TECHNICAL REPORT
R-67
First version: Jun, 2020
Last revision: Nov, 2020

Causal Discovery from Soft Interventions with Unknown Targets: Characterization and Learning

Amin Jaber
Department of Computer Science
Purdue University, USA
jaber0@purdue.edu

Murat Kocaoglu
MIT-IBM Watson AI Lab
IBM Research MA, USA
murat@ibm.com

Karthikeyan Shanmugam
MIT-IBM Watson AI Lab
IBM Research NY, USA
karthikeyan.shanmugam2@ibm.com

Elias Bareinboim
Department of Computer Science
Columbia University, USA
eb@cs.columbia.edu

Abstract

One fundamental problem in the empirical sciences is of reconstructing the causal structure that underlies a phenomenon of interest through observation and experimentation. While there exists a plethora of methods capable of learning the equivalence class of causal structures that are compatible with observations, it is less well-understood how to systematically combine observations and experiments to reconstruct the underlying structure. In this paper, we investigate the task of structural learning in non-Markovian systems (i.e., when latent variables affect more than one observable) from a combination of observational and soft experimental data when the interventional targets are unknown. Using causal invariances found across the collection of observational and interventional distributions (not only conditional independences), we define a property called Ψ-Markov that connects these distributions to a pair consisting of (1) a causal graph $\mathcal{D}$ and (2) a set of interventional targets $\mathcal{I}$. Building on this property, our main contributions are two-fold: First, we provide a graphical characterization that allows one to test whether two causal graphs with possibly different sets of interventional targets belong to the same Ψ-Markov equivalence class. Second, we develop an algorithm capable of harnessing the collection of data to learn the corresponding equivalence class. We then prove that this algorithm is sound and complete, in the sense that it is the most informative in the sample limit, i.e., it discovers as many tails and arrowheads as can be oriented within a Ψ-Markov equivalence class.

1 Introduction

Learning cause-and-effect relationships is one of the fundamental problems for various fields, including biology [28, 6], epidemiology [26], and economics [12]. A prominent approach for causal discovery models the underlying system as a causal graph represented by a directed acyclic graph (DAG), where nodes denote random variables (measured or latent) and directed edges denote causal effects from tails to arrowheads [22, 29, 24]. Accordingly, the task of structural learning entails piecing together the constraints found in the data (and implied by the underlying, unobserved causal system) to infer the corresponding causal graph. In practice, however, these constraints are almost never sufficient to determine the true causal graph, and a collection of compatible graphs ends up being the target of the analysis, which forms what is known as an *equivalence class* (EC).

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

The formal understanding and characterization of equivalence classes have been an important part of the causal discovery literature for various reasons. For instance, one needs to understand how the output of a learning algorithm relates to the true underlying system that they are trying to infer. Also, ECs are defined with respect to certain constraints implied by the underlying structure in the data (e.g., conditional independences (CIs)), which need to be made explicit and fully understood if one wants to learn from the data (including due to false positives, negatives). Whenever only observational (non-experimental) data is available, the *Markov equivalence class* (for short, MEC) characterizes the causal graphs that imply, by the d-separation criterion, the same set of conditional independences (CIs) over the measured variables [32]. The availability of interventional (i.e., experimental) data opens up new opportunities to reduce the size of the equivalence class down, possibly to recover the true causal graph [10, 18, 8]. An intervention on a (measured) variable X modifies the mechanism by which it is generated, inducing an interventional distribution over the measured variables $\mathbf{V}$, denoted as $P_X(\mathbf{V})$ or P_X [22]. The works in [9, 34, 17] characterize the so called $\mathcal{I}$-Markov equivalence class, which uses distributional invariances within and across the available mixture of observational and interventional distributions. For instance, the graphs $\mathcal{D}_1 = \{X \to Y, X \leftarrow L \to Y\}$, where L is latent, and $\mathcal{D}_2 = \{X \leftarrow Y\}$ are indistinguishable from observational data alone as no CI is implied (i.e., $X \not\perp\!\!\!\perp Y$). Still, they are immediately distinguishable given $\langle P, P_X \rangle$ by contrasting $P(Y)$ and $P_X(Y)$.

In this paper, we investigate *soft interventions* such that the mechanism of an intervened node V_i is modified without fully eliminating the effect of its parents. This operation is also known as a mechanism change [31] or a parameter change [5], and it presents in many settings a more realistic model than *hard* or *perfect* interventions, where variables are forced to a fixed value (see also [2, 3, 33, 22, Sec. 3.2.2]). Furthermore, we relax the interventional setting by assuming the targets of the intervention to be unknown. For example, in molecular biology, the effects of various added chemicals to the cell are not set to one specific value nor they are precisely known [4].

The unknown interventional target setting requires a separate treatment than the known one since it's certainly less informative, i.e., the equivalence class of causal graphs is usually larger (never smaller), and many of the proposed characterizations and algorithms do not immediately apply. For concreteness, consider the two causal graphs mentioned above ($\mathcal{D}_1$, $\mathcal{D}_2$) that are distinguishable under a known interventional target set $\mathcal{I} = \langle \emptyset, \{X\} \rangle$, where $\emptyset$ denotes the observational setting and $\{X\}$ denotes an intervention on the variable. However, they turn out to be indistinguishable when $\mathcal{D}_1$ is associated with $\mathcal{I}_1 = \langle \emptyset, \{X\} \rangle$ but $\mathcal{D}_2$ is associated with $\mathcal{I}_2 = \langle \emptyset, \{X, Y\} \rangle$. In other words, the distributional invariances (to be formally defined in Section 3) accept both hypotheses that a pair of distributions with unknown intervention targets $\langle P_1, P_2 \rangle$ is generated by $\langle \mathcal{D}_1, \mathcal{I}_1 \rangle$ or $\langle \mathcal{D}_2, \mathcal{I}_2 \rangle$, where $P_1 = P(\mathbf{V})$ in both, $P_2 = P_X(\mathbf{V})$ for $\mathcal{D}_1$, and $P_2 = P_{X,Y}(\mathbf{V})$ for $\mathcal{D}_2$. Since the data is compatible with both $\mathcal{I}_1, \mathcal{I}_2$ for different graphs, the EC is underdetermined relative to known interventional targets.

Various approaches have been proposed to learn the causal graph from interventional distributions with unknown interventional targets. The works in [4, 23, 38, 30, 15] assume Markovianity (the absence of latent confounders). Another approach described in [27] learns cyclic causal graphs assuming linearity from unknown shift interventions, which is a specific type of soft interventions. Finally, [20] presents a framework called JCI, which pools the various distributions together by constructing context variables and then running traditional learning algorithms to identify the corresponding EC.[1]

In this work, we take a more fundamental approach and explicitly formalize the constraints that are being tested among the mixture of observational (when available) and interventional distributions, as well as provide a characterization of the equivalence class with respect to those constraints. Assuming a tuple of distributions $\mathbf{P} = \langle P_i \rangle_{i=1}^m$ is generated by the same system, i.e., causal graph with latents, we define a property called Ψ-Markov that connects $\mathbf{P}$ to a pair consisting of (1) a causal graph $\mathcal{D}$ and (2) a set of interventional targets $\mathcal{I}$. Building on this property in Section 3, we provide a graphical characterization that allows one to test whether two causal graphs with possibly different sets of intervention targets belong to the same Ψ-Markov equivalence class. We show that a graphical characterization for the causally sufficient case follows as a special case of this result. In Section 4, we develop Ψ-FCI, a constraint-based algorithm capable of harnessing the distributional invariances found across the combined data to learn the corresponding equivalence class. Finally, we prove that this algorithm is sound and complete, in the sense that it is the most informative in the sample limit. In other words, Ψ-FCI discovers as many tails and arrowheads as can be oriented within the corresponding Ψ-Markov equivalence class. In summary, our contributions are as follow:

[1] We provide detailed discussion on how some of these works compare to ours in the full report [13, Appx. D].

1. We formulate a graphical characterization to test whether two pairs of causal graphs and their corresponding interventional target sets, $\langle \mathcal{D}_1, \mathcal{I}_1 \rangle$ and $\langle \mathcal{D}_2, \mathcal{I}_2 \rangle$, are in the same Ψ-Markov equivalence class, i.e., they are indistinguishable with respect to the available datasets.
2. We develop a sound and complete algorithm to learn equivalence classes of causal graphs from a collection of observational and experimental distributions with unknown interventional targets.

2 Preliminaries

We introduce in this section the necessary concepts and notation used throughout the paper. Upper case letters denote random variables and lower case letters denote an assignment. Also, bold letters denote sets. For $\mathbf{X}, \mathbf{Y}, \mathbf{Z}$, the CI relation $\mathbf{X}$ *is independent of* $\mathbf{Y}$ *conditioned on* $\mathbf{Z}$ is written as $\mathbf{X} \perp\!\!\!\perp \mathbf{Y} | \mathbf{Z}$. The d-separation statement $\mathbf{X}$ *is d-separated from* $\mathbf{Y}$ *given* $\mathbf{Z}$ *in graph* $\mathcal{D}$ is written as $(\mathbf{X} \perp\!\!\!\perp \mathbf{Y} | \mathbf{Z})_{\mathcal{D}}$. $\mathcal{D}_{\overline{\mathbf{X}}}$ denotes the graph obtained from $\mathcal{D}$ where all the incoming edges to the nodes in $\mathbf{X}$ are removed. Similarly, $\mathcal{D}_{\underline{\mathbf{X}}}$ denotes the removal of outgoing edges. We assume there is no selection bias. A star on edge endpoints is used as a wildcard to denote circle, arrowhead, or tail.

Causal Bayesian Network (CBN): Let $P(\mathbf{V})$ be a probability distribution over a set of variables $\mathbf{V}$, and $P_{\mathbf{x}}(\mathbf{V})$ denote the distribution resulting from the *hard intervention* $do(\mathbf{X} = \mathbf{x})$, which sets $\mathbf{X} \subseteq \mathbf{V}$ to constants $\mathbf{x}$. Let $\mathbf{P}^*$ denote the set of all interventional distributions $P_{\mathbf{x}}(\mathbf{V})$, for all $\mathbf{X} \subseteq \mathbf{V}$, including $P(\mathbf{V})$. A directed acyclic graph (DAG) over $\mathbf{V}$ is said to be a *causal Bayesian network* compatible with $\mathbf{P}^*$ if and only if, for all $\mathbf{X} \subseteq \mathbf{V}$, $P_{\mathbf{x}}(\mathbf{v}) = \prod_{\{i|V_i \notin \mathbf{X}\}} P(v_i|\mathbf{pa}_i)$, for all $\mathbf{v}$ consistent with $\mathbf{x}$, and where $\mathbf{Pa}_i$ is the set of parents of V_i [22, 1, pp. 24]. Given that a subset of the variables are unmeasured or latent, $\mathcal{D}(\mathbf{V} \cup \mathbf{L}, \mathbf{E})$ will represent the causal graph where $\mathbf{V}$ and $\mathbf{L}$ denote the measured and latent variables, respectively, and $\mathbf{E}$ denotes the edges. Following the convention in [22], for simplicity, a dashed bi-directed edge is used instead of the corresponding latent variables. The CI relations can be read from the graph using a graphical criterion known as *d-separation* [21].

Soft Interventions: Under this type of interventions, the original conditional distributions of the intervened variables $\mathbf{X}$ are replaced with new ones, without completely eliminating the causal effect of the parents. Accordingly, the interventional distribution $P_{\mathbf{X}}(\mathbf{v})$ for $\mathbf{X} \subseteq \mathbf{V}$ is such that $P^*(X_i|Pa_i) \neq P(X_i|Pa_i)$, $\forall X_i \in \mathbf{X}$. We refer to the mixture of observational and interventional distributions as interventional for simplicity, which factorizes as follow:

$$P_{\mathbf{X}}(\mathbf{v}) = \sum_{\mathbf{L}} \prod_{\{i|X_i \in \mathbf{X}\}} P^*(x_i|\mathbf{pa}_i) \prod_{\{j|T_j \notin \mathbf{X}\}} P(t_j|\mathbf{pa}_j) \tag{1}$$

Ancestral Graphs: A *mixed* graph can contain directed and bi-directed edges. A is an ancestor of B if there is a directed path from A to B. A is a *spouse* of B if $A \leftrightarrow B$ is present. If A is both a spouse and an ancestor of B, this creates an *almost directed cycle*. A path is a sequence of edges joining a unique sequence of nodes. An *inducing path* relative to $\mathbf{L}$ is a path on which every non-endpoint node $X \notin \mathbf{L}$ is a collider on the path (i.e., both edges incident to the node are into it) and every collider is an ancestor of an endpoint of the path. A mixed graph is *ancestral* if it does not contain directed or almost directed cycles. It is *maximal* if there is no inducing path (relative to the empty set) between any two non-adjacent nodes. A *Maximal Ancestral Graph* (MAG) is a graph that is both ancestral and maximal [25]. Given a causal graph $\mathcal{D}(\mathbf{V} \cup \mathbf{L}, \mathbf{E})$, a unique MAG $\mathcal{M}_{\mathcal{D}}$ over $\mathbf{V}$ can be constructed such that both the independence and the ancestral relations among $\mathbf{V}$ are retained; see, [36, p. 6].

A triple $\langle X, Y, Z \rangle$ is an unshielded triple if X and Y are adjacent, Y and Z are adjacent, and X and Z are not adjacent. If both edges are into Y, then the triple is referred to as *unshielded collider*. A path between X and Y, $p = \langle X, \dots, W, Z, Y \rangle$, is *discriminating* for Z if every node between X and Z is a collider on p and is a parent of Y. Two MAGs are Markov equivalent if and only if (1) they have the same adjacencies; (2) the same unshielded colliders; and (3) if a path p is a discriminating path for Z in both graphs, then Z is a collider on p in one graph if and only if it is a collider on p in the other. A *PAG* represents an MEC of a MAG and is learnable from data. The output of the celebrated FCI algorithm is a PAG, which is proven sound and complete for the corresponding MEC [37].

3 Interventional Equivalence with Unknown Targets

In this section, we formalize the notion of interventional equivalence class when the interventional targets are unknown. Let V_i^j denote an intervention on V_i with a unique mechanism identified by

j. Hence, interventions denoted by V_i^j and V_i^k force different mechanisms such that $P_{V_i^j}(V_i|Pa_i) \neq P_{V_i^k}(V_i|Pa_i)$. Accordingly, each interventional target $\mathbf{I} = \{V_i^j\}_{i\in|\mathbf{V}|}$ is defined by a set of variables with corresponding mechanism identifiers denoted by $j \in \mathbb{N}$. We drop the mechanism identifier whenever it is not necessary. Next, we define an important operation between two interventional targets.

Definition 1 (Symmetrical Difference Δ)**.** *Given two interventional targets* $\mathbf{I}$ *and* $\mathbf{J}$*, let* $\mathbf{I}\Delta\mathbf{J}$ *denote the symmetrical difference set such that* $V_i \in \mathbf{I}\Delta\mathbf{J}$ *if* $V_i^j \in \mathbf{I}$ *and* $V_i^j \notin \mathbf{J}$ *or vice versa.*

In words, the operation identifies the set of variables that have a unique interventional mechanism across two interventional targets. For example, given $\mathbf{I} = \{X^1, Y, Z\}$ and $\mathbf{J} = \{X^2, Y\}$, then $\mathbf{I}\Delta\mathbf{J} = \{X, Z\}$.

Next, we generalize the interventional Markov property ($\mathcal{I}$-Markov) [17] for the case when the targets are unknown, which we call Ψ-Markov. This property features prominently the different tests that emerge when a combination of observational and experimental distributions is available.

