Intervention Target Estimation in the Presence of Latent Variables (Supplementary Material)

Burak Varıcı[1] **Karthikeyan Shanmugam** [*2] **Prasanna Sattigeri**[2] **Ali Tajer**[1]

[1]Rensselaer Polytechnic Institute
[2]IBM Research AI

Abstract

This paper considers the problem of estimating unknown intervention targets in causal directed acyclic graphs from observational and interventional data in the presence of latent variables. The focus is on linear structural equation models with soft interventions. The existing approaches to this problem involve performing extensive conditional independence tests, and they estimate the unknown intervention targets alongside learning the structure of the causal model in its entirety. This joint learning approach results in algorithms that are not scalable as graph sizes grow. This paper proposes an approach that does not necessitate learning the entire causal model and focuses on learning only the intervention targets. The key idea of this approach is leveraging the property that interventions impose sparse changes in the precision matrix of a linear model. The proposed framework consists of a sequence of precision difference estimation steps. Furthermore, the necessary knowledge to refine an observational Markov equivalence class (MEC) to an interventional MEC is inferred. Simulation results are provided to illustrate the scalability of the proposed algorithm and compare it with those of the existing approaches.

1 INTRODUCTION

Enabling modern machine learning systems to reason involves predicting the effect of an intervention and counterfactual estimation [Pearl, 2009]. Forming such predictions crucially depends on the knowledge of causal models [Pearl and Mackenzie, 2018]. One approach to represent causal knowledge is through a causal Bayesian network, which is a directed graphical model specified by a directed acyclic graph (DAG). The nodes of a DAG represent random variables, and its directed edges represent the cause-and-effect relationships among the random variables. Such a model facilitates factorizing the observed distribution, where each factor is a conditional distribution of a variable given its causal parents. These conditionals specify the local causal mechanisms of the variables. However, based on purely observational data, a causal DAG is identifiable only up to an equivalence class of DAGs. Such uncertainty is because different DAGs can encode different ways of factorizing the same observed distribution into conditionals. The equivalence class of DAGs that can be identified from the observational data alone is called the Markov equivalence class (MEC) [Peters et al., 2017].

To reduce the ambiguity in the MEC obtained from the observational data, interventional data can be leveraged. Intervening on a variable refers to modifying the causal mechanism (the conditional distribution) that connects this variable and its parents in the true causal DAG while leaving the other factors unchanged. The combination of observational and interventional data reduces the number of possible factorizations that are consistent with both data types. In this paper, we perform *soft* interventions. A soft intervention induces a change in the causal mechanism by replacing it with a different one without requiring the causal effects on the target node to be removed. While hard interventions, e.g., assigning fixed values to intervention targets, can be performed too, there are applications in which soft interventions are better suited for modeling the experiments. For instance, soft interventions can effectively model altering the gene expressions for cellular reprogramming [Zhang et al., 2021].

In a broad range of applications, when interventional data is available, the variables whose causal mechanisms have been changed, called the *intervention targets*, are unknown. For instance, there is a recent growing interest in using causal discovery for fault localization in microservices systems in cloud-native applications [Bogatinovski et al., 2021,

[*] Author is currently affiliated with Google Research India. The work was done while the author was at IBM Research AI, NY.

Aggarwal et al., 2020]. These systems are built as an interconnected set of loosely coupled services across various layers [Kim et al., 2013, Mariani et al., 2018]. Such systems are vulnerable to unwanted changes (e.g., equipment failure and attacks). During the faulty operation of these systems, it is imperative to localize the faults quickly. The root causes of the faulty operations are modeled as interventions to the system. Hence, the data is collected under (unknown) faults, rendering fault localization a causal discovery task from interventional data of unknown intervention targets. Furthermore, a fault in the operation of a node, e.g., a delay, is closer to soft interventions than to hard interventions since the causal parents can still affect the operation of the node. Another example is gene knockout experiments in biology. In these experiments, a target set of genes is knocked out in an assay, and gene expressions are collected. These are known to affect off-target genome sites [Fu et al., 2013]. Sometimes drugs are injected into protein signaling networks, and the expression levels are measured. In these settings, the intervention targets are unknown [Sachs et al., 2005, Ness et al., 2017].

Identifying unknown intervention targets in fully observed graphs was recently explored [Varici et al., 2021]. However, in this study, all variables of a true causal DAG are typically not observed. This induces confounding between observed variables due to unobserved or latent variables. A model with such confounding is called a *causally insufficient* model. Recent studies have characterized the interventional MEC for causally insufficient models and have provided algorithms for learning their structures. These algorithms leverage invariance testing and conditional independence testing by using both interventional and observational data and accommodate both settings of known and unknown intervention targets [Mooij et al., 2020, Kocaoglu et al., 2019, Jaber et al., 2020]. In these algorithms, the intervention targets are usually learned along with the interventional MEC. In this paper, we focus on the following question: **is there an efficient way to learn only the intervention targets given interventional and observational data?**

Our Contributions: We address the above question in linear structural equation models (SEMs) under soft interventions. We first show that the difference in the precision matrices of the interventional datasets can be used to deduce the intervention status of a node. Next, we use the fact that these precision differences have sparse support to narrow down our interest to the nodes directly affected by the interventions. Then, we show how to refine this sparse set by repeated precision difference estimations to obtain the intervention targets. In the process, we also infer the causal knowledge newly induced by the interventions. Finally, using these elements, we propose a scalable algorithm to estimate the intervention targets.

There are two studies whose scopes are close to that of this paper. Jaber et al. [2020] characterize the interventional MEC for soft interventions and proves that the intervention targets can be identified only up to a superset that they graphically describe. Noting this result, in this paper, we focus on estimating this superset, which we call the *effective intervention targets*. In a different study, Varici et al. [2021] address a related problem. However, their method is limited to only causally sufficient models. We present theoretical results for causally insufficient models, which are non-trivial generalizations that combine the precision difference approach to the problem and the graphical characterization of the soft interventions on causally insufficient models.

The existing interventional causal discovery algorithms for insufficient systems jointly learn the causal structure and the intervention targets. These approaches require performing a significant number of conditional independence and invariance tests, a major impediment to these algorithms for being scalable to large graphs [Jaber et al., 2020]. However, unlike interventional causal discovery, there exist highly efficient algorithms for causal discovery with observational data. One of the byproducts of our results is that our scalable algorithm for intervention target discovery can be used in conjunction with any observational learning algorithm for insufficient systems to refine the observational MEC efficiently to an interventional MEC. Finally, we perform experiments on real and synthetic datasets to illustrate the scalability of the proposed algorithm.

2 RELATED WORK

At its core, this paper infers causal knowledge from interventional settings through an invariance criterion. The existing literature on related topics is discussed next.

Interventional causal learning for causally sufficient systems. There is extensive literature on interventional learning for causally sufficient models. Among them, Eaton and Murphy [2007] proposed a dynamic programming approach to interventional learning. Hauser and Bühlmann [2012] considers the interventional MEC under hard interventions and provides a score-based algorithm for interventional learning. Rothenhäusler et al. [2015] learn causal cyclic graphs using shift interventions. Ghassami et al. [2018] consider multi-domain data without explicitly formulating the different domains via interventions. Its method estimates the causal order by generalizing the invariance of parameters to the independence of the changes in the parameters across domains. Huang et al. [2020] use the distribution shifts that can be the results of interventions to determine the causal directions. Their method works under a pseudo-causal sufficiency condition in which the values of the unobserved confounders are fixed in each domain. Yang et al. [2018] characterize interventional MEC under hard and soft interventions using invariance testing and provides a learning algorithm when the intervention targets are known. The algorithm of Squires et al. [2020] greedily searches over the space of permutations to score DAGs when the intervention targets are unknown. Ke et al. [2019] and Brouillard et al. [2020] leverage differentiable methods through continuous optimization to learn the causal structure from interventional data. For linear SEMs and causally sufficient models, Wang et al. [2018] propose to learn the difference graph, which is the set of edge weights in the linear SEM that have been changed across two environments. Ghoshal et al. [2021] leverage precision difference estimates to address the same problem under more stringent assumptions. Varici et al. [2021] use precision difference estimates and achieves a higher level of scalability through a hierarchical grouping of the nodes.

Learning from observational data for causally insufficient systems. The fast causal inference (FCI) algorithm of Spirtes et al. [2000] is a classic constraint-based method for learning causally insufficient models from observational data. Many efficient variants such as the really fast causal inference (RFCI) algorithm of Colombo et al. [2012] and the greedy fast causal inference (GFCI) algorithm of Ogario et al. [2016] have been proposed to improve scalability. Bernstein et al. [2020] extend the greedy permutation search to partially ordered sets to include the effects of latent variables in ordering.

Learning from interventions on causally insufficient systems. Triantafillou and Tsamardinos [2015] consider multiple interventions for causally insufficient systems. Their algorithm applies ideal hard interventions and provides a solution based on constraint satisfaction and conditional independence testing. Mooij et al. [2020] propose a joint causal inference framework to pool interventional datasets to learn the causal graph. Jaber et al. [2020] characterize the interventional MEC and propose a variant of FCI to learn from soft interventional data in causally insufficient systems. The key shortcoming of these methods is that their runtime becomes prohibitive for large graphs.

3 PRELIMINARIES

We introduce some concepts and notations pertinent to causal discovery in causally insufficient systems.

Let $\mathcal{D} \triangleq (\mathbf{W}, \mathbf{E})$ denote a causal graph in which $\mathbf{W}$ represents the set of nodes and $\mathbf{E}$ represents the set of edges. Denote the number of nodes by $p \triangleq |\mathbf{W}|$. We associate the random variable X_i to node i, for $i \in [p] \triangleq \{1, \dots, p\}$, and accordingly define the random vector $X \triangleq (X_1, \dots, X_p)^\top$[1]. We consider a linear SEM, according to which

$$X = B^\top X + \epsilon\,, \tag{1}$$

where $B \in \mathbb{R}^{p \times p}$ is the edge weights matrix in which $B_{i,j} \neq 0$ if and only if $X_i \to X_j \in \mathbf{E}$. The random noise vector $\epsilon \in \mathbb{R}^{p \times 1}$ has zero mean with covariance matrix $\Omega \triangleq \mathsf{diag}(\sigma_1^2, \dots, \sigma_p^2)$. We denote the covariance matrix of X by Σ, and the precision matrix by Θ, which satisfies $\Theta = (I - B)\Omega^{-1}(I - B)^\top$. For the entries of Θ we have

$$\Theta_{i,j} = -\frac{B_{i,j}}{\sigma_j^2} - \frac{B_{j,i}}{\sigma_i^2} + \sum_{k \in \mathsf{ch}(i) \cap \mathsf{ch}(j)} \frac{B_{i,k} B_{j,k}}{\sigma_k^2}\,, \quad \forall i \neq j\,, \tag{2}$$

$$\Theta_{i,i} = \frac{1}{\sigma_i^2} + \sum_{j \in \mathsf{ch}(i)} \sigma_j^{-2} B_{i,j}^2\,, \qquad \forall i \in [p]\,, \tag{3}$$

where $\mathsf{ch}(i)$ denotes the set of children of node $i \in [p]$. In the causal graph $\mathcal{D}$, we have two sets of nodes: a set of observed variables denoted by $\mathbf{V}$, and a set of latent variables denoted by $\mathbf{L}$. Clearly, $\mathbf{V} \cup \mathbf{L} = \mathbf{W}$. The observational data, consequently, is represented by $\{X_i : i \in \mathbf{V}\}$.

From the observational data alone, a DAG with only observed variables can be identified up to its MEC [Verma and Pearl, 1992]. For causally insufficient systems with latent variables $\mathbf{L}$, we can only describe the MEC in terms of a family of graphs called *maximal ancestral graphs* (MAGs), which we formally specify later in this section. The MAG associated with $\mathbf{V}$ represents the pairwise ancestral and confounding relationships among the observed variables $\{X_i : i \in \mathbf{V}\}$ that cannot be made conditionally independent. Therefore, for the true causal graph $\mathcal{D}$, there exists a unique MAG. This MAG cannot be identified uniquely. However, it is possible to recover it up to a family of equivalent MAGs that contains the true one. Next, we describe how a MAG is obtained from a DAG and then proceed to describe the MEC of MAGs and how they are represented.

