Characterization and Learning of Causal Graphs with Latent Variables from Soft Interventions

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

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

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

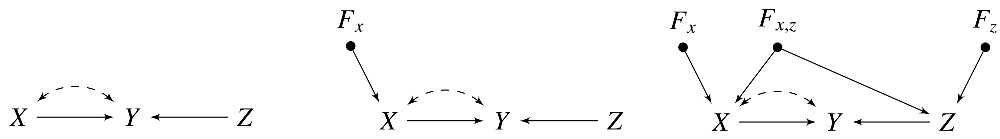
The challenge of learning the causal structure underlying a certain phenomenon is undertaken by connecting the set of conditional independences (CIs) readable from the observational data, on the one side, with the set of corresponding constraints implied over the graphical structure, on the other, which are tied through a graphical criterion known as d-separation (Pearl, 1988). In this paper, we investigate the more general setting where multiple observational and experimental distributions are available. We start with the simple observation that the invariances given by CIs/d-separation are just one special type of a broader set of constraints, which follow from the careful comparison of the different distributions available. Remarkably, these new constraints are intrinsically connected with do-calculus (Pearl, 1995) in the context of soft-interventions. We then introduce a novel notion of interventional equivalence class of causal graphs with latent variables based on these invariances, which associates each graphical structure with a set of interventional distributions that respect the do-calculus rules. Given a collection of distributions, two causal graphs are called interventionally equivalent if they are associated with the same family of interventional distributions, where the elements of the family are indistinguishable using the invariances obtained from a direct application of the calculus rules. We introduce a graphical representation that can be used to determine if two causal graphs are interventionally equivalent. We provide a formal graphical characterization of this equivalence. Finally, we extend the FCI algorithm, which was originally designed to operate based on CIs, to combine observational and interventional datasets, including new orientation rules particular to this setting.

1 Introduction

Explaining a complex system through their cause and effect relations is one of the fundamental challenges in science. Data is collected and experiments are performed with the intent of understanding how a certain phenomenon comes about, or how the underlying system works, which could be social, biological, artificial, among others. The study of causal relations can be seen through the lens of learning and inference [16, 21]. The learning component is concerned with discovering the causal structure, which is the very subject of interest in many domains, since they can provide insight about

*Equal contribution.

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.



(a) Causal Diagram $\mathcal{D}$. (b) $Aug(\mathcal{D})$ for P and P_x. (c) $Aug(\mathcal{D})$ for P, P_x, and $P_{x,z}$.

Figure 1: (a) Causal graph where the bidirected edge represents a latent confounder. (b) Given P_x, P, we can use F_x to capture information such as *"there is a backdoor path from X to Y"* in terms of m-separation $F_x \not\perp\!\!\!\perp Y|X$. (c) Given $P, P_x, P_{x,z}$, under controlled experiment assumption, we can add F_z although P_z is not available. This allows us to discover that Z is a cause of Y and there is no confounder between them. Without adding F_z this relation cannot be identified.

how a complex system works and lead to better understanding about the phenomenon under investigation. The latter, inference, attempts to leverage the causal structure to compute quantitative claims about the effect of interventions and retrospective counterfactuals, which are critical to assign credit, understand blame and responsibility, and perform judgement about fairness in decision-making.

One of the most popular languages used to encode the invariances needed to reason about causal relations, for both learning and inference, is based on graphical models, and appears under the rubric of *causal graphs* [16, 21, 2]. A causal graph is a directed acyclic graph (DAG) with latent variables, where each edge encodes a causal relationship between its endpoints: X is a direct cause of Y, i.e., $X \to Y$, if, when the remaining factors are held constant, forcing X to take a specific value affects the realization of Y, where X, Y are random variables representing some relevant features of the system.

The task of learning the causal structure entails a search over the space of causal graphs that are compatible with the observed data; the collection of these graphs forms what is called an *equivalence class*. The most popular mark imprinted on the data by the underlying causal structure that is used to delineate an equivalence class are *conditional independence* (CI) relations. These relations are the most basic type of probabilistic invariances used in the field and have been studied at large in the context of graphical models since, at least, [15] (see also [5]). While CIs are powerful and have been the driving force behind some of the most prominent structural learning algorithms in the field [16, 21], including the PC, FCI, these are constraints specific for one distribution.

In this paper, we start by noting something very simple, albeit powerful, that happens when a combination of observational and experimental distributions are available: There are constraints over the graphical structure that emerge by comparing these different distributions, and which are not of CI-type[2]. Remarkably, and unknown until our work, the converse of the causal calculus developed by Pearl [18] offers a systematic way of reading these constraints and tying them back to the underlying graphical structure. In reference to their connection to the do-calculus rules (or a generalization, as discussed later), we call these constraints the *do-constraints*. For concreteness, consider the graph in Fig. 1(a), where the dashed-bidirected arrow represents hidden variables that generate variations of the two observed variables, X, Y in this case. Suppose the observational (conditional) distribution and an interventional distribution on X are available, which are written as $P(y|x)$, $P(y|do(x))$, respectively. Suppose we contrast these two distributions and the test evaluating the expression $P(y|do(x)) = P(y|x)$ comes out as *false*. This is called a do-see test since the experimental (or "do") and observational ("see") distributions are contrasted. Based on the second rule of do-calculus, one can infer that there is an open *backdoor path* from X to Y, where the edge adjacent to X on this path has an arrowhead into X. In our setting, we do not have access to the true graph, but we leverage this and the other do-constraints to reverse engineer the process and try to learn the structure. Broadly speaking, do-constraints will play a critical role for learning, in the same way CI/d-separation plays in learning when only observational data is available. To the best of our knowledge, this type of constraints appeared first at the very definition of causal Bayesian networks (CBNs) in [1] and then were leveraged to design efficient experiments to learn the causal graph in [12].

We assume throughout this work that interventions are *soft*. A soft intervention affects the mechanism that generates the variable, while keeping the causal connections intact. Soft-interventions are widely employed in biology and medicine, where it is hard to change the underlying system, but possibly

[2]Recall that a CI represents a constraint readable from one specific distribution saying that the value of Z is irrelevant for computing the likelihood of Y once we know the value of X, i.e., $P(Y|X,Z) = P(Y|X), \forall X, Y, Z$.

easier to perturb it. For our characterization, we utilize an extension of the causal calculus to soft interventions introduced in [4]. Under soft-interventions, the do-see test can be written as checking if $P_x(y|x) = P(y|x)$, where P_x is the distribution obtained after a soft intervention on X.

The second observation leveraged here follows from another realization by Pearl that interventions can be represented explicitly in the graphical model [17]. He then introduced what we call *F-nodes*, which graphically encode the changes due to an intervention and the corresponding parametrization (see also [16, Sec. 3.2.2]). This is important in our context since the do-calculus tests will be visible more explicitly in the graph. The graph obtained by adding F-nodes to the causal graph is called the *augmented graph*. The same construct was used more prominently in [6] in the context of inference and identification. Going back to Fig. 1b, the existence of the backdoor path from X to Y, as detected by rule 2 of the calculus, can be captured by the statement F_X *is not d-separated from* Y *given* X. In the context of structure learning, similar constructions have been leveraged in the literature [13, 24].

We further make a specific assumption throughout the paper about the soft-interventions. We call it the *controlled experiment setting*, where each variable is intervened with the *same mechanism* change across different interventions. For example, in Fig. 1c, suppose we are given distributions from two controlled experiments $P_x, P_{x,z}$ along with observational data. We can then use F_z to capture the invariances between $P_{x,z}$ and P_x. For example, if $P_{x,z}(y) \neq P_x(y)$, for some y, we can read that $F_Z \not\perp\!\!\!\perp Y \,\big|\, F_X, F_{X,Z}$. Accordingly, given a set of interventional distributions, we construct an augmented graph by introducing an F-node for every unique set difference between pairs of controlled intervention sets (more on that later on). Without the controlled experiment assumption, our machinery can still be used if one knows which mechanism changes are identical and by constructing F-nodes to reflect and capture the mechanism difference across two interventions. For simplicity of presentation, however, we restrict ourselves to the controlled experiment setting and do not pursue this route explicitly.

To encapsulate the distributional invariants directly induced by the causal calculus rules[3], we call a set of interventional distributions $\mathcal{I}$-Markov to a graph, if these distributions respect the causal calculus rules relative to that graph. Note that the notion of $\mathcal{I}$-Markov is first introduced in [9, 10] for causally sufficient systems without the use of do-constraints[4]. For our characterization, we first extend the causal calculus rules to operate between arbitrary sets of interventions. We call two causal graphs $\mathcal{D}_1, \mathcal{D}_2$ $\mathcal{I}$-Markov equivalent if the set of distributions that are $\mathcal{I}$-Markov to $\mathcal{D}_1$ and $\mathcal{D}_2$ are the same. Using the augmented graph, we identify a graphical condition that is necessary and sufficient for two CBNs with latents to be $\mathcal{I}$-Markov equivalent. Finally, we propose a sound algorithm for learning the augmented graph from interventional data. Our contributions can be summarized as follows:

- We propose a characterization of $\mathcal{I}$-Markov equivalence between two causal graphs with latent variables for a given intervention set $\mathcal{I}$ that is based on a generalization of do-calculus rules to arbitrary subsets of interventions.
- We show a graphical characterization of $\mathcal{I}$-Markov equivalence of causal graphs with latents.
- We introduce a learning algorithm for inferring the graphical structure using a combination of observational and interventional data and utilizing the corresponding new constraints. This procedure comes with a new set of orientation rules. We formally show its soundness.

2 Background and Related Work

In this section, we introduce necessary concepts that we use throughout the paper. Upper case letters denote variables and lower case letters denote an assignment. Also, bold letters denote sets.

Causal Bayesian Network (CBN): Let $P(\mathbf{v})$ be a probability distribution over a set of variables $\mathbf{V}$, and let $P_{\mathbf{x}}(\mathbf{v})$ 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 [16, 1, pp. 24]. If so, we refer to the DAG as *causal*.

[3]There may be constraints that can be obtained by applying the rules multiple times we do not consider here.

[4]In the causally sufficient case, name is in reference to both global and local Markov conditions. However, in our work, the name stems from our observation that the do-constraints correspond to the global Markov conditions in the augmented graph.

Given that a subset of the variables are unmeasured or latent, $\mathcal{D}(\mathbf{V} \cup \mathbf{L}, \mathbf{E})$ represents the causal graph where $\mathbf{V}$ and $\mathbf{L}$ denote the measured and latent variables, respectively, and $\mathbf{E}$ denotes the edges. A dashed bi-directed edge is used instead of $\leftarrow L \rightarrow$, where $L \in \mathbf{L}$, whenever L is a root node with exactly two children. The observed distribution $P(\mathbf{v})$ is obtained by marginalizing $\mathbf{L}$ out.

$$P(\mathbf{v}) = \sum_{\mathbf{L}} \prod_{\{i | T_i \in \mathbf{V} \cup \mathbf{L}\}} P(t_i | \mathbf{pa}_i)$$

Clearly, the joint distribution over $\mathbf{V}$ does not factorize relative to $\mathcal{D}$ in a typical fashion, since Markovianity is no longer valid, but it does relative to both $\mathbf{V}$ and $\mathbf{L}$. Still, CI relations can be read from the graph using a graphical criterion known as *d-separation*. Also, two causal graphs are called *Markov equivalent* whenever they share the same set of conditional independences over $\mathbf{V}$.

Soft Interventions: Another common type of intervention is *soft*, where 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 $P_{\mathbf{x}}(\mathbf{v})$ becomes as follows, where $P'(X_i | Pa_i) \neq P(X_i | Pa_i)$ is the new conditional distribution set by the intervention:

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

In this work, we assume that all the soft interventions are *controlled*. This means that for any two interventions $\mathbf{I}, \mathbf{J} \subseteq \mathbf{V}$ where $X_i \in \mathbf{I} \cap \mathbf{J}$, we have $P_{\mathbf{I}}(X_i | Pa_i) = P_{\mathbf{J}}(X_i | Pa_i)$.

Ancestral graphs: We now introduce a graphical representation of equivalence classes of causal graphs with latent nodes. 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*. 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 a directed or almost directed cycle. 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 [19]. Given a causal graph $\mathcal{D}(\mathbf{V}, \mathbf{L})$, a MAG $\mathcal{M}_{\mathcal{D}}$ over $\mathbf{V}$ can be constructed such that both the independence and the ancestral relations among variables in $\mathbf{V}$ are retained, see, for example, [27, 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 adjancent. If both edges are into Y, then the triple is referred to as *unshielded collider*. A path between X and Y, $p = \langle X, \ldots, W, Z, Y \rangle$, is a *discriminating path* for Z if (1) p includes at least three edges; (2) Z is a non-endpoint node on p, and is adjacent to Y on p; and (3) X is not adjacent to Y, and 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) they have the same unshielded colliders; and (3) if a path p is a discriminating path for a vertex Z in both graphs, then Z is a collider on the path in one graph if and only if it is a collider on the path in the other. A *PAG*, which represents a Markov equivalence class of a MAG, is learnable from the independence model over the observed variables, and the FCI algorithm is a standard sound and complete method to learn such an object [28].

Related Work: Learning causal graphs from a combination of observational and interventional data has been studied in the literature [3, 11, 7, 20, 8, 12, 23]. For causally sufficient systems, the notion and characterization of interventional Markov equivalence has been introduced in [9, 10]. More recently, [24] showed that the same characterization can be used for both hard and soft interventions. For causally insufficient systems, [22] uses SAT solvers to learn a summary graph over the observed variables given data from different experimental conditions. [13] introduces an algorithm to pool experimental datasets together and runs a modification of FCI to learn an augmented graph; however, they do not consider characterizing an equivalence class.

Notations: For random variables X, Y, Z, the CI relation X *is independent of* Y *conditioned on* Z is shown by $X \perp\!\!\!\perp Y | Z$. The d-separation statement *node* X *is d-separated from* Y *given* Z *in graph* $\mathcal{D}$ is shown by $(X \perp\!\!\!\perp Y | Z)_{\mathcal{D}}$. $\mathcal{I} \subseteq 2^{\mathbf{V}}$ is reserved for a set of interventions, where $2^{\mathbf{V}}$ is the power set of $\mathbf{V}$. We show the symmetric difference by $\mathbf{I} \triangle \mathbf{J} := (\mathbf{I} \setminus \mathbf{J}) \cup (\mathbf{J} \setminus \mathbf{I})$. $\mathcal{D}_{\overline{\mathbf{X}}}$ denotes the graph obtained from $\mathcal{D}$ where all the incoming edges to the set of nodes in $\mathbf{X}$ are removed. Similarly, $\mathcal{D}_{\underline{\mathbf{X}}}$ denotes the removal of outgoing edges. We assume that there is no selection bias. A star on an endpoint of an edge $*\!\!-\!\!*$ is used as a wildcard to denote circle, arrowhead, or tail.

