TECHNICAL REPORT
R-130
May, 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 structural causal models and provide a basis from which causal inferences can be performed. Models that encode stronger constraints tend to have higher expressive power at the expense of lower empirical falsifiability. In this paper, we introduce two new collections of distributions which include counterfactual quantities that become experimentally accessible under the counterfactual randomization action. Correspondingly, we provide 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 inference in these two new models with rules that are also fall within the empirically falsifiable boundary. In addition, 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 scientific disciplines and human decision-making, and it is increasingly recognized as a key element for advancing AI and machine learning in enhancing robustness, interpretability, and generalizability [16, 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* [16, 2]. 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 in fact followed diet x'.

When the true underlying 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 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 model. The inferential process can be implemented through the *causal inference engine* [1], as illustrated in Fig. 1. The engine takes three inputs: $\{$(1)*Query*, (2)*Data*, (3)*Model*$\}$, where each represents a different aspect of the underlying SCM. The *Query* specifies the causal quantity of interest, the *Data* consists of data gathered through interactions with the environment like random samplings or

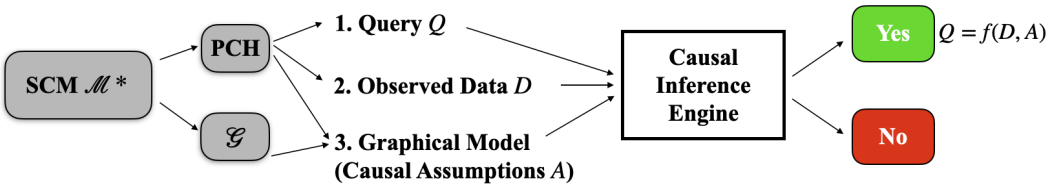


Figure 1: Unobserved SCM and the causal inference engine. The engine takes as input a query, a model, and datasets, and returns whether the query is computable from the assumptions and data.

randomized experiments, and the *Model* encodes assumptions about the SCM. A common method for articulating such assumptions is via *graphical models*, particularly *causal diagrams* [14, 15, 2, 6], which encode constraints describing how different quantities within the PCH relate to one another. For example, Pearl in his celebrated 1988 book gave a comprehensive study of Bayesian Networks (BN) as encoders of $\mathcal{L}_1$ equality constraints between observational distributions implied by *conditional independence*, like $P(y|x) = P(y)$ [14]. In contrast, Causal Bayesian Networks (CBN) encode equality constraints across distributions in both $\mathcal{L}_1$ and $\mathcal{L}_2$, like $P(y|do(x)) = P(y|x)$ [15].

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 matching reflects Nancy Cartwright's famous motto "no causes in, no causes out" [4], which has been formalized by the *Causal Hierarchy Theorem* (CHT): to perform inferences on a quantity in layer i, one needs knowledge from layer i or above [2]. 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 Counterfactual Bayesian Network (CTFBN) encoding $\mathcal{L}_3$ constraints, while sufficient for the target query in $\mathcal{L}_2$, impose assumptions that are stronger than necessary [5, 6].

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* from the philosophy of science. As advocated by Popper, a system is scientific only if it is refutable by empirical tests [18]. The falsifiability of an assumption in a graphical model depends on the feasibility to draw samples from its underlying distributions (also known as *realizability* of the distributions [19]. 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* [8]. $\mathcal{L}_3$, in contrast, 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 sampled via any randomized experiments. However, a recent work showed a surprising result that an $\mathcal{L}_3$ quantity known as the effect of the treatment on the treated (ETT), $P(y_x|x')$, can be sampled from using a new experimental procedure called *counterfactual randomization* [3]. Subsequent work further refined and characterized the set of $\mathcal{L}_3$ distributions that are realizable in principle [19].

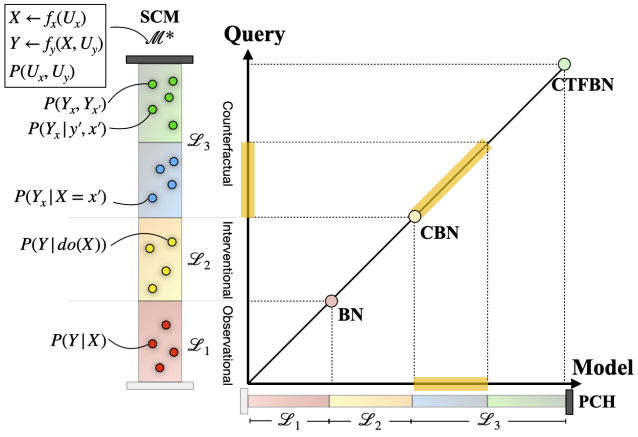


Figure 2: Expressive power of queries and graphical models along the PCH. The model's constraints should be at least as expressive as the query for the causal inference engine to work. Layer 3 is partitioned into two sub-regions: the green region represents $\mathcal{L}_3$ distributions that cannot be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

This empirical heterogeneity of distributions between $\mathcal{L}_2$ and $\mathcal{L}_3$ leads to an important question: what assumptions are sufficient and necessary to answer a query in $\mathcal{L}_3$? In this paper, we address this question by providing a finer-grained understanding of the region between $\mathcal{L}_2$ and $\mathcal{L}_3$ in PCH. In particular, we focus on the orange region in Fig. 2, where models extend beyond the typical (Fisherian) interventional world, which allows certain counterfactual queries to be answered. To this end, we will formally define language/model/inferential machinery for two families of realizable distributions. Together, they offer a precise formalization of Cartwright's principle where "causes in" (models/assumptions) are matched with "causes out" (queries). Our main contributions are:

(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 hierarchy from two angles: (a) Expressive Power (b) Empirical Falsifiability. We show that models higher in the hierarchy encode stronger assumptions that permit more expressive queries, but are increasingly harder to empirically falsify.

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).

Background and Definitions. We use *Structural Causal Models* (SCM) as the underlying semantical framework [15]. An SCM $\mathcal{M}$ is a 4-tuple $\langle \mathbf{V}, \mathbf{V}, \mathcal{F}, P(\mathbf{u}) \rangle$, where $\mathbf{U}$ is a set of exogenous (latent) variables, distributed according to $P(\mathbf{u})$; $\mathbf{V}$ is a set of endogenous (observable) variables; $\mathcal{F}$ is a set of functions such that for each $V_i \in \mathbf{V}$, f_i maps from a set of exogenous variables $\mathbf{U}_i \subseteq \mathbf{U}$ and a set of endogenous variables $\mathbf{Pa}_i \subseteq \mathbf{V}$ to the $Val(V_i)$ [2]. An SCM $\mathcal{M}$ induces a *causal diagram* $\mathcal{G}$ over $\mathbf{V}$ where directed edges reflect functional arguments and bidirected edges reflect shared or correlated latent confounders. We assume the model has no cyclic dependencies among variables. Two variables belong to the same *c-component* if they are connected by a path made entirely of bidirected edges.

Intervention $do(\mathbf{x})$ in an SCM $\mathcal{M}$ creates a *submodel* $\mathcal{M}_{\mathbf{x}}$, where functions generating $\mathbf{X}$ are replaced with constant values $\mathbf{x}$. The functions in $\mathcal{M}_{\mathbf{x}}$ are denoted as $\mathcal{F}_{\mathbf{x}}$. Given a set of variables $\mathbf{Y} \in \mathbf{V}$, the solution for $\mathbf{Y}$ in $\mathcal{M}_{\mathbf{x}}$ defines a *potential outcome* denoted as $\mathbf{Y}_{\mathbf{x}}(u)$. $\|Y_{\mathbf{x}}\|$ denotes the *exclusion operator* such that $\|Y_{\mathbf{x}}\| = Y_{\mathbf{z}}$ with $\mathbf{Z} = \mathbf{X} \cap An(Y)_{\mathcal{G}_{\overline{\mathbf{X}}}}$, $\mathbf{z} = \mathbf{x} \cap \mathbf{Z}$ and $\mathcal{G}_{\overline{\mathbf{X}}}$ is $\mathcal{G}$ with all incoming edges into X removed. An SCM $\mathcal{M}$ also induces all quantities within the *Pearl Causal Hierarchy* (PCH): for any $\mathbf{Y}, \mathbf{Z}, ..., \mathbf{X}, \mathbf{W} \subseteq \mathbf{V}$, $\mathcal{L}_1$ (Observational): $\mathbf{P}^{\mathcal{M}}(\mathbf{y}) = \sum_{\mathbf{u}} \mathbf{1}[\mathbf{Y}(\mathbf{u}) = \mathbf{y}] P(\mathbf{u})$; $\mathcal{L}_2$ (Interventional): $\mathbf{P}^{\mathcal{M}}(\mathbf{y}_{\mathbf{x}}) = \sum_{\mathbf{u}} \mathbf{1}[\mathbf{Y}_{\mathbf{x}}(\mathbf{u}) = \mathbf{y}] P(\mathbf{u})$; $\mathcal{L}_3$ (Counterfactual): $\mathbf{P}^{\mathcal{M}}(\mathbf{y}_{\mathbf{x}}, ..., \mathbf{z}_{\mathbf{w}}) = \sum_{\mathbf{u}} \mathbf{1}[\mathbf{Y}_{\mathbf{x}}(\mathbf{u}) = \mathbf{y}, ..., \mathbf{Z}_{\mathbf{w}}(\mathbf{u}) = \mathbf{z}] P(\mathbf{u})$. We denote the collection of all $\mathcal{L}_1$ distributions as $\mathbf{P}^{\mathcal{L}_1}$, the collection of all $\mathcal{L}_2$ distributions as $\mathbf{P}^{\mathcal{L}_2}$, and the collection of all $\mathcal{L}_3$ distributions as $\mathbf{P}^{\mathcal{L}_3}$.

Equalities or inequalities between polynomials over $\mathcal{L}_i$ terms represent special marks an SCM imprints on its distributions, called *invariance constraints*. A *graphical model* (also known as a *compatibility relation*) is a pair $\langle \mathcal{G}, \mathbf{P} \rangle$, where $\mathcal{G}$ is a graph and $\mathbf{P}$ is a collection of distributions over $\mathbf{V}$. The missing edges in $\mathcal{G}$ represent certain invariance constraints within $\mathbf{P}$. Some examples of graphical models corresponding to three layers of the PCH are *Bayesian Network* (BN) [14], *Causal Bayesian Network* (CBN) [2], and *Counterfactual Bayesian Network* (CTFBN, [6]).

The *counterfactual randomization* action (CTF-RAND$(X \rightarrow \mathbf{C})^{(i)}$) [3, 19] is an experimental procedure to fix the value of X as an input to functions generating $\mathbf{C} \subseteq Ch(X)$ using a randomising device having support over $Val(X)$, for unit i, where $Ch(X)$ stands for variables taking X as an argument in their functions. A *feasible action set* describes all experimental actions allowed in a system. The *maximal feasible action set* contains all sampling, intervention and CTF-RAND actions over all variables and gives the agent the most granular experimental capabilities. More detailed background definitions and examples are provided in Appendix A for reference.

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

In this section, we provide a finer-grained analysis of the counterfactual layer ($\mathcal{L}_3$) by circumscribing subsets of distributions that become realizable, assuming the feasible action set includes all actions required to sample from any $\mathcal{L}_2$ distributions, along with some counterfactual randomization capabilities. Specifically, we define two distinct collections of distributions that are realizable based on different degrees of flexibility in how counterfactual randomization influences downstream variables (Sec. 2.1). We then introduce the corresponding graphical models encoding constraints in these distribution sets (Sec. 2.2), followed by the inferential machinery for each model (Sec. 2.3).

2.1 Formal Languages for Distribution Families

The first collection of realizable distributions is defined under the assumption that counterfactual randomization is allowed for all variables in $\mathbf{V}$, subject to the constraint that each CTF-RAND on X fixes a single value of X across all its children. The symbolic representation and valuation given an SCM for distributions in this collection are provided below.

Definition 1 (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 intervention value set $\mathbf{x}$. For each $\mathbf{X}, \mathbf{Y} \subseteq \mathbf{V}, \mathbf{x} \in Val(\mathbf{X})$:*

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

such that (i) $\mathbf{x}_i \subseteq \mathbf{x}$ and $\bigcup_i \mathbf{x}_i = \mathbf{x}$; and (ii) For any $v_i \in \mathbf{x}$, for 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 as $\mathbf{P}^{\mathcal{L}_{2.25}}$.

Cond. (i) of Def. 1 ensures that only value assignments from the intervention value set $\mathbf{x}$ appear in the subscript, and each value in $\mathbf{x}$ appears at least once to prevent redundancy of representing the same distribution under different interventions where $\mathbf{x}_1 \subset \mathbf{x}_2$. Cond. (ii) enforces all descendants of the intervened variable X to share the common value of x, unless the path from X to the descendant is blocked by another variable in the intervention set. Both conditions arise from the limited flexibility imposed on the counterfactual randomization action.

The second collection of distributions is defined under the maximal feasible action set with a more flexible counterfactual randomization that allows each children of X to take a potentially different value. This relaxation expands the set of realizable distributions beyond those in the first collection. Next, we define the symbolic representation and valuation of these distributions.