Definition 2 (Ψ-Markov Property)**.** *Let* $\mathcal{D} = (\mathbf{V} \cup \mathbf{L}, \mathbf{E})$ *denote a causal graph, let* $\mathbf{P}$ *denote an ordered tuple of distributions, and let* $\boldsymbol{\mathcal{I}}$ *denote an ordered tuple of interventional targets such that* $|\mathbf{P}| = |\boldsymbol{\mathcal{I}}|$*. Tuple* $\mathbf{P}$ *satisfies the* Ψ*-Markov property with respect to the pair* $\langle \mathcal{D}, \boldsymbol{\mathcal{I}} \rangle$ *if the following holds for disjoint* $\mathbf{Y}, \mathbf{Z}, \mathbf{W} \subseteq \mathbf{V}$*:*

(a) For $\mathbf{I}_i \in \boldsymbol{\mathcal{I}}$*:* $\quad P_i(\mathbf{y}|\mathbf{w}, \mathbf{z}) = P_i(\mathbf{y}|\mathbf{w}) \quad$ *if* $\mathbf{Y} \perp\!\!\!\perp \mathbf{Z}\,|\mathbf{W}$ *in* $\mathcal{D}$

(b) For $\mathbf{I}_i, \mathbf{I}_j \in \boldsymbol{\mathcal{I}}$*:* $\quad P_i(\mathbf{y}|\mathbf{w}) = P_j(\mathbf{y}|\mathbf{w}) \quad$ *if* $\mathbf{Y} \perp\!\!\!\perp \mathbf{K}\,|\mathbf{W} \setminus \mathbf{W}_\mathbf{K}$ *in* $\mathcal{D}_{\underline{\mathbf{W}_\mathbf{K}}, \overline{\mathbf{R}(\mathbf{W})}}$*,*

where $\mathbf{K} := \mathbf{I}_i \Delta \mathbf{I}_j$*,* $\mathbf{W}_\mathbf{K} =: \mathbf{W} \cap \mathbf{K}$*,* $\mathbf{R} := \mathbf{K} \setminus \mathbf{W}_\mathbf{K}$*, and* $\mathbf{R}(\mathbf{W}) \subseteq \mathbf{R}$ *are non-ancestors of* $\mathbf{W}$ *in* $\mathcal{D}$*.*

$\Psi_{\boldsymbol{\mathcal{I}}}(\mathcal{D})$ *denotes set of distribution tuples that satisfy the* Ψ*-Markov property with respect to* $\langle \mathcal{D}, \boldsymbol{\mathcal{I}} \rangle$*.*

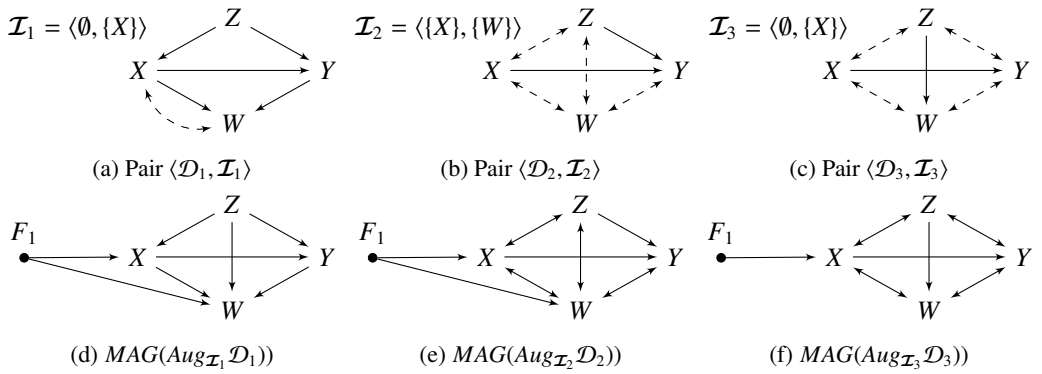
For concreteness and to illustrate this definition, we provide two examples with tuples of distributions that satisfy and do not satisfy the corresponding Ψ-Markov property, respectively.

Example 1. *Consider the causal graph* $\mathcal{D}^* = \{X \to Y, X \leftarrow L \to Y\}$ *where* L *is a latent node, and let the pair of distributions* $\langle P_1, P_2 \rangle$ *be the result of intervening on the targets* $\boldsymbol{\mathcal{I}}^* = \langle \emptyset, \{X\} \rangle$*. It is easy to check that* $\mathbf{P}$ *satisfies the* Ψ*-Markov property with respect to* $\langle \mathcal{D}^*, \boldsymbol{\mathcal{I}}^* \rangle$ *as no constraint of type (a) or (b) is applicable. For example, if* $(Y \perp\!\!\!\perp X)_{\mathcal{D}^*_{\underline{X}}}$*, then the invariance* $P_1(y|x) = P_2(y|x)$ *must hold. Since the d-separation fails, the invariance is not required. Similarly,* $\mathbf{P}$ *satisfies the* Ψ*-Markov property with respect to* $\langle \mathcal{D}, \boldsymbol{\mathcal{I}} \rangle$ *where* $\mathcal{D} = \{X \leftarrow Y\}$ *and* $\boldsymbol{\mathcal{I}} = \langle \emptyset, \{X, Y\} \rangle$ *or* $\boldsymbol{\mathcal{I}} = \langle \{X\}, \{Y\} \rangle$*.*

Example 2. *Consider the pair* $\langle \mathcal{D}^*, \boldsymbol{\mathcal{I}}^* \rangle$ *and the corresponding tuple of distributions* $\mathbf{P}$ *from Ex. 1. We check if* $\mathbf{P}$ *satisfies the* Ψ*-Markov property with respect to* $\langle \mathcal{D}^*, \boldsymbol{\mathcal{I}} \rangle$ *for* $\boldsymbol{\mathcal{I}} = \langle \emptyset, \{Y\} \rangle$*. Now,* $\mathbf{K} = \emptyset \Delta \{Y\} = \{Y\}$ *and we have* $(X \perp\!\!\!\perp Y)_{\mathcal{D}_{\overline{Y}}}$*, so* $P_1(X) = P_2(X)$ *should hold according to Constraint (b). However, the invariance does not hold simply because* P_2*, in truth, corresponds to the interventional distribution on* X*. Therefore,* $\mathbf{P}$ *does not satisfy the* Ψ*-Markov property with respect to* $\langle \mathcal{D}^*, \boldsymbol{\mathcal{I}} \rangle$*.*

A few remarks are relevant about the Ψ-Markov property at this point. First, an ordered tuple of interventional distributions $\mathbf{P}$ with unknown interventional targets is said to satisfy the Ψ-Markov property if two qualitatively different types of constraints hold – (a) the "traditional" Markov property, where separation in the causal graph $\mathcal{D}$ implies CI in the corresponding distribution (including the interventional ones); (b) invariances across pairs of distributions given separation statements in the mutilated graph. These mutilations depend on the symmetrical difference set ($\mathbf{K}$) of the interventional targets. Intuitively, should $\mathbf{I}_i, \mathbf{I}_j$ correspond to the true interventional targets of P_i, P_j, respectively, (b) verifies distributional invariances between $P_{\mathbf{I}_i}$ and $P_{\mathbf{I}_j}$ if the corresponding separation holds.

Second, the importance of the property stems from the fact that a tuple of interventional distributions generated by a causal graph $\mathcal{D}$ satisfies the Ψ-Markov property relative to it and the corresponding true interventional targets. See [13, Thm. 4 in Appx. A] for an explicit statement. Third, we note that if the interventional targets are known (i.e., $\boldsymbol{\mathcal{I}}_1 = \boldsymbol{\mathcal{I}}_2$), the Ψ-Markov property still generalizes $\mathcal{I}$-Markov [17] by relaxing the assumption of *controlled experiment setting*. In practice, it may be hard to ascertain that interventions over the same variable are performed exactly in the same way, which makes this more refined characterization potentially interesting even to when the interventional target is known. Finally, decoupling the distributions from the corresponding interventional targets is instrumental to formulate interventional equivalence when the targets are unknown as shown below.



(a) Pair $\langle \mathcal{D}_1, \mathcal{I}_1 \rangle$ (b) Pair $\langle \mathcal{D}_2, \mathcal{I}_2 \rangle$ (c) Pair $\langle \mathcal{D}_3, \mathcal{I}_3 \rangle$

(d) $MAG(Aug_{\mathcal{I}_1}\mathcal{D}_1))$ (e) $MAG(Aug_{\mathcal{I}_2}\mathcal{D}_2))$ (f) $MAG(Aug_{\mathcal{I}_3}\mathcal{D}_3))$

Figure 1: Pairs of causal graphs and intervention-target sets and the corresponding $\mathcal{I}$-MAGs.

Definition 3 (Ψ-Markov Equivalence)**.** *Given the causal graphs $\mathcal{D}_1 = (\mathbf{V} \cup \mathbf{L}_1, \mathbf{E}_1)$ and $\mathcal{D}_2 = (\mathbf{V} \cup \mathbf{L}_2, \mathbf{E}_2)$, and the corresponding interventional targets $\mathcal{I}_1, \mathcal{I}_2$, the pairs $\langle \mathcal{D}_1, \mathcal{I}_1 \rangle$ and $\langle \mathcal{D}_2, \mathcal{I}_2 \rangle$ are said to be Ψ-Markov equivalent if $\Psi_{\mathcal{I}_1}(\mathcal{D}_1) = \Psi_{\mathcal{I}_2}(\mathcal{D}_2)$.*

In words, two pairs of causal graphs and their corresponding sets of interventional targets $\langle \mathcal{D}_1, \mathcal{I}_1 \rangle$ and $\langle \mathcal{D}_2, \mathcal{I}_2 \rangle$ are Ψ-Markov equivalent if they can induce the same set of distribution tuples. In practice, it may be challenging to evaluate whether the premises of the Ψ-Markov property hold since they entail different graph mutilations of $\mathcal{D}$. In order to ameliorate this task, we build on the graph augmentation construction following [17].

Definition 4 (Augmented graph)**.** *Consider a causal graph $\mathcal{D} = (\mathbf{V} \cup \mathbf{L}, \mathbf{E})$ and a set of interventional targets $\mathcal{I}$. Let the multiset $\mathcal{K}$ be defined as such $\mathcal{K} = \{\mathbf{K}_1, \mathbf{K}_2, \ldots, \mathbf{K}_k\} = \{\mathbf{K} | \mathbf{I}, \mathbf{J} \in \mathcal{I} \wedge \mathbf{I} \Delta \mathbf{J} = \mathbf{K}\}$. The augmented graph of $\mathcal{D}$ with respect to $\mathcal{I}$, denoted as $Aug_{\mathcal{I}}(\mathcal{D})$, is the graph constructed as follows: $Aug_{\mathcal{I}}(\mathcal{D}) = (\mathbf{V} \cup \mathbf{L} \cup \mathcal{F}, \mathbf{E} \cup \mathcal{E})$ where $\mathcal{F} := \{F_i\}_{i \in [k]}$ and $\mathcal{E} = \{(F_i, j)\}_{i \in [k], j \in \mathbf{K}_i}$.*

For each pair of interventional targets $\mathbf{I}, \mathbf{J} \in \mathcal{I}$ such that $\mathbf{K} = \mathbf{I} \Delta \mathbf{J}$, the augmented graph appends the causal graph $\mathcal{D}$ with a utility F-node that is a parent to each node in $\mathbf{K}$. The significance of this construction follows from Proposition 1 where separation statements in the Ψ-Markov definition are tied (shown to be equivalent, formally speaking) to ones in the augmented graph, with no need to perform any graphical mutilation.The result is illustrated in the following example.

Proposition 1. *Consider a causal graph $\mathcal{D} = (\mathbf{V} \cup \mathbf{L}, \mathbf{E})$, a set of interventional targets $\mathcal{I}$, and the augmented graph $Aug_{\mathcal{I}}(\mathcal{D})$, where $\mathcal{F} = \{F_i\}_{i \in [k]}$. Let $\mathbf{K}_i$ be the set of nodes adjacent to $F_i, \forall i \in [k]$. The following equivalence relations hold for disjoint $\mathbf{Y}, \mathbf{Z}, \mathbf{W} \subseteq \mathbf{V}$, where $\mathbf{W}_i := \mathbf{W} \cap \mathbf{K}_i$, $\mathbf{R} := \mathbf{K}_i \setminus \mathbf{W}_i$.*[2]

$$(\mathbf{Y} \perp\!\!\!\perp \mathbf{Z} \,|\mathbf{W})_{\mathcal{D}} \iff (\mathbf{Y} \perp\!\!\!\perp \mathbf{Z} \,\big|\mathbf{W}, F_{[k]})_{Aug_{\mathcal{I}}(\mathcal{D})} \tag{2}$$

$$(\mathbf{Y} \perp\!\!\!\perp \mathbf{K}_i \,|\mathbf{W} \setminus \mathbf{W}_i)_{\mathcal{D}_{\underline{\mathbf{W}_i}, \overline{\mathbf{R}(\mathbf{W})}}} \iff (\mathbf{Y} \perp\!\!\!\perp F_i \,\big|\mathbf{W}, F_{[k] \setminus \{i\}})_{Aug_{\mathcal{I}}(\mathcal{D})} \tag{3}$$

Example 3. *Consider $\mathcal{D} = \{X \to Y \leftarrow L \to Z\}$ where L is latent and let $\mathcal{I} = \langle \{X, Z^1\}, \{Z^2\} \rangle$. The corresponding augmented graph $\mathcal{D}'$ is composed of $\mathcal{D}$ appended with $X \leftarrow F_1 \to Z$. By Prop. 1, $(X \perp\!\!\!\perp Z)_{\mathcal{D}}$ can be tested by $(X \perp\!\!\!\perp Z \,|F_1)_{\mathcal{D}'}$. Also, $(Y \perp\!\!\!\perp \{X, Z\})_{\mathcal{D}_{\underline{X}, \overline{Z}}}$ can be tested as $(F_1 \perp\!\!\!\perp Y \,|X)_{\mathcal{D}'}$.*

Maximal Ancestral Graphs (MAGs) provide a convenient representation capable of preserving all the tested constraints in augmented graphs represented by d-separations [25]; see also [36, p. 6]. This is formalized in Definition 5 and the construct is referred to as an *$\mathcal{I}$-MAG*; see Example 4 below.

Definition 5 ($\mathcal{I}$-MAG)**.** *Given a causal graph $\mathcal{D} = (\mathbf{V} \cup \mathbf{L}, \mathbf{E})$ and a set of interventional targets $\mathcal{I}$, an $\mathcal{I}$-MAG is the MAG constructed over $\mathbf{V}$ from $Aug_{\mathcal{I}}(\mathcal{D})$, i.e., $MAG(Aug_{\mathcal{I}}(\mathcal{D}))$.*

Example 4. *Consider $\mathcal{D}^*$ from Ex. 1. $Aug_{\mathcal{I}}(\mathcal{D}^*) = \{F_1 \to X \to Y, X \leftarrow L \to Y\}$ for $\mathcal{I} = \langle \emptyset, \{X\} \rangle$. Then, the corresponding $\mathcal{I}$-MAG is $MAG(Aug_{\mathcal{I}}(\mathcal{D}^*)) = \{X \leftarrow F_1 \to Y, X \to Y\}$.*

Putting these results together, we derive next a graphical characterization for two causal graphs with corresponding sets of interventional targets to be Ψ-Markov equivalent.

[2] All the proofs can be found in Appendices A & B of the full report [13].

Theorem 1 (Ψ-Markov Characterization). *Given causal graphs $\mathcal{D}_1 = (\mathbf{V} \cup \mathbf{L}_1, \mathbf{E}_2), \mathcal{D}_2 = (\mathbf{V} \cup \mathbf{L}_2, \mathbf{E}_2)$ and corresponding sets of interventional targets $\mathcal{I}_1, \mathcal{I}_2$, $\langle \mathcal{D}_1, \mathcal{I}_1 \rangle$ and $\langle \mathcal{D}_2, \mathcal{I}_2 \rangle$ are Ψ-Markov equivalent if and only if for $\mathcal{M}_1 = MAG(Aug_{\mathcal{I}_1}(\mathcal{D}_1))$ and $\mathcal{M}_2 = MAG(Aug_{\mathcal{I}_2}(\mathcal{D}_2))$:*[3]

1. *$\mathcal{M}_1$ and $\mathcal{M}_2$ have the same skeleton;*
2. *$\mathcal{M}_1$ and $\mathcal{M}_2$ have the same unshielded colliders;*
3. *If a path p is a discriminating path for a node Y in both $\mathcal{M}_1$ and $\mathcal{M}_2$, then Y is a collider on the path in one graph if and only if it is a collider on the path in the other.*

Theorem 1 states that the pairs $\langle \mathcal{D}_1, \mathcal{I}_1 \rangle$ and $\langle \mathcal{D}_2, \mathcal{I}_2 \rangle$ are Ψ-Markov equivalent if their corresponding $\mathcal{I}$-MAGs satisfy the corresponding three conditions, as illustrated in the example below.

Example 5. *Consider the pairs $\langle \mathcal{D}_1, \mathcal{I}_1 \rangle$ and $\langle \mathcal{D}_2, \mathcal{I}_2 \rangle$ in Figs. 1a and 1b, respectively. The corresponding $\mathcal{I}$-MAGs are shown in Figs. 1d and 1e satisfy the three conditions in Thm. 1, hence the pairs are Ψ-Markov equivalent. Note $\mathcal{K}$, according to Def. 4, is $\{\{X\}\}$ for $\mathcal{I}_1$ and $\{\{X, W\}\}$ for $\mathcal{I}_2$. Hence, F_1 is adjacent to $\{X\}$ in $Aug_{\mathcal{I}_1}(\mathcal{D}_1)$ and F_1 is adjacent to $\{X, W\}$ in $Aug_{\mathcal{I}_2}(\mathcal{D}_2)$. However, F_1 is adjacent to W in Fig. 1d due to the inducing path $\langle F_1, X, W \rangle$ in $Aug_{\mathcal{I}_1}(\mathcal{D}_1)$. On the other hand, $\langle \mathcal{D}_3, \mathcal{I}_3 \rangle$ in Fig. 1c is not Ψ-Markov equivalent to either one of the other pairs as we violate all three conditions of Thm. 1. For instance, the $\mathcal{I}$-MAGs in Figs. 1e and 1f do not share the same skeleton, $\langle F_1, X, W \rangle$ is an unshielded collider only in Fig. 1f, and $p = \langle F_1, X, Z, Y \rangle$ is a discriminating path for Z in both; however, Z is a collider along p in Fig. 1f while it is a non-collider in Fig. 1e.*

In a setting where the observational distribution is available and identified among the available distributions, it becomes necessary to fix $\emptyset$ across $\mathcal{I}_1, \mathcal{I}_2$, which is a special case of Thm. 1. Further, note that the graphical characterization introduced in [17] for causal graphs with known interventional targets is a special case of Thm. 1 whenever $\mathcal{I}_1 = \mathcal{I}_2$ with the controlled experiment setting.