3 Do-Constraints – Combining Observational and Experimental Distributions

One of the most celebrated results in causal inference comes under the rubric of do-calculus (or causal calculus) [18, 16]. The calculus consists of a set of inference rules that allows one to create a map between distributions generated by a causal graph when certain graphical conditions hold in the graph. The calculus was developed in the context of hard interventions, and recent work presented a generalization of this result for soft interventions [4], which we state next:

Theorem 1 (Special case of Thm. 1 in [4])**.** *Let $\mathcal{D} = (\mathbf{V} \cup \mathbf{L}, \mathbf{E})$ be a causal graph. Then, the following holds for any strictly positive distribution consistent with $\mathcal{D}$.*

Rule 1 (see-see): For any $\mathbf{X} \subseteq \mathbf{V}$ and disjoint $\mathbf{Y}, \mathbf{Z}, \mathbf{W} \subseteq \mathbf{V}$

$$P_x(y|w, z) = P_x(y|w) \qquad \text{if } Y \perp\!\!\!\perp Z\,|W \text{ in } \mathcal{D}.$$

Rule 2 (do-see): For any disjoint $\mathbf{X}, \mathbf{Y}, \mathbf{Z} \subseteq \mathbf{V}$ and $\mathbf{W} \subset \mathbf{V} \setminus (\mathbf{Z} \cup \mathbf{Y})$

$$P_{x,z}(y|z, w) = P_x(y|z, w) \qquad \text{if } Y \perp\!\!\!\perp Z\,|W \text{ in } \mathcal{D}_{\underline{Z}}.$$

Rule 3 (do-do): For any disjoint $\mathbf{X}, \mathbf{Y}, \mathbf{Z} \subseteq \mathbf{V}$ and $\mathbf{W} \subset \mathbf{V} \setminus (\mathbf{Z} \cup \mathbf{Y})$

$$P_{x,z}(y|w) = P_x(y|w) \qquad \text{if } Y \perp\!\!\!\perp Z\,|W \text{ in } \mathcal{D}_{\overline{Z(W)}},$$

where $\mathbf{Z}(\mathbf{W}) \subseteq \mathbf{Z}$ are non-ancestors of $\mathbf{W}$ in $\mathcal{D}$.

The first rule of the calculus is a d-separation type of statement relative to a specific interventional distribution P_x, which says that $Y \perp\!\!\!\perp Z\,|W$ in $\mathcal{D}$ implies the corresponding conditional independence $P_x(y|w, z) = P_x(y|w)$. Note that the converse of this rule is the work horse underlying most of the structure learning algorithms found in practice, which says that if some independence hold in P, this would imply a corresponding graphical separation (under faithfulness). In the case just mentioned, this would imply that Y and Z should be separated in $\mathcal{D}$, meaning, they have neither a directed nor a bidirected arrow connecting them.

From this understanding, we make a very simple, albeit powerful observation – i.e., the converse of the other two rules should offer insights about the underlying graphical structure as well. To witness, consider the causal graph $\mathcal{D} = \{X \to Y, X \dashleftarrow\!\dashrightarrow Y\}$, and suppose we have the observational and interventional distributions $P(Y, X)$ and $P_X(Y, X)$, respectively. Using the CI tests $P(Y, X) \neq P(Y) \cdot P(X)$ and $P_X(Y, X) \neq P_X(Y) \cdot P_X(X)$, we infer that the two variables are dependent (or not independent) and consequently d-connected in the graph, while no claim can be made about the causal relation between them. Given the inequality $P_X(Y) \neq P(Y)$, we infer that the condition for rule 3 does not hold and $Y \not\perp\!\!\!\perp X$ in $\mathcal{D}_{\overline{X}}$. Hence, X must be a cause of Y – changing the value of X has a downstream effect on Y. Similarly, given the inequality $P_X(Y|X) \neq P(Y|X)$, the condition related to rule 2 does not hold, and $Y \not\perp\!\!\!\perp X$ in $\mathcal{D}_{\underline{X}}$. The implication in this case is that there is an unblockable backdoor path between X and Y that is into X, i.e., a latent variable. Alternatively, if $\mathcal{D} = \{X \to Y\}$, then $P_X(Y|X) = P(Y|X)$, under faithfulness, implies the absence of a latent variable by the converse of rule 2.

Broadly speaking, rule 3 allows one to infer causal relations between variables, and consequently directed edges in the causal graph. Since the compared interventional distributions differ by a subset of interventions (Z), we call this the *do-do* test. On the other hand, rule 2 allows one to infer spurious relations between variables, and consequently latent variables in the causal graph[5]. The *do-see* naming of the test stems from the fact that we compare a distribution with an intervention on a subset Z (do) versus another which only conditions on Z (see). Naturally, rule 1 is the usual conditional independence test that allows one to detect that neither directed nor bidirected arrow exists.

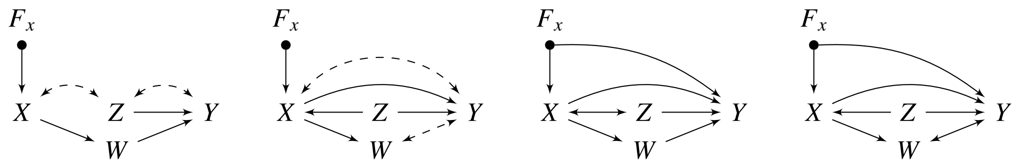
Putting together these rules, we show in Corollary 1 a generalization of rules 2 and 3 . Note that rule 2 appears when $\mathbf{J} \subset \mathbf{I}$ and $\mathbf{I} \setminus \mathbf{J} \subseteq \mathbf{W}$; similarly, rule 3 can be seen when $\mathbf{J} \subset \mathbf{I}$ and $(\mathbf{I} \setminus \mathbf{J}) \cap \mathbf{W} = \emptyset$.

Corollary 1 (mixed do-do/do-see)**.** *Let $\mathcal{D} = (\mathbf{V} \cup \mathbf{L}, \mathbf{E})$ be a causal graph. Under the controlled intervention assumption, for any $\mathbf{I}, \mathbf{J} \subseteq \mathbf{V}$ and disjoint $\mathbf{Y}, \mathbf{W} \subseteq \mathbf{V}$, we have the following:*

$$P_{\mathbf{I}}(\mathbf{y}|\mathbf{w}) = P_{\mathbf{J}}(\mathbf{y}|\mathbf{w}) \qquad \text{if } \mathbf{Y} \perp\!\!\!\perp \mathbf{K}\,|\mathbf{W} \setminus \mathbf{W_k} \text{ in } \mathcal{D}_{\underline{\mathbf{W_k}}, \overline{\mathbf{R}(\mathbf{W})}},$$

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

[5]More precisely, rule 2 allows us to detect inducing paths that are into both variables.



(a) $Aug_{\mathcal{I}}(\mathcal{D})$ (b) $Aug_{\mathcal{I}}(\mathcal{D}')$ (c) $MAG(Aug_{\mathcal{I}}(\mathcal{D}))$. (d) $MAG(Aug_{\mathcal{I}}(\mathcal{D}'))$.

Figure 2: Augmented graphs with respect to $\mathcal{I} = \{\emptyset, \{X\}\}$ and the corresponding augmented MAGs.

In general, the proposed rule is a mixture of rules 2 and 3 as we could be conditioning in $\mathbf{W}$ on a subset of the symmetrical difference set $\mathbf{I} \triangle \mathbf{J}$. For instance, consider the causal graph $\mathcal{D} = \{C \leftarrow\!\!\dashrightarrow A \rightarrow B, C \leftarrow\!\!\dashrightarrow B\}$ and suppose we have the interventional distributions $P_{A,B}$ and $P_{C,B}$. Since $B \perp\!\!\!\perp \{A, C\}$ in $\mathcal{D}_{\underline{A},\overline{C}}$, then $P_{A,B}(B|A) = P_{B,C}(B|A)$. This generalization will soon play a significant role in the characterization and learning of the interventional equivalence class.

4 Interventional Markov Equivalence under Do-constraints

In this section, the new do-constraints will be used to define the notion of interventional Markov equivalence. Then, we will characterize when two causal graphs are equivalent in accordance to the proposed definition. We start by defining the notion of interventional Markov as shown below.

Definition 1. *Consider the tuples of absolutely continuous probability distributions $(P_{\mathbf{I}})_{\mathbf{I}\in\mathcal{I}}$ over a set of variables $\mathbf{V}$. A tuple $(P_{\mathbf{I}})_{\mathbf{I}\in\mathcal{I}}$ satisfies the $\mathcal{I}$-Markov property with respect to a graph $\mathcal{D} = (\mathbf{V} \cup \mathbf{L}, \mathbf{E})$ if the following holds for disjoint $\mathbf{Y}, \mathbf{Z}, \mathbf{W} \subseteq \mathbf{V}$:*

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

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

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

The set of all tuples that satisfy the $\mathcal{I}$-Markov property with respect to $\mathcal{D}$ are denoted by $\mathcal{P}_{\mathcal{I}}(\mathcal{D}, \mathbf{V})$.

The two conditions used in the definition correspond to rule 1 of Theorem 1 and that of Corollary 1. Notice that the traditional Markov definition only considers the first condition over the observational distribution $P(\mathbf{V})$; a case included in the $\mathcal{I}$-Markov whenever $\emptyset \in \mathcal{I}$. Accordingly, two causal graphs are said to be $\mathcal{I}$-Markov equivalent if they license the same set of distribution tuples. This notion is formalized in the following definition.

Definition 2. *Given two 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 an intervention set $\mathcal{I} \subseteq 2^{\mathbf{V}}$, $\mathcal{D}_1$ and $\mathcal{D}_2$ are called $\mathcal{I}$-Markov equivalent if $\mathcal{P}_{\mathcal{I}}(\mathcal{D}_1, \mathbf{V}) = \mathcal{P}_{\mathcal{I}}(\mathcal{D}_2, \mathbf{V})$.*

One challenge with Definition 1 is that testing for the d-separation statement in condition (2) requires a mutilated graph where we cut some of the edges in $\mathcal{D}$. This makes it harder to represent all the constraints imposed by a causal graph compactly. Accordingly, we use the notion of an *augmented graph* that is introduced below (Definition 3). In words, the construction of the augmented graph goes as follows. First, initialize the augmented graph to the input causal graph. Then, for every distinct symmetric set difference between $\mathbf{I}, \mathbf{J} \in \mathcal{I}$, denoted by $\mathbf{S}_i$, introduce a new node F_i and make it a parent to each node in $\mathbf{S}_i$, i.e., $F_i \rightarrow S \in \mathbf{S}_i$. Note that this type of construction has been used in the literature to model interventions [17, 6]. For example, for $\mathcal{I} = \{\emptyset, \{X\}\}$, Figure 2a presents the augmented graph corresponding to the causal graph, which is the induced subgraph over $\{X, W, Z, Y\}$. Node F_x is added in accordance with the symmetrical difference set $(\emptyset \setminus \{X\}) \cup (\{X\} \setminus \emptyset) = \{X\}$.

Definition 3 (Augmented graph)**.** *Consider a causal graph $\mathcal{D} = (\mathbf{V} \cup \mathbf{L}, \mathbf{E})$ and an intervention set $\mathcal{I} \subseteq 2^{\mathbf{V}}$. Let $\mathcal{S} = \{\mathbf{S}_1, \mathbf{S}_2, \ldots, \mathbf{S}_k\} = \{S : \exists \mathbf{I}, \mathbf{J} \in \mathcal{I} \text{ s.t. } \mathbf{I} \triangle \mathbf{J} = S\}$. 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 $\mathcal{E} = \{(F_i, j)\}_{i\in[k], j\in\mathbf{S}_i}$.*

The significance of the augmented graph construction is illustrated by Proposition 1, which provides criteria to test the d-separation statements in Definition 1 equivalently from the corresponding augmented graph of a causal graph. Back to the example in Figure 2a, the statement $Y \perp\!\!\!\perp X | Z$ in

$\mathcal{D}_{\overline{X}}$ can be equivalently tested by the statement $Y \perp\!\!\!\perp F_x | Z$ in the corresponding augmented graph. Similarly, $Y \perp\!\!\!\perp X$ in $\mathcal{D}_{\underline{X}}$ can be equivalently tested by $Y \perp\!\!\!\perp F_x | X$ in $Aug_{\mathcal{I}}(\mathcal{D})$.

Proposition 1. *Consider a causal graph $\mathcal{D} = (\mathbf{V} \cup \mathbf{L}, \mathbf{E})$ and the corresponding augmented graph $Aug_{\mathcal{I}}(\mathcal{D}) = (\mathbf{V} \cup \mathbf{L} \cup \mathcal{F}, \mathbf{E} \cup \mathcal{E})$ with respect to an intervention set $\mathcal{I}$, where $\mathcal{F} = \{F_i\}_{i \in [k]}$. Let $\mathbf{S}_i$ be the set of nodes adjacent to $F_i, \forall i \in [k]$. We have the following equivalence relations.*

For disjoint $\mathbf{Y}, \mathbf{Z}, \mathbf{W} \subseteq \mathbf{V}$:

$$(\mathbf{Y} \perp\!\!\!\perp \mathbf{Z} | \mathbf{W})_{\mathcal{D}} \iff (\mathbf{Y} \perp\!\!\!\perp \mathbf{Z} \Big| \mathbf{W}, F_{[k]})_{Aug(\mathcal{D})} \tag{1}$$

For disjoint $\mathbf{Y}, \mathbf{W} \subseteq \mathbf{V}$, where $\mathbf{W}_i := \mathbf{W} \cap \mathbf{S}_i, \mathbf{R} := \mathbf{S}_i \setminus \mathbf{W}_i$:

$$(\mathbf{Y} \perp\!\!\!\perp \mathbf{S}_i | \mathbf{W} \setminus \mathbf{W}_i)_{\mathcal{D}_{\underline{\mathbf{W}_i}, \overline{\mathbf{R(W)}}}} \iff (\mathbf{Y} \perp\!\!\!\perp F_i \Big| \mathbf{W}, F_{[k] \setminus \{i\}})_{Aug(\mathcal{D})} \tag{2}$$

In order to characterize causal graphs that are $\mathcal{I}$-Markov equivalent, we draw some insight from the Markov equivalence of causal graphs with latents. Ancestral graphs, and more specifically MAGs, were proposed as a representation to encode the d-separation statements of a causal graph among the measured variables while not explicitly encoding the latent nodes. The definition below (Def. 4) introduces the *augmented MAG* that is constructed over an augmented graph. Since all the constraints in the $\mathcal{I}$-Markov definition can be tested by d-separation statements in the augmented graph, then an augmented MAG preserves all those constraints. For example, Figs. 2c and 2d present the augmented MAGs corresponding to the augmented graphs in Figs. 2a and 2b, respectively. Notice that F_x and Y are adjacent in both MAGs since they are not separable by any set in the augmented graphs.

Definition 4 (Augmented MAG)**.** *Given a causal graph $\mathcal{D} = (\mathbf{V} \cup \mathbf{L}, \mathbf{E})$ and an intervention set $\mathcal{I}$, the augmented MAG is the MAG constructed over $\mathbf{V}$ from $Aug_{\mathcal{I}}(\mathcal{D})$, i.e., $MAG(Aug_{\mathcal{I}}(\mathcal{D}))$.*

Below, we derive a characterization for two causal graphs to be $\mathcal{I}$-Markov equivalent – two causal graphs are $\mathcal{I}$-Markov equivalent if their corresponding augmented MAGs satisfy the three conditions given in Theorem 2. For example, the two augmented MAGs in Figures 2c and 2d satisfy the three conditions, hence the original causal graphs are in the same $\mathcal{I}$-Markov equivalence class.

Theorem 2. *Two causal graphs $\mathcal{D}_1 = (\mathbf{V} \cup \mathbf{L}_1, \mathbf{E}_1)$ and $\mathcal{D}_2 = (\mathbf{V} \cup \mathbf{L}_2, \mathbf{E}_2)$ are $\mathcal{I}$-Markov equivalent for a set of controlled experiments $\mathcal{I}$ if and only if for $\mathcal{M}_1 = MAG(Aug_{\mathcal{I}}(\mathcal{D}_1))$ and $\mathcal{M}_2 = MAG(Aug_{\mathcal{I}}(\mathcal{D}_2))$:*

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.*

5 Learning by Combining Observations and Experiments

In this section, we develop an algorithm to learn the augmented graph from a combination of observational and interventional data, which consequently recovers the causal graph. However, similar to the observational case, it is typically impossible to completely determine the causal graph from the available measured data, especially when latents are present. Then, the objective is to learn a class of augmented MAGs consistent with data. For this, we define an augmented PAG as follows.

