fig. 4. Following the design of Kaltenpoth and Vreeken (2023b), we keep PKC hidden and use the data over the remaining variables in nine contexts. As we illustrate in Fig. 4, CoCo correctly discovers a confounder Z and all of its outgoing edges (green) as well as two spurious ones (dashed). To ensure that CoCo does not return spurious confounding, we repeat the experiment while keeping each other node in turn hidden. Notably, we always discover Raf and Mek to be confounded, suggesting the possibility of unmeasured confounding in this highly controlled study. Other than that, however, CoCo returns only one more false positive edge, and correctly rejects confounding in all other cases. Running JCI-FCI on the same data, we discover multiple false positive confounded edges. We show the discovered network in the supplement.



(a) Consensus Network

(b) Confounder that CoCo discovers

Figure 4: *Confounding in Cell Signalling data.* On the Sachs et al. (2005) data CoCo recovers confounding effects of PKC. Solid green edges are correctly discovered as confounded, dashed edges are spurious.

7 DISCUSSION & CONCLUSION

Real-world data is often heterogeneous in nature and may be subject to unmeasured confounding effects, challenging the common assumptions of causal sufficiency and i.i.d.-ness. We study the intersection of both assumptions, and show how to relax them under the principle of independent causal mechanisms. Our main insight is that latent variables introduce dependencies in observed causal mechanism changes. We show how to measure such a dependence using mutual information, give identifiability results for confounding with known and unknown causal structure, and evaluate our approach in practice. While our approach purely relies on detecting causal mechanism shifts and hence is completely nonparametric, this comes with the limitation that we need multiple context distributions with a measurable shift of the latent variable. Hence, combining CoCo with existing approaches for the i.i.d. case is a promising direction for future work.

Acknowledgements

We thank the anonymous reviewers for their helpful feedback.

Bibliography

Bhattacharya, R., Nagarajan, T., Malinsky, D., and Shpitser, I. (2021). Differentiable causal discovery under unmeasured confounding. In *International Conference on Artificial Intelligence and Statistics*, pages 2314–2322. PMLR.

Chandrasekaran, V., Parrilo, P. A., and Willsky, A. S. (2010). Latent variable graphical model selection via convex optimization. In *2010 48th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 1610–1613. IEEE.

Chickering, D. M. (2002). Learning equivalence classes of bayesian-network structures. *Journal of Machine Learning Research*, 2:445–498.

Colombo, D., Maathuis, M. H., Kalisch, M., and Richardson, T. S. (2012). Learning high-dimensional directed acyclic graphs with latent and selection variables. *The Annals of Statistics*, 40(1):294–321.

Dominguez, A. A., Lim, W. A., and Qi, L. S. (2016). Beyond editing: repurposing crispr–cas9 for precision genome regulation and interrogation. *Nature Reviews Molecular Cell Biology*, 17(1):5–15.

Donath, W. E. and Hoffman, A. J. (1972). Algorithms for partitioning of graphs and computer logic based on eigenvectors of connection matrices. *IBM Technical Disclosure Bulletin*, 15(3):938–944.

Elidan, G., Lotner, N., Friedman, N., and Koller, D. (2000). Discovering hidden variables: A structure-based approach. *Advances in Neural Information Processing Systems*, 13.

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

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. (2012). A Kernel Two-Sample Test. *Journal of Machine Learning Research*, 13(25):723–773.

Guo, S., Tóth, V., Schölkopf, B., and Huszár, F. (2022). Causal de finetti: On the identification of invariant causal structure in exchangeable data. *arXiv*.

Huang, B., Zhang, K., Zhang, J., Ramsey, J., Sanchez-Romero, R., Glymour, C., and Schölkopf, B. (2020). Causal discovery from heterogeneous/nonstationary data. *Journal of Machine Learning Research*, 21(1).

Janzing, D. and Schölkopf, B. (2010). Causal inference using the algorithmic markov condition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 56:5168–5194.

Kaltenpoth, D. and Vreeken, J. (2023a). Causal Discovery with Hidden Confounders using the Algorithmic Markov Condition. In *Uncertainty in Artificial Intelligence*, pages 1016–1026. PMLR.

Kaltenpoth, D. and Vreeken, J. (2023b). Nonlinear Causal Discovery with Latent Confounders. In *International Conference on Machine Learning*, pages 15639–15654. PMLR.

Karlsson, R. and Krijthe, J. (2023). Detecting hidden confounding in observational data using multiple environments. In Oh, A., Neumann, T., Globerson, A., Saenko, K., Hardt, M., and Levine, S., editors, *Advances in Neural Information Processing Systems*, volume 36, pages 44280–44309. Curran Associates, Inc.

Mameche, S., Kaltenpoth, D., and Vreeken, J. (2023). Learning causal mechanisms under independent changes.

Mooij, J., Magliacane, S., and Claassen, T. (2020). Joint causal inference from multiple contexts. *Journal of Machine Learning Research*, 21:99:1–99:108.

Ogarrio, J. M., Spirtes, P., and Ramsey, J. (2016). A hybrid causal search algorithm for latent variable models. In *Proceedings of the Eighth International Conference on Probabilistic Graphical Models*, pages 368–379.

Park, J., Shalit, U., Schölkopf, B., and Muandet, K. (2021). Conditional distributional treatment effect with kernel conditional mean embeddings and u-statistic regression. In *Proceedings of 38th International Conference on Machine Learning (ICML)*, volume 139 of *Proceedings of Machine Learning Research*, pages 8401–8412. PMLR.