Definition 2 (Layer 2.5 ($\mathcal{L}_{2.5}$)). [1] *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 intervention variable set $\mathbf{X}$. For each $\mathbf{Y}, \mathbf{X} \subseteq \mathbf{V}$:*

$$\begin{aligned} &P^{\mathcal{M}}\Big(\bigwedge_{V_i \in \mathbf{Y} \setminus \mathbf{X}} V_{i_{[\mathbf{x}_i]}} = v_i, \bigwedge_{V_i \in \mathbf{Y} \cap \mathbf{X}, v_i = V_i \cap \mathbf{x}} V_{i_{[\mathbf{x}_i \setminus v_i]}} = v_i\Big) \\ &= \sum_{\mathbf{u}} \mathbf{1}\Big[\bigwedge_{V_i \in \mathbf{Y} \setminus \mathbf{X}} V_{i_{[\mathbf{x}_i]}}(\mathbf{u}) = v_i, \bigwedge_{V_i \in \mathbf{Y} \cap \mathbf{X}, v_i = V_i \cap \mathbf{x}} V_{i_{[\mathbf{x}_i \setminus v_i]}}(\mathbf{u}) = v_i\Big] P(\mathbf{u}) \end{aligned} \tag{2}$$

such that (i) $\mathbf{X}_i \subseteq \mathbf{X}$, $\mathbf{x}_i \in Val(\mathbf{X}_i)$ and $\bigcup_i \mathbf{X}_i = \mathbf{X}$; and (ii) 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 $\mathcal{M}_{\mathbf{x}_j}$, then $\mathbf{x}_i \cap B = \mathbf{x}_j \cap B$. The collection of all such distributions is denoted as $\mathbf{P}^{\mathcal{L}_{2.5}}$.

The key difference between $\mathcal{L}_{2.25}$ and $\mathcal{L}_{2.5}$ lies in how the distributions are indexed: $\mathcal{L}_{2.25}$ are indexed by specific interventional value sets $\mathbf{x} \in Val(\mathbf{X})$, while $\mathcal{L}_{2.5}$ are indexed by interventional variable sets $\mathbf{X}$. The more refined index for $\mathcal{L}_{2.25}$ creates more restrictions on the expressiveness for its distributions, which is also reflected in the differences between conditions of the two definitions. Similar to Def. 1, Cond. (i) of Def. 2 ensures that each variable in the intervention set appears at least once in the subscript. In addition, it relaxes Def. 1 by allowing multiple value assignments for

[1] Some nested counterfactuals also belong to this layer, provided that their unnested formula contain only distributions in this layer (Appendix B).

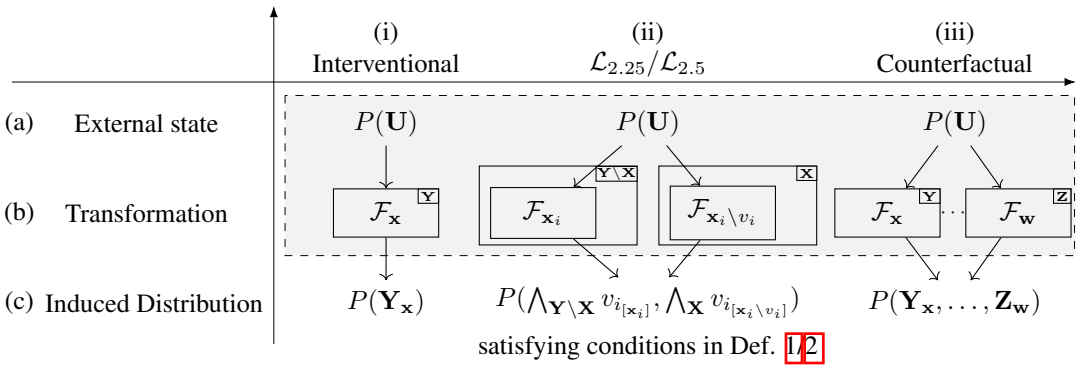


Figure 3: Given an SCM's initial state (i.e., population) (a), we show the different functional transformations (b) and the corresponding induced distribution (c) of each layer of the hierarchy. (i) represents the transformation (i.e., $\mathcal{F}$) from the natural state of the system ($P(\mathbf{U})$) to an 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.

the interventional variable set $\mathbf{X}$. Cond. (ii) is also relaxed such that value consistency enforcement for downstream variables start at children of the intervened X, instead of at X itself.

The evaluation processes for distributions in these two new layers are illustrated in Fig. 3, which are contrasted with $\mathcal{L}_2$ and $\mathcal{L}_3$. A variable in $\mathbf{Y}$ can only go through one submodel: each variable in the intervention set $\mathbf{X}$ is evaluated in its own submodel $\mathcal{M}_{\mathbf{x}_i \setminus v_i}$, corresponding to the values of $\mathbf{X}$ it takes, and each variable outside the intervention set is evaluated in its own submodel $\mathcal{M}_{\mathbf{x}_i}$ based on the value of $\mathbf{X}$ it receives. The submodels in $\mathcal{L}_{2.25}$ and $\mathcal{L}_{2.5}$ are jointly constrained to satisfy the two conditions in Def. 1 and Def. 2, respectively. By comparing the evaluation processes across different layers within PCH, we see that $\mathcal{L}_{2.25}$ and $\mathcal{L}_{2.5}$ are more expressive than $\mathcal{L}_2$ as all variables in $\mathbf{Y}$ are evaluated in a single submodel in $\mathcal{L}_2$ while $\mathcal{L}_{2.25}/\mathcal{L}_{2.5}$ are evaluated from joining multiple submodels (or counterfactual worlds). On the other hand, they are less expressive than the full-blown $\mathcal{L}_3$, as they impose conditions on which specific submodels are allowed to be joined.

Example 1 (SCM inducing $\mathcal{L}_{2.25}/\mathcal{L}_{2.5}$)**.** *Consider the SCM $\mathcal{M} = \langle \mathbf{U} = \{U_x, U_y, U_z\}, \mathbf{V} = \{X, Y, Z\}, \mathcal{F}, P(\mathbf{u})\rangle$, where $\mathcal{F} = \{X \leftarrow f_x(U_x); Z \leftarrow f_z(X, U_z); Y \leftarrow f_y(X, U_y)\}$ and $U_x \perp\!\!\!\perp U_z \perp\!\!\!\perp U_y$. $P(X, Y_x, Z_x)$, indexed by the interventional value set x, belongs to $\mathcal{L}_{2.25}$ as it satisfies Cond. (i) of Def. 1 by having only x in the subscript and Cond. (ii) by sharing consistent subscript between all children of X (i.e., Y, Z). $P(X, Y_x, Z_{x'})$, in contrast, does not belong to $\mathcal{L}_{2.25}$ as it contains conflicting value assignments for X which makes it not indexable by any specific interventional value set. However, it belongs to $\mathcal{L}_{2.5}$ as the conditions in Def. 2 allow different value assignments for the same variable in the intervention set. The $\mathcal{L}_3$ distribution $P(Y_x, Y)$ falls outside both languages as it contains the same variable Y under two different submodels, which are not allowed in $\mathcal{L}_{2.25}/\mathcal{L}_{2.5}$*[2].

2.2 Graphical Models

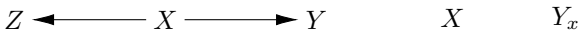
With the new collections of distributions properly defined, we introduce two graphical models to encode the corresponding constrains and compatibility relations.

Definition 3 (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:*

(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 $\mathbf{v}$, $\mathbf{C}_1, ..., \mathbf{C}_l$ 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:

$$P(\bigwedge_{W_{pa_w} \in \mathbf{W}_*} W_{pa_w}) = \prod_{j=1}^{l} P(\bigwedge_{W_{pa_w} \in \mathbf{C}_{j*}} W_{pa_w}) \tag{3}$$

[2] More examples of $\mathcal{L}_{2.25}/\mathcal{L}_{2.5}$ are provided in B.



(a) (b)

Figure 4: Sample causal diagram $\mathcal{G}$ and its AMWN $\mathcal{G}_A(\mathcal{G}, \{Y_x, X\})$

(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* $P(Y_{\mathbf{pa}_y,\mathbf{z}}, \mathbf{W}_*) \in \mathbf{P}^{\mathcal{L}_{2.25}}$*:*

$$P(Y_{\mathbf{pa}_y,\mathbf{z}}, \mathbf{W}_*) = P(Y_{\mathbf{pa}_y}, \mathbf{W}_*) \tag{4}$$

(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* $\mathbf{W}_*$ *such that* $P(Y_{\mathbf{xz}} = y, \mathbf{X_z} = \mathbf{x}, \mathbf{W}_*) \in \mathbf{P}^{\mathcal{L}_{2.25}}$*:*

$$P(Y_{\mathbf{z}} = y, \mathbf{X_z} = \mathbf{x}, \mathbf{W}_*) = P(Y_{\mathbf{xz}} = y, \mathbf{X_z} = \mathbf{x}, \mathbf{W}_*) \tag{5}$$

It can be seen that CBN2.25 closely resembles CTFBN, sharing the same types of constraints but restricted to the subset of distributions circumscribed to $\mathcal{L}_{2.25}$. Cond. (i) requires variables not sharing latent confounders to be jointly independent once their parents are fixed by intervention. Cond. (ii) state that once the parents of a variable Y have been fixed, no other intervention could affect the value of Y, regardless of any other observation. Finally, Cond. (iii) connects observations and interventions. Intuitively, if a parent X of Y has been observed taking the value x while both X and Y are under the same intervention $do(Z = z)$, then it is the same as having Y being intervened by $do(Z = z, X = x)$. Importantly, the next proposition states that a causal diagram $\mathcal{G}$ induced by an SCM $\mathcal{M}$ is a CBN2.25 for the $\mathcal{L}_{2.25}$ distribution $\mathcal{M}$ generates.

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. 11 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 2 (CBN2.25)**.** *Given the SCM in Example 1, the CBN2.25 induced =* $\langle \mathcal{G}, \mathbf{P}^{\mathcal{L}_{2.25}} \rangle$ *with* $\mathcal{G}$ *being the causal diagram in Fig. 4(a) and* $\mathbf{P}^{\mathcal{L}_{2.25}}$ *satisfying the following constraints:*

(i)[*Independence Restrictions*] $P(X, Y_x, Z_x) = P(X)P(Y_x)(Z_x)$ (6)

(ii)[*Exclusion Restrictions*] $P(X_{\mathbf{a}} = x, \mathbf{W}_*) = P(X = x, \mathbf{W}_*), \mathbf{a} \subseteq \{z, y\}$ (7)

$P(Y_{xz} = y, \mathbf{W}_*) = P(Y_x = y, \mathbf{W}_*)$ (8)

$P(Z_{xy} = z, \mathbf{W}_*) = P(Z_x = z, \mathbf{W}_*)$ (9)

(iii)[*Local Consistency*] $P(Y = y, X = x) = P(Y_x = y, X = x)$ (10)

$P(Y_z = y, X_z = x, \mathbf{W}_*) = P(Y_{zx} = y, X_z = x, \mathbf{W}_*)$ (11)

$P(Z = z, X = x, \mathbf{W}_*) = P(Z_x = z, X = x, \mathbf{W}_*)$ (12)

$P(Z_y = z, X_y = x, \mathbf{W}_*) = P(Z_{yx} = z, X_y = x, \mathbf{W}_*)$ (13)

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

Similarly for $\mathcal{L}_{2.5}$, a graphical model can be defined by imposing the same type of constraints on distributions circumscribed to $\mathcal{L}_{2.5}$. The causal diagram $\mathcal{G}$ induced by an SCM $\mathcal{M}$ is also a CBN2.5 for the $\mathcal{L}_{2.5}$ distributions $\mathcal{M}$ generates. Detailed definition and theorem are given in Appendix C.

2.3 Inferential Machinery

From the definitions of CBN2.25 and CBN2.5, we observe that constraints listed are *local*, namely, they involve counterfactual variables with their parents in the subscripts. These local constraints serve as a basis to be translated and combined to generate *global* constraints involving other variables, possibly far away in $\mathcal{G}$.

Example 3 (Local to Global Constraints)**.** *Consider the CBN2.25 from Example 2. One global constraint implied by it, which is not in the set of local constraints in CBN2.25 definition is* $P(y_z, x) = P(y, x)$*. Still, it can be derived from composition of several local constraints as shown in equations (on the right).*

$$\begin{aligned}
P(y_z, x) &= P(y_z, x_z) && (Eq.(7)) && (14)\\
&= P(y_{xz}, x_z) && (Eq.(11)) && (15)\\
&= P(y_x, x_z) && (Eq.(8)) && (16)\\
&= P(y_x, x) && (Eq.(7)) && (17)\\
&= P(y, x) && (Eq.(10)) && (18)
\end{aligned}$$

The inferential machinery associated with a graphical model is exactly to facilitate the process of composing the local constraints defined in the model to determine whether a given query can be expressed as a function of the available data. In the case of a CBN, one such machinery is Pearl's celebrated *do-calculus* [15], while for CTFBN, an example is the *ctf-calculus* [6]. As discussed earlier, the key difference between CBN2.25/CBN2.5 and CTFBN lies in the distributions where the local constraints are imposed on. Building on *ctf-calculus*, we provide a machinery for inferences within CBN2.25/CBN2.5, by restricting the rules to distributions in the corresponding layers.

Definition 4 (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)*

$$P(\mathbf{y}_{\mathbf{T}_*\mathbf{x}}, \mathbf{x}_{\mathbf{T}_*}, \mathbf{w}_*) = P(\mathbf{y}_{\mathbf{T}_*}, \mathbf{x}_{\mathbf{T}_*}, \mathbf{w}_*) \tag{19}$$

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

$$P(\mathbf{y}_\mathbf{r}|\mathbf{x}_\mathbf{t}, \mathbf{w}_*) = P(\mathbf{y}_\mathbf{r}|\mathbf{w}_*) \text{ if } (\mathbf{Y}_\mathbf{r} \perp\!\!\!\perp \mathbf{X}_\mathbf{t}|\mathbf{W}_*) \text{ in } \mathcal{G}_A \tag{20}$$

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

$$P(\mathbf{y}_{\mathbf{xz}}, \mathbf{w}_*) = P(\mathbf{y}_\mathbf{z}, \mathbf{w}_*) \text{ if } (\mathbf{X} \cap An(\mathbf{Y}) = \emptyset) \text{ in } \mathcal{G}_{\overline{\mathbf{Z}}} \tag{21}$$

where $\mathcal{G}_A$ is the AMWN $\mathcal{G}_A(\mathcal{G}, \mathbf{Y}_\mathbf{r} \cup \mathbf{X}_\mathbf{t} \cup \mathbf{W}_*)$ [3] 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 thought of as the global counterparts to the three conditions in the definitions of CBN2.25 and CBN2.5. To ensure all $P(\cdot)$ in the rules belong to $\mathbf{P}^{\mathcal{L}_{2.25}}$/$\mathbf{P}^{\mathcal{L}_{2.5}}$ given $\mathcal{G}$, we introduce a graphical criterion. For $\mathcal{L}_{2.5}$, the criterion checks the counterfactual ancestor set, but for $\mathcal{L}_{2.25}$, it also needs to check descendants of ancestors as the more restrictive CTF-RAND imposes stronger constraints over all downstream variables sharing the same intervened parents.