Mixed Graphs: From a structure learning perspective, causally insufficient systems are often represented by *mixed* graphs. A mixed graph can contain both directed ($\rightarrow$) and bi-directed ($\leftrightarrow$) edges. In our notations, we use $\leftarrow\!\!\circ$ to emphasize that an edge represents either a directed or a bi-directed edge. If there is a directed path from node A to node B, then A is an ancestor of B, and B is a descendant of A. Bi-directed edges create *spouses*, that is, A is a spouse of B if $A \leftrightarrow B$ is present. A node on a path is a *collider* if both of its edges on the path are into the node. A triple $\langle X, Y, Z \rangle$ is an *unshielded collider* if $X \circ\!\!\rightarrow Y \leftarrow\!\!\circ Z$, and X and Z are not adjacent. A path $\langle X, \dots, W, Z, Y \rangle$ is a *discriminating path* for Z if every node between X and Z is a collider on the path, and is also a parent of Y. An *inducing path* relative to $\mathbf{L}$ is a path on $\mathcal{D}$ such that on this path, every non-endpoint node $X \notin \mathbf{L}$ is a collider on the path, and every collider is an ancestor of an endpoint of the path.

Maximal Ancestral Graphs: Consider the causal graph $\mathcal{D} = (\mathbf{V} \cup \mathbf{L}, \mathbf{E})$. A unique mixed graph called the MAG [Richardson and Spirtes, 2002] $\mathcal{M}_{\mathcal{D}}$ over $\mathbf{V}$ has the following three properties: (i) in a MAG, there exists an edge between two nodes if and only if their associated variables

[1] Throughout the paper, we use X_i to represent node $i \in [p]$.

cannot be made conditionally independent (or d-separated) by conditioning on any subset of observed variables in the true $\mathcal{D}$; (ii) if there is an edge in the skeleton that represents the ancestral relationships among the variables in $\mathbf{V}$ in the true $\mathcal{D}$ [Zhang, 2008], then a directed edge is used to represent this edge; and (iii) if there is an edge in a MAG that connects two variables that do not have any ancestral relationship in $\mathcal{D}$, then a bi-directed edge $\leftrightarrow$ is used to represent it. We note that the relationships between DAGs and MAGs are many-to-one, i.e., different DAGs can have the same MAG. Similar to the DAGs, a MAG can be identified only up to a family of MAGs that are Markov equivalent. This Markov equivalence class is represented by a *partial ancestral graph* (PAG).

Markov Equivalence: Two MAGs are Markov equivalent if and only if they have (i) the same adjacencies; (ii) the same unshielded colliders; and (iii) if a path π is a discriminating path for Z in both graphs, then Z is a collider on π in one graph if and only if it is a collider on π in the other graph as well. A PAG represents the MEC of a MAG that can be learned from the observational data. The skeletons of all MAGs in the MEC are identical. Therefore, the PAG has the same skeleton as all members of the MEC. If an edge is oriented as $\rightarrow$ or $\leftrightarrow$, this orientation is fixed for that edge in all MAGs of the MEC. If an edge in a PAG is oriented as $\leftarrow\!\circ$, this implies that there are at least two MAGs in the MEC, such that for the first MAG, this edge is oriented as $\leftrightarrow$ and for the second MAG, this edge is oriented as $\leftarrow$. An edge with circles on both ends means there are three MAGs in the MEC with three distinct orientations $\leftarrow$, $\rightarrow$, and $\leftrightarrow$.

We denote the MAG corresponding to the DAG $\mathcal{D} = (\mathbf{V} \cup \mathbf{L}, \mathbf{E})$ by $\mathcal{M}_{\mathcal{D}}$. Let $\mathsf{pa}(A)$, $\mathsf{ch}(A)$, $\mathsf{sp}(A)$, $\mathsf{an}(A)$, and $\mathsf{de}(A)$ denote the sets of parents, children, spouses, ancestors, and descendants of a node A. We also create the set $\mathsf{ps}(A) = \mathsf{pa}(A) \cup \mathsf{sp}(A)$ to denote the union of parents and spouses of a node A. We denote these relationships with respect to a graph, e.g., $\mathsf{pa}_{\mathcal{D}}(A)$. The subscript is dropped if the specified graph is clear from the context.

4 PROBLEM STATEMENT

Interventions on causal models improve the identifiability of the underlying causal structure. We consider a soft intervention model, which changes the conditional distributions of an intervention target node given its true parents (both observed and unobserved) in the causal DAG $\mathcal{D}$ without completely removing the causal effects of its parents.

Soft Intervention Model. Assume that we have n interventional settings, and denote the collection of the intervention target sets by $\mathcal{I} \triangleq \{\mathbf{I}^{(j)} : j \in [n]\}$. In the j-th setting, the nodes in $\mathbf{I}^{(j)} \subset \mathbf{V}$ are targeted for intervention. Soft interventions in the linear SEM specified in (1) change the conditional distributions of variables $\{X_i : i \in \mathbf{I}^{(j)}\}$. Under these changes (i) the variances of the noise terms $\{\epsilon_i : i \in \mathbf{I}^{(j)}\}$ change, and (ii) the weights connecting the parents of the nodes associated with $\{X_i : i \in \mathbf{I}^{(j)}\}$ in the linear SEM *may* change. In other words, $\{B_{\mathsf{pa}(i),i} : i \in \mathbf{I}^{(j)}\}$, where $B_{\mathsf{pa}(i),i} \triangleq \{B_{u,i} : X_u \in \mathsf{pa}(X_i)\}$, may vary freely. We also note that this formulation can readily work with mean-shift interventions that change the mean of the noise variables (see supplementary material Section D.1 for details).

Post-intervention linear SEMs have new parameters. We denote the linear SEM parameters associated with interventions $\mathbf{I}^{(j)}$ by $B^{(j)}$ and $\Omega^{(j)} \triangleq \mathsf{diag}\left((\sigma_1^{(j)})^2, \ldots, (\sigma_p^{(j)})^2\right)$. Since the noise variance terms change under soft interventions, $\mathbf{I}^{(j)}$ is described as follows:

$$\mathbf{I}^{(j)} \triangleq \{i : i \in \mathbf{V},\ \sigma_i^{(j)} \neq \sigma_i\}\ . \tag{4}$$

A node can be targeted in multiple interventional settings, e.g., $i \in \mathbf{I}^{(j)} \cap \mathbf{I}^{(l)}$. We assume that each target set have different mechanisms such that upon interventions on sets $\mathbf{I}^{(j)}$ and $\mathbf{I}^{(l)}$, we have $\sigma_i^{(j)} \neq \sigma_i^{(l)}$ for all $i \in \mathbf{I}^{(j)} \cup \mathbf{I}^{(l)}$. We note that this assumption is purely for simplicity in the notation, and can be dropped by denoting the exact mechanism applied on a node under each setting.

Identifiability conditions of the causal graphs with unknown soft interventions and the corresponding graphical characterization are established by Jaber et al. [2020]. Importantly, causal graphs with the same observed variables but different latent variables and intervention targets can still belong to the same MEC. We follow the augmented graph construction of Kocaoglu et al. [2019] and Jaber et al. [2020] to represent the MEC's under interventions graphically. First, we construct the augmented graph as follows: for each pair of intervention targets $\mathbf{I}, \mathbf{J} \in \mathcal{I}$, the augmented graph $\mathsf{Aug}_{\mathcal{I}}(\mathcal{D})$ appends the causal graph $\mathcal{D}$ with an auxiliary node and assign directed edges from this node to each node in $\mathbf{H} = \mathbf{I} \cup \mathbf{J}$. We denote the set of these auxiliary nodes by $\mathcal{F}$, and refer to the members of $\mathcal{F}$ as F-nodes. In the example in Fig. 1, we have observational setting $\emptyset$ and intervention target set $\{X\}$, so there is just one pair of target sets. An F-node is created corresponding to this pair, and the edge $F \rightarrow X$ is drawn since X is the only node in the set $\emptyset \cup \{X\}$.

Definition 1 (Augmented Graph) *Consider a causal graph $\mathcal{D} = (\mathbf{V} \cup \mathbf{L}, \mathbf{E})$ and a set of intervention targets $\mathcal{I}$. Define the multiset $\mathcal{H}$ as $\mathcal{H} = \{\mathbf{I} \cup \mathbf{J} : \mathbf{I}, \mathbf{J} \in \mathcal{I}\}$. Given $\mathcal{H}$, generate $h \triangleq |\mathcal{H}|$ nodes $\mathcal{F} \triangleq \{F_i : i \in [h]\}$ and define the augmented graph of $\mathcal{D}$ as $\mathsf{Aug}_{\mathcal{I}}(\mathcal{D}) \triangleq (\mathbf{V} \cup \mathbf{L} \cup \mathcal{F}, \mathbf{E} \cup \mathcal{E})$, where $\mathcal{E} \triangleq \{(F_i, V) : i \in [h],\ V \in \mathbf{H}_i\}$.*

The study in Jaber et al. [2020] shows that the augmented graph exactly represents the separation statements among the random variables in interventional settings. Similar to obtaining a unique MAG from a DAG, a corresponding maximal ancestral graph for the augmented graph is constructed next. In the example in Fig. 1, $F \rightarrow W$ and $Z \rightarrow W$

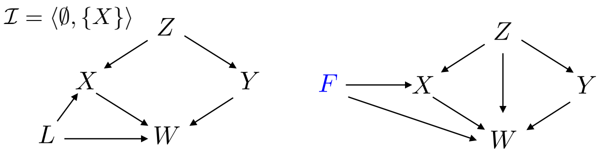

Figure 1: An example of a $\langle \mathcal{D}, \mathcal{I} \rangle$ with $\mathbf{L} = \{L\}$, and the corresponding $\mathcal{I}$-MAG, $\mathcal{M} = \mathsf{MAG}(\mathsf{Aug}_{\mathcal{I}}(\mathcal{D}))$. Note that $F \to X$ is constructed in $\mathsf{Aug}_{\mathcal{I}}(\mathcal{D})$. $F \to W$ edge on $\mathcal{M}$ is due to the inducing path $F \to X \leftarrow L \to W$. Similarly, $Z \to W$ is due to the inducing path $Z \to X \leftarrow L \to W$.

edges are drawn due to the inducing paths existing in the augmented graph $\mathsf{Aug}_{\mathcal{I}}(\mathcal{D})$.

Definition 2 ($\mathcal{I}$-MAG) *Given a causal graph $\mathcal{D} = (\mathbf{V} \cup \mathbf{L}, \mathbf{E})$ and a set of intervention targets $\mathcal{I}$, we define $\mathcal{I}$-MAG to represent the maximal ancestral graph constructed over $\mathbf{V}$ from $\mathsf{Aug}_{\mathcal{I}}(\mathcal{D})$, i.e., $\mathsf{MAG}(\mathsf{Aug}_{\mathcal{I}}(\mathcal{D}))$, and denote its edges by $\mathcal{E}_{\mathcal{I}}$.*

Corresponding to every pair of intervention sets $\mathbf{I}^{(j)}$ and $\mathbf{I}^{(l)}$, define the set $\mathbf{I}_{jl} \triangleq \mathbf{I}^{(j)} \cup \mathbf{I}^{(l)}$. Denote the single F-node associated with $\mathbf{I}^{(j)}$ and $\mathbf{I}^{(l)}$ by $F_{jl} \in \mathcal{F}$, and denote the set of nodes adjacent to F_{jl} in $\mathcal{I}$-MAG by

$$\mathbf{K}_{jl} \triangleq \{i : (F_{jl}, i) \in \mathcal{E}_{\mathcal{I}}\} \; . \qquad (5)$$

We remark that, in causally sufficient systems, $\mathbf{K}_{jl} = \mathbf{I}_{jl}$. However, in the presence of latent variables, one cannot distinguish between the nodes in $\mathbf{K}_{jl} \setminus \mathbf{I}_{jl}$ and $\mathbf{I}_{jl}$ according to the $\mathcal{I}$-MAG. Therefore, we will focus on estimating $\mathbf{K}_{jl}$, which we call the *effective intervention targets*.

We note that the observational setting can be considered as an interventional setting with an empty target set. When there exist more than two interventional settings, there are multiple F-nodes and intervention targets. Accordingly, we denote the set of intervention targets by

$$\mathcal{K} \triangleq \left\{ \mathbf{K}_{jl} : \forall j, l \in [n], \; j \neq l \right\} \; . \qquad (6)$$

Problem Statement. We focus on two estimation problems. In the first problem, we estimate the set of intervention targets $\mathcal{K}$ given the data from linear SEMs with latent variables under soft interventions. We denote the estimate of $\mathcal{K}$ by $\hat{\mathcal{K}}$. Our objective is to design the estimator $\phi : \left(\mathbb{R}^{m \times |\mathbf{V}|}\right)^n \to \left(2^{\mathbf{V}}\right)^n$, in which $|\mathbf{V}|$ denotes the number of the observed variables, n denotes the interventional settings, and m denotes the number of samples in a setting.