Pearl, J. (2009a). *Causality: Models, Reasoning and Inference*. Cambridge University Press, 2nd edition.

Pearl, J. (2009b). *Causality: Models, Reasoning and Inference*. Cambridge University Press.

Perry, R., Kügelgen, J. V., and Schölkopf, B. (2022). Causal discovery in heterogeneous environments under the sparse mechanism shift hypothesis.

Peters, J., Bühlmann, P., and Meinshausen, N. (2016). Causal inference using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(5):947–1012. (with discussion).

Peters, J., Janzing, D., and Schlkopf, B. (2017). *Elements of Causal Inference: Foundations and Learning Algorithms*. The MIT Press.

Ramsey, J., Glymour, M., Sanchez-Romero, R., and Glymour, C. (2017). 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.

Reddy, A. G., Dash, S., Sharma, A., and Balasubramanian, V. N. (2022). Counterfactual generation under confounding. In *NeurIPS 2022 Workshop on Causality for Real-world Impact*.

Richardson, T. S., Evans, R. J., Robins, J. M., and Shpitser, I. (2017). Nested markov properties for acyclic directed mixed graphs. *arXiv preprint arXiv:1701.06686*.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. 66:688–701.

Sachs, K., Perez, O., Pe'er, D., Lauffenburger, D., and Nolan, G. (2005). Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, pages 523–9.

Scanagatta, M., de Campos, C. P., Corani, G., and Zaffalon, M. (2015). Learning bayesian networks with thousands of variables. In *NIPS*, pages 1864–1872.

Schölkopf, B., Janzing, D., Peters, J., Sgouritsa, E., Zhang, K., and Mooij, J. (2012). On causal and anticausal learning. ICML'12, pages 459–466, Madison, WI, USA. Omnipress.

Schölkopf, B., Locatello, F., Bauer, S., Ke, N. R., Kalchbrenner, N., Goyal, A., and Bengio, Y. (2021). Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634.

Shpitser, I., Evans, R. J., and Richardson, T. S. (2018). Acyclic linear sems obey the nested markov property. In *Uncertainty in artificial intelligence: proceedings of the... conference. Conference on Uncertainty in Artificial Intelligence*, volume 2018. NIH Public Access.

Shpitser, I., Evans, R. J., Richardson, T. S., and Robins, J. M. (2014). Introduction to nested markov models. *Behaviormetrika*, 41(1):3–39.

Sinha, M., Tadepalli, P., and Ramsey, S. A. (2021). Voting-based integration algorithm improves causal network learning from interventional and observational data: An application to cell signaling network inference. *PLoS One*, 16(2):e0245776.

Spirtes, P. (2001). An Anytime Algorithm for Causal Inference. In *International Workshop on Artificial Intelligence and Statistics*, pages 278–285. PMLR.

Spirtes, P., Glymour, C. N., and Scheines, R. (2000). *Causation, prediction, and search*. MIT press.

Spirtes, P., Meek, C., and Richardson, T. (1999). An algorithm for causal inference in the presence of latent variables and selection bias in computation, causation and discovery. MIT Press.

Vinh, N. X., Epps, J., and Bailey, J. (2009). Information theoretic measures for clusterings comparison: is a correction for chance necessary? In *Proceedings of the 26th annual international conference on machine learning*, pages 1073–1080.

Vinh, N. X., Epps, J., and Bailey, J. (2010). Information Theoretic Measures for Clusterings Comparison: Variants, Properties, Normalization and Correction for Chance. *Journal of Machine Learning Research*, 11(95):2837–2854.

Watanabe, S. (1960). Information theoretical analysis of multivariate correlation. *IBM Journal of Research and Development*, 4:66–82.

Zhang, J. (2008). On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias. 172:1873–1896.

Zhang, K., Peters, J., Janzing, D., and Schölkopf, B. (2011). Kernel-based conditional independence test and application in causal discovery. In *Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence*, UAI'11, page 804–813, Arlington, Virginia, USA. AUAI Press.

Zheng, X., Aragam, B., Ravikumar, P., and Xing, E. P. (2018). Dags with no tears: Continuous optimization for structure learning. *arXiv preprint arXiv:1803.01422*.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes, we defer all proofs to the supplementary material.]
   (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Yes]
   (b) The license information of the assets, if applicable. [Yes]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Yes]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

A Proofs

Lemma 3.1 (Significance and Power) Let X, Y be unconfounded and $X \to Y$. Let Π_X, Π_Y be the corresponding partitions.

$$\lim_{n_c \to \infty} \mathcal{P}(t > q_{1-\alpha}) \to \alpha \,,$$

where $q_{1-\alpha}$ is the $1-\alpha$-quantile of standard normal distribution. Conversely, if X, Y are confounded by Z, then for any $\alpha > 0$ we obtain a power of 1 in the limit,

$$\lim_{n_c \to \infty} \mathcal{P}(t > q_{1-\alpha}) \to 1 \,.$$

Proof. Since t is asymptotically normal (Vinh et al., 2009), the first assertion follows directly.