Definition 5. *Given a causal graph $\mathcal{D}$ and an intervention set $\mathcal{I}$, let $\mathcal{M} = MAG(Aug_{\mathcal{I}}(\mathcal{D}))$ and let $[\mathcal{M}]$ be the set of augmented MAGs corresponding to all the causal graphs that are $\mathcal{I}$-Markov equivalent to $\mathcal{D}$. An Augmented PAG for $\mathcal{D}$, denoted $\mathcal{G} = PAG(Aug_{\mathcal{I}}(\mathcal{D}))$, is a graph such that:*

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

As with any learning algorithm, some faithfulness assumption is needed to infer graphical properties from the corresponding distributional constraints. Hence, we assume that the given interventional distributions are *c-faithful* to the causal graph $\mathcal{D}$ as defined below.

Algorithm 1 Algorithm for Learning Augmented PAG

function LearnAugPAG($\mathcal{I}, (P_{\mathbf{I}})_{\mathbf{I}\in\mathcal{I}}, \mathbf{V}$)
  $(\mathcal{F}, \mathcal{S}, \sigma) \leftarrow$ CreateAugmentedNodes($\mathcal{I}, \mathbf{V}$)
  $\mathbf{V} \leftarrow \mathbf{V} \cup \mathcal{F}$
  Phase I: Learn Adjacencies and Seperating Sets
  Form the complete graph $\mathcal{G}$ on $\mathbf{V}$ where between every pair of nodes there is an edge $\circ\!\!-\!\!\circ$.
  for Every pair $X, Y \in \mathbf{V}$ **do**
    if $X \in \mathcal{F} \wedge Y \in \mathcal{F}$ **then**
      $SepSet(X, Y) \leftarrow \emptyset, SepFlag(X, Y) = True$
    else
      $(SepSet(X, Y), SepFlag) \leftarrow$ Do-Constraints($(P_{\mathbf{I}})_{\mathbf{I}\in\mathcal{I}}, X, Y, V, \mathcal{F}, \sigma$)
    if $SepFlag = True$ **then**
      Remove the edge between X, Y in $\mathcal{G}$.
  Phase II: Learn Unshielded Colliders
  For every unshielded triple $\langle X, Z, Y\rangle$ in $\mathcal{G}$, orient it as $X{*\!\!\rightarrow} Z \leftarrow\!\!* Y$ iff $Z \notin SepSet(X, Y)$
  Phase III: Apply Orientation Rules
  Apply 7 FCI rules in [28] together with the following 2 additional rules until none applies.
  Rule 8: For any $F_k \in \mathcal{F}$, orient adjacent edges out of F_k.
  Rule 9: For any $F_k \in \mathcal{F}$ that is adjacent to a node $Y \notin \mathbf{S}_k$
    If $|\mathbf{S}_k| = 1$, orient $X *\!\!-\!\!* Y$ as $X \rightarrow Y$ for $X \in \mathbf{S}_k$.

Algorithm 2 Creating F-nodes.

function CreateAugmentedNodes($\mathcal{I}, \mathbf{V}$)
  $\mathcal{F} = \emptyset, \mathcal{S} = \emptyset, k = 0, \sigma : \mathbb{N} \rightarrow 2^{\mathbf{V}} \times 2^{\mathbf{V}}$
  for all pairs $\mathbf{I}, \mathbf{J} \in \mathcal{I}$, if $\mathbf{I}\triangle\mathbf{J} \notin \mathcal{S}$ **do**
    Set $k \leftarrow k + 1$, set $\mathbf{S}_k = \mathbf{I}\triangle\mathbf{J}$, add F_k to $\mathcal{F}$, add $\mathbf{S}_k$ to $\mathcal{S}$, set $\sigma(k) = (\mathbf{I}, \mathbf{J})$.
  return $\mathcal{F}, \mathcal{S}, \sigma$

Definition 6. *Consider a causal graph $\mathcal{D} = (\mathbf{V} \cup \mathbf{L}, \mathbf{E})$. A tuple of distributions $(P_{\mathbf{I}})_{\mathbf{I}\in\mathcal{I}} \in \mathcal{P}(\mathcal{D}, \mathbf{V})$ is called c-faithful to graph $\mathcal{D}$ if the converse for each of the conditions given in Definition 1 holds.*

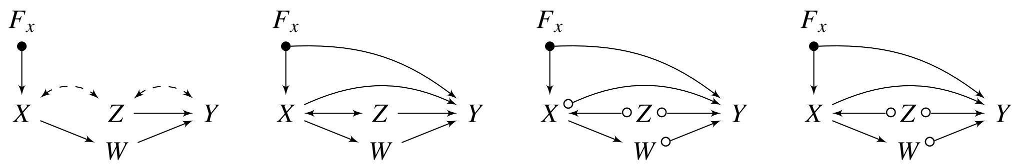
Algorithm 1 presents a modification of the FCI algorithm to learn augmented PAGs. To explain the algorithm, we first describe FCI which, given an independence model over the measured variables, proceeds in three phases [25]: In phase I, the algorithm initializes a complete graph with circle edges ($\circ\!\!-\!\!\circ$), then it removes the edge between any pair of nodes if a separating set between the pair exists and records the set. In phase II, the algorithm identifies unshielded triples $\langle A, B, C\rangle$ and orients the edges into B if B is not in the separating set of A and C. Finally, in phase III, FCI applies the orientation rules. Only one of the rules uses separating sets while the rest use MAG properties, and soundness and completeness of the previous phases – the skeleton is correct and all the unshielded colliders are discovered. We note that FCI looks for any separating sets, and not necessarily the minimal ones. We also observe that if two nodes X, Y are separated given $\mathbf{Z}$ in $Aug_{\mathcal{I}}(\mathcal{D})$, they are also separated given $\mathbf{Z} \cup \mathcal{F}$ since $\mathcal{F}$ are root nodes by construction, i.e., all the edges incident on F-nodes are out of them.

Algorithm 1 follows a similar flow to that of the FCI. In phase I, it learns the skeleton of the augmented PAG. Function CreateAugmentedNodes($\cdot$) in Alg. 2 creates the F-nodes by computing the set $\mathcal{S}$ of unique symmetric difference sets from all pairs of interventions in $\mathcal{I}$. Sigma (σ) maps every F-node to a source pair of interventions, which is used later on to perform the do-tests. The algorithm starts by creating a complete graph of circle edges between $\mathbf{V} \cup \mathcal{F}$. Then, it removes the edge between any two nodes X and Y if a separating set exists. If the two nodes are F-nodes, then they are separated by the empty set by construction. Otherwise, it calls the function Do-Constraints($\cdot$) in Alg. 3 to search for a separating set using the corresponding do-constraints. The function routine works as follows: If the two nodes are random variables (and not F-nodes), then an arbitrary distribution is chosen and we find a subset $\mathbf{W}$ that establishes conditional independence between X and Y (rule 1 of Thm. 1). Else, one of the two nodes is an F-node; without loss of generality, we choose it to be X. The algorithm then looks for a subset $\mathbf{W}$ that satisfies the invariance of Corollary 1, i.e., $P_{\mathbf{I}}(y|\mathbf{w}) = P_{\mathbf{J}}(y|\mathbf{w})$.

Phase II of Alg. 1 is similar to the FCI counterpart. For the edge orientation phase, note that the augmented MAG is a MAG indeed, hence all the FCI orientation rules still apply. Therefore, phase III

Algorithm 3 Find m-separation sets via Calculus Tests.

function Do-Constraints($\mathcal{I}, (P_{\mathbf{I}})_{\mathbf{I}\in\mathcal{I}}, X, Y, \mathbf{V}, \mathcal{F}, \sigma$)
  $SepSet = \emptyset, SepFlag = False$
  if $X \notin \mathcal{F} \wedge Y \notin \mathcal{F}$ **then**
    Pick $\mathbf{I} \in \mathcal{I}$ arbitrarily.
    for $\mathbf{W} \subseteq \mathbf{V} \setminus \mathcal{F}$ **do**
      if $P_{\mathbf{I}}(y|\mathbf{w}, x) = P_{\mathbf{I}}(y|\mathbf{w})$ **then** $SepSet = \mathbf{W} \cup \mathcal{F}, SepFlag = True$
  else
    Suppose $X \in \mathcal{F}, Y \notin \mathcal{F}$ and $X = F_i$ without loss of generality.
    $(\mathbf{I}, \mathbf{J}) = \sigma(i)$
    for $\mathbf{W} \subseteq \mathbf{V} \setminus (\mathcal{F} \cup \mathbf{Y})$ **do**
      if $P_{\mathbf{I}}(y|\mathbf{w}) = P_{\mathbf{J}}(y|\mathbf{w})$ **then** $SepSet = \mathbf{W}, \mathcal{F} \setminus \{F_i\}, SepFlag = True$
  return $(SepSet, SepFlag)$



(a) Causal Graph $\mathcal{D}$ (b) $MAG(Aug(\mathcal{D}))$ (c) Before rule 9. (d) After rule 9.

Figure 3: An example of learning the augmented PAG from the distributions P, P_x consistent with the given causal graph. Rule 9 allows orienting the tail at $X \circ\!\!\rightarrow Y$.

uses the FCI orientation rules along with the following two new ones. The algorithm keeps applying the rules until none applies anymore.

Rule 8 (F-node Edges): For any edge adjacent to an F node, orient the edge out of the F node.

Rule 9 (Inducing Paths): If $F_k \in \mathcal{F}$ is adjacent to a node $Y \notin \mathbf{S}_k$ and $|\mathbf{S}_k| = 1$, e.g., $\mathbf{S}_k = \{X\}$, then orient $X *\!\!-\!\!* Y$ out of X, i.e., $X \rightarrow Y$. The intuition for this rule is as follows: If F_k is adjacent to a node $Y \notin \mathbf{S}_k$ in $\mathcal{G}$, then there is an inducing path p between F_k and Y in $Aug_{\mathcal{I}}(\mathcal{D})$, where $\mathcal{D}$ is any causal graph in the equivalence class. Since F_k is a root node and by the properties of inducing paths, the subpath of p from X to Y is an inducing path as well and X is an ancestor of Y in $Aug_{\mathcal{I}}(\mathcal{D})$. Hence, the edge between X and Y is out of X and into Y in $MAG(Aug_{\mathcal{I}}(\mathcal{D}))$ and consequently in $\mathcal{G}$.

We give an example to illustrate the steps of the algorithm in Figure 3, where $\mathcal{I} = \{\emptyset, \{X\}\}$. Figure 3a shows the augmented causal graph, i.e., $Aug_{\mathcal{I}}(\mathcal{D})$, and Figure 3b shows the corresponding augmented MAG, i.e., $MAG(Aug_{\mathcal{I}}(\mathcal{D}))$. Nodes F_x and Z are separable in $Aug_{\mathcal{I}}(\mathcal{D})$ given the empty set and this can be tested by the do-constraint $P(Z) = P_X(Z)$. Similarly, we can infer the separation of F_x and W by the test $P(W|X) = P_X(W|X)$. Figure 3c shows the graph obtained after applying the seven rules of the FCI together with Rule 8. Finally, by applying Rule 9, we infer that the edge between X and Y has a tail at X and we obtain the graph in Figure 3d. The soudness of the algorithm is shown next.

Theorem 3. *Consider a set of interventional distributions $(P_{\mathbf{I}})_{\mathbf{I}\in\mathcal{I}}$ c-faithful to a causal graph $\mathcal{D} = (\mathbf{V} \cup \mathbf{L})$, where $\mathcal{I}$ is a set of controlled experiments. Algorithm 1 is sound, i.e., every adjacency and orientation is common for all $MAG(Aug(\mathcal{D}'))$ where $\mathcal{D}'$ is $\mathcal{I}$-Markov equivalent to $\mathcal{D}$.*

6 Conclusions

We investigate the problem of learning the causal structure underlying a phenomenon of interest from a combination of observational and experimental data. We pursue this endeavor by noting that a generalization of the converse of Pearl's do-calculus (Thm. 1) leads to new tests that can be evaluated against data. These tests, in turn, translate into constraints over the structure itself. We then define an interventional equivalence class based on such criteria (Def. 1), and then derive a graphical characterization for the equivalence of two causal graphs (Thm. 2). Finally, we develop an algorithm to learn an interventional equivalence class from data, which includes new orientation rules.

Acknowledgements

Bareinboim and Jaber are supported in parts by grants from NSF IIS-1704352, IIS-1750807 (CAREER), IBM Research, and Adobe Research. 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] Elias Bareinboim and Judea Pearl. Causal inference and the data-fusion problem. *Proceedings of the National Academy of Sciences*, 113(27):7345–7352, July 2016.

[3] David Maxwell Chickering. Optimal structure identification with greedy search. *Journal of Machine Learning Research*, 3:507–554, 2002.

[4] Juan Correa and Elias Bareinboim. A calculus for stochastic interventions: Causal effect identification and surrogate experiments. Technical report, R-51, Causal Artificial Intelligence Laboratory, Columbia University, New York, 2019.

[5] A Philip Dawid. Conditional independence in statistical theory. *Journal of the Royal Statistical Society, Series B*, 41(1):1–31, 1979.

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

[7] Frederick Eberhardt. *Causation and Intervention*. PhD thesis, Department of Philosophy, Carnegie Mellon University, 2007.

[8] AmirEmad Ghassami, Saber Salehkaleybar, Negar Kiyavash, and Elias Bareinboim. Budgeted experiment design for causal structure learning. In *International Conference on Machine Learning (ICML)*, pages 1719–1728, 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] Antti Hyttinen, Frederick Eberhardt, and Patrik Hoyer. Experiment selection for causal discovery. *Journal of Machine Learning Research*, 14:3041–3071, 2013.

[12] 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.

[13] Sara Magliacane, Tom Claassen, and Joris M Mooij. Joint causal inference on observational and experimental datasets. *arXiv preprint arXiv:1611.10351*, 2016.

[14] Christopher Meek. Strong completeness and faithfulness in bayesian networks. In *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence*, pages 411–418. Morgan Kaufmann Publishers Inc., 1995.

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

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

[17] Judea Pearl. Aspects of graphical models connected with causality. *Proceedings of the 49th Session of the International Statistical Institute*, 1(August):399–401, 1993.

[18] Judea Pearl. Causal diagrams for empirical research. *Biometrika*, 82(4):669–688, 1995.

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

[20] Karthikeyan Shanmugam, Murat Kocaoglu, Alexandros G Dimakis, and Sriram Vishwanath. Learning causal graphs with small interventions. In *Advances in Neural Information Processing Systems*, pages 3195–3203, 2015.

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

[22] Sofia Triantafillou and Ioannis Tsamardinos. Constraint-based causal discovery from multiple interventions over overlapping variable sets. *J. Mach. Learn. Res.*, 16(1):2147–2205, January 2015.

[23] Yuhao Wang, Liam Solus, Karren Yang, and Caroline Uhler. Permutation-based causal inference algorithms with interventions. In *Advances in Neural Information Processing Systems*, pages 5822–5831, 2017.

[24] Karren Yang, Abigail Katoff, and Caroline Uhler. Characterizing and learning equivalence classes of causal dags under interventions. In *ICML*, 2018.

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

[26] Jiji Zhang. A characterization of markov equivalence classes for directed acyclic graphs with latent variables. In *Proceedings of the Twenty-Third Conference on Uncertainty in Artificial Intelligence*, UAI'07, pages 450–457. AUAI Press, 2007.

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

[28] 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.

7 Appendix

7.1 Do-calculus rules for soft interventions

A recent work developed an extension of the do-calculus rules to soft interventions in structural causal models (SCMs) [4]. We reproduce a variation of this result for CBNs for completeness.