In the second problem, based on the estimate $\hat{\mathcal{K}}$, for any set $\mathbf{K} \in \hat{\mathcal{K}}$, we consider the problem of estimating the parents and spouses of $\mathbf{K}$ in the augmented MAG ($\mathcal{I}$-MAG). For any $\mathbf{K} \in \hat{\mathcal{K}}$, we denote the set of parents and spouses of the nodes in $\mathbf{K}$ by $\mathsf{ps}(\mathbf{K})$, and denote its estimate by $\hat{\mathsf{ps}}(\mathbf{K})$. Therefore, our second objective is to design the estimator $\phi_{\mathsf{ps}(\mathbf{K})} : 2^{\mathbf{V}} \to \left(2^{\mathbf{V}}\right)^{|\mathbf{K}|}$. These estimates (i.e., $\hat{\mathcal{K}}$ and $\{\hat{\mathsf{ps}}(\mathbf{K}) : \mathbf{K} \in \hat{\mathcal{K}}\}$) are sufficient to refine the observational PAG to the MEC of the $\mathcal{I}$-MAG. In the rest of the paper, we denote this interventional refinement of observational PAG by ψ-PAG.

5 MAIN RESULTS AND ALGORITHM

Overview. In this section, we provide our theoretical results and our **Pre**cision **D**ifference-based **I**ntervention **T**arget **E**stimato**r** (PreDITEr) algorithm. With scalability as the central objective, we focus on estimating only the effective intervention targets. This is a computationally simpler task compared to learning the causal structure of a DAG, and, consequently, facilitates scalability.

The pivotal idea in our algorithm's design is that soft interventions result in only sparse changes in the precision matrix of the linear SEM. Hence, the precision matrix differences have traces of the identities of the intervention sites. We analytically establish how to use the precision matrix differences between a pair of interventional settings to identify the underlying intervened sites. Upon establishing this property, we then devise an algorithm that successively identifies pairs of intervention settings and estimates the difference between their associated precision matrices. These successive estimates are aggregated to identify the intervention targets. Given the extensive literature on estimating precision matrix differences, we can adopt any generic precision difference estimation (PDE) algorithm to generate the estimates that we need in our algorithm.

Once we estimate the target intervention sites, we also provide an estimate for the set of parents and spouses of each of the nodes deemed to be an intervened node. Theoretically, this information enables the increased identifiability of the causal structure due to the interventions. We start describing the details by introducing the precision difference estimation procedure.

Precision Difference Estimation (PDE). When the difference between two linear SEMs is sparse, the difference of their respective precision matrices will also be sparse. Hence, for the two intervention target sets $\mathbf{I}^{(j)}$ and $\mathbf{I}^{(l)}$, the difference between their precision matrices $\Delta_{jl} \triangleq \Theta^{(j)} - \Theta^{(l)}$ will be sparse. In this paper, we use the algorithm of Jiang et al. [2018] to estimate Δ_{jl}. The algorithm computes sample covariance matrices $\hat{\Sigma}^{(j)}$ and $\hat{\Sigma}^{(l)}$ from the data. Then, it solves the following convex optimization problem with the alternating direction method of multipliers (ADMM):

$$\hat{\Delta}_{jl} = \underset{\Delta_{jl}}{\operatorname{argmin}} \left\{ \frac{1}{2}\mathsf{Tr}(\Delta_{jl}^{\top}\hat{\Sigma}^{(j)}\Delta_{jl}\hat{\Sigma}^{(l)}) - \mathsf{Tr}(\Delta_{jl}(\hat{\Sigma}^{(j)} - \hat{\Sigma}^{(l)})) + \lambda\|\Delta_{jl}\|_1 \right\} , \qquad (7)$$

where λ is a tuning parameter. We note that there exist alternative approaches to PDE [Zhao et al., 2014, Yuan et al., 2017]. Any method that is guaranteed to converge to the correct solution can be used as our PDE subroutine in a modular way. We have chosen the method of Jiang et al. [2018] due to its significant advantage in computational complexity compared to the others ($O(p^3)$ vs. $O(p^4)$). Next, we define the marginal SEM over a subset of observed variables.

Definition 3 (Marginal SEM) *Corresponding to a subset of nodes $S \subseteq \mathbf{V}$, we define (B_S, ϵ_S) as the marginal SEM that characterizes the relationship among the random variables $X_S \triangleq \{i : i \in S\}$. Accordingly, the corresponding precision matrix is denoted by Θ_S. The parametrization of a marginal SEM is given by the following lemma.*

Lemma 1 (Ghoshal et al. [2021]) *Corresponding to a subset $S \subseteq \mathbf{W}$, denote the removed set of nodes by $U \triangleq \mathbf{W} \setminus S$ and define $U_i \triangleq U \cap \mathsf{an}(i)$, for $i \in S$. For $i, j \in S$, we have*

$$\sigma^2_{S,i} = \sigma^4_i \left(\sigma^2_i - B^\top_{U_i,i}[\Theta_{\mathsf{an}(i)}]^{-1}_{U_i,U_i} B_{U_i,i}\right)^{-1}, \tag{8}$$

$$[B_S]_{j,i} = \frac{\sigma^2_{S,i}}{\sigma^2_i}\left(B_{j,i} - B^\top_{U_i,i}[\Theta_{\mathsf{an}(i)}]^{-1}_{U_i,U_i}[\Theta_{\mathsf{an}(i)}]_{U_i,j}\right). \tag{9}$$

Before describing the theoretical results, we need the following faithfulness assumption. This assumption rules out the pathological cases in which the effect of an intervention is canceled by other changes in the system. Faithfulness assumptions are generally needed for successful learning.

Assumption 1 ($\mathcal{I}$-faithfulness) *For any choice of $i, j \in S \subseteq \mathbf{V}$, we have the following properties:*

- *If $\sigma^{(1)}_i \neq \sigma^{(2)}_i$, then $\sigma^{(1)}_{S,i} \neq \sigma^{(2)}_{S,i}$.*
- *If $\sigma^{(1)}_{S,i} \neq \sigma^{(2)}_{S,i}$, then $[\Theta^{(1)}_S]_{i,i} \neq [\Theta^{(2)}_S]_{i,i}$. If further $[B_S]_{j,i} \neq 0$ in either model, then $[\Theta^{(1)}_S]_{i,j} \neq [\Theta^{(2)}_S]_{i,j}$.*

5.1 THEORETICAL RESULTS

For the rest of the discussion, we consider a pair of interventional settings. Without loss of generality, let them be $\mathbf{I}^{(1)}$ and $\mathbf{I}^{(2)}$. Denote the difference in their precision matrices by $\Delta_{12} = \Theta^{(1)} - \Theta^{(2)}$, and the difference in marginal precision matrices for set S by $\Delta_{12_S} = \Theta^{(1)}_S - \Theta^{(2)}_S$. For simplicity in the notation, we denote the corresponding F-node F_{12} by F, $\mathbf{K}_{12}$ by $\mathbf{K}$, Δ_{12} by Δ, and Δ_{12_S} by Δ_S. We also denote the set of *affected nodes* among the observed variables by $S_\Delta \triangleq \{i : [\Delta_{\mathbf{V}}]_{i,i} \neq 0\}$.

Separation Property for Invariance. For a non-intervened node $i \in \mathbf{V} \setminus \mathbf{K}$, there is no edge between F and i in $\mathcal{I}$-MAG. Therefore, there exists a set S that separates F and i, and the conditional probability distribution of X_i is invariant given $S \setminus \{X_i\}$. Then, the conditional mean and variance of X_i, and subsequently $\sigma_{S,i}$, are invariant. Finally, applying the result of Wang et al. [2018], $[\Theta_S]_{i,i} = \sigma^{-2}_{S,i}$ is also invariant. Therefore, the set S that separates F and i yields $[\Delta_S]_{i,i} = 0$ by the definition of Δ_S.

Theorem 1 *Consider an $F \in \mathcal{F}$ and an observed node $V \in \mathbf{V}$ in the augmented MAG ($\mathcal{I}$-MAG). Then, $(F, V) \in \mathcal{E}_\mathcal{I}$ if and only if $\nexists\, S \subseteq \mathbf{V}$ such that $[\Delta_S]_{V,V} = 0$.*

Theorem 1 states the existence of a conditioning set S for any non-intervened node V, that makes the corresponding diagonal entry of the precision matrix invariant. In the following lemma, we show that the ancestors of V within the set of affected nodes S_Δ suffice to separate F and V.

Lemma 2 *For a node $V \in S_\Delta \setminus \mathbf{K}$, consider the set $S = S_\Delta \cap \mathsf{an}(V)$. Diagonal entry corresponding to V in the precision matrix of the marginal SEM over S is invariant, i.e., $[\Delta_S]_{V,V} = 0$.*

Lemma 2 implies that we can eliminate all the non-intervened nodes (i.e., nodes not in the effective intervention target set $\mathbf{K}$) in S_Δ by computing PDE for each subset of S_Δ. Therefore, we can identify $\mathbf{K}$ with $2^{|S_\Delta|}$ number of PDEs. Now that we have a way to recover $\mathbf{K}$, we show how to identify the parents and/or spouses of the intervened nodes. This property will play a critical role in improving the identifiability of the MAGs under interventions.

Lemma 3 *Consider $K \in \mathbf{K}$ and $J \in \mathbf{V} \setminus \mathbf{K}$. If $K \leftarrow\!\circ J$ in $\mathcal{I}$-MAG, there does not exist $S \subseteq S_\Delta$ containing $\{K, J\}$ such that $[\Delta_S]_{K,J} = 0$. On the other hand, if $K \rightarrow J$, or there is no edge between them in $\mathcal{I}$-MAG, there exists a set $S \subseteq S_\Delta$ containing $\{K, J\}$ such that $[\Delta_S]_{K,J} = 0$.*

Lemma 2 and Lemma 3 are sufficient to design our algorithm for learning $\mathcal{K}$.

5.2 LEARNING ALGORITHM

We leverage the results in Lemma 2 and Lemma 3 to learn the intervention targets $\mathcal{K}$ from a tuple of interventional distributions generated by some unknown pair $\langle \mathcal{D}, \mathcal{I} \rangle$. Algorithm 1 presents our main learning algorithm PreDITEr that uses the results to learn $\mathcal{K}$, and subsequently $\mathsf{ps}(\mathbf{K})$ for $\mathbf{K} \in \mathcal{K}$. We briefly describe PreDITEr and the rationale underlying its design.

Algorithm 1 (PreDITEr) takes sample covariance matrices of interventional data as inputs. Since estimating intervention targets $\mathbf{K}$ for each pair of interventional settings is independent, we investigate each pair individually. For each

Algorithm 1 Precision Difference-based Intervention Target Estimator (PreDITEr)

```
1: Input: Observed nodes $\mathbf{V}$, sample covariance matrices $\hat{\Sigma}^{(1)}, \ldots, \hat{\Sigma}^{(n)}$
2: Output. Intervention targets $\mathcal{K}$, and $\mathsf{ps}(K)$ $\forall K \in \mathbf{K}$, $\forall \mathbf{K} \in \mathcal{K}$
3: $\mathcal{K} \leftarrow \emptyset$, $\mathcal{F} \leftarrow \emptyset$
4: for $V \in \mathbf{V}$ do $\mathsf{ps}(V) \leftarrow \emptyset$ end for
5: for all pairs $j, l \in [n]$ do
6:     $\mathcal{F} \leftarrow \mathcal{F} \cup \{F_{jl}\}$, $\mathbf{K}_{jl} \leftarrow \emptyset$
7:     Estimate $\Delta_{jl} \leftarrow$ PDE $(\hat{\Sigma}^{(j)}, \hat{\Sigma}^{(l)})$
8:     $S_\Delta \leftarrow \{V : V \in \mathbf{V},\ [\Delta_{jl}]_{V,V} \neq 0\}$
9:     For all $S \subseteq S_\Delta$, estimate $\Delta_{jl_S} \leftarrow$ PDE $(\hat{\Sigma}^{(j)}_{S,S}, \hat{\Sigma}^{(l)}_{S,S})$
10:    for $V \in \mathbf{V}$ do
11:        if $\nexists\ S \subseteq S_\Delta$, such that $V \in S$, and $[\Delta_S]_{V,V} = 0$ then
12:            $\mathbf{K}_{jl} \leftarrow \mathbf{K}_{jl} \cup \{V\}$
13:        end if
14:    end for
15:    $\mathcal{K} \leftarrow \mathcal{K} \cup \mathbf{K}_{jl}$
16:    for all pairs $K \in \mathbf{K}_{jl}$, $J \in S_\Delta \setminus \mathbf{K}_{jl}$ do
17:        if $\nexists\ S \subseteq S_\Delta$, such that $K, J \in S$, and $[\Delta_S]_{K,J} = 0$ then
18:            $\mathsf{ps}(K) \leftarrow \mathsf{ps}(K) \cup \{J\}$
19:        end if
20:    end for
21: end for
```

Precision Difference Estimation (PDE) $(\hat{\Sigma}^{(j)}, \hat{\Sigma}^{(l)})$

```
1: Estimate $\Delta_{jl} = (\hat{\Sigma}^{(j)})^{-1} - (\hat{\Sigma}^{(l)})^{-1}$ using algorithm of Jiang et al. [2018].
2: Symmetrize $\Delta_{jl}$: set $\Delta_{jl} = (\Delta_{jl} + \Delta_{jl}^\top)/2$.
3: Threshold $\Delta_{jl}$: set $[\Delta_{jl}]_{u,v} = 0$ if $|[\Delta_{jl}]_{u,v}| < \varepsilon$.
4: Return $\Delta_{jl}$
```

pair of interventional distributions (or the corresponding F-node), we first estimate the set of affected nodes S_Δ (lines 7 and 8). Then, we estimate precision difference Δ_S for each subset S of S_Δ. If there does not exist a set S for a node $V \in S_\Delta$ such that $[\Delta_S]_{V,V} = 0$, then by Lemma 2, V is an intervened node and belongs to $\mathbf{K}$ (lines 10-15).

