Causal Abstraction Inference under Lossy Representations

Kevin Xia [1] **Elias Bareinboim** [1]

Abstract

The study of causal abstractions bridges two integral components of human intelligence: the ability to determine cause and effect, and the ability to interpret complex patterns into abstract concepts. Formally, causal abstraction frameworks define connections between complicated low-level causal models and simple high-level ones. One major limitation of most existing definitions is that they are not well-defined when considering lossy abstraction functions in which multiple low-level interventions can have different effects while mapping to the same high-level intervention (an assumption called the abstract invariance condition). In this paper, we introduce a new type of abstractions called *projected abstractions* that generalize existing definitions to accommodate lossy representations. We show how to construct a projected abstraction from the low-level model and how it translates equivalent observational, interventional, and counterfactual causal queries from low to high-level. Given that the true model is rarely available in practice we prove a new graphical criteria for identifying and estimating high-level causal queries from limited low-level data. Finally, we experimentally show the effectiveness of projected abstraction models in high-dimensional image settings.

1. Introduction

The ability to determine cause and effect, and the ability to interpret complex patterns into abstract concepts, are two integral components of human intelligence. From the causality perspective, causal reasoning is vital in planning courses of actions, determining blame and responsibility, and generalizing across changing environments. From the abstraction perspective, humans generally grasp better intuition when understanding something at a high-level. For example, a human can easily parse the object in an image as a dog or a car instead of interpreting it as a collection of pixel values. Combining these two modes of reasoning is vital for building more advanced AI systems.

Causal inference is often studied under the semantics of structural causal models (SCMs) (Pearl, 2000). An SCM models reality with a collection of mechanisms and exogenous distributions. Each SCM induces a collection of distributions categorized into three successively more descriptive layers known as the Ladder of Causation or Pearl Causal Hierarchy (PCH) (Pearl & Mackenzie, 2018; Bareinboim et al., 2022). These three layers refer to the observational ($\mathcal{L}_1$), interventional ($\mathcal{L}_2$), and counterfactual ($\mathcal{L}_3$) distributions. In many causal inference tasks, the goal is to infer a quantity from a higher layer using data from lower layers, a problem known as *cross-layer inference*. It is understood that it is generally impossible to infer higher layer information without additional assumptions (a result known as the Causal Hierarchy Theorem or CHT (Bareinboim et al., 2022)), so understanding the necessary assumptions for performing inferences is a key component of any causal inference task.

Existing works on causal abstractions have made significant progress in defining abstraction principles, proving insightful properties, and learning abstraction functions in practice (Rubenstein et al., 2017; Beckers & Halpern, 2019; Beckers et al., 2019; Geiger et al., 2023; Massidda et al., 2023; Zennaro et al., 2023; Felekis et al., 2024). Causal abstractions are typically studied by comparing a high-level model $\mathcal{M}_H$, defined over high-level variables $\mathbf{V}_H$, with its low-level counterpart $\mathcal{M}_L$, defined over $\mathbf{V}_L$. An abstraction function τ maps from $\mathbf{V}_L$ to $\mathbf{V}_H$, and $\mathcal{M}_H$ is formally defined as an abstraction of $\mathcal{M}_L$ if it satisfies key properties with respect to τ such as commutativity with interventions. More recently, this notion has been relaxed to only enforcing properties between distributions of $\mathcal{M}_H$ and $\mathcal{M}_L$ from the PCH (Xia & Bareinboim, 2024). For example, rather than saying $\mathcal{M}_H$ is a full abstraction of $\mathcal{M}_L$, one can say that $\mathcal{M}_H$ is an abstraction of $\mathcal{M}_L$ specifically for interventional quantities in $\mathcal{L}_2$ or for a single causal effect $P(y \mid do(x)) \in \mathcal{L}_2$. Xia & Bareinboim (2024) also shows the synergy between causal abstraction theory and representation learning (Bengio et al., 2013), which has shown great success in many deep learning

[1]CausalAI Lab, Columbia University. Correspondence to: Kevin Xia <kmx2000@columbia.edu>.

CausalAI Lab Technical Report, New York, United States.

applications by mapping high-dimensional data like images or text to simpler representation spaces. These definitions of causal abstractions have accomplished formalizing a broad topic of human intelligence into mathematical language.

One particular limitation of existing definitions of abstractions is known as the Abstract Invariance Condition (AIC), which states, informally, that two values cannot be abstracted together if they have different downstream impacts. This is illustrated in Fig. 1. For example, a nutritionist may have collected data on two types of cholesterol, HDL and LDL, and are studying their impact on heart disease (Steinberg, 2007; Truswell, 2010). They would like to abstract the two together by summing them as total cholesterol (TC). However, this violates the AIC, as it is known that HDL decreases rate of heart disease while LDL increases it, so the sum is ambiguous (a lossy representation).[1] Nonetheless, it may still be desirable to have a consistent formalism in which these kinds of ambiguous abstractions are well-defined, since in many practical settings (where representation learning or dimensionality reduction is needed), the AIC is clearly violated or is impossible to verify.

In this paper, we study this extension of causal abstractions, which we later define as *projected abstractions*, referring to the idea that an abstraction that violates the AIC results in a loss of information that is then characterized in the exogenous space. The proposed formalism generalizes abstractions both on the SCM and on the PCH level to allow for mathematically consistent abstractions even with AIC violations. Projected abstractions have many uses in practice, resulting in tractable causal inference and high-quality causal sampling even in the presence of extreme dimensionality reduction, a result which we show in the experiments.

To summarize, in Sec. 2, we generalize abstractions to settings which the AIC does not hold and provide an algorithm for constructing the high-level model. In Sec. 3, we show how to perform causal inference from data within this class of abstractions when the true model is not observed. In Sec. 4, we empirically demonstrate the power of abstractions at performing causal inference in high-dimensional image settings. All proofs can be found in App. A.

1.1. Preliminaries

We now introduce the notation and definitions used throughout the paper. We use uppercase letters (X) to denote random variables and lowercase letters (x) to denote corresponding values. Similarly, bold uppercase ($\mathbf{X}$) and lowercase ($\mathbf{x}$) letters denote sets of random variables and values respectively. We use $\mathcal{D}_X$ to denote the domain of X and $\mathcal{D}_{\mathbf{X}} = \mathcal{D}_{X_1} \times \cdots \times \mathcal{D}_{X_k}$ for the domain of $\mathbf{X} = \{X_1, \ldots, X_k\}$. We denote $P(\mathbf{X} = \mathbf{x})$ (often shortened to $P(\mathbf{x})$) as the probability of $\mathbf{X}$ taking the values $\mathbf{x}$ under the distribution $P(\mathbf{X})$.

[1] See App. C Ex. 6 for a more concrete explanation.

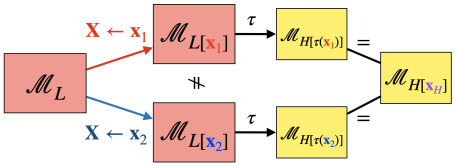

Figure 1: An illustration of AIC violations. On the low level, two different interventions may be performed (e.g., $\mathbf{X} \leftarrow \mathbf{x}_1$ and $\mathbf{X} \leftarrow \mathbf{x}_2$). However, after applying the abstraction function τ to obtain the high-level model, both interventions are mapped to the same result ($\tau(\mathbf{x}_1) = \tau(\mathbf{x}_2) = \mathbf{x}_H$). If $\mathcal{M}_L$ behaves differently under $\mathbf{x}_1$ compared to $\mathbf{x}_2$, $\mathcal{M}_H$ cannot stay consistent with both models.

We utilize the basic semantic framework of structural causal models (SCMs) (Pearl, 2000), following the presentation in Bareinboim et al. (2022).

Definition 1 (Structural Causal Model (SCM))**.** An SCM $\mathcal{M}$ is a 4-tuple $\langle \mathbf{U}, \mathbf{V}, \mathcal{F}, P(\mathbf{U}) \rangle$, where $\mathbf{U}$ is a set of exogenous variables (or "latents") that are determined by factors outside the model; $\mathbf{V}$ is a set $\{V_1, V_2, \ldots, V_n\}$ of (endogenous) variables of interest that are determined by other variables in the model – that is, in $\mathbf{U} \cup \mathbf{V}$; $\mathcal{F}$ is a set of functions $\{f_{V_1}, f_{V_2}, \ldots, f_{V_n}\}$ such that each f_{V_i} is a mapping from (the respective domains of) $\mathbf{U}_{V_i} \cup \mathbf{Pa}_{V_i}$ to V_i, where $\mathbf{U}_{V_i} \subseteq \mathbf{U}$, $\mathbf{Pa}_{V_i} \subseteq \mathbf{V} \setminus V_i$, and the entire set $\mathcal{F}$ forms a mapping from $\mathbf{U}$ to $\mathbf{V}$. That is, for $i = 1, \ldots, n$, each $f_{V_i} \in \mathcal{F}$ is such that $v_i \leftarrow f_{V_i}(\mathbf{pa}_{V_i}, \mathbf{u}_{V_i})$; and $P(\mathbf{U})$ is a probability function defined over the domain of $\mathbf{U}$. ■

Each $\mathcal{M}$ induces a causal diagram $\mathcal{G}$, where every $V_i \in \mathbf{V}$ is a vertex, there is a directed arrow $(V_j \rightarrow V_i)$ for every $V_i \in \mathbf{V}$ and $V_j \in \mathbf{Pa}_{V_i}$, and there is a dashed-bidirected arrow $(V_j \dashleftarrow\!\dashrightarrow V_i)$ for every pair $V_i, V_j \in \mathbf{V}$ such that $\mathbf{U}_{V_i}$ and $\mathbf{U}_{V_j}$ are not independent (Markovianity is not assumed). Our treatment is constrained to *recursive* SCMs, which implies acyclic causal diagrams, with finite discrete domains over endogenous variables $\mathbf{V}$.

Counterfactual (and also interventional and observational) quantities can be computed from SCM $\mathcal{M}$ as follows:

Definition 2 (Layer 3 Valuation (Bareinboim et al., 2022, Def. 7))**.** An SCM $\mathcal{M}$ induces layer $\mathcal{L}_3(\mathcal{M})$, a set of distributions over $\mathbf{V}$, each with the form $P(\mathbf{Y}_*) = P(\mathbf{Y}_{1[\mathbf{x}_1]}, \mathbf{Y}_{2[\mathbf{x}_2],\ldots})$ such that

$$P^{\mathcal{M}}(\mathbf{y}_{1[\mathbf{x}_1]}, \mathbf{y}_{2[\mathbf{x}_2]}, \ldots) = \int_{\mathcal{D}_{\mathbf{U}}} \mathbf{1}\left[\mathbf{Y}_{1[\mathbf{x}_1]}(\mathbf{u}) = \mathbf{y}_1, \mathbf{Y}_{2[\mathbf{x}_2]}(\mathbf{u}) = \mathbf{y}_2, \ldots\right] dP(\mathbf{u}) \quad (1)$$

where $\mathbf{Y}_{i[\mathbf{x}_i]}(\mathbf{u})$ is evaluated under $\mathcal{F}_{\mathbf{x}_i} := \{f_{V_j} : V_j \in \mathbf{V} \setminus \mathbf{X}_i\} \cup \{f_X \leftarrow x : X \in \mathbf{X}_i\}$. $\mathcal{L}_2$ is the subset of $\mathcal{L}_3$ for

which all $\mathbf{x}_i$ are equal, and $\mathcal{L}_1$ is the subset for which all $\mathbf{X}_i = \emptyset$. ■

Each $\mathbf{Y}_i$ corresponds to a set of variables in a world where the original mechanisms f_X are replaced with constants $\mathbf{x}_i$ for each $X \in \mathbf{X}_i$; this is also known as the mutilation procedure. This procedure corresponds to interventions, and we use subscripts to denote the intervening variables (e.g. $\mathbf{Y}_{\mathbf{x}}$) or subscripts with brackets when the variables are indexed (e.g. $\mathbf{Y}_{1[\mathbf{x}_1]}$). For instance, $P(y_x, y'_{x'})$ is the probability of the joint counterfactual event $Y = y$ had X been x and $Y = y'$ had X been x'.

We use the notation $\mathcal{L}_i(\mathcal{M})$ to denote the set of $\mathcal{L}_i$ distributions from $\mathcal{M}$. We use $\mathbb{Z}$ to denote a set of quantities from Layer 2 (i.e. $\mathbb{Z} = \{P(\mathbf{V}_{\mathbf{z}_k})\}_{k=1}^{\ell}$), and $\mathbb{Z}(\mathcal{M})$ denotes those same quantities induced by SCM $\mathcal{M}$ (i.e. $\mathbb{Z}(\mathcal{M}) = \{P^{\mathcal{M}}(\mathbf{V}_{\mathbf{z}_k})\}_{k=1}^{\ell}$).

The theory of causal abstractions developed in this paper build on the foundations of constructive abstraction functions, under which individual distributions of the PCH are well-defined between low and high-level models.

Definition 3 (Inter/Intravariable Clusterings (Xia & Bareinboim, 2024, Def. 5))**.** Let $\mathcal{M}$ be an SCM over $\mathbf{V}$.

1. A set $\mathbb{C}$ is said to be an intervariable clustering of $\mathbf{V}$ if $\mathbb{C} = \{\mathbf{C}_1, \mathbf{C}_2, \dots \mathbf{C}_n\}$ is a partition of a subset of $\mathbf{V}$. $\mathbb{C}$ is further considered admissible w.r.t. $\mathcal{M}$ if for any $\mathbf{C}_i \in \mathbb{C}$ and any $V \in \mathbf{C}_i$, no descendent of V outside of $\mathbf{C}_i$ is an ancestor of any variable in $\mathbf{C}_i$. That is, there exists a topological ordering of the clusters of $\mathbb{C}$ relative to the functions of $\mathcal{M}$.

2. A set $\mathbb{D}$ is said to be an intravariable clustering of variables $\mathbf{V}$ w.r.t. $\mathbb{C}$ if $\mathbb{D} = \{\mathbb{D}_{\mathbf{C}_i} : \mathbf{C}_i \in \mathbb{C}\}$, where $\mathbb{D}_{\mathbf{C}_i} = \{\mathcal{D}^1_{\mathbf{C}_i}, \mathcal{D}^2_{\mathbf{C}_i}, \dots, \mathcal{D}^{m_i}_{\mathbf{C}_i}\}$ is a partition (of size m_i) of the domains of the variables in $\mathbf{C}_i$, $\mathcal{D}_{\mathbf{C}_i}$ (recall that $\mathcal{D}_{\mathbf{C}_i}$ is the Cartesian product $\mathcal{D}_{V_1} \times \mathcal{D}_{V_2} \times \cdots \times \mathcal{D}_{V_k}$ for $\mathbf{C}_i = \{V_1, V_2, \dots, V_k\}$, so elements of $\mathcal{D}^j_{\mathbf{C}_i}$ take the form of tuples of the value settings of $\mathbf{C}_i$). ■

Definition 4 (Constructive Abstraction Function (Xia & Bareinboim, 2024, Def. 6))**.** A function $\tau : \mathcal{D}_{\mathbf{V}_L} \to \mathcal{D}_{\mathbf{V}_H}$ is said to be a constructive abstraction function w.r.t. inter/intravariable clusters $\mathbb{C}$ and $\mathbb{D}$ iff

1. There exists a bijective mapping between $\mathbf{V}_H$ and $\mathbb{C}$ such that each $V_{H,i} \in \mathbf{V}_H$ corresponds to $\mathbf{C}_i \in \mathbb{C}$;

2. For each $V_{H,i} \in \mathbf{V}_H$, there exists a bijective mapping between $\mathcal{D}_{V_{H,i}}$ and $\mathbb{D}_{\mathbf{C}_i}$ such that each $v^j_{H,i} \in \mathcal{D}_{V_{H,i}}$ corresponds to $\mathcal{D}^j_{\mathbf{C}_i} \in \mathbb{D}_{\mathbf{C}_i}$; and

3. τ is composed of subfunctions $\tau_{\mathbf{C}_i}$ for each $\mathbf{C}_i \in \mathbb{C}$ such that $\mathbf{v}_H = \tau(\mathbf{v}_L) = (\tau_{\mathbf{C}_i}(\mathbf{c}_i) : \mathbf{C}_i \in \mathbb{C})$, where $\tau_{\mathbf{C}_i}(\mathbf{c}_i) = v^j_{H,i}$ if and only if $\mathbf{c}_i \in \mathcal{D}^j_{\mathbf{C}_i}$. We also apply the same notation for any $\mathbf{W}_L \subseteq \mathbf{V}_L$ such that $\mathbf{W}_L$ is a union of clusters in $\mathbb{C}$ (i.e. $\tau(\mathbf{w}_L) = (\tau_{\mathbf{C}_i}(\mathbf{c}_i) : \mathbf{C}_i \in \mathbb{C}, \mathbf{C}_i \subseteq \mathbf{W}_L)$). ■

Finally, we state the AIC formally below.

Definition 5 (Abstract Invariance Condition (AIC))**.** Let $\mathcal{M}_L = \langle \mathbf{U}_L, \mathbf{V}_L, \mathcal{F}_L, P(\mathbf{U}_L) \rangle$ be an SCM and $\tau : \mathcal{D}_{\mathbf{V}_L} \to \mathcal{D}_{\mathbf{V}_H}$ be a constructive abstraction function relative to $\mathbb{C}$ and $\mathbb{D}$. The SCM $\mathcal{M}_L$ is said to satisfy the abstract invariance condition (AIC, for short) with respect to τ if, for all $\mathbf{v}_1, \mathbf{v}_2 \in \mathcal{D}_{\mathbf{V}_L}$ such that $\tau(\mathbf{v}_1) = \tau(\mathbf{v}_2)$, $\forall \mathbf{u} \in \mathcal{D}_{\mathbf{U}_L}, \mathbf{C}_i \in \mathbb{C}$, the following holds:

$$\begin{aligned} &\tau_{\mathbf{C}_i}\left(\left(f^L_V(\mathbf{pa}^{(1)}_V, \mathbf{u}_V) : V \in \mathbf{C}_i\right)\right) \\ &= \tau_{\mathbf{C}_i}\left(\left(f^L_V(\mathbf{pa}^{(2)}_V, \mathbf{u}_V) : V \in \mathbf{C}_i\right)\right), \end{aligned} \tag{2}$$

where $\mathbf{pa}^{(1)}_V$ and $\mathbf{pa}^{(2)}_V$ are the values corresponding to $\mathbf{v}_1$ and $\mathbf{v}_2$. Then, $\widetilde{\mathbf{pa}}_V$ is used to denote any arbitrary value s.t. $\tau(\widetilde{\mathbf{pa}}_V) = \tau(\mathbf{pa}^{(1)}_V) = \tau(\mathbf{pa}^{(2)}_V)$. ■

2. Abstractions under AIC Violations

The abstract invariance condition (AIC) states, in words, that two low-level values cannot map to the same high-level value if they have different downstream effects. This is a critical property that must hold for existing definitions of abstractions to be well-defined. In this paper, we will use the following running example to illustrate the key points.