3.1 Markovian Case

One special class of causal graphs that is of high interest in the literature is known as *Markovian*, where there is no latent variable affecting more than one observable node (i.e., no bidirected arrows). It follows from Theorem 1 the following graphical characterization for this class of models.

Corollary 1. *Given causal graphs without latents, $\mathcal{D}_1 = (\mathbf{V}, \mathbf{E}_2), \mathcal{D}_2 = (\mathbf{V}, \mathbf{E}_2)$, and the corresponding interventional targets $\mathcal{I}_1, \mathcal{I}_2$, the pairs $\langle \mathcal{D}_1, \mathcal{I}_1 \rangle$ and $\langle \mathcal{D}_2, \mathcal{I}_2 \rangle$ are Ψ-Markov equivalent if and only if $Aug_{\mathcal{I}_1}(\mathcal{D}_1)$ and $Aug_{\mathcal{I}_2}(\mathcal{D}_2)$ have (1) the same skeleton and (2) the same unshielded colliders.*

Note that under known interventional targets (i.e., $\mathcal{I}_1 = \mathcal{I}_2$), Corol. 1 recovers and generalizes the characterization in [34, Thm. 3.9] by encoding different interventional mechanisms and thus identifying a smaller equivalence class. For a more detailed comparison, we refer readers to Appendix D.1.

4 Learning Algorithm: Soundness and Completeness

We investigate in this section the problem of how to learn the Ψ-Markov EC (Def. 3) from a tuple of interventional distributions generated by some unknown pair $\langle \mathcal{D}, \mathcal{I} \rangle$. The characterization provided in Thm. 1 together with PAGs motivate the following definition of *Ψ-PAG*.

Definition 6 (Ψ-PAG). *Given a pair of causal graph and interventional target, $\langle \mathcal{D}, \mathcal{I} \rangle$, let $\mathcal{M} = MAG(Aug_{\mathcal{I}}(\mathcal{D}))$, and let $[\mathcal{M}]$ be the set of $\mathcal{I}$-MAGs corresponding to all the pairs $\langle \mathcal{D}', \mathcal{I}' \rangle$ that are Ψ-Markov equivalent to $\langle \mathcal{D}, \mathcal{I} \rangle$. The Ψ-PAG for $\langle \mathcal{D}, \mathcal{I} \rangle$, denoted $\mathcal{P}$, is a graph such that:*

1. *$\mathcal{P}$ has the same adjacencies as $\mathcal{M}$, and any member of $[\mathcal{M}]$ does; and*
2. *every non-circle mark (tail or arrowhead) in $\mathcal{P}$ is an invariant mark in $[\mathcal{M}]$.*

Some remarks follow immediately from this definition. First, Ψ-PAG generalizes PAGs, as used in the observational case. Second, even though the augmented F-nodes are part of the Ψ-PAG, which is the very target of the learning process, they never transpire as random variables, and are

[3] We assume the symmetrical difference sets $\mathcal{K}_1$, corresponding to $\mathcal{I}_1$, and $\mathcal{K}_2$, corresponding to $\mathcal{I}_2$, are indexed following the same pattern such that $\mathcal{K}_1 \ni \mathbf{K}_k = \mathbf{I}_i \Delta \mathbf{I}_j$ where $\mathbf{I}_i, \mathbf{I}_j \in \mathcal{I}_1$ iff $\mathcal{K}_2 \ni \mathbf{K}_k = \mathbf{I}_i \Delta \mathbf{I}_j$ where $\mathbf{I}_i, \mathbf{I}_j \in \mathcal{I}_2$. This is required to maintain the correspondence between the F-nodes in $\mathcal{M}_1$ and $\mathcal{M}_2$.

Algorithm 1 Ψ-FCI: Algorithm for Learning a Ψ-PAG

Input: Tuple of distributions $\mathbf{P} = \langle P_1, \ldots, P_m \rangle$
Output: Ψ-PAG $\mathcal{P}$

1: $\mathcal{F} \leftarrow \emptyset, k \leftarrow 0, \sigma : \mathbb{N} \rightarrow \mathbb{N} \times \mathbb{N}$
2: **for** all pairs $P_i, P_j \in \mathbf{P}$ **do** $k \leftarrow k + 1, \mathcal{F} \leftarrow \mathcal{F} \cup \{F_k\}, \sigma(k) \rightarrow (i, j)$
3: **Phase I: Skeleton**
4: Form a complete graph $\mathcal{P}$ over $\mathbf{V} \cup \mathcal{F}$ with $\circ\!\!-\!\!\circ$ edges between every pair of nodes.
5: **for** every pair $X, Y \in \mathbf{V} \cup \mathcal{F}$ **do**
6: **if** $X \in \mathcal{F} \wedge Y \in \mathcal{F}$ **then** $\texttt{SepSet}(X, Y) \leftarrow \emptyset$, $\texttt{SepFlag} \leftarrow True$
7: **else** $(\texttt{SepSet}(X, Y), \texttt{SepFlag}) \leftarrow$ InvToSep$(\mathbf{P}, X, Y, \mathbf{V}, \mathcal{F}, \sigma)$
8: **if** $\texttt{SepFlag} = True$ **then** Remove the edge between X, Y in $\mathcal{P}$.
9: **Phase II: Unshielded Colliders**
10: $\mathcal{R}_0$: For every unshielded triple $\langle X, Z, Y \rangle$ in $\mathcal{P}$, orient it as $X \ast\!\!\rightarrow Z \leftarrow\!\!\ast Y$ iff $Z \notin \texttt{SepSet}(X, Y)$
11: **Phase III: Orientation Rules**
12: *Rule* $\mathcal{R}^+$*:* For any $F_k \in \mathcal{F}$, orient adjacent edges out of F_k.
13: Apply the seven FCI rules in [37] ($\mathcal{R}_1 - \mathcal{R}_4, \mathcal{R}_8 - \mathcal{R}_{10}$) until none applies.
14: **function** InvToSep$(\mathbf{P}, X, Y, \mathbf{V}, \mathcal{F}, \sigma)$
15: $\texttt{SepSet} \leftarrow \emptyset$, $\texttt{SepFlag} \leftarrow False$
16: **if** $X \notin \mathcal{F} \wedge Y \notin \mathcal{F}$ **then** Pick $P_i \in \mathbf{P}$ arbitrarily.
17: **for** $\mathbf{W} \subseteq \mathbf{V} \setminus \mathcal{F}$ **do**
18: **if** $P_i(y|\mathbf{w}, x) = P_i(y|\mathbf{w})$ **then** $\texttt{SepSet} \leftarrow \mathbf{W} \cup \mathcal{F}$, $\texttt{SepFlag} \leftarrow True$, **break**
19: **else** Suppose $X \in \mathcal{F}, Y \notin \mathcal{F}$, and let F_k denote X.
20: $(i, j) \leftarrow \sigma(k)$
21: **for** $\mathbf{W} \subseteq \mathbf{V} \setminus \{Y\}$ **do**
22: **if** $P_i(y|\mathbf{w}) = P_j(y|\mathbf{w})$ **then** $\texttt{SepSet} \leftarrow \mathbf{W} \cup \mathcal{F} \setminus \{F_k\}$, $\texttt{SepFlag} \leftarrow True$, **break**
return $(\texttt{SepSet}, \texttt{SepFlag})$

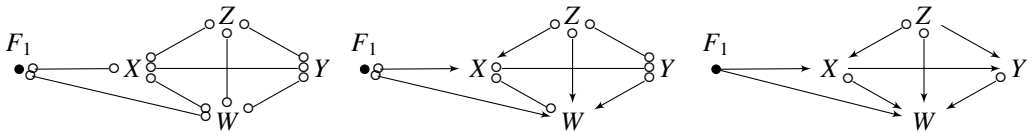
merely graphical instruments used to represent the equivalence class. In fact, the real invariance tests across distributions are stated in the Ψ-Markov property (Def. 2). Third, as expected in any learning setting, some type of faithfulness assumption is needed to infer graphical properties from the corresponding distributional constraints [37, 34, 17, 30]. Hence, we assume that the given collection of interventional distributions is *c-faithful* to the true generating causal graph $\mathcal{D}$ as defined next.

Definition 7 (c-faithfulness)**.** *Consider a causal graph $\mathcal{D}$. A tuple of distributions $\langle P_{\mathbf{I}} \rangle_{\mathbf{I} \in \boldsymbol{\mathcal{I}}} \in \Psi_{\boldsymbol{\mathcal{I}}}(\mathcal{D})$ is called c-faithful to $\mathcal{D}$ if the converse of each of the Ψ-Markov conditions (Def. 2) holds.*

The new algorithm is called Ψ-FCI and is shown in Alg. 1. Ψ-FCI starts by mapping every pair of distributions in $\mathbf{P}$ to a constructed F-node (line 2). In Phase I, Ψ-FCI learns the skeleton of the Ψ-PAG $\mathcal{P}$. It starts by creating a complete graph of circle edges ($\circ\!\!-\!\!\circ$) over $\mathbf{V} \cup \mathcal{F}$, and then uses the function InvToSep$(\cdot)$ at line 7, which we discuss next, to infer a separation set for every pair of nodes, if such a set exists. Line 6 handles a special case in which both nodes are F-nodes and are separable by the empty set, by construction. Phase II recovers the unshielded colliders $\langle X, Z, Y \rangle$ by checking that Z does not belong to the corresponding separation set $\texttt{SepSet}(X, Y)$. Finally, the algorithm orients all the edges incident on F-nodes out of them in $\mathcal{R}^+$ followed by a subset of the FCI rules until none applies anymore. Note that we drop three of the FCI rules ($\mathcal{R}_5 - \mathcal{R}_7$) as they are only applicable in the presence of selection bias which we do not consider.

InvToSep$(\cdot)$ can be considered as the most fundamental part of Ψ-FCI. This function infers separation sets for pairs of nodes in $\mathcal{P}$ from the invariances found across the distributions.[4] The separation sets are key in Ψ-FCI to learn the skeleton and orient the edges of $\mathcal{P}$. If both X and Y are not F-nodes, then we pick an arbitrary distribution $P_i \in \mathbf{P}$ and check if there exists a subset of the variables $\mathbf{W}$ such that $(X \perp\!\!\!\perp Y | W)$ in P_i (lines 3-5). The reason we choose an arbitrary distribution in $\mathbf{P}$ is that the set of conditional independences that can be read from an observational or interventional distribution is the same under soft interventions. For the next step, recall that every F-node is mapped to a unique pair of distributions in $\mathbf{P}$. If one of the two nodes is an F-node, denoted F_k, then we search for a subset of variables $\mathbf{W}$ such that $P_i(y|\mathbf{w}) = P_j(y|\mathbf{w})$ where $(i, j) \leftarrow \sigma(k)$. If such an invariance exists,

[4] There are different ways of implementing hypothesis testing for the distributional invariances, as required in line 22 of Ψ-FCI. In fact, these tests can be seen as evaluating statements in the form $|\hat{P}_i(y|\mathbf{w}) - \hat{P}_j(y|\mathbf{w})| \leq \epsilon$, where the hat represents the empirical distribution. Ψ-FCI is agnostic to the particular implementation of the test, which is in general chosen based on the specific details of the setting.



(a) Learn the skeleton (b) Recover unshielded colliders (c) After orientation rules

Figure 2: Different phases of Ψ-FCI to learn Ψ-PAG $\mathcal{P}$ given a tuple of distributions $\langle P_1, P_2 \rangle$ that is generated by the unknown pair $\langle \mathcal{D}, \mathcal{I} \rangle$, shown in Fig. 1a.

we mark $\mathbf{W} \cup \mathcal{F} \setminus \{F_k\}$ as a separating set between F_k and Y. The validity of this function follows from the constraints of the Ψ-Markov property (Def. 2) and the equivalences in Proposition 1 coupled with the c-faithfulness assumption in Def. 7. We illustrate the use of Ψ-FCI in the example below.

Example 6. *Consider a tuple of distributions $\langle P_1, P_2 \rangle$ and let the pair $\langle \mathcal{D}, \mathcal{I} \rangle$ in Fig. 1a be the true and unknown causal graph and set of corresponding interventional targets. $\mathcal{I}$-MAG $\mathcal{M}$ is shown in Fig. 1d and the aim is to recover the corresponding Ψ-PAG $\mathcal{P}$. Fig. 2a shows the output of Phase I. For instance, F_1 and Y are separable by $\{X, Z\}$, which is inferred by the distributional invariance $P_1(Y|X, Z) = P_2(Y|X, Z)$. Phase II recovers the unshielded colliders as shown in Fig. 2b. For example, $\langle F_1, W, Y \rangle$ is oriented as a collider since $W \notin \texttt{SepSet}(F_1, Y) = \{X, Z\}$. Finally, Phase III applies the orientation rules which gives the graph in Fig. 2c. The edges incident on F_1 are oriented out of it by $\mathcal{R}^+$, $X \to Y$ by $\mathcal{R}_1$, the arrowhead on $X \circ\!\!\to W$ by $\mathcal{R}_2$, and $Z \to Y$ by $\mathcal{R}_4$.*

Putting these observations together, finally, the next theorem ascertains the soundness of Ψ-FCI.

Theorem 2 (Ψ-FCI Soundness)**.** *Assuming tuple $\mathbf{P}$ is generated by unknown pair $\langle \mathcal{D}, \mathcal{I} \rangle$, then Ψ-FCI is sound in the sample limit, i.e., $MAG(Aug_{\mathcal{I}}(\mathcal{D}))$ has the same skeleton as $\mathcal{P}_{\Psi\text{-FCI}}$, the Ψ-PAG learned by Ψ-FCI, and shares all its tail and arrowhead orientations.*

4.1 Ψ-FCI Completeness

One common question for any learning algorithm is how close it can get to the underlying causal structure. In the limit, one would like to discover all the invariant features of the corresponding Ψ-Markov EC, a property called completeness. Concretely, for every circle mark on an edge end in $\mathcal{P}_{\Psi\text{-FCI}}$, we need to establish the following. There exist two pairs $\langle \mathcal{D}_1, \mathcal{I}_1 \rangle$ and $\langle \mathcal{D}_2, \mathcal{I}_2 \rangle$ that are Ψ-Markov equivalent to the true pair $\langle \mathcal{D}, \mathcal{I} \rangle$ such that the corresponding $\mathcal{I}$-MAGs $\mathcal{M}_1$ and $\mathcal{M}_2$ have different marks for that end (i.e., one has a tail while the other has an arrowhead), as illustrated next.

Example 7. *Consider Ψ-PAG $\mathcal{P}$ in Fig. 2c from Ex. 6. $\mathcal{I}$-MAGs in Figs. 1d and 1e are both in the corresponding equivalence class represented by $\mathcal{P}$. Notice that for every circle mark in $\mathcal{P}$, the mark is a tail in one of the $\mathcal{I}$-MAGs while it is an arrowhead in the other. Hence, the orientations are complete. If the observational distribution is known, then consider the graph $\mathcal{D}_2$ in Fig. 1b with $\mathcal{I}_2^* = \langle \emptyset, \{X, W\} \rangle$. The corresponding $\mathcal{I}$-MAG is the one in Fig. 1e, so we obtain the same result.*

To understand the challenge of establishing Ψ-FCI's completeness, denote by $\mathcal{M}'$ a complete orientation of $\mathcal{P}_{\Psi\text{-FCI}}$. Note that even though $\mathcal{M}'$ may satisfy the three conditions of Thm. 1 with respect to the true $\mathcal{I}$-MAG $\mathcal{M}$, it is not implied that $\mathcal{M}'$ is a valid $\mathcal{I}$-MAG. Following the further requirements of Def. 5, we show next that there exists a pair $\langle \mathcal{D}', \mathcal{I}' \rangle$ such that $MAG(Aug_{\mathcal{I}'}(\mathcal{D}')) = \mathcal{M}'$.

