TECHNICAL REPORT
R-130
First version: May, 2025
Last version: Nov, 2025

A Hierarchy of Graphical Models for Counterfactual Inferences

Hongshuo Yang **Elias Bareinboim**
Causal Artificial Intelligence Lab
Columbia University
hy2712@columbia.edu eb@cs.columbia.edu

Abstract

Graphical models have been widely used as parsimonious encoders of assumptions of the underlying causal system and provide a basis from which causal inferences can be performed. Models that encode stronger constraints tend to require higher expressive power, which are also harder, and sometimes impossible to empirically falsify. In this paper, we introduce two new collections of distributions that include counterfactual quantities which are experimentally accessible under what is known as counterfactual randomizations. Correspondingly, we define two new classes of graphical models for encoding empirically testable constraints in these distributions. We further present a sound and complete calculus, based on counterfactual calculus, which licenses inferences in these two new models with rules that are within the empirically falsifiable boundary. Finally, we formulate a hierarchy over several graphical models based on the constraints they encode and study the fundamental trade-off between the expressive power and empirical falsifiability of different models across the hierarchy.

1 Introduction

Causal information is fundamental across a wide range of scientific disciplines and human decision-making, and it is increasingly recognized as a necessary ingredient for advancing AI and machine learning in enhancing robustness, interpretability, and generalizability [21, 1]. The *Pearl Causal Hierarchy* (PCH) organizes such information into three layers: the *observational*, the *interventional*, and the *counterfactual*, corresponding roughly to the ordinary human capabilities of *seeing, doing, and imagining* [21, 3]. Each layer is formalized through a distinct symbolic language and encodes causal quantities with increasingly expressive semantics. For example, consider a system with two observed variables, X (*treatment*, e.g., diet) and Y (*outcome*, e.g., BMI). Layer 1 ($\mathcal{L}_1$) includes *observational* distributions, like $P(y|x)$, which represents the probability of observing BMI y among individuals who naturally follow diet x. Layer 2 ($\mathcal{L}_2$) contains *interventional* distributions, like $P(y|do(x))$, which represents the probability of having BMI y among individuals who were externally assigned to diet x. Layer 3 ($\mathcal{L}_3$) comprises *counterfactual* distributions, like $P(y_x|x')$, which represents the probability of having BMI y if the diet had been set to x among those who would naturally follow diet x'.

When the true causal mechanism underpinning a phenomenon of interest — formally represented by a *Structural Causal Model (SCM)* — is known, all layers of the PCH are immediately computable. Unfortunately, it is rare for SCMs to be known at this level of precision in most real-world scenarios. This limitation gives rise to the field of *causal inference*, which seeks to understand the conditions under which valid inferences can be made given access to limited features and data from the causal system. The inferential process can be illustrated through the *causal inference engine* [1, Sec. 1.3.4], as illustrated in Fig. 1. The engine takes three inputs: {(1) *Query*, (2) *Data*, (3) *Model*}, each reflecting a different aspect of the underlying SCM. The *Query* specifies the causal quantity of

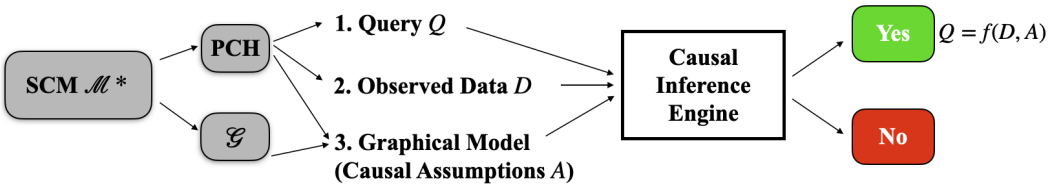


Figure 1: Unobserved SCM and the causal inference engine. The engine takes as input a query, a model, and datasets, and returns whether the query is computable from the assumptions and data.

interest, the *Data* consists of data gathered through interactions with the environment like random samplings or randomized experiments, and the *Model* encodes assumptions about the SCM. A common language for articulating such assumptions is provided by *graphical models*, particularly *causal diagrams* [18, 20, 3, 1] and their variants, which encode constraints describing how different quantities within the PCH relate to one another. For example, Pearl's celebrated treatise *Probabilistic Reasoning in Intelligent Systems* developped a comprehensive account of Bayesian Networks (BN) as encoders of conditional independencies — that is, $\mathcal{L}_1$ equality constraints within an observational distribution, such as $P(y \mid x) = P(y)$ [18]. In contrast, Causal Bayesian Networks (CBN) encode equality constraints across distributions in both $\mathcal{L}_1$ and $\mathcal{L}_2$, like $P(y|do(x)) = P(y|x)$ [20, 2, 3]. Counterfactual Bayesian Networks (CTFBN) further extend this framework to encode constraints across $\mathcal{L}_3$ distributions, like $P(y_x, x) = P(y, x)$ [1, Sec. 13.2].

For a graphical model to be sufficient for supporting inference on a query, there must be a match in *expressiveness* between the model's constraints and the query, as illustrated in Fig. 2. This requirement aligns with Nancy Cartwright's famous motto "no causes in, no causes out" [5], as mathematically formalized by the *Causal Hierarchy Theorem* (CHT): to perform inferences on a quantity in layer i, one needs knowledge from layer i or above [3, Corollary 1]. For instance, given an $\mathcal{L}_2$ query, a BN encoding only $\mathcal{L}_1$ constraints is insufficient, while a CBN encoding both $\mathcal{L}_1$ and $\mathcal{L}_2$ constraints is both sufficient and necessary for inference. A CTFBN encoding $\mathcal{L}_3$ constraints, while sufficient for the target query in $\mathcal{L}_2$, imposes assumptions that are stronger than necessary [6, 1].

While models that encode constraints higher in the PCH support inferences about more expressive queries, it is also generally preferable to avoid unnecessary assumptions for a given query. This notion of parsimony is grounded by the concept of *empirical falsification* in the sciences. As Popper emphasized, a system is scientific only if it is refutable by empirical tests [23]. At the same time, the degree of falsifiability varies across domains. In some fields, opportunities for direct refutation are abundant; in others, such as cognitive science and AI, they may be scarcer, motivating the introduction of additional assumptions that render counterfactuals articulable. This does not weaken the principle of falsifiability, but reflects a continuum of empirical accessibility across scientific disciplines. [1]

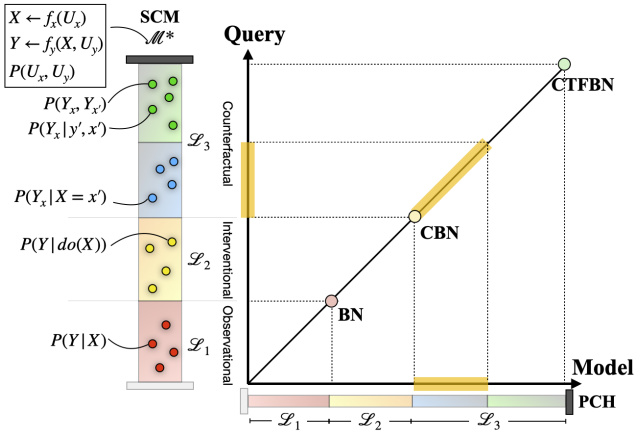


Figure 2: Expressive power of queries and graphical models along the PCH. The model's constraints should be at least as expressive as the query for the causal inference engine to work. Layer 3 is partitioned into two sub-regions: the green region represents $\mathcal{L}_3$ distributions that cannot be accessed via any experiments, while the blue region represents those that can, at least in principle, be sampled via experiments.

[1] Even well-established disciplines illustrate the nuanced nature of falsifiability. In theoretical physics, for instance, String theory has been critiqued for its impressive mathematical structure yet still lacking empirically testable predictions [27]. And recent observations with the James Webb Space Telescope have identified ultra-massive galaxies at very high redshifts, initially thought to conflict with standard cosmological models, which famously generated headlines about potentially "breaking cosmology" before alternative explanations were found [31]. These examples suggest that falsifiability in science is not an all-or-nothing criterion, but instead

For graphical models, this continuum becomes a concrete question of empirical accessibility: can the assumptions encoded in a model be subjected to empirical testing with available data? Concretely, the falsifiability of an assumption in a graphical model depends on the feasibility of drawing samples from its underlying distributions (also known as *realizability*,[24]). Among the three layers of the PCH, it is generally understood that data from $\mathcal{L}_1$ and $\mathcal{L}_2$ distributions are, at least in principle, attainable via *random sampling* and *randomized controlled trials* [11]. In contrast, $\mathcal{L}_3$ encodes counterfactual knowledge traditionally considered beyond the reach of physical experimentation. For example, the probability of necessity and sufficiency (PNS), $P(y_x, y'_{x'})$, is an $\mathcal{L}_3$ quantity that cannot be obtained via any randomized experiments. Yet, recent work by Bareinboim, Forney and Pearl have revealed that an $\mathcal{L}_3$ quantity known as the effect of the treatment on the treated (ETT), $P(y_x|x')$, can be sampled through a new experimental procedure called *counterfactual randomization* [4]. Subsequent work further refined and characterized the set of $\mathcal{L}_3$ distributions that are realizable in principle [24]. These advances reveal that $\mathcal{L}_3$ is not monolithic but instead contains distributions with varying degrees of empirical accessibility. This heterogeneity naturally raises the question of what assumptions are sufficient and necessary to support counterfactual inference — a question we now address directly.

The empirical heterogeneity of $\mathcal{L}_3$ distributions naturally raises a central question: what assumptions are sufficient and necessary to support counterfactual inference? In this paper, we addresses that question by analyzing the region between $\mathcal{L}_2$ and $\mathcal{L}_3$ in the PCH, illustrated in the orange zone of Fig. 2. We introduce formal languages, models, and inferential machinery for two families of realizable distributions that extend beyond the Fisherian interventional world yet remain empirically accessible. In doing so, we give a precise mathematical form to Cartwright's principle: only when the assumptions built into a model ("causes in") are adequate can the corresponding counterfactual queries ("causes out") be answered.

Our main contributions are as follows:

(1) **Graphical Models & Inferential Machinery:** We introduce symbolic languages and valuation semantics for two new collections of distributions, each entail quantities that become experimentally accessible by a distinct implementation of *counterfactual randomization*. We then define two new classes of graphical models, CBN2.25 and CBN2.5, that encode constraints within these distributions which are amenable to empirical testing. We prove that counterfactual calculus with graphical checks form a sound and complete inferential machinery for CBN2.25 and CBN2.5.

(2) **Hierarchy of Graphical Models:** We formally define a hierarchical structure for graphical models based on constraints they encode and analyze this structure from two angles: (a) expressive power and (b) empirical falsifiability. We show that models higher in the hierarchy encode stronger assumptions that permit more expressive queries, but are increasingly harder to falsify.

2 Causality and Graphical Model Review

In this section, we review key concepts and definitions needed throughout the paper.

First, we start with the notations. We denote *variables* by capital letters, X, and *values* by small letters, x. Bold letters, $\mathbf{X}$ represent a set of variables and $\mathbf{x}$ a set of values. The *domain* of X is denoted by $Val(X)$. Two values $\mathbf{x}$ and $\mathbf{z}$ are *consistent* if they share common values for $\mathbf{X} \cap \mathbf{Z}$. We denote by $\mathbf{x}\backslash\mathbf{Z}$ the value of $\mathbf{X}\backslash\mathbf{Z}$ consistent with $\mathbf{x}$ and by $\mathbf{x} \cap \mathbf{Z}$ the subset of $\mathbf{x}$ corresponding to variables in $\mathbf{Z}$. We assume the domain of every variable is finite. $\mathbf{W}_*$ denotes an arbitrary counterfactual event, and $\mathbf{V}(\mathbf{W}_*) = \{W \in \mathbf{V} | W_t \in \mathbf{W}_*\}$. $\mathcal{G}[\mathbf{W}]$ denotes a vertex-induced subgraph over $\mathbf{W}$. We use kinship notation for variable relationships: parents (Pa), children (Ch), descendants (De), ancestors (An).

We adopt *Structural Causal Models* (SCMs) as the baseline generative framework, following the presentation in [3]. The discussion there is more detailed and may be consulted if additional context on these foundational notions is needed.

operates along a nuanced continuum, a theme that, in the context of graphical models, manifests in questions of which distributions are empirically realizable.

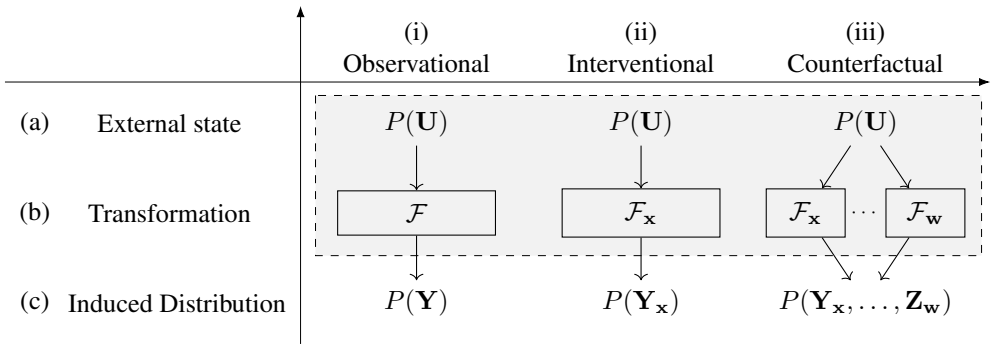


Figure 3: Given an SCM's initial state (i.e., population) (a), we show the different functional transformations (b) and the corresponding induced distribution (c) of each layer of the hierarchy. (i) represents the transformation (i.e., $\mathcal{F}$) from the natural state of the system ($P(\mathbf{U})$) to an observational world, (ii) to an interventional world (i.e., with modified mechanisms $\mathcal{F}_{\mathbf{x}}$), and (iii) to multiple counterfactual worlds (i.e., with multiple modified mechanisms).

Definition 1 (Structural Casual Model (SCM) [3])**.** *A structural causal model $\mathcal{M}$ is a 4-tuple $\langle \mathbf{U}, \mathbf{V}, \mathcal{F}, P(\mathbf{U}) \rangle$, where*

- $\mathbf{U}$ *is a set of background variables, also called exogenous variables, that are determined by factors outside the model;*
- $\mathbf{V}$ *is a set $\{V_1, V_2, \ldots, V_n\}$ of variables, called endogenous, that are determined by other variables in the model — that is, variables in $\mathbf{U} \cup \mathbf{V}$;*
- *$\mathcal{F}$ is a set of functions $\{f_1, f_2, \ldots, f_n\}$ such that each f_i is a mapping from (the respective domains of) $U_i \cup \mathbf{Pa}_i$ to V_i, where $U_i \subseteq \mathbf{U}$, $\mathbf{Pa}_i \subseteq \mathbf{V} \setminus V_i$, and the entire set $\mathcal{F}$ forms a mapping from $\mathbf{U}$ to $\mathbf{V}$. That is, for $i = 1, \ldots, n$, each $f_i \in \mathcal{F}$ is such that*

$$v_i \leftarrow f_i(\mathbf{pa}_i, \mathbf{u}_i), \tag{1}$$

  i.e., it assigns a value to V_i that depends on (the values of) a select set of variables in $\mathbf{U} \cup \mathbf{V}$; and
- $P(\mathbf{U})$ *is a probability function defined over the domain of* $\mathbf{U}$.

Intervention in an SCM can be viewed as a modification of the model by changing the mechanism of the intervened variables, while keeping all other components of the SCM intact.

Definition 2 (Submodel — "Interventional SCM" [20])**.** *Let $\mathcal{M}$ be a structural causal model, $\mathbf{X}$ a set of variables in $\mathbf{V}$, and $\mathbf{x}$ a particular realization of $\mathbf{X}$. A submodel $\mathcal{M}_{\mathbf{x}}$ of $\mathcal{M}$ is the causal model*

$$\mathcal{M}_{\mathbf{x}} = \langle \mathbf{U}, \mathbf{V}, \mathcal{F}_{\mathbf{x}}, P(\mathbf{U}) \rangle, \tag{2}$$

where

$$\mathcal{F}_{\mathbf{x}} = \{f_i : V_i \notin \mathbf{X}\} \cup \{\mathbf{X} \leftarrow \mathbf{x}\}. \tag{3}$$

The impact of the intervention on an outcome variable Y is commonly called the potential outcome:

Definition 3 (Potential Outcomes [20])**.** *Let $\mathbf{X}$ and $\mathbf{Y}$ be two sets of variables in $\mathbf{V}$, and $\mathbf{u}$ be a unit. The potential outcome $\mathbf{Y}_{\mathbf{x}}(\mathbf{u})$ is defined as the solution for $\mathbf{Y}$ of the set of equations $\mathcal{F}_{\mathbf{x}}$ with respect to SCM $\mathcal{M}$ (or, $\mathbf{Y}_{\mathcal{M}_{\mathbf{x}}}(\mathbf{u})$). That is, $\mathbf{Y}_{\mathbf{x}}(\mathbf{u}) \triangleq \mathbf{Y}_{\mathcal{M}_{\mathbf{x}}}(\mathbf{u})$.*

An SCM induces observational, interventional, and counterfactual distributions over the endogenous variables, which form three layers known as the Pearl Causal Hierarchy (PCH).

Definition 4 (Pearl Causal Hierarchy (PCH) ([3]))**.** *An SCM $\mathcal{M} = \langle \mathbf{U}, \mathbf{V}, \mathcal{F}, P(\mathbf{u}) \rangle$ induces three layers of probability distributions that form the Pearl Causal Hierarchy. For any $\mathbf{Y}, \mathbf{Z}, ..., \mathbf{X}, \mathbf{W} \subseteq \mathbf{V}$, the three layers of distributions are given by:*

- *$\mathcal{L}_1$ (Observational):*

$$\mathbf{P}^{\mathcal{M}}(\mathbf{y}) = \sum_{\mathbf{u}} \mathbf{1}[\mathbf{Y}(\mathbf{u}) = \mathbf{y}] P(\mathbf{u}) \tag{4}$$

- $\mathcal{L}_2$ *(Interventional):*

$$\mathbf{P}^{\mathcal{M}}(\mathbf{y_x}) = \sum_{\mathbf{u}} \mathbf{1}[\mathbf{Y_x}(\mathbf{u}) = \mathbf{y}]P(\mathbf{u}) \tag{5}$$

- $\mathcal{L}_3$ *(Counterfactual):*

$$\mathbf{P}^{\mathcal{M}}(\mathbf{y_x}, ..., \mathbf{z_w}) = \sum_{\mathbf{u}} \mathbf{1}[\mathbf{Y_x}(\mathbf{u}) = \mathbf{y}, ..., \mathbf{Z_w}(\mathbf{u}) = \mathbf{z}]P(\mathbf{u}) \tag{6}$$

The collection of all $\mathcal{L}_1$ (Observational) is denoted as $\mathbf{P}^{\mathcal{L}_1}$, the collection of all $\mathcal{L}_2$ (Interventional) is denoted as $\mathbf{P}^{\mathcal{L}_2}$, and the collection of all $\mathcal{L}_3$ (Counterfactual) is denoted as $\mathbf{P}^{\mathcal{L}_3}$.

PCH specifies both the symbolic representation and the valuation of each probabilistic quantity given an underlying SCM. If the SCM is fully specified, every quantity in any layer of the PCH can be computed directly via Def. 4 (Fig.3). In the causal inference engine (Fig. 1), this correspondence is depicted by the arrow from $\mathcal{M}^*$ to the PCH.

In practice, however, only partial knowledge of the SCM is available, so only a subset of the PCH can be observed. For example, the observational distribution may be available (Fig. 1, item (2)), while the interventional distribution remains unobserved and must be queried (item (1)). Each causal inference task therefore rests on assumptions about the structural "marks" left by the SCM on its distributions. These assumptions take the form of invariance constraints, defined as follows:

Definition 5 (Invariance Constraint)**.** *Given an SCM $\mathcal{M}^*$, an invariance constraint is an equality or inequality between polynomials over $\mathcal{L}_i$ terms of the PCH.*

A common example is conditional independence in the observational distribution. For instance, $P(y \mid x) = P(y)$ encodes that X is probabilistically independent of Y.

Invariance constraints can be seen as coarsening the PCH: they abstract away from specific numerical values and instead capture relationships among distributions. As more invariance constraints are included, the granularity of knowledge about the underlying SCM increases. To avoid enumerating constraints individually, we exploit graphical models, which encode them systematically and parsimoniously by linking invariance constraints to graph topology (e.g., relations among parents, neighbors, and ancestors). The natural first step in this process is to construct a causal diagram from a given SCM, since the diagram directly captures the topological relations that determine which invariances hold.

Definition 6 (Causal Diagram [3])**.** *Consider an SCM $\mathcal{M} = \langle \mathbf{U}, \mathbf{V}, \mathcal{F}, P(\mathbf{u}) \rangle$. Then $\mathcal{G}$ is a causal diagram of $\mathcal{M}$ if constructed as follows:*

(1) add a vertex for every endogenous variable in the set $\mathbf{V}$

(2) add an edge $V_i \longrightarrow V_j$ for every $V_i, V_j \in \mathbf{V}$ if V_i appears as an argument of f_j

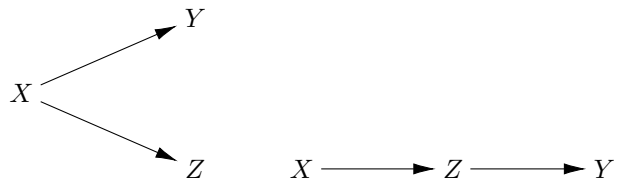
(3) Add a bidirected edge $V_i \dashleftarrow\!\dashrightarrow V_j$ for every $V_i, V_j \in \mathbf{V}$ if

(a) the corresponding functions f_i, f_j share some common $U \in \mathbf{U}$ as an argument, or
(b) the corresponding $U_i, U_j \in \mathbf{U}$ are correlated.

The causal diagram $\mathcal{G}$ can be viewed as a non-parametric coarsening of the SCM $\mathcal{M}$: it preserves the structural signatures (i.e., the arguments of the functions and dependency relationships among exogenous variables) while abstracting away from their specific parametrization. In Fig. 1, this is depicted by the arrow from $\mathcal{M}^*$ to $\mathcal{G}$.

Pairing a causal diagram with the set of invariance constraints it encodes over a collection of distributions defines a graphical model, also referred to as a compatibility relation. For any SCM, its induced causal diagram $\mathcal{G}$ together with the corresponding PCH distributions naturally form such a relation. In this way, the invariance constraints encoded in $\mathcal{G}$ serve as surrogates for the empirical content of the underlying SCM, summarizing knowledge about the different layers of distributions it induces.

Definition 7 (Graphical Model)**.** *A graphical model is a pair $\langle \mathcal{G}, \mathbf{P} \rangle$, where $\mathcal{G}$ is a graph and $\mathbf{P}$ is a collection of distributions over the same set of endogenous variables $\mathbf{V}$, such that the missing edges in $\mathcal{G}$ encode invariance constraints within $\mathbf{P}$.*



(a) Fork (b) Chain

Figure 4: Two causal diagrams encoding knowledge about the causal mechanisms governing three observable variables X, Z and Y.

Layer	Graphical Model	Distribution	Physical Sampling Procedure	Inferential Machinery	Related Lit.
$\mathcal{L}_1$	BN	$P(\mathbf{V})$	random sampling	d-separation	[18, 20]
$\mathcal{L}_2$	CBN	$P(\mathbf{V}\|do(\mathbf{x})), \forall \mathbf{x}$	randomized controlled trials (RCT)	do-calculus	[20, 2, 3]
$\mathcal{L}_3$	CTFBN	$P(\mathbf{W}_*), \forall \mathbf{W}_*$	counterfactual randomization	ctf-calculus	[6, 1]

Table 1: Sample graphical models and their corresponding distributions, physical sampling procedures, and inferential machinery for each layer of the PCH.

Depending on the assumptions made on different layers of the PCH, a different graphical model can be defined. As alluded to earlier, some examples of models corresponding to the three layers of the PCH are Bayesian Network (BN) [18], Causal Bayesian Network (CBN) [3], and Counterfactual Bayesian Network (CTFBN) [1, Sec. 13.2]. These graphical models are powerful tools for encoding assumptions to perform causal inference tasks such as identification (Fig. 1), with each model accompanied by its own inferential machinery like *d-separation* for BNs, *do-calculus* for CBNs, and *ctf-calculus* for CTFBNs [18, 20, 6].

As we ascend the PCH, the corresponding graphical models encode invariance constraints over increasingly richer sets of distributions. These enlarging sets of constraints naturally induce a hierarchy: models higher in the hierarchy support more powerful inferences but depend on stronger assumptions, which are correspondingly harder to verify empirically.

Example 1 (SCM and Graphical Models)**.** *Consider the SCM* $\mathcal{M} = \langle \mathbf{U} = \{U_x, U_z, U_y\}, \mathbf{V} = \{X, Z, Y\}, \mathcal{F}, P(\mathbf{u})\rangle$*, where*

$$\mathcal{F} = \begin{cases} X \leftarrow U_x \\ Z \leftarrow X \oplus U_z \\ Y \leftarrow X \oplus U_y \end{cases} \tag{7}$$

$$P(\mathbf{u}) : U_x \sim Bernoulli(0.2), U_z \sim Bernoulli(0.4), U_y \sim Bernoulli(0.3) \tag{8}$$

The endogenous variables $\mathbf{V}$ *represent, respectively, a treatment* X *(e.g., a diet) and outcomes* Z *and* Y *(e.g. patient's BMI and cholesterol level). The exogenous variables* U_x*,* U_z*, and* U_y *represent other variables outside the model that affect* X*,* Z*, and* Y*, respectively.*

Given the SCM, all quantities from the PCH can be computed following Def. 4, by mapping from each $\mathbf{U} = \mathbf{u}$ *to all potential outcomes derived from* $\mathcal{M}$*, as shown in Table 2.* $\mathcal{L}_1$ *distributions will only involve the observed variables* X*,* Z*, and* Y*, and* $\mathcal{L}_2$ *distributions will have additional access to the all potential outcome variables like* Z_x *and* $Y_{x'}$*, individually.* $\mathcal{L}_3$ *includes joint distributions over all potential outcomes in the table, capturing the full range of counterfactual dependencies.*

However, in practice, the SCM is often not observed with such details, and we analyze the invariance constraints that hold in the distributions it induces. For example, it can be calculated from Table 2 that

$$P(x) = P(x_y), \forall (x, y) \tag{9}$$

$$P(y_x, x) = P(y, x), \forall (x, y) \tag{10}$$

	1	2	3	4	5	6	7	8
U_x	0	0	0	0	1	1	1	1
U_z	0	0	1	1	0	0	1	1
U_y	0	1	0	1	0	1	0	1
X	0	0	0	0	1	1	1	1
Z	0	0	1	1	1	1	0	0
Y	0	1	0	1	1	0	1	0
$Z_{X=0}$	0	0	1	1	0	0	1	1
$Y_{X=0}$	0	1	0	1	0	1	0	1
$Z_{X=1}$	1	1	0	0	1	1	0	0
$Y_{X=1}$	1	0	1	0	1	0	1	0
$X_{Z=0}$	0	0	0	0	1	1	1	1
$Y_{Z=0}$	0	1	0	1	1	0	1	0
$X_{Z=1}$	0	0	0	0	1	1	1	1
$Y_{Z=1}$	0	1	0	1	1	0	1	0
$X_{Y=0}$	0	0	0	0	1	1	1	1
$Z_{Y=0}$	0	0	1	1	1	1	0	0
$X_{Y=1}$	0	0	0	0	1	1	1	1
$Z_{Y=1}$	0	0	1	1	1	1	0	0
$X_{Z=0,Y=0}$	0	0	0	0	1	1	1	1
$X_{Z=0,Y=1}$	0	0	0	0	1	1	1	1
$X_{Z=1,Y=0}$	0	0	0	0	1	1	1	1
$X_{Z=1,Y=1}$	0	0	0	0	1	1	1	1
$Z_{X=0,Y=0}$	0	0	1	1	0	0	1	1
$Z_{X=0,Y=1}$	0	0	1	1	0	0	1	1
$Z_{X=1,Y=0}$	1	1	0	0	1	1	0	0
$Z_{X=1,Y=1}$	1	1	0	0	1	1	0	0
$Y_{X=0,Z=0}$	0	1	0	1	0	1	0	1
$Y_{X=0,Z=1}$	0	1	0	1	0	1	0	1
$Y_{X=1,Z=0}$	1	0	1	0	1	0	1	0
$Y_{X=1,Z=1}$	1	0	1	0	1	0	1	0
$P(\mathbf{u})$	0.336	0.144	0.224	0.096	0.084	0.036	0.056	0.024

Table 2: Mapping of events in the space of $\mathbf{U}$ to potential outcomes in the context of Example 1.

These invariance constraints can be represented using the causal diagram shown in Fig. 4(a). When this causal diagram is interpreted as the graphical model for different layers of the PCH, it encodes different constraints according to the definitions of models:

- $\mathcal{L}_1$ *BN:*

$$P(x, y, z) = P(x)P(z|x)P(y|x) \tag{11}$$

- $\mathcal{L}_2$ *CBN:*

  (i)

$$P(x, y, z) = P(x)P(z|x)P(y|x) \tag{12}$$
$$P(y, z|do(x)) = P(y|do(x))P(z|do(x)) \tag{13}$$

  (ii)

$$P(x|do(\mathbf{a})) = P(x), \forall \mathbf{a} \subseteq \{z, y\} \tag{14}$$
$$P(z|do(x, y)) = P(z|do(x)) \tag{15}$$
$$P(y|do(x, z)) = P(y|do(x)) \tag{16}$$

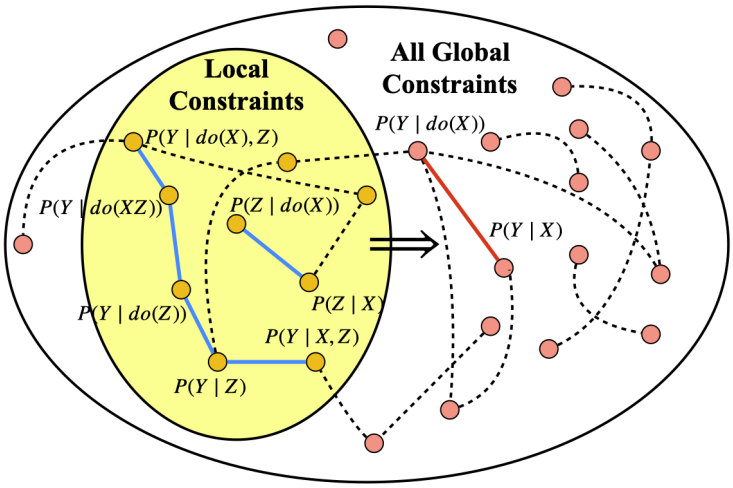


Figure 5: Constraints listed in the definition of a graphical model serves as a local basis that implies all constraints encoded in the model. Blue lines represent a set of local invariance constraints that can be composed to imply the global constraint represented by the red line.

(iii)

$$P(z|do(x)) = P(z|x) \tag{17}$$
$$P(z|do(x,y)) = P(z|do(y),x) \tag{18}$$
$$P(y|do(x)) = P(y|x) \tag{19}$$
$$P(y|do(x,z)) = P(y|do(z),x) \tag{20}$$

- $\mathcal{L}_3$ *CTFBN:*

(i)

$$P(X, Z_x, Z_{x'}, Y_x, Y_{x'}) = P(X)P(Z_x, Z_{x'})P(Y_x, Y_{x'}) \tag{21}$$

(ii)

$$P(x_{\mathbf{a}}, \mathbf{w}_*) = P(x, \mathbf{w}_*), \forall \mathbf{a} \subseteq \{z, y\} \tag{22}$$
$$P(z_{xy}, \mathbf{w}_*) = P(z_x, \mathbf{w}_*) \tag{23}$$
$$P(y_{xz}, \mathbf{w}_*) = P(y_x, \mathbf{w}_*) \tag{24}$$

(iii)

$$P(z, x, \mathbf{w}_*) = P(z_x, x, \mathbf{w}_*) \tag{25}$$
$$P(z_y, x_y, \mathbf{w}_*) = P(z_{xy}, x_y, \mathbf{w}_*) \tag{26}$$
$$P(y, x, \mathbf{w}_*) = P(y_x, x, \mathbf{w}_*) \tag{27}$$
$$P(y_z, x_z, \mathbf{w}_*) = P(y_{xz}, x_z, \mathbf{w}_*) \tag{28}$$

Interestingly, the constraints explicitly listed in the definitions of graphical models represent only a subset of all the constraints implied by the model. These explicitly stated constraints are typically local, involving variables and their immediate parents. Importantly, these local constraints form a "basis" from which all other (global) constraints in the model can be spanned, as illustrated in Fig. 5. The process of composing local constraints to derive global ones underpins the operation of the causal inference engine.

For example, consider the constraint:

$$P(y|do(z)) = P(y) \tag{29}$$

This is a constraint that does not involve the parents and does not appear in the local basis, since X which is in the parents of Y does not appear in the expression. Still, it can be derived by combining

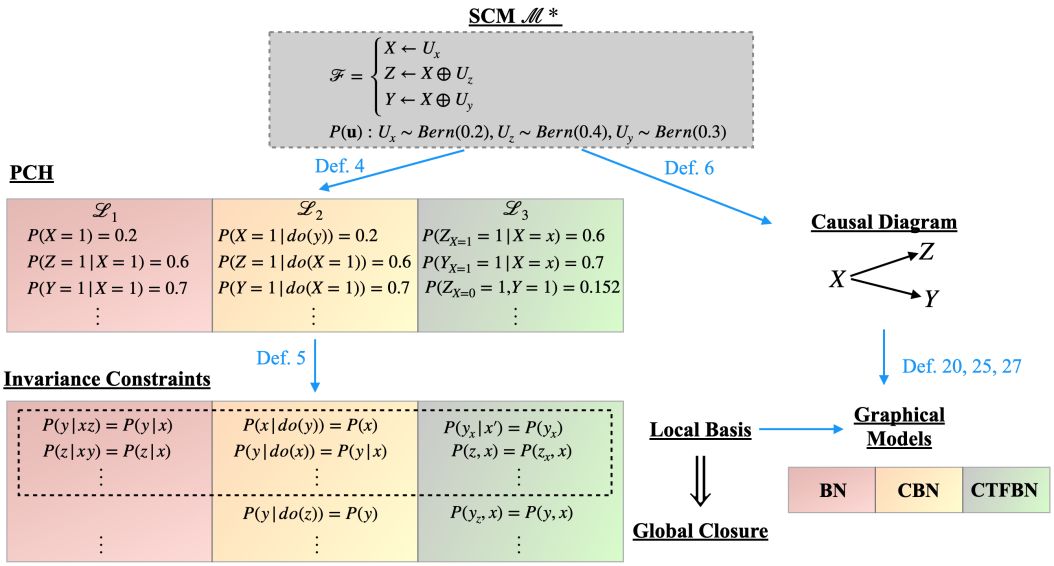


Figure 6: Illustration of how an SCM induces the PCH, invariance constraints, causal diagram, and graphical models, following Example 1.

various local constraints as shown next:

$$P(y|do(z)) = \sum_x P(y|do(z),x)P(x|do(z)) \tag{30}$$

$$= \sum_x P(y|do(zx))P(x|do(z)) \qquad (Eq.20) \tag{31}$$

$$= \sum_x P(y|do(x))P(x|do(z)) \qquad (Eq.16) \tag{32}$$

$$= \sum_x P(y|x)P(x|do(z)) \qquad (Eq.19) \tag{33}$$

$$= \sum_x P(y|x)P(x) \qquad (Eq.14) \tag{34}$$

$$= P(y) \tag{35}$$

The connections among these different moving parts – the SCM, the PCH, the causal diagram, the invariance constraints and the graphical model – are illustrated in Fig. 6. The constraints in each model determines its inferential power. Given the $\mathcal{L}_1$ constraints, the only inference can be drawn is that Y and Z are independent conditional on X in the observational distributions. However, with the $\mathcal{L}_2$ constraints, the causal effect from the treatment to the outcome can be inferred, and in this case it coincides with their observational correlation (i.e. $P(y|do(x)) = P(y|x)$). If we are able to interpret the causal diagram as an $\mathcal{L}_3$ object, say a CTFBN, the local constraints can be leveraged to infer that the effect of the treatment on the treated (ETT). To witness, the ETT is also equal to the conditional distribution, $P(y_x|x') = P(y|x)$. ■

An agent may sometimes interact with a system of interest through experiments, thereby collecting data from different layers of the PCH. Counterfactual randomization is an experimental procedure that enables an agent to observe the value of a variable before an intervention takes effect [4]. For instance, a doctor may be able to determine a patient's natural choice of drug prior to randomly assigning treatment in a clinical trial. This extension of experimental capability is formalized in the following definition of a new type of physical action that an agent may be able to perform in an environment.

Definition 8 (Counterfactual (ctf-) Randomization (Def. 2.3 [24]))**.** *For a variable X and some particular unit i[2] in the target population of the environment, the operation*

$$\text{CTF-RAND}(X \to \mathbf{C})^{(i)} \tag{36}$$

denotes fixing the value of X as an input to the mechanisms generating $\mathbf{C} \subseteq Ch(X)$ for this particular unit, where $Ch(X)$ is the set of variables whose mechanisms take X as an argument.

The value of X is assigned by a randomizing device with support over $\text{Domain}(X)$.

The essential differences between Fisherian randomization and $\text{CTF-RAND}(X \to \mathbf{C})^{(i)}$ are:

1. CTF-RAND does not erase unit i's natural decision $X^{(i)}$[3].
2. While Fisherian randomization affects all children of X, CTF-RAND only affects the chosen subset $\mathbf{C} \subseteq Ch(X)$, leaving $Ch(X) \setminus \mathbf{C}$ untouched.

Importantly, CTF-RAND can only be enacted under certain structural conditions. These include environments where one can measure a unit's natural decision while simultaneously randomizing its actual decision [4], or settings where counterfactual mediators allow altering how a subset of children perceive the value of X [24]. In either case, ctf-randomization enables multiple randomizations on the same variable X for a single unit i. Further, CTF-RAND must always be applied with respect to a graphical child variable; it is not possible to bypass a child and directly alter the perception of a descendant.

Example 2 (CTF-RAND)**.** *Consider the SCM from Example 1. Counterfactual randomization on X allows an agent to observe the natural value of X, say x', while simultaneously assigning a specific value x as input to its children Z and Y. This is illustrated graphically in Fig. 8 (b), and as a result, the $\mathcal{L}_3$ distribution $P(X = x', Z_{X=x}, Y_{X=x})$ becomes experimentally accessible (i.e., realizable).* ■

3 CBN2.25 and CBN2.5: Graphical Models for Realizable Constraints

In this section, we provide a fine-grained analysis of the counterfactual layer ($\mathcal{L}_3$) by circumscribing the subsets of distributions that are realizable, given the feasible action set. We assume that all actions required to sample from any $\mathcal{L}_2$ distribution are available, together with certain counterfactual randomization capabilities. Specifically, we define two distinct collections of realizable distributions, each determined by a different degree of flexibility in how counterfactual randomization propagates to downstream variables (Sec.3.1). We then introduce the corresponding graphical models that encode the invariance constraints inherent to these distribution sets (Sec.3.2), followed by the inferential machinery tailored to each model class (Sec. 3.3).

3.1 Formal Languages for Realizable Counterfactual Distributions

Before introducing the two layers of language, we first provide two definitions of interventional sets that help distinguish between them.

Definition 9 (Interventional Variable Set)**.** *Given a set of random variables $\mathbf{V}$, an* interventional variable set *is a subset of $\mathbf{V}$ on which an intervention is performed.*

Definition 10 (Interventional Value Set)**.** *Given a set of random variables $\mathbf{V}$, an* interventional value set*, $\mathbf{x}$, is a specific assignment of values to an interventional variable set $\mathbf{X} \subseteq \mathbf{V}$. That is, $\mathbf{x} \in Val(\mathbf{X})$.*

For each interventional variable set $\mathbf{X}$, there may exist multiple corresponding interventional value sets $\mathbf{x}$ drawn from $Val(\mathbf{X})$.

Example 3 (Interventional Variable Set and Interventional Value Set)**.** *Consider the variables from the SCM in Example 1. An example of an interventional variable set is $\{X\}$, with interventional value sets $\{X = 0\}$ and $\{X = 1\}$.* ■

[2]This definition discusses a unit-specific experimental procedure, as it takes a physical perspective on how an agent interacts with the units in a system.

[3]Another way to understand this difference is that the unit's natural inclination is taken into account.

The first collection of realizable distributions is defined under the assumption that each CTF-RAND on X fixes a single value of x across all its children. [4] The symbolic representation and valuation of distributions in this collection, given an SCM, are provided below.

Definition 11 (Layer 2.25 ($\mathcal{L}_{2.25}$))**.** *An SCM $\mathcal{M} = \langle \mathbf{U}, \mathbf{V}, \mathcal{F}, P(\mathbf{u}) \rangle$ induces a family of joint distributions over $\mathbf{V}$, indexed by each interventional value set $\mathbf{x}$. For each $\mathbf{X}, \mathbf{Y} \subseteq \mathbf{V}$ and $\mathbf{x} \in Val(\mathbf{X})$:*

$$P^{\mathcal{M}}\left(\bigwedge_{V_i \in \mathbf{Y} \setminus \mathbf{X}} V_{i_{[\mathbf{x}_i]}} = v_i, \bigwedge_{V_i \in \mathbf{Y} \cap \mathbf{X},\, v_i = V_i \cap \mathbf{x}} V_{i_{[\mathbf{x}_i \setminus v_i]}} = v_i\right) = \sum_{\mathbf{u}} \mathbf{1}\left[\bigwedge_{V_i \in \mathbf{Y} \setminus \mathbf{X}} V_{i_{[\mathbf{x}_i]}}(\mathbf{u}) = v_i, \bigwedge_{V_i \in \mathbf{Y} \cap \mathbf{X},\, v_i = V_i \cap \mathbf{x}} V_{i_{[\mathbf{x}_i \setminus v_i]}}(\mathbf{u}) = v_i\right] P(\mathbf{u}). \tag{37}$$

subject to the following conditions:

(i) $\mathbf{x}_i \subseteq \mathbf{x}$ *and* $\bigcup_i \mathbf{x}_i = \mathbf{x}$*;*

(ii) for any $v_i \in \mathbf{x}$ *and all* $V_j \in \mathbf{Y}$*, if* $V_i \in An(V_j)$ *in* $\mathcal{M}_{\mathbf{x} \setminus V_j}$*, then* $v_i \in \mathbf{x}_j$*.*

The collection of all such distributions is denoted $\mathbf{P}^{\mathcal{L}_{2.25}}$*.*

Cond. (i) of Def. 11 ensures that only assignments from the intervention value set $\mathbf{x}$ appear in the subscripts, and that each value in $\mathbf{x}$ appears at least once. This prevents redundancy in representing the same distribution under different intervention value sets, e.g., when $\mathbf{x} \subset \mathbf{x}'$.

Example 4 (Cond. (i) of Def.11)**.** *Consider the SCM from Example1 and two interventional value sets: $\emptyset$ and $Y = y$.*

The empty intervention set $\emptyset$ indexes the distribution $P(Z, X, Y)$, where no variables are intervened on and all subscripts are omitted. This aligns with the interpretation that an empty intervention corresponds to observational data.