Example 1. For concreteness, consider a setting in which different insurance companies (Z) offer various insurance plans (X), which affect whether an insurance claim is approved (Y). For simplicity, suppose there are two insurance companies (z_1 and z_2) that offer three insurance plans (x_1, x_2, and x_3), and the claim is either approved ($Y = 1$) or not approved ($Y = 0$). Suppose the true model $\mathcal{M}^* = \mathcal{M}_L = \langle \mathbf{U}_L, \mathbf{V}_L, \mathcal{F}_L, P(\mathbf{U}_L) \rangle$ is described as

$$\begin{aligned} \mathbf{U}_L &= \{U_Z, U^{z_1}_X, U^{z_2}_X, U^{x_1}_Y, U^{x_2}_Y, U^{x_3}_Y\} \\ \mathbf{V}_L &= \{Z, X, Y\} \\ \mathcal{F}_L &= \begin{cases} f^L_Z(u_Z) & = u_Z \\ f^L_X(z, u^{z_1}_X, u^{z_2}_X) & = u^z_X \\ f^L_Y(x, u^{x_1}_Y, u^{x_2}_Y, u^{x_3}_Y) & = u^x_Y \end{cases} \\ P(\mathbf{U}_L) &= \begin{cases} P(U_Z = z_1) = 0.5 \\ P(U^{z_1}_X) = \{x_1 \to 0.4; x_2 \to 0.1; x_3 \to 0.5\} \\ P(U^{z_2}_X) = \{x_1 \to 0.1; x_2 \to 0.4; x_3 \to 0.5\} \\ P(U^{x_1}_Y = 1) = 0.9, P(U^{x_2}_Y = 1) = 0.1, \\ P(U^{x_3}_Y = 1) = 0.9 \end{cases} \end{aligned} \tag{3}$$

The interpretation of the model is as follows: Insurance plans x_1 and x_3 are very effective, with 0.9 probability of

claim acceptance, while x_2 is very ineffective at only 0.1 probability. Insurance company z_1 is more reputable than z_2 and is more likely to offer plan x_1 over x_2, while company z_2 prefers to offer plan x_2 over x_1.

Suppose an important factor of consideration not shown in the model is that x_1 and x_2 are cheaper insurance plans, while x_3 is more expensive. A data scientist who is studying this model may choose to abstract the different plans away, categorizing them simply as "cheap" and "expensive" plans. Formally, they would study a set of higher-level variables $\mathbf{V}_H = \{Z_H, X_H, Y_H\}$, where $Z_H = Z$, $Y_H = Y$, and X_H has a domain $\mathcal{D}_{X_H} = \{x_C, x_E\}$ corresponding to cheap and expensive plans respectively. There exists an abstraction function $\tau : \mathcal{D}_{\mathbf{V}_L} \to \mathcal{D}_{\mathbf{V}_H}$ such that τ maps x_1 and x_2 to x_C (cheap) and maps x_3 to x_E (expensive). We will use the notation Z and Y instead of Z_H and Y_H since the variables are the same.

This immediately brings the AIC into question. If the data scientist is interested in the causal effect of cheap plans on claim acceptance (i.e., $P(Y_{X_H=x_C} = 1)$), whether x_C refers to x_1 or x_2 is ambiguous. To witness, note that

$$P(Y_{X_L=x_1} = 1) = 0.9 \tag{4}$$
$$P(Y_{X_L=x_2} = 1) = 0.1. \tag{5}$$

Since $\tau(x_1) = \tau(x_2) = x_C$, but $P(Y_{x_1}) \neq P(Y_{x_2})$, the AIC is clearly violated, leaving the intervention on x_C ambiguous. ■

Fundamentally, the issue with AIC violations is clear: formal definitions of abstractions expect an equality between low-level and corresponding high-level quantities, but it is not well-defined when one high-level quantity corresponds to multiple differing low-level quantities. In practice, the AIC can be a difficult restriction. Generally, it is assumed to be true whenever abstractions are applied, but it is difficult to verify given that the true SCM and functions are rarely available in real-world settings. The assumption is also likely to be incorrect when applying abstractions naïvely, for example, by performing representation learning or dimensionality reduction without taking the AIC into account. By definition, dimensionality reduction is a lossy transformation of the original data, and the AIC is violated if any of the lost information is relevant for downstream functions.

Even when the AIC does not hold, it does not necessarily mean that these lossy transformations should not be used. Representation learning and dimensionality reduction are often performed to improve tractability or interpretability at the cost of some lost information. Hence, it would still be desirable to perform causal inferences in the high-level space even under AIC violations. To address the issue of different low-level quantities matching the same high-level quantity, one can reinterpret the high-level quantity as a distribution over its corresponding low-level quantities, where the randomness in the distribution results from the lost information from the abstraction (i.e., a hard intervention on the high-level translates to a soft intervention on the low-level).

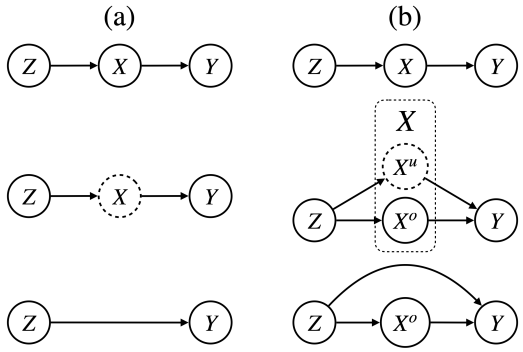

Figure 2: Comparison between (a) full SCM projections and (b) partial SCM projections. When X is fully projected away, its function is subsumed by its child's function f_Y. When X is partially projected, it is split into observed portion X^o and unobserved portion X^u. The role of X^o is preserved, while X^u is subsumed into the function f_Y.

2.1. Projected Abstractions

The discussion on relaxing the AIC begins with the concept of SCM projections (Lee & Bareinboim, 2019), which can be viewed as a primitive form of abstraction. An SCM $\mathcal{M}$ projected to a subset of variables $\mathbf{W} \subseteq \mathbf{V}$ is a functionally identical SCM defined over $\mathbf{W}$, where the functions of $\mathbf{V} \setminus \mathbf{W}$ are subsumed by other downstream functions (see App. A Def. 4 for the full definition and App. C Ex. 7 for an example). In the context of constructive abstraction functions, the act of projecting away a variable can be viewed as excluding the variable from all intervariable clusters. This brings the first major insight in addressing AIC violations. In general, when reducing the granularity of a variable, some parts of the variable deemed less important are abstracted away while others are retained. While by definition, SCM projections only allow for entire variables to be included or excluded, one could conceive of SCM projections in which variables are only partially projected away (see App. C Ex. 8 for an example). Formally, partial SCM projections can be defined as follows.

Proposition 1 (Partial SCM Projection)**.** *Let $\mathbf{V}$ be a set of variables and $\mathbf{W} \subseteq \mathbf{V}$ be a subset. For each $W_i \in \mathbf{W}$, let $\delta_i : \mathcal{D}_{W_i^o} \times \mathcal{D}_{W_i^v} \to \mathcal{D}_{W_i}$ be a surjective function mapping new variables W_i^o and W_i^v to W_i. W_i^o and W_i^v are called the observed and unobserved projections of W_i respectively. Denote $\delta(\mathbf{W}^o, \mathbf{W}^u) = \mathbf{W}$, where $\mathbf{W}^o = \{W_i^o : W_i \in \mathbf{W}\}$ and $\mathbf{W}^u = \{W_i^u : W_i \in \mathbf{W}\}$. For any SCM $\mathcal{M} = \langle \mathbf{U}, \mathbf{V}, \mathcal{F}, P(\mathbf{U}) \rangle$, there exists an SCM $\mathcal{M}' = \langle \mathbf{U}' = \mathbf{U} \cup \mathbf{W}^u, \mathbf{V}' = \mathbf{W}^o, \mathcal{F}', P(\mathbf{U}') \rangle$ such that,*

for all $\mathbf{u} \in \mathcal{D}_{\mathbf{U}}$, $\mathbf{X} \subseteq \mathbf{W}$, *and* $\mathbf{x} \in \mathcal{D}_{\mathbf{X}}$,

$$\mathbf{w}^o_{\mathbf{x}} = \mathcal{M}'_{[\mathbf{x}^o]}(\mathbf{u}, \mathbf{x}^u, \mathbf{z}^u), \tag{6}$$

where $\delta(\mathbf{w}^o_{\mathbf{x}}, \mathbf{w}^u_{\mathbf{x}}) = \mathbf{W}_{\mathbf{x}}(\mathbf{u})$, $\delta(\mathbf{x}^o, \mathbf{x}^u) = \mathbf{x}$, $\mathbf{Z}^u = \mathbf{W}^u \setminus \mathbf{X}^u$, *and* $\mathbf{z}^u$ *are the corresponding values from* $\mathbf{w}^u_{\mathbf{x}}$. $\mathcal{M}'$ *is called a partial SCM projection of* $\mathcal{M}$ *over* $\mathbf{W}^o$. ■

In words, a partial SCM projection of $\mathcal{M}$ over $\mathbf{W}^o$ is essentially a smaller version of $\mathcal{M}$ defined only on the variables of $\mathbf{W} \subseteq \mathbf{V}$, where each $W_i \in \mathbf{W}$ is only partially represented in the projection. A function δ splits W_i's domain into its observed (W^o_i) and unobserved (W^u_i) portions. Eq. 6 ensures that any value of $\mathbf{W}^o$ obtained from an intervention on the original SCM $\mathcal{M}_{\mathbf{x}}$ will match the corresponding output from $\mathcal{M}'$, when the observed portion of the intervention $\mathbf{x}^o$ is applied to $\mathcal{M}'$, while the unobserved portions of $\mathbf{x}^u$ and $\mathbf{w}^u$ are passed as unobserved arguments to the functions. A comparison between regular SCM projections and partial SCM projections is shown in Fig. 2. The definition of projected abstractions follow.

Definition 6 (Projected Abstraction)**.** An SCM $\mathcal{M}_H$ is a projected abstraction of $\mathcal{M}_L$ if and only if it is a partial SCM projection of a τ-abstraction (Beckers & Halpern, 2019, Def. 3.13) of $\mathcal{M}_L$. ■

Projected abstractions make an important step to working around the AIC as Eq. 6 allows for quantities to be well-defined between low and high-level variables by simply obtaining a partial projection of the original SCM $\mathcal{M}_L$ over the high-level variables $\mathbf{V}_H$. However, unlike full SCM projections, partial SCM projections are not unique in terms of the induced PCH distributions. Prop. 1 guarantees its existence but is underspecified in a couple of ways. First, $P(\mathbf{U}')$ is not fully defined, and it is not clear how $\mathbf{W}^u$ should be sampled. Second, Eq. 6 does not specify what behavior $\mathcal{M}'$ should follow when $\mathbf{z}^u$ does not match $\mathbf{w}^{\mathbf{u}}_{\mathbf{x}}$.

The specific choice of partial SCM projection that best serves as an abstraction can be determined by understanding how low-level interventions relate to high-level interventions. In other words, given a high-level intervention $\mathbf{X}_H \leftarrow \mathbf{x}_H$, it is important to define the corresponding low-level soft-intervention $\sigma_{\mathbf{X}_L}$, which is a distribution over all possible interventions $\mathbf{x}_L$ that map to $\mathbf{x}_H$. The consequence of the underspecification of partial SCM projections is that there are many possible choices of defining $\sigma_{\mathbf{X}_L}$. For a full discussion on how $\sigma_{\mathbf{X}_L}$ should be decided, see App. B. A useful general form of $\sigma_{\mathbf{X}_L}$ is defined as follows. Split $\sigma_{\mathbf{X}_L}$ into individual soft interventions $\sigma_{\mathbf{C}_i}$ for each intervariable cluster $\mathbf{C}_i \subseteq \mathbf{X}_L$. Then define each $\sigma_{\mathbf{C}_i}$ as

$$P(\sigma_{\mathbf{C}_i} = \mathbf{c}_i) = P(\mathbf{c}_i \mid \tau(\mathbf{c}_i) = v_{H,i}, \mathbf{pa}_{V_{H,i}}, \mathbf{u}^c_{V_{H,i}}). \tag{7}$$

In words, a high-level intervention should be equivalent to a distribution over the corresponding low-level interventions that assigns probability to each possible intervention based on their prior probabilities given their parents.[2]

Example 2. Continuing Example 1, suppose the data scientist is interested in the causal effect of choosing a cheap insurance plan on claim approval. In other words, she would like to study the intervention $X_H \leftarrow x_C$, which is ambiguous on the low-level as it could refer to either $X_L \leftarrow x_1$ or $X_L \leftarrow x_2$. More specifically, according to Eq. 7, $X_H \leftarrow x_C$ corresponds to a soft intervention σ_{X_C} on the low level, defined as

$$\sigma_{X_L} = \begin{cases} x_1 & \text{w.p. } P(x_1 \mid X_L \in \{x_1, x_2\}, z) \\ x_2 & \text{w.p. } P(x_2 \mid X_L \in \{x_1, x_2\}, z) \end{cases} \tag{8}$$

While there are many ways to disambiguate whether x_C is referring to x_1 or x_2, this choice of σ_{X_L} will assign probabilities based on the prior probabilities of X_L being one of x_1 or x_2. Moreover, the probabilities change depending on the value of z. This makes intuitive sense, since under the intervention $X_H \leftarrow x_C$, we expect that if $Z = z_1$, then X_L is more likely to be x_1 than x_2, or vice-versa when $Z = z_2$. From a query perspective, this implies that

$$\begin{aligned} &P(Y_{X_H = x_C} = 1 \mid Z = z_1) \\ &= P(Y_{\sigma_{X_L}(x_C, Z)} = 1 \mid Z = z_1) \\ &= \sum_{x_i \in \{x_1, x_2\}} P(x_i \mid X_L \in \{x_1, x_2\}, z_1) P(Y_{x_i} = 1) = 0.74 \end{aligned} \tag{9}$$

$$\text{Likewise, } P(Y_{X_H = x_C} = 1 \mid Z = z_2) = 0.26 \tag{10}$$

■

While projected abstractions are defined over the entire SCM, the mapping between low and high-level interventions are more clear at the query-level (i.e., individual interventional and counterfactual distributions of interest). Such quantities can be defined as follows.

Definition 7 (Generalized Query)**.** Denote $\mathbf{Y}_{L,*}$ as a set of counterfactual variables over $\mathbf{V}_L$. That is,

$$\mathbf{Y}_{L,*} = \left(\mathbf{Y}_{L,1[\sigma_{\mathbf{X}_{L,1}}]}, \mathbf{Y}_{L,2[\sigma_{\mathbf{X}_{L,2}}]}, \dots \right), \tag{11}$$

where each $\mathbf{Y}_{L,i[\sigma_{\mathbf{X}_{L,i}}]}$ corresponds to the potential outcomes of the variables $\mathbf{Y}_{L,i}$ under the (possibly soft) intervention $\sigma_{\mathbf{X}_{L,i}}$ over $\mathbf{X}_{L,i}$. Each $\mathbf{Y}_{L,i}$ and $\mathbf{X}_{L,i}$ must be unions of clusters from $\mathbb{C}$ (i.e. $\mathbf{Y}_{L,i} = \bigcup_{\mathbf{C} \in \mathbb{C}'} \mathbf{C}$ for some $\mathbb{C}' \subseteq \mathbb{C}$) such that $\tau(\mathbf{Y}_{L,i})$ and $\tau(\mathbf{X}_{L,i})$ are well-defined (i.e. $\tau(\mathbf{Y}_{L,i}) = \left(\bigwedge_{\mathbf{C} \in \mathbb{C}'} \tau_{\mathbf{C}}(\mathbf{C})\right)$). For the high-level counterpart, denote

$$\mathbf{Y}_{H,*} = \tau(\mathbf{Y}_{L,*}) = \left(\mathbf{Y}_{H,1[\mathbf{x}_{H,1}]}, \mathbf{Y}_{H,2[\mathbf{x}_{H,2}]}, \dots \right), \tag{12}$$

[2] Here, $\mathbf{u}^c_{V_{H,i}}$ can informally be thought of as the confounded exogenous parents of $V_{H,i}$. The full definition is somewhat involved, and the subtleties are discussed in App. B.2. Due to space constraints, the main body provides intuition in Markovian settings, where unobserved confounding is not present.

such that $\mathbf{Y}_{H,i} = \tau(\mathbf{Y}_{L,i})$, and $\mathbf{X}_{H,i} = \tau(\mathbf{X}_{L,i})$ for all i. For any value $\mathbf{y}_{H,*} \in \mathcal{D}_{\mathbf{Y}_{H,*}}$, denote

$$\mathcal{D}_{\mathbf{Y}_{L,*}}(\mathbf{y}_{H,*}) = \{\mathbf{y}_{L,*} : \mathbf{y}_{L,*} \in \mathcal{D}_{\mathbf{Y}_{L,*}}, \tau(\mathbf{y}_{L,*}) = \mathbf{y}_{H,*}\}, \tag{13}$$

that is, the set of all values $\mathbf{y}_{L,*}$ such that $\tau(\mathbf{y}_{L,*}) = \mathbf{y}_{H,*}$. ■

This query definition connects the distributions of $\mathcal{L}_3(\mathcal{M}_H)$ to corresponding distributions of $\mathcal{L}_3(\mathcal{M}_L)$. Compared to earlier definitions, Eq. 11 has been generalized to account for soft interventions in addition to hard interventions. Under constructive abstractions functions τ, a notion of Q-τ consistency was established for certain queries $Q \in \mathcal{L}_3(\mathcal{M}_L)$ (App. A Def. 13), which still apply under this generalized definition. In short, for a low level query $Q = \sum_{\mathbf{y}_{L,*} \in \mathcal{D}_{\mathbf{Y}_{L,*}}(\mathbf{y}_{H,*})} P(\mathbf{Y}_{L,*} = \mathbf{y}_{L,*})$ and its high-level counterpart $\tau(Q) = P(\mathbf{Y}_{H,*} = \mathbf{y}_{H,*})$, $\mathcal{M}_H$ is said to be Q-τ consistent with $\mathcal{M}_L$ if $Q^{\mathcal{M}_L} = \tau(Q)^{\mathcal{M}_H}$. One can then say that $\mathcal{M}_H$ is an abstraction of $\mathcal{M}_L$ specifically for the query Q, even if $\mathcal{M}_H$ may not be Q'-τ consistent with $\mathcal{M}_L$ for other query choices Q'. If $\mathcal{M}_H$ is Q-τ consistent with $\mathcal{M}_L$ for all $\tau(Q) \in \mathcal{L}_i(\mathcal{M}_H)$, then $\mathcal{M}_H$ is said to be $\mathcal{L}_i$-τ consistent with $\mathcal{M}_L$.