Lemma 1. *Let $\mathcal{D}(\mathbf{V} \cup \mathbf{L}, \mathbf{E})$ denote a causal graph, $\mathcal{I}$ denote a set of interventional targets, $\mathcal{M} = MAG(Aug_{\mathcal{I}}(\mathcal{D}))$, and $\mathcal{M}'$ denote an arbitrary MAG over $\mathbf{V} \cup \mathcal{F}$. If the following holds:*

1. *All the edges incident on $\mathcal{F}$ in $\mathcal{M}'$ are out of $\mathcal{F}$; and,*
2. *$\mathcal{M}$ and $\mathcal{M}'$ share the same separation statements over $\mathbf{V} \cup \mathcal{F}$,*

then there exists a pair $\langle \mathcal{D}', \mathcal{I}' \rangle$, including when $\emptyset \in \mathcal{I}$ and is fixed, such that $MAG(Aug_{\mathcal{I}'}(\mathcal{D}')) = \mathcal{M}'$. In other words, $\mathcal{M}'$ is an $\mathcal{I}$-MAG and the pair $\langle \mathcal{D}', \mathcal{I}' \rangle$ is Ψ-Markov equivalent to $\langle \mathcal{D}, \mathcal{I} \rangle$.

Based on this result, completeness can be finally proved as shown next.

Theorem 3 (Ψ-FCI Completeness)**.** *Assuming tuple $\mathbf{P}$ is generated by unknown pair $\langle \mathcal{D}, \mathcal{I} \rangle$, then Ψ-FCI is complete, i.e., $\mathcal{P}$ contains all the common edge marks in the Ψ-Markov equivalence class.*

A few compelling connections emerge from this proposition. Leveraging Corollary 1, one can show that a variant of Ψ-FCI constrained to Meek's rules (which we called Ψ-PC) is also complete in the Markovian case for both known and unknown interventional targets. On the other hand, perhaps surprisingly, it can also be shown that the same result does not hold when sufficiency cannot be ascertained. For a more detailed discussion on these subtleties, see [13, Appendix C].

5 Conclusion

In this work, we investigated the problem of learning causal graphs with latent variables from a mixture of observational and interventional distributions with unknown interventional targets. We started by defining the Ψ-Markov property that connects a tuple of distributions with unknown targets to a pair of causal graph $\mathcal{D}$ and a corresponding possible interventional target set $\mathcal{I}$. Accordingly, two pairs $\langle \mathcal{D}_1, \mathcal{I}_1 \rangle$ and $\langle \mathcal{D}_2, \mathcal{I}_2 \rangle$ are said to be Ψ-Markov equivalent if they license the same tuples of distributions. Based on this refined equivalence relation, we derived a graphical characterization to evaluate whether two pairs are in the same Ψ-Markov equivalence class. Finally, we developed a sound and complete algorithm that recovers a Ψ-Markov equivalence class given a tuple of distributions. This work grounds the theoretical aspects of learning from unknown soft-interventions, thus, as we envision, paving the way for a new family of more robust and scalable methods that can address issues of computational and sample complexity, including score-based and approximation algorithms.

Broader Impact

Learning cause-and-effect relationships is one of the fundamental problems for various fields, including biology [28, 6], epidemiology [26], and economics [12]. The introduced characterization and algorithm provide a clear understanding on how to accomplish this task while leveraging interventional data, even when the interventional targets are unknown. Moreover, the proposed approach can be instrumental towards explainability in artificial intelligence, which has been a topic of increasing importance recently. On the other hand, performing experiments to obtain interventional data poses some ethical challenges, such as randomizing the smoking factor which would require forcing individuals to smoke. Therefore, such limitations and concerns should be taken into consideration.

Acknowledgments and Disclosure of Funding

Bareinboim and Jaber are supported in parts by grants from NSF IIS-1704352 and IIS-1750807 (CAREER). Kocaoglu and Shanmugam are supported by the MIT-IBM Watson AI Lab.

References

[1] Elias Bareinboim, Carlos Brito, and Judea Pearl. Local characterizations of causal bayesian networks. In *Graph Structures for Knowledge Representation and Reasoning (IJCAI)*, pages 1–17. Springer Berlin Heidelberg, 2012.

[2] N. Cartwright. *Hunting Causes and Using Them: Approaches in Philosophy and Economics*. Cambridge University Press, 2007.

[3] A Philip Dawid. Influence diagrams for causal modelling and inference. *International Statistical Review*, 70:161–189, 2002.

[4] Daniel Eaton and Kevin Murphy. Exact bayesian structure learning from uncertain interventions. In *Artificial intelligence and statistics*, pages 107–114, 2007.

[5] Frederich Eberhardt, Clark Glymour, and Richard Scheines. On the number of experiments sufficient and in the worst case necessary to identify all causal relations among n variables. In *Proceedings of the 21st Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 178–184, 2005.

[6] Kyle Kai-How Farh, Alexander Marson, Jiang Zhu, Markus Kleinewietfeld, William J Housley, Samantha Beik, Noam Shoresh, Holly Whitton, Russell JH Ryan, Alexander A Shishkin,

et al. Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature*, 518(7539):337–343, 2015.

[7] Dan Geiger, Thomas Verma, and Judea Pearl. d-separation: From theorems to algorithms. In *Machine Intelligence and Pattern Recognition*, volume 10, pages 139–148. Elsevier, 1990.

[8] AmirEmad Ghassami, Saber Salehkaleybar, Negar Kiyavash, and Elias Bareinboim. Budgeted experiment design for causal structure learning. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1724–1733. PMLR, 2018.

[9] 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.

[10] Alain Hauser and Peter Bühlmann. Two optimal strategies for active learning of causal networks from interventional data. In *Proceedings of Sixth European Workshop on Probabilistic Graphical Models*, 2012.

[11] Patrik O Hoyer, Dominik Janzing, Joris Mooij, Jonas Peters, and Bernhard Schölkopf. Nonlinear causal discovery with additive noise models. In *Proceedings of NIPS 2008*, 2008.

[12] Paul Hünermund and Elias Bareinboim. Causal inference and data-fusion in econometrics. *arXiv preprint arXiv:1912.09104*, 2019.

[13] Amin Jaber, Murat Kocaoglu, Karthikeyan Shanmugam, and Elias Bareinboim. Causal discovery from soft interventions with unknown targets: Characterization and learning. Technical report, R-67, Columbia CausalAI Lab, Department of Computer Science, Columbia University, 2020.

[14] Dominik Janzing, Joris Mooij, Kun Zhang, Jan Lemeire, Jakob Zscheischler, Povilas Daniušis, Bastian Steudel, and Bernhard Schölkopf. Information-geometric approach to inferring causal directions. *Artificial Intelligence*, 182-183:1–31, 2012.

[15] Nan Rosemary Ke, Olexa Bilaniuk, Anirudh Goyal, Stefan Bauer, Hugo Larochelle, Chris Pal, and Yoshua Bengio. Learning neural causal models from unknown interventions. *arXiv preprint arXiv:1910.01075*, 2019.

[16] Murat Kocaoglu, Alexandros G. Dimakis, Sriram Vishwanath, and Babak Hassibi. Entropic causal inference. In *AAAI'17*, 2017.

[17] Murat Kocaoglu, Amin Jaber, Karthikeyan Shanmugam, and Elias Bareinboim. Characterization and learning of causal graphs with latent variables from soft interventions. In *Advances in Neural Information Processing Systems*, pages 14346–14356, 2019.

[18] Murat Kocaoglu, Karthikeyan Shanmugam, and Elias Bareinboim. Experimental design for learning causal graphs with latent variables. In *Advances in Neural Information Processing Systems*, pages 7018–7028, 2017.

[19] Christopher Meek. Causal inference and causal explanation with background knowledge. In *Proceedings of the eleventh conference on uncertainty in artificial intelligence*, 1995.

[20] Joris M Mooij, Sara Magliacane, and Tom Claassen. Joint causal inference from multiple contexts. *arXiv preprint arXiv:1611.10351*, 2016.

[21] J. Pearl. *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann, San Mateo, CA, 1988.

[22] J. Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, New York, 2000. 2nd edition, 2009.

[23] Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(5):947–1012, 2016.

[24] Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: foundations and learning algorithms*. MIT press, 2017.

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

[26] James M Robins, Miguel Angel Hernan, and Babette Brumback. Marginal structural models and causal inference in epidemiology, 2000.

[27] Dominik Rothenhäusler, Christina Heinze, Jonas Peters, and Nicolai Meinshausen. Backshift: Learning causal cyclic graphs from unknown shift interventions. In *Advances in Neural Information Processing Systems*, pages 1513–1521, 2015.

[28] 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.

[29] Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, Prediction, and Search*. A Bradford Book, 2001.

[30] Chandler Squires, Yuhao Wang, and Caroline Uhler. Permutation-based causal structure learning with unknown intervention targets. *arXiv preprint arXiv:1910.09007*, 2019.

[31] Jin Tian and Judea Pearl. Causal discovery from changes. In *Proceedings of UAI2013*, 2013.

[32] Thomas Verma and Judea Pearl. An algorithm for deciding if a set of observed independencies has a causal explanation. In *Proceedings of the Eighth international conference on uncertainty in artificial intelligence*, 1992.

[33] J. Woodward, J.F. Woodward, and Oxford University Press. *Making Things Happen: A Theory of Causal Explanation*. Oxford scholarship online. Oxford University Press, 2003.

[34] Karren Yang, Abigail Katcoff, and Caroline Uhler. Characterizing and learning equivalence classes of causal DAGs under interventions. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pages 5541–5550. PMLR, 2018.

[35] Jiji Zhang. *Causal inference and reasoning in causally insufficient systems*. PhD thesis, Department of Philosophy, Carnegie Mellon University, 2006.

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

[37] Jiji Zhang. On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias. *Artificial Intelligence*, 172(16):1873–1896, 2008.

[38] Kun Zhang, Biwei Huang, Jiji Zhang, Clark Glymour, and Bernhard Schölkopf. Causal discovery from nonstationary/heterogeneous data: Skeleton estimation and orientation determination. In *IJCAI: Proceedings of the Conference*, volume 2017, page 1347, 2017.

"Causal Discovery From Soft Interventions with Unknown Targets: Characterization and Learning" – Supplementary Material

A Proofs of Section 3

Theorem 4 (CBN Invariances)**.** *Let $\mathcal{D} = (\mathbf{V} \cup \mathbf{L}, \mathbf{E})$ denote the causal graph of a CBN, and let $\mathbf{P}$ be a tuple of interventional distributions generated by $\mathcal{D}$. Then, the following distributional invariances hold for disjoint $\mathbf{Y}, \mathbf{Z}, \mathbf{W} \subseteq \mathbf{V}$:*

(a) For $P_{\mathbf{I}} \in \mathbf{P}$: $P_{\mathbf{I}}(\mathbf{y}|\mathbf{w}, \mathbf{z}) = P_{\mathbf{I}}(\mathbf{y}|\mathbf{w})$ *if* $\mathbf{Y} \perp\!\!\!\perp \mathbf{Z} \,|\mathbf{W}$ *in* $\mathcal{D}$

(b) For $P_{\mathbf{I}}, P_{\mathbf{J}} \in \mathbf{P}$: $P_{\mathbf{I}}(\mathbf{y}|\mathbf{w}) = P_{\mathbf{J}}(\mathbf{y}|\mathbf{w})$ *if* $\mathbf{Y} \perp\!\!\!\perp \mathbf{K} \,|\mathbf{W} \setminus \mathbf{W}_{\mathbf{K}}$ *in* $\mathcal{D}_{\underline{\mathbf{W}_{\mathbf{K}}}, \overline{\mathbf{R}(\mathbf{W})}}$,

where $\mathbf{K} := \mathbf{I} \Delta \mathbf{J}$, $\mathbf{W}_{\mathbf{K}} =: \mathbf{W} \cap \mathbf{K}$, $\mathbf{R} := \mathbf{K} \backslash \mathbf{W}_{\mathbf{K}}$, and $\mathbf{R}(\mathbf{W}) \subseteq \mathbf{R}$ are non-ancestors of $\mathbf{W}$ in $\mathcal{D}$.

Proof. Constraint (a) follows from the factorization of $P_{\mathbf{I}}(\mathbf{V})$ according to Equation 1 and the d-separation criterion [7, Thm. 2].

To prove (b), we construct a hypothetical CBN while modeling the intervention on each variable with an endogenous root node/variable. This is valid because we assume the soft interventions are triggered by exogenous agents that are not affected by any variable in $\mathcal{D}$. Let $\mathbf{I}^n, \mathbf{J}^n$ denote the set of nodes in $\mathbf{I}, \mathbf{J}$ without the mechanism identifiers. We augment $\mathcal{D}$ with the set of nodes $\mathcal{F} = \{F_i | V_i \in \mathbf{I}^n \cup \mathbf{J}^n\}$ and edges $\mathcal{E} = \{F_i \to V_i | F_i \in \mathcal{F}\}$. We refer to the constructed causal graph as $\mathcal{D}'$. For each variable V_i with F_i, we have a new set of parents $Pa_i' = Pa_i \cup \{F_i\}$. The distribution of $P(V_i|Pa_i')$ is given as follows where $P^j(V_i|Pa_i)$ is a unique conditional probability for each identifier j.

$$P(V_i|Pa_i') = \begin{cases} P(V_i|Pa_i), & \text{if } F_i = 0/idle \\ P^j(V_i|Pa_i), & \text{if } F_i = j. \end{cases}$$

Finally, each F_i has an arbitrary prior distributions over its domain. This induces a new distribution P' over $\mathbf{V} \cup \mathbf{L} \cup \mathcal{F}$ and P' factorizes according to $\mathcal{D}'$. Then, $P_{\mathbf{I}}(\mathbf{V})$ relates to P' as follows where we condition on every $F_i \in \mathcal{F}$ such that (1) $F_i =$ idle if $V_i \notin \mathbf{I}^n$ and (2) $F_i = k$ if $V_i^k \in \mathbf{I}$.

$$P_{\mathbf{I}}(\mathbf{V}) = \sum_{\mathbf{L}} P'(\mathbf{V} \cup \mathbf{L} | F_i = j, \dots) \tag{4}$$

A similar expression applies for $P_{\mathbf{J}}(\mathbf{V})$. Now, let $\mathbf{F}_{\mathbf{K}} = \{F_i | V_i \in \mathbf{I}\Delta\mathbf{J}\}$. If $(\mathbf{F}_{\mathbf{K}} \perp\!\!\!\perp \mathbf{Y} \,|\mathbf{W})_{\mathcal{D}'}$, then changing the conditioning values of $\mathbf{F}_{\mathbf{K}}$ is irrelevant to $\mathbf{Y}$ and we get $P_{\mathbf{I}}(\mathbf{y}|\mathbf{w}) = P_{\mathbf{J}}(\mathbf{y}|\mathbf{w})$.

Finally, let $\mathcal{G} = Aug_{\{\mathbf{I},\mathbf{J}\}}(\mathcal{D})$. By Prop. 1, the d-separation statement in Constraint (b) is equivalent to $(F_1 \perp\!\!\!\perp \mathbf{Y} \,|\mathbf{W})_{\mathcal{G}}$. The differences between $\mathcal{D}'$ and $\mathcal{G}$ are: (1) $\mathcal{D}'$ has the additional F nodes for $(\mathbf{I}^n \cup \mathbf{J}^n) \setminus (\mathbf{I}\Delta\mathbf{J})$, and (2) $\mathcal{G}$ merges $\mathbf{F}_{\mathbf{K}}$ in node F_1. It is easy to see that the differences do not affect the separation statement, so $(\mathbf{F}_{\mathbf{K}} \perp\!\!\!\perp \mathbf{Y} \,|\mathbf{W})_{\mathcal{D}'} \iff (F_1 \perp\!\!\!\perp \mathbf{Y} \,|\mathbf{W})_{\mathcal{G}}$. This concludes the proof. □

Proof of Proposition 1. This result has been established in [17, Proposition 1]. □

Proof of Theorem 1. **(If)** If $\mathcal{M}_1$ and $\mathcal{M}_2$ satisfy the three graphical conditions, then they entail the same separation statements ([37, Def. 5 & Prop. 2]). This in turn implies that the corresponding augmented graphs entail the same separation statements as well ([25, Theorem 4.18]). It follows by Proposition 1 that the pairs $\langle \mathcal{D}_1, \mathcal{I}_1 \rangle$ and $\langle \mathcal{D}_2, \mathcal{I}_2 \rangle$ impose the same set of graphical constraints mandated by Definition 2. Hence, a tuple of distributions $\mathbf{P}$ satisfies the Ψ-Markov property with respect to $\langle \mathcal{D}_1, \mathcal{I}_1 \rangle$ if and only if it satisfies the Ψ-Markov property with respect to $\langle \mathcal{D}_2, \mathcal{I}_2 \rangle$. Therefore, $\langle \mathcal{D}_1, \mathcal{I}_1 \rangle$ and $\langle \mathcal{D}_2, \mathcal{I}_2 \rangle$ satisfy Definition 3 and they are Ψ-Markov equivalent.

