TECHNICAL REPORT
R-114
Aug, 2024

Characterizing and learning multi-domain causal structures from observational and experimental data

Adam Li ADAM.LI@COLUMBIA.EDU
Causal Artificial Intelligence Lab
Columbia University
New York, NY 10025, USA

Amin Jaber AMIN.JABER@SYNLICO.COM
Synlico Inc.

Elias Bareinboim EB@CS.COLUMBIA.EDU
Causal Artificial Intelligence Lab
Columbia University
New York, NY, 10025, USA

Abstract

A fundamental problem throughout the sciences is the learning of causal structure underlying a system, by combining passive observations and active experimentation. Commonly, one may also collect such data across multiple domains, such as gene sequencing from different labs, or neural recordings from different species under resting-state and stimulation. Although there exist methods for learning the equivalence class of causal diagrams from observational and experimental data, they are meant to operate only in a single domain. In this paper, we develop a fundamental approach to structure learning in non-Markovian systems (i.e. when latent confounders are not ruled out apriori) leveraging observational and interventional data collected from multiple domains. Specifically, we first show an equivalence between learning from observational data in multiple domains and learning from interventional data with unknown targets in a single domain. Still there are subtleties when considering observational and experimental data. Using causal invariances derived from do-calculus, we define a property called multi-domain (MD) Markov that connects interventional distributions from multiple-domains to graphical criteria on a selection diagram. Leveraging the MD-Markov property, we introduce a new constraint-based causal discovery algorithm, MD-FCI, that can learn from observational and interventional data from different domains. We prove that the algorithm is sound and subsumes existing constraint-based causal discovery algorithms.

Keywords: causal discovery, Markov equivalence, structure learning, graphical model, interventions, multiple environments, transportability

1 Introduction

Causal discovery is the process of learning cause-and-effect relationships between variables in a given system. This learning is sometimes the final goal of the data scientist or a necessary step towards a more refined qualitative causal understanding of the underlying system [1, 2]. The learning process typically leverages constraints found in the data to infer the corresponding causal diagram. In practice, it is common that the data constraints do not uniquely identify the true, generative diagram. Therefore, the target of analysis is often an equivalence class (EC) of causal diagrams that encodes the constraints found in the data; as implied by the underlying unknown causal system.

Equivalence Classes An EC encodes invariances in the form of graphical constraints, and thus is used to represent the collection of causal diagrams that imply the corresponding invariances. Formal characterizations of ECs are fundamental for understanding the output of a learning algorithm and how it relates to the true, underlying causal system that the scientist aims to explain.

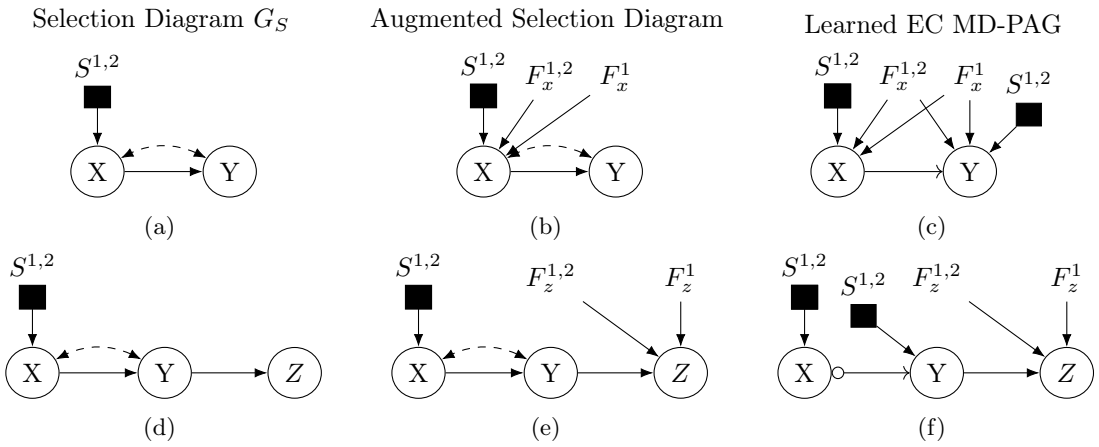


Figure 1: Example selection diagrams (a,d) and their respective augmented graphs (b,e). The top row corresponds to intervention set $\mathbf{\Psi}^{\Pi} = \langle\{\}^1, \{\}^2, \{X\}^1\rangle$ and the bottom row corresponds to intervention set $\mathbf{\Psi}^{\Pi'} = \langle\{\}^1, \{\}^2, \{Z\}^1\rangle$. The resulting EC learned from the MD-FCI algorithm (c,f) assuming the interventions are all known-target $\mathcal{K} = [1, 1, 1]$.

ECs are defined with respect to distributional invariances that are implied by the structure of the graph. For example, under the Markov property, conditional independences (CI) can be read off through the d-separation criterion from the causal graph [1]. CI relations can be discovered in principle from observational data. With interventional data, conditional invariances between pairs of distributions can be mapped to augmented d-separation statements over surgically modified causal graph [3].

An early example of an EC when only observational data is available in a single domain is the Markov equivalence class (MEC). The MEC characterizes the set of causal diagrams with the same conditional independences over the observed variables [2, 4–6]. Given interventional (i.e. experimental) data, naturally, one can reduce the size of the equivalence class [3, 7]. There are EC characterizations for the case of known intervention targets, which fall under the rubric of $\boldsymbol{\mathcal{I}}$-MEC [3, 8, 9] and in the case of unknown targets, it is called the Ψ-MEC [7].

Interventions vs Domains One important observation to our learning problem is that prior works have treated the concepts of interventions and domain-changes interchangeably [10–14]. However, various examples across scientific disciplines in the real world aim to leverage observational and interventional data across different domains (see last row of Table 1). Thus, we highlight their distinction and study its intricacies in depth in this paper. Consider Figure 1(d) for an illustration of this case when considering the extrapolation of experimental conclusions from bonobos to humans. Notably, humans and bonobos are known to have different kidney function, which is graphically represented by a special, (S) squared node pointing to X [15]. When applying a CRISPR intervention to a gene linked to kidney protein production (X), researchers investigate the impact of medication (Y) on fluid balance in the body (Z). This intervention is explicitly different from the kidney-function differences between bonobos and humans because the change-in-domain is there regardless of whether or not an intervention is made. This differentiation between interventions and domains holds significance, and will have implications for causal discovery and the characterization of the corresponding EC. By leveraging invariances across observational and interventional data from both bonobos and humans, one can learn additional causal relationships, as it will become clearer later on. One can also learn the commonalities and disparities between the different domains. In this case, this can lead to an improved qualitative understanding about humans and bonobos and how they relate. Moreover, conflating these qualitatively distinct settings of interventions and domains is

Domain	Obs.	Interv. $\mathcal{K}$	Interv. $\mathcal{U}$	Property	FCI-variant	Related Lit.
1	✓	x	x	Markov [20]	[2, 21–25]	[26, 27]
1	✓	✓	x	I-Markov [3, 28]	[3, 8, 28]	[26, 29]
1	✓	x	✓	Ψ-Markov [7]	[7, 13, 30]	[29, 31, 32]
k	✓	x	x	Ψ-Markov (Thm. 13)	[7] (Cor. 51)	[33–38]
k	✓	✓	✓	MD-Markov (Thm. 8)	MD-FCI (Thm. 18)	[26, 27, 29, 31, 33–43]

Table 1: Summary of Markov property results, and related algorithms that learn the ancestral graph based on number of domains and types of interventional (interv.) data provided such as observational (obs.), and known ($\mathcal{K}$) and unknown ($\mathcal{U}$) targets. The last column indicates a brief survey of different fields in ecology, economics, genomics, neurosciences, neurology and medicine that attempt to answer questions at each level. The rows highlighted in "red" are new concepts.

generally invalid, as pointed out in transportability analysis [16]. Pearl and Bareinboim introduced formal semantics for S-nodes (environments), providing a unified representation in the form they called a selection diagram [17–19].

Structure Learning We investigate structure learning when mixtures of observational and interventional data (known and unknown targets) across multiple domains are available. The multi-domain setting has been analyzed from the lens of selection diagrams, where selection nodes (or S-nodes) encode distributional changes in the mechanisms or exogenous variables due to a change in domain [17, 44, 45]. We will introduce in this paper a characterization of the EC for selection diagrams. Generalizing the structure learning setting to multiple domains requires a formal treatment and represents a common scenario found in the sciences [26, 27, 29, 32, 34–36, 38, 40–42, 46, 47]; see Table 1 for an example of different settings and related literature. For example, in single-cell sequencing analysis, scientists are interested in analyzing the causal effects of proteins on one another. However, they may typically collect observational and/or experimental data from multiple labs (domains) and wish to combine them into one dataset. Also, scientists may collect observational and experimental data over multiple species in order to learn more about one specific species, or the relationships among them [32, 41, 48].

The FCI algorithm and its variants learn a partial ancestral graph (PAG), an MEC of causal diagrams given purely observational data in a single domain [1, 2, 23]. The $\mathcal{I}$-FCI (with known targets) and Ψ-FCI (with unknown targets) generalize these results to interventional data, and further reduce the size of the EC to an $\mathcal{I}$-PAG and Ψ-PAG, respectively [3, 7]. However, these algorithms operate with data from a single domain, and do not account for combining known/unknown interventional targets across distinct domains.

Various approaches have been proposed throughout the literature for causal discovery from multiple domains. They impose various assumptions on the data generating process. For instance, the works in [10, 13, 49–53] assume Markovianity, a functional model (e.g. linearity) holds, and/or do not take into account arbitrary combinations of observational and interventional data with known and unknown targets. Alternatively, a method known as JCI pools data together and performs learning on the combined dataset [14]. However, it has been shown that the pooling of data is an incomplete procedure when considering interventional data within a single domain, which also leads to an incomplete characterization in the multi-domain setting (for details, see [7, Appendix D.2]).

Proposed Work In this paper, we take a principled approach to the multi-domain structure learning problem and formally characterize MD-PAGs, the object of learning that is expressive and capable of encoding constraints across both domains and interventional regimes. This paper extends our previous work in [54] and advances the characterization and learning algorithm in various directions. We graphically characterize general multi-domain invariances and provide a general

Algorithm	Graphical Characterization	UC	Interv. $\mathcal{K}$	Interv. $\mathcal{U}$	Nonparametric	General Multi Domain
Single-Distribution						
[6, PC],[1, p. IC]	✓	x	x	x	✓	x
[2, FCI], [1, IC^*]	✓	✓	x	x	✓	x
Multi-Distribution						
[7, I-FCI]	✓	✓	✓	x	✓	x
[3, Ψ-FCI]	✓	✓	x	✓	✓	x
[28, IGSP]	✓	x	✓	x	x	x
[13, 50, ICP]	x	x	✓/x	✓/x	✓	x
[14, JCI]	✓/x	✓	✓/x	✓/x	✓	x
[11, 12, 51, NSC]	x	x	✓/x	✓/x	✓	x
[52, MDLS]	x	x	✓/x	✓/x	x	x
This work	✓	✓	✓	✓	✓	✓

Table 2: **Comparison of proposed algorithms that may be used in a multi-domain setting** - Various algorithms exist in the literature concerning multi-domain, or interventional distributions. We summarize: i) whether a graphical characterization exists, ii) do they account for unobserved confounding (UC), iii) can the methods handle known-target ($\mathcal{K}$), or unknown-target ($\mathcal{U}$) interventions, iv) is the method nonparametric and v) does the method handle the MD in the general setting (i.e. not make the MDIX assumption)? The table is sectioned into methods that handle single-distribution, or multiple distributions.

structure learning algorithm compared to existing work (see Table 2 [1]). We provide a characterization and learning algorithm when observations are not present in all domains and when interventions may have the same mechanisms across domains. Specifically, we contribute the following:

1. **Multi-Domain Markov (MD-Markov) Property** - We introduce the MD-Markov property (Def. 3), which extends and generalizes the observational Markov, $\mathcal{I}$-Markov, and Ψ-Markov properties to the setting of multiple domains with arbitrary mixtures of observational and/or interventional data with known and unknown targets. We provide a complete graphical characterization (Thm. 8), that enables i) an efficient representation of the distributional invariances in the data and ii) the ability to translate these data-invariances to graphical constraints (e.g. d-separation).

2. **Learning algorithm** - We develop a sound algorithm, MD-FCI (Thm. 18), for learning an equivalence class of selection diagrams with observational and/or interventional data across different domains. Moreover, our algorithm is provably more informative than any of its FCI-variant predecessors. [2]

1.1 Preliminaries and Notation

In this section, we introduce the notation and background used throughout the manuscript. Uppercase letters (X) represent random variables, lowercase letters (x) signify assignments, and bold ones ($\mathbf{X}$) indicate sets. The CI relation $\mathbf{X}$ being independent of $\mathbf{Y}$ given $\mathbf{Z}$ is denoted as $\mathbf{X} \perp\!\!\!\perp \mathbf{Y}|\mathbf{Z}$. The d-separation (or m-separation) of $\mathbf{X}$ from $\mathbf{Y}$ given $\mathbf{Z}$ in graph G is expressed as $(\mathbf{X} \perp\!\!\!\perp \mathbf{Y}|\mathbf{Z})_G$. $G_{\overline{X}}$ depicts G with incoming edges to X removed, while $G_{\underline{X}}$ omits all edges outgoing from X. Superscripts and subscripts will be dropped where feasible to simplify notation.

1. There are many methods in single-domain methods [23, 55–58] and this list is incomprehensive.
2. Our algorithm is implemented in open-source MIT-Licensed https://github.com/py-why/dodiscover.

Structural Causal Models

We use Structural Causal Models (SCMs) [1] as our basic semantical framework. A SCM is a 4-tuple $\langle \mathbf{U}, \mathbf{V}, \mathbf{F}, P(\mathbf{u}) \rangle$, where 1) $\mathbf{U}$ is a set of exogenous (latent) variables, 2) $\mathbf{V}$ is a set of endogenous (observed) variables, 3) $\mathbf{F}$ is a set of functions that determine the values of endogenous variables (i.e. $v \leftarrow f_V(\mathbf{pa}_V, \mathbf{u}_V)$ is a function with $\mathbf{pa}_V \subseteq \mathbf{V} \setminus \{V\}$ and $\mathbf{U}_V \subseteq \mathbf{U}$ and 4) $P(\mathbf{u})$ is a joint distribution over exogenous variables, $\mathbf{U}$.

Each SCM induces a causal diagram, G, where every variable $V \in \mathbf{V}$ is a vertex and directed edges in G correspond to functional relationships specified by $\mathbf{F}$ and bidirected edges represent common exogenous variables between two vertices [59]. Within the structural semantics, an intervention by setting $X = x$ is represented with the do-operator, which encodes the operation of replacing the original functions of X (i.e. $f_X(\mathbf{pa}_X, \mathbf{u}_X)$) by the constant x and then induces a submodel M_x and corresponding interventional distribution $P(\mathbf{v}|do(x))$. When considering a collection of SCMs, another related graphical object known as the selection diagram is used.

Definition 1 (Selection Diagrams adapted from [19]) *Let $\mathcal{M} = \langle M^1, M^2, ..., M^N \rangle$ be a collection of SCMs relative to the N domains $\langle \Pi^1, , \Pi^2, ..., \Pi^N \rangle$, sharing a causal diagram G. $\mathcal{M}$ is said to induce a **selection diagram** G_S, if G_S is constructed as follows: every edge in G is also an edge in G_S; G_S contains an extra node $S^{i,j}$ and corresponding edge $S^{i,j}_k \rightarrow V_k$ whenever there exists a discrepancy $f^i_k \neq f^j_k$, or $P^i(U_k) \neq P^j(U_k)$ between M^i and M^j.* ■

Selection diagrams are causal graphs imbued with extra selection "S-nodes". Selection diagrams are induced from tuples of SCMs rather than a single one, since they represent different underlying SCMs. Here, S-nodes can be differentiated with a subscript indicating which variable one is pointing to. For example, $S^{i,j}_k \rightarrow V_k$ denotes another S-node for domain pair Π^i, Π^j. All S-nodes for a domain pair, Π^i, Π^j are part of the set $\mathbf{S}^{i,j}$. A single S-node only points to at most one variable [3].

Causal Bayesian Network (CBN)

Let $P(\mathbf{V})$ be a probability distribution over a set of variables $\mathbf{V}$, and $P_{\mathbf{x}}(\mathbf{V})$ denote the distribution resulting from the *hard intervention* $do(\mathbf{X} = \mathbf{x})$, which sets $\mathbf{X} \subseteq \mathbf{V}$ to constants $\mathbf{x}$. Let $\mathbf{P}^*$ denote the set of all interventional distributions $P_{\mathbf{x}}(\mathbf{V})$, for all $\mathbf{X} \subseteq \mathbf{V}$, including $P(\mathbf{V})$. A directed acyclic graph (DAG) over $\mathbf{V}$ is said to be a *causal Bayesian network* compatible with $\mathbf{P}^*$ if and only if, for all $\mathbf{X} \subseteq \mathbf{V}$, $P_{\mathbf{x}}(\mathbf{v}) = \prod_{\{i|V_i \notin \mathbf{X}\}} P(v_i|\mathbf{pa}_i)$, for all $\mathbf{v}$ consistent with $\mathbf{x}$, and where $\mathbf{Pa}_i$ is the set of parents of V_i [1]. Given that a subset of the variables are unmeasured or latent, $G = (\mathbf{V} \cup \mathbf{L}, \mathbf{E})$ is a mixed-edge graph representing the causal graph where $\mathbf{V}$ and $\mathbf{L}$ denote the measured and latent variables, respectively, and $\mathbf{E}$ denotes the edges. Following the convention in [1], for simplicity, a dashed bi-directed edge is used instead of the corresponding latent variables. CI relations can be read from the graph using a graphical criterion known as *d-separation*[4] [60].

Ancestral Graphs

A mixed graph can contain directed and bi-directed edges. A is a child of B if $B \rightarrow A$ and is denoted $A \in Ch(B)$. A is an ancestor of B if there is a directed path from A to B, denoted $A \in Anc(B)$. A is a spouse of B if $A \leftrightarrow B$ is present. If A is both a spouse and an ancestor of B, this creates an almost directed cycle. A mixed graph is ancestral if it does not contain directed or almost directed cycles. It is maximal if there is no inducing path (relative to the empty set) between any two non-adjacent nodes. 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 Maximal Ancestral Graph (MAG) is a graph that is both

3. In the EC representation of selection diagrams, S-nodes may point to more than one variable, which does not change the meaning, but is simply for convenience.
4. A generalization of d-separation is known as "m-separation", which we will use interchangably in this paper.

ancestral and maximal [61]. Given a causal graph $G = (\mathbf{V} \cup \mathbf{L}, \mathbf{E})$, a unique MAG M_G over $\mathbf{V}$ can be constructed such that both the independence and the ancestral relations among $\mathbf{V}$ are retained; see, [61]. Ancestral graphs, such as MAGs represent a Markov equivalence class (MEC) of a DAG, G. PAGs represent the unique MEC of a MAG. Therefore, a DAG maps to a unique PAG, representing the MEC of the conditional independence statements. However, a PAG represents a class of different DAGs that encode the same conditional independence statements.

Soft Interventions

Under this type of interventions, the original conditional distributions of the intervened variables $\mathbf{X}$ are replaced with new ones, without completely severing the effect of the corresponding parents. Accordingly, the interventional distribution $P_{\mathbf{X}}(\mathbf{v})$ for $\mathbf{X} \subseteq \mathbf{V}$ is such that $P_{\mathbf{X}}(X_i|Pa_i) \neq P(X_i|Pa_i)$, $\forall X_i \in \mathbf{X}$, and factorizes as follows:

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

In this work, we assume no selection bias and solely consider soft interventions. In the presence of multiple domains, a selection diagram captures commonalities and disparities between domains [17, 18, 59]. Represented as $G_S = (\mathbf{V} \cup \mathbf{L} \cup \mathbf{S}, \mathbf{E} \cup \mathbf{E_S})$, it extends a causal diagram by incorporating S-nodes and their edges. $\binom{N}{2}$ S-nodes, $S^{i,j}$, indicate distribution changes across pairs among N domains, by pointing to nodes in $\mathbf{V}$ whose mechanism, or the exogenous variables affecting $\mathbf{V}$ are altered between domains i and j. An example is shown in Figure 1(a), where the S-node is pointing to X, indicating that the distribution of X changes, or that of the distribution of latent variable of X is different across the two domains.

Similarly, "F-nodes" are auxiliary nodes used in [1, 3, 62–64] to represent invariances with respect to interventions within the same domain. We can use F-nodes, such as $F_X^{i,j}$ to capture the invariances between a pair of distributions from domain i and j, such that their symmetric difference is X (to be defined rigorously in Def. 2). For example, $F_X^{i,j}$ can be used to capture the invariances between $P_{X,Z}^i$ and P_Z^j. Similarly, F_X^i can be used to capture the invariances for a pair of distributions arising from a single domain i. For example, Figure 1(b; top row) shows $F_x^{1,2}$, representing invariances between the observational distribution, $\{\}^2$, in domain Π^2 against the interventional distribution, $\{X\}^1$ in domain Π^1, while Figure 1(b; bottom row) shows $F_z^{1,2}$, representing invariances between the observational distribution, $\{\}^2$, in domain Π^2 against the interventional distribution, $\{Z\}^1$ in domain Π^1. Unlike interventions, the change from domain Π^i to domain Π^j (i.e. also known as domain-shift) potentially alter latent variable distributions or functional relationships and persist irrespective of whether or not external intervention occurs. Distinguishing these concepts enables learning what is invariant across domains (rather than just across certain pairs of distributions), vital for transportability analysis [17].

Capturing the intricacies of multiple distributions, interventions, and domains results in a non-trivial notation. To gently introduce the notation that is required, the following section highlights common elements required to discuss the general multi-domain setting.

Multi-domain setup

The following objects are utilized throughout the paper, and are summarized here, building from the notation in [7, 54] and the transportability literature [19].

1. **Domains**: $\mathbf{\Pi} = \{\Pi^1, \Pi^2, ..., \Pi^N\}$ denotes a set of N domains, where N is finite.
   - For example, consider three domains $\mathbf{\Pi} = \{\Pi^1, \Pi^2, \Pi^3\}$ denoting humans, bonobos, and mice.
2. **Distributions**: $\mathbf{P}^{\Pi} = \langle P_1^1, P_2^1 ..., P_M^N \rangle$ is an ordered tuple of probability distributions. Denote $\mathbf{P}^i$ as the distributions associated with domain i. There is a one-to-one correspondence between $\mathbf{P}$

(we drop the superscript Π unless necessary) and $\mathbf{\Psi}$, such that P_j^i is the distribution associated with targets $\mathbf{\Psi}_j^i$ in domain Π^i.

- For example, consider the three domains: humans (Π^1), bonobos (Π^2), and mice (Π^3). One may passively observe as well as experiment within each domain, evoking distributions $\mathbf{P} = \langle P_1^1, P_1^2, P_2^2, P_1^3, P_2^3 \rangle$. Here, P_1^1, P_1^2, P_1^3 corresponds to observations in humans (Π^1), bonobos (Π^2), and mice (Π^3), respectively, and P_2^2, P_2^3 corresponds to experiments on bonobos (Π^2), and mice (Π^3).

3. **Interventional targets**: $\mathbf{\Psi}^\Pi = \langle \Psi_1^1, \Psi_2^1, ..., \Psi_M^N \rangle$ is an ordered tuple of sets of intervention targets, with different sets of intervention targets occurring within each of the N domains for a total of M intervention target sets. We will denote $\mathbf{\Psi}^i$ as the intervention targets associated with domain i.

   - Interventional targets are matched one-to-one with distributions. Considering the previous example, intervention targets $\mathbf{\Psi} = \langle \Psi_1^1, \Psi_1^2, \Psi_2^2, \Psi_1^3, \Psi_2^3 \rangle$ correspond to $\mathbf{P}$. $\Psi_1^1 = \{\}^1, \Psi_1^2 = \{\}^2, \Psi_1^3 = \{\}^3$, corresponding to observations in humans, bonobos and mice, respectively. Ψ_2^2, Ψ_2^3 here correspond to sets of intervention-targets; for example, they could be on certain genes. By explicitly intervening in these domains by performing a gene-knockout experiment, the corresponding distributions P_2^2, P_2^3 are obtained.

4. **Known target indices**: $\mathcal{K}$ is a vector of binary variables indicating which sets of interventions are known-targets, where 1 indicates the target is known and 0 that the target is unknown. $\mathcal{U} := 1 - \mathcal{K}$ represents therefore an index vector selecting the distributions and interventions with unknown targets. $\mathbf{P}_\mathcal{K}$ and $\mathbf{\Psi}_\mathcal{K}$ denotes the set of distributions and intervention targets corresponding to the known target interventions.

   - For example, we may consider the vector $\mathcal{K} = [1, 1, 0, 1, 1]$ corresponding to the distributions $\mathbf{P}$ and intervention targets $\mathbf{\Psi}$ in the previous examples. Here, all observational distributions are "known-target" by convention. The interventional distribution in bonobos P_2^2 induced by intervention-targets on certain genes Ψ_2^2 correspond to an unknown-target meaning one does not know that the specific gene Ψ_2^2 was explicitly intervened. The interventional distribution on mice P_2^3 induced by intervention-targets on the set of genes Ψ_2^3 correspond to a known-target. This means one knows that P_2^3 was generated due to an intervention on the set of genes Ψ_2^3. Whenever knowledge of this sort is unavailable, $\mathcal{K} = [0, 0, ..., 0]$, which means that all targets are unknown.

5. **Causal diagram**: $G = (\mathbf{V} \cup \mathbf{L}, \mathbf{E})$, is a shared diagram over the N domains.

   - For example, a scientist might posit the following causal diagram describing how certain gene expression (X), kidney function (Y), and drug-metabolism (Z) are related: $G = (X \to Y \leftarrow Z; X \leftrightarrow Y)$. This means gene expression may cause drug-metabolism mediated by the kidney. Moreover, there is possibly a latent confounder that affects both gene expression and kidney function. In practice, a possible confounder in this instance could be diet, or sleep.

6. **Selection diagram**: $G_S = (\mathbf{V} \cup \mathbf{L} \cup \mathbf{S}, \mathbf{E} \cup \mathbf{E_S})$, extends G with the corresponding S-nodes and their edges to represent each pair of domains. Let $\mathbf{V}_{S^{i,j}}$ denote the set of nodes that S-node $S^{i,j}$ points to and $\mathbf{V_S}$ as the set of children for all S-nodes of G_S.

   - A selection diagram relating gene expression (X), kidney function (Y), and drug-metabolism (Z), while also capturing differences across humans (Π^1) and bonobos (Π^2) could be posited by a scientist. For example, consider Figure 1(a), where now a S-node $S^{1,2} \to X$ indicates that the mechanism for gene expression is possibly different across the two species, while all other mechanisms are invariant across species. This is true regardless of whether the scientist performs a gene-knockout experiment, or passively sequence the genome.

$\mathbf{X}^i$ denotes the set of variables $\mathbf{X}$ in the ith domain Π^i, and $X_j \in \mathbf{X}$ indicates the jth variable within $\mathbf{X}$. When discussing intervention targets, $X_j^{i,(k)}$ refers to the jth variable with the kth mechanism change in domain Π^i. For instance, $X^{i,(k)}, X^{i,(l)}$ represent two interventions with distinct mechanisms (k and l) on variable X in domain Π^i. $\{\}^i \in \mathbf{\Psi}$ explicitly denotes the observational distribution (empty intervention) for domain Π^i and is by convention here a "known-target" (more discussion on this assumption later).

Let $\mathbf{S} = \{\mathbf{S}^{1,2}, \mathbf{S}^{1,3}, ..., \mathbf{S}^{N-1,N}\}$ represent S-nodes for distribution changes across all domain pairs. When $i = j$, $\mathbf{S}^{i,j} = \emptyset$, indicating there is no S-node. Each S-node is connected to exactly one variable when added, and thus $S_k^{i,j}$ indicates the S-node for domains Π^i, Π^j pointing to variable $V_k \in \mathbf{V}$. We will typically drop the superscript, and subscript when possible to simplify notation.

We further illustrate the interplay between selection diagrams and multi-domain distributions and interventions in the following example.

Example 1 (Selection Diagram and Multi-domain Distributions) *Consider Figure 1(a), which is a selection diagram G_S over two domains, $\mathbf{\Pi} = \{\Pi^1, \Pi^2\}$. The S-node points to X indicating that there is a possible difference in mechanism between domain 1 to 2. Let the tuple of distributions $\mathbf{P} = \langle P_1^1, P_2^1, P_1^2 \rangle$ be the result of the corresponding intervention sets $\mathbf{\Psi}^{\Pi} = \langle \{\}^1, \{X\}^1, \{\}^2 \rangle$ with known-target indices $\mathcal{K} = [1, 0, 1]$.*

The S-node, $S^{1,2} \to X$ implies a change in mechanism and corresponds to the lack of an invariance; i.e. $P_1^1(X) \neq P_1^2(X)$. ■

Given this introduction of the multi-domain notation, an example of the proposed output of the new algorithm (MD-FCI) is shown in Figure 1(c). This model class represents an EC of the ground-truth selection diagrams given data from the intervention sets $\mathbf{\Psi}^{\Pi} = \langle \{\}^1, \{\}^2, \{X\}^1 \rangle$ and $\mathbf{\Psi}^{\Pi'} = \langle \{\}^1, \{\}^2, \{Z\}^1 \rangle$ for the top and bottom row respectively. We observe from the top row of Figure 1 (a-c) that it is possible to orient all possible edges. In the bottom row of the Figure, full orientation is not achieved and the circular endpoint in the arrow indicates an uncertainty such that there is a selection diagram within the EC where the endpoint is either a tail or an arrowhead. This work will characterize what is learnable given the selection diagram, and tuple of distributions arising from arbitrary sets of interventions from multiple domains.

1.1.1 Organization

Table 3 provides a summary of the notation in terms of the distributions. Each domain $\mathbf{\Pi} = \{\Pi^1, \Pi^2, ..., \Pi^N\}$ may contain observational and interventional distributions. Based on our interpretation, the first row and column are considered exchangeable in the existing literature, under what we call the multi-domain intervention exchangeability assumption (MDIX) [3, 7, 10, 11, 14, 28, 50, 52, 53].

In this paper, we develop a general framework that includes a formal characterization of the EC and a learning algorithm for data arising from multiple domains. In Section 2, we introduce and define the MD-Markov property to establish a mapping from graphical conditions to distributional invariances. We prove a graphical characterization, and define the MD Markov EC. In Section 3, we discuss characterization and a learning algorithm in the simplified Markovian setting, where one assumes unobserved confounding is absent. In Section 4, we discuss characterization and a learning algorithm in the general non-Markovian setting, where there may be unobserved confounding. We demonstrate a duality between single-domain interventions and multi-domain observations in this setting, and introduce the MDIX assumption. In addition, we introduce in Alg. 2 the so-called MD-FCI causal discovery algorithm for learning an EC of selection diagrams. Finally, in Section 5, we elaborate on how the MD-Markov and MD-FCI algorithm improves upon previous work [11, 13, 14, 51].

Domain	Observational	Interventional			
Π^1	$P^1_{\{\}}(V)$	$P^1_{v_i}(V)$	$P^1_{v_j}(V)$	$P^1_{v_i,v_j}(V)$	$\dots$
Π^2	$P^2_{\{\}}(V)$	$P^2_{v_i}(V)$	$P^2_{v_k}(V)$	$P^2_{v_i,v_k}(V)$	$\dots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
Π^N	$P^N_{\{\}}(V)$	$P^N_{v_l}(V)$	$P^N_{v_m}(V)$	$P^N_{v_l,v_j}(V)$	$\dots$

Table 3: **Possible distributions observed in the Multi-domain setup** - Each domain $\mathbf{\Pi} = \{\Pi^1, \Pi^2, ..., \Pi^N\}$ may contain observational and interventional distributions. Based on our interpretation, the first row and column are studied in the existing literature under the lens of the multi-domain intervention exchangeability assumption (MDIX) [3, 7, 10, 11, 14, 28, 50, 52, 53]. Section 4.1 demonstrates an equivalence between the two settings of multi-domain observational and single-domain interventional. This paper discusses a more general setting when additional distributions are available in multiple-domains (i.e. including the cells highlighted in yellow).

2 Multi-domain Markov Equivalence Class

The first step to understand and design the behavior of a learning agent is to obtain a characterization of what can be learned from a given data collection. This section explores ECs in a multi-domain setting with arbitrary mixtures of observational and interventional data.

We will start with the following assumptions.

Assumption A1 (Shared causal structure) *Assume that each environment shares the same causal diagram. That is, the S-nodes do not change the underlying structure of the causal diagram.* ■

This means that, in our context, the S-nodes will not represent structural changes such as when V_i has a different parent set across domains. This is reflected in Def. 1 [5].

Assumption A2 (Soft interventions without altering the causal structure) *Assume that interventions do not alter the causal diagram. That is for each intervention set in the tuple of interventions $\mathbf{I} \in \mathbf{\Psi}$, the intervention does not remove, or add any edges to the graph.* ■

This assumption precludes any interventions that modify the graphical structure of the causal diagram. For instance, hard interventions cut all incoming edges for a specific node. As such, hard interventions are not considered in this work. In addition, more general interventions, that may add edges, or remove some edges are also not considered. Section 5.3 discusses the nuances of modeling hard versus soft interventions.

Assumption A3 (Distinct interventions across domains) *Assume that interventions over the same variable in various domains possess distinct mechanisms. For instance, if $X^{(m)} \in \Psi^i$ and $X^{(n)} \in \Psi^j$, where $i \neq j$, then $m \neq n$. Simply put, X operates under different mechanisms in the distributions P_{Ψ^i} and P_{Ψ^j}.* ■

This assumption states that whenever two interventions on the same set of variables occur in different domains, they occur with different mechanisms. This is a realistic assumption that precludes the possibility that any interventions that occur in different domains result in the same

5. The assumption that there are no structural changes between domains can be relaxed and is considered in the context of inference, as discussed in [17]. This is an interesting topic for future explorations, and we do not consider this avenue here in the context of structure learning.

exact mechanism. For example, even if medication $\mathbf{\Psi} = \langle \{X\}^1, \{X\}^2 \rangle$ is given to humans Π^1 and bonobos Π^2, it is not unusual to expect the intervention has different mechanisms of action in each domain; $\{X\}^1$ and $\{X\}^2$ occur with distinct mechanisms $\{X^{(i)}\}^1$ and $\{X^{(j)}\}^2$ respectively where the mechanisms are different $i \neq j$. The superscript denoting the mechanism is dropped for simplicity given this assumption.

Next, we define an important operation when comparing two different intervention sets.

Definition 2 (Symmetrical Difference Operator (Δ) in Multiple Domains) *For two domains Π^i, Π^j (possibly $i = j$), given two sets of intervention targets, Ψ^i and Ψ^j, let $\Psi^i \Delta \Psi^j$ denote the symmetrical difference set such that $X \in \Psi^i \Delta \Psi^j$ if $X^{(k)} \in \Psi^i$ and $X^{(k)} \notin \Psi^j$ or vice versa.* ■

This operation identifies the set of variables with unique interventional mechanisms across two intervention targets and also tracks the domain indices. For example, given two interventions sets, $\mathbf{\Psi}^1_j = \{X^{(1)}, Y^{(1)}, Z^{(1)}\}^1$ and $\mathbf{\Psi}^1_i = \{X^{(2)}, Y^{(1)}\}^1$, then $\mathbf{\Psi}^1_j \Delta \mathbf{\Psi}^1_i = \{X, Z\}^{1,2}$. By convention, we drop the mechanism superscript if there is no distinction to be made. In addition, an implication of the above definition and Assumption A3 is that the symmetrical difference of two intervention target sets from two different domains is the union of all the variables in both sets since the mechanisms is always unique. For example, one can consider the intervention set from domain Π^1, $\mathbf{\Psi}^1 = \{X, Y, Z\}^1$, and an intervention set from domain Π^2, $\mathbf{\Psi}^2 = \{X, Y\}^2$. Then, $\mathbf{\Psi}^1 \Delta \mathbf{\Psi}^2 = \{X, Y, Z\}^{1,2}$ because the mechanisms for X, Y in $\mathbf{\Psi}^1$ and $\mathbf{\Psi}^2$ are distinct.