For the converse statement, note that for two confounded variables X_1, X_2, their partitions satisfy

$$\mathbb{E}I(\Pi_1, \Pi_2) \geq \frac{n_c}{2} H(p) \gg \mathbb{E}I(\Pi_1', \Pi_2')$$

where $H(p) = -p\log(p) - (1-p)\log(1-p)$ is the binary entropy of the probability p of two different contexts belonging to different sets of the partition as defined in Assumption 3. Note that the relation $\frac{n_c}{2} H(p) \gg \mathbb{E}I(\Pi_1', \Pi_2')$ follows from the fact that $\lim_{n_c \to \infty} \frac{1}{n_c} I(\Pi_1', \Pi_2') = 0$ $\mathcal{P}$-almost surely so that $\mathbb{E}I(\Pi_1', \Pi_2')$ cannot be extensive in n_c. Since $I(\Pi_1, \Pi_2)$ also concentrates around its mean, the result follows. □

Theorem 3.1 Let X_S be a set of variables such that all $X_i, X_j \in X_S$ are pair-wise confounded. Then X_S is jointly confounded if and only if for each triple $X_i, X_j, X_k \in S$ we have

$$\lim_{n_c \to \infty} \mathcal{P}(T(\Pi_i, \Pi_j, \Pi_k) < I(\Pi_i, \Pi_j) + I(\Pi_j., \Pi_k)) = \begin{cases} 1, & X_1, X_j, X_k \text{ jointly confounded} \\ 0, & \text{otherwise} \,. \end{cases}$$

Proof. As we have seen, the condition $T(\Pi_i, \Pi_j, \Pi_k) < I(\Pi_i, \Pi_j) + I(\Pi_j, \Pi_k)$ is equivalent to $I(\Pi_j, \Pi_k \mid \Pi_i) < I(\Pi_j, \Pi_k)$, which is true if and only if the correlations between the partitions are shared, which can only happen due to joint confounding of more than two variables at a time. Now, let us assume that some set S of $s \geq 4$ pair-wise confounded nodes satisfying this inequality, does not share the same confounder between all nodes. W.l.o.g. let us call these variables $X_1, \ldots, X_s$. Then the way for *every* triplet to have a shared confounder, and that requires the least number of edges into, is for three distinct confounders to affect the sets $\{X_1, \ldots, X_{s-1}\}$, $\{X_2, \ldots, X_s\}$, and $\{X_1, X_2, X_s\}$. This requires $2(s-1)+3 = 2s+1$ edges into the set $X_1, \ldots, X_s$ in contradiction with Assumption 6. □

Proposition 3.1 (Consistency for Pairs of Variables) If a variable pair X, Y is confounded by a variable Z, then there exists some $\rho > 0$ such that

$$\mathcal{P}(\mathrm{I}(\Pi_X^*, \Pi_Y^*) < \mathrm{I}(\Pi_X, \Pi_Y)) = 1 - O(e^{-\rho n_c}) \,.$$

Proof. Following Perry et al. (2022) we show more precisely that

$$\mathcal{P}(\mathrm{I}(\Pi_X^*, \Pi_Y^*) < \mathrm{I}(\Pi_X, \Pi_Y)) = 1 - O\left((p + (1-p)(1-p+p^2))^{\lfloor n_c/2 \rfloor}\right) \,,$$

by splitting the contexts into pairs c_{2i}, c_{2i+1} and note we will get a wrong result if and only if for *all* these pairs of contexts we have that a change in the mechanism of Y does not introduce an additional change in the mechanism of $X \mid Y$.

The probability of this not happening for any one pair is given by three parts: either the mechanism of Z already changes between the environments (probability p), or it does not (probability $1-p$) *and* either Y does not change (probability $1-p$) or both X and Y change (probability p^2).

Since the changes between any two environments c_{2i}, c_{2i+1} are independent of each other, the probability of this happening in *all* environments is therefore $(p + (1-p)(1-p+p^2))^{\lfloor n_c/2 \rfloor}$ and since $p + (1-p)(1-p+p^2) \leq 1$ as convex combination of 1 and $(1-p+p^2)$, the result follows. □

Proposition 3.2 (Recovering Parents) Let X_i be a target variable and let $\mathcal{G}$ and $\mathcal{G}'$ be two graphs in the MEC of the marginal distribution $P^c(X)$. Assume that only one of the two graphs correctly recovers the parents of X_i, $\text{pa}_i = \text{pa}_i^*$ and $\text{pa}_i' \neq \text{pa}_i^*$, and further assume that the number of latent confounders affecting X_i plus spurious siblings is bounded by $\frac{\log(0.5)}{\log(1-p)}$. Then

$$\begin{aligned}\mathcal{P}(\mathrm{I}(\Pi_i, \{\Pi_j : j \in \text{pa}_i\}) < \mathrm{I}(\Pi_i', \{\Pi_j' : j \in \text{pa}_i'\})) \\ = 1 - O(e^{-\rho n_c})\end{aligned}$$

Proof. More precisely, we will show that

$$\mathcal{P}(\mathrm{I}(\Pi_i, \{\Pi_j : j \in \text{pa}_i\}) < \mathrm{I}(\Pi_i', \{\Pi_j' : j \in \text{pa}_i'\})) = 1 - O\left(\left((1-(1-p)^r) + (1-p)^r(1-p+p^2)\right)^{n_c/2}\right),$$

where r is the number of latent parents of X_i plus the number of other variables with which it is pair-wise confounded. In essence, these variables are precisely those which could make us not detect changes between two environments, just as in the previous proof changes in the mechanism of Z between environments could prevent us from detecting changes in the mechanisms of X or Y.