(Only if) This proof follows exactly as the necessity condition (only if) of [17, Theorem 2]. □

Proof of Corollary 1. In the absence of latent nodes, $Aug_{\mathcal{I}}(\mathcal{D}) = MAG(Aug_{\mathcal{I}}(\mathcal{D}))$ which justifies the use of the augmented graphs directly instead of the $\mathcal{I}$-MAGs. Note that the augmented graph is a special $\mathcal{I}$-MAG with no bi-directed edges. It follows that for every discriminating path for a node Y in $Aug_{\mathcal{I}}(\mathcal{D})$, Y has to be a non-collider. This trivially satisfies the third condition of Theorem 1. This concludes the proof. □

B Proofs of Section 4

Proof of Theorem 2 (soundness). Assuming the data is generated by some unknown pair $\langle \mathcal{D}, \mathcal{I} \rangle$, then $\mathbf{P}$ satisfies the Ψ-Markov property with respect to $\langle \mathcal{D}, \mathcal{I} \rangle$ by Theorem 4. This implies that the Ψ-PAG learned in Ψ-FCI over $\mathbf{P}$ is defined relative to $\langle \mathcal{D}, \mathcal{I} \rangle$. The correctness of Phase I of Ψ-FCI follows from Proposition 1 and the faithfulness assumption (Def. 7) with regard to the Ψ-Markov property in Def. 2. Orienting the unshielded colliders follows from the soundness of InvToSep(·) and recovering the separation sets (`SepSet`). Finally, $\mathcal{R}^+$ is valid by construction of every $\mathcal{I}$-MAG in the equivalence class. The FCI orientation rules are sound since an $\mathcal{I}$-MAG is a MAG as well and does not contain cycles or almost directed cycles. This concludes the proof. □

Proof of Lemma 1. First, we construct an interventional target set $\mathcal{I}^*$ such that the adjacencies of $\mathcal{F}$ in both $Aug_{\mathcal{I}^*}(\mathcal{D})$ and $\mathcal{M}$ are the same. We construct $\mathcal{I}^*$ as follows:

1. Let $\mathcal{I}^* = \langle \mathbf{I}^*_1, \ldots, \mathbf{I}^*_{|\mathcal{I}|} \rangle$ and $\mathbf{I}^*_i = \{V^i_1, \ldots, V^i_{|\mathbf{V}|}\}$ where V^i_j denotes an intervention on V_j with a mechanism unique for $\mathbf{I}^*_i$ as opposed to the other interventional targets.
2. Choose any pair of nodes $F_k \in \mathcal{F}, V_p \in \mathbf{V}$ that are adjacent in $Aug_{\mathcal{I}^*}(\mathcal{D})$ but not adjacent in $\mathcal{M}$. For $\mathbf{I}^*_i$,$\mathbf{I}^*_j$ such that $\mathbf{I}^*_i \Delta \mathbf{I}^*_j = \mathbf{K}_k$, switch the mechanism V^i_p to V^j_p in $\mathbf{I}^*_i$ and in every interventional target that has the mechanism V^i_p.
3. Repeat step 2 until it is not applicable anymore.

Step 1 initializes $\mathcal{I}^*$ such that, in $Aug_{\mathcal{I}^*}(\mathcal{D})$, every node in $\mathcal{F}$ is adjacent to all the nodes in $\mathbf{V}$. Step 2 is applied recursively to remove adjacencies of F-nodes in $Aug_{\mathcal{I}^*}(\mathcal{D})$ that are not present in $\mathcal{M}$. Note that Step 2 does not add any adjacencies to $Aug_{\mathcal{I}^*}(\mathcal{D})$. It is left to show that Step 2 does not remove a required adjacency that is in $\mathcal{M}$.

For the sake of contradiction, consider the first iteration of Step 2 such that a required adjacency is lost and let it be between the pair F_l and V_m where F_l is adjacent to $\mathbf{K}_l = \mathbf{I}^*_a \Delta \mathbf{I}^*_b$, following the construction in Def. 4. Obviously, $\mathbf{I}^*_a, \mathbf{I}^*_b$ are not the modified targets in Step 2 since F_l and V_m do not satisfy the condition. Then, there exists a subset of $\mathcal{I}^*$, denoted $\langle \mathbf{I}^*_a, \mathbf{I}^*_i, \mathbf{I}^*_j, \mathbf{I}^*_b \rangle$, such that (1) $\mathbf{I}^*_i, \mathbf{I}^*_j$ are the pair of targets in which the mechanism of V_m is modified to be the same in Step 2, and (2) the mechanism of V_m in the pairs $\mathbf{I}^*_a, \mathbf{I}^*_i$ and $\mathbf{I}^*_j, \mathbf{I}^*_b$ are the same from previous iterations of Step 2 which makes them consistent (required to be the same mechanisms). If $\mathbf{I}^*_a, \mathbf{I}^*_j$ or $\mathbf{I}^*_i, \mathbf{I}^*_b$ require the same mechanism for V_m so that it is not adjacent to the corresponding F-node, then we obtain a triple $\langle \mathbf{I}^*_{h_1}, \mathbf{I}^*_{h_2}, \mathbf{I}^*_{h_3} \rangle$, where $h_1 = a$, $h_2 = i$ (or $h_2 = j$), and $h_3 = b$, and all the targets have the same mechanism for V_m; yet, the mechanism is only required to be different between $\mathbf{I}^*_a$ and $\mathbf{I}^*_b$ such that F_l and V_m are adjacent. Otherwise, $\mathbf{I}^*_a, \mathbf{I}^*_j$ (and $\mathbf{I}^*_i, \mathbf{I}^*_b$) require a different mechanism for V_m and we obtain a similar triple with $h_2 = j$ (or $h_2 = i$). Next, we prove that such a triple in $\mathcal{I}^*$ cannot exist.

Now, the required difference in the mechanism of V_m between $\mathbf{I}^*_{h_1}$ and $\mathbf{I}^*_{h_3}$ is due to the adjacency of F_l and V_m in $\mathcal{M}$. If this adjacency is in $Aug_{\mathcal{I}}(\mathcal{D})$, then the difference in the mechanism of V_m between $\mathbf{I}^*_{h_1}$ and $\mathbf{I}^*_{h_3}$ exists in $\mathcal{I}$. This leads to a contradiction as the inconsistent pattern of mechanisms in the subset of $\mathcal{I}^*$ would exist in $\mathcal{I}$ as well; which is impossible.

Alternatively, F_l and V_m are not adjacent in $Aug_{\mathcal{I}}(\mathcal{D})$ but they are adjacent in $\mathcal{M}$. It follows that F_l is adjacent to some node V_s in $Aug_{\mathcal{I}}(\mathcal{D})$ and there is an inducing path between F_l and V_m through V_s. Consider the interventional targets $\langle \mathbf{I}_{h_1}, \mathbf{I}_{h_2}, \mathbf{I}_{h_3} \rangle$ in $\mathcal{I}$ which correspond to $\langle \mathbf{I}^*_{h_1}, \mathbf{I}^*_{h_2}, \mathbf{I}^*_{h_3} \rangle$ in $\mathcal{I}^*$. The mechanism of V_s in $\mathbf{I}_{h_1}, \mathbf{I}_{h_3}$ is different based on the previous observation. Hence, the mechanism of V_s in $\mathbf{I}_{h_2}$ must be different than $\mathbf{I}_{h_1}$ or $\mathbf{I}_{h_3}$. Without loss of generality, let the mechanism of V_s in $\mathbf{I}_{h_2}$ be different than that in $\mathbf{I}_{h_1}$. Thus, the corresponding F-node of $\mathbf{I}_{h_1}, \mathbf{I}_{h_2}$ is adjacent to V_s in $Aug_{\mathcal{I}}(\mathcal{D})$, and consequently adjacent to V_m in $\mathcal{M}$ due to the inducing path. But, V_m has the same mechanism in

$\mathbf{I}^*_{h_1}, \mathbf{I}^*_{h_2}$ consistently with the absence of an adjacency between the corresponding F-node and V_m in $\mathcal{M}$; a contradiction. Therefore, the given algorithm to construct $\boldsymbol{\mathcal{I}}^*$ is sound.

If $\emptyset \in \boldsymbol{\mathcal{I}}$ for some $\mathbf{I}_r$ (could be more than one) and it is fixed/known, then apply the following adjustment to $\boldsymbol{\mathcal{I}}^*$. For every $V^j_i \in \mathbf{I}^*_t$ such that $V^j_i \in \mathbf{I}^*_r$, remove V^j_i from $\mathbf{I}^*_t$. Note this leads to $\mathbf{I}^*_r = \emptyset$. This gives a valid construction for $\boldsymbol{\mathcal{I}}^*$ where the observational target set is fixed relative to $\boldsymbol{\mathcal{I}}$.

Finally, consider the pair $\langle \mathcal{D}^*, \boldsymbol{\mathcal{I}}^* \rangle$, where $\mathcal{D}^*$ is the induced subgraph of $\mathcal{M}'$ over $\mathbf{V}$ while replacing every bidirected edge $i \leftrightarrow j$ with $i \leftarrow U_{i,j} \rightarrow j$, and $\boldsymbol{\mathcal{I}}^*$ is the set of interventional targets constructed earlier. It is easy to show that $MAG(Aug_{\boldsymbol{\mathcal{I}}^*}(\mathcal{D}^*)) = \mathcal{M}'$. First, $MAG(\mathcal{D}^*) = \mathcal{M}'_{\mathbf{V}}$, the induced subgraph of $\mathcal{M}'$ over $\mathbf{V}$. Second, $\boldsymbol{\mathcal{I}}^*$ generates F-nodes and corresponding adjacencies as those in $\mathcal{M}$ by construction. Also, $\mathcal{M}$ and $\mathcal{M}'$ share the same skeleton following the second condition of the lemma at hand, and the edges incident on the F-nodes in $\mathcal{M}'$ are out of them by the first condition. Hence, the F-node adjacencies in $Aug_{\boldsymbol{\mathcal{I}}^*}(\mathcal{D}^*)$ are the same as $\mathcal{M}$, and consequently $\mathcal{M}'$. Third, there are no inducing paths between non-adjacent nodes in $Aug_{\boldsymbol{\mathcal{I}}^*}(\mathcal{D}^*)$ except for $i \leftarrow U_{i,j} \rightarrow j$. Assume for the sake of contradiction that such an inducing path exists. Then, the same path, while replacing $i \leftarrow U_{i,j} \rightarrow j$ with $\leftrightarrow$, would exist in $\mathcal{M}'$. However, this is not possible since $\mathcal{M}'$ is a MAG (maximal property); a contradiction. It follows that $MAG(Aug_{\boldsymbol{\mathcal{I}}^*}(\mathcal{D}^*)) = \mathcal{M}'$ and $\mathcal{M}'$ is an $\boldsymbol{\mathcal{I}}$-MAG. Therefore, $\langle \mathcal{D}^*, \boldsymbol{\mathcal{I}}^* \rangle$ is Ψ-Markov equivalent to $\langle \mathcal{D}, \boldsymbol{\mathcal{I}} \rangle$ as their corresponding $\boldsymbol{\mathcal{I}}$-MAGs satisfy the three conditions of Theorem 1 by the second condition of the current lemma. □

B.1 Ψ-FCI (Algorithm 1) Completeness

To prove completeness of the orientations, we need to establish that none of the circle marks left in a Ψ-PAG output by Ψ-FCI, denoted $\mathcal{P}_{\Psi\text{-FCI}}$, hide an arrowhead or tail orientation that is common for all the $\boldsymbol{\mathcal{I}}$-MAGs in the equivalence class. The result is established in two phases. The first one proves that the arrowheads in $\mathcal{P}_{\Psi\text{-FCI}}$ are complete while the second proves it for the tails.

B.1.1 Arrowhead Completeness

This proof follows the same procedure in [37] to prove the completeness of the PAG (FCI output). We start with the following graphical property which is crucial for the completeness proof.

Lemma 2. *In $\mathcal{P}_{\Psi\text{-FCI}}$, the following property holds:*

for any three nodes A, B, C, if $A *\!\!\rightarrow B \circ\!\!-\!\!* C$, then there is an edge between A and C with an arrowhead at C, namely, $A *\!\!\rightarrow C$. Furthermore, if the edge between A and B is $A \rightarrow B$, then the edge between A and C is either $A \rightarrow C$ or $A \circ\!\!\rightarrow C$ (i.e., it is not $A \leftrightarrow C$).

Proof. For the sake of contradiction, suppose the property does not hold and there exists at least one triple $\langle A, B, C \rangle$ such that the edge between A and B is into B and the edge between B and C has a circle incident on B, but the consequent does not hold. First, it is easy to see that A and C have to be adjacent, otherwise the circle mark incident on B would be oriented by $\mathcal{R}_0$ or $\mathcal{R}_1$. The rest of the proof extends that of [37, Lemma A.1] which exhausts the orientation rules ($\mathcal{R}_0 - \mathcal{R}_4$) that could orient the edge between A and B into B while violating the property, and reaches a contradiction in all. In the last case, suppose $A *\!\!\rightarrow B$ is oriented by $\mathcal{R}^+$ which means that A is an F node. It trivially follows by $\mathcal{R}^+$ that the edge between A and C is out of A and into C which is a contradiction. This concludes the proof. □

The following property follows from Lemma 2 with exactly the same proof of [35, Lemma 3.3.2].

Lemma 3. *In $\mathcal{P}_{\Psi\text{-FCI}}$, for any two nodes A and B, if there is a circle path, i.e., a path consisting of $\circ\!\!-\!\!\circ$ edges, between A and B, then:*

- *(i) if there is an edge between A and B, the edge is $A \circ\!\!-\!\!\circ B$.*
- *(ii) for any other node C, $C *\!\!\rightarrow A$ if and only if $C *\!\!\rightarrow B$. Furthermore, $C \leftrightarrow A$ if and only if $C \leftrightarrow B$.*

Proof. The proof follows that of [35, Lemma 3.3.2]. □

We refer to the subgraph of $\mathcal{P}_{\Psi\text{-FCI}}$ consisting of circle edges ($\circ\!\!-\!\!\circ$) as $\mathcal{P}^C$. We have the following property.

Lemma 4. *For every $A \circ\!\!-\!\!\circ B$ in $\mathcal{P}^C$, $\mathcal{P}^C$ can be oriented into a DAG with no unshielded colliders in which $A \to B$ appears, and can also be oriented into a DAG with no unshielded colliders in which $A \leftarrow B$ appears.*

Proof. The work in [19] shows that all chordal undirected graphs have the desired property. Hence, it is sufficient for the sake of establishing the property to show that $\mathcal{P}^C$ is chordal. Suppose for the sake of contradiction that there is a chordless cycle of four or more circle edges in $\mathcal{P}^C$. Let $\langle V_0, V_1, V_2, V_3, \ldots, V_0\rangle$ be the shortest such cycle. Note that no two non-consecutive nodes on the cycle are adjacent in $\mathcal{P}_{\Psi\text{-FCI}}$ as the edge would have to be a circle edge by Lemma 3 which creates a shorter cycle. This leads to a contradiction as Ψ-FCI would have detected at least one collider by $\mathcal{R}_0$ along this undirected cycle; otherwise, we have a directed cycle in the causal graph and the corresponding $\mathcal{I}$-MAG which is not possible. This concludes the proof. □

Using the above properties, we show that the following procedure generates a special $\mathcal{I}$-MAG that is in the equivalence class of a given $\mathcal{P}_{\Psi\text{-FCI}}$ output by Ψ-FCI.

Lemma 5. *Let $\mathcal{H}$ be the result of applying the following procedure to $\mathcal{P}_{\Psi\text{-FCI}}$:*

1. *orient the circles on $\circ\!\!\to$ edges as tails; and*
2. *orient $\mathcal{P}^C$ into a DAG with no unshielded colliders.*

Then $\mathcal{H}$ is an $\mathcal{I}$-MAG in the equivalence class of $\mathcal{P}_{\Psi\text{-FCI}}$.

Proof. The proof extends that of [35, Lemma 3.3.4] which proves that the resulting graph, denoted $\mathcal{M}$, is a MAG and shares the same independence model as that of the true ($\mathcal{I}$-MAG) MAG. Note here the true $\mathcal{I}$-MAG is generated by some pair $\langle \mathcal{D}, \mathcal{I}\rangle$. Also, all the edges incident on the F-nodes in $\mathcal{M}$ are out of them since the orientations of $\mathcal{M}$ are consistent with $\mathcal{P}_{\Psi\text{-FCI}}$ which applies $\mathcal{R}^+$. It follows by Lemma 1 that $\mathcal{M}$ is an $\mathcal{I}$-MAG in the corresponding equivalence class. This concludes the proof. □

Finally, the construction in Lemma 5 along with the property in Lemma 4 establishes the following result of arrowhead completeness. This concludes the first phase of the completeness proof. It is left to prove that the tail marks are complete as well.

Proposition 2. *$\mathcal{P}_{\Psi\text{-FCI}}$ is arrowhead complete, i.e., each circle in $\mathcal{P}_{\Psi\text{-FCI}}$ is not an invariant arrowhead, and it is a tail in some $\mathcal{I}$-MAG in the equivalence class of $\mathcal{P}_{\Psi\text{-FCI}}$.*

Proof. This result follows from Lemmas 4 and 5. □

B.1.2 Tail Completeness

Lemmas 4 and 5 together establish that for every circle edge in $\mathcal{P}_{\Psi\text{-FCI}}$, a circle mark can be oriented as an arrowhead in some $\mathcal{I}$-MAG in the corresponding equivalence class. To prove the tail completeness, it is left to show that a circle mark on a partially directed edge $\circ\!\!\to$ can be oriented as an arrowhead in some $\mathcal{I}$-MAG in the equivalence class. This is established by the following proposition.