Proof of Theorem 1. Note that a soft-intervention does not change the underlying causal graph. Since interventional distribution factorizes with respect to the original graph, any m-separation statement in graph D implies conditional independence. Under the strict positivity, conditional independence is equivalent to the invariance given in the rule, which concludes the proof.

For the proof of next two rules, similar to [18], we introduce F-nodes as random variables. Notice that this is different than the augmented graph construction we have in the main text, where we treat F-nodes as parameters. This is allowed as only a single F-node is introduced to show the result here, which is explained next.

Construct the probability distribution P^* on $V \cup \{F\}$ as follows: $P^*(V|F = 0) = P_{x,z}(V), P^*(V|F = 1) = P_x(V)$, where $P_{x,z}$ is the interventional distribution after a soft intervention on the set $X \cup Z$ and P_x is the interventional distribution after a soft intervention on the set X is performed. Marginal distribution of $P^*(F)$ can be picked arbitrarily from the set of strictly positive distributions for our purposes. Assume that interventions are controlled, i.e., $P_{x,z}(x|pa_x) = P_x(x|pa_x)$, where pa_x is the set of parents of node X.

The desired equality in Rule 2 can be rewritten as $P^*(y|z, w, F = 0) = P^*(y|z, w, F = 1)$. Under the assumption of strictly positive distributions, this invariance is implied by the conditional independence statements $(Y \perp\!\!\!\perp Z|W)_{P^*}$. Therefore, we need to show that the graph separation statement given in the rule implies the desired conditional independence statement.

For this, observe that P^* can be factorized as follows:

$$P^*(V, F) = P^*(F)P^*(V|F) = P^*(F)P^*(z|pa_z, F) \prod_{u \neq z} P(u|pa_u). \tag{3}$$

where pa_x are the parents of x in D. Note that in G the set of parents of Z is $pa_Z \cup F$. Therefore, P^* factorizes according to the graph G. This implies that any d-separation statement on G implies conditional independence [16][Theorem 1.2.4]. Therefore, we only need to show that the separation statement given in the rule on mutilated graph implies d-separation statement between F_z and Y given Z, W.

If $Y \perp\!\!\!\perp Z|W$ in $D_{\underline{Z}}$, this means there is no backdoor path from Z to Y that is active conditioned on W. Since F_z only has an edge into Z, conditioned on W, Z any d-connecting path to Y must go through a backdoor from Z. However the statement $Y \perp\!\!\!\perp Z|W$ in $D_{\underline{Z}}$ implies this cannot happen, implying that $F_z \perp\!\!\!\perp Y|Z, W$ in G, completing the proof.

For the proof of rule 3, we use a similar argument under strict positivity. Consider the same P^* construction. Similarly, this distribution factorizes with respect to graph G which means and d-separation statement implies conditional independence. Therefore we only need to show that the given separation statement in the mutilated graph implies the desired d-separation statement in G. Suppose $Y \perp\!\!\!\perp Z|W$ in $D_{\overline{Z(W)}}$. This implies that given W, there is no active path from the nodes in $Z - Z(W)$ to Y. Moreover there is no front-door path from the elements of $Z(W)$ to Y given W. Suppose for the sake of contradiction that $F_Z \not\perp\!\!\!\perp Y|W$ in G. Since F_Z only has edges into Z, any active path must go through an element in Z. Suppose it goes through an element in $Z(W)$. Since no descendant of $Z(W)$ is conditioned on, the active path must go through a backdoor in $Z(W)$. However this would imply $Y \not\perp\!\!\!\perp Z|W$ in $D_{\overline{Z(W)}}$, which leads to contradiction. Now suppose active path goes through an element in $Z - Z(W)$. However, these nodes are not mutilated in G, hence the same active path would persist in D as well, contradicting with the statement $Y \perp\!\!\!\perp Z|W$ in $D_{\overline{Z(W)}}$. Therefore we have $F_Z \perp\!\!\!\perp Y|W$ in G which concludes the proof. $\square$

7.2 Generalized Do-calculus Rules

In this section, we extend the do-calculus rules to be able to apply them across two arbitrary interventions. This is essential for the characterizing of our equivalence class, when arbitrary sets of interventional distributions are available.

Proposition 2 (Generalized do-calculus for soft interventions)**.** *Let $\mathcal{D} = (\mathbf{V} \cup \mathbf{L}, \mathbf{E})$ be a causal graph with latents. Then, the following holds for any tuple of strictly positive soft-interventional distributions $(P_I)_{I\in\mathcal{I}}$ consistent with $\mathcal{D}$, where $\mathcal{I} \subseteq 2^{\mathbf{V}}$.*

Rule 1 (conditional independence): *For any $I \subseteq \mathbf{V}$ and disjoint $Y, Z, W \subseteq \mathbf{V}$*

$$P_I(y|w,z) = P_I(y|w) \text{ if } Y \perp\!\!\!\perp Z\,|W \text{ in } D. \tag{4}$$

Rule 2 (do-see): *For any $I, J \subseteq \mathbf{V}$ and disjoint $Y, W \subseteq \mathbf{V}\backslash K$, where $K := I \triangle J$*

$$P_I(y|w,k) = P_J(y|w,k) \text{ if } Y \perp\!\!\!\perp K\,|W \text{ in } D_{\underline{K}}. \tag{5}$$

Rule 3 (do-do): *For any $I, J \subseteq \mathbf{V}$ and disjoint $Y, W \subseteq \mathbf{V}\backslash K$, where $K := I \triangle J$*

$$P_I(y|w) = P_J(y|w) \text{ if } Y \perp\!\!\!\perp K\,|W \text{ in } D_{\overline{K(W)}}. \tag{6}$$

Rule 4 (mixed do-do/do-see): *For any $I, J \subseteq \mathbf{V}$ and disjoint $Y, W \subseteq \mathbf{V}$, where $K := I \triangle J$*

$$P_I(y|w) = P_J(y|w) \text{ if } Y \perp\!\!\!\perp K\,|W \setminus W_k \text{ in } D_{\underline{W_k},\overline{R(W)}}, \tag{7}$$

where $W_k =: W \cap K$ and $R := K\backslash W_k$.

Note that Rule 2 and Rule 3 are special cases of Rule 4. We present all three to make the connection to standard causal calculus rules more explicit.

Proof. Let $K_I := I\backslash J, K_J := J\backslash I, T := I \cap J$.

Rule 1: The result follows from the rule 1 of Theorem 1.

Rule 2: We have the following lemma:

Lemma 1. *If $Y \perp\!\!\!\perp K\,|W$ in $G_{\underline{K}}$ then $Y \perp\!\!\!\perp K_I\,|W, K_J$ in $G_{\underline{K_I}}$ and $Y \perp\!\!\!\perp K_J\,|W, K_I$ in $G_{\underline{K_J}}$.*

Proof. Suppose for the sake of contradiction that $Y \perp\!\!\!\perp K_I\,|W, K_J$ in $G_{\underline{K_I}}$ does not hold, then there exist a corresponding active path, denoted p. If every collider along p is active due to a node in W and not K_J, then p is active in $G_{\underline{K}}$ as well which contradicts the input. Otherwise, let $K_J^* \in K_J$ be the node activating the last collider S along p (where possibly $K_J^* = S$) starting from K_I. The path p' composed of the directed path from S to K_J^* concatenated with the subpath of p from S to Y is active in $G_{\underline{K}}$ which contradicts the input. Hence, $Y \perp\!\!\!\perp K_I\,|W, K_J$ in $G_{\underline{K_I}}$. Similarly, we can show that $Y \perp\!\!\!\perp K_J\,|W, K_I$ in $G_{\underline{K_J}}$. □

Therefore we can apply rule 2 of Theorem 1 to obtain $P_I(y|w,k) = P_T(y|w,k)$. Furthermore, we can apply rule 2 of Theorem 1 once more to obtain $P_T(y|w,k) = P_J(y|w,k)$, which concludes the proof.

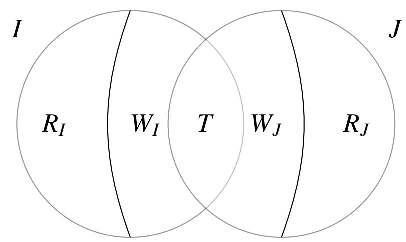
Rule 3: We have the following lemma:

Lemma 2. *If $Y \perp\!\!\!\perp K\,|W$ in $G_{\overline{K(W)}}$, then $Y \perp\!\!\!\perp K_I\,|W$ in $G_{\overline{K_I(W)}}$ and $Y \perp\!\!\!\perp K_J\,|W$ in $G_{\overline{K_J(W)}}$.*

Proof. If $Y \perp\!\!\!\perp K\,|W$ in $G_{\overline{K(W)}}$, then clearly $Y \perp\!\!\!\perp K_I\,|W$ in $G_{\overline{K(W)}}$. Suppose for the sake of contradiction, we have $Y \not\perp\!\!\!\perp K_I\,|W$ in $G_{\overline{K_I(W)}}$. Notice that the only difference between $G_{\overline{K(W)}}$ and $G_{\overline{K_I(W)}}$ are the incoming edges into $K_J(W)$. Therefore, the active path p between K_I and Y in $G_{\overline{K_I(W)}}$ must include a vertex $S \in K_J(W)$ and also must pass through an edge that is into S. Otherwise, p would be active in the graph $G_{\overline{K(W)}}$ which contradicts the input.Since no descendant of $K_J(W)$ is conditioned on by definition, no descendant of S is conditioned on. Also, since p is active, then S cannot be a collider on p. This implies that the other edge that is adjacent to S must be out of it. Moreover, along the subpath of p that is out of S, denoted p', none of the nodes is a collider. Suppose otherwise for the sake of contradiction and let X be the first collider. since p is active, then we condition on a descendant of X. Since the path from S to X is a directed path out of S, this contradicts the condition that S is not an ancestor of a node in W. Therefore, p' is a directed path out of K_J and can be either into K_I or into Y. If p' is into $A \in K_I$, then it must be that $A \notin K_I(W)$.So, S is an ancestor of A and A is an ancestor of some node in W which contradicts the condition $S \in K_J(W)$. Hence, p' must be a directed path out of S and into Y. This path is active in $G_{\overline{K_I(W)}}$ and consequently in $G_{\overline{K(W)}}$ which contradicts the separation statement in the assumption. Hence, $Y \perp\!\!\!\perp K_I \,|W$ in $G_{\overline{K_I(W)}}$. Similarly, we can show that $Y \perp\!\!\!\perp K_J \,|W$ in $G_{\overline{K_J(W)}}$. □

Since, $Y \perp\!\!\!\perp K_I \,|W$ in $G_{\overline{K_I(W)}}$, then we have $P_I(y|w) = P_T(y|w)$ by rule 3 of Theorem 1. Similarly, since $Y \perp\!\!\!\perp K_J \,|W$ in $G_{\overline{K_J(W)}}$ we have $P_T(y|w) = P_J(y|w)$. This concludes the proof.

Rule 4: In addition to the notation defined in rule 4, let $W_I := W_k \cap I$, $W_J := W_k \cap J$, $R_I := R \cap I$, $R_J := R \cap J$. The following venn diagram summarizes those relations.



First, we establish the following. Note that $R_I \cup R_J = R$ and $W_I \cup W_J = W_k$.

Lemma 3. *If $Y \perp\!\!\!\perp K \,|W \setminus W_k$ in $D_{\underline{W_k},\overline{R(W)}}$, then $Y \perp\!\!\!\perp R \,|W$ in $D_{\overline{R(W)}}$ and $Y \perp\!\!\!\perp W_k \,|W \setminus W_k$ in $D_{\underline{W_k}}$.*

Proof. If $Y \perp\!\!\!\perp K \,|W \setminus W_k$ in $D_{\underline{W_k},\overline{R(W)}}$, then $Y \perp\!\!\!\perp R \,|W \setminus W_k$ in $D_{\underline{W_k},\overline{R(W)}}$ since $R \subset K$. Suppose for the sake of contradiction that $Y \not\perp\!\!\!\perp R \,|W$ in $D_{\overline{R(W)}}$ and let p be one active path between Y and R. The difference between $D_{\underline{W_k},\overline{R(W)}}$ and $D_{\overline{R(W)}}$ is cutting the edges out of W_k. Hence, p is discontinued or blocked in $D_{\underline{W_k},\overline{R(W)}}$ conditioned on $W \setminus W_k$ due to one of two conditions: (1) p includes a non-collider node in W_k, or (2) p has a collider S that is active because it has a descendant in W_k (possibly $S \in W_k$). Case (1) is not possible because $W_k \subset W$ and p would be blocked in $D_{\overline{R(W)}}$ which contradicts the assumption that p is active. Consider the collider along p closest to Y that is consistent with case (2). The directed path from S to the node in W_k concatenated with the subpath of p from S to Y is active given $W \setminus W_k$ in $D_{\underline{W_k},\overline{R(W)}}$ which contradicts the input condition. Thus, $Y \perp\!\!\!\perp R \,|W$ in $D_{\overline{R(W)}}$ and this concludes the proof of first part.

If $Y \perp\!\!\!\perp K \,|W \setminus W_k$ in $D_{\underline{W_k},\overline{R(W)}}$, then $Y \perp\!\!\!\perp W_k \,|W \setminus W_k$ in $D_{\underline{W_k},\overline{R(W)}}$ since $W_k \subset K$. Suppose for the sake of contradiction that $Y \not\perp\!\!\!\perp W_k \,|W \setminus W_k$ in $D_{\underline{W_k}}$ and let p denote any active path. The only difference between the two graphs is the set of incoming edges to $R(W)$. Therefore, p contains an edge into a node $S \in R(W)$ so that p is active in $D_{\underline{W_k}}$ and blocked in $D_{\underline{W_k},\overline{R(W)}}$. Since $R(W)$ are by definition non-ancestors of W, S cannot be a collider in $D_{\underline{W_k}}$ otherwise it would be blocked. Since S is a non-collider, the other edge adjacent to S must be out of S. Moreover, along the subpath of p that is out of S, denoted p', none of the nodes is a collider. Suppose otherwise for the sake of contradiction and let X be the first collider. since p is active, then we condition on a descendant of X. Since the path from S to X is a directed path out of S, this contradicts the condition that S is not an ancestor of a node in W ($S \in R(W)$). Therefore, p' is a directed path out of S and can be either into Y or into a node in W_k. It p' is into W_k, then S is an ancestor of a node in W which is a contradiction since $S \in R(W)$. If p' is into Y then p' is active in $D_{\underline{W_k},\overline{R(W)}}$ which contradicts the input condition that $Y \perp\!\!\!\perp K \,|W \setminus W_k$. This concludes the proof of the second claim. □

We establish the following equivalences which prove rule 4. Note that the first and the last equivalences follows by definition.

$$P_I(y|w) = P_{R_I \cup W_I \cup T}(y|w) = P_{R_J \cup W_I \cup T}(y|w) = P_{R_J \cup W_J \cup T}(y|w) = P_J(y|w)$$

The second equality is an application of rule 3 since $Y \perp\!\!\!\perp R|W$ in $D_{\overline{R(W)}}$ and the third equality is an application of rule 2 since $Y \perp\!\!\!\perp W_k|W \setminus W_k$ in $D_{\underline{W_k}}$. This concludes the proof. □

Proof of Corollary 1. The correctness follows by Rule 4 of Proposition 2. □

7.3 Generalized do-calculus Graph Mutilations and F-node Equivalence

We show graphical conditions on the augmented graph that are equivalent to those given in the generalized causal calculus rules.