Definition 5 (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.*

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) Does not contain any pair of potential outcomes $R_\mathbf{s}, W_\mathbf{t}$ with inconsistent subscripts where $\mathbf{s} \cap \mathbf{T} \neq \mathbf{t} \cap \mathbf{S}$.*

Example 4. *Consider the causal diagram in Fig. 4(a) and whether $P(Z_x, Y_{x'})$ belongs to layer 2.25 induced by the corresponding SCMs. The reachability set $CRS(Z_x, Y_{x'}) = \{X, Z_x, Y_x, Z_{x'}, Y_{x'}\}$. The joint counterfactual $\{Z_x, Z_{x'}\}$ is in the reachability set with Z under different regimes. Applying Lemma 1, we conclude that $P(Z_x, Y_{x'})$ is not in the $\mathcal{L}_{2.25}$ distributions.*

With Lemma 1 to ensure that the distributions are in the corresponding layers, we can apply ctf-calculus in CBN2.25 and CBN2.5. This calculus guarantees the correctness of derivations from the available $\mathcal{L}_1$ or $\mathcal{L}_2$ distributions to a counterfactual query.

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.*

Example 5 (Effect of the Treatment on the Treated)**.** *Consider the causal diagram $\mathcal{G}$ in Fig. 4(a) and the effect of treatment on the treated (ETT), $P(y_x|x')$, given observational distributions $P(\mathbf{v})$ is available as input. The derivation following the ctf-calculus rules are:*

$$\begin{aligned} P(y_x|x') &= P(y_x) && \text{(Rule 2: } Y_x \perp\!\!\!\perp X \text{ in } \mathcal{G}_A(\mathcal{G}, \{Y_x, X\}) \text{ Fig. 4(b))} && (22)\\ &= P(y_x|x) && \text{(Rule 2: } Y_x \perp\!\!\!\perp X \text{ in } \mathcal{G}_A(\mathcal{G}, \{Y_x, X\}) \text{ Fig. 4(b))} && (23)\\ &= P(y|x) && \text{(Rule 1: Consistency)} && (24) \end{aligned}$$

where Eq. (22) and (23) are justified by Lemma 1, as $CRS(X, Y_x) = \{X, Z_x, Y_x\}$ is in $\mathcal{L}_{2.25}$.

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

3 Hierarchy of Graphical Models

In this section, we formally introduce a more refined hierarchy of graphical models defined at different layers of the PCH. We then illustrate how models in the hierarchy vary in terms of the types of queries they support and the empirical falsifiability of assumptions they encode. First, we note that the two collection of distributions defined earlier (Def. 1 and Def. 2) can be placed within the PCH.

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

The hierarchy of distributions is graphically illustrated on the left side of Fig. 5. Building on this hierarchy of distributions, we consider the constraints encoded by each graphical model. Given a causal diagram $\mathcal{G}$, the constraints it encodes arise from the interpretation of missing edges. As we move higher up in the hierarchy of graphical models, the missing edges corresponds to increasingly stronger constraints, as illustrated in example below.

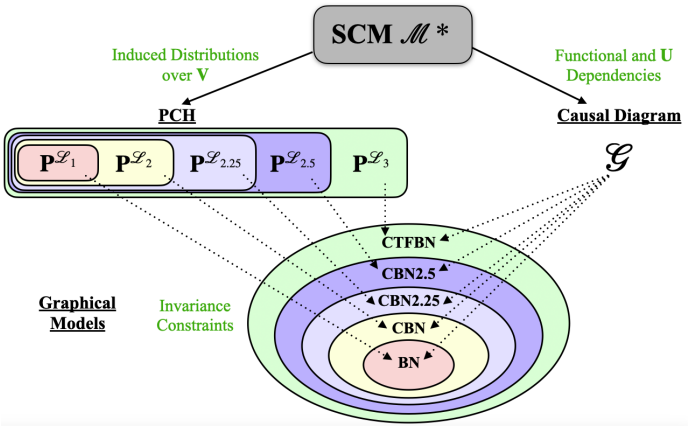


Figure 5: Pearl Causal Hierarchy (PCH*) and hierarchy of graphical models induced by an SCM

Example 6. *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):*

$$\textit{BN:} \quad P(Y|X,Z) = P(Y|X) \tag{25}$$

$$\textit{CBN:} \quad P(Y_{xz}) = P(Y_x) \tag{26}$$

$$\textit{CBN2.25:} \quad P(Y_{xz}, \mathbf{W}_*) = P(Y_x, \mathbf{W}_*), \forall \mathbf{W}_* \in \mathcal{P}(\{X, Z_x\}) \tag{27}$$

$$\textit{CBN2.5:} \quad P(Y_{xz}, \mathbf{W}_*) = P(Y_x, \mathbf{W}_*), \forall \mathbf{W}_* \in \mathcal{P}(\{X, Z_x\}) \cup \mathcal{P}(\{X, Z_{x'}\}) \tag{28}$$

$$\textit{CTFBN:} \quad P(Y_{xz}, \mathbf{W}_*) = P(Y_x, \mathbf{W}_*), \forall \mathbf{W}_* \tag{29}$$

Moving from BN to CBN adds $\mathcal{L}_2$ constraints to the assumptions set, and moving from CBN to the other three models introduces $\mathcal{L}_3$ constraints. Among the three models encoding counterfactual constraints, increasing flexibility to express richer forms of $\mathbf{W}_*$ as we move up the hierarchy corresponds to the stronger assumptions. Missing bidirected edges encode independence constraints at different layers:

$$\textit{CBN:} \quad P(Z_x) = P(Z|X=x) \tag{30}$$

$$P(Y_x) = P(Y|X=x) \tag{31}$$

$$\textit{CBN2.25:} \quad P(Z_x, Y_x, X) = P(Z_x)P(Y_x)P(X) \tag{32}$$

$$\textit{CBN2.5:} \quad P(Z_x, Y_{x'}, X) = P(Z_x)P(Y_{x'})P(X) \tag{33}$$

$$\textit{CTFBN:} \quad P(\bigwedge_{x \in Val(X)} Z_x, \bigwedge_{x' \in Val(X)} Y_{x'}, X) = P(\bigwedge_{x \in Val(X)} Z_x) P(\bigwedge_{x' \in Val(X)} Y_{x'}) P(X) \tag{34}$$

As we move up the hierarchy, independence constraints involve richer sets of variables, reflecting the increase in strength of the assumptions.

In fact, constraints encoded by graphical models higher in the hierarchy always imply those of models lower in the hierarchy. This property defines a hierarchy of graphical models, as illustrated in Fig. 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 $i \leq j$.*

As discussed earlier, the causal inference engine works when a query is matched with a model that encodes sufficient – and ideally necessary – set of assumptions, at least in principle. This matching logic can be understood from two complementary perspectives.

The first one concerns the expressiveness of the query a model can support for valid inferences. In other words, it is about the sufficiency of the assumptions in the model to enable inference based on the query. When the expressiveness of the query exceeds that of the model's assumptions, the causal inference engine fails to proceed. For example, a BN has all its constraints in $\mathcal{L}_1$, and thus it is helpless in making inference on an $\mathcal{L}_2$ query like $P(y|do(x))$. Similarly, a CBN only encodes constraints on $\mathcal{L}_2$ and it cannot help for making inference on $\mathcal{L}_3$ queries like $P(Y_x, X)$. In contrast, when a model's assumptions are expressive enough to support the query, we say there is a match between the query and the model. A CTFBN, currently sitting at the top of the graphical hierarchy, can match with the most expressive queries in the PCH[4]. This dimension of a model's capabilities is referred to as its *expressive power*: models higher in the hierarchy support more expressive queries.

The second perspective concerns whether the model encodes only the necessary assumptions for a given query, or if it is parsimonious enough. As we move up the hierarchy, models encode increasingly stronger assumptions that are harder to empirically falsify, often requiring more sophisticated experimental capabilities. If a model contains assumptions not required for the inference task, these assumptions may be unnecessarily burdensome, especially in terms of empirical falsification. Therefore, given a query, the preferred model is typically the one with the fewest unnecessary assumptions, while it can still potentially answer the query. This dimension is referred to as the model's *empirical falsifiability*: models higher in the hierarchy encode stronger, and often less falsifiable, assumptions.

Example 7 (Natural Direct Effect (NDE)). *Consider $\mathcal{G}$ in Fig. 6. The natural direct effect from X to Y, $NDE_{x,x'}(y) = P(y_{x',Z_x}) - P(y_x)$. Applying unnesting, the first term becomes $\sum_z P(y_{x'z}, z_x)$, which is ID if and only if NDE is ID. Let Q be $P(y_{x'z}, z_x)$, which is an $\mathcal{L}_{2.25}$ query in this case. Q can be identified in the CBN2.5 associated with $\mathcal{G}$ via ctf-calculus as $P(y_{x'z}, z_x)$=$P(y|x', z)P(z|x)$. The CTFBN, encoding stronger constraints than CBN2.5, can also identify Q; but it includes unnecessary assumptions like $P(Z_x, Z_{x'}, X) = P(Z_x, Z_{x'})P(X)$, with cannot be tested given the current experimental limits. The CBN associated with $\mathcal{G}$, in contrast, is not expressive enough to support inference on Q.*

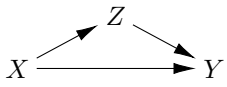


Figure 6: Causal diagram for NDE

This example further highlights the tension between the expressiveness of the queries and the models, where the optimal match occurs when the assumptions in the model are both sufficient and necessary for inference. With the two new models, the necessity boundaries in $\mathcal{L}_3$ are refined such that queries in $\mathcal{L}_{2.25}/\mathcal{L}_{2.5}$ better match with the more parsimonious, and potentially falsifiable, models CBN2.25/CBN2.5. In a nutshell, models higher in the hierarchy gain power by encoding constraints over increasingly expressive distributions. However, these gains come at the expense of increased difficulty for empirical falsification. It is therefore crucial for researchers to understand this trade-off and choose a model that appropriately balances inferential capabilities and testability for the task at hand.

Q Layer	GM	Suff.	Nec.
$\mathcal{L}_{2.5}$	CBN	x	✓
$\mathcal{L}_{2.5}$	CBN2.5	✓	✓
$\mathcal{L}_{2.5}$	CTFBN	✓	x

Table 1: Examples of Matching between Graphical Models and Queries. 'Suff.':= Sufficient and 'Nec.':= Necessary

4 Conclusions

In this paper, we introduced two new classes of graphical models, CBN2.25/CBN2.5, encoding constraints in two distinct collections of distributions that are realizable given counterfactual randomization. We proved that the models are induced naturally by SCMs (Thm. 1) and provided a sound and complete inferential machinery within them (Thm. 2). We placed the new distribution collections within the PCH (Thm. 3) and proved that graphical models encoding constraints in the PCH also form a hierarchy (Thm. 4). We then highlighted the tension between expressive power and empirical falsifiability of models in this hierarchy. We hope this work supports a deeper understanding of graphical models, and guides researchers in making more informed model selection decisions.

[4]This does not immediately imply identification, but at least in principle, certain queries can be answered.

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, 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.

[3] 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.

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

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

[6] Juan D Correa and Elias Bareinboim. Counterfactuals — A Graphical Perspective. 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] Ronald Fisher. *The Design of Experiments*. Oliver and Boyd, 1935.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

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

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

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

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

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

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

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

[21] 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.

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

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

[24] 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.

Appendices

Contents

A Background and Definitions

In this section, we introduce the basic definitions and concepts that are fundamental to this work.