Proposition 3. *$\mathcal{P}_{\Psi\text{-FCI}}$ is tail complete, i.e., each circle in $\mathcal{P}_{\Psi\text{-FCI}}$ is not an invariant tail, and it is a arrowhead in some $\mathcal{I}$-MAG in the equivalence class of $\mathcal{P}_{\Psi\text{-FCI}}$.*

Proof. This proof follows almost exactly the same procedure in [37, Subsection 4.2] to establish the analogous result in PAGs, while assuming the absence of selection bias which rules out the orientation rules $\mathcal{R}_5 - \mathcal{R}_7$ and the presence of undirected edges in the $\mathcal{I}$-MAGs. For each MAG $\mathcal{M}$ constructed in the proof, we use the utility result in Lemma 5 to establish that $\mathcal{M}$ is in fact an $\mathcal{I}$-MAG generated by a pair $\langle \mathcal{D}, \mathcal{I}\rangle$ that is in the Ψ-Markov equivalence class. □

Proof of Theorem 3. This follows from Propositions 2 and 3. □

Algorithm 2 Ψ-PC: Algorithm for Learning a Ψ-PDAG

Input: Tuple of distributions $\mathbf{P} = \langle P_1, \ldots, P_m \rangle$
Output: Ψ-PDAG $\mathcal{P}$

$\mathcal{F} \leftarrow \emptyset, k \leftarrow 0, \sigma : \mathbb{N} \rightarrow |\mathbf{P}| \times |\mathbf{P}|$
for all pairs $P_i, P_j \in \mathbf{P}$ **do** $k \leftarrow k + 1, \mathcal{F} \leftarrow \mathcal{F} \cup \{F_k\}, \sigma(k) \leftarrow (i, j)$
Phase I: Skeleton
Form a complete graph $\mathcal{P}$ over $\mathbf{V} \cup \mathcal{F}$ with $\circ\!\!-\!\!\circ$ edges between every pair of nodes.
for every pair $X, Y \in \mathbf{V} \cup \mathcal{F}$ **do**
    if $X \in \mathcal{F} \wedge Y \in \mathcal{F}$ **then** $\texttt{SepSet}(X, Y) \leftarrow \emptyset$, $\texttt{SepFlag} \leftarrow True$
    else $(\texttt{SepSet}(X, Y), \texttt{SepFlag}) \leftarrow \textsc{InvToSep}(\mathbf{P}, X, Y, \mathbf{V}, \mathcal{F}, \sigma)$
    if $\texttt{SepFlag} = True$ **then** Remove the edge between X, Y in $\mathcal{P}$.
Phase II: Unshielded Colliders
$\mathcal{R}_0$: For every unshielded triple $\langle X, Z, Y \rangle$ in $\mathcal{P}$, orient it as $X *\!\!\rightarrow Z \leftarrow\!\!* Y$ iff $Z \notin \texttt{SepSet}(X, Y)$
Phase III: Orientation Rules
Rule $\mathcal{R}^+$*:* For any $F_k \in \mathcal{F}$, orient adjacent edges out of F_k.
Apply the three orientation rules from [19], shown in Fig. 3, until none applies.

C Complete Algorithm for Causally Sufficient Models

In this section, we present a sound and complete algorithm to learn an equivalence class of causal graphs under causal sufficiency from interventional data with unknown interventional targets.

Definition 8 (Ψ-PDAG)**.** *Given a pair of causal graph with no latents and interventional target, $\langle \mathcal{D}, \mathcal{I} \rangle$, let $\mathcal{G} = Aug_{\mathcal{I}}(\mathcal{D})$, and let $[\mathcal{G}]$ be the set of augmented graphs corresponding to all the pairs $\langle \mathcal{D}', \mathcal{I}' \rangle$ that are Ψ-Markov equivalent to $\langle \mathcal{D}, \mathcal{I} \rangle$. The Ψ-PDAG for $\langle \mathcal{D}, \mathcal{I} \rangle$, denoted $\mathcal{P}$, is a graph such that:*

1. *$\mathcal{P}$ has the same adjacencies as $\mathcal{G}$, and any member of $[\mathcal{G}]$ does; and*
2. *every non-circle mark (tail or arrowhead) in $\mathcal{P}$ is an invariant mark in $[\mathcal{G}]$.*

The algorithm is named Ψ-PC, following the PC algorithm for the Markov equivalence class, and is shown in Alg. 2. Ψ-PC is almost identical to Ψ-FCI except for the orientation phase where we apply the three rules in Fig. 3 from Meek [19] instead of the FCI rules. The soundness of Ψ-PC follows from the completeness of Ψ-FCI and [19, Thm. 2]. Our main objective is to prove the completeness of Ψ-PC under causal sufficiency akin to Ψ-FCI. This claim is established in Thm. 5 below.

Theorem 5 (Ψ-PC Completeness)**.** *Assuming tuple $\mathbf{P}$ is generated by unknown pair $\langle \mathcal{D}, \mathcal{I} \rangle$, then Ψ-PC is complete, i.e., $\mathcal{P}$ contains all the common edge marks in the Ψ-Markov equivalence class.*

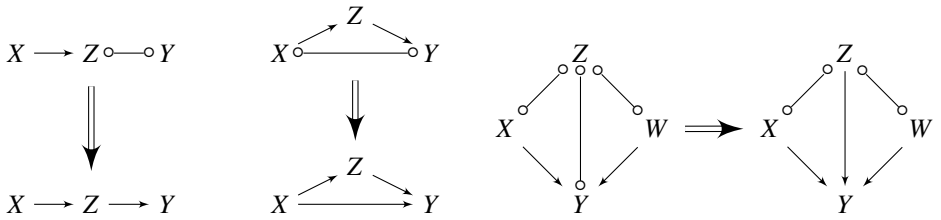
Proof. Let $\mathcal{P}^C$ denote the subgraph of $\mathcal{P}_{\Psi\text{-PC}}$ consisting of undirected/circle edges and their corresponding nodes. By Lemma 6, it is easy to show that $\mathcal{P}^C$ is chordal using a proof similar to that of Lemma 4. It follows by [19, Lemma 5] that we can orient $\mathcal{P}^C$ as a DAG with no unshielded colliders with any circle edge $A \circ\!\!-\!\!\circ B$ oriented as $A \rightarrow B$ in one and $A \leftarrow B$ in another.

Consider $\mathcal{G}'$ to be any fully oriented graph starting from $\mathcal{P}_{\Psi\text{-PC}}$ and consistent with the previous property. Assume for the sake of contradiction that $\mathcal{G}'$ contains a cycle and consider the shortest one denoted p. Then, we must have $A \rightarrow B \circ\!\!-\!\!\circ C$ along p in $\mathcal{P}_{\Psi\text{-PC}}$. By Lem. 6, we have $A \rightarrow C$ and we obtain a shorter cycle in $\mathcal{G}'$; contradiction. Similarly, assume $\mathcal{G}'$ contains an unshielded collider that is not in $\mathcal{P}_{\Psi\text{-PC}}$. Then, we must have $A \rightarrow B \circ\!\!-\!\!\circ C$ in $\mathcal{P}_{\Psi\text{-PC}}$ such that A is not adjacent to C and $B \circ\!\!-\!\!\circ C$ is oriented $B \leftarrow C$ in $\mathcal{G}'$. This is not possible as A, C are adjacent in $\mathcal{P}_{\Psi\text{-PC}}$ by Lem. 6. Hence, $\mathcal{G}'$ is a DAG with the same skeleton and unshielded colliders as the true augmented graph $\mathcal{G}$.

Finally, let $\mathcal{D}'$ denote the induced subgraph of $\mathcal{G}'$ over $\mathbf{V}$. In the absence of latents, it is easy to see that $Aug_{\mathcal{I}}(\mathcal{D}') = \mathcal{G}'$, where $\langle \mathcal{D}, \mathcal{I} \rangle$ is the true pair. It follows by Corollary 1 that $\langle \mathcal{D}', \mathcal{I} \rangle$ is Ψ-Markov equivalent to $\langle \mathcal{D}, \mathcal{I} \rangle$ and $\mathcal{G}' \in [\mathcal{G}]$. This concludes the proof. □

Lemma 6. *In $\mathcal{P}_{\Psi\text{-PC}}$, the output of Ψ-FCI, the following property holds: if $A \rightarrow B \circ\!\!-\!\!\circ C$, then $A \rightarrow C$.*

Proof. Suppose for the sake of contradiction that there is a triple A, B, C such that the above property does not hold. If A and C are not adjacent, then $B \circ\!\!-\!\!\circ C$ will be oriented by $\mathcal{R}_1$. Also, we can't have



(a) Rule $\mathcal{R}_1$. (b) Rule $\mathcal{R}_2$. (c) Rule $\mathcal{R}_3$.

Figure 3: The three orientation rules used in Ψ-PC.

$A \leftarrow C$ as $B \circ\!\!-\!\!\circ C$ will be oriented by $\mathcal{R}_2$. The rest of the proof extends that of [19, Lemma 1] where we consider every orientation rule that could orient $A \rightarrow B$ and reach a contradiction in each. Last option is $A \rightarrow B$ is oriented by $\mathcal{R}^+$. This means that A is an F-node and we also have $A \rightarrow C$, which is a contradiction to the initial assumption. This concludes the proof. □

It is relevant to note that Meek [19] introduces a fourth rule that is only applicable when background knowledge is available. The additional rule is necessary to prove the completeness of the PDAG for the Markov equivalence class with arbitrary background knowledge. However, given the restricted nature of the background knowledge we have, represented by $\mathcal{R}^+$, this rule turns out to be not needed, i.e., not applicable.

C.1 Known Interventional Targets

Given a pair $\langle \mathcal{D}, \mathcal{I} \rangle$, we can see by Corol. 1 that the corresponding Ψ-Markov equivalence class with $\mathcal{I}$ known/fixed ($\mathcal{I}$-Markov) is a subset of the Ψ-Markov equivalence class with unknown interventional targets. Hence, the completeness result for unknown interventional targets raises the question of whether the algorithm is also complete for known interventional targets. The following lemma is key to answer this question.

Lemma 7. *Let $\langle \mathcal{D}_1, \mathcal{I}_1 \rangle$ and $\langle \mathcal{D}_2, \mathcal{I}_2 \rangle$ be two arbitrary pairs, each with a causal graph with no latents and an interventional target sets. If $\langle \mathcal{D}_1, \mathcal{I}_1 \rangle$ is Ψ-Markov equivalent to $\langle \mathcal{D}_2, \mathcal{I}_2 \rangle$, then $\langle \mathcal{D}_1, \mathcal{I}_1 \rangle$ is Ψ-Markov equivalent to $\langle \mathcal{D}_2, \mathcal{I}_1 \rangle$ and $\langle \mathcal{D}_1, \mathcal{I}_2 \rangle$ is Ψ-Markov to $\langle \mathcal{D}_2, \mathcal{I}_2 \rangle$.*

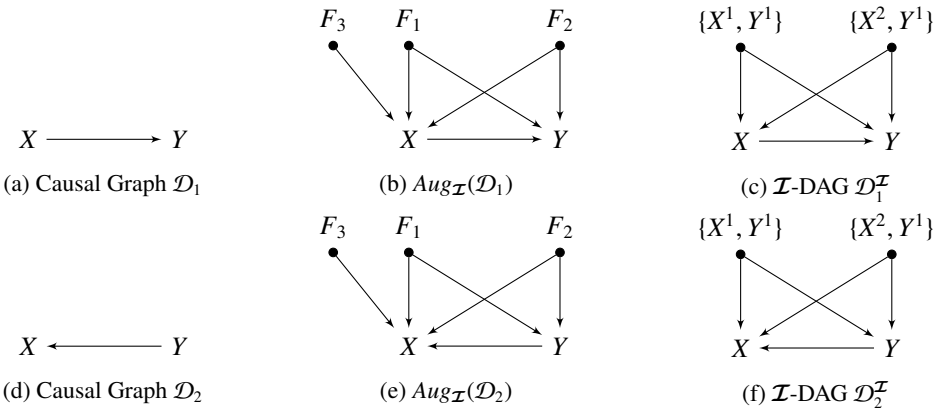
Proof. If $\langle \mathcal{D}_1, \mathcal{I}_1 \rangle$ is Ψ-Markov equivalent to $\langle \mathcal{D}_2, \mathcal{I}_2 \rangle$, then $Aug_{\mathcal{I}_1}(\mathcal{D}_1)$ and $Aug_{\mathcal{I}_2}(\mathcal{D}_1)$ share the same skeleton and unshielded colliders (Corol. 1). By Def. 4, it is easy to see that $Aug_{\mathcal{I}_2}(\mathcal{D}_1)$ and $Aug_{\mathcal{I}_1}(\mathcal{D}_2)$ share the same skeleton and unshielded colliders with $Aug_{\mathcal{I}_1}(\mathcal{D}_1)$ and $Aug_{\mathcal{I}_2}(\mathcal{D}_1)$. Hence, the result in the lemma follows. □

In words, Lemma 7 establishes that the size, i.e., number of causal graphs, of a Ψ-Markov equivalence class under unknown interventional targets is the same as that with known interventional targets. This implies that knowing the interventional targets does not have an impact on the causal discovery task. Hence, the next result follows easily.

Theorem 6 (Completeness with Known Targets)**.** *Assuming tuple* **P** *is generated by unknown $\mathcal{D}$ with known interventional targets $\mathcal{I}$, then Ψ-PC is complete, i.e., $\mathcal{P}$ contains all the common edge marks in the Ψ-Markov equivalence class.*

Proof. This follows from Lemma 7 and Theorem 5. □

An important remark here is that Lemma 7 does not hold in the presence of latents. For instance, the pairs $\langle \mathcal{D}_1, \mathcal{I}_1 \rangle$ and $\langle \mathcal{D}_2, \mathcal{I}_2 \rangle$ in Fig. 1a and 1b are Ψ-Markov equivalent as shown in Ex. 5. However, $\langle \mathcal{D}_2, \mathcal{I}_1 \rangle$ is not Ψ-Markov equivalent to $\langle \mathcal{D}_1, \mathcal{I}_1 \rangle$. This is also evident by the additional orientation rule in [17, Alg. 1, Rule 9] which is not sound for unknown interventional targets. Hence, the question of completeness for causal discovery from interventional data with known interventional targets remains open.



(a) Causal Graph $\mathcal{D}_1$ (b) $Aug_{\mathcal{I}}(\mathcal{D}_1)$ (c) $\mathcal{I}$-DAG $\mathcal{D}_1^{\mathcal{I}}$

(d) Causal Graph $\mathcal{D}_2$ (e) $Aug_{\mathcal{I}}(\mathcal{D}_2)$ (f) $\mathcal{I}$-DAG $\mathcal{D}_2^{\mathcal{I}}$

Figure 4: Two causal graph $\mathcal{D}_1$ and $\mathcal{D}_2$ with the interventional target set $\boldsymbol{\mathcal{I}} = \langle \emptyset, \{X^1, Y^1\}, \{X^2, Y^1\} \rangle$.

D Connections with Previous Work

In this section, we explore how our framework compares with some of the related work in the area of causal discovery from interventional data. In general, we show that the proposed work identifies equivalence classes that are strictly smaller than the ones recovered by other methods.

D.1 Yang et al. [34] & Squires et al. [30]

The work in [34] investigates learning equivalence classes of causal graphs without latents from interventional data with known interventional targets. They use interventional nodes, which are analogous to F-nodes in our approach, to graphically characterize causal graphs that are interventionally equivalent (i.e., $\boldsymbol{\mathcal{I}}$-Markov equivalent). We note they do not encode the different mechanisms for variables within each interventional target in $\boldsymbol{\mathcal{I}}$. The work in [30] builds on this characterization and develop a learning algorithm under unknown interventional targets. First, we consider the case when the observational distribution is available, i.e., $\emptyset \in \boldsymbol{\mathcal{I}}$, as this is handled separately in [34, Thm. 3.9]. The next example shows that our equivalence class is strictly more informative under causal sufficiency (Corol. 1) and known interventional targets.