To this end, note that if $\text{pa}_i' \neq \text{pa}_i^*$ then there exists either a variable in pa_i^* that is missing in pa_i' or a child of X_i in pa_i'. In either case, additional joint shifts are introduced between X_i and these variables and therefore the mutual information increased. This increase in mutual information is guaranteed by the fact that $r \leq \frac{\log(0.5)}{\log(1-p)}$, so that the probability of mechanism between shifts in X_i is less than 0.5. □

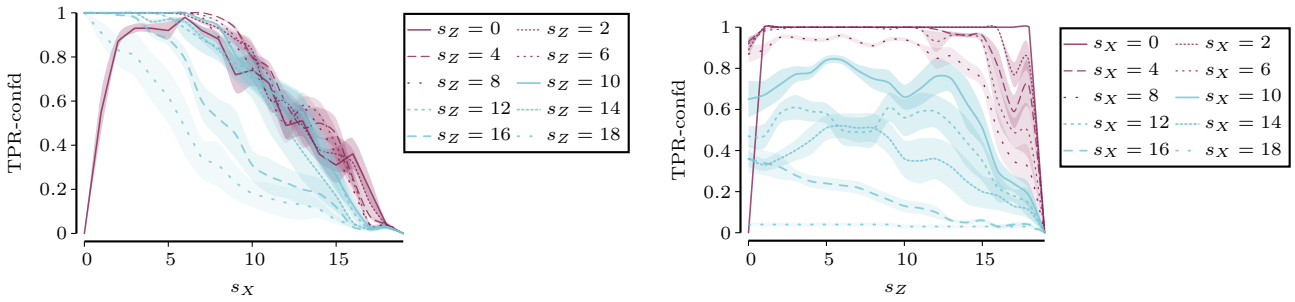
Theorem 3.2 (Consistency) Let $\mathcal{G}^*$ be the true graph over V and let $\mathcal{G}_x^*$ be the induced graph on X, and assume that for all X_i the number of latent confounders plus spurious siblings is bounded by $\frac{\log(0.5)}{\log(1-p)}$. Then with high probability, $\mathcal{G}_x^*$ and its partitions $\Pi_1^*, \ldots, \Pi_k^*$ are the unique minimum of total correlation,

$$\mathcal{P}(\arg\min_{\mathcal{G}} T(\Pi_1, \ldots, \Pi_m) = \{\mathcal{G}_x^*\}) = 1 - O(e^{-\rho n_c}) .$$

Proof. Let m be the number of observed variables and r be an upper bound on all the $r = \max\{r_i\}$ from the above Proposition. Then we specifically show that

$$\mathcal{P}\left(\arg\min_{\mathcal{G}} T(\Pi_1, \ldots, \Pi_m) = \{\mathcal{G}_x^*\}\right) = 1 - O\left(\frac{m^2(m-1)}{2}\left((1-(1-p)^r) + (1-p)^r(1-p+p^2)\right)^{n_c/2}\right).$$

To this end let us assume that the true causal ordering over X is given by $X_1 \leq \cdots \leq X_m$. Then note that by construction $T(\Pi_1, \ldots, \Pi_m) = \sum_i \mathrm{I}(\Pi_i, \{\Pi_j : j \in \text{pa}_i\})$ so that the inside of our statement here is simply the sum of all terms in Proposition 3.2. As such, the total correlation is the unique minimum if the above proposition holds for all i and when compared against *any* other graph $\mathcal{G}$, resulting in the union bound above. □



(a) Effect of increasing the number of shifts of the observed variables.

(b) Effect of increasing the number of shifts of the confounder.

Figure 5: *Identifiability depending on the number of mechanism shifts of the observed and latent variables.* We show the power of our confounding test (the true positive rate of decisions over node pairs, higher is better) depending on the observed mechanism shifts s_X (left) and latent shifts s_Z (right) over $n_c = 20$ contexts. We can identify confounding when observed mechanism shifts are sparse ($s_X < 10$, red plots on the right) unless the confounder changes in almost every context ($s_Z > 15$, blue plots on the left) or does not change at all ($s_Z = 0$).

B Empirical Analysis

In this section, we revisit our theoretical results and assumptions and support them with empirical results.

B.1 Sparse Mechanism Shift (Assumption 3)

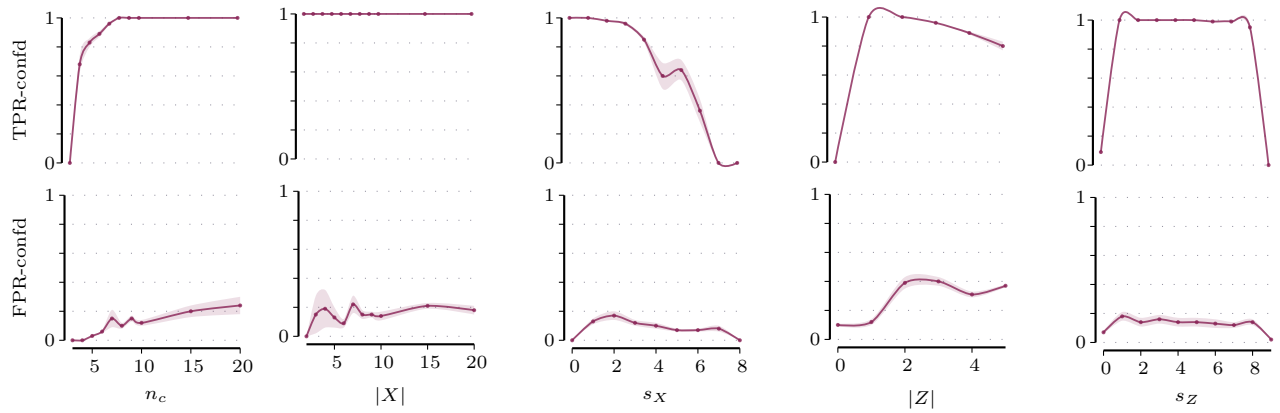
The key assumption in our analysis is the sparse mechanism shift hypothesis (Guo et al., 2022; Schölkopf et al., 2012). It states that distribution changes are a result of only a *small* number of changes in causal mechanisms. This is consistent with the view of causal mechanisms as independent modules that do not influence each other, and is closely related to the invariance principle whereby causal mechanisms remain the same even in different contexts (Peters et al., 2016; Huang et al., 2020). While sparsity has recently been proposed as a relaxation of the i.i.d. assumption (Perry et al., 2022), it is not easily testable in practice. Hence, we want to investigate empirically how sensitive our confounding test is to an increasing number of causal mechanism changes.