A.1 SCMs and Graphical Models

Definition 6 (Structural Casual Model (SCM) [2]). *An SCM $\mathcal{M}$ is a 4-tuple $\langle \mathbf{V}, \mathbf{V}, \mathcal{F}, P(\mathbf{u}))\rangle$, 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 $\{V_1, V_2, ..., V_n\}$ 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 $\{f_1, f_2, ..., f_n\}$ such that each f_i is a mapping from (the respective domains of) $\mathbf{U}_i \cup \mathbf{Pa}_i$ to V_i, where $\mathbf{U}_i \subseteq \mathbf{U}, \mathbf{Pa}_i \subseteq \mathbf{V} \backslash V_i$, and the entire set $\mathcal{F}$ forms a mapping from $\mathbf{U}$ to $\mathbf{V}$. That is, for $i = 1, ..., n$, each $f_i \in \mathcal{F}$ is such that*

  $$v_i \leftarrow fi(\mathbf{pa}_i, \mathbf{u}_i) \tag{35}$$

  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 7 (Submodel — "Interventional SCM" [15]). *Consider an SCM $\mathcal{M} = \langle \mathbf{U}, \mathbf{V}, \mathcal{F}, P(\mathbf{u})\rangle$. Given $\mathbf{X} \subseteq \mathbf{V}$ and $\mathbf{x}$ being a particular realization of $\mathbf{X}$. A submodel $\mathcal{M}_\mathbf{x}$ of $\mathcal{M}$ is the causal model*

$$\mathcal{M}_\mathbf{x} = \langle \mathbf{U}, \mathbf{V}, \mathcal{F}_\mathbf{x}, P(\mathbf{u})\rangle, \textit{where} \tag{36}$$
$$\mathcal{F}_\mathbf{x} = \{f_i : V_i \notin \mathbf{X}\} \cup \{\mathbf{X} = \mathbf{x}\} \tag{37}$$

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

Definition 8 (Potential Outcome [15]). *Let $\mathbf{X}$ and $\mathbf{Y}$ be two sets of variables in $\mathbf{V}$, and u be a unit. The potential outcome $\mathbf{Y}_\mathbf{x}(u)$ is defined as the solution for $\mathbf{Y}$ of the set of equations $\mathcal{F}_\mathbf{x}$ with respect to SCM $\mathcal{M}$ (for short, $\mathbf{Y}_{\mathcal{M}_\mathbf{x}}(u)$). That is, $\mathbf{Y}_\mathbf{x}(u) = \mathbf{Y}_{\mathcal{M}_\mathbf{x}}(u)$.*

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

Definition 9 (Pearl Causal Hierarchy (PCH) ([2])). *An SCM $\mathcal{M} = \langle \mathbf{U}, \mathbf{V}, \mathcal{F}, P(\mathbf{u})\rangle$ induces three layers of probability distributions which 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):*

  $$\mathbf{P}^\mathcal{M}(\mathbf{y}) = \sum_\mathbf{u} \mathbf{1}[\mathbf{Y}(\mathbf{u}) = \mathbf{y}]P(\mathbf{u}) \tag{38}$$

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

  $$\mathbf{P}^\mathcal{M}(\mathbf{y}_\mathbf{x}) = \sum_\mathbf{u} \mathbf{1}[\mathbf{Y}_\mathbf{x}(\mathbf{u}) = \mathbf{y}]P(\mathbf{u}) \tag{39}$$

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

  $$\mathbf{P}^\mathcal{M}(\mathbf{y}_\mathbf{x}, ..., \mathbf{z}_\mathbf{w}) = \sum_\mathbf{u} \mathbf{1}[\mathbf{Y}_\mathbf{x}(\mathbf{u}) = \mathbf{y}, ..., \mathbf{Z}_\mathbf{w}(\mathbf{u}) = \mathbf{z}]P(\mathbf{u}) \tag{40}$$

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 $\mathbf{P}^{\mathcal{L}_3}$.

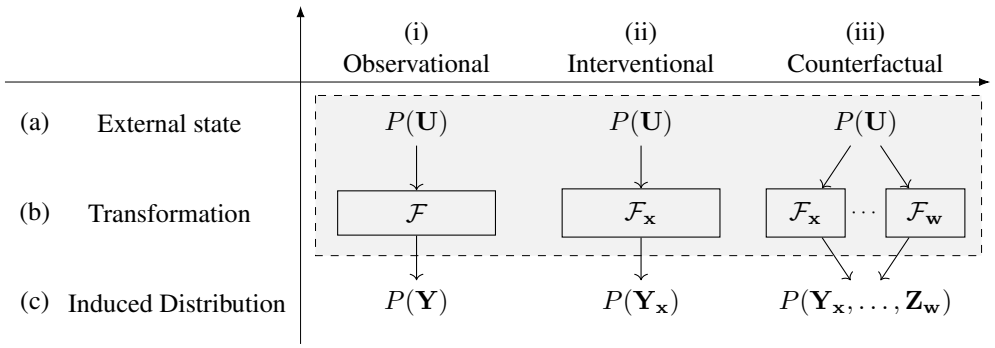


Figure 7: Given an SCM's initial state (i.e., population) (a), we show the different functional transformations (b) and the corresponding induced distribution (c) of each layer of the hierarchy. (i) represents the transformation (i.e., $\mathcal{F}$) from the natural state of the system ($P(\mathbf{U})$) to an 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).

PCH specifies both the symbolic representation and the valuations of each probabilistic quantity given an underlying SCM. If the SCM is fully specified, all conceivable quantities from any layer of the PCH are immediately computable (Fig. 7). However, in most real-life applications, only partial knowledge of the SCM is available. In order to understand what causal inference tasks are possible given this partial knowledge, we first need to analyze the marks an SCM imprints on its distributions. This type of information is called invariance constraints as defined below:

Definition 10 (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.*

For example, a common type of invariance constraint used in graphical models is conditional independence over observational distribution [2], such as $P(y|x) = P(y)$, which represents that X is probabilistically independent of Y. Invariance constraints coarsen the PCH by zooming into the relationships among different distributions while abstracting away from their specific numerical values. As more invariance constraints are included in the assumption set, the granularity of the encoded knowledge about the underlying SCM increases. Rather than enumerating each constraint individually, we leverage graphs to encode them – capitalizing on the close relationship between invariance constraints and the topological properties of graphs (specifically kinship among nodes, like parents, neighbors and ancestors, etc.).

Given an SCM $\mathcal{M}$, a graph can be constructed to capture the topological information among endogenous and exogenous variables.

Definition 11 (Causal Diagram [2])**.** *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 the corresponding functions f_i, f_j share some common $U \in \mathbf{U}$ as an argument, or the corresponding $U_i, U_j \in \mathbf{U}$ are correlated.

The pairing of such a graph with the set of invariance constraints it encodes over a collection of distributions defines a graphical model (also known as a compatibility relation).

Definition 12 (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 $\mathbf{V}$. Further, the missing edges in $\mathcal{G}$ represent certain invariance constraints within $\mathbf{P}$.*

Depending on the assumptions made on different layers of the PCH, different graphical models can be derived. Some examples of models corresponding to the three layers of the PCH are Bayesian Network (BN) [14], Causal Bayesian Network (CBN) [2] and Counterfactual Bayesian Network

(CTFBN) [6]. 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 *do-calculus* for CBN and *ctf-calculus* for CTFBN [14, 15, 5].

As we move up the ladder of PCH, the corresponding graphical models encode invariance constraints over increasingly richer sets of distributions. Naturally, these growing sets of constraints induce a hierarchy among the graphical models, where models higher in the hierarchy offer greater inferential power. However, adding more assumptions to the set of invariance constraints can also increase the difficulty of empirically verifying them.

Example 8 (Graphical Models)**.** *Consider the SCM* $\mathcal{M} = \langle \mathbf{U} = \{U_x, U_y\}, \mathbf{V} = \{X, Y\}, \mathcal{F}, P(\mathbf{u})\rangle$*, where*

$$\mathcal{F} = \begin{cases} X \leftarrow U_x \\ Y \leftarrow X \oplus U_y \end{cases} \tag{41}$$

$$P(\mathbf{u}) : U_x \sim Bernoulli(0.2), U_y \sim Bernoulli(0.3) \tag{42}$$

The endogenous variables $\mathbf{V}$ *represent, respectively, a certain treatment* X *(e.g., drug) and an outcome* Y *(survival). The exogenous variables* U_x *and* U_y *represents other variables outside the model that affect* X *and* Y*, respectively.*

The SCM $\mathcal{M}$ *is unobserved, yet it imprints constraints over the distributions it induces. And this information can be encoded as invariance constraints using the causal diagram shown in Fig. 8(c). 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: No invariance constraints*
- $\mathcal{L}_2$ *CBN:*

$$P(y|do(x)) = P(y|x) \tag{43}$$

$$P(x|do(y)) = P(x) \tag{44}$$

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

$$P(y_x, y'_{x'}, x'') = P(y_x, y'_{x'})P(x'') \tag{45}$$

$$P(x_y) = P(x) \tag{46}$$

$$P(y_x, x) = P(y, x) \tag{47}$$

The constraints in each model determines its inferential power. Given the $\mathcal{L}_1$ *constraints, no inference can be drawn as the constraint set is empty. 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* $\mathcal{L}_3$ *CTFBN, we can leverage the local constraints and infer that the effect of the treatment on the treated (ETT) is also the observational correlation (i.e.* $P(y_x|x') = P(y|x)$*).*

Another observation from this example is that as we move to graphical models in higher layers, the assumptions become richer and stronger. To verify stronger assumptions, it requires higher experimental capabilities so that the agent can obtain data corresponding to distributions in the assumptions set. However, such capabilities may not always be available. For example, consider the $\mathcal{L}_2$ *constraint* $P(y|do(x)) = P(y)$*, if intervention on* X *is not feasible in the system, then it is not possible to verify it empirically. The assumptions in* $\mathcal{L}_3$ *are even more demanding. For example, verifying the constraint* $P(y_x, y'_{x'}, x'') = P(y_x, y'_{x'})P(x'')$ *requires access to data from the counterfactual distribution* $P(Y_x, Y_{x'}, X)$*, which cannot be obtained through standard Fisherian randomization.*

In the following sections, we give the definitions and examples for graphical models introduced in previous works ([14, 15, 2, 6]).

A.2 $\mathcal{L}_1$: 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.

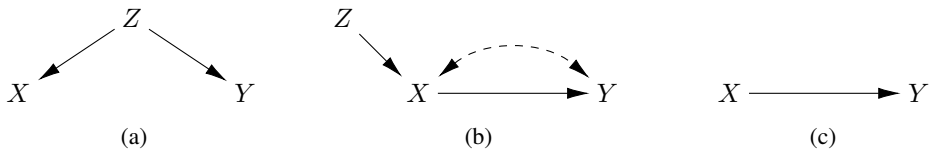


Figure 8: SCM Induced DAG or Causal Diagrams

Definition 13 (Confounded Component of an SCM [2]). *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 14 (SCM-induced DAG [2]). *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 15 (Markov Relative to [14]). *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:*

$$P(\mathbf{V}) = \prod_i P(v_i | \mathbf{pa}_i) \tag{48}$$

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

Definition 16 (Bayesian Network [14]). *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 9 (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*

$$\mathcal{F} = \begin{cases} Z \leftarrow U_z \\ X \leftarrow Z \vee U_x \\ Y \leftarrow Z \oplus U_y \end{cases} \tag{49}$$

$$P(\mathbf{u}) : U_z \sim Bernoulli(0.5), U_x \sim Bernoulli(0.5), U_y \sim Bernoulli(0.5) \tag{50}$$

Its SCM-induced DAG is shown in Fig. 8(a) and its induced observational distribution $P(\mathbf{v})$ satisfies:

$$P(\mathbf{v}) = P(z)P(x|z)P(y|z) \tag{51}$$

for all x, y, z in $Val(X) \times Val(Y) \times Val(Z)$. The DAG in Fig. 8(a) is a BN for $P(\mathbf{v})$.

A.3 $\mathcal{L}_2$: Causal Bayesian Networks

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

Definition 17 (CBN Markovian [2]). *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 $\mathcal{G}$:*

$$P(v_i | do(pa_i), do(\mathbf{x})) = P(v_i | do(pa_i)) \tag{52}$$

3. [*Parents do/see*] *For every* $V_i \in \mathbf{V}, V_i \notin \mathbf{X}$*:*

$$P(v_i|do(pa_i), do(\mathbf{x})) = P(v_i|pa_i, do(\mathbf{x})) \tag{53}$$

Example 10 (SCM-induced CBN Markovian)**.** *Consider the SCM from Example 9. Its induced causal diagram is shown in Fig. 8(a) and its induced set of interventional distributions* $\mathbf{P}_*$ *satisfy:*

1. [*Markov*]

$$P(\mathbf{v}) = P(z)P(x|z)P(y|z) \tag{54}$$
$$P(\mathbf{v}|do(x)) = P(z|do(x))P(y|z, do(x)) \tag{55}$$
$$P(\mathbf{v}|do(y)) = P(z|do(y))P(x|z, do(y)) \tag{56}$$
$$P(\mathbf{v}|do(z)) = P(x|do(z))P(y|do(z)) \tag{57}$$

2. [*Missing-link*]

$$P(x|do(y, z)) = P(x|do(z)) \tag{58}$$
$$P(y|do(x, z)) = P(y|do(z)) \tag{59}$$
$$P(z|do(\mathbf{a})) = P(z), \forall \mathbf{a} \subseteq \{x, y\} \tag{60}$$

3. [*Parents do/see*]

$$P(x|do(z)) = P(x|z) \tag{61}$$
$$P(x|do(y, z)) = P(x|z, do(y)) \tag{62}$$
$$P(y|do(z)) = P(y|z) \tag{63}$$
$$P(y|do(x, z)) = P(y|z, do(x)) \tag{64}$$

The causal diagram in Fig. 8(a) is a CBN Markovian for $\mathbf{P}_*$*.*

Definition 18 (Confounded Component [23])**.** *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 19 (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:*

$$Pa_i^+ = \cup_{j|V_j \in \mathbf{T_i}} Pa_j \backslash \{V_i\} \tag{65}$$

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 11 (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*

$$\mathcal{F} = \begin{cases} Z \leftarrow U_z \\ X \leftarrow Z \vee U \\ Y \leftarrow X \oplus U \end{cases} \tag{66}$$
$$P(\mathbf{u}) : U_z \sim Bernoulli(0.5), U \sim Bernoulli(0.5) \tag{67}$$

The causal diagram $\mathcal{G}$ *it induces is shown in Fig. 8(b). The respective augmented parents of* X, Y, Z *in* $\mathcal{G}$ *are:*

$$Pa_z^+ = \{\} \tag{68}$$
$$Pa_x^+ = \{Z\} \tag{69}$$
$$Pa_y^+ = \{X, Z\} \tag{70}$$