Proposition 3. *Consider a CBN $(D = (V \cup L, E), p)$ with latent variables L and its augmented graph $Aug_{\mathcal{I}}(D) = (V \cup L \cup \mathcal{F}, E \cup \mathcal{E})$ with respect to an intervention set $\mathcal{I}$, where $\mathcal{F} = \{F_i\}_{i \in [k]}$. Let S_i be the set of nodes adjacent to $F_i, \forall i \in [k]$. We have the following equivalence relations:*

Suppose Y, Z, W are disjoint subsets of V. We have

$$(Y \perp\!\!\!\perp Z|W)_D \iff (Y \perp\!\!\!\perp Z \Big| W, F_{[k]})_{Aug(D)} \tag{8}$$

For each S_i, suppose Y, W are disjoint subsets of $V \setminus S_i$. We have

$$(Y \perp\!\!\!\perp S_i|W)_{D_{\underline{S_i}}} \iff (Y \perp\!\!\!\perp F_i \Big| W, S_i, F_{[k]\setminus\{i\}})_{Aug(D)} \tag{9}$$

$$(Y \perp\!\!\!\perp S_i|W)_{D_{\overline{S_i(W)}}} \iff (Y \perp\!\!\!\perp F_i \Big| W, F_{[k]\setminus\{i\}})_{Aug(D)} \tag{10}$$

For each S_i, let $Y \subseteq V$ and $W \subseteq V$. Let $W_i := W \cap S_i, R := S_i \setminus W_i$. Then we have

$$(Y \perp\!\!\!\perp S_i|W \setminus W_i)_{D_{\underline{W_i},\overline{R(W)}}} \iff (Y \perp\!\!\!\perp F_i \Big| W, F_{[k]\setminus\{i\}})_{Aug(D)} \tag{11}$$

Proof. Conditioning on a source node is equivalent to removing it from the graph in terms of the graph separation statements. Hence, conditioning on $F_{[k]\setminus\{i\}}$ in the right-hand side eliminates them. Therefore, equations (8), (9), and (10) follow from [18, Proof of Th. 4.1] by Pearl. In what follows, we prove (11).

We first consider the case when $Y \cap S_i \neq \emptyset$. Then the relation is trivially true since it implies that for some $U \in S_i$, U and F_i are adjacent in $Aug(D)$ and Y is dependent with U since $U \subseteq Y$.

In the rest of the proof, suppose $Y \subseteq V \setminus S_i$.

Suppose $(Y \not\perp\!\!\!\perp S_i|W \setminus W_i)_{D_{\underline{W_i},\overline{R(W)}}}$, and let p denote any active path from $A \in S_i$ to Y. Note that the same path is active in $Aug(D)$ given $W, F_{[k]\setminus\{i\}}$. If p is into A, then either (1) $A \in W_i$ or (2) $A \notin R(W)$. Hence, the concatenation of p with $F_i \rightarrow A$ is active in $Aug(D)$ given $W, F_{[k]\setminus\{i\}}$ since $A \in W$ for case (1) and A has a descendant in W for case(2). Hence, $(Y \not\perp\!\!\!\perp F_i \Big| W, F_{[k]\setminus\{i\}})_{Aug(D)}$.

Next, suppose $(Y \not\perp\!\!\!\perp F_i \Big| W, F_{[k]\setminus\{i\}})_{Aug(D)}$ and let p denote any active path. Also, let A be the closest node to Y along p such that A is active due to a node in S_i, i.e., $A \in S_i$ is along p or $A \notin S_i$ is an active collider due to a descendant in $W_i \subseteq S_i$. If A is a non-collider along p, then $A \in R \subseteq S_i$ else p is blocked. If the subpath from A to Y is out of A, then this subpath is active in $D_{\underline{W_i},\overline{R(W)}}$ given $W \setminus W_i$ and $(Y \not\perp\!\!\!\perp S_i|W \setminus W_i)_{D_{\underline{W_i},\overline{R(W)}}}$. Otherwise, the subpath between A and F_i is out of A. In this case, we argue that $A \notin R(W)$, hence the subpath from A to Y along p is active in $D_{\underline{W_i},\overline{R(W)}}$ given $W \setminus W_i$ and $(Y \not\perp\!\!\!\perp S_i|W \setminus W_i)_{D_{\underline{W_i},\overline{R(W)}}}$. Since all the edges incident on F_i are out of it, then there exist at least one collider between A and F_i along p. Let X denote such a collider closest to A. Since X is active, then X has a descendant in W, thus A has a descendant in W through X and $A \notin R(W)$. Alternatively, A is an active collider along p. If $A \in W_i$ or $A \notin R(W)$, then the path from A to Y is active and $(Y \not\perp\!\!\!\perp S_i|W \setminus W_i)_{D_{\underline{W_i},\overline{R(W)}}}$. Not that A can't be in $R(W)$, else A would be blocked along p. Finally, $A \notin S_i$ and it has a descendant in W_i. In this case, the directed path from A to the node in W_i concatenated with the subpath of p from A to Y is active in $D_{\underline{W_i},\overline{R(W)}}$ given $W \setminus W_i$ and $(Y \not\perp\!\!\!\perp S_i|W \setminus W_i)_{D_{\underline{W_i},\overline{R(W)}}}$. This concludes the proof. □

Proof of Proposition 1. This follows from Proposition 3. □

7.4 Proof of Theorem 2

Proof. **(If)** Suppose $MAG(Aug_{\mathcal{I}}(D_1))$ and $MAG(Aug_{\mathcal{I}}(D_2))$ satisfy the three conditions. Then, they induce the same m-separation statements and vice-versa [26, Prop. 1 & Def. 5]. It follows by Proposition 1 that $\mathcal{D}_1$ and $\mathcal{D}_2$ impose the same constraints over the distribution tuples in Definition 1. Therefore, $\mathcal{P}_{\mathcal{I}}(D_1, V) = \mathcal{P}_{\mathcal{I}}(D_2, V)$.

(Only if) Suppose $\mathcal{M}_1 = MAG(Aug_{\mathcal{I}}(D_1))$ and $\mathcal{M}_2 = MAG(Aug_{\mathcal{I}}(D_2))$ do not satisfy the three conditions. Then, they must induce at least one different m-separation statement. What is left is to establish that if the two graphs induce different m-separation statements, then they are not $\mathcal{I}$-Markov equivalent. We first introduce some necessary definitions for types of separation sets, where $\mathcal{M}$ is an arbitrary augmented MAG.

$$\mathcal{U} = \{(X \perp\!\!\!\perp Y|\mathbf{Z},\mathbf{F})_{\mathcal{M}} : X, Y \in \mathbf{V} \cup \mathcal{F},\ \mathbf{Z} \subseteq \mathbf{V} - \{X, Y\},\ \mathbf{F} \subsetneq \mathcal{F} - \{X, Y\}\} \tag{12}$$
$$\mathcal{O} = \{(X \perp\!\!\!\perp Y|\mathbf{Z},\mathbf{F})_{\mathcal{M}} : X, Y \in \mathbf{V} \cup \mathcal{F},\ \mathbf{Z} \subseteq \mathbf{V} - \{X, Y\},\ \mathbf{F} = \mathcal{F} - \{X, Y\}\} \tag{13}$$
$$\mathcal{T} = \{(X \perp\!\!\!\perp Y|\mathbf{Z},\mathbf{F})_{\mathcal{M}} : X \in \mathbf{V},\ Y \in \mathbf{V} \cup \mathcal{F},\ \mathbf{Z} \subseteq \mathbf{V} - \{X, Y\},\ \mathbf{F} = \mathcal{F} - \{X, Y\}\} \tag{14}$$

$\mathcal{U}$ are the set of m-separation statements between any two nodes given $\mathbf{Z}$, a subset of the nodes corresponding to the measured variables, and a strict subset of all the remaining F nodes. $\mathcal{O}$ are the set of m-separation statements between any two nodes given $\mathbf{Z}$ and all the remaining F nodes. $\mathcal{T}$ are the set of m-separation statements between a node corresponding to a measured variable and any other node given $\mathbf{Z}$ and all the remaining F nodes. Note that $\mathcal{U}, \mathcal{O}$ are disjoint, whereas $\mathcal{T}$ is a subset of $\mathcal{O}$. By Prop. 1 and Def. 1, we see that an m-separation statement is in $\mathcal{T}$ if and only if it corresponds to a graphical condition in the definition of $\mathcal{I}$-Markov property. Hence, we refer to a separation statement in $\mathcal{T}$ as a *testable separation statement*. Also, if an m-separation between arbitrary subsets of nodes holds in causal graph $\mathcal{D}_1$ but not in $\mathcal{D}_2$, then by definition, there is at least one pair of singletons for which the corresponding m-separation holds in $\mathcal{D}_1$ but not in $\mathcal{D}_2$. Therefore, it is sufficient to consider m-separation statements between singletons which are included in $\mathcal{U} \cup \mathcal{O} \cup \mathcal{T}$. Accordingly, the following lemma is central to establish the **(only if)** side.

Lemma 4. *Suppose $(A \perp\!\!\!\perp B|\mathbf{C})_{\mathcal{M}_1}, (A \not\perp\!\!\!\perp B|\mathbf{C})_{\mathcal{M}_2}$, where $A, B \in \mathbf{V} \cup \mathcal{F}$ and $\mathbf{C} \subseteq (\mathbf{V} \cup \mathcal{F}) \setminus \{A, B\}$. Then at least one of the following is true:*

$$(a)\ \exists X, Y \in \mathbf{V}, \mathbf{Z} \subseteq \mathbf{V} \text{ such that } (X \perp\!\!\!\perp Y|\mathbf{Z}, \mathcal{F})_{\mathcal{M}_1} \wedge (X \not\perp\!\!\!\perp Y|\mathbf{Z}, \mathcal{F})_{\mathcal{M}_2} \tag{15}$$
$$(b)\ \exists T \in \mathbf{V}, \mathbf{W} \subseteq \mathbf{V}, F_i \in \mathcal{F} \text{ such that } (F_i \perp\!\!\!\perp T|\mathbf{W}, \mathcal{F} \setminus \{F_i\})_{\mathcal{M}_1} \wedge (F_i \not\perp\!\!\!\perp T|\mathbf{W}, \mathcal{F} \setminus \{F_i\})_{\mathcal{M}_2} \tag{16}$$

Proof. The statement of the lemma can be rephrased as follows: Any difference in the truth value of any m-separation statement from the set $\mathcal{U} \cup \mathcal{O} \cup \mathcal{T}$ between $\mathcal{M}_1$ and $\mathcal{M}_2$ implies a difference between truth value of some m-separation statement in $\mathcal{T}$. We show this in two steps:

1. **Step 1**: If there is any difference in the truth value of any m-separation statement from the set $\mathcal{U} \cup \mathcal{O} \cup \mathcal{T}$ between $\mathcal{M}_1$ and $\mathcal{M}_2$, then this implies a difference between the truth value of some m-separation statement in $\mathcal{O}$.

2. **Step 2**: If there is any difference in the truth value of any m-separation statement in $\mathcal{O}$ between $\mathcal{M}_1$ and $\mathcal{M}_2$, then this implies a difference in the truth value of some m-separation statement in $\mathcal{T}$ between the two augmented MAGs.

The detailed proof goes as follows.

Proof of Step 1: The main result in this step is given by Corollary 2. We have the following Lemma that relates m-separation statements from $\mathcal{U}$ to other m-separation statements that are 'closer' to $\mathcal{O}$. Recursively applying this lemma proves the result in this step.

Lemma 5. *Let $\mathcal{M}$ be the augmented MAG with respect to a causal graph $\mathcal{D}$. Consider any m-separation statement of the form $(X \perp\!\!\!\perp Y|\mathbf{Z},\mathbf{F})_{\mathcal{M}}$ where $X, Y \in \mathbf{V} \cup \mathcal{F}$, $\mathbf{Z} \subseteq \mathbf{V} - \{X, Y\}$, and $\mathbf{F} \subsetneq \mathcal{F} - \{X, Y\}$. For any $F_i \in \mathcal{F} - (\mathbf{F} \cup \{X, Y\})$, the following statements are equivalent.*

$$(X \perp\!\!\!\perp Y|\mathbf{Z},\mathbf{F})_{\mathcal{M}} \tag{17}$$
$$(X \perp\!\!\!\perp Y|\mathbf{Z},\mathbf{F} \cup \{F_i\})_{\mathcal{M}} \wedge \left[(F_i \perp\!\!\!\perp Y|\mathbf{Z},\mathbf{F},\{X\})_{\mathcal{M}} \vee (F_i \perp\!\!\!\perp X|\mathbf{Z},\mathbf{F},\{Y\})_{\mathcal{M}}\right] \tag{18}$$

Proof. We prove (17) $\rightarrow$ (18) by contrapositive. If $(X \not\perp\!\!\!\perp Y|\mathbf{Z},\mathbf{F}\cup\{F_i\})_{\mathcal{M}}$, then the statement $(X \not\perp\!\!\!\perp Y|\mathbf{Z},\mathbf{F})_{\mathcal{M}}$ follows since F_i is a root node by construction so not conditioning on it does not block any active paths. Otherwise, we have that $(F_i \not\perp\!\!\!\perp Y|\mathbf{Z},\mathbf{F},\{X\})_{\mathcal{M}}$ and $(F_i \not\perp\!\!\!\perp X|\mathbf{Z},\mathbf{F},\{Y\})_{\mathcal{M}}$. Accordingly, let p_Y denote an active path between F_i and Y and let A be the first node along the path starting from Y that is active due to conditioning on X. If A is defined, then there exist an active path between X and Y conditioned on $\mathbf{Z},\mathbf{F}$ and $(X \not\perp\!\!\!\perp Y|\mathbf{Z},\mathbf{F})_{\mathcal{M}}$. If A is not defined, then p_Y is active conditioned on $\mathbf{F},\mathbf{Z}$ only and $(F_i \not\perp\!\!\!\perp Y|\mathbf{Z},\mathbf{F})_{\mathcal{M}}$. A similar argument applies for an active path p_X between F_i and X with A defined relative to Y. If A is not defined for both p_Y and p_X, then we have $(F_i \not\perp\!\!\!\perp Y|\mathbf{Z},\mathbf{F})_{\mathcal{M}}$ and $(F_i \not\perp\!\!\!\perp X|\mathbf{Z},\mathbf{F})_{\mathcal{M}}$. It follows that $(X \not\perp\!\!\!\perp Y|\mathbf{Z},\mathbf{F})_{\mathcal{M}}$ since F_i is a root node by construction and thus a non-collider along any path.