2.1 Multi-distributional invariances: interventions and change-of-domain

This section elaborates on distributional invariances in a multi-domain setting. Starting from an SCM M^i, three qualitatively different sets of distributions arise as illustrated in Figure 2. These distributions are related to the human concepts of "seeing" (called observational), "doing" (interventional), and "imagining" (counter-factual), collectively known as Pearl's Causal Hierarchy (PCH), or the Ladder of Causation [22, 59]. The PCH is a containment hierarchy in which each of these distribution sets can be increasingly refined, going from observational distributions in layer 1 ($\mathcal{L}_1$), interventional in layer 2 ($\mathcal{L}_2$), and counterfactual in layer 3 ($\mathcal{L}_3$). In causal discovery tasks, we observe relevant distributions within each layer $\mathbf{P}^i, \mathbf{P}^j$ [6]. In addition to the PCH and its relevant distributions, each SCM also induces a causal diagram, which captures the causal and unobserved confounding relationships between variables. Formally, this data structure (causal diagram) encodes the constraints discussed above in an economical fashion. There are three types of constraints, which we discuss below.

Type 1. Single-domain and single-distribution invariances When given only observational data in domain Π^i, $P^i_{\{\}}$, invariances in the form of conditional independences (CI) arise from the SCM; $P(\mathbf{Y}|\mathbf{X}, \mathbf{Z}) = P(\mathbf{Y}|\mathbf{Z})$. This considers only one probability distribution, $P(\mathbf{V})$ – whether the value of Y is probabilistically invariant to X once the value of Z is known [2, 6, 22, 66]. The CI statements can be observed within a single distribution corresponding to any single cell in Table 3. Importantly, CI type invariances are reserved to a single-distribution and hence a single-domain. A collection of such CI statements exist in the standard Markov EC characterization, as shown in the left-bottom side in Figure 2 denoted as CI^i (in green). These invariances, or CI statements can be mapped to separation statements in a graphical model, and where some CI holds, the connection (directed and bidirected arrows) between Y and X are removed. The resulting object forms a Markov EC, known as the partial ancestral graph (PAG), where the CI and separation statements are consistent, represents the EC when only observational data is given within a single domain. The FCI algorithm leverages these CI invariances to learn the PAG given observational data $P^i_{\{\}}$ within domain Π^i.

6. The exception is $\mathcal{L}_3$, since counterfactuals are mostly unobserved [65]. Thus causal discovery tasks typically deal with $\mathcal{L}_1$ and $\mathcal{L}_2$ data.

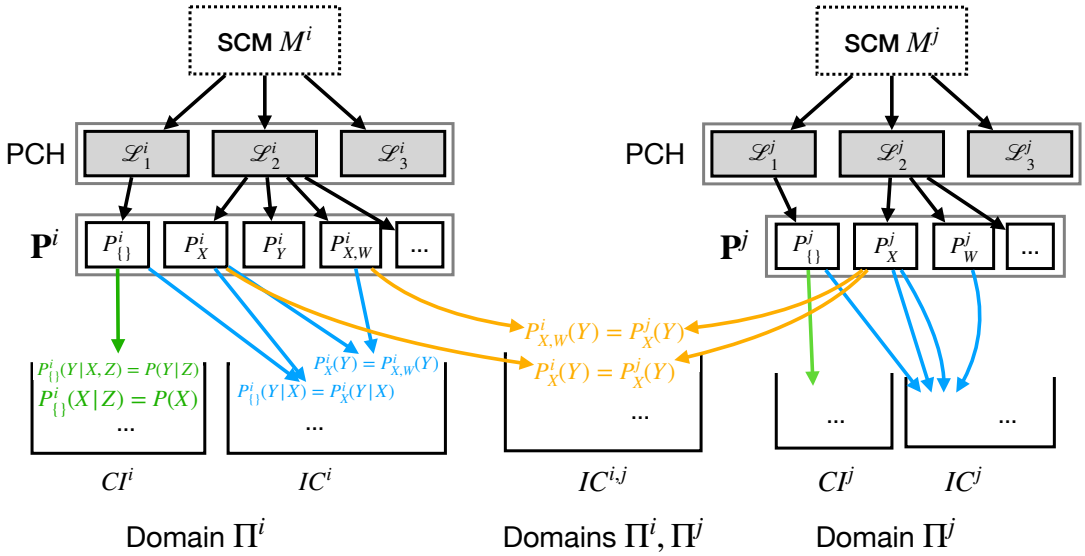


Figure 2: **Overview of how distributional constraints arise from SCMs** - The true, but unobserved SCMs M^i, M^j give rise to Pearl's Causal Hierarchy (PCH), or Ladder of Causation. Each layer of the PCH gives rise to distributions $\mathbf{P}^i, \mathbf{P}^j$ that may be used for causal discovery (counterfactuals are not observed and thus one typically cannot collect data from $\mathcal{L}_3$). From these distributions, different invariances can be seen. In a single distribution, invariances of conditional independences (CI^i, CI^j) can be seen. Within a single SCM, different distributions may be compared giving rise to single-domain invariance constraints (IC^i, IC^j). Across different SCMs, different distributions may be compared giving rise to multi-domain invariance constraints ($IC^{i,j}$).

Type 2. Single-domain and multi-distribution invariances Other invariances arise when considering interventional data from $\mathcal{L}_2$. For example, in SCM M^i of Figure 2, data arising from interventional distributions on X (P^i_X), Y (P^i_Y), X and W simultaneously ($P^i_{X,W}$) can be collected. In domain Π^i, one can compare the different distributions to derived invariance constraints (IC^i for short), which are shown as the bucket in the bottom of Figure 2 (in blue). These correspond to comparing distributions across a single row in Table 3 (not across rows). For example, comparing observations and an intervention on X arising from the SCM M^i, one may observe that Y is invariant between these two distributions (i.e. $P^i_{\{\}}(Y) = P^i_X(Y)$). This IC results in learning that $X \not\to Y$ in the resulting graphical model. In addition, one may compare what happens when we condition on X and observe the IC $P^i_X(Y|X) = P^i_{\{\}}(Y|X)$, as is elaborated below. This IC results in learning that also $X \not\leftrightarrow Y$. These are remarkably very different types of invariances. Other ICs may arise due to the structure in the SCM M^i by plugging in different variables and comparing different distributions. In words, these are different probability distributions that remain invariant under different interventions [7]. These ICs are related to missing directed and bidirected arrows resulting in separation statements in a graphical model. However, this graphical model is not the same as specified in the single-domain single-distributional setting. The causal graph is modified as a function of the variables in the conditioning set and intervention. The works in [3, 7, 8, 28] build upon the Markov EC to characterize these types of ICs in an object called the interventional Markov EC. Importantly, these invariances are markedly different from that of the CI statements, where only a

7. For a more detailed discussion, see discussion on CBNs in [59]

single distribution is present, because one is now comparing probabilities across *different distributions* (regimes) [8].

Type 3. Multiple-domains and multi-distribution invariances Similar invariances (CIs and ICs) in domain Π^j may also be observed by when considering another domain, represented by the SCM M^j. However, when considering distributions across the two SCMs M^i, M^j corresponding to two different domains Π^i, Π^j, different invariances arise that characterize the similarities and disparities between the two SCMs. The IC described now leverage comparisons across any two cells in Table 3, where any two distributions arising from different domains can be compared. For example, one may compare the observational distributions in both domains to obtain an IC $P^i_{\{\}}(Y) = P^j_{\{\}}(Y)$. This implies that there is no S-node pointing to Y, $S^{i,j} \not\to Y$. Another example of an invariance is one may compare the intervention on X in domain Π^i, $\mathbf{\Psi}^i = \{X\}^i$ with the intervention on X in domain Π^j, $\mathbf{\Psi}^j = \{X\}^j$. If one observes that the distribution of Y is invariant between these two distributions $P^i_X(Y) = P^j_X(Y)$, then one is able to also discern that $X \not\to Y$ and $S^{i,j} \not\to Y$ in some graphical model. Critically, since Y is not intervened in either distribution, one can reason that Y has the same mechanism within both domains [9]. Again, there may be other invariance constraints ($IC^{i,j}$) that can be observed by comparing any pair of distributions between the two domains.

This discussion illustrates the differences between domain and interventions from first principles derived from the SCMs. These constraints generalizes the single-domain setting and considers as an input a set of distributions from (possibly) different domains to characterize a more general type of "Markov" condition [10]. Consider distributional invariances of the form $P^i_{\mathbf{W}}(\mathbf{Y}|\mathbf{X}) = P^j_{\mathbf{K}}(\mathbf{Y}|\mathbf{X})$ such that distributions could stem from different domains ($i \neq j$). Such an invariance implies that the conditional distribution of $\mathbf{Y}|\mathbf{X}$ remains the same across domains i and j under interventions on $\mathbf{W}$ and $\mathbf{K}$, respectively. Whenever $i = j$, the invariances reduce to the ones considered in the interventional Markov EC. From a syntactic perspective, it is immediate to see that multi-domain invariances generalize the single-domain invariances analyzed in observational and interventional data. Building on the concepts of invariances, we introduce in the next section the general MD-Markov property that contracts CI and IC statements to separation statements within a graphical model.

2.2 MD-Markov Property

We start by introducing the multi-domain Markov property and explain it below.

Definition 3 (MD-Markov Property) *Given the Multi-domain setup, the tuple of distributions* $\mathbf{P}$ *satisfies the MD-Markov property with respect to the pair* $\langle G_S, \mathbf{\Psi} \rangle$ *if the following two conditions holds for any disjoint* $\mathbf{Y}, \mathbf{W}, \mathbf{Z} \subseteq \mathbf{V}$*:*

1. *(Conditional) independences: For any intervention set,* $\mathbf{\Psi}^i_j \in \mathbf{\Psi}$*:*
   $P^i_j(\mathbf{y}|\mathbf{w}, \mathbf{z}) = P^i_j(\mathbf{y}|\mathbf{w})$ *if* $(\mathbf{Y} \perp\!\!\!\perp \mathbf{Z}|\mathbf{W})_{G_S}$
2. *(Conditional) distributional invariance: For any pairs of intervention sets,* $\mathbf{\Psi}^i_m, \mathbf{\Psi}^j_l \in \mathbf{\Psi}$*:*
   $P^i_m(\mathbf{y}|\mathbf{w}) = P^j_l(\mathbf{y}|\mathbf{w})$ *if* $(\mathbf{Y} \perp\!\!\!\perp \mathbf{K} \cup \mathbf{S}^{i,j}|\mathbf{W} \backslash \mathbf{W}_{\mathbf{K}})_{G_{S_{\underline{\mathbf{W}_{\mathbf{K}}}, \overline{\mathbf{R}(\mathbf{W})}}}}$

where $\mathbf{K} = \mathbf{\Psi}^i_m \Delta \mathbf{\Psi}^j_l$, $\mathbf{W}_{\mathbf{K}} = \mathbf{W} \cap \mathbf{K}, \mathbf{R} = \mathbf{K} \backslash \mathbf{W}_{\mathbf{K}}$ *and* $\mathbf{R}(\mathbf{W}) \subseteq \mathbf{R}$ *are non-ancestors of* $\mathbf{W}$ *in* G_S*.*

8. Note that when conditional independence is found, which is a layer 1-type of invariance, both directed and bidirected arrows should be removed. On the other hand, the invariances involving constraints from layers 1 and 2 are more fine-grained, capable of detecting the absence of individual arrows, whether they are directed or bidirected.
9. This is a crucial point that underlies the transportabiliy and external validity literature, allowing scientists to make out-of-domain generalizations [67].
10. We use the term Markov here to indicate graphical conditions that imply invariances within the probability distributions.

Let $S^{\mathbf{\Pi}}_{\mathcal{K}}(G_S, \mathbf{\Psi})$ denote the set of distribution tuples that satisfy the MD-Markov property with respect to $\langle G_S, \mathbf{\Psi} \rangle$ where $\mathcal{K}$ denotes the known intervention targets. ■

The MD-Markov property is related to the inverse of the causal calculus, known as σ-calculus [63] when considering soft interventions that do not alter the causal structure. For instance, the inverse of R1 of the calculus gives condition 1 in the MD-Markov property. The inverse of R2 and R3 of the calculus gives condition 2 in the MD-Markov property. Given Assumption A2, the MD-Markov property only models soft interventions that do not alter the graphical structure. Note though that condition 2 of the MD-Markov property involves cutting incoming and outgoing edges of certain nodes. These modifications of the graphical model are not due to the intervention actually modifying the structure, but are a tool for querying a submodel of the SCM (i.e. where certain edges are modified) to determine if separation statement holds in the modified graph $G_{\overline{S_{W_K}}, \overline{R(W)}}$. The interventions themselves do not change the graphical structure. Thus, we see that the MD-Markov property definition does not contradict Assumption A2.

For example, $G_{\overline{S_{W_K}}, \overline{R(W)}}$ cuts the outgoing edges of the set W_K and the incoming edges for the set $R(W)$. The MD-Markov property will modify G_S given different tests of the form $P^i_K(Y|W) \stackrel{?}{=} P^j_{K'}(Y|W)$. Depending on the interventions compared (K, K'), the conditioning set $\mathbf{W} \subset \mathbf{V}$, and the domains, the resulting condition 2 will check different submodels. Specifically, sets W_K and $R(W)$ will contain different variables, and in the context of comparing distributions occurring in different domains Π^i, Π^j, one would also check separation with respect to the corresponding S-node $S^{i,j}$. In summary, condition 2 of the MD-Markov property will query various submodels of the SCM (represented by the modifications of certain edges) to determine if IC constraints are present when comparing soft interventions that do not change the underlying causal structure.

A valid question at this point is how can we characterize all kinds of interventions, and specifically hard interventions through the lens of do-calculus [1]? Characterizing hard interventions in non-Markovian setting is still an open research problem, and we discuss some of its nuances in Section 5.3. In this paper, all interventions are soft and do not alter the causal structure.

Next, we illustrate several examples about this property. In particular, we will illustrate the connection to a particular version of the σ-calculus that preserves the graphical structure. The inverse of each of rules map to the missingness of edges between two variables. When dealing with multiple domains, invariances implied by the do-calculus rules are related to the missingness of an S-node. First, let us consider the MD-Markov property in the single-domain setting with a few examples.

Notationally, we will say that $V_i \not\to V_j$ means there is no directed edge between V_i and V_j. Similarly, $V_i \not\leftrightarrow V_j$ will indicate there is no bidirected edge between V_i and V_j.

2.2.1 Single-domain MD-Markov property

Single-distribution invariances First, when there is only a single domain $\mathbf{\Pi} = \{\Pi^1\}$ and a single distribution, the first constraint reduces to standard separation on a causal diagram, that maps to CI invariances. Consider the selection diagram in Figure 1(d). The first condition in the MD-Markov property is related to R1 of the do-calculus rules, where the CI constraint $P(Z|Y) = P(Z|Y, X)$ is encoded via $(Z \perp\!\!\!\perp X|Y)_{G_S}$. This invariance implies that there is a missing edge between X and Z. More specifically, $X \not\to Z$, $X \not\leftarrow Z$, and $X \not\leftrightarrow Z$. Here, the subscripts and superscripts on P were dropped since this constraint holds within the same intervention set and same domain.

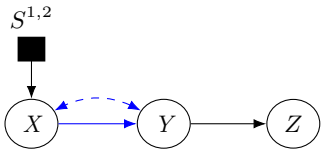
Multi-distribution invariances Next, consider when there are multiple distributions from the same domain to consider.

The second condition in the MD-Markov property in a single domain considers two interventional distributions P^i_m and P^i_l, corresponding to intervention sets $\mathbf{\Psi}^i_m$ and $\mathbf{\Psi}^i_l$, respectively, within domain Π^i. Based on the intervention sets compared and the conditioning set, the graph will get surgically modified according to the second condition of Def. 3.

The first example compares an intervention with an observation (in this case the empty intervention set). We will illustrate the interplay between the symmetric difference among the intervention sets (in red), the outgoing edges that are cut (in orange), and the incoming edges that are cut (in blue).

Example 2 (R3; do-do) *Consider again the selection diagram in Figure 1(d) (shown below for convenience), the tuple of intervention sets* $\mathbf{\Psi}^{\Pi} = \langle \mathbf{\Psi}_1, \mathbf{\Psi}_2 \rangle = \langle \{\}, \{Y\} \rangle$*, and the test* $P_{\{\}}(X) \stackrel{?}{=} P_Y(X)$*. In words, we are interested in testing the invariance of the distribution* $P(X)$ *between these two intervention sets. The domain index superscripts were dropped since we consider only distributions arising from the same domain. In this case, we construct the sets described in the second condition of the MD-Markov property:*

$K = \Psi_1 \Delta \Psi_2 = \{Y\}$	*(Symmetric difference of interventions)*
$W = \{\}$	*(Conditioning set)*
$W_K = W \cap K = \{\}$	*(Overlap of interventions and conditioning set)*
$R = K \backslash W_K = \{Y\}$	*(What is intervened, but not conditioned)*
$R(W) = R \backslash Anc(W) = \{Y\}$	*(Not an ancestor of W)*

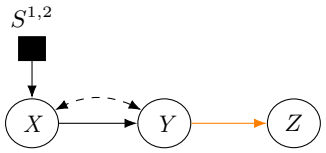


$P_{\{\}}(X) = P_Y(X)$ *since* $(X \perp\!\!\!\perp Y)_{G_{\overline{Y}}}$ *holds. This IC implies there is a missing directed edge between* Y *and* X *(*$Y \not\to X$*). This is because if there was an edge between the intervened variable (*Y*) and* X*, then the invariance would not hold since the effect of the intervention would cause some change in the distribution.* ■

The resulting invariance property is related to R3 of the σ-calculus [63]. The missing edge arises because one compares an intervention on Y vs the empty intervention set (i.e. observational distribution). R3 allows one to infer the lack of causal relations between variables, and consequently, directed edges in the causal graph. Next, consider the various examples illustrating the application of R2, where we compare an intervention to when we condition on the intervened variable.

Example 3 (R2; do-see) *Consider the same selection diagram, the tuple of intervention sets* $\mathbf{\Psi} = \langle \mathbf{\Psi}_1, \mathbf{\Psi}_2 \rangle = \langle \{\}, \{Y\} \rangle$*, and the test* $P_{\{\}}(Z|Y) \stackrel{?}{=} P_Y(Z|Y)$*. In words, we will consider the distribution* $P(Z|Y)$ *and test for its invariance across two distributions.*

$K = \Psi_1 \Delta \Psi_2 = \{Y\}$
$W = \{Y\}$
$W_K = W \cap K = \{Y\}$
$R = K \backslash W_K = \{\}$
$R(W) = R \backslash Anc(W) = \{\}$



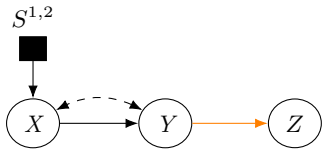
The resulting invariance property is related to R2 of the do-calculus rules, and since $(Z \perp\!\!\!\perp Y)_{G_{\underline{Y}}}$ *holds the invariance equation holds (*$P_{\{\}}(Z|Y) = P_Y(Z|Y)$*). This IC implies that there is no bidirected edge between* Y *and* Z *(*$Y \not\leftrightarrow Z$*).* ■

The implication of R2 allows one to infer the existence of an unblockable back-door path. By comparing an intervention on Y conditioned on the value of Y, if the IC fails to hold, then there must be a back-door path that causes the difference in distributions. Alternatively, if $Y \to Z$ only,

then as we see in the example, the IC holds. Another example considers comparing two interventions on the same set of variables with different mechanisms. Remember that as a result of Assumption A3, comparing two intervention sets with the same variables across different domains will always result in a different mechanism.

Example 4 (R2; do-see with different mechanisms) *Consider the selection diagram in Figure 1(d), the tuple of intervention sets* $\mathbf{\Psi} = \langle \mathbf{\Psi}_1, \mathbf{\Psi}_2 \rangle = \langle \{Y^{(i)}\}, \{Y^{(j)}\} \rangle$*, and the test* $P_{Y^{(i)}}(Z|Y) \stackrel{?}{=} P_{Y^{(j)}}(Z|Y)$*. In words, we will consider the distribution* $P(Z|Y)$ *and test for its invariance across two distributions.*

$K = \Psi_1 \Delta \Psi_2 = \{Y\}$
$W = \{Y\}$
$W_K = W \cap K = \{Y\}$
$R = K \backslash W_K = \{\}$
$R(W) = R \backslash Anc(W) = \{\}$

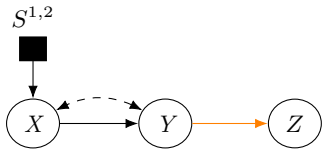


The resulting invariance property is related to R2 of the do-calculus rules. $P_{Y^{(i)}}(Z|Y) = P_{Y^{(j)}}(Z|Y)$ *since* $(Z \perp\!\!\!\perp Y)_{G_{\underline{Y}}}$ *holds. This IC implies that there is no bidirected edge between* Y *and* Z $(Y \not\leftrightarrow Z)$*.* ■

In summary, we see R3 allows one to infer causal relations between variables by comparing different interventions (do-do), and R2 allows one to infer spurious relations between variables, and consequently latent variables in the causal diagram. This inference stems from comparing an intervention on some set of variables versus another where the set of variables is conditioned on (do-see). Finally, consider the combined R2 + R3 of the do-calculus, where we combine the previous two concepts.

Example 5 (R2 + R3; do-see and do-do) *Consider the selection diagram in Figure 1(a), the tuple of intervention sets* $\mathbf{\Psi} = \langle \mathbf{\Psi}_1, \mathbf{\Psi}_2 \rangle = \langle \{X\}, \{Y\} \rangle$*, and the test* $P_X(Z|Y) \stackrel{?}{=} P_Y(Z|Y)$*. We will consider the distribution* $P(Z|Y)$ *and test for its invariance between two distributions.*

$K = \Psi_1 \Delta \Psi_2 = \{X, Y\}$
$W = \{Y\}$
$W_K = W \cap K = \{Y\}$
$R = K \backslash W_K = \{X\}$
$R(W) = R \backslash Anc(W) = \{\}$

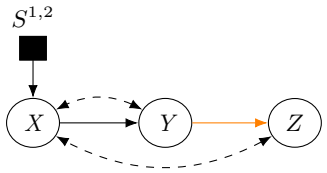


The resulting invariance property leverages both R2+R3 of the do-calculus rules. $P_X(Z|Y) = P_Y(Z|Y)$ *since* $(Z \perp\!\!\!\perp X, Y)_{G_{\underline{Y}}}$ *holds. This invariance implies that there is no bidirected edge between* Z *and* Y $(Y \not\leftrightarrow Z)$*, and* X *and* Y $(X \not\leftrightarrow Z)$*. The missing bidirected edge arises because we condition on* X*, and compare with an intervention where* X *is intervened. If a bidirected edge existed between* X *and* Z*, or* Y *and* Z*, then a back-door path would be opened when conditioning on* Y*.* ■

In the previous example, note the $R(W)$ set is empty as we do not cut edges for nodes that are ancestors of W (X in this case is an ancestor of Y). This is because conditioning on a variable that is a descendant of a collider will open up the collider path.

Example 6 (R2 + R3; with lack of IC due to backdoor path) *Consider the selection diagram in Figure 1(a) with an extra bidirected edge between* X *and* Z*, the tuple of intervention sets* $\mathbf{\Psi} = \langle \mathbf{\Psi}_1, \mathbf{\Psi}_2 \rangle = \langle \{X\}, \{Y\} \rangle$*, and the test* $P_X(Z|Y) \stackrel{?}{=} P_Y(Z|Y)$*. We consider the distribution* $P(Z|Y)$ *and test for its IC.*

$K = \Psi_1 \Delta \Psi_2 = \{X, Y\}$
$W = \{Y\}$
$W_K = W \cap K = \{Y\}$
$R = K \backslash W_K = \{X\}$
$R(W) = R \backslash Anc(W) = \{\}$



Since $(Z \not\perp\!\!\!\perp X, Y)_{G_{\underline{Y}}}$ *holds, then the equality does not necessarily hold (i.e.* $P_X(Z|Y) \neq P_Y(Z|Y)$*). The inequality implies that there is a backdoor path between* Z *and* Y*.* ■

The above examples demonstrate that the reversal of the do-calculus R2 and R3 from IC comparing different distributions imply additional constraints on the graphical structure. That is, if some IC implying do-calculus R2 is observed, then this implies the absence of a backdoor path. If some IC implying do-calculus R3 is observed, then this implies the absence of a directed path.

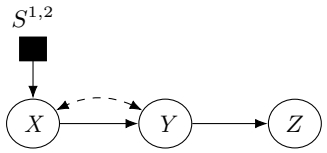
2.2.2 Multi-domain MD-Markov Property

In multiple domains, the second condition for the MD-Markov property is where one considers two interventional distributions P^i_m and P^j_l, corresponding to intervention sets $\mathbf{\Psi}^i_m$ and $\mathbf{\Psi}^j_l$, respectively within two distinct domains, $\Pi^i \neq \Pi^j$. Here the superscripts are retained since we are comparing different distributions across domains. We illustrate various examples of invariances across domains.

Observational data in both Π^i and Π^j First, consider the setting where observational data is in both domains, Π^1, Π^2. That is, $\mathbf{\Psi}^\Pi = \langle \mathbf{\Psi}^1_1, \mathbf{\Psi}^2_1 \rangle = \langle \{\}^1, \{\}^2 \rangle$ (see-see across domains). Let us investigate whether the distribution $P(Y)$ is invariant within $\mathbf{\Psi}^1_1, \mathbf{\Psi}^2_1$.

Example 7 (see-see across domains) *Consider the selection diagram in Figure 1(d), the tuple of intervention sets* $\mathbf{\Psi}^\Pi = \langle \{\}^1, \{\}^2 \rangle$*, and the test* $P^1_{\{\}}(Z|Y) \stackrel{?}{=} P^2_{\{\}}(Z|Y)$*. We will determine if the distribution* $P(Z|Y)$ *is invariant across two distributions from different domains* Π^1 *and* Π^2*.*

$K = \Psi^i_1 \Delta \Psi^j_1 = \{\}$
$W = \{\}$
$W_K = W \cap K = \{\}$
$R = K \backslash W_K = \{\}$
$R(W) = R \backslash Anc(W) = \{\}$



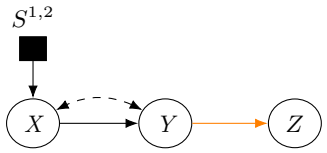
Thus $P^i_{\{\}}(Z|Y) = P^j_{\{\}}(Z|Y)$ *since* $(Z \perp\!\!\!\perp S^{1,2}|Y)_{G_S}$ *holds. In words, this means a missing S-node for the variable* Z $(S^{1,2} \not\to Z)$*. This implies that the distribution of* Z *given* Y *is invariant across domains* Π^1 *and* Π^2*.* ■

The observational distribution in domains Π^i, Π^j allows one to determine which mechanisms are different due to a domain change. The next example illustrates that observational data is not necessary to determine such an invariance.

Observational data in Π^i, but not Π^j Next, consider when observational data is not present in one of the domains, For example, $\mathbf{\Psi}^\Pi = \langle \{Y\}^1, \{\}^2 \rangle$.

Example 8 (do-see across domains) *Consider the selection diagram in Figure 1(d), the tuple of intervention sets* $\mathbf{\Psi}^\Pi = \langle \{Y\}^1, \{\}^2 \rangle$*, and the test* $P^1_Y(Z|Y) \stackrel{?}{=} P^2_{\{\}}(Z|Y)$*.*

$K = \Psi_1^i \Delta \Psi_2 = \{Y\}$
$W = \{Y\}$
$W_K = W \cap K = \{Y\}$
$R = K \backslash W_K = \{\}$
$R(W) = R \backslash Anc(W) = \{\}$



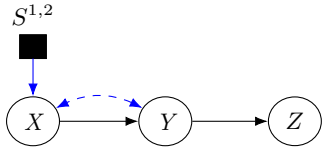
Thus $P^i_Y(Z|Y) = P^j_{\{\}}(Z|Y)$ *since* $(Z \perp\!\!\!\perp Y)_{G_{S_{\underline{Y}}}}$ *and* $(Z \perp\!\!\!\perp S^{1,2})_{G_{S_{\underline{Y}}}}$ *holds (i.e.* $(Z \perp\!\!\!\perp Y, S^{1,2})_{G_{S_{\underline{Y}}}}$*). This results in a missing S-node for the variable* Z *(*$S^{1,2} \not\to Z$*), and a missing bidirected edge between* Z *and* Y *(*$Z \not\leftrightarrow Y$*). As a result, the distribution* $P(Z|Y)$ *is invariant across domains* Π^1 *and* Π^2*, and invariant when comparing an intervention on* Y *with conditioning on* Y*.* ■

This example illustrates that one can compare interventional distributions across domains, while conditioning on the intervened variable to determine both if a S-node, and a backdoor path exists. Here, one of the intervention sets was still the observational distribution. The next example illustrates there are IC even when comparing two non-empty intervention sets from different domains.

No observational data in Π^i or Π^j Finally, consider when observational data is not present in either domains.

Example 9 (do-do across domains) *Consider the selection diagram in Figure 1(d), the tuple of intervention sets* $\mathbf{\Psi}^{\Pi} = \langle \{X\}^i, \{X\}^j \rangle$*, and the test* $P^i_{X^{(1)}}(Z|Y) \stackrel{?}{=} P^j_{X^{(2)}}(Z|Y)$*.*

$K = \Psi_1^i \Delta \Psi_2 = \{X\}$
$W = \{Y\}$
$W_K = W \cap K = \{\}$
$R = K \backslash W_K = \{X\}$
$R(W) = R \backslash Anc(W) = \{X\}$



Note that the symmetrical difference results in X *because of Assumption A3, such that the intervention sets* $\{X\}^i$ *for domain* Π^i *and* $\{X\}^j$ *for domain* Π^j *have distinct mechanisms (i.e.* $\{X^{(1)}\}$ *and* $\{X^{(2)}\}$*).*

Thus $P^i_{X^{(1)}}(Z|Y) = P^j_{X^{(2)}}(Z|Y)$ *since* $(Z \perp\!\!\!\perp X, S^{i,j}|Y)_{G_{S_{\overline{X}}}}$ *holds. This results in a missing S-node for the variable* Z *(*$S^{i,j} \not\to Z$*), and a missing directed edge between* X *and* Z *(*$X \not\to Z$*). The distribution* $P(Z|Y)$ *is invariant across domains* Π^i *and* Π^j*, and invariant when comparing interventions on* X *with different mechanisms.* ■

From these examples, one can see that IC constraints compare different pairs of distributions. In the same domain, this amounts to comparing interventions in different settings to determine the lack of a causal (missing directed edge), or a confounding relation (missing bidirected edge). When comparing distributions across different domains, one can leverage observations to determine the differences in mechanisms due to the change in domain. In addition, one can compare intervention sets across domains. However, when comparing intervention sets across domains, an IC will imply both a separation statement with respect to the intervened nodes, and to the corresponding S-node.

Def. 3 shows how to map graphical conditions of the selection diagram G_S for a given intervention tuple $\mathbf{\Psi}$ to invariances found in the distributions $\mathbf{P}$. The examples illustrated above draw a connection between the rules of σ-calculus, graphical implications and the distributional invariances implied by the underlying SCM(s).

Building in this uncertainty, we define MD-Markov equivalence next.

Definition 4 (MD-Markov Equivalence) *Let* $\mathbf{\Pi}$ *and* $\mathcal{K}$ *denote fixed sets of domains and indices of known intervention targets, respectively. Given selection diagrams* G_S, G'_S *defined over* $\mathbf{V} \cup \mathbf{S}$

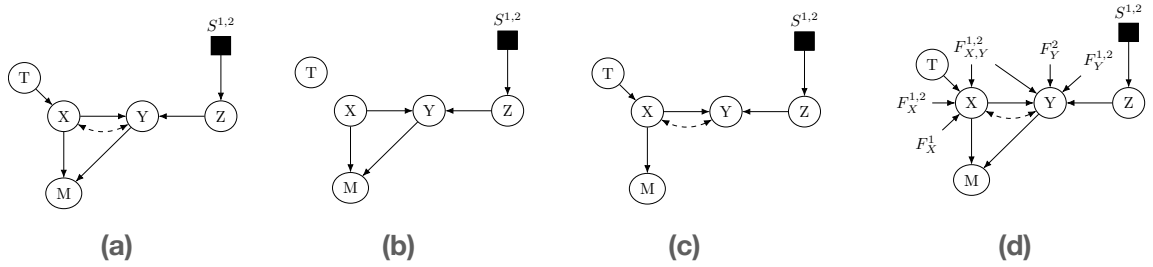


Figure 3: Example selection diagram G_S (a), along with surgically modified selection diagrams: $G_{S_{\overline{X}}}$ (b), $G_{S_{\underline{Y}}}$ (c). A representative augmented selection diagram (d) with interventions $\mathbf{\Psi} = \langle \{\}^{\overline{1}}, \{\}^{2}, \{X\}^{\overline{1}}, \{Y\}^{2} \rangle$ and corresponding distributions $\mathbf{P}$.

and the corresponding intervention targets $\mathbf{\Psi}, \mathbf{\Psi}'$*, the pairs* $\langle G_S, \mathbf{\Psi} \rangle$ *and* $\langle G'_S, \mathbf{\Psi}' \rangle$ *are said to be MD-Markov equivalent if, the set of distribution tuples that satisfy the MD-Markov property for both pairs are the same (i.e.* $S^{\Pi}_{\mathcal{K}}(G_S, \mathbf{\Psi}) = S^{\Pi}_{\mathcal{K}}(G'_S, \mathbf{\Psi}')$*).* ■

The $\mathcal{K}$ term encodes known-target interventions that fix the intervention sets for those distributions. This allows the MD-Markov property to compare tuples of distributions with known *and* unknown intervention targets paired with a selection diagram to other tuples of intervention sets and selection diagrams.

2.3 Graphical Characterization of MD Equivalence Class

The MD-Markov equivalence, in principle, allows one to compare pairs of selection diagrams and intervention targets to determine if they satisfy the MD-Markov property with respect to the same set of distributions. Testing CI and IC constraints would thus allow one to learn about the graphical structure. However, as seen in Ex.2-9 this process requires performing various manipulations of the graph for each distribution pair and conditioning set argument. As another example, consider the selection diagram in Figure 3(a). In Figure 3(b), one can determine that $P^1_{\{\}}(Z) = P^1_X(Z)$ because $(Z \perp\!\!\!\perp X)_{G_{S_{\overline{X}}}}$. In addition, in Figure 3(c), one can determine that $P^1_{\{\}}(M|Y,X) = P^2_Y(M|Y,X)$ because $(M \perp\!\!\!\perp Y, S^{1,2}|X)_{G_{S_{\underline{Y}}}}$.

We see that different IC constraints lead to different submodels of the original graphical model. When considering the learning task (i.e. causal discovery) given data, it is desirable to be able to map invariances within distributions to some graphical separation in the graphical model. It is thus desirable to *characterize* all testable CI and IC constraints with a separation statement on a single graphical model. This would allow one to infer separation statements in the underlying selection diagram, and facilitate learning an EC. Besides facilitating learning, a completely graphical characterization would facilitate a more efficient data structure for representing the different invariances present given a set of distributions and interventional targets. Figure 3(d) shows a refined treatment that will i) facilitate the learning task, and ii) provides an efficient representation of all the invariances implied by the dataset given. The graph shown allows one to map all CI and IC constraints to a separation statement on the graphical model. For example, to determine if $P^1_{\{\}}(Z) = P^1_X(Z)$, we check that $(Z \perp\!\!\!\perp F^1_X)$ in Figure 3(d). In addition, we determine that $P^1_{\{\}}(M|Y,X) = P^2_Y(M|Y,X)$ by checking that $(M \perp\!\!\!\perp F^{1,2}_Y, S^{1,2})$ in Figure 3(d).