To this end, we we vary the number of changes for the observed (s_X) and latent variables (s_Z) in a fixed set of contexts, here $n_c = 20$. We generate data as in our main experiments and test with CoCo(Π^*) for confounding between all node pairs in a causal DAG. To show the empirical power of our confounding test, we show the true positive rate (TPR-confd) over these decisions in Fig. 5.

Observed Shifts In Fig. 5a, we show the effect of increasing s_X. We run the experiment for each s_Z, and color plots red if latent shifts are sparse ($s_Z < 10$) and blue otherwise. We observe a tipping point at $s_X = 9$ where the observed nodes have partitions with 10 different groups of the 20 contexts, that is, exactly when mechanism shifts are no longer sparse, the power of our test decreases. For $s_X < 10$, we have perfect power in most cases. We note that in the special case where all variables are identically distributed in all contexts, $s_Z = 0, s_X = 0$, the confounding effect is not measurable using our method.

Latent Shifts In Fig. 5a, we show the same result when we increase s_Z instead. Sparse shifts $s_X < 10$ are now colored red, dense shifts blue. We can see a clear separation of the two cases, confirming our observations above. In particular, under sparse shifts of s_X, we can tolerate up to $s_Z = 15$ shifts of the confounder.

We conclude that our approach works best in settings where the sparse shift assumption holds for the observed variables, while we can handle more shifts for the latent variables. We obtained similar trends for $n_c = 5, 10$.



(a) Contexts (b) Variables (c) Observed Shifts (d) Latent Variables (e) Latent Shifts

Figure 6: *Empirical Power and Significance.* We show the true positive rate (top, higher is better) and false positive rate (bottom, lower is better) of our confounding test depending on the number of contexts, variables, confounders, and mechanism shifts, starting from $n_c = 10, |X| = 10, |Z| = 1, s_X = 1, s_Z = 2$.

B.2 Empirical Significance and Power (Lemma 1)

Next, we revisit Lemma 1, which guarantees a power of 1 of our test as we observe more contexts, $n_c \to \infty$. To give a more practical result for fewer contexts, we investigate the power and significance of our confounding test empirically.

We consider CoCo(Π^*) to study our confounding test in isolation, and show true positive rates (TPR-confd) and false positive rates (FPR-confd) to show the power, respectively significance, of the test. As in our main experiment, we test for confounding between all pairs of nodes in a causal DAG and consider $n_X = 10$ nodes in $n_c = 10$ contexts, where one confounder influences a random set of between two and n_X nodes, and where nodes undergo $s_X = 1$ mechanism change and the confounder $s_Z = 2$ changes. We show the results in Fig. 6. We note that we consider up to $\frac{n_X}{2} = 5$ confounders because each confounder always affects at least two variables, and up to $n_c - 1 = 9$ mechanism changes because this corresponds to a change in every context.

Power We find that our test already works well given few contexts, with perfect power starting from $n_c = 8$ contexts (Fig. 6a). We point out the special case $s_Z = n_c - 1$, where the confounder changes in *every* context. In this case, the mutual information of the observed (single-group) partitions does not differ from the expected one and we cannot detect confounding, as we can see for $n_c = 3$ (Fig. 6a) and $n_c = 10$ (Fig. 6e). Otherwise, the number of latent shifts does not affect the results significantly (Fig. 6e) and only the shifts of the observed variables do (Fig. 6c), as we discussed in the previous section. The sensitivity of our test is not affected by the number of variables (Fig. 6b) and decreases slightly when we add more confounders to the system (Fig. 6d).

Significance As the false positive rates show, our test rarely detects unconfounded variable pairs as confounded, with FPR-confd remaining around 0.1 and below in most experiments. We notice a change when there is more than one confounder (Fig. 6d). To explain, in this case we also check whether variables are affected by the *same* confounder, and our method may discover a variable pair X_i, X_j as confounded when they are each affected by a *different* confounder, $Z_k \to X_i, Z_l \to X_j$. In particular, this happens if Z_k, Z_l have joint mechanism shifts coincidentally, in which case the mechanism shifts of X_i, X_j also appear correlated. However, even in this case, FPR-confd remains below 0.2 (Fig. 6d), suggesting that our method can mostly separate which variables are affected by which latent variable.

content...

Algorithm 2: CoCo(Π)

input : Partitions Π for each observed variable.
output: Subsets of X that are jointly confounded by a latent variable Z_j.

1 **foreach** pair of variables X_i, X_j **do**
2 Test whether there is a common confounder, $p_{ij} = p\text{-val}(I(\Pi_i, \Pi_j) \neq \mathbb{E}\left[\mathrm{I}(\Pi_i^\star, \Pi_j^\star)\right])$.
3 Construct an affinity matrix $\mathcal{M}$ with entries $\mathcal{M}_{ij} = I(\Pi_i, \Pi_j)$
4 Discover subsets X_S of X that are connected components in $\mathcal{M}$, using spectral clustering
5 **return** Subsets X_S

Algorithm 3: CoCo($\mathcal{G}$)

input : Data over X, C; causal DAG $\mathcal{G}$.
output: Subsets of X that are jointly confounded by a latent variable Z_j.