Next, we prove (18) $\rightarrow$ (17) by contrapositive. We have $(X \not\perp\!\!\!\perp Y|\mathbf{Z},\mathbf{F})_{\mathcal{M}}$. If $(X \not\perp\!\!\!\perp Y|\mathbf{Z},\mathbf{F}\cup\{F_i\})_{\mathcal{M}}$ then the negation of (18) holds and we are done. Otherwise, we have $(X \not\perp\!\!\!\perp Y|\mathbf{Z},\mathbf{F})_{\mathcal{M}}$ and assume $(X \perp\!\!\!\perp Y|\mathbf{Z},\mathbf{F}\cup\{F_i\})_{\mathcal{M}}$. Thus, every active path between X and Y conditioning on $\mathbf{Z},\mathbf{F}$ goes through F_i. It follows that $(F_i \not\perp\!\!\!\perp Y|\mathbf{Z},\mathbf{F})_{\mathcal{M}}$ and $(F_i \not\perp\!\!\!\perp X|\mathbf{Z},\mathbf{F})_{\mathcal{M}}$. If $(F_i \perp\!\!\!\perp Y|\mathbf{Z},\mathbf{F},\{X\})_{\mathcal{M}}$, then X is a non-collider along every active path between F_i and Y conditioned on $\mathbf{Z},\mathbf{F}$. This implies that there is an active path between X and Y conditioned on $\mathbf{Z},\mathbf{F}\cup\{F_i\}$ and $(X \not\perp\!\!\!\perp Y|\mathbf{Z},\mathbf{F}\cup\{F_i\})_{\mathcal{M}}$ which contradicts our assumption. Thus, $(F_i \not\perp\!\!\!\perp Y|\mathbf{Z},\mathbf{F},\{X\})_{\mathcal{M}}$. A similar argument applies for F_i and X, thus $(F_i \not\perp\!\!\!\perp X|\mathbf{Z},\mathbf{F},\{Y\})_{\mathcal{M}}$. This concludes the proof. □

Corollary 2. *Let $\mathcal{M}_1$ and $\mathcal{M}_2$ denote two augmented MAGs and suppose $(X \perp\!\!\!\perp Y|\mathbf{Z},\mathbf{F})_{\mathcal{M}_1}$ and $(X \not\perp\!\!\!\perp Y|\mathbf{Z},\mathbf{F})_{\mathcal{M}_2}$ where $X,Y \in \mathbf{V}\cup\mathcal{F}$, $\mathbf{Z}\subseteq\mathbf{V}\setminus\{X,Y\}$, and $\mathbf{F}\subset\mathcal{F}\setminus\{X,Y\}$. Then, there exists a differing separation statement between $\mathcal{M}_1,\mathcal{M}_2$ in the form of $\mathcal{O}$.*

Proof. Let F_i denote an arbitrary node in $\mathcal{F}\setminus(\{X,Y\}\cup\mathbf{F})$ and consider the equivalence in Lemma 5. We have $(X \perp\!\!\!\perp Y|\mathbf{Z},\mathbf{F})_{\mathcal{M}_1} \equiv (X \perp\!\!\!\perp Y|\mathbf{Z},\mathbf{F}\cup\{F_i\})_{\mathcal{M}_1} \wedge \left[(F_i \perp\!\!\!\perp Y|\mathbf{Z},\mathbf{F},\{X\})_{\mathcal{M}_1} \vee (F_i \perp\!\!\!\perp X|\mathbf{Z},\mathbf{F},\{Y\})_{\mathcal{M}_1}\right]$ and $(X \not\perp\!\!\!\perp Y|\mathbf{Z},\mathbf{F})_{\mathcal{M}_2} \equiv (X \not\perp\!\!\!\perp Y|\mathbf{Z},\mathbf{F}\cup\{F_i\})_{\mathcal{M}_2} \vee \left[(F_i \not\perp\!\!\!\perp Y|\mathbf{Z},\mathbf{F},\{X\})_{\mathcal{M}_2} \wedge (F_i \not\perp\!\!\!\perp X|\mathbf{Z},\mathbf{F},\{Y\})_{\mathcal{M}_2}\right]$. If $(X \not\perp\!\!\!\perp Y|\mathbf{Z},\mathbf{F}\cup\{F_i\})_{\mathcal{M}_2}$, then $(X \perp\!\!\!\perp Y|\mathbf{Z},\mathbf{F}\cup\{F_i\})$ is a differing separation statement between $\mathcal{M}_1,\mathcal{M}_2$ that is 'closer' to $\mathcal{O}$ with F_i included. Alternatively, either $(F_i \perp\!\!\!\perp Y|\mathbf{Z},\mathbf{F},\{X\})$ or $(F_i \perp\!\!\!\perp X|\mathbf{Z},\mathbf{F},\{Y\})$ is a differing separation statement 'closer' to $\mathcal{O}$. By recursively applying this equivalence step to a differing separation statement, we finally obtain a differing statement in $\mathcal{O}$. □

Proof of Step 2: We only need to focus on the separation statements in $\mathcal{O}$ that are not in $\mathcal{T}$. Those are precisely the separation statements involving two F-nodes given a subset of the nodes corresponding to the observed variables $\mathcal{V}$, and all the other F-nodes. This is handled by the following lemma.[6]

Lemma 6. *Let $\mathcal{M}_1$ and $\mathcal{M}_2$ be two augmented MAGs where $(F_i \perp\!\!\!\perp F_j \big|\mathbf{W},\mathcal{F}\setminus\{F_i,F_j\})_{\mathcal{M}_1}$ and $(F_i \not\perp\!\!\!\perp F_j \big|\mathbf{W},\mathcal{F}\setminus\{F_i,F_j\})_{\mathcal{M}_2}$. Then, there exists a differing testable separation statement, i.e., a statement of the form in $\mathcal{T}$.*

We can finally prove Lemma 4. Suppose we have $(A \perp\!\!\!\perp B|\mathbf{C})_{\mathcal{M}_1}$ and $(A \not\perp\!\!\!\perp B|\mathbf{C})_{\mathcal{M}_2}$. Recall that any m-separation statement belongs to one of the sets $\mathcal{O}$, $\mathcal{U}$, or $\mathcal{T}$.

(a) If $(A \perp\!\!\!\perp B|\mathbf{C})$ belongs to $\mathcal{T}$, we are done.

(b) If $(A \perp\!\!\!\perp B|\mathbf{C})$ belongs to $\mathcal{O}\setminus\mathcal{T}$, then by Lemma 6, there exists an m-separation statement in $\mathcal{T}$ with different truth values for $\mathcal{M}_1,\mathcal{M}_2$. Then, the result follows from *(a)*.

(c) If $(A \perp\!\!\!\perp B|\mathbf{C})$ belongs to $\mathcal{U}$, then by Corollary 2, there exists an m-separation statement in $\mathcal{O}$ that evaluates differently for $\mathcal{M}_1,\mathcal{M}_2$. Then, the result follows from *(a)* and *(b)*. □

We showed that if $\mathcal{M}_1$ and $\mathcal{M}_2$ do not satisfy the three graphical conditions in the theorem, then there exists a separation statement consistent with the form in $\mathcal{T}$ that holds in one graph but not the other. Suppose the testable statement holds in $\mathcal{M}_1$ but not in $\mathcal{M}_2$. Then, there is a testable separation statement that holds in $Aug(\mathcal{D}_1)$ but not in $Aug(\mathcal{D}_2)$. By Prop. 1, the differing separation statement is equivalent to a graphical constraint in the definition of the $\mathcal{I}$-Markov property that evaluates differently with respect to the two causal graphs $\mathcal{D}_1,\mathcal{D}_2$. For the last step of the **(only if)** proof, we utilize the following theorem by Meek [14] to construct a tuple of distributions $\mathbf{P}$ such that $\mathbf{P}\in\mathcal{P}_{\mathcal{I}}(\mathcal{D}_2,\mathbf{V})$ and $\mathbf{P}\notin\mathcal{P}_{\mathcal{I}}(\mathcal{D}_1,\mathbf{V})$.

[6] The proof is provided later.

Theorem 5 in [14]. For all directed acyclic graphs $\mathcal{G}$, there exists a multinomial distribution P which is faithful to $\mathcal{G}$.

First, we consider the case where the differing statement is such that $(X \perp\!\!\!\perp Y|\mathbf{Z},\mathcal{F})_{Aug(\mathcal{D}_1)}$ and $(X \not\perp\!\!\!\perp Y|\mathbf{Z},\mathcal{F})_{Aug(\mathcal{D}_2)}$ where $X, Y \in \mathbf{V}, \mathbf{Z} \subset \mathbf{V}$. It follows that $(X \perp\!\!\!\perp Y|\mathbf{Z})_{\mathcal{D}_1}$ and $(X \not\perp\!\!\!\perp Y|\mathbf{Z})_{\mathcal{D}_2}$. By Meek's theorem, there exists a distribution P^* that is faithful to $\mathcal{D}_2$ over $\mathbf{V} \cup \mathbf{L}$. We construct the tuple of distributions $\mathbf{P}$ by replicating P^* and marginalizing out the latent variables $\mathbf{L}$. It is easy to see that $\mathbf{P} \notin \mathcal{P}_{\mathcal{I}}(\mathcal{D}_1, \mathbf{V})$ since for every distribution $P \in \mathbf{P}$, we have $(X \not\perp\!\!\!\perp Y|\mathbf{Z})_P$ while $(X \perp\!\!\!\perp Y|\mathbf{Z})_{\mathcal{D}_1}$. On the other hand, $\mathbf{P} \in \mathcal{P}_{\mathcal{I}}(\mathcal{D}_2, \mathbf{V})$ for the following two reasons. First, every $P \in \mathbf{P}$ is faithful to $\mathcal{D}_2$ which satisfies the first condition of the $\mathcal{I}$-Markov definition (Def. 1) with respect to $\mathcal{D}_2$. Second, for any pair $P^i, P^j \in \mathbf{P}$, we have $P^i(\mathbf{Y}|\mathbf{W}) = P^j(\mathbf{Y}|\mathbf{W})$ since P^i, P^j are identical distributions which satisfies the second condition of the $\mathcal{I}$-Markov definition with respect to $\mathcal{D}_2$. This concludes the proof for the first case.

Next, we have the case where the differing statement is such that $(F_i \perp\!\!\!\perp Y|\mathbf{W},\mathcal{F} \setminus \{F_i\})_{Aug(\mathcal{D}_1)}$ and $(F_i \not\perp\!\!\!\perp Y|\mathbf{W},\mathcal{F} \setminus \{F_i\})_{Aug(\mathcal{D}_2)}$ where $Y \in \mathbf{V}, \mathbf{W} \subset \mathbf{V}, F_i \in \mathcal{F}$. Let p be an active path in $Aug(\mathcal{D}_2)$ such that $(F_i \not\perp\!\!\!\perp Y|\mathbf{W},\mathcal{F} \setminus \{F_i\})_{Aug(\mathcal{D}_2)}$, and let X denote the first node along p after F_i. We construct $\mathcal{D}^*$, a supergraph of $\mathcal{D}_2$, by adding one node, denoted F^*, and such that F^* is a parent of X. By Meek's theorem, there exists a distribution P^* that is faithful to $\mathcal{D}^*$ over $\mathbf{V} \cup \mathbf{L} \cup \{F^*\}$. We construct the tuple of distributions $\mathbf{P}$ as follows:

1. Marginalize out $\mathbf{L}$ over P^*.
2. Consider $\mathbf{I}, \mathbf{J} \in \mathcal{I}$ such that $\mathbf{I} \triangle \mathbf{J} = \mathbf{S}_i$, F_i is the F-node adjacent to $\mathbf{S}_i$ in $Aug_{\mathcal{I}}(\mathcal{D}_2)$, and where $X \in \mathbf{S}_i$. Let $P_{\mathbf{I}} := P^*(\mathbf{V}|F^* = 0)$ and $P_{\mathbf{J}} := P^*(\mathbf{V}|F^* = 1)$.
3. For all the other $\mathbf{K} \in \mathcal{I}$, Let $P_{\mathbf{K}} := P_{\mathbf{I}}$ if $X \notin \mathbf{I} \triangle \mathbf{K}$, else $P_{\mathbf{K}} := P_{\mathbf{J}}$.

Note that $\mathbf{P} \notin \mathcal{P}_{\mathcal{I}}(\mathcal{D}_1, \mathbf{V})$ for the following reason. Recall that $(F_i \perp\!\!\!\perp Y|\mathbf{W},\mathcal{F} \setminus \{F_i\})_{Aug(\mathcal{D}_1)}$ and $(F_i \not\perp\!\!\!\perp Y|\mathbf{W},\mathcal{F} \setminus \{F_i\})_{Aug(\mathcal{D}_2)}$. It follows by the construction of $\mathcal{D}^*$ from $\mathcal{D}_2$ that $(F^* \not\perp\!\!\!\perp Y|\mathbf{W})_{\mathcal{D}^*}$. Hence, $P_{\mathbf{I}}(Y|\mathbf{W}) = P^*(Y|\mathbf{W}, F^* = 0) \neq P^*(Y|\mathbf{W}, F^* = 1) = P_{\mathbf{J}}(Y|\mathbf{W})$, and $\mathbf{P}$ violates the $\mathcal{I}$-Markov property with respect to $\mathcal{D}_1$. Finally, we show that $\mathbf{P} \in \mathcal{P}_{\mathcal{I}}(\mathcal{D}_2, \mathbf{V})$ which concludes the proof. First, $\mathbf{P}$ satisfies the first condition of the $\mathcal{I}$-Markov property relative to $\mathcal{D}_2$ since P^* is faithful to $\mathcal{D}^*$. Next, consider any $\mathbf{S}, \mathbf{T} \in \mathcal{I}$. Recall that $P_{\mathbf{S}}, P_{\mathbf{T}}$ are either set to $P_{\mathbf{I}}$ or $P_{\mathbf{J}}$ depending on the conditions in steps two and three of the construction above. If $P_{\mathbf{S}} = P_{\mathbf{T}}$, then we are done since any invariance will be satisfied irrespective of the set difference $\mathbf{S} \triangle \mathbf{T}$. Suppose otherwise and without loss of generality that $P_{\mathbf{S}} = P_{\mathbf{I}}$ and $P_{\mathbf{T}} = P_{\mathbf{J}}$. This means that $X \notin \mathbf{S} \triangle \mathbf{I}$ and $X \in \mathbf{T} \triangle \mathbf{J}$. Hence, it can be shown that $X \in \mathbf{S} \triangle \mathbf{T}$ because $X \in \mathbf{I} \triangle \mathbf{J}$. Let F' denote the F-node in $Aug_{\mathcal{I}}(\mathcal{D}_2)$ that is adjacent to $\mathbf{S} \triangle \mathbf{T}$. Note here that the graph $\mathcal{D}^*$ is a subgraph of $Aug_{\mathcal{I}}(\mathcal{D}_2)$. Hence, if $(F' \perp\!\!\!\perp Y|\mathbf{W},\mathcal{F} \setminus \{F_i\})_{Aug_{\mathcal{I}}(\mathcal{D}_2)}$, then $(F^* \perp\!\!\!\perp Y|\mathbf{W})_{\mathcal{D}^*}$ and $P_{\mathbf{I}}(Y|\mathbf{W}) = P_{\mathbf{J}}(Y|\mathbf{W})$. This concludes the proof.

□

Lemma 6. *Let $\mathcal{M}_1$ and $\mathcal{M}_2$ be two augmented MAGs where $(F_i \perp\!\!\!\perp F_j \,\big|\, \mathbf{W},\mathcal{F} \setminus \{F_i, F_j\})_{\mathcal{M}_1}$ and $(F_i \not\perp\!\!\!\perp F_j \,\big|\, \mathbf{W},\mathcal{F} \setminus \{F_i, F_j\})_{\mathcal{M}_2}$. Then, there exists a differing testable separation statement, i.e., a statement of the form in $\mathcal{T}$.*

Proof. First, if the two MAGs have different skeletons, then by the maximality property of MAGs, every pair of nodes that are not adjacent are separable. Consider any pair A, B that are adjacent in one graph but not the other. Without loss of generality, let $\mathcal{M}_2$ be the graph where A, B are adjacent. If $A, B \in \mathbf{V}$, then we have $(A \perp\!\!\!\perp B|\mathbf{Z},\mathcal{F})_{\mathcal{M}_1}$ and $(A \not\perp\!\!\!\perp B|\mathbf{Z},\mathcal{F})_{\mathcal{M}_2}$ for some $\mathbf{Z} \subset \mathbf{V}$. Note here that the conditioning on $\mathcal{F}$, even if it is not necessary, does not affect the separation statement since all the edges incident on F-nodes are out of them and conditioning on them does not open any paths. Otherwise, either A or B is an F-node which we assume to be A. Note that we cannot have $A, B \in \mathcal{F}$ since, by construction, all pairs of F-nodes are separable by the empty set. Hence, we have $(A \perp\!\!\!\perp B|\mathbf{Z},\mathcal{F} \setminus \{A\})_{\mathcal{M}_1}$ and $(A \not\perp\!\!\!\perp B|\mathbf{Z},\mathcal{F} \setminus \{A\})_{\mathcal{M}_2}$ for some $\mathbf{Z} \subset \mathbf{V}$. Similar to the argument earlier, conditioning on the remaining F-nodes in $\mathcal{F}$ does not affect the separation statement. Both cases provide differing statements of the form in $\mathcal{T}$ which concludes the proof.