In particular, we will leverage a graphical approach that encodes the symmetric differences of interventions using F-nodes [3]. As discussed in Section 2.1, invariance constraints (ICs) arise from the SCM. Interestingly, these distributional invariances can be characterized graphically by an extended separation property on an augmented graph with "F-nodes". These special, "virtual" nodes serve as graphical representations of the differences in distributions due to the interventions and are not part

of the causal system itself. That is, $F_k^i \to V_k$ does not imply any causal relationship because F_k^i is not a random variable, but a node used to represent an intervention k that occurs in domains Π^i. The following definition extends to the multi-domain setting for selection diagrams, and introduces an augmented diagram that will represent all testable CI and IC constraints in a single graphical model.

Definition 5 (Augmented selection diagram) *Consider the Multi-domain setup with at least one intervention set and corresponding distribution per domain that is considered. Let the multiset $\mathcal{I}$ be defined as:*

$$\mathcal{I} = \{\mathbf{K}_1, \mathbf{K}_2, ...\mathbf{K}_k\} = \{\mathbf{K} \mid \mathbf{\Psi}_m^i \Delta \mathbf{\Psi}_l^j = \mathbf{K} \text{ for } \mathbf{\Psi}_m^i, \mathbf{\Psi}_l^j \in \mathbf{\Psi}\}$$

.

Let the set $\boldsymbol{\mathcal{D}}_{ij} = \bigcap_{m,l} \psi_{m,l}^{i,j}$, where $\psi_{m,l}^{i,j} = (\mathbf{\Psi}_m^i \cup \mathbf{\Psi}_l^j)$, be the intersection of all symmetric differences between pairwise intervention sets for a pair of domains Π^i, Π^j.

The augmented graph of $G_S = (\mathbf{V} \cup \mathbf{L} \cup \mathbf{S}, \mathbf{E})$ with respect to $\mathbf{\Psi}$ is denoted as $Aug_{\mathbf{\Psi}}(G_S) = (\mathbf{V} \cup \mathbf{L} \cup \mathbf{S} \cup \boldsymbol{\mathcal{F}} \cup \boldsymbol{\mathcal{S}}, \mathbf{E} \cup \boldsymbol{\mathcal{E}}_{\boldsymbol{\mathcal{S}}} \cup \boldsymbol{\mathcal{E}}_{\boldsymbol{\mathcal{F}}})$ and constructed as follows starting with a copy of G_S:

1. *add a new F-node pointing to node $v \in K_l$, $F_l^{i,j} \to v$, for every pair of distributions compared in set $\mathcal{I}$, where the superscript i, j indicates the pair of domain indices. If $i = j$, then the superscript is dropped.*
2. *add a new S-node pointing to $k \in \boldsymbol{\mathcal{D}}_{ij}$, $S^{i,j} \to k$ for every pair of domains $\Pi^i \neq \Pi^j$.*

Denote the sets $\boldsymbol{\mathcal{F}} = \{F_l^{i,j}\}^{i,j \in [N]}$, $\boldsymbol{\mathcal{S}} = \{S_k^{i,j}\}_{k \in \boldsymbol{\mathcal{D}}_{ij}}$, $\boldsymbol{\mathcal{E}}_{\boldsymbol{\mathcal{F}}} = \{(F_l^{j,k}, v)\}_{v \in \mathbf{K}_l}$, and $\boldsymbol{\mathcal{E}}_{\boldsymbol{\mathcal{S}}} = \{(S^{i,j}, k)\}_{k \in \boldsymbol{\mathcal{D}}_{i,j}}$. ■

The F-nodes graphically encode the symmetrical difference sets between every pair of intervention targets in $\mathbf{\Psi}^{\Pi}$ (i.e. $\mathbf{\Psi}_m^i \Delta \mathbf{\Psi}_l^j$) within and across the different domains in Π. $F_k^{i,i} = F_k^i$ denotes an F-node representing the kth symmetric difference of intervention targets within domain i and $F_k^{i,j}$ denotes an F-node from comparing intervention targets between domains i and j. The result is an augmented selection diagram with the original causal structure augmented with S-nodes, F-nodes, and their corresponding edges.

We ground this construction with a few examples [11].

Example 10 (Multi-domain with observational data in both domains) *Consider the selection diagram in Figure 1(d) (shown below for convenience), and the tuple of intervention sets $\mathbf{\Psi} = \langle \{\}^1, \{\}^2 \rangle$. Following the construction in Def. 5, we define the following sets of variables:*

1. $\mathbf{K} = \{K | K = \{\} \Delta \{\}\} = \{\}$
2. $\boldsymbol{\mathcal{F}} = \{\}$
3. $\boldsymbol{\mathcal{D}}_{ij} = \{\}$
4. $\boldsymbol{\mathcal{S}} = \{\}$
5. $\boldsymbol{\mathcal{E}}_{\boldsymbol{\mathcal{F}}} = \{\}$
6. $\boldsymbol{\mathcal{E}}_{\boldsymbol{\mathcal{S}}} = \{\}$

We add these sets of nodes and edges to the existing selection diagram G_S, which results in the augmented selection diagram $Aug_{\Psi}(G_S)$ shown in Figure 4(a). In this simple setting, no additional S-nodes and edges (beyond the one that is already there), or F-nodes and edges are added. ■

11. Note the Augmented Selection Diagram is only well defined over domains with a distribution. That is, if one has a selection diagram over domains Π^1, Π^2, but no intervention sets and associated distributions in one of the domains, then there is no augmented selection diagram.

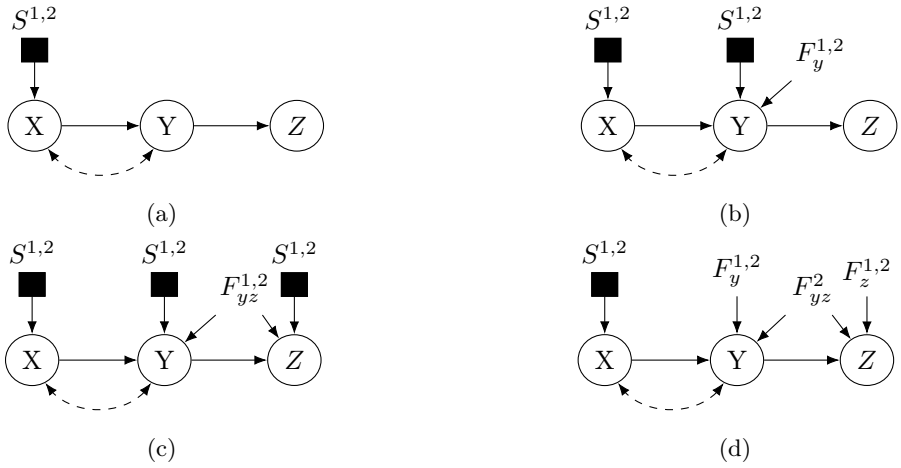


Figure 4: **Examples of augmented selection diagrams for G_S** - Selection diagram G_S (a), and the Augmented Selection Diagrams resulting from intervention sets $\mathbf{\Psi}^b = \langle \{\}^1, \{Y\}^2 \rangle$ (b), $\mathbf{\Psi}^c = \langle \{Z\}^1, \{Y\}^2 \rangle$ (c), $\mathbf{\Psi}^d = \langle \{\}^1, \{Y\}^2, \{Z\}^2 \rangle$ (d).

Thus, when observational data is available in each domain Π^i and Π^j, the S-nodes representing the mechanism differences between the two domains is the same in the augmented selection diagram as it is in the original selection diagram. The next three examples illustrates when observational data is not necessarily available.

Example 11 (Multi-domain with observational data in one domain) *Consider the selection diagram in Figure 4(a), and the tuple of intervention sets $\mathbf{\Psi} = \langle \{\}^1, \{Y\}^2 \rangle$. Again, following the construction in Def. 5, we define the following sets:*

1. $\mathbf{K} = \{K | K = \{\}^1 \Delta \{Y\}^2\} = \{Y\}$
2. $\boldsymbol{\mathcal{F}} = \{F_y^{1,2}\}$
3. $\boldsymbol{\mathcal{D}}_{ij} = \bigcap \{Y\} = \{Y\}$
4. $\boldsymbol{\mathcal{S}} = \{S_y^{1,2}\}$
5. $\boldsymbol{\mathcal{E}}_{\boldsymbol{\mathcal{F}}} = \{F_y^{1,2} \to Y\}$
6. $\boldsymbol{\mathcal{E}}_{\boldsymbol{\mathcal{S}}} = \{S^{1,2} \to Y\}$

The extra sets of nodes and edges are added to the original selection diagram resulting in the augmented selection diagram shown in Figure 4(b). ■

Example 12 (Multi-domain without observational data) *Consider the selection diagram in Figure 4(a), and the tuple of intervention sets $\mathbf{\Psi} = \langle \{Z\}^1, \{Y\}^2 \rangle$. Define the following sets of variables:*

1. $\mathbf{K} = \{K | K = \{Z\}^1 \Delta \{Y\}^2\} = \{Y, Z\}$
2. $\boldsymbol{\mathcal{F}} = \{F_{yz}^{1,2}\}$
3. $\boldsymbol{\mathcal{D}}_{ij} = \bigcap \{Z, Y\} = \{Z, Y\}$
4. $\boldsymbol{\mathcal{S}} = \{S_y^{1,2}, S_z^{1,2}\}$

5. $\boldsymbol{\mathcal{E}}_{\boldsymbol{\mathcal{F}}} = \{F_{yz}^{1,2} \to Y, F_{yz}^{1,2} \to Z\}$
6. $\boldsymbol{\mathcal{E}}_{\boldsymbol{\mathcal{S}}} = \{S^{1,2} \to Y, S^{1,2} \to Z\}$

An F-node is added pointing to Y and Z, which represents the two distributions with interventions $\{Z\}^1$ and $\{Y\}^2$, and an S-node is added pointing to Y and Z resulting in the augmented selection diagram shown in Figure 4(c). ■

In these examples, S-nodes are added to different nodes to capture the uncertainty in the cross-domain invariances. The added S-nodes are a byproduct of what intervention sets are available within each domain pair. However, the next example illustrates that even without observational data in each domain, S-nodes are not necessarily added.

Example 13 (Observational data is not required to characterize all S-node edges) *Consider the selection diagram in Figure 4(a), and the tuple of intervention sets $\boldsymbol{\Psi} = \langle\{\}^1, \{Y\}^2, \{Z\}^2\rangle$.*

Define the following sets of variables:

1. $\mathbf{K} = \{\{\}^1\Delta\{Y\}^2, \{\}^1\Delta\{Z\}^2, \{Y\}^2\Delta\{Z\}^2\} = \{\{Y\}^{1,2}, \{Z\}^{1,2}, \{Y,Z\}^2\}$
2. $\boldsymbol{\mathcal{F}} = \{F_y^{1,2}, F_z^{1,2}, F_{yz}^2\}$
3. $\boldsymbol{\mathcal{D}}_{ij} = \bigcap\{Z\}, \{Y\}, \{Z,Y\} = \{\}$
4. $\boldsymbol{\mathcal{S}} = \{\}$
5. $\boldsymbol{\mathcal{E}}_{\boldsymbol{\mathcal{F}}} = \{F_{yz}^{1,2} \to Y, F_{yz}^{1,2} \to Z, , F_y^{1,2} \to Y, , F_z^{1,2} \to Z\}$
6. $\boldsymbol{\mathcal{E}}_{\boldsymbol{\mathcal{S}}} = \{\}$

Thus no additional S-nodes beyond the one in Figure 4(a) are added even though no observational data is provided in domain Π^2. F-nodes are added representing each pair of distributions and their associated interventions resulting in the augmented selection diagram shown in Figure 4(d). ■

In words, the set of S-node edges are increased to a superset that comprises the smallest symmetric difference among the different pairs of interventions between domain Π^i and Π^j. When observational data is given in both Π^i and Π^j, then there exists $\{\}^i \in \boldsymbol{\Psi}^i$ and $\{\}^j \in \boldsymbol{\Psi}^j$. As a result, the set $\mathbf{D}_{i,j} = \bigcap_{\forall m,l} (\boldsymbol{\Psi}_m^i \Delta \boldsymbol{\Psi}_n^j)$ is simply the original nodes with S-nodes for domain Π^i, Π^j. In Section 4.2.1, we will provide additional remarks when observational data is missing in some domains. The significance of this construction follows from the next proposition where separation statements in the MD-Markov definition are formally tied to ones in the augmented selection diagram. This means that there is no need to perform any graphical manipulations in order to evaluate the CI and ICs described in Def.3. We are now ready to state our graphical characterization, which links separation statements in the augmented selection diagram to the MD-Markov property CI and IC conditions. In practice, the next result demonstrates how to perform comparisons of distributions to determine CI and IC constraints in an efficient and economical manner using a purely graphical characterization of the MD-Markov property.

Proposition 6 (Graphical MD-Markov Characterization) *Consider the Multi-domain setup. Let $Aug_{\boldsymbol{\Psi}}(G_S) = (\mathbf{V} \cup \mathbf{L} \cup \mathbf{S} \cup \boldsymbol{\mathcal{F}} \cup \boldsymbol{\mathcal{S}}, \mathbf{E} \cup \boldsymbol{\mathcal{E}}_{\boldsymbol{\mathcal{F}}} \cup \boldsymbol{\mathcal{E}}_{\boldsymbol{\mathcal{S}}})$ be the augmented selection diagram of G_S with respect to $\boldsymbol{\Psi}^{\Pi}$, where $\boldsymbol{\mathcal{S}}$ and $\boldsymbol{\mathcal{F}}$ are defined as they are in Definition 5.*

Let $\mathbf{K}_l^{i,j} = F_l^{i,j} \cup \mathbf{S}^{i,j}$ be the union of the set of nodes adjacent to the F-node $F_l^{i,j}$ and the set of S-nodes for domains Π^i, Π^j. The following equivalence relations hold for disjoint $\mathbf{Y}, \mathbf{Z}, \mathbf{W} \subseteq \mathbf{V}$, where $\mathbf{W}_l = \mathbf{W} \cap \mathbf{K}_l$ and $\mathbf{R} = \mathbf{K}_l \backslash \mathbf{W}_l$:

$$(\mathbf{Y} \perp\!\!\!\perp \mathbf{Z}|\mathbf{W})_{G_S} \iff (\mathbf{Y} \perp\!\!\!\perp \mathbf{Z}|\mathbf{W}, \mathcal{F})_{Aug_{\boldsymbol{\Psi}}(G_S)} \tag{2}$$

$$(\mathbf{Y} \perp\!\!\!\perp \mathbf{K}_l^{i,j}|\mathbf{W}\backslash\mathbf{W}_l)_{G_{S_{\underline{\mathbf{W}_l},\overline{\mathbf{R}(\mathbf{W})}}}} \iff (\mathbf{Y} \perp\!\!\!\perp \{F_i^{j,k}, \mathbf{S}^{j,k}\}|\mathbf{W}, F_{[k]\backslash\{i\}})_{Aug_{\boldsymbol{\Psi}}(G_S)} \tag{3}$$

■

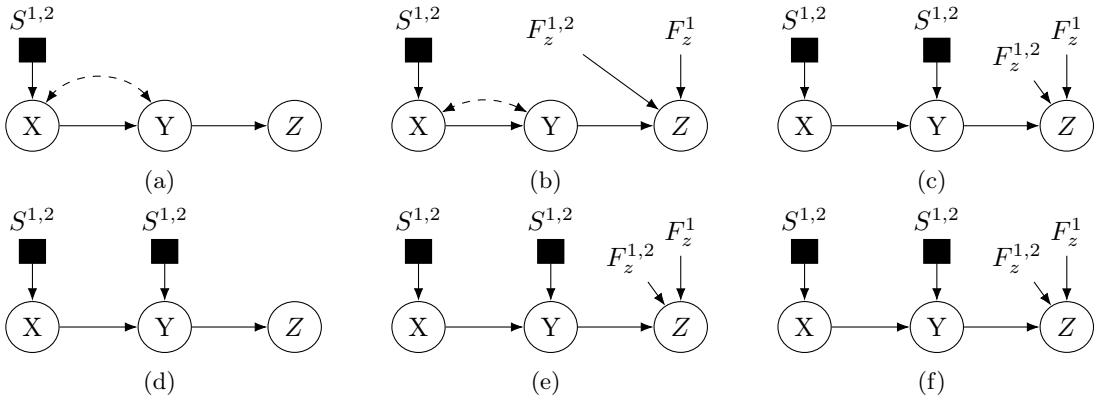


Figure 5: **MD-Markov equivalence from the augmented selection diagrams and MD-MAGs** - Selection diagrams across domains Π^1 and Π^2 (a,d), their augmented selection diagrams given interventions $\mathbf{\Psi} = \langle\{\}^1, \{Z\}^1, \{\}^2\rangle$ (b,e) and the MD-MAG EC of the selection diagrams (c,f).

Proposition 6 provides a way of characterizing the invariances of the MD-Markov Property without altering the graph. Instead, the augmented selection diagram constructed via Def. 5 will encode the relevant invariances. The result in the proposition is illustrated in the following example.

Example 14 (Testing single-domain invariances) *Consider the selection diagram in Fig. 5(a; top row) with intervention targets* $\mathbf{\Psi} = \langle\{\}^1, \{Z\}^1, \{\}^2\rangle$. *By Prop. 6, we can evaluate the MD-Markov property in the corresponding augmented diagram in Fig. 5(b). For example,* $(Y \perp\!\!\!\perp Z)_{G_{S\overline{Z}}}$ *can be tested by* $(Y \perp\!\!\!\perp F_z^1 | F_z^{1,2})_{Aug_\Psi(G_S)}$ *so as to determine that the invariance* $P^1_{\{\}}(Y) = P^1_Z(Y)$ *holds.* ■

Example 15 (Testing multi-domain invariances) *Consider the selection diagram in Fig. 5(a; top row) with intervention targets* $\mathbf{\Psi} = \langle\{\}^1, \{Z\}^1, \{\}^2\rangle$ *resulting in the augmented selection diagram shown in (b). We can test if cross-domain distributional invariances should hold for* $P^1_Z(Y|X,Z)$ *vs* $P^2_{\{\}}(Y|X,Z)$. *The MD-Markov property would inspect the separation on a modified graph* $(Y \not\perp\!\!\!\perp Z, S^{1,2}|X)_{G_{S_{\underline{Z},\overline{X}}}}$. *The graphical characterization would inspect the separation on the augmented selection diagram* $(Y \not\perp\!\!\!\perp F_z^{1,2}, S^{1,2}|X,Z,F_z^1)_{Aug_\Psi(G_S)}$. *Since the separation does not hold, the invariance* $P^1_Z(Y|X,Z) = P^2(Y|X,Z)$ *is not required.* ■

The previous two examples illustrate how one can leverage the augmented selection diagram to test CI and IC constraints defined by the MD-Markov property using separation in the graph. However, the graphical structure that encodes these constraints are not necessarily unique. Maximal Ancestral Graphs (MAGs) provide a compact and convenient representation for constraints present in an EC of causal graphs, see also [61, p. 6] [68]. Similarly, we can construct a MAG of the Augmented Selection Diagram. The invariances represented can typically be encoded in different causal graphs. Thus an EC of augmented selection diagrams can represent all the constraints given. This is formalized through an *MD-MAG*.

Definition 7 (MD-MAG) *Given a selection diagram* G_S *and a set of intervention targets* $\mathbf{\Psi}$, *an MD-MAG is the MAG constructed from* $Aug_\Psi(G_S)$. *That is* $MAG(Aug_\Psi(G_S))$. ■

The MD-MAG is a MAG that is constructed from the augmented selection diagram.

Example 16 *Consider the selection diagram in Figure 5(a) and let* $\mathbf{\Psi} = \langle\{\}^1, \{Z\}^1, \{\}^2\rangle$ *be the corresponding intervention tuple. The corresponding augmented selection diagram* $Aug_\Psi(G_S)$ *is shown in Fig. 5(b). Finally, the corresponding MD-MAG is* $MAG(Aug_\Psi(G_S))$ *shown in Fig. 5(c).*

The bottom row of Figure 5 also shows a similar construction from selection diagram, to augmented graph to the corresponding MD-MAG. ■

Finally, putting these results together, we derive a graphical characterization for two selection diagrams with corresponding tuple of intervention target sets to be MD-Markov equivalent.

Theorem 8 (Graphical MD-Markov Equivalence) *Let $\mathbf{\Pi}$ and $\mathcal{K}$ denote fixed sets of domains and indices of known intervention targets, respectively. Given selection diagrams G_S, G'_S defined over $\mathbf{V} \cup \mathbf{S}$ and the corresponding intervention targets $\mathbf{\Psi}, \mathbf{\Psi'}$, the pairs $\langle G_S, \mathbf{\Psi} \rangle$ and $\langle G'_S, \mathbf{\Psi'} \rangle$ are MD-Markov equivalent if and only if for $M = MAG(Aug_{\mathbf{\Psi}}(G_S))$ and $M' = MAG(Aug_{\mathbf{\Psi'}}(G'_S))$:*[12]

1. *M and M' have the same skeleton;*
2. *M and M' have the same unshielded colliders; and,*
3. *If a path p is a discriminating path for a node Y in both M and M', then Y is a collider on the path in one graph if and only if it is a collider on the path in the other.* ■

Theorem 8 states that the pairs $\langle G_S, \mathbf{\Psi} \rangle$ and $\langle G'_S, \mathbf{\Psi'} \rangle$ are MD-Markov equivalent if their corresponding MD-MAGs satisfy the corresponding three conditions, as illustrated in the example below.

Example 17 (MD-Markov Equivalent) *Consider the two selection diagrams given in Figure 5(a), G_S, G'_S. They have the intervention tuples $\mathbf{\Psi} = \langle \{\}^1, \{Z\}^1, \{\}^2 \rangle$ and $\mathbf{\Psi'} = \langle \{\}^1, \{Z\}^1, \{\}^2 \rangle$, respectively. The corresponding MD-MAGs, M_1, M_2 shown in panel (c) have the same skeleton, unshielded colliders and colliders on discriminating paths. Therefore, the pairs $\langle G_S, \mathbf{\Psi} \rangle$ and $\langle G'_S, \mathbf{\Psi'} \rangle$ are MD-Markov equivalent according to Theorem 8.* ■

Example 18 (Not MD-Markov Equivalent) *Consider the intervention sets for G_S and G'_S are the same as in Example 17. G_S is the selection diagram in Figure 5(a; top row), while G'_S is the selection diagram in Figure 5(a; bottom row) **without** the $S^{1,2} \to Y$ edge. In this case, the augmented selection diagram and the MD-MAG will also not have that edge. Therefore, the MD-MAG of this selection diagram has a different skeleton compared to Figure5(c; top row) and thus the pairs $\langle G_S, \mathbf{\Psi} \rangle$ and $\langle G'_S, \mathbf{\Psi'} \rangle$ are not MD-Markov equivalent.* ■

Given this characterization, we can devise an algorithm to learn the corresponding equivalence class of a true, underlying selection diagram. First, we will analyze the MD-Markov property and its characterization and learning in Markovian SCMs (i.e. no unobserved confounders).

3 Multi-domain Markov Characterization and Learning Without Unobserved Confounding

There is a simplified characterization and learning algorithm when one assumes there are no unobserved confounders (UC) present (i.e. also known as the causal sufficiency assumption). The core idea is that there are no more bidirected edges in the graph, resulting in the absence of inducing paths.

Corollary 9 (Markovian MD-Markov Equivalence) *Given Markovian selection diagrams, $G_{S_1} = (\mathbf{V} \cup \mathbf{S}, \mathbf{E_1} \cup \mathbf{E_S})$ and $G_{S_2} = (\mathbf{V} \cup \mathbf{S}, \mathbf{E_2} \cup \mathbf{E_S})$ and corresponding interventional targets $\mathbf{\Psi}_1, \mathbf{\Psi}_2$, the pairs $\langle G_{S_1}, \mathbf{\Psi}_1 \rangle$ and $\langle G_{S_2}, \mathbf{\Psi}_2 \rangle$ are MD-Markov equivalent if and only if $Aug_{\Psi_1}(G_{S_1})$ and $Aug_{\Psi_2}(G_{S_2})$ have (1) the same skeleton and (2) the same unshielded colliders.* ■

Next, we will summarize how the MD-Markov property is required even in the Markovian setting and compare against existing literature while highlighting the distinction of domain and intervention.

12. We assume that the symmetrical difference sets are indexed for both diagrams in the same pattern such that the correspondence between F-nodes and S-nodes is the same in M and M'.

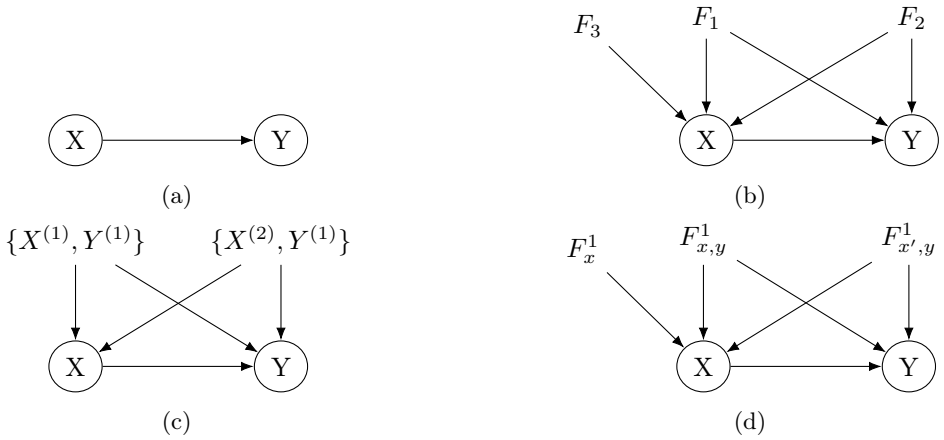


Figure 6: **Single-domain characterizations with interventional data without unobserved confounding** - Causal graph G (a) with intervention set $\mathbf{\Psi} = \langle\{\}^1, \{X^{(1)}, Y^{(1)}\}, \{X^{(2)}, Y^{(1)}\}\rangle$. The augmented interventional causal graph $Aug_\Psi(G)$ from [7] (b), the $\mathcal{I}$-DAG from [28] and the augmented selection diagram (d).

3.1 Characterization

Existing works deal with characterization given data arising from a Markovian SCM. For instance, [28] proposes the $\mathcal{I}$-MEC, which can be viewed as an EC given observational and/or interventional data.

First, we will introduce some related characterizations, which may be tempting to use to characterize differences. The works of [28] propose an EC of causal graphs without unobserved confounders from interventional data with known-targets. They use interventional nodes, which are analogous to F-nodes, to graphically characterize causal graphs that are interventionally equivalent (i.e., I-Markov equivalent). Another work that generalizes, but still solely for single-domain interventional data is [7].

Single-domain First, consider a single domain where different distributions consisting of observational and/or interventional data are given. The MD-Markov property and the MD-Markov characterization simplify to the Ψ-Markov property and the Ψ-Markov characterization given in [7].

Example 19 *Consider the single-domain selection diagram in Figure 6(a) with interventions* $\mathbf{\Psi} = \langle\{\}^1, \{X^{(1)}, Y^{(1)}\}, \{X^{(2)}, Y^{(1)}\}\rangle$.

The graphical characterization of [28] does not encode different mechanisms among interventions and appends to the graph one interventional node per interventional target and adds directed edges from the interventional nodes to the corresponding targets. The constructed graph is referred to as the $\mathcal{I}$-DAG, shown in Figure 6(c).

The graphical characterization of [7] does encode different mechanisms among interventions and constructs an F-node per pair of interventional distributions, including the observational distribution and points the nodes to the symmetric difference (similar to ours). In this case, F_3 maps to $\{X^{(1)}, Y^{(1)}\}\Delta\{X^{(2)}, Y^{(1)} = \{X\}$. This results in the augmented interventional diagram, $Aug_\Psi(G)$, shown in Figure 6(b). The separation $F_3 \perp\!\!\!\perp Y|X$ holds in $Aug_\Psi(G)$, which corresponds to the invariance $P_{X^{(1)},Y^{(1)}}(Y|X) = P_{X^{(2)},Y^{(1)}}(Y|X)$. This is not characterized in the $\mathcal{I}$-DAG.

The MD-Markov characterization results in Figure 6(d). The resulting augmented selection diagram provides the same graphical structure as that of $Aug_\Psi(G)$ and all the invariances represented there are also here. ■

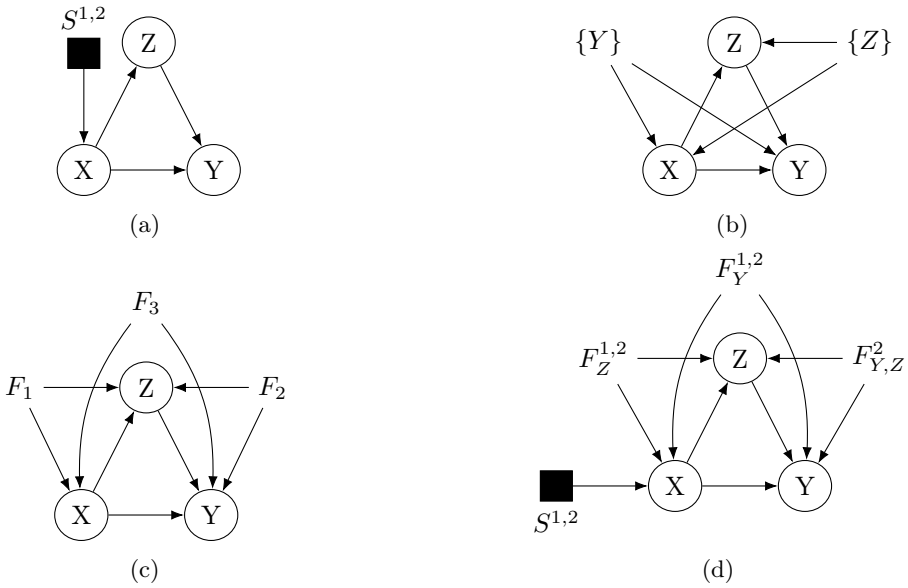


Figure 7: **Markovian SCM characterization comparisons** - Different characterizations given a ground truth selection diagram (a), and intervention targets $\mathbf{\Psi} = \langle \{\}^1, \{Y\}^2, \{Z\}^2 \rangle$: $\mathcal{I}$-MEC [28] (b), the Augmented Diagram from [7], and the Augmented Selection Diagram (d).

Multiple-domains Next, consider distributions arising from multiple domains. In this case, to our knowledge, there is no characterization that exists for multi-domain data that handles arbitrary mixtures of interventional data. However, a possible approach is collapsing the domain index and treating interventions and domains interchangeably. The next example demonstrates that this approach to characterization does not represent all possible invariances given a selection diagram and a mixture of multi-domain distributions.

Example 20 (Multi-domain vs purely interventional characterization) *Consider the selection diagram in Figure 7(a) over domains* $\mathbf{\Pi} = \{\Pi^1, \Pi^2\}$. *Let* $\mathbf{\Psi} = \langle \{\}^1, \{X\}^2, \{Z\}^2 \rangle$ *be the tuple of intervention targets, along with the corresponding distributions* $\mathbf{P}$.

The $\mathcal{I}$-DAG in Figure 7(b) is constructed by adding an extra node per intervention, $\{X\}$ and $\{Z\}$ [28].

Similarly, the augmented interventional causal graph $Aug_\Psi(G)$ is constructed in Figure 7(c) by ignoring the domain index and treating the domain as a separate intervention.

Finally, the augmented selection diagram is shown in Figure 7(d) via Definition 5.

The $\mathcal{I}$-DAG does not capture the invariance $P_Z^2(X|Z) = P_Y^2(X|Z)$, which is represented by the separation $F_{Y,Z}^2 \perp\!\!\!\perp X|Z$. This invariance is represented in the augmented diagrams of (c) and (d). However, by conflating interventions and domains, the augmented diagram in (c) does not represent the invariance between domain Π^1 and Π^2, e.g. $P^1(Z|X) = P^2(Z|X)$. ■

This example demonstrates that the MD-Markov property and its graphical characterization represents more invariances than prior work that ignore the domain index. Moreover, the additional information is useful to distinguish because one can clearly see the invariances present between domains 1 and 2. This is very useful when for example, domains 1 is humans and domain 2 is bonobos. These invariances can be leveraged in the context of transportability for instance [44].

Algorithm 1 MD-PC: Algorithm for Learning a MD-PDAG - *SepSet* the separating sets, $\mathbf{S}$ is the S-node set, $\mathcal{F}^{\Pi}$ the F-node set, and σ maps each pair of distributions to a pair of domains.

Input: Tuple of distributions $\mathbf{P}^{\Pi} = \langle P_1^1, ..., P_m^N \rangle$, intervention targets $\mathbf{\Psi}^{\Pi}$, and indices of known-targets $\mathcal{K}$.
Output: MD-PDAG, $\mathcal{P}$
1: $\mathbf{S}, \mathcal{F} \leftarrow \emptyset, k \leftarrow 0, \sigma : \mathbb{N} \rightarrow \mathbb{N} \times \mathbb{N}$
2: $(\mathbf{S}, \mathcal{F}, , \sigma) \leftarrow$ CreateAugmentedNodes$(\mathbf{\Psi}^{\Pi}, V, \mathcal{K})$ (see Alg. F.3)
3: **Phase I: Learn skeleton**
4: **for** all pairs $X, Y \in \mathbf{V} \cup \mathcal{F} \cup \mathbf{S}$ **do**
5: $\quad SepSet(X, Y), SepFlag \leftarrow$ GeneralizedDoConstraints$(X, Y, \mathcal{F}, \mathbf{S}, \sigma, \mathbf{\Psi}^{\Pi}, \mathcal{K}, \mathbf{V})$ (see Alg. F.5)
6: $\quad$ **if** SepFlag = True **then**
7: $\quad\quad$ Remove edge between X and Y
8: **Phase IIa: Orient unshielded colliders**
9: For every unshielded triple $\langle X, Y, Z \rangle$ in $\mathcal{P}$ orient it as a collider iff $Z \notin SepSet(X, Y)$
10: **Phase IIb: Apply logical orientation rules**
11: R1-4: Apply 4 PC rules from [6] and following two rules until none apply.
12: Rule 5': For $F_k^{i,j} \in \mathcal{F}^{\Pi}$ and for $S^{i,j} \in \mathbf{S}$, orient adjacent edges out of $F_k^{i,j}$ and $S^{i,j}$.

3.2 Learning

We present a sound and complete algorithm to learn an equivalence class of selection diagrams under causal sufficiency from multi-domain data. In this setting, instead of dealing with MAGs, one deals with PDAGs.

Definition 10 (MD-PDAG) *Given a selection diagram with no latents, G_S and interventional targets, $\mathbf{\Psi}$ defined over domains $\mathbf{\Pi}$, let $G = Aug_{\Psi}(G_S)$ and let $[G]$ be the set of augmented selection diagrams corresponding to all the pairs $\langle G'_S, \mathbf{\Psi}' \rangle$ that are MD-Markov equivalent to $\langle G_S, \mathbf{\Psi} \rangle$. The MD-PDAG for $\langle G_S, \mathbf{\Psi} \rangle$ denoted as $\mathcal{P}$ is a graph such that:*

1. $\mathcal{P}$ has the same adjacencies as M and any member of $[G_S]$ does; and

2. every non-circle mark (tail or arrowhead) in $\mathcal{P}$ is an invariant mark in $[G_S]$. ■

A MD-PDAG differs from a MD-PAG in that there are no bidirected paths possible in the underlying DAG. Therefore, any orientation of $\circ\!\!-\!\!\circ$ will result in either $\leftarrow$, or $\rightarrow$. When there is a lack of sufficient information, the $\circ\!\!-\!\!\circ$ can be represented as an undirected edge indicating an uncertainty of whether $\leftarrow$, or $\rightarrow$ is the underlying causal direction. The MD-PDAG will be the EC target of the algorithm, MD-PC algorithm. The MD-PC algorithm is similar to the PC algorithm where one applies the three rules from Meek [6].