With $\sigma_{\mathbf{X}_{L,i}}$ defined in Eq. 7, one can then algorithmically construct a projected abstraction consistent in all queries. Given $\mathcal{M}_L$ and a constructive abstraction function τ (which may not satisfy the AIC), Alg. 1 can be used to construct the high-level abstraction $\mathcal{M}_H$. In line 4, each $W \in \mathbf{V}_L$ is split into its observed and unobserved counterparts W^o and W^u. Line 8 assigns each W^u a distribution based on Eq. 7. Line 9 builds the high-level function using the low-level function with inputs reconstructed using δ. Finally, the full high-level model $\mathcal{M}_H$ is assembled and returned in line 10. Under these inputs, Alg. 1 constructs a projected abstraction $\mathcal{M}_H$ that is Q-τ consistent with $\mathcal{M}_L$ for all possible high-level $\mathcal{L}_3$ queries, as shown by the following result.

Theorem 1. *The SCM $\mathcal{M}_H$ constructed by Alg. 1 is a projected abstraction of $\mathcal{M}_L$ that is Q-τ consistent with $\mathcal{M}_L$ for all $\tau(Q) \in \mathcal{L}_3(\mathcal{M}_H)$.* ■

3. Projected Abstraction Inference

Alg. 1 finds an abstraction model $\mathcal{M}_H$ that is consistent with its low-level counterpart $\mathcal{M}_L$ for all queries, but it requires the full specification of $\mathcal{M}_L$. In practice, $\mathcal{M}_L$ typically represents the true model of reality and will not be observed. Inferences of $\mathcal{L}_2$ and $\mathcal{L}_3$ queries must be made through limited available data, usually observational ($\mathcal{L}_1$).

The Causal Hierarchy Theorem (Bareinboim et al., 2022, Thm. 1) states that cross-layer inference, or inferring higher layer quantities (e.g., $\mathcal{L}_2$, $\mathcal{L}_3$) from lower layer data (e.g., $\mathcal{L}_1$), is generally impossible without additional assumptions.

Algorithm 1 Constructing $\mathcal{M}_H$ from $\mathcal{M}_L$.

input $\mathcal{M}_L = \langle \mathbf{U}_L, \mathbf{V}_L, \mathcal{F}_L, P(\mathbf{U}_L) \rangle$, constructive abstraction function τ from clusters $\mathbb{C}$ and $\mathbb{D}$
1: $\mathbf{U}_H \leftarrow \mathbf{U}_L, P(\mathbf{U}_H) \leftarrow P(\mathbf{U}_L)$
2: $\mathbf{V}_H \leftarrow \mathbb{C}, \mathcal{D}_{\mathbf{V}_H} \leftarrow \mathbb{D}$
3: **for** $W \in \mathbf{V}_L$ **do**
4: $\quad W^o, W^u \leftarrow \texttt{project}(W)$ {from Prop. 1}
5: $\quad \mathbf{U}_H \leftarrow \mathbf{U}_H \cup \{W^u\}$
6: **end for**
7: **for** $\mathbf{C}_i \in \mathbb{C}$ (and corresponding $V_i \in \mathbf{V}_H$) **do**
8: $\quad P(\delta(\mathbf{c}_i^o, \mathbf{C}_i^u) = \mathbf{c}_i \mid \mathbf{U}_L) \leftarrow P(\mathbf{C}_i = \mathbf{c}_i \mid \tau(\mathbf{c}_i) = v_i, \mathbf{pa}_{V_i}, \mathbf{u}^c_{V_{H,i}})$ {from Eq. 7}
9: $\quad f_i^H \leftarrow \tau(f_V^L(\delta(\mathbf{pa}_V^o, \mathbf{pa}_V^u), \mathbf{u}_V) : V \in \mathbf{C}_i)$
10: **end for**
11: $\mathcal{F}_H \leftarrow \{f_i^H : \mathbf{C}_i \in \mathbb{C}\}$
12: **return** $\mathcal{M}_H = \langle \mathbf{U}_H, \mathbf{V}_H, \mathcal{F}_H, P(\mathbf{U}_H) \rangle$

Many such assumptions take the form of a graphical model, such as a causal diagram (Pearl, 1995), which imply constraints between causal distributions from causal (Bareinboim et al., 2022) and counterfactual Bayesian networks (Correa & Bareinboim, 2024). In the context of abstractions, when τ is a constructive abstraction function that satisfies the AIC, it has been shown that one can avoid assuming the entire causal diagram of the low-level model in favor of a cluster causal diagram (C-DAG) (Anand et al., 2023) w.r.t. the intervariable clusters $\mathbb{C}$. Unfortunately, this graphical model is insufficient for the case when the AIC is violated.

Proposition 2 (C-DAG Insufficiency (Informal))**.** *For a constructive abstraction function τ over intervariable clusters $\mathbb{C}$ in which the AIC does not hold, the C-DAG $\mathcal{G}_{\mathbb{C}}$ implies constraints that may be unsound.* ■

To witness why this is the case, Fig. 2(b) shows the issue clearly. Attempting an abstraction in violation of the AIC is akin to performing a partial SCM projection, which may introduce new dependencies between SCM functions, therefore implying new edges in the graph. Ex. 2 explains this dependence numerically. Since no variables are clustered together in the example, both the original causal diagram $\mathcal{G}$ and the C-DAG $\mathcal{G}_{\mathbb{C}}$ are represented by the top graph in Fig. 3. However, this graph implies that $P(Y_{x_H} \mid z) = P(Y_{x_H})$. Evidently, this is not true since Eq. 9 is not equal to Eq. 10. As hinted by the construction in Alg. 1, the high-level function f_Y^H requires some additional information from Z to decide between interpreting x_C as x_1 or x_2. This information adds a dependence from Z to the function of f_Y^H, which requires adding a directed edge from Z to Y.

While the original C-DAG construction is not valid for projected abstraction inferences, one can use a modified version that adds the new required dependencies into the C-DAG.

Definition 8 (Partially Projected C-DAG)**.** Let $\tau : \mathcal{D}_{V_L} \to \mathcal{D}_{V_H}$ be a constructive abstraction function w.r.t. intervariable clusters $\mathbb{C}$ and intravariable clusters $\mathbb{D}$. Let $\mathcal{G}_{\mathbb{C}} =$

$\langle \mathbf{V}_H, \mathbf{E}_{\mathbb{C}} \rangle$ be a C-DAG (with nodes $\mathbf{V}_H$ and edges $\mathbf{E}_{\mathbb{C}}$), of graph $\mathcal{G}$ w.r.t. $\mathbb{C}$. Let $\mathbf{V}_H^\dagger \subseteq \mathbf{V}_H$ be the set of AIC violation variables (App. A Def. 16). Then, construct $\mathcal{G}_{\mathbb{C}}^\dagger = \langle \mathbf{V}_H, \mathbf{E}_{\mathbb{C}}^\dagger \rangle$ as follows. Start by setting $\mathbf{E}_{\mathbb{C}}^\dagger \leftarrow \mathbf{E}_{\mathbb{C}}$. Then apply the following rules for all $X \in \mathbf{V}_H^\dagger$.

(1) If $Z \to X \to Y$ in $\mathbf{E}_{\mathbb{C}}$, then add $Z \to Y$ into $\mathbf{E}_{\mathbb{C}}^\dagger$.

(2) If $Z \dashleftarrow\!\dashrightarrow X \to Y$ in $\mathbf{E}_{\mathbb{C}}$, then add $Z \dashleftarrow\!\dashrightarrow Y$ and $X \dashleftarrow\!\dashrightarrow Y$ into $\mathbf{E}_{\mathbb{C}}^\dagger$.

(3) If $Z \leftarrow X \to Y$ in $\mathbf{E}_{\mathbb{C}}$, then add $Z \dashleftarrow\!\dashrightarrow Y$ into $\mathbf{E}_{\mathbb{C}}^\dagger$.

Repeat iteratively to accommodate new edges.[3] $\mathcal{G}_{\mathbb{C}}^\dagger$ is called the partially projected C-DAG of $\mathcal{G}$ w.r.t. $\mathbb{C}$ and $\mathbf{V}_H^\dagger$. ■

The steps correspond to the intuition discussed earlier–when performing a partial projection, parts of the variables in $\mathbf{V}_H^\dagger$ are projected into the exogenous space, resulting in additional dependences that require additional edge connections. Examples of C-DAGs and their corresponding projected C-DAGs are shown in Fig. 3. In the figure, rows (a), (b), and (c) correspond to examples of steps 1, 2, 3 respectively. It turns out that this new definition is precisely what is needed for abstraction inference in the absence of the AIC.

Theorem 2 (Projected C-DAG Sufficiency and Necessity (Informal))**.** *Let $\mathcal{M}_L$ be an SCM over variables $\mathbf{V}_L$, $\tau : \mathcal{D}_{\mathbf{V}_L} \to \mathcal{D}_{\mathbf{V}_H}$ be a constructive abstraction function w.r.t. clusters $\mathbb{C}$ and $\mathbb{D}$, and $\mathbf{V}_H^\dagger$ be the AIC violation set. The partially projected C-DAG $\mathcal{G}_{\mathbb{C}}^\dagger$ w.r.t. $\mathbb{C}$ and $\mathbf{V}_H^\dagger$ completely describes all constraints over $\mathbf{V}_H$.* ■

In other words, the projected C-DAG provides exactly the constraints necessary to solve the task of performing causal inferences across abstractions, even when the AIC is violated. In particular, certain interventional and counterfactual distributions may be inferrable from a combination of the projected C-DAG $\mathcal{G}_{\mathbb{C}}^\dagger$ and the available datasets from $\mathcal{M}_L$. Determining precisely which queries can be inferred is known as the identification problem, which is defined below in the context of abstract identification.

Definition 9 (Abstract Identification (General))**.** Let $\tau : \mathcal{D}_{\mathbf{V}_H} \to \mathcal{D}_{\mathbf{V}_L}$ be a constructive abstraction function. Consider projected C-DAG $\mathcal{G}_{\mathbb{C}}^\dagger$, and let $\mathbb{Z} = \{P(\mathbf{V}_{L[\mathbf{z}_k]})\}_{k=1}^{\ell}$ be a collection of available interventional (or observational if $\mathbf{Z}_k = \emptyset$) distributions over $\mathbf{V}_L$. Let Ω_L and Ω_H be the space of SCMs defined over $\mathbf{V}_L$ and $\mathbf{V}_H$, respectively, and let $\Omega_L(\mathcal{G}_{\mathbb{C}}^\dagger)$ and $\Omega_H(\mathcal{G}_{\mathbb{C}}^\dagger)$ be their corresponding subsets that induce $\mathcal{G}_{\mathbb{C}}^\dagger$. A query Q is said to be τ-ID from $\mathcal{G}_{\mathbb{C}}^\dagger$ and $\mathbb{Z}$ iff for every $\mathcal{M}_L \in \Omega_L(\mathcal{G}_{\mathbb{C}}^\dagger), \mathcal{M}_H \in \Omega_H(\mathcal{G}_{\mathbb{C}}^\dagger)$ such that $\mathcal{M}_H$ is $\mathbb{Z}$-τ consistent with $\mathcal{M}_L$, $\mathcal{M}_H$ is also Q-τ consistent with $\mathcal{M}_L$. ■

In words, a query Q is considered τ-ID if, for any pair of models $\mathcal{M}_L$ and $\mathcal{M}_H$ such that both are compatible with

[3] Procedure can be applied algorithmically in one pass by applying all rules for each node in $\mathbf{V}_H^\dagger$ in topological order.

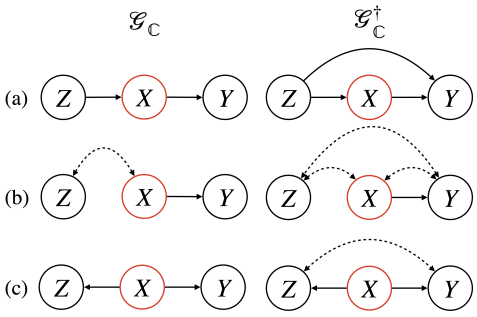

Figure 3: Examples of C-DAGs (left) and their corresponding projected C-DAGs (right), with AIC violation variables $\mathbf{V}_H^\dagger$ outlined in red.

$\mathcal{G}_{\mathbb{C}}^\dagger$ and $\mathbb{Z}$, they also match in Q. In contrast, Q is not τ-ID if there exist $\mathcal{M}_L$ and $\mathcal{M}_H$ that are compatible with both $\mathcal{G}_{\mathbb{C}}^\dagger$ and $\mathbb{Z}$ but disagree on Q (i.e., $Q^{\mathcal{M}_L} \neq \tau(Q)^{\mathcal{M}_H}$). Abstract identification may seem like a difficult property to check, but it turns out that there is a natural connection with the classical identification problem, as shown below.

Theorem 3 (Dual Abstract ID (General))**.** *Consider a counterfactual query Q over $\mathbf{V}_L$, a constructive abstraction function τ w.r.t. clusters $\mathbb{C}$ and $\mathbb{D}$, a projected C-DAG $\mathcal{G}_{\mathbb{C}}^\dagger$, and data $\mathbb{Z}$ from $\mathbf{V}_L$. Q is τ-ID from $\mathcal{G}_{\mathbb{C}}^\dagger$ and $\mathbb{Z}$ if and only if $\tau(Q)$ is ID from $\mathcal{G}_{\mathbb{C}}^\dagger$ and $\tau(\mathbb{Z})$.* ■

In words, τ-identification across abstractions is equivalent to classic identification on the high-level space.

Example 3. Continuing Ex. 1, note that X_H is the only AIC violator in $\mathbf{V}_H$, since x_1 and x_2 both map to x_C but have different effects on Y. Hence, $\mathbf{V}_H^\dagger = \{X_H\}$, and the C-DAG $\mathcal{G}_{\mathbb{C}}$ and projected C-DAG $\mathcal{G}_{\mathbb{C}}^\dagger$ are the two graphs in Fig. 3(a). To answer the query of interest $P(Y_{X_H=x_C} = 1)$, one can apply Thm. 3 to simply identify the quantity w.r.t. $P(\mathbf{V}_H)$ and $\mathcal{G}_{\mathbb{C}}^\dagger$. In this case, note that the causal effect of X_H on Y can be computed via backdoor adjustment on Z, so $P(Y_{X_H=x_C} = 1)$ is equal to

$$\sum_z P(Y = 1 \mid X_H = x_C, Z = z)P(Z = z) \tag{14}$$

$$= \sum_z P(Y = 1 \mid X_L \in \{x_1, x_2\}, z)P(z) \tag{15}$$

$$= (0.7)(0.74) + (0.3)(0.26) = 0.596. \tag{16}$$

■

Thm. 3 implies that, in practice, τ-ID can be checked by performing any classical ID procedure on the high-level space. This may include algorithmic approaches or other optimization-based approaches.

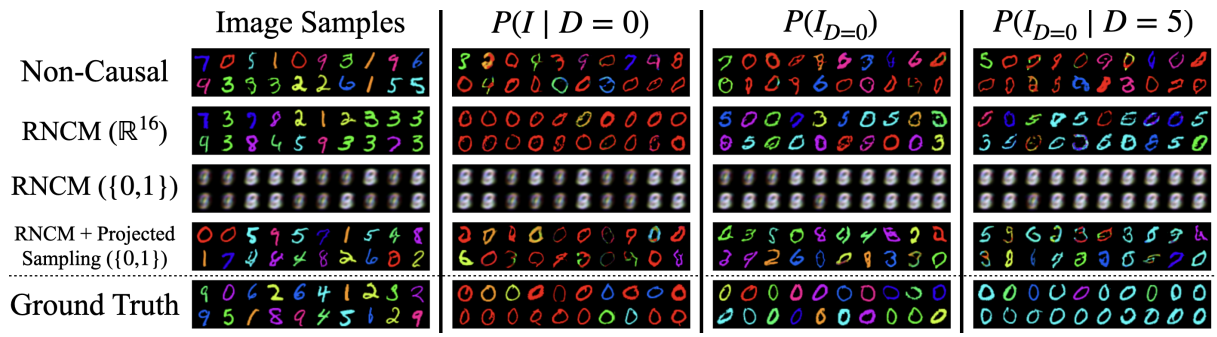

Figure 4: Colored MNIST results. Samples from different causal queries (top) are collected from competing approaches (left). The expressions in parentheses are the representation sizes. The left column shows direct image samples from each of the models, while the second, third, and fourth columns show samples generated from an $\mathcal{L}_1$, $\mathcal{L}_2$, and $\mathcal{L}_3$ query, respectively.

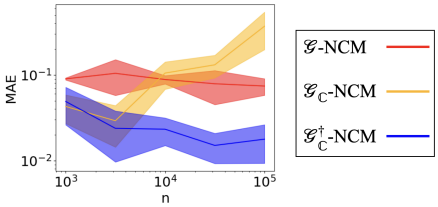

Figure 5: Mean absolute error (MAE) v. number of samples for the MNIST estimation task. Comparisons between an abstractionless approach (red), a C-DAG approach (yellow), and a projected C-DAG approach (blue).

Figure 6: (Left) Graph of Colored MNIST experiment. (Right) Correlation shown between color and digit.

4. Experiments

We perform two experiments to demonstrate the benefits of projected abstractions. The models in the experiments leverage Neural Causal Models (NCMs) (Xia et al., 2021; 2023), specifically the generative adversarial implementation called GAN-NCMs. Details of the experiment setup can be found in App. D, and code will be released upon paper acceptance.

In the first experiment, we test the necessity of the projected C-DAGs when the AIC does not hold. The high-level query $\tau(Q) = P(y_x \mid z)$ is estimated in the graph setting shown in Fig. 3(a), where Z is a digit from 0 to 9, X is a corresponding colored MNIST image, and Y is a label denoting the color prediction of X. $\tau(X)$ maps the image to a binary variable representing the shade (light or dark) of X.

The results are shown in Fig. 5. Three different GAN-NCMs are trained: one directly on the low-level data that does not use abstractions (red), an abstracted one constrained by the C-DAG (yellow), and an abstracted one constrained by the projected C-DAG (blue). 95% confidence intervals of the errors are plotted in the figure. Note that the abstractionless model and the projected C-DAG model have decreasing error with more samples, but the regular C-DAG model is unable to learn the correct query. The abstractionless model has higher error than the projected C-DAG model since it operates in a higher-dimensional space.