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:

$$Pa_z^{x+} = \{\} \tag{71}$$
$$Pa_y^{x+} = \{\} \tag{72}$$

Definition 20 (Semi-Markov Relative to [2])**.** *A probability $P(\mathbf{V})$ is said to be semi-Markov relative to a graph $\mathcal{G}$ if for any topological order $<$ of $\mathcal{G}$:*

$$P(\mathbf{V}) = \prod_i P(v_i|pa_i^+) \tag{73}$$

Definition 21 (CBN Semi-Markovian [2])**.** *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 $V_i \in \mathbf{V}\backslash\mathbf{X}$, $\mathbf{W} \subseteq \mathbf{V}\backslash(Pa_i^{\mathbf{x}+} \cup \mathbf{X} \cup \{V_i\})$:*
$$P(v_i|do(\mathbf{x}), pa_i^{\mathbf{x}+}, do(\mathbf{w})) = P(v_i|do(\mathbf{x}), pa_i^{\mathbf{x}+}) \tag{74}$$
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 $\mathcal{G}_{\overline{\mathbf{x}}}$::*
$$P(v_i|do(\mathbf{x}), pa_i^c, do(pa_i^u)) = P(v_i|do(\mathbf{x}), pa_i^c, pa_i^u) \tag{75}$$

Example 12 (SCM-induced CBN Semi-Markovian)**.** *Consider the SCM from Example 11. Its induced causal diagram is shown in Fig. 8(b) and its induced set of interventional distributions $\mathbf{P}_*$ satisfy:*

1. [*Semi-Markov*]
$$P(\mathbf{v}) = P(z)P(x|z)P(y|x,z) \tag{76}$$
$$P(\mathbf{v}|do(x)) = P(z|do(x))P(y|do(x)) \tag{77}$$
$$P(\mathbf{v}|do(y)) = P(z|do(y))P(x|z,do(y)) \tag{78}$$
$$P(\mathbf{v}|do(z)) = P(x|do(z))P(y|x,do(z)) \tag{79}$$
2. [*Missing directed-link*]
$$P(x|z,do(y)) = P(x|z) \tag{80}$$
$$P(x|do(z),do(y)) = P(x|do(z)) \tag{81}$$
$$P(y|do(x),do(z)) = P(y|do(x)) \tag{82}$$
$$P(z|do(\mathbf{a})) = P(z), \forall \mathbf{a} \subseteq \{x, y\} \tag{83}$$
3. [*Missing bidirected-link*]
$$P(x|do(z)) = P(x|z) \tag{84}$$
$$P(x|do(y,z)) = P(x|z,do(y)) \tag{85}$$
$$P(y|x,do(z)) = P(y|x,z) \tag{86}$$

The causal diagram in Fig. 8(b) is a CBN Semi-Markovian for $\mathbf{P}_*$.

A.4 $\mathcal{L}_3$: Counterfactual Bayesian Networks

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

Definition 22 (CTFBN Markovian [6])**.** *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*

$$P(\bigwedge_{W_{pa_w} \in \mathbf{W}_*} W_{pa_w}) = \prod_{V \in \mathbf{V}(\mathbf{W}_*)} P(\bigwedge_{W_{pa_w} | W \in \mathbf{V}(\mathbf{W}_*)} W_{pa_w}) \tag{87}$$

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*

$$P(Y_{\mathbf{pa}_y, \mathbf{z}}, \mathbf{W}_*) = P(Y_{\mathbf{pa}_y}, \mathbf{W}_*) \tag{88}$$

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*

$$P(Y_{\mathbf{z}} = y, \mathbf{X}_{\mathbf{z}} = \mathbf{x}, \mathbf{W}_*) = P(Y_{\mathbf{xz}} = y, \mathbf{X}_{\mathbf{z}} = \mathbf{x}, \mathbf{W}_*) \tag{89}$$

Example 13 (SCM-induced CTFBN Markovian)**.** *Consider the SCM from Example 9. Its induced causal diagram is shown in Fig. 8(a) and its induced set of counterfactual distributions* $\mathbf{P}_{\#}$ *satisfy:*

1. [*Independence Restrictions*]

$$P(z, x_z, x'_{z'}, y_{z''}, y'_{z'''}) = P(z) P(x_z, x'_{z'}) P(y_{z''}, y'_{z'''}) \tag{90}$$

2. [*Exclusion Restrictions*]

$$P(x_{yz}, \mathbf{w}_*) = P(x_z, \mathbf{w}_*) \tag{91}$$
$$P(y_{xz}, \mathbf{w}_*) = P(y_z, \mathbf{w}_*) \tag{92}$$
$$P(z_{\mathbf{a}}, \mathbf{w}_*) = P(z, \mathbf{w}_*), \forall \mathbf{a} \subseteq \{x, y\} \tag{93}$$

3. [*Local Consistency*]

$$P(x, z) = P(x_z, z) \tag{94}$$
$$P(x_y, z_y) = P(x_{yz}, z_y) \tag{95}$$
$$P(y, z) = P(y_z, z) \tag{96}$$
$$P(y_x, z_x) = P(y_{xz}, z_x) \tag{97}$$

The causal diagram in Fig. 8(a) is a CTFBN Markovian for $\mathbf{P}_{\#}$*.*

Definition 23 (CTFBN Semi-Markovian [6])**.** *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*

$$P(\bigwedge_{W_{pa_w} \in \mathbf{W}_*} W_{pa_w}) = \prod_{j=1}^{l} P(\bigwedge_{W_{pa_w} \in \mathbf{C}_{j*}} W_{pa_w}) \tag{98}$$

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*

$$P(Y_{\mathbf{pa}_y, \mathbf{z}}, \mathbf{W}_*) = P(Y_{\mathbf{pa}_y}, \mathbf{W}_*) \tag{99}$$

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*

$$P(Y_{\mathbf{z}} = y, \mathbf{X}_{\mathbf{z}} = \mathbf{x}, \mathbf{W}_*) = P(Y_{\mathbf{xz}} = y, \mathbf{X}_{\mathbf{z}} = \mathbf{x}, \mathbf{W}_*) \tag{100}$$

Example 14 (SCM-induced CTFBN Semi-Markovian)**.** *Consider the SCM from Example 11. Its induced causal diagram is shown in Fig. 8(b) and its induced set of counterfactual distributions* $\mathbf{P}_{\#}$ *satisfy:*

1. [*Independence Restrictions*]

$$P(z, x_z, x'_{z'}, y_{x''}, y'_{x'''}) = P(z)P(x_z, x'_{z'}, y_{x''}, y'_{x'''}) \tag{101}$$

2. [*Exclusion Restrictions*]

$$P(x_{yz}, \mathbf{w}_*) = P(x_z, \mathbf{w}_*) \tag{102}$$
$$P(y_{xz}, \mathbf{w}_*) = P(y_x, \mathbf{w}_*) \tag{103}$$
$$P(z_{\mathbf{a}}, \mathbf{w}_*) = P(z, \mathbf{w}_*), \forall \mathbf{a} \subseteq \{x, y\} \tag{104}$$

3. [*Local Consistency*]

$$P(x, z) = P(x_z, z) \tag{105}$$
$$P(x_y, z_y) = P(x_{yz}, z_y) \tag{106}$$
$$P(y, x) = P(y_x, x) \tag{107}$$
$$P(y_z, x_z) = P(y_{xz}, x_z) \tag{108}$$

The causal diagram in Fig. 8(b) is a CTFBN Semi-Markovian for $\mathbf{P}_{\#}$.

A.5 Counterfactual Randomization

Counterfactual randomization is an experimental procedure that allows an agent to access the value of variable before an intervention takes effect [3]. For example, the doctor may be able to learn the patient's natural choice of drug before randomly assigning a treatment to the patient in a clinical trial. The formal definition of this action in an SCM is given below.

Definition 24 (Counterfactual (ctf-) Randomization (Def. 2.3 [19])). *CTF-RAND*$(X \to \mathbf{C})^{(i)}$*: fixing the value of* X *as an input to the mechanisms generating* $\mathbf{C} \subseteq Ch(X)$ *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.*

The implementation of counterfactual randomization can be achieved under certain structural conditions, like when electroencephalogram recordings are available to measure a unit's natural decision while simultaneously intervening on the actual decision [3], or when counterfactual mediators are present to change how children of X perceive it [19]. Whether counterfactual randomization can be performed depends on the specific experimental settings.

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 25 (Maximal Feasible Action Set (SCM) [19]). *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*$(X \to C)^{(i)}, \forall X, \forall C \in Ch(X)$*: 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, READand 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 Languages for $\mathcal{L}_{2.25}$ and $\mathcal{L}_{2.5}$

B.1 Nested Counterfactuals

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 $do(\mathbf{X} = \mathbf{X_z})$, where the variable $\mathbf{X}$ is set to behave as another counterfactual variable, say $\mathbf{X_z}$. A random variable Y in such a system is represented with a counterfactual of the form $Y_{\mathbf{X_z}}$, which is called a *nested counterfactual*.

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) [5])**.** *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:*

$$P(Y_{\mathbf{T}_* X_\mathbf{z}} = y) = \sum_x P(Y_{\mathbf{T}_* x} = y, X_\mathbf{z} = x) \tag{109}$$

Given a nested counterfactual, to determine if it belongs to $\mathcal{L}_{2.25}$ and $\mathcal{L}_{2.5}$, we need to check if the unnested expression contains only distributions that belong to $\mathcal{L}_{2.25}$ and $\mathcal{L}_{2.5}$.

Definition 26 (Nested Counterfactuals in $\mathcal{L}_{2.25}$/$\mathcal{L}_{2.5}$)**.** *We say that a nested counterfactual is in $\mathcal{L}_{2.25}$/$\mathcal{L}_{2.5}$ if and only if there exists a sequence of applications of the CUT procedure that reduces it to a function of unnested counterfactuals in $\mathcal{L}_{2.25}$/$\mathcal{L}_{2.5}$.*

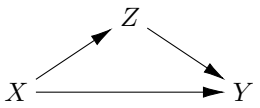


Figure 9: Causal Diagram: path $X \to Y$ represents the Natural Direct Effect (NDE)

Example 15 (Natural Direct Effect (NDE))**.** *Consider the causal diagram in Fig. 9. The natural direct effect from X to Y can be written in counterfactual language as*

$$NDE_{x,x'}(y) = P(y_{x',Z_x}) - P(y_x) \tag{110}$$

The first term is a nested counterfactual, and we can derive its unnested expression by applying CUT.

$$P(y_{x',Z_x}) = \sum_z P(y_{x'z}, z_x) \tag{111}$$

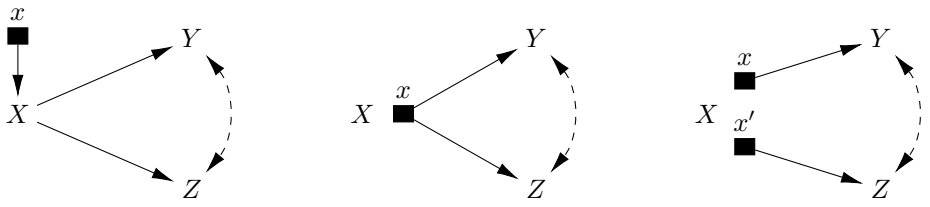
From this unnested expression, we can conclude that it is in $\mathcal{L}_{2.5}$ as $P(Y_{x'z}, Z_x)$ satisfies the conditions in Def. 2. However, it is not in $\mathcal{L}_{2.25}$ due to the conflicting subscript x and x' in the two counterfactual variables joint.

B.2 Examples for $\mathcal{L}_{2.25}$ and $\mathcal{L}_{2.5}$

Def. 1 and Def. 2 can be viewed as the template to enumerate distributions in $\mathcal{L}_{2.25}$ and $\mathcal{L}_{2.5}$. The key difference between the two layers is that $\mathcal{L}_{2.25}$ is indexed by specific interventional values, while $\mathcal{L}_{2.5}$ is indexed by interventional variables. This difference is illustrated in Example 1 where different value assignments for the interventional variable set $\mathbf{X}$ is allowed in $\mathcal{L}_{2.5}$ but not in $\mathcal{L}_{2.25}$. We further illustrate this difference in another example below.

Example 16 (Difference in Indexing between $\mathcal{L}_{2.25}$ and $\mathcal{L}_{2.5}$)**.** *Consider the SCM from Example 11 where the variables $\mathbf{V} = \{Z, X, Y\}$ form a chain $Z \to X \to Y$ topologically. Let the interventional variable set be $\{Y\}$.*

- *For $\mathcal{L}_{2.25}$, it is indexed by a specific interventional value. So we need to fix the value assignment of Y to be $y \in Val(Y)$. Then by Def. 1, $P(Z_y, X_y, Y)$, where Z and X share the same subscript, is a distribution in $\mathcal{L}_{2.25}$.*



(a) $\mathcal{L}_2$ (b) $\mathcal{L}_{2.25}$ (c) $\mathcal{L}_{2.5}$

Figure 10: Differences in how intervention on X affects downstream variables in $\mathcal{L}_2$, $\mathcal{L}_{2.25}$ and $\mathcal{L}_{2.5}$

- *For $\mathcal{L}_{2.25}$, the interventional variable can take any value in its domain unless it is constrainted by Cond. (ii) of Def. 2 when two variables are descendants of the same child of an intervened value. In this example, Z and X are not in descendants of Y. As a result, there is no constraint on value assignment to Y for Z and X. Taking any $y, y' \in Val(Y)$, $P(Z_y, X_{y'}, Y)$ is a distribution in $\mathcal{L}_{2.5}$.*

This example shows that the increased flexibility of indexing in $\mathcal{L}_{2.5}$ compared to $\mathcal{L}_{2.25}$ allows it to include more distributions.

Cond. (i) of Def. 1 and Def. 2 ensures that all variables in the intervention set must appear at least once as subscript in the counterfactuals joint. This avoids any redundant symbolic representation to appear during the enumeration of distributions in the languages, as illustrated in example below.