Alg. 1 proceeds by first adding augmented nodes in L2 of the algorithm, which is a function of the different distributions provided. Then Phase I proceeds to learn a skeleton of the selection diagram by testing the CI and IC distributional constraints from the MD-Markov property. Then Phase IIa and IIb proceeds as the PC algorithm, and orients edges accordingly to logical rules. An additional Rule 5' is added, which orients all edges out of F-nodes and S-nodes. The next example illustrates how the MD-PC algorithm is different from algorithms where one may treat interventions and domains interchangably.

Example 21 (Causal discovery over multiple domains without latent confounding) *Consider the selection diagram and setting in Figure 8(a), with interventions $\mathbf{\Psi} = \langle \{\}^1, \{Y\}^2 \rangle$ and corresponding distributions $\mathbf{P}$. This is the common bivariate causal setting, where one is interested in determining the causal relationship among two variables. To ground the example, consider a*

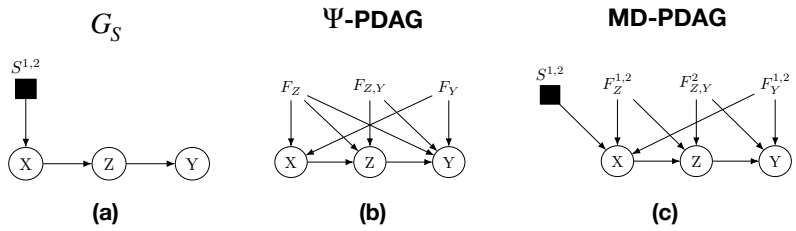


Figure 8: Comparing multi-domain causal discovery algorithms under the setting of causal sufficiency - The ground truth selection diagram (a) is given over domains Π^1, Π^2 with intervention targets $\mathbf{\Psi} = \langle \{\}^1, \{Y\}^2, \{Z\}^2 \rangle$ with all interventions being known-target, $\mathcal{K} = [1, 1, 1]$. The Ψ-PDAG learned from the Ψ-PC algorithm [7] (b) and the MD-PDAG (c) learned from the MD-PC algorithm.

setting where one may have access to observational data in one hospital setting Π^1, but only access to clinical trial data (i.e. interventional) in another hospital Π^2. Furthermore in hospital 2, there is no observational data. One would like to determine if smoking (X) causes a difference in cancer outcomes (Y) with also tar measurements (Z) [13]. In hospital 2, the clinical trial added a new patient monitoring machine for the duration of the trial to evaluate the efficacy of a new machine. Besides the causal relationship between varaibles, one may also be interested in determining the invariances between these two hospitals (i.e. domains).

Treating interventions and domains interchangeably *Applying the Ψ-PC algorithm introduced in [7], one obtains the learned Ψ-PDAG in Figure 8(b). Though the causal relationships among the observed variables are fully oriented, there is no information that infers across-domain invariance between Π^1 and Π^2.*

Treating domains and interventions separately *Applying the MD-PC algorithm in Alg. 1, one obtains the MD-PDAG in Figure 8(c). The causal relationships are fully learned among the variables with the MD-PC. In addition, the invariance between hospital Π^1 and Π^2 is learned - indicated by the lack of S-node edges for Z and Y.*

By comparing the distributions $P_Z^2(.)$, $P_Y^2(.)$ and $P_{\{\}}^1(.)$, one will observe that $P_Z^2(X|\mathbf{W}) \neq P_{\{\}}^1(X|\mathbf{W})$ and $P_Y^2(X|\mathbf{W}) \neq P_{\{\}}^1(X|\mathbf{W})$ for any conditioning set $\mathbf{W} \in \mathbf{V} \backslash \{X\}$. However, we know the intervention targets are $\{Z\}^2$ and $\{Y\}^2$ for $P_Z^2(.)$ and $P_Y^2(.)$ respectively, and can leverage this information. The conditional distribution of $X|\mathbf{W}$ is not invariant when comparing against the observational distribution of domain Π^1, $P_{\{\}}^1(X|\mathbf{W})$ then we know the change in distribution is due entirely to the domain differences between Π^1 and Π^2 encoded by the S-node, $S^{1,2}$, that is $S^{1,2} \to X$. In addition, since $P_Z^2(Y|Z) = P_{\{\}}^1(Y|Z)$ and $P_Y^2(Z|X) = P_{\{\}}^1(Z|X)$, it is possible to determine that $S^{1,2} \not\to Y$ and $S^{1,2} \not\to Z$. ■

The MD-PC algorithm is complete in the sense that it learns the most about the underlying selection diagram one can possibly learn without further assumptions.

Theorem 11 (MD-PC Completeness) *Let the tuple of distributions $\mathbf{P}$ be generated by an unknown pair $\langle G_S, \mathbf{\Psi} \rangle$ over domains $\mathbf{\Pi}$, then MD-PC is complete, i.e. $\mathcal{P}$ contains all common edge marks in the MD-Markov equivalence class.* ■

4 Non-Markovian Multi-Domain Markov Equivalence

In this section, the MD Markov characterization will be discussed in a non-Markovian setting in various settings that are possible for multi-domain data.

13. For simplicity, in this example, we are ignoring the possibility of latent confounding.

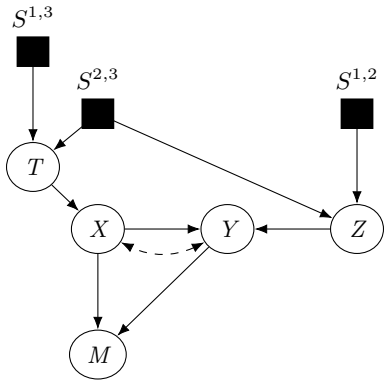
4.1 Multi-domain observational data

S-nodes introduced through the lens of selection diagrams can be thought of as augmentations of the causal graph to represent different domains and changes in distributions that may occur [3, 16, 62, 69]. In a sense, S-nodes are graphically similar to F-nodes, which have been commonly used to represent interventions [3, 7, 62]. F-nodes are augmented nodes where each one is a parent to (each element in) a symmetrical difference set, and they are used to represent – graphically – invariances between interventional distributions. The significance of these F-nodes will be further emphasized in Section 4.2 (see Def. 5 and Proposition 6). Despite the similarity between F-nodes and S-nodes, it is worthy to distinguish S-nodes since most challenges related to transportability, rely on knowing the S-node structure [16, 18].

Before deriving the graphical characterization for the MD-Markov equivalence class, we consider the setting where there is only observational data across different domains.

Definition 12 (Corresponding Tuple of Intervention Sets) *Consider the Multi-domain setup. For a selection diagram G_S over N domains. $\langle \mathbf{V}_{S^{i,j}} \rangle = \langle \mathbf{V}_{S^{1,2}}, \mathbf{V}_{S^{1,3}}, ..., \mathbf{V}_{S^{N-1,N}}, \rangle \forall i \neq j \in [N]$ is an ordered tuple of the children of each S-node. A corresponding tuple of intervention sets for $\mathbf{V_S}$ is $\langle \mathbf{\Psi}^1, \mathbf{\Psi}^2, ..., \mathbf{\Psi}^N \rangle$, such that $\mathbf{\Psi}^i \Delta \mathbf{\Psi}^j = \mathbf{V}_{S^{i,j}}$ for all $i \neq j$.* ■

Example 22 *Consider the selection diagram shown in Figure 3(a) with an additional two S-nodes: $(S^{1,3} \to T)$ and $(Z \leftarrow S^{2,3} \to T)$ representing three domains, Π^1, Π^2, Π^3 (shown below for convenience). Denote the selection diagram G_S.*



Thus, there are three S-nodes pointing to $\{T\}$, $\{T, Z\}$, and $\{Z\}$. A corresponding intervention set for G_S is $\langle \mathbf{\Psi}^1, \mathbf{\Psi}^2, \mathbf{\Psi}^3 \rangle = \langle \{Z, T\}, \{T\}, \{Z\} \rangle$. The symmetric difference of each corresponding intervention set results in the children of the S-nodes:

1. $\{Z, T\} \Delta \{Z\} = \{T\}$ *for* $S^{1,3}$
2. $\{Z, T\} \Delta \{T\} = \{Z\}$ *for* $S^{1,2}$
3. $\{T\} \Delta \{Z\} = \{T, Z\}$ *for* $S^{2,3}$

Thus, the corresponding intervention tuple provides a set of interventions that represent a similar change in distribution as the change in mechanism due to the domain change (i.e. the effect of the S-node edge). ■

The corresponding tuple of intervention sets simply represents the different mechanisms due to a change of domain as interventions instead. The following theorem uses this to present a duality between characterization of interventions and multiple domains when only observational data across domains is given.

Theorem 13 (Equivalence of Ψ and MD-Markov characterization) *Consider the Multi-domain setup and let G_S be a selection diagram among N domains, and G be the corresponding causal diagram without S-nodes. Let $\mathbf{\Psi}^{\Pi} = \langle \{\}^1, ..., \{\}^N \rangle$ for each of the N domains, and $\mathcal{K} = [1, 1, ..., 1]$, such that only observational data is available. Let $\mathbf{I}_S$ be the corresponding intervention set for $\mathbf{V_S}$. Let $\mathbf{P}^{\Pi}$ be an arbitrary set of distributions generated by the corresponding interventions. $\mathbf{P}^{\Pi}$ satisfies the MD-Markov property with respect to $\langle G_S, \mathbf{\Psi} \rangle$ if and only if it satisfies the Ψ-Markov property with respect to $\langle G, \mathbf{I}_S \rangle$.*[14] ■

Observations collected from multiple domains is equivalent to collecting distributions with unknown-target interventions. This result coincides with other works that treat different domains and interventions interchangably [10, 13]. In this setting, S-nodes have a correspondence to the augmented graph's F-nodes in [7]. In some sense, the change-in-domain can be viewed as "Nature's" intervention on the causal system. However, this simplification is not warranted when considering interventions that occur in different domains, as elaborated in the sequel.

Given this duality between interventions and domains when only considering observational data across domains, we state a core assumption when considering single-domain interventional, or multi-domain observational data.

Assumption 14 (Multi-domain & interventional exchangeability (MDIX)) *Let $\mathbf{P} = \{P^1(\mathbf{V}), P^2(\mathbf{V}), ..., P^N(\mathbf{V})\}$ be a set of N distributions over $\Pi^1, \Pi^2, ..., \Pi^N$ domains and $\mathbf{P}' = \{P'_1(\mathbf{V}), P'_2(\mathbf{V}), ..., P'_N(\mathbf{V})\}$ be a set of N interventional distributions within a single domain Π'. The multi-domain and interventional exchangeability (MDIX) assumption states that we assume these two settings are exchangeable in the sense that applying a causal discovery algorithm, or characterization of the implied invariances will produce the same result.* ■

This assumption states one is treating interventions as domain-changes and vice-versa (i.e. the yellow section in Table 3 are seen as equivalent). In the context of Theorem 13, this turns out to be fine since the distributional invariances implied are the same and the resulting characterization and learning algorithm are the same up to re-labeling of an intervention and domain-change. In fact, this assumption is commonly implicitly assumed in many prior works, as discussed in Section 5. Stating this assumption explicitly is necessary as we will see in the following sections because when one considers the general setting of interventional distributions across different domains, the MDIX assumption is not valid. We next discuss the general MD setting with arbitrary mixtures of interventional and observational data in multiple domains.

4.2 Arbitrary mixtures of multi-domain observational and interventional data

Next, we analyze the general setting with possibly arbitrary mixtures of multi-domain observational and interventional data. In previous sections, we assumed observational data was provided. In this next section, we describe some of the nuances when observational data is not available in every domain.

4.2.1 MD-Markov Property Without Observational Data

This section elaborates on the intricacies of having observational data vs not. This can occur, for example, when there are multiple clinical trial data occurring in different hospitals. Each hospital is a different environment, and the clinical trials are interventions. Observational data may be missing, or unable to be used due to HIPAA violations. Consider the following example, which provides insight into the MD-Markov property in this setting.

14. Due to space constraints, all the proofs are provided in the Appendix F.1

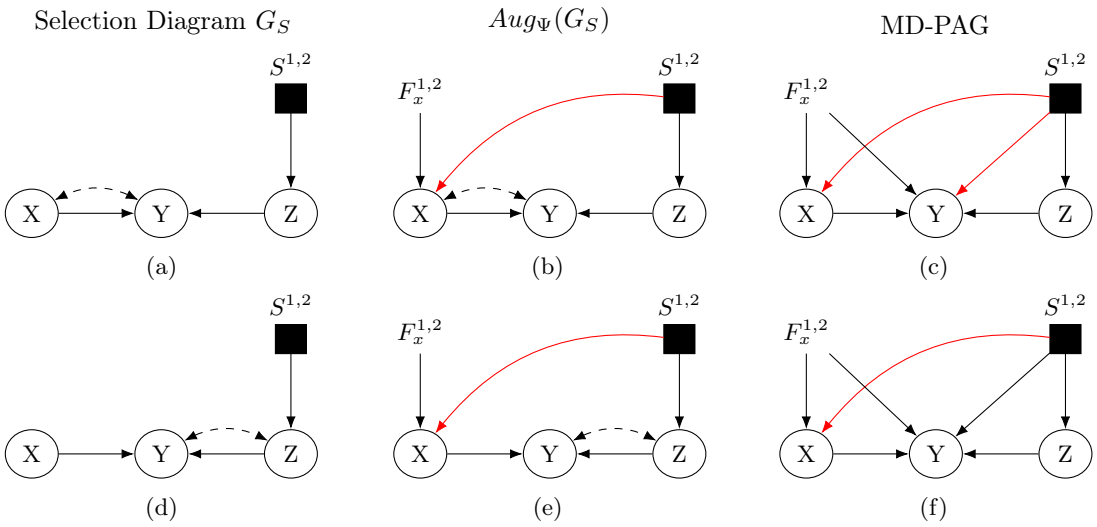


Figure 9: Example diagrams without observational data defined for interventions $\mathbf{\Psi} = \langle\{\}^1, \{X\}^2\rangle$ over domains Π^1, Π^2. Each row has selection diagrams, G_S (a,d), their corresponding augmented diagrams, $Aug_\Psi(G_S)$ (b,e) and the resulting MAG, $MAG(Aug_\Psi(G_S))$ (c,f). The red arrows from the S-node $S^{1,2}$ indicate additional testable edges if observational data was given in domain Π^2.

Example 23 (MD-Markov Property Without Obs. Data) *Consider the two selection diagrams in Figure 9(a), with interventions $\mathbf{\Psi} = \langle\{\}^1, \{X\}^2\rangle$ over domains Π^1, Π^2 and their corresponding distributions $\mathbf{P}$. It is clear that these two selection diagrams are different because in the top row, X and Y have an unobserved confounder, whereas in the bottom row, Y and Z have an unobserved confounder.*

The given distributions do not contain observations in Π^2 and the resulting augmented selection diagram is given in panel (b). In both diagrams, from Def. 5, the added "S-node edges" $\boldsymbol{\mathcal{E}}_S = \{(S^{1,2}, X)\}$ are added because $\mathbf{K}_{S^{1,2}} = (\{\}^1 \Delta \{X\}^2) \cup \{Z\} = \{X, Z\}$. This in turn results in the MAG of both diagrams in Figure 9(c). In fact, this is due to the lack of any invariances of the form $P^1_{\{\}}(X) = P^2_X(X)$ given the distributions we have. This can be seen since comparing observational data in domain Π^1 to the interventional data in Π^2, it is clear that $P^1_{\{\}}(X) \neq P^2_X(X)$, which maps to the MD-Markov property condition where $X \not\perp\!\!\!\perp S^{1,2}$. Proposition 6 in turn implies that the two selection diagrams are indistinguishable because they are within the same EC.

This example illustrates that the MD-Markov property and its resulting graphical characterization is general enough to handle the setting where observational data is not present in all domains. However, the next example illustrates the importance of observational data in discerning the invariances across domains.

Example 24 (MD-Markov Property With Obs. Data) *Consider the two selection diagrams in Figure 9(a), with interventions $\mathbf{\Psi} = \langle\{\}^1, \{\}^2, \{X\}^2\rangle$ over domains Π^1, Π^2 and their corresponding distributions $\mathbf{P}$. In the top row's selection diagram, X and Y have an unobserved confounder, whereas in the bottom row's selection diagram, Y and Z have an unobserved confounder.*

*In this instance, the given distributions contain observations in Π^2 and the resulting augmented selection diagram is the diagram in panel (b) **without** the red edge $S^{1,2} \to X$. In both diagrams, from Def. 5, the added "S-node edges" $\boldsymbol{\mathcal{E}}_S = \{\}$. In this example, the MAG of both diagrams are shown in panel (c) **without** the red edges. In this case, by the MD-Markov property, the separation $X \perp\!\!\!\perp S^{1,2}$ implies the invariance $P^1_{\{\}}(X) = P^2_{\{\}}(X)$. Proposition 6 in turn implies that the two selection diagrams from Figure 9(a) are distinguishable because they are not within the same EC.*

Algorithm 2 MD-FCI: Algorithm for Learning a MD-PAG - *SepSet* the separating sets, $\mathbf{S}$ is the S-node set, $\mathcal{F}^{\Pi}$ the F-node set, $\mathbf{H}$ maps each pair of known-targets symmetric diffs., and σ maps each pair of distributions to a pair of domains.

Input: Tuple of distributions $\mathbf{P}^{\Pi} = \langle P_1^1, ..., P_m^N \rangle$, vector of known intervention targets $\mathcal{K}$ and $\mathbf{\Psi}^{\Pi}$.
Output: MD-PAG, $\mathcal{P}$

$\mathbf{S}, \mathcal{F} \leftarrow \emptyset, k \leftarrow 0, \sigma : \mathbb{N} \rightarrow \mathbb{N} \times \mathbb{N}, \mathbf{H} \leftarrow \emptyset$
$(\mathbf{S}, \mathcal{F}, \mathbf{H}, \sigma) \leftarrow$ CreateAugmentedNodes$(\mathbf{\Psi}^{\Pi}, V)$ (see Alg. F.3)
Phase I: Learn skeleton
for all pairs $X, Y \in \mathbf{V} \cup \mathcal{F} \cup \mathbf{S}$ **do**
    $SepSet(X, Y), SepFlag \leftarrow$ GeneralizedDoConstraints$(X, Y, \mathcal{F}, \mathbf{S}, \sigma, \mathbf{\Psi}^{\Pi}, \mathcal{K}, \mathbf{V})$ (see Alg. F.5)
    if SepFlag = True **then**
        Remove edge between X and Y
Phase IIa: Orient unshielded colliders
For every unshielded triple $\langle X, Y, Z \rangle$ in $\mathcal{P}$ orient it as a collider iff $Z \notin SepSet(X, Y)$
Phase IIb: Apply logical orientation rules
R1-7: Apply 7 FCI rules from [20] and following two rules until none apply.
Rule 8': For $F_k^{i,j} \in \mathcal{F}^{\Pi}$ and for $S^{i,j} \in \mathbf{S}$, orient adjacent edges out of $F_k^{i,j}$ and $S^{i,j}$.
Rule 9': For $F_k^{i,j} \in \mathcal{F}^{\Pi}$ with $X \in H_k^{i,j}$, that is adjacent to a node $Y \notin H_k^{i,j}$, if $|H_k^{i,j}| = 1$, then orient $X \rightarrow Y$.

Remark 15 (The importance of observational data) *The previous example demonstrates that observational data in two domains Π^i, Π^j allows one to determine across-domain invariances. For example, if we are considering humans (Π^1) and bonobos (Π^2) in Figure 9(a; top row) and how a drug (X), cardiovascular health (Y) and diet (Z) are related. Drug effects and cardiovascular health may be confounded by an unobserved variable (e.g. age). Then observational data in humans and bonobos enables one to determine that the distributions of drug effects and cardiovascular health conditioned on diet in bonobos are transportable to humans. This is illustrated by being able to test that $P^1_{\{\}}(X) = P^2_{\{\}}(X)$ and $P^1_{\{\}}(Y|Z) = P^2_{\{\}}(Y|Z)$, which is implied by the separations $X \perp\!\!\!\perp S^{1,2}$ and $Y \perp\!\!\!\perp S^{1,2}|Z$, respectively. However, these invariances can not be tested with the provided dataset. But it could be tested if observational data was provided for bonobos.*

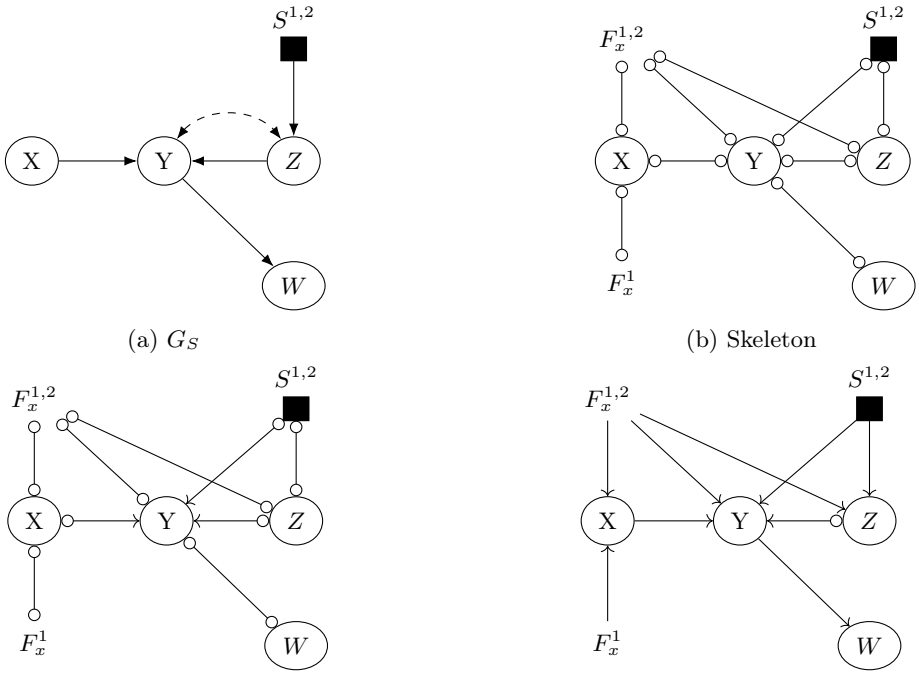
4.3 Causal Discovery From Multiple Domains

We investigate in this section how to learn an EC of selection diagrams from a mixture of observational and interventional data that is generated across multiple domains. The graphical characterization of MD-Markov equivalence in Theorem 8 and the significance of ancestral graphs (MAGs) in deriving this result motivate the following definition of MD-PAG.

Definition 16 (MD-PAG) *Consider the Multi-domain setup and let $M = MAG(Aug_{\mathbf{\Psi}}(G_S))$, and $[M]$ be the set of MD-MAGs corresponding to all the tuples $\langle G'_S, \mathbf{\Psi'}^{\Pi} \rangle$ that are MD-Markov equivalent to $\langle G_S, \mathbf{\Psi}^{\Pi} \rangle$. The MD-PAG for $\langle G_S, \mathbf{\Psi}^{\Pi} \rangle$, denoted $\mathcal{P}$ is a graph such that:*

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

The MD-PAG is a valid PAG by construction. Moreover, MD-PAGs generalize PAGs and Ψ-PAGs from the single-domain to the multiple-domain setting[7, 24]. There are two important aspects that have changed compared to the traditional PAG. Firstly, the extra S and F nodes and



(a) G_S (b) Skeleton

(c) After Orienting Unshielded Colliders (d) MD-PDAG

Figure 10: Example of MD-FCI applied with $\mathbf{\Psi} = \langle\{\}^1, \{X\}^1, \{\}^2\rangle$ and $\mathcal{K} = [1, 1, 1]$. The S-node representing domain-shift between domains 1 and 2 is the black square in (a).

their corresponding edges help graphically represent differences across domains and interventional distributions in this EC. Secondly, the characterizations are now not only tied to a graph G_S, but also to a set of interventions $\mathbf{\Psi}$ that give rise to the data distributions available. The PAG only considers a single observational distribution along with the graph. Similar to the PAG, the MD-PAG is the goal for causal discovery.

Next, a generalization of the standard faithfulness assumptions for multi-domain data is introduced that enables causal discovery [7, 70].

Definition 17 (MD-faithfulness) *Consider a selection diagram G_S over N domains. A tuple of distributions $\langle \mathbf{P}_\mathbf{I} \rangle_{\mathbf{I} \in \mathbf{\Psi}^\Pi} \in S^\Pi_\mathcal{K}(G_S, \mathbf{\Psi}^\Pi)$ is called MD-faithful to G_S if the converse of each of the MD-Markov conditions (Definition 3) holds.* ■

MD-faithfulness states that the CI and IC constraints presented in Def. 3 only arise due to the graphical conditions stated, and do not otherwise appear. For instance, if $P(X|Y, Z) = P(X|Y)$, then this implies that $(X \perp\!\!\!\perp Z|Y)_{G_S}$. MD-faithfulness taken together with the MD-Markov property, means that the CI and IC constraints are necessary and sufficient for a corresponding graphical separation. With this assumption, we are ready to propose our causal discovery algorithm that learns an equivalence class of selection diagrams in the general non-Markovian setting with possible latent confounding.

The new algorithm, called MD-FCI is shown in Alg. 2. The algorithm proceeds by first constructing the augmented graph using Alg. F.3, by adding S-nodes and F-nodes to represent every pair of domains and interventions. Then it uses hypothesis testing to learn invariances in the skeleton (Alg. F.4) and finally applies orientation rules (Alg. F.6). MD-FCI learns the skeleton by mapping pairs of distributions in $\mathbf{P}^\Pi$ to F-nodes, or S-nodes by testing for the distributional invariances discussed in Section 2.1. Def. 3 and Prop. 6 connect these invariances to graphical criterion, which allow us to

reconstruct the skeleton of the causal diagram. Interventional distributions across domains are used to learn F-node structure, and whereas observational distributions across domains are used to learn S-node structure. Besides the standard FCI rules that apply in the absence of selection bias, the algorithm also applies the following rules R8'-9'.

Rule 8' (Augmented Node Edges) - Edges out of F-nodes and S-nodes are oriented.

Rule 9' (Identifiable Inducing Paths) - If $F_k^{i,j} \in \mathcal{F}$ is adjacent to a $Y \notin H_k^{i,j}$ known-target node and the intervention target is known to be node X, $X \to Y$ can be oriented since the $F_k^{i,j} \to Y$ is only present due to an inducing path between X and Y.

In Figure 10, the different stages of the MD-FCI algorithm are shown. The selection diagram G_S Figure 10(a), encodes the causal structure across domains 1 and 2. Line 2 of Alg. 2 constructs the complete MD-PAG by adding all relevant F-nodes to the graph. The initial skeleton has an edge between every single node. Then, phase I of Alg. 2 proceeds to learn the skeleton of the MD-PAG, resulting in Figure 10(b). For every pair of nodes $X, Y \in \mathbf{V}$, one can leverage standard CI tests to determine if there is a separating set X and Y. If so, the edge (X, Y) is removed from the skeleton. This removal corresponds to the first condition of the MD-Markov Property. Without loss of generality, let $X \in \mathbf{S} \cup \mathcal{F}$, and $Y \in \mathbf{V}$, then here the second condition of the MD-Markov Property is tested via the graphical characterization in Proposition 6. One tests null hypotheses of the form $P_l^i(Y|W) = P_m^j(Y|W)$, where W is a hypothesized separator. These tests are known also as two-sample conditional tests [71–74]. If one failes to reject the null hypothesis, then it is implied that there is an invariance such that Y is d-separated from X given W. Finally, if both $X, Y \in \mathbf{S} \cup \boldsymbol{\mathcal{F}}$, then the edge (X, Y) is removed from the skeleton. After learning the skeleton, unshielded colliders are oriented in Phase IIa, resulting in Figure 10(c). Finally, orientation rules are applied in Phase IIb given the existing skeleton and separating sets learned in the earlier phases of the algorithm. This results in the MD-PAG shown in Figure 10(d).

Next, we prove the proposed MD-FCI algorithm is sound, and results in a valid MD-PAG.

Theorem 18 (MD-FCI Soundness) *Given $\mathcal{K}$, let $\mathbf{P}^{\mathbf{\Pi}}$ be generated by some unknown tuple $\langle G_S, \mathbf{\Psi}^{\mathbf{\Pi}} \rangle$ from domains $\mathbf{\Pi}$ with a corresponding selection diagram G_S and is MD-faithful to the selection diagram G_S. MD-FCI algorithm is sound (i.e. every adjacency and orientation in $\mathcal{P}_{\text{MD-FCI}}$, the MD-PAG learned by MD-FCI, is common for $MAG(Aug_{\Psi}(G_S))$).* ■

Known vs Unknown Targets We elaborate briefly on the usefulness of known and unknown targets in the learning of the EC. When considering Markovian SCMs, the known targets do not provide any additional information with respect to causal orientations. However, when considering non-Markovian SCMs with possible latent confounders, it is possible to leverage known-targets to orient inducing paths.

Next, we illustrate some subtleties between the MD-FCI and related algorithms that assume the MDIX assumption pooling observational and interventional distributions regardless of the domain. The example is motivated from biomedical sciences, where interventions are commonly performed in different domains and the goal is to leverage all datasets for learning. A group of scientists are trying to determine the causal structure of a set of proteins, but leverage data across the lab and hospital setting. Different experiments are run in each setting and combined into a single dataset [40].

Example 25 (MDIX Assumption May Result in Unsound Results) *Let G_S be a selection diagram as shown in Figure 11(a). Let $\mathbf{\Pi} = \langle \Pi^1, \Pi^2 \rangle$ be the set of domains representing the lab (Π^1) and the hospital (Π^2). These are a tuple of distributions $\mathbf{P} = \langle P_1^1, P_1^2 \rangle$ with intervention targets $\mathbf{\Psi}^{\mathbf{\Pi}} = \langle \{\}^1, \{Y\}^1, \{\}^2 \rangle$ and $\mathcal{K} = [1, 1, 1]$, where X represents some protein in the dataset.*

In this setting, let G_S be the true selection diagram as shown in Figure 11(a). Given the interventional and observational data, one may be tempted to use the $\boldsymbol{\mathcal{I}}$-FCI algorithm and simply pool the observational data, while ignoring the domain differences [3]. Still, this would learn the graph

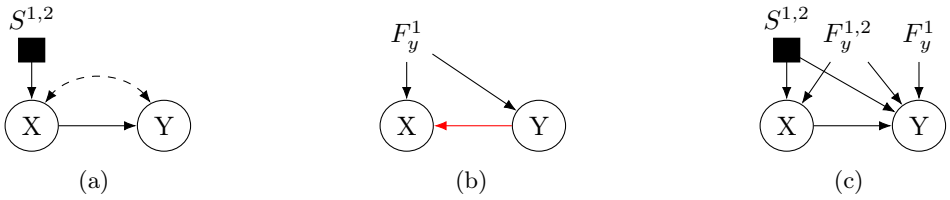


Figure 11: Causal graphs related to example 25 - The ground truth selection diagram (a). Causal discovery algorithm is applied given interventions $\mathbf{\Psi} = \langle\{\}^1, \{Y\}^2\rangle$, and known-target indices $\mathcal{K} = [1, 1]$. Applying $\mathcal{I}$-FCI without considering domain-changes under the MDIX assumption, one learns the equivalence class shown in (b). Applying the MD-FCI algorithm, one learns the MD-PAG in (c).

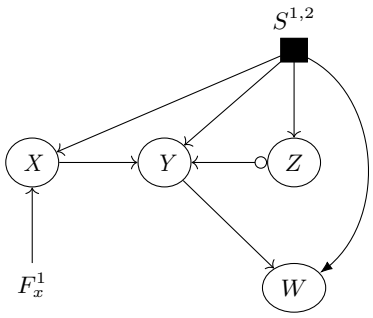
in Figure 11(b) with an incorrect orientation (shown as the red edge). This I-PAG only contains one F-node because there is only two distributions: i) the pooled observational data and ii) the data resulting from intervention on Y. Applying R9 of the $\mathcal{I}$-FCI algorithm incorrectly orients the edge $X \leftarrow Y$. Thus, R9 of the $\mathcal{I}$-FCI algorithm is not sound when the domains are ignored [3, 28].

Figure 11(c) contains what MD-FCI would recover. Intuitively, one should learn (c) instead of (b) because even though there is a change in distribution among X and Y, one cannot ascertain whether there is an inducing path from F_y^1 to X, or a change in distribution due to the domain. ■

The next few examples then illustrate when we relax the MDIX Assumption and proceed by applying the MD-FCI algorithm on multiple distributions arising from interventions and different domains. We begin with the single-domain setting.

Example 26 (Causal discovery in a single domain) *Consider the selection diagram shown in Figure 10(a), and the tuple of intervention sets $\mathbf{\Psi} = \langle\{\}^1, \{X\}^1\rangle$ with the corresponding distributions and known-target indices $\mathcal{K} = [1, 1]$. Our goal is to learn an equivalence class of selection diagrams over domains Π^1, Π^2.*

By applying the MD-FCI algorithm, we learn the resulting equivalence class shown below for convenience. Note that the S-node points to all variables in $\mathbf{V}$. In essence, we are unable to infer any invariances between domains Π^1 and Π^2 since we do not have any data from domain Π^2.

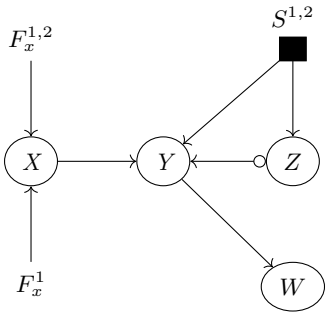


■

Example 27 (Causal discovery in multiple domains with observational data) *Consider the selection diagram shown in Figure 10(a), and the tuple of intervention sets $\mathbf{\Psi} = \langle\{\}^1, \{X\}^1, \{\}^2\rangle$ with the corresponding distributions and known-target indices $\mathcal{K} = [1, 1, 1]$. Our goal is to learn an equivalence class of selection diagrams over domains Π^1, Π^2.*

The output of the MD-FCI algorithm is the EC graph shown below. As we can see, with observational data in domain Π^2, we are able to refine our knowledge of what is invariant between domains

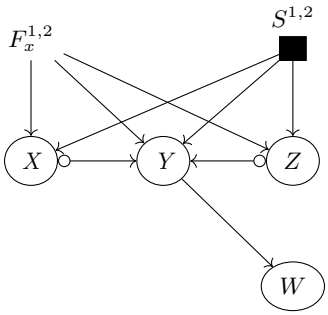
Π^1 *and* Π^2. *This implies for instance that the distribution of* X *is invariant in both domains, but for instance* Y *is not.*



■

Example 28 (Causal discovery in multiple domains without observational data) *Consider the selection diagram shown in Figure 10(a), and the tuple of intervention sets* $\mathbf{\Psi} = \langle \{X\}^1, \{\}^2 \rangle$ *with the corresponding distributions and known-target indices* $\mathcal{K} = [1, 1]$. *Our goal is to learn an equivalence class of selection diagrams over domains* Π^1, Π^2.

The output of the MD-FCI algorithm is the EC graph shown below. Even without observational data in domain Π^2, *we are still able to refine our knowledge of what is invariant between domains* Π^1 *and* Π^2. *This implies for instance that the distribution of* Z *is invariant in both domains, but for instance* Y *is not necessarily so. We also do not necessarily know that* X *is invariance across domains* Π^1 *and* Π^2. *This uncertainty is reflected in the learned EC that contains the S-node edge* $S^{1,2} \to X$.



■

These examples illustrate that the MDIX assumption is not always valid as sometimes

5 Comparisons with Previous Works

This paper extends the work from [54] in a number of important directions. Firstly, the assumption that observational data is present is dropped. This means that even when there are only experimental data present in domains, we can still leverage that domain to learn aspects of the selection diagram EC. For example, consider the setting where experimental data is collected in different labs studying how different genes affect stem-cell maturation. Each lab conducts a series of gene-knockout experiments and collects data on the resulting stem-cells. Each lab is considered a separate domain that has results in mechanism changes due to different operating protocols and equipment. The scientists would still like to leverage the combined experimental data to perform causal discovery. Even when observational data is available, one may not wish to use it for various reasons, such as selection bias, artifacts, missing data, or more.