In the second experiment, we test an interesting consequence of the projected abstraction theory: the soft intervention definition in Eq. 7 can be directly modeled and sampled if attempting to reconstruct the low-level data. We show this in the colored MNIST experiment (Xia & Bareinboim, 2024). In the model, digit D and color C both cause the image I, but they are confounded (e.g., 0's are red, 5's are cyan, see Fig. 6). Three different queries are tested (the right three columns of Fig. 4). $P(I \mid D = 0)$ is an $\mathcal{L}_1$ query representing images conditioned on digit $= 0$, resulting in red 0's. $P(I_{D=0})$ is an $\mathcal{L}_2$ query representing images with the digit intervened as 0, cutting the confounding and resulting in 0's of all colors. $P(I_{D=0} \mid D=5)$ is an $\mathcal{L}_3$ query representing images with digit intervened as 0, conditioned on the digit originally being 5. This results in 0's with colors of images that were originally 5's, resulting in cyan 0's.

Four methods are compared on these queries in Fig. 4, with the ground truth shown on row 5. The non-causal approach (row 1) simply directly models the conditional distribution between digit and image and therefore fails to model anything higher than $\mathcal{L}_1$. The representational NCM or RNCM (Xia & Bareinboim, 2024) (row 2) is able to decently reproduce all queries, but it uses a 16-dimensional representation space, which cannot shrink much further due to AIC limitations. When forced to take a binary representation (row 3), the RNCM clearly lacks the representation power to properly generate images. In contrast, using a projected sampling approach (row 4) can reproduce the images even with a representation size as small as a binary digit.

5. Conclusions

This paper introduced projected abstractions (Def. 6), which can be constructed algorithmically (Alg. 1, Thm. 1), to overcome the AIC limitation. When the full model was not available, we leveraged a new graphical model (Def. 8, Thm. 2) that allowed for causal inferences through the abstract-ID problem (Def. 9, Thm. 3). Finally, we demonstrated the ability of projected abstractions to leverage representation learning within difficult causal inference settings through high-dimensional image experiments.

Impact Statement

This paper presents work whose goal is to advance the field of causal inference, a subfield of machine learning. The results in this paper may have implications bringing together strong practical results in representation learning and computer vision research with the explainability and generalizability of causal inference results. The trend is that this will lead to smarter AI, which itself has many consequences out of the scope of this work, but the benefit of understanding causal inference is that it can lead to less bias and more accountability of AI models.

Acknowledgements

This research was supported in part by the NSF, ONR, AFOSR, DoE, Amazon, JP Morgan, and the Alfred P. Sloan Foundation.

References

Anand, T. V., Ribeiro, A. H., Tian, J., and Bareinboim, E. Causal effect identification in cluster dags. In *Proceedings of the 37th AAAI Conference on Artificial Intelligence*. AAAI Press, 2023.

Balke, A. and Pearl, J. Bounds on treatment effects from studies with imperfect compliance. *Journal of the American Statistical Association*, 92(439):1172–1176, 9 1997.

Bareinboim, E., Correa, J. D., Ibeling, D., and Icard, T. On pearl's hierarchy and the foundations of causal inference. In *Probabilistic and Causal Inference: The Works of Judea Pearl*, pp. 507–556. Association for Computing Machinery, New York, NY, USA, 1st edition, 2022.

Beckers, S. and Halpern, J. Y. Abstracting causal models. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'19/IAAI'19/EAAI'19. AAAI Press, 2019. ISBN 978-1-57735-809-1. doi: 10.1609/aaai.v33i01.33012678. URL https://doi.org/10.1609/aaai.v33i01.33012678.

Beckers, S., Eberhardt, F., and Halpern, J. Y. Approximate causal abstraction. In *Proceedings of the 35th Conference on Uncertainty in Artificial Intelligence*. 2019.

Bengio, Y., Courville, A., and Vincent, P. Representation learning: A review and new perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(8):1798–1828, aug 2013. ISSN 0162-8828. doi: 10.1109/TPAMI.2013.50. URL https://doi.org/10.1109/TPAMI.2013.50.

Correa, J. and Bareinboim, E. Counterfactual graphical models: Constraints and inference. Technical Report R-115, Causal Artificial Intelligence Lab, Columbia University, August 2024.

Felekis, Y., Zennaro, F. M., Branchini, N., and Damoulas, T. Causal optimal transport of abstractions. In *Conference on Causal Learning and Reasoning, CLeaR 2024*, 2024.

Geiger, A., Potts, C., and Icard, T. Causal abstraction for faithful model interpretation, 2023.

Lee, S. and Bareinboim, E. Structural Causal Bandits with Non-manipulable Variables. In *Proceedings of the 33rd AAAI Conference on Artificial Intelligence*, 2019.

Massidda, R., Geiger, A., Icard, T., and Bacciu, D. Causal abstraction with soft interventions. In van der Schaar, M., Zhang, C., and Janzing, D. (eds.), *Proceedings of the Second Conference on Causal Learning and Reasoning*, volume 213 of *Proceedings of Machine Learning Research*, pp. 68–87. PMLR, 11–14 Apr 2023. URL https://proceedings.mlr.press/v213/massidda23a.html.

Pearl, J. Causal diagrams for empirical research. *Biometrika*, 82(4):669–688, 1995.

Pearl, J. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, New York, NY, USA, 2nd edition, 2000.

Pearl, J. and Mackenzie, D. *The Book of Why*. Basic Books, New York, 2018.

Rubenstein, P. K., Weichwald, S., Bongers, S., Mooij, J., Janzing, D., Grosse-Wentrup, M., and Schölkopf, B. Causal Consistency of Structural Equation Models. In *Proceedings of the Thirty-Third Conference on Uncertainty in Artificial Intelligence*, 2017.

Steinberg, D. Copyright. In *The Cholesterol Wars*. Academic Press, Oxford, 2007. ISBN 978-0-12-373979-7. doi: https://doi.org/10.1016/B978-0-12-373979-7.50003-0. URL https:

//www.sciencedirect.com/science/article/pii/B9780123739797500030.

Truswell, A. *Cholesterol and Beyond: The Research on Diet and Coronary Heart Disease 1900-2000*. 01 2010. ISBN 978-90-481-8874-1. doi: 10.1007/978-90-481-8875-8.

Xia, K. and Bareinboim, E. Neural causal abstractions. In *Proceedings of the 38th AAAI Conference on Artificial Intelligence*. AAAI Press, 2024.

Xia, K., Lee, K.-Z., Bengio, Y., and Bareinboim, E. The causal-neural connection: Expressiveness, learnability, and inference. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 10823–10836. Curran Associates, Inc., 2021.

Xia, K., Pan, Y., and Bareinboim, E. Neural causal models for counterfactual identification and estimation. In *Proceedings of the 11th International Conference on Learning Representations (ICLR-23)*, 2023.

Zennaro, F. M., Drávucz, M., Apachitei, G., Widanage, W. D., and Damoulas, T. Jointly learning consistent causal abstractions over multiple interventional distributions. In van der Schaar, M., Zhang, C., and Janzing, D. (eds.), *Conference on Causal Learning and Reasoning, CLeaR 2023, 11-14 April 2023, Amazon Development Center, Tübingen, Germany, April 11-14, 2023*, volume 213 of *Proceedings of Machine Learning Research*, pp. 88–121. PMLR, 2023. URL https://proceedings.mlr.press/v213/zennaro23a.html.

Zhang, J., Jin, T., and Bareinboim, E. Partial counterfactual identification from observational and experimental data. In *Proceedings of the 39th International Conference on Machine Learning (ICML-22)*, 2022.

A. Proofs

This section contains all proofs of the results described in the main body.

A.1. Additional Definitions

The following definitions about τ-abstractions from Beckers & Halpern (2019) set the groundwork for many discussions on abstraction theory.

Definition 10 (τ-Abstraction (Beckers & Halpern, 2019, Def. 3.13))**.** Let $\mathcal{M}_L = \langle \mathbf{U}_L, \mathbf{V}_L, \mathcal{F}_L, P(\mathbf{U}_L)\rangle$ and $\mathcal{M}_H = \langle \mathbf{U}_H, \mathbf{V}_H, \mathcal{F}_H, P(\mathbf{U}_H)\rangle$ be two SCMs. Let $\mathcal{I}_L$ and $\mathcal{I}_H$ be the sets of allowed interventions respectively. Given $\tau : \mathcal{D}_{\mathbf{V}_L} \to \mathcal{D}_{\mathbf{V}_H}$, we say that $(\mathcal{M}_H, \mathcal{I}_H)$ is a τ-abstraction of $(\mathcal{M}_L, \mathcal{I}_L)$ if:

1. τ is surjective;
2. There exists surjective $\tau_{\mathbf{U}} : \mathcal{D}_{\mathbf{U}_L} \to \mathcal{D}_{\mathbf{U}_H}$ that is compatible with τ, i.e.
$$\tau(\mathcal{M}_{L[\mathbf{X}_L \leftarrow \mathbf{x}_L]}(\mathbf{u}_L)) = \mathcal{M}_{H[\omega_\tau(\mathbf{X}_L \leftarrow \mathbf{x}_L)]}(\tau_{\mathbf{U}}(\mathbf{u}_L)), \tag{17}$$
for all $\mathbf{u}_L \in \mathcal{D}_{\mathbf{U}_L}$ and all $(\mathbf{X}_L \leftarrow \mathbf{x}_L) \in \mathcal{I}_L$;
3. $\mathcal{I}_H = \omega_\tau(\mathcal{I}_L)$.

■

Further, we will assume that if $(\mathcal{M}_H, \mathcal{I}_H)$ is a τ-abstraction of $(\mathcal{M}_L, \mathcal{I}_L)$, then $P(\mathbf{U}_H) = \tau_{\mathbf{U}}(P(\mathbf{U}_L)) = P(\tau_{\mathbf{U}}(\mathbf{U}_L))$, that is, the distribution of $P(\mathbf{U}_H)$ can be obtained from $P(\mathbf{U}_L)$ via the push-forward measure through $\tau_{\mathbf{U}}$. While it is not explicitly stated in the definition, this property aligns with the intention of linking the spaces of $\mathbf{U}_L$ and $\mathbf{U}_H$ through $\tau_{\mathbf{U}}$.

Definition 11 (Strong τ-Abstraction (Beckers & Halpern, 2019, Def. 3.15))**.** We say that $\mathcal{M}_H$ is a strong τ-abstraction of $\mathcal{M}_L$ if $(\mathcal{M}_H, \mathcal{I}_H)$ is a τ-abstraction of $(\mathcal{M}_L, \mathcal{I}_L)$ and $\mathcal{I}_H = \mathcal{I}_H^*$. ■

Definition 12 (Constructive τ-Abstraction (Beckers & Halpern, 2019, Def. 3.19))**.** $\mathcal{M}_H$ is a constructive τ-abstraction of $\mathcal{M}_L$ if $\mathcal{M}_H$ is a strong τ-abstraction of $\mathcal{M}_L$, and there exists a partition of $\mathbf{V}_L$, $\mathbb{C} = \{\mathbf{C}_1, \mathbf{C}_2, \ldots, \mathbf{C}_{n+1}\}$ (where $n = |\mathbf{V}_H|$) with nonempty $\mathbf{C}_1$ to $\mathbf{C}_n$, such that τ can be decomposed as $\tau = (\tau_{\mathbf{C}_1}, \tau_{\mathbf{C}_2}, \ldots, \tau_{\mathbf{C}_n})$, where each $\tau_{\mathbf{C}_i} : \mathcal{D}_{\mathbf{C}_i} \to \mathcal{D}_{V_{H,i}}$ maps the ith partition to the ith variable of $\mathbf{V}_H$. ■

For constructive abstraction functions, there is a notion of Q-τ consistency that connects low and high-level quantities. The formal definition is below.

Definition 13 (Q-τ Consistency (Xia & Bareinboim, 2024, Def. 7))**.** Let $\mathcal{M}_L$ and $\mathcal{M}_H$ be SCMs defined over variables $\mathbf{V}_L$ and $\mathbf{V}_H$, respectively. Let $\tau : \mathcal{D}_{\mathbf{V}_L} \to \mathcal{D}_{\mathbf{V}_H}$ be a constructive abstraction function w.r.t. clusters $\mathbb{C}$ and $\mathbb{D}$. Let

$$Q = \sum_{\mathbf{y}_{L,*} \in \mathcal{D}_{\mathbf{Y}_{L,*}}(\mathbf{y}_{H,*})} P(\mathbf{Y}_{L,*} = \mathbf{y}_{L,*}) \tag{18}$$

be a low-level Layer 3 quantity of interest (for some $\mathbf{y}_{H,*} \in \mathcal{D}_{\mathbf{Y}_{H,*}}$), as expressed in Eq. 11, and let

$$\tau(Q) = P(\mathbf{Y}_{H,*} = \mathbf{y}_{H,*}) \tag{19}$$

be its high level counterpart. We say that $\mathcal{M}_H$ is Q-τ consistent with $\mathcal{M}_L$ if

$$\begin{aligned} &\sum_{\mathbf{y}_{L,*} \in \mathcal{D}_{\mathbf{Y}_{L,*}}(\mathbf{y}_{H,*})} P^{\mathcal{M}_L}(\mathbf{Y}_{L,*} = \mathbf{y}_{L,*}) \\ &= P^{\mathcal{M}_H}(\mathbf{Y}_{H,*} = \mathbf{y}_{H,*}), \end{aligned} \tag{20}$$

that is, the value of Q induced by $\mathcal{M}_L$ is equal to the value of $\tau(Q)$ induced by $\mathcal{M}_H$[4]. Furthermore, if $\mathcal{M}_H$ is Q-τ consistent with $\mathcal{M}_L$ for all $Q \in \mathcal{L}_i(\mathcal{M}_L)$ of the form of Eq. 18, then $\mathcal{M}_H$ is said to be $\mathcal{L}_i$-τ consistent with $\mathcal{M}_L$. ■

[4]Note that the equality in Eq. 20 is consistent with the push-forward measure through τ.

The following result relates constructive abstraction functions and the concept of Q-τ consistency with τ-abstractions.

Proposition 3 (Abstraction Connection (Xia & Bareinboim, 2024, Prop. 1)). *Let $\tau : \mathcal{D}_{\mathbf{V}_L} \to \mathcal{D}_{\mathbf{V}_H}$ be a constructive abstraction function (Def. 4). $\mathcal{M}_H$ is $\mathcal{L}_3$-τ consistent (Def. 13) with $\mathcal{M}_L$ if and only if there exists SCMs $\mathcal{M}'_L$ and $\mathcal{M}'_H$ s.t. $\mathcal{L}_3(\mathcal{M}'_L) = \mathcal{L}_3(\mathcal{M}_L)$, $\mathcal{L}_3(\mathcal{M}'_H) = \mathcal{L}_3(\mathcal{M}_H)$, and $\mathcal{M}'_H$ is a constructive τ-abstraction of $\mathcal{M}'_L$.*

■

For abstraction inference, C-DAGs can often be leveraged in place of causal diagrams, defined below.

Definition 14 (Cluster Causal Diagram (C-DAG) (Anand et al., 2023, Def. 1)). Given a causal diagram $\mathcal{G} = \langle \mathbf{V}, \mathbf{E} \rangle$ and an admissible clustering $\mathbb{C} = \{\mathbf{C}_1, \ldots, \mathbf{C}_k\}$ of $\mathbf{V}$, construct a graph $\mathcal{G}_{\mathbb{C}} = \langle \mathbb{C}, \mathbf{E}_{\mathbb{C}} \rangle$ over $\mathbb{C}$ with a set of edges $\mathbf{E}_{\mathbb{C}}$ defined as follows:

1. A directed edge $\mathbf{C}_i \to \mathbf{C}_j$ is in $\mathbf{E}_{\mathbb{C}}$ if there exists some $V_i \in \mathbf{C}_i$ and $V_j \in \mathbf{C}_j$ such that $V_i \to V_j$ is an edge in $\mathbf{E}$.
2. A dashed bidirected edge $\mathbf{C}_i \dashleftarrow\dashrightarrow \mathbf{C}_j$ is in $\mathbf{E}_{\mathbb{C}}$ if there exists some $V_i \in \mathbf{C}_i$ and $V_j \in \mathbf{C}_j$ such that $V_i \dashleftarrow\dashrightarrow V_j$ is an edge in $\mathbf{E}$. ■

This paper shows that they are insufficient for inferences when the AIC does not hold, but they are used as the base graph for constructing projected C-DAGs.

Most of the experiments in this paper leverage the $\mathcal{G}$-constrained neural causal model for practical implementations, defined below.

Definition 15 ($\mathcal{G}$-Constrained Neural Causal Model ($\mathcal{G}$-NCM) (Xia et al., 2021, Def. 7)). Given a causal diagram $\mathcal{G}$, a $\mathcal{G}$-constrained Neural Causal Model (for short, $\mathcal{G}$-NCM) $\widehat{M}(\boldsymbol{\theta})$ over variables $\mathbf{V}$ with parameters $\boldsymbol{\theta} = \{\theta_{V_i} : V_i \in \mathbf{V}\}$ is an SCM $\langle \widehat{\mathbf{U}}, \mathbf{V}, \widehat{\mathcal{F}}, P(\widehat{\mathbf{U}}) \rangle$ such that

- $\widehat{\mathbf{U}} = \{\widehat{U}_{\mathbf{C}} : \mathbf{C} \in \mathbb{C}(\mathcal{G})\}$, where $\mathbb{C}(\mathcal{G})$ is the set of all maximal cliques over bidirected edges of $\mathcal{G}$;
- $\widehat{\mathcal{F}} = \{\hat{f}_{V_i} : V_i \in \mathbf{V}\}$, where each $\hat{f}_{V_i}$ is a feedforward neural network parameterized by $\theta_{V_i} \in \boldsymbol{\theta}$ mapping values of $\mathbf{U}_{V_i} \cup \mathbf{Pa}_{V_i}$ to values of V_i for $\mathbf{U}_{V_i} = \{\widehat{U}_{\mathbf{C}} : \widehat{U}_{\mathbf{C}} \in \widehat{\mathbf{U}} \text{ s.t. } V_i \in \mathbf{C}\}$ and $\mathbf{Pa}_{V_i} = Pa_{\mathcal{G}}(V_i)$;
- $P(\widehat{\mathbf{U}})$ is defined s.t. $\widehat{U} \sim \text{Unif}(0, 1)$ for each $\widehat{U} \in \widehat{\mathbf{U}}$. ■

A.2. Proofs of Sec. 2

In this section, we prove the theoretical results stated in Sec. 2.

The first observation is that although the AIC is a property of the entire abstraction, one can clearly distinguish individual high-level variables that violate the AIC, as shown in the following definition.