After identifying $\mathbf{K}$, consider a $K \in \mathbf{K}$ and $J \in S_\Delta \setminus \mathbf{K}$. If there does not exist a set S such that $[\Delta_S]_{K,J} = 0$, by Lemma 3, J belongs to $\mathsf{ps}(K)$ (lines 16-20).

Algorithm 1 uses PDE as a subroutine. Hence, the quality of the estimate formed by Algorithm 1 hinges on those of the precision difference estimates. To assess the accuracy of Algorithm 1 in estimating the intervention targets irrespectively of the PDE subroutine used, we provide population-level results. In the following theorem, we establish that Algorithm 1 has perfect estimation if the underlying PDE subroutine performs perfectly. This result allows decoupling the accuracy of Algorithm 1 from that of the PDE subroutine used. In practice, however, PDE subroutines are imperfect, which is imposed by having access to only finite samples. To address the convergence to the correct estimates, we discuss the sample complexity and convergence guarantees of the algorithm of Jiang et al. [2018] in supplementary material Section B.

Theorem 2 *When the covariance estimates are perfect and Assumption 1 holds, Algorithm 1 perfectly estimates the set of effective intervention targets $\mathcal{K}$ under soft interventions with probability* 1. *Furthermore, Algorithm 1 recovers non-intervened parents and/or spouses (i.e.,* $\mathsf{ps}(K)$*) of an intervened node K with probability* 1.

5.3 RECOVERING ψ-MARKOV EQUIVALENCE

Next, we show how we can use the intervention target recovery of Algorithm 1 to refine the observational MEC represented by a PAG to the interventional MEC for soft interventions ψ-PAG. We first review the interventional equivalence characterization approaches in the existing literature.

The ψ-Markov equivalence property, i.e., the conditions for two $\mathcal{I}$-MAGs to be Markov equivalent, is characterized by Jaber et al. [2020, Theorem 1]. For two MAGs $\mathcal{M}_1$ and $\mathcal{M}_2$ to be ψ-Markov equivalent:

- $\mathcal{M}_1$ and $\mathcal{M}_2$ must have the same skeleton.
- $\mathcal{M}_1$ and $\mathcal{M}_2$ must have the same unshielded colliders.
- If a path π is a discriminating path for a node V in both $\mathcal{M}_1$ and $\mathcal{M}_2$, then V is a collider on the path in one graph if and only if it is a collider on the path in the other.

The following theorem builds on the results of Theorem 2 and Lemma 3 to obtain ψ-PAG.

Theorem 3 (ψ-PAG) *Given the PAG for the MAG $\mathcal{M}$, and the results of Algorithm 1, i.e., the sets $\mathcal{K}$,* $\mathsf{ps}(\mathbf{K})$ $\forall \mathbf{K} \in \mathcal{K}$*, we can obtain ψ-PAG of $\mathcal{I}$-MAG.*

6 EMPIRICAL RESULTS

First, we run our PreDITEr algorithm on synthetically generated data from linear SEMs to recover intervention targets. Next, we provide comparisons with the state-of-the-art method. Finally, we apply our method to a biological dataset to illustrate its applicability to real data. [2]

[2] Codebase for reproducing the simulations are available at https://https://github.com/bvarici/uai2022-intervention-estimation-latents

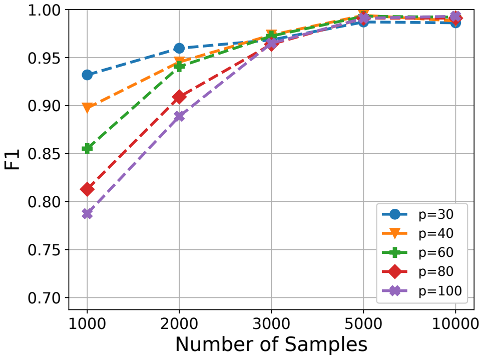

Figure 2: Average F1 scores at estimating **K** for $|\mathbf{L}| = 5$, $|\mathbf{I}| = 5$ intervention targets.

6.1 SYNTHETIC DATA

We test the efficiency of PreDITEr for recovering the intervention targets. We generate 100 realizations of Erdős-Rényi random DAGs with the expected neighborhood size $c = 2$. We consider one interventional setting in addition to the observational one, i.e., $\mathcal{I} = \langle \emptyset, \mathbf{I} \rangle$. Therefore, we are estimating a single target set **K**. For each model, we set the number of latent variables to $|\mathbf{L}| = 5$, and the number of intervened nodes to $|\mathbf{I}| = 5$. The edge weights of the causal model, i.e., the entries of B, are sampled independently at random according to the uniform distribution on $[-1, -0.25] \cup [0.25, 1]$. The additive Gaussian noise terms have distribution $\mathcal{N}(0, I_p)$. The intervention targets are selected randomly from the observed variables **V**. For the intervened nodes $I \in \mathbf{I}$, upon intervention, the variance of the noise term ϵ_I changes to 2.

We run PreDITEr with a varying number of samples on graphs with varying sizes p. Figure 2 illustrates the target recovery performance. Specifically, it shows that our method recovers the intervention target with high F1 scores. We emphasize that PreDITEr can easily process large graphs (e.g., $p = 100$ nodes), and have less than 1 second average runtime for the simulations shown in Figure 2. This scalability is due to its computational complexity of $O(2^{|S_\Delta|})$. Since the size of S_Δ is determined only by the number of intervened nodes and their parents/spouses, our method is not directly affected by the graph size p.

6.2 COMPARISON TO THE RELATED WORK

We compare the scalability and accuracy of PreDITEr to those of two competing methods under various settings: the ψ-FCI algorithm of Jaber et al. [2020] and the FCI-JCI123 algorithm of Mooij et al. [2020]. We note that both of these algorithms solve a more general problem than the linear SEMs we are considering. To the best of our knowledge, there is no algorithm specifically designed for linear SEMs, and these are the only two methods that can be applied to our setting. Therefore, we compare our results to those of these two methods.

Jaber et al. [2020] do not provide simulations for graphs that have more than a few nodes since ψ-FCI requires an exponentially growing number of conditional independence and invariance tests. Mooij et al. [2020] report experiments with larger graphs, and we compare our algorithm to their FCI-JCI123 algorithm. We focus on scalability and provide additional experiments on small graphs and MEC refinement results in supplementary material section D.2.

To enable comparisons under soft interventions, we adopt *mechanism changes* of Mooij et al. [2020], in which a constant offset is added to the intervention targets (see page 53, Section 5.2 for details). We note that this is different from our model of soft interventions and results in slight degrading of the performance of our algorithm, but since we are using FCI-JCI123 as our benchmark, we adopt its setting.

We consider two environments and one intervention target for simplicity of the comparisons. We generate 30 Erdős-Rényi random DAGs. The probability of an edge being present in the random graphs is set to $2/p$ where p is the number of observed and latent variables. We report the precision and recall rates of both algorithms along with their runtimes in Table 1. While both methods have similar performance, there is a significant discrepancy in their runtime. More importantly, the runtime of FCI-JCI123 becomes prohibitive very quickly, even with graphs with as few as 40 nodes. In contrast, PreDITEr has a significantly lower runtime even though the considered setting (i.e., mechanism changes) is not the setting for which it is designed. All simulations are run on a computer with i7-4960HQ, 16GB 1600MHz RAM.

Table 1: Intervention recovery results and median runtime.

Method	p	Precision	Recall	Runtime (s)
PreDITEr	20	1.0	0.83	< 1
FCI-JCI123	20	1.0	1.0	80.9
PreDITEr	30	1.0	0.80	< 1
FCI-JCI123	30	1.0	0.97	318.0
PreDITEr	40	1.0	0.87	< 1
FCI-JCI123	40	0.96	0.96	1301.9

6.3 BIOLOGICAL DATA

We apply the PreDITEr algorithm to a real dataset with data from observational and multiple interventional settings. Since PreDITEr estimates the intervention targets and their corresponding parent-spouse sets for each pair of available settings, we combine the findings from each pair and yield a mixed graph estimate of the associated causal structure.

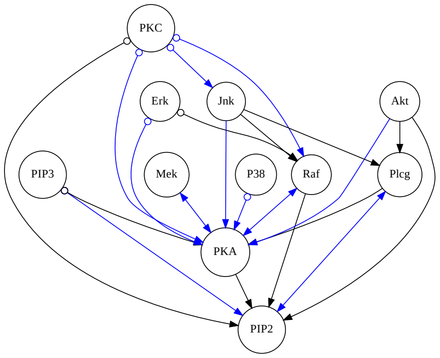

Figure 3: Recovered causal structure using Algorithm 1. Blue edges represent the edges that are in the skeleton of the reference network [Ness et al., 2017].

Protein signaling data. We consider the dataset of Sachs et al. [2005], which is a standard benchmark in causal inference literature. The data is obtained from measurements of the proteins involved in T-4 cell signaling. The protein signaling network consists of 11 nodes. In each interventional setting, various drugs are injected into the cells to inhibit or activate different signaling proteins. The target proteins are considered sites of intervention. Data from observational and five interventional settings are provided. The true ground truth network is not exactly known, and the accepted ground truth has been updated over the years. Notably, it is represented by a DAG without latent confounders. We use the recent version of Ness et al. [2017], which consists of 16 edges, and use the preprocessed real data provided by Squires et al. [2020].

Figure 3 shows the output of our algorithm. For a pair of nodes, if they are found to be in the parent-spouse sets of one another, they must be spouses, and we assign a bi-directed edge. If only one of them lies in the parent-spouse set of the other node, it must be the parent, and we assign a directed edge to them. If we do not have either of the above results, the relationship can be either parent or spouse, and we denote it by $\circ\!\!\rightarrow$ on the graph. The recovered edges that are also present in the skeleton of the ground truth DAG are marked in blue. This result illustrates that even though our algorithm is designed for linear models, it has the potential to be applied to real datasets with non-linear models.

7 CONCLUSION

In this paper, we have considered the problem of estimating intervention targets for causally insufficient systems in linear structural equation models (SEMs). We have assumed a soft intervention model that is more realistic than hard interventions, which eradicate all causal effects on targets. We have shown the usage of invariance of precision matrix entries and proposed an algorithm to identify intervention targets. The algorithm can also be used to refine the observational MEC to interventional MEC for maximal ancestral graphs. Since there exist efficient algorithms for the former, our algorithm provides scalability for the latter as well. We support our analytical results through simulations and compare them with competing methods. The limitation of our approach is that it only applies to linear SEMs. However, we have demonstrated strong performance in real and synthetic datasets, which shows its applicability to other settings.

Acknowledgements

This work was supported by the Rensselaer-IBM AI Research Collaboration (http://airc.rpi.edu), part of the IBM AI Horizons Network (http://ibm.biz/AIHorizons).

References

Pooja Aggarwal, Ajay Gupta, Prateeti Mohapatra, Seema Nagar, Atri Mandal, Qing Wang, and Amit Paradkar. Localization of operational faults in cloud applications by mining causal dependencies in logs using golden signals. In *Proc. International Conference on Service-Oriented Computing*, pages 137–149, December 2020.

Daniel Bernstein, Basil Saeed, Chandler Squires, and Caroline Uhler. Ordering-based causal structure learning in the presence of latent variables. In *Proc. International Conference on Artificial Intelligence and Statistics*, pages 4098–4108, August 2020.