Algorithm	Graphical Characterization	UC	Interv. $\mathcal{K}$	Interv. $\mathcal{U}$	Nonparametric	General Multi Domain
Single-Distribution						
[6, PC], [1, p. IC]	✓	x	x	x	✓	x
[2, FCI], [1, IC]	✓	✓	x	x	✓	x
Multi-Distribution						
[7, I-FCI]	✓	✓	✓	x	✓	x
[3, Ψ-FCI]	✓	✓	x	✓	✓	x
[28, IGSP]	✓	x	✓	x	x	x
[13, 50, ICP]	x	x	✓/x	✓/x	✓	x
[14, JCI]	✓/x	✓	✓/x	✓/x	✓	x
[11, 12, 51, NSC]	x	x	✓/x	✓/x	✓	x
[52, MDLS]	x	x	✓/x	✓/x	x	x
This work	✓	✓	✓	✓	✓	✓

Table 4: **A copy of Table 2 for convenience of the reader**

Secondly, we have provided an extensive introduction and discussion on the MD-Markov property, its characterizations and learning in the Markovian and non-Markovian settings. Third, we provide more extensive simulations, and a detailed comparison with other works.

Next, the similarities and differences among various Markov properties are highlighted in relation to the MD-Markov Property.

5.1 Comparing Markov Properties

The single-domain observational Markov property maps graphical d-separation to invariances in the decomposition of the joint probability distribution. The standard Markov property takes a DAG's d-separation statements and maps them to conditional independences. Compared to the MD-Markov property from Definition 3, the Markov property only captures invariances present in a single distribution. However, in the complex real world, problems may be modeled with different distributions. For example, in machine learning, a common problem is generalizing learning to out-of-distribution settings.

[3, 28] introduced a new characterization that extends the Markov property to account for experimental data arising from known-target interventions.

Definition 19 (I-Markov Property [3]) *Consider the tuple of absolutely continuous probability distributions $(P_I)_{I\in\mathcal{I}}$ over a set of variables $\mathbf{V}$. A tuple $(P_I)_{I\in\mathcal{I}}$ satisfies the I-Markov property with respect to a graph $G = (\mathbf{V} \cup \mathbf{L}, \mathbf{E})$ if the following holds for disjoint $\mathbf{Y}, \mathbf{Z}, \mathbf{W} \subseteq \mathbf{V}$:*

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

(2) For $\mathbf{I}, \mathbf{J} \in \mathcal{I}$: $P_I(y|w) = P_J(y|w)$ if $(Y \perp\!\!\!\perp K|W\backslash W_k)_{G_{\underline{W_k},\overline{R(W)}}}$

Remark 20 *We see that the I-Markov property fixes the intervention targets, $\mathbf{I} \in \mathcal{I}$ and then allows the graphical structure to change fitting the Markov property with respect to a **tuple** of distributions now rather than a single distribution.*

Similarly, the MD-Markov property allows one to fix the intervention targets in the case of known-target interventions, but more importantly generalizes to the setting with different domains and unknown-targets at the same time.

Experimental data can come with either known-target interventions, where the targets are explicitly perturbed, or from unknown-target interventions, where one knows an intervention took place, but is unsure of what nodes it possibly affects. This resulted in the Ψ-Markov property [7].

Definition 21 (Ψ-Markov Property [7]) *Let $G = (\mathbf{V} \cup \mathbf{L}, \mathbf{E})$ denote a causal graph, let $\mathbf{P}$ denote an ordered tuple of distributions and let $\mathcal{I}$ denote an ordered tuple of intervention targets such that $|\mathbf{P}| = |\mathcal{I}|$. Tuple $\mathbf{P}$ satisfies the Ψ-Markov property with respect to the pair $\langle G, \mathcal{I} \rangle$ if the following holds for disjoint $\mathbf{Y}, \mathbf{Z}, \mathbf{W} \subseteq \mathbf{V}$:*

(1) For $\mathbf{I}_i \in \mathcal{I}$: $P_i(y|w,z) = P_i(y|w)$ if $(\mathbf{Y} \perp\!\!\!\perp \mathbf{Z}|W)_G$

(2) For $\mathbf{I}_i, \mathbf{I}_j \in \mathcal{I}$: $P_i(y|w) = P_j(y|w)$ if $(\mathbf{Y} \perp\!\!\!\perp \mathbf{K}|\mathbf{W}\backslash\mathbf{W}_\mathbf{K})_{G_{\overline{W_K}, \overline{R(W)}}}$

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

Remark 22 *Compared to the MD-Markov property, the Ψ-Markov property does not allow us to characterize invariances with known-target interventions. More importantly, the Ψ-Markov property does not characterize invariances for distributions that occur in different domains.*

Thus, we reiterate that interventional invariances are distinctly different from cross-domain invariances. As we showed in Example 25, treating interventions separately from domain-changes is an important distinction to make also in structure learning.

5.2 Comparing Structure Learning Algorithms

In this section, we explicitly discuss some subtleties compared to work that could be also seen as learning over multiple domains. We survey a few works in the area of causal discovery that touch upon structure learning in the presence of multiple distributions of data. We illustrate similarities and differences via examples over a variety of different settings and previously proposed works that may be used for multi-domain structure learning.

5.2.1 Single-domain interventions with known-targets: $\mathcal{I}$-FCI [3, 28]

[28] characterize a MEC under interventions with known-targets. [3] further refines the characterization and shows an improved EC and learning algorithm, the $\mathcal{I}$-FCI algorithm. As shown in Ex. 25 and related simulation experiment in Proof of Main Text Thm. 18 [S-FCI Soundness], MD-FCI not only learns additional details when possible, but also does not learn incorrect statements (i.e. the algorithm is sound).

5.2.2 Invariant Causal Prediction [13, 50]

Invariant causal prediction (ICP) can identify the causal parents of a target variable under the assumption that the target's causal mechanism is invariant across environments [13, 50]. The work in [13] treats interventions and different regimes (i.e. domains) as similar concepts, whereas in this work, as explicitly noted by the MD-Markov property, they are in fact quite different in subtle ways.

The work proposes for a target variable Y, to identify subsets S such that $P_i(Y|S)$ are invariant across all distributions $P_i(.)$ for all "environments" i. Interestingly, the paper provides sufficient conditions for the ICP framework to uniquely identify the true causal parents of Y. However, this requires the assumption of linear SCMs, the absence of latent confounders, and certain constraints on the set of interventions. It is interesting future work to explore the ideas introduced in this paper in the context of functional assumptions on the causal structure. However, we contrast our approach mainly with the idea of leveraging invariances across distributions.

Mainly, the authors in [13] suggest looking for invariances that hold across *all* domains, whereas we look for invariances across pairs of domains. Moreover, we also leverage different pairs of distributions to learn different information. For example, comparing interventions across domains allows one to learn invariances with respect to both the domain change and the intervention set. However, comparing interventions within a domain allows one to learn invariances with respect to the intervention set, with the implicit assumption that there are no other changes induced by a changing environment.

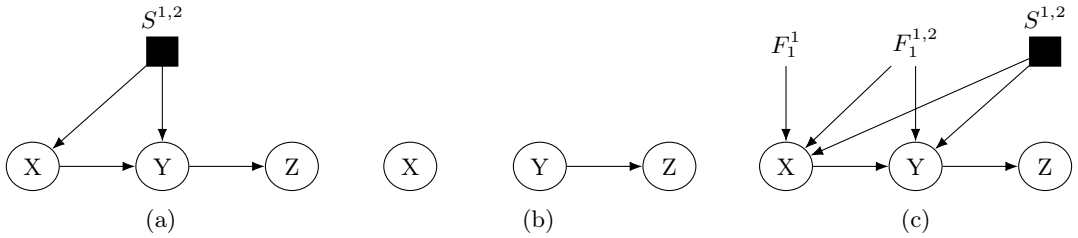


Figure 12: **Comparison of ICP vs MD-FCI** given ground-truth graph (a). Assume that we are given interventions $\mathbf{\Psi} = \langle \{\}^1, \{X\}^1, \{\}^2 \rangle$ and their corresponding distributions with known-targets $\mathcal{K} = [1, 0, 1]$. (b) is the graph learned by ICP. (c) is the MD-PAG learned by MD-FCI.

As an example, consider the graph and setting shown in Figure 12(a). We have known-target interventions $\mathbf{\Psi} = \langle \{\}^1, \{\}^2, \{X\}^1 \rangle$, $\mathcal{K} = [1, 1, 1]$ and their associated distributions $\mathbf{P}$. When applying ICP, one would recursively say discover parents and say we start with Z. Across all distributions, one would see that $P(z|y)$ is invariant, and thus $Y \to Z$. However, say one moves to the node Y next. There is no invariance for $P(Y|S)$ across all distributions, since for example the domain-shift from domain 1 to 2 through the S-node, $S^{1,2}$ affects Y. Thus, ICP may learn the graph in Figure 12(b). On the other hand, Figure 12(c) show the result of applying MD-FCI and even the S-node structure is recovered.

5.2.3 Causal Discovery with Joint Causal Inference [14]

The work in [14] proposed "Joint Causal Inference" (JCI) as a framework that pools multiple datasets/distributions with unknown intervention targets and then employs a standard causal discovery algorithm to learn the causal graph, such as FCI. Namely, FCI-JCI is an adaptation of the FCI algorithm that learns causal graphs over the pooled datasets, combining different observational and interventional datasets. In [7] Appendix Section D.2, it is shown that Ψ-FCI explicitly can learn more than the JCI procedure. Moreover, [7] Appendix Section D.2 Proposition 6 demonstrates a proof that this holds in general for settings with at least three distributions. The basic intuition is that JCI compares everything relative to the observational distribution, which can miss invariances. On the other hand, comparing every pair of distributions is important for characterizing all possible invariances. Since JCI is already shown to characterize and learn less in a single-domain setting, the same will hold when we consider the multi-domain setting. We direct the readers to [7] for additional discussion on the single-domain setting comparing Ψ-FCI to JCI.

The pooling procedure constructs auxiliary context variables, $\mathbf{C} = \{C_i\}_{i=1}^M$ given datasets $\langle D_0, ..., D_M \rangle$, where D_0 corresponds to the "observational distribution and D_i corresponds to an "interventional" distribution. The algorithm pools the datasets into one, D^* and then appends context variables such that $\mathbf{C} = 0$ for D_0 and $C_i = 1$ if the sample corresponds to D_i, else $C_i = 0$. Thus there are an additional M columns in the dataset, which result in added nodes to the causal graph. When context nodes are added, $C_i \leftrightarrow C_j$ for all i, j and then $C_i \to V_j$ if there is a dependency among the C_i variable and the V_j variable.

In Figure 13, MD-FCI is shown to learn more than JCI. JCI learns the graph in Figure 13(b). MD-FCI learns the graph in (c) and importantly also estimates the S-node structure.

5.2.4 Causal Discovery with Nonstationary Changes [11, 51]

[11, 51] also uses auxiliary random variables to capture mechanism changes. JCI can be seen as an extension of this idea. Similarly to JCI, our approach differs in how we treat these auxiliary nodes and characterize the pairwise-distribution invariances in a more complete manner.

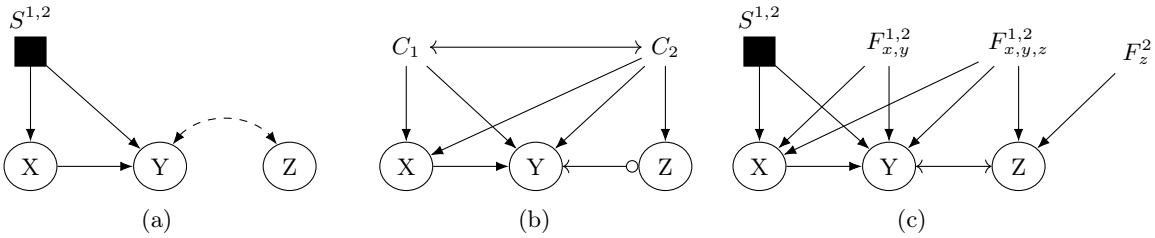


Figure 13: **Comparison of JCI vs MD-FCI** given ground-truth graph (a). Assume that we are given interventions $\mathbf{\Psi} = \langle \{\}^1, \{X, Y\}^2, \{X, Y, Z\}^2 \rangle$ and their corresponding distributions with known-targets $\mathcal{K} = [1, 1, 1]$. (b) is the graph is the MD-PAG learned by MD-FCI and (c) is the graph learned by JCI. The results hold if the interventions are unknown-targets as well, and even if there was observational data in both domains.

5.2.5 Multi-Domain Causal Structure Learning in Linear Systems [52]

The paper [52] proposes a causal discovery method that accounts for observations across multiple domains. However, the setting relies on the absence of latent confounders and also linearity in the SCM. In this work, we characterize the EC in the non-Markovian and nonparametric setting.

5.2.6 Single-domain interventions with unknown-targets: Ψ-FCI [7]

[7] generalize the work of the $\mathcal{I}$-FCI and its EC characterization to the setting with unknown intervention targets and the authors propose a constraint-based learning algorithm for learning a Ψ-MEC, the Ψ-FCI algorithm. The work here is the most similar to ours given the results from Corr. 51. As a result, one might expect that the MD-FCI algorithm and the MD-Markov characterization is just a relabeling of the Ψ-FCI and Ψ-Markov characterization. However, as demonstrated in previous sections, accounting for interventional data across domains is not a simple application of the Ψ-Markov characterization and Ψ-FCI learning algorithm.

Here, we present an additional example demonstrating that when considering the domain setting, one can learn more than just naively applying the Ψ-FCI algorithm. Moreover, this demonstrates that the MD-Markov characterization is a more refined EC characterization.

Example 29 (Pooling non-interventional data) *Consider the selection diagram in Figure 14(a) over two domains* $\mathbf{\Pi} = \{\Pi^1, \Pi^2\}$. *The S-node pointing to Z indicates that there is a possible change in mechanism going from domain 1 to domain 2. We are given interventions* $\mathbf{\Psi}^{\Pi} = \langle \{\}^1, \{X\}^1, \{\}^2 \rangle$, *and corresponding distributions* $\mathbf{P}^{\Pi} = \langle P^1_1, P^1_2, P^2_1 \rangle$ *with known-intervention targets* $\mathcal{K} = [1, 0, 1]$. *Assume we have access to an oracle to query for d-separation.*

If one runs the Ψ-FCI algorithm, then there is no notion of multiple domains in the Ψ-Markov characterization. Therefore, we would ignore the domain superscript, and combine the two observational datasets. Running the algorithm results in the Ψ-PAG in Figure 14(b). Observe that the skeleton of the variables $\{X, Y, Z\}$ *is correct. However, no orientations are learned. In contrast, Figure 14(c) shows the results of running the MD-FCI algorithm. Observe that there is not only improved orientation by learning that* $Y \to Z$, *but also the augmented nodes provide additionally rich structure. For example, the MD-PAG indicates that the only S-node present in the true selection diagram is one that points to Z.* ■

This example demonstrates that the characterization and MD-FCI algorithm proposed in this paper improves upon the work of [7]. Note that we demonstrate subtle differences that show we improve upon the Ψ-FCI algorithm. The appendix of [7] also shows similar examples that illustrate subtle differences of the Ψ-FCI algorithm with respect to other works, such as [13, 14, 28, 75].

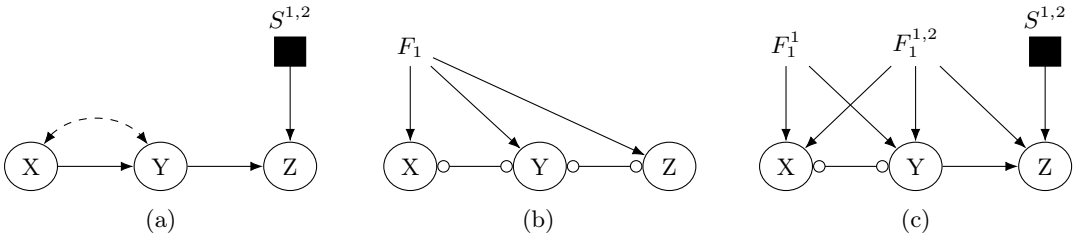


Figure 14: Augmented graph representing an intervention and a S-node over domains 1 and 2 (a) with interventions $\mathbf{\Psi}^{\Pi} = \langle \{\}^1, \{X\}^1, \{\}^2 \rangle$. The resulting Ψ-PAG one can learn using Ψ-FCI (b), and the resulting MD-PAG one can learn using MD-FCI (c).

5.3 Hard vs Soft Interventions: Characterization and Learning

This paper focuses on the setting of soft interventions, but hard interventions may occur in practice, where the experimentalist has control to set the exact value of a variable, or randomize the setting of the variable's value (i.e. in a randomized control trial). In this case, the causal graph is modified such that the dependency on the variable's parents are cut. It is a reasonable question to ask if the characterization and learning holds in the setting of *hard* interventions. We provide a brief discussion here on the challenges involved and what can be done.

Markovian SCMs First, we direct our attention towards the Markovian SCMs. Section 3 discussed the characterization and learning in the context of soft interventions. And now, we will discuss some of its nuances when considering hard interventions. [28, Yang et al.] and [8, Hauser et al.] was the first to characterize the hard interventional EC in the single-domain setting when there are no latent confounders assuming the interventions are "conservative" meaning $\forall j \in \mathbf{V}$, $\exists I \in \mathbf{\Psi}$ such that $j \notin I$. In words, conservative interventions do not overlap.

Definition 23 (Conservative family of intervention targets from [8]) *A family of intervention targets $\mathbf{\Psi}$ is called conservative if for all $a \in \mathbf{V}$, there is some $I \in \mathbf{\Psi}$ such that $a \notin I$.* ■

This gives rise to a theorem from [8] that characterizes an EC in the context of hard interventions from a conservative family of targets.

Theorem 24 (Interventional Markov Equiv. from [8]) *Let G_1 and G_2 be two causal diagrams over variables $\mathbf{V}$ and $\mathbf{\Psi}$ be a conservative family of intervention targets under hard interventions. Then the following statements are equivalent:*

1. *$G_1 \sim_{\mathbf{\Psi}} G_2$;*
2. *for all $I \in \mathbf{\Psi}$, $G_1^{(I)} \sim G_2^{(I)}$ (Markov equivalent in the observational sense);*
3. *For all $I \in \mathbf{\Psi}$, $G_1^{(I)}$ and $G_1^{(I)}$ have the same skeleton and the same v-structures;*
4. *G_1 and G_2 have the same skeleton and the same v-structures and $G_1^{(I)}$ and $G_1^{(I)}$ have the same skeleton for all $I \in \mathbf{\Psi}$*

■

However, hard interventions in the context of non-Markovian SCMs have additional complexity, which is not captured in the existing graphical characterizations.

Non-Markovian SCMs When not all causal variables are observed and there exist latent confounders, the characterization of $\mathcal{I}$-MEC and Ψ-MEC are insufficient at representing all possible invariances present in the data. As a result, this translates to a challenge in developing a learning algorithm that utilizes all possible invariances present in the data. We describe some of the challenges. With latent confounding, the causal graph contains bidirected edges, $X \leftrightarrow Y$, which possibly results in inducing paths within the graph. Inducing paths complicate structure learning in non-Markovian setting because an inducing path between two variables means they are unable to be m-separated. In the MD-MAG characterization, this means two variables are connected with an edge even if they are not adjacent in the true underlying causal graph. It is thus, not trivial to determine whether an adjacency comes from a direct causal relationship, or from an inducing paths. Hard interventions cut incoming edges to variables and thus may provide information on inducing path structures. It becomes difficult to represent this larger space of invariances that hard interventions provide and thus the complete characterization of an EC in the context of hard interventions remains elusive. Further research on this exciting topic is warranted.

6 Conclusions

In this paper, we introduced a generalized Markov property called MD-Markov, which defines a new EC of selection diagrams, the MD-PAG, representing the constraints found across observational and experimental distributions collected from multiple domains. Building on this new characterization, we develop a causal discovery algorithm called MD-FCI, which subsumes FCI, $\boldsymbol{\mathcal{I}}$-FCI, and Ψ-FCI, and accepts as input a mixture of observational and interventional data from multiple domains.

Future interesting work would involve relaxing the assumptions made in this paper, and leveraging the characterization of the EC for downstream causal ID and estimation tasks. An interesting line of work could be extending the transportability-ID algorithms to work on the MD-PAG [76].

Acknowledgements

AL was supported by the NSF Computing Innovation Fellowship (#2127309). EB was supported in part by the NSF, ONR, AFOSR, DoE, Amazon, JP Morgan, and The Alfred P. Sloan Foundation.

References

[1] Judea Pearl. *Causality: Models, reasoning, and inference.* 2nd. Cambridge University Press, 2009.

[2] P. Spirtes, C. Glymour, and R. Scheines. "Causation, Prediction, and Search." In: 81 (1993). Place: New York, NY Publisher: Springer New York.

[3] M. Kocaoglu, A. Jaber, K. Shanmugam, and E. Bareinboim. "Characterization and Learning of Causal Graphs with Latent Variables from Soft Interventions." In: *Advances in Neural Information Processing Systems* 32 (2019).

[4] T. S. Verma and J. Pearl. *An Algorithm for Deciding if a Set of Observed Independencies Has a Causal Explanation.* arXiv:1303.5435 [cs]. 2013.

[5] P. L. Spirtes, C. Meek, and T. S. Richardson. *Causal Inference in the Presence of Latent Variables and Selection Bias.* arXiv:1302.4983 [cs]. 2013.

[6] C. Meek. *Causal Inference and Causal Explanation with Background Knowledge.* arXiv:1302.4972 [cs]. 2013.

[7] A. Jaber, M. Kocaoglu, K. Shanmugam, and E. Bareinboim. "Causal Discovery from Soft Interventions with Unknown Targets: Characterization and Learning." In: *Advances in Neural Information Processing Systems*. Vol. 33. Curran Associates, Inc., 2020, pp. 9551–9561.

[8] A. Hauser and P. Bühlmann. *Characterization and Greedy Learning of Interventional Markov Equivalence Classes of Directed Acyclic Graphs.* arXiv:1104.2808 [cs, math, stat]. 2012.

[9] A. Hauser and P. Bühlmann. "Two Optimal Strategies for Active Learning of Causal Models from Interventional Data." In: *International Journal of Approximate Reasoning* 55.4 (2014). arXiv:1205.4174 [cs, stat], pp. 926–939.

[10] R. Perry, J. von Kügelgen, and B. Schölkopf. *Causal Discovery in Heterogeneous Environments Under the Sparse Mechanism Shift Hypothesis.* arXiv:2206.02013 [cs, stat]. 2022.

[11] B. Huang, K. Zhang, M. Gong, and C. Glymour. "Causal Discovery and Forecasting in Nonstationary Environments with State-Space Models." en. In: *Proceedings of the 36th International Conference on Machine Learning*. ISSN: 2640-3498. PMLR, 2019, pp. 2901–2910.

[12] B. Huang, C. J. H. Low, F. Xie, C. Glymour, and K. Zhang. "Latent hierarchical causal structure discovery with rank constraints." In: *arXiv preprint arXiv:2210.01798* (2022).

[13] J. Peters, P. Bühlmann, and N. Meinshausen. *Causal inference using invariant prediction: identification and confidence intervals.* arXiv:1501.01332 [stat]. 2015.

[14] J. M. Mooij, S. Magliacane, and T. Claassen. "Joint causal inference from multiple contexts." In: *The Journal of Machine Learning Research* 21.1 (2020), 99:3919–99:4026.

[15] G. Eder. "A longitudinal study of the kidney function of the chimpanzee (Pan troglodytes) in comparison with humans." eng. In: *European Journal of Clinical Chemistry and Clinical Biochemistry: Journal of the Forum of European Clinical Chemistry Societies* 34.11 (1996), pp. 889–896.

[16] J. Pearl and E. Bareinboim. "Transportability of Causal and Statistical Relations: A Formal Approach." en. In: *Proceedings of the AAAI Conference on Artificial Intelligence* 25.1 (2011). Number: 1, pp. 247–254.

[17] E. Bareinboim and J. Pearl. "Transportability of Causal Effects: Completeness Results." In: *Proceedings of the AAAI Conference on Artificial Intelligence* 26.1 (2012), pp. 698–704.

[18] E. Bareinboim and J. Pearl. "Meta-Transportability of Causal Effects: A Formal Approach." en. In: *Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics*. ISSN: 1938-7228. PMLR, 2013, pp. 135–143.

[19] E. Bareinboim and J. Pearl. "Causal inference and the data-fusion problem." In: *Proceedings of the National Academy of Sciences* 113.27 (2016). Publisher: National Academy of Sciences, pp. 7345–7352.

[20] J. Zhang. "On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias." In: *Artificial Intelligence* 172.16-17 (2008). Publisher: Elsevier, pp. 1873–1896.

[21] P. Spirtes, C. Glymour, and R. Scheines. "From probability to causality." In: *Philosophical Studies* 64 (1991), pp. 1–36.

[22] J. Pearl and D. Mackenzie. *The book of why : the new science of cause and effect*. Pages: 418. 2019.

[23] D. Colombo, M. H. Maathuis, M. Kalisch, and T. S. Richardson. "Learning high-dimensional directed acyclic graphs with latent and selection variables." In: *Annals of Statistics* 40.1 (2011), pp. 294–321.

[24] D. Colombo, M. H. Maathuis, M. Kalisch, and T. S. Richardson. "Learning high-dimensional directed acyclic graphs with latent and selection variables." In: *The Annals of Statistics* 40.1 (2012). arXiv:1104.5617 [cs, math, stat].

[25] D. Colombo and M. H. Maathuis. "Order-Independent Constraint-Based Causal Structure Learning." In: *Journal of Machine Learning Research* 15 (2014), pp. 3921–3962.

[26] T. R. Hughes, M. J. Marton, A. R. Jones, C. J. Roberts, R. Stoughton, C. D. Armour, H. A. Bennett, E. Coffey, H. Dai, Y. D. He, M. J. Kidd, A. M. King, M. R. Meyer, D. Slade, P. Y. Lum, S. B. Stepaniants, D. D. Shoemaker, D. Gachotte, K. Chakraburtty, J. Simon, M. Bard, and S. H. Friend. "Functional discovery via a compendium of expression profiles." eng. In: *Cell* 102.1 (2000), pp. 109–126.

[27] X. Shen, S. Ma, P. Vemuri, and G. Simon. "Challenges and Opportunities with Causal Discovery Algorithms: Application to Alzheimer's Pathophysiology." en. In: *Scientific Reports* 10.1 (2020). Number: 1 Publisher: Nature Publishing Group, p. 2975.

[28] K. D. Yang, A. Katcoff, and C. Uhler. *Characterizing and Learning Equivalence Classes of Causal DAGs under Interventions*. arXiv:1802.06310 [math, stat]. 2019.

[29] K. Sachs, O. Perez, D. Pe'er, D. A. Lauffenburger, and G. P. Nolan. "Causal protein-signaling networks derived from multiparameter single-cell data." eng. In: *Science (New York, N.Y.)* 308.5721 (2005), pp. 523–529.

[30] D. Eaton and K. Murphy. "Exact Bayesian structure learning from uncertain interventions." en. In: *Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics*. ISSN: 1938-7228. PMLR, 2007, pp. 107–114.

[31] R. J. Smith, M. A. Hays, G. Kamali, C. Coogan, N. E. Crone, J. Y. Kang, and S. V. Sarma. "Stimulating native seizures with neural resonance: a new approach to localize the seizure onset zone." In: *Brain* 145.11 (2022), pp. 3886–3900.

[32] C. D. Stimpson, N. Barger, J. P. Taglialatela, A. Gendron-Fitzpatrick, P. R. Hof, W. D. Hopkins, and C. C. Sherwood. "Differential serotonergic innervation of the amygdala in bonobos and chimpanzees." In: *Social Cognitive and Affective Neuroscience* 11.3 (2016), pp. 413–422.

[33] A. Li, P. Myers, N. Warsi, K. M. Gunnarsdottir, S. Kim, V. Jirsa, A. Ochi, H. Otusbo, G. M. Ibrahim, and S. V. Sarma. *Neural Fragility of the Intracranial EEG Network Decreases after Surgical Resection of the Epileptogenic Zone*. en. Pages: 2021.07.07.21259385. 2022.

[34] A. Li, C. Huynh, Z. Fitzgerald, I. Cajigas, D. Brusko, J. Jagid, A. O. Claudio, A. M. Kanner, J. Hopp, S. Chen, J. Haagensen, E. Johnson, W. Anderson, N. Crone, S. Inati, K. A. Zaghloul, J. Bulacio, J. Gonzalez-Martinez, and S. V. Sarma. "Neural fragility as an EEG marker of the seizure onset zone." en. In: *Nature Neuroscience* 24.10 (2021). Number: 10 Publisher: Nature Publishing Group, pp. 1465–1474.

[35] A. Li, S. Inati, K. Zaghloul, and S. Sarma. "Fragility in Epileptic Networks : the Epileptogenic Zone." In: 2017, pp. 1–8.

[36] A. Palepu, A. Li, Z. Fitzgerald, K. Hu, J. Costacurta, J. Bulacio, J. Martinez-Gonzalez, and S. V. Sarma. "Evaluating Invasive EEG Implantations with Structural Imaging Data and Functional Scalp EEG Recordings from Epilepsy Patients." eng. In: *Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual International Conference* 2019 (2019), pp. 3866–3869.

[37] K. M. Gunnarsdottir, A. Li, R. J. Smith, J.-Y. Kang, A. Korzeniewska, N. E. Crone, A. G. Rouse, J. J. Cheng, M. J. Kinsman, P. Landazuri, U. Uysal, C. M. Ulloa, N. Cameron, I. Cajigas, J. Jagid, A. Kanner, T. Elarjani, M. M. Bicchi, S. Inati, K. A. Zaghloul, V. L. Boerwinkle, S. Wyckoff, N. Barot, J. Gonzalez-Martinez, and S. V. Sarma. "Source-sink connectivity: a novel interictal EEG marker for seizure localization." In: *Brain* 145.11 (2022), pp. 3901–3915.

[38] J. M. Bernabei, A. Li, A. Y. Revell, R. J. Smith, K. M. Gunnarsdottir, I. Z. Ong, K. A. Davis, N. Sinha, S. Sarma, and B. Litt. "Quantitative approaches to guide epilepsy surgery from intracranial EEG." In: *Brain* (2023), awad007.

[39] K. Jo Black and M. Richards. "Eco-gentrification and who benefits from urban green amenities: NYC's high Line." en. In: *Landscape and Urban Planning* 204 (2020), p. 103900.

[40] P.-Y. Tung, J. D. Blischak, C. J. Hsiao, D. A. Knowles, J. E. Burnett, J. K. Pritchard, and Y. Gilad. "Batch effects and the effective design of single-cell gene expression studies." en. In: *Scientific Reports* 7.1 (2017). Number: 1 Publisher: Nature Publishing Group, p. 39921.

[41] S. Nolte and J. Call. "Targeted helping and cooperation in zoo-living chimpanzees and bonobos." eng. In: *Royal Society Open Science* 8.3 (2021), p. 201688.

[42] E. A. Petersen, T. G. Stauss, J. A. Scowcroft, E. S. Brooks, J. L. White, S. M. Sills, K. Amirdelfan, M. N. Guirguis, J. Xu, C. Yu, A. Nairizi, D. G. Patterson, K. C. Tsoulfas, M. J. Creamer, V. Galan, R. H. Bundschu, C. A. Paul, N. D. Mehta, H. Choi, D. Sayed, S. P. Lad, D. J. DiBenedetto, K. A. Sethi, J. H. Goree, M. T. Bennett, N. J. Harrison, A. F. Israel, P. Chang, P. W. Wu, G. Gekht, C. E. Argoff, C. E. Nasr, R. S. Taylor, J. Subbaroyan, B. E. Gliner, D. L. Caraway, and N. A. Mekhail. "Effect of High-frequency (10-kHz) Spinal Cord Stimulation in Patients With Painful Diabetic Neuropathy: A Randomized Clinical Trial." eng. In: *JAMA neurology* 78.6 (2021), pp. 687–698.

[43] A. M. Lozano, N. Lipsman, H. Bergman, P. Brown, S. Chabardes, J. W. Chang, K. Matthews, C. C. McIntyre, T. E. Schlaepfer, M. Schulder, Y. Temel, J. Volkmann, and J. K. Krauss. "Deep brain stimulation: current challenges and future directions." In: *Nature reviews. Neurology* 15.3 (2019), pp. 148–160.

[44] J. Pearl and E. Bareinboim. "Transportability across studies: A formal approach." In: (2018).

[45] J. D. Correa and E. Bareinboim. "From Statistical Transportability to Estimating the Effect of Stochastic Interventions." In: ().

[46] T. R. Frieden. "Evidence for Health Decision Making — Beyond Randomized, Controlled Trials." en. In: *New England Journal of Medicine* 377.5 (2017). Ed. by J. M. Drazen, D. P. Harrington, J. J. McMurray, J. H. Ware, and J. Woodcock, pp. 465–475.

[47] C. J. Kelly, A. Karthikesalingam, M. Suleyman, G. Corrado, and D. King. "Key challenges for delivering clinical impact with artificial intelligence." In: *BMC Medicine* 17.1 (2019), pp. 195–195.

[48] D. Ehrens, A. Li, F. Aeed, Y. Schiller, and S. V. Sarma. "Network Fragility for Seizure Genesis in an Acute in vivo Model of Epilepsy." eng. In: *Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual International Conference* 2020 (2020), pp. 3695–3698.

[49] A. Ghassami, S. Salehkaleybar, N. Kiyavash, and K. Zhang. *Learning Causal Structures Using Regression Invariance*. arXiv:1705.09644 [cs, stat]. 2017.

[50] C. Heinze-Deml, J. Peters, and N. Meinshausen. *Invariant Causal Prediction for Nonlinear Models*. arXiv:1706.08576 [stat]. 2018.

[51] B. Huang, K. Zhang, J. Zhang, J. Ramsey, R. Sanchez-Romero, C. Glymour, and B. Schölkopf. *Causal Discovery from Heterogeneous/Nonstationary Data with Independent Changes*. arXiv:1903.01672 [cs, stat]. 2020.

[52] A. Ghassami, N. Kiyavash, B. Huang, and K. Zhang. "Multi-domain Causal Structure Learning in Linear Systems." In: *Advances in Neural Information Processing Systems*. Vol. 31. Curran Associates, Inc., 2018.

[53] Y. Zeng, S. Shimizu, R. Cai, F. Xie, M. Yamamoto, and Z. Hao. "Causal Discovery with Multi-Domain LiNGAM for Latent Factors." en. In: *Proceedings of The 2021 Causal Analysis Workshop Series*. ISSN: 2640-3498. PMLR, 2021, pp. 1–4.

[54] A. Li, A. Jaber, and E. Bareinboim. "Causal discovery from observational and interventional data across multiple environments." In: *Thirty-seventh Conference on Neural Information Processing Systems*. 2023.

[55] D. M. Chickering. "Optimal Structure Identification With Greedy Search." In: *Journal of Machine Learning Research* 3.Nov (2002), pp. 507–554.

[56] M. Chickering. "Statistically Efficient Greedy Equivalence Search." en. In: *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI)*. ISSN: 2640-3498. PMLR, 2020, pp. 241–249.

[57] T. Claassen and I. G. Bucur. "Greedy equivalence search in the presence of latent confounders." en. In: *Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence*. ISSN: 2640-3498. PMLR, 2022, pp. 443–452.

[58] D. M. Chickering and C. Meek. *Selective Greedy Equivalence Search: Finding Optimal Bayesian Networks Using a Polynomial Number of Score Evaluations*. arXiv:1506.02113 [cs]. 2015.

[59] E. Bareinboim, J. D. Correa, D. Ibeling, and T. Icard. "On Pearl's Hierarchy and the Foundations of Causal Inference." In: *Probabilistic and Causal Inference: The Works of Judea Pearl*. 1st ed. Vol. 36. New York, NY, USA: Association for Computing Machinery, 2022, pp. 507–556.

[60] D. Geiger, T. Verma, and J. Pearl. "d-Separation: From Theorems to Algorithms." en. In: *Machine Intelligence and Pattern Recognition*. Ed. by M. Henrion, R. D. Shachter, L. N. Kanal, and J. F. Lemmer. Vol. 10. Uncertainty in Artificial Intelligence. North-Holland, 1990, pp. 139–148.