Definition 16 (AIC Violation Set). Let $\mathcal{M}_L$ be an SCM defined over $\mathbf{V}_L$ and $\tau : \mathcal{D}_{\mathbf{V}_L} \to \mathcal{D}_{\mathbf{V}_H}$ be a constructive abstraction function w.r.t. clusters $\mathbb{C}$ and $\mathbb{D}$. Let $\mathbf{V}_H^\dagger \subseteq \mathbf{V}_H$ be the set of high-level variables such that $V_{H,i} \in \mathbf{V}_H^\dagger$ iff there exists $V_{H,j} \in \mathbf{V}_H$ with $V_{H,i} \in \mathbf{Pa}_{V_{H,j}}$ such that Eq. 2 is violated for $\mathbf{C}_j$, some $\mathbf{u} \in \mathcal{D}_{\mathbf{U}}$, and some $\mathbf{v}_1, \mathbf{v}_2 \in \mathcal{D}_{\mathbf{V}_L}$ where $\mathbf{v}_1$ and $\mathbf{v}_2$ only differ in the values associated with $\mathbf{C}_i$ ($\mathbf{C}_i$ and $\mathbf{C}_j$ are the corresponding clusters of $\mathbb{C}$ respectively). $\mathbf{V}_H^\dagger$ is called the AIC violation set of τ w.r.t. $\mathcal{M}_L$. ■

In words, a high-level variable is an AIC violator if two of its values that have different effects on its children are clustered together (e.g., X is an AIC violator in Ex. 1). Now recall the definition of an SCM projection.

Proposition 4 (SCM Projection). *Given an SCM $\mathcal{M} = \langle \mathbf{U}, \mathbf{V}, \mathcal{F}, P(\mathbf{U}) \rangle$, there exists an SCM $\mathcal{M}' = \langle \mathbf{U}, \mathbf{W}, \mathcal{F}', P(\mathbf{U}) \rangle$ such that, for all $\mathbf{u} \in \mathcal{D}_{\mathbf{U}}$, $\mathbf{X} \subseteq \mathbf{W}$, and $\mathbf{x} \in \mathcal{D}_{\mathbf{X}}$,*

$$\mathcal{M}_{\mathbf{x}}(\mathbf{u})[\mathbf{W}] = \mathcal{M}'_{\mathbf{x}}(\mathbf{u}) \tag{21}$$

$\mathcal{M}'$ is called an SCM projection of $\mathcal{M}$ over $\mathbf{W}$. ■

Proof. We show how to construct $\mathcal{M}'$. For each $Y \in \mathbf{V}$ in topological order according to the inputs of the functions of $\mathcal{F}$, choose $f'_Y \leftarrow f_Y(\mathbf{U}_Y, \mathbf{Pa}_Y)$, where for each $X \in \mathbf{Pa}_Y$,

1. If $X \in \mathbf{W}$, then keep X as an input of f'_Y;
2. Otherwise if $X \notin \mathbf{W}$, then replace X with $f'_X(\mathbf{U}_X, \mathbf{Pa}_X)$. Denote $\mathbf{U}'_Y$ and $\mathbf{Pa}'_Y$ as the new exogenous variables and parents of Y after recursively applying this rule until all endogenous inputs are in $\mathbf{W}$.

Then, construct $\mathcal{F}' = \{f'_Y; Y \in \mathbf{W}\}$ and $\mathcal{M}' = \langle \mathbf{U}, \mathbf{W}, \mathcal{F}', P(\mathbf{U})\rangle$. Note that for all $\mathbf{u} \in \mathcal{D}_{\mathbf{U}}$, $\mathbf{X} \subseteq \mathbf{W}$, and $\mathbf{x} \in \mathcal{D}_{\mathbf{X}}$,

$$\begin{aligned}
&\mathcal{M}_{\mathbf{x}}(\mathbf{u})[\mathbf{W}] && (22)\\
&= \mathbf{W}_{\mathbf{x}}(\mathbf{u}) && (23)\\
&= \{f_Y(\mathbf{u}_Y, \mathbf{Pa}_{Y[\mathbf{x}]}(\mathbf{u})) : Y \in \mathbf{W}\} && (24)\\
&= \{f'_Y(\mathbf{u}'_Y, \mathbf{Pa}'_{Y[\mathbf{x}]}(\mathbf{u})) : Y \in \mathbf{W}\} && (25)\\
&= \mathcal{M}'_{\mathbf{x}}(\mathbf{u}). && (26)
\end{aligned}$$

□

Note that a version of this proposition was proven in Lee & Bareinboim (2019), specifically for all $\mathcal{L}_2$ queries of $\mathcal{M}'$. Our proof uses a similar argument, but we show that the implied result is stronger: $\mathcal{M}'$ matches $\mathcal{M}$ on the SCM level for all exogenous settings $\mathbf{u}$ and interventions $\mathbf{x}$. This implies not only matching in $\mathcal{L}_2$ query but also $\mathcal{L}_3$ queries.

Proposition 1 (Partial SCM Projection)**.** *Let $\mathbf{V}$ be a set of variables and $\mathbf{W} \subseteq \mathbf{V}$ be a subset. For each $W_i \in \mathbf{W}$, let $\delta_i : \mathcal{D}_{W_i^o} \times \mathcal{D}_{W_i^v} \to \mathcal{D}_{W_i}$ be a surjective function mapping new variables W_i^o and W_i^v to W_i. W_i^o and W_i^v are called the observed and unobserved projections of W_i respectively. Denote $\delta(\mathbf{W}^o, \mathbf{W}^u) = \mathbf{W}$, where $\mathbf{W}^o = \{W_i^o : W_i \in \mathbf{W}\}$ and $\mathbf{W}^u = \{W_i^u : W_i \in \mathbf{W}\}$. For any SCM $\mathcal{M} = \langle \mathbf{U}, \mathbf{V}, \mathcal{F}, P(\mathbf{U})\rangle$, there exists an SCM $\mathcal{M}' = \langle \mathbf{U}' = \mathbf{U} \cup \mathbf{W}^u, \mathbf{V}' = \mathbf{W}^o, \mathcal{F}', P(\mathbf{U}')\rangle$ such that, for all $\mathbf{u} \in \mathcal{D}_{\mathbf{U}}$, $\mathbf{X} \subseteq \mathbf{W}$, and $\mathbf{x} \in \mathcal{D}_{\mathbf{X}}$,*

$$\mathbf{w}^o_{\mathbf{x}} = \mathcal{M}'_{[\mathbf{x}^o]}(\mathbf{u}, \mathbf{x}^u, \mathbf{z}^u), \tag{6}$$

where $\delta(\mathbf{w}^o_{\mathbf{x}}, \mathbf{w}^u_{\mathbf{x}}) = \mathbf{W}_{\mathbf{x}}(\mathbf{u})$, $\delta(\mathbf{x}^o, \mathbf{x}^u) = \mathbf{x}$, $\mathbf{Z}^u = \mathbf{W}^u \setminus \mathbf{X}^u$, and $\mathbf{z}^u$ are the corresponding values from $\mathbf{w}^u_{\mathbf{x}}$. $\mathcal{M}'$ is called a partial SCM projection of $\mathcal{M}$ over $\mathbf{W}^o$. ■

Proof. $\mathcal{M}'$ can be created through Alg. 1, and Thm. 1 proves that Alg. 1 is sound. See the proof of Thm. 1 for details on this construction. □

Theorem 1. *The SCM $\mathcal{M}_H$ constructed by Alg. 1 is a projected abstraction of $\mathcal{M}_L$ that is Q-τ consistent with $\mathcal{M}_L$ for all $\tau(Q) \in \mathcal{L}_3(\mathcal{M}_H)$.* ■

Proof. Let $\mathcal{M}_H$ be the output from Alg. 1 given $\mathcal{M}_L$ and τ constructed from clusters $\mathbb{C}$ and $\mathbb{D}$. Let $\tau(Q) = P(Y_{H,*} = \mathbf{y}_{H,*})$ be any arbitrary high-level query from $\mathcal{L}_3(\mathcal{M}_H)$, and let $Q = \sum_{\mathbf{y}_{L,*} \in \mathcal{D}_{Y_{L,*}}(\mathbf{y}_{H,*})} P(\mathbf{Y}_{L,*} = \mathbf{y}_{L,*})$ be its low-level counterpart. Without loss of generality, we can assume that the set of AIC violators $\mathbf{V}^\dagger_H = \mathbf{V}_H$, since any variable $V \notin \mathbf{V}^\dagger_H$ can be mapped by a trivial δ that ignores V^u.

We first show that $\mathcal{M}_H$ is a projected abstraction of $\mathcal{M}_L$. First, consider the SCM $\mathcal{M}'_H$ defined over the variables $\mathbf{V}'_H = \tau'(\mathbf{V}_L)$, where τ' is the constructive abstraction function constructed from the same intervariable clusters $\mathbb{C}$ and the trivial intravariable clusters $\mathbb{D}' = \mathcal{D}_{\mathbf{V}_L}$ (i.e., each value of $\mathbf{V}_L$ is its own cluster, and $\mathbb{D}$ is ignored). Note that each $V_j \in \mathbf{V}'_H$ corresponds to a cluster $\mathbf{C}_j \in \mathbb{C}$. Suppose $\mathcal{M}'_H$ is constructed such that it is $\mathcal{L}_3$-τ consistent with $\mathcal{M}_L$, which is possible through Alg. 1 of Xia & Bareinboim (2024). Then, Prop. 3 states that $\mathcal{M}'_H$ must be $\mathcal{L}_3$-consistent with a constructive τ-abstraction of $\mathcal{M}_L$. Without loss of generality, suppose $\mathcal{M}'_H$ is this constructive τ-abstraction.

For any $\mathbf{C}_j \in \mathbb{C}$, one can construct variables X^o_H and X^u_H such that $X^o_H = \tau(\mathbf{C}_j)$ and there exists a function δ such that $\delta(X^o_H, X^u_H) = \mathbf{C}_j$, as done so in line 4 of the algorithm. This can be done by simply giving X^u_H an arbitrarily large domain and using X^u_H to disambiguate any information lost in the transformation from $\mathbf{C}_j$ to X^o_H when constructing δ. Note that in the construction of $\mathcal{M}_H$, the variables of X^u_H are placed into $\mathbf{U}_H$.

To show that $\mathcal{M}_H$ is a projected abstraction, we must show that it is a partial SCM projection of $\mathcal{M}'_H$. Looking at Prop. 1, simply choose $\mathbf{W} = \mathbf{V}'_H$. In Alg. 1, each $\mathbf{C}_j \in \mathbb{C}$ is split into $\mathbf{C}^o_j, \mathbf{C}^u_j$ such that $\delta(\mathbf{C}^o_j, \mathbf{C}^u_j) = \mathbf{C}_j$. Denote $\mathbf{V}^o_H$ and $\mathbf{V}^u_H$ as the corresponding sets of variables in $\mathbf{V}'_H$. By construction, indeed $\mathbf{U}_H = \mathbf{U}_L \cup \mathbf{V}^u_H$, and $\mathbf{V}_H = \mathbf{V}^o_H$. Fix $\mathbf{u}_L \in \mathbf{U}_L$,

$\mathbf{X} \subseteq \mathbf{V}'_H$, and $\mathbf{x} \in \mathcal{D}_{V'_H}$. Let $\delta(\mathbf{v}^o_H, \mathbf{v}^u_H) = \mathbf{V}'_{H[\mathbf{x}]}(\mathbf{u}_L)$. Let $\mathbf{Z}^u = \mathbf{V}^u_H \setminus \mathbf{X}^u$ and $\mathbf{z}^u$ be the corresponding values from $\mathbf{w}^u_{\mathbf{x}}$. Then, observe that

$$
\begin{aligned}
&\mathbf{v}^o_H && (27)\\
&= \tau(\mathbf{V}'_{H[\mathbf{x}]}(\mathbf{u}_L)) && (28)\\
&= \tau(\{f'_Y(\mathbf{u}_Y, \mathbf{Pa}'_{Y[\mathbf{x}]}(\mathbf{u})) : Y \in \mathbf{V}'_H\}) && (29)\\
&= \tau(\{f'_Y(\mathbf{u}_Y, \delta(\mathbf{Pa}^o_{Y[\mathbf{x}]}(\mathbf{u}), \mathbf{Pa}^u_{Y[\mathbf{x}]}(\mathbf{u}))) : Y \in \mathbf{V}'_H\}) && (30)\\
&= \tau(\{f_{H,Y}(\mathbf{u}_Y, \mathbf{Pa}^o_{Y[\mathbf{x}]}(\mathbf{u}), \mathbf{Pa}^u_{Y[\mathbf{x}]}(\mathbf{u})) : Y \in \mathbf{V}_H\}) && (31)\\
&= \mathcal{M}_{H[\mathbf{x}_o]}(\mathbf{u}_L, \mathbf{x}^u, \mathbf{z}^u), && (32)
\end{aligned}
$$

matching Eq. 6.

Now we show that $\mathcal{M}_H$ is $\mathcal{L}_3$-τ consistent with $\mathcal{M}_L$. Denote $\mathbf{x}_{L,*}$ and $\mathbf{x}_{H,*}$ as the corresponding sets of interventions of Q and $\tau(Q)$ respectively, and denote $\mathbf{y}_{L,[\mathbf{x}_{L,*}]}$ as the the values of $\mathbf{y}_{L,*}$ specifically under the hard intervention $\mathbf{x}_{L,*}$ (as opposed to the soft interventions under $\sigma_{\mathbf{X}_{L,i}}$). Denote $\mathcal{D}_{\mathbf{U}_L}(\mathbf{y}_{L,[\mathbf{x}_{L,*}]}) \subseteq \mathcal{D}_{\mathbf{U}_L}$ as the values of $\mathbf{U}$ such that $\mathbf{Y}_{L,[\mathbf{x}_{L,*}]}(\mathbf{u}_L) = \mathbf{y}_{L,*}$ (similar notation applies to $\mathcal{D}_{\mathbf{U}_H}$).

Now observe that

$$
\begin{aligned}
&Q^{\mathcal{M}_L} && (33)\\
&= \sum_{\mathbf{y}_{L,*} \in \mathcal{D}_{Y_{L,*}}(\mathbf{y}_{H,*})} P^{\mathcal{M}_L}(\mathbf{Y}_{L,*} = \mathbf{y}_{L,*}) && (34)\\
&= \sum_{\mathbf{y}_{L,*} \in \mathcal{D}_{Y_{L,*}}(\mathbf{y}_{H,*})} \sum_{\mathbf{x}_{L,*} \in \mathcal{D}_{X_{L,*}}(\mathbf{x}_{H,*})} P^{\mathcal{M}_L}(\mathbf{Y}_{L,*} = \mathbf{y}_{L,*} \mid \sigma_{\mathbf{X}_{L,*}} = \mathbf{x}_{L,*}) P(\sigma_{\mathbf{X}_{L,*}} = \mathbf{x}_{L,*}) && (35)\\
&= \sum_{\mathbf{y}_{L,*}, \mathbf{x}_{L,*}} P(\mathbf{U}_L \in \mathcal{D}_{\mathbf{U}_L}(\mathbf{y}_{L,[\mathbf{x}_{L,*}]})) P(\sigma_{\mathbf{X}_{L,*}} = \mathbf{x}_{L,*}) && (36)\\
&= \sum_{\mathbf{y}_{L,*}, \mathbf{x}_{L,*}} P(\mathbf{U}_L \in \mathcal{D}_{\mathbf{U}_L}(\mathbf{y}_{L,[\mathbf{x}_{L,*}]})) \prod_{\mathbf{c}_j \in \mathbf{x}_{L,*}} P(\sigma_{\mathbf{C}_j} = \mathbf{c}_j) && (37)\\
&= \sum_{\mathbf{y}_{L,*}, \mathbf{x}_{L,*}} P(\mathbf{U}_L \in \mathcal{D}_{\mathbf{U}_L}(\mathbf{y}_{L,[\mathbf{x}_{L,*}]})) \prod_{\mathbf{c}_j \in \mathbf{x}_{L,*}} P(\mathbf{c}_j \mid \tau(\mathbf{c}_j) = v_{H,j}, \mathbf{Pa}_{V_{H,j}}(\mathbf{U}_{\mathbf{C}_j}), \mathbf{R}_{\mathbf{V}^c_H(V_{H,j})}(\mathbf{U}_{\mathbf{C}_j})) && (38)\\
&= \sum_{\mathbf{y}_{L,*}, \mathbf{x}_{L,*}} P(\mathbf{U}_L \in \mathcal{D}_{\mathbf{U}_L}(\mathbf{y}_{L,[\mathbf{x}_{L,*}]})) \prod_{\mathbf{c}_j \in \mathbf{x}_{L,*}} P(\delta(\mathbf{c}^o_j, \mathbf{C}^u_j) = \mathbf{c}_j \mid \mathbf{U}_L) && (39)\\
&= \sum_{\mathbf{y}_{L,*}, \mathbf{x}_{L,*}} P(\mathbf{U}_L \in \mathcal{D}_{\mathbf{U}_L}(\mathbf{y}_{L,[\mathbf{x}_{L,*}]})) P(\delta(\mathbf{x}_{H,*}, \mathbf{X}^u_{H,*}) = \mathbf{x}_{L,*} \mid \mathbf{U}_L) && (40)\\
&= \sum_{\mathbf{y}_{L,*}, \mathbf{x}_{L,*}} P(\mathbf{U}_L \in \mathcal{D}_{\mathbf{U}_L}(\mathbf{y}_{L,[\mathbf{x}_{L,*}]}), \delta(\mathbf{x}_{H,*}, \mathbf{X}^u_{H,*}) = \mathbf{x}_{L,*}) && (41)\\
&= P\left(\bigvee_{\mathbf{y}_{L,*} \in \mathcal{D}_{Y_{L,*}}(\mathbf{y}_{H,*}), \mathbf{x}_{L,*} \in \mathcal{D}_{X_{L,*}}(\mathbf{x}_{H,*})} \mathbf{U}_L \in \mathcal{D}_{\mathbf{U}_L}(\mathbf{y}_{L,[\mathbf{x}_{L,*}]}), \delta(\mathbf{x}_{H,*}, \mathbf{X}^u_{H,*}) = \mathbf{x}_{L,*} \right) && (42)\\
&= P((\mathbf{U}_L, \mathbf{V}^u_H) \in \mathcal{D}_{\mathbf{U}_H}(\mathbf{y}_{H,*})) && (43)\\
&= P(\mathbf{U}_H \in \mathcal{D}_{\mathbf{U}_H}(\mathbf{y}_{H,*})) && (44)\\
&= \tau(Q)^{\mathcal{M}_H}. && (45)
\end{aligned}
$$

Explaining each line, line 34 starts by applying the definition of Q. Since the interventions $\mathbf{x}_{L,*}$ are determined through soft interventions σ, we can expand the soft intervention through all possible values of $\mathbf{x}_{L,*}$ (via marginalization), as done so in line 35. The first term is simply computed as the probability of all values of $\mathbf{U}_L$ where $\mathbf{Y}_{L,[\mathbf{x}_{L,*}]}(\mathbf{u}_L) = \mathbf{y}_{L,*}$ (Def. 1), resulting in line 36. The second term can be broken down into the soft interventions of each intravariable cluster (line 37), whose probability is computed through Eq. 89, resulting in line 38. Line 39 is true by construction of $\mathcal{M}_H$, through line 8 of the algorithm. Finally, we consolidate all terms back into $\mathbf{x}_{L,*}$ and merge back into the joint distribution in lines 40 and 41.