Jasmin Bogatinovski, Sasho Nedelkoski, Alexander Acker, Florian Schmidt, Thorsten Wittkopp, Soeren Becker, Jorge Cardoso, and Odej Kao. Artificial Intelligence for IT Operations Workshop White Paper. *arXiv:2101.06054*, November 2021.

Philippe Brouillard, Sébastien Lachapelle, Alexandre Lacoste, Simon Lacoste-Julien, and Alexandre Drouin. Differentiable causal discovery from interventional data. In *Proc. Advances in Neural Information Processing Systems*, pages 21865–21877, December 2020.

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

Daniel Eaton and Kevin Murphy. Exact Bayesian structure learning from uncertain interventions. In *Proc. International Conference on Artificial Intelligence and Statistics*, pages 107–114, San Juan, Puerto Rico, March 2007.

Yanfang Fu, Jennifer A Foden, Cyd Khayter, Morgan L Maeder, Deepak Reyon, J Keith Joung, and Jeffry D Sander. High-frequency off-target mutagenesis induced by CRISPR-Cas nucleases in human cells. *Nature Biotechnology*, 31(9):822–826, 2013.

AmirEmad Ghassami, Negar Kiyavash, Biwei Huang, and Kun Zhang. Multi-domain causal structure learning in linear systems. *Advances in neural information processing systems*, 31, 2018.

Asish Ghoshal, Kevin Bello, and Jean Honorio. Direct learning with guarantees of the difference dag between structural equation models. *arXiv:1906.12024*, 2021.

Alain Hauser and Peter Bühlmann. Characterization and greedy learning of interventional Markov equivalence classes of directed acyclic graphs. *Journal of Machine Learning Research*, 13(1):2409–2464, 2012.

Biwei Huang, Kun Zhang, Jiji Zhang, Joseph D Ramsey, Ruben Sanchez-Romero, Clark Glymour, and Bernhard Schölkopf. Causal discovery from heterogeneous/nonstationary data. *J. Mach. Learn. Res.*, 21(89): 1–53, 2020.

Amin Jaber, Murat Kocaoglu, Karthikeyan Shanmugam, and Elias Bareinboim. Causal discovery from soft interventions with unknown targets: Characterization and learning. In *Proc. Advances in Neural Information Processing Systems*, pages 9551–9561, December 2020.

Binyan Jiang, Xiangyu Wang, and Chenlei Leng. A direct approach for sparse quadratic discriminant analysis. *The Journal of Machine Learning Research*, 19(1):1098–1134, 2018.

Nan Rosemary Ke, Olexa Bilaniuk, Anirudh Goyal, Stefan Bauer, Hugo Larochelle, Bernhard Schölkopf, Michael C Mozer, Chris Pal, and Yoshua Bengio. Learning neural causal models from unknown interventions. *arXiv:1910.01075*, 2019.

Myunghwan Kim, Roshan Sumbaly, and Sam Shah. Root cause detection in a service-oriented architecture. *ACM SIGMETRICS Performance Evaluation Review*, 41(1): 93–104, 2013.

Murat Kocaoglu, Amin Jaber, Karthikeyan Shanmugam, and Elias Bareinboim. Characterization and learning of causal graphs with latent variables from soft interventions. In *Proc. Advances in Neural Information Processing Systems*, page 14346–14356, Vancouver, Canada, December 2019.

Leonardo Mariani, Cristina Monni, Mauro Pezzé, Oliviero Riganelli, and Rui Xin. Localizing faults in cloud systems. In *2018 IEEE 11th International Conference on Software Testing, Verification and Validation (ICST)*, pages 262–273. IEEE, 2018.

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

Robert Osazuwa Ness, Karen Sachs, Parag Mallick, and Olga Vitek. A Bayesian active learning experimental design for inferring signaling networks. In *Proc. Research in Computational Molecular Biology*, pages 134–156, Hong Kong, May 2017.

Juan Miguel Ogarrio, Peter Spirtes, and Joe Ramsey. A hybrid causal search algorithm for latent variable models. In *Proc. International Conference on Probabilistic Graphical Models*, pages 368–379, Lugano, Switzerland, September 2016.

Judea Pearl. *Causality*. Cambridge University Press, Cambridge, UK, 2009.

Judea Pearl and Dana Mackenzie. *The book of why: The new science of cause and effect*. Basic Books, New York, NY, 2018.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: Foundations and learning algorithms*. The MIT Press, Cambridge, MA, 2017.

Mohsen Pourahmadi. Covariance estimation: The GLM and regularization perspectives. *Statistical Science*, 26 (3):369–387, 2011.

Thomas Richardson and Peter Spirtes. Ancestral graph Markov models. *Annals of Statistics*, 30(4):962–1030, 2002.

Dominik Rothenhäusler, Christina Heinze, Jonas Peters, and Nicolai Meinshausen. Backshift: Learning causal cyclic graphs from unknown shift interventions. In *Proc. Advances in Neural Information Processing Systems*, page 14346–14356, Montreal, Canada, December 2015.

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.

Peter Spirtes, Clark N Glymour, Richard Scheines, and David Heckerman. *Causation, prediction, and search*. MIT Press, Cambridge, MA, 2000.

Chandler Squires, Yuhao Wang, and Caroline Uhler. Permutation-based causal structure learning with unknown intervention targets. In *Proc. Conference on Uncertainty in Artificial Intelligence*, pages 1039–1048, August 2020.

Sofia Triantafillou and Ioannis Tsamardinos. Constraint-based causal discovery from multiple interventions over overlapping variable sets. *Journal of Machine Learning Research*, 16(1):2147–2205, 2015.

Burak Varici, Karthikeyan Shanmugam, Prasanna Sattigeri, and Ali Tajer. Scalable intervention target estimation in linear models. In *Advances in Neural Information Processing Systems*, December 2021.

Thomas Verma and Judea Pearl. An algorithm for deciding if a set of observed independencies has a causal explanation. In *Proc. Conference on Uncertainty in Artificial Intelligence*, pages 323–330, Stanford, CA, July 1992.

Yuhao Wang, Chandler Squires, Anastasiya Belyaeva, and Caroline Uhler. Direct estimation of differences in causal graphs. In *Proc. Advances in Neural Information Processing Systems*, pages 3770–3781, Montreal, Canada, December 2018.

Karren Yang, Abigail Katcoff, and Caroline Uhler. Characterizing and learning equivalence classes of causal DAGs under interventions. In *Proc. International Conference on Machine Learning*, pages 5541–5550, Stockholm, Sweden, July 2018.

Huili Yuan, Ruibin Xi, Chong Chen, and Minghua Deng. Differential network analysis via lasso penalized d-trace loss. *Biometrika*, 104(4):755–770, 2017.

Jiji Zhang. Causal reasoning with ancestral graphs. *Journal of Machine Learning Research*, 9(47):1437–1474, 2008.

Vicky Zhang, Chandler Squires, and Caroline Uhler. Matching a desired causal state via shift interventions. In *Advances in Neural Information Processing Systems*, December 2021.

Sihai Dave Zhao, T Tony Cai, and Hongzhe Li. Direct estimation of differential networks. *Biometrika*, 101(2): 253–268, 2014.

Supplementary Material

A EFFICIENT CONDITIONING SETS

Under certain topological assumptions on the true augmented MAG, we can refine the result of Lemma 2 by identifying a smaller conditioning set. Consider the different tuples of $\langle \mathcal{D}, \mathcal{I} \rangle$ that give rise to the $\mathcal{I}$-MAG of the true model $\langle \mathcal{D}, \mathcal{I} \rangle$. Suppose that there exists a causal graph $\mathcal{D}'$ that has the same set of observed variables, has the same $\mathcal{I}$-MAG under a set of interventions $\mathcal{K}$, and contains each directed edge of the $\mathcal{I}$-MAG. In other words, $\mathcal{M} = \mathsf{MAG}(\mathsf{Aug}_{\mathcal{I}}(\mathcal{D})) = \mathsf{MAG}(\mathsf{Aug}_{\mathcal{K}}(\mathcal{D}'))$ and each directed edge of $\mathcal{M}$ is also an edge in $\mathcal{D}'$. Since our goal is to recover $\mathcal{K}$, we can assume that $\mathcal{D}'$ is the underlying causal model. This assumption leads to the parent-child relationships among the observed variables being the same in $\mathcal{D}'$ and the corresponding $\mathcal{I}$-MAG.

Assumption 2 *There is a causal model $\mathcal{D}'$ such that the set of edges among the observed variables of the causal model $\mathcal{D}'$ is the same as the set of directed edges among the observed variables in $\mathcal{I}$-MAG of the true model $\mathcal{D}$. Furthermore, let* $\mathcal{M} = \mathsf{MAG}(\mathsf{Aug}_{\mathcal{I}}(\mathcal{D})) = \mathsf{MAG}(\mathsf{Aug}_{\mathcal{K}}(\mathcal{D}'))$.

Under the above topological assumption on the augmented MAG, we can identify a (potentially) smaller conditioning set. In a mixed graph, a *c-component* refers to any maximal subset of nodes that are connected using only bi-directed edges. Accordingly, $\mathsf{cc}(A)$ denotes the c-component that contains node A. Let $\mathsf{pa}^+(S)$ denote the union of S and the parents of the elements of S.

Lemma 4 *Under Assumption 2, for a node $V \in S_\Delta \setminus \mathbf{K}$, define* $S_0 = \mathsf{pa}_{\mathcal{M}}{}^+(\mathsf{cc}_{\mathcal{M}}(\mathsf{an}_{\mathcal{M}}(V) \cap \mathbf{K}))$. *Then, for the set $S = S_0 \cap \mathsf{an}(V)$, we have $[\Delta_S]_{V,V} = 0$.*

In other words, we are considering the set of ancestors of V that are either in the same c-component with an intervened ancestor of V, or have children in such a c-component that contains an intervened ancestor of V. The described set is an invariance set for V. For an application in which interventions are distributed over the graph, the set described in Lemma 4 can be much smaller than S_Δ. Therefore, we need to estimate the precision difference matrix for not all subsets of S_Δ, but only subsets up to a certain size.

B TECHNICAL PROOFS

In a mixed graph, if a node A is both an ancestor and spouse of B, this creates an *almost directed cycle*. A mixed graph is called an *ancestral graph* if it does not contain any directed or almost directed cycles. An *inducing path* relative to $\mathbf{L}$ is a path on $\mathcal{D}$ such that on this path, every non-endpoint node $X \notin \mathbf{L}$ is a collider on the path, and every collider is an ancestor of an endpoint of the path. An ancestral graph is *maximal* if every missing edge corresponds to a conditional independence statement. In MAGs, there is an edge between two nodes if and only if there is an inducing path between them [Richardson and Spirtes, 2002]. We use this equivalence between inducing paths and the edges of a MAG repeatedly in the following proofs. A path π between A and B on $\mathcal{D}$ is *active* given $\mathbf{Z}$ if (i) none of the non-colliders on π is in $\mathbf{Z}$, and (ii) for every collider on π, either the collider or a descendant of the collider is in $\mathbf{Z}$.

Proof of Theorem 1. ($\Rightarrow$) Assume that $(F, V) \in \mathcal{E}_{\mathcal{I}}$ which means that there exists an inducing path π between F and V on the augmented graph $\mathsf{Aug}_{\mathcal{I}}(\mathcal{D}) = (\mathcal{F} \cup \mathbf{V} \cup \mathbf{L}, \mathbf{E} \cup \mathcal{E})$. If $V \in \mathbf{I}$, by Assumption 1, $[\Delta_S]_{V,V} \neq 0$ for any set S. Suppose $V \notin \mathbf{I}$.

First, we will show that any latent node on π is an ancestor of V. If a latent node has observed children on both sides on π, it must be an ancestor of V since those observed children are colliders on π and ancestors of V. Suppose $L_1 - L_2$ is the leftmost adjacent pair of latent nodes on π. If $L_1 \leftarrow L_2$, then $L_2 \in \mathsf{an}(V)$. Let $L_1 \rightarrow L_2$ and suppose $L_2 \notin \mathsf{an}(V)$. If $L_2 \rightarrow \ldots L_n \rightarrow V_1$ on π for some observed V_1, then L_2 becomes an ancestor of V. If $L_2 \rightarrow \cdots \rightarrow L_n \leftarrow L_{n+1}$ on π, L_n is a collider and must be an ancestor of V. Hence, $L_2 \in \mathsf{an}(V)$. Similarly, all the latent nodes on π must be an ancestor of V.