Alternatively, $\mathcal{M}_1$ and $\mathcal{M}_2$ share the same skeleton. Let p_2 be a shortest active path between F_i and F_j given $\mathbf{W},\mathcal{F} \setminus \{F_i, F_j\}$ in $\mathcal{M}_2$, and let p_1 be the corresponding path in $\mathcal{M}_1$. Now p_1 is blocked since

$(F_i \perp\!\!\!\perp F_j \,\big|\mathbf{W}, \mathcal{F} \setminus \{F_i, F_j\})_{\mathcal{M}_1}$. Note here that p_1 is blocked as opposed to p_2 because either (1) one of the non-endpoint nodes along the path has a different status across p_1 and p_2, i.e., it is a collider in one while it is a non-collider in the other, or (2) a collider on p_2 has a descendant in $\mathbf{W}$ while the corresponding collider on p_1 does not have a descendant in $\mathbf{W}$. Also, none of the non-endpoint nodes along p_2 can be an F-node since p_2 is active and we are conditioning on all the remaining F-nodes. Similar to [25, Lemma 5.1.7], we handle this proof in two steps. But first, we establish the following result which proves useful throughout the proof.

Lemma 0: Let $\mathcal{M}_1$ and $\mathcal{M}_2$ be two augmented MAGs with the assumption that there is no testable separation statement that evaluates differently between them. If $\langle U, \ldots, X, Y, Z\rangle$ is a discriminating path for Y in $\mathcal{M}_2$, and the corresponding subpath between U and Y in $\mathcal{M}_1$ is also a collider path, then the path is also a discriminating path for Y in $\mathcal{M}_1$.

Proof. First note that only U, Y can be F-nodes since all the other nodes have arrowheads incident on them by definition of a discriminating path. Also, both $\mathcal{M}_1$ and $\mathcal{M}_2$ have the same skeleton, otherwise there exists a differing testable separation statement between them exists as shown earlier. Hence, U and Z are not adjacent in $\mathcal{M}_1$ as well. We show by induction that every collider between U and Y in $\mathcal{M}_1$ is a parent of Z. In the base case, V, the node after U, is a non-collider along $\langle U, V, Z\rangle$ in $\mathcal{M}_2$ and U and Z are not adjacent. If $\langle U, V, Z\rangle$ is a collider in $\mathcal{M}_1$, then we have a differing testable separation statement of the form $(U \perp\!\!\!\perp Z | \mathbf{T}, \mathcal{F})_{\mathcal{M}_2}$ and $(U \not\perp\!\!\!\perp Z | \mathbf{T}, \mathcal{F})_{\mathcal{M}_1}$ where $V \in \mathbf{T} \subset \mathbf{V}$; a contradiction. Hence, $\langle U, V, Z\rangle$ must be a non-collider in $\mathcal{M}_1$, i.e., $V \to Z$ since we have $U{*\!\!\to} V$ in $\mathcal{M}_1$. Now, for the inductive step, consider an arbitrary node W between U and Y. Suppose every node between U and W is a parent of Z in $\mathcal{M}_1$ with S denoting the node before W. Then, $\langle U, \ldots, S, W, Z\rangle$ is a discriminating path for W in $\mathcal{M}_1$ and $\mathcal{M}_2$. If $W \leftarrow\!\!* Z$, then we have a differing testable separation statement of the form $(U \perp\!\!\!\perp Z | \mathbf{T}, \mathcal{F})_{\mathcal{M}_2}$ and $(U \not\perp\!\!\!\perp Z | \mathbf{T}, \mathcal{F})_{\mathcal{M}_1}$ where $\mathbf{T} \subset \mathbf{V}$ contains all the nodes between U and W including W; a contradiction. Hence, we have $W \to Z$ in $\mathcal{M}_1$. This concludes the induction step. It follows that the subpath between U and Y is a collider path in $\mathcal{M}_1$ and every non-endpoint node along this path is a parent of Z. Thus, $\langle U, \ldots, X, Y, Z\rangle$ is a discriminating path for Y in $\mathcal{M}_1$. This concludes the proof. □

Lemma I: Let p_2 be a shortest active path between F_i and F_j given $\mathbf{W}, \mathcal{F} \setminus \{F_i, F_j\}$ in $\mathcal{M}_2$, and let p_1 be the corresponding path in $\mathcal{M}_1$. If there exists a non-endpoint node along p_2 that has a different status along p_1, then there exists a testable separation statement that is different between $\mathcal{M}_1$ and $\mathcal{M}_2$.

Proof. Suppose for the sake of contradiction that there are no testable separation statements that are different between $\mathcal{M}_1$ and $\mathcal{M}_2$. Let $\langle F_i = V_0, \ldots, V_n = F_j\rangle$ be the sequence of nodes along p_2 and p_1. We first show that the following property holds.

> **claim:** For every $1 \le k \le n-1$, if V_k has different status on p_2 versus p_1, then V_{k+1} is a parent of V_{k-1} in $\mathcal{M}_2$ and V_{k+1} is a collider on p_2 in $\mathcal{M}_2$.

We prove the claim by induction. In the base case, let V_K be the node along p_2 closest to F_i that has a different status compared to p_1. If V_{K-1} and V_{K+1} are not adjacent on both p_1 and p_2, then we have a differing testable statement of the form $(V_{K-1} \perp\!\!\!\perp V_{K+1} | \mathbf{T}, \mathcal{F})$, $V_K \in \mathbf{T}$, where the statement only holds when V_K is a non-collider. Therefore, V_{K-1} and V_{K+1} are adjacent in both $\mathcal{M}_1$ and $\mathcal{M}_2$ since they share the same skeleton. Since p_2 is the shortest active path in $\mathcal{M}_2$ between F_i, F_j, then the concatenated path $p_2' = \langle F_i, \ldots, V_{K-1}\rangle \oplus \langle V_{K-1}, V_{K+1}\rangle \oplus \langle V_{K+1}, \ldots, F_j\rangle$ is not active and it must be blocked at either V_{K-1} or V_{K+1}. Below, we rule out three of the four possibilities for the concatenated path to be blocked in $\mathcal{M}_2$.

Case 1: V_{K-1} is a non-collider along p_2 but a collider along p_2' and neither V_{K-1} nor any of its descendants are in $\mathbf{W}$. This means that we have $V_{K-1} \leftarrow\!\!* V_{K+1}$ and $V_{K-1} \to V_K$. Hence, the edge between V_K and V_{K+1} is into V_K, otherwise we violate the ancestral property in $\mathcal{M}_2$. Now, V_K is a collider along p_2 so a descendant of it must be in $\mathbf{W}$. But, this activates the collider at V_{K-1} along p_2' since V_K is a child of V_{K-1}; a contradiction. Hence, case 1 is not possible.

Case 2: V_{K+1} is a non-collider along p_2 but a collider along p_2' and neither V_{K+1} nor any of its descendants are in $\mathbf{W}$. This case is symmetric to case 1 and hence it is not possible.

Case 3: V_{K-1} is a collider along p_2 but a non-collider along p_2' and $V_{K-1} \in \mathbf{W}$. Note this directly implies that $V_{K-1} \neq F_i$. Then, we have that $V_{K-1} \to V_{K+1}$ and $V_{K-1} \leftarrow\!\!* V_K$. Hence, the edge between

V_K and V_{K+1} is into V_{K+1}, otherwise the ancestral property of $\mathcal{M}_2$ is violated. Next, we prove by induction that for every $0 < l < K$, V_l is a collider and a parent of V_{K+1}. In the base case, the claim holds for V_{K-1} by assumption of case 3. In the inductive step, suppose the claim holds for V_{L+1} and we prove it for V_L. If V_L is not adjacent to V_{K+1}, then $\langle V_L, \dots, V_{K-1}, V_K, V_{K+1}\rangle$ is a discriminating path for V_K in $\mathcal{M}_2$. Since V_K is the first node starting from F_i that has a different status along p_1 compared to p_2, then the nodes $V_{L+1}, ..., V_{K-1}$ are also colliders along p_1 in $\mathcal{M}_1$. Therefore, by Lemma 0, $\langle V_L, \dots, V_{K-1}, V_K, V_{K+1}\rangle$ is a discriminating path for V_K in $\mathcal{M}_1$ as well. This path does not involve any F-nodes as stated earlier except if $V_L = V_0 = F_i$. Hence, this discriminating path implies a differing testable statement of the form $(V_L \perp\!\!\!\perp V_{K+1} | \mathbf{T}, \mathcal{F})$, where all the nodes between V_L and V_{K+1} are conditioned on in $\mathbf{T}$. The statement holds where V_K is a non-collider and it does not hold when V_K is a collider, thus we have a contradiction. Thus, V_L and V_{K+1} have to be adjacent. Also, the edge is into V_{K+1}, otherwise we have $V_{L+1} \rightarrow V_{K+1} \rightarrow V_L$ and $V_L {*\!\!\rightarrow} V_{L+1}$ which violates the ancestral property of $\mathcal{M}_2$ in both orientations $V_L \rightarrow V_{L+1}$ and $V_L \leftrightarrow V_{L+1}$. Now, consider the concatenated path $p_2'' = \langle F_i, \dots, V_L\rangle \oplus \langle V_L, V_{K+1}\rangle \oplus \langle V_{K+1}, \dots, F_j\rangle$. This path must be blocked as it is shorter than p_2 and it can only be blocked at either V_L or V_{K+1}.Note that V_{K+1} has the same status along p_2 and p_2'' due to $V_L {*\!\!\rightarrow} V_{K+1}$ and $V_K {*\!\!\rightarrow} V_{K+1}$. Hence, V_L must be blocked along p_2'', and thus has a different status along p_2'' compared to p_2. If V_L is a non-collider along p_2 while it is a collider along p_2'', then we have $V_L \rightarrow V_{L+1} \rightarrow V_{K+1}$ and $V_L \leftrightarrow V_{K+1}$ which violates the ancestral property of $\mathcal{M}_2$, hence it is not possible. Therefore, V_L is a collider along p_2 and a non-collider along p_2'' which concludes the induction step. Finally, $\langle F_i, \dots, V_{K-1}, V_K, V_{K+1}\rangle$ is a discriminating path for V_K in $\mathcal{M}_2$ and consequently in $\mathcal{M}_1$ by Lemma 0, which leads to a differing testable statement, and hence contradicts our assumption. Hence, F_i and V_{K+1} are adjacent. Similar to the inductive step, the edge between F_i and V_{K+1} is into V_{K+1} by the ancestral property. However, this creates a shorter connecting path composed of $F_i {*\!\!\rightarrow} V_{K+1} \oplus \langle V_{K+1}, \dots, F_j\rangle$ which contradicts the choice of p_2 as a shortest active path between F_i, F_j in $\mathcal{M}_2$. Therefore, case 3 is not possible.

This leaves us with only one case where V_{K+1} is a collider along p_2 and V_{K+1} is a parent of V_{K-1} which concludes the base case.

For the inductive step, suppose the claim holds for all $V_i, i \leq l$, such that V_i has different status along p_1 than p_2, and let V_{l+r} be the next node with a different status. We argue the following holds: (1) the edge between V_{l-1} and V_l in $\mathcal{M}_2$ is into V_{l-1}, and (2) for every $1 \leq m \leq r$, V_{l+m} is a collider along p_2 and it is a parent of V_{l-1}. Starting with (1), recall we have $V_{l+1} \rightarrow V_{l-1}$ and V_{l+1} is a collider along p_2 by the inductive step. It follows that the edge between V_{l-1} and V_l in $\mathcal{M}_2$ is into V_{l-1}, for otherwise we violate the ancestral property of $\mathcal{M}_2$.

The argument for (2) is by induction. In the base case, V_{l+1} satisfies the property by the inductive step for **claim**. In the inductive step, assume the property holds for V_{l+p-1} where $1 < p \leq r$. If V_{l+p} is not adjacent to V_{l-1}, then $\langle V_{l+p}, \dots, V_{l+1}, V_l, V_{l-1}\rangle$ is a discriminating path for V_l in $\mathcal{M}_2$, and consequently in $\mathcal{M}_1$ by Lemma 0 since all the nodes between V_l and V_r have the same status in both graphs. This establishes a differing testable separation statement which is not possible. Hence, V_{l+p} and V_{l-1} must be adjacent. This edge is into V_{l-1}, otherwise it violates the ancestral property in $\mathcal{M}_2$ due to $V_{l+p-1} \rightarrow V_{l-1} \rightarrow V_{l+p}$ and $V_{l+p-1} \leftarrow\!\!* V_{l+p}$. Consider the concatenated path $p_2' = \langle F_i, \dots, V_{l-1}\rangle \oplus \langle V_{l-1}, V_{l+p}\rangle \oplus \langle V_{l+p}, \dots, F_j\rangle$. Notice that V_{l-1} has the same status along p_2 and p_2' in $\mathcal{M}_2$. Hence, V_{l+p} has to be blocking along p_2', otherwise we would have a shorter active path which violates the choice of p_2 in $\mathcal{M}_2$. If V_{l+p} is a collider along p_2' and a non-collider along p_2, then we have $V_{l+p} \rightarrow V_{l+p-1} \rightarrow V_{l-1}$ and $V_{l+p} \leftrightarrow V_{l-1}$ which violates the ancestral property. Hence, V_{l+p} must be a collider along p_2 and a non-collider along p_2', and we get $V_{l+p} \rightarrow V_{l-1}$. This concludes the proof for point (2).

Now, we go back to the inductive step in the proof of **claim**. Similar to the argument in the base step, V_{l+r-1} and V_{l+r+1} must be adjacent since they have different status along p_2 and p_1, and the concatenated path $p_2' = \langle F_i, \dots, V_{l+r-1}\rangle \oplus \langle V_{l+r-1}, V_{l+r+1}\rangle \oplus \langle V_{l+r+1}, \dots, F_j\rangle$ must be blocked at V_{r+l-1} or V_{r+l+1} with four possible cases:

Cases 1 and 2: Since we already showed that V_{l+r-1} is a collider on p_2 in point (2), Case 1 cannot happen. The argument for Case 2 goes exactly as in the base step.

Case 3: Suppose V_{l+r-1} is a collider along p_2 and a non-collider along p_2' and $V_{l+r-1} \in \mathbf{W}$. Note that we have established previously that every node starting from V_{l+1} until V_{l+r} are colliders and parents of V_{l-1}. First, the edge between V_{l+r} and V_{l+r+1} is into the latter, else we get an almost directed cycle due to $V_{l+r-1} \rightarrow V_{l+r+1} \rightarrow V_{l+r}$ and $V_{l+r-1} \leftrightarrow V_{l+r}$ which violates the ancestral property.