Line 42 holds because the probabilities of each individual value of $\mathbf{y}_{L,[\mathbf{x}_{L,*}]}$ are disjoint since $\mathbf{X}_{L,*}$ and $\mathbf{Y}_{L,*}$ can both only be equal to one value at a time. Line 43 holds since $\mathbf{U}_H = \mathbf{U}_L \cup \mathbf{V}_H^u$ and by construction of line 9 in the algorithm, we have exhausted every possible value that $\mathcal{M}_H((\mathbf{u}_L, \mathbf{v}_H^u)) = \mathbf{y}_{H,*}$. This allows us to finish the comparison with $\tau(Q)$ on lines 44 and 45.

Therefore, $Q^{\mathcal{M}_L} = \tau(Q)^{\mathcal{M}_H}$ for all $\tau(Q) \in \mathcal{L}_3(\mathcal{M}_H)$, concluding the proof. □

A.3. Proofs of Sec. 3

The proofs in this section are concerned with the properties of the partially projected C-DAG in Def. 8.

First, we must define what it means for a causal graph to be "sufficient". In general, the role of causal graphs in causal inference tasks is typically to encode the constraints of the model, which are useful for allowing one to make inferences of higher layers using lower layer data. These constraints, on layers 1, 2, and 3 of the PCH, can be described by the Counterfactual Bayesian Network, defined below.

Definition 17 (Counterfactual Bayesian Network (CTF-BN) (Correa & Bareinboim, 2024, Def. D.1, D.2))**.** Let $\mathbf{P}_{**}$ be the collection of all distributions of the form $P(W_{1[x_1]}, W_{2[x_2]}, \dots)$, where $W_i \in \mathbf{V}$, $\mathbf{X}_i \subseteq \mathbf{V}$, $\mathbf{x}_i \in \mathcal{D}_{\mathbf{X}_i}$. A directed acyclic graph (possibly with bidirected edges) $\mathcal{G}$ is a Counterfactual Bayesian Network for $\mathbf{P}_{**}$ if:

(i) (Independence Restrictions) Let $\mathbf{W}_*$ be a set of counterfactuals of the form $W_{\mathbf{pa}_w}$, $\mathbf{Z}_1, \dots, \mathbf{Z}_l$ the c-components of $\mathcal{G}[\mathbf{V}(\mathbf{W}_*)]$ (two variables are in the same c-component if there is a bidirected path between them in $\mathcal{G}$ within the variables $\mathbf{V}(\mathbf{W}_*)$), and $\mathbf{Z}_{1*}, \dots, \mathbf{Z}_{l*}$ the corresponding partition over $\mathbf{W}_*$. Then $P(\mathbf{W}_*)$ factorizes as

$$P\left(\bigwedge_{W_{\mathbf{pa}_w} \in \mathbf{W}_*} W_{\mathbf{pa}_w}\right) = \prod_{j=1}^{l} P\left(\bigwedge_{W_{\mathbf{pa}_w} \in \mathbf{Z}_{j*}} W_{\mathbf{pa}_w}\right). \tag{46}$$

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

$$P(Y_{\mathbf{pa}_y, \mathbf{z}}, \mathbf{W}_*) = P(Y_{\mathbf{pa}_y}, \mathbf{W}_*). \tag{47}$$

(iii) (Local Consistency) For every variable Y with parents $\mathbf{Pa}_y$, let $\mathbf{X} \subseteq \mathbf{Pa}_y$, then for every set $\mathbf{Z} \subseteq \mathbf{V} \setminus (\mathbf{X} \cup \{Y\})$, and any set of counterfactuals $\mathbf{W}_*$, we have

$$P(Y_{\mathbf{z}} = y, \mathbf{X}_{\mathbf{z}} = \mathbf{x}, \mathbf{W}_*) = P(Y_{\mathbf{zx}} = y, \mathbf{X}_{\mathbf{z}} = \mathbf{x}, \mathbf{W}_*). \tag{48}$$

■

Although a full discussion of these constraints is out of the scope of this paper, the two that are particularly of insight in regards to the difference between C-DAGs and projected C-DAGs are points (i) and (ii) in the definition. In words, point (ii) is stating that a lack of a directed edge implies a lack of an interventional effect, and point (i) is says that a lack of a bidirected edge (or lack of unobserved confounding) implies independence of functions. One particularly useful result is that causal diagrams are guaranteed to satisfy the CTF-BN constraints of the distributions induced by the SCM that generated the graph, shown below.

Lemma 1 ((Correa & Bareinboim, 2024))**.** *For any SCM $\mathcal{M}$ inducing causal diagram $\mathcal{G}$, $\mathcal{G}$ is a CTF-BN for $\mathcal{L}_3(\mathcal{M})$.* ■

Using the constraints of the CTF-BN, we can state the results of Sec. 3 more formally.

Proposition 2 (C-DAG Insufficiency (Formal))**.** *There exists SCM $\mathcal{M}_L$ and constructive abstraction function τ defined over clusters $\mathbb{C}$ and $\mathbb{D}$ with C-DAG $\mathcal{G}_{\mathbb{C}}$ such that, for $\mathcal{M}_H$ that is $\mathcal{L}_3$-τ consistent with $\mathcal{M}_L$, $\mathcal{G}_{\mathbb{C}}$ is not a CTF-BN for $\mathcal{L}_3(\mathcal{M}_H)$.* ■

Proof. Since there exist projected C-DAGs that have a superset of the edges of the corresponding C-DAG, this result is a consequence of the necessity of projected C-DAGs, stated in Thm. 2. □

The above result states that C-DAGs are insufficient for the general case abstraction problem, where the AIC may be violated. The below result shows that projected C-DAGs have precisely the correct constraints.

Lemma 2. *Let $\mathcal{M}_H$ be the SCM generated from running Alg. 1 on $\mathcal{M}_L$ and τ. Then, the causal diagram of $\mathcal{M}_H$ is the projected C-DAG of $\mathcal{M}_L$ over $\mathbf{V}_H$.* ■

Proof. This proof considers a slight modification of Alg. 1 that incorporates AIC violators $\mathbf{V}_H^\dagger$. Specifically, in line 9, $\mathbf{Pa}_V$ can be split into $\mathbf{Pa}_V^0$ and $\mathbf{Pa}_V^\dagger$, where $\tau(\mathbf{Pa}_V^0) \cap \mathbf{V}_H^\dagger = \emptyset$, and $\delta(\mathbf{pa}_V^o, \mathbf{pa}_V^u)$ is only applied to $\mathbf{Pa}_V^\dagger$, while for parents in $\mathbf{Pa}_V^0$, δ is replaced by an arbitrary $\mathbf{pa}_V^0$ such that $\tau(\mathbf{pa}_V^0) = \mathbf{pa}_{V_H}$, since they all map to the same value due to the lack of AIC violation.

The causal diagram of $\mathcal{M}_H$ is a graph $\mathcal{G}_H = \langle \mathbf{V}_H, \mathbf{E} \rangle$. $\mathbf{E}$ must at least contain the edges of the C-DAG $\mathcal{G}_\mathbb{C}$, since every function of $\mathcal{M}_L$ is incorporated into $\mathcal{M}_H$. Extra edges are only added through line 9 of the algorithm, where δ may introduce new dependencies through $\mathbf{pa}_V^u$. If, for some $W \in \mathbf{Pa}_V$, $W \notin \mathbf{V}_H^\dagger$, then new edges are not added w.r.t. W. Otherwise, the existence of W^u may confound other functions that also take W^u as an input, implying rule 3 of Def. 8. For the other rules, as stated in line 8, W^u depends on its own parents $\mathbf{Pa}_W$ (or the grandparents of V), which implies rule 1 of Def. 8. Additionally, W^u also depends on $\mathbf{u}^c_{\tau(W)}$, which implies a dependence on any unobserved confounder that influences the parents of W, implying rule 2 of Def. 8. No other dependences are introduced through lines 8 and 9 of the algorithm, meaning that $\mathbf{E}$ contains precisely the edges of $\mathcal{G}^\dagger$ plus those introduced by the rules of Def. 8. Hence, $\mathcal{G}_H = \mathcal{G}^\dagger_\mathbb{C}$. □

Theorem 2 (Projected C-DAG Sufficiency and Necessity (Formal))**.** *Let $\mathcal{M}_L = \langle \mathbf{U}_L, \mathbf{V}_L, \mathcal{F}_L, P(\mathbf{U}_L) \rangle$ be a low-level model with causal diagram $\mathcal{G}$, and let $\tau : \mathcal{D}_{\mathbf{V}_L} \to \mathcal{D}_{\mathbf{V}_H}$ be a constructive abstraction function defined over clusters $\mathbb{C}$ and $\mathbb{D}$. Let $\mathbf{V}_H^\dagger$ be the set of AIC violators of τ. Let $\mathcal{G}^\dagger_\mathbb{C} = \langle \mathbf{V}_H, \mathbf{E} \rangle$ be the partially projected C-DAG of $\mathcal{G}$ w.r.t. $\mathbb{C}$ and $\mathbf{V}_H^\dagger$. Let $\mathcal{M}_H = \langle \mathbf{U}_H, \mathbf{V}_H, \mathcal{F}_H, P(\mathbf{U}_H) \rangle$ be a high-level model that is $\mathcal{L}_3$-τ consistent with $\mathcal{M}_L$. Then*

1. *(Sufficiency) $\mathcal{G}^\dagger_\mathbb{C}$ is a CTF-BN for $\mathcal{L}_3(\mathcal{M}_H)$*
2. *(Necessity) For any other graph $\mathcal{G}' = \langle \mathbf{V}_H, \mathbf{E}' \rangle$ such that $\mathcal{G}' \neq \mathcal{G}$ and $\mathcal{G}'$ is a CTF-BN for $\mathcal{L}_3(\mathcal{M}_H)$, it must be the case that $\mathbf{E} \subset \mathbf{E}'$.*

■

Proof. The proof for sufficiency is straightforward. Alg. 1 generates $\mathcal{M}_H$ that is $\mathcal{L}_3$-τ consistent with $\mathcal{M}_L$ by Thm 1. By Lemma 2, the causal diagram of $\mathcal{M}_H$ is $\mathcal{G}^\dagger_\mathbb{C}$. Then, by Lemma 1, $\mathcal{G}^\dagger_\mathbb{C}$ must be a CTF-BN for $\mathcal{L}_3(\mathcal{M}_H)$.

The proof for necessity is more involved. We argue that every single edge in $\mathbf{E}$ must be included for $\mathcal{G}^\dagger_\mathbb{C}$ to maintain the correct CTF-BN constraints.

First, at least every edge in the C-DAG $\mathcal{G}_\mathbb{C}$ is necessary. This is because every edge in the C-DAG corresponds to an edge in the original graph $\mathcal{G}$. This edge cannot be removed without adding new constraints to the original graph generated by $\mathcal{M}_L$.

Next, we step through each of the three rules of Def. 8 and argue that they must hold. For each of the rules, consider the basic case with $Z, X, Y \in \mathbf{V}_H$ (and their corresponding clusters $\mathbf{C}_Z, \mathbf{C}_X, \mathbf{C}_Y \in \mathbb{C}$).

1. If $Z \to X \to Y$ in $\mathcal{G}_\mathbb{C}$ and $X \in \mathbf{V}_H^\dagger$, consider the query $P(y_{\mathbf{pa}_y, z})$. According to Eq. 47, $P(y_{\mathbf{pa}_y, z}) = P(y_{\mathbf{pa}_y})$ if there is no edge from $Z \to Y$. However, we see that

$$
\begin{aligned}
& P(y_{\mathbf{pa}_y, z}) && (49)\\
&= P(\mathbf{c}_{Y[\sigma_{\mathbf{Pa}_{\mathbf{C}_Y}}, \sigma_{\mathbf{C}_Z}]}) && (50)\\
&= P(\mathbf{c}_{Y[\sigma_{\mathbf{Pa}_{\mathbf{C}_Y} \setminus \mathbf{C}_X}, \sigma_{\mathbf{C}_X}, \sigma_{\mathbf{C}_Z}]}) && (51)\\
&= \sum_{\mathbf{c}_X \in \mathcal{D}_{\mathbf{C}_X}} P(\mathbf{c}_{Y[\sigma_{\mathbf{Pa}_{\mathbf{C}_Y} \setminus \mathbf{C}_X}, \mathbf{c}_X, \sigma_{\mathbf{C}_Z}]}) P(\sigma_{\mathbf{C}_X} = \mathbf{c}_X) && (52)\\
&= \sum_{\mathbf{c}_X \in \mathcal{D}_{\mathbf{C}_X}} P(\mathbf{c}_{Y[\sigma_{\mathbf{Pa}_{\mathbf{C}_Y} \setminus \mathbf{C}_X}, \mathbf{c}_X, \sigma_{\mathbf{C}_Z}]}) P(\mathbf{c}_X \mid \tau(\mathbf{c}_X) = x, \mathbf{pa}_{\mathbf{C}_X}, \mathbf{u}^c_{\mathbf{C}_X}). && (53)
\end{aligned}
$$

   Clearly, if $Z \notin \mathbf{Pa}_Y$, then including $\sigma_{\mathbf{C}_Z}$ into the low-level query can impact the value of the query, since $Z \in \mathbf{Pa}_X$, so the right term in Eq. 53 depends on $\sigma_{\mathbf{C}_Z}$. This would break Eq. 47.

2. If $Z \dashleftarrow\!\!\dashrightarrow X \rightarrow Y$ in $\mathcal{G}_{\mathbb{C}}$ and $X \in \mathbf{V}_H^{\dagger}$, two types of edges must be considered.

   (a) If there is no bidirected edge between Z and Y, then according to Eq. 46, $P(y_{\mathbf{pa}_y}, z_{\mathbf{pa}_z}) = P(y_{\mathbf{pa}_y})P(z_{\mathbf{pa}_z})$. However, we see that

$$
\begin{aligned}
&P(y_{\mathbf{pa}_y}, z_{\mathbf{pa}_z}) && (54)\\
&= P(\mathbf{c}_{Y[\sigma_{\mathbf{Pa}_{\mathbf{C}_Y}}]}, \mathbf{c}_{Z[\sigma_{\mathbf{Pa}_{\mathbf{C}_Z}}]}) && (55)\\
&= P(\mathbf{c}_{Y[\sigma_{\mathbf{Pa}_{\mathbf{C}_Y}\setminus\mathbf{C}_X}, \sigma_{\mathbf{C}_X}]}, \mathbf{c}_{Z[\sigma_{\mathbf{Pa}_{\mathbf{C}_Z}}]}) && (56)\\
&= \sum_{\mathbf{c}_X \in \mathcal{D}_{\mathbf{C}_X}, \mathbf{pa}_{\mathbf{C}_Z} \in \mathcal{D}_{\mathbf{Pa}_{\mathbf{C}_Z}}} P(\mathbf{c}_{Y[\sigma_{\mathbf{Pa}_{\mathbf{C}_Y}\setminus\mathbf{C}_X}, \mathbf{c}_X]}, \mathbf{c}_{Z[\mathbf{pa}_{\mathbf{C}_Z}]}) P(\sigma_{\mathbf{C}_X} = \mathbf{c}_X) P(\sigma_{\mathbf{Pa}_{\mathbf{C}_Z}} = \mathbf{pa}_{\mathbf{C}_Z}). && (57)
\end{aligned}
$$

   From here,

$$
P(\sigma_{\mathbf{C}_X} = \mathbf{c}_X) = P(\mathbf{c}_X \mid \tau(\mathbf{c}_X) = x, \mathbf{pa}_{\mathbf{C}_X}, \mathbf{u}^c_{\mathbf{C}_X}) \tag{58}
$$

   and

$$
P(\mathbf{c}_{Y[\sigma_{\mathbf{Pa}_{\mathbf{C}_Y}\setminus\mathbf{C}_X}, \mathbf{c}_X]}, \mathbf{c}_{Z[\mathbf{pa}_{\mathbf{C}_Z}]}) = P\left(\mathbf{c}_{Y[\sigma_{\mathbf{Pa}_{\mathbf{C}_Y}\setminus\mathbf{C}_X}, \mathbf{c}_X]}, \bigwedge_{V \in \mathbf{C}_Z} f_V(\mathbf{pa}_V, \mathbf{u}_V) = v\right) \tag{59}
$$

   for v consistent with $\mathbf{c}_Z$. However, since there is a bidirected edge between Z and X, there may be a dependence between $\mathbf{U}_V$ for some $V \in \mathbf{C}_Z$ and $\mathbf{u}^c_{\mathbf{C}_X}$. This would make the independence between the two terms $y_{\mathbf{pa}_y}$, $z_{\mathbf{pa}_z}$ impossible, violating Eq. 46.

   (b) If there is no bidirected edge between X and Y, then according to Eq. 46, $P(y_{\mathbf{pa}_y}, x_{\mathbf{pa}_x}) = P(y_{\mathbf{pa}_y})P(x_{\mathbf{pa}_x})$. However, following the same argument as above, this cannot hold either because

$$
P(\mathbf{c}_{Y[\sigma_{\mathbf{Pa}_{\mathbf{C}_Y}\setminus\mathbf{C}_X}, \mathbf{c}_X]}, \mathbf{c}'_{X[\mathbf{pa}_{\mathbf{C}_X}]}) = P\left(\mathbf{c}_{Y[\sigma_{\mathbf{Pa}_{\mathbf{C}_Y}\setminus\mathbf{C}_X}, \mathbf{c}_X]}, \bigwedge_{V \in \mathbf{C}_X} f_V(\mathbf{pa}_V, \mathbf{u}_V) = v\right) \tag{60}
$$

   for v consistent with $\mathbf{c}'_X$. Certainly, there could be a dependence between $\mathbf{U}_V$ for some $V \in \mathbf{C}_X$ and $\mathbf{u}^c_{\mathbf{C}_X}$, since both terms influence the functionality of $\mathbf{C}_X$.