For the interventional set $\{Y = y\}$, if condition (i) were not imposed, $P(Z, X, Y)$ could also be indexed, since Y is not an ancestor of either X or Z. This would lead to the same distribution being redundantly enumerated under different interventional sets.

Condition (i) avoids this by requiring the union of all subscripts to cover the entire intervention value set. In other words, y must appear as a subscript in at least one of the counterfactual terms. As a result, the enumeration based on $\{Y = y\}$ does not index $P(Z, X, Y)$, but instead produces distributions such as $P(Z_y, X, Y)$, $P(Z, X_y, Y)$, and $P(Z_y, X_y, Y)$. ■

Condition (ii) requires all descendants of an intervened variable X to share the same value x, unless the path from X to a descendant is blocked by another variable in the intervention set. This restriction, together with condition (i), reflects the limited flexibility allowed under the counterfactual randomization action.

Example 5 (Cond. (ii) of Def.11)**.** *Consider again the SCM from Example1 with the interventional value set $\{X = x\}$. Since X is an ancestor of both Y and Z, the value x must appear in the subscripts of both variables in order to satisfy condition (ii) of Def. 11. Consequently, $\{X = x\}$ indexes only the distribution $P(X, Z_x, Y_x)$.* ■

The second collection of distributions relaxes this restriction by allowing each child of X to receive a potentially different randomized value. This more flexible form of counterfactual randomization expands the class of realizable distributions beyond those in the first collection.

Definition 12 (Layer 2.5 ($\mathcal{L}_{2.5}$))**.** *An SCM $\mathcal{M} = \langle \mathbf{U}, \mathbf{V}, \mathcal{F}, P(\mathbf{u}) \rangle$ induces a family of probability distributions over $\mathbf{V}$ indexed by each interventional variable set $\mathbf{X}$. For each $\mathbf{Y}, \mathbf{X} \subseteq \mathbf{V}$:*

$$P^{\mathcal{M}}\Big(\bigwedge_{V_i \in \mathbf{Y} \setminus \mathbf{X}} V_{i_{[\mathbf{x}_i]}} = v_i, \bigwedge_{V_i \in \mathbf{Y} \cap \mathbf{X},\, v_i = V_i \cap \mathbf{x}} V_{i_{[\mathbf{x}_i \setminus v_i]}} = v_i\Big) = \sum_{\mathbf{u}} \mathbf{1}\Big[\bigwedge_{V_i \in \mathbf{Y} \setminus \mathbf{X}} V_{i_{[\mathbf{x}_i]}}(\mathbf{u}) = v_i, \bigwedge_{V_i \in \mathbf{Y} \cap \mathbf{X},\, v_i = V_i \cap \mathbf{x}} V_{i_{[\mathbf{x}_i \setminus v_i]}}(\mathbf{u}) = v_i\Big] P(\mathbf{u}) \tag{38}$$

[4] For both $\mathcal{L}_{2.25}$ and $_{2.5}$, we assume that counterfactual randomization is allowed for all variables in $\mathbf{V}$.

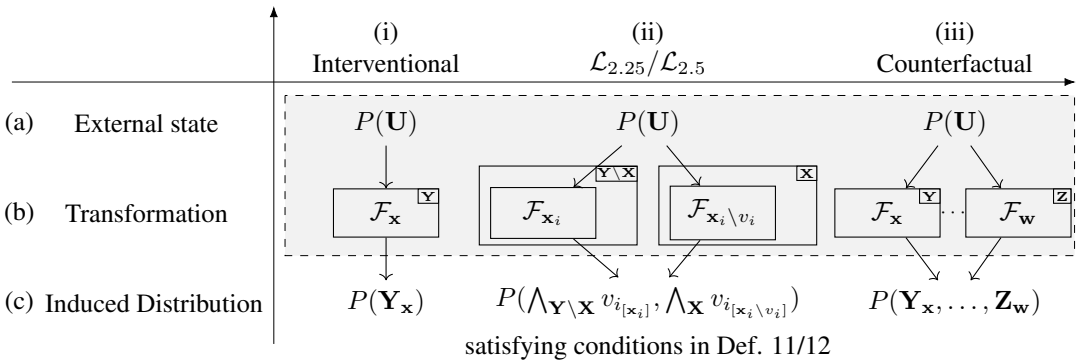


Figure 7: Given an SCM's initial state (i.e., population) (a), we show the different functional transformations (b) and the corresponding induced distribution (c) of each layer of the hierarchy. (i) represents the transformation (i.e., $\mathcal{F}$) from the natural state of the system ($P(\mathbf{U})$) to an interventional world (i.e., with modified mechanisms $\mathcal{F}_{\mathbf{x}}$), (ii) to multiple counterfactual worlds representing $\mathcal{L}_{2.25}/\mathcal{L}_{2.5}$, and (iii) to multiple counterfactual worlds with no constraints on the worlds joint.

subject to the following conditions:

(i) $\mathbf{X}_i \subseteq \mathbf{X}$, $\mathbf{x}_i \in Val(\mathbf{X}_i)$, *and* $\bigcup_i \mathbf{X}_i = \mathbf{X}$.

(ii) For any V_i *and any* $B \in \mathbf{X} \cap \mathbf{Pa}(V_i)$, *and for all* $V_j \in \mathbf{Y}$: *if* $V_i \notin \mathbf{X}_j$ *and* $V_i \in An(V_j)$ *in* $\mathcal{M}_{\mathbf{x}_j}$, *then* $\mathbf{x}_i \cap B = \mathbf{x}_j \cap B$.

The collection of all such distributions is denoted by $\mathbf{P}^{\mathcal{L}_{2.5}}$.

Def. 11 and Def. 12 serve as templates for enumerating the distributions in $\mathcal{L}_{2.25}$ and $\mathcal{L}_{2.5}$. The key distinction lies in the way distributions are indexed: $\mathcal{L}_{2.25}$ distributions are indexed by specific interventional value sets $\mathbf{x} \in Val(\mathbf{X})$, whereas $\mathcal{L}_{2.5}$ distributions are indexed by interventional variable sets $\mathbf{X}$.

The more specific indexing in $\mathcal{L}_{2.25}$ imposes stronger restrictions on the expressiveness of its distributions, which is reflected in the corresponding conditions. Similar to Def. 11, condition (i) of Def. 12 ensures that each variable in the intervention set appears at least once in the subscript. However, it relaxes Def. 11 by allowing multiple value assignments for the interventional variable set $\mathbf{X}$. Condition (ii) is likewise relaxed: instead of requiring value consistency to propagate from X itself, it enforces consistency only at the level of X's children.

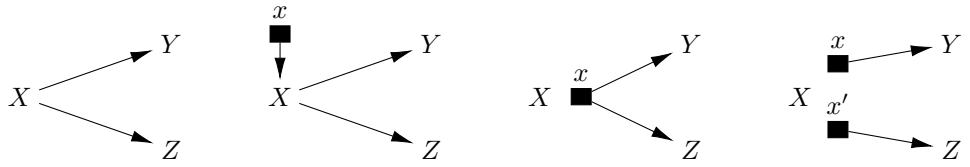
Example 6 (SCM inducing $\mathcal{L}_{2.25}/\mathcal{L}_{2.5}$)**.** *Consider the SCM from Example 1.*

The distribution $P(X, Y_x, Z_x)$, *indexed by the interventional value set* $\{X = x\}$, *belongs to* $\mathcal{L}_{2.25}$. *It satisfies condition (i) of Def. 11 by having only* x *in the subscript, and condition (ii) by enforcing consistent subscripts across all children of* X, *i.e.,* Y *and* Z.

In contrast, the distribution $P(X, Y_x, Z_{x'})$ *does not belong to* $\mathcal{L}_{2.25}$ *because it contains conflicting value assignments for* X, *making it unindexable by any specific interventional value set. However, it does belong to* $\mathcal{L}_{2.5}$ *since the conditions in Def. 12 allow different value assignments for the same variable in the intervention set. This difference between the two layers is illustrated in Fig. 8(b) and (c).*

Finally, the $\mathcal{L}_3$ *distribution* $P(Y_x, Y)$ *lies outside both languages, as it includes the same variable* Y *under two different submodels, which is not permitted in* $\mathcal{L}_{2.25}$ *or* $\mathcal{L}_{2.5}$. ■

The counterfactual variables in the symbolic representation of $\mathcal{L}_{2.25}$ and $\mathcal{L}_{2.5}$ are all of the form $Y_{\mathbf{x}}$, where the subscript $\mathbf{x}$ indicates that an intervention $do(\mathbf{X} = \mathbf{x})$ has been performed in the system. There is another type of counterfactual variables which represents interventions like $do(\mathbf{X} = \mathbf{X}_{\mathbf{z}})$, where the variable $\mathbf{X}$ is set to behave as another counterfactual variable, say $\mathbf{X}_{\mathbf{z}}$. A random variable Y in such a system is represented with a counterfactual of the form $Y_{\mathbf{X}_{\mathbf{z}}}$, which is called a *nested counterfactual*.



(a) Causal Diagram $\mathcal{G}$ (b) $\mathcal{L}_2$ (c) $\mathcal{L}_{2.25}$ (d) $\mathcal{L}_{2.5}$

Figure 8: Differences in how intervention on X affects downstream variables in $\mathcal{L}_2$, $\mathcal{L}_{2.25}$ and $\mathcal{L}_{2.5}$.

All nested counterfactuals can be unnested via the Counterfactual Unnesting (CUT) process below and be transformed into non-nested ones.

Corollary 1 (Counterfactual Unnesting (CUT) [6])**.** *Let $Y, X \in \mathbf{V}, \mathbf{T}, \mathbf{Z} \subseteq \mathbf{V}$, and let z be a set of values for Z. Then,the nested counterfactual $P(Y_{\mathbf{T}_* X_\mathbf{z}} = y)$ can be written with one less level of nesting as:*

$$P(Y_{\mathbf{T}_* X_\mathbf{z}} = y) = \sum_x P(Y_{\mathbf{T}_* x} = y, X_\mathbf{z} = x) \tag{39}$$

Nested counterfactuals may also belong to $\mathcal{L}_{2.25}$ and $\mathcal{L}_{2.5}$, provided that their unnested form, derived via the Counterfactual Unnesting Theorem, contains only distributions admissible within the corresponding layer.

Lemma 1 (Nested Counterfactuals in $\mathcal{L}_{2.25}$/$\mathcal{L}_{2.5}$)**.** *A nested counterfactual belongs to $\mathcal{L}_{2.25}$/$\mathcal{L}_{2.5}$ if and only if there exists a sequence of applications of the CUT procedure that reduces it to a function of unnested counterfactuals in $\mathcal{L}_{2.25}$/$\mathcal{L}_{2.5}$.*

Example 7 (Nested Counterfactual - Natural Direct Effect (NDE))**.** *Consider the causal diagram in Fig. 11. The natural direct effect from X to Y can be expressed in counterfactual notation as*

$$NDE_{x,x'}(y) = P(y_{x',Z_x}) - P(y_x). \tag{40}$$

The first term, $P(y_{x',Z_x})$, is a nested counterfactual. By applying the CUT procedure, we obtain its unnested expression:

$$P(y_{x',Z_x}) = \sum_z P(y_{x'z}, z_x). \tag{41}$$

From this expression, we can conclude that $P(Y_{x'z}, Z_x)$ lies in $\mathcal{L}_{2.5}$ since satisfies the conditions in Def. 12. However, it does not belong to $\mathcal{L}_{2.25}$ because of the conflicting subscripts x and x' across the joint counterfactuals.

The evaluation processes for distributions in these two new layers are illustrated in Fig. 7, with interventional distributions ($\mathcal{L}_2$) shown on the left and full counterfactual distributions ($\mathcal{L}_3$) on the right. In $\mathcal{L}_{2.25}$ and $\mathcal{L}_{2.5}$, a variable in $\mathbf{Y}$ is always evaluated within one submodel: each intervened variable in $\mathbf{X}$ is evaluated in its own submodel $\mathcal{M}_{\mathbf{x}_i \setminus v_i}$, while each non-intervened variable is evaluated in $\mathcal{M}_{\mathbf{x}_i}$ according to the value of $\mathbf{X}$ it receives. The submodels in $\mathcal{L}_{2.25}$ and $\mathcal{L}_{2.5}$ are further constrained by the conditions in Def. 11 and Def. 12, respectively.

Comparing across layers in the PCH, $\mathcal{L}_{2.25}$ and $\mathcal{L}_{2.5}$ are more expressive than $\mathcal{L}_2$, since $\mathcal{L}_2$ evaluates all variables in $\mathbf{Y}$ within a single submodel, while $\mathcal{L}_{2.25}$ and $\mathcal{L}_{2.5}$ allow joint evaluation across multiple submodels (i.e., counterfactual worlds). At the same time, they are less expressive than the full $\mathcal{L}_3$, which imposes no such restrictions on which submodels may be joined.

3.2 Graphical Models

With these new collections of distributions defined, we now introduce two graphical models that encode the corresponding constraints and compatibility relations. This construction builds on the $\mathcal{L}_3$ models introduced and discussed in detail in [1, Sec. 13].

Definition 13 (Causal Bayesian Network 2.25 (CBN2.25, Semi-Markovian))**.** *Given a graph with directed and bidirected edges, $\mathcal{G}$, and let $\mathbf{P}^{\mathcal{L}_{2.25}}$ be the collection of all $\mathcal{L}_{2.25}$ distributions over $\mathbf{V}$. Then, $\mathcal{G}$ is a CBN2.25 for $\mathbf{P}^{\mathcal{L}_{2.25}}$ if the following hold:*

(i) ***Independence Restrictions.*** *For a fixed intervention value set $\mathbf{v} \in Val(\mathbf{V})$ and a subset of variables $\mathbf{W} \subseteq \mathbf{V}$, let $\mathbf{W}_*$ be the set of counterfactuals of the form $W_{\mathbf{pa}_w}$ with $\mathbf{pa}_w$ taking values in $\mathbf{v}$. Let $\mathbf{C}_1, ..., \mathbf{C}_l$ be the c-components of $G[\mathbf{V}(\mathbf{W}_*)]$, and $\mathbf{C}_{1*}, ..., \mathbf{C}_{l*}$ the corresponding partition over $\mathbf{W}_*$. Then $P(\mathbf{W}_*)$ factorizes as*

$$P\left(\bigwedge_{W_{pa_w} \in \mathbf{W}_*} W_{pa_w}\right) = \prod_{j=1}^{l} P\left(\bigwedge_{W_{pa_w} \in \mathbf{C}_{j*}} W_{pa_w}\right). \tag{42}$$

(ii) ***Exclusion Restrictions.*** *For every variable $Y \in \mathbf{V}$ with parents $\mathbf{Pa}_y$, for every set $\mathbf{Z} \subseteq \mathbf{V} \backslash (\mathbf{Pa}_y \cup \{Y\})$, and any counterfactual set $\mathbf{W}_*$ such that $P(Y_{\mathbf{pa}_y, \mathbf{z}}, \mathbf{W}_*) \in \mathbf{P}^{\mathcal{L}_{2.25}}$,*

$$P(Y_{\mathbf{pa}_y, \mathbf{z}}, \mathbf{W}_*) = P(Y_{\mathbf{pa}_y}, \mathbf{W}_*). \tag{43}$$

(iii) ***Consistency Restrictions.*** *For every variable $Y \in \mathbf{V}$ with parents $\mathbf{Pa}_y$, $\mathbf{X} \subseteq \mathbf{Pa}_y$, for every set $\mathbf{Z} \subseteq \mathbf{V} \backslash (\mathbf{X} \cup \{\mathbf{Y}\})$, and any counterfactual set $\mathbf{W}_*$ such that $P(Y_{\mathbf{xz}} = y, \mathbf{X}_\mathbf{z} = \mathbf{x}, \mathbf{W}_*) \in \mathbf{P}^{\mathcal{L}_{2.25}}$,*

$$P(Y_\mathbf{z} = y, \mathbf{X}_\mathbf{z} = \mathbf{x}, \mathbf{W}_*) = P(Y_{\mathbf{xz}} = y, \mathbf{X}_\mathbf{z} = \mathbf{x}, \mathbf{W}_*). \tag{44}$$

This definition closely resembles CTFBN, sharing the same types of constraints but restricted to the subset of distributions circumscribed by $\mathcal{L}_{2.25}$. Condition (i) requires that variables not sharing latent confounders be jointly independent once their parents are fixed by intervention. Condition (ii) states that once the parents of a variable Y have been fixed, no further intervention can affect the value of Y, regardless of any other observation. Finally, condition (iii) connects observations and interventions: if a parent X of Y is observed to take the value x while both X and Y are under the same intervention $do(Z = z)$, this is equivalent to intervening on Y by $do(Z = z, X = x)$. Importantly, the next proposition establishes that a causal diagram $\mathcal{G}$ induced by an SCM $\mathcal{M}$ is a CBN2.25 for the $\mathcal{L}_{2.25}$ distribution generated by $\mathcal{M}$.

Theorem 1 ($\mathcal{L}_{2.25}$-Connection — CBN2.25 (Markovian and Semi-Markovian))**.** *The causal diagram $\mathcal{G}$ induced by the SCM $\mathcal{M}$, following the constructive procedure in Def. 6, is a CBN2.25 for $\mathbf{P}^{\mathcal{L}_{2.25}}$, the collection of all $\mathcal{L}_{2.25}$ distributions induced by $\mathcal{M}$.*

Example 8 (CBN2.25)**.** *Given the SCM in Example 1, the fork, the pair $\langle \mathcal{G}, \mathbf{P}^{\mathcal{L}_{2.25}} \rangle$ is a CBN2.25, where $\mathcal{G}$ denotes the causal diagram in Fig. 4(a) and $\mathbf{P}^{\mathcal{L}_{2.25}}$ satisfies the following constraints:*

(i) ***Independence Restrictions***

$$P(X, Y_x, Z_x) = P(X)P(Y_x)P(Z_x) \tag{45}$$

(ii) ***Exclusion Restrictions***

$$P(X_\mathbf{a} = x, \mathbf{W}_*) = P(X = x, \mathbf{W}_*), \quad \mathbf{a} \subseteq \{z, y\} \tag{46}$$
$$P(Y_{xz} = y, \mathbf{W}_*) = P(Y_x = y, \mathbf{W}_*) \tag{47}$$
$$P(Z_{xy} = z, \mathbf{W}_*) = P(Z_x = z, \mathbf{W}_*) \tag{48}$$

(iii) ***Local Consistency***

$$P(Y = y, X = x, \mathbf{W}_*) = P(Y_x = y, X = x, \mathbf{W}_*) \tag{49}$$
$$P(Y_z = y, X_z = x, \mathbf{W}_*) = P(Y_{zx} = y, X_z = x, \mathbf{W}_*) \tag{50}$$
$$P(Z = z, X = x, \mathbf{W}_*) = P(Z_x = z, X = x, \mathbf{W}_*) \tag{51}$$
$$P(Z_y = z, X_y = x, \mathbf{W}_*) = P(Z_{yx} = z, X_y = x, \mathbf{W}_*) \tag{52}$$

Here, $\mathbf{W}_*$ can be any set of counterfactual variables such that $P(\cdot) \in \mathbf{P}^{\mathcal{L}_{2.25}}$. ■

Similarly, a graphical model for $\mathcal{L}_{2.5}$ can be defined by imposing the same types of constraints on distributions restricted to $\mathcal{L}_{2.5}$. In this case, the causal diagram $\mathcal{G}$ induced by an SCM $\mathcal{M}$ is also a CBN2.5 for the $\mathcal{L}_{2.5}$ distributions generated by $\mathcal{M}$.

Definition 14 (Causal Bayesian Network 2.5 (CBN2.5, Semi-Markovian))**.** *Given a graph with directed and bidirected edges, $\mathcal{G}$, and let $\mathbf{P}^{\mathcal{L}_{2.5}}$ be the collection of all $\mathcal{L}_{2.5}$ distributions over $\mathbf{V}$. Then, $\mathcal{G}$ is a CBN2.5 for $\mathbf{P}^{\mathcal{L}_{2.5}}$ if the following hold:*

(i) ***Independence Restrictions.*** *Let $\mathbf{W}_*$ be a set of counterfactuals of the form $W_{\mathbf{pa}_w}$ with distinct W. Let $\mathbf{C}_1, ..., \mathbf{C}_l$ be the c-components of $G[\mathbf{V}(\mathbf{W}_*)]$, and $\mathbf{C}_{1*}, ..., \mathbf{C}_{l*}$ the corresponding partition over $\mathbf{W}_*$ such that $P(\mathbf{W}_*) \in \mathbf{P}^{\mathcal{L}_{2.5}}$. Then $P(\mathbf{W}_*)$ factorizes as*

$$P\left(\bigwedge_{W_{pa_w} \in \mathbf{W}_*} W_{pa_w}\right) = \prod_{j=1}^{l} P\left(\bigwedge_{W_{pa_w} \in \mathbf{C}_{j*}} W_{pa_w}\right). \tag{53}$$

(ii) ***Exclusion Restrictions.*** *For every variable $Y \in \mathbf{V}$ with parents $\mathbf{Pa}_y$, for every set $\mathbf{Z} \subseteq \mathbf{V} \backslash (\mathbf{Pa}_y \cup \{Y\})$, and any counterfactual set $\mathbf{W}_*$ such that $P(Y_{\mathbf{pa}_y, \mathbf{z}}, \mathbf{W}_*) \in \mathbf{P}^{\mathcal{L}_{2.5}}$,*

$$P(Y_{\mathbf{pa}_y, \mathbf{z}}, \mathbf{W}_*) = P(Y_{\mathbf{pa}_y}, \mathbf{W}_*). \tag{54}$$

(iii) ***Consistency Restrictions.*** *For every variable $Y \in \mathbf{V}$ with parents $\mathbf{Pa}_y$, let $\mathbf{X} \subseteq \mathbf{Pa}_y$. Then, for every set $\mathbf{Z} \subseteq \mathbf{V} \backslash (\mathbf{X} \cup \{\mathbf{Y}\})$, and any counterfactual set $\mathbf{W}_*$ such that $P(Y_{\mathbf{xz}} = y, \mathbf{X}_\mathbf{z} = \mathbf{x}, \mathbf{W}_*) \in \mathbf{P}^{\mathcal{L}_{2.5}}$,*

$$P(Y_\mathbf{z} = y, \mathbf{X}_\mathbf{z} = \mathbf{x}, \mathbf{W}_*) = P(Y_{\mathbf{xz}} = y, \mathbf{X}_\mathbf{z} = \mathbf{x}, \mathbf{W}_*). \tag{55}$$

Theorem 2 ($\mathcal{L}_{2.5}$-Connection — CBN2.5 (Markovian and Semi-Markovian))**.** *The causal diagram $\mathcal{G}$ induced by the SCM $\mathcal{M}$, following the constructive procedure in Def. 6, is a CBN2.5 for $\mathbf{P}^{\mathcal{L}_{2.5}}$, the collection of all $\mathcal{L}_{2.5}$ distributions induced by $\mathcal{M}$.*

Example 9 (CBN2.5)**.** *Consider the SCM from Example 6. The causal diagram it induces is shown in Fig. 4(a), and the collection of realizable distributions $\mathbf{P}^{\mathcal{L}_{2.5}}$ satisfies the following constraints:*

(i) ***Independence Restrictions.***

$$P(X, Y_x, Z_{x'}) = P(X)\, P(Y_x)\, P(Z_{x'}). \tag{56}$$

(ii) ***Exclusion Restrictions.***

$$P(X_\mathbf{a} = x, \mathbf{W}_*) = P(X = x, \mathbf{W}_*), \quad \mathbf{a} \subseteq \{z, y\}, \tag{57}$$
$$P(Y_{xz} = y, \mathbf{W}_*) = P(Y_x = y, \mathbf{W}_*), \tag{58}$$
$$P(Z_{xy} = z, \mathbf{W}_*) = P(Z_x = z, \mathbf{W}_*). \tag{59}$$

(iii) ***Consistency Restrictions.***

$$P(Y = y, X = x, \mathbf{W}_*) = P(Y_x = y, X = x, \mathbf{W}_*), \tag{60}$$
$$P(Y_z = y, X_z = x, \mathbf{W}_*) = P(Y_{zx} = y, X_z = x, \mathbf{W}_*), \tag{61}$$
$$P(Z = z, X = x, \mathbf{W}_*) = P(Z_x = z, X = x, \mathbf{W}_*), \tag{62}$$
$$P(Z_y = z, X_y = x, \mathbf{W}_*) = P(Z_{yx} = z, X_y = x, \mathbf{W}_*). \tag{63}$$

Here, $\mathbf{W}_*$ denotes any set of counterfactual variables such that $P(\cdot) \in \mathbf{P}^{\mathcal{L}_{2.5}}$. ■

Comparing Examples 8 and 9, both models share the same types of invariance constraints, but differ in the distributions to which those constraints apply. For instance, the independence restrictions for CBN2.25 apply to the $\mathcal{L}_{2.25}$ distribution $P(X, Z_x, Y_x)$, where Z and Y share the same subscript x. In contrast, the independence restrictions for CBN2.5 applies to richer $\mathcal{L}_{2.5}$ distributions such as $P(X, Z_x, Y_{x'})$, where the subscripts for Z and Y may differ. The same pattern also characterizes the $\mathcal{L}_3$ model, CTFBN, which uses the same types of constraints but applies them to even more expressive joint distributions over counterfactuals. For example, the corresponding independence restriction in the CTFBN induced by the SCM of Example 8 applies to the $\mathcal{L}_3$ distribution $P(X, Z_x, Z_{x'}, Y_x, Y_{x'})$, where potential outcomes for the same variables are jointly represented.

3.3 Inferential Machinery

As discussed earlier, the listed constraints in the definitions of CBN2.25 and CBN2.5 are *local*, in the sense that they involve counterfactual variables together with their parents (or augmented parents in semi-Markovian models). These local constraints serve as the building blocks from which more *global* statements can be derived, involving variables that may be far apart in the system. This deductive process provides the foundation for what is known as causal inference.

Example 10 (Local to Global Constraints — Fork)**.** *Consider the CBN2.25 from Example 8 and the distributions $P(y_z, x)$ and $P(y, x)$. One may ask how these two distributions are related, for instance, whether $P(y_z, x) = P(y, x)$. This relation cannot be read off directly from the model, since it does not appear as a local constraint in the basis (Example 8).*

However, it can be derived by composing several local constraints, as shown below:

$$\begin{aligned} P(y_z, x) &= P(y_z, x_z) && \textit{(Eq. 46)} && (64)\\ &= P(y_{xz}, x_z) && \textit{(Eq. 50)} && (65)\\ &= P(y_x, x_z) && \textit{(Eq. 47)} && (66)\\ &= P(y_x, x) && \textit{(Eq. 46)} && (67)\\ &= P(y, x) && \textit{(Eq. 49)} && (68) \end{aligned}$$

Example 11 (Local to Global Constraints — Chain)**.** *Consider the SCM $\mathcal{M} = \langle \mathbf{U} = \{U_x, U_z, U_y\}, \mathbf{V} = \{X, Z, Y\}, \mathcal{F}, P(\mathbf{u})\rangle$, where*

$$\mathcal{F} = \begin{cases} X \leftarrow U_x \\ Z \leftarrow X \oplus U_z \\ Y \leftarrow Z \vee U_y \end{cases} \tag{69}$$

$$P(\mathbf{u}) : U_x \sim Bernoulli(0.2),\ U_z \sim Bernoulli(0.4),\ U_y \sim Bernoulli(0.3). \tag{70}$$

The causal diagram $\mathcal{G}$ induced by $\mathcal{M}$ is a causal chain, shown in Fig. 4(b). The CBN2.25 induced by $\mathcal{M}$ includes the constraints listed in Def. 13, following from $\mathcal{G}$, namely:

(i) ***Independence Restrictions.***

$$P(X, Z_x, Y_z) = P(X)\, P(Z_x)\, P(Y_z) \tag{71}$$

(ii) ***Exclusion Restrictions.***

$$\begin{aligned} P(X_{\mathbf{a}} = x, \mathbf{W}_*) &= P(X = x, \mathbf{W}_*), && \mathbf{a} \subseteq \{z, y\} && (72)\\ P(Z_{xy} = z, \mathbf{W}_*) &= P(Z_x = z, \mathbf{W}_*) && && (73)\\ P(Y_{xz} = y, \mathbf{W}_*) &= P(Y_z = y, \mathbf{W}_*) && && (74) \end{aligned}$$

(iii) ***Local Consistency.***

$$\begin{aligned} P(Z = z, X = x, \mathbf{W}_*) &= P(Z_x = z, X = x, \mathbf{W}_*) && (75)\\ P(Z_y = z, X_y = x, \mathbf{W}_*) &= P(Z_{yx} = z, X_y = x, \mathbf{W}_*) && (76)\\ P(Y = y, Z = z, \mathbf{W}_*) &= P(Y_z = y, Z = z, \mathbf{W}_*) && (77)\\ P(Y_x = y, Z_x = z, \mathbf{W}_*) &= P(Y_{zx} = y, Z_x = z, \mathbf{W}_*) && (78) \end{aligned}$$

where $\mathbf{W}_*$ can be any set of counterfactual variables such that $P(\cdot) \in \mathbf{P}^{\mathcal{L}_{2.25}}$.

Now consider whether $P(y_x, x) = P(y, x)$. As in Example 11, this relation cannot be read off directly from the CBN2.25, since it is not a local constraint (here X is not a parent of Y in $\mathcal{G}$). However, it can be derived by composing several local constraints, as follows:

$$\begin{aligned} P(y_x, x) &= \sum_z P(y_x, z, x) && && (79)\\ &= \sum_z P(y_x, z_x, x) && (75) && (80)\\ &= \sum_z P(y_{zx}, z_x, x) && (78) && (81)\\ &= \sum_z P(y_z, z_x, x) && (74) && (82)\\ &= \sum_z P(y_z, z, x) && (75) && (83)\\ &= \sum_z P(y, z, x) && (77) && (84)\\ &= P(y, x). && && (85) \end{aligned}$$

The inferential machinery associated with a graphical model facilitates the process of composing the local constraints defined in the model and determining whether a given query can be expressed as a function of the available data. For $\mathcal{L}_1$ assumptions, the standard machinery for the probabilistic constraints encoded in a BNs is *d-separation* [18]. For CBNs ($\mathcal{L}_2$), Pearl's celebrated *do-calculus* serves this role [20], while for CTFBNs ($\mathcal{L}_3$), the corresponding tool is the *ctf-calculus* [6]. As discussed earlier, the key distinction between CBN2.25/CBN2.5 and CTFBN lies in the distributions to which the local constraints apply. Building on *ctf-calculus*, we develop an inferential machinery for CBN2.25 and CBN2.5 by restricting the rules to distributions in their respective layers.

Definition 15 (Counterfactual Calculus (ctf-calculus) for CBN2.25/CBN2.5)**.** *Let $\mathcal{G}$ be a CBN2.25/CBN2.5 for $\mathbf{P}^{\mathcal{L}_{2.25}}$/$\mathbf{P}^{\mathcal{L}_{2.5}}$, then $\mathbf{P}^{\mathcal{L}_{2.25}}$/$\mathbf{P}^{\mathcal{L}_{2.5}}$ satisfies the Counterfactual-Calculus rules according to $\mathcal{G}$. Namely, for any disjoint sets $\mathbf{X}, \mathbf{Y}, \mathbf{Z}, \mathbf{W}, \mathbf{T}, \mathbf{R} \subseteq \mathbf{V}$ the following three rules hold:*

Rule 1 *(Consistency Rule - Observation/Intervention Exchange)*

$$P(\mathbf{y}_{\mathbf{T}_*\mathbf{x}}, \mathbf{x}_{\mathbf{T}_*}, \mathbf{w}_*) = P(\mathbf{y}_{\mathbf{T}_*}, \mathbf{x}_{\mathbf{T}_*}, \mathbf{w}_*) \tag{86}$$

Rule 2 *(Independence Rule - Adding/Removing Counterfactual Observations)*

$$P(\mathbf{y}_\mathbf{r}|\mathbf{x}_\mathbf{t}, \mathbf{w}_*) = P(\mathbf{y}_\mathbf{r}|\mathbf{w}_*) \text{ if } (\mathbf{Y}_\mathbf{r} \perp\!\!\!\perp \mathbf{X}_\mathbf{t}|\mathbf{W}_*) \text{ in } \mathcal{G}_A \tag{87}$$

Rule 3 *(Exclusion Rule - Adding/Removing Interventions)*

$$P(\mathbf{y}_{\mathbf{xz}}, \mathbf{w}_*) = P(\mathbf{y}_\mathbf{z}, \mathbf{w}_*) \text{ if } (\mathbf{X} \cap An(\mathbf{Y}) = \emptyset) \text{ in } \mathcal{G}_{\overline{\mathbf{Z}}} \tag{88}$$

where $\mathcal{G}_A$ is the AMWN $\mathcal{G}_A(\mathcal{G}, \mathbf{Y}_\mathbf{r} \cup \mathbf{X}_\mathbf{t} \cup \mathbf{W}_*)$ [5], and all $P(\cdot)$ in the rules belong to $\mathbf{P}^{\mathcal{L}_{2.25}}$/$\mathbf{P}^{\mathcal{L}_{2.5}}$.

The three rules of the calculus can be viewed as global counterparts to the three conditions in the definitions of CBN2.25 and CBN2.5. Whereas the graphical model definitions restrict local constraints to counterfactual variables whose parents appear in the subscript, the calculus rules capture more general counterfactual forms. For instance, the independence rule applies not only to counterfactuals of the form $W_{\mathbf{pa}_w}$, but also to broader classes of counterfactual variables.

It is important to note that the ctf-calculus rules must, for each model, be restricted so that all distributions involved belong to the appropriate layer. For example, the calculus for CBN2.25 is limited to distributions in $\mathbf{P}^{\mathcal{L}_{2.25}}$, while for CBN2.5 it is limited to $\mathbf{P}^{\mathcal{L}_{2.5}}$. To enforce this, we introduce a graphical check to verify that all $P(\cdot)$ appearing in the rules correspond to valid distributions in the given model. In $\mathcal{L}_{2.5}$, this criterion checks only the counterfactual ancestor set, whereas in $\mathcal{L}_{2.25}$ it must also examine the descendants of ancestors, since the more restrictive CTF-RAND imposes stronger consistency requirements across all downstream variables sharing the same intervened parents.

Definition 16 (Counterfactual Reachability Set)**.** *Given a graph $\mathcal{G}$ and a potential outcome $Y_\mathbf{x}$, the counterfactual reachability set of $Y_\mathbf{x}$, denoted $CRS(Y_\mathbf{x})$, consists of:*

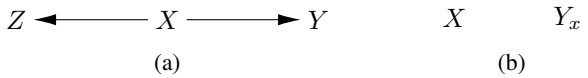
- $\|W_\mathbf{x}\|$ *such that* $W \in (An(Y) \cup \{De(V) : \forall V \in \mathbf{X}\}) \setminus \mathbf{X}$*, and*
- $\|W_{\mathbf{x}\setminus w}\|$ *such that* $W \in (An(Y) \cup \{De(V) : \forall V \in \mathbf{X}\}) \cap \mathbf{X}$.

For a set $\mathbf{W}_*$, $CRS(\mathbf{W}_*)$ is defined as the union of the CRS of each potential outcome in the set, with the following merging rule: if $\{W_{i_{[\mathbf{x}_i]}}\}_i \subseteq \mathbf{W}_*$ have CRS sets containing counterfactual variables $\{R_{[\mathbf{x}_i]}\}_i$ over the same variable R, then $\{R_{[\mathbf{x}_i]}\}_i$ are merged into a single variable $\|R_{[\cup_i \mathbf{x}_i]}\|$ whenever $\|W_{i_{[\cup_i \mathbf{x}_i]}}\| = W_{i_{[\mathbf{x}_i]}}$ for all i.

Lemma 2. *A distribution $Q = P(\mathbf{W}_*)$ induced by any SCM compatible with a given graph $\mathcal{G}$ belongs to:*

1. *$\mathcal{L}_{2.25}$. Q is an $\mathcal{L}_{2.25}$ distribution if and only if $CRS(\mathbf{W}_*)$ satisfies:*
   - *(i) it does not contain any pair of potential outcomes $W_\mathbf{s}, W_\mathbf{t}$ of the same variable W under different regimes ($\mathbf{s} \neq \mathbf{t}$); and*
   - *(ii) it does not contain any pair of potential outcomes $R_\mathbf{s}, W_\mathbf{t}$ with inconsistent subscripts, i.e., $\mathbf{s} \cap \mathbf{T} \neq \mathbf{t} \cap \mathbf{S}$.*

[5] Definition and algorithm for Ancestral Multi-World Network (AMWN) is given in Appendix. B



(a) (b)

Figure 9: Sample causal diagram $\mathcal{G}$ and its AMWN $\mathcal{G}_A(\mathcal{G}, \{Y_x, X\})$

2. $\mathcal{L}_{2.5}$. *Q is an $\mathcal{L}_{2.5}$ distribution if and only if $An(\mathbf{W}_*)$ satisfies:*

   (i) it does not contain any pair of potential outcomes $W_{\mathbf{s}}, W_{\mathbf{t}}$ of the same variable W under different regimes ($\mathbf{s} \neq \mathbf{t}$).

Example 12 (CRS Check - Not in $\mathcal{L}_{2.25}$). *Consider the causal diagram in Fig. 4(a) and whether $P(Z_x, Y_{x'})$ belongs to layer 2.25 induced by the corresponding SCM. The reachability set is $CRS(Z_x, Y_{x'}) = \{X, Z_x, Y_x, Z_{x'}, Y_{x'}\}$. Since the joint counterfactual $\{Z_x, Z_{x'}\}$ appears in this set with Z under different regimes, Lemma 2 implies that $P(Z_x, Y_{x'})$ does not belong to the $\mathcal{L}_{2.25}$ distributions.* ■

Example 13 (CRS Check - In $\mathcal{L}_{2.5}$). *Now consider the same query $P(Z_x, Y_{x'})$ but under $\mathcal{L}_{2.5}$. In this case, the conditions in Lemma 2 for $\mathcal{L}_{2.5}$ only forbid conflicting potential outcomes of the* same *variable under different regimes. Here, Z_x and $Y_{x'}$ involve different variables (Z and Y), so no violation occurs. Therefore, $P(Z_x, Y_{x'})$ is admissible in $\mathcal{L}_{2.5}$, illustrating how $\mathcal{L}_{2.5}$ permits richer distributions than $\mathcal{L}_{2.25}$.*

With Lemma 2 ensuring that the relevant distributions lie within the corresponding layers, we can apply the ctf-calculus in CBN2.25 and CBN2.5.

Theorem 3 (Soundness and Completeness for CBN2.25/CBN2.5 Identifiability). *An $\mathcal{L}_{2.25}$ or $\mathcal{L}_{2.5}$ quantity Q is identifiable from a given set of observational and interventional distributions and a causal diagram $\mathcal{G}$ if and only if there exists a sequence of applications of the rules of the ctf-calculus for CBN2.25/CBN2.5, together with the probability axioms restricted to $\mathcal{L}_{2.25}$/$\mathcal{L}_{2.5}$, that reduces Q to a function of the available distributions.*

Example 14 (Effect of the Treatment on the Treated). *Consider the causal diagram $\mathcal{G}$ in Fig. 9(a) and the effect of treatment on the treated (ETT), defined as $Q = P(y_x \mid x')$, with observational distribution $P(\mathbf{v})$ as input. Q can be derived using the ctf-calculus rules as follows:*

$$
\begin{aligned}
P(y_x|x') &= P(y_x) && (\textit{Rule 2: } Y_x \perp\!\!\!\perp X \textit{ in } \mathcal{G}_A(\mathcal{G}, \{Y_x, X\}) \textit{ Fig. 9(b)}) && (89)\\
&= P(y_x|x) && (\textit{Rule 2: } Y_x \perp\!\!\!\perp X \textit{ in } \mathcal{G}_A(\mathcal{G}, \{Y_x, X\}) \textit{ Fig. 9(b)}) && (90)\\
&= P(y|x) && (\textit{Rule 1: Consistency}). && (91)
\end{aligned}
$$

Steps Eq. (103) and (104) follow from Lemma 2, since $CRS(X, Y_x) = \{X, Z_x, Y_x\}$ is in $\mathcal{L}_{2.25}$. ■

4 Hierarchy of Graphical Models

In this section, we develop a refined view of the PCH by incorporating the two new graphical models that allow counterfactual inference between layer 2 and the full layer 3. We then illustrate how models in the hierarchy differ in the types of queries they support and in the falsifiability of the assumptions they encode.

First, note that the two collections of distributions defined earlier (Def. 11 and Def. 12) can be positioned naturally within the PCH.

Theorem 4 (PCH*, or Augmented PCH). *Given an SCM $\mathcal{M}$ and its induced collections of observational ($\mathbf{P}^{\mathcal{L}_1}$), interventional ($\mathbf{P}^{\mathcal{L}_2}$), $\mathcal{L}_{2.25}$ ($\mathbf{P}^{\mathcal{L}_{2.25}}$), $\mathcal{L}_{2.5}$ ($\mathbf{P}^{\mathcal{L}_{2.5}}$), and counterfactual ($\mathbf{P}^{\mathcal{L}_3}$) distributions:*

$$
\mathbf{P}^{\mathcal{L}_1} \subseteq \mathbf{P}^{\mathcal{L}_2} \subseteq \mathbf{P}^{\mathcal{L}_{2.25}} \subseteq \mathbf{P}^{\mathcal{L}_{2.5}} \subseteq \mathbf{P}^{\mathcal{L}_3}. \tag{92}
$$

The illustration in Fig. 10 provides a global view of the components involved in the analysis. The SCM $\mathcal{M}^*$ sits at the top of the generative process and induces both the PCH distributions on the left (Def. 4) and the causal diagram on the right (Def. 6). This augments the original PCH (Fig. 6) by explicitly incorporating the intermediate layers $\mathbf{P}^{\mathcal{L}_{2.25}}$ and $\mathbf{P}^{\mathcal{L}_{2.5}}$. The connection between each collection of distributions (left) and its associated graphical model (right) defines the corresponding

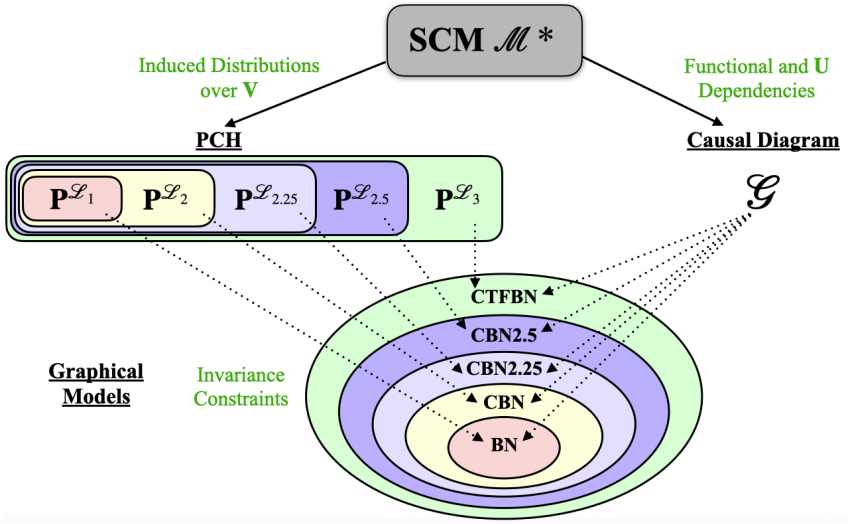


Figure 10: Pearl Causal Hierarchy (PCH*) and hierarchy of graphical models induced by an SCM

compatibility relations. These relations in turn define increasingly fine-grained equivalence classes of SCMs within the space Ω. As we move upward in the hierarchy, the missing edges in the graphical models encode progressively stronger invariance constraints, supporting richer queries but also imposing stronger assumptions that are correspondingly harder to falsify.