[61] J. Zhang and G. F. Cooper. "Causal Reasoning with Ancestral Graphs." In: *Journal of Machine Learning Research* 9 (2008), pp. 1437–1474.

[62] A. P. Dawid. "Influence Diagrams for Causal Modelling and Inference." In: *International Statistical Review / Revue Internationale de Statistique* 70.2 (2002). Publisher: [Wiley, International Statistical Institute (ISI)], pp. 161–189.

[63] J. Correa and E. Bareinboim. "A Calculus for Stochastic Interventions:Causal Effect Identification and Surrogate Experiments." en. In: *Proceedings of the AAAI Conference on Artificial Intelligence* 34.06 (2020). Number: 06, pp. 10093–10100.

[64] J. Pearl. "Causal Diagrams for Empirical Research." In: *Biometrika* 82.4 (1995). Publisher: [Oxford University Press, Biometrika Trust], pp. 669–688.

[65] E. Bareinboim, A. Forney, and J. Pearl. "Bandits with Unobserved Confounders: A Causal Approach." In: *Advances in Neural Information Processing Systems*. Vol. 28. Curran Associates, Inc., 2015.

[66] D. Koller and N. Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

[67] J. Pearl and E. Bareinboim. "External Validity: From Do-Calculus to Transportability Across Populations." In: *Statistical Science* 29.4 (2014). arXiv:1503.01603 [cs, stat].

[68] T. Richardson and P. Spirtes. "Ancestral graph Markov models." In: *The Annals of Statistics* 30.4 (2002). Publisher: Institute of Mathematical Statistics, pp. 962–1030.

[69] J. D. Correa and E. Bareinboim. "General Transportability of Soft Interventions: Completeness Results." In: ().

[70] J. Zhang and P. Spirtes. "The three faces of faithfulness." In: *Synthese* 193.4 (2016). Publisher: Springer, pp. 1011–1027.

[71] X. Wang, W. Pan, W. Hu, Y. Tian, and H. Zhang. "Conditional Distance Correlation." en. In: *Journal of the American Statistical Association* 110.512 (2015), pp. 1726–1734.

[72] J. Park, U. Shalit, B. Schölkopf, and K. Muandet. *Conditional Distributional Treatment Effect with Kernel Conditional Mean Embeddings and U-Statistic Regression*. arXiv:2102.08208 [cs, stat]. 2021.

[73] X. Hu and J. Lei. "A Two-Sample Conditional Distribution Test Using Conformal Prediction and Weighted Rank Sum." In: *Journal of the American Statistical Association* (2023). arXiv:2010.07147 [stat], pp. 1–19.

[74] P. Hall and J. D. Hart. "Bootstrap Test for Difference Between Means in Nonparametric Regression." In: *Journal of the American Statistical Association* 85.412 (1990). Publisher: [American Statistical Association, Taylor & Francis, Ltd.], pp. 1039–1049.

[75] C. Squires, Y. Wang, and C. Uhler. *Permutation-Based Causal Structure Learning with Unknown Intervention Targets*. arXiv:1910.09007 [stat]. 2020.

[76] J. Correa and E. Bareinboim. "General Transportability of Soft Interventions: Completeness Results." In: *Advances in Neural Information Processing Systems*. Vol. 33. Curran Associates, Inc., 2020, pp. 10902–10912.

[77] S. L. Lauritzen, A. P. Dawid, B. N. Larsen, and H.-G. Leimer. "Independence properties of directed markov fields." en. In: *Networks* 20.5 (1990). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/net.3230 pp. 491–505.

[78] J. Zhang. *A Characterization of Markov Equivalence Classes for Directed Acyclic Graphs with Latent Variables*. arXiv:1206.5282 [cs, stat]. 2012.

[79] C. Meek. *Strong Completeness and Faithfulness in Bayesian Networks*. arXiv:1302.4973 [cs]. 2013.

[80] A. Jaber, J. Zhang, and E. Bareinboim. "Causal Identification under Markov Equivalence: Completeness Results." In: (2019). Publisher: PMLR, pp. 2981–2989.

[81] J. M. Robins, M. A. Hernán, and B. Brumback. "Marginal structural models and causal inference in epidemiology." eng. In: *Epidemiology (Cambridge, Mass.)* 11.5 (2000), pp. 550–560.

[82] A. Jaber, M. Kocaoglu, K. Shanmugam, and E. Bareinboim. "Causal Discovery from Soft Interventions with Unknown Targets: Characterization and Learning." In: ().

[83] A. P. Dawid. "Conditional Independence in Statistical Theory." en. In: *Journal of the Royal Statistical Society: Series B (Methodological)* 41.1 (1979). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.2517 6161.1979.tb01052.x, pp. 1–15.

[84] A. A. Hagberg, D. A. Schult, and P. J. Swart. "Exploring Network Structure, Dynamics, and Function using NetworkX." In: *Proceedings of the 7th Python in Science Conference*. Ed. by G. Varoquaux, T. Vaught, and J. Millman. Pasadena, CA USA, 2008, pp. 11–15.

[85] A. Li, J. Lee, F. Montagna, C. Trevino, and R. Ness. *Dodiscover: Causal discovery algorithms in Python.*

[86] A. Li, J. Lee, and A. Roy. *Pywhy-Graphs: Causal graphs that are networkx-compliant for the py-why ecosystem.*

[87] J. M. Mooij and T. Claassen. "Constraint-Based Causal Discovery using Partial Ancestral Graphs in the presence of Cycles." en. In: *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI)*. ISSN: 2640-3498. PMLR, 2020, pp. 1159–1168.

[88] R. Nagarajan, M. Scutari, and S. Lèbre. *Bayesian Networks in R: with Applications in Systems Biology*. en. New York, NY: Springer New York, 2013.

[89] A. Ankan and A. Panda. "pgmpy: Probabilistic graphical models using python." In: *Proceedings of the 14th Python in Science Conference (SCIPY 2015)*. Citeseer, 2015.

[90] P. Hünermund and E. Bareinboim. *Causal Inference and Data Fusion in Econometrics*. arXiv:1912.09104 [econ]. 2023.

[91] D. T. Campbell, J. C. Stanley, and N. L. Gage. *Experimental and quasi-experimental designs for research*. Experimental and quasi-experimental designs for research. Pages: ix, 84. Boston, MA, US: Houghton, Mifflin and Company, 1963.

[92] C. F. Manski. *Identification for Prediction and Decision:* Cambridge, MA: Harvard University Press, 2008.

[93] S. Wasserman. "Review of Statistical Methods for Meta-Analysis." In: *Journal of Educational Statistics* 13.1 (1988). Publisher: [Sage Publications, Inc., American Educational Research Association, American Statistical Association], pp. 75–78.

[94] W. R. Shadish, T. D. Cook, and D. T. Campbell. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Cengage Learning, 2002.

[95] S. L. Morgan and C. Winship. *Counterfactuals and Causal Inference: Methods and Principles for Social Research*. Analytical Methods for Social Research. Cambridge: Cambridge University Press, 2007.

[96] B. Schölkopf, F. Locatello, S. Bauer, N. R. Ke, N. Kalchbrenner, A. Goyal, and Y. Bengio. *Towards Causal Representation Learning*. arXiv:2102.11107 [cs]. 2021.

Appendix

Contents

F.1 Proofs

Here, we provide the detailed proofs of theoretical results in the main paper. First, we review some fundamental definitions and results that guide the main results. Readers familiar with the literature can skip to Section Appendix.

F.1.1 Background results

In this section, we centralize theoretical results in relation to the theory presented in this paper.

Definition 25 ("Global" Markov property of DAGs [77]) *Consider a joint probability distribution, P over a set of variables $\boldsymbol{V}$ satisfies the **Markov property** with respect to a graph $G = (V \cup L, E)$ if the following holds for, $(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$ disjoint subsets of V:*

$$P(y|x, z) = P(y|z) \quad \text{if } Y \perp\!\!\!\perp X|Z \text{ in } G \text{ (that is } Y \text{ is d-separated from } X \text{ given } Z)$$

■

The global Markov property maps graphical structure in causal directed acyclic graphs (DAGs) to conditional independence (CI) statements in the relevant probability distributions from data.

Definition 26 (Maximal Ancestral Graphs (MAGs) [61]) *A mixed-edge graph is a maximal ancestral graph (MAG) if:*

1. *there is no directed cycles (acyclicity) and*
2. *there are no almost directed cycles (ancestrality) and*
3. *there is no primitive inducing path between any two non-adjacent vertices (maximal)*

■

Thus acyclicity, ancestrality and maximality are graphical properties of any equivalence class that stems from a MAG. As we will see, this includes the PAG and its multi-distributional generalizations. Many DAGs may encode the same CI statements, and a MAG encodes an equivalence class of these CI statements that has desirable properties such as maximality and ancestrality. To compare different MAGs, one can leverage Definition 27.

Definition 27 (General Markov Equivalence from [78]) *Two MAGs $G_1 = (V, E_1)$, $G_2 = (V, E_2)$ are Markov equivalent if for any three disjoint sets of vertices, X, Y, Z, X and Y are m-separated by Z in G_1 if and only if X and Y are m-separated by Z in G_2.* ■

Checking Definition 27 is quite tedious because it involves explicitly comparing every single m-separation statement possible in both graphs. An equivalent completely graphical criterion in Proposition 28 can be instead used.

Proposition 28 (Graphical Criterion for Markov Equivalence from [78]) *Two MAGs over the same set of vertices are Markov equivalent if and only if*

1. *Skeleton: They have the same adjacencies*
2. *V-structures: They have the same unshielded colliders*
3. *Discriminating Paths: If a path p is a discriminating path for a vertex Y in both graphs, 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.*

■

Unfortunately, a MAG is not uniquely identifiable (i.e. learnable) from observational data in general. Therefore, a partial ancestral graph (PAG) is defined as the object of interest instead.

Definition 29 (Partial Ancestral Graph [61]) *Let [M] be the MEC of an arbitrary MAG M. The PAG for [M], $\mathcal{P}_{[M]}$ is a partial mixed graph such that:*

1. *$\mathcal{P}_{[M]}$ has the same adjacencies as M (and any member of [M]) does*
2. *A mark of arrowhead is in $\mathcal{P}_{[M]}$ if and only if it is shared by all MAGs in [M]*
3. *A mark of tail is in $\mathcal{P}_{[M]}$ if and only if it is shared by all MAGs in [M].*

■

We note that do-calculus is complete for learning PAGs [20]. As noted in [7], the FCI algorithm really only leverages the inversion of R1 of do-calculus within a single domain. If we have access to interventional distributions, the inversion of R2 and R3 of the do-calculus enable one to further characterize and learn a more detailed EC [7].

A final lemma due to [79] is useful for proving properties about distributions that satisfy certain graph constraints, but not others. Meek uses the following result to show that set of unfaithful distributions has Lebesgue measure zero.

Lemma 30 (Meek [79]) *Let $D = (V, E)$ be a causal DAG where $(A \not\perp\!\!\!\perp B|C)_D$. Let $D_s = (V_s, E_s)$ be the subgraph that contains all the nodes in the m-connecting path that induces $(A \not\perp\!\!\!\perp B|C)_D$. Then any distribution p over V_s where every adjacent pair of variables are dependent satisfies $(A \not\perp\!\!\!\perp B|C)_p$.*
■

F.1.2 Multi-Domain Causal Bayesian Network Invariances

[63] developed an extension of Pearl's do-calculus rules to soft interventions in SCMs. In [76], it was shown that for the general problem of transportability, the generalized do-calculus rules are complete. In this section, we take the do-calculus rules and extend them to invariances present in a Causal Bayesian Network (CBN) that can apply across two arbitrary interventions and two arbitrary domains. As such, we generalize the Theorem 2 and 3 in [67]. This is essential for motivating the MD-Markov property characterization and the corresponding equivalence class. This result leads to the Definition 3 presented in Section 2. Here, we present a special case of σ-calculus rules in Thm. 1 of [63, 80] for the setting of multiple domains.

Proof of Theorem 31.

Theorem 31 (σ-calculus rules for multiple domains) *Let $G_S = (\mathbf{V} \cup \mathbf{L} \cup S, \mathbf{E} \cup \mathbf{E_S})$ be the corresponding causal selection diagram with latents and S-nodes defined between arbitrary N domains $\mathbf{\Pi} = \{\Pi^1, \Pi^2, ..., \Pi^N\}$. Then the following holds for any strictly positive distribution consistent with G_S.*

Rule 1 (multi-domain see-see): For any $X \subseteq V$ and disjoint $Y, Z, W \subseteq V$ and any $\Pi^i \in \mathbf{\Pi}$,

$$P^i_X(y|w,z) = P^i_X(y|w) \quad \text{if } Y \perp Z|W, \mathbf{S} \text{ in } G_S.$$

Rule 2 (multi-domain do-see): For any disjoint $X, Y, Z \subseteq V$ and $W \subset V \setminus (Z \cup Y)$ and any $\Pi^i, \Pi^j \in \mathbf{\Pi}$,

$$P^i_{X,Z}(y|z,w) = P^j_X(y|z,w) \quad \text{if } Y \perp \{Z, S^{i,j}\}|W \text{ in } G_{S_{\underline{Z}}}.$$

Rule 3 (multi-domain do-do): For any disjoint $X, Y, Z \subseteq V$ and $W \subset V \setminus (Z \cup Y)$ and any $\Pi^i, \Pi^j \in \mathbf{\Pi}$,

$$P^i_{X,Z}(y|w) = P^j_X(y|w) \quad \text{if } Y \perp \{Z, S^{i,j}\}|W \text{ in } G_{S_{\overline{Z(W)}}},$$

Proof

R1 Since regardless of domain, Π^i, and soft-interventions the graph does not change.

Any distribution factorizes with respect to the original graph and any m-separation statement in the graph G_S implies conditional independence, or m-separation with respect to an S-node. However, m-separation with respect to an S-node implies an invariance across different domains [67, Thm. 2]. Since, R1 only considers a single domain, the invariance is implied by the strict positivity, the conditional independence is equivalent to the invariance in R1.

R2 If $i = j$, then $S^{i,j} = \emptyset$, and therefore the invariance is proven in [80, Thm. 1]. If $i \neq j$, then we note that $(Y \perp\!\!\!\perp \{Z, S^{i,j}\}|W)_{G_{S_{\underline{Z}}}}$ implies $(Y \perp\!\!\!\perp Z|W)_{G_{S_{\underline{Z}}}}$ and $(Y \perp\!\!\!\perp S^{i,j}|W)_{G_{S_{\underline{Z}}}}$. $S^{i,j} \in \{i, j\}$ selecting the domain mechanism for the distribution. Then we have the following:

$$P^i_{X,W}(Y|Z,W) = P_{X,W}(Y|Z,W,S^{i,j} = i)$$
$$= P_{X,W}(Y|Z,W) \quad \text{Y is m-separated from } S^{i,j} \text{ given W}$$

Similarly, we have:

$$P^j_X(Y|Z,W) = P_X(Y|Z,W,S^{i,j} = j)$$
$$= P_X(Y|Z,W) \quad \text{Y is m-separated from } S^{i,j} \text{ given W}$$

Thus the invariance follows from [80, Thm. 1].

R3 A similar logic applies, where $(Y \perp\!\!\!\perp \{Z, S^{i,j}\}|W)_{G_{S_{\overline{Z(W)}}}}$ implies $(Y \perp\!\!\!\perp Z|W)_{G_{S_{\overline{Z(W)}}}}$ and $(Y \perp\!\!\!\perp S^{i,j}|W)_{G_{S_{\overline{Z(W)}}}}$. ■

Next, we extend the do-calculus rules to apply them across two arbitrary interventions and domains. This leads to the characterization of our EC, where arbitrary sets of interventional distributions across arbitrary domains are available.

Proposition 32 (Generalized multi-domain σ-calculus) *Let $G_S = (\mathbf{V} \cup \mathbf{L} \cup S, \mathbf{E} \cup \mathbf{E_S})$ be the corresponding causal selection diagram with latents and S-nodes defined between arbitrary N domains $\mathbf{\Pi} = \{\Pi^1, \Pi^2, ..., \Pi^N\}$. Then the following holds for any strictly positive distributions, $\mathbf{P} = \langle P_1^1, P_2^1, ..., P_m^N \rangle$ corresponding one-to-one to intervention targets $\mathbf{\Psi} = \langle \mathbf{\Psi}_1^1, \mathbf{\Psi}_2^1, ..., \mathbf{\Psi}_m^N \rangle$ where $\mathbf{\Psi} \subseteq 2^{\mathbf{V}}$:*

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

$$p_I^m(y|w, z) = p_I^m(y|w) \text{ if } Y \perp Z|W, \mathbf{S} \text{ in } G_S. \tag{4}$$

Rule 2 (mixed do-do/do-see): For any $I, J \subseteq V$ and disjoint $Y, W \subseteq V$, where $K := I\Delta J$,

$$p_I^m(y|w) = p_J^n(y|w) \text{ if } Y \perp K \cup \{S^{i,j}\}|W \setminus W_k, \mathbf{S} \setminus \{S^{i,j}\} \text{ in } \mathcal{D}_{\underline{W_k}, \overline{R(W)}}, \tag{5}$$

where $W_k := W \cap K$ and $R := K \setminus W_k$

Proof **R1** The first rule follows from Thm. 31.

R2 Define the following sets:

$W_I := W_k \cap I, W_J := W_k \cap J, R_I := R \cap I, R_J := R \cap J$.

■

Lemma 33 (Generalized CBN Invariances Across Domains) *Let G be a causal diagram and $G_S = (\mathbf{V} \cup \mathbf{L} \cup S, \mathbf{E} \cup \mathbf{E_S})$ be the corresponding causal selection diagram with latents and S-nodes defined between two domains $\mathbf{\Pi} = \{\Pi^1, \Pi^*\}$ of a CBN. Let $\mathbf{P^{\Pi}}$ be a tuple of interventional distributions generated by G. Let $\mathbf{V_S}$ be the set of nodes that have an edge with respect to $\mathbf{S}$. Then the following distributional invariances hold for disjoint $\mathbf{Y}, \mathbf{Z}, \mathbf{W} \subseteq \mathbf{V}$*

(a) For $P_{\mathbf{I}}^i \in \mathbf{P^{\Pi}}$, we have $P_{\mathbf{I}}^i(y|w, z) = P_{\mathbf{I}}^i(y|w)$ if $\mathbf{Y} \perp\!\!\!\perp \mathbf{Z}|\mathbf{W}$ in G.

(b) For $P_{\mathbf{I}}^i, P_{\mathbf{J}}^j \in \mathbf{P}$, we have $P_{\mathbf{I}}^i(y|w) = P_{\mathbf{J}}^j(y|w)$ if $\mathbf{Y} \perp\!\!\!\perp \mathbf{K}|\mathbf{W}\setminus\mathbf{W_K}$ in $G_{\underline{W_K}, \overline{R(W)}}$, where $\mathbf{K} = (\mathbf{I}\Delta\mathbf{J}) \cup \mathbf{V_S^{i,j}}, \mathbf{W_K} = \mathbf{W} \cap \mathbf{K}, \mathbf{R} = \mathbf{K}\setminus\mathbf{W_K}$ and $\mathbf{R}(\mathbf{W}) \subseteq \mathbf{R}$ are non-ancestors of $\mathbf{W}$ in G.

Proof Whenever $i = j$, for domain indicators i, j, then there is no S-node by definition, since the S-node is added to select between different domains i and j. Therefore in constraint (a), because of the shared causal structure assumption A1, $P_{\mathbf{I}}^i(\mathbf{V})$ can just be written as $P_{\mathbf{I}}(\mathbf{V})$ and factorized according to the following equation:

$$P_X(\mathbf{v}) = \sum_{\mathbf{L}} \prod_{i|X_i \in \mathbf{X}} P^*(x_i|\mathbf{pa_i}) \prod_{j|T_j \notin \mathbf{X}} P(t_j|\mathbf{pa_j})$$

which is also known as the truncated factorization formula [1], or the g-formula [81]. Then applying d-separation criterion, constraint (a) follows [60].

Constraint (b) is proven in [82] Thm 4, when $i = j$. So we prove the case when $i \neq j$. To prove this, we take a similar strategy to the proof of the do-calculus rules [1]. We construct a hypothetical CBN that models the selection of a domain on each variable with an endogenous root node/variable along with the intervention on each variable. We assume the change in domain is not caused by any

variable in G. Moreover, we assume that soft interventions are triggered by exogenous variables and not affected by any variable in G.

Let $\mathbf{I}^i, \mathbf{J}^j$ denote set of nodes in $\mathbf{I}$ and $\mathbf{J}$ that occur in domains i and j respectively. We can augment G with $\mathcal{F}^{i,j} = \{F_k^{i,j} | V_i \in \mathbf{I}^i \cup \mathbf{J}^j\}$ and edges $\mathcal{E}^{i,j} = \{F_k^{i,j} \to V_k | F_k \in \mathcal{F}\}$.

G_S has an S-node $S^{i,j}$, representing the selection between domain i and j. The edges from $S^{i,j}$ are in $\mathbf{E_S}$ and their direct children are $\mathbf{V_{S^{i,j}}}$. Thus, we constraint b) holds by definition of the selection diagram, if we can remove the effect of the S-node, $S^{i,j}$.

The constructed augmented causal graph is G'. Let Pa_i denote the parents of variable $V_i \in \mathbf{V}$ that excludes nodes in S. Let Pa'_i denote the parents of variables $V_i \in \mathbf{V}$ that can include nodes in S. For each variable V_k with $F_k^{i,j}$, there are a new set of parents $Pa''_i = Pa'_i \cup \{F_k^{i,j}\}$. The distribution of $P(V_i|Pa''_i)$ is given as follows where $P^l(V_i|Pa_i)$ is a unique conditional probability for each identifier l. We have:

$$P(V_k|Pa''_k) = \begin{cases} P(V_k|Pa'_k), & \text{if } F_k^{i,j} = 0 \\ P^l(V_k|Pa'_k), & \text{if } F_k^{i,j} = l. \end{cases} \tag{6}$$

Furthermore, we can decompose each of those conditional probabilities into ones that are a function of just the Pa_k.

$$P(V_k|Pa'_k) = P(V_k|Pa_k), \quad \text{if } S^{i,j} \to V_k \notin E_S \tag{7}$$

and

$$P^l(V_k|Pa'_k) = P^l(V_k|Pa_k), \quad \text{if } S^{i,j} \to V_k \notin E_S \tag{8}$$

Thus, each $F_k^{i,j}$ has an arbitrary prior distribution over its domain, which induces a new distribution P'' over $\mathbf{V} \cup \mathbf{L} \cup \mathbf{S} \cup \mathcal{F}$ and P'' factorizes according to G'. Then $P'_{\mathbf{I}^i}(\mathbf{V})$ relates to P'' as follows where we condition on every $F_k^{i,j} \in \mathcal{F}$ such that 1) $F_k^{i,j} = 0$ if $V_k \notin \mathbf{I}^i$ and 2) $F_k^{i,j} = l$ if $V_k^l \in \mathbf{I}$.

$$P'_{\mathbf{I}}(\mathbf{V}) = \sum_{\mathbf{L}} P''(\mathbf{V} \cup \mathbf{L} \cup \mathbf{S} | F_k^{i,j} = l, ...)$$

We can similarly decompose P' and relate it to P following the same logic for the S-node. In this sense, we see that the S-nodes and the F-nodes play a similar graphical role in selecting the distribution that applies based on the selection of the S-nodes and F-nodes.

We can repeat the logic for $P_{\mathbf{J}}(\mathbf{V})$. Now, let $\mathbf{F}_{\mathbf{K}}^{i,j} = \{F_k^{i,j} | V_k \in \mathbf{I}^i \Delta \mathbf{J}^j\}$. If $(\{\mathbf{F_K}^{i,j}, S^{i,j}\} \perp \!\!\!\perp \mathbf{Y} | W)_{G'}$, then changing the conditioning values of $\mathbf{F_K}$ and $S^{i,j}$ is irrelevant to $\mathbf{Y}$ and we get $P^i_{\mathbf{I}}(y|w) = P^j_{\mathbf{J}}(y|w)$. Thus we have successfully factorized the two distributions to show they are equivalent when the corresponding graphical criterion holds. ■

The result implies that d-separation from S-nodes and their corresponding direct children represent invariances in the conditional probability distributions of observational data assuming the Markov property. Since all S-nodes are source nodes, then to be d-separated from $\mathbf{V_S}$ is equivalent to d-separation from the S-nodes $\mathbf{S}$.

F.1.3 Markovian SCM MD-Markov Property Proofs

This section contains the proofs for the theorems in Section 3.

Proof of Corollary 9

Corollary 34 (Markovian MD-Markov Equivalence) *Given selection diagrams without latents,* $G_{S_1} = (\mathbf{V} \cup \mathbf{S}, \mathbf{E_1} \cup \mathbf{E_S})$ *and* $G_{S_2} = (\mathbf{V} \cup \mathbf{S}, \mathbf{E_2} \cup \mathbf{E_S})$ *and corresponding interventional targets*

$\mathbf{\Psi}_1, \mathbf{\Psi}_2$, the pairs $\langle G_{S_1}, \mathbf{\Psi}_1 \rangle$ and $\langle G_{S_2}, \mathbf{\Psi}_2 \rangle$ are MD-Markov equivalent if and only if $Aug_{\Psi_1}(G_{S_1})$ and $Aug_{\Psi_2}(G_{S_2})$ have (1) the same skeleton and (2) the same unshielded colliders. ■

Proof In the absence of latent nodes, $Aug_{\Psi}(G_{S_1}) = MAG(Aug_{\Psi}(G_{S_1})$. This means, one can use the augmented selection diagram instead of the MD-MAG. Since the augmented selection diagram is simply a specific MD-MAG with no bidirected edges, it follows that every discriminating path for a node Y in $Aug_{\Psi}(G_{S_1})$, Y must be a non-collider.

To see, this consider the path $\pi = \langle x, q_1, ..., q_p, b, y \rangle$ that is a discriminating path for b in a graph without bidirected edges. Then (q_p, b, y), can have the following orientation $q_p \leftarrow b \rightarrow y$, or $q_p \leftarrow b \leftarrow y$. This is because q_p must be a collider that is also a parent of y. Thus, b is always a non-collider.

Since every discriminating path for Y results in Y being a non-collider, this trivially satisfies the third condition of Theorem 8. The other two conditions are trivially met. ■

To prove the completeness of MD-PC algorithm, we next have the following lemma that characterizes the orientation of a triple (A, B, C). The proof is an extension of Lemma 1 [6].

Lemma 35 (Shielded Paths in $\mathcal{P}_{MD-PC}$) *In $\mathcal{P}_{MD-PC}$, the output of MD-PC algorithm, the following property holds: if $A \rightarrow B \circ\!\!-\!\!\circ C$, then $A \rightarrow C$.*

Proof Suppose by way of contradiction that there is a triple (A, B, C) such that the property does not hold even though $A \rightarrow B \circ\!\!-\!\!\circ C$. If A and C are not adjacent, then R1 of the MD-PC algorithm would orient $A \rightarrow B \circ\!\!-\!\!\circ C$ as $A \rightarrow B \rightarrow C$, reaching a contradiction.

If A and C were adjacent, but with $A \leftarrow C$, then R2 would orient $A \rightarrow B \circ\!\!-\!\!\circ C$ as $A \rightarrow B \leftarrow C$ thus reaching a contradiction.

Next, we consider all possible orientation rules that resulted in $A \rightarrow B$, but leave $B \circ\!\!-\!\!\circ C$.

A partial ordering from the vertices is aavailable where $X < Y$ if X is an ancestor of Y. Since $B \circ\!\!-\!\!\circ C$, we can consider B the minimum vertex order since A is adjacent to C.

Case 1: $A \rightarrow B$ is oriented via R1. Thus there is an edge $D \rightarrow A$ such that $D \notin adj(B)$. Thus there are edges $A \rightarrow B \circ\!\!-\!\!\circ C$. However, now A also satisfies the definition of a minimal vertex.

Case 2: $A \rightarrow B$ because it is part of an unshielded collider. In this case, there is an edge $D \rightarrow B$ such that $D \notin adj(A)$. If $D \notin adj(C)$, then $B \circ\!\!-\!\!\circ C$ would be oriented by R1. If $D \in adj(C)$, and $D \circ\!\!-\!\!\circ C$ unoriented, then $B \circ\!\!-\!\!\circ C$ is oriented via R3. If $D \circ\!\!-\!\!\circ C$ is oriented, and $D \rightarrow C$, then $B \circ\!\!-\!\!\circ C$ is oriented by R2, else if $D \leftarrow C$, then $A \circ\!\!-\!\!\circ C$ is oriented $A \leftarrow C$ by R1 and $B \circ\!\!-\!\!\circ C$ oriented by R2.

Case 3: $A \rightarrow B$ is oriented by R3. Then there is an unshielded collider colliding at B. This results in Case 2.

Case 4: $A \rightarrow B$ is oriented by R2. There would be a vertex D such that $A \rightarrow D$ and $D \rightarrow B$. $D \in adj(C)$ because otherwise $B \circ\!\!-\!\!\circ C$ oriented by R1. $D \circ\!\!-\!\!\circ C$ is oriented by construction with $D < B$. If $C \rightarrow D$, the $B \circ\!\!-\!\!\circ C$ is oriented by R2. Otherwise if $D \rightarrow C$, then $A \rightarrow C$ by R2 contradicting original hypothesis.

Case 5: A is a F-node, or S-node and oriented by R5' from Alg. 1, which would also result then in $A \rightarrow C$ contradicting the initial assumption.

Thus all cases that assume the hypothesis false all lead to contradiction. ■

Next, we prove the completeness of the MD-PC algorithm. This relies on results from [6].

Proof of Thm 11

Theorem 36 (MD-PC Completeness) *Let the tuple of distributions $\mathbf{P}$ be generated by an unknown pair $\langle G_S, \mathbf{\Psi} \rangle$ over domains $\mathbf{\Pi}$, then MD-PC is complete, i.e. $\mathcal{P}$ contains all common edge marks in the MD-Markov equivalence class.*

Proof ■

F.1.4 MD-Markov Property Results

In this section, we prove some results about the MD-Markov property. The first result proves that the MD-Markov property generalizes other interventional Markov properties.

Lemma 37 (MD-Markov property generalizes the Ψ-Markov property) *Let* $\mathbf{\Pi} = \{\Pi^1\}$ *and* $G = (\mathbf{V} \cup \mathbf{L}, \mathbf{E})$. *Say* $\mathbf{\Psi}^{\Pi}$ *and* $\mathbf{P}^{\Pi}$ *be an arbitrary set of interventions and distributions with* $\mathcal{K} = []$. *Given* $\mathcal{K}$, $\mathbf{P}^{\Pi}$ *satisfies the MD-Markov property with respect to* $\langle G_S, \mathbf{\Psi}^{\Pi} \rangle$, *then* $\mathbf{P}^{\Pi}$ *also satisfies the* Ψ*-Markov property with respect to* $\langle G, \mathbf{\Psi}^{\Pi} \rangle$.

Proof By assumption, we have only distributions with unknown intervention targets. Given $\mathcal{K}$, $\mathbf{P}^{\Pi}$ satisfies the MD-Markov property, so we will show that it also simultaneously satisfies the Ψ-Markov property. Moreover, since there is only one domain $\mathbf{V_S} = \emptyset$, the empty set.

$$\text{For } \mathbf{I}_i^j \in \mathbf{\Psi}^{\Pi}: \quad P_i^j(y|w,z) = P_i^j(y|w) \quad \text{if } Y \perp\!\!\!\perp Z|W \text{ in D}$$

is satisfied by the first condition of the MD-Markov property in Def. 3.

$$\text{For } \mathbf{I}_i, \mathbf{I}_j \in \mathbf{\Psi}^{\Pi}: \quad P_i(y|w,z) = P_j(y|w) \quad \text{if } Y \perp\!\!\!\perp Z|W \backslash W_K \text{ in } G_{\underline{W_K},\overline{R(W)}}$$

is satisfied by the second condition of the MD-Markov property. There is only a single domain, so there is by definition no S-node, and thus the condition reduces to the Ψ-Markov property condition two. Therefore, $\mathbf{P}^{\Pi}$ satisfies the Ψ-Markov property with respect to $\langle G, \mathbf{\Psi}^{\Pi} \rangle$. ■

Lemma 37 demonstrates that the MD-Markov property generalizes the Ψ-Markov property. Since the Ψ-Markov property itself has been shown to generalize the I-Markov and Global Markov property, we have the following corollaries.

Corollary 38 (MD-Markov property generalizes the I-Markov property) *Let* $\mathbf{\Pi} = \{\Pi^1\}$, $G = (\mathbf{V} \cup \mathbf{L}, \mathbf{E})$, $\mathbf{\Psi}^{\Pi}$ *be an arbitrary set of inteventions and* $\mathbf{P}^{\Pi}$ *an arbitrary set of distributions induced by* $\mathcal{Z}^{\Pi}$. *Let* $\mathcal{K}$ *be a vector of 1's, such that all distributions have a known intervention target. If* $\mathbf{P}^{\Pi}$ *satisfies the MD-Markov property with respect to* $\langle G_S, \mathbf{\Psi}^{\Pi} \rangle$, *then it also satisfies the I-Markov property with respect to* G. ■

Corollary 39 (MD-Markov property generalizes the Markov property) *Let* $\mathbf{\Pi} = \{\Pi^1\}$, $G = (\mathbf{V} \cup \mathbf{L}, \mathbf{E})$, $\mathbf{\Psi}^{\Pi} = \langle \{\}^1 \rangle$ *and* $\mathbf{P}^{\Pi}$ *an arbitrary set of distributions. If* $\mathbf{P}^{\Pi}$ *satisfies the MD-Markov property with respect to* $\langle G, \mathbf{\Psi}^{\Pi} \rangle$, *then it also satisfies the Markov property with respect to* G. ■

Example 30 (Markov vs MD-Markov property) *Let* G *be the causal diagram in Figure 1(b). For an arbitrary set of interventions set* $\mathbf{\Psi}^{\Pi}$, *we have that* $(X \perp\!\!\!\perp Z|Y)_G$ *implies that* $P_j^i(Z|Y,X) = P_j^i(Z|Y)$ *for all* $\Pi^i \in \mathbf{\Pi}$ *and distributions. Thus, the MD-Markov property includes the Markov property invariance. However, the Markov property does not capture other invariances that are presented in Def. 3.* ■

Example 31 (Ψ-Markov vs MD-Markov property) *Let* G *be a selection diagram as shown in Figure 1(b). Let* $\mathbf{\Psi}^{\Pi} = \langle \{\}^1, \{\}^2, \{X\}^1, \{Y\}^1 \rangle$ *and* $\mathcal{K} = [1,1,0,0]$ *for a corresponding* $\mathbf{P}^{\Pi}$. *The MD-Markov property states that there is an invariance* $P_1^1(z|y) = P_2^1(z|y) = P_3^1(z|y)$. *The I-Markov states the equivalence between* $P_1^1(z|y) = P_3^1(z|y)$ *and the* Ψ*-Markov property* $P_1^1(z|y) = P_2^1(z|y)$. ■

We defined a joint selection diagram in Definition **??**. Here, we show that there is no information loss when we construct the joint selection diagram, which is easier to analyze. The joint selection diagram (as defined in Definition **??**) is a valid representation of a collection of selection diagrams stemming from different domains. Thus we refer to joint selection diagrams as selection diagrams in the main paper.

Lemma 40 (Joint selection diagrams are valid representations) *A joint selection diagram preserves transportability phenomena. That is, if a causal effect is transportable in the non-joint selection diagram if and only if it is transportable in the joint selection diagram.*

Proof Since S-nodes are defined as pointing out of S-nodes by construction, then in the joint selection diagram they can act as "confounders" when viewed graphically. Define A as the node that an S-node points to originally. An S-node by definition only has additional edges if there is an inducing path between the A and another node B. If such a path exists, then there is an unblockable subpath from A to B and conditioning on the S-node would not change the m-separation statements. ■

In Definition 3, we define the Markov property for a (joint) selection diagram. The MD-Markov property generalizes the Markov property and extends the conditions of d-separation (condition i) and distributional invariances (condition ii) to selection diagrams. Condition ii is no longer a conditional independence statement, but rather a different type of invariance [83]. Note that compared to the Ψ-Markov property, there are some subtle differences. Namely, there is always the question of whether or not nodes are d-separated with respect an S-node. D-separation with respect to an S-node representing a pair of domains allows one to map invariances across those two domains.