Example 8. *Consider the causal graphs in Figs. 4a, 4d, denoted by $\mathcal{D}_1, \mathcal{D}_2$, respectively, and the known set of interventional targets $\boldsymbol{\mathcal{I}} = \langle \emptyset, \{X^1, Y^1\}, \{X^2, Y^1\} \rangle$. Note X is intervened on with different mechanisms (X^1, X^2) while Y is intervened on with the same mechanism (Y^1). The graphical characterization of [34, Thm. 3.9] does not encode different mechanisms, and appends graph $\mathcal{D}$ with one interventional node per interventional target, excluding the observational ($\emptyset$), and adds directed edges from the interventional nodes to the corresponding interventional targets. The constructed graph is referred to as $\boldsymbol{\mathcal{I}}$-DAG $\mathcal{D}^{\boldsymbol{\mathcal{I}}}$. Back to the example, the corresponding $\boldsymbol{\mathcal{I}}$-DAGs for $\mathcal{D}_1, \mathcal{D}_2$ are shown in Figs. 4c, 4f, respectively. The two graphs are said to be $\boldsymbol{\mathcal{I}}$-Markov equivalent as their corresponding $\boldsymbol{\mathcal{I}}$-DAGs share the same skeleton and unshielded colliders. On the other hand, our approach constructs the corresponding augmented graphs shown in Figs. 4b, 4e, following Def. 4, where F_3 maps to $\{X^1, Y^1\} \Delta \{X^2, Y^1\} = \{X\}$. By Corol. 1, $\mathcal{D}_1$ and $\mathcal{D}_2$ are not Ψ-Markov equivalent under the same interventional target set $\boldsymbol{\mathcal{I}}$ as the augmented graphs do not share the same unshielded colliders – $\langle F_3, X, Y \rangle$ is a collider in Fig. 4e while it is not in Fig. 4b. So, the separation statement $(F_3 \perp\!\!\!\perp Y)$ holds in $Aug_{\boldsymbol{\mathcal{I}}}(\mathcal{D}_2)$ but not in $Aug_{\boldsymbol{\mathcal{I}}}(\mathcal{D}_1)$. More fundamentally, this corresponds to testing for the invariance $P_{X^1,Y^1}(Y) = P_{X^2,Y^1}(Y)$, which holds for data generated by $\mathcal{D}_2$, but not for $\mathcal{D}_1$.*

In contrast to [34, Defs. 3.3& 3.6], the Ψ-Markov property tests for distributional invariances over variable Y between two interventional distributions even if Y is intervened on in both given that the mechanism is the same; e.g., $P_{X^1,Y^1}(Y) = P_{X^2,Y^1}(Y)$ in Example 8. More formally, the above example coupled with the following proposition establish that Corol. 1 subsumes that in [34, Thm. 3.9].

Proposition 4. *Consider two causal graphs without latents $\mathcal{G}_1, \mathcal{G}_2$, and a set of interventional targets $\boldsymbol{\mathcal{I}}$, where $\emptyset \in \boldsymbol{\mathcal{I}}$. The causal graphs $\mathcal{G}_1$ and $\mathcal{G}_2$ are $\boldsymbol{\mathcal{I}}$-Markov equivalent by [34, Thm. 3.14] if $\langle \mathcal{G}_1, \boldsymbol{\mathcal{I}} \rangle$ and $\langle \mathcal{G}_2, \boldsymbol{\mathcal{I}} \rangle$ are Ψ-Markov equivalent by Corol. 1.*

Proof. We prove the contrapositive of the statement. If $\mathcal{G}_1$ and $\mathcal{G}_2$ are not $\mathcal{I}$-Markov equivalent by [34, Thm. 3.14], then the corresponding $\mathcal{I}$-DAGs $\mathcal{G}_1^{\mathcal{I}}$ and $\mathcal{G}_2^{\mathcal{I}}$ do not have the same skeletons or they do not have the same unshielded colliders. By the construction of the $\mathcal{I}$-DAG in [34, Def. 3.5] and the augmented graph in Def. 4, it is easy to see that the interventional nodes and their adjacencies in $\mathcal{I}$-DAGs are a node-induced subset of the F-nodes and their adjacencies when $\emptyset \in \mathcal{I}$. Hence, $Aug_{\mathcal{I}}(\mathcal{G}_1)$ and $Aug_{\mathcal{I}}(\mathcal{G}_2)$ must have different skeletons or different unshielded colliders. □

Next, when the observational distribution is not available, i.e., $\emptyset \notin \mathcal{I}$, [34, Theorem 3.14] presents a generalization of Theorem 3.9 to handle sets of interventional targets that are *conservative*, i.e., $\forall V_i \in \mathbf{V}, \exists \mathbf{I} \in \mathcal{I}$ s.t. $V_i \notin \mathbf{I}$ for all mechanisms. Whenever the mechanism of each intervened node is different across the interventional targets, i.e., $\forall \mathbf{I}, \mathbf{J} \in \mathcal{I}$, if $V_i^j \in \mathbf{I}$ and $V_i^k \in \mathbf{J}$, then $j \neq k$, Prop. 5 below shows that the graphical characterization in [34, Thm. 3.14] and Corol. 1 are equivalent.

Proposition 5. *Consider two causal graphs without latents $\mathcal{G}_1, \mathcal{G}_2$, and a conservative set of interventional targets $\mathcal{I}$. If the mechanism changes for each intervened node are different across the interventional targets, then $\langle \mathcal{G}_1, \mathcal{I} \rangle$ and $\langle \mathcal{G}_2, \mathcal{I} \rangle$ are Ψ-Markov equivalent by Corol. 1 if and only if $\mathcal{G}_1$ and $\mathcal{G}_2$ are $\mathcal{I}$-Markov equivalent by [34, Thm. 3.14].*

Proof. (**only if**) Pick $\mathbf{I} \in \mathcal{I}$. Let $F_{\mathbf{I},\mathbf{J}}$ for $\mathbf{J} \in \mathcal{I} - \{\mathbf{I}\}$ be the F-node obtained for the pair of interventional targets $(\mathbf{I}, \mathbf{J})$. Consider the induced subgraph of $Aug_{\mathcal{I}}(\mathcal{G}_1)$ and $Aug_{\mathcal{I}}(\mathcal{G}_2)$ on the nodes $\mathbf{V} \cup \{F_{\mathbf{I},\mathbf{J}} : \mathbf{J} \in \mathcal{I} - \{\mathbf{I}\}\}$. Since $Aug_{\mathcal{I}}(\mathcal{G}_1)$ and $Aug_{\mathcal{I}}(\mathcal{G}_2)$ have the same skeleton and same unshielded colliders, these two induced subgraphs have the same skeleton and unshielded colliders as well. Since the mechanism changes are assumed to be different, then $\mathbf{I}\Delta\mathbf{J} = \mathbf{I} \cup \mathbf{J}$ (ignoring the variable mechanisms in the union) as per Def. 1. Hence, the two induced graphs are identical to $\mathcal{G}_1^{\tilde{\mathcal{I}}_{\mathbf{I}}}$ and $\mathcal{G}_2^{\tilde{\mathcal{I}}_{\mathbf{I}}}$, which implies $\mathcal{G}_1$ and $\mathcal{G}_2$ are $\mathcal{I}$-Markov equivalent. This concludes this direction of the proof.

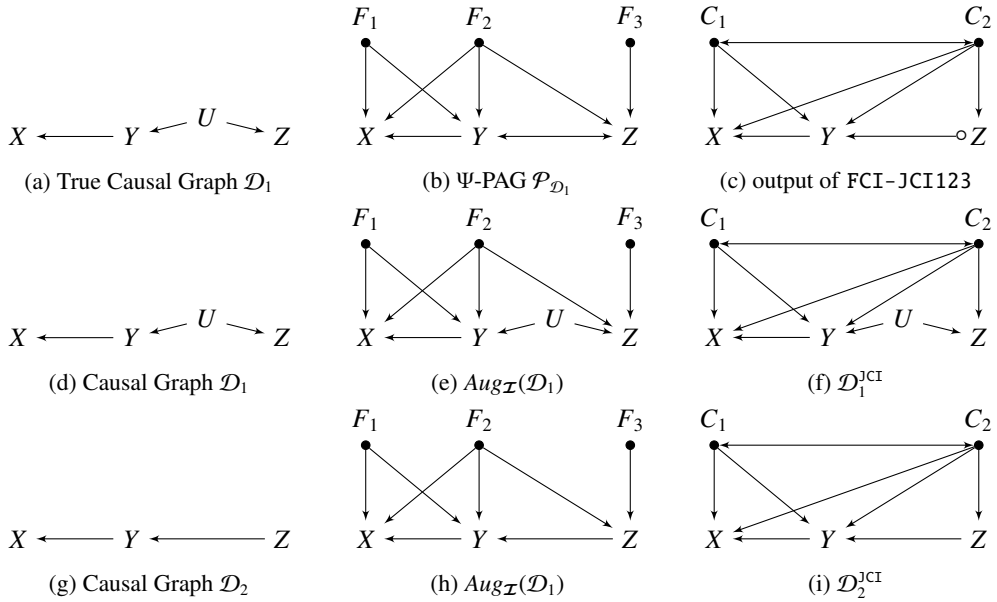
(**if**) Since $\mathcal{G}_1$ and $\mathcal{G}_2$ are $\mathcal{I}$-Markov equivalent, then $\mathcal{G}_1^{\tilde{\mathcal{I}}_{\mathbf{I}}}$ and $\mathcal{G}_2^{\tilde{\mathcal{I}}_{\mathbf{I}}}$ have the same skeleton and unshielded colliders for all $\mathbf{I} \in \mathcal{I}$ (by [34, Thm. 3.14]). From the argument above (in **only if**), the corresponding graph union of $\mathcal{G}_1^{\tilde{\mathcal{I}}_{\mathbf{I}}}$ and $\mathcal{G}_2^{\tilde{\mathcal{I}}_{\mathbf{I}}}$ across all $\mathbf{I} \in \mathcal{I}$ is $Aug_{\mathcal{I}}(\mathcal{G}_1)$ and $Aug_{\mathcal{I}}(\mathcal{G}_2)$, which must have the same skeleton as well. Furthermore, both augmented graphs must have the same set of unshielded colliders of the form $X \to Y \leftarrow Z$ for $X, Y, Z \in \mathbf{V}$, otherwise $\mathcal{G}_1^{\tilde{\mathcal{I}}_{\mathbf{I}}}$ and $\mathcal{G}_2^{\tilde{\mathcal{I}}_{\mathbf{I}}}$ would have different unshielded colliders for all $\mathbf{I}$. Also, by construction, $Aug_{\mathcal{I}}(\mathcal{G}_1)$ and $Aug_{\mathcal{I}}(\mathcal{G}_2)$ have the same set of unshielded colliders of the form $F_i \to X \leftarrow F_j$, irrespective of the topology of $\mathcal{G}_1$ and $\mathcal{G}_2$. It is left to show that unshielded colliders of the form $F_{\mathbf{I},\mathbf{J}} \to X \leftarrow Y$ are identical across $Aug_{\mathcal{I}}(\mathcal{G}_1)$ and $Aug_{\mathcal{I}}(\mathcal{G}_2)$. This must be true, otherwise $\mathcal{G}_1^{\tilde{\mathcal{I}}_{\mathbf{I}}}$ and $\mathcal{G}_2^{\tilde{\mathcal{I}}_{\mathbf{I}}}$ would have different unshielded colliders. Thus, $\langle \mathcal{G}_1, \mathcal{I} \rangle$ and $\langle \mathcal{G}_2, \mathcal{I} \rangle$ are Ψ-Markov equivalent. This concludes the proof. □

Finally, it is important to note that the algorithms proposed here, Ψ-FCI (Algorithm 1) and Ψ-PC (Algorithm 2), do not require knowledge of the interventional targets and their mechanisms, yet they are able to leverage their existence to learn a smaller equivalence class.

D.2 Mooij et al. [20]

The work in [20] proposes JCI as a framework to pool multiple datasets with unknown interventional targets and then employ traditional causal discovery algorithms to learn the causal graph. Under this framework, FCI-JCI [20, Sec. 4.2.4] is an adaptation of the FCI algorithm to learn causal graphs with latents over the pooled datasets, which combine observational and interventional datasets.

In this comparison, we consider the formulation denoted in [20, Sec. 4.2.4] by `FCI-JCI123` since it is the most comparable to the role of the F-nodes in our characterization and subsequent algorithms. They assume that C-nodes are source variables as background knowledge to help orient the graph. We assume data from the observational distribution is available since this seems to be the assumption in [20]. Hence, we have the datasets $\langle D_0, D_1, \ldots, D_M \rangle$, where D_0 denotes samples from the observational distribution P_0 and D_i denotes samples from the interventional distribution P_i. The algorithm pools all the datasets into one, denoted D^*, and appends D^* with context variables $\mathbf{C} = \{C_i\}_{i=1}^M$ such that $\mathbf{C} = \mathbf{0}$ for D_0, and $C_i = 1$ iff the sample is in D_i, otherwise $C_i = 0$. Then, `FCI-JCI` runs FCI on D^* (combining the observational and all the interventional datasets). According to the assumptions in `FCI-JCI123`, $C_i \leftrightarrow C_j, \forall i, j$ and $C_i \to V_j$ if C_i and V_j are adjacent after the skeleton phase of FCI.



(a) True Causal Graph $\mathcal{D}_1$ (b) Ψ-PAG $\mathcal{P}_{\mathcal{D}_1}$ (c) output of `FCI-JCI123`

(d) Causal Graph $\mathcal{D}_1$ (e) $Aug_{\mathcal{I}}(\mathcal{D}_1)$ (f) $\mathcal{D}_1^{\texttt{JCI}}$

(g) Causal Graph $\mathcal{D}_2$ (h) $Aug_{\mathcal{I}}(\mathcal{D}_1)$ (i) $\mathcal{D}_2^{\texttt{JCI}}$

Figure 5: Comparing Ψ-FCI and `FCI-JCI` with the interventional target set $\mathcal{I} = \langle \emptyset, \{X, Y\}, \{X, Y, Z\} \rangle$.

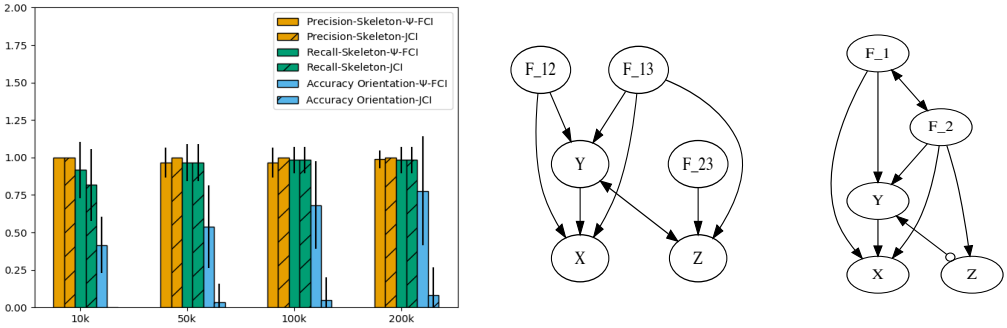
In the following example, we illustrate how Ψ-FCI powered by the characterization in Thm. 1 learns more about the true generating causal graph than `FCI-JCI`, thus identifying a more informative equivalence class. In fact, the previous example in Fig. 4 applies here as well, however, we choose another example to illustarte the use of latent variables.

Example 9. *Consider the causal graph in Fig. 5a, denoted $\mathcal{D}_1$ with U latent, and let $\langle P_0, P_1, P_2 \rangle$ be distributions generated by $\mathcal{D}_1$ with the corresponding set of interventional targets $\mathcal{I} = \{\emptyset, \mathbf{I} = \{X, Y\}, \mathbf{J} = \{X, Y, Z\}\}$. Following our notation, the absence of a mechanism identifier for each variable implies that X in $\mathbf{I}, \mathbf{J}$ is being intervened on with the same mechanism (same applies for Y). The output of Ψ-FCI is shown in Fig. 5b, where it is able to identify the edge marks of both edges (i.e., Y causes X and the existence of a latent variable between X and Z). On the other hand, `JCI-FCI123` constructs one context variable per interventional dataset D_i sampled from P_i, where $i \neq 0$. The output from `FCI-JCI123` is shown in Fig. 5c, where C_1 is adjacent to X, Y and C_2 is adjacent to X, Y, Z. Ψ-FCI is able to orient the edge between Y and Z into the latter while `FCI-JCI123` cannot.*

To understand why Ψ-FCI is more informative, we consider the augmented graphs in Figs. 5e and 5h, which correspond to the graphs in Figs. 5d and 5g. Node F_3 creates an unshielded collider $\langle F_3, Z, Y \rangle$ in Fig. 5e while the same is a non-collider in Fig. 5h. This can be tested by the invariance $P_{X,Y}(Y) = P_{X,Y,Z}(Y)$, which holds in distributions generated by $\mathcal{D}_1$. On the other hand, the causal graphs corresponding to the meta-system, according to [20], are shown in Figs. 5f and 5i. Node Z is not an unshielded collider across $\langle C_2, Z, Y \rangle$ in neither graph, hence the two graphs are indistinguishable.

In general, the advantage of the proposed approach is that distributional invariances across all the available distributions are tested, following the definition of the Ψ-Markov property (Def. 2). These invariances are identified graphically using the F-nodes and their separation statements in the augmented graph (Prop. 1). This is not the case for `FCI-JCI123` as illustrated in Example 9.