Example 17 (Cond. (i) of Def. 1 and Def. 2)**.** *Consider the same SCM from Example 11 where the variables $\mathbf{V} = \{Z, X, Y\}$ form a chain $Z \to X \to Y$ topologically. Given two interventional variable sets $\emptyset$ and $\{Y\}$.*

The empty interventional set gives the distribution $P(Z, X, Y)$, where all subscripts are empty. This is consistent with our understanding that empty intervention is equivalent to observation. For the interventional variable set $\{Y\}$, if Cond. (i) is not imposed, $P(Z, X, Y)$ would also be compatible with the symbolic representation for distributions in these layers. This means that the same distribution is repetitively enumerated under different interventional variable sets. To avoid this redundancy, we impose Cond. (i) to require the union of all subscripts to cover the interventional variable set. In other words, y must appear as a subscript in at least one of the counterfactuals joint. As a result, the enumeration would not produce $P(Z, X, Y)$, but rather, produces distributions like $P(Z_y, X, Y)$, $P(Z, X_y, Y)$ or $P(Z_y, X_y, Y)$ for $\mathcal{L}_{2.25}$, and also $P(Z_y, X_{y'}, Y)$ for $\mathcal{L}_{2.5}$.

Cond. (ii) of Def. 1 and Def. 2 reflects how counterfactual randomization enforces consistent values over downstream variables. For $\mathcal{L}_{2.25}$, counterfactual randomization on variable X is restrained such that all children of X shares the same value x. As a result, all descendants of X share the same value x. In contrast, counterfactual randomization in $\mathcal{L}_{2.5}$ allow each child of X to interpret X differently. Yet, given that counterfactual randomization cannot bypass a child to affect descendants directly, it still imposes a consistent value constraint over the descendants of X. This constraint starts at the children of X, instead of at X itself.

Example 18 (Cond. (ii) of Def. 1 and Def. 2)**.** *Consider the causal diagram in Fig. 4(a) and the intervention on X. In $\mathcal{L}_2$, the submodel fixes $X = x$ and we obtain the distribution $P(Y, Z|do(x))$. In $\mathcal{L}_{2.25}$, all downstream variables of X must include x in its subscript, i.e., Y_x, Z_x. At the same time, counterfactual randomization allows us to join the natural value of X with the other counterfactual variables and obtain the distribution $P(X, Y_x, Z_x)$. In $\mathcal{L}_{2.25}$, the downstream variable consistency is only enforced at the child level. In this example, different subscripts of X for Y and Z are allowed and we obtain the distribution $P(X, Y_x, Z_{x'})$. The difference between the three layers are illustrated graphically in Fig. 10.*

C Details on Models and Inferential Machinery

C.1 Details on CBN2.5

In this section, we give the detailed definition and theorem for CBN2.5.

Definition 27 (CBN2.5 Semi-Markovian)**.** *Given a mixed graph $\mathcal{G}$ and let $\mathbf{P}^{\mathcal{L}_{2.5}}$ be the collection of all $\mathcal{L}_{2.5}$ distributions. $\mathcal{G}$ is a Causal Bayesian Network 2.5 for $\mathbf{P}^{\mathcal{L}_{2.5}}$ if:*

1. [*Independence Restrictions*] *Let $\mathbf{W}_*$ be a set of counterfactuals of the form $W_{\mathbf{pa}_w}$ with distinct 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 $\mathbf{W}_*$ such that $P(\mathbf{W}_*) \in \mathbf{P}^{\mathcal{L}_{2.5}}$. Then $P(\mathbf{W}_*)$ factorizes as*

$$P(\bigwedge_{W_{pa_w} \in \mathbf{W}_*} W_{pa_w}) = \prod_{j=1}^{l} P(\bigwedge_{W_{pa_w} \in \mathbf{C}_{j*}} W_{pa_w}) \tag{112}$$

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 $\mathbf{W}_*$ such that $P(Y_{\mathbf{pa}_y, \mathbf{z}}, \mathbf{W}_*) \in \mathbf{P}^{\mathcal{L}_{2.5}}$, we have*

$$P(Y_{\mathbf{pa}_y, \mathbf{z}}, \mathbf{W}_*) = P(Y_{\mathbf{pa}_y}, \mathbf{W}_*) \tag{113}$$

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 $\mathbf{W}_*$ such that $P(Y_{\mathbf{xz}} = y, \mathbf{X}_\mathbf{z} = \mathbf{x}, \mathbf{W}_*) \in \mathbf{P}^{\mathcal{L}_{2.5}}$, we have*

$$P(Y_\mathbf{z} = y, \mathbf{X}_\mathbf{z} = \mathbf{x}, \mathbf{W}_*) = P(Y_{\mathbf{xz}} = y, \mathbf{X}_\mathbf{z} = \mathbf{x}, \mathbf{W}_*) \tag{114}$$

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. 11 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 19 (CBN2.5)**.** *Consider the SCM from Example 1.*

The Causal Diagram is induces is shown in Fig. 4(a) and the collection of realizable distributions $\mathbf{P}^{\mathcal{L}_{2.5}}$ it induces satisfies the following constraints:

1. [*Independence Restrictions*]

$$P(X, Y_x, Z_{x'}) = P(X)P(Y_x)(Z_{x'}) \tag{115}$$

2. [*Exclusion Restrictions*]

$$P(X_\mathbf{a} = x, \mathbf{W}_*) = P(X = x, \mathbf{W}_*), \mathbf{a} \subseteq \{z, y\} \tag{116}$$
$$P(Y_{xz} = y, \mathbf{W}_*) = P(Y_x = y, \mathbf{W}_*) \tag{117}$$
$$P(Z_{xy} = z, \mathbf{W}_*) = P(Z_x = z, \mathbf{W}_*) \tag{118}$$

3. [*Local Consistency*]

$$P(Y = y, X = x) = P(Y_x = y, X = x) \tag{119}$$
$$P(Y_z = y, X_z = x) = P(Y_{zx} = y, X_z = x) \tag{120}$$
$$P(Z = z, X = x) = P(Z_x = z, X = x) \tag{121}$$
$$P(Z_y = z, X_y = x) = P(Z_{yx} = z, X_y = x) \tag{122}$$

C.2 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 [19] below.

Definition 28 (Ancestors of a Counterfactual[5])**.** *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 2 (Corollary 3.7 in [19])**.** *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 20. *Consider the causal diagram in Fig. 4(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 2, we conclude that $P(Z_x, Y_{x'})$ is in the $\mathcal{L}_{2.5}$ distributions.

C.3 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 [5]) 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 $X_{\mathbf{t}}, Y_{\mathbf{r}}, \mathbf{Z}_*$,*

$$(\|X_{\mathbf{t}}\| \perp\!\!\!\perp \|Y_{\mathbf{r}}\| \mid \|\mathbf{Z}_*\|)_{\mathcal{G}_A} \implies (\|X_{\mathbf{t}}\| \perp\!\!\!\perp \|Y_{\mathbf{r}}\| \mid \|\mathbf{Z}_*\|)_{\mathbf{P}^{\mathcal{L}_3}} \tag{123}$$

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 1
- CBN2.5: $An(\mathbf{W}_*)$ satisfies Lemma 2

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

D 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.

D.1 Hierarchy of SCMs compatible with Graphs

Given a causal diagram $\mathcal{G}$ and the set of constraints it encodes, it 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 29 (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 21. *Consider the causal diagram $\mathcal{G}$ in Fig. 8 (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 $\mathcal{L}_1$. For example, all five SCMs from Example 22 are compatible with $\mathcal{G}$ on $\mathcal{L}_1$.*
- *When $\mathcal{G}$ is interpreted as an $\mathcal{L}_2$ model, it encodes the following constraints:*

$$P(y|do(x)) = P(y|x) \tag{124}$$
$$P(x|do(y)) = P(x) \tag{125}$$
$$\tag{126}$$

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 $\mathcal{M}^{(1)}, \mathcal{M}^{(2)*}$ and $\mathcal{M}^{(3)*}$ from Example 22 are all compatible with $\mathcal{G}$ on $\mathcal{L}_2$. However, SCMs $\mathcal{M}^{(2)}$ and $\mathcal{M}^{(3)}$ from Example 22 do not satisfy these constraints and are therefore not compatible with $\mathcal{G}$ on $\mathcal{L}_2$.

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.

D.2 Hierarchy of Constraints from Realizability

Another way to compare graphical models across different layers is by analyzing the empirical falsifiability of the constraints they encode. The falsifiability of a constraint does not depend on the way 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 appear 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.

Based on the results from [19], 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 the counterfactual randomization action allows us to access distributions that lie between $\mathcal{L}_2$ and $\mathcal{L}_{2.5}$. As discussed earlier, the counterfactual randomization action needed to obtain $\mathcal{L}_{2.25}$ restricts all children to have 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. 11. Specifically, the action 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 further reinforces the hierarchical structure of distributions from the perspective of realizability.

For $\mathcal{L}_3$ distributions that lie outside $\mathbf{P}^{\mathcal{L}_{2.5}}$, there is currently no known experimental procedure to sample from them. This is related to discussions about some independence constraints in CTFBN being never empirically falsifiable because they are cross-world constraints [20]. 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

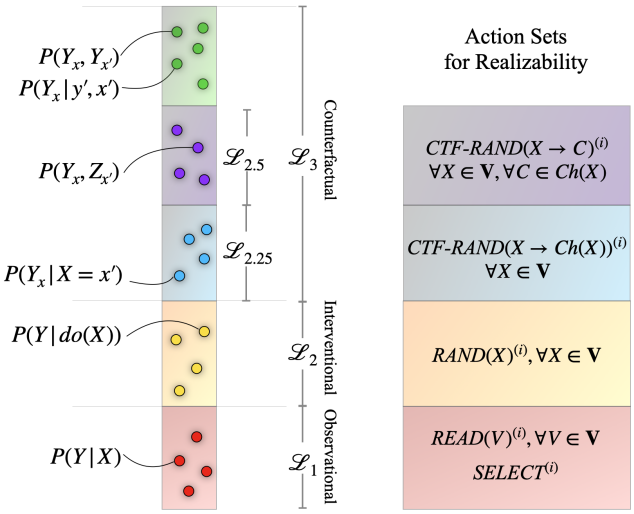


Figure 11: Hierarchy of action sets to realize distributions in different layers

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.

E Other Graphical Models

Another graphical model inducing constraints in $\mathbf{P}^{\mathcal{L}_{2.25}}$ is the Fully Randomized Causally Interpretable Structured Tree Graphs (FFRCISTG) [21, 20]. We denote the diagram used in this model as the FFRCISTG diagram. The difference between CBN2.25 and FFRCISTG stems from the diagram construction process: while an FFRCISTG diagram adds a bidirected edge between variables only when they share a common latent confounder, a causal diagram also includes a bidirected edge between two variables when there is a nonzero correlation between their latent parents.

Definition 30 (FFRCISTG Diagram (Semi-Markovian Models))**.** *Consider an SCM^F $\mathcal{M} = \langle \mathbf{U}, \mathbf{V}, \mathcal{F}, P(\mathbf{u}) \rangle$. Then $\mathcal{G}^F$ is an FFRCISTG diagram of $\mathcal{M}$ if constructed as follows:*

- *add a vertex for every endogenous variable in the set $\mathbf{V}$*
- *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*
- *Add a bidirected edge $V_i \dashleftarrow\!\dashrightarrow V_j$ for every $V_i, V_j \in \mathbf{V}$ if the corresponding functions f_i, f_j share some common $U \in \mathbf{U}$ as an argument.*

Based on the definition of FFRCISTG, it also encodes independence constraints based on the c-components of the graph [20],. However, given the difference in graph construction, FFRCISTG models do not cover the whole space of SCMs as it requires a special property on the exogenous distribution.

Definition 31 (SCM^F)**.** *Consider an SCM $\mathcal{M} = \langle \mathbf{U}, \mathbf{V}, \mathcal{F}, P(\mathbf{u}) \rangle$, it is an SCM^F if it satisfies the following constraint:*

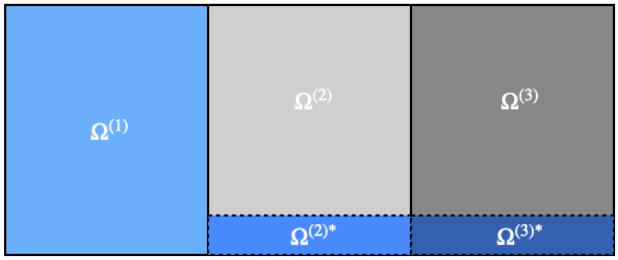


Figure 12: Space of SCMs over 2 variables X, Y.

- *for every two variables $X, Y \in \mathbf{V}$ such that their corresponding functions f_x, f_y do not share any common $U \in \mathbf{U}$ as an argument, $Y_{pa_y} \perp\!\!\!\perp X_{pa_x}$ must hold in $\mathbf{P}_F^{\mathcal{M}}$, the collection of counterfactual distributions it induces. I.e. $P(\mathbf{u})$ must satisfy:*

$$\sum_{\mathbf{u}} \mathbf{1}[X_{pa_x}(\mathbf{u}) = x \wedge Y_{pa_y}(\mathbf{u}) = y]P(\mathbf{u}) = \left[\sum_{\mathbf{u}} \mathbf{1}[X_{pa_x}(\mathbf{u}) = x]P(\mathbf{u})\right]\left[\sum_{\mathbf{u}} \mathbf{1}[Y_{pa_y}(\mathbf{u}) = y]P(\mathbf{u})\right] \tag{127}$$

We denote the subspace of all SCM^F by Ω^F.