To prove Thm. 13, we first prove a few useful lemmas. The first lemma relates m-separation statements with a conditioning set of S-nodes to other m-separation statements that contain "more" S-nodes.

Lemma 41 (M-separation statements can arbitrarily add S-node singletons) *Let G be the joint selection diagram with respect to a causal Bayesian network with latents, $G = (V \cup \mathbf{S}, E \sup E_S)$. Consider m-separation statement with respect to G with $X \perp\!\!\!\perp Y | Z, S_i$ where $X, Y \subset V \cup \mathbf{S}$ and $Z \subseteq V - \{X, Y\}$ and $S_i \subset \mathbf{S} - \{X, Y\}$ (that is S_i is a set of S-nodes).*

For any $S_i \in \mathbf{S} - (S \cup \{X, Y\})$, the following statements are equivalent:

1. *$X \perp\!\!\!\perp Y | Z, S_i$ in G*
2. *$X \perp\!\!\!\perp Y | Z, S_i \cup \{S_j\}$ in G and ($S_j \perp\!\!\!\perp Y | Z, S_i$ or $S_j \perp\!\!\!\perp X | Z, S_i$)*

Proof The first statement states that X and Y are d-separated given Z and the ith set of S-nodes in the joint selection diagram.

The second statement states that if we augment the m-separation statement with a conditioning set of the jth S-node, then either the jth S-node is m-separated from Y given Z and S_i or the jth S-node is m-separated from X given Z and S_i.

We show the equivalence of the m-separation statements by analyzing the paths that are m-connecting.

We are given that $X, Y \neq S_j$ and $S_j \notin S_i$. Suppose that there is a m-connecting path between X and Y given Z and S_i (the converse of the first statement in the lemma). Either it passes through S_j S-node or it does not.

If it does not pass through S_j, then since all S_j are oriented out of S_j, then $X \not\perp\!\!\!\perp Y | Z, S_i \cup \{S_j\}$ in G.

If it does pass through S_j, then there are two m-connecting paths that lead from X to S_j given Z and S_i and from S_j to Y given Z and S_i.

If there are no m-connecting paths between X and Y given Z and S_i, then all the paths have to be m-separating. ■

Next, we show that when there is a difference in m-separation statements between two selection diagrams, these can be mapped to m-separation statements from U, O, or T, sets that are defined as follows:

We define the following sets of m-separation statements:

$$U = \{(X \perp\!\!\!\perp Y|Z,S)_G \ : X,Y \in V \cup \mathbf{S}, Z \subseteq V - \{X,Y\},\ S \subset \mathbf{S} - \{X,Y\}$$
$$O = \{(X \perp\!\!\!\perp Y|Z,S)_G \ : X,Y \in V \cup \mathbf{S}, Z \subseteq V - \{X,Y\},\ S = \mathbf{S} - \{X,Y\}$$
$$T = \{(X \perp\!\!\!\perp Y|Z,S)_G \ : X \in V, Y \in V \cup \mathbf{S}, Z \subseteq V - \{X,Y\},\ S = \mathbf{S} - \{X,Y\}$$

Intuitively, U, O and T are m-separation statement sets that contain all possible sets of m-separation statements inside a MAG.

Lemma 42 (Arbitrary differences in m-separation statements induce a difference in U, O, or T) *Let $G_1 = (V \cup \mathcal{S}, E_1 \cup E_S\}$ and $G_2 = (V \cup \mathcal{S}, E_2 \cup E_S\}$ be selection diagrams over the same sets of variables V. Suppose X, Y, Z are disjoint subsets of $V \cup \mathcal{S}$.*

$X \perp\!\!\!\perp Y|Z$ in G_1, $X \not\perp\!\!\!\perp Y|Z$ in G_2, then at least one of the following is true:

i) there exists $X, Y, Z \subseteq V$ such that $X \perp\!\!\!\perp Y|Z, \mathcal{S}$ in G_1 and $X \not\perp\!\!\!\perp Y|Z, \mathcal{S}$

ii) There exists $A, B \subseteq V$ and $S_i \in \mathcal{S}$ such that $(S_i \perp\!\!\!\perp A|B, \mathcal{S}\backslash S_i)$ in G_1 and $(S_i \not\perp\!\!\!\perp A|B, \mathcal{S}\backslash S_i)$ in G_2

In other words: Any difference in m-separation statement from the set of statements $U \cup O \cup T$ between G_1 and G_2 can be stated as just a difference between m-separation statements in T between G_1 and G_2.

Proof Given m-separation statement in U, we can write these as m-separation statements in O. This is done by repeatedly applying Lemma 41 to m-separation statements in U until all m-separation statements lie in O.

Now, we prove that all m-separation statements in O that are not in T can be mapped to T. First, note that T is a subset of O, since there is the additional constraint that $X \in V$, rather than $X \in V \cup \mathbf{S}$.

Define $W = T\backslash O = \{(X \perp\!\!\!\perp Y|Z,S)_G \ : \ X \in \mathbf{S}, Y \in \mathbf{S}, Z \subseteq V - \{X,Y\}$ as the set of m-separation statements that are in O, but not in T. These are m-separation statements then between S-nodes of the selection diagram. We consider any m-separation statement where $S_i \perp\!\!\!\perp S_j|Z, \mathbf{S} - \{S_i, S_j\}$ in G_1, but $S_i \not\perp\!\!\!\perp S_j|Z, \mathbf{S} - \{S_i, S_j\}$ in G_2.

S-nodes are by Definition 1 pointing out, there must be at least one collider along paths between S_i, S_j. First we consider a path that is active in G_2, but not in G_1. If $S_i \perp\!\!\!\perp S_j|Z, \mathbf{S} - \{S_i, S_j\}$ in G_1, but $S_i \not\perp\!\!\!\perp S_j|Z, \mathbf{S} - \{S_i, S_j\}$ in G_2 for some $Z \subset V$, then this can only happen if in G_2, there exists a node in Z that is a descendent of both S_i and S_j. That is $v \in Z$ such that $v \in Desc(S_i) \cap Desc(S_j)$, which makes the collider active in G_2. In G_1, we have simultaneously that for all nodes in Z, there does not exist any descendants of both S_i and S_j. That is $\nexists v \in Z$ such that $v \in Desc(S_i) \cap Desc(S_j)$. This then means that v is either not a descendant of S_i, or it is not a descendant of S_j. Suppose WLOG that v is not a descendant of S_i.

Then this implies that $S_i \perp\!\!\!\perp v|\mathbf{S} - \{S_i\}$ in G_1, and $S_i \not\perp\!\!\!\perp v|\mathbf{S} - \{S_i\}$ in G_2.

Now, suppose $(X \perp\!\!\!\perp Y|Z)_{D_1}$ and $(X \not\perp\!\!\!\perp Y|Z)_{D_2}$. Any m-separation statment belongs to one of the sets O, U or T. Since G_1 and G_2 share the same vertex set, V, then the m-separation statement would be in the same set.

If this m-separation statement set belongs to T, then we are done.

If it belongs to O, then by our earlier result, any m-separation statement with differences imply an m-separation statement difference in T and the result follows.

If it belongs to U, then by Lemma 41 and the above, the m-separation statement can be mapped to m-separation statements in O. Then by the previous statement, the result follows.

This proves the lemma. ■

M-separation given S-node set can be arbitrarily given by the entire S-node set

Lemma 43 (M-separations given a subset of S-nodes implies m-separation with respect to any other S-n
Let $Aug_\Psi(G_S)$ be the augmented selection diagram with respect to a selection diagram, G_S and tuple of interventions $\mathbf{\Psi}$ defined over domains $\mathbf{\Pi}$. Consider any m-separation statement with respect to $Aug_\Psi(G_S)$ of the form:

$$(X \perp\!\!\!\perp Y | Z, S_A, \mathcal{F})_{Aug_\Psi(G_S)}$$

where $X, Y \in \mathbf{V} \cup \mathcal{F} \cup \mathbf{S}$, $Z \subseteq \mathbf{V} - \{X, Y\}$ and $S_A \subsetneq S - \{X, Y\}$. Then for any $S_i \in S - (S_A \cup \{X\} \cup \{Y\})$, the following two statements are equivalent:

- $(X \perp\!\!\!\perp Y | Z, \mathcal{F}, S_A)_{Aug_\Psi(G_S)}$
- $(X \perp\!\!\!\perp Y | Z, \mathcal{F}, S_A \cup \{S_i\})_{Aug_\Psi(G_S)}$ *AND* $[(S_A \perp\!\!\!\perp Y | Z, \mathcal{F}, S_A)_{Aug_\Psi(G_S)}$ *or* $(S_A \perp\!\!\!\perp X | Z, \mathcal{F}, S_A)_{Aug_\Psi(G_S)}]$

Proof

■

This lemma demonstrates that any m-separation statements given a subset of S-node between X, Y chosen from the entire set of possible nodes in the augmented selection diagram can equivalently add any subset of the other S-nodes. In addition, these other S-nodes are in fact also m-separated from either X, Y, or both.

F.1.5 Results from Section 4.1 on observational multi-environment Markov equivalence

In the following results leading up to Theorem 13, we assume that we only have access to observational data across multiple domains. In this setting, the MD-Markov property and the relevant graphical MD-Markov equivalence properties are much simpler. We show that there is a mapping at this point between the MD-Markov property and the Ψ-Markov property [7]. We are ready to prove an equivalent graphical condition for MD-Markov equivalence.

Theorem 44 (Graphical MD-Markov Equivalence Among Selection Diagrams With Only Observational I
Let G^1 and G^2 be two causal diagrams. Let $G^1_S = (V \cup L_1 \cup S_1, E_1)$ and $G^2_S = (V \cup L_2 \cup S_2, E_2)$ be their corresponding selection diagrams over N environments with S-nodes $\mathbf{S_1}$, $\mathbf{S_2}$, $\mathbf{\Pi} = \{\Pi^1, ..., \Pi^N\}$, with interventions $\mathbf{\Psi}^\Pi = \langle \{\}^1, \{\}^2, ..., \{\}^N \rangle$ and associated distributions $\mathbf{P}^\Pi$. Let $\mathcal{K} = [1, 1, ..., 1]$ be the vector of known interventions.

We say $\langle G^1_S, \mathbf{\Psi} \rangle$ and $\langle G^2_S, \mathbf{\Psi} \rangle$ are MD-Markov equivalent if and only if for $M_1 = MAG(Aug_\Psi(G^1_S)$ and $M_2 = MAG(Aug_\Psi(G^2_S)$:

1. *M_1 and M_2 have the same skeleton;*
2. *M_1 and M_2 have the same unshielded colliders;*
3. *If a path p is a discriminating path for a node Y in both M_1 and 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.*

Proof (=>) Assuming that $MAG(D_1)$ and $MAG(D_2)$ satisfy the three conditions, we will show they are MD-Markov equivalent. Then by Definition 27 and Proposition 28, the two MAGs have the same m-separation statements and vice versa, thereby satisfying the MD-Markov equivalence condition, where both G^1_S and G^2_S impose the same constraints over the set of distributions defined in 27.

(<=) We prove this direction by contradiction. Suppose $MAG(D_1)$ and $MAG(D_2)$ do not satisfy the three conditions, then we want to show that the two graphs are not MD-Markov equivalent.

By definition of a MAG, if the two MAGs have one of the conditions different, then there is at least one different m-separation statement. Without loss of generality, we consider only m-separation statements among pairs of singletons. If an m-separation statement holds between arbitrary sets of nodes in one selection diagram, G_1, but not G_2, then there is at least one pair of singletons where the m-separation statement differs between G_1 and G_2.

Consider the sets U, O and T again of m-separation statements in Lemma 42. U, and O, are m-separation statements between any two nodes given a strict subset of all remaining S-nodes, all remaining S-nodes. T is the set of m-separation statements between normal nodes and any other node given all remaining S-nodes.

By Definition 3, an m-separation statement is in T if and only if it appears in the MD-Markov equivalence class of distributions for G.

By Lemma 42, we show that if the two MAGs of the selection diagram are not Markov equivalent, then there is a m-separation statement in the definition of MD-Markov equivalence that is different in the two graphs. As a result, we are able to show that there is a tuple $\langle G^2_S, \mathbf{\Psi}, \mathbf{S}\rangle$ that contains tuples of distributions $\mathbf{P}$ that is not MD-Markov with respect to $\langle G^2_S, \mathbf{\Psi}, \mathbf{S}\rangle$. ■

Thus, graphically, the two selection diagrams over multiple domains of only observational data are Markov-equivalent if the MAGs of their augmented diagrams fulfill certain similarity constraints.

Proof of Main Text Thm. 13 [Equivalence of Ψ and MD-Markov property given multi-domain observational distributions] Next, we state an equivalence between the Ψ-Markov characterization in Theorem 13 of [7] and MD-Markov characterization.

Theorem 45 (Equivalence of Ψ and MD-Markov property given multi-domain observational distributions) *Let G be a causal diagram and $G_S = (\mathbf{V} \cup \mathbf{L} \cup \mathbf{S}, \mathbf{E} \cup \mathbf{E_S})$ the selection diagram over N domains $\mathbf{\Pi} = \{\Pi^1, \Pi^2, ...\Pi^N\}$. Let $\mathbf{S}^\Pi$ be set of S-nodes and $\mathbf{E_S}$ their set of edges. Let $\mathbf{\Psi}^\Pi = \langle \{\{\}\}^1, ..., \{\{\}\}^N\rangle$ and $\mathcal{K} = [1, 1, ..., 1]$, such that for each of the N domains, there is only observational data. Let $\mathbf{P}^\Pi$ be an arbitrary set of distributions. If $\mathbf{P}^\Pi$ satisfies the Ψ-Markov property with respect to $\langle G, \mathbf{\Psi}, \mathbf{S}\rangle$, then it also satisfies the MD-Markov property with respect to $\langle G, \mathbf{\Psi}, \mathbf{S}\rangle$.*

Proof If $\mathbf{P}$ satisfies the Ψ-Markov property, then for disjoint $Y, Z, W \subseteq V$ the condition related to d-separation of is held for each distribution in the joint selection diagram given the shared causal structure assumption.

For the second condition relating pairs of distributions to each other in the Ψ-Markov property, we know that this is equivalent to d-separation in the augmented graph with the augmented graph nodes added from pairs of different distributions given the Definition 5. In our case, each pair of distibutions correspond to a pair of different domains, and thus the augmented F-node has a similar meaning to the S-node.

Let Z be a S-node (that is represented by an F-node in the augmented graph) and $Y \perp\!\!\!\perp Z|W, S_{[N]\setminus[i]}$. This then shows that if the Ψ-Markov property holds, then the MD-Markov property holds. Similarly if the MD-Markov property holds with respect to $\langle G, \mathbf{\Psi}, \mathbf{S}\rangle$, then it implies the Ψ-Markov property with respect to $\langle G, \mathbf{\Psi}\rangle$. ■

This proves the result stated in Thm. 13 and shows that the Ψ-Markov property implies the MD-Markov property in the case where only observational data is present in multiple domains. This can be seen conceptually that the domain change can be viewed as an intervention on the data distributions with unknown targets (i.e. we do not know where the environment targets). In fact they are equivalent in this setting.

Example 32 *Let G_S be the selection diagram in Figure 1(b), among two domain $\mathbf{\Pi} = \{\Pi^1, \Pi^2\}$. Let $\mathbf{S}^{\Pi} = \{S_x^{1,2}\}$ and $\mathbf{\Psi}^{\Pi}$ and $\mathcal{K}$ be defined with just observational distributions from domains 1 and 2. Consider an arbitrary $\mathbf{P}$ that satisfies the Ψ-Markov property with respect to $\langle G, \mathbf{\Psi}\rangle$. This implies the distribution of Z is the same between domains 1 and 2 through the invariance $P^1(Z) = P^2(Z)$. This is the only invariance that is required. Observe that is also the only invariance required by the MD-Markov property and thus $\mathbf{P}$ satisfies the MD-Markov property with respect to $\langle G_S, \mathbf{\Psi}^{\Pi}\rangle$.* ■

Corollary 46 (An equivalence of MD-Markov Equivalence and Ψ-Markov Equivalence) *Let G be a causal diagram and $G_S = (\mathbf{V} \cup \mathbf{L} \cup \mathbf{S}, \mathbf{E} \cup \mathbf{E_S})$ the selection diagram over N domains $\mathbf{\Pi} = \{\Pi^1, \Pi^2, ...\Pi^N\}$. Let $\mathbf{S}^{\Pi}$ be set of S-nodes and $\mathbf{E_S}$ their set of edges. Let $\mathbf{\Psi}^{\Pi} = \langle \{\{\}\}^1, ..., \{\{\}\}^N\rangle$ and $\mathcal{K} = [1, 1, ..., 1]$, such that for each of the N domains, there is only observational data. Let $\mathbf{P}^{\Pi}$ be an arbitrary set of distributions. $\mathbf{P}^{\Pi}$ satisfies the Ψ-Markov property with respect to $\langle G, \mathbf{\Psi}, \mathbf{S}\rangle$, if and only if it satisfies the MD-Markov property with respect to $\langle G, \mathbf{\Psi}, \mathbf{S}\rangle$.*

Proof The result follows from 6 and 45. ■

This proves an interesting equivalence mapping between multi-domain observational setting and single-domain unknown interventional setting. One can view the change in domain as an unknown intervention that occurs via nature. However, knowing the domain change is still import information as not only does nature induce an intervention, but there may also be various interventional datasets collected explicitly in the domain. Thus one would know that these interventions in this domain are different from similar interventions in another domain. Note there are a few subtle differences that one should be aware of.

1. In the case of Ψ-Markov equivalence, one works with an augmented graph with the symmetric difference between all pairs of different intervention target sets. The sets of variables from the symmetric difference of intervention targets form the "F-nodes" of the augmented graph, which can then be viewed analagously to S-nodes. In MD-Markov equivalence, one has S-nodes on all variables that have differences between the source and target domains. These S-nodes represent possible distribution differences between pairs of domains regardless of whether or not the distributions are observational or interventional. Thus S-nodes can be seen as "nature's intervention" that is always present.

2. S-nodes represent a difference in distribution between the source and target domain at the nodes it points to. An F-node represents a difference between a pair of distributions due to a symmetric difference in intervention target. These are important subtleties, which allow the user to utilize such qualitative information in downstream transportability ID tasks [44]. The extra information that comes from the knowledge that each observational distribution comes from a different environment manifests purely in the interpretation of the nodes. However, transportability ID in an EC is still an open problem, and thus it is unclear how to leverage the results of the learning algorithm.

Based on this equivalence of MD-Markov and Ψ-Markov property for multi-domain observational data, the Algorithm S-FCI introduced in this paper is sound and complete. That is every adjacency and orientation is common for all $MAG(G')$ where G' is a selection diagram MD-Markov equivalent to G. Moreover the recovered graph is the most informative it can be (i.e. discovers as many tails and arrowheads that can be oriented within a MD-Markov equivalence class).

Corollary 47 (Modified Ψ-FCI algorithm to learn an S-PAG) *Define the modified Ψ-FCI algorithm with two modifications: i) represent S as the set of intervention targets and ii) take the graph learned and remove all S-nodes that represent a pairing between distributions from two target domains. The modified Ψ-FCI algorithm is complete for learning an S-PAG given only observational data.*

Proof If we run Ψ-FCI, with the S-nodes represented as our intervention targets, then we will learn a supergraph of the graph of interest. The supergraph will contain extra F-nodes due to symmetric differences among the combinations of source domains. By removing those, we have a Ψ-PAG with only F-nodes representing the source and target domain, which is the S-PAG. ■

F.1.6 Results from Section 4.2 obs. + interv. data in multiple domains

In this section, we prove results related to causal discovery in the setting of multiple domains with observational and interventional data. First, we show some equivalence relations when going from the non-augmented graph to the augmented graph.

Proposition 48 (augmented graph Equivalence Relations) *Let $G = (\mathbf{V} \cup \mathbf{L} \cup \mathbf{S}, \mathbf{E} \cup \mathbf{E_S})$ be a joint selection diagram, with latent variables L and its augmented graph $Aug_{\mathbf{\Psi},\mathbf{S}}(G) = (\mathbf{V} \cup \mathbf{L} \cup \mathbf{S} \cup \boldsymbol{\mathcal{F}}, \mathbf{E} \cup \mathbf{E_S} \cup \boldsymbol{\mathcal{E}})$ with respect to the intervention set across all N domains $\mathbf{\Pi}$, where $\mathcal{F} = \{F_i^j\}_{i\in[k]}^{j\in[N]}$. Let A_{ij} be the set of nodes adjacent to F_i^j for all $i \in [k]$ and all $j \in [N]$. And denote B_i as the set of nodes adjacent to $S^{i,j} \in \mathbf{S}$. We have the following equivalence relations:*

For disjoint $Y, Z, W \subseteq V$, we have:

$$(Y \perp\!\!\!\perp Z|W)_G \iff (Y \perp\!\!\!\perp Z|W, F_{[k],[N]})_{Aug_{\mathbf{\Psi},\mathbf{S}}(G)} \tag{9}$$

For each A_{ij} suppose $Y, W \subseteq V \backslash G_{ij}$, we have:

$$(Y \perp\!\!\!\perp A_{ij}|W)_{G_{\underline{A_{ij}}}} \iff (Y \perp\!\!\!\perp F_{ij}|W, A_{ij}, F_{[k]\backslash\{i\}}^{[N]\backslash\{j\}})_{Aug_{\mathbf{\Psi},\mathbf{S}}(G)} \tag{10}$$

$$(Y \perp\!\!\!\perp A_{ij}|W)_{G_{\underline{A_{ij}(W)}}} \iff (Y \perp\!\!\!\perp F_{ij}|W, F_{[k],[N]})_{Aug_{\mathbf{\Psi},\mathbf{S}}(G)} \tag{11}$$

For each A_{ij}, let $Y, W \subseteq V$, and let $W_{ij} = W \cap A_{ij}, R = A_{ij} \backslash W_{ij}$, then:

$$(Y \perp\!\!\!\perp A_{ij}|W\backslash W_{ij})_{G_{\underline{A_{ij}}}} \iff (Y \perp\!\!\!\perp F_{ij}|W, A_{ij}, F_{[k]\backslash\{i\},[N]\backslash\{k\}})_{Aug_{\mathbf{\Psi},\mathbf{S}}(G)} \tag{12}$$

For each $S_i \in S$, let $Y, W \subseteq V$, then

$$(Y \perp\!\!\!\perp S^{i,j}|W)_G \implies (Y \perp\!\!\!\perp B_{ij}|W, S^{[N]\times[N]})_{Aug_{\mathbf{\Psi},\mathbf{S}}(G)} \tag{13}$$

Proof Conditioning on a source node is equivalent to removing it from the graph in terms of the graph separation statements. Hence, conditioning on $F_{[k]\backslash i,[N]\backslash\{j\}}$ in the right-hand side eliminates them. Therefore, equations 9, 10, 11 and 13 follow from [[64], Proof of Th. 4.1] by Pearl.

Note that $S_j \perp\!\!\!\perp F_{ij}$ for all $i \in [k], j \in [N]$ because S-nodes and F-nodes in this setting are source nodes and thus will always have a collider due to the multi-domain intervention assumption.

The rest of the proof is exactly as it is in [Proposition 3 of [3]]. ■

The proof for Prop. 6 follows.

Proof of Main Text Proposition 6 [Graphical MD-Markov Property]

Proposition 49 (Graphical MD-Markov Property) *Consider the multi-domain setup 1.1. Let* $M = MAG(Aug_{\Psi}(G))$ *and let* $[M]$ *be the set of S-MAGs corresponding to all the tuples* $\langle G'_S, \mathbf{\Psi'}^{\Pi} \rangle$ *that are MD-Markov equivalent to* $\langle G_S, \mathbf{\Psi}^{\Pi} \rangle$*. The S-PAG for* $\langle G, \mathbf{\Psi}^{\Pi} \rangle$*, denoted* $\mathcal{P}$ *is a graph such that:*

$$(Y \perp\!\!\!\perp Z|W)_{G_S} \iff (Y \perp\!\!\!\perp Z|W, F_{\mathcal{E}}, \mathbf{S})_{Aug_{\Psi}(G)} \tag{14}$$

$$(Y \perp\!\!\!\perp \{S^{j,k}, \mathbf{K_i^{j,k}}\}|W \backslash W_i^{j,k})_{G_{S_{\underline{W_i^{j,k}}, \overline{R(W)}}}} \iff$$
$$(Y \perp\!\!\!\perp \{S^{j,k}, F_i^{j,k}\}|W, \mathbf{S} \backslash S^{j,k}, F_{\mathcal{E}} \backslash F_i^{j,k})_{Aug_{\Psi}(G)} \tag{15}$$

where $\mathbf{W}_i^{j,k} = \mathbf{W} \cap \mathbf{K}_i^{j,k}, \mathbf{R} = \mathbf{K}_i^{j,k} \backslash \mathbf{W}_i^{j,k}$.

Proof The proof follows from Proposition 48. ■

We see that graphical equivalence is nicely modular using the augmented graph framework, where we add nodes indicating change in distributions due to domain, or interventions. Similarly, since the augmented graph is still a DAG and the MAG of the augmented graph is a MAG, and the corresponding PAG of the augmented graph is a PAG, we can leverage existing theory that analyzes properties of those graphs. Next, we prove Thm. 8 showing a graphical criterion for determining the MD-Markov equivalence among two graphs.

Proof of Main Text Thm. 8 [MD-Markov Characterization]

Theorem 50 (MD-Markov Characterization) *Let there be two causal graphs* $G^1 = (\mathbf{V} \cup \mathbf{L}_1, \mathbf{E}_1)$, $G^2 = (\mathbf{V} \cup \mathbf{L}_2, \mathbf{E}_2)$ *with* G_S^1 *and* G_S^2 *the selection diagrams and a corresponding set of intervention targets,* $\mathbf{\Psi_1}^{\Pi}$, $\mathbf{\Psi_2}^{\Pi}$*, a corresponding set of S-nodes set* $\mathbf{S_1}^{\Pi}, \mathbf{S_2}^{\Pi}$ *and a fixed index vector of known intervention targets* $\mathcal{K}$*. Assume that the symmetrical difference sets are indexed in both sets in the same pattern such that correspondence between F-nodes and S-nodes are the same in* M_1 *and* M_2*. Then* $\langle G_S^1, \mathbf{\Psi_1}^{\Pi}, \mathbf{S_1}^{\Pi} \rangle$ *and* $\langle G_S^2, \mathbf{\Psi_2}^{\Pi}, \mathbf{S_2}^{\Pi} \rangle$ *are MD-Markov equivalent if and only if for* $M_1 = MAG(Aug_{\mathbf{\Psi_1}, \mathbf{S_1}}(G^1)$ *and* $M_2 = MAG(Aug_{\mathbf{\Psi_2}, \mathbf{S_2}}(G^2)$*:*

1. M_1 *and* M_2 *have the same skeleton*
2. M_1 *and* M_2 *have the same unshielded colliders*
3. *If a path* p *is a discriminating path for a node* Y *in both* M_1 *and* M_2*, then* Y *i s a collider on the path in one graph if and only if it is a collider on the path in the other.*

Proof We proved a similar version earlier in Thm. 44 for only multi-domain observational data.

(<=) Suppose the two MAGs, M_1, M_2 satisfy the three conditions stated. Then, they induce the same m-separation statements [78]. Therefore, by Prop. 6, G_1 and G_2 impose the same constraints over the distributions in the MD-Markov property definition (Def. 3). Therefore, $S_{\mathcal{K}}^{\Pi}(G_1) = S_{\mathcal{K}}^{\Pi}(G_2)$.

(=>) Suppose by way of contradiction that the two MAGs do not satisfy the three conditions. Then at least one different m-separation statement is present, since the MAGs encode m-separation statements. With a different m-separation statement, we want to show they are also therefore not MD-Markov equivalent.

In Lemma 42, we demonstrated that any m-separation statement is included in the defined sets $U \cup O \cup T$. Therefore, there is an m-separating path in one graph that is m-connecting in the other. In the final step, we demonstrate that the distribution tuple of Def. 3 is different in $Aug(G_1)$ vs $Aug(G_2)$.

We do this by construction.

Suppose $X, Y, Z \subseteq V$ such that $(X \perp\!\!\!\perp Y | Z, \mathcal{F}, \mathbf{S})_{AugG_1}$ and $(X \not\perp\!\!\!\perp Y | Z, \mathcal{F}, \mathbf{S})_{AugG_2}$. Any tuple of distributions (observational or interventional) across any domain obtained is faithful to the selection diagram with latent variables will suffice to demonstrate the proof.

Suppose $X = F_i^j$ for some $i \in [k]$ and $Y \in V$. Therefore an F-node is m-connected to an observed variable in $Aug(G_2)$ but not in $Aug(G_1)$. Now, consider $G_{path} = (V_{path}, E_{path})$, the subgraph of G_2 that includes all the variables that contribute to the m-connecting path of $(X \not\perp\!\!\!\perp Y | Z, \mathcal{F}, \mathbf{S})_{Aug(G_2)}$.

Consider now a jointly Gaussian distribution, p_{path}, on V_{path} that is faithful to G_{path}. Thm. 7 of [79] shows that this is possible.

We proceed now by considering two interventions I, J on the graph where $I \Delta J = A_i$, where the distributions p_I, p_J from the same domain are responsible for the graphical separation of F_i^j. Different from the rest of the paper, for this proof we will treat F_i^j as a regime variable that indicates when we switch to p_I and when we switch to p_J. Note that we can do this since we only add this single F node and no others in this domain j. Consider the distribution $p*$ defined as follows: $p*(.|F_i^j = 0) = p_I(.), p*(.|F_i^j = 1) = p_J$.

We will now show that the variable F_i^j is dependent with Y given Z on the distribution $p*$. So, we construct a SCM that induces an interventional distribution and the relevant graph in question.

Consider the following linear SCM: $\mathbf{x} = \mathbf{A}\mathbf{x} + \mathbf{e}$, where A is a lower-triangular matrix that captures the DAG structure and parent-child relationships in G_{path} and $\mathbf{e} \in \mathbb{R}^d$ is an exogenous noise vector and d is the number of observed variables in the graph. Let p_I be the distribution obtained by adding noise vector $\mathbf{e}_I$ to the system. $\mathbf{e}_I$ is non-zero in the rows corresponding to the nodes that it perturbs. Therefore p_I is a valid soft-interventional distribution. Let $\mathbf{e}_J$ be the noise vector now for adding an intervention on J.

Next, we show that every adjacent variable is dependent. The correlation of variables in G_{path} is computed as:

$$\mathbf{x} = \mathbf{A}\mathbf{x} + \mathbf{e} + \mathbf{e}_I \implies (\mathbf{I} - \mathbf{A})\mathbf{x} = \mathbf{e}_1 \implies \mathbf{x} = (\mathbf{I} - \mathbf{A})^{-1}\mathbf{e}_1$$
$$\mathbf{x} = \mathbf{A}\mathbf{x} + \mathbf{e} + \mathbf{e}_J \implies (\mathbf{I} - \mathbf{A})\mathbf{x} = \mathbf{e}_2 \implies \mathbf{x} = (I - A)^{-1}\mathbf{e}_2$$

with $\mathbf{e}_I = \mathbf{e} + \mathbf{e}_1$ and $\mathbf{e}_J = \mathbf{e} + \mathbf{e}_2$. Note when $\mathbf{e}_1$ and $\mathbf{e}_2$ are different, then the F-variable is dependent with the variables in $K := I \Delta J$, since $p(K|F = 0) \neq p(K|F = 1)$, implying $(K \not\perp\!\!\!\perp F)_{p*}$. We can compute the correlation matrix between observed variables with respect to $p*(.)$, since the binary regime variable can be marginalizd out:

$$E[\mathbf{x}\mathbf{x}^T] = 0.5(\mathbf{I} - \mathbf{A})^{-1} E[\mathbf{e_1}\mathbf{e_1}^T](\mathbf{I} - \mathbf{A})^{-1^T} + 0.5(\mathbf{I} - \mathbf{A})^{-1} E[\mathbf{e_2}\mathbf{e_2}^T](\mathbf{I} - \mathbf{A})^{-1^T} \tag{16}$$
$$= 0.5\mathbf{A})^{-1}(\mathbf{D_1} + \mathbf{D_2})(\mathbf{I} - \mathbf{A})^{-1^T} \tag{17}$$

where $\mathbf{D}_i = E[\mathbf{e_i}\mathbf{e_i}^T]$ are the diagonal covariance matrices of the noise that is added due to the interventions. Now, consider two adjacent variables $x_i, x_j \in V_{path}$. Since x_i, x_j are jointly Gaussian, they are dependent if and only if they are correlated. Therefore, we want to show $E[x_i x_j] \neq 0$ for any arbitrary adjacent pairs. By assumption, any pair of adjacent variables are dependent since the original distribution is chosen to be faithful to the graph G_{path}. Therefore, if we randomly pick variances of the added noise terms, with probability 1, any adjacent pairs of variables will be dependent still after a union bound.

Therefore, in the graph G_{path} plus the F-variable for interventions specifically in domain j, every pair of adjacent variables are dependent. Then, using Meek's Lemma 30, we have that $(F_i^j \not\perp\!\!\!\perp Y | Z)_p$. Since we have not added any other F-variables in this domain as regime variables, we do not need to condition on them. Now, we can augment this distribution to cover variables outside G_{path}. Pick all remaining variables independent from the variables in G_{path} and construct interventional distributions by adding extra noise terms to the intervened variables. Note, even with different

domains, we only need to construct a distribution valid within that specific domain. That is, being adjacent to an F-node not associated with the particular domain in question, is irrelevant.

We can repeat the same procedure as described for S-nodes that are now adding extra noise terms to the nodes in which it changes the distribution due to the domain. We simply pick one of the observational distributions as a reference and fix it. Then we arbitrarily add noise terms to each node that has an S-node that perturbs the distribution with respect to another domain. We continue until all domains have an associated distribution.

All that is left is to account for the case, where we are comparing interventions between different domains. That is, we must account for simultaneously a S-node and F-node change in regime. Without loss of generality, we can consider the case of just two domains, i and j and interventions I from domain i and J from domain j. In this case, we can define $\mathbf{e}_1 = \mathbf{e} + \mathbf{e}_I + \mathbf{e}_S$ and $\mathbf{e}_2 = \mathbf{e} + \mathbf{e}_J + \mathbf{e}_S$, where $\mathbf{e}_S$ is the noise vector added due to the change in domain. It is defined with non-zero values at the rows of nodes that are affected by the S-node $S^{i,j}$. Since $\mathbf{e}_S$ is constant between both $\mathbf{e}_1$ and $\mathbf{e}_2$, we can simply redefine $\mathbf{e}' = \mathbf{e} + \mathbf{e}_S$ and the result still follows. Then to generalize across all possible domains, we fix a domain and observational distribution and repeat the process until all domains have an associated set of observational and interventional distributions.

The corresponding tuple of distributions across interventions and domains belong to $S^{\Pi}_{\mathcal{K}}(G_1)$, but not $S^{\Pi}_{\mathcal{K}}(G_2)$ since m-separation should have implied invariance between the interventional and domain-change distributions whereas we constructed the distributions such that this is not true. ■

The difference between this statement and the one is Thm. 44 is simply what data is available. But the Lemma 42 does not care what sort of data is available, but is simply a result of the graphical structure.

Given Thm. 13, we can leverage the Ψ-FCI algorithm in the multi-domain observational data setting.

Corollary 51 (Modified Ψ-FCI algorithm given multi-domain observational data) *Let* $\mathbf{\Pi} = \{\Pi^1, ..., \Pi^N\}$ *be N domains with* $\mathbf{P}^{\Pi}$ *generated from* $\mathbf{\Psi}^{\Pi} = \langle \{\}^1, \{\}^2, ..., \{\}^N \rangle$ *consists of N observational distributions. Define the modified Ψ-FCI algorithm with the following modification: represent S as the set of intervention targets. The resulting Ψ-PAG learned is the same as the S-PAG.* ■

Therefore, the Ψ-FCI algorithm is applicable to the multi-domain setting when there is only observational data.