Building on this hierarchy of distributions, we turn to the constraints encoded by each graphical model. Given a causal diagram $\mathcal{G}$, the constraints it encodes arise from the interpretation of its missing edges. As we move higher in the hierarchy of graphical models, the missing edges correspond to increasingly stronger constraints on the distributions, and, by implication, to a finer partitioning of the space of SCMs, Ω. This progression is illustrated in the example below.

Example 15 (Constraints from Missing Edges)**.** *Consider the causal diagram in Fig. 4(a). The constraints encoded by the* missing directed edge *from Z to Y across different layers are ($\mathcal{P}(\cdot)$ denotes the power set):*

$$\textit{BN:} \quad P(Y \mid X, Z) = P(Y \mid X) \tag{93}$$

$$\textit{CBN:} \quad P(Y_{xz}) = P(Y_x) \tag{94}$$

$$\textit{CBN2.25:} \quad P(Y_{xz}, \mathbf{W}_*) = P(Y_x, \mathbf{W}_*),\ \forall \mathbf{W}_* \in \mathcal{P}(\{X, Z_x\}) \tag{95}$$

$$\textit{CBN2.5:} \quad P(Y_{xz}, \mathbf{W}_*) = P(Y_x, \mathbf{W}_*),\ \forall \mathbf{W}_* \in \mathcal{P}(\{X, Z_x\}) \cup \mathcal{P}(\{X, Z_{x'}\}) \tag{96}$$

$$\textit{CTFBN:} \quad P(Y_{xz}, \mathbf{W}_*) = P(Y_x, \mathbf{W}_*),\ \forall \mathbf{W}_* \tag{97}$$

Moving from BN to CBN augments the model with $\mathcal{L}_2$ constraints, and moving further to CBN2.25, CBN2.5, and CTFBN introduces $\mathcal{L}_3$ constraints. Among the counterfactual models, higher layers allow increasingly flexible forms of $\mathbf{W}_*$, corresponding to stronger assumptions.

The missing bidirected edge *encodes independence constraints at different layers:*

$$\textit{CBN:} \quad P(Z_x) = P(Z \mid X = x), \quad P(Y_x) = P(Y \mid X = x) \tag{98}$$

$$\textit{CBN2.25:} \quad P(Z_x, Y_x, X) = P(Z_x)\, P(Y_x)\, P(X) \tag{99}$$

$$\textit{CBN2.5:} \quad P(Z_x, Y_{x'}, X) = P(Z_x)\, P(Y_{x'})\, P(X) \tag{100}$$

$$\textit{CTFBN:} \quad P\left(\bigwedge_{x \in Val(X)} Z_x, \bigwedge_{x' \in Val(X)} Y_{x'},\ X \right) \tag{101}$$

$$= P\left(\bigwedge_{x \in Val(X)} Z_x \right) P\left(\bigwedge_{x' \in Val(X)} Y_{x'} \right) P(X) \tag{102}$$

As we move up the hierarchy, independence constraints involve richer sets of counterfactual variables, reflecting the stronger assumptions imposed. CBN encodes the parent do/see constraints restricted

to $\mathbf{P}^{\mathcal{L}_2}$. In contrast, CBN2.25 introduces counterfactual constraints beyond $\mathcal{L}_2$, and together with consistency conditions, these imply the parent do/see restrictions of CBN. CBN2.5 uses the same constraint forms as CBN2.25 but permits more flexible subscripts in joint counterfactuals. At the top, CTFBN allows the most expressive independence constraints, spanning broad joint counterfactual distributions and implying those in CBN2.5 via marginalization. ■

In fact, the constraints encoded by graphical models higher in the hierarchy always subsume those of the models lower in the hierarchy. This monotonicity property defines the hierarchy of graphical models, as illustrated in Fig. 10.

Theorem 5 (Hierarchy of Graphical Models, PCH*)**.** *Given a causal diagram $\mathcal{G}$, the set of constraints it encodes when interpreted as a graphical model at layer i is always a subset of the constraints it encodes when interpreted at a higher layer j, for all $i \leq j$.*

As discussed earlier in the context of Fig. 1, the causal inference engine operates by matching a query with a model that encodes a sufficient — and ideally only necessary — set of assumptions. This matching logic can be understood from two complementary perspectives.

The first concerns the *expressive power* of a model: the extent to which its assumptions are sufficient to support valid inference for a given query. When the expressiveness of the query exceeds that of the model's assumptions, the causal inference engine lacks the necessary ingredients to proceed. For example, a BN encodes only $\mathcal{L}_1$ constraints and is therefore blind to $\mathcal{L}_2$ structure, making it unable to evaluate queries such as $P(y \mid do(x))$. Similarly, a CBN, which encodes only $\mathcal{L}_2$ constraints, cannot support inference for $\mathcal{L}_3$ queries like $P(Y_x, X)$. In contrast, when a model's assumptions are expressive enough to support the query, we say that the query and the model are *matched*. A CTFBN, which sits at the top of the hierarchy in terms of expressive power, can in principle match the most demanding queries in the PCH.[6]

The second perspective concerns the *empirical falsifiability* of a model: whether the assumptions it encodes are not only sufficient but also necessary for the query at hand. As one ascends the hierarchy, models impose increasingly stronger counterfactual constraints, many of which cannot be directly falsified with the given data collections, whether observational ($\mathbf{P}^{\mathcal{L}_1}$), interventional ($\mathbf{P}^{\mathcal{L}_2}$), or those realizable via counterfactual randomization ($\mathbf{P}^{\mathcal{L}_{2.25}}$, $\mathbf{P}^{\mathcal{L}_{2.5}}$). If a model encodes assumptions beyond what is strictly required, these may become empirically untestable and ontologically burdensome. Accordingly, the preferred model for a given query is the most parsimonious one that still supports valid inference, striking a balance between expressive power and empirical falsifiability.

Example 16 (Effect of the Treatment on the Treated — Expressive Power)**.** *Consider the causal diagram $\mathcal{G}$ in Fig. 9(a) and the effect of treatment on the treated (ETT), defined as $Q = P(y_x \mid x')$, with observational distribution $P(\mathbf{v})$ as input.*

A BN encodes only $\mathcal{L}_1$ constraints and is therefore unable to represent Q, which requires $\mathcal{L}_{2.25}$ structure. By contrast, CBN2.25 has sufficient expressive power: the query can be derived using the ctf-calculus rules as follows:

$$P(y_x \mid x') = P(y_x) \qquad \textit{(Rule 2: } Y_x \perp\!\!\!\perp X \textit{ in } \mathcal{G}_A(\mathcal{G}, \{Y_x, X\})\textit{, Fig. 9(b))} \tag{103}$$

$$= P(y_x \mid x) \qquad \textit{(Rule 2: } Y_x \perp\!\!\!\perp X \textit{ in } \mathcal{G}_A(\mathcal{G}, \{Y_x, X\})\textit{, Fig. 9(b))} \tag{104}$$

$$= P(y \mid x) \qquad \textit{(Rule 1: Consistency).} \tag{105}$$

Steps (103) *and* (104) *follow from Lemma 2, since $CRS(X, Y_x) = \{X, Z_x, Y_x\}$ lies in $\mathcal{L}_{2.25}$. This example illustrates that CBN2.25 provides just enough expressive power to support inference about Q, whereas weaker models such as BNs or CBNs cannot.*

Example 17 (Natural Direct Effect — Empirical Falsifiability)**.** *Consider $\mathcal{G}$ in Fig. 11. The natural direct effect from X to Y is defined as*

$$NDE_{x,x'}(y) = P(y_{x', Z_x}) - P(y_x). \tag{106}$$

Applying unnesting, the first term becomes

$$P(y_{x', Z_x}) = \sum_z P(y_{x'z}, z_x), \tag{107}$$

which is an $\mathcal{L}_{2.5}$ query.

[6] This does not immediately imply that the query is identifiable, only that it can be represented in principle.

Query Layer	Graphical Model	Sufficient	Necessary
$\mathcal{L}_1$	BN	✓	✓
$\mathcal{L}_1$	CBN	✓	x
$\mathcal{L}_2$	BN	x	✓
$\mathcal{L}_2$	CBN	✓	✓
$\mathcal{L}_2$	CBN2.25	✓	x
$\mathcal{L}_{2.25}$	CBN	x	✓
$\mathcal{L}_{2.25}$	CBN2.25	✓	✓
$\mathcal{L}_{2.25}$	CBN2.5	✓	x
$\mathcal{L}_{2.5}$	CBN2.25	x	✓
$\mathcal{L}_{2.5}$	CBN2.5	✓	✓
$\mathcal{L}_{2.5}$	CTFBN	✓	x
$\mathcal{L}_3$	CBN2.5	x	✓
$\mathcal{L}_3$	CTFBN	✓	✓

Table 3: Examples of Matching between Graphical Models and Queries. Rows highlighted in green represent a match between the model and the query such that the assumptions in the model are both sufficient and necessary for making inference about the query.

This query $Q = P(y_{x'z}, z_x)$ can be identified in the CBN2.5 associated with $\mathcal{G}$ via ctf-calculus:

$$P(y_{x'z}, z_x) = P(y \mid x', z)P(z \mid x). \tag{108}$$

A CTFBN, which encodes stronger constraints than CBN2.5, can also identify Q, but it brings in unnecessary assumptions such as $P(Z_x, Z_{x'}, X) = P(Z_x, Z_{x'})P(X)$, which cannot be empirically falsified with current data collection methods. In contrast, a CBN is not expressive enough to represent Q at all.

This example highlights the parsimony dimension: although both CBN2.5 and CTFBN can in principle support the query, the more parsimonious CBN2.5 is preferable since it avoids unnecessary, unfalsifiable commitments.

This example highlights the trade-off between the expressiveness of queries and the parsimony of models. The optimal match occurs when the assumptions in the model are both sufficient and necessary for the intended inference (illustrated as green rows in Table 3). The introduction of CBN2.25 and CBN2.5 refines the necessity boundaries in $\mathcal{L}_3$, ensuring that queries in $\mathcal{L}_{2.25}$ and $\mathcal{L}_{2.5}$ can be matched with more parsimonious and empirically testable models. In short, higher models in the hierarchy gain inferential power by encoding constraints over increasingly expressive distributions, but at the expense of falsifiability. It is therefore crucial for researchers to choose models that strike an appropriate balance between expressive power and empirical testability for the task at hand.

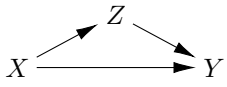


Figure 11: Causal diagram for NDE

5 Conclusions

In this paper, we introduced two new classes of graphical models, CBN2.25 and CBN2.5, which encode constraints over distinct collections of distributions realizable under counterfactual randomization. We showed that these models are naturally induced by SCMs (Thm. 1) and established a sound and complete inferential machinery for them (Thm. 3). We then placed the new distribution classes within the PCH (Thm. 4) and proved that graphical models over the PCH form a hierarchy (Thm. 5). Finally, we highlighted the trade-off between expressive power and empirical falsifiability across the hierarchy.

Taken together, these results refine the landscape of graphical models and provide a more nuanced map of the space between $\mathcal{L}_2$ and $\mathcal{L}_3$. They also offer guidance for researchers in selecting models that balance inferential capability with empirical testability, depending on the goals and constraints of a given application.

Acknowledgments

This research is supported in part by the NSF, ONR, AFOSR, DoE, Amazon, JP Morgan, and The Alfred P. Sloan Foundation. We thank Arvind Raghavan for their thoughtful comments.

References

[1] Elias Bareinboim. *Causal Artificial Intelligence: A Roadmap for Building Causally Intelligent Systems*. https://causalai-book.net/, 2025.

[2] Elias Bareinboim, Carlos Brito, and Judea Pearl. Local Characterizations of Causal Bayesian Networks. In *Lecture Notes in Artificial Intelligence*. Springer, 2012.

[3] Elias Bareinboim, Juan D Correa, Duligur Ibeling, and Thomas Icard. On Pearl's Hierarchy and the Foundations of Causal Inference. In *Probabilistic and Causal Inference: The Works of Judea Pearl (ACM, Special Turing Series)*, 2022.

[4] Elias Bareinboim, Andrew Forney, and Judea Pearl. Bandits with Unobserved Confounders: A Causal Approach. In *Proceedings of the 28th Annual Conference on Neural Information Processing Systems*, 2015.

[5] Nancy Cartwright. *Nature's Capacities and Their Measurement*. Clarendon Press, Oxford, 1989.

[6] Juan D Correa and Elias Bareinboim. Counterfactual Graphical Models: Constraints and Inference. In *Proceedings of the 42nd International Conference on Machine Learning*, 2025.

[7] Juan D Correa, Sanghack Lee, and Elias Bareinboim. Nested Counterfactual Identification from Arbitrary Surrogate Experiments. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

[8] A. P. Dawid. Influence Diagrams for Causal Modelling and Inference. *International Statistical Review / Revue Internationale de Statistique*, 70(2):161–189, 2002. Publisher: [Wiley, International Statistical Institute (ISI)].

[9] A. Philip Dawid. Beware of the DAG! In *Proceedings of Workshop on Causality: Objectives and Assessment at NIPS 2008*, pages 59–86. PMLR, February 2010. ISSN: 1938-7228.

[10] Dawid, A. Philip. What Is a Causal Graph?, January 2024. arXiv:2402.09429.

[11] Ronald Fisher. *The Design of Experiments*. Oliver and Boyd, 1935.

[12] S Geneletti. Identifying direct and indirect effects in a non-counterfactual framework. *Journal of the Royal Statistical Society, Series B (Methodological)*, 2007.

[13] Yimin Huang and Marco Valtorta. On the completeness of an identifiability algorithm for semi-Markovian models. *Annals of Mathematics and Artificial Intelligence*, 54(4):363–408, 2008.

[14] Amin Jaber, Murat Kocaoglu, Karthikeyan Shanmugam, and Elias Bareinboim. Causal discovery from soft interventions with unknown targets: Characterization and learning. *Advances in neural information processing systems*, 33:9551–9561, 2020.

[15] Murat Kocaoglu, Amin Jaber, Karthikeyan Shanmugam, and Elias Bareinboim. Characterization and learning of causal graphs with latent variables from soft interventions. *Advances in Neural Information Processing Systems*, 32, 2019.

[16] Murat Kocaoglu, Karthikeyan Shanmugam, and Elias Bareinboim. Experimental Design for Learning Causal Graphs with Latent Variables. In *Advances in Neural Information Processing Systems 30*, 2017.

[17] Adam Li, Amin Jaber, and Elias Bareinboim. Causal discovery from observational and interventional data across multiple environments. *Advances in Neural Information Processing Systems*, 36:16942–16956, 2023.

[18] Judea Pearl. *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann Publishers, 1988.

[19] Judea Pearl. Aspects of graphical models connected with causality. In *Proceedings of the 49th Session of the International Statistical Institute*, 1993.

[20] Judea Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, second edition, 2009.

[21] Judea Pearl and Dana Mackenzie. *The Book of Why*. Basic Books, 2018.

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

[23] Karl Popper. *The Logic of Scientific Discovery*. Routledge, 2002.

[24] Arvind Raghavan and Elias Bareinboim. Counterfactual Realizability. In *Proceedings of the 13rd International Conference on Learning Representations*, 2025.

[25] Thomas Richardson and James Robins. Single World Intervention Graphs (SWIGs): A Unification of the Counterfactual and Graphical Approaches to Causality. 2013.

[26] James Robins. A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Mathematical Modelling*, 7(9-12):1393–1512, 1986.

[27] L. Smolin. *The Trouble with Physics: The Rise of String Theory, the Fall of a Science, and what Comes Next*. A Mariner Book. Houghton Mifflin, 2006.

[28] P Spirtes, C N Glymour, and R Scheines. *Causation, Prediction, and Search*. MIT Press, Cambridge, MA, 2nd edition, 2000.

[29] Jin Tian and Judea Pearl. A General identification condition for causal effects. In *Proceedings of the 18th AAAI Conference on Artificial Intelligence*, 2002.

[30] Julius von Kügelgen, Michel Besserve, Liang Wendong, Luigi Gresele, Armin Kekić, Elias Bareinboim, David Blei, and Bernhard Schölkopf. Nonparametric identifiability of causal representations from unknown interventions. *Advances in Neural Information Processing Systems*, 36, 2024.

[31] Mengyuan Xiao, Pascal A. Oesch, David Elbaz, Longji Bing, Erica J. Nelson, Andrea Weibel, Garth D. Illingworth, Pieter van Dokkum, Rohan P. Naidu, Emanuele Daddi, Rychard J. Bouwens, Jorryt Matthee, Stijn Wuyts, John Chisholm, Gabriel Brammer, Mark Dickinson, Benjamin Magnelli, Lucas Leroy, Daniel Schaerer, Thomas Herard-Demanche, Seunghwan Lim, Laia Barrufet, Ryan Endsley, Yoshinobu Fudamoto, Carlos Gómez-Guijarro, Rashmi Gottumukkala, Ivo Labbé, Dan Magee, Danilo Marchesini, Michael Maseda, Yuxiang Qin, Naveen A. Reddy, Alice Shapley, Irene Shivaei, Marko Shuntov, Mauro Stefanon, Katherine E. Whitaker, and J. Stuart B. Wyithe. Accelerated formation of ultra-massive galaxies in the first billion years. *Nature*, 635(8038):311–315, November 2024. Publisher: Nature Publishing Group.

[32] Jiji Zhang. Causal Reasoning with Ancestral Graphs. *Journal of Machine Learning Research*, 9(47):1437–1474, 2008.

Appendices

Contents

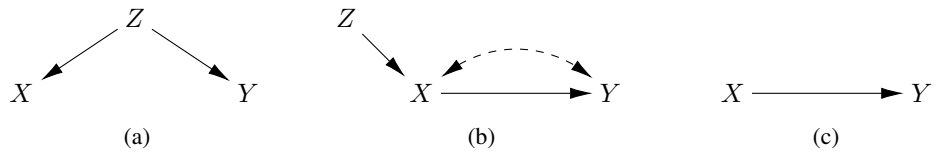


Figure 12: SCM Induced DAG or Causal Diagrams

A Background and Definitions

In the following sections, we give the definitions and examples for graphical models introduced in previous works ([18, 20, 3, 1]).

A.1 $\mathcal{L}_1$: Bayesian Networks

The first graphical model encodes invariance constraints in the observational distributions. Firstly, we formally define how to construct a graph from an SCM.

Definition 17 (Confounded Component of an SCM [3]). *Given an SCM $\mathcal{M} = \langle \mathbf{U}, \mathbf{V}, \mathcal{F}, P(\mathbf{u}) \rangle$, let $\mathbf{U}_1^c, \mathbf{U}_1^c, ..., \mathbf{U}_l^c \subseteq \mathbf{U}$ be disjoint maximal subsets of the exogenous variables in $\mathcal{M}$ such that $P(\mathbf{u}) = \prod_{k=1}^{l} P(\mathbf{U}_k^c)$. Then, we say that $V_i, V_j \in \mathbf{V}$ are in the same confounded component (for short, C-component) of $\mathcal{M}$ if $|\{\mathbf{U}_k^c | \mathbf{U}_k^c \cap \mathbf{U}_i \neq \emptyset, \mathbf{U}_k^c \cap \mathbf{U}_j \neq \emptyset\}| > 0$, that is, if f_i and f_j have both latent arguments in some common $\mathbf{U}_k^c$.*

Definition 18 (SCM-induced DAG [3]). *Consider an SCM $\mathcal{M} = \langle \mathbf{U}, \mathbf{V}, \mathcal{F}, P(\mathbf{u}) \rangle$. Then $\mathcal{G}$ is a DAG induced by $\mathcal{M}$ if it:*

- *has a vertex for every endogenous variable in the set $\mathbf{V}$*
- *has an edge $V_i \longrightarrow V_j$ for every $V_i, V_j \in \mathbf{V}$ if V_i appears as an argument of f_j*
- *there exists an order over the functions in $\mathcal{F}$ such that for every pair V_i, V_j in the same C-component of $\mathcal{M}$ such that $f_i < f_j$, the edge $V_i \longrightarrow V_j$ and the edges $V_k \longrightarrow V_j, V_k \in \mathbf{Pa}_i$ are in $\mathcal{G}$.*

Definition 19 (Markov Relative to [18]). *A probability distribution $P(\mathbf{V})$ over a set of observed variables $\mathbf{V}$ is said to be Markov relative to a graph $\mathcal{G}$ if:*

$$P(\mathbf{V}) = \prod_i P(v_i | \mathbf{pa}_i) \tag{109}$$

where $\mathbf{Pa}_i = \{V_j \in \mathbf{V} | (V_j \to V_i) \in \mathcal{G}\}$.

Definition 20 (Bayesian Network [18]). *A directed acyclic graph (DAG) $\mathcal{G}$ is a Bayesian Network for a probability distribution P over the variables in $\mathbf{V}$ if P is Markov relative to $\mathcal{G}$.*

Example 18 (SCM-induced BN). *Consider the SCM $\mathcal{M} = \langle \mathbf{U} = \{U_z, U_x, U_y\}, \mathbf{V} = \{Z, X, Y\}, \mathcal{F}, P(\mathbf{u}) \rangle$ where*

$$\mathcal{F} = \begin{cases} Z \leftarrow U_z \\ X \leftarrow Z \vee U_x \\ Y \leftarrow Z \oplus U_y \end{cases} \tag{110}$$

$$P(\mathbf{u}) : U_z \sim Bernoulli(0.5), U_x \sim Bernoulli(0.5), U_y \sim Bernoulli(0.5) \tag{111}$$

Its SCM-induced DAG is shown in Fig. 12(a) and its induced observational distribution $P(\mathbf{v})$ satisfies:

$$P(\mathbf{v}) = P(z)P(x|z)P(y|z) \tag{112}$$

for all x, y, z in $Val(X) \times Val(Y) \times Val(Z)$. The DAG in Fig. 12(a) is a BN for $P(\mathbf{v})$.

A.2 $\mathcal{L}_2$: Causal Bayesian Networks

The second graphical model encodes invariance constraints in the interventional distributions.

Definition 21 (CBN Markovian [3]). *Let $\mathbf{P}_*$ be the collection of all interventional distributions $P(\mathbf{V}|do(\mathbf{x})), \mathbf{X} \subseteq \mathbf{V}, \mathbf{x} \in Val(\mathbf{X})$, including the null intervention, $P(\mathbf{V})$, where $\mathbf{V}$ is the set of observed variables. A directed acyclic graph $\mathcal{G}$ is called a Causal Bayesian Network for $\mathbf{P}_*$ if:*

1. [*Markov*] $P(\mathbf{V})|do(\mathbf{x})$ *is Markov relative to* $\mathcal{G}$*;*
2. [*Missing-link*] *For every* $V_i \in \mathbf{V}, V_i \notin \mathbf{X}$ *such that there is no arrow from* $\mathbf{X}$ *to* V_i *in* $\mathcal{G}$*:*
$$P(v_i|do(pa_i), do(\mathbf{x})) = P(v_i|do(pa_i)) \tag{113}$$
3. [*Parents do/see*] *For every* $V_i \in \mathbf{V}, V_i \notin \mathbf{X}$*:*
$$P(v_i|do(pa_i), do(\mathbf{x})) = P(v_i|pa_i, do(\mathbf{x})) \tag{114}$$

Example 19 (SCM-induced CBN Markovian). *Consider the SCM from Example 18. Its induced causal diagram is shown in Fig. 12(a) and its induced set of interventional distributions $\mathbf{P}_*$ satisfy:*

1. [*Markov*]
$$P(\mathbf{v}) = P(z)P(x|z)P(y|z) \tag{115}$$
$$P(\mathbf{v}|do(x)) = P(z|do(x))P(y|z, do(x)) \tag{116}$$
$$P(\mathbf{v}|do(y)) = P(z|do(y))P(x|z, do(y)) \tag{117}$$
$$P(\mathbf{v}|do(z)) = P(x|do(z))P(y|do(z)) \tag{118}$$
2. [*Missing-link*]
$$P(x|do(y, z)) = P(x|do(z)) \tag{119}$$
$$P(y|do(x, z)) = P(y|do(z)) \tag{120}$$
$$P(z|do(\mathbf{a})) = P(z), \forall \mathbf{a} \subseteq \{x, y\} \tag{121}$$
3. [*Parents do/see*]
$$P(x|do(z)) = P(x|z) \tag{122}$$
$$P(x|do(y, z)) = P(x|z, do(y)) \tag{123}$$
$$P(y|do(z)) = P(y|z) \tag{124}$$
$$P(y|do(x, z)) = P(y|z, do(x)) \tag{125}$$

The causal diagram in Fig. 12(a) is a CBN Markovian for $\mathbf{P}_*$.

Similar to the confounded components in an SCM (Def. 17), there is also a corresponding set of confounded components in the causal diagram induced.

Definition 22 (Confounded Component [29]). *Let $\mathbf{C}_1, \mathbf{C}_2, ...\mathbf{C}_k$ be a partition over the set of variables $\mathbf{V}$, where $\mathbf{C}_i$ is said to be a confounded component (for short, C-component) of $\mathcal{G}$ if for every $V_i, V_j \in \mathbf{C}_i$ there exists a path made entirely of bidirected edges between V_i and V_j in $\mathcal{G}$ and $\mathbf{C}_i$ is maximal.*

Definition 23 (Augmented Parents). *Let $<$ be a topological order over the variables $V_1, ..., V_n$ in $\mathcal{G}$, let $\mathcal{G}(V_i)$ be the subgraph of $\mathcal{G}$ consists only of variables in $V_1, ..., V_i$, and let $\mathbf{C}(V_i)$ be the C-component of V_i in $\mathcal{G}(V_i)$. The augmented parents of V_i, denoted as Pa_i^+, is the union of parents of all variables in $\mathbf{C}(V_i)$ that comes before V_i in topological order:*
$$Pa_i^+ = \cup_{j|V_j \in \mathbf{T_i}} Pa_j \backslash \{V_i\} \tag{126}$$
where $\mathbf{T}_i = \{X \in \mathbf{C}(V_i) : X \leq V_i\}$.

We use $\mathcal{G}_{\overline{\mathbf{X}}}$ to denote the mutilated graph with all incoming edges to $\mathbf{X}$ removed from $\mathcal{G}$. The augmented parent of V_i in $\mathcal{G}_{\overline{\mathbf{X}}}$ is denoted $Pa_i^{\mathbf{x}+}$.

Example 20 (Augmented Parents). *Consider the SCM $\mathcal{M} = \langle \mathbf{U} = \{U_z, U\}, \mathbf{V} = \{Z, X, Y\}, \mathcal{F}, P(\mathbf{u})\rangle$ where*
$$\mathcal{F} = \begin{cases} Z \leftarrow U_z \\ X \leftarrow Z \vee U \\ Y \leftarrow X \oplus U \end{cases} \tag{127}$$
$$P(\mathbf{u}) : U_z \sim Bernoulli(0.5), U \sim Bernoulli(0.5) \tag{128}$$

The causal diagram $\mathcal{G}$ it induces is shown in Fig. 12(b). The respective augmented parents of X, Y, Z in $\mathcal{G}$ are:

$$Pa_z^+ = \{\} \tag{129}$$
$$Pa_x^+ = \{Z\} \tag{130}$$
$$Pa_y^+ = \{X, Z\} \tag{131}$$

If we consider the induced subgraph $\mathcal{G}(Y, Z)$ where there are no edges at all, it is the same graph as $\mathcal{G}_{\overline{X}}$. In this graph, nodes Y and Z form their own c-components respectively, so their augmented parents are both empty:

$$Pa_z^{x+} = \{\} \tag{132}$$
$$Pa_y^{x+} = \{\} \tag{133}$$

Definition 24 (Semi-Markov Relative to [3])**.** *A probability $P(\mathbf{V})$ is said to be semi-Markov relative to a graph $\mathcal{G}$ if for any topological order $<$ of $\mathcal{G}$:*

$$P(\mathbf{V}) = \prod_i P(v_i | pa_i^+) \tag{134}$$

Definition 25 (CBN Semi-Markovian [3])**.** *Let $\mathbf{P}_*$ be the collection of all interventional distributions $P(\mathbf{V}|do(\mathbf{x})), \mathbf{X} \subseteq \mathbf{V}, \mathbf{x} \in Val(\mathbf{X})$, including the null intervention, $P(\mathbf{V})$, where $\mathbf{V}$ is the set of observed variables. A directed acyclic graph $\mathcal{G}$ is called a Causal Bayesian Network for $\mathbf{P}_*$ if, considering $Pa_i^{\mathbf{x}+}$ in all compatible topological orders over $\mathbf{V}$:*

1. [*Semi-Markov*] *$P(\mathbf{V}|do(\mathbf{x}))$ is semi-Markov relative to $\mathcal{G}$;*
2. [*Missing directed-link*] *For every $V_i \in \mathbf{V} \backslash \mathbf{X}$, $\mathbf{W} \subseteq \mathbf{V} \backslash (Pa_i^{\mathbf{x}+} \cup \mathbf{X} \cup \{V_i\})$:*
$$P(v_i | do(\mathbf{x}), pa_i^{\mathbf{x}+}, do(\mathbf{w})) = P(v_i | do(\mathbf{x}), pa_i^{\mathbf{x}+}) \tag{135}$$
3. [*Missing bidirected-link*] *For every $V_i \in \mathbf{V} \backslash \mathbf{X}$, let $Pa_i^{\mathbf{x}+}$ be partitioned into two sets of confounded and unconfounded parents, Pa_i^c and Pa_i^u in $\mathcal{G}_{\overline{\mathbf{x}}}$::*
$$P(v_i | do(\mathbf{x}), pa_i^c, do(pa_i^u)) = P(v_i | do(\mathbf{x}), pa_i^c, pa_i^u) \tag{136}$$

Example 21 (SCM-induced CBN Semi-Markovian)**.** *Consider the SCM from Example 20. Its induced causal diagram is shown in Fig. 12(b) and its induced set of interventional distributions $\mathbf{P}_*$ satisfy:*

1. [*Semi-Markov*]
$$P(\mathbf{v}) = P(z)P(x|z)P(y|x,z) \tag{137}$$
$$P(\mathbf{v}|do(x)) = P(z|do(x))P(y|do(x)) \tag{138}$$
$$P(\mathbf{v}|do(y)) = P(z|do(y))P(x|z,do(y)) \tag{139}$$
$$P(\mathbf{v}|do(z)) = P(x|do(z))P(y|x,do(z)) \tag{140}$$
2. [*Missing directed-link*]
$$P(x|z,do(y)) = P(x|z) \tag{141}$$
$$P(x|do(z),do(y)) = P(x|do(z)) \tag{142}$$
$$P(y|do(x),do(z)) = P(y|do(x)) \tag{143}$$
$$P(z|do(\mathbf{a})) = P(z), \forall \mathbf{a} \subseteq \{x, y\} \tag{144}$$
3. [*Missing bidirected-link*]
$$P(x|do(z)) = P(x|z) \tag{145}$$
$$P(x|do(y,z)) = P(x|z,do(y)) \tag{146}$$
$$P(y|x,do(z)) = P(y|x,z) \tag{147}$$

The causal diagram in Fig. 12(b) is a CBN Semi-Markovian for $\mathbf{P}_*$.

A.3 $\mathcal{L}_3$: Counterfactual Bayesian Networks

If we climb further up the PCH, we get another graphical model that encodes structural constraints in the counterfactual distributions.

Definition 26 (CTFBN Markovian [1, Def. 13.2.1]). *A directed acyclic graph $\mathcal{G}$ is a Counterfactual Bayesian Network for $\mathbf{P}_{\#}$ if:*

1. [*Independence Restrictions*] *Let $\mathbf{W}_*$ be a set of counterfactuals of the form $W_{\mathbf{pa}_w}$, then $P(\mathbf{W}_*)$ factorizes as*

$$P(\bigwedge_{W_{pa_w} \in \mathbf{W}_*} W_{pa_w}) = \prod_{V \in \mathbf{V}(\mathbf{W}_*)} P(\bigwedge_{W_{pa_w} | W \in \mathbf{V}(\mathbf{W}_*)} W_{pa_w}) \tag{148}$$

2. [*Exclusion Restrictions*] *For every variable $Y \in \mathbf{V}$ with parents $\mathbf{Pa}_y$, for every set $\mathbf{Z} \subseteq \mathbf{V} \backslash (\mathbf{Pa}_y \cup \{Y\})$, and any counterfactual set $\mathbf{W}_*$, we have*

$$P(Y_{\mathbf{pa}_y, \mathbf{z}}, \mathbf{W}_*) = P(Y_{\mathbf{pa}_y}, \mathbf{W}_*) \tag{149}$$

3. [*Local Consistency*] *For every variable $Y \in \mathbf{V}$ with parents $\mathbf{Pa}_y$, let $\mathbf{X} \subseteq \mathbf{Pa}_y$, then for every set $\mathbf{Z} \subseteq \mathbf{V} \backslash (\mathbf{X} \cup \{Y\})$, and any counterfactual set $\mathbf{W}_*$, we have*

$$P(Y_{\mathbf{z}} = y, \mathbf{X}_{\mathbf{z}} = \mathbf{x}, \mathbf{W}_*) = P(Y_{\mathbf{xz}} = y, \mathbf{X}_{\mathbf{z}} = \mathbf{x}, \mathbf{W}_*) \tag{150}$$

Example 22 (SCM-induced CTFBN Markovian). *Consider the SCM from Example 18. Its induced causal diagram is shown in Fig. 12(a) and its induced set of counterfactual distributions $\mathbf{P}_{\#}$ satisfy:*

1. [*Independence Restrictions*]

$$P(z, x_z, x'_{z'}, y_{z''}, y'_{z'''}) = P(z) P(x_z, x'_{z'}) P(y_{z''}, y'_{z'''}) \tag{151}$$

2. [*Exclusion Restrictions*]

$$P(x_{yz}, \mathbf{w}_*) = P(x_z, \mathbf{w}_*) \tag{152}$$
$$P(y_{xz}, \mathbf{w}_*) = P(y_z, \mathbf{w}_*) \tag{153}$$
$$P(z_{\mathbf{a}}, \mathbf{w}_*) = P(z, \mathbf{w}_*), \forall \mathbf{a} \subseteq \{x, y\} \tag{154}$$

3. [*Local Consistency*]

$$P(x, z) = P(x_z, z) \tag{155}$$
$$P(x_y, z_y) = P(x_{yz}, z_y) \tag{156}$$
$$P(y, z) = P(y_z, z) \tag{157}$$
$$P(y_x, z_x) = P(y_{xz}, z_x) \tag{158}$$

The causal diagram in Fig. 12(a) is a CTFBN Markovian for $\mathbf{P}_{\#}$.

Definition 27 (CTFBN Semi-Markovian [1, Def. 13.2.2]). *A directed acyclic graph $\mathcal{G}$ is a Counterfactual Bayesian Network for $\mathbf{P}_{\#}$ if:*

1. [*Independence Restrictions*] *Let $\mathbf{W}_*$ be a set of counterfactuals of the form $W_{\mathbf{pa}_w}$, $\mathbf{C}_1, ..., \mathbf{C}_l$ the c-components of $G[\mathbf{V}(\mathbf{W}_*)]$, and $\mathbf{C}_{1*}, ..., \mathbf{C}_{l*}$ the corresponding partition over $\mathbf{W}_*$. Then $P(\mathbf{W}_*)$ factorizes as*

$$P(\bigwedge_{W_{pa_w} \in \mathbf{W}_*} W_{pa_w}) = \prod_{j=1}^{l} P(\bigwedge_{W_{pa_w} \in \mathbf{C}_{j*}} W_{pa_w}) \tag{159}$$

2. [*Exclusion Restrictions*] *For every variable $Y \in \mathbf{V}$ with parents $\mathbf{Pa}_y$, for every set $\mathbf{Z} \subseteq \mathbf{V} \backslash (\mathbf{Pa}_y \cup \{Y\})$, and any counterfactual set $\mathbf{W}_*$, we have*

$$P(Y_{\mathbf{pa}_y, \mathbf{z}}, \mathbf{W}_*) = P(Y_{\mathbf{pa}_y}, \mathbf{W}_*) \tag{160}$$

3. [*Local Consistency*] *For every variable $Y \in \mathbf{V}$ with parents $\mathbf{Pa}_y$, let $\mathbf{X} \subseteq \mathbf{Pa}_y$, then for every set $\mathbf{Z} \subseteq \mathbf{V} \backslash (\mathbf{X} \cup \{Y\})$, and any counterfactual set $\mathbf{W}_*$, we have*

$$P(Y_{\mathbf{z}} = y, \mathbf{X}_{\mathbf{z}} = \mathbf{x}, \mathbf{W}_*) = P(Y_{\mathbf{xz}} = y, \mathbf{X}_{\mathbf{z}} = \mathbf{x}, \mathbf{W}_*) \tag{161}$$

Example 23 (SCM-induced CTFBN Semi-Markovian)**.** *Consider the SCM from Example 20. Its induced causal diagram is shown in Fig. 12(b) and its induced set of counterfactual distributions $\mathbf{P}_{\#}$ satisfy:*

1. [*Independence Restrictions*]

$$P(z, x_z, x'_{z'}, y_{x''}, y'_{x'''}) = P(z)P(x_z, x'_{z'}, y_{x''}, y'_{x'''}) \tag{162}$$

2. [*Exclusion Restrictions*]

$$P(x_{yz}, \mathbf{w}_*) = P(x_z, \mathbf{w}_*) \tag{163}$$
$$P(y_{xz}, \mathbf{w}_*) = P(y_x, \mathbf{w}_*) \tag{164}$$
$$P(z_{\mathbf{a}}, \mathbf{w}_*) = P(z, \mathbf{w}_*), \forall \mathbf{a} \subseteq \{x, y\} \tag{165}$$

3. [*Local Consistency*]

$$P(x, z) = P(x_z, z) \tag{166}$$
$$P(x_y, z_y) = P(x_{yz}, z_y) \tag{167}$$
$$P(y, x) = P(y_x, x) \tag{168}$$
$$P(y_z, x_z) = P(y_{xz}, x_z) \tag{169}$$

The causal diagram in Fig. 12(b) is a CTFBN Semi-Markovian for $\mathbf{P}_{\#}$.

A.4 Counterfactual Randomization

An agent may sometimes interact with a system of interest through experiments, thereby collecting data from different layers of the PCH. Counterfactual randomization is an experimental procedure that enables an agent to observe the value of a variable before an intervention takes effect [4]. For instance, a doctor may be able to determine a patient's natural choice of drug prior to randomly assigning treatment in a clinical trial. This extension of experimental capability is formalized in the following definition of a new type of physical action that an agent may be able to perform in an environment.

Definition 28 (Counterfactual (ctf-) Randomization (Def. 2.3 [24]))**.** *For a variable X and some particular unit i[7] in the target population of the environment, the operation*

$$\text{CTF-RAND}(X \to \mathbf{C})^{(i)} \tag{170}$$

denotes fixing the value of X as an input to the mechanisms generating $\mathbf{C} \subseteq Ch(X)$ for this particular unit, where $Ch(X)$ is the set of variables whose mechanisms take X as an argument.

The value of X is assigned by a randomizing device with support over $\text{Domain}(X)$.

The essential differences between Fisherian randomization and CTF-RAND$(X \to \mathbf{C})^{(i)}$ are:

1. CTF-RAND does not erase unit i's natural decision $X^{(i)}$[8].
2. While Fisherian randomization affects all children of X, CTF-RAND only affects the chosen subset $\mathbf{C} \subseteq Ch(X)$, leaving $Ch(X) \setminus \mathbf{C}$ untouched.

Importantly, CTF-RAND can only be enacted under certain structural conditions. These include environments where one can measure a unit's natural decision while simultaneously randomizing its actual decision [4], or settings where counterfactual mediators allow altering how a subset of children perceive the value of X [24]. In either case, ctf-randomization enables multiple randomizations on the same variable X for a single unit i. Further, CTF-RAND must always be applied with respect to a graphical child variable; it is not possible to bypass a child and directly alter the perception of a descendant.

[7] This definition discusses a unit-specific experimental procedure, as it takes a physical perspective on how an agent interacts with the units in a system.

[8] Another way to understand this difference is that the unit's natural inclination is taken into account.

Example 24 (CTF-RAND). *Consider the SCM from Example 1. Counterfactual randomization on X allows an agent to observe the natural value of X, say x', while simultaneously assigning a specific value x as input to its children Z and Y. This is illustrated graphically in Fig. 8 (b), and as a result, the $\mathcal{L}_3$ distribution $P(X = x', Z_{X=x}, Y_{X=x})$ becomes experimentally accessible (i.e., realizable).* ■

By including the counterfactual randomization action into our experimental toolkit, we obtain the action set that gives the agent the most granular experimental capabilities.

Definition 29 (Maximal Feasible Action Set (SCM) [24]). *Given an SCM $\mathcal{M} = \langle \mathbf{U}, \mathbf{V}, \mathcal{F}, P(\mathbf{u}) \rangle$. The maximal feasible action set $\mathbb{A}^{\dagger}(\mathcal{M})$ is the set of all actions the agent can perform in $\mathcal{M}$ with the most granular interventional capabilities:*

(i) $\textsc{Select}^{(i)}$*: randomly choosing, without replacement, a unit i from the target population, to observe in the system;*

(ii) $\textsc{Read}(V)^{(i)}, \forall V \in \mathbf{V}$*: measuring the way in which a causal mechanism $f_V \in \mathcal{F}$ has physically affected unit i, by observing its realised feature $V^{(i)}$;*

(iii) $\textsc{Rand}(X)^{(i)}, \forall X \in \mathbf{V}$*: erasing and replacing i's natural mechanism f_X for a decision variable X with an enforced value drawn from a randomising device having support over Domain(X);*

(iv) $\textsc{ctf-Rand}(X \rightarrow C)^{(i)}, \forall X, \forall C \in Ch(X)$*: fixing the value of X as an input to the mechanisms generating $C \in Ch(X)$ using a randomising device having support over Domain(X), for unit i, where $Ch(X)$ stands for the set of variables that take X as an argument in their mechanisms.*

SELECT with READ correspond to random sampling. When SELECT and READ are permitted over all units and variables, all distributions in $\mathcal{L}_1$ are realizable. Adding RAND to the action set gives the agent the ability to perform randomized experiments. When SELECT, READ and RAND are permitted over all units and variables, all distributions in $\mathcal{L}_2$ are realizable. With CTF-RAND, some distributions in $\mathcal{L}_3$ also become realizable. These distributions are the ones that lie within $\mathbf{P}^{\mathcal{L}_{2.25}}$ and $\mathbf{P}^{\mathcal{L}_{2.5}}$. If we can perform all actions from the maximal feasible action set in an environment, we are able to draw samples from any distributions in $\mathbf{P}^{\mathcal{L}_{2.5}}$.

B Details on Models and Inferential Machinery

B.1 Graphical Criterion for Distributions in $\mathcal{L}_{2.5}$

We reproduce the sound and complete graphical criterion for checking a distribution is in $\mathcal{L}_{2.5}$ from [24] below.