1 **foreach** variable X_i **do**
2 **foreach** pair of contexts c, c' **do**
3 Test whether there is a causal mechanism change for X_i between the contexts,
4 $p_{c,c'} = p\text{-val}\left(P^c(X_i \mid \mathrm{pa}_i) \neq P^{c'}(X_i \mid \mathrm{pa}_i)\right)$.
5 Convert $\{p_{c,c'}\}$ to a partition Π_i, ensuring that $p_{c,c'} \geq \alpha \Rightarrow \Pi_i(c) \neq \Pi_i(c')$
6 **return** CoCo (Π)

Algorithm 4: CoCo($\mathcal{M}$)

input : Data over X, C; causal MEC $\mathcal{M}$.
output: Subsets of X that are jointly confounded by a latent variable Z_j.

1 Discover the best DAG in $\mathcal{M}$ using MSS,
2 $\mathcal{G} = \arg\min_{\mathcal{G} \in \mathcal{M}} \mathrm{MSS}(\mathcal{G})$
3 **return** CoCo ($\mathcal{G}$)

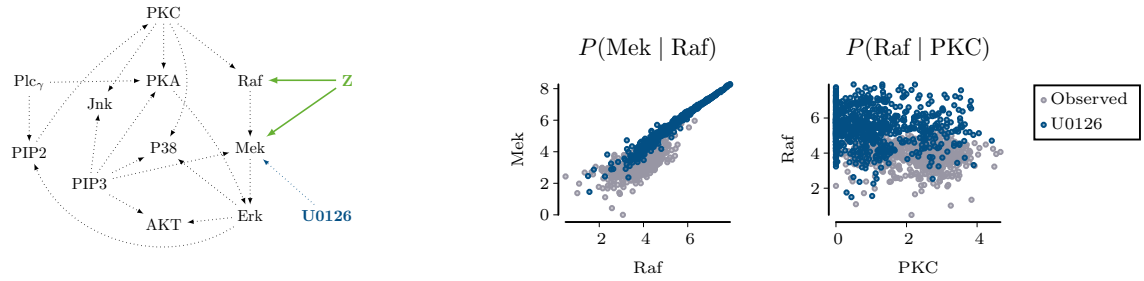
C Additional Details: Methodology

In our evaluation, we add oracles in different stages of our algorithm and combine CoCo with the Mechanism Shift Score (MSS, Perry et al. (2022)). For completeness, we include the pseudo-code for all versions here.

Alg. 2 implements our main confounding test. It starts from a set of partitions Π that encode the mechanism changes for each variable. For each pair of variables, we test whether the mutual information of their partitions is higher than expected, obtaining p-values indicating whether the variables are likely confounded (line 2). Using the pairwise mutual information as a measure of "distance" between nodes, we apply spectral clustering, which is commonly used to identify connected components in graphs (Donath and Hoffman, 1972). In this way, we discover subsets X_S that are likely affected by the *same* confounder (line 4). We obtain an oracle version of Alg. 2 by providing it with the true partitions $\Pi^\star$ according to our data generation process. Throughout our experiments, we color the oracle version CoCo($\Pi^\star$) purple.

To discover the partitions from data, Alg. 3 a performs conditional distribution discrepancy test between each pair of contexts (line 4). Unless otherwise stated, we use the Kernel Conditional Independence test (KCI, Zhang et al. (2011)). We convert the pairwise tests to a clustering, where we ensure that contexts are assigned to different groups whenever we detect a mechanism change between them (line 5). While we present Alg. 3 as our main algorithm, it needs to start from the true structure $\mathcal{G}^\star$, and we color the oracle CoCo($\mathcal{G}^\star$) light blue.

Finally, to discover $\mathcal{G}$ from data, we can in principle use any DAG discovery method. We give a proof-of-concept implementation in Alg. 4. It uses the MSS estimand (line 2), which determines causal directions as the ones that result in the fewest causal mechanism shifts (Perry et al., 2022). We color this version green in our experiments.



(a) Consensus Network, intervention with U0126 and confounder that CoCo discovers

(b) Joint Mechanism Shift of Raf and Mek that CoCo discovers

Figure 7: *Hidden confounding between Raf and Mek.* In the cell signalling network (Sachs et al., 2005), CoCo discovers confounding between the molecules Raf and Mek (a). Although the consensus network only contains the edge Raf $\to$ Mek, many causal discovery methods also report a pathway Mek $\to$ Raf (Perry et al., 2022), suggesting that there may be feedback. We illustrate this in (b), where we show the data in the observational context (gray) and in an interventional context (blue) where the reagent U0126 was added (Mooij et al., 2020). While U0126 is presumed to only directly influence Mek, we see a change in the abundance of Raf. With CoCo, we discovered a joint mechanism change of both conditionals P(Raf | PKC) and P(Mek | Raf) in the interventional context, and overall found that partitions for Raf and Mek are correlated.

D Additional Details: Experiments

We end with a more detailed discussion of our experiment on real-world data.