In the final theorem, we show that the S-FCI algorithm is sound, in that it learns a valid S-PAG (i.e. PAG with additional orientations).

Proof of Main Text Thm. 18 [S-FCI Soundness]

Theorem 52 (S-FCI Soundness) *Assuming tuple* $\mathbf{P}^{\Pi}$ *is generated by some unknown tuple* $\langle G, \mathbf{\Psi}^{\Pi}, \mathbf{S}^{\Pi} \rangle$ *with known intervention target* $\mathcal{K}$ *from domains* $\mathbf{\Pi}$ *and is s-faithful, where* $\mathbf{\Psi}$ *is a tuple of set of interventions with known/unknown targets,* $\mathbf{S}$ *and its corresponding edges indicate the S-nodes and their edges and G is the causal diagram, with* G_S *being the selection diagram. S-FCI algorithm is sound (i.e. every adjacency and orientation in* $\mathcal{P}_{S-FCI}$ *is common for* $MAG(Aug_{\mathbf{\Psi},\mathbf{S}}(G))$*).*

Proof [Proof Idea] In order to prove soundness that the result of S-FCI is a valid S-PAG, we will show that the algorithm's inferred separating sets between pairs of nodes are valid.

We determine:

1. Are all pairs of separable nodes in the graph correctly identified? I.e. all edges in the PAG are the result of an adjacency in the underlying DAG, or a primitive inducing path.

2. Do the augmented separating sets affect (negatively) the application of FCI rules?

3. Are the additional orientation rules sound?

The proof idea for the additional orientation rule R9' is as follows: adjacencies in a MAG are due to either adjacency in the true underlying selection diagram, or an inducing path between two nodes. Determining when this inducing path is the case across multiple domains and different interventions with known-targets allows one to then orient this inducing path. ■

Proof

(1) All pairs of F-nodes and S-nodes are separable with the empty set by construction of the augmented graph. Hence, after phase I of the S-FCI algorithm, they are non-adjacent with $SepSet(F_{ij}, F_{kl}) = \emptyset$ and the same for the S-nodes.

(2) To validate that the existing FCI rules are sound, we simply need to check that the rules that rely on separating sets are still valid given our augmented separating sets. The orientation of unshielded colliders and discriminating paths is sound based on the same reasoning as that in [3], since S-nodes are also in fact source nodes.

(3) Finally, we address the soundness of orientation rules. In [3] R9 of the $\mathcal{I}$-FCI algorithm is proved sound, which we follow a similar logic.

Define A_{ijk} as the set of nodes that are children of the F-nodes $F_i^{j,k}$.

We consider a pair of nodes $F_i^{j,k}, Y$, where $F_i^{j,k} \in \mathcal{F}^{\Pi}, Y \in V$ that are not adjacent, but $Y \in Neigh(A_{ij})$, indicating that there is no separating set between $F_i^{j,k}$ and Y in the augmented graph. Since they are not adjacent by construction, then there must be an inducing path between the two nodes relative to latent variables L. The same argument applies to separate Y from $S^{j,k}$. Therefore the MAG of the augmented diagram, $MAG(Aug_{\mathbf{\Psi},\mathbf{S}}(G))$ contains an edge from this node to Y. ■

F.1.7 Results improving efficiency of skeleton discovery phase

The skeleton discovery phase of the $\mathcal{I}$-FCI and Ψ-FCI algorithm require testing every possible combinations of nodes with every possible combination of conditioning sets. Constraint-based causal discovery algorithms typically searches for invariances by testing for example conditional independences among existing node pairs. These algorithms then typically may test all possible nodes as part of the separating set.

In the $\mathcal{I}$-FCI [3] and Ψ-FCI algorithms [7], the algorithms compare run through every single possible conditioning set when comparing distributions similar to the SGS algorithm [2]. However, this is obviously very inefficient.

This strategy while sound and works in theory, is very inefficient. Other strategies involve considering only neighbors, such as in the PC algorithm [6]. In addition, the FCI algorithm has been extended to be more computationally efficient, by only testing the possibly d-separating sets [24]. When dealing with the augmented graphs, we would like to ignore the augmented nodes that are irrelevant in the conditioning set. This is possible because we will see graphically that none of the augmented F-nodes constructed in Def. 5 are part of the possibly d-separating sets between nodes.

This enables one to speed up the S-FCI algorithm during the skeleton discovery stage using the same techniques.

Definition 53 (Possible-D-sep sets) *Let G be a mixed-edge graph with circular endpoints, and bidirected edges. $pds(X, Y)$ in G is defined as follows:*

$X \in pds(G, X, Y)$ if and only if there is a path π between X and Y in G such that every subpath (X_i, X_j, X_k) of π, X_j is a collider on the subpath in G, or (X_i, X_j, X_k) is a triangle in G. ■

The $pds(X, Y)$ set is useful because $pds(X, Y) \supseteq dsep(X, Y)$.

Lemma 54 (S-nodes are not required to be part of a d-separating sets) *Let $G = (V \cup S, E \cup E_S)$ be a joint selection diagram. Define $PDS(X, Y)$ as the possibly d-separating sets of X and Y as defined in [24]. For all $X, Y \subseteq V$ disjoint, no S-nodes are required to be part of d-sep(X, Y).*

Proof Assume an S-node, S_i is only pointing to one node, $Z \in V$. Then it is always an ancestor of Z. Consider disjoint $X, Y \subseteq V \backslash \{Z, S_i\}$. For X and Y to be d-separated, all d-connected paths must be blocked. Consider the path from X to Y through Z. If the path is a collider at $A \star\!\!\rightarrow Z \leftarrow\!\!\star B$, then the triplet (A, Z, B) is blocked as long Z, or descendants of Z are not conditioned on. If the path is a non-collider at Z, then it is blocked as long as Z is conditioned on. In both scenarios, S_i may be added to the conditioning set without changing the blocked/unblocked status of the triplet.

Consider now $S_j \in S$ that is another S-node. If that S-node is not along the path from X to Y, then it can be conditioned on arbitrarily since it is never a descendant of a collider and therefore would not open up a collider path.

Say S_i is pointing to now multiple nodes due to inducing paths. The argument is the same now for each node it is pointing to. If S_i is pointing to multiple nodes, the presence of an inducing path between X and Y indicates that there is no d-separating set between X and Y, so even if S_i is a graphical "confounder", adding it or not would not change the d-connectedness between X and Y. ■

This is useful to know as the skeleton search phase of the FCI algorithm and its variants typically rely on defining a superset of the d-separating set between variables, such as the $PDS(X, Y)$ set. Based on this lemma, we do not need to include any S-nodes ever in the conditioning set. This results in a faster skeleton discovery stage in S-FCI, which we incorporate into our implementation.

The skeleton phase proceeds as follows:

1. Run the FCI skeleton discovery phase among the non S-node variables using neighbors to select the conditioning sets
2. Orient unshielded colliders
3. Compute the $pds(X, Y)$ for all disjoint $X, Y \in V$
4. Orient all edges into circular endpoints
5. Re-run the FCI skeleton discovery phase using the $pds(X, Y)$ to select conditioning sets
6. Repeat the above now among S-node variables and non-S node variables

See Algorithm F.4 where we can leverage the strategy of possibly d-separating sets in the "CondSel" function. Moreover, we can limit the PDS set further by always removing all S-nodes and F-nodes from the PDS set.

F.2 Experimental Results - Simulations

F.2.1 Learning Selection Diagrams Across More Than Two Domains

The traditional selection diagram is presented with S-nodes that represent a change in mechanism between a pair of domains [44]. When we extend this to allow more than two domains, we can add additional S-nodes for each pair of domains. Consider Figure S1(a), where there are three S-nodes representing domains 1, 2 and 3. The presence of an S-node edge means there is not necessarily an invariance of X: i.e. $P^i(X) = P^j(X)$ is not necessarily true for $i \neq j$. However, in Figure S1(b), removing the edge between $S^{1,2}$ and X indicates that an invariance is present in the marginal distribution, $P^1(X) = P^2(X)$. However, if we also remove the edge $S^{1,3} \rightarrow X$, then this implies the invariance $P^1(X) = P^3(X)$. Then by transitivity, $P^2(X) = P^3(X)$ must also be true and the S-node edge $S^{2,3} \rightarrow X$ should also be removed in order for the graph to be valid.

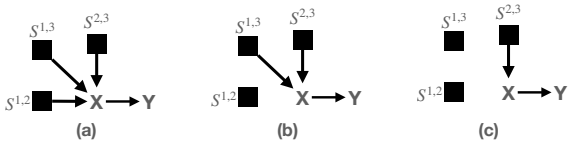


Figure S1: (a) shows a selection diagram with 3 domains with the distribution of X changing across any pair of domains. If we remove an edge $S^{1,2} \to X$, then this implies that for domain 1 and 2, the distribution of X is invariant (b). However, if we also remove the edge $S^{1,3} \to X$, then this additionally implies for domain 1 and 3 the distribution of X is invariant (c). Without explicitly testing the invariance, one can remove the edge $S^{2,3} \to X$ by transitivity. The reasoning is described in more detail in Section F.2.1.

This removal means that with higher number of domains, the learning of invariances across domains due to the lack of S-node edges can be accelerated. Say we have observational data across three domains and the selection diagram indicates that X is d-separated from all the S-nodes. Then as soon S-FCI determines the invariance such that the corresponding F-node, $F^{i,j}_{\{\}}$ is removed for two pairs of domains (1,2) and (1,3), then it can immediately remove the F-node $F^{2,3}_{\{\}}$ since the invariance must be true as well. To determine the invariant domains per node in the graph, one simply needs to construct an undirected graph among the domain IDs of the removed S-node edges and compute the connected components, which can be done in $\bigoplus(N)$ time, where N is the number of domains. This is a common graph algorithm that uses a disjoint set and is implemented in a variety of different packages, such as networkx [84]. This enables one to efficiently compute the invariant domains during the skeleton removal phase of Algorithm F.4.

This improvement due to limiting the necessary CI tests needed to be run can help improve runtime of the S-FCI algorithm.

F.2.2 Experiments

All experiments are reproducible using the algorithm implementations at https://github.com/py-why/dodiscover and https://github.com/py-why/pywhy-graphs [85, 86].

Chain-Graph Experiment In this section, we demonstrate empirically through computational experiments that S-FCI learns more, or more accurate graphs relative to the true selection diagram.

In the first simulation, a very simple setup is done to confirm the presentation of Ex. 25. In this example, $G = \{Y \to X, S^{1,2} \to X\}$ is the selection diagram with the augmented-selection diagram shown in Figure S2(c). The ground-truth causal diagram and augmented graph are shown in Figure S2(a-b), neither of which encode the change in domain.

Data is generated using a linear SCM, where nodes have exogenous noise generated from a Gaussian distribution (μ, σ) where μ is generated uniformly in [-5, 5] and σ is generated uniformly in [0.01, 1.5], and edge weights are generated uniformly in [-5, 5]. Each node is a linear combination of its parents, where edge functions are generated uniformly from the following choices with "x" as the input: linear (x), quadratic (x^2), sin (sin(x)), or negative (-x). We repeat the experiment 10 times with sample sizes ranging from 500 to 5000 linearly spaced. At each parametrization, we repeat the experiment 5 times. We simulate two different domains, with the S-node pointing to Y indicating a possible change in mechanism between domain 1 and 2. In total, we generate two different distributions, $\mathbf{P}^{\Pi} = \langle P^1_1, P^2_2 \rangle$, one per domain. Each distribution is interventional. We simulate a soft intervention on the node X by additively perturbing the values of X. We encode different soft interventions in domain 1 and 2 (i.e. the mechanisms have the same target, but different mechanisms; $\mathbf{\Psi}^{\Pi} = \langle \{X^a\}^1, \{X^b\}^2 \rangle$). We assume the targets are known, $\mathcal{K} = [1, 1]$.

Using the ground-truth diagram as an oracle for conditional independence and conditional invariance testing (of the form listed in Def. 3), we can get different ECs, which are shown in Figure S2(d-f). As we expect from Ex. 25, the $\mathcal{I}$-FCI arrives at the incorrect causal conclusion, $X \to Y$. Next, using partial correlation and the Kernel conditional discrepancy test [72], we test this setting with finite data. In Figure S3, we see that it is always the case that S-FCI learns the correct graph even with finite data.

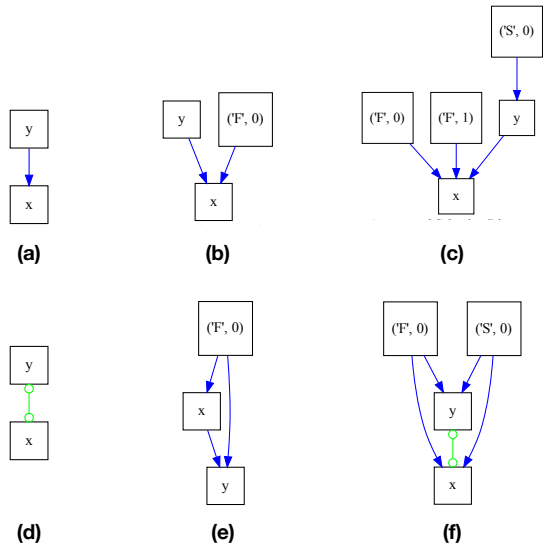


Figure S2: Comparing S-FCI vs FCI vs $\mathcal{I}$-FCI in a simulation with two known-target interventions with different mechanisms on X - The top row shows the true diagrams: (a) is the true causal diagram, (b) is the augmented diagram encoding the intervention on X, (c) is the augmented graph that shows the interventions on X in each domain and the S-node indicating a possible change in mechanism for Y. The bottom row shows the learned EC with an oracle for querying d-separation - (d) the PAG learned by the FCI algorithm, (e) the I-PAG learned by the $\mathcal{I}$-FCI algorithm and (f) the S-PAG learned by the S-FCI algorithm.

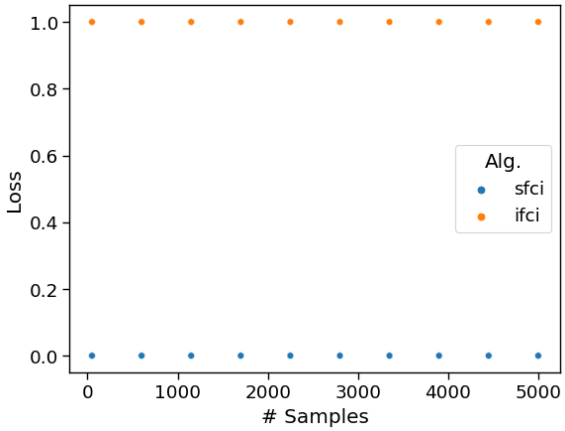


Figure S3: S-FCI learns the correct orientation consistently given known target interventions in multiple domains compared to $\mathcal{I}$-FCI in linear SCMs following the two-node setting.

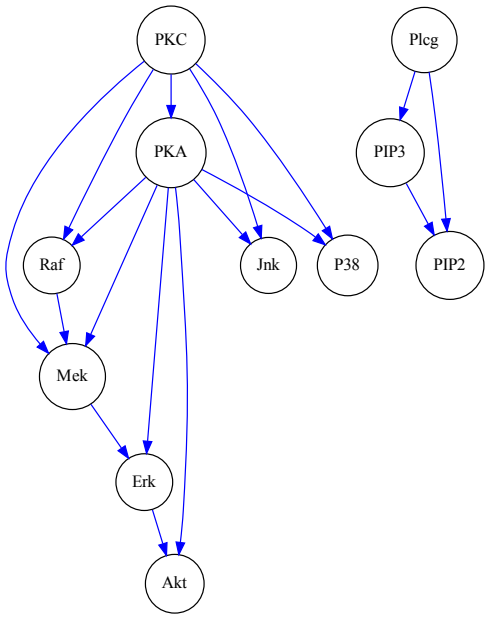


Figure S4: The presumed ground-truth graph for the protein experiments from [29]. Imported from bnlearn [88].

F.2.3 Analysis of Protein Sequencing

As motivated by Ex. 31 and 25, we next analyze a protein sequencing Sachs dataset [29], where different perturbations of proteins were made, and then responses from other proteins were observed. The ground truth graph is given by [29] and is shown in Supplemental Figure S4. We utilize this dataset because it is a commonly used dataset to evaluate causal discovery in many papers [3, 7, 53, 87]. We run S-FCI and get the results shown in Figure S5, where various structures such as the cluster among (PIP3, PIP2, Plcg) is detected and certain orientations in the larger graph are also correctly detected. As a result, these two experiments provide a realistic setting in which S-FCI could plausibly be used [15].

F.2.4 Simulated Data

In this next section, we present some experiments validating that adding additional data across multiple domains improves upon the structure by helping orient additional edges.

The ground-truth graph is shown in Figure S6(a). We forward-sample discrete data according to the graphical model and implement categorical data with cardinality of "3" per node. We then sample a random conditional probability distribution (CPD) for each node in topological order using pgmpy [89]. By specifying the full conditional distributions for each node as a function of its parents, this now specifies the full SCM. We then proceed with four different settings:

15. Real world data with ground truth selection diagrams and observational and interventional data collected over multiple domains is a big challenge that is necessary to evaluate multi-domain causal discovery algorithms. This paper partially addresses this need by leveraging real single-domain data and using that data to generate plausible datasets to simulate the multi-domain setting. Additional research is needed that generates this dataset in the real world from experiments and observations.

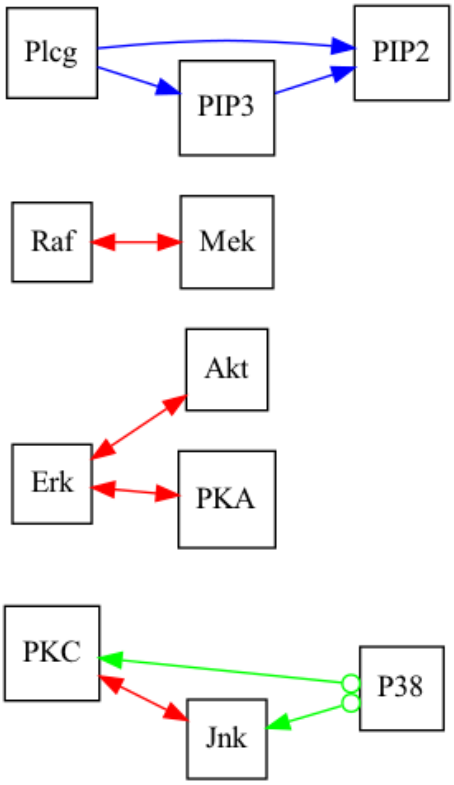


Figure S5: Shows the learned S-PAG of the Sachs dataset.

1. From this SCM, we sample 30,000 samples to denote the observational distribution, obs. We will denote this SCM as coming from domain 1. We run FCI on the data and obtain Figure S6(b).
2. Next, we generate 30,000 samples of interventional data by intervening on the 'D' node, generating a new CPD for node D. Then we run the $\mathcal{I}$-FCI, or Ψ-FCI algorithm depending on if we assume the intervention is a known-target or not. Regardless of the algorithm, the graph learned is in Figure S6(c).
3. Next we generate a domain-shift that changes the distribution of node X and C. I.e. in the corresponding selection diagram of (a), this would have the additional edges $X \leftarrow S^{1,2} \rightarrow C$. This generates a new SCM that represents domain 2. Combining the observational datasets from domain 1 and the domain 2, we can run FCI again and obtain the Figure S6(d).
4. We also simulate an intervention that occurs on node D again this time in domain 2. By pooling the interventional datasets and the observational datasets and naively ignoring the difference in domain, we can re-run the Ψ-FCI algorithm and obtain the graph in Figure S6(e).
5. Finally, taking all datasets together and applying the S-FCI algorithm, we obtain the result in Figure S6(f).

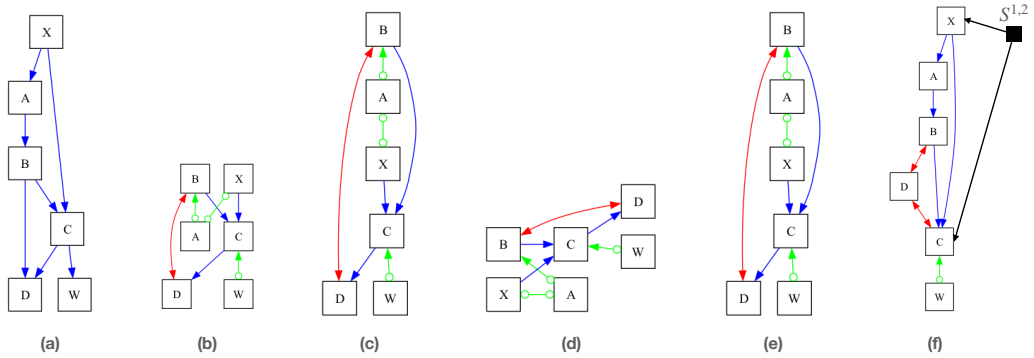


Figure S6: **Example simulation comparing FCI, Ψ-FCI and S-FCI using the same datasets** with ground-truth graph in (a). Running FCI on single-domain observational data results in (b). Running Ψ-FCI on single-domain observational and interventional data results in (c). Stacking the multi-domain observational data, ignoring the domains and running FCI results in (d). Stacking the multi-domain observational and interventional data, ignoring the domains and running Ψ-FCI results in (e). Running S-FCI on the same dataset as (e) without ignoring domains results in (f). (f) learns the most correct graph relative to ground-truth (a).

F.3 Background and Additional Preliminaries

In this section, we provide additional background notation and concepts relevant for the proofs and theoretical concepts introduced in this paper.

Additional Notation

A path p from X to Y in G is a sequence of distinct nodes $\langle X, ..., Y \rangle$ where each pair of consecutive nodes is adjacent in G. A directed path (also known as a causal path) from X to Y is a path where all edges are directed $X \rightarrow \cdots \rightarrow Y$. A possibly directed path from X to Y is a path where no arrowhead is pointing to X. A star on edge endpoints is used as a wildcard to denote circle, arrowhead, or tail.

We say if $X \to Y$, then X is a parent of Y. If there is a (possibly) directed path from X to Y, then X is a (possible) ancestor of Y and Y is a (possible) descendant of X. The convention is that every node is also a descendant and ancestor of itself. The sets of parents and (possible) descendants of X in G are denoted by $Pa(X, G)$ (or just $Pa(X)$ when it is unambiguous) and (Poss)De(X, G) respectively. Similarly, we would also write $PossCh(X)$ as the possible children of X, and $NonDesc(X)$ as the definite non-descendants of X. A definite non-descendant, Z, is one where there is no possibility of Z being a descendant of X. This can occur if there is a arrow-endpoint ending at X, or a tail-endpoint ending at Z.

A triple $\langle X, Y, Z\rangle$ is an unshielded triple if X and Y are adjacent, Y and Z are adjacent, and X and Z are not adjacent. If both edges are into Y, then the triple is referred to as unshielded collider. A path between X and Y, $p = \langle X, ..., W, Z, Y\rangle$, is discriminating for Z if every node between X and Z is a collider on p and is a parent of Y. Two MAGs are Markov equivalent if and only if (1) they have the same adjacencies; (2) the same unshielded colliders; and (3) if a path p is a discriminating path for Z in both graphs, then Z is a collider on p in one graph if and only if it is a collider on p in the other. A PAG represents an MEC of a MAG and is learnable from data. The output of the celebrated FCI algorithm is a PAG, which is proven sound and complete for the corresponding MEC [20].

m-separation and m-connectedness

In this section, we briefly review the graphical criteria m-separation and m-connectedness, which is a generalization of d-separation and d-connectedness [61]. First, a necessary definition to fully understand and characterize m-separation is the notion of definite colliders and definite non-colliders in PAGs as these are a way to determine definite status along a path that does not need oriented end points.

Definition 55 (Definite collider and non-colliders) *Let $\langle A, B, C\rangle$ be a consecutive triple along a path p in a PAG, G. B is a definite collider on p if both edges are into B (i.e. arrowhead endpoint at B). B is a definite non-collider on p if either one of these is true:*

i) One of the edges is out of B ($A \leftarrow B \ast\!\!\rightarrow C$, or $A \ast\!\!\rightarrow B \rightarrow C$. ii) Both edges have a circle-endpoint at B, and there is no edge between A and C (i.e. $\langle A, B, C\rangle$ is unshielded). This looks like $A \ast\!\!-\!\!o B o\!\!-\!\!\ast C$.

Otherwise B has a non-definite status along p.

A definite status path p between nodes X and Y is m-connecting given a set of nodes $\mathbf{Z}$ (with $X, Y \notin \mathbf{Z}$) if every definite non-collider on p is not in $\mathbf{Z}$ and every collider in p has a descendant in $\mathbf{Z}$. A possibly m-connecting path between X and Y given $\mathbf{Z}$ is a path where every definite non-collider on the path is not in $\mathbf{Z}$ and every collider has a possible-descendant in $\mathbf{Z}$.

If $\mathbf{Z}$ blocks all definite status paths between X and Y, we say that X and Y are m-separated given $\mathbf{Z}$. Otherwise X and Y are m-connected. If $\mathbf{Z}$ blocks all possibly m-connecting paths between X and Y, we say that X and Y are $\hat{m}$-separated given $\mathbf{Z}$.

Chordal Graphs

An undirected graph H is chordal if and only if every undirected cycle of length four or more has an edge between two non-consecutive vertices on the cycle.

F.4 Broader Impact and Forward Looking Statements

The development of new causal discovery algorithms has the potential to improve our understanding of complex systems, and to help identify the causal factors underlying important societal issues. By improving our ability to learn causal relationships from observational and interventional data across multiple domains, your work could ultimately lead to more effective interventions to address these issues that are transportable across operating domains. Beyond the causal inference community, we

expect that our results will enable fundamental contributions in various fields, including biology [29], epidemiology [81], economics [90] and neuroscience [34].

One significant research direction is to study how to relax the assumption that the joint selection diagram does not contain structural differences among the different domains. Additionally, it will be important in future research to develop new benchmarks that reflect this emerging multi-domain causal discovery paradigm to evaluate algorithms. Another important research question is how to perform transportability inference within this newly introduced equivalence class. Transportability of causal effects, also known as "external validity" [91, 92], "meta-analysis" [93], "quasi-experiments" [94], "heterogeneity" [95], is a critical task that has been studied under the assumption that a well-specified selection diagram is available. It will be important to develop algorithms for transportability inference given the selection diagrams' EC and develop algorithms for computing causal effects from an EC of selection diagrams. This would enable scientists to perform completely data-driven causal analysis across multiple domains.

F.5 MD-FCI Algorithm Additional Details

Here, we expand on the MD-FCI algorithm and its details. The inner-workings of the MD-FCI algorithm are introduced in Algorithm 2. Here, we provide details for the rest of the algorithm.

F.5.1 MD-FCI Algorithm Details

Creating augmented graph Alg. F.3 describes how the augmented graph is created by adding nodes that map to pairs of distributions, and optionally symmetric difference targets.

Algorithm F.3 Generalized Augmenting Nodes - $\mathbf{S}$ is the set of S-nodes over N domains $\mathbf{\Pi}$, $\mathcal{F}^{\Pi}$ is the set of F-nodes over each domain, $\mathcal{K}$ is the vector of known intervention targets, $\mathbf{H}$ is the set of intervention targets mapping each pair of known-target interventions, σ is the mapping of each pair of distributions within each domain and $\mathbf{V}$ is the set of nodes in the graph.

function CreateAugmentedNodes($\mathbf{\Psi}^{\Pi}, \mathcal{K}, \mathbf{V}$)
  $(\mathbf{S} = \emptyset, \mathcal{F}^{\Pi} = \emptyset, \mathbf{H} = \emptyset, \sigma : \mathbf{N} \times \mathbf{N} \to 2^V \times 2^V$
  $k \leftarrow 0$
  Add S-nodes
  for all pairs $\Pi^i \neq \Pi^j \in \mathbf{\Pi}$ **do**
    Add S^{ij} to $\mathbf{S}$
  Add F-nodes
  for all pairs $\mathbf{I}_l^i, \mathbf{J}_m^j \in \mathbf{\Psi}^{\Pi}$ **do**
    if $\mathbf{I}_l^i = \{\}^i, \mathbf{J}_m^j = \{\}^j, i \neq j$ **then**
      do nothing
    else
      $k \leftarrow k + 1$
      Add F_k^{ij} to $\mathcal{F}^{\Pi}$
      if $\mathbf{I}_l^i$ and $\mathbf{J}_m^j$ are known-targets and $i = j$ **then**
        $H_k^{ij} = \mathbf{I}^i \Delta \mathbf{J}^j$
        Add H_k^{ij} to $\mathbf{H}$
    $\sigma(k) = (l, m)$: **(Maps the kth F-node to distributions l and m)**
  return $\mathbf{S}, \mathcal{F}^{\Pi}, \mathbf{H}, \sigma$

Generalized Multi-Domain Skeleton Discovery Alg. F.4 describes a generalized algorithm for performing constraint-based skeleton discovery, which allows our algorithm to choose a method for choosing candidate conditioning sets, *CondSel*. For example, one may use all possible combinations of nodes (e.g. the SGS algorithm does this [21]), or the neighbors of the nodes (e.g. the PC algorithm

does this [6]), or the possibly d-separating sets (e.g. in RFCI algorithm [24]). Alg. F.5 describes how to infer the skeleton structure using constraints found in the data. For instance, the first else-if statement states that all F-nodes are by construction separated. The second else-if statement states that an F-node will be separated from another node given a specific kind of invariance described in Condition 2 of Def. 3.

Algorithm F.4 Generalized Skeleton Discovery - G is the augmented causal diagram from Def. 5, $CondSel$ is the conditioning selection function for determining how to select candidate separating sets Z, P_{max} is a hyperparameter controlling the maximum size of the conditioning set

function GeneralizedSkeletonDiscovery($G, CondSel, P_{max}$)
  $G = (V \cup \mathcal{F}, E \cup E_{\mathcal{F}})$
  while $p < P_{max}$ **do**
    for $X \in V$ **do**
      for $Z \in CondSel(X, p)$ **do**
        if $(X \in \mathcal{F} \cap Y \in \mathcal{F})$ **then**
          $SepSet(X, Y) \leftarrow \emptyset, Sep(X, Y) = True$
        else
          $(SepSet(X, Y), Sep(X, Y) \leftarrow$ Generalized Do-constraints (see Alg. F.5)
        if Sep(X,Y) = True **then**
          Remove (X, Y) edge in graph G

Algorithm F.5 Generalized Do-Constraints - $\mathbf{\Psi}^{\Pi}$ is the intervention targets per N domains, $\mathbf{\Pi}$; $\mathcal{K}$ are the known targets; $\mathbf{V}$ are the relevant causal variables.

function GeneralizedDoConstraints($X, Y, \mathbf{S}, \mathcal{F}^{\Pi}, \sigma, \mathbf{\Psi}^{\Pi}, \mathcal{K}, \mathbf{V}$)
  ($\mathcal{F}^{\Pi} = \emptyset, SepSet = \emptyset, \sigma : \mathbf{N} \to 2^V \times 2^V$
  $\mathbf{V} \leftarrow \mathbf{V} \cup \mathcal{F}^{\Pi}$
  if $X, Y \notin \mathcal{F}^{\Pi}$, and $X, Y \notin \mathbf{S}$ **then**
    for $I^i \in \mathbf{\Psi}^{\Pi}$ **do**
      for $W \subseteq V \backslash \mathcal{F}$ **do**
        if $P^i_I(y|w,x) = P^i_I(y|w)$ **then**
          $SepSet = W \cup \mathcal{F}^{\Pi} \cup \mathbf{S}$
          SepFlag=True
  else if $X \in \mathbf{S}, Y \in V$ **then**
    $(l, m) \leftarrow \sigma(k)$
    for $W \subseteq V \backslash \mathcal{F}$ **do**
      if $P^i_l(y|w,x) = P^j_m(y|w)$ **then**
        $SepSet = W \cup \mathcal{F}^{\Pi} \cup \mathbf{S} \backslash S^{i,j}$
        SepFlag=True
  else if $X, Y \in \mathcal{F}^{\Pi}$ **then** (X and Y are both F-nodes)
    SepSet = $\mathcal{F}^{\Pi} \cup \mathbf{S} \backslash \{X, Y\}$
    SepFlag=True
  else if $X, Y \in \mathbf{S}$ **then** (X and Y are both S-nodes)
    SepSet = $\mathcal{F}^{\Pi} \cup \mathbf{S} \backslash \{X, Y\}$
    SepFlag=True
  else if ($X \in \mathcal{F}^{\Pi}$ and ($Y \in V$), so let $F^{i,j}_k$ denote X (X is a F-node representing a distribution between domains i and j, and Y is a normal node in $\mathbf{V}$) **then**
    $(l, m) \leftarrow \sigma(k)$
    for $W \subseteq V \backslash \mathcal{F}$ **do**
      if $P^i_l(y|w,x) = P^j_m(y|w)$ **then**
        $SepSet = W \cup \mathcal{F}^{\Pi} \backslash \{F^{i,j}_k\} \cup \mathbf{S}$
        SepFlag=True

Generalized Multi-Domain Orientation Rules - We restate the orientation rules presented in 4.3 for completeness of the appendix.

Algorithm F.6 Generalized Orientation Rules - G is the causal diagram, $SepSet$ are the separating sets that were learned, $\mathcal{F}^{\Pi}$ is the set of F-nodes, H is the set of known-intervention targets and S are the S-nodes.

For every unshielded triple (X, Y, Z), if $Z \notin SepSet(X, Y)$ orient it as $X \ast\!\!\rightarrow Y \leftarrow\!\!\ast Z$
Phase IIb: Apply logical orientation rules
R1-7: Apply 7 FCI rules from [20] and following two rules until none apply.
Rule 8': For any $F^{i,j}_k \in \mathcal{F}^{\Pi}$, orient adjacent edges out of $F^{i,j}_k$.
Rule 9': For any $F^{i,j}_k \in \mathcal{F}^{\Pi}$, that is adjacent to a node $Y \notin H^{i,j}_k$, if $i = j$ and $X \in H^{i,j}_k$ and $|H^{i,j}_k| = 1$, orient $X \to Y$.

F.6 Additional Comparisons

Mechanism Shift Score [10] Consider the setup in Example 21.

A key assumption commonly leveraged in the causality literature is known as the Sparse Mechanism Shift (SMS) hypothesis [10, 96], which states that changes in mechanisms between observed domains

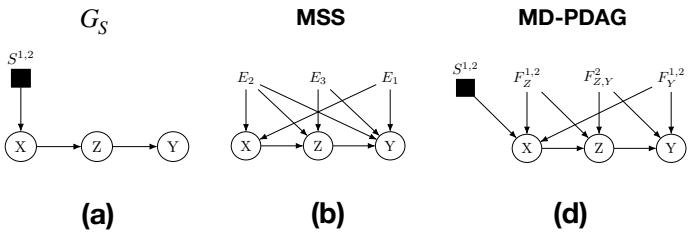


Figure S7: Comparing multi-domain causal discovery algorithms under the setting of causal sufficiency - The ground truth selection diagram (a) is given over domains Π^1, Π^2 with intervention targets $\mathbf{\Psi} = \langle \{\}^1, \{Y\}^2, \{Z\}^2 \rangle$ with all interventions being known-target, $\mathcal{K} = [1, 1, 1]$. The Mechanism Shift Score (MSS) estimand [10] (b), and the MD-PDAG (d) learned from the MD-PC algorithm. Note the MSS estimand shows both arrow directions indicating that both DAGs could theoretically be recovered. This is due to the fact that the MSS returns a DAG rather than an EC.

are sparse. The SMS hypothesis allows one to leverage data across different domains and environments, but it does not state how to interpret interventional data across domains. Nor does it emphasize the necessity to distinguish the two related, yet different concepts.

By applying the Mechanism Shift Score in Example 21, one would learn the graph shown in Figure 8(b), which does not provide any information about the invariance between hospital 1 and 2.