Definition 30 (Ancestors of a Counterfactual[6]). *Given a potential response $Y_{\mathbf{x}}$ with $Y \in \mathbf{V}, \mathbf{X} \subseteq \mathbf{V}$, the set of counterfactual ancestors of $Y_{\mathbf{x}}$, denoted by $An(Y_{\mathbf{x}})$), consist of each $W_{\mathbf{z}}$ such that $W \in An(Y)_{\mathcal{G}_{\overline{\mathbf{x}}}} \backslash \mathbf{X}$ (which includes Y itself), and $\mathbf{z} = \mathbf{x} \cap An(W)_{\mathcal{G}_{\overline{\mathbf{x}}}}$. For a set of counterfactuals $\mathbf{W}_*$, $An(\mathbf{W}_*)$ is defined to be the union of the ancestors of each potential response in the set.*

Lemma 3 (Corollary 3.7 in [24]). *Given a causal diagram $\mathcal{G}$, an $\mathcal{L}_3$-distribution $Q = P(\mathbf{W}_*)$ is in maximally realizable distributions induced by any SCM compatible with a given graph $\mathcal{G}$ if and only if the ancestor set $An(\mathbf{W}_*)$ does not contain a pair of potential responses $W_{\mathbf{s}}, W_{\mathbf{t}}$ of the same variable W under different regimes where $\mathbf{s} \neq \mathbf{t}$.*

Example 25. *Consider the causal diagram in Fig. 9(a), we check if $P(Z_x, Y_{x'})$ is in the $\mathcal{L}_{2.5}$ distributions induced by SCMs compatible with it.*

- $An(Z_x) = \{Z_x\}$
- $An(Y_{x'}) = \{Y_{x'}\}$

Applying Lemma 3, we conclude that $P(Z_x, Y_{x'})$ is in the $\mathcal{L}_{2.5}$ distributions.

B.2 Independence Constraints and AMWN

The independence rule in ctf-calculus requires the construction of another graphical object, known as the *Ancestral Multi-World Network* (AMWN). We reproduce the algorithm for AMWN construction and the theorem stating its soundness.

Algorithm 1 AMWN-CONSTRUCT($\mathcal{G}, \mathbf{W}_*$)

Input: Causal Diagram $\mathcal{G}$ and a set of counterfactual variables $\mathbf{W}_*$
Output: $\mathcal{G}_A(\mathbf{W}_*)$, the AMWN constructed from $\mathcal{G}$ and $\mathbf{W}_*$

1: Initialise $\mathcal{G}'$ by adding variables in $An(\mathbf{W}_*)$ together with the directed arrows witnessing the ancestrality
2: **for** each node $V \in \mathbf{V}$ appearing more than once in $\mathcal{G}'$ **do**
3: Add a node U_V and an edge $U_V \to V_{\mathbf{x}}$ for every instance $V_{\mathbf{x}}$ of V.
4: **end for**
5: **for** each bidirected $V \dashleftarrow\!\!\dashrightarrow W$ where V and W are in $\mathcal{G}'$ **do**
6: Add a node U_{VW} and edges from it to $V_{\mathbf{x}}$ and $W_{\mathbf{x}}$ for every instance $V_{\mathbf{x}}$ of V or $W_{\mathbf{x}}$ of W in $\mathcal{G}'$.
7: **end for**
return $\mathcal{G}'$.

Theorem 6 ($\mathcal{L}_3$ Independence Constraints – Counterfactual d-separation)**.** *(Theorem 1 in [6]) Consider a causal diagram $\mathcal{G}$ and a collection of counterfactual distributions, $\mathbf{P}^{\mathcal{L}_3}$, induced by the SCM associated with $\mathcal{G}$. For counterfactual variables $X_{\mathbf{t}}, Y_{\mathbf{r}}, \mathbf{Z}_*$,*

$$(\|X_{\mathbf{t}}\| \perp\!\!\!\perp \|Y_{\mathbf{r}}\| \mid \|\mathbf{Z}_*\|)_{\mathcal{G}_A} \implies (\|X_{\mathbf{t}}\| \perp\!\!\!\perp \|Y_{\mathbf{r}}\| \mid \|\mathbf{Z}_*\|)_{\mathbf{P}^{\mathcal{L}_3}} \tag{171}$$

In words, if $\|X_{\mathbf{t}}\|$ and $\|Y_{\mathbf{r}}\|$ are d-separated given $\|\mathbf{Z}_*\|$ in the diagram $\mathcal{G}_A(X_{\mathbf{t}}, Y_{\mathbf{r}}, \mathbf{Z}_*)$, then $\|X_{\mathbf{t}}\|$ and $\|Y_{\mathbf{r}}\|$ are independent given $\|\mathbf{Z}_*\|$ in every distribution $\mathbf{P}^{\mathcal{L}_3}$ compatible with the causal diagram $\mathcal{G}$.

When adapting ctf-calculus to CBN2.25 and CBN2.5, there is an extra step to ensure that the distributions belong to the corresponding layers. This can be added as an extra step before Step 1 of Alg. 1 to check that:

- CBN2.25: $CRS(\mathbf{W}_*)$ satisfies Lemma 2
- CBN2.5: $An(\mathbf{W}_*)$ satisfies Lemma 3

The same check applies to the other two rules of ctf-calculus too.

C Discussion on Hierarchy of Graphical Models

In this section, we offer further insights into the hierarchy of graphical models by examining (a) the set of compatible SCMs and (b) the action sets required to render their encoded constraints empirically falsifiable.

C.1 Hierarchy of SCMs compatible with Graphs

Considering a causal diagram $\mathcal{G}$ and its constraints, $\mathcal{G}$ can be viewed as a representation of an equivalence class of SCMs that induce distributions that are compatible with these constraints. We formally define this notion of compatibility below.

Definition 31 (SCM compatible with $\mathcal{G}$ on $\mathcal{L}_i$)**.** *Given a causal diagram $\mathcal{G}$, an SCM $\mathcal{M}$ is said to be compatible with $\mathcal{G}$ on $\mathcal{L}_i$, if all constraints encoded by $\mathcal{G}$, when interpreted as a graphical model on $\mathcal{L}_i$ hold in the $\mathcal{L}_i$ distributions induced by $\mathcal{M}$.*

Example 26. *Consider the causal diagram $\mathcal{G}$ in Fig. 12 (c).*

- *When $\mathcal{G}$ is interpreted as an $\mathcal{L}_1$ model, it encodes no constraints. As a result, any SCM with an endogenous variable set consisting of two variables is compatible with $\mathcal{G}$ on $\mathcal{L}_1$.*

- *When $\mathcal{G}$ is interpreted as an $\mathcal{L}_2$ model, it encodes the following constraints:*

$$P(y|do(x)) = P(y|x) \tag{172}$$
$$P(x|do(y)) = P(x) \tag{173}$$
$$\tag{174}$$

Thus, any SCM compatible with $\mathcal{G}$ on $\mathcal{L}_2$ must induce a collection of interventional distributions that satisfy these constraints. For example, SCMs with Y being an argument in the mechanism f_x will typically induce distributions that fail to satisfy these two constraints (unless the SCM has very peculiar parameterizations where the effect from Y to X are zeroed out). The same applies to SCMs with f_x and f_y sharing some common or correlated exogenous factors.

The example above provides a glimpse of how the set of SCMs compatible with a causal diagram shrinks as we transition from $\mathcal{L}_1$ to $\mathcal{L}_2$. In fact, this property generalizes across all layers: as we move to higher layers, additional constraints are imposed, further restricting the set of compatible SCMs.

C.2 Hierarchy of Constraints from Realizability

Another way to compare graphical models across different layers is by analyzing the empirical constraints, and whether these constraints are falsifiable or not. The falsifiability of a constraint does not depend how it is encoded in a particular model, but rather on the realizability of the distributions involved. In particular, a constraint is empirically falsifiable only if the agent has the experimental capability to sample from all distributions that are evoked in the constraint. Therefore, the hierarchy of graphical models can be understood by examining the action sets required to make the corresponding distributions realizable at each layer.

First, we define the collection of all actions an agent is allowed to perform in a system in principle.

Definition 32 (Maximal Feasible Action Set (SCM) [24])**.** *Given an SCM $\mathcal{M} = \langle \mathbf{U}, \mathbf{V}, \mathcal{F}, P(\mathbf{u}) \rangle$. The maximal feasible action set $\mathbb{A}^{\dagger}(\mathcal{M})$ is the set of all actions the agent can perform in $\mathcal{M}$ with the most granular interventional capabilities:*

(i) SELECT$^{(i)}$: randomly choosing, without replacement, a unit i from the target population, to observe in the system;

(ii) READ$(V)^{(i)}, \forall V \in \mathbf{V}$: measuring the way in which a causal mechanism $f_V \in \mathcal{F}$ has physically affected unit i, by observing its realised feature $V^{(i)}$;

(iii) RAND$(X)^{(i)}, \forall X \in \mathbf{V}$: erasing and replacing i's natural mechanism f_X for a decision variable X with an enforced value drawn from a randomising device having support over Domain(X);

(iv) CTF-RAND$(X \to C), \forall X \in \mathbf{V}, \forall C \in Ch(X)$: fixing the value of X as an input to the mechanisms generating $C \in Ch(X)$ using a randomizing device having support over $Val(X)$, for unit i, where $Ch(X)$ stands for the set of variables that take X as an argument in their mechanisms.

Based on the results from [24], we know that

- With the first three actions from the maximal feasible action set, we can access all distributions in $\mathbf{P}^{\mathcal{L}_2}$, which allows us to empirically test all constraints encoded by a CBN.
- With all four actions from the maximal feasible action set, we can access all distributions in $\mathbf{P}^{\mathcal{L}_{2.5}}$, which allows us to empirically test all constraints encoded by a CBN2.5.

The ability for an agent to perform counterfactual randomization allows it to access distributions that lie between $\mathcal{L}_2$ and $\mathcal{L}_{2.5}$. As discussed earlier, the randomization procedures associated with $\mathcal{L}_{2.25}$ assumes all children to take the same value. It is clear that the action set to realize $\mathcal{L}_{2.25}$ lie between $\mathcal{L}_2$ and $\mathcal{L}_{2.5}$.

From the definitions of action sets, we observe a hierarchical structure in the feasible actions an agent can perform to access distributions at different layers, as illustrated in Fig. 13. Specifically, the action

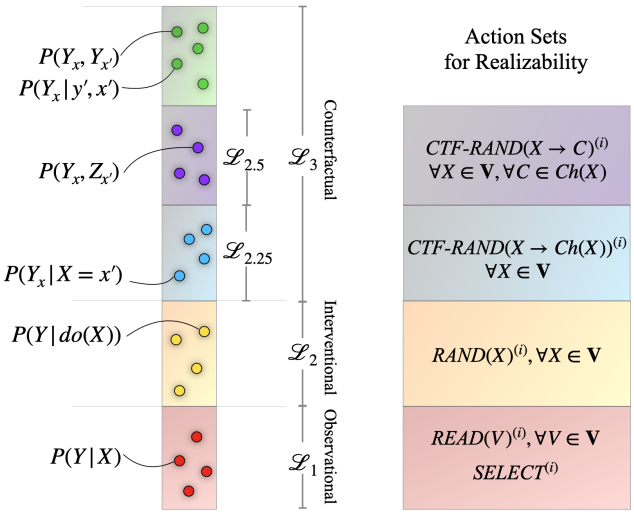


Figure 13: Hierarchy of action sets to realize distributions in different layers

set for $\mathbf{P}^{\mathcal{L}_2}$ is a subset of the action set for $\mathbf{P}^{\mathcal{L}_{2.25}}$, which in turn is a subset of the action set for $\mathbf{P}^{\mathcal{L}_{2.5}}$. This hierarchy of action sets match the hierarchical structure of distributions from the perspective of realizability.

For $\mathcal{L}_3$ distributions that lie beyond $\mathbf{P}^{\mathcal{L}_{2.5}}$, there is currently no known experimental procedure to sample from them. In fact, there are some independence constraints in CTFBNs that are not falsifiable, which involve cross-world constraints [25]. While we acknowledge the validity of this claim, we emphasize that the difference in empirical testability between constraints in CBN2.25 and CTFBN does not arise from whether the constraints are cross-world. Instead, it stems from the set of actions permitted to access the counterfactual variables. For instance, the ETT distribution $P(Y_x = y|X = x')$ in $\mathbf{P}^{\mathcal{L}_{2.25}}$ is technically a 'cross-world' quantity, as Y_x is derived from the interventional regime $\mathcal{M}_x$, while X originates from the natural regime $\mathcal{M}$. However, under the assumption that FFRCISTG randomization on X is a valid action within the system – allowing observation of the X's value before the intervention takes effect – constraints involving the ETT distribution can be empirically falsifiable.

To highlight its practical implications, the falsifiability of a constraint depends on whether we can access the distributions it is defined on, and the realizability of these distributions hinges on the set of physical actions available to the agent in the environment. Therefore, it is crucial for researchers to understand the limitations of their actions when assessing whether the assumptions they make can be justified through experiments or expert knowledge. Researchers who prioritize empirically testable assumptions can choose a graphical model from the hierarchy that encodes only falsifiable constraints based on the available actions. Conversely, those willing to incorporate assumptions based on background knowledge can do so with a clear understanding of which constraints remain untested given the permissible actions. In summary, graphical models are tools, and the realizability of the distributions underlying the constraints they encode is one of many important criteria that helps researchers assess their appropriateness for specific applications.

D Relationship with Other Graphical Models

Besides the graphical models discussed in this paper, there are other graphic-based approaches. In this section, we examine various graph-based approaches and provide a mathematical account of how they relate to the structural framework to causality developed so far. We begin in Sec. D.1 with a brief review of the causal inference pipeline that was discussed earlier. This pipeline encompasses both the modeling component and the inferential machinery. In Sec. D.2, we discuss the Fully Randomized Causally Interpretable Structured Tree Graph (FFRCISTG) proposed by Jamie Robins and colleagues [26, 25], and in Sec. D.3 we turn to the augmented DAGs/decision-theoretic approach developped by

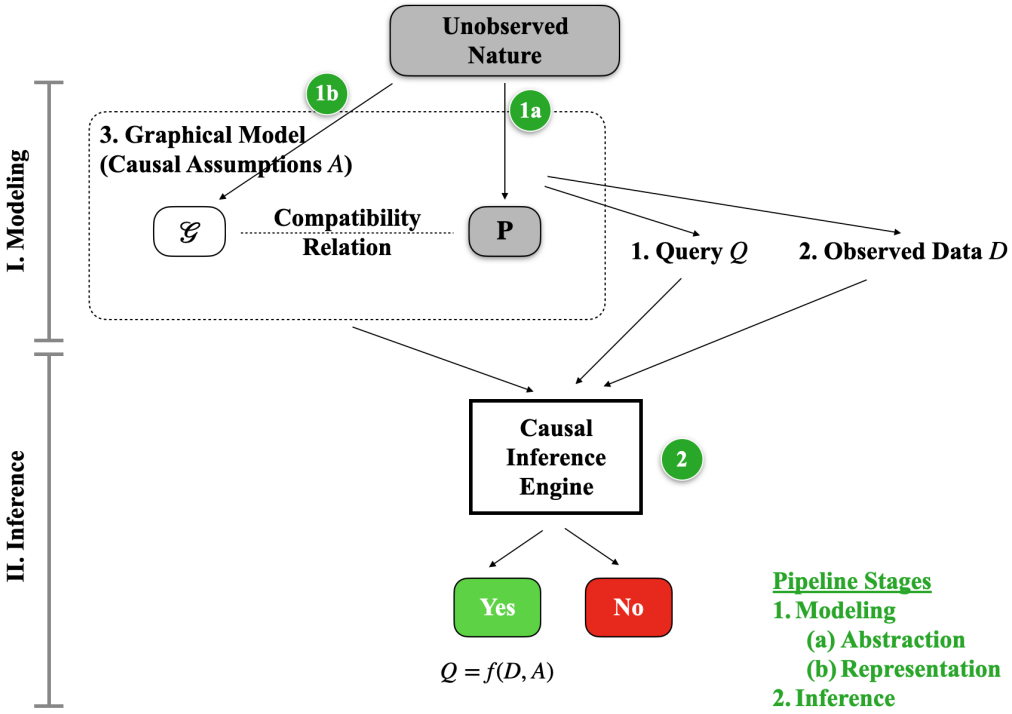


Figure 14: Stages of the causal inference pipeline

Philip Dawid [9, 10]. We then analyze how these approaches align with different stages of the causal pipeline and discuss when they agree or disagree with models grounded in the SCM framework.

D.1 Causal Inference Pipeline

We start by revisiting the causal inference pipeline, which can be framed as a two-stage process: first comes the modeling stage, which proceeds from abstraction to explicit model construction, followed by the inference stage, which are illustrated in Fig. 14.

1. *Modeling*. This stage encompasses both the ontological commitments we make about how the world generates data and the explicit representations we construct to encode those commitments. It proceeds in two stages:

   (1a) *Abstraction (Ontological Commitment)*. At this stage we posit that the data we observe are generated by underlying mechanisms, even if these mechanisms are not directly observable. The generative model is typically represented as distributions over variables, some of which may remain unobserved. For example, under the SCM framework, the data-generating model is a fully instantiated SCM (Def. 1), and the induced distributions correspond to the PCH (Def. 4), which is only partially observable (e.g., the observational distribution). Abstraction also includes qualitative commitments, such as invariance assumptions (Def. 5), which discard numerical details while retaining qualitative relationships among distributions. These commitments bridge what is unobserved with what is empirically available.

   (1b) *Representation (Model Construction)*. Once the abstraction is in place, we construct explicit causal models that parsimoniously encode assumptions about the generative processes. Most commonly, this representation takes the form of a causal diagram. The input may come directly from the full data-generating model or from the qualitative assumptions abstracted at that stage (step 1a). Different families of graphical models are defined based on the type of compatibility relation between the diagram and the distributions. For instance, a causal diagram (Def. 6) is constructed from an SCM by retaining only the functional dependencies among variables and the independence

relations among exogenous variables. Graphical models built on top of such a diagram, such as a CBN (Def. 21) or a CTFBN (Def. 26), encode the causal assumptions relevant to specific layers of the PCH.

2. *Inference*. This stage is carried out by the *causal inference engine*, illustrated in Fig. 14. Given three inputs, (1) *Query*, (2) *Data*, and (3) *Model*, the engine decides whether the target query, typically unobserved, can be expressed in terms of the available (observed) data, relying on the assumptions encoded in the model (step 1b). For example, given a query $Q = P(y \mid do(x))$, observational data $P(\mathbf{V})$, and a CBN over the causal diagram $X \to Y$, *do-calculus* can be used to derive that $P(y \mid do(x)) = P(y \mid x)$.

Based on the definitions and examples provided earlier in this chapter, all stages of the causal inference pipeline are well defined within the SCM framework. In contrast, we now examine whether the same level of definitional completeness holds for graphical models developed under other perspectives. To better understand these alternative models, we will (1) reproduce the key definitions underlying each model, (2) analyze which stage(s) of the pipeline these definitions occupy, and (3) identify any stages where essential definitions are missing or under-specified.

D.2 Graphical Models from PO Framework: FFRCISTG and SWIG

The first graphical model that we analyze is the Fully Randomized Causally Interpretable Structured Tree Graph (FFRCISTG) developed by Robins and Richardson [25, 26], henceforth referred to as RR.

D.2.1 Construction of FFRCISTG (Modeling stage)

In this section, we examine the components and processes involved in the first stage of the pipeline as applied to FFRCISTG models. Specifically, we will analyze four different components: (1) the underlying data-generating model, (2) the distributions it induces, (3) the causal assumptions derived from coarsening these distributions, and (4) the graphical representation constructed to encode those assumptions.

(1) Data-generating model

The underlying data-generating model of an FFRCISTG is a structural equation model, which we refer to as SEM-RR. While a formal definition of the SEM-RR is not provided in the original reference, it is described as follows:

> We wish to emphasize here that FFRCISTGs may also be explicitly defined via a system of structural equations (with possibly dependent errors – though any such dependence is undetectable via randomized experiments) [25, Page 1].
>
> As we have described above, in an NPSEM [9] each variable is given by a function f_V expressed as an arbitrary function of the variables that are its parents in the graph, together with an error term [25, Page 17].

We reproduce a specific SEM-RR from one example in the paper here.

Example 27 (Example of SEM-RR from [25, Sec. 6.1.1])**.** *An SEM-RR with dependent errors is:*

$$Z = \varepsilon_Z; \tag{175}$$
$$M(z) = (1 - z) \cdot \mathbb{I}(\varepsilon_M > 1/2); \tag{176}$$
$$Y(z, m) = z \cdot \mathbb{I}(\varepsilon_Y > 1/2); \tag{177}$$
$$\varepsilon_Z \perp\!\!\!\perp \{\varepsilon_M, \varepsilon_Y\}; \tag{178}$$
$$\varepsilon_Z \sim Ber(1/2); \tag{179}$$
$$\varepsilon_M = \varepsilon_Y \sim Unif(0, 1). \tag{180}$$

[9] The name used for such a structural equation model in the original reference is NPSEM, which stands for non-parametric structural equation model. The term 'non-parametric' in the name refers to models where the functions are not parametrized and are left in the form $V_i \leftarrow f_i(Pa_i, \varepsilon_i)$. However, many examples presented in the paper specify full parametric details of the functions, much like done in SCMs. Thus, we choose to stay with the parametrized version of the model and use the name SEM-RR instead.

Based on these descriptions and the examples provided by RR, we formulate a more precise definition of an SEM-RR that parallels the structure of an SCM.

Definition 33 (Structural Equation Model - RR (SEM-RR))**.** *A structural equation model - RR $\mathcal{N}$ is a 4-tuple $\langle \mathbf{V}', \varepsilon_{\mathbf{V}'}, \mathcal{F}, P(\varepsilon_{\mathbf{V}'}) \rangle$, where*

- $\mathbf{V}' = \{\mathbf{V}, \mathbf{U}\}$ *is a set of random variables, either endogenous ($\mathbf{V}$) or exogenous ($\mathbf{U}$);*
- $\varepsilon_{\mathbf{V}'} = \{\varepsilon_i, \forall V_i \in \mathbf{V}'\}\rangle$ *is a set of error terms, such that for each $V_i \in \mathbf{V}'$, ε_i directly affects the value of V_i.*
- *$\mathcal{F}$ is a set of functions $\{f_1, f_2, ..., f_n\}$ such that each f_i is a mapping from (the respective domains of) $\mathbf{Pa}_i \cup \varepsilon_i$ to $V_i \in \mathbf{V}'$, where $\mathbf{Pa}_i \subseteq \mathbf{V}' \backslash V_i$. That is, for $i = 1, ..., n$, each $f_i \in \mathcal{F}$ is such that*

$$v_i \leftarrow fi(\mathbf{pa}_i, \varepsilon_i) \tag{181}$$

- *$P(\varepsilon_{\mathbf{V}'})$ contains the probability functions defined over each ε_i.*

We now compare the components of an SEM-RR and an SCM. In an SCM, the set of variables is partitioned into exogenous $\mathbf{U}$ and endogenous $\mathbf{V}$, where functions are defined for $\mathbf{V}$ and distributions are specified over $\mathbf{U}$ to capture all randomness arising from factors external to the model. In contrast, SEM-RRs as modeled by RR do not make this partition explicit. Their set $\mathbf{V}'$ includes both observed and unobserved variables, with functions defined for all of them. A different set of exogenous variables is grouped under the label of *error terms*, over which probability distributions are defined.

Example 28 (SCM vs SEM-RR)**.** *Consider the SEM-RR in Example 27. If Z is a latent variable, the SEM-RR will remain mostly the same, except that Z is now labeled as 'unobserved' and becomes a member of $\mathbf{U}$.*

However, the SCM $\mathcal{M}$ that represents the same level details will group Z and all error terms together under $\mathbf{U}$, as follows:

$$\mathbf{V} = \{M, Y\} \tag{182}$$
$$\mathbf{U} = \{U, U_m, U_y\} \tag{183}$$
$$\mathcal{F} = \begin{cases} M \leftarrow (1 - U) \cdot \mathbb{I}(U_M > 1/2) \\ Y \leftarrow U \cdot \mathbb{I}(U > 1/2) \end{cases} \tag{184}$$
$$P(\mathbf{U}) : \begin{cases} U \sim Ber(1/2) \\ U_m = U_y \sim Unif(0, 1) \end{cases} \tag{185}$$

From the descriptions of SEM-RR, we observe that it is naturally associated with a directed acyclic graph (DAG). While no formal definition or construction procedure for this DAG is provided by RR, we offer such a formal definition based on the examples and discussions, which we call *FFRCISTG diagram*. This formalization will facilitate a direct comparison between this new object and the causal diagram as discussed in the literature (Def. 6).

Definition 34 (FFRCISTG Diagram (From SEM-RR))**.** *Consider an SEM-RR $\mathcal{N} = \langle \mathbf{V}', \varepsilon_{\mathbf{V}'}, \mathcal{F}, P(\varepsilon_{\mathbf{V}'}) \rangle$. A graph $\mathcal{G}^F$ is said to be an FFRCISTG diagram of $\mathcal{N}$ if constructed as follows:*

(1) add a vertex for every variable in the set $\mathbf{V}'$,

(2) color all vertices representing latent variables in $\mathbf{V}'$ gray,

(3) add an edge $V_i \longrightarrow V_j$ for every $V_i, V_j \in \mathbf{V}'$ if V_i appears as an argument of $f_j \in \mathcal{F}$.

An FFRCISTG diagram can be viewed as a coarsening over the space of SEM-RRs: they retain the structural skeleton (i.e., variable dependencies) while omitting the precise functional parameters and distribution over exogenous factors. Comparing the definitions of the FFRCISTG diagram and the causal diagram (Def. 6) reveals two key differences. First, in an FFRCISTG diagram, exogenous variables that are shared across functions are explicitly represented as gray nodes; in contrast, such shared exogenous influences are often captured by bidirected edges between the

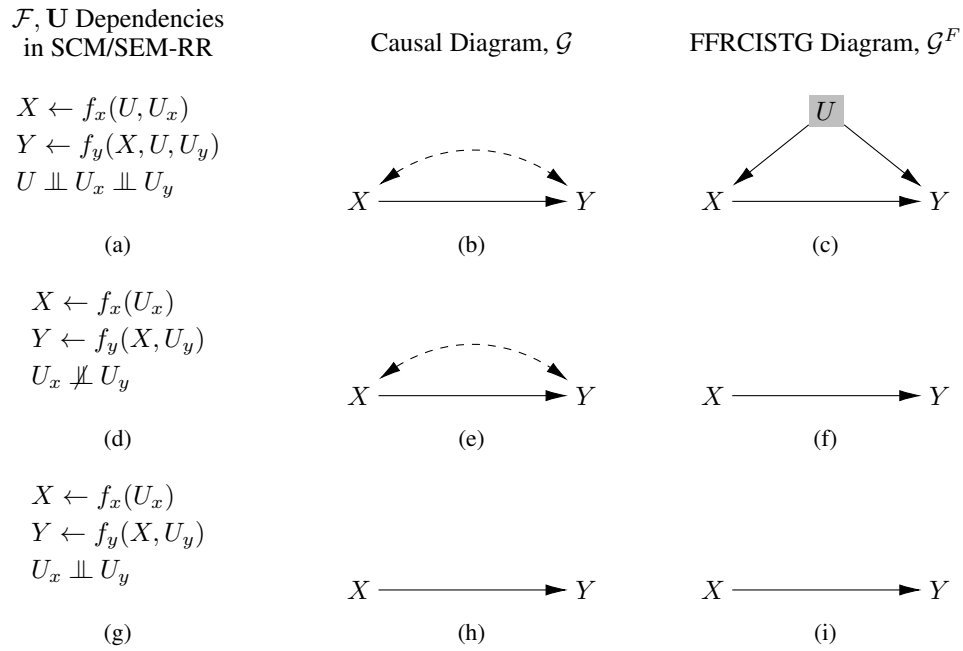


Figure 15: Comparison of how SCMs/SEM-RRs with different functional and latent dependencies induce different causal diagrams and FFRCISTG diagrams.

affected endogenous variables in a causal diagram. [10] Second, a causal diagram includes bidirected edges between endogenous variables to represent nonzero correlations between their error terms. In contrast, FFRCISTG diagrams do not graphically represent this type of latent dependency at all. In the original reference, the authors explained that there is no need represent such error term correlations based on the ground that they are never empirically testable.

> Presumably Pearl, in insisting on the NPSEM-IE model, would view the DAG in Figure 9(i) as mis-specified since the dependence between the errors ε_M and ε_Y is not represented on the graph. In contrast, since the FFRCISTG model allows for such dependence, there is no need for the graph to represent the dependence, especially when, as discussed above, there is no experimental test (involving the variables Z, M and Y) that could distinguish between the two models.

However, we will see that even though the error correlations are not testable themselves, they will induce correlations among counterfactuals over the endogenous variables which are also not correctly captured by the FFRCISTG diagram.

Example 29 (Causal diagram vs FFRCISTG diagram (Explicit vs Implicit Latent Variable))**.** *Consider the SEM-RR $\mathcal{N}^{(1)}$ in Example 27 with Z being a latent variable and its equivalent SCM $\mathcal{M}$ in Example 28.*

The functional and exogenous variable dependencies in $\mathcal{M}$ are illustrated in Fig. 15(a), where directed edges represent functional relationships among all exogenous and endogenous variables.

The causal diagram induced by $\mathcal{M}$ is shown in Fig. 15(b), where a bidirected edge captures the shared latent confounder U. In contrast, the FFRCISTG diagram induced by $\mathcal{N}^{(1)}$, shown in Fig. 15(c), includes a gray node to represent latent confounding.

Example 30 (Causal diagram vs FFRCISTG diagram (Correlated Errors))**.** *Consider an SEM-RR*

$$\mathcal{N}^{(2)} = \langle \mathbf{V}' = \{X, Y\}, \varepsilon_{\mathbf{V}'} = \{\varepsilon_X, \varepsilon_Y\}, \mathcal{F}, P(\varepsilon_{\mathbf{V}'}) \rangle, \tag{186}$$

[10] The use of bidirected arrows is the current standard for operationalizing model construction in the literature, following early work in the field. This convention is generally sufficient for most tasks, such as identification, but there are subtleties. In particular, challenges emerge in bounding problems and in situations where the structure of latent variables must be taken into account.

where

$$\mathcal{F} = \begin{cases} X = \varepsilon_X; \\ Y = X \vee \varepsilon_Y; \end{cases} \tag{187}$$

$$P(\varepsilon_{\mathbf{V}'}) : \begin{cases} \varepsilon_X \sim Ber(0.4); \\ \varepsilon_Y \mid \varepsilon_X = 0 \sim Ber(0.3); \\ \varepsilon_Y \mid \varepsilon_X = 1 \sim Ber(0.7). \end{cases} \tag{188}$$

The SCM $\mathcal{M}$ that is equivalent to $\mathcal{N}^{(2)}$ will have both error terms grouped under $\mathbf{U}$, with $P(\mathbf{U}) = P(\varepsilon_{\mathbf{V}'})$. The functional and exogenous variable dependencies in $\mathcal{M}$ and $\mathcal{N}^{(2)}$ are illustrated in Fig. 15(d), where the bidirected edge between U_x and U_y represent nonzero correlation between them. The causal diagram induced by $\mathcal{M}$ is shown in Fig. 15(e), where a bidirected edge captures the nonzero correlation between the latent parents of X and Y. In contrast, the FFRCISTG diagram induced by $\mathcal{N}^{(2)}$ is shown in Fig. 15(f), where the nonzero correlation between the error terms are not shown at all.

The construction from an SEM-RR to an FFRCISTG diagram aligns naturally with representation stage of the causal inference pipeline. However, we will see that issues will arise when the FFRCISTG diagram is used to partition the space of underlying models, exactly due to its lack in modeling expressiveness to represent all latent correlations. Before elaborating on these inferential issues, we continue focusing on the abstraction stage of the pipeline, examining definitions related to data generation and qualitative coarsening of distributions.

(2) Counterfactuals and Induced Distributions

The distributions generated from the underlying models are over the set of counterfactual variables defined as follows.

Definition 35 (SEM-RR Counterfactual Existence Assumption [25, Def. 1])**.** *Let $\mathbf{V}'$ be an underlying set of random variables.*

(i) For each variable $V \in \mathbf{V}'$ [11] *and assignment $\mathbf{pa}_V$ to $\mathbf{Pa}_V$, the parents of V in $\mathcal{G}$, we assume the existence of a counterfactual variable $V(\mathbf{pa}_V)$.*

(ii) For any set $\mathbf{R}$, with $\mathbf{R} \neq \mathbf{Pa}_V$, $V(\mathbf{r})$ is defined recursively via:

$$V(\mathbf{r}) = V(\mathbf{Pa}_V \cap \mathbf{r}, (\mathbf{Pa}_V \backslash \mathbf{R})(\mathbf{r})), \text{[12]} \tag{189}$$

where $(\mathbf{Pa}_V \backslash \mathbf{R})(\mathbf{r}) \equiv \{V^*(\mathbf{r}) \mid V^* \in \mathbf{Pa}_V, V^* \notin \mathbf{R}\}$.

The bracket notation used in the above definition serves a similar role to the subscript notation employed in the potential outcome variables from Def. 3. However, there are several key differences in semantics worth highlighting.

First, the counterfactual variables in the definition above are defined based on the topological order in the FFRCISTG diagram ("parents of V in $\mathcal{G}$"), rather than being derived directly from the underlying data-generating SEM-RR. Thus, this definition is more syntactic in nature, specifying how counterfactual expressions should be written, or expanded, rather than how the model should be constructed. In particular, the semantics of these counterfactual variables are provided separately in another paragraph of the original reference:

> The resulting counterfactual distribution $P(\mathbf{V}(\mathbf{a}))$ is obtained from the SEM-RR by simply replacing each variable $A_i \in \mathbf{A}$ by the value assigned a_i in the function f_V for any variable V of which A_i is a parent in $\mathcal{G}$ [25, Page 15].

This description suggests the same approach as the one used in the SCM framework are used to derive counterfactual variables through submodels by modifying the functional relationships accordingly to

[11] We use $\mathbf{V}'$ to denote the set of variables in FFRCISTG, reserving $\mathbf{V}$ for endogenous/observed variables.

[12] Similar to nested counterfactuals.

a specific substitution [13]. There are also no detailed descriptions on how the functions in SEM-RRs are evaluated to induce distributions over the counterfactual variables. Based on examples given in the paper, they also seem to follow the same approach as SCMs, and we denote the collection of all distributions over counterfactual variables induced by an SEM-RR by $\mathbf{P}^F$.

Second, this definition not only specifies which counterfactual variables are allowed but also implicitly encodes two properties: exclusion and consistency. Both are consequences of the recursive operation in clause (ii), as stated in the original paper by RR:

> This is referred to as the individual level 'exclusion restriction'... Assumption (ii) combines the assumptions described in other works as 'consistency' and 'recursive substitution' [25, Page 21].

The exclusion property allows the removal of variables in $\mathbf{R} \setminus \mathbf{Pa}_V$ from the bracket, indicating that interventions on non-parents of V do not affect its value. The consistency property is reflected in the inclusion of variables over $\mathbf{Pa}_V \setminus \mathbf{R}$ sharing the same value assignment $\mathbf{r}$ in the bracket. These two properties already encode some assumptions from qualitative coarsening over the distributions of the counterfactual variables. In other words, this definition lies within step 1a of the causal inference pipeline.

Example 31 (Counterfactuals in FFRCISTG)**.** *Consider the FFRCISTG diagram $\mathcal{G}^F$ in Fig. 15 (c). The counterfactual variables defined by clause (i) of Def. 35 are X and $Y(x)$. Clause (ii) of Def. 35 further implies:*

$$X(y) = X \tag{190}$$

$$Y = Y(X) \tag{191}$$

Eq. 190 illustrates the exclusion restriction where non-parents of X (Y in this case) are removed from the bracket. Eq. 191 implies the consistency restriction. As when the parent of Y, i.e. $X = x$, is conditioned on or joint with Y, it implies that

$$X = x, Y = y \equiv X = x, Y(x) = y \tag{192}$$

This equation states that if a unit is observed to receive treatment $X = x$, then its observed outcome Y must coincide with the counterfactual outcome $Y(x)$ under the same treatment. This is the same as the consistency restriction in its more commonly used form

$$X = x \implies Y = Y_x. \tag{193}$$

In addition to the exclusion and consistency restrictions, distributions in the FFRCISTG models also follow the independence assumption below.

Definition 36 (FFRCISTG Independence Assumption [25, Def. 2])**.** *For every value assignment $\mathbf{v}' \in Val(\mathbf{V}')$, the variables $\{V(\mathbf{pa}_V) \mid V \in \mathbf{V}', \mathbf{pa}_V = \mathbf{Pa}_V \cap \mathbf{v}'\}$ are mutually independent.*

We note that if all variables in $\mathbf{V}$ are observed, this independence assumption coincides with the independence restriction stated in condition (i) of the CBN2.25 definition (Def. 13). Importantly, these independence assumptions are defined only for counterfactual variables that share consistent value assignments in their brackets, reflecting an implicit restriction on the set of counterfactual variables that can be combined in FFRCISTG models.

Example 32. *Consider the FFRCISTG diagram $\mathcal{G}^F$ in Fig. 15 (i). The independence assumption based on Def. 36 is*

$$X \perp\!\!\!\perp Y(x), \forall x \in Val(X) \tag{194}$$

However, the assumption $X \perp\!\!\!\perp \{Y(x), Y(x')\}$ is not implied by the definition given the inconsistent values x and x' in the brackets.

[13] There is a subtle difference in how the model is altered to obtain a corresponding submodel: SCMs modify the functions associated with the intervened variables by replacing them with constant values; in contrast, SEM-RRs as proposed by RR replace the arguments corresponding to the intervened variables in the functions of their children. Despite this procedural difference, the two approaches are equivalent with respect to results considered in the context here.

Like exclusion and consistency, the independence restriction in Def. 36 constitutes part of the causal assumptions obtained from the qualitative coarsening of counterfactual distributions. These three assumptions define a compatibility relation between the FFRCISTG diagram and the set of distributions. In other words, Defs. 35 and 36, taken together, formally specify the FFRCISTG model. To facilitate comparison with the other graphical models discussed earlier, we provide an alternative definition of the FFRCISTG model, which is more explicit and aligned with typical compatibility relations.

Definition 37 (FFRCISTG)**.** *Given an FFRCISTG diagram, $\mathcal{G}^F$, and let $\mathbf{P}^F$ be the collection of all distributions on counterfactual variables defined in the set of variables $\mathbf{V}'$. $\mathcal{G}^F$ is an FFRCISTG for $\mathbf{P}^F$ if:*

(i) [Independence Restrictions] For a fixed intervention value set $\mathbf{v}'$ in $Val(\mathbf{V}')$ and a subset of variables $\mathbf{W} \subseteq \mathbf{V}'$. Let $\mathbf{W}_*$ be the set of counterfactuals of the form $W(\mathbf{pa}_w)$ with $\mathbf{pa}_w$ taking values in $\mathbf{v}'$. Then $P(\mathbf{W}_*)$ factorizes as:

$$P\Big(\bigwedge_{W(pa_w)\in\mathbf{W}_*} W(pa_w)\Big) = \prod_{W(pa_w)\in\mathbf{W}_*} P(W(pa_w)) \tag{195}$$

(ii) [Exclusion Restrictions] For every variable $Y \in \mathbf{V}'$ with parents $\mathbf{Pa}_y$, for every set $\mathbf{Z} \subseteq \mathbf{V}'\backslash(\mathbf{Pa}_y \cup \{Y\})$, and any counterfactual set $\mathbf{W}_*$:

$$P(Y(\mathbf{pa}_y, \mathbf{z}), \mathbf{W}_*) = P(Y(\mathbf{pa}_y), \mathbf{W}_*) \tag{196}$$

(iii) [Consistency Restrictions] For every variable $Y \in \mathbf{V}'$ with parents $\mathbf{Pa}_y$, $\mathbf{X} \subseteq \mathbf{Pa}_y$, for every set $\mathbf{Z} \subseteq \mathbf{V}'\backslash(\mathbf{X} \cup \{Y\})$, and any counterfactual set $\mathbf{W}_*$:

$$P(Y(\mathbf{z}) = y, \mathbf{X}(\mathbf{z}) = \mathbf{x}, \mathbf{W}_*) = P(Y(\mathbf{xz}) = y, \mathbf{X}(\mathbf{z}) = \mathbf{x}, \mathbf{W}_*) \tag{197}$$

With the formal definitions of how an SEM-RR induces both an FFRCISTG diagram $\mathcal{G}^F$ and a corresponding set of counterfactual distributions $\mathbf{P}^F$, it is natural to expect that the pair $\langle \mathcal{G}^F, \mathbf{P}^F \rangle$ form an FFRCISTG, just as models in the SCM framework do. However, as the following example demonstrates, this expectation does not always hold.

Example 33 (SEM-RR not inducing FFRCISTG)**.** *Consider the SEM-RR $\mathcal{N}^{(2)}$ from Example 30, and the FFRCISTG diagram it induces in Fig. 15 (f). By the independence restriction, the FFRCISTG diagram encodes the constraint $P(X, Y_x) = P(X)P(Y_x)$. However, by evaluating the SEM-RR, it can be checked that $P(X = 1, Y_{X=1} = 1) = 0.28$ while $P(X = 1)P(Y_{X=1} = 1) = 0.4 \times (0.6 \times 0.3 + 0.4 \times 0.7) = 0.184$. This inequality stems from the correlation of the error terms, which are not captured in either the FFRCISTG diagram or the definition of the independence restriction. Therefore, the FFRCISTG diagram induced by $\mathcal{N}^{(2)}$ following the constructive procedure in Def. 34 does not form an FFRCISTG model with its induced distributions.*

In fact, for an SEM-RR to induce an FFRCISTG model following the diagram construction procedure in Def. 34, it must satisfy some extra constraints in its distribution of the error terms.

Proposition 1 (FFRCISTG-Connection – SEM-RR-FFRCISTG)**.** *The FFRCISTG diagram $\mathcal{G}^F$ induced by an SEM-RR $\mathcal{N}$ following the constructive procedure in Def. 34 is an FFRCISTG for the collection of counterfactual distributions $\mathbf{P}^F$ induced by $\mathcal{N}$ if and only if $\mathcal{N}$ satisfies:*

- *For every pair $X, Y \in \mathbf{V}'$, and all value assignments $\mathbf{pa}_X, \mathbf{pa}_Y \subseteq Val(\mathbf{V}')$:*

$$\begin{aligned}\sum_{\varepsilon_X,\varepsilon_Y} \mathbb{I}[X(\mathbf{pa}_X, \varepsilon_X) = x \wedge Y(\mathbf{pa}_Y, \varepsilon_X) = y] P(\varepsilon_X, \varepsilon_Y) \\ = \Big[\sum_{\varepsilon_X} \mathbb{I}[X(\mathbf{pa}_X, \varepsilon_X)] P(\varepsilon_X)\Big] \Big[\sum_{\varepsilon_Y} \mathbb{I}[Y(\mathbf{pa}_Y, \varepsilon_X) = y] P(\varepsilon_Y)\Big]\end{aligned} \tag{198}$$

We denote the collection of SEM-RRs satisfying this condition as SEM-RR^F.