For a subset $S \subseteq \mathbf{V}$, let us denote the ancestors of node V out of the set S by U_V. All of the latent nodes on π will be contained in U_V. Without loss of generality, denote these latent nodes by $L_V = \{L_1, \ldots, L_n\}$ according to their order of appearance on π. For any $L_i - L_{i+1}$ pair for $i \in [n-1]$, either L_i and L_{i-1} have a common observed child on π, or one of them is the parent of the other. This is due to the fact that any non-collider node on an inducing path must be in $\mathbf{L}$. Hence, π cannot contain two consecutive observed nodes. Therefore, the corresponding entry in the precision matrix for a $L_i - L_{i+1}$ pair will be non-zero. Subsequently, $[\Theta_{\mathsf{an}(V)}]_{L_V,L_V}$ will have a non-zero 3-wide diagonal band.

Note that π starts with $F \rightarrow I \leftarrow L_1$ where $I \in \mathbf{I} \cap \mathsf{an}(V)$, and ends with $L_n \rightarrow V$. By Assumption 1, diagonal entry corresponding to L_1 in any precision matrix changes with the intervention since $I \in \mathbf{I}$. This means that, when we take the inverse of $[\Theta_{\mathsf{an}(V)}]_{U_V,U_V}$, the diagonal elements associated with L_V will change, i.e., $\left([\Theta_{\mathsf{an}(V)}]^{-1}_{U_V,U_V}\right)_{L_n,L_n}$ is not invariant. Since $B_{L_n,V} \neq 0$, Lemma 1 states that $\sigma_{S,V}$ changes, and subsequently $[\Delta_S]_{V,V} \neq 0$. This result proves that if $(F, V) \in \mathcal{E}_{\mathcal{I}}$, then there does not exist $S \subseteq \mathbf{V}$ such that $[\Delta_S]_{V,V} = 0$.

($\Leftarrow$) We will prove it by contradiction. Assume that there is no inducing path from F to V. Then, F and V are separated given $\mathsf{an}_{\mathcal{M}}(\{F, V\}) \setminus \{F, V\} = \mathsf{an}_{\mathcal{M}}(V) \setminus \{F, V\}$. Then, probability distribution of V is invariant given observed ancestors of V which in turn implies that both conditional mean and variance of the node V is invariant. Applying the results of Pourahmadi [2011] and Wang et al. [2018], for $S = \{\mathsf{an}_{\mathcal{M}}(V)\}$, $[\Theta_S]_{V,V}$ is also invariant and $[\Delta_S]_{V,V} = 0$. This contradicts with the initial statement. Hence, there exists an inducing path from F to V, and $(F, V) \in \mathcal{E}_{\mathcal{I}}$. ■

Proof of Lemma 2. Consider a node $V \in S_\Delta \setminus \mathbf{K}$, and let $S = S_\Delta \cap \mathsf{an}(V)$. For $[\Delta_S]_{V,V} \neq 0$, there needs to be an active path $\pi : \langle F \rightarrow I \ldots V\rangle$ given S. Denote the generic observed variables by V_i, latent variables by L_i, and variables that can be either observed or latent by X_i.

- Suppose I has a tail end on π. If $I \in S$, it blocks the path. If $I \notin S$, then $I \notin \mathsf{an}(V)$. Subsequently, there is a collider on the path from I to V, that is also not in S and neither are its descendants. This collider blocks π.
- Suppose I is a collider on π. If $I \notin S$, neither are its descendants, and I blocks π. Then $I \in \mathsf{an}(V)$. If $I \leftarrow V_1$ on π , $V_1 \in S$ and blocks the path. Therefore, only possible path contains $I \leftarrow L_1$. L_1 is not a parent of V, otherwise $F \rightarrow I \leftarrow L_1 \rightarrow V$ would be an inducing path and contradict with the assumption on V.
- Suppose $F \rightarrow I \leftarrow L_1 \rightarrow V_1$. Note that $V_1 \in S_\Delta$. If $V_1 \rightarrow X_1$ on π, either $V_1 \in \mathsf{an}(V)$ and blocks the path, or a descendant of V_1 is a collider on π and blocks the path. Therefore, suppose $F \rightarrow I \leftarrow L_1 \rightarrow V_1 \leftarrow X_1$. If $V_1 \notin \mathsf{an}(V)$, it blocks π. If $V_1 \leftarrow V_2$, then V_2 is also in $S_\Delta \cap \mathsf{an}(V)$, and blocks π. Therefore, π must start with $F \rightarrow I \leftarrow L_1 \rightarrow V_1 \leftarrow L_2$.
- Note that we have reached another latent node on π, that also cannot be adjacent to V. We can also rule out paths of the form $F \rightarrow I \leftarrow L_1 \rightarrow L_2 \cdots \rightarrow L_n \rightarrow V$ with no observed nodes between I and V, since it would be an inducing path.
- For the remaining cases, suppose $F \rightarrow I \leftarrow L_1 \rightarrow L_2 \cdots \rightarrow L_n \leftarrow X_1$, where $n \geq 1$. If $L_n \notin \mathsf{an}(V)$, it would block π. Therefore, $L_i \in \mathsf{an}(V)$ for $i \in [n]$, and L_n is not adjacent to V. If $L_n \leftarrow V_1$, V_1 is also in S_Δ since $I \leftarrow L_1 \rightarrow \cdots \rightarrow L_n \leftarrow V_1$ path is active given $\mathbf{V}$. Therefore, $V_1 \in S_\Delta \cap \mathsf{an}(V)$ and blocks π. Final case is $L_n \leftarrow L_{n+1}$. Note that we have reached another latent node on π, that also cannot be adjacent to V.

In all cases, we reach a recursive pattern, and cannot reach to V. Therefore, given $S = S_\Delta \cap \mathsf{an}(V)$, there is no active path between F and V. Equivalently, $[\Delta_S]_{V,V} = 0$. ■

Proof of Lemma 3. Consider $K \in \mathbf{K}$, and $J \in \mathbf{V} \setminus \mathbf{K}$. Suppose $K \leftarrow\!\!\circ J$ is in $\mathcal{M}$. If $K \leftarrow J$, by Assumption 1, $[\Delta_S]_{K,J} \neq 0$ for any S. Suppose $K \leftrightarrow J$. Then, there is an inducing path π between K and J. Any non-collider on π must be a latent node. Any collider is an ancestor of at least one of K and J. Therefore, π starts with $K \leftarrow L$ for a $L \in \mathbf{L}$. This active inducing path involves changed parameters due to the intervened K. This leads to $[\Delta_S]_{K,J} \neq 0$ for any S. Let $S = S_\Delta \cap \mathsf{an}(J)$. If $K \rightarrow J$, then $K \in S$ and by Lemma 2, $[\Delta_S]_{Y,Y} = 0$, which also implies the row and column corresponding to Y in $[\Delta_S]$ is zero. Finally, if there is no edge between K and J in $\mathcal{M}$, then taking S the separating set of them gives $[\Delta_S]_{K,J} = 0$. ■

Proof of Lemma 4. Consider a node $V \notin \mathbf{K}$. Recalling the separation property for invariance, if all paths from F to V are blocked given S, then $[\Delta_S]_{V,V} = 0$. Let $S_0 = \mathsf{pa}_{\mathcal{M}}{}^{+}(\mathsf{cc}_{\mathcal{M}}(\mathsf{an}_{\mathcal{M}}(V) \cap \mathbf{K}))$ and $S = S_0 \cap \mathsf{an}(V)$. For $[\Delta_S]_{V,V} \neq 0$, there should be an active path between F and V given S. Suppose that such a path on the augmented graph exists and π is the one among those paths that has the smallest number of colliders on it. Also assume that Assumption 2 holds, in other words, all edges in $\mathcal{M} = \mathsf{MAG}(\mathsf{Aug}(\mathcal{D}))$ are also in $\mathsf{Aug}(\mathcal{D})$.

Let $\pi = \langle F \rightarrow I - \pi_{L_0} - V_1 - \pi_{L_1} - \cdots - V_{n-1} - \pi_{L_{n-1}} - V_n - \pi_{L_n} - V\rangle$, in which, $I \in \mathbf{K}$, and each of the subpaths $\pi_{L_0}, \ldots, \pi_{L_n}$ consists of latent nodes, and $V_1, \ldots, V_n$ are observed nodes on π. Any latent collider must have a descendant on S to not block π. Therefore, any latent collider on π is in $\mathsf{an}(V)$. Each node in S is contained in $\mathsf{pa}_{\mathcal{M}}{}^{+}(\mathsf{cc}_{\mathcal{M}}(I))$ for some $I \in \mathsf{an}(V) \cap \mathbf{I}$. Accordingly, if a node $V_i \in \mathsf{pa}_{\mathcal{M}}{}^{+}(\mathsf{cc}_{\mathcal{M}}(I)) \cap \mathsf{an}(V)$, we will say that $V_i \in S$ *due to I*. We will prove that each of $V_1, \ldots, V_n$ will be in S due to I, which results in an inducing path between F and V, which is a contradiction.

Firstly, if $I \notin \mathsf{an}(V)$, either I or a descendant of I is a collider on π, and blocks π. If $I \in \mathsf{an}(V)$, then $I \in S$ and to not block π, it must be a collider. Note that if $n = 0$, the path becomes $\pi = \langle F \rightarrow I \leftarrow \pi_{L_0} - V\rangle$. Any latent collider on π_{L_0} must have a descendant on S to not block π. Therefore, any such latent collider is in $\mathsf{an}(Y)$ and π becomes an inducing path, which contradicts the assumption on V.

Next, we are not interested in non-collider latent nodes on π_{L_0} since they cannot block π. Denote the latent colliders on π_{L_0} by $L_{c_1}, \ldots, L_{c_m}$. Denote the *eldest* observed descendant of L_{c_i} within $\mathsf{an}(V)$ by V_{c_i} for $i \in [m]$. In other words, $L_{c_i} \rightarrow \cdots \rightarrow V_{c_i} \rightarrow \cdots \rightarrow V$ and there is no observed nodes between L_{c_i} and V_{c_i}. In this case, for the subpath $\langle I - V_{c_1} - V_{c_2} - \ldots V_{c_m} - V_1\rangle$, there is an inducing path that consists of non-collider latent nodes between any adjacent pair. Therefore, this subpath will be contained in $\mathcal{M}$.

Now, we will show that $V_1 \in S$ due to I. First, we will show that $V_{c_1}, \ldots, V_{c_m}$ are in S due to I. Suppose $I \rightarrow V_{c_1} \in \mathcal{M}$. Then, there is an inducing path from F to V_{c_1} passing through L_{c_1}, and we have $F \rightarrow V_{c_1}$. As a result, the path $\langle F \rightarrow V_{c_1} \leftarrow \cdots \leftarrow L_{c_1} \ldots L_{c_2} \ldots V_1 \ldots V\rangle$ is active, and contains one fewer collider than π, which contradicts with the assumption on π. We will use this *shorter path* contradiction repeatedly in the rest of the proof. Subsequently, we have $I \leftarrow\!\!\circ V_{c_1} \in \mathcal{M}$. Let us consider the $V_{c_1} - V_{c_2}$ edge next.

- Suppose $V_{c_1} \rightarrow V_{c_2} \in \mathcal{M}$, then $I - V_{c_2} \in \mathcal{M}$ due to inducing path $\langle I \ldots L_{c_1} \rightarrow \ldots V_{c_1} \rightarrow V_{c_2}\rangle$. Suppose $I \leftrightarrow V_{c_1}$. $I \leftarrow V_{c_2}$ would create an almost directed cycle, $I \rightarrow V_{c_2}$ would create $F \rightarrow V_{c_2}$ due to inducing path $\langle F \rightarrow I \ldots L_{c_1} \ldots L_{c_2} \rightarrow \cdots \rightarrow V_{c_2}\rangle$. This implies a shorter active path than π with fewer number of colliders starting with $F \rightarrow Z_{c_2} \leftarrow \cdots \leftarrow L_{c_2}$. Therefore, $I \leftrightarrow V_{c_2}$ and $V_{c_2} \in S$ due to I. In the other case, suppose $I \leftarrow V_{c_1}$. If $I \rightarrow V_{c_2}$, there is an inducing path for $F \rightarrow V_{c_2}$, and subsequently, a shorter path. Therefore, it should be $I \leftarrow\!\circ V_{c_2}$, and $V_{c_2} \in S$ due to I.

- Suppose $V_{c_1} \leftarrow\!\circ V_{c_2} \in \mathcal{M}$. If $I \leftrightarrow V_{c_1}$, then $V_{c_2} \in S$ due to I. If $I \leftarrow V_{c_1}$, either $I \leftarrow V_{c_2}$ which results in a shorter path, or $X \leftarrow\!\circ V_{c_2}$. Therefore, $V_{c_2} \in S$ due to I.