Theorem 7 (FF-Connection — SCM FFRCISTG (Semi-Markovian))**.** *The FFRCISTG diagram $\mathcal{G}^F$ induced by an SCM $\mathcal{M}$ following the constructive procedure in Def. 30 is an FFRCISTG for $\mathbf{P}^{\mathcal{L}_{2.25}}$, the collection of $\mathcal{L}_{2.25}$ distributions induced by $\mathcal{M}$, if and only if $\mathcal{M}$ is an SCM^F.*

Proof. The proof is the same as Theorem. 1, with the extra constraint on $P(\mathbf{u})$ in SCM^F ensuring that independence constraints hold between variables without bidirected edges. □

To understand the difference between CBN2.25 and FFRCISTG, we zoom in to the space of Ω^F and compare the structural constraints encoded in FFRCISTG diagram versus causal diagram. First, we examine the space of SCMs over 2 variables. All SCMs with $\mathbf{V} = \{X, Y\}$ can be partitioned into 3 subsets, as shown in Fig. :

- $\Omega^{(1)}$: SCMs with no latent confounding (i.e. $U_x \cap U_y = \emptyset$), and $U_x \perp\!\!\!\perp U_y$
- $\Omega^{(2)}$: SCMs with no latent confounding (i.e. $U_x \cap U_y = \emptyset$), but $U_x \not\perp\!\!\!\perp U_y$
- $\Omega^{(3)}$: SCMs with latent confounding (i.e. $U_x \cap U_y \neq \emptyset$)

Within $\Omega^{(2)}$ and $\Omega^{(3)}$, there are two subsets of SCMs such that weak ignorability holds from the parametrization of U_x, U_y:

$$Y_x \perp\!\!\!\perp X, \forall x \in Val(X) \tag{128}$$

We denote them as $\Omega^{(2)*}$ and $\Omega^{(3)*}$.

The following example provides a collection of sample SCMs falling within each subset.

Example 22.

$\mathcal{M}^{(1)} \in \Omega^{(1)}$:
$\mathbf{V} = \{X, Y\}$
$\mathbf{U} = \{U_x, U_y\}$
$\mathcal{F} = \begin{cases} X \leftarrow U_x \\ Y \leftarrow X \oplus U_y \end{cases}$
$P(\mathbf{u}) : \begin{cases} U_x \sim \textit{Bernoulli}(0.7) \\ U_y \sim \textit{Bernoulli}(0.5) \end{cases}$

$\mathcal{M}^{(2)} \in \Omega^{(2)}$:
$\mathbf{V} = \{X, Y\}$
$\mathbf{U} = \{U_x, U_y, U\}$
$\mathcal{F} = \begin{cases} X \leftarrow U_x \\ Y \leftarrow X \vee U_y \end{cases}$
$P(\mathbf{u}) : \begin{cases} U_x \sim \textit{Bernoulli}(0.6) \\ U_y | U_x = 1 \sim \textit{Bernoulli}(0.7) \\ U_y | U_x = 0 \sim \textit{Bernoulli}(0.3) \end{cases}$

$\mathcal{M}^{(3)} \in \Omega^{(3)}$:
$\mathbf{V} = \{X, Y\}$
$\mathbf{U} = \{U_x, U_y\}$
$\mathcal{F} = \begin{cases} X \leftarrow U \oplus U_x \\ Y \leftarrow (X \oplus U) \vee U_y \end{cases}$
$P(\mathbf{u}) : \begin{cases} U_x \sim \textit{Bernoulli}(0.7) \\ U_y \sim \textit{Bernoulli}(0.5) \\ U \sim \textit{Bernoulli}(0.2) \end{cases}$

$$\mathcal{M}^{(2)*} \in \Omega^{(2)*}:$$
$$\mathbf{V} = \{X, Y\}$$
$$\mathbf{U} = \{U_a, U_b, U_x, U_y\}$$
$$\mathcal{F} = \begin{cases} X \leftarrow U_a \oplus U_x \\ Y \leftarrow \begin{cases} U_b \oplus U_y \text{ if } X = 0 \\ U_y \text{ if } X = 1 \end{cases} \end{cases}$$
$$P(\mathbf{u}): \begin{cases} U_x \sim \textit{Bernoulli}(0.7) \\ U_y \sim \textit{Bernoulli}(0.5) \\ U_a \sim \textit{Bernoulli}(0.5) \\ U_b = U_a \end{cases}$$

$$\mathcal{M}^{(3)*} \in \Omega^{(3)*}:$$
$$\mathbf{V} = \{X, Y\}$$
$$\mathbf{U} = \{U_x, U_y, U\}$$
$$\mathcal{F} = \begin{cases} X \leftarrow U \oplus U_x \\ Y \leftarrow \begin{cases} U \oplus U_y \text{ if } X = 0 \\ U_y \text{ if } X = 1 \end{cases} \end{cases}$$
$$P(\mathbf{u}): \begin{cases} U_x \sim \textit{Bernoulli}(0.7) \\ U_y \sim \textit{Bernoulli}(0.5) \\ U \sim \textit{Bernoulli}(0.5) \end{cases}$$

These models correspond to the ones shown in Fig. 12.

- $\mathcal{M}^{(1)} \in \Omega^{(1)}$ *is a Markovian SCM with no latent confounders and independent exogenous variables.*
- $\mathcal{M}^{(2)} \in \Omega^{(2)}$ *is an SCM with no latent confounding between the endogenous variables but the error terms are correlated. This correlation induces correlation among the counterfactual variables which results in no constraints in* $\mathbf{P}_{\#}$*. This model cannot induce any FFRCISTG given it does not satisfy the condition in Def. 31.*
- $\mathcal{M}^{(3)} \in \Omega^{(3)}$ *is an SCM with latent confounding between the two endogenous variables, but no constraint holds. The causal diagram and FFRCISTG diagram have a bidirected edge.*
- $\mathcal{M}^{(2)*} \in \Omega^{(2)*}$ *is an SCM with no latent confounders but have non-zero correlation between the error terms. It induces the constraint* $Y_x \perp\!\!\!\perp X, \forall x \in Val(X)$ *in* $\mathbf{P}_{\#}$ *from the parametrization of the functions. This model induces a Causal Diagram with bidirected edge from the correlated error terms, while the FFRCISTG diagram it induces does not have any bidirected edge.*
- $\mathcal{M}^{(3)*} \in \Omega^{(3)*}$ *is an SCM with latent confounding between the two endogenous variables. However, due to the parametrization, the constraint* $Y_x \perp\!\!\!\perp X, \forall x \in Val(X)$ *still holds. The causal diagram and FFRCISTG diagram it induces are the same with a bidirected edge between the variables.*

Given an SCM in each subset, we can construct the FFRCISTG diagram and the causal diagram it induces following Def. 30 and Def. 11. The results are summarized in Table 2, and we state four key observations from the table:

1. SCMs in $\Omega^{(2)} \backslash \Omega^{(2)*}$ fall outside Ω^F and do not satisfy Theorem 7.
   - The FFRCISTG diagram induced by any SCM in this subset does not include a bidirected edge between X and Y. However, the $\mathcal{L}_{2.25}$ distributions induced by these SCMs does not satisfy the ignorability constraint $Y_x \perp\!\!\!\perp X$. As a result, the absence of the bidirected edge creates a mismatch between the FFRCISTG diagram and the $\mathcal{L}_{2.25}$ distribution. In contrast, the causal diagram correctly represent the correlation between the exogenous parents of X and Y and the correlation between X and Y_x with its bidirected edge.
2. SCMs in $\Omega^{(1)}$, $\Omega^{(2)*}$ and $\Omega^{(3)*}$ induce the same constraints in the $\mathcal{L}_{2.25}$ distributions $\mathbf{P}^{\mathcal{L}_{2.25}}$; and among all these SCMs, those in $\Omega^{(2)*} \cup \Omega^{(3)*}$ have Lebesgue measure zero.
   - All SCMs in these subsets induce the same ignorability constraint as given in Equation (128). SCMs in $\Omega^{(1)}$ naturally satisfy this constraint from the independence of exogenous parents of X and Y. However, for SCMs in $\Omega^{(2)*} \cup \Omega^{(3)*}$ to induce such an independence constraint, a specific parametrization of the model is required. This requirement restricts these models to a set of Lebesgue measure zero.
3. Despite having the same constraints in the $\mathcal{L}_{2.25}$ distributions $\mathbf{P}^{\mathcal{L}_{2.25}}$, SCMs in $\Omega^{(2)*}$ and $\Omega^{(3)*}$ induce different FFRCISTG diagrams.

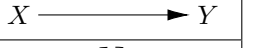
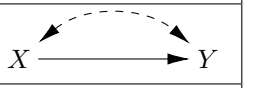
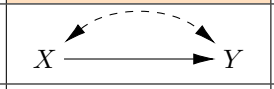
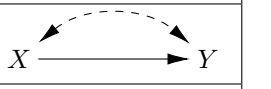
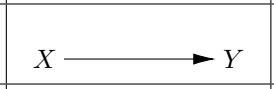
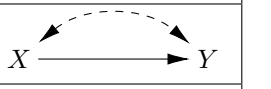
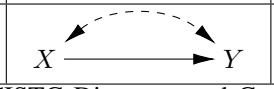
SCM Space	FFRCISTG Diagram	Causal Diagram
$\Omega^{(1)}$		
$\Omega^{(2)} \setminus \Omega^{(2)*}$		
$\Omega^{(3)} \setminus \Omega^{(3)*}$		
$\Omega^{(2)*}$		
$\Omega^{(3)*}$		

Table 2: Comparison of FFRCISTG Diagrams and Causal Diagrams induced by SCMs over 2 variables. The FFRCISTG diagram highlighted in orange fails to represent the correct constraints induced by the SCM.

- All SCMs in these two subsets induce the ignorability constraint via parametrization. However, only the FFRCISTG diagram induced by SCMs in $\Omega^{(2)*}$ encode this constraint.

4. The key difference between FFRCISTG diagram and causal diagram lies in the subset $\Omega^{(2)}$.

   - The causal diagram includes a bidirected edge to represent the correlation between the exogenous parents of X and Y. In contrast, the FFRCISTG diagram omits this bidirected edge. For SCMs in $\Omega^{(2)*}$, the omission correctly encodes the ignorability constraint. However, for SCMs in $\Omega^{(2)} \setminus \Omega^{(2)*}$, where the ignorability constraint do not hold, the omission results in a mismatch between the FFRCISTG diagram and the $\mathcal{L}_{2.25}$ distribution.

Several of the points discussed above stem from constraints imposed by the parametrization of SCMs. To formally differentiate between constraints arising from the topological structure of an SCM and those resulting from the parametrization of functions or error distributions, we extend Pearl's Definition 6.4.2 (Structurally Stable No-Confounding) [15].

Definition 32 (Structurally Stable Constraints)**.** *Given a graph $\mathcal{G}$ and a collection of distributions $\mathbf{P}$, and let $\mathcal{C}$ denote a constraint in $\mathbf{P}$. We define $\mathcal{C}$ as a structurally stable constraint if it holds in every SCM in the same NPSEM which shares the same functional arguments and dependency among exogenous variables.*

In other words, structurally stable constraints capture functional dependencies between variables in the SCM, rather than mere probabilistic dependencies. Clearly, the ignorability constraint induced by SCMs in $\Omega^{(1)}$ is a structurally stable constraint. In contrast, the same constraint induced by SCMs in $\Omega^{(2)*}$ is not structurally stable, as SCMs in $\Omega^{(2)} \setminus \Omega^{(2)*}$ belong to the same NPSEM but do not induce the ignorability constraint. The same argument applies to $\Omega^{(3)*}$.

Observation 1 confirms Theorem 7 by presenting additional examples of SCMs that fail to induce an FFRCISTG model via Def. 30 and Def. 1. Furthermore, it establishes that Ω^F, the set of SCMs that induce FFRCISTG models, is a strict subset of Ω, the space of all SCMs.

From observation 2 above, we see that with a set of SCMs that share the same constraint, the ones that induce it as a structurally stable constraint dominates. Leveraging observation 3 and 4 above, we can formally characterize the difference between FFRCISTG diagrams and causal diagrams in terms of structurally stable constraints: given an SCM $\mathcal{M}$, its causal diagram only encodes all structurally stable constraints on the collection of distributions induced by $\mathcal{M}$, whereas its FFRCISTG diagram also encode structurally unstable constraints arising from correlated error terms.

F Proofs for Theorems

F.1 Supporting Lemmas

Lemma F.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. 7, $\mathcal{M}_{\mathbf{x}}$ 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. 11, 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 2. *Condition (ii) of Def. 1 and Def. 2 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 $v_i \in \mathbf{x}_j$.*

$\mathcal{L}_{2.25}$*: 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 F.1 □

Lemma F.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. 1. We prove that after applying the exclusion operator to $\mathbf{W}_*$, the distribution still satisfies both conditions of Def. 1.

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. 1 still holds.

Given that $P(\mathbf{W}_*)$ also satisfies Cond. (ii) of Def. 1 and by Cor. 2, 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. 1. As a result, whenever, the antecedent of Cond. (ii) of Def. 1 holds, t still belongs to the subscript of W_i. So Cond. (ii) of Def. 1 still holds.

Given that $P(\|\mathbf{W}_*\|)$ satisfies both conditions of Def. 1, it is in $\mathbf{P}^{\mathcal{L}_{2.25}}$. □

Lemma F.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 F.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. 2. □

Lemma F.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 1 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. 1 and belongs to $\mathbf{P}^{\mathcal{L}_{2.25}}$. □

Lemma F.5 (ctf-calculus — do-calculus reduction (Lemma 6 in [5]))**.** *ctf-calculus subsumes do-calculus.*

Lemma F.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 F.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* [23] 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 33 (Ancestral components [5])**.** *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 $\mathbf{A}_{1*}, \mathbf{A}_{2*}, \dots$ 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 F.7 (Ancestral Set Factorization (Lemma 3 in [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,*

$$P(\mathbf{W}_* = \mathbf{w}_*) = P(\bigwedge_{W_\mathbf{t} \in \mathbf{W}_*} W_{\mathbf{pa}_w} = w) \tag{129}$$

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 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 F.8 (C-component Factorization (Lemma 4 in [5]))**.** *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*

$$P(\mathbf{W}_* = \mathbf{w}_*) = \prod_j P(\mathbf{C}_{j_*} = \mathbf{c}_{j_*}) \tag{130}$$

Lemma F.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. 5, and for $\mathcal{L}_{2.5}$, $An(An(\mathbf{W}_*)) = An(\mathbf{W}_*)$ by Def. 28. Thus, $\mathbf{W}_*$ satisfies Lemma 1 if and only if $An(\mathbf{W}_*)$ satisfies Lemma 1. □

Lemma F.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. 1/Def. 2 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. □

Algorithm 2 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 $P(\mathbf{Y}_* = \mathbf{y}_*) \leftarrow \sum_{\mathbf{w}_* \setminus \mathbf{y}_*} \prod_i P(\mathbf{C}_{i*} = \mathbf{c}_{i*})$.