This restriction ensures mutual independence among counterfactual variables of the form $X(\mathbf{pa}_X)$ and $Y(\mathbf{pa}_Y)$, as defined in Def. 36. When all error terms in the model are mutually independent, this condition is automatically satisfied. However, it can also hold in cases where some error terms are

correlated, provided that the parameters of the SEM-RR yield exact cancellations that satisfy Eq. 198. Such cancellations are considered *unstable* because small perturbations in the parameters can break the equality, causing the constraint to disappear.

To formally distinguish between constraints that stem from the structural features of a model and those that arise from specific, potentially fragile parameterizations of functions or error distributions, consider the following definition: [14]

Definition 38 (Structurally Stable Constraints)**.** *Let $\mathcal{C}(\mathbf{P})$ denotes the set of all invariance constraints embodied in $\mathbf{P}$. An SEM-RR $\mathcal{N}$ generates a set of structurally stable constraints $\mathcal{C}(\mathbf{P}^F)$ if and only if its induced distributions $\mathbf{P}^F$ contain no extraneous independences – that is, if and only if $\mathcal{C}(\mathbf{P}^F) \subseteq \mathcal{C}(\mathbf{P}^{F'})$ for all possible parametrizations of $\mathcal{N}$ that induces $\mathbf{P}^{F'}$.*

In other words, a constraint is considered *structurally stable* if it is robust to parameter changes and remains valid as long as the structure of causal connections and error dependencies in the model is preserved.

Example 34 (SEM-RR^F inducing FFRCISTG)**.** *Consider the two SEM-RRs below.*

$\mathcal{N}^{(3)} = \langle \mathbf{V}' = \{X, Y\}, \varepsilon_{\mathbf{V}'} = \{\varepsilon_X, \varepsilon_Y\}, \mathcal{F}, P(\varepsilon_{\mathbf{V}'})\rangle$*, where*

$$\mathcal{F} = \begin{cases} X = \varepsilon_X; \\ Y = X \oplus \varepsilon_Y; \end{cases} \tag{199}$$

$$P(\varepsilon_{\mathbf{V}'}) : \begin{cases} \varepsilon_X \sim Ber(0.7); \\ \varepsilon_Y \sim Ber(0.5). \end{cases} \tag{200}$$

$\mathcal{N}^{(4)} = \langle \mathbf{V}' = \{X, Y\}, \varepsilon_{\mathbf{V}'} = \{\varepsilon_X^1, \varepsilon_X^2, \varepsilon_Y^1, \varepsilon_Y^2\}, \mathcal{F}, P(\varepsilon_{\mathbf{V}'})\rangle$*, where*

$$\mathcal{F} = \begin{cases} X = \varepsilon_X^1 \oplus \varepsilon_X^2; \\ Y = \begin{cases} \varepsilon_Y^1 \oplus \varepsilon_Y^2 \text{ if } X = 0; \\ \varepsilon_Y^2 \text{ if } X = 1; \end{cases} \end{cases} ; \tag{201}$$

$$P(\varepsilon_{\mathbf{V}'}) : \begin{cases} \varepsilon_X^1 = \varepsilon_Y^1 \sim Ber(0.5); \\ \varepsilon_X^2 \sim Ber(0.7); \\ \varepsilon_Y^2 \sim Ber(0.5). \end{cases} \tag{202}$$

The FFRCISTG diagram $\mathcal{G}^{F(3)}$ induced by $\mathcal{N}^{(3)}$ is Fig. 15 (i) with the variables in $\mathbf{V}'$ shown as nodes and the bidirected edge and the FFRCISTG diagram $\mathcal{G}^{F(4)}$ induced by $\mathcal{N}^{(4)}$ is Fig. 15 (f). The two diagrams are identical, since both models share the same variable set $\mathbf{V}'$ and functional dependencies among $\mathbf{V}'$.

Although the error terms in $\mathcal{N}^{(3)}$ are independent, while those in $\mathcal{N}^{(4)}$ are not due to $\varepsilon_X^1 = \varepsilon_Y^1$, it can be checked that the distributions induced by both models satisfy the independence restrictions:

$$P(X, Y_x) = P(X)P(Y_x), \forall x \in Val(X) \tag{203}$$

*This independence constraint is a structurally stable constraint in $\mathcal{N}^{(3)}$ because it stems from the functional and exogenous variable independencies. On the other hand, the same independence constraint is considered **structurally unstable** in $\mathcal{N}^{(4)}$ because it comes from the special parametrization of the correlated error terms.*

The argument for stability usually stems from the fact that strict equalities among products of parameters have zero measure (e.g., see [28, Thm. 3.2]). We can extend the same argument to SEM-RRs inducing independence constraints over counterfactual variables.

Proposition 2. *Given a set of invariance constraints $\mathcal{C}$, let $\Omega^{\mathcal{C}}$ be the set of SEM-RRs that induce $\mathbf{P}^F$ with $\mathcal{C} \subseteq \mathcal{C}(\mathbf{P}^F)$, and let $\Omega_S^{\mathcal{C}}$ be the subset in which $\mathcal{C}$ holds as structurally stable constraints. Then $\Omega^{\mathcal{C}} \setminus \Omega_S^{\mathcal{C}}$ has Lebesgue measure zero.*

[14] This extends the notion of *Stability* [20, Def. 2.4.1].

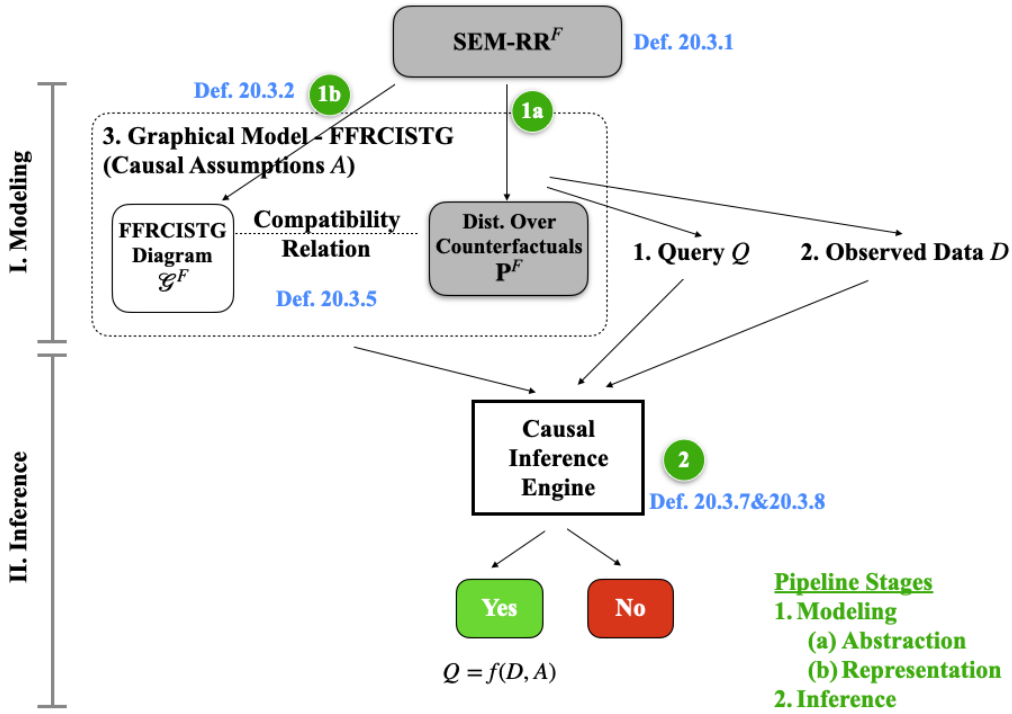


Figure 16: Causal inference pipeline for FFRCISTG models.

With the formal definitions established for the modeling stage of the causal pipeline, Fig. 16 grounds the different aspects in terms of an FFRCISTG model. Throughout the introduction of these definitions, we have noted several differences between the components of FFRCISTG models and those of other graphical models discussed earlier. We step back to reflect on these distinctions from a broader perspective and highlight additional differences that emerge when considering the modeling stage of the pipeline as an integrated whole.

First, consider the treatment of latent variables. As discussed earlier, neither the FFRCISTG diagram nor its associated counterfactual variables excludes latent variables. Consequently, the FFRCISTG model may encode constraints that involve unobserved variables, which are not empirically verifiable even in principle. In contrast, the graphical models introduced earlier in Sec. 3.2 – CBN2.25 and CBN2.5 – define constraints solely over distributions involving endogenous variables, ensuring that all encoded constraints are, at least in principle, testable through experiment. This difference has practical implications: not only does it increase the complexity of model specification, but it also introduces challenges during the inference stage, where both the query and the available data are typically restricted to observed variables. We will discuss the implications for downstream inferences in the next section.

Second, consider the distributions used to define the compatibility relations in the FFRCISTG model. In contrast to the PCH framework, which introduces a formal stratification of distributions across different layers (Thm. 4), the FFRCISTG model lacks such a formal mechanism for distinguishing among these layers. Specifically, the clauses in Def. 35 do not impose constraints on which counterfactual variables may appear jointly within a distribution. As a result, although the independence restrictions in FFRCISTG resemble those in CBN2.25, its consistency and exclusion restrictions extend beyond the $\mathcal{L}_{2.25}$ layer. For example, there is no restriction in FFRCISTG models against writing the consistency constraint over the counterfactual distributions $P(y_x, y_{x'}, x) = P(y, y_{x'}, x)$. However, these distributions fall outside $\mathcal{L}_{2.25}$ as they join two counterfactual instances of Y with different subscripts x and x', and the corresponding constraint over them are not included in CBN2.25 with the distributions in CBN2.25 circumscribed to $\mathcal{L}_{2.25}$. This lack of stratification makes it challenging to clearly situate the FFRCISTG model within a potential hierarchy of graphical models based on the strength and testability of empirical claims.

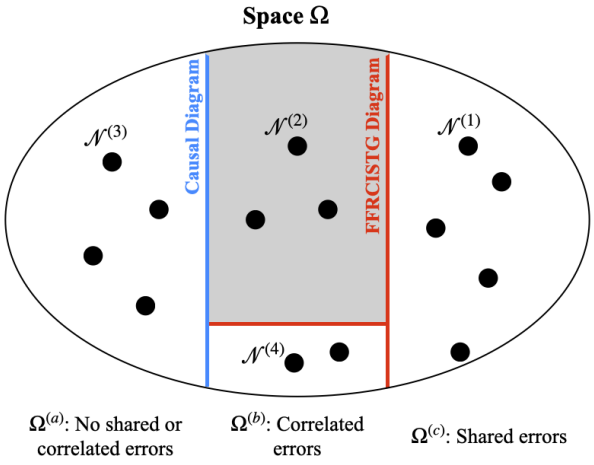


Figure 17: Visualization of different coarsening by FFRCISTG diagrams versus causal diagrams. Each black point in the space Ω represents an SCM or SEM-RR over two endogenous variables, where X is a cause of Y. Points labeled as $\mathcal{N}^{(i)}$ correspond to example SEM-RRs discussed in the text. The blue line partitions Ω according to the coarsening induced by causal diagrams: models to the left induce diagrams without a bidirected edge between X and Y, while those to the right induce diagrams with a bidirected edge, indicating latent confounding. The red line represents the coarsening induced by FFRCISTG diagrams. The gray shaded region denotes the subset of SEM-RRs that cannot induce any valid FFRCISTG diagram, highlighting that FFRCISTG models do not fully cover the space of SEM-RRs.

Third, consider how different graphical models partition the space of underlying data-generating models. Graphical models such as CBN2.25 and CBN2.5 cover the full space of SCMs, with every SCM inducing a corresponding diagram (Defs. 6, 4). By contrast, FFRCISTG models do not fully cover the space of SEM-RRs: as shown in Example 33, there exist SEM-RRs that induce no valid FFRCISTG model. Moreover, some SEM-RRs induce counterfactual independencies only through parameter-specific cancellations (e.g., correlations among error terms). Graphical models like CBN2.25 conservatively group these cases with models that contain latent confounding, while FFRCISTG groups them with models free of confounding, since it does not distinguish between structurally stable and unstable constraints. These differences are illustrated in Fig. 17: the causal diagram partition (blue) separates models by the presence of correlated errors, while the FFRCISTG partition (red) groups all models satisfying certain counterfactual independencies, regardless of whether those arise from structure or parameterization.

D.2.2 SWIGs (Inference stage)

Having formally defined the FFRCISTG graphical model, we now proceed to the inference stage and examine the machinery through which causal conclusions are derived from a combination of assumptions and data. In Fig. 16, data is shown as 'Observed Data *D*', assumptions are encoded in 'Graphical Model - FFRCISTG (Causal Assumptions *A*)', and causal conclusions of interest are represented as 'Query *Q*'.

The FFRCISTG machinery is grounded at this stage in a graphical representation called the Single-World Intervention Graph (SWIG), which is constructed from the underlying FFRCISTG diagram. The procedure for constructing a SWIG is reproduced as Alg. 2, and inference is subsequently carried out based on two key properties of the SWIG: factorization and modularity.

Definition 39 (SWIG Factorization [25, Def. 10])**.** *A joint distribution $P(\mathbf{V}'(\mathbf{a}))$ factorizes with respect to a SWIG $G^F(\mathbf{a})$ if*

$$P(\mathbf{V}'(\mathbf{a})) = \prod_{Y \in \mathbf{V}'} P(Y(\mathbf{a}_Y) \mid \mathbf{Pa}_Y(\mathbf{a}) \backslash \mathbf{A}(\mathbf{a})), \tag{204}$$

whenever the right-hand side is well-defined. Here

$$Pa_Y(\mathbf{a}) \backslash \mathbf{A}(\mathbf{a}) = \{V(\mathbf{a}_V) \mid V \in \mathbf{Pa}_Y \backslash \mathbf{A}\}. \tag{205}$$

Algorithm 2 SWIG-CONSTRUCT($\mathcal{G}^F, do(\mathbf{x})$) [25, Sec 3.3.1]

Input: FFRCISTG Diagram $\mathcal{G}^F$ and intervention set $do(\mathbf{x})$
Output: $\mathcal{G}^F(\mathbf{x})$, the SWIG constructed from $\mathcal{G}^F$ and $do(\mathbf{x})$
1: Initialise $\mathcal{G}^F(\mathbf{x})$ to be equal to $\mathcal{G}^F$
2: For each node $X \in \mathbf{X}$, split the node into a random and fixed component, labeled as X and x respectively.
3: **for** each node $X \in \mathbf{X}$ **do**
4: Split the node into a random and fixed component, labeled as X and x respectively.
5: Random component inherits all edges directed into X in $\mathcal{G}^F$
6: Fixed component inherits all edges directed out of X in $\mathcal{G}^F$
7: **end for**
8: **for** each node $V \in \mathbf{V}'$ **do**
9: update its label to $V(\mathbf{a})$, where $\mathbf{a} \subseteq \mathbf{x}$ denotes the subset of fixed nodes that are ancestors of V in $\mathcal{G}^F(\mathbf{x})$.
10: **end for**
return $\mathcal{G}^F(\mathbf{x})$.

Definition 40 (SWIG Modularity [25, Def. 15])**.** *Pairs $\langle \mathcal{G}^F, P(\mathbf{V}') \rangle$ and $\langle \mathcal{G}^F(\mathbf{a}), P(\mathbf{V}'(\mathbf{a})) \rangle$ are said to satisfy the modularity property if for every $Y \in \mathbf{V}'$,*

$$P(Y(\mathbf{a}_Y) = y \mid \mathbf{Pa}_Y(\mathbf{a}) \backslash \mathbf{A}(\mathbf{a}) = \mathbf{q}) = P(Y = y \mid \mathbf{Pa}_Y \backslash \mathbf{A} = \mathbf{Q}, \mathbf{Pa}_Y \cap \mathbf{A} = \mathbf{Pa}_Y \cap \mathbf{a}), \quad (206)$$

whenever both sides of the equation are well-defined.

Here, we note that a SWIG plays a role very similar to that of an AMWN, with the factorization property corresponding to counterfactual separation (Theorem 6). The modularity property, by contrast, combines all three types of restrictions to link distributions across different layers. Together, these two properties can be regarded as the FFRCISTG analogue of inference rules, akin in spirit to the counterfactual calculus. In particular, given a query Q, if factorization and modularity allow Q to be expressed as a function of observed data, then Q is identifiable under the assumptions.

Example 35 (Inference with a SWIG)**.** *Consider the FFRCISTG induced by the SEM-RR $\mathcal{N}^{(3)}$ in Example 34, whose associated FFRCISTG diagram $\mathcal{G}^F$ is shown in Fig. 18 (a).*

Its corresponding SWIG for intervention $do(x)$ is constructed following Alg. 2 by splitting the X node into a random component X and a fixed component x. The random component X inherits all incoming edges (which are empty in this example), while the fixed component x inherits all outgoing edges (specifically, the edge directed towards Y). Any variable $V \in \mathbf{V}'$ that has x as its ancestor in the graph (in this case, Y) will be relabeled as $V(x)$ (i.e., $Y(x)$). The output graph $\mathcal{G}^F(x)$ is shown in Fig. 18 (d).

From $\mathcal{G}^F(x)$, the constraints corresponding to factorization and modularity are:

$$P(Y_x, X) = P(Y_x)P(X) \quad (207)$$
$$P(Y_x) = P(Y|X = x) \quad (208)$$

Eq. 207 is implied by SWIG factorization with Y_x d-separated from X, and Eq. 208 is implied by SWIG modularity with X being the only parent of Y. These constraints can then be used in the inference task to check if a given query is identifiable from the data available. For example, given the query of the ETT, $Q = P(Y_x|X = x')$, and assume the available data is the observational distribution $P(\mathbf{V})$. Q can be identified by applying the constraints as follows:

$$P(Y_x|X = x') = P(Y_x) \quad (Eq.\ 207) \quad (209)$$
$$= P(Y|X = x) \quad (Eq.\ 208) \quad (210)$$

Thus, it can be concluded that the ETT is identifiable given the observational distribution with the constraints encoded in this FFRCISTG model.

The soundness and completeness of the factorization property are established in Proposition 14 and Theorem 19, while the soundness of the modularity property is established in Proposition 16 of the

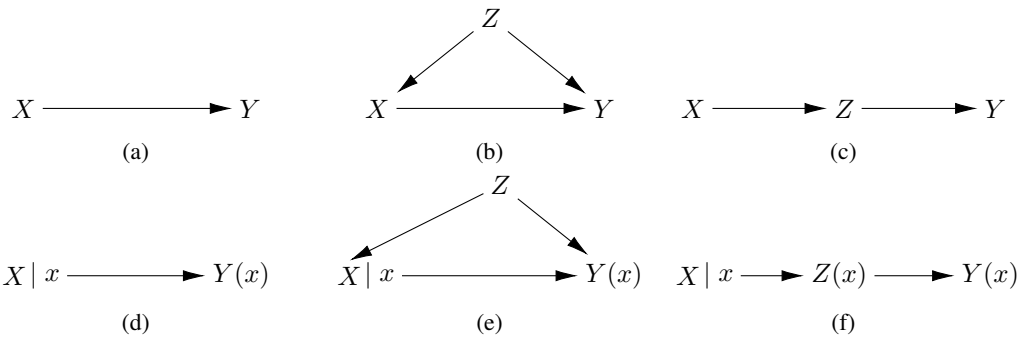


Figure 18: FFRCISTG diagrams (row 1) and their corresponding SWIGs for $do(x)$ (row 2)

original paper [25]. However, there is no result on the completeness of the modularity property. Based on these propositions, the follow corollary on the soundness of inference based on SWIGs can be derived.

Corollary 2 (Soundness for Identification with SWIGs [25, Prop. 14 and 16])**.** *A query Q is identifiable from the observational distribution and an FFRCISTG diagram $\mathcal{G}^F$, if there exists a sequence of applications of factorization and modularity properties of SWIGs constructed from $\mathcal{G}^F$ and the probability axioms that reduces Q into a function of the available distributions.*

We discuss below on several important observations about SWIGs. First, a SWIG can be used to check whether a distribution belongs to $\mathcal{L}_{2.25}$. Second, while the exclusion and consistency constraints in FFRCISTG extend beyond $\mathcal{L}_{2.25}$, the modularity property may be incomplete. Third, because the diagrams may include latent variables, these must be removed in the final ID function, which requires extra caution. Finally, for $\mathcal{L}_2$ queries that in practice are very important and pervasive across different domains and applications, no assumptions about counterfactuals are required at all, implying that SWIGs may impose unnecessary commitments. We will elaborate on each in the following.

Remark 1. SWIGs as a tool to check $\mathcal{L}_{2.25}$ membership

We note that the node splitting and edge inheritance steps in the SWIG construction algorithm force all children of the intervened variable $X = x$ to inherit the same value x. This is consistent with the restriction on the counterfactual randomization procedure to obtain distributions in $\mathcal{L}_{2.25}$. In other words, the SWIG algorithm can be applied on causal diagrams (not only FFRCISTG diagrams) and use it as a graphical tool to check if a given distribution is in $\mathcal{L}_{2.25}$, as formalized next.

Proposition 3. *Given a causal diagram $\mathcal{G}$ and a distribution $P(\mathbf{W}_*)$ over variables $\mathbf{V}$, where $\mathbf{W}_*$ is a set of counterfactuals of the form $W_{i[\mathbf{x}_i]}$. The $P(\mathbf{W}_*)$ is in $\mathbf{P}^{\mathcal{L}_{2.25}}$ induced by any SCM compatible with G if:*

(i) The subscripts $\mathbf{x}_i$ are consistent across all i, and

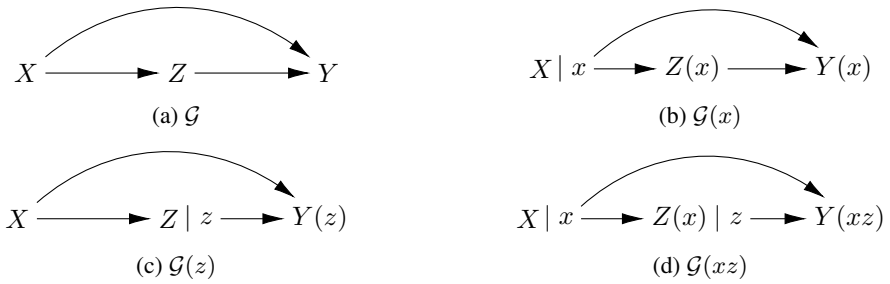
(ii) In the SWIG $(\mathcal{G}, do(\bigcup_i \mathbf{x}_i))$, $\|W_{i[\mathbf{x}_i]}\|$ appears as a node for all i.

where $\|\cdot\|$ represents the exclusion operator.

Example 36. *Consider the diagram $\mathcal{G}$ in Fig. 19 (a), and its corresponding SWIGs for different interventional sets in Fig. 19 (b-d).*

Prop. 3 can be invoked to check if the following distributions are in $\mathbf{P}^{\mathcal{L}_{2.25}}$ for SCMs compatible with $\mathcal{G}$.

- *$P(Z_x, Y_{x'}) \notin \mathbf{P}^{\mathcal{L}_{2.25}}$: Condition (i) is not satisfied due to inconsistent values for X.*
- *$P(Y_x, Y_z) \notin \mathbf{P}^{\mathcal{L}_{2.25}}$: Condition (i) is satisfied but condition (ii) is not, as the SWIG for the interventional set $\{xz\}$ (Fig. 19 (d)) does not have either nodes $\|Y_x\|, \|Y_z\|$.*
- *$P(Z, Y_x) \notin \mathbf{P}^{\mathcal{L}_{2.25}}$:Condition (i) is satisfied but condition (ii) is not, as the SWIG for the interventional set $\{x\}$ (Fig. 19 (b)) does not have node Z.*



(a) $\mathcal{G}$ (b) $\mathcal{G}(x)$ (c) $\mathcal{G}(z)$ (d) $\mathcal{G}(xz)$

Figure 19: Causal diagram G (a) and its corresponding SWIGs for different interventional sets (b,c,d)

- $P(Z, Y_z) \in \mathbf{P}^{\mathcal{L}_{2.25}}$*: Both conditions are satisfied with its SWIG for the interventional set* $\{z\}$ *(Fig. 19 (c)) containing both nodes* Z *and* Y_z*.*

This proposition suggests that the SWIG construction, along with its associated factorization and modularity properties effectively encodes constraints within $\mathcal{L}_{2.25}$, despite the absence of an explicit discussion on the layered structure of distributions in the original reference.

Remark 2. Exclusion and consistency constraints beyond layer $\mathcal{L}_{2.25}$

We note that SWIGs modularity property encodes exclusion and consistency constraints within $\mathcal{L}_{2.25}$, even though the underlying FFRCISTG model may impose these constraints beyond $\mathcal{L}_{2.25}$.

Example 37. *Consider the FFRCISTG diagram in Fig. 18 (c). Given that the exclusion restriction is not restricted to* $\mathcal{L}_{2.25}$*, it can be applied to the joint distribution* $P(Z(x), Y(zx'))$ *despite the inconsistent interventional values and obtain the constraint:*

$$P(Z(x), Y(zx')) = P(Z(x), Y(z)). \tag{211}$$

However, due to the inconsistent interventional values for X*, the counterfactuals* $Z(x), Y(zx')$ *can never appear in the same SWIG. Thus, constraints related to distributions joining them are not encoded by the modularity properties of any SWIGs.*

The example above illustrates that, without an explicit restriction against inconsistent subscripts among counterfactual variables in the exclusion and consistency conditions, FFRCISTG model may encode constraints that lie beyond what SWIGs can encode. This mismatch between the constraints encoded by the underlying FFRCISTG model and those used in its inferential SWIG can be resolved by explicitly distinguishing between different layers of distributions and restricting the encoded constraints to $\mathcal{L}_{2.25}$, as is done in the definition of CBN2.25. With such refinement, both the completeness of the modularity property and the inferential power of SWIGs could be restored.[15]

Remark 3. Explicit latent variables in constraints

We further note that the constraints derived from SWIG's factorization and modularity properties may involve latent variables if such variables are present in the underlying FFRCISTG model. However, since these latent variables are not observed, any valid identification must yield a final expression that does not include them. Thus, even though intermediate derivation steps may involve latent variables, a query is only considered identifiable if they are eliminated from the final formula.

Example 38 (Identification with latent variables)**.** *Given the query* $Q = P(y_x)$*, we check if they are identifiable given different FFRCISTG diagrams.*

[15]Further work is needed to establish a formal proof of this completeness result.

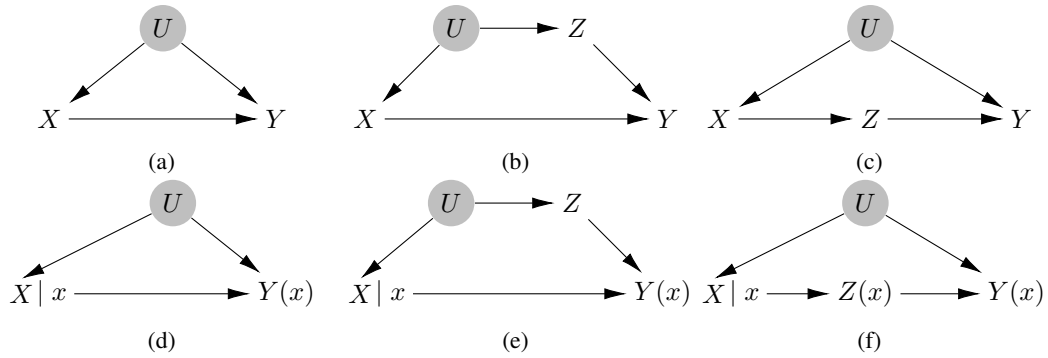


Figure 20: FFRCISTG diagrams with latent variables (row 1) and their corresponding SWIGs for $do(x)$ (row 2)

If the FFRCISTG diagram is the one shown in Fig. 20 (a), the derivation following the factorization and modularity of its SWIG in Fig. 20 (d) as follows:

$$P(y_x) = \sum_{u,x'} P(y_x, u, x') \qquad (\textit{Summing over } U, X) \tag{212}$$

$$= \sum_{u,x'} P(u)P(x'|u)P(y_x|u) \qquad (\textit{SWIG factorization}) \tag{213}$$

$$= \sum_{u,x'} P(u)P(x'|u)P(y|u,x) \qquad (\textit{SWIG modularity}) \tag{214}$$

$$= \sum_{u} P(u)P(y|u,x) \qquad (\textit{Marginalization}) \tag{215}$$

There are no further rules that can be applied to remove U from the expression, and $P(y_x)$ is not identifiable in this case.

However, if the FFRCISTG diagram is the one shown in Fig. 20 (b), the derivation following the factorization and modularity of its SWIG in Fig. 20 (e) as follows:

$$P(y_x) = \sum_{u,z,x'} P(y_x, u, z, x') \qquad (\textit{Summing over } U, Z, X) \tag{216}$$

$$= \sum_{u,z,x'} P(u)P(x'|u)P(z|u)P(y_x|z) \qquad (\textit{SWIG factorization}) \tag{217}$$

$$= \sum_{u,z,x'} P(u)P(x'|u)P(z|u)P(y|z,x) \qquad (\textit{SWIG modularity}) \tag{218}$$

$$= \sum_{u,z} P(z,u)P(y|z,x) \qquad (\textit{Marginalization}) \tag{219}$$

$$= \sum_{z} P(z)P(y|z,x) \qquad (\textit{Marginalization}) \tag{220}$$

Although there is a latent variable in the FFRCISTG diagram and its SWIG, the final formula does not include it. Thus, the query $P(y_x)$ is identifiable in this case.

This example illustrates that the identifiability status of a query does not depend solely on the existence of a latent variable, but rather on its position within the graph and the specific constraints associated with it. When compared to the structural approach, we note that graphical models like CBN, CTFBN, and their inferential machinery like *do-calculus* and *ctf-calculus*, are all defined entirely over the observed variables. This advantage in design allows the data scientist to focus solely on transforming the query from higher-layer distributions to those at lower layers, without needing to track or eliminate latent variables from the final identification formula.

Example 39 (Identification of Front-door using SWIG vs do-calculus)**.** *Given the same inferential query $Q = P(y_x)$, and considering the FFRCISTG diagram and its corresponding SWIG be the ones in Fig. 20 (c) and (f). The identification of the query using the SWIG is as follows:*

$$P(y_x) = \sum_{u,z,x'} P(y_x, u, z_x, x') \qquad (\textit{Summing over } U, Z_x, X) \tag{221}$$

$$= \sum_{u,z,x'} P(u)P(x'|u)P(z_x)P(y_x|z_x, u) \qquad (\textit{SWIG factorization}) \tag{222}$$

$$= \sum_{u,z,x'} P(u)P(x'|u)P(z|x)P(y|z, u) \qquad (\textit{SWIG modularity}) \tag{223}$$

$$= \sum_{z} P(z|x) \sum_{u,x'} P(u|x')P(x')P(y|z, u) \qquad (\textit{Bayes' Rule}) \tag{224}$$

$$= \sum_{z} P(z|x) \sum_{u,x'} P(u|x')P(x')P(y|z, u, x') \qquad (Y \perp\!\!\!\perp X|Z, U)_{G^F} \tag{225}$$

$$= \sum_{z} P(z|x) \sum_{u,x'} P(u|x', z)P(x')P(y|z, u, x') \qquad (U \perp\!\!\!\perp Z|X)_{G^F} \tag{226}$$

$$= \sum_{z} P(z|x) \sum_{x'} P(x')P(y|z, x') \qquad (\textit{Marginalization}) \tag{227}$$

If the same model is represented by a causal diagram with a bidirected edge between X and Y instead of having an explicit latent variable U, its derivation using do-calculus is as follows:

$$P(y|do(x)) = \sum_{z} P(y|do(x), z)P(z|do(x)) \qquad (\textit{Summing over } Z) \tag{228}$$

$$= \sum_{z} P(y|do(x), do(z))P(z|do(x)) \qquad (\textit{Rule 2:}(Y \perp\!\!\!\perp Z|X)_{\mathcal{G}_{\overline{X}\underline{Z}}}) \tag{229}$$

$$= \sum_{z} P(y|do(z))P(z|do(x)) \qquad (\textit{Rule 3:}(Y \perp\!\!\!\perp X|Z)_{\mathcal{G}_{\overline{X}\overline{Z}}}) \tag{230}$$

$$= \sum_{z} P(y|do(z))P(z|x) \qquad (\textit{Rule 2:}(Z \perp\!\!\!\perp X)_{\mathcal{G}_{\underline{Z}}}) \tag{231}$$

$$= \sum_{z} P(z|x) \sum_{x'} P(y|do(z), x')P(x'|do(z)) \qquad (\textit{Summing over } X) \tag{232}$$

$$= \sum_{z} P(z|x) \sum_{x'} P(y|z, x')P(x'|do(z)) \qquad (\textit{Rule 2:}(Y \perp\!\!\!\perp Z|X)_{\mathcal{G}_{\underline{Z}}}) \tag{233}$$

$$= \sum_{z} P(z|x) \sum_{x'} P(y|z, x')P(x') \qquad (\textit{Rule 3:}(X \perp\!\!\!\perp Z)_{\mathcal{G}_{\overline{Z}}}) \tag{234}$$

The same identification formula is obtained from both methods. However, the latent variable U is present in the intermediate steps in the derivation using SWIG[16] *while all steps in the derivation by do-calculus only involves the endogenous variables.*

The example above also demonstrates that identification of the ATE query, which lies within the $\mathcal{L}_2$ layer of the PCH, can be accomplished using a CBN (Def. 21) that encodes constraints solely over $\mathcal{L}_2$ distributions. In contrast, the derivation using SWIGs relies on constraints involving counterfactual distributions that lie beyond $\mathcal{L}_2$. This is certainly not aligned with the goal of parsimony and making minimal assumptions as discussed in Sec. 3.2.

Remark 4. Inferential Power of FFRCISTG vs CBN

We note that the inferential power of FFRCISTG is on par with that of CBN2.25 (Def. 13), and the counterfactual assumptions they make are unnecessary for queries in $\mathcal{L}_2$. In other words, they are not quite parsimonious for the class of identification problems over queries in $\mathcal{L}_2$. We illustrate this point using another example known as the *back-door adjustment*.

[16] Pearl's initial derivation on the front-door did evoke latent variables and the same idea.

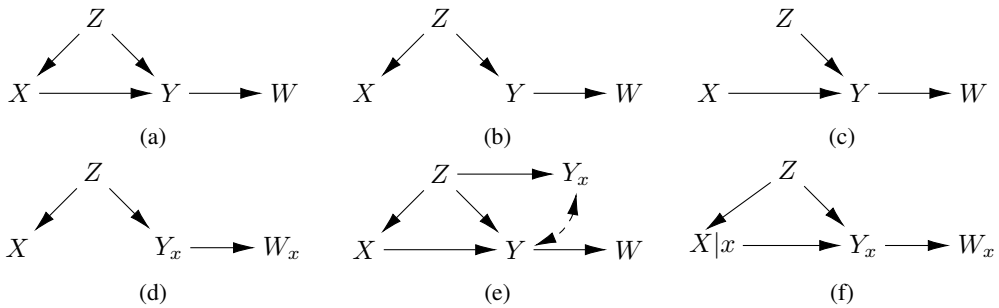


Figure 21: Causal Diagram $\mathcal{G}$ (a), its mutilated diagrams $\mathcal{G}_{\underline{X}}$ (b) and $\mathcal{G}_{\underline{X}}$ (c), its AMWNs over $\{X, Y_x, Z, W_x\}$ (d) and $\{X, Y_x, Z, W\}$ (e), and its SWIG over $do(x)$ (f).

Definition 41 (Back-door Criterion and Adjustment [20, Def.3.3.1&3.3.2])**.** *Let $\mathcal{G}$ be a causal diagram and $\mathbf{X}$ and $\mathbf{Y}$ be the sets of treatment and outcomes variables, respectively. A set of variables $\mathbf{Z}$ is said to satisfy the back-door criterion relative to the pair $(\mathbf{X}, \mathbf{Y})$ in $\mathcal{G}$ if:*

(i) No node in $\mathbf{Z}$ is a descendant of $\mathbf{X}$, and

(ii) $\mathbf{Z}$ blocks every path between $\mathbf{X}$ and $\mathbf{Y}$ that contains an arrow into $\mathbf{X}$.

If such $\mathbf{Z}$ exists, the causal effect of $\mathbf{X}$ on $\mathbf{Y}$ is identifiable and given by the expression:

$$P(\mathbf{y}|do(\mathbf{x})) = \sum_{\mathbf{z}} P(\mathbf{y}|\mathbf{x}, \mathbf{z})P(\mathbf{z}) \tag{235}$$

The back-door can be derived immediately from *do-calculus*, as Cond. (i) corresponds to rule 3 and Cond. (ii) corresponds to rule 2 of the calculus. This correspondence explains the practice of checking d-separation in mutilated diagrams when applying the back-door criterion.

$$P(\mathbf{y}|do(\mathbf{x})) = \sum_{\mathbf{z}} P(\mathbf{y}|do(\mathbf{x}), \mathbf{z})P(\mathbf{z}|do(\mathbf{x})) \tag{236}$$

$$= \sum_{\mathbf{z}} P(\mathbf{y}|do(\mathbf{x}), \mathbf{z})P(\mathbf{z}) \qquad (\textbf{Rule 3: } (\mathbf{Z} \perp\!\!\!\perp \mathbf{X})_{\mathcal{G}_{\overline{\mathbf{X}}}}) \tag{237}$$

$$= \sum_{\mathbf{z}} P(\mathbf{y}|\mathbf{x}, \mathbf{z})P(\mathbf{z}) \qquad (\textbf{Rule 2: } (\mathbf{Y} \perp\!\!\!\perp \mathbf{X} \mid \mathbf{Z})_{\mathcal{G}_{\underline{\mathbf{X}}}}) \tag{238}$$

Given that do-calculus is the inferential machinery for CBNs, such that all its rules are implied by constraints encoded by the underlying CBN, the assumptions in a CBN are sufficient to justify the use of the back-door criterion.

Example 40 (Back-door criterion)**.** *Consider the causal diagram $\mathcal{G}$ in Fig. 21 (a). The variable Z satisfies the back-door criterion relative to X and Y as it satisfies both conditions in Def. 41. So, the adjustment follows:*

$$P(y|do(x)) = \sum_{z} P(y|x, z)P(z) \tag{239}$$

However, the variable set $\{Z, W\}$ does not satisfy the back-door criterion as it violates Cond. (i) by having W as a descendant of X. Or equivalently, $W \not\perp\!\!\!\perp X$ in $\mathcal{G}_{\overline{X}}$ in Fig. 21 (c).

A similar adjustment criterion is proposed based on counterfactual independence under the FFRCISTG models by RR.

Definition 42 (Counterfactual Adjustment Criterion [25])**.** *Given $\mathbf{L} \in \mathbf{V}$, if $X \perp\!\!\!\perp Y_x \mid \mathbf{L}$ is implied by the SWIG over $do(x)$, then*

$$P(y_x) = \sum_{\mathbf{l}} P(y|x, \mathbf{l})P(\mathbf{l}) \tag{240}$$

This derivation of the back-door formula based on this criterion relies on a constraint over the counterfactual distributions, which is stronger than the original criterion in Def. 41. However, since

do-calculus is complete for identification using $\mathcal{L}_1$ data, it suffices to rely on a graphical model that encodes only $\mathcal{L}_1$ and $\mathcal{L}_2$ constraints. Therefore, applying the back-door adjustment does not require invoking a model that assumes any counterfactual relationships at all, such as the FFRCISTG model and its associated counterfactual adjustment criterion. The key concerns raised by RR against the back-door criterion are:

> The graph $\mathcal{G}_{\underline{X}}$ does not appear to offer an explanation as to why d-separation of X and Y given $\mathbf{L}$ in $\mathcal{G}_{\underline{X}}$ should ensure that Eq. 235 holds (even though it does) when X has an effect on Y [25, Page 12].
>
> The backdoor criterion requires that in addition to X and Y being d-separated given $\mathbf{L}$ in $\mathcal{G}_{\underline{X}}$, no variable in $\mathbf{L}$ may be a descendant of X. The reason for this additional condition is not transparent, since the inclusion of such a variable does not preclude that X and Y may be d-separated in $\mathcal{G}_{\underline{X}}$ [25, Page 12-13].

The reason for checking d-separation in the mutilated diagrams corresponds directly to the constraints used in the proof of the back-door adjustment formula via *do-calculus*. Specifically, the condition that 'd-separation of X and Y given $\mathbf{L}$ in $\mathcal{G}_{\underline{X}}$' licenses the constraint corresponding to rule 3 in Eq. 237, and the requirement of "no variable in $\mathbf{L}$ may be a descendant of X" licenses the constraint corresponding to rule 3 in Eq. 238. Even though the conditions in Def. 41 may involve reasoning over multiple mutilated diagrams, it still relies on weaker, non-counterfactual assumptions, which makes it a more parsimonious and preferable choice over its counterfactual counterparts.

In this section, we aligned the FFRCISTG model with the two stages of the causal pipeline, and hope this pipeline offers a clearer picture of how the model enables causal inference. We highlight a few key points from the discussion above, which we believe researchers should take note when applying FFRCISTG models to their tasks at hand.

First, it is unclear where to place FFRCISTG and SWIG within the hierarchy of the graphical models, since they encode constraints over latent variables and other non-empirically falsifiable distributions. As discussed in Sec. 4, all $\mathcal{L}_2$ queries can be answered with an CBN without invoking any counterfactual assumptions. For counterfactual queries, researchers are free to choose CBN2.25/CBN2.5 if they are more concerned with empirical falsifiability of the assumptions, or opt for CTFBN if they proritize maximal inferential power. FFRCISTG, on the other hand, mixes constraints that are empirically falsifiable and those that are not, with its explicit inclusion of latent variables as well as counterfactual distributions beyond $\mathcal{L}_{2.5}$. This makes it hard to clearly see the trade-off between empirical falsifiability and inferential power when applying FFRCISTG.

Second, the constructive procedure used to induce an FFRCISTG (Def. 34) by coarsening the underlying model creates several issues. In particular, the graph construction process does not distinguish between models that induce constraints from stable structures and those that induce the same constraints from peculiar parameterizations. We showed that this second group of models have measure zero (Prop. 2). Yet, by grouping them with other models inducing structurally stable constraints, the procedure imposes an implicit restriction over models that can induce FFRCISTGs (Prop. 1). As a result, a much larger and more substantive subset of SEM-RRs or SCMs are completely excluded from the whole pipeline as they cannot induce any FFRCISTG representations. In other words, any causal inference based on FFRCISTG becomes entirely helpless for models that fall within this subset.

We hope the discussion in this section can help researchers better understand how the FFRCISTG model compares to the other graphical models introduced in this section.

D.3 Graphical Models from DT Framework: Augmented DAG

This section examines the Decision-Theoretic (DT)[17] framework introduced by Philip Dawid [8], which we refer to as PD. The analysis relies primarily on the definitions and descriptions of augmented DAGs developed in papers from 2010 and 2024 [9, 10].

[17]The name 'decision-theoretic' comes from PD's emphasis on the decision-making aspects of causal queries, where effects of different treatments are evaluated to decide which one to take. For instance, a patient may consider and decide whether or not to take medication when having a headache. PD then used the name 'decision-theoretic' as key concepts and tools in his approach closely resemble those in standard statistical decision analysis.