Next, we further establish that every node between V_l and V_{l+r-1}, inclusive, is a collider and a parent of V_{l+r+1}. The argument follows similar to that of point (2). For each node V_{l+i}, where $i = r-2$ down to $i = 0$, we argue by induction that V_{l+i} is adjacent to V_{l+r+1}, otherwise we have a differing testable statement between $\mathcal{M}_1$ and $\mathcal{M}_2$. The differing statement is due to the discriminating path $\langle V_{l+i}, \ldots, V_{l+r-1}, V_{l+r}, V_{l+r+1}\rangle$ for V_{l+r} where V_{l+r} has a different status along p_1 and p_2. Also, V_{l+i} is a parent of V_{l+r+1}, otherwise we violate the ancestral property or we obtain a shorter active path between F_i, F_j than p_2 in $\mathcal{M}_2$. At this point, we have a path of bi-directed edges along p_2 between V_{l-1} and V_{l+r+1} and every non-endpoint node along the path is a parent of V_{l-1} or V_{l+r+1}. This creates an inducing path between V_{l-1} and V_{l+r+1}, hence the two nodes must be adjacent by the maximality property of MAGs. Also, the edge can only be a bi-directed edge, else we violate the ancestral property due to $V_l \rightarrow V_{l+r+1} \rightarrow V_{l-1}$ and $V_{l-1} \leftrightarrow V_l$, or $V_{l+r} \rightarrow V_{l-1} \rightarrow V_{l+r+1}$ and $V_{l+r} \leftrightarrow V_{l+r+1}$. Note now that the concatenated path $p_2' = \langle F_i, \ldots, V_{l-1}\rangle \oplus \langle V_{l-1}, V_{l+r+1}\rangle \oplus \langle V_{l+r+1}, \ldots, F_j\rangle$ is shorter than p_2 and it must be active since V_{l-1} and V_{l+r+1} have the same status along both p_2 and p_2'. Therefore, we reach a contradiction for case 3. **This concludes the proof for the claim**. Now that **claim** is proven, we consider another claim to establish Lemma I.

> **claim*:** For every $1 \leq k \leq n-1$, if V_k has different status on p_2 versus p_1, then V_{k-1} is a parent of V_{k+1} in $\mathcal{M}_2$ and V_{k-1} is a collider on p_2 in $\mathcal{M}_2$.

The proof for **claim*** follows exactly that of **claim** while starting the induction argument with the node closest to F_j, instead of F_i, that has a different status along p_2 than p_1. Now, obviously those two claims are contradicting as long as there exists a single node along p_2 in $\mathcal{M}_2$ that has a different status than p_1 in $\mathcal{M}_1$. Therefore, either all the nodes along p_2 have the same status along p_1 or our assumption at the beginning of the proof does not hold and there exist differing testable separation statements. This concludes the proof of Lemma I. □

Given any active path p between X and Y given $\mathbf{Z}$ in $\mathcal{M}$, for every collider Q on p, there is a directed path (possibly of length 0) from p to a member of $\mathbf{Z}$. Define the distance-from-$\mathbf{Z}$ of Q to be the length of a shortest directed path (possibly of length 0) from Q to $\mathbf{Z}$, and define the distance-from-$\mathbf{Z}$ of p to be the sum of the distances from $\mathbf{Z}$ of the colliders on p.

Lemma II: Let p_2 be a shortest active paths between F_i and F_j given $\mathbf{W}, \mathcal{F} \setminus \{F_i, F_j\}$ in $\mathcal{M}_2$ such that no equally short active path has a shorter distance-from-($\mathbf{W}, \mathcal{F} \setminus \{F_i, F_j\}$) than p_2 does, and let p_1 be the corresponding path in $\mathcal{M}_1$. Then, there exists at least one differing testable separation statement between $\mathcal{M}_1$ and $\mathcal{M}_2$.

Proof. If there exists at least one node along p_2 in $\mathcal{M}_2$ that has a different status along p_1 in $\mathcal{M}_1$, then the result follows by Lemma I. Otherwise, every non-endpoint node along p_1 and p_2 has the same status. Since p_1 is not active in $\mathcal{M}_1$, then there exists at least one collider Q along p_1 and p_2 which does not have any descendants in $\mathbf{W}$ in $\mathcal{M}_1$ while Q has a descendant in $\mathbf{W}$ in $\mathcal{M}_2$. Note that the nodes in $\mathcal{F}$ do not activate the collider Q since they are not descendants of any node by construction.

Let Q denote any such collider on p_1, p_2 and let p_2^* denote the shortest directed path from Q to $\mathbf{W}$ in $\mathcal{M}_2$, and let p_1^* be the corresponding path in $\mathcal{M}_1$. Note that none of the non-consecutive nodes along p_2^* are adjacent, otherwise the edge must be a directed edge away from Q by the ancestral property, and we have a shorter directed path than p_2^* which is not possible. This applies to $\mathcal{M}_1$ as well, else we have different skeletons for $\mathcal{M}_1, \mathcal{M}_2$ which leads to a differing testable statement as shown in the beginning of the proof. If the first edge along p_1^* is out of Q in $\mathcal{M}_1$, then there exists at least one unshielded collider ($A {*\!\!\rightarrow} B \leftarrow\!\!* C$) along p_1^* as the path cannot be directed. The same triple in $\mathcal{M}_2$ is part of the directed path ($A \rightarrow B \rightarrow C$). Therefore, we have a differing testable statement ($A \perp\!\!\!\perp C \,|\mathbf{Z}, \mathcal{F}$) where $B \in \mathbf{Z} \subset \mathbf{V}$.

Let S denote the node adjacent to A along p_1^*, p_2^*. Alternatively to the previous case, we have $Q \rightarrow S$ along p_2^* and $Q \leftarrow\!\!* S$ along p_1^*. Suppose S is not on p_2 for the moment and suppose for the sake of contradiction that there is no differing testable statement between $\mathcal{M}_1$ and $\mathcal{M}_2$. In what follows, we prove that there exists a node C along the subpath of p_2 from F_i to Q such that (1) C and S are adjacent in $\mathcal{M}_2$ and the edge is into S, and (2) the status of C along p_2 is the same as that of C on $\langle F_i \ldots, C\rangle \oplus \langle C, S\rangle$. We argue that if no node between F_i and Q satisfies those conditions, then F_i itself must satisfy those conditions.

Suppose no node between F_i and Q satisfies both conditions. We argue by induction that all the nodes between F_i and Q along p_2 are colliders and parents of S. In the base case, S and Q_l, the node closest to Q along p_2 starting from F_i, are adjacent else we have a differing testable statement due to unshielded $Q_l{*\!\!\rightarrow} Q \rightarrow S$ in $\mathcal{M}_2$ and $Q_l{*\!\!\rightarrow} Q \leftarrow\!\!* S$ in $\mathcal{M}_1$. Also, the edge is into S in $\mathcal{M}_2$, otherwise we violate the ancestral property. Hence, we established condition (1). If $Q_l = F_i$, then Q_l trivially satisfies condition (2). Otherwise, $Q_l \neq F_i$ and Q_l must fail condition (2) by the assumption that no node satisfies both conditions. If Q_l is a non-collider along p_2 while it is a collider along $\langle F_i \ldots, Q_l\rangle \oplus \langle Q_l, S\rangle$, then we have $Q_l \rightarrow Q \rightarrow S$ and $Q_l \leftrightarrow S$ which violates the ancestral property and is not possible. Hence, Q_l is a collider along p_2 and the edge between Q_l and S is out of Q_l.

In the inductive step, let Q_m be some node between F_i and Q such that every node between Q_m and Q is a collider along p_2 and a parent of S in $\mathcal{M}_2$. If Q_m is not adjacent to S, then $\langle Q_m, \ldots, Q_l, Q, S\rangle$ is a discriminating path for Q in $\mathcal{M}_2$. Since $\mathcal{M}_1$ and $\mathcal{M}_2$ have the same skeleton and every node along p_2 has the same status as that on p_1, it follows by Lemma 0 that $\langle Q_m, \ldots, Q_l, Q, S\rangle$ is a discriminating path for Q in $\mathcal{M}_1$ as well. Node Q has a different status along the discriminating path in $\mathcal{M}_1$ compared to $\mathcal{M}_2$, thus we have a differing testable separation statement which contradicts the assumption. Therefore, Q_m and S are adjacent in both graphs and the edge is into S otherwise we violate the ancestral property. This establishes condition (1). Again, Q_m must fail condition (2) by the assumption that no node satisfies both conditions. Hence, we conclude that Q_m is a collider along p_2 with $Q_m \rightarrow S$ by a similar argument as in the base case. This concludes the inductive step.

Now we have established that every node between F_i and Q in $\mathcal{M}_2$ is a collider and a parent of S. Hence, $\langle F_i, \ldots, Q_l, Q, S\rangle$ is a discriminating path for Q in $\mathcal{M}_2$. By Lemma 0, it follows that the same is a discriminating path for Q in $\mathcal{M}_1$ as well which creates a differing testable separation statement leading to a contradiction. Hence, F_i and S are adjacent and the edge is into S to satisfy the ancestral property. Thus, F_i satisfies condition (1) and trivially condition (2).

This establishes that either there exists a $C \neq F_i$ that satisfies the conditions $(1, 2)$ or F_i itself satisfies these conditions. By symmetry, it follows that there exists a node B on $\langle Q, Q_r, \ldots, F_j\rangle$ (distinct from Q and possibly F_j itself) that satisfies both conditions discussed earlier. Now, the concatenated path $p_2' = \langle F_i, \ldots, C\rangle \oplus \langle C, S, B\rangle \oplus \langle B, \ldots, F_j\rangle$ is active given $\mathbf{W}, \mathcal{F} \setminus \{F_i, F_j\}$, since C, B have the same status on p_2' as they do on p_2, and p_2' is either shorter than p_2 or as long but has a shorter distance-from-($\mathbf{W}, \mathcal{F} \setminus \{F_i, F_j\}$) than p_2. Either case is a contradiction with our assumption about p_2.

Finally, if S is on p_2, it either lies on the subpaths $\langle F_i, Q\rangle$ or $\langle Q, F_j\rangle$, excluding F_i, F_j, and Q. Without loss of generality, suppose it is on $\langle Q, F_j\rangle$. The same previous argument goes through to establish that there exists a node E (distinct from Q) on $\langle F_i, Q\rangle$ such that (1) there is an edge between E and S in $\mathcal{M}_2$ that is into S, and (2) the status of E on p_2 is the same as the status of E on $\langle F_i, \ldots, E\rangle \oplus E{*\!\!\rightarrow} S$. Let T be the node adjacent to S along p_2 and closer to F_j. If we have $S \leftarrow\!\!* T$, then the concatenated path $\langle F_i, \ldots, E\rangle \oplus E{*\!\!\rightarrow} S \oplus \langle S, \ldots, F_j\rangle$ is active since E has the same status along p_2, p_2' and S has a descendant in $\mathbf{W}$. Moreover, this path is shorter than p_2 as $\langle E, \ldots, S\rangle$, the subpath of the latter, is at least of length 2; contradiction. Alternatively, we have $S \rightarrow T$ and the status of S along p_2 and the path $\langle F_i, \ldots, E\rangle \oplus E{*\!\!\rightarrow} S \oplus \langle S, \ldots, F_j\rangle$ is the same, hence both are active. Again, the latter path is shorter than p_2 and we have a contradiction.

Therefore, our initial supposition does not hold and there is at least one differing testable statement between $\mathcal{M}_1$ and $\mathcal{M}_2$. □

It follows by Lemma II that if there is an active path between F_i and F_j given $\mathbf{W}, \mathcal{F} \setminus \{F_i, F_j\}$ in $\mathcal{M}_2$ but not in $\mathcal{M}_1$, then there exists at least one testable separation statement that is different between the two augmented MAGs. This concludes the proof. □

7.5 Proof of Theorem 3

The main idea of the algorithm is to infer the separating sets between pairs of nodes using the invariance tests. Using c-faithfulness assumption, it is easy to see that the invariances that are checked imply m-separation statements between the nodes of the augmented graph. However, the separating sets that are found always include all the F-nodes. There are few questions we need to address to prove soundness of the algorithm:

(1) Are all pairs of separable nodes in Aug(D) correctly identified by the algorithm?

(2) Does the choice of separating set affect the application of FCI rules.

(3) Are the orientation rules sound?

We first address *(1)*: Note that all pairs of F-nodes are separable with the empty set by construction of $Aug(D)$. This is captured in Line 8 of the algorithm by setting $SepSet(F_i, F_j) = \emptyset$ for all pairs of F-nodes. This assures that after Phase I, they become non-adjacent.

Next consider all pairs X, Y where at least one is not an F-node. Suppose two nodes are separable in $Aug(D)$. Then there is a set W that makes them separable. There is no restriction on W: It may or may not have some of the F-nodes. However, since F-nodes of $Aug(D)$ are always source nodes, adding the remaining F-nodes cannot open new paths. Therefore, the set $W \cup \mathcal{F}$ is also a separating set. Formally, we have the following lemma:

Lemma 7. *For any pair $X, Y \in V \cup \mathcal{F}$, if $(X \perp\!\!\!\perp Y | W)_{Aug(D)}$, then $(X \perp\!\!\!\perp Y | W \cup (\mathcal{F} - X \cup Y))_{Aug(D)}$.*

Proof. Proof follows from the fact that F-nodes are source nodes in $Aug(D)$ and the rules of m-separation. □

Therefore, any separable pair imply a testable separation statement by Theorem 1 and it will be identified by the algorithm. This addresses *(1)*. We next address *(2)*.

We make use of the following simple observation: Although there may be more than one separating set for a pair of variables in the graph, FCI algorithm is sound and complete irrespective of which separating set is chosen. From the phrasing of the algorithm and its soundness, this is obvious since which separating set should be used is not specified. Here, we verify this by checking how the rules that require use of separating sets are affected by our choice of separating set:

Orienting unshielded colliders: Suppose we consider an ordered triple $\langle X, Y, Z \rangle$ where X, Z are non-adjacent. An F-node can never be a collider. Then the only case where the application of the rule may be affected by which separating system is used is when $X, Z \in V, Y \in \mathcal{F}$. Since by construction of $SepSet$, $Y \in SepSet(X, Z)$ algorithm does not orient it as a collider, which is correct. No collider will be missed by the algorithm due to the choice of $SepSet$.

Discriminating paths: By definition of discriminating path [28] and construction of augmented graph, there cannot be discriminating paths between pairs of F-nodes. We can have discriminating paths between an F-node and an observed node as $\langle X, \ldots, W, U, Y \rangle$, where $X \in \mathcal{F}$ and $Y \in V$. First, no F-node can be between X and U since by definition of discriminating path, they should be colliders. If U is not an F-node, then the change in separating system, i.e., adding extra F-nodes does not affect how the rule is applied. Suppose U is an F-node. Then by construction of the separating set, it has to be in the separating set. Then the rule is applied to orient $U \to Y$, which is consistent with Rule 8 and the augmented graph construction.

Finally, we address the soundness of orientation rules to address *(3)*. The rules of FCI are sound as shown by [27]. This is applicable in our setting, as one can see the augmented graph as a CBN with latents, ignoring how F-nodes are constructed, since m-separation statements implied by this CBN, which are purely graph theoretic criteria, are identical to those implied by the augmented graph. Moreover, previous phases of our algorithm are shown to be sound and complete, which is required for the soundness of this step: Skeleton is correctly identified. Moreover, if there is an unshielded collider, previous phases will correctly identify it. This is necessary for the correctness of the orientation rules of FCI. Therefore, we only need to check the soundness of the additional rules Rule 8,9. Soundness of Rule 8 is trivial since in any augmented graph $Aug(D)$, F-nodes are source nodes.

Soundness of Rule 9: Consider a pair F_i, Y where $F_i \in \mathcal{F}, Y \in V$ that are adjacent and $Y \ni nS_i$. This means there is no separating set for F_i, Y in $Aug(D)$, although by construction, they are not adjacent. This can only happen if there is an inducing path between F_i and Y relative to the latent variables L. An inducing path relative to latents L is defined as follows [28]: A path l in $Aug(D)$ is an inducing path if *i*) every non-endpoint that is not in L is a collider and *ii*) every collider is an ancestor of either endpoints. Since F-nodes by construction do not have ancestors, every collider on the inducing path between F_i, Y must be an ancestor of Y. Therefore in $MAG(Aug(D))$, the observed node must be an ancestor of Y. If $|S_k| = 1$, then any inducing path must go through the node in S_i, since in $Aug(D)$, F_i is only adjacent to the node in S_i. Since this node is on an inducing path, it must be an ancestor of Y. Therefore $MAG(Aug(D))$ contains an edge from this node to Y. This concludes the proof. □