Now, we perform a small set of experiments to confirm the behavior of this example (i.e., the causal graph in Fig. 5a) with different sample sizes. We start by discussing the parametrization used in the simulations. For simplicity, all observable variables are binary. For each variable V_i, $P(V_i = 0|pa_i)$ is obtained through the logistic function. Specifically, we have $P(V_i = 0|pa_i) = \sigma(\sum_{z \in pa_i} c_z z + c_0)$, for all configurations of pa_i, where $\sigma(.)$ is the logistic function. The coefficients c_i are randomly chosen from the interval $[0.8, 0.95]$ except the intercept term c_0, which is set to 0.01. Let $\mathbf{c}$ denote the vector of $\{c_z\}_{z \in pa_i}$ along with the constant c_0. The last element (in order) of $\mathbf{c}$ is c_0. The interventional set is given by $\mathcal{I} = \langle \emptyset, \{X, Y\}, \{X, Y, Z\} \rangle$. For a soft intervention on X, the logistic model governing X is set to $\mathbf{c} = (0.8, 0.2)$, for Y, it is set to $\mathbf{c} = (0.3, 1)$, and for Z, it is changed according to $\mathbf{c} = (0.2, 1)$. This ensures that the nodes that are connected in the graph are statistically dependent, which is essential



(a) (b) (c)

Figure 6: Comparison of Ψ-FCI and `FCI-JCI123` using data from $\langle \mathcal{D}_1, \mathcal{I} \rangle$ in Fig. 5. (a) Bar plots for *precision* and *recall* for skeleton discovery and *accuracy* for edge orientations. (b) and (c) show the most informative graph that are output by Ψ-FCI and `FCI-JCI123`, respectively across all the runs and all sample sizes.

for using conditional independence and dependence statements to reverse-engineer the graph (i.e., to avoid faithfulness violations). We use `disCItest` tester for discrete Data from `pcalg` package and all results are thresholded at the p-value of 0.05 for producing the results.

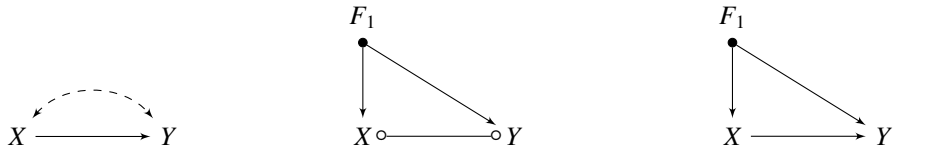
We compare our algorithm with `FCI-JCI123` and report the results for *precision* and *recall* for discovering the true skeleton (undirected version of the underlying graph) as well as the *accuracy* for correctly detecting edge orientations. The results are shown in Fig. 6(a). The errors bars are after averaging with 30 runs. We observe that in terms of discovering the graph skeleton, both algorithms perform similarly. Correctly detecting edge orientations is harder with small number of samples for both methods. The result is not surprising in the sense that they use virtually the same tests to discover the skeleton (conditional independences in the observational and experimental distributions). However, as the number of samples is increased, our method can learn the most informative equivalence class (Fig. 6b), whereas `FCI-JCI123` does not show performance improvement beyond a certain point (Fig. 6c). This is expected as well since, as we point out in Example 9, the invariances across interventional distributions are not fully utilized, which is informative in this case.

One might surmise that it may be possible to choose a different configuration for the context variables (i.e., a different initialization of values) and have all the tests covered. To answer this question formally, Proposition 6 proves that there exists no configuration for the context variables that allows `FCI-JCI123` to test for the invariances prescribed by the Ψ-Markov property, i.e., those of the form $P_i(y|w) = P_j(y|w)$, where $i, j \neq 0$.

Proposition 6. *If there are at least three distributions, then there does not exist a configuration for the context variables that enables `FCI-JCI123` to perform all the distributional invariance tests across every pair of distributions required by the Ψ-Markov property.*

Proof. Let the distributions be labeled as $\{1, \ldots, k\}$. First, note that [20, Lemma 19] shows that, for CI tests involving C-variables, it is sufficient to condition on all the remaining C-variables. Suppose that the design matrix (i.e., the pooled dataset with context variables) is such that for every pair of distributions $(1, i)$, there exists a j such that $(C_j \perp\!\!\!\perp X | W, C_{[k]-\{j\}}) \iff p_1(x|w) = p_i(x|w)$. We can assume without loss of generality $i = j$, i.e., the C-node used to check invariance across distributions $(1, i)$ is $C_i, \forall i$. We show that this set of constraints fixes the design matrix irrespective of the number of C variables/nodes introduced: It has to be either T or $1 - T$ where T is the following matrix. The first row of $\mathbf{T}$ is all-zeros. The i^{th} row of $\mathbf{T}$ is 1 only at the $i - 1^{th}$ column. Note that a necessary condition for an independence test between a node C_l and X to check the invariance $P_i(X) = P_j(X)$ is to have the design matrix satisfy the following constraint: Row i and row j differs only in column l. It is easy to see that neither $\mathbf{T}$ nor $1 - \mathbf{T}$ satisfies this condition. □

To summarize, we have shown in Example 9 that the equivalence class (EC) captured by `FCI-JCI123` is strictly less informative than the one discovered by Ψ-FCI. In fact, we utilized the smallest possible causal model (with 3 observable and 1 latent variable) such that this fact could be highlighted, and so



(a) Causal Graph $\mathcal{D}$ (b) Output of Ψ-FCI (c) Output of [17, Alg. 1]

Figure 7: Comparing Ψ-FCI and [17, Alg. 1] given causal graph $\mathcal{D}$ and $\mathcal{I} = \langle \emptyset, \{X\} \rangle$.

that it would be more clear from the characterization, and the corresponding tests, why the Ψ-Markov property leads to a more informative EC with respect to the same collection of distributions. We then showed a small set of experiments that support this findings, i.e., that asymptotically Ψ-FCI find the most informative EC. Finally, we showed that this gap is not specific to this example, but is a general phenomenon (Prop. 6), more formally, there is no design matrix compatible with JCI that can capture all the tests entailed by the Ψ-Markov-equivalence class.

D.3 Kocaoglu et al. [17]

The work in [17] considers the problem of learning causal graphs with latents from soft interventions when the interventional targets are known. They present a graphical characterization for when two causal graphs are in the same interventional equivalence class, denoted $\mathcal{I}$-Markov. Building on the characterization, they propose a sound algorithm to learn $\mathcal{I}$-Markov equivalence classes of causal graphs with latents represented by an augmented PAG, which is analogous to the Ψ-PAG.

In this work, we consider the problem of learning causal graphs with latents from soft interventions when the interventional targets are *unknown*. The Ψ-Markov property generalizes the $\mathcal{I}$-Markov property by detaching the distributions from the interventional targets and considering the set of interventional targets $\mathcal{I}$ to be a variable along with the causal graph $\mathcal{D}$. Accordingly, the graphical characterization of Ψ-Markov equivalence generalizes the one in [17, Thm. 2] by considering different sets of interventional targets with the corresponding causal graphs. It is easy to see that both theorems are equivalent given the same interventional target set as shown below.

Proposition 7. *Consider two causal graphs with latents $\mathcal{D}_1, \mathcal{D}_2$, and the set of interventional targets $\mathcal{I}$. Assuming controlled experimental setting, $\langle \mathcal{D}_1, \mathcal{I} \rangle$ and $\langle \mathcal{D}_2, \mathcal{I} \rangle$ are Ψ-Markov equivalent if and only if $\mathcal{D}_1$ and $\mathcal{D}_2$ are $\mathcal{I}$-Markov equivalent according to [17, Thm. 2].*

Proof. Under the controlled experimental setting, we can drop the mechanism identifier for each intervened variable and the symmetrical difference (Def. 1) reduces to that in [17]. In this case, the construction of the augmented graph in Def. 4 differs from that in [17, Def. 3] by having duplicate symmetrical difference sets in $\mathcal{K}$, and thus additional redundant F-nodes. It follows that the corresponding augmented MAGs according to [17] are induced subgraphs of the respective $\mathcal{I}$-MAGs with the difference being redundant F-nodes. Hence, the claim of the proposition follows easily. □

As for the learning algorithm, we point out the following differences between Ψ-FCI and that of [17, Alg. 1]. First, running [17, Alg. 1] is not possible under unknown interventional targets since the construction of the F-nodes takes the set $\mathcal{I}$ as input, which is not available in the problem considered here. Moreover, the output could have incorrect orientations even if we fix the F-node construction due to *Rule 9* in [17, Alg. 1], which is only sound under known interventional targets (as shown in Example 10 below). Finally, Ψ-FCI is complete for learning with unknown targets as shown in Thm. 3 while [17, Alg. 1] is sound with no claim of completeness.

Example 10. *Consider the causal graph in Fig. 7a with a pair of distributions $\langle P_1, P_2 \rangle$ and the corresponding set of interventional targets $\mathcal{I} = \langle \emptyset, \{X\} \rangle$. When the interventional targets are known, [17, Alg. 1] learns the augmented PAG shown in Fig. 7c. The edge between X and Y is oriented due to Rule 9 in [17, Alg. 1]. On the other hand, the output of Ψ-FCI is shown in Fig. 7b when the interventional targets are unknown. Notice the edge between X and Y is not oriented here. That is because $\langle \mathcal{D}, \mathcal{I} \rangle$ is Ψ-Markov equivalent to $\langle \mathcal{D}', \mathcal{I}' \rangle$ where $\mathcal{D}' = \{X \leftarrow Y\}$ and $\mathcal{I}' = \langle \emptyset, \{X, Y\} \rangle$.*

In summary, we consider the problem of learning from interventional data with *unknown interventional targets*, a setting which cannot be handled by the work in [17]. When considering the results in [17], we make the following contributions:

1. We formulate the Ψ-Markov property and derive a graphical characterization that subsumes that of [14], which is formally shown in Proposition 7, Appendix D.3.
2. We show that the algorithm introduced in [14] is not applicable under unknown interventional targets and present a complete algorithm (Ψ-FCI) under unknown interventional targets (Example 10).
3. We handle causal sufficiency as a special case of the derived results. Subsection 3.1 establishes a graphical characterization for this case, and Section C, in the Appendix, presents an algorithm for learning an equivalence class from interventional data under causal sufficiency. We prove this algorithm to be complete for both known and unknown interventional targets. The work in [14] does not discuss the causally sufficient case nor completeness.

D.4 Peters et al. [23]

The work in [23] proposes using invariances among conditional distributions of a specific target variable, given subsets of the predictor variables. Specifically, for a target variable Y, they identify subsets S of the predictor variables such that $P_i(Y|S)$ are identical across all distributions $P_i(.)$, for all environments i. It is easy to show that, as long as the target variable is not intervened on, the causal parents of Y is a valid set S, i.e., $P_i(Y|Pa_Y)$ does not change across different environments.

The advantage of this framework is that it does not require the knowledge of interventional targets. The authors provides sufficient conditions for their framework to uniquely identify the true causal parents of Y. They assume linear SCMs without latent variables and also impose certain conditions on the set of interventions. A follow-up work in [39] relaxes the linearity assumption. In the following, we only discuss the idea of using invariances across distributions, ignoring the parametric assumptions of the framework.

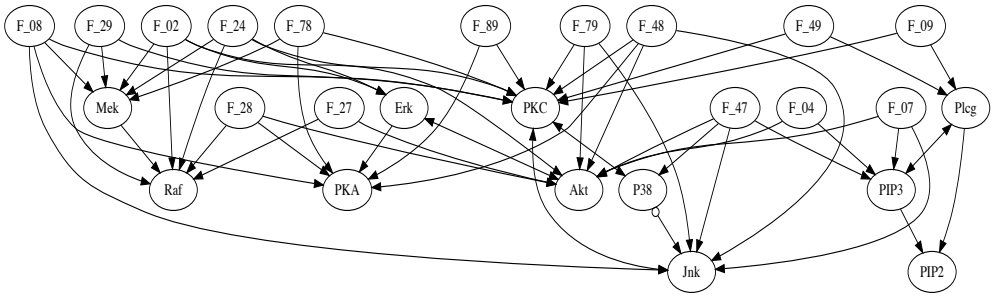
One can imagine applying this framework recursively to identify the parents of each node to eventually learn the underlying causal graph. However, there are some difficulties with this approach: Suppose in iteration i, we seek for the set of parents of X_i. Given an arbitrary set of interventional targets, it's not possible, in general, to identify a subset of interventions where X_i is not intervened on.

To illustrate this point, consider the simple causal chain $Z \to Y \to X$ with the distributions P, P_Y, i.e., the observational distribution and the interventional distribution on Y. Suppose these interventional targets are not known. Let us assume we start with X. Then we identify $P(X|Y) = P_Y(X|Y)$. Then Y can be declared as the parent of X. Suppose we consider Y as the next node. It can be seen that there is no invariance $P(Y|S)$ for any S. Therefore, one might either wrongfully infer $Y \to Z$ or halt the algorithm. This is, of course, due to the fact that the new target variable Y is intervened on.

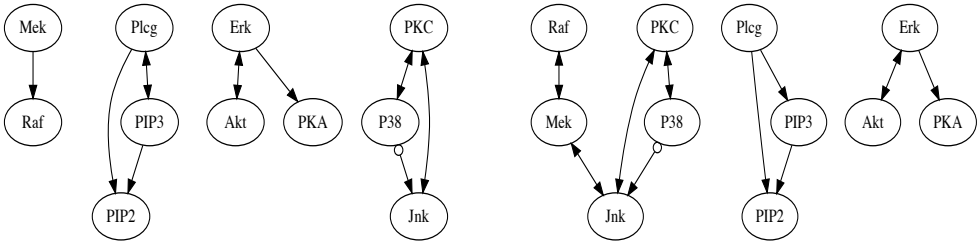
In general, one could expect that looking for invariances that hold across all environments should be less informative than looking for invariances across pairs of environments. For example, in the same graph above, consider the interventional distributions $P_{X,Y}, P_{X,Z}, P_{Y,Z}$, where each variable is intervened with the same mechanism change, if it is an interventional target. It is easy to see that there is no conditional distribution that is invariant across all environments. However, we have $P_{X,Y}(X|Y) = P_{X,Z}(X|Y)$, which can be used for discovering the edge between X and Y using our framework.

D.5 Zhang et al. [38]

Another related work that uses changes in the mechanism for learning the underlying causal graph is [38]. Without distinguishing whether this change happens in time or across different contexts, the authors propose using auxiliary random variables to capture the mechanism changes. In that sense, JCI [20] can be seen as an extension of this idea. Therefore, similar to JCI, our approach differs in how we treat these auxiliary nodes as parameters, rather than random variables. Later, authors also connect this approach to independence of cause and mechanism principle used previously for causal discovery from observational data [14], in order to discover some of the edges that are otherwise not identifiable, such as the ones adjacent to a context variable. In that sense, it would be interesting to invoke this and other observational discovery methods [11, 14, 16] to identify edges that are otherwise not identifiable using our work.



(a) Ψ-FCI output with $\alpha = 0.05$



(b) $\alpha = 0.05$ (c) $\alpha = 0.1 \& \alpha = 0.15$

Figure 8: Comparison of the Ψ-FCI output after running it on Sachs' dataset [28] under different significance levels. Fig. 8a presents the full output of Ψ-FCI while Figs. 8b& 8c remove the F-nodes for clarity.

D.6 Rothenhäusler et al. [27]

We do not conduct a thorough comparison between the work in [27] and Ψ-FCI given the very nature of both approaches, they are not really comparable. On the one hand, [27] considers the broader class of cyclic causal models while ours is restricted to acyclic models. On the other hand, our work makes no assumption about the functional form or type of soft intervention, while [27] considers linear causal relations and shift interventions.

D.7 Sachs et al. [28]

We also run our Algorithm on the interventional datasets obtained from single cell fluorescence based measurements of proteins involved in T-4 cell signalling. This is a standard benchmark used in other causal inference studies. We used the pre-processed and quantized dataset from the link in [41]. Interventions are various drugs injected to inhibit or activate various signalling proteins involved in the pathways. We provide the output of our algorithm for various significance levels ($0.05, 0.1$ and 0.15) used for the CI testers in Figure 8. We display the full output along with F nodes one each for a pair of interventions in Figure 8a. To enhance readability, we provide the induced subgraph on the protein nodes alone in Figures 8b and 8c. While the true ground truth network is not exactly known, a plausible one has been used in [22]. In that network, Plcg, PIP3, PIP2 form a sub-network which is recovered by our outputs (specifically in Figure 8c).

Appendix References

[39] Christina Heinze-Deml, Jonas Peters, and Nicolai Meinshausen. Invariant causal prediction for nonlinear models. *Journal of Causal Inference*, 6(2), 2018.

[40] Robert Osazuwa Ness, Karen Sachs, Parag Mallick, and Olga Vitek. A bayesian active learning experimental design for inferring signaling networks. In *International Conference on Research in Computational Molecular Biology*, pages 134–156. Springer, 2017.

[41] `https://www.bnlearn.com/book-useR/`.