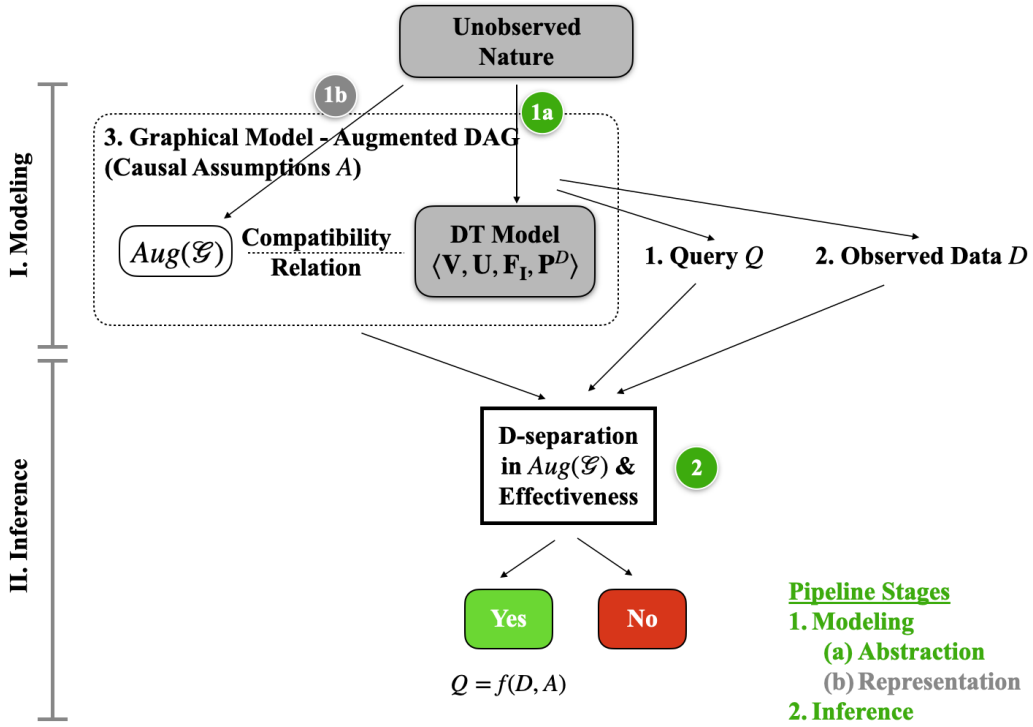


Figure 22: Causal inference pipeline for graphical models under the decision-theoretic approach. Stage 1b is shaded gray to reflect that formal definitions and procedures are unclear or missing in this stage.

D.3.1 DT Causal Models and Augmented DAGs (Modeling stage)

We start by examining the components and processes involved in the modeling stage of the causal inference pipeline as applied to augmented DAGs.

First, we note that the data-generating model in the DT approach is not a functional model like SCM or SEM-RR, but rather the abstraction starts directly at a set of distributions over 'a collection of domain variables that together describe relevant aspects of the behaviour of a system under each regime of interest'[9, Sec. 9.1]. The *domain variables* may be observed or latent, and the *regimes* describe the settings under which the distributions over these variables are defined. The *regime of interest* typically includes both the observational regime and some interventional regimes, where some variables are externally manipulated.[18] These regimes are indexed by *regime indicator variables* or *interventional variables* $\{F_i\}$, where each F_i indicates whether and how a domain variable V_i is intervened upon. Importantly, these regime indicators are not random, but similar to statistical parameters to index different probabilistic regimes under consideration [9].

To facilitate further analysis and comparison with other graphical models, we provide a formal definition putting together the key components of a causal model under PD's approach.

Definition 43 (Decision-Theoretic (DT) Causal Model)**.** *A decision-theoretic causal model, M^D, is a 4-tuple $\langle \mathbf{V}, \mathbf{U}, \mathbf{F}, \mathbf{P}^D \rangle$, where*

- $\mathbf{V}$ *is a collection of endogenous domain variables;*
- $\mathbf{U}$ *is a collection of exogenous domain variables, that together with* $\mathbf{V}$ *describe relevant aspects of the behavior of a system;*

[18] Only hard interventions are considered in this case.

- $\mathbf{F_I} = \{F_i \mid V_i \in \mathbf{I}, \mathbf{I} \subseteq \mathbf{V}$[19][20] *is a set of regime indicator variables, with the domain of each* F_i *being* $Val(V_i) \cup \{\emptyset\}$*, representing the different regimes;*
- $\mathbf{P}^D$ *is a collection of joint distributions of the form* $P(\mathbf{V}, \mathbf{U} \mid \mathbf{F_I})$*, which describes relevant aspects of the behavior of a system under each regime of interest.*

When $F_i = \emptyset$, it indicates that V_i is not intervened and take its natural value. When $F_i = \emptyset$ for all $V_i \in \mathbf{I}$, the regime corresponds to the setting with no interventions, and the resulting distribution coincides with the distribution $P(\mathbf{V}, \mathbf{U})$.[21] PD also emphasizes that domain variables represent 'real-world variables' and regime variables represent 'real-world regimes' in such models, for instance:

> For application to modeling a particular external system, we must fully understand what real-world variables are supposed represented by the domain variables in the model, and what real-world regimes by the regime variables in the model [9, Sec. 9].
>
> And to be able to approach this task in a meaningful way, we must be able to identify the unobserved explanatory variables U_P and U_Q as real-world quantities [9, Sec. 11]. [22]

To ground this definition, an example from the original work is reproduced below to illustrate these various components.

Example 41 (DT Causal Model [9, Example 9.1])**.** *Let the observed domain variables* $\mathbf{V}$ *be* $\{X, Y\}$*, and assume there is no latent domain variable, i.e.,* $\mathbf{U} = \emptyset$*. The variable of which the intervention is considered of interest* $\mathbf{I}$ *is* X *and its corresponding regime indicator is* F_X*. The collection of distributions in this model, denoted* $\mathbf{P}^D$*, includes those of the form:*

$$P(X = x, Y = y \mid F_X = f). \tag{241}$$

There are a few points worth highlighting about these distributions. First, the regimes represented using the F variables here are from hard interventions, where the intervened variables are set to a constant value by external manipulation. In fact, the same F variable representations were used in earlier papers by Pearl [19], and subsequently replaced by equivalent $do()$ operator representations. As a result, each such distribution has an equivalent representation using the $do()$ operator, by replacing each $F_i = v_i$ to $do(v_i)$.

Example 42 (Example 41 Continued)**.** *Consider the DT causal model in Example 41. Each distribution* $P(X = x, Y = y \mid F_X = f)$ *can be represented as an equivalent distribution with the* $do()$*-operator as* $P(X = x, Y = y \mid do(X = f))$ *when* $f \in Val(X)$*, or* $P(X = x, Y = y)$ *when* $f = \emptyset$*.*

Second, since domain variables may include latent variables, the joint distributions in $\mathbf{P}^D$ can also involve these unobserved components. Consequently, such distributions go beyond Layer 2 of the PCH, which only includes interventional distributions over endogenous variables. Any distribution involving latent variables is not empirically accessible, even in principle. Any causal assumptions formulated on them impose additional challenges during inference, as extra care must be given to ensure that these latent variables are either marginalized out or otherwise removed from the final expressions. These issues will be further discussed in the sections on assumptions and inference. The example below illustrates distributions involving latent variables in the DT model.

Example 43 (DT Causal Model [9, Example 9.2])**.** *Let the observed domain variables* $\mathbf{V}$ *be* $\{X, Z, Y\}$ *and latent domain variable* $\mathbf{U}$ *be* $\{U\}$*. The variable of which the intervention is considered of interest* $\mathbf{I}$ *is* X*, and its corresponding regime indicator is* F_X*. The collection of distributions*

[19] There is no explicit description of whether interventions on latent variables are allowed. Here, we assume that interventions are only meaningful for endogenous variables.

[20] Examples given by PD usually have only one F-node linked to the treatment variable. Here, we generalize this to allow multiple F-nodes.

[21] This is not the same as the observational distribution $P(\mathbf{V})$ in PCH, as it also joins the latent variables $\mathbf{U}$.

[22] This is a substantive difference as compared to the structural approach. Model construction in the structural approach can be viewed as a coarsening of the real underlying system, using knowledge that is only qualitative and not precise. This makes model building possible, given the hardness of pursuing precision and details in expert knowledge about the system.

in this model, denoted $\mathbf{P}^D$, includes those of the form:

$$P(Y = y, X = x, Z = z, U = u \mid F_X = f). \tag{242}$$

Given U is unobserved, any distribution that does not have U marginalized out, like $P(Y = y, U = u \mid F_X = \emptyset)$, are not observed.

Third, given that $\mathbf{I}$ is only a subset of domain variables $\mathbf{V}$, the collection of distributions over observed variables in $\mathbf{P}^D$ constitutes only a subset of the collection of interventional distributions in $\mathbf{P}^{\mathcal{L}_2}$.[23]

Example 44 (Example 41 Continued)**.** *Consider the causal model in Example 41. Given the same variable set $\{X, Y\}$, the collection of all interventional distributions $\mathbf{P}^{\mathcal{L}_2}$ would also include those where Y is directly intervened on. As a result, $\mathbf{P}^{\mathcal{L}_2}$ would contain distributions of the form*

$$P(X = x, Y = y | F_X = f_X, F_Y = f_y) \tag{243}$$

It follows that the collection $\mathbf{P}^D$ in this model, which includes only interventions on X (i.e., through F_X), forms a strict subset of $\mathbf{P}^{\mathcal{L}_2}$.[24]

Given the collection of distributions $\mathbf{P}^D$ in the DT causal model, further abstraction is performed to obtain qualitative constraints over these distributions. The first type of constraint resembles the *effectiveness* property of CBN [20, 2], which imposes a restriction over the regime indicators, i.e.:

$$F_X = x \implies X = x, \forall X \in \mathbf{I}. \tag{244}$$

This can be equivalently expressed as a constraint on the distributions in $\mathbf{P}^D$ as follows:

$$P(\mathbf{Z} = \mathbf{z}, X = x \mid F_X = x) = 1, \forall \mathbf{Z} \subseteq (\mathbf{V} \backslash X) \cup \mathbf{U}, \forall X \in \mathbf{I}, \forall x \in \mathit{Val}(X) \tag{245}$$

The same *effectiveness* property is imposed on the interventional distributions with the do-notation. In other words, when restricted to observed variables, Equation (245) is equivalent to:

$$P(\mathbf{Z} = \mathbf{z}, \mathbf{X} = \mathbf{x} \mid do(\mathbf{X} = \mathbf{x})) = 1, \forall \mathbf{Z} \subseteq \mathbf{V} \backslash \mathbf{X}, \forall \mathbf{X} \subseteq \mathbf{I}, \forall \mathbf{x} \in \mathit{Val}(\mathbf{X})^{25} \tag{246}$$

The second type of constraint is called *extended conditional independence* (ECI), and it resembles the *conditional independence* constraints over domain variables, by extending also to include regime indicators in $\mathbf{F_I}$. These constraints are of the form:

$$\mathbf{A} \perp\!\!\!\perp \mathbf{B}, \mathbf{F^1} \mid \mathbf{C}, \mathbf{F^2}, \tag{247}$$

where $\mathbf{A}, \mathbf{B}, \mathbf{C} \subseteq \mathbf{V} \cup \mathbf{U}$ and $\mathbf{F}^1, \mathbf{F}^2 \subseteq \mathbf{F_I}$ with $\mathbf{F}^1 \cap \mathbf{F}^2 = \emptyset$, and can be written as:

$$P(\mathbf{A} \mid \mathbf{B}, \mathbf{C}, \mathbf{F}^1, \mathbf{F}^2) = P(\mathbf{A} \mid \mathbf{C}, \mathbf{F}^2)^{26} \tag{248}$$

To ground these notions, consider the following example:

Example 45 (Example 41 Continued)**.** *Consider the causal model in Example 41.*

One example of an ECI constraint over $\mathbf{P}^D$ is

$$Y \perp\!\!\!\perp F_X \mid X, \tag{249}$$

which encodes the assumption that the distribution of Y given X is invariant, irrespective of the regimes. The equivalent expression in $\mathbf{P}^D$ is

$$\begin{aligned} P(Y = y \mid X = x, F_X = f) = P(Y = y \mid X = x, F_X = f'), \\ \forall (x, y) \in \mathit{Val}(X) \times \mathit{Val}(Y), \forall f, f' \in \mathit{Val}(X) \cup \{\emptyset\}. \end{aligned} \tag{250}$$

[23] Any distributions in $\mathbf{P}^D$ involving latent variables in $\mathbf{U}$ will not be found in $\mathbf{P}^{\mathcal{L}_2}$, since all distributions in PCH are over the observed variables.

[24] This set of representational constraints would have implications on downstream inference, as discussed further in the pipeline.

[25] Given that distributions in PCH only involve endogenous variables, $\mathbf{Z}$ comes from the endogenous set $\mathbf{V}$.

[26] The ECI constraints considered are the ones that hold for all values of the variables. Context-specific ECI constraints, like $Y \perp\!\!\!\perp X | F_X = x$, will need to be handled differently for graphical representations.

Based on the interpretation of the ECI constraints, they can be grouped into two types:

- (Type 1) Within each regime: when $\mathbf{F}^1$ is empty, the ECI implies a conditional independence constraint that holds within each regime specified under $\mathbf{F}^2$.

  For example, $Y \perp\!\!\!\perp X|F_X$ means that the conditional distribution of Y given X is the same as the marginal distribution of Y, under each regime $F_X = f$ (including the obervational regime when $f = \emptyset$). That is, $P(Y \mid X, F_X = f) = P(Y \mid F_X = f), \forall f \in Val(X) \cup \{\emptyset\}$.
- (Type 2) Cross regimes: when $\mathbf{F}^1$ is not empty, the ECI implies a conditional distribution that is invariant across all regimes labeled by $\mathbf{F}^1$, under each specific regime assignment for $\mathbf{F}^2$.

  For example, $Y \perp\!\!\!\perp F_X|X$ means that the conditional distribution of Y given X is the same across all regimes that are labeled by F_X. That is, $P(Y \mid X, F_X = f) = P(Y \mid X, F_X = f'), \forall f, f' \in Val(X) \cup \{\emptyset\}$.

Type 1 encodes assumptions within each regime, while type 2 encodes assumptions across different regimes. In fact, type 2 constraints are the ones that help bridge the gap across distributions and enable causal inference and identification (which are related to rules 2 and 3 of do-calculus).

We also highlight some ECI constraints that have special meanings:

- $Y \perp\!\!\!\perp F_X$: this is interpreted as X has *no causal effect* on Y, or $P(Y|do(x)) = P(Y)$;
- $Y \perp\!\!\!\perp F_X|X$: this is defined as the *ignorability* in [10] under the DT approach, or $P(Y|do(x)) = P(Y|x)$. [27]

These ECI constraints are not only the primary output of the abstraction stage (Stage 1a in Fig. 22), they also serve as the building blocks for the model construction stage of the DT approach (Stage 1b in Fig. 22). Unlike in the structural approach (e.g., SCM) where graphical representations (e.g., causal diagrams) are constructed by coarsening structural equations, model construction in the DT approach requires researchers to directly hypothesize a set of ECI constraints they consider valid for the system under study. A graphical representation can then be constructed from these assumptions to enable a more parsimonious encoding and efficient manipulation of causal information. PD motivates this approach with the following argument from the original reference:

> In contrast, we do not seek to impose any particular modularity requirements, nor do we require that the problem be representable by a DAG. We simply provide a language for expressing and manipulating any modularity properties that we might think it appropriate, on the basis of subject matter understanding, to impose or hypothesise.
>
> This purely formal approach does, of necessity, leave entirely untouched such essential questions as "Where do we get our causal assumptions from?" and "How can they be justified?" It is at this point, entirely removed from representational issues, that we might find a place for more informal arguments, based on intuitive understandings of cause and effect.

The first point emphasizes the flexibility of expressing assumptions without any additional restrictions, and the second highlights the benefit of separating assumption-making from justification. However, the first of these supposed advantages actually introduces several challenges in the modeling process, which will be discussed later. The second is not unique to the DT framework at all. As shown in Sec. 4, the structural approach naturally induces a hierarchy of graphical models that differ in both inferential power and empirical falsifiability, thereby already offering detailed insights into balancing assumption strength and justification. The main argumend given by PD in favor of ECI contraints over CBNs (referred to as 'Pearlian DAG' in his work) is that the assumptions encoded by CBNs are too strong and difficult to justify:

> Pearlian representability requires many strong relationships to hold between the behaviours of the system under various kinds of interventions.

[27] The comparison is subtle, but this assumption is weaker than ignorability in the traditional PO framework, $Y_x \perp\!\!\!\perp X$, which is at least as strong as an $\mathcal{L}_{2.25}$ constraint.

> In my view, the strong assumptions needed even to get started with causal interpretation of a DAG are far from self-evident as a matter of course,18 and whenever such an interpretation is proposed in a real-world context these assumptions should be carefully considered and justified. Without such justification, why should we have any faith at all in, say, the application of Pearl's causal theory, or in the output of causal discovery algorithms?

However, if all the assumptions are empirically testable, [28] in the sense that the required data are directly available, then there is no need to construct a graphical model, since the target query can be evaluated directly from the data. As discussed earlier, the core purpose of causal inference is to articulate and leverage assumptions about the underlying causal system in order to bridge the gap between the query of interest (well-defined but unobserved) and the available data. A model serves as the formal carrier of such assumptions, and the bridge it enables is only necessary when the query is not already empirically accessible. If the required data for the target query are available, then the query is already contained within the data itself, and the entire causal inference process becomes unnecessary.

Next, we elaborate on the issues caused by the lack of structure in the constraints.

Issue 1. Difficulty in Constraint Evaluation

First, the flexibility of including any constraints without a structured approach imposes significant barrier and resistance to modeling in practice. Specifically, the amount of knowledge required to articulate and assess each assumption one by one is far more detailed than is often assumed. To make this point more concrete, consider the following example.

Example 46 (ECI Constraints without Structure)**.** *Consider the models in Fig. 23 (a-d), where X is the treatment, Y is the outcome, $\{Z_1, Z_2\}$ is a set of observed confounders, and $\{U_1, U_2, U_3\}$ is a set of latent variables affecting the observed variables.*

Consider the situation where a data scientist is making ECI assumptions over the model with these variables and wants to analyze if the following independence relation holds

$$Y \perp\!\!\!\perp F_X \mid X, Z_1, Z_2. \tag{251}$$

The apparent upside of this approach is its relative simplicity – the data scientist is expected to assess the independence statement without the need to elicit more causal knowledge than necessary. However, evaluations of such independence statements are much more involved than expected if they are not supported by a collection of causal mechanisms.

Consider the tow models in Fig. 23 (a) and (b), it can be confirmed that the ECI constraint (251) holds in both, since every path between Y and F_X is blocked by X, Z_1, Z_2. This implies that the set of variables $\mathbf{Z} = \{Z_1, Z_2\}$ is a valid adjustment set for restoring the conditional ignorability condition between X on Y (i.e., $P(Y|F_X = x, Z_1, Z_2) = P(Y|X = x, Z_1, Z_2)$). It further implies that the set of variables $\mathbf{Z}$ may be treated as a block to signal that the causal relations within the set $\mathbf{Z}$ are not of importance to the data scientist, and can safely be abstracted away. The two examples (cases a, b), for both of which the ECI constraint holds, seem to confirm this intuition, since the evaluation of the independence does not depend on the existence of the latent confounder between Z_1 and Z_2 (which was absent in (a) but present in (b)).

When we shift focus to the two models in Fig. 23 (c) and (d), it can be confirmed that the ECI constraint (251) holds in (c) despite the existence of latent confounding between X, Z_1 and Y, Z_2. Interestingly, the same ECI constraint does not hold in (d) due to the existence of latent confounding between variables Z_1, Z_2. Therefore, in the second row of Fig. 23, the structure within the set of variables $\mathbf{Z}$ in fact plays a critical role in determining whether the set $\mathbf{Z}$ is valid for adjustment.

Crucially, note that for anyone who is evaluating the ECI constraint (251), they (i) would not be able to distinguish between the two columns of Fig. 23, since the set of variables $\mathbf{Z}$ is treated as a single block; and (ii) would not be able to distinguish between the two rows of Fig. 23, since the language of ECI constraints inherently does not support such distinctions. However, any time a data scientist

[28] This discussion is focused on the practical considerations, distinct from the debate on testability *in principle*, which is treated separately.

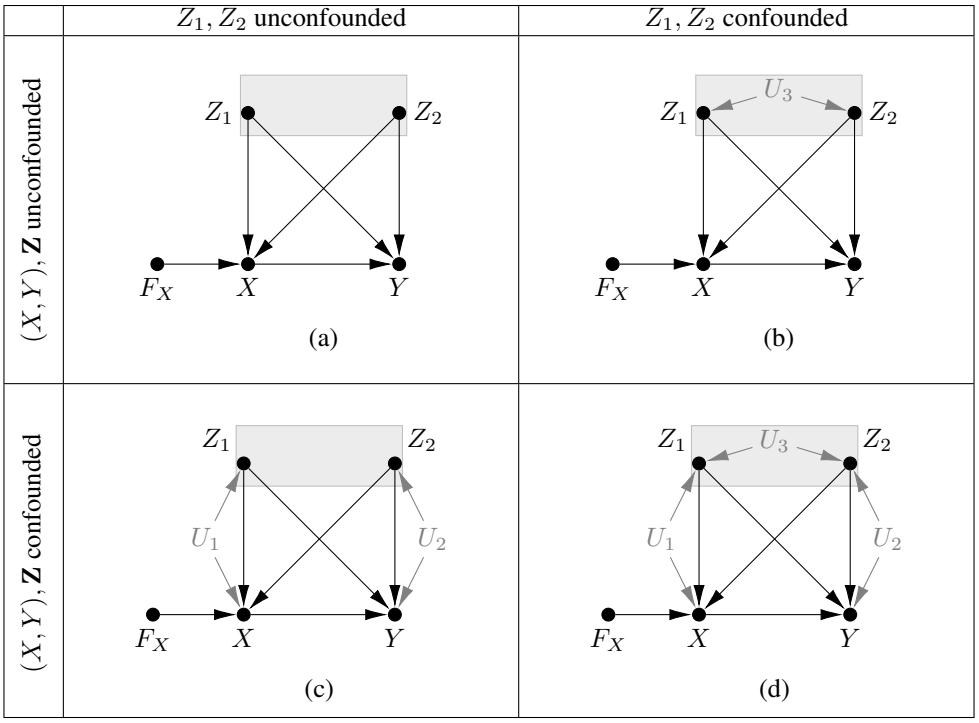


Figure 23: Augmented DAGs, where X and Y represent the treatment and the outcome, $\{Z_1, Z_2\}$ a set of observed confounders, and $\{U_1, U_2, U_3\}$ a set of latent variables affecting the observed variables. F_X is the regime indicator variable for interventions on X.

invokes the ECI constraint, they must implicitly distinguish between the four cases in Fig. 23. To argue that this happens naturally, in absence of any formal tools, is almost impossible.

From the example above, it can be seen that to make any judgment about an ECI constraint, the data scientist must have a fine-grained level of knowledge about the relationships among all variables in the conditioning set. For a set of covariates $\mathbf{Z}$ with k variables, there are $\binom{k}{2}$ pairwise relationships that must be specified. For any pair $Z_i, Z_j \in \mathbf{Z}$, there may be : (i) no relation between them; (ii) a causal relation; (iii) a confounding relation; or (iv) both a causal and a confounding relationship. This implies that the data scientist must implicitly evaluate an order of 4^{n^2} possible configurations to determine the validity of Eq. (251). In fact, for a set of two covariates, there are only four possible relationships to consider, but for a set of five covariates, this number already grows to over a million configurations. [29] This suggests that, without structural organization or systematic search algorithms, the evaluation of ECI constraints becomes considerably more complex than is commonly assumed.

Issue 2. Graphical Representation Incompatibility

Second, the lack of structural organization among constraints makes the DT model less compatible with graphical approaches to encoding assumptions. In the DT framework, the graphical object used to encode assumptions is the *augmented DAG*, where d-separations among nodes correspond to conditional independence or ECI constraints in the associated distributions. This correspondence can be formally captured by the Markov property of augmented DAGs, stated below.

Definition 44 (Global Markov Property of Augmented DAGs)**.** *Let $\mathbf{P}^D$ be a collection of distributions of the form $P(\mathbf{V}, \mathbf{U} \mid \mathbf{F_I})$, and $Aug(\mathcal{G})$ be an augmented DAG over nodes $\mathbf{V} \cup \mathbf{U} \cup \mathbf{F_I}$. $\mathbf{P}^D$ is said to be Markov relative to $Aug(\mathcal{G})$ if:*

$$\mathbf{X} \perp\!\!\!\perp_{Aug(\mathcal{G})} \mathbf{Y} \mid \mathbf{Z} \implies \mathbf{X} \perp\!\!\!\perp_{\mathbf{P}^D} \mathbf{Y} \mid \mathbf{Z}, \tag{252}$$

where $\mathbf{X} \subseteq \mathbf{V} \cup \mathbf{U}$ and $\mathbf{Y}, \mathbf{Z} \subseteq \mathbf{V} \cup \mathbf{U} \cup \mathbf{F_I}$.

[29] This example also applied to ignorability constraints in PO framework, which was discussed in details in [1, Sec. 5.7.1].

When an augmented DAG is given, the encoded ECI constraints can be read directly via the Markov property. However, when starting from an unstructured collection of ECI constraints, constructing or identifying an augmented DAG that encodes them exactly is far more subtle and involved. As noted in the original reference by PD:

> The collection of CI properties that can be represented by a DAG are very special [9, Sec. 4.1].
>
> Given a collection of conditional independence relations for a set of variables, there may be 0,1, or several DAGs that represent the identical conditional independencies [10, Sec. 2.2].

The goal here is to construct an augmented DAG whose d-separations form a perfect map for the given set of ECI constraints. Specifically, it requires a bijective correspondence between d-separations in the graph and the specified constraints. This task is closer in spirit to causal discovery, and accomplishing it implicitly requires assuming an additional property for augmented DAGs:

Definition 45 (Faithfulness in Augmented DAG)**.** *Let $\mathbf{P}^D$ be a collection of distributions of the form $P(\mathbf{V}, \mathbf{U} \mid \mathbf{F_I})$, and $Aug(\mathcal{G})$ be an augmented DAG over nodes $\mathbf{V} \cup \mathbf{U} \cup \mathbf{F_I}$. $Aug(\mathcal{G})$ is a faithful representation of $\mathbf{P}^D$ if:*

$$\mathbf{X} \perp\!\!\!\perp_{\mathbf{P}^D} \mathbf{Y} \mid \mathbf{Z} \implies \mathbf{X} \perp\!\!\!\perp_{Aug(\mathcal{G})} \mathbf{Y} \mid \mathbf{Z}. \tag{253}$$

Faithfulness, together with the Markov property, ensures that an augmented DAG encodes exactly the same set of ECI constraints posited by the researcher over $\mathbf{P}^D$. However, since the original reference provides almost no details about how augmented DAGs are to be constructed, several issues arise regarding the compatibility between DT causal models and their graphical representations. In particular, three key issues emerge: given a set of ECI constraints,

1. It is unclear whether a compatible augmented DAG exists that encodes the given set of constraints.
2. If such a DAG exists, there is no clear procedure for constructing it from the constraints.
3. If multiple compatible augmented DAGs exist, there is no guidance on how to select among them as the graphical representation.

These points will be elaborated one by one next.

Issue 2.1: No Compatible Augmented DAGs

When a data scientist is free to hypothesize any ECI constraints without structure, it is possible that the resulting assumption set is not compatible with any augmented DAG, as illustrated in the example below.

Example 47 (No Augmented DAG with Perfect Mapping)**.** *Consider a DT causal model*

$$\langle \mathbf{V} = \{X, Y, Z\}, \mathbf{U} = \emptyset, \mathbf{F_I} = \{F_X\}, \mathbf{P}^D = \{P(X, Y, Z \mid F_X\}\}\rangle. \tag{254}$$

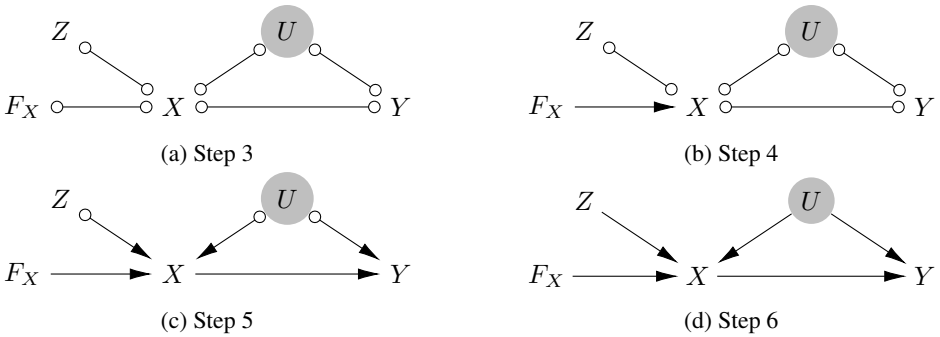
If the set of assumptions made by a data scientist only includes these two ECI constraints:

$$X \perp\!\!\!\perp Y \mid Z, F_X \tag{255}$$

$$X \perp\!\!\!\perp Y \mid F_X, \tag{256}$$

then, there is no augmented DAG that can encode these constraints exactly (i.e. a perfect map between d-separation and ECI).

First of all, any missing edge in the augmented DAG implies some additional independence constraints. For example, if there is no edge between X and Z, there must exist an independence constraint between them of the form $X \perp\!\!\!\perp Z|\mathbf{W}$ with $\mathbf{W} \subseteq \{Y, F_X\}$. However, no such ECI constraints are found in the assumption set above, and this implies that there must be edges between X and Z. The same argument applies to the edge between Y and Z. On the other hand, if an edge is present between two nodes, then these two nodes can never be separated in the graph and no independence constraints between the corresponding variables are encoded by the graph. For example, if there is an edge between X and Y, then X and Y are always d-connected and no ECI constraints over



(a) Step 3 (b) Step 4

(c) Step 5 (d) Step 6

Figure 24: Output from steps in Algorithm 3 in Example 48.

these two variables are implied. However, given that there are ECI constraints between X and Y in the assumptions set, there must be no edge between X and Y. As for the orientation of the edges, in order to encode the marginal independence between X and Y, Z should be an unshielded collider. However, this would imply that the independence conditional on Z will not hold. Hence, no augmented DAG can faithfully represent this set of assumptions.

Therefore, for a DT causal model to admit a corresponding graphical representation, it must encode ECI assumptions following some structure. However, this is not required by the current approach, where ECI constraints can be specified freely. This may explain why PD advocated using ECI constraints as the primary modeling tool, with graphical representations serving only as an optional means of organizing these constraints. As stated by Geneletti:

> Directed acyclic graphs (DAGs) are useful in the decision theoretic framework as visual representations of conditional independences and are not essential for inference [12].

However, we respectfully disagree with this perspective. As detailed throughout this chapter, graphical models play a crucial role in the pipeline, serving as parsimonious encoders of constraints and as inputs to the inferential machinery. The claim that graphical representations are optional and merely supplementary overlooks the significant benefits and efficiencies they provide. Without the structure embedded in graphical models, the space complexity would increase substantially, as constraints would need to be listed explicitly. Furthermore, the time complexity of inference would also rise, since it would require exhaustive search over all feasible graphoid axioms.[30]

Issue 2.2: No Construction Process for Augmented DAGs

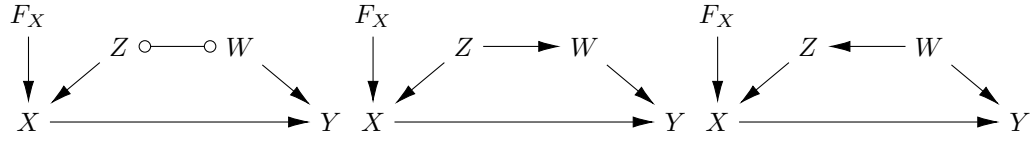
Given the goal of comparing different graphical models, we will focus on DT causal models that are compatible with augmented DAGs. Again, there is no formal definition or systematic procedure for constructing the augmented DAG, in contrast to the structural approach (Def. 6). However, to facilitate discussion and comparison, we propose an augmented DAG construction algorithm (Alg. 3, AG-CONSTRUCT), based on the examples and descriptions in [8, 9].

This algorithm works similar as the FCI algorithm in structural learning, using a constraint set as input to construct and orient edges in the graph [32]. Step 1 to 3 initializes a complete graph over all nodes in $\mathbf{V}, \mathbf{U}$, with circles as edge marks to indicate orientation uncertainty. Step 4 adds directed edges from each regime indicator variable F_i to its manipulation target V_i. The orientations of these edges are fully determined as they are always fixed in the augmented DAGs. Step 5 invokes the FCI algorithm as a subroutine to remove and orient edges using the constraint set. However, since the output from the FCI algorithm is a PAG, which represents an equivalence class of mixed graphs, some edges may remain unoriented with circle marks. Finally, Step 6 then orients any edge of the form $\circ\!\!\rightarrow$ as $\rightarrow$, given that there are no bidirected edges in augmented DAGs.

[30] Refer to Pearl's book *Probabilistic Reasoning in Intelligent Systems* for details of these axioms.

Algorithm 3 AG-CONSTRUCT($\langle \mathbf{V}, \mathbf{U}, \mathbf{F_I}, \mathbf{P}^D \rangle$)

Input: DT Causal Model $\langle \mathbf{V}, \mathbf{U}, \mathbf{F_I}, \mathbf{P}^D \rangle$
Output: $Aug(\mathcal{G})$, the augmented DAG encoding ECIs in $\mathbf{P}^D$
1: Initialize $Aug(\mathcal{G})$ with nodes for each variable in $\mathbf{V}, \mathbf{U}, \mathbf{F_I}$
2: Color each node in $\mathbf{U}$ gray to reflect their latent states
3: Add circle edges $\circ - \circ$ between nodes in $\mathbf{V}$ and $\mathbf{U}$ to form a complete graph
4: Add directed edges $\rightarrow$ from each $F_i \in \mathbf{F_I}$ to $V_i \in \mathbf{I}$
5: Run the FCI Algorithm with ECIs as the oracle to remove edges among nodes
6: Orient $\circ \rightarrow$ as $\rightarrow$
return $Aug(\mathcal{G})$.



(a) Step 6 (b) Step 7, Option 1 (c) Step 7, Option 2

Figure 25: Output from steps in Algorithm 3 in Example 49.

Example 48 (Construct Augmented DAG)**.** *Consider a DT causal model with four variables* $\mathbf{V} = \{Z, X, Y\}, \mathbf{U} = \{U\}$*, and one regime indicator* F_X*. If ECI constraints on the distributions* $\mathbf{P}^D$ *are:*

$$(U, Z) \perp\!\!\!\perp F_X \tag{257}$$
$$U \perp\!\!\!\perp Z \mid F_X \tag{258}$$
$$Y \perp\!\!\!\perp F_X \mid (X, U) \tag{259}$$
$$Y \perp\!\!\!\perp Z \mid (X, U; F_X) \tag{260}$$

The output of each intermediate step of AG-CONSTRUCT is shown in Fig. 24 (a-c), and the final augmented DAG output is shown in Fig. 24 (d).

From the example above, it can be seen that the diagram output from step 5 of AG-CONSTRUCT may still contain unoriented edges with circle marks (Fig. 24 (c)). Step 6 is invoked to orient such edges to ensure all edges are directed in the final output. In this way, the algorithm formalizes the construction of augmented DAGs and fills a key gap in the DT approach.

Issue 2.3: Multiple Compatible Augmented DAGs

However, the AG-CONSTRUCT algorithm does not work for all ECI constraint sets. This is because for some constraint sets, multiple compatible augmented DAGs may exist, as illustrated in the example below.

Example 49 (Multiple augmented DAGs encoding the same ECI constraints)**.** *Consider a DT causal model with four variables* $\mathbf{V} = \{X, W, Y, Z\}, \mathbf{U} = \emptyset$*, and one regime indicator* F_X*. If ECI constraints on the distributions* $\mathbf{P}^D$ *are:*

$$Z \perp\!\!\!\perp F_X \tag{261}$$
$$W \perp\!\!\!\perp X \mid Z, F_X \tag{262}$$
$$Y \perp\!\!\!\perp Z \mid X, W, F_X \tag{263}$$
$$Y \perp\!\!\!\perp F_X \mid X, W \tag{264}$$
$$W \perp\!\!\!\perp F_X \mid Z \tag{265}$$

The output from AG-CONSTRUCT is shown in Fig. 25 (a), which is not an augmented DAG at all due to the presence of the unoriented edge between Z, W*. This is because there are two possible orientations of this edge which are both compatible with the constraints, as shown in Fig. 25 (b) and (c).*

This example shows that when multiple augmented DAGs are compatible with the input constraints, the AG-CONSTRUCT algorithm can only output a mixed graph. The data scientist must then manually select one of the compatible DAGs to serve as the graphical model. This ambiguity in representation arises from the fact that there exists an equivalence class of augmented DAGs that encode the same ECI constraints, unlike in CBNs, where a unique causal diagram corresponds to the collection of $\mathcal{L}_2$ constraints over all interventional distributions. All augmented DAGs in such an equivalence class share the same skeleton and the same set of unshielded colliders.

As defined in the Markov property of augmented DAGs (Def. 44), the compatibility relation between the augmented DAG and the set of distributions is defined through d-separation, similar to BN. In other words, unlike in causal diagrams or FFRCISTG diagrams where arrows carry causal interpretations from functional dependencies, the arrows in the augmented DAGs serve only as a construct to support d-separation among the variables. As stated in the original reference:

> The arrows in the DAG have no intrinsic meaning, being there only to support the moralization procedure [10, Sec. 2.5].
> It is the whole structure that, with ECI, imparts causal meaning to the arrow from A to B – the direction of that arrow would not otherwise be meaningful by itself [10, Sec. 5].

More specifically, the presence of an arrow $X \to Y$ in an augmented DAG does not necessarily imply the presence of a causal influence of X on Y.

Although not explicitly defined in the original reference, there is a local counterpart to the global Markov property of augmented DAGs, analogous to that of Layer 1 BNs.

Definition 46 (Local Markov Property of Augmented DAGs)**.** *Let $\mathbf{P}^D$ be a collection of distributions of the form $P(\mathbf{V}, \mathbf{U} \mid \mathbf{F_I})$, and $Aug(\mathcal{G})$ be an augmented DAG over nodes $\mathbf{V} \cup \mathbf{U} \cup \mathbf{F_I}$. $\mathbf{P}^D$ is said to be locally Markov relative to $Aug(\mathcal{G})$ if all distributions in $\mathbf{P}^D$ factorize as:*

$$P(\mathbf{V}, \mathbf{U}|\mathbf{F_I}) = \prod_{V_i \in \mathbf{I}} P(V_i|\mathbf{Pa}_i, F_i) \prod_{V_i \in \mathbf{V} \cup \mathbf{U} \setminus \mathbf{I}} P(V_i|\mathbf{Pa}_i), \tag{266}$$

where $\mathbf{Pa}_i \subseteq \mathbf{V} \cup \mathbf{U}$ are the set of variables having a directed edge pointing towards V_i.

This local Markov property, together with the effectiveness property, encodes the collection of constraints that form the basis for inference in augmented DAGs. We therefore propose a formal definition of augmented DAGs grounded in these local properties:

Definition 47 (Local Causal Condition of Augmented DAGs)**.** *An augmented DAG $Aug(\mathcal{G})$ is said to be locally compatible with a set of distributions $\mathbf{P}^D$ of the form $P(\mathbf{V}, \mathbf{U}|\mathbf{F_I})$ if and only if the following conditions hold for every distribution in $\mathbf{P}^D$:*

(i) [*Effectiveness*] $P(V_i = v_i', \mathbf{V} \setminus V_i, \mathbf{U}|F_i = v_i, \mathbf{F_I} \setminus F_i) = 1$ *when* $v_i' = v_i$

(i) [*Markov*] $P(\mathbf{V}, \mathbf{U}|\mathbf{F_I}) = \prod_{V_i \in \mathbf{I}} P(V_i|\mathbf{Pa}_i, F_i) \prod_{V_i \in \mathbf{V} \cup \mathbf{U} \setminus \mathbf{I}} P(V_i|\mathbf{Pa}_i)$

Example 50 (Augmented DAG)**.** *Consider the augmented DAG in Fig. 24 (d). The local causal conditions it encodes are:*

(i) [*Effectiveness*] $P(X = x, Z, U, Y \mid F_X = x) = 1, \forall x \in Val(X)$

(i) [*Markov*] $P(X, Z, U, Y \mid F_X) = P(X \mid Z, U, F_X)P(Z)P(U)P(Y \mid X, U)$

The Markov conditions encode the ECI constraints:

$$Z \perp\!\!\!\perp F_X \tag{267}$$

$$U \perp\!\!\!\perp Z, F_X \tag{268}$$

$$Y \perp\!\!\!\perp Z, F_X \mid X, U \tag{269}$$

Besides the issues with augmented DAG construction, there are several important differences between augmented DAGs and CBNs, particularly in how they encode constraints:

(1) In augmented DAGs, the compatibility relation with distributions is defined over both observed and latent variables, whereas in CBNs and $\mathcal{L}_2$ distributions it is restricted to endogenous variables.

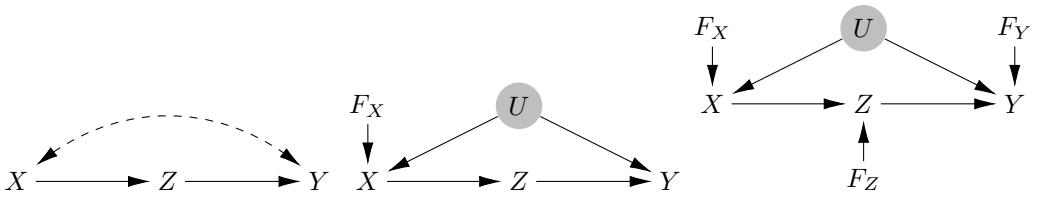


Figure 26: Causal diagram and augmented DAGs representing the front-door example.

(2) Augmented DAGs encode constraints only for regimes involving interventions on variables in $\mathbf{I}$, while CBNs, by definition, encode constraints for all interventional regimes $\mathbf{X} \subseteq \mathbf{V}$.

(3) Edges in augmented DAGs do not carry direct causal interpretations, whereas in CBNs each missing edge has a specific causal interpretation and gives rise to testable implications.

These differences will lead to further distinctions in the corresponding inference machinery, as we elaborate next.