3. If $Z \leftarrow X \rightarrow Y$ in $\mathcal{G}_{\mathbb{C}}$ and $X \in \mathbf{V}_H^{\dagger}$, consider the query $P(y_{\mathbf{pa}_y}, z_{\mathbf{pa}_z})$. According to Eq. 46, $P(y_{\mathbf{pa}_y}, z_{\mathbf{pa}_z}) = P(y_{\mathbf{pa}_y})P(z_{\mathbf{pa}_z})$ if there is no bidirected edge between Z and Y. However, we see that

$$
\begin{aligned}
&P(y_{\mathbf{pa}_y}, z_{\mathbf{pa}_z}) && (61)\\
&= P(\mathbf{c}_{Y[\sigma_{\mathbf{Pa}_{\mathbf{C}_Y}}]}, \mathbf{c}_{Z[\sigma_{\mathbf{Pa}_{\mathbf{C}_Z}}]}) && (62)\\
&= P(\mathbf{c}_{Y[\sigma_{\mathbf{Pa}_{\mathbf{C}_Y}\setminus\mathbf{C}_X}, \sigma_{\mathbf{c}_X}]}, \mathbf{C}_{Z[\sigma_{\mathbf{Pa}_{\mathbf{C}_Z}\setminus\mathbf{C}_X}, \sigma_{\mathbf{C}_X}]}). && (63)
\end{aligned}
$$

   Note that $\sigma_{\mathbf{C}_X}$ is computed once for both terms, so clearly the two terms cannot be independent as they both depend on $\sigma_{\mathbf{C}_X}$. This would break Eq. 46.

With all rules covered, no edge can be removed without breaking the CTF-BN condition, ensuring that the edge set of $\mathcal{G}^{\dagger}_{\mathbb{C}}$ is minimal. □

Finally, we prove how the projected C-DAG can be used for cross-layer inferences by solving the abstraction identification problem. First consider the classical identification problem.

For the following proofs, consider the classical definition of identifiability.

Definition 18. Let Ω^* be the space containing all SCMs defined over endogenous variables $\mathbf{V}$. We say that a causal query Q is identifiable (ID) from the available data $\mathbb{Z}$ and the causal diagram $\mathcal{G}$ if $Q(\mathcal{M}_1) = Q(\mathcal{M}_2)$ for every pair of models $\mathcal{M}_1, \mathcal{M}_2 \in \Omega^*$ such that $\mathcal{M}_1$ and $\mathcal{M}_2$ both induce $\mathcal{G}$ and $\mathbb{Z}(\mathcal{M}_1) = \mathbb{Z}(\mathcal{M}_2)$. ■

Now we show how abstract identification is equivalent.

Theorem 3 (Dual Abstract ID (General))**.** *Consider a counterfactual query Q over $\mathbf{V}_L$, a constructive abstraction function τ w.r.t. clusters $\mathbb{C}$ and $\mathbb{D}$, a projected C-DAG $\mathcal{G}^\dagger_{\mathbb{C}}$, and data $\mathbb{Z}$ from $\mathbf{V}_L$. Q is τ-ID from $\mathcal{G}^\dagger_{\mathbb{C}}$ and $\mathbb{Z}$ if and only if $\tau(Q)$ is ID from $\mathcal{G}^\dagger_{\mathbb{C}}$ and $\tau(\mathbb{Z})$.* ■

Proof. Let Ω_L and Ω_H be the space of SCMs defined over $\mathbf{V}_L$ and $\mathbf{V}_H$ respectively, and let $\Omega_L(\mathcal{G}^\dagger_{\mathbb{C}})$ and $\Omega_H(\mathcal{G}^\dagger_{\mathbb{C}})$ be their corresponding subsets that induce graph $\mathcal{G}^\dagger_{\mathbb{C}}$. For clarity, $\mathcal{M}_L \in \Omega_L(\mathcal{G}^\dagger_{\mathbb{C}})$ if $\mathcal{G}^\dagger_{\mathbb{C}}$ is a partially projected C-DAG of its causal diagram $\mathcal{G}$ w.r.t. $\mathbb{C}$ and AIC violation set $\mathbf{V}^\dagger_H$. $\mathcal{M}_H \in \Omega_H(\mathcal{G}^\dagger_{\mathbb{C}})$ if $\mathcal{M}_H$ induces $\mathcal{G}^\dagger_{\mathbb{C}}$ as its causal diagram.

If Q is τ-ID from $\mathcal{G}^\dagger_{\mathbb{C}}$ and $\mathbb{Z}$, then every pair of $\mathcal{M}_L \in \Omega_L(\mathcal{G}^\dagger_{\mathbb{C}}), \mathcal{M}_H \in \Omega_H(\mathcal{G}^\dagger_{\mathbb{C}})$ such that $\mathcal{M}_H$ is $\mathbb{Z}$-τ consistent with $\mathcal{M}_L$ must have $\mathcal{M}_H$ be Q-τ consistent with $\mathcal{M}_L$. For all such $\mathcal{M}_H$, $\mathbb{Z}$-τ consistency and Q-τ consistency with $\mathcal{M}_L$ implies that $\mathcal{M}_H$ is $\tau(\mathbb{Z})$-consistent and $\tau(Q)$-consistent by Def. 13. For any pair $\mathcal{M}_1, \mathcal{M}_2 \in \Omega_H$ that induce $\mathcal{G}^\dagger_{\mathbb{C}}$, $\tau(\mathbb{Z})(\mathcal{M}_1) = \tau(\mathbb{Z})(\mathcal{M}_2)$ therefore implies that both $\mathcal{M}_1$ and $\mathcal{M}_2$ must be $\mathbb{Z}$-τ consistent with $\mathcal{M}_L$ and must therefore both be Q-τ consistent, so $\tau(Q)(\mathcal{M}_1) = \tau(Q)(\mathcal{M}_2)$. Hence, $\tau(Q)$ is ID from $\mathcal{G}_{\mathbb{C}}$ and $\tau(\mathbb{Z})$ by Def. 18.

Conversely, if $\tau(Q)$ is ID from $\mathcal{G}^\dagger_{\mathbb{C}}$ and $\tau(\mathbb{Z})$, then for any $\mathcal{M}_1, \mathcal{M}_2 \in \Omega_H$ that induces $\mathcal{G}^\dagger_{\mathbb{C}}$ such that $\tau(\mathbb{Z})(\mathcal{M}_1) = \tau(\mathbb{Z})(\mathcal{M}_2)$, it must be the case that $\tau(Q)(\mathcal{M}_1) = \tau(Q)(\mathcal{M}_2)$. For every $\mathcal{M}_L \in \Omega_L(\mathcal{G}^\dagger_{\mathbb{C}})$, Thm. 1 and Lemma 2 state that there exists some $\mathcal{M}_H \in \Omega_H(\mathcal{G}^\dagger_{\mathbb{C}})$ that is $\mathcal{L}_3$-τ consistent with $\mathcal{M}_L$, implying that $\mathcal{M}_H$ is both $\mathbb{Z}$-τ consistent and Q-τ consistent with $\mathcal{M}_L$. Since all $\mathcal{M}_H \in \Omega_H(\mathcal{G}^\dagger_{\mathbb{C}})$ that match in $\tau(\mathbb{Z})$ must also match in $\tau(Q)$, it must be the case that all such $\mathcal{M}_H$ that are $\mathbb{Z}$-τ consistent with $\mathcal{M}_L$ must also be Q-τ consistent with $\mathcal{M}_L$. Hence, by definition, Q is τ-ID from $\mathcal{G}^\dagger_{\mathbb{C}}$ and $\mathbb{Z}$. □

B. Additional Results

In this section, we add additional technical results and expand on the ideas presented in the main body.

B.1. Choosing an Intervention Mapping Definition

Consider a basic setting with only two variables $\mathbf{V}_L = \{X_L, Y\}$, where X_L is ternary ($\mathcal{D}_{X_L} = \{x_0, x_1, x_2\}$), and Y is binary ($\mathcal{D}_Y = \{y_0, y_1\}$). Let $\mathbf{V}_H = \{X_H, Y\}$, where $\mathcal{D}_{X_H} = \{x_A, x_B\}$, such that

$$\tau(x_L, y) = \begin{cases} (x_A, y) & x_L = x_0 \\ (x_B, y) & x_L = x_1, x_2, \end{cases} \tag{64}$$

that is, x_1 and x_2 are both mapped to the same high-level value x_B. Naturally, one may be interested in causal queries on the high-level model such as $P(Y_{X_H = x_B} = y_1)$. However, making no assumptions about the AIC or the structural equations and probability distributions of the low-level model, how would such a quantity be defined on the low-level?

When the AIC holds, the answer is simple, since the AIC would imply that $P(Y_{X_L = x_1} = y_1) = P(Y_{X_L = x_2} = y_1)$. Since both of these values are equal, it must be the case that $P(Y_{X_H = x_B} = y_1) = P(Y_{X_L = x_1} = y_1) = P(Y_{X_L = x_2} = y_1)$. When the AIC does not hold, however, the answer is ambiguous. It is possible that $P(Y_{X_L = x_1} = y_1) \neq P(Y_{X_L = x_2} = y_1)$, so the choice of $P(Y_{X_H = x_B} = y_1)$ is not clear.

To illustrate the full range of possible options of $P(Y_{X_H = x_B} = y_1)$, consider a perspective of the problem akin to the canonical model formulation used for causal partial identification (Balke & Pearl, 1997; Zhang et al., 2022). Note that there are eight possible functions from X_L to Y, since there are three possible values of X_L and two possible values for Y. Define $R_X = f_X(\mathbf{U})$, $R^0_Y = f_Y(X_L = x_0, \mathbf{U})$, $R^1_Y = f_Y(X_L = x_1, \mathbf{U})$, $R^2_Y = f_Y(X_L = x_2, \mathbf{U})$, all of which are random variables that depend on $\mathbf{U}$. Now define

$$p_{ijk\ell} = P(R_X = x_i, R^0_Y = y_j, R^1_Y = y_k, R^2_Y = y_\ell). \tag{65}$$

Note that $Y_{X_L = x_1} = y_1$ holds as long as $R^1_Y = y_1$ and $Y_{X_L = x_2} = y_1$ holds as long as $R^2_Y = y_1$. Expanding this result, we get

$$
\begin{aligned}
&P(Y_{X_L=x_1} = y_1) \qquad (66)\\
&= p_{0010} + p_{0011} + p_{0110} + p_{0111}\\
&+ p_{1010} + p_{1011} + p_{1110} + p_{1111}\\
&+ p_{2010} + p_{2011} + p_{2110} + p_{2111},\\
&P(Y_{X_L=x_2} = y_1) \qquad (67)\\
&= p_{0001} + p_{0011} + p_{0101} + p_{0111}\\
&+ p_{1001} + p_{1011} + p_{1101} + p_{1111}\\
&+ p_{2001} + p_{2011} + p_{2101} + p_{2111}.
\end{aligned}
$$

The terms that are colored black are terms that are contained in both equations. This implies that $P(Y_{X_H=x_B} = y_1)$ must at least contain all of the black terms and may potentially contain any of the colored terms to any proportion. In other words,

$$
\begin{aligned}
&P(Y_{X_H=x_B} = y_1) \qquad (68)\\
&\geq p_{0011} + p_{0111}\\
&+ p_{1011} + p_{1111}\\
&+ p_{2011} + p_{2111},
\end{aligned}
$$

and

$$
\begin{aligned}
&P(Y_{X_H=x_B} = y_1) \qquad (69)\\
&\leq p_{0010} + p_{0110} + p_{0001} + p_{0101} + p_{0011} + p_{0111}\\
&+ p_{1010} + p_{1110} + p_{1001} + p_{1101} + p_{1011} + p_{1111}\\
&+ p_{2010} + p_{2110} + p_{2001} + p_{2101} + p_{2011} + p_{2111}.
\end{aligned}
$$

The question is then how to choose which of these colored terms to include in the definition of $P(Y_{X_H=x_B} = y_1)$. It is entirely possible to define $P(Y_{X_H=x_B} = y_1)$ as simply being equal to $P(Y_{X_L=x_1} = y_1)$ or $P(Y_{X_L=x_2} = y_1)$ (i.e., choosing Eq. 66 or 67). It could also be defined as Eq. 68 or Eq. 69, which can be interpreted as the minimum or maximum possible value of the query. However, these choices are somewhat arbitrary and extreme—it is unlikely that a practitioner would intuitively mean one of these definitions when studying the high-level query $P(Y_{X_H=x_B} = y_1)$.

More specifically, the reason that the above definitions are undesirable is because they do not take into account the nuance of when $X_H = x_B$ should be interpreted as $X_L = x_1$ or as $X_L = x_2$. Indeed, all of the colored terms in the above equations show a disconnect between $Y_{X_L=x_1}$ and $Y_{X_L=x_2}$. For example, p_{1010} represents a case where $R^1_Y = y_1$ and $R^2_Y = y_0$, which means that Y will take the value of y_1 if $X_L = x_1$ and y_0 if $X_L = x_2$. In such cases, it is important to distinguish whether $X_L = x_1$ or $X_L = x_2$. In contrast, both $R^1_Y = y_1$ and $R^2_Y = y_1$ for the black terms. By interpreting $X_H = x_B$ as the disjunctive intervention $X_H = x_1 \vee x_2$, it becomes clear that the ambiguity largely has to do with which particular value is used as X_H. From the unit-level perspective, how should x_B be interpreted for any particular individual datapoint?

One answer to making this decision is to look at the natural value of the intervened variable. In this case, if the intervention $X_H = x_B$ is applied, one can check if X_L was originally going to be x_1 or x_2. In such cases, the blue terms would be included, while the red terms would be excluded. For example, in p_{1010}, $R^1_Y = y_1$ and $R^2_Y = y_0$. However, since $R_X = x_1$, we know that the natural value of $X_L = x_1$, so we would apply the value of R^1_Y instead of R^2_Y, implying that $Y = y_1$. In contrast, in p_{2010}, R^1_Y and R^2_Y are identical, but this time $R_X = x_2$, so we would apply the value of R^2_Y instead of R^1_Y, implying that $Y = y_0$. Such a definition would look like

$$
\begin{aligned}
&P(Y_{X_H=x_B} = y_1) \qquad (70)\\
&= p_{0011} + p_{0111} + p_{1011} + p_{1111} + p_{2011} + p_{2111}\\
&+ p_{1010} + p_{1110} + p_{2001} + p_{2101}\\
&+ \beta_1 p_{0010} + \beta_2 p_{0110} + \beta_3 p_{0001} + \beta_4 p_{0101},
\end{aligned}
$$

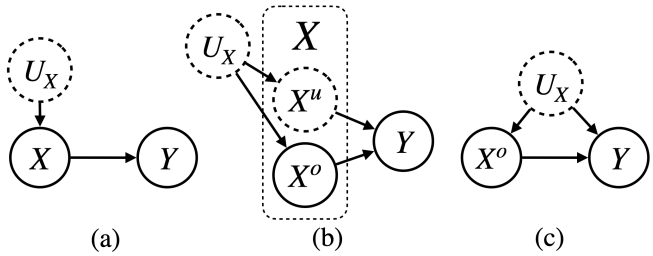

Figure 7: The confounding issue of using the natural value in partial projections. (a) In the original system, there is no confounding between X and Y. (b) When X is partially projected into X^o and X^u , there is now a path from U_X to Y both through X^o and X^u which is used for calculating the natural value of X. (c) The path through X^u results in confounding between the remaining X^o and Y.

for $\beta_1, \beta_2, \beta_3, \beta_4 \in [0, 1]$.

This seems like an appealing method of deciding between x_1 and x_2, but it has many issues. For one, it is unclear what should happen with the purple terms, in which $R^1_Y \neq R^2_Y$, but $R_X = x_0$. In other words, the natural value takes a value that does not map to the high-level value $X_H = x_B$, so it is unhelpful for deciding between whether x_B implies x_1 or x_2. Another issue is that the mechanisms of deciding the natural value of X_L are usually unobserved, so requiring this information adds a layer of unobserved confounding. Specifically, all variables will now be confounded with their parents since they will need the exogenous information used to generate the parents to find their natural value (see Fig. 7). This extra confounding adds difficulty in practical applications that require identifying the high-level quantities from minimal data.

In general, it is preferable to avoid the extra level of complication added by considering the natural value of variables given that the natural value is typically not observable in practice if an intervention was performed. For that reason, the interpretation of whether $X_H = x_B$ should be disambiguated as x_1 or x_2 should allow for either possibility without depending on the mechanism of X_L. This leads to the following formulation.

$$
\begin{aligned}
&P(Y_{X_H=x_B} = y_1) \\
&= \alpha p_{0010} + \alpha p_{0110} + (1-\alpha) p_{0001} + (1-\alpha) p_{0101} + p_{0011} + p_{0111} \\
&+ \alpha p_{1010} + \alpha p_{1110} + (1-\alpha) p_{1001} + (1-\alpha) p_{1101} + p_{1011} + p_{1111} \\
&+ \alpha p_{2010} + \alpha p_{2110} + (1-\alpha) p_{2001} + (1-\alpha) p_{2101} + p_{2011} + p_{2111},
\end{aligned} \tag{71}
$$

where $\alpha \in [0, 1]$. In other words,

$$
P(Y_{X_H=x_B} = y_1) = \alpha P(Y_{X_L=x_1} = y_1) + (1-\alpha) P(Y_{X_L=x_2} = y_1). \tag{72}
$$

This formulates the high-level intervention as a soft intervention over the low-level interventions, where α determines the probability of each possible low-level value. From the canonical model perspective, every possible value from Eq. 69 is considered, but its weight is determined by α. This resolves issues arising from using the natural value, since Eq. 72 can be computed for a fixed α as long as $P(Y_{X_L=x_1} = y_1)$ and $P(Y_{X_L=x_2} = y_1)$ can be computed. However, the question of choosing α remains. Indeed, an arbitrary value such as $\alpha = 0.5$ may not make the most sense. For example, if $P(X_L = x_1) >>> P(X_L = x_2)$, it may make more sense to pick a choice of α that is biased towards x_1. By this line of reasoning, the ideal choice of α should be

$$
\alpha = P(X_L = x_1 \mid X_L \in \{x_1, x_2\}) = \frac{P(X_L = x_1, X_L \in \{x_1, x_2\})}{P(X_L \in \{x_1, x_2\})} = \frac{P(X_L = x_1)}{P(X_L = x_1) + P(X_L = x_2)}. \tag{73}
$$

More generally, this choice of α implies that a high-level intervention should be a soft intervention over the corresponding low-level interventions with probabilities based on their proportions. Formally, for a high-level intervention $\mathbf{X}_H \leftarrow \mathbf{x}_H$, there is a corresponding soft intervention $\sigma_{\mathbf{X}_L}$ that is a distribution over all low level interventions $\mathbf{X}_L \leftarrow \mathbf{x}_L$ where $\tau(\mathbf{x}_L) = \mathbf{x}_H$.