Lemma F.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. 1/Def. 2. 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. 1/Def. 2. □

Lemma F.12 (Consistency (Lemma 1 in [5]))**.** *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,*

$$P(Y_{\mathbf{T}_*}, X_{\mathbf{T}_*} = x) = P(Y_{\mathbf{T}_* x}, X_{\mathbf{T}_*} = x), \tag{131}$$

where $\mathbf{T}_*$ represent any combination of counterfactuals based on $\mathbf{T}$.

Lemma F.13 (Exclusion operator (Lemma 2 in [5]))**.** *Let $Y_{\mathbf{x}}$ be a counterfactual variable, $\mathcal{G}$ a causal diagram, and*

$$Y_{\mathbf{z}} \text{ such that } \mathbf{Z} = \mathbf{X} \cap An_{\mathcal{G}_{\overline{\mathbf{X}}}}(Y) \text{ and } \mathbf{z} = \mathbf{x} \cap \mathbf{Z}. \tag{132}$$

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 F.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 F.9 and Lemma F.10. □

Algorithm 3 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 $\langle \mathcal{G}, \mathbb{Z} \rangle$.
1: Let $\mathbf{A}_{1*}, \mathbf{A}_{2*}, \ldots$ be the ancestral components of $\mathbf{Y}_* \cup \mathbf{X}_*$ given $\mathbf{X}_*$.
2: Let $\mathbf{D}_*$ be the union of the ancestral components containing a variable in $\mathbf{Y}_*$ and $\mathbf{d}_*$ the corresponding set of values.
3: let $Q \leftarrow$ CTFIDU$(\bigcup_{D_t \in \mathbf{D}_*} \mathbf{D}_d, \mathbf{d}_*, \mathbb{Z}, \mathcal{G})$.
4: **return** $\sum_{\mathbf{d}_* \setminus (\mathbf{y}_* \cup \mathbf{x}_*)} Q / \sum_{\mathbf{d}_* \setminus \mathbf{x}_*} Q$.

F.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. 11 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. 3 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. 1. 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. 1,

(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. 1, for any counterfactual set $\mathbf{W}_*$ such that $P(Y_{\mathbf{pa}_y, \mathbf{z}} = y, \mathbf{W}_* = \mathbf{w}_*) \in \mathbf{P}^{\mathcal{L}_{2.25}}$,

$$P(Y_{\mathbf{pa}_y, \mathbf{z}} = y, \mathbf{W}_* = \mathbf{w}_*) = \sum_{\mathbf{u}} \mathbf{1}(Y_{\mathbf{pa}_y, \mathbf{z}}(\mathbf{u}) = y, \mathbf{W}_*(\mathbf{u}) = \mathbf{w}_*) P(\mathbf{u}) \tag{133}$$

$$= \sum_{\mathbf{u}} \mathbf{1}(Y_{\mathbf{pa}_y}(\mathbf{u}) = y, \mathbf{W}_*(\mathbf{u}) = \mathbf{w}_*) P(\mathbf{u}) \tag{134}$$

$$= P(Y_{\mathbf{pa}_y} = y, \mathbf{W}_* = \mathbf{w}_*) \tag{135}$$

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 $Y \in \mathbf{V}, \mathbf{X} \subseteq \mathbf{Pa}_y, \mathbf{Z} \subseteq \mathbf{V} \setminus (\mathbf{X} \cup \{Y\})$, $\mathbf{R} = \mathbf{Pa}_y \setminus (\mathbf{X} \cup \mathbf{Z})$, we have

$$Y_{\mathbf{z}}(\mathbf{u}) = f_Y(\mathbf{z} \cap \mathbf{pa}_y, \mathbf{X}_{\mathbf{z}}(\mathbf{u}), \mathbf{R}_{\mathbf{z}}(\mathbf{u}), \mathbf{u}(\mathbf{U}_y)) \tag{136}$$

$$= f_Y(\mathbf{z} \cap \mathbf{pa}_y, \mathbf{x}, \mathbf{R}_{\mathbf{z}}(\mathbf{u}), \mathbf{u}(\mathbf{U}_y)) \tag{137}$$

$$= Y_{\mathbf{zx}}(\mathbf{u}) \tag{138}$$

Then by Def. 1, for any counterfactual set $\mathbf{W}_*$ such that $P(Y_{\mathbf{z}} = y, \mathbf{X}_{\mathbf{z}} = \mathbf{x}, \mathbf{W}_* = \mathbf{w}_*) \in \mathbf{P}^{\mathcal{L}_{2.25}}$,

$$P(Y_{\mathbf{z}} = y, \mathbf{X}_{\mathbf{z}} = \mathbf{x}, \mathbf{W}_* = \mathbf{w}_*) = \sum_{\mathbf{u}} \mathbf{1}(Y_{\mathbf{z}}(\mathbf{u}) = y, \mathbf{X}_{\mathbf{z}}(\mathbf{u}) = \mathbf{x}, \mathbf{W}_*(\mathbf{u}) = \mathbf{w}_*) P(\mathbf{u}) \tag{139}$$

$$= \sum_{\mathbf{u}} \mathbf{1}(Y_{\mathbf{xz}}(\mathbf{u}) = y, \mathbf{X}_{\mathbf{z}}(\mathbf{u}) = \mathbf{x}, \mathbf{W}_*(\mathbf{u}) = \mathbf{w}_*) P(\mathbf{u}) \tag{140}$$

$$= P(Y_{\mathbf{xz}} = y, \mathbf{X}_{\mathbf{z}} = \mathbf{x}, \mathbf{W}_* = \mathbf{w}_*) \tag{141}$$

which proves the consistency restrictions are satisfied. □

Definition 34 (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 1.

(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 F.4.

Therefore, $P(\mathbf{W}_*)$ satisfies conditions in Def. 1 and belongs to $\mathbf{P}^{\mathcal{L}_{2.25}}$.

The graphical check for $\mathcal{L}_{2.5}$ is proved in Corollary 3.7 of [19]. □

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 F.12 for Rule 1, Lemma F.14 for Rule 2 and Lemma F.13 for Rule 3.

To prove that it is complete, we rely on the completeness of the CTFID algorithm reproduced as Algo. 3 and Algo. 2 [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. 3 are justfied by Lemma F.9 and Lemma F.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. 2 is justified by rule 3 of the ctf-calculus and Lemma F.2 and Lemma F.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 [23] which soundness can be justified with do-calculus [9], which in turn is subsumed by ctf-calculus 2.25 by Lemma F.6. At line 17, the algorithm returns the result as a product that is justified by Lemma F.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 $\mathcal{L}_3$ [2], 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 $\mathbf{P}^{\mathcal{L}_{2.25}}$:

$$P(\mathbf{Y}=\mathbf{y}|do(\mathbf{X}=\mathbf{x})) = \sum_{X \in \mathbf{Y} \cap \mathbf{X}} P(\bigwedge_{V_i \in \mathbf{Y} \setminus \mathbf{X}} V_{i_{[\mathbf{x}]}} = v_i, \bigwedge_{V_i \in \mathbf{Y} \cap \mathbf{X}, v_i = V_i \cap \mathbf{x}} V_{i_{[\mathbf{x} \setminus v_i]}} = v_i) \quad (142)$$

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. 1.

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. 1, it also satisfies Def. 2. First, note that the key difference between Def. 1 and Def. 2 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. 1 must also satisfy the two conditions in Def. 2.

For Cond. (i), both languages require the subscripts to cover the whole space of $\mathbf{X}$. However, Def. 1 is stronger by restricting the value assignments to the set $\mathbf{x}$, while Def. 2 allows $\mathbf{x}_i$ to take different values from $Val(\mathbf{X}_i)$. Thus, if Cond. (i) of Def. 1 holds, Cond. (i) of Def. 2 immediately holds.

For Cond. (ii) and by Cor. 2, the antecedent in Def. 2 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. 1 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. 1 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. 2 is triggered, Cond. (ii) of Def. 1 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 F.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.

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 3: Summary of how missing edges are interpreted in graphical models at different layers

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 3. 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. 11 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 $\_LL_{2.5}$ instead of $\_LL_{2.25}$. □

G 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 [15, 22, 17]. 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 [2, 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 [12, 11, 10, 13, 24].

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 [14]. 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 [2]. 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. 13, 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 [2, 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. 13, the invariance constraint $P(Y|do(X)) = P(Y|X)$ 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 [3]. This work was further extended in [19]. 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. 14. 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

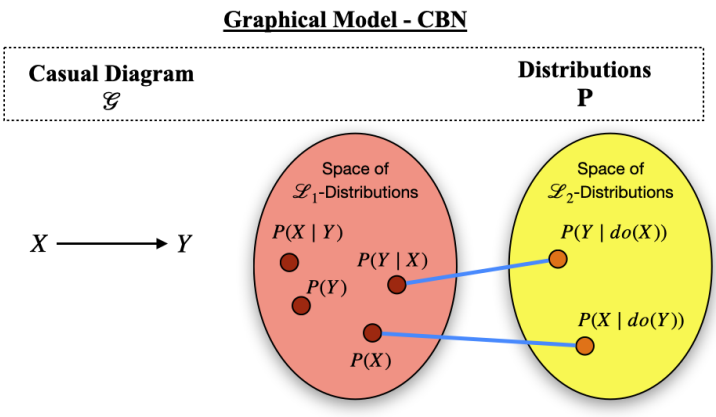


Figure 13: A CBN is a pair $\langle \mathcal{G}, \mathbf{P} \rangle$. Blue lines represent invariant constraints in $\mathbf{P}$, which are represented by features from $\mathcal{G}$: missing directed edge from Y to X corresponds to the invariance constraint $P(X|do(Y)) = P(X)$ and missing bidirected edge between X and Y corresponds to the invariance constraint $P(Y|do(X)) = P(Y|X)$.

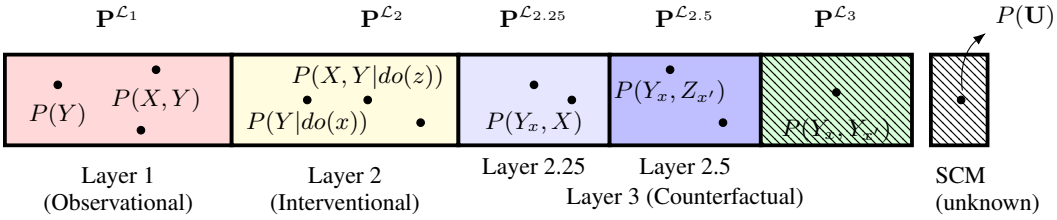


Figure 14: Pearl Causal Hierarchy (PCH*) induced by an unknown SCM $\mathcal{M}$. Layers 1 and 2 are realizable, and Layer 3 is partially realizable. The realizable portion of Layer 3 are further refined into two new layers: 2.25 and 2.5.

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 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. 10. 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 [19].

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 4: 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.

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 [19]. 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 4 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

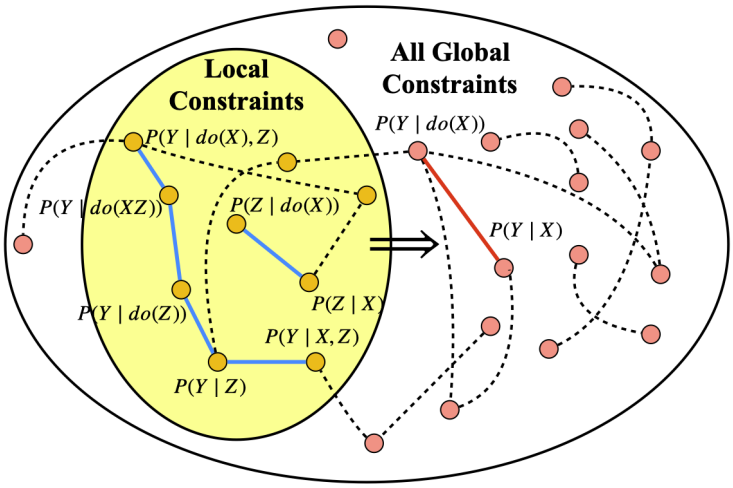


Figure 15: Constraints listed in the definition of a graphical model serves as a local basis that implies all constraints encoded in the model. Blue lines represent a set of local invariance constraints that can be composed to imply the global constraint represented by the red line.

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. 5, 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 [18]. 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. 15. 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 3. We revisit this idea with a new example in Fig. 15. 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.

$$
\begin{aligned}
P(y|do(x)) &= \sum_z P(y|do(x), z)P(z|do(x)) && \text{(Probability Axiom)} && (143)\\
&= \sum_z P(y|do(xz))P(z|do(x)) && \text{(Cond. (iii) of Def. 17)} && (144)\\
&= \sum_z P(y|do(z))P(z|do(x)) && \text{(Cond. (ii) of Def. 17)} && (145)\\
&= \sum_z P(y|z)P(z|x) && \text{(Cond. (iii) of Def. 17)} && (146)\\
&= \sum_z P(y|xz)P(z|x) && \text{(Cond. (i) of Def. 17)} && (147)\\
&= P(y|x) && \text{(Probability Axiom)} && (148)
\end{aligned}
$$

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.