D.3.2 Inference in Augmented DAGs (Inference stage)

In this section, we discuss the inferential machinery for DT models with compatible augmented DAGs. Given that constraints in the DT framework are encoded via d-separation relations in the augmented DAG, the inferential machinery built upon the corresponding graphical model also relies on these d-separation properties. In addition, the effectiveness properties implicitly assumed for all interventional distributions are also crucial for inference. This is illustrated in the second stage of Fig. 22. The soundness of the inference steps, and underlying machinery, are directly implied by the Markov property of the augmented DAG. However, its completeness has not been formally established. To demonstrate how constraints in augmented DAGs are utilized in the identification task, consider the example below:

Example 51 (ID in Front-Door with Augmented DAGs)**.** *Consider the augmented DAG in Fig. 26 (b) and the query $P(y|do(x))$. The query can be identified by composing ECI constraints implied by d-separation in the augmented DAG and effectiveness constraints as follows:*

$$P(y|F_X = x) \tag{270}$$

$$= \sum_z P(y \mid F_X = x, z) P(z \mid F_X = x) \qquad (\textit{Prob. Axioms}) \tag{271}$$

$$= \sum_{z,u} P(y \mid F_X = x, z, u) P(u \mid F_X = x, z) P(z \mid F_X = x) \qquad (\textit{Prob. Axioms}) \tag{272}$$

$$= \sum_{z,u} P(y \mid z, u) P(u \mid F_X = x, z) P(z \mid F_X = x) \qquad (Y \perp\!\!\!\perp F_X \mid Z, U) \tag{273}$$

$$= \sum_{z,u} P(y \mid z, u) P(u \mid F_X = x, x, z) P(z \mid F_X = x) \qquad (F_X = x \implies X = x) \tag{274}$$

$$= \sum_{z,u} P(y \mid z, u) P(u \mid F_X = x, x) P(z \mid F_X = x) \qquad (U \perp\!\!\!\perp Z \mid X, F_X) \tag{275}$$

$$= \sum_{z,u} P(y \mid z, u) P(u \mid F_X = x) P(z \mid F_X = x, x) \qquad (F_X = x \implies X = x) \tag{276}$$

$$= \sum_{z,u} P(y \mid z, u) P(u) P(z|x) \qquad (Z \perp\!\!\!\perp F_X \mid X) \tag{277}$$

$$\tag{278}$$

$$= \sum_{z,u} P(y \mid z,u)P(u)P(z|x) \qquad (U \perp\!\!\!\perp F_X) \tag{279}$$

$$= \sum_{z,u,x'} P(y \mid z,u,x')P(u \mid x')P(x')P(z \mid x) \qquad (Y \perp\!\!\!\perp X \mid Z,U) \tag{280}$$

$$= \sum_{z,u,x'} P(y \mid z,u,x')P(u \mid z,x')P(x')P(z \mid x) \qquad (U \perp\!\!\!\perp Z \mid X) \tag{281}$$

$$= \sum_{z,x'} P(y \mid z,x')P(x')P(z \mid x) \qquad (\textit{Prob. Axioms}) \tag{282}$$

From this example, we see that latent variables $\mathbf{U}$ may appear in the intermediate steps but they are removed from the final identification formula by applying independence constraints over $\mathbf{U}$. In fact, this derivation is identical to the original formulation of the front-door adjustment by Pearl, where constraints over the latent variables are also invoked explicitly. However, latent variables were then deliberately removed from both the causal diagram and the invariance constraints to achieve a more powerful symbolic machinery for identification (i.e., do-calculus). For example, the derivation of the front-door identification using do-calculus is as follows:

$$P(y|do(x)) = \sum_{z} P(y|do(x),z)P(z|do(x)) \tag{283}$$

$$= \sum_{z} P(y \mid do(x),do(z))P(z \mid do(x)) \qquad \text{Rule 2: } (Y \perp\!\!\!\perp Z \mid X)_{\mathcal{G}_{\overline{X}\underline{Z}}} \tag{284}$$

$$= \sum_{z} P(y \mid do(z))P(z \mid do(x)) \qquad \text{Rule 3: } (Y \perp\!\!\!\perp X|Z)_{\mathcal{G}_{\overline{X}\overline{Z}}} \tag{285}$$

$$= \sum_{z} P(y \mid do(z))P(z \mid x) \qquad \text{Rule 2: } (Z \perp\!\!\!\perp X)_{\mathcal{G}_{\underline{X}}} \tag{286}$$

$$= \sum_{z,x'} P(y \mid do(z),x')P(x' \mid do(z))P(z \mid x) \qquad \text{(Prob. Axioms)} \tag{287}$$

$$= \sum_{z,x'} P(y \mid z,x')P(x' \mid do(z))P(z \mid x) \qquad \text{Rule 2: } (Y \perp\!\!\!\perp Z \mid X)_{\mathcal{G}_{\underline{Z}}} \tag{288}$$

$$= \sum_{z,x'} P(y \mid z,x')P(x')P(z \mid x) \qquad \text{Rule 3: } (X \perp\!\!\!\perp Z)_{\mathcal{G}_{\overline{Z}}} \tag{289}$$

Comparing the two derivations, it can be seen that the derivation from constraints in the augmented DAG seems to operate only within the regimes captured by the variable F_X, while the derivation from do-calculus invokes constraints from both regimes $do(x)$ and $do(z)$. Nevertheless, all these constraints ultimately arise from the same graphical features in the causal diagram. To demonstrate the connection among these constraints, we first introduce the Latent Truncated Factorization Product in augmented DAGs, analogous to the theorem on causal diagrams.

Theorem 7 (Latent Truncated Factorization Product (Augmented DAG))**.** *Let $Aug(\mathcal{G})$ be an augmented DAG over $\mathbf{V}, \mathbf{U}, \mathbf{F_I}$ and let $\mathbf{X} \subseteq \mathbf{I}, \mathbf{Y} \subseteq \mathbf{V}$. The distribution $P(\mathbf{V}|\mathbf{F_X} = \mathbf{x})$ can be expressed through the latent truncated factorization product as*

$$P(\mathbf{v}|\mathbf{F_X} = \mathbf{x}) = \sum_{\mathbf{u}} \prod_{V_i \in \mathbf{V}\setminus\mathbf{X}} P(v_i|\mathbf{pa}_i, \mathbf{u}_i)p(\mathbf{u})\Big|_{\mathbf{X}=\mathbf{x}}. \tag{290}$$

Moreover, the marginal interventional distribution of $\mathbf{X}$ on $\mathbf{Y}$ is given by:

$$P(\mathbf{y}|\mathbf{F_X} = \mathbf{x}) = \sum_{\mathbf{v}\setminus\mathbf{y}\cup\mathbf{x}} \sum_{\mathbf{u}} \prod_{V_i \in \mathbf{V}\setminus\mathbf{X}} P(v_i|\mathbf{pa}_i, \mathbf{u}_i)p(\mathbf{u})\Big|_{\mathbf{X}=\mathbf{x}}. \tag{291}$$

The latent truncated factorization product (LTF product) in augmented DAGs is a direct consequence of the Markov property and the effectiveness property. It can be viewed as a shortcut in the identification process, allowing multiple independence constraints to be composed in order to eliminate all

F-nodes and express the query solely in terms of $\mathbf{V}$ and $\mathbf{U}$. To illustrate this concept, consider again the front-door example: the query $P(y \mid F_X = x)$ can be directly expressed using the LTF product as

$$P(y|F_X = x) = \sum_{u,z} P(y|z,u)P(u)P(z|x), \tag{292}$$

which exactly matches Eq. (278) in the derivation using d-separation in the augmented DAG.

Similarly, the distribution $P(y|F_X = x, z)$ in Eq. (272) can also be written using LTF product as

$$P(y|F_X = x, z) = \frac{\sum_u P(y|z,u)P(u)P(z|x)}{\sum_{uy} P(y|z,u)P(u)P(z|x)} = \sum_u P(y|z,u)P(u). \tag{293}$$

Further, if the regime indicator F_Z is added to the augmented DAG (as illustrated in Fig. 26 (c)), the distribution $P(y|F_Z = z)$ can also be written using LTF product as:

$$P(y|F_Z = z) = \sum_u P(y|u,z)P(u). \tag{294}$$

Comparing the LTF product expression for $P(y|F_X = x, z)$ and $P(y|F_Z = z)$, it can be seen that they are identical. Thus, we can replace $P(y|F_X = x, z)$ in Eq. (272) and write the query $P(y|F_x = x)$ in another equivalent expression as:

$$P(y|F_x = x) = \sum_z P(y|F_Z = z)P(z|F_X = x), \tag{295}$$

which now matches Eq. (286) in the derivation using do-calculus. This shows that both expressions are derived from the same constraints over the variables, but do-calculus abstracts away the constraints involving latent variables, replacing them with equivalent constraints in interventional distributions defined solely over the observed variables.

There are several advantages to excluding latent variables from inference. First, distributions over observed variables provide a better framework to understand the empirical falsifiability of the assumptions in the model. As discussed in Sec. 4, the graphical models also form a hierarchy in terms of the empirical falsifiability of constraints based on the feasibility of physical actions required to access distributions in the PCH. However, the same hierarchy cannot be fully realized in augmented DAGs, because latent variables render all constraints involving them experimentally inaccessible. Specifically, constraints over latent variables encoded in augmented DAGs are never directly falsifiable, since these variables cannot be observed. Yet, such constraints may imply invariance conditions in $\mathcal{L}_2$ distributions that become testable when experiments allow access to the relevant distributions. For example, in the front-door scenario, it is not possible to verify the equation $P(y|F_Z = z) = \sum_u P(y|u,z)P(u)$ due to the presence of the unobserved variable U. However, if the data scientist can intervene on X and Z, then it becomes possible to empirically test $P(y|F_Z = z) = P(y|F_X = x, z)$. Thus, modeling constraints over observed variables gives the data scientist a clearer understanding of which constraints are empirically justifiable, based on the available experimental capabilities to intervene on variables.

Second, do-calculus provide a more systematic approach to identification than the algebraic manipulation required to remove latent variables. As discussed earlier, each do-calculus rule effectively composes multiple low-level constraints over latent variables, offering a compact way to organize them. Furthermore, algorithmic identification complements do-calculus by providing a computationally efficient method for solving the identification problem [1, Sec. 4.4&5.4]. In contrast, constraints over latent variables in augmented DAGs are more difficult to handle, as there is no systematic procedure to determine whether a sequence of constraints exists that reduces a query to an expression involving only observed distributions. This difficulty is illustrated in the following example.

Example 52 (Napkin)**.** *Consider the Napkin graph in Fig. 27(a) and the query $P(y|do(x))$ or equivalently $P(y|F_X = x)$. Its identification based on do-calculus and algorithmic ID was illustrated in Chapter 4.*

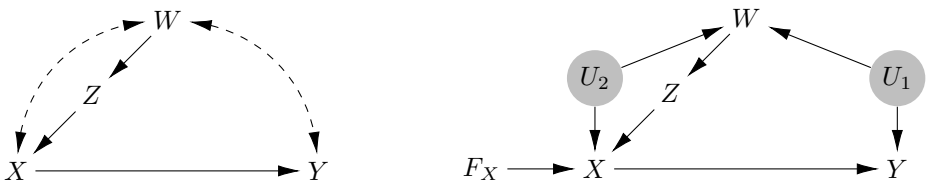


Figure 27: The Napkin graph (a) and its corresponding augmented DAG over interventions on X (b).

If we consider its corresponding augmented DAG with F_X as shown in Fig. 27(b), the identification proceeds by composing ECI and effectiveness constraints as follows:

$$P(y|F_X = x) = \sum_{u_1} P(y|u_1, F_X = x)P(u_1|F_X = x) \qquad \textit{(Prob. Axioms)} \tag{296}$$

$$= \sum_{u_1} P(y|u_1, F_X = x)P(u_1) \qquad (U_1 \perp\!\!\!\perp F_X) \tag{297}$$

$$= \sum_{u_1} P(y|u_1, x, F_X = x)P(u_1) \qquad (F_X = x \implies X = x) \tag{298}$$

$$= \sum_{u_1} P(y|u_1, x)P(u_1) \qquad (Y \perp\!\!\!\perp F_X \mid X) \tag{299}$$

$$= \frac{\sum_{u_1} P(y|u_1, x)P(u_1) \sum_{u_2} P(x|u_2, z)P(u_2)}{\sum_y \sum_{u_1} P(y|u_1, x)P(u_1) \sum_{u_2} P(x|u_2, z)P(u_2)} \qquad \textit{(Prob. Axioms)} \tag{300}$$

$$= \frac{\sum_{u_1,u_2} P(y|u_1, u_2, x, z)P(u_1) \sum_{u_2} P(x|u_2, z)P(u_2)}{\sum_y \sum_{u_1} P(y|u_1, u_2, x, z)P(u_1) \sum_{u_2} P(x|u_2, z)P(u_2)} \qquad Y \perp\!\!\!\perp Z, U_2 \mid X, U_1 \tag{301}$$

$$= \frac{\sum_{u_1,u_2} P(y|u_1, u_2, x, z)P(u_1) \sum_{u_2} P(x|u_1, u_2, z)P(u_2)}{\sum_y \sum_{u_1} P(y|u_1, u_2, x, z)P(u_1) \sum_{u_2} P(x|u_2, z)P(u_2)} \qquad X \perp\!\!\!\perp U_1 \mid Z, U_2 \tag{302}$$

$$= \frac{\sum_{u_1,u_2} P(y, x|u_1, u_2, z)P(u_1, u_2)}{\sum_y \sum_{u_1,u_2} P(y, x|u_1, u_2, z)P(u_1, u_2)} \qquad U_1 \perp\!\!\!\perp U_2 \tag{303}$$

$$= \frac{\sum_{u_1,u_2,w} P(y, x|u_1, u_2, z)P(u_1, u_2|w)P(w)}{\sum_y \sum_{u_1,u_2,w} P(y, x|u_1, u_2, z)P(u_1, u_2|w)P(w)} \qquad \textit{(Prob. Axioms)} \tag{304}$$

$$= \frac{\sum_{u_1,u_2,w} P(y, x|u_1, u_2, z, w)P(u_1, u_2|w)P(w)}{\sum_y \sum_{u_1,u_2,w} P(y, x|u_1, u_2, z, w)P(u_1, u_2|w)P(w)} \qquad Y, X \perp\!\!\!\perp W \mid U_1, U_2, Z \tag{305}$$

$$= \frac{\sum_{u_1,u_2,w} P(y, x|u_1, u_2, z, w)P(u_1, u_2|w, z)P(w)}{\sum_y \sum_{u_1,u_2,w} P(y, x|u_1, u_2, z, w)P(u_1, u_2|w, z)P(w)} \qquad U_1, U_2 \perp\!\!\!\perp Z \mid W \tag{306}$$

$$= \frac{\sum_w P(y, x|z, w)P(w)}{\sum_{y,w} P(y, x|z, w)P(w)} \qquad \textit{(Prob. Axioms)} \tag{307}$$

The step from Eq. (300) to Eq. (301) involves multiplying by a factor over the variables that equals 1. Among all possible choices, only the factor in Eq. (301) leads to the final identification of the query.

To determine this correct factor, the data scientist must enumerate all potential factors and verify whether they yield an expression over the observed distributions. In order to find this correct factor, the data scientist must enumerate all the potential factors and check if they can derive an expression over the observed distributions. This lack of systematic structure in iterating over constraints creates a significant efficiency barrier when performing inference using augmented DAGs. In contrast, the tree structures employed in the algorithmic ID approach enable much more efficient identification [1, Sec. 4.4].

For DT models where a regime indicator F_i is included for each variable $V_i \in \mathbf{V}$, the constraints encoded about $\mathbf{V}$ in the augmented DAG are equivalent to those encoded by the CBN over $\mathbf{V}$. As a result, do-calculus can be used as the inferential tool for such models. The corresponding do-calculus rules in F-notations are given as follows:

Definition 48 (Do-Calculus in Augmented DAGs)**.** *Given an augmented DAG $Aug(\mathcal{G})$ for a collection of distributions $\mathbf{P}^D$, where there is a regime indicator F_i for all $V_i \in \mathbf{V}$. Then for any disjoint sets $\mathbf{X}, \mathbf{Y}, \mathbf{Z}, \mathbf{W} \subseteq \mathbf{V}$ and disjoint sets $\mathbf{F_X}, \mathbf{F_Z} \subseteq \mathbf{F_I}$, the following three rules hold:*

Rule 1

$$P(\mathbf{y} \mid \mathbf{F_X} = \mathbf{x}, \mathbf{z}, \mathbf{w}) = P(\mathbf{y} \mid \mathbf{F_X} = \mathbf{x}, \mathbf{w}) \text{ if } \mathbf{Y} \perp\!\!\!\perp \mathbf{Z} \mid \mathbf{X}, \mathbf{F_X} \neq \emptyset, \mathbf{W} \text{ in } Aug(\mathcal{G})_{\overline{\mathbf{X}}} \tag{308}$$

Rule 2

$$P(\mathbf{y} \mid \mathbf{F_X} = \mathbf{x}, \mathbf{F_Z} = \mathbf{z}, \mathbf{w}) = P(\mathbf{y} \mid \mathbf{F_X} = \mathbf{x}, \mathbf{z}, \mathbf{w}) \text{ if } \mathbf{Y} \perp\!\!\!\perp \mathbf{F_Z} \mid \mathbf{X}, \mathbf{F_X} \neq \emptyset, \mathbf{Z}, \mathbf{W} \text{ in } Aug(\mathcal{G})_{\overline{\mathbf{X}}} \tag{309}$$

Rule 3

$$P(\mathbf{y} \mid \mathbf{F_X} = \mathbf{x}, \mathbf{F_Z} = \mathbf{z}, \mathbf{w}) = P(\mathbf{y} \mid \mathbf{F_X} = \mathbf{x}, \mathbf{w}) \text{ if } \mathbf{Y} \perp\!\!\!\perp \mathbf{F_Z} \mid \mathbf{X}, \mathbf{F_X} \neq \emptyset, \mathbf{W} \text{ in } Aug(\mathcal{G})_{\overline{\mathbf{X}}} \tag{310}$$

It can be seen from the definition above that same as the do-calculus rules in do-notation, all rules in F-notation also only involve the observed variables $\mathbf{V}$ and regime indicators $\mathbf{F_I}$, while all latent variables $\mathbf{U}$ are excluded. Another important note about these rules is that the F-nodes represent hard interventions, so graph mutilation is still required when conditioning on non-empty intervention sets. For example, in Rule 1, when $\mathbf{F_X} \neq \emptyset$ is conditioned on, the d-separation check must happen in the mutilated diagram where all incoming arrows to $\mathbf{X}$ are removed. The next example below illustrates how the do-calculus rules in F-notations can be applied to derive the same front-door expression.

Example 53 (ID in Front-Door with Augmented DAGs using do-calculus)**.** *Consider the augmented DAG in Fig. 26 (c) and the query $P(y|do(x))$. The query can be identified by composing do-calculus rules implied by the augmented DAG as follows:*

$$P(y|F_X = x) \tag{311}$$

$$= \sum_z P(y|F_X = x, z)P(z|F_X = x) \tag{312}$$

$$= \sum_z P(y \mid F_X = x, z)P(z \mid x) \qquad \textit{Rule 2: } (Z \perp\!\!\!\perp F_X|X)_{Aug(\mathcal{G})} \tag{313}$$

$$= \sum_z P(y \mid F_X = x, F_Z = z)P(z \mid x) \qquad \textit{Rule 2: } (Y \perp\!\!\!\perp F_Z|Z, X, F_X = x)_{Aug(\mathcal{G})_{\overline{X}}} \tag{314}$$

$$= \sum_z P(y \mid F_Z = z)P(z \mid x) \qquad \textit{Rule 3: } (Y \perp\!\!\!\perp F_X|Z, F_Z = z)_{Aug(\mathcal{G})_{\overline{Z}}} \tag{315}$$

$$= \sum_{z,x'} P(y \mid F_Z = z, x')P(x' \mid F_Z = z)P(z \mid x) \tag{316}$$

$$= \sum_{z,x'} P(y \mid z, x')P(x' \mid F_Z = z)P(z \mid x) \qquad \textit{Rule 2: } (Y \perp\!\!\!\perp F_Z \mid X, Z)_{Aug(\mathcal{G})} \tag{317}$$

$$= \sum_{z,x'} P(y \mid z, x')P(x')P(z \mid x) \qquad \textit{Rule 3: } (X \perp\!\!\!\perp F_Z)_{Aug(\mathcal{G})} \tag{318}$$

Unlike the derivation in Example 51, none of the steps in this derivation involve latent variables. Moreover, the derivation closely mirrors the do-calculus derivation in do-notation, as there is a one-to-one correspondence between the rules in do-notation and those in F-notation.

We hope the discussions in this section provide researchers with a deeper understanding of augmented DAGs. In particular, several key points should be kept in mind when applying augmented DAGs in practice:

- Since graphical representations are treated as merely 'visual aids' in the DT approach, the modeling process becomes considerably more complex and unstructured. Specifically, researchers must articulate and evaluate assumptions individually, including constraints that are extremely difficult to assess because they require detailed knowledge of the relationships among variables in the conditioning set.
- There is no formal constructive procedure for generating augmented DAGs. Instead, researchers must rely on alternative approaches – such as the AG-CONSTRUCT algorithm we proposed – to obtain graphical representations for their constraint sets. Moreover, some constraint sets are compatible with multiple augmented DAGs, leaving the choice of representation arbitrary and dependent on researcher preference.
- The explicit inclusion of latent variables renders it impossible to empirically verify all assumptions in the model. Unlike graphical models in the SCM framework, where empirical falsifiability depends on the feasibility of intervening on observed variables, augmented DAGs lack a well-defined hierarchy for organizing their empirical content. Researchers must therefore be prepared to shoulder the additional burden of justifying assumptions in DT models.
- Inference with augmented DAGs is substantially less systematic and efficient than with do-calculus. Researchers must not only reduce queries to distributions in the idle regime but also ensure that latent variables are eliminated from the final expression. Furthermore, no algorithmic method currently exists for solving the identification problem in augmented DAGs efficiently.

E Proofs

E.1 Supporting Lemmas

Lemma E.1 (Casual Diagram of Submodel)**.** *Given an SCM $\mathcal{M}$ and its causal diagram $\mathcal{G}$, the causal diagram induced by its submodel $\mathcal{M}_{\mathbf{x}}$ is $\mathcal{G}_{\overline{\mathbf{X}}}$, i.e., $\mathcal{G}$ with all incoming edges to $\mathbf{X}$ removed.*

Proof. By Def. 2, $\mathcal{M}_{\mathbf{x}}$ replaces f_x with $X \leftarrow x$ for all $X \in \mathbf{X}$. As a result, $\mathbf{X}$ have no endogenous or exogenous parents. By the causal diagram construction in Def. 6, edges that point to $\mathbf{X}$ are added only when $\mathbf{X}$ have parents. Thus, there is no edges incoming to $\mathbf{X}$ in the causal diagram induced by $\mathcal{M}_{\mathbf{x}}$. In addition, given that $\mathcal{M}_{\mathbf{x}}$ keeps all other components of $\mathcal{M}$ intact, all other edges remain the same. Therefore, the causal diagram induced by $\mathcal{M}_{\mathbf{x}}$ is $\mathcal{G}$ with all incoming edges to $\mathbf{X}$ removed, denoted as $\mathcal{G}_{\overline{\mathbf{X}}}$. □

Corollary 3. *Condition (ii) of Def. 11 and Def. 12 can be translated to an equivalent graphical condition:*

$\mathcal{L}_{2.25}$*: For any $v_i \in \mathbf{x}$, for all $V_j \in \mathbf{Y}$, if $V_i \in An(V_j)$ in $\mathcal{G}_{\overline{\mathbf{X} \setminus V_j}}$, then $v_i \in \mathbf{x}_j$.*

$\mathcal{L}_{2.5}$*: For any V_i, $B \in \mathbf{X} \cap \mathbf{Pa}(V_i)$, for all $V_j \in \mathbf{Y}$, if $V_i \notin \mathbf{X}_j$ and $V_i \in An(V_j)$ in $\mathcal{G}_{\overline{\mathbf{x}_j}}$, then $\mathbf{x}_i \cap B = \mathbf{x}_j \cap B$.*

Proof. It follows from Lemma E.1. □

Lemma E.2. *Given a causal diagram $\mathcal{G}$ over $\mathbf{V}$ and a set of counterfactual events $\mathbf{W}_*$, if $P(\mathbf{W}_*)$ is in $\mathbf{P}^{\mathcal{L}_{2.25}}$ of all SCMs compatible with $\mathcal{G}$, then $P(\|\mathbf{W}_*\|)$ is also in $\mathbf{P}^{\mathcal{L}_{2.25}}$ of all SCMs compatible with $\mathcal{G}$.*

Proof. If $P(\mathbf{W}_*)$ is in $\mathbf{P}^{\mathcal{L}_{2.25}}$ of all SCMs compatible with $\mathcal{G}$, it satisfies both conditions of Def. 11. We prove that after applying the exclusion operator to $\mathbf{W}_*$, the distribution still satisfies both conditions of Def. 11.

Let the set of potential outcome variables in $\mathbf{W}_*$ be denoted as $\{W_{1_{[\mathbf{t}_1]}}, ..., W_{n_{[\mathbf{t}_n]}}\}$. $P(\mathbf{W}_*)$ is indexed by the union of subscripts of all $W_{i_{\mathbf{t}_i}} \in \mathbf{W}_*$, and we denote this index by $\mathbf{t} \triangleq \bigcup_i \mathbf{t}_i$. The exclusion operator does not add subscripts to the variable, so let the new index set be the union of subscripts of all $\|W_{i_{\mathbf{t}_i}}\| \in \|\mathbf{W}_*\|$ and denote it as $\mathbf{t}' \triangleq \bigcup_i \mathbf{t}'_i$. Cond. (i) of Def. 11 still holds.

Given that $P(\mathbf{W}_*)$ also satisfies Cond. (ii) of Def. 11 and by Cor. 3, it means that whenever there is a directed path from $T \in \mathbf{T}$ to $W_i \in V[\mathbf{W}_*]$ in $G_{\overline{\mathbf{T}\setminus W}}$, t is in the subscript of W_i, i.e. $t \in \mathbf{t}_i$. Applying the exclusion operator on $W_{i_{\mathbf{t}_i}}$ removes variables in $\mathbf{t}_i$ that does not have a directed edge to W_i in $G_{\overline{\mathbf{T}_i}}$. Thus, it does not affect those that satisfy the antecedent of Cond. (ii) of Def. 11. As a result, whenever, the antecedent of Cond. (ii) of Def. 11 holds, t still belongs to the subscript of W_i. So Cond. (ii) of Def. 11 still holds.

Given that $P(\|\mathbf{W}_*\|)$ satisfies both conditions of Def. 11, it is in $\mathbf{P}^{\mathcal{L}_{2.25}}$. □

Lemma E.3. *Given a causal diagram $\mathcal{G}$ over $\mathbf{V}$ and a set of counterfactual events $\mathbf{W}_*$, if $P(\mathbf{W}_*)$ is in $\mathbf{P}^{\mathcal{L}_{2.5}}$ of all SCMs compatible with $\mathcal{G}$, then $P(\|\mathbf{W}_*\|)$ is also in $\mathbf{P}^{\mathcal{L}_{2.5}}$ of all SCMs compatible with $\mathcal{G}$.*

Proof. The proof is very similar to Lemma E.2, with the key point being that the exclusion operator on $W_{i_{\mathbf{t}_i}}$ removes variables in $\mathbf{t}_i$ that does not have a directed edge to W_i in $G_{\overline{\mathbf{T}_i}}$. Thus, it does not affect those that satisfy the antecedent of Cond. (ii) of Def. 12. □

Lemma E.4. *Given a causal diagram $\mathcal{G}$ over $\mathbf{V}$ and a set of counterfactual events $\mathbf{W}_* = \{W_{i_{[\mathbf{x}_i]}}\}$ with all subscripts taking consistent values from the same set $\mathbf{v} \in Val(\mathbf{V})$, if $\|W_{i_{[\mathbf{x}_i]}}\| = \|W_{i_{[\cup_i \mathbf{x}_i]}}\|$ for all i, then $P(\mathbf{W}_*)$ is in $\mathbf{P}^{\mathcal{L}_{2.25}}$ of all SCMs compatible with $\mathcal{G}$.*

Proof. The exclusion operator removes subscripts x from W_i if there is no directed path from X to W_i in $G_{\overline{\cup_i \mathbf{X}_i}}$. Thus, the subscripts that remain after exclusion capture precisely the cases in which the antecedent of Cond.(ii) in Definition 11 holds. If $\|W_{i_{[\mathbf{x}_i]}}\| = \|W_{i_{[\cup_i \mathbf{x}_i]}}\|$, the subscript in $\mathbf{x}_i$ accounts for all instances in $\cup_i * x_i$ that are restricted by Cond. (ii). Therefore, $P(\mathbf{W}_*)$ satisfies Def. 11 and belongs to $\mathbf{P}^{\mathcal{L}_{2.25}}$. □

Lemma E.5 (ctf-calculus — do-calculus reduction (Lemma 6 in [6]))**.** *ctf-calculus subsumes do-calculus.*

Lemma E.6 (ctf-calculus 2.25 — do-calculus reduction)**.** *ctf-calculus restricted to $\mathbf{P}^{\mathcal{L}_{2.25}}$ subsumes do-calculus.*

Proof. This result follows from the proof of Lemma E.5 where all steps in the reduction only involves quantities within $\mathbf{P}^{\mathcal{L}_{2.25}}$. □

Given a graphical model with bidirected edges, $\mathcal{G}$, the set $\mathbf{V}$ of observable variables represented as vertex can be partitioned into subsets called *c-components* [29] such that two variables belong to the same c-component if they are connected in $\mathcal{G}$ by a path made entirely of bidirected edges.

Definition 49 (Ancestral components [6])**.** *Let $\mathbf{W}_*$ be a set of counterfactual variables, $\mathbf{X}_* \subseteq \mathbf{W}_*$, and $\mathcal{G}$ be a causal diagram. Then the ancestral components induced by $\mathbf{W}_*$, given $\mathbf{X}_*$, are sets $\mathbf{A}_{1*}, \mathbf{A}_{2*}, \ldots$ that form a partition over $An\mathbf{W}_*$, made of unions of ancestral sets $An[\mathcal{G}_{\underline{\mathbf{X}_*(W_\mathbf{t})}}]W_\mathbf{t}, W_\mathbf{t} \in \mathbf{W}_*$. Sets $An[\mathcal{G}_{\underline{\mathbf{X}_*(W_{1[\mathbf{t}_1]})}}]W_{1[\mathbf{t}_1]}$ and $An[\mathcal{G}_{\underline{\mathbf{X}_*(W_{2[\mathbf{t}_2]})}}]W_{2[\mathbf{t}_2]}$ are put together if they are not disjoint or there exists a bidirected arrow in $\mathcal{G}$ connecting variables in those sets.*

Lemma E.7 (Ancestral Set Factorization (Lemma 3 in [6]))**.** *Let $\mathbf{W}_*$ be an ancestral set, that is, $An(\mathbf{W}_*) = \mathbf{W}_*$, and let $\mathbf{w}_*$ be a vector with a value for each variable in $\mathbf{W}_*$. Then,*

$$P(\mathbf{W}_* = \mathbf{w}_*) = P(\bigwedge_{W_\mathbf{t} \in \mathbf{W}_*} W_{\mathbf{pa}_w} = w) \tag{319}$$

where each w is taken from $\mathbf{w}_*$ and $\mathbf{pa}_w$ is determined for each $W_\mathbf{t} \in \mathbf{W}_*$ as follows:

(i) the values for variables in $\mathbf{Pa}_w \cap \mathbf{T}$ *are the same as in* $\mathbf{t}$*, and*

(ii) the values for variables in $\mathbf{Pa}_w \backslash \mathbf{T}$ *are taken from* $\mathbf{w}_*$ *corresponding to the parents of* $W_{\mathbf{t}}$*.*

Lemma E.8 (C-component Factorization (Lemma 4 in [6]))**.** *Let* $P(\mathbf{W}_* = \mathbf{w}_*)$ *be a distribution such that each variable in* $\mathbf{W}_*$ *has the form* $W_{\mathbf{pa}_w}$*, let* $W_1 < W_2 < \cdots$ *be a topological order over the variables in* $\mathcal{G}[\mathbf{V}(\mathbf{W}_*)]$*, and let* $\mathbf{C}_1, ..., \mathbf{C}_k$ *be the c-components of the same graph. Define* $\mathbf{C}_{j_*} = \{W_{\mathbf{pa}_w} \in \mathbf{W}_* \mid W \in \mathbf{C}_j\}$ *and* $\mathbf{c}_{j_*}$ *as the values in* $\mathbf{w}_*$ *corresponding to* $\mathbf{C}_{j_*}$*, then* $P(\mathbf{W}_* = \mathbf{w}_*)$ *decomposes as*

$$P(\mathbf{W}_* = \mathbf{w}_*) = \prod_j P(\mathbf{C}_{j_*} = \mathbf{c}_{j_*}) \tag{320}$$

Lemma E.9 (Ancestral Set in $\mathcal{L}_{2.25}/\mathcal{L}_{2.5}$)**.** $P(\mathbf{W}_*)$ *is in* $\mathcal{L}_{2.25}/\mathcal{L}_{2.5}$ *if and only if the distribution over its ancestral set* $P(An(\mathbf{W}_*))$ *is also in* $\mathcal{L}_{2.25}/\mathcal{L}_{2.5}$*.*

Proof. For $\mathcal{L}_{2.25}$, $CRS(An(\mathbf{W}_*)) = CRS(\mathbf{W}_*)$ by Def. 16; and for $\mathcal{L}_{2.5}$, $An(An(\mathbf{W}_*)) = An(\mathbf{W}_*)$ by Def. 30. Thus, $\mathbf{W}_*$ satisfies Lemma 2 if and only if $An(\mathbf{W}_*)$ satisfies Lemma 2. □

Lemma E.10 (Ancestral Set Factor in $\mathcal{L}_{2.25}/\mathcal{L}_{2.5}$)**.** *Let* $\mathbf{W}_*$ *be an ancestral set, that is,* $An(\mathbf{W}_*) = \mathbf{W}_*$*, and let* $\mathbf{w}_*$ *be a vector with a value for each variable in* $\mathbf{W}_*$*. Then,* $P(\mathbf{W}_*)$ *is in* $\mathcal{L}_{2.25}/\mathcal{L}_{2.5}$ *only if its ancestral set factor* $P(\bigwedge_{W_{\mathbf{t}} \in \mathbf{W}_*} W_{\mathbf{pa}_w} = w)$ *is in* $\mathcal{L}_{2.25}/\mathcal{L}_{2.5}$*.*

Proof. If $P(\mathbf{W}_*)$ is in $\mathcal{L}_{2.25}/\mathcal{L}_{2.5}$, then there does not exist two variables $W_{\mathbf{t}}$ and $W_{\mathbf{s}}$ in $\mathbf{W}_*$ with inconsistent subscripts. Therefore, the ancestral set factorization will also have distinct W for each $W_{\mathbf{pa}_w}$. It satisfies conditions in Def. 11/Def. 12 with consistent values from $\mathbf{w}_*$ for $\mathcal{L}_{2.25}$ and with $\mathbf{Pa}_w$ blocking all directed path from other variables to W. □

Lemma E.11 (C-component Factor in $\mathcal{L}_{2.25}/\mathcal{L}_{2.5}$)**.** *Let* $P(\mathbf{W}_* = \mathbf{w}_*)$ *be a distribution such that each variable in* $\mathbf{W}_*$ *has the form* $W_{\mathbf{pa}_w}$*, with its c-component factorization* $P(\mathbf{W}_* = \mathbf{w}_*) = \prod_j P(\mathbf{C}_{j_*} = \mathbf{c}_{j_*})$*. Then,* $P(\mathbf{W}_*)$ *is in* $\mathcal{L}_{2.25}/\mathcal{L}_{2.5}$ *only if its c-component factors* $P(\mathbf{C}_{j_*} = \mathbf{c}_{j_*})$ *are in* $\mathcal{L}_{2.25}/\mathcal{L}_{2.5}$*.*

Proof. If $P(\mathbf{W}_*)$ is in $\mathcal{L}_{2.25}/\mathcal{L}_{2.5}$, then it has distinct W for each counterfactual in the set and satisfies Def. 11/Def. 12. This property is not affected by c-component factorization as it only partitions $\mathbf{W}_*$ into subsets connected by bidirected paths. As a result, each $P(\mathbf{C}_{j_*} = \mathbf{c}_{j_*})$ will also satisfy Def. 11/Def. 12. □

Lemma E.12 (Consistency (Lemma 1 in [6]))**.** *Given SCM* $\mathcal{M}$ *and* $X, Y \in \mathbf{V}$*,* $\mathbf{T}, \mathbf{R} \subseteq \mathbf{V}$*, and let* x *be a value in the domain of* X*. Then,*

$$P(Y_{\mathbf{T}_*}, X_{\mathbf{T}_*} = x) = P(Y_{\mathbf{T}_* x}, X_{\mathbf{T}_*} = x), \tag{321}$$

where $\mathbf{T}_*$ *represent any combination of counterfactuals based on* $\mathbf{T}$*.*

Lemma E.13 (Exclusion operator (Lemma 2 in [6]))**.** *Let* $Y_{\mathbf{x}}$ *be a counterfactual variable,* $\mathcal{G}$ *a causal diagram, and*

$$Y_{\mathbf{z}} \text{ such that } \mathbf{Z} = \mathbf{X} \cap An_{\mathcal{G}_{\overline{\mathbf{X}}}}(Y) \text{ and } \mathbf{z} = \mathbf{x} \cap \mathbf{Z}. \tag{322}$$

Then, $Y_{\mathbf{z}} = Y_{\mathbf{x}}$ *holds for any model compatible with* $\mathcal{G}$*. Moreover, this transformation is denoted as* $\|(Y_{\mathbf{x}})\| := Y_{\mathbf{z}}$*.*

Lemma E.14 (Independence in $\mathcal{L}_{2.25}/\mathcal{L}_{2.5}$)**.** *Given a CBN2.25/CBN2.5, Theorem 6 is sound when the AMWN is constructed over* $\mathbf{W}_*$ *where* $P(\mathbf{W}_*)$ *is in* $\mathcal{L}_{2.25}/\mathcal{L}_{2.5}$*.*

Proof. The soundness follows from soundness of Theorem 6, where the ancestral set factorization constructed over $\{\mathbf{X}_{\mathbf{t}}, \mathbf{Y}_{\mathbf{r}}, \mathbf{Z}\}$ in the proof is also in the corresponding layers $\mathcal{L}_{2.25}/\mathcal{L}_{2.5}$ by Lemma E.9 and Lemma E.10. □

Algorithm 4 CTFIDU$(\mathbf{Y}_*, \mathbf{y}_*, \mathbb{Z}, \mathcal{G})$

Input: $\mathcal{G}$ causal diagram over variables $\mathbf{V}$; $\mathbf{Y}_*$ a set of counterfactual variables in $\mathbf{V}$; $\mathbf{y}_*$ a set of values for $\mathbf{Y}_*$; and available distribution specification $\mathbb{Z}$.
Output: $P(\mathbf{Y}_* = \mathbf{y}_*)$ in terms of available distributions or FAIL if not identifiable from $\langle \mathcal{G}, \mathbb{Z} \rangle$

let $\mathbf{Y}_* \leftarrow \|\mathbf{Y}_*\|$.
if there exists $Y_{\mathbf{x}} \in \mathbf{Y}_*$ with two or more different values in $\mathbf{y}_*(Y_{\mathbf{x}})$ or $Y_y \in \mathbf{Y}_*$ with $\mathbf{y}_*(Y_y) \neq y$ **then return** 0.
end if
if there exists $Y_{\mathbf{x}} \in \mathbf{Y}_*$ with two consistent values in $\mathbf{y}_*(Y_{\mathbf{x}})$ or $Y_y \in \mathbf{Y}_*$ with $\mathbf{y}_*(Y_y) = y$ **then** remove repeated variables from $\mathbf{Y}_*$ and values $\mathbf{y}_*$.
end if
let $\mathbf{W}_* \leftarrow An(\mathbf{Y}_*)$, and let $\mathbf{C}_{1*}, \ldots, \mathbf{C}_{k*}$ be corresponding ctf-factors in $\mathcal{G}[\mathbf{V}(\mathbf{W}_*)]$.
for each $\mathbf{C}_i$ s.t. $(\mathbf{C}_{i*} = \mathbf{c}_{i*})$ is not inconsistent, $\mathbf{Z} \in \mathbb{Z}$ s.t. $\mathbf{C}_i \cap \mathbf{Z} = \emptyset$ **do**
    let $\mathbf{B}_i$ be the c-component of $\mathcal{G}_{\overline{\mathbf{Z}}}$ such that $\mathbf{C}_i \subseteq \mathbf{B}_i$, compute $P_{\mathbf{V}\setminus\mathbf{B}_i}(\mathbf{B}_i)$ from $P_{\mathbf{Z}}(\mathbf{V})$.
    if IDENTIFY$(\mathbf{C}_i, \mathbf{B}_i, P_{\mathbf{V}\setminus\mathbf{B}_i}(\mathbf{B}_i), \mathcal{G})$ does not FAIL **then**
        let $P_{\mathbf{V}\setminus\mathbf{C}_i}(\mathbf{C}_i) \leftarrow$ IDENTIFY$(\mathbf{C}_i, \mathbf{B}_i, P_{\mathbf{V}\setminus\mathbf{B}_i}(\mathbf{B}_i), \mathcal{G})$.
        let $P(\mathbf{C}_{i*} = \mathbf{c}_{i*}) \leftarrow P_{\mathbf{V}\setminus\mathbf{C}_i}(\mathbf{C}_i)$ evaluated with values $(\mathbf{c}_{i*} \cup \bigcup_{C_{\mathbf{t}} \in \mathbf{C}_{i*}} \mathbf{pa}_c)$.
        move to the next $\mathbf{C}_i$.
    end if
end for
if any $P(\mathbf{C}_{i*} = \mathbf{c}_{i*})$ is inconsistent or was not identified from $\mathbb{Z}$ **then return** FAIL.
end if
return $P(\mathbf{Y}_* = \mathbf{y}_*) \leftarrow \sum_{\mathbf{w}_* \setminus \mathbf{y}_*} \prod_i P(\mathbf{C}_{i*} = \mathbf{c}_{i*})$.

Algorithm 5 CTFID$(\mathbf{Y}_*, \mathbf{y}_*, \mathbf{X}_*, \mathbf{x}_*, \mathbb{Z}, \mathcal{G})$

Input: $\mathcal{G}$ causal diagram over variables $\mathbf{V}$; $\mathbf{Y}_*, \mathbf{X}_*$ a set of counterfactual variables in $\mathbf{V}$; $\mathbf{y}_*, \mathbf{x}_*$ a set of values for $\mathbf{Y}_*$ and $\mathbf{X}_*$; and available distribution specification $\mathbb{Z}$.
Output: $P(\mathbf{Y}_* = \mathbf{y}_* \mid \mathbf{X}_* = \mathbf{x}_*)$ in terms of available distributions or FAIL if non-ID from $\langle \mathcal{G}, \mathbb{Z} \rangle$.

Let $\mathbf{A}_{1*}, \mathbf{A}_{2*}, \ldots$ be the ancestral components of $\mathbf{Y}_* \cup \mathbf{X}_*$ given $\mathbf{X}_*$.
Let $\mathbf{D}_*$ be the union of the ancestral components containing a variable in $\mathbf{Y}_*$ and $\mathbf{d}_*$ the corresponding set of values.
let $Q \leftarrow$ CTFIDU$(\bigcup_{D_{\mathbf{t}} \in \mathbf{D}_*} \mathbf{D}_{\mathbf{pa}_d}, \mathbf{d}_*, \mathbb{Z}, \mathcal{G})$.
return $\sum_{\mathbf{d}_* \setminus (\mathbf{y}_* \cup \mathbf{x}_*)} Q / \sum_{\mathbf{d}_* \setminus \mathbf{x}_*} Q$.