In general, $\mathbf{X}_L$ must be a union of clusters for the abstraction mapping to be well defined, that is, $\mathbf{X}_L = \bigcup_{\mathbf{C}\in\mathbb{C}'} \mathbf{C}$ for some $\mathbb{C}' \subseteq \mathbb{C}$. $\sigma_{\mathbf{X}_L}$ must be decomposed into $\{\sigma_{\mathbf{C}} : \mathbf{C} \in \mathbb{C}'\}$, which must be sampled independently, otherwise having multiple interventions may introduce unintentional confounding. Following the above example, $\sigma_{\mathbf{C}_i}$ for some $\mathbf{C}_i \in \mathbb{C}$ (and corresponding $V_{H,i} \in \mathbf{V}_H$) would be defined as

$$P(\sigma_{\mathbf{C}_i} = \mathbf{c}_i) = P(\mathbf{c}_i \mid \tau(\mathbf{c}_i) = v_{H,i}). \tag{74}$$

Still, Eq. 74 may not be expressive enough for many applications. While this choice of α in Eq. 73 works for the two-variable study shown here, it may fail to hold in general cases with more variables. Consider the following example.

Example 4. Recall the setting discussed in Ex. 1. For convenience, the setting is described again here. Different insurance companies (Z) offer various insurance plans (X), which affect whether an insurance claim is approved (Y). For simplicity, suppose there are two insurance companies (z_1 and z_2) which offer three different insurance plans (x_1, x_2, and x_3), and the claim is either approved ($Y = 1$) or not approved ($Y = 0$). Suppose the true model $\mathcal{M}^* = \mathcal{M}_L = \langle \mathbf{U}_L, \mathbf{V}_L, \mathcal{F}_L, P(\mathbf{U}_L)\rangle$ is described as follows.

$$\begin{aligned}
\mathbf{U}_L &= \{U_Z, U_X^{z_1}, U_X^{z_2}, U_Y^{x_1}, U_Y^{x_2}, U_Y^{x_3}\} \\
\mathbf{V}_L &= \{Z, X, Y\} \\
\mathcal{F}_L &= \begin{cases} f_Z^L(u_Z) & = u_Z \\ f_X^L(z, u_X^{z_1}, u_X^{z_2}) & = u_X^z \\ f_Y^L(x, u_Y^{x_1}, u_Y^{x_2}, u_Y^{x_3}) & = u_Y^x \end{cases} \\
P(\mathbf{U}_L) &= \begin{cases} P(U_Z = z_1) = 0.7 \\ P(U_X^{z_1} = x_1) = 0.4, P(U_X^{z_1} = x_2) = 0.1, P(U_X^{z_1} = x_3) = 0.5 \\ P(U_X^{z_2} = x_1) = 0.1, P(U_X^{z_2} = x_2) = 0.4, P(U_X^{z_2} = x_3) = 0.5 \\ P(U_Y^{x_1} = 1) = 0.9, P(U_Y^{x_2} = 1) = 0.1, P(U_Y^{x_3} = 1) = 0.9 \end{cases}
\end{aligned} \tag{75}$$

The interpretation of the model is as follows: Insurance plans x_1 and x_3 are very effective, with 0.9 probability of claim acceptance, while x_2 is very ineffective at only 0.1 probability. Insurance company z_1 is more reputable than z_2 and is more likely to offer plan x_1 over x_2, while company z_2 prefers to offer plan x_2 over x_1.

Moreover, an important factor of consideration not shown in the model is that x_1 and x_2 are cheaper plans, while x_3 is more expensive. A data scientist who is studying this model may choose to abstract the different plans away, categorizing them simply as "cheap" and "expensive" plans. Formally, they would study a set of higher-level variables $\mathbf{V}_H = \{Z_H, X_H, Y_H\}$, where $Z_H = Z, Y_H = Y$, and X_H has a domain $\mathcal{D}_{X_H} = \{x_C, x_E\}$ corresponding to cheap and expensive plans respectively. There exists an abstraction function $\tau : \mathcal{D}_{\mathbf{V}_L} \to \mathcal{D}_{\mathbf{V}_H}$ such that τ maps x_1 and x_2 to x_C and maps x_3 to x_E. We will use the notation Z and Y instead of Z_H and Y_H since the variables are the same. One question that the data scientist may have is "What is the causal effect of choosing a cheap plan on claim acceptance rate?", denoted as $P(Y_{X_H=x_C} = 1)$.

First, we note that τ is a constructive abstraction function with clusters $\mathbb{C}$ and $\mathbb{D}$, where $\mathbb{C}$ and $\mathbb{D}$ trivially leaves the original variables and values in their own clusters, except x_1 and x_2 are clustered together. Under this choice of τ, observe that the AIC does not hold, notably that

$$0.9 = P(Y_{X=x_1} = 1) \neq P(Y_{X=x_2} = 1) = 0.1. \tag{76}$$

Then, how could one compute $P(Y_{X_H=x_C} = 1)$ given that both x_1 and x_2 map to x_C? The answer is that the intervention $X_H = x_C$ should correspond to a soft intervention on X, denoted as σ_X and assigning a different probability to x_1 and x_2 (but not to x_3, since $\tau(x_3)$ does not map to x_C). If $P(\sigma_X = x_1) = \alpha$ and $P(\sigma_X = x_2) = 1 - \alpha$, then it is clear that

$$P(Y_{X_H=x_C} = 1) = \alpha P(Y_{X=x_1} = 1) + (1 - \alpha)P(Y_{X=x_2} = 1), \tag{77}$$

for some choice of $\alpha \in [0, 1]$.

Still, this leaves the question of how to choose α. As a first attempt, it may be appealing to choose

$$\begin{aligned}\alpha = P(X = x_1 \mid X = x_1 \vee X = x_2) = \frac{P(X = x_1)}{P(X = x_1) + P(X = x_2)} = 0.5 \\ (1-\alpha) = P(X = x_2 \mid X = x_1 \vee X = x_2) = \frac{P(X = x_2)}{P(X = x_1) + P(X = x_2)} = 0.5,\end{aligned} \tag{78}$$

implying that $P(Y_{X_H=x_C} = 1) = 0.5$. Indeed, this choice has some appealing properties, notably that

$$P(Y_{X_H=x_C} = 1) = P(Y = 1 \mid X_H = x_C) = P(Y = 1 \mid X = x_1 \vee X = x_2). \tag{79}$$

One immediate observation that arises from this choice of σ_X is that the insurance company, Z, is not taken into account. Indeed, consider another query $P(Y_{X_H=x_C} = 1 \mid Z = z_1)$, which answers the question "What is the causal effect of choosing a cheap plan on claim acceptance rate given that the plan was provided by company z_1?" We would expect that, while $P(Y_{X_H=x_C} = 1) = 0.5$, it is very obvious that conditioning on $Z = z_1$ should change the result given that company z_1 is much more likely to recommend plan x_1 over x_2. However, with the choice of σ_X from Eq. 78, we would evaluate $P(Y_{X_H=x_C} = 1 \mid Z = z_1)$ to be equal to $P(Y_{X_H=x_C} = 1)$, since neither σ_X nor f_Y takes Z into account. To resolve this issue, it seems that a better choice of α may be deduced as follows

$$\begin{aligned}\alpha = P(X = x_1 \mid X = x_1 \vee X = x_2, Z) = \frac{P(x_1 \mid Z)}{P(x_1 \mid Z) + P(x_2 \mid Z)} \\ (1-\alpha) = P(X = x_2 \mid X = x_1 \vee X = x_2, Z) = \frac{P(x_2 \mid Z)}{P(x_1 \mid Z) + P(x_2 \mid Z)},\end{aligned} \tag{80}$$

which evaluates as $\alpha = 0.8$ when $Z = z_1$ and $\alpha = 0.2$ when $Z = z_2$. This translates to $P(Y_{X_H=x_C} = 1 \mid Z = z_1) = 0.74$, whereas $P(Y_{X_H=x_C} = 1 \mid Z = z_2) = 0.26$. ■

As illustrated in the example, the soft intervention applied to the low-level should not be agnostic of the parents of the intervened variables. This is shown visually in Fig. 2. Given that important information about X is lost through the abstraction, information from Z may be required in downstream functions to supplement the lost information. This brings us to the more general definition:

$$P(\sigma_{\mathbf{C}_i} = \mathbf{c}_i) = P(\mathbf{c}_i \mid \tau(\mathbf{c}_i) = v_{H,i}, \mathbf{pa}_{V_{H,i}}). \tag{81}$$

Note that this definition of $\sigma_{\mathbf{X}_L}$ does not take into account any exogenous variables of $\mathbf{U}_L$. This is sufficient in cases where the high-level model $\mathcal{M}_H$ is Markovian (i.e., no unobserved confounders), but this is not a reasonable assumption in most settings, as even Markovian low-level models can translate to non-Markovian high-level models after the abstraction function is applied (see row (c) of Fig. 3 for an example). The next section discusses why this definition is insufficient for non-Markovian cases and presents ideas on possible generalizations.

B.2. Non-Markovian Considerations

Eq. 81 is reasonable for Markovian cases, where there is no unobserved confounding (i.e., all $U_H \in \mathbf{U}_H$ are independent and are only parents for at most one $V_H \in \mathbf{V}_H$). The interpretation of Eq. 81 is that the disambiguation of low-level interventions for any given high-level intervention should depend on the endogenous parents of the intervened variables, but the exogenous parents should be ignored and resampled. Having a dependence on the exogenous variables would result in identifiability issues (e.g., from newly generated confounding as visualized in Fig. 7). However, if the exogenous parents are causing unobserved confounding, it does not make sense to simply ignore them. Consider the following example.

Example 5. Consider the same setting as Ex. 4, except in the data collection process, data is collected on hospitals instead of insurance companies. That is, for each person, Z is recorded as their registered hospital instead of their insurance company, which is now left unobserved. For the sake of simplicity, Z will stay as a binary variable, representing two possible hospitals z_1 and z_2. It turns out that people will choose their hospital based on which hospitals are covered by their insurance company, which now serves as an unobserved confounder between hospital choice and insurance plan. The full SCM $\mathcal{M}^* = \mathcal{M}_L = \langle \mathbf{U}_L, \mathbf{V}_L, \mathcal{F}_L, P(\mathbf{U}_L) \rangle$ is described as follows.

$$
\begin{aligned}
\mathbf{U}_L &= \{U_Z, U_X^{z_1}, U_X^{z_2}, U_Y^{x_1}, U_Y^{x_2}, U_Y^{x_3}\} \\
\mathbf{V}_L &= \{Z, X, Y\} \\
\mathcal{F}_L &= \begin{cases} f_Z^L(u_Z) & = u_Z \\ f_X^L(u_Z, u_X^{z_1}, u_X^{z_2}) & = u_X^{u_Z} \\ f_Y^L(x, u_Y^{x_1}, u_Y^{x_2}, u_Y^{x_3}) & = u_Y^x \end{cases} \\
P(\mathbf{U}_L) &= \begin{cases} P(U_Z = z_1) = 0.7 \\ P(U_X^{z_1} = x_1) = 0.4, P(U_X^{z_1} = x_2) = 0.1, P(U_X^{z_1} = x_3) = 0.5 \\ P(U_X^{z_2} = x_1) = 0.1, P(U_X^{z_2} = x_2) = 0.4, P(U_X^{z_2} = x_3) = 0.5 \\ P(U_Y^{x_1} = 1) = 0.9, P(U_Y^{x_2} = 1) = 0.1, P(U_Y^{x_3} = 1) = 0.9 \end{cases}
\end{aligned} \tag{82}
$$

Note that the only different between this SCM and the one from Eq. 75 is that instead of Z, f_X now takes U_Z as input. However, the behavior of the two SCMs are identical on the observational level, and moreover, if Z is projected away, the rest of the SCM is completely the same. Therefore, $P(Y_{X_H=x_C} = 1 \mid z)$ should be the same as the result computed in Eq. 80. However, this is obviously not the case when applying Eq. 81, since Z is no longer a parent of X.

Indeed, the computation of $P(Y_{X_H=x_C} = 1 \mid z)$ according to Eq. 81 can now be shown as follows.

$$
\begin{aligned}
& P(Y_{X_H=x_C} = 1 \mid z) && (83)\\
& = P(Y_{\sigma_{X_L}(x_C)=1} \mid z) && (84)\\
& = P(X_L = x_1 \mid \tau(X_L) = x_C) P(Y_{X_L=x_1} \mid z) + P(X_L = x_2 \mid \tau(X_L) = x_C) P(Y_{X_L=x_2} \mid z) && (85)\\
& = P(X_L = x_1 \mid \tau(X_L) = x_C) P(Y_{X_L=x_1}) + P(X_L = x_2 \mid \tau(X_L) = x_C) P(Y_{X_L=x_2}) && (86)\\
& = 0.5. && (87)
\end{aligned}
$$

Notably, line 85 applies Eq. 81, which no longer includes z in the probability of choosing the low-level intervention on X_L, and line 86 follows since Z and Y are independent when intervening on X_L. ■

The discrepancy in the above example follows from the issue that Eq. 81 makes a distinction between whether a variable's parent is endogenous or exogenous. In this particular example, the issue could be solved by modifying Eq. 81 to include U_Z instead of Z. However, it is unclear why U_Z should be included but not $U_X^{z_1}$ or $U_X^{z_2}$. Even in this example, SCM $\mathcal{M}_L$ could be designed in a way that behaves identically, but the exogenous space is chosen differently. For example U_Z, $U_X^{z_1}$, and $U_X^{z_2}$ could be subsumed into a Gaussian distribution, and their behavior can be mimicked using the inverse integral transform. In such a case, one could not pick and choose individual variables from $\mathbf{U}_L$ to include in the low-level soft intervention.

The key insight for solving this problem in the non-Markovian setting is to find a way to disentangle the confounded parts of the exogenous variables from the parts that are only influencing individual variables. For example, perhaps $\mathbf{U}_X$ could be split into $\mathbf{U}_X^c$ and $\mathbf{U}_X^u$, where $\mathbf{U}_X^c$ are all the exogenous variables that affect X and also some other variable, while $\mathbf{U}_X^u$ only affects X. Moreover, $\mathbf{U}_X^u$ needs to be chosen in a way that is "maximal", so as to not allow arbitrary flexibility between whether a variable belongs in $\mathbf{U}_X^u$ or $\mathbf{U}_X^c$.

Once again, to solve this problem, one can leverage the principles of canonical models (Balke & Pearl, 1997; Zhang et al., 2022). For any high-level variable X_H, define R_{X_H} as a random variable, where $\mathcal{D}_{R_{X_H}}$ consists of all possible functions of f_{X_H} w.r.t. $\mathbf{Pa}_{X_H}$. Note that for a fixed choice of $\mathbf{U}_{X_H}$, f_{X_H} is a deterministic function w.r.t. $\mathbf{Pa}_{X_H}$. Hence,

$$
P(R_{X_H} = r_{X_H}) = \sum_{\mathbf{u} \in \mathcal{D}_{\mathbf{U}_{X_H}} : f_{X_H}(\cdot, \mathbf{u}) = r_{X_H}} P(\mathbf{u}). \tag{88}
$$

For any high-level variable X_H, denote $\mathbf{V}_H^c(X_H) \subseteq \mathbf{V}_H \setminus \{X_H\}$ as the set of variables of $\mathbf{V}_H$ that share an confounding exogenous variable with X_H. Finally, denote $R_{X_H}(\mathbf{u}')$ as the random variable R_{X_H} over the distribution $P(R_{X_H} = r_{X_H} \mid \mathbf{U}' = \mathbf{u}')$ for some $\mathbf{U}' \subseteq \mathbf{U}$.

Now redefine $\sigma_{\mathbf{C}_i}$ as

$$
P(\sigma_{\mathbf{C}_i} = \mathbf{c}_i) = P(\mathbf{c}_i \mid \tau(\mathbf{c}_i) = v_{H,i}, \mathbf{pa}_{V_{H,i}}, \mathbf{R}_{\mathbf{V}_H^c(V_{H,i})}(\mathbf{u}_{\mathbf{C}_i})). \tag{89}
$$

This now matches Eq. 7 in Sec. 2, with $\mathbf{u}^c_{V_{H,i}}$ being used as a shorthand for $\mathbf{R}_{\mathbf{V}^c_H(V_{H,i})}(\mathbf{u}_{\mathbf{C}_i})$. Intuitively, the soft intervention over $\mathbf{C}_i$ now also depends on $\mathbf{U}_{\mathbf{C}_i}$ but only in the way that it affects the functions of the confounded neighbors of $V_{H,i}$. Notably $\mathbf{V}^c_H(V_{H,i})$ does not contain $V_{H,i}$ itself, so $\mathbf{U}_{\mathbf{C}_i}$ is still free to vary in ways that affect $f_{V_{H,i}}$ but not any other function.

Two important points must be clarified to avoid ambiguity when considering queries that contain interventions over multiple variables. First, $\sigma_{\mathbf{C}_i}(V_{H,i}, \mathbf{pa}_{V_{H,i}}, \mathbf{u}_{\mathbf{C}_i})$ is applied at most once for each value of $v_{H,i}$ in $\tau(Q)$, so if there are multiple terms $\mathbf{Y}_{H,i[\mathbf{x}_{H,i}]}$ that share the same intervention (e.g., $V_{H,i} = v_{H,i}$ in both $\mathbf{x}_{H,1}$ and $\mathbf{x}_{H,2}$), then $\sigma_{\mathbf{C}_i}$ is only sampled once and is used for both terms. However, if $V_{H,i} = v_{H,i}$ in $\mathbf{x}_{H,1}$ but $V_{H,i} = v'_{H,i}$ in $\mathbf{x}_{H,2}$, then it is sampled separately even though $V_{H,i}$ is in both terms. Second, if two high-level variables in the same intervention are confounded, interventions on both variables are performed according to $\sigma_{\mathbf{C}_i}$ with $\mathbf{V}^c_H(V_{H,i})$ remaining the same, ignoring the fact that the confounded neighbor is being intervened. These conditions are set to allow low-level queries to match corresponding high-level queries without generating semantic differences between identical high-level queries that are written in different forms (e.g., $P(Y_x, Z_x) = P(\{Y, Z\}_x)$).

C. Additional Examples

In this section, we provide additional examples to illustrate the key points of the paper.

The main limitation that this paper aims to address is the requirement of the abstract invariance condition (AIC) in Def. 5. A commonly cited example of this issue is about the abstraction of the two types of cholesterol, HDL and LDL, as shown below.

Example 6. Consider a study on the effects of diet on heart disease. Having an unhealthy diet (X) can raise the risk of heart disease (Y) depending on its cholesterol content. Cholesterol comes in two forms, called high-density and low-density lipoproteins (HDL and LDL, respectively), where HDL is believed to lower heart disease risk while LDL increases it (Steinberg, 2007; Truswell, 2010). Suppose the study is simplified to binary variables, and the true model $\mathcal{M}_L$ is:

$$\mathbf{U}_L = \{U_X, U_{C1}, U_{C2}, U_Y\} \tag{90}$$

$$\mathbf{V}_L = \{X, HDL, LDL, Y\} \tag{91}$$

$$\mathcal{F}_L = \begin{cases} X \leftarrow f^L_X(