D.1 Confounded Edges in the Flow Cytometry Data (Fig. 3)

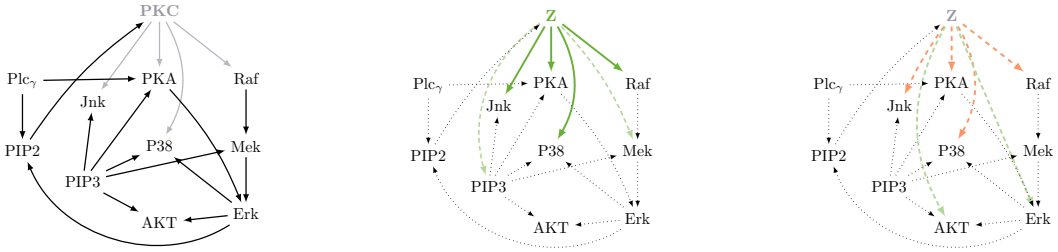
In our experiment on the flow cytometry data by Sachs et al. (2005), we mimic confounding by removing one variable at a time from the dataset and testing whether CoCo discovers its causal children as confounded.

When we keep the most influential variable PKC hidden, we find that CoCo captures its confounding effect on all of its four causal children (Fig. 8). As most other variables have at most one incoming edge, only PIP3 and Erk serve as potential other confounders, but CoCo does not detect their effects on JNK, P38, and AKT. However, CoCo returns remarkably few confounding effects that disagree with the consensus, and we discuss these cases here.

- PIP3: We discover PIP3 in the confounded set when we keep PKC hidden todo (Fig. 8) and when we keep PIP2 hidden. As a likely explanation, we note that Sachs et al. (2005) include a *bidirected* pathway between PIP2 and PIP3, and indeed also other causal discovery methods report edges between both signaling molecules (Sinha et al., 2021; Perry et al., 2022).
- Raf $\leftrightarrow$ Mek: We discover confounding between Raf and Mek regardless of which other variable we keep hidden. Mooij et al. (2020) discuss the relationship between these signaling molecules in detail as an example suggesting that the consensus network may be incomplete. As shown in Fig. 7, this network includes the pathway Raf $\to$ Mek, and the only intervention targeting either of the molecules is the Mek inhibitor U0126 (Sachs et al., 2005). Consider however the data shown in Fig. 7b. We show the observational context (gray) and the interventional context where the reagent U0126 was added (blue), and can see that there is a distribution shift of Raf under U0126. This suggests that either U0126 also targets Raf, or there is a feedback loop from Mek to Raf; for example, a path from Erk to Raf was suggested (Mooij et al., 2020).

  We found that CoCo detects this observation. In the partitions for Mek and Raf, reflecting changes in the conditional distributions P(Mek | Raf) and P(Raf | PKC), we discover a joint mechanism shift of both signaling molecules in the interventional context U0126, and higher than expected mutual information of the partitions, hence deciding that Raf and Mek are confounded.

In conclusion, CoCo discovers a *dependent* mechanism shift of Raf and Mek under the intervention with U0126, thus pointing to potential hidden confounding between the cells that is consistent with the data.



(a) Consensus Network

(b) Confounder that CoCo discovers

(b) Confounder that JCI-FCI discovers

Figure 8: *Confounding effects of PKC.* On the Sachs et al. (2005) data CoCo recovers confounding effects of PKC. Solid green edges are correctly discovered as confounded, dashed green edges are additional edges discovered as confounded, and orange edges are confounded but not discovered.

D.2 Competitors on the Flow Cytometry Data

Finally, we also include the result of JCI-FCI on this dataset. As we show in Fig. 8, JCI-FCI does not discover the confounding edges (dashed orange), and reports two false positives (dashed green).

Bibliography

Bhattacharya, R., Nagarajan, T., Malinsky, D., and Shpitser, I. (2021). Differentiable causal discovery under unmeasured confounding. In *International Conference on Artificial Intelligence and Statistics*, pages 2314–2322. PMLR.

Chandrasekaran, V., Parrilo, P. A., and Willsky, A. S. (2010). Latent variable graphical model selection via convex optimization. In *2010 48th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 1610–1613. IEEE.

Chickering, D. M. (2002). Learning equivalence classes of bayesian-network structures. *Journal of Machine Learning Research*, 2:445–498.

Colombo, D., Maathuis, M. H., Kalisch, M., and Richardson, T. S. (2012). Learning high-dimensional directed acyclic graphs with latent and selection variables. *The Annals of Statistics*, 40(1):294–321.

Dominguez, A. A., Lim, W. A., and Qi, L. S. (2016). Beyond editing: repurposing crispr–cas9 for precision genome regulation and interrogation. *Nature Reviews Molecular Cell Biology*, 17(1):5–15.

Donath, W. E. and Hoffman, A. J. (1972). Algorithms for partitioning of graphs and computer logic based on eigenvectors of connection matrices. *IBM Technical Disclosure Bulletin*, 15(3):938–944.

Elidan, G., Lotner, N., Friedman, N., and Koller, D. (2000). Discovering hidden variables: A structure-based approach. *Advances in Neural Information Processing Systems*, 13.

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

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. (2012). A Kernel Two-Sample Test. *Journal of Machine Learning Research*, 13(25):723–773.

Guo, S., Tóth, V., Schölkopf, B., and Huszár, F. (2022). Causal de finetti: On the identification of invariant causal structure in exchangeable data. *arXiv*.

Huang, B., Zhang, K., Zhang, J., Ramsey, J., Sanchez-Romero, R., Glymour, C., and Schölkopf, B. (2020). Causal discovery from heterogeneous/nonstationary data. *Journal of Machine Learning Research*, 21(1).

Janzing, D. and Schölkopf, B. (2010). Causal inference using the algorithmic markov condition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 56:5168–5194.

Kaltenpoth, D. and Vreeken, J. (2023a). Causal Discovery with Hidden Confounders using the Algorithmic Markov Condition. In *Uncertainty in Artificial Intelligence*, pages 1016–1026. PMLR.