E.2 Proofs for Main Theorems

Theorem 1 ($\mathcal{L}_{2.25}$-Connection — SCM-CBN2.25)**.** *The Causal diagram $\mathcal{G}$ induced by the SCM $\mathcal{M}$ following the constructive procedure in Def. 6 is a CBN2.25 for $\mathbf{P}^{\mathcal{L}_{2.25}}$, the collection of all $\mathcal{L}_{2.25}$ distributions induced by $\mathcal{M}$.*

Proof. Let $\mathcal{M}$ be an SCM, $\mathbf{P}^{\mathcal{L}_{2.25}}$ the $\mathcal{L}_{2.25}$ distributions it induces and $\mathcal{G}$ its causal diagram. We prove that $\langle \mathcal{G}, \mathbf{P}^{\mathcal{L}_{2.25}} \rangle$ is a CBN2.25, by showing that the 3 conditions defined in Def. 13 holds in $\mathbf{P}^{\mathcal{L}_{2.25}}$ according to $\mathcal{G}$.

(Independence Restrictions) Given a potential response of the form $W_{\mathbf{pa}_w}$, its value only depends on the exogenous variables $\mathbf{U}_w$ which appear as arguments in f_W. Let $\mathbf{W}_*$ be the set of counterfactuals of the form $W_{\mathbf{pa}_w}$ with $\mathbf{pa}_w$ taking consistent values from $\mathbf{v} \in Val(\mathbf{V})$, $P(\mathbf{W}_*)$ falls in $\mathcal{L}_{2.25}$ as it satisfy conditions of Def. 11. Let $\mathbf{C}_1, ..., \mathbf{C}_l$ be the c-components of $G[\mathbf{V}(\mathbf{W}_*)]$, and $\mathbf{C}_{1*}, ..., \mathbf{C}_{l*}$ the corresponding partition over $\mathbf{W}_*$. Then the set of exogenous variables $\mathbf{U}(\mathbf{W}_*)$ can be partitioned as $\mathbf{U}(\mathbf{C}_{1*}), ..., \mathbf{U}(\mathbf{C}_{l*})$ where $\mathbf{U}(\mathbf{C}_{i*})$ and $\mathbf{U}(\mathbf{C}_{j*})$ are disjoint for all $i, j = 1, ..., l, i \neq j$, due to the absence of bidirected paths between variables in $\mathbf{C}_i$ and and variables $\mathbf{C}_j$. Then by Def. 11,

(Exclusion restrictions) Given a potential response of the form $Y_{\mathbf{pa}_y, \mathbf{z}}$, its value only depends on the exogenous variables $\mathbf{U}_y$ which appear as arguments in f_Y as $\mathbf{pa}_y$ are fixed. Thus, $Y_{\mathbf{pa}_y, \mathbf{z}}(\mathbf{u}) = Y_{\mathbf{pa}_y}(\mathbf{u})$. Then by Def. 11, for any counterfactual set $\mathbf{W}_*$ such that $P(Y_{\mathbf{pa}_y, \mathbf{z}} = y, \mathbf{W}_* = \mathbf{w}_*) \in$

$\mathbf{P}^{\mathcal{L}_{2.25}}$,

$$P(Y_{\mathbf{pa}_y,\mathbf{z}} = y, \mathbf{W}_* = \mathbf{w}_*) = \sum_{\mathbf{u}} \mathbf{1}(Y_{\mathbf{pa}_y,\mathbf{z}}(\mathbf{u}) = y, \mathbf{W}_*(\mathbf{u}) = \mathbf{w}_*)P(\mathbf{u}) \tag{323}$$

$$= \sum_{\mathbf{u}} \mathbf{1}(Y_{\mathbf{pa}_y}(\mathbf{u}) = y, \mathbf{W}_*(\mathbf{u}) = \mathbf{w}_*)P(\mathbf{u}) \tag{324}$$

$$= P(Y_{\mathbf{pa}_y} = y, \mathbf{W}_* = \mathbf{w}_*) \tag{325}$$

which proves the exclusion restrictions are satisfied.

(Consistency restrictions) Given $\mathbf{u} \in Val(\mathbf{U})$ such that $Y_{\mathbf{z}}(\mathbf{u}) = y, \mathbf{X}_{\mathbf{z}}(\mathbf{u}) = \mathbf{x}, \mathbf{W}_*(\mathbf{u}) = \mathbf{w}_*$, for some $Y \in \mathbf{V}$, $\mathbf{X} \subseteq \mathbf{Pa}_y$,$\mathbf{Z} \subseteq \mathbf{V}\backslash(\mathbf{X} \cup \{Y\})$, $\mathbf{R} = \mathbf{Pa}_y\backslash(\mathbf{X} \cup \mathbf{Z})$, we have

$$Y_{\mathbf{z}}(\mathbf{u}) = f_Y(\mathbf{z} \cap \mathbf{pa}_y, \mathbf{X}_{\mathbf{z}}(\mathbf{u}), \mathbf{R}_{\mathbf{z}}(\mathbf{u}), \mathbf{u}(\mathbf{U}_y)) \tag{326}$$

$$= f_Y(\mathbf{z} \cap \mathbf{pa}_y, \mathbf{x}, \mathbf{R}_{\mathbf{z}}(\mathbf{u}), \mathbf{u}(\mathbf{U}_y)) \tag{327}$$

$$= Y_{\mathbf{zx}}(\mathbf{u}) \tag{328}$$

Then by Def. 11, for any counterfactual set $\mathbf{W}_*$ such that $P(Y_{\mathbf{z}} = y, \mathbf{X}_{\mathbf{z}} = \mathbf{x}, \mathbf{W}_* = \mathbf{w}_*) \in \mathbf{P}^{\mathcal{L}_{2.25}}$,

$$P(Y_{\mathbf{z}} = y, \mathbf{X}_{\mathbf{z}} = \mathbf{x}, \mathbf{W}_* = \mathbf{w}_*) = \sum_{\mathbf{u}} \mathbf{1}(Y_{\mathbf{z}}(\mathbf{u}) = y, \mathbf{X}_{\mathbf{z}}(\mathbf{u}) = \mathbf{x}, \mathbf{W}_*(\mathbf{u}) = \mathbf{w}_*)P(\mathbf{u}) \tag{329}$$

$$= \sum_{\mathbf{u}} \mathbf{1}(Y_{\mathbf{xz}}(\mathbf{u}) = y, \mathbf{X}_{\mathbf{z}}(\mathbf{u}) = \mathbf{x}, \mathbf{W}_*(\mathbf{u}) = \mathbf{w}_*)P(\mathbf{u}) \tag{330}$$

$$= P(Y_{\mathbf{xz}} = y, \mathbf{X}_{\mathbf{z}} = \mathbf{x}, \mathbf{W}_* = \mathbf{w}_*) \tag{331}$$

which proves the consistency restrictions are satisfied. □

Definition 50 (Counterfactual Reachability Set)**.** *Given a graph $\mathcal{G}$ and a potential outcome $Y_{\mathbf{x}}$, the counterfactual reachability set of $Y_{\mathbf{x}}$, denoted $CRS(Y_{\mathbf{x}})$, consists of each $\|W_{\mathbf{x}}\|$ s.t. $W \in (An(Y) \cup \{De(V) : \forall V \in \mathbf{X}\})\backslash\mathbf{X}$ and $\|W_{\mathbf{x}\backslash w}\|$ s.t. $W \in (An(Y) \cup \{De(V) : \forall V \in \mathbf{X}\}) \cap \mathbf{X}$. For a set $\mathbf{W}_*$, $CRS(\mathbf{W}_*)$ is defined to be the union of the CRS of each potential outcome in the set, such that for any set of variables $\{W_{i_{[\mathbf{x}_i]}}\}_i \subseteq \mathbf{W}_*$ with their CRS set having counterfactual variables $\{R_{[\mathbf{x}_i]}\}_i$ over the same variable R, $\{R_{[\mathbf{x}_i]}\}_i$ is merged into one variable $\|R_{[\cup_i \mathbf{x}_i]}\|$ if $\|W_{i_{[\cup_i \mathbf{x}_i]}}\| = W_{i_{[\mathbf{x}_i]}}$ for all i.*

Lemma 1. *A distribution $Q = P(\mathbf{W}_*)$ is in the $\mathcal{L}_{2.25}/\mathcal{L}_{2.5}$ distributions induced by any SCM compatible with a given graph $\mathcal{G}$ if and only if the set $CRS(\mathbf{W}_*)$ satisfies (i) and (ii) / $An(\mathbf{W}_*)$ satisfies (i): (i) Does not contain any pair of potential outcomes $W_{\mathbf{s}}, W_{\mathbf{t}}$ of the same variable W under different regimes where $\mathbf{s} \neq \mathbf{t}$; (ii) $\mathbf{W}_*$ does not contain any pair of potential outcomes $R_{\mathbf{s}}, W_{\mathbf{t}}$ with inconsistent subscripts where $\mathbf{s} \cap \mathbf{T} \neq \mathbf{t} \cap \mathbf{S}$.*

Proof. Consistent values across the variables are enforced by (ii). Each CRS set corresponding to a potential outcome Y_* includes all variables that must remain consistent with Y_* under the regime $*$. When taking the union of CRS sets over multiple potential outcomes, and if the union does not contain any pair of potential outcomes $W_{\mathbf{s}}, W_{\mathbf{t}}$ for the same variable W under different regimes, then two cases arise:

(a) All CRS sets are disjoint with respect to the variables from which their potential outcomes are derived. This implies that the ancestral and descendant sets of these variables are also disjoint, so there is no directed path crossing the CRS sets in a way that would trigger the antecedent of Cond. (ii) in Definition 11.

(b) Any overlapping CRS sets must involve counterfactuals over the same variable, which are merged as $|W_{i_{[\cup_i * x_i]}}| = |W_{i_{[* x_i]}}|$ for all i. This condition implies that the variables underlying these merged CRS sets are consistent, by Lemma E.4.

Therefore, $P(\mathbf{W}_*)$ satisfies conditions in Def. 11 and belongs to $\mathbf{P}^{\mathcal{L}_{2.25}}$.

The graphical check for $\mathcal{L}_{2.5}$ is proved in Corollary 3.7 of [24]. □

Theorem 2 (Soundness and Completeness for CBN2.25/CBN2.5 Identifiability)**.** *An $\mathcal{L}_{2.25}$/$\mathcal{L}_{2.5}$ quantity Q is identifiable from a given set of observational and interventional distributions and a CBN2.25/CBN2.5 if and only if there exists a sequence of applications of the rules of ctf-calculus for CBN2.25/CBN2.5 and the probability axioms restrained within $\mathcal{L}_{2.25}$/$\mathcal{L}_{2.5}$ that reduces Q into a function of the available distributions.*

Proof. The soundness of the calculus for $\mathcal{L}_{2.25}$/$\mathcal{L}_{2.5}$ follows from the soundness of the ctf-calculus rules. The soundness of the ctf-calculus rules in turn follows from Lemma E.12 for Rule 1, Lemma E.14 for Rule 2 and Lemma E.13 for Rule 3.

To prove that it is complete, we rely on the completeness of the CTFID algorithm reproduced as Algo. 5 and Algo. 4 [7]. Specifically, we show that if the query is in $\mathcal{L}_{2.25}$/$\mathcal{L}_{2.5}$, all steps of the CTFID algorithm can be justified by the rules of ctf-calculus for CBN2.25/CBN2.5 and the probability axioms restrained within $\mathcal{L}_{2.25}$/$\mathcal{L}_{2.5}$.

Line 1 and 2 of Algo. 5 are justfied by Lemma E.9 and Lemma E.10: if the input query $P(\mathbf{Y}_* = \mathbf{y}_*|\mathbf{X}_* = \mathbf{x}_*)$ is in $\mathcal{L}_{2.25}$/$\mathcal{L}_{2.5}$, then the ancestral set factorization $P(\bigcup_{D_t \in \mathbf{D}_*} D_{\mathbf{pa}_d} = d)$ over $\mathbf{D}_* = An(\mathbf{Y}_*, \mathbf{X}_*)$ and $\mathbf{d}_* \in Val(\mathbf{D}_*)$ consistent with $\mathbf{y}_*, \mathbf{x}_*$ is also in $\mathcal{L}_{2.25}$/$\mathcal{L}_{2.5}$. Thus the probability axioms underlying the marginalization step have all quantities within the corresponding layers.

Line 1 of Algo. 4 is justified by rule 3 of the ctf-calculus and Lemma E.2 and Lemma E.3 where both $\mathbf{D}_*$ and $\|\mathbf{D}_*\|$ are in the corresponding layers. Line 2 to 3 are justified by quantities in $\mathcal{L}_{2.25}$/$\mathcal{L}_{2.5}$ having consistent values. Line 4 to 5 follow from probability axiom to remove redundant variables. From line 6 to 14, the algorithm identifies the factors based on c-componentes using IDENTIFY [29] which soundness can be justified with do-calculus [13], which in turn is subsumed by ctf-calculus 2.25 by Lemma E.6. At line 17, the algorithm returns the result as a product that is justified by Lemma E.11.

Therefore, given a query in $\mathcal{L}_{2.25}$/$\mathcal{L}_{2.5}$, CTFID is both sound and complete to determine if it is identifiable from the available data without any intermediate step having quantities outside the layer. □

Theorem 3 (PCH*)**.** *Given an SCM $\mathcal{M}$ and its induced collections of observational ($\mathbf{P}^{\mathcal{L}_1}$), interventional($\mathbf{P}^{\mathcal{L}_2}$), $\mathcal{L}_{2.25}$ ($\mathbf{P}^{\mathcal{L}_{2.25}}$), $\mathcal{L}_{2.5}$ ($\mathbf{P}^{\mathcal{L}_{2.5}}$), and counterfactual ($\mathbf{P}^{\mathcal{L}_3}$) distributions: $\mathbf{P}^{\mathcal{L}_1} \subseteq \mathbf{P}^{\mathcal{L}_2} \subseteq \mathbf{P}^{\mathcal{L}_{2.25}} \subseteq \mathbf{P}^{\mathcal{L}_{2.5}} \subseteq \mathbf{P}^{\mathcal{L}_3}$.*

Proof. With PCH already established and proved for $\mathcal{L}_1$,$\mathcal{L}_2$ and $\mathcal{L}_3$ [3], we prove that (1) $\mathbf{P}^{\mathcal{L}_2} \subseteq \mathbf{P}^{\mathcal{L}_{2.25}}$, (2) $\mathbf{P}^{\mathcal{L}_{2.25}} \subseteq \mathbf{P}^{\mathcal{L}_{2.5}}$ and (3) $\mathbf{P}^{\mathcal{L}_{2.5}} \subseteq \mathbf{P}^{\mathcal{L}_3}$.

It is easy to show that $\mathbf{P}^{\mathcal{L}_2} \subseteq \mathbf{P}^{\mathcal{L}_{2.25}}$, because each distribution in $\mathbf{P}^{\mathcal{L}_2}$ can be derived from a marginalization of a distribution in $\mathbf{P}^{\mathcal{L}_{2.25}}$:

$$P(\mathbf{Y} = \mathbf{y}|do(\mathbf{X} = \mathbf{x})) = \sum_{X \in \mathbf{Y} \cap \mathbf{X}} P(\bigwedge_{V_i \in \mathbf{Y} \setminus \mathbf{X}} V_{i_{[\mathbf{x}]}} = v_i, \bigwedge_{V_i \in \mathbf{Y} \cap \mathbf{X}, v_i = V_i \cap \mathbf{x}} V_{i_{[\mathbf{x} \setminus v_i]}} = v_i) \quad (332)$$

where the subscripts for all variables take the whole set $\mathbf{x}$. Clearly, it is in $\mathbf{P}^{\mathcal{L}_{2.25}}$ as the consistent subscripts satisfy conditions of Def. 11.

It is also easy to see that $\mathbf{P}^{\mathcal{L}_{2.5}} \subseteq \mathbf{P}^{\mathcal{L}_3}$ because $\mathbf{P}^{\mathcal{L}_3}$ contains all possible joint distributions over all counterfactual variables, whereas $\mathbf{P}^{\mathcal{L}_{2.5}}$ imposes additional constraints over the joint of counterfactual variables.

To prove that $\mathbf{P}^{\mathcal{L}_{2.25}} \subseteq \mathbf{P}^{\mathcal{L}_{2.5}}$, we show that if a distribution satisfies Def. 11, it also satisfies Def. 12. First, note that the key difference between Def. 11 and Def. 12 lies in the two conditions. Thus, we only need to prove that a distribution of the form $P(\bigwedge_{V_i \in \mathbf{Y} \setminus \mathbf{X}} V_{i_{[\mathbf{x}_i]}} = v_i, \bigwedge_{V_i \in \mathbf{Y} \cap \mathbf{X}, v_i = V_i \cap \mathbf{x}} V_{i_{[\mathbf{x}_i \setminus v_i]}} = v_i)$ satisfying the two conditions in Def. 11 must also satisfy the two conditions in Def. 12.

For Cond. (i), both languages require the subscripts to cover the whole space of $\mathbf{X}$. However, Def. 11 is stronger by restricting the value assignments to the set $\mathbf{x}$, while Def. 12 allows $\mathbf{x}_i$ to take different values from $Val(\mathbf{X}_i)$. Thus, if Cond. (i) of Def. 11 holds, Cond. (i) of Def. 12 immediately holds.

Graphical Model	Meaning of Missing Directed Edge	Meaning of Missing Bidirected Edge
$\mathcal{L}_1$: BN	$P(v_i\|\mathbf{pa}_i, \mathbf{nd}_i) = P(v_i\|\mathbf{pa}_i)$	
$\mathcal{L}_2$: CBN	$P(v_{i_{\mathbf{pa}_i,\mathbf{z}}}) = P(v_{i_{\mathbf{pa}_i}})$	$P(v_i\|do(\mathbf{x}), \mathbf{pa}_i^c, do(\mathbf{pa}_i^u))$ $= P(v_i\|do(\mathbf{x}), \mathbf{pa}_i^c, \mathbf{pa}_i^u)$
$\mathcal{L}_{2.25}$: CBN2.25	$P(v_{i_{\mathbf{pa}_i,\mathbf{z}}}, \mathbf{w}_*) = P(v_{i_{\mathbf{pa}_i}}, \mathbf{w}_*)$, with $P(v_{i_{\mathbf{pa}_i,\mathbf{z}}}, \mathbf{w}_*) \in \mathbf{P}^{\mathcal{L}_{2.25}}$	$P(v_{i_{\mathbf{pa}_i}}, v_{j_{\mathbf{pa}_j}}) = P(v_{i_{\mathbf{pa}_i}})P(v_{j_{\mathbf{pa}_j}})$, with $V_i \neq V_j$ and $\mathbf{pa}_i$ and $\mathbf{pa}_j$ taking consistent values
$\mathcal{L}_{2.5}$: CBN2.5	$P(v_{i_{\mathbf{pa}_i,\mathbf{z}}}, \mathbf{w}_*) = P(v_{i_{\mathbf{pa}_i}}, \mathbf{w}_*)$, with $P(v_{i_{\mathbf{pa}_i,\mathbf{z}}}, \mathbf{w}_*) \in \mathbf{P}^{\mathcal{L}_{2.5}}$	$P(v_{i_{\mathbf{pa}_i}}, v_{j_{\mathbf{pa}_j}}) = P(v_{i_{\mathbf{pa}_i}})P(v_{j_{\mathbf{pa}_j}})$, with $V_i \neq V_j$
$\mathcal{L}_3$: CTFBN	$P(v_{i_{\mathbf{pa}_i,\mathbf{z}}}, \mathbf{w}_*) = P(v_{i_{\mathbf{pa}_i}}, \mathbf{w}_*)$, for any $\mathbf{w}_*$	$P(v_{i_{\mathbf{pa}_i}}, v_{j_{\mathbf{pa}_j}}) = P(v_{i_{\mathbf{pa}_i}})P(v_{j_{\mathbf{pa}_j}})$ with $\mathbf{pa}_i \neq \mathbf{pa}_j$ if $V_i = V_j$

Table 4: Summary of how missing edges are interpreted in graphical models at different layers

For Cond. (ii) and by Cor. 3, the antecedent in Def. 12 checks if there is a directed path from $B \in \mathbf{X}$ to $V_i \in Ch(B)$ to V_j in $G_{\overline{\mathbf{X}_j}}$. If such a path exists, we denote it by p. There are two possibilities: (a) p is in $G_{\overline{\mathbf{X}\setminus V_j}}$; (b) p is not in $G_{\overline{\mathbf{X}\setminus V_j}}$. For (a), Cond. (ii) of Def. 11 will enforce b to appear in the subscript of both V_i and V_j. For (b), it implies that there exists a variable $X \in \mathbf{X}\setminus\mathbf{X}_j$ that lies on p between V_i and V_j. We focus on the subpath p' of p directed from X to V_j. If X is in $An(V_j)$ in $G_{\overline{\mathbf{X}_j}}$, then X must be in $\mathbf{X}_j$ by Cond. (ii) of Def. 11 which leads to a contradiction. If X is not in $An(V_j)$ in $G_{\overline{\mathbf{X}_j}}$, then there exists another $X' \in \mathbf{X}\setminus\mathbf{X}_j$ that lies on p' between X and V_j. We can apply the same logic to shorten p until there is no more variable in $\mathbf{X}\setminus\mathbf{X}_j$ that fulfills the same condition. When this terminal condition is hit, the final subpath enforces the variable in $\mathbf{X}\setminus\mathbf{X}_j$ on the path to be in the subscript of V_j. The same contradiction is achieved. As a result, there cannot be any variable $X \in \mathbf{X}\setminus\mathbf{X}_j$ that lies on p between V_i and V_j. Therefore, whenever the antecedent of Cond. (ii) of Def. 12 is triggered, Cond. (ii) of Def. 11 also holds to enforce consistent subscripts between V_i and V_j.

This proves that all distributions in $\mathbf{P}^{\mathcal{L}_{2.25}}$ are also in $\mathbf{P}^{\mathcal{L}_{2.5}}$, or equivalently $\mathbf{P}^{\mathcal{L}_{2.25}} \subseteq \mathbf{P}^{\mathcal{L}_{2.5}}$.

□

Theorem 4 (Hierarchy of Graphical Models, PCH*)**.** *Given a causal diagram $\mathcal{G}$, the set of constraints it encodes when it is interpreted as a graphical model on layer i is a subset of the constraints it encodes when it is interpreted as a graphical model on layer j, when $i \leq j$.*

Proof. The constraints encoded by a BN are included as Cond. (i) of the corresponding CBN, making the containment relationship is straightforward. The hierarchical relationship among the constraints encoded by CBN2.25, CBN2.5, and CTFBN is also straightforward, as they share the same structural form while progressively increasing the flexibility of distributions allowed at each level in the model hierarchy. The containment relationship between CBN and CBN2.25 follows from the fact that do-calculus is subsumed by the ctf-calculus 2.25 (Lemma E.6), and that the constraints defined in CBN imply all rules of do-calculus, while those in CBN2.25 imply all rules of ctf-calculus 2.25.

Since the constraints encoded by graphical models are encoded by the missing edges in $\mathcal{G}$, we can alternatively establish the hierarchy by comparing how different models interpret these missing edges, as summarized in Table 4. For missing directed edges, the constraint forms are consistent across layers, but higher layers allow increasing flexibility in the sets $\mathbf{w}_*$ that can be jointly conditioned on. Similarly, for missing bidirected edges, the independence constraints in CBN2.25s, CBN2.5s, and CTFBNs share a common structure, with each successive model relaxing the limitations on how these independencies are expressed:

- Independence constraints in CBN2.25s only apply to distributions over distinct variables that share consistent parent values.
- Independence constraints in CBN2.5s extend to distributions over distinct variables, allowing their parents' values to vary freely.

- Independence constraints in CTFBNs apply to distributions over any variables, including those of the form $P(W_{\mathbf{pa}_w}, W_{\mathbf{pa}'_w})$ as long as $\mathbf{pa}_w \neq \mathbf{pa}'_w$.

□

Theorem 5 ($\mathcal{L}_{2.5}$-Connection — CBN2.5 (Markovian and Semi-Markovian))**.** *The Causal diagram $\mathcal{G}$ induced by the SCM $\mathcal{M}$ following the constructive procedure in Def. 6 is a CBN2.5 for $\mathbf{P}^{\mathcal{L}_{2.5}}$, the collection of all $\mathcal{L}_{2.5}$ distributions induced by $\mathcal{M}$.*

Proof. The proof is similar to the proof for Theorem 1 with the independence restrictions expanded to allow inconsistent parent values, and the exclusion and consistency restrictions expanded to join more $\mathbf{W}_*$ such that the distributions are within $\mathcal{L}_{2.5}$ instead of $\mathcal{L}_{2.25}$. □

F Frequently Asked Questions

Q1. Where is the causal diagram coming from? Is it reasonable to expect the data scientist to create one?

Answer. First, the assumption of the causal diagram is made out of necessity. The causal diagram is a well-known flexible data structure that is used throughout the literature to encode a qualitative description of the generating model, which is often much easier to obtain than the actual mechanisms of the underlying SCM [20, 28, 22]. The goal of this paper is not to decide which set of assumptions is the best but rather to provide tools to perform the inferences once the assumptions have already been made, as well as understanding the trade-off between assumptions and the guarantees provided by the method.

Second, the true underlying causal diagrams cannot be learned only from the observational distribution in general. More specifically, there almost surely exist situations that $\mathcal{M}_1$ and $\mathcal{M}_2$ induce the same observational distribution but are compatible with different causal diagrams (see [3, Sec. 1.3] for details). With higher layer distributions (such as distributions from $\mathcal{L}_2$), it is possible to recover a more informative equivalence class of diagrams that encode additional constraints present in the input layer [16, 15, 14, 17, 30].

Q2. What is a graphical model and how can it help us in causal inference?

Answer. A graphical model is a modeling tool that allows one to represent a compatibility relationship between a causal diagram $\mathcal{G}$ and a collection of distributions $\mathbf{P}$. Specifically, it encodes how the topological structure of the diagram can be interpreted to impose constraints on the associated distributions. For instance, when restricting attention to $\mathcal{L}_1$ distributions (i.e., purely observational), Bayesian Networks (BNs) are the most prominent graphical models to encode conditional independence constraints of the observational distribution [18]. As we climb up the PCH and include more distributions into the collection, more constraints start to emerge. To encode the richer set of causal constraints in $\mathcal{L}_2$ distributions (i.e., interventional), the Causal Bayesian Network (CBN) was introduced [3]. More recently, CTFBN is introduced to encode the compatibility relationship between the causal diagram and $\mathcal{L}_3$ distributions (i.e., counterfactual) [1]. The models defined in this work further refine the space of $\mathcal{L}_3$ distributions by restricting to constraints that are, at least in principle, empirically falsifiable. In a nutshell, a graphical model should not be viewed merely as a causal diagram, but rather as a formal specification of the compatibility relationship between a pair $\langle \mathcal{G}, \mathbf{P} \rangle$. An example of a CBN is illustrated in Fig. 28, where missing edges in the causal diagram represent invariance constraints in the distributions.

The causal diagram in the graphical model offers a compact representation for constraints in the associated distributions. These constraints are fundamental to causal inference, as they constitute one of the three core inputs to the causal inference engine (Fig. 1). As discussed ealier, the main task in causal inference is to determine whether a query from a higher layer of the PCH can be identified as a function of observed data from lower layers. For example, the task may be to identify a causal effect $P(y|do(x))$ when only the observational data $P(\mathbf{v})$ is available. According to the Causal Hierarchy Theorem (CHT), these layers are strictly distinct, and it is impossible to ascend to a higher layer without additional assumptions about that layer [3, Thm. 1]. The constraints encoded by graphical models serve precisely this role – they encode the assumptions about higher layers that enable us to bridge the gap and make such inferences possible. Given the CBN in Fig. 28, the invariance constraint

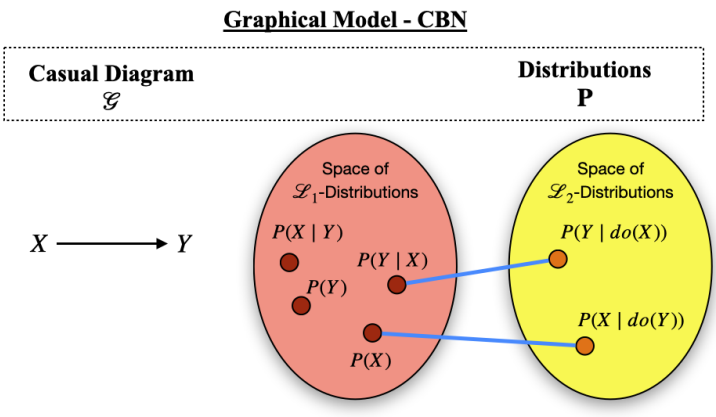


Figure 28: A CBN is a pair $\langle \mathcal{G}, \mathbf{P} \rangle$. Blue lines represent invariant constraints in $\mathbf{P}$, which are represented by features from $\mathcal{G}$: missing directed edge from Y to X corresponds to the invariance constraint $P(X|do(Y)) = P(X)$ and missing bidirected edge between X and Y corresponds to the invariance constraint $P(Y|do(X)) = P(Y|X)$.

$P(Y|do(X)) = P(Y|X)$ allows us to identify the $\mathcal{L}_2$ query $P(y|do(x))$ as $P(y|x)$, which only involves observational distributions. Question 9 below will provide further details on the inferential process by explaining how the local constraints defined in a graphical model can be composed to derive additional constraints implied by the model.

Q3. Why do we need to introduce new layers to the PCH, besides the existing ones?

Answer. The original three layers of the PCH, capturing observational, interventional, and counterfactual distributions, provide a natural partition among distinct capabilities in causal reasoning. Layers 1 and 2 correspond to well-understood physical procedures: random sampling for observational distributions and random experimentation for interventional distributions. In contrast, Layer 3 consists of purely counterfactual quantities, that are traditionally considered detached from empirical data collection in principle. In addition, while Layers 1 and 2 are well-structured and homogeneous (each quantity within a layer having a similar interpretation), Layer 3 is more heterogeneous and contains quantities that represent different aspects of the underlying data-generating process.

More recently, Bareinboim, Forney and Pearl introduced a new experimental procedure, counterfactual randomization, that allowed one to sample directly from an $\mathcal{L}_3$ distribution [4]. This work was further extended in [24]. The introduction of counterfactual randomization reveals a finer structure within Layer 3, distinguishing between counterfactual distributions that are empirically accessible and those that are not. This fine-graining of Layer 3 is illustrated in Fig. 29. Notably, these new families of distributions have attractive properties, including well-defined symbolic languages as well as a closed set of inferential rules, as shown in this work. This new view opened up a natural way of partitioning $\mathcal{L}_3$. In this work, we studied the interplay between graphical models that inherent these features of the PCH and have the property of empirical falsifiability.

To answer the question, the new layers introduced in the refined PCH may not be necessary for all researchers. The original PCH already represents a major milestone in formalizing the logic of causal inference. Still, for some researchers, the refinement and further partitioning of Layer 3 can offer valuable insights. In particular, it allows for a more precise understanding of the trade-off between empirical falsifiability and the inferential power of graphical models, and provides a tighter feedback loop between theoretical assumptions and experimental capabilities.

Q4. What is the difference between layers 2.25 and 2.5?

Answer. The main difference between $\mathcal{L}_{2.25}$ and $\mathcal{L}_{2.5}$ lies in the type of counterfactual randomization allowed. For $\mathcal{L}_{2.25}$, a counterfactual randomization applied to a variable X assigns the same value x across all its children and descendants. As a result, distributions in this layer cannot contain pairs of potential outcomes $W_{\mathbf{s}}, R_{\mathbf{t}}$ with conflicting subscripts where $x \in \mathbf{s}, x' \in \mathbf{t}$ and $x \neq x'$. In contrast, the counterfactual randomization action on a

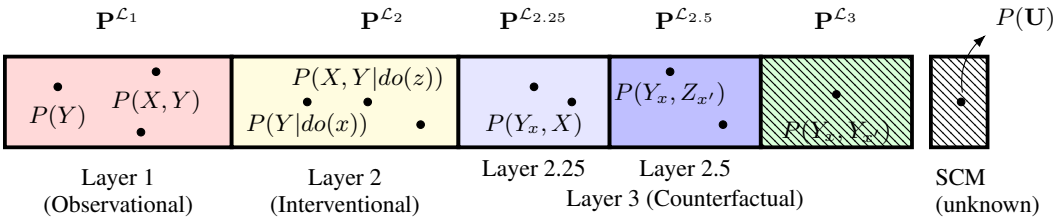


Figure 29: Pearl Causal Hierarchy (PCH*) induced by an unknown SCM $\mathcal{M}$. Layers 1 and 2 are realizable, and Layer 3 is partially realizable. The realizable portion of Layer 3 are further refined into two new layers: 2.25 and 2.5.

variable X in $\mathcal{L}_{2.5}$ is more flexible and allows each outgoing edge from X to take a different value. This flexibility leads to the possibility of some distributions in the layer to include potential outcomes with different subscripts. This difference is graphically illustrated in Fig. 8. However, all descendants of each child of X must still share the same value of x, unless all directed paths from X to the descendant are blocked by other intervened variables. This restriction stems from the rules of counterfactual randomization, which prohibit an intervention to bypass a child and directly affect a descendant's perception of X. In summary, the constraint on consistent subscript begins at the intervened variable X in $\mathcal{L}_{2.25}$, but shifts to the children of X in $\mathcal{L}_{2.5}$. These differences are reflected in the relaxed conditions that define the symbolic language of $\mathcal{L}_{2.5}$, relative to those of $\mathcal{L}_{2.25}$.

Q5. Are all distributions within Layers 2.5 realizable?

Answer. Theoretically, all distributions in $\mathcal{L}_{2.5}$ are realizable if every action in the maximal feasible action set is permitted. That is, *in principle*, an agent could draw samples from any distribution in this layer through experimental procedures. However, whether a distribution is realizable *in practice* depends on the physical constraints of the system. If certain actions – such as counterfactual randomization on specific variables – are not feasible, then some distributions in $\mathcal{L}_{2.5}$ will not be realizable in real-world settings [24].

The same principle applies to other layers of the PCH. For example, all distributions in $\mathcal{L}_2$ are realizable *in principle*, assuming the agent can freely intervene on all variables. However, practical constraints – such as cost, ethics, or technological barriers – may render some interventions infeasible, thereby restricting the subset of $\mathcal{L}_2$ distributions that can be realized.

Given a causal diagram and a specification of the allowed actions, one can determine whether a given set of distributions is realizable [24]. Viewed this way, the full collection of distributions in $\mathcal{L}_{2.5}$ can be interpreted as the theoretical boundary of what is empirically accessible through physical experimentation.

Q6. How does the hierarchical structure defined over graphical models provide useful information on the models?

Answer. The hierarchical structure over graphical models offers a clear picture of the differences in the strength of assumptions encoded by each model. In causal inference specifically, the strength of the assumptions determines what queries the model may in principle support – specifically, whether the causal inference engine can proceed and provide useful insights about the query. For instance, an $\mathcal{L}_2$ query $P(y|do(x))$ cannot be answered by a BN, which only encods $\mathcal{L}_1$ constraints that does not have the power to bridge the gap between the two layers. This limitation is formally captured by the Causal Hierarchy Theorem (CHT), which states that to answer questions at one layer, one needs assumptions at the same layer or even higher. This understanding allows practitioners to select models from the hierarchy with sufficient inferential power for the query at hand.

On the other hand, the hierarchy also provides guidance in the opposite direction – helping to identify when a model might be stronger than necessary. For instance, while any model at or above a CBN in the hierarchy can answer an $\mathcal{L}_2$ query $P(y|do(x))$, using a model that makes counterfactual assumptions (e.g., a CBN2.5) would be unnecessarily strong and harder to falsify. Therefore, knowing the hierarchy of graphical models also allows practitioners to avoid choosing models that make extra assumptions not required in the target inferential task.

Putting these observations together, Table 3 summarizes when a model is sufficient and/or necessary for queries from each layer of the PCH. In short, the hierarchy serves as a practical guide for selecting models that are both sufficient and necessary – maximizing inferential power while minimizing unfalsifiable assumptions.

Q7. What is the difference between the hierarchical structure of languages and graphical models?

Answer. The hierarchical structure of the languages (i.e., the PCH) defines how different families of distributions are related – specifically, each layer's distributions form a subset of those in the layer above. In parallel, the hierarchy of graphical models reflects how constraints on these distributions are encoded through the topological properties of the causal diagram. Each graphical model at layer i encodes constraints over the corresponding family of distributions in layer i of the PCH. Therefore, the hierarchy of the languages directly informs the hierarchy of graphical models.

However, since a graphical model is defined as a compatibility relationship between a pair $\langle \mathcal{G}, \mathbf{P} \rangle$, the expressiveness of the topological features in $\mathcal{G}$ also plays a critical role. As we move up the hierarchy, the causal diagrams must support richer or more expressive interpretations of missing edges to capture the increasingly complex constraints required by higher-layer distributions. Both hierarchies are illustrated in Fig. 10, where square boxes depict the hierarchy over distributions, and round boxes represent the hierarchy over the constraints encoded by graphical models.

Q8. Why should a data scientist care about the trade-off between expressive power and empirical falsifiability of the graphical models?

Answer. In any modeling task, it is generally desirable to construct a model that accurately reflects the underlying generative process while also supporting future inferential tasks. Achieving stronger inferential power often requires incorporating stronger assumptions into the model. However, these assumptions can make the model more prone to errors that does not match with reality. Empirical falsifiability acts as a form of regularization, enabling the data scientist to identify, falsify and possibly correct wrong assumptions using empirical evidence. As a result, the model can yield more reliable and trustworthy causal conclusions. The importance of falsifiability echoes Karl Popper's philosophy, which argues that scientific theories must be testable and refutable – setting science apart from pseudoscience [23]. Thus, understanding where each graphical model falls on the spectrum of expressive power versus empirical falsifiability is essential for practitioners who align with Popper's principle.

Q9. What are the differences between local constraints and global constraints?

Answer. As discussed earlier when we introduce the inferential machinery for CBN2.25/CBN2.5, local constraints refer to those that are defined over distributions involving a variable and its parents, and they are the constraints that are explicitly stated in the definitions of graphical models. For example, the local constraints in a BN are the conditional independencies of the form $P(v_i|\mathbf{pa}_i, \mathbf{nd}_i) = P(v_i|\mathbf{pa}_i)$, where $\mathbf{pa}_i$ denotes the parents and $\mathbf{nd}_i$ the non-descendants of V_i. Given a BN over the chain diagram $X \to Z \to W \to Y$, the local constraints include $P(w|z, x) = P(w|z)$ and $P(y|w, z, x) = P(y|w)$.

Global constraints, on the other hand, involve arbitrary subsets of variables, possibly far apart in the causal diagram. These constraints are not explicitly listed in the model's definition but can be derived by composing local constraints. For example, given the same BN over the chain above, a global constraint is $P(y|z, x) = P(y|z)$, where the direct parent of Y, namely W, is no longer explicitly conditioned on.

This distinction highlights the role of local constraints as a basis for implying the full set of global constraints that a graphical model implies, as illustrated in Fig. 5. This relationship is mirrored in the connection between a graphical model and its associated inferential calculus: the calculus rules form the closure of all global constraints that logically follow from the local ones encoded in the model.

The process by which local constraints can be composed to yield global constraints was illustrated in Example 11. We revisit this idea with a new example in Fig. 5. Consider a CBN over the chain diagram $X \to Z \to Y$. The local constraints specified in the definition of the CBN are depicted as connecting lines between nodes within the small yellow circle. These local constraints can imply additional constraints not explicitly listed in the definition. One such global constraint is $P(y|do(x)) = P(y|x)$, represented by the red connection line in the figure. This global constraint can be derived by composing – or "gluing" – a sequence

of local invariance constraints, shown as blue connection lines.

$$P(y|do(x)) = \sum_z P(y|do(x), z)P(z|do(x)) \quad \text{(Probability Axiom)} \tag{333}$$

$$= \sum_z P(y|do(xz))P(z|do(x)) \quad \text{(Cond. (iii) of Def. 21)} \tag{334}$$

$$= \sum_z P(y|do(z))P(z|do(x)) \quad \text{(Cond. (ii) of Def. 21)} \tag{335}$$

$$= \sum_z P(y|z)P(z|x) \quad \text{(Cond. (iii) of Def. 21)} \tag{336}$$

$$= \sum_z P(y|xz)P(z|x) \quad \text{(Cond. (i) of Def. 21)} \tag{337}$$

$$= P(y|x) \quad \text{(Probability Axiom)} \tag{338}$$

In summary, although not all constraints are explicitly included in the local basis of a graphical model definition, many are implied through its structure. Since the 1980s, this ability to encode a parsimonious, polynomial-sized set of local constraints that implicitly represent an exponential number of global constraints has been an attractive feature contributing to the popularity and usefulness of graphical models in inferential tasks.

Q10. What is the connection between realizability and empirical falsifiability?

Answer. Realizability is a property of distributions, indicating that an agent can draw samples from them through physical experimentation. For example, if an agent can intervene on a variable X and fix it to a value x, it gains access to the interventional distribution $P(\mathbf{v} \mid do(x))$ in layer $\mathcal{L}_2$.

In the context of graphical models, empirical falsifiability is property of constraints over these distributions. To empirically falsify a constraint, the agent must have the experimental capabilities to draw samples from all distributions involved in the constraint. In other words, the constraint's falsifiability requires the realizability of the associated distributions. For instance, testing the constraint $P(y \mid do(x, z)) = P(y \mid do(x))$ requires the ability to sample from both $P(y \mid do(x, z))$ and $P(y \mid do(x))$. Whether this is feasible depends on the experimental capabilities and limitations of the system in question.

Q11. What is the difference between an SCM and Layer 3 distributions or Layer 3 graphical models?

Answer. An SCM is a more granular level model with details about the exogenous variables $\mathbf{U}$, which induces the full set of distributions over the endogenous variables $\mathbf{V}$ in the PCH, as illustrated in Fig. 7 and Fig. 3. Specifically, given a distribution over the exogenous variables $P(\mathbf{u})$ and the structural equations that determine each endogenous variable as a function of its parents (both endogenous and exogenous), we can compute all distributions over the endogenous variables following the formula in the PCH definition (Def. 4). In contrast, the PCH abstracts away from the exogenous variables, treating them as hidden background factors unobserved by the agent. As a result, Layer 3 distributions and its corresponding graphical models are defined solely in terms of the endogenous variables.

Given an SCM, it is also possible to evaluate individual level effects when the exogenous state of a specific unit $\mathbf{u} \in Val(\mathbf{U})$ is known, by solving the set of mechanisms following the topological order of evaluation. Layer 3 distributions and graphical models, on the other hand, offer population-level descriptions of causal relationships, without access to individual-level information.

In a nutshell, an SCM provides full access to Layer 3 distributions and graphical models, as it encodes the necessary generative mechanisms. However, the reverse does not hold: Layer 3 distributions or graphical models do not determine a unique SCM, since an SCM requires additional, often unobservable, information about the exogenous variables and structural mechanisms.

Q12. Given that the constructive procedure for the causal diagram is the same, why do we need, or even have, different layers of graphical models?

Answer. Even though the same causal diagram $\mathcal{G}$ is shared across many different models, the compatibility relationships it represents differ depending on the model. As discussed

earlier, a graphical model is a pair $\langle \mathcal{G}, \mathbf{P} \rangle$, where graphical feature in $\mathcal{G}$ are interpreted to represent constraints in $\mathbf{P}$. As $\mathbf{P}$ expands to include distributions from higher layers of the PCH, the set of constraints that the graph must represent also becomes richer. As a result, each missing edge is required to encode stronger and more expressive constraints over a broader class of distributions. This is illustrated in Example 15 and Table 4.