Now, suppose that $V_{c_1}, \ldots, V_{c_t}$ are in $\mathsf{pa}_{\mathcal{M}}^{+}(\mathsf{cc}_{\mathcal{M}}(I))$. We will show that $V_{c_{t+1}}$ is also in $\mathsf{pa}_{\mathcal{M}}^{+}(\mathsf{cc}_{\mathcal{M}}(I))$.

- Suppose $V_{c_t} \in \mathsf{cc}_{\mathcal{M}}(I)$. For convenience of the notation, denote the nodes on the path $X \leftrightarrow \ldots \leftrightarrow V_{c_t}$ in $\mathcal{M}$ by $W_1, \ldots, W_m$ in which $\{W_1, \ldots, W_m\} \subseteq \{V_{c_1}, \ldots, V_{c_t}\}$, and let $W_{m+1} = V_{c_{n+1}}$. Note that bi-directed edges are only in $\mathcal{M}$, and we previously defined c-components in MAGs accordingly. If $W_m \leftarrow\!\circ W_{m+1}$, then $W_{m+1} \in \mathsf{pa}_{\mathcal{M}}^{+}(\mathsf{cc}_{\mathcal{M}}(I))$. Else, $W_m \rightarrow W_{m+1} \in \mathcal{M}$ and there is an inducing path between W_{m-1} and W_{m+1}. If $W_{m-1} \leftarrow\!\circ W_{m+1}$, then $W_{m+1} \in \mathsf{pa}_{\mathcal{M}}^{+}(\mathsf{cc}_{\mathcal{M}}(I))$. We continue recursively until reaching an edge $W_i \leftarrow\!\circ W_{m+1}$, which results in $W_{m+1} \in \mathsf{pa}_{\mathcal{M}}^{+}(\mathsf{cc}_{\mathcal{M}}(I))$. If there does not exist such $W_i \leftarrow\!\circ W_{m+1}$, then all of $X, W_1, \ldots, W_m$ are ancestors of W_{m+1}. This final case is invalid since it results in $F \rightarrow W_{m+1}$ that leads to a shorter path than π. Therefore, W_{m+1}, or $V_{c_{t+1}}$, is in $\mathsf{pa}_{\mathcal{M}}^{+}(\mathsf{cc}_{\mathcal{M}}(I))$.
- Suppose $V_{c_t} \notin \mathsf{cc}_{\mathcal{M}}(I)$ but $V_{c_t} \in \mathsf{pa}_{\mathcal{M}}^{+}(\mathsf{cc}_{\mathcal{M}}(I))$. Again, for the sake of convenience, we denote the path between I and V_{c_t} by $I \leftrightarrow W_1 \leftrightarrow \cdots \leftrightarrow W_{m-1} \leftarrow W_m - W_{m+1}$ in which $\{W_1, \ldots, W_m\} \subseteq \{V_{c_1}, \ldots, V_{c_t}\}$, $W_m = V_{c_t}$, and $W_{m+1} = Z_{c_{t+1}}$. There is an inducing path between W_{m-1} and W_{m+1}. If $W_m \leftarrow W_{m+1}$, then $W_{m-1} \leftarrow W_{m+1}$ and $W_{m+1} \in \mathsf{pa}_{\mathcal{M}}^{+}(\mathsf{cc}_{\mathcal{M}}(X))$. If $W_m \leftrightarrow W_{m+1}$, we have $W_{m-1} \leftarrow\!\circ W_{m+1}$ which results in $W_{m+1} \in \mathsf{pa}_{\mathcal{M}}^{+}(\mathsf{cc}_{\mathcal{M}}(X))$. Finally, suppose $W_m \rightarrow W_{m+1}$. If $W_{m-1} \leftarrow\!\circ W_{m+1}$, we have $W_{m+1} \in \mathsf{pa}_{\mathcal{M}}^{+}(\mathsf{cc}_{\mathcal{M}}(I))$. Otherwise, we have $W_{m-1} \rightarrow W_{m+1}$. We continue recursively until reaching an edge $W_i \leftarrow\!\circ W_{m+1}$ or $I \leftarrow\!\circ W_{m+1}$, which makes $W_{m+1} \in \mathsf{pa}_{\mathcal{M}}^{+}(\mathsf{cc}_{\mathcal{M}}(I))$. If there does not exist such an edge, then all of $X, W_1, \ldots, W_m$ are ancestors of W_{m+1}. This final case is invalid since it results in $F \rightarrow W_{m+1}$ that leads to a shorter path than π. Therefore, W_{m+1}, or $V_{c_{t+1}}$, is in $\mathsf{pa}_{\mathcal{M}}^{+}(\mathsf{cc}_{\mathcal{M}}(I))$.

By induction, we conclude that all of the nodes $V_{c_1}, \ldots, V_{c_m}$, and V_1 are in $\mathsf{pa}_{\mathcal{M}}^{+}(\mathsf{cc}_{\mathcal{M}}(I))$. Now, if $V_1 \notin \mathsf{an}(V)$, then V_1 is not a collider for π to be active. Then, there exists a descendant of V_1 on the path, that is a collider, and also not in $\mathsf{an}(V)$. This collider blocks π. Therefore, V_1 is a collider and an ancestor of V. Hence, $V_1 \in \mathsf{pa}_{\mathcal{M}}^{+}(\mathsf{cc}_{\mathcal{M}}(I)) \cap \mathsf{an}(V)$ and $V_1 \in S$.

Now suppose that $V_1, V_2, \ldots, V_t$ are in $\mathsf{pa}_{\mathcal{M}}^{+}(\mathsf{cc}_{\mathcal{M}}(X)) \cap \mathsf{an}(V)$, and consider V_{t+1}. With a slight abuse of notation, by using previously used auxiliary node names, denote the latent colliders on the path L_t by $L_{c_1}, \ldots, L_{c_m}$. Similarly, denote their eldest observed descendants among $\mathsf{an}(V)$ by V_{c_i} for $i \in [m]$. Similarly, path $V_t - V_{c_1} - \cdots - V_{c_m} - V_{t+1}$ will be contained in $\mathcal{M}$.

- Suppose $V_t \in \mathsf{cc}_{\mathcal{M}}(I)$. If $V_t \leftarrow\!\circ V_{c_1}$, then $V_{c_1} \in \mathsf{pa}_{\mathcal{M}}^{+}(\mathsf{cc}_{\mathcal{M}}(X))$. Else, we have $V_t \rightarrow V_{c_1}$. Similarly to the steps of the proof for $V_1 \in S$, we find either $V_{c_1} \in \mathsf{pa}_{\mathcal{M}}^{+}(\mathsf{cc}_{\mathcal{M}}(I))$, or an active path shorter than π exists, which is invalid.
- Suppose $V_t \notin \mathsf{cc}_{\mathcal{M}}(I)$ but $V_t \in \mathsf{pa}_{\mathcal{M}}^{+}(\mathsf{cc}_{\mathcal{M}}(I))$. By following the same steps, we obtain either $V_{t+1} \in \mathsf{pa}_{\mathcal{M}}^{+}(\mathsf{cc}_{\mathcal{M}}(I))$, or reach an invalid path.

By induction, we conclude that all of the nodes $V_1, \ldots, V_n$ are in $\mathsf{pa}_{\mathcal{M}}^{+}(\mathsf{cc}_{\mathcal{M}}(I)) \cap \mathsf{an}(V)$. Since this results implies that π is an inducing path, which contradicts with the assumption, there does not exist such an active path between F and V given $S = \mathsf{pa}_{\mathcal{M}}^{+}(\mathsf{cc}_{\mathcal{M}}(\mathsf{an}(V) \cap \mathbf{K})) \cap \mathsf{an}(V)$, and $[\Delta_S]_{V,V} = 0$. ■

Proof of Theorem 2. The theorem directly follows from Theorem 1, Lemma 2, and Lemma 3 provided that we can estimate the precision difference correctly from the correct covariance estimates. Any node J that is in $\mathbf{K}$, or in $\mathsf{ps}(K)$ for some $K \in \mathbf{K}$ will be in S_Δ since there will be an active path in one of the forms: (i) $F \rightarrow J$, (ii) $F \rightarrow K \leftarrow J$, or (iii) $F \rightarrow K \leftarrow L \rightarrow J$. By Theorem 1 and Lemma 2, for any $J \in S_\Delta \setminus \mathbf{K}$, we would find a set $S \subset S_\Delta$ that allows us to identify J as non-intervened, and the remaining nodes as intervened. Similarly, by Lemma 3, we can identify $\mathsf{ps}(K)$ set through checking Δ_S for every S subset of S_Δ. ■

PDE algorithm of Jiang et al. [2018]. Let d denote the maximum degree of an intervened node, $\Gamma \triangleq \Sigma^{(l)} \otimes \Sigma^{(j)}$, and define $M \triangleq \max\{\|\Sigma^{(j)}\|_\infty, \|\Sigma^{(l)}\|_\infty \}$, $M_\Sigma \triangleq \max\{\|\Sigma^{(j)}\|_{1,\infty}, \|\Sigma^{(l)}\|_{1,\infty}\}$, $M_{\Gamma,\Gamma^T} \triangleq \max\{\|\Gamma_{S,S}\|_{1,\infty}, \|\Gamma^T_{S,S}\|_{1,\infty}\}$, where S is the support of $(\Sigma^{(l)})^{-1} - (\Sigma^{(j)})^{-1}$. Theorem 1 of Jiang et al. [2018] states that, if irrepresentability condition holds, i.e. $\alpha \triangleq 1 - \max_{e \notin \text{supp}} |\Gamma_{e,\text{supp}}\Gamma^{-1}_{\text{supp},\text{supp}}|_1 > 0$, and λ is set to $Cd^2 \sqrt{\frac{\log p}{n}}$ for some constant $C > 0$, then their PDE algorithm has sample complexity of $O(M_\Sigma M_{\Gamma,\Gamma^T} d^4 \log p)$. Under these conditions, we can derive a similar sample complexity result to that of Theorem 3 of Varici et al. [2021]. However, our focus in this paper is leveraging the precision differences to estimate intervention targets in the presence of latent variables. The discussion of the algorithm of Jiang et al. [2018] is intended to emphasize that it converges to the perfect estimates with sample complexity that scales with $\log p$.

Proof of Theorem 3. Suppose we are given the partial ancestral graph (PAG) corresponding to the observational MEC for $\mathsf{MAG}(\mathcal{D})$, and skeleton of the MAG remains the same after the interventions. Theorem 2 guarantees to recover the intervened nodes $\mathbf{K}$ (or F-adjacent nodes). Lemma 3 guarantees to recover all $J \circ\!\!\rightarrow K$ pairs for $K \in \mathbf{K}$ and $J \in \mathbf{V} \setminus \mathbf{K}$. Each such pair creates a new unshielded collider $F \rightarrow K \leftarrow\!\!\circ J$. Subsequently, we distinguish between the $\mathcal{I}$-MAGs that are identifiable through the second condition.

Next, consider a discriminating path π. If π does not start with $F \rightarrow K$, its effect to the equivalence class is in the PAG before the interventions. Suppose $\pi = \langle F \rightarrow K \leftrightarrow V_1 \leftrightarrow \cdots \leftrightarrow V_n \leftarrow\!\!\circ Y \circ\!\!\rightarrow Z$ is a discriminating path for Y, where $Z \in \mathbf{V} \setminus \mathbf{K}$, and $K, V_1, \ldots, V_n$ are in $\mathsf{pa}_{\mathcal{M}}(Z)$. Note that if Y is a collider on the path, $V_n \leftrightarrow Y \leftrightarrow Z$. Otherwise, $Y \rightarrow Z$. This means that any set S that makes Z invariant must contain $K, V_1, \ldots, V_n$. Moreover, if Y is not a collider, S also has to contain Y. If Y is a collider, there exists a set S that does not contain Y and makes Z invariant. Since we can find all sets S that make Z invariant, we can determine whether Y is a collider on a discriminating path or not. Subsequently, we can distinguish two $\mathcal{I}$-MAGs that are identifiable through the third condition, and refine PAG to ψ-PAG. ■

C IMPLEMENTATION DETAILS

We make a remark here regarding the hyperparameters in the implementation of Algorithm 1. The PDE algorithm uses an ℓ_1 regularization parameter λ. We use two hyperparameters: λ_1 for the estimation of S_Δ, and λ_2 for all the other subsets S. We note that generally λ_1 is supposed to be greater than λ_2 since $\Delta_{\mathbf{V}}$ over $\mathbf{V}$ is expected to be sparser than Δ_S of marginal SEMs. We have used $\{\lambda_1 = 0.2, \lambda_2 = 0.1\}$ in Figure 2, $\{\lambda_1 = 0.3, \lambda_2 = 0.2\}$ in Figure 3, and $\{\lambda = 0.1, \lambda_2 = 0.05\}$ in Figure 4. We note that the algorithm performs well without the need for extensive hyperparameter tuning.