Kaltenpoth, D. and Vreeken, J. (2023b). Nonlinear Causal Discovery with Latent Confounders. In *International Conference on Machine Learning*, pages 15639–15654. PMLR.

Karlsson, R. and Krijthe, J. (2023). Detecting hidden confounding in observational data using multiple environments. In Oh, A., Neumann, T., Globerson, A., Saenko, K., Hardt, M., and Levine, S., editors, *Advances in Neural Information Processing Systems*, volume 36, pages 44280–44309. Curran Associates, Inc.

Mameche, S., Kaltenpoth, D., and Vreeken, J. (2023). Learning causal mechanisms under independent changes.

Mooij, J., Magliacane, S., and Claassen, T. (2020). Joint causal inference from multiple contexts. *Journal of Machine Learning Research*, 21:99:1–99:108.

Ogarrio, J. M., Spirtes, P., and Ramsey, J. (2016). A hybrid causal search algorithm for latent variable models. In *Proceedings of the Eighth International Conference on Probabilistic Graphical Models*, pages 368–379.

Park, J., Shalit, U., Schölkopf, B., and Muandet, K. (2021). Conditional distributional treatment effect with kernel conditional mean embeddings and u-statistic regression. In *Proceedings of 38th International Conference on Machine Learning (ICML)*, volume 139 of *Proceedings of Machine Learning Research*, pages 8401–8412. PMLR.

Pearl, J. (2009a). *Causality: Models, Reasoning and Inference*. Cambridge University Press, 2nd edition.

Pearl, J. (2009b). *Causality: Models, Reasoning and Inference*. Cambridge University Press.

Perry, R., Kügelgen, J. V., and Schölkopf, B. (2022). Causal discovery in heterogeneous environments under the sparse mechanism shift hypothesis.

Peters, J., Bühlmann, P., and Meinshausen, N. (2016). Causal inference using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(5):947–1012. (with discussion).

Peters, J., Janzing, D., and Schlkopf, B. (2017). *Elements of Causal Inference: Foundations and Learning Algorithms*. The MIT Press.

Ramsey, J., Glymour, M., Sanchez-Romero, R., and Glymour, C. (2017). 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.

Reddy, A. G., Dash, S., Sharma, A., and Balasubramanian, V. N. (2022). Counterfactual generation under confounding. In *NeurIPS 2022 Workshop on Causality for Real-world Impact*.

Richardson, T. S., Evans, R. J., Robins, J. M., and Shpitser, I. (2017). Nested markov properties for acyclic directed mixed graphs. *arXiv preprint arXiv:1701.06686*.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. 66:688–701.

Sachs, K., Perez, O., Pe'er, D., Lauffenburger, D., and Nolan, G. (2005). Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, pages 523–9.

Scanagatta, M., de Campos, C. P., Corani, G., and Zaffalon, M. (2015). Learning bayesian networks with thousands of variables. In *NIPS*, pages 1864–1872.

Schölkopf, B., Janzing, D., Peters, J., Sgouritsa, E., Zhang, K., and Mooij, J. (2012). On causal and anticausal learning. ICML'12, pages 459–466, Madison, WI, USA. Omnipress.

Schölkopf, B., Locatello, F., Bauer, S., Ke, N. R., Kalchbrenner, N., Goyal, A., and Bengio, Y. (2021). Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634.

Shpitser, I., Evans, R. J., and Richardson, T. S. (2018). Acyclic linear sems obey the nested markov property. In *Uncertainty in artificial intelligence: proceedings of the... conference. Conference on Uncertainty in Artificial Intelligence*, volume 2018. NIH Public Access.

Shpitser, I., Evans, R. J., Richardson, T. S., and Robins, J. M. (2014). Introduction to nested markov models. *Behaviormetrika*, 41(1):3–39.

Sinha, M., Tadepalli, P., and Ramsey, S. A. (2021). Voting-based integration algorithm improves causal network learning from interventional and observational data: An application to cell signaling network inference. *PLoS One*, 16(2):e0245776.

Spirtes, P. (2001). An Anytime Algorithm for Causal Inference. In *International Workshop on Artificial Intelligence and Statistics*, pages 278–285. PMLR.

Spirtes, P., Glymour, C. N., and Scheines, R. (2000). *Causation, prediction, and search*. MIT press.

Spirtes, P., Meek, C., and Richardson, T. (1999). An algorithm for causal inference in the presence of latent variables and selection bias in computation, causation and discovery. MIT Press.

Vinh, N. X., Epps, J., and Bailey, J. (2009). Information theoretic measures for clusterings comparison: is a correction for chance necessary? In *Proceedings of the 26th annual international conference on machine learning*, pages 1073–1080.

Vinh, N. X., Epps, J., and Bailey, J. (2010). Information Theoretic Measures for Clusterings Comparison: Variants, Properties, Normalization and Correction for Chance. *Journal of Machine Learning Research*, 11(95):2837–2854.

Watanabe, S. (1960). Information theoretical analysis of multivariate correlation. *IBM Journal of Research and Development*, 4:66–82.

Zhang, J. (2008). On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias. 172:1873–1896.

Zhang, K., Peters, J., Janzing, D., and Schölkopf, B. (2011). Kernel-based conditional independence test and application in causal discovery. In *Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence*, UAI'11, page 804–813, Arlington, Virginia, USA. AUAI Press.

Zheng, X., Aragam, B., Ravikumar, P., and Xing, E. P. (2018). Dags with no tears: Continuous optimization for structure learning. *arXiv preprint arXiv:1803.01422*.