Comparison to ψ-FCI. To our knowledge, ψ-FCI algorithm of Jaber et al. [2020] is the most general method that can learn causal structure from interventions in the presence of latent nodes. Even if we restrict the goal of ψ-FCI to learning only the edges $F \rightarrow K$, which is the focus of our paper, it still requires $O(2^{|\mathbf{V}|})$ tests, whereas we perform $O(2^{|S_\Delta|})$ PDEs, which is significantly smaller. This improvement allows us to work in conjunction with faster observational algorithms for high-dimensional settings.

D ADDITIONAL EXPERIMENTS

In this section, we conduct more detailed simulations to expand the discussions in Section 6.

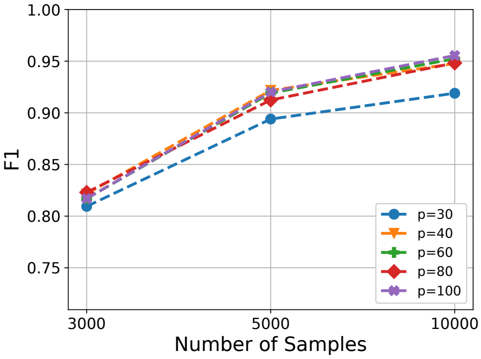

Figure 4: Average F1 scores at estimating $\mathbf{K}$ for $|\mathbf{L}| = 5$, $|\mathbf{I}| = 5$ intervention targets for shift interventions.

D.1 SHIFT INTERVENTIONS

We note that mean shift interventions are commonly used in causal inference from intervention data [Squires et al., 2020, Zhang et al., 2021]. Hence, we conduct experiments in this setting. In our soft intervention formulation, we assumed a change in variance where the means of noise terms remain the same for simplicity. We can slightly change the model to allow the mean shift interventions as follows: Suppose $\epsilon_i \sim \mathcal{N}(\mu_i, \sigma_i)$, and denote $c_i^2 \triangleq \mu_i^2 + \sigma_i^2$, and $C = \mathsf{diag}(c_1^2, \ldots, c_p^2)$. Then, we have

$$A \triangleq \left(\mathbb{E}[XX^T]\right)^{-1} = (I - B)C^{-1}(I - B^T)\,, \tag{10}$$

$$A_{i,i} = \frac{1}{c_i^2} + \sum_{k \in \mathsf{ch}(i)} \frac{B_{i,k}^2}{c_k^2}\,. \tag{11}$$

In this more general case, the only change from Eq. (3) is that we use the inverse of correlation matrix instead of the inverse of the covariance matrix. Therefore, the difference estimation becomes $(A^{(1)})^{-1} - (A^{(2)}))^{-1}$. Via this slight modification, we are able to identify the intervention targets in which there is a change in $c_i^2 = \mu_i^2 + \sigma_i^2$.

Under the same conditions as in Section 6.1, for intervened nodes $I \in \mathbf{I}$, we shift the mean of the noise ϵ_I from 0 to 1 while leaving its variance unchanged. Figure 4 illustrates that PreDITEr recovers the intervention target with high F1 scores for shift interventions.

D.2 COMPARISON TO RELATED WORK

In this section, we extend the simulations in Section 6.2. We first study a small graph in which both of ψ-FCI and FCI-JCI123 can be run.

D.2.1 Small Graphs

We first describe the experimental setting that allows us to compare all three methods. Jaber et al. [2020] compare the

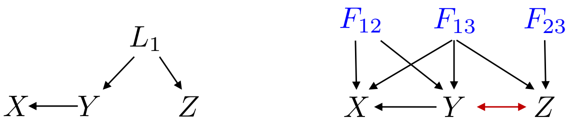

Figure 5: A causal graph with $\mathbf{L} = \{L_1\}$ and its $\mathcal{I}$-MAG for $\mathcal{I} = \{\emptyset, \{X, Y\}, \{X, Y, Z\}\}$.

results of ψ-FCI to that of FCI-JCI123, in Section D.2 of their supplementary document. They only work with discrete data and do not comment on the feasibility of working with Gaussian data. We note that our characterization of the graphs is closer to Jaber et al. [2020] than to Mooij et al. [2020]. On the other hand, our formulation is for normal distributions. To find a middle ground, we generate data from normal distributions with increased variance soft interventions, discretize the data into 5 levels, and run ψ-FCI, FCI-JCI123 and PreDITEr.

Jaber et al. [2020] use the following simple graph to illustrate the difference between ψ-Markov formulation and JCI method:

$$X \leftarrow Y \leftarrow L_1 \rightarrow Z.$$

The observational MEC of this graph consists of 5 graphs: (i) $X \leftarrow Y \leftrightarrow Z$, (ii) $X \leftarrow Y \leftarrow Z$, (iii) $X \leftarrow Y \rightarrow Z$, (iv) $X \rightarrow Y \rightarrow Z$, (v) $X \leftrightarrow Y \rightarrow Z$. Consider the interventions $\mathcal{I} = \{\emptyset, \{X, Y\}, \{X, Y, Z\}\}$. The corresponding $\mathcal{I}$-MAG is given at Fig. 5. We simulate this graph 20 times with different edge weights, and generated 20.000 samples for each set of weights. The results in Table 2 show that all methods have high precision whereas PreDITEr and ψ-FCI has higher recall rates than FCI-JCI123.

Table 2: Intervention targets recovery in Fig. 5

Method	Precision	Recall
PreDITEr	1	0.96
ψ-FCI	1	0.97
FCI-JCI123	1	0.75

The important difference between ψ-Markov formulation and JCI method is that edge $Y \leftrightarrow Z$ can only be identified in the former one. Correctly recovering edge $Y \leftrightarrow Z$ would eliminate the other possible graphs in the observational MEC and immediately give us the correct underlying graph: $X \leftarrow Y \leftrightarrow Z$. The results of recovering edge $Y \leftrightarrow Z$ is given at Table 3.

Table 3: Recovery of edge $Y \leftrightarrow Z$ in Fig. 5

Method	Skeleton	Orientation
PreDITEr	0.6	0.3
ψ-FCI	0	0
FCI-JCI123	0.15	0

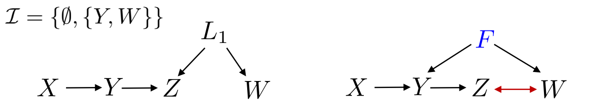

Figure 6: A causal graph with $\mathbf{L} = \{L_1\}$ and its $\mathcal{I}$-MAG for $\mathcal{I} = \{\emptyset, \{Y, W\}\}$.

Poor results of ψ-FCI and FCI-JCI123 are striking. Theoretically, FCI-JCI123 can only identify the edge up to $Y \leftarrow\!\circ Z$. Therefore, the best FCI-JCI123 can do is recovering the skeleton of the edge. At this point, we want to refer to Fig. 6 of Jaber et al. [2020]. There are two main differences: (i) We use discretized Gaussian data while experiments in their paper were for SEMs with logistic model generating binary data, (ii) they needed as much as 200.000 samples to report their results. Furthermore, in their Fig. 6b. and 6c., the best results out of 30 trials are reported. Comparing with the intervention recovery task, we observe that ψ-FCI is able to find intervention targets on a small number of variables but struggle for recovering the edges among observed variables when the number of samples is not extremely large and the discrete nature of the data is not as simple as binary.

D.2.2 Causally Sufficient Models

Varici et al. [2021] recently proposed an algorithm for causally sufficient models that leverages precision matrix differences and decomposes the affected nodes into ancestral equivalence classes. Their computational complexity scales exponential in the size of the largest class. In the worst case, the largest class size can be $|S_\Delta|$. In this section, we give an example to show that their algorithm does not work in the presence of latent variables.

Consider the causal graph in Fig. 6. The algorithm of Varici et al. [2021] starts by identifying $S_\Delta = \{X, Y, Z, W\}$, and non-intervened source nodes $J_0 = \{X\}$. Subsequently, $\mathcal{A}_1 = \{W\}$ and $\mathcal{A}_2 = \{Y, Z\}$ since W does not share an ancestor with X where Y and Z do. Then, their algorithm estimates precision differences for only $\{W, Z\}$ and $\{W, Y, Z\}$ to decide if $Z \in \mathcal{I}$, and declare Z is intervened. However, PDE for $\{Y, Z\}$ would reveal that Z is not intervened. This discrepancy is caused by missing the non-intervened node L_1 from J_0. Therefore, their algorithm that promises scalability for causally sufficient models, does not work when there are latent variables.

D.2.3 Larger Models

In this section, we compare PreDITEr with FCI-JCI123 in our model of soft intervention. As in Section 6.1, the variance of the noise ϵ_I changes to 2 for the intervened nodes $I \in \mathbf{I}$. We consider two environments, and three intervened nodes. We generate 30 random DAGs, and the probability of the presence of an edge in the random graphs is set to

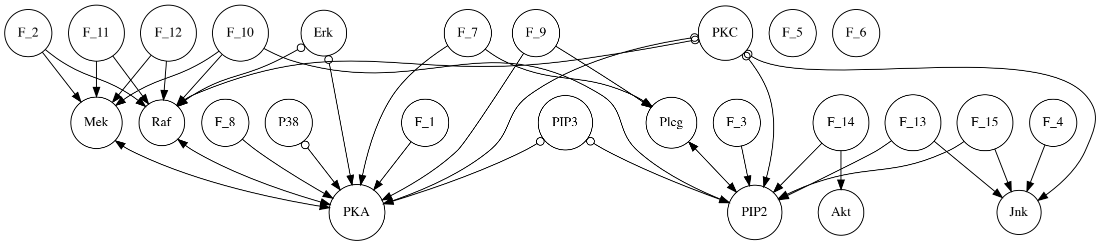

Figure 7: Recovered causal structure of protein signaling network including augmented F-nodes.

$2/p$. We note that these values are close to the setting of Mooij et al. [2020]. The results are reported in Table 4.

Table 4: Intervention recovery results and average runtime.

Method	p	Precision	Recall	Runtime (s)
PreDITEr	10	0.99	0.95	< 1
FCI-JCI123	10	1.0	0.87	3.4
PreDITEr	20	0.98	0.91	< 1
FCI-JCI123	20	1.0	0.92	6.5
PreDITEr	30	0.97	0.97	< 1
FCI-JCI123	30	0.99	0.98	51.3
PreDITEr	40	0.94	0.96	< 1
FCI-JCI123	40	0.99	0.95	192.5

Both algorithms have similar precision and recall rates at intervention recovery. On the other hand, average runtime of our algorithm remains under 1 second for graphs as large as $p = 50$ whereas the runtime of FCI-JCI123 grows rapidly.

D.3 PROTEIN SIGNALING DATA DETAILS

We extend the discussion for real data experiments in Section 6.3. We highlight some properties of the protein signaling network of Sachs et al. [2005], interpret and compare our results with that of the competing methods.

We re-iterate that the true network for the protein signaling is not known. Under this uncertainty about the true network, learning the intervention targets helps us to understand the underlying network better. Mooij et al. [2020] illustrate the need for interventional methods for this network in their Section 5.8. For instance, in one of the settings, it is observed that both Raf and Mek proteins are affected when only Raf was targeted. This effect cannot be explained through the consensus network or the observational data. However, in $\mathcal{I}$-MAG formulation, we find that there exist F-nodes, representing intervention settings, that are adjacent to both Raf and Mek (see Fig. 7). Similar results can be found in Fig. 8 of the supplementary document of Jaber et al. [2020]. Therefore, learning the soft interventions explain some properties of the protein signaling network that the observational data cannot address.

Table 5: Skeleton recovery of protein signaling network

	True	**False**	**False**
Method	**Positives**	**Positives**	**Negatives**
PreDITEr	11	8	5
ψ-FCI	7	2	9
FCI-JCI123	5	1	11

None of the algorithms are thoroughly optimized for the best results. For instance, changing the parameters of the CI tests for Jaber et al. [2020] and Mooij et al. [2020], or changing the regularization of PDE for our method directly affect the final results. What we aim here is to provide conceptual understanding. Since the true network is unknown, it is difficult to comment on the edge orientation results. Therefore, we report skeleton recovery results with respect to the consensus network of Ness et al. [2017]. The estimated networks for competing methods are taken from Fig. 39 of Mooij et al. [2020] and Fig. 8 of Jaber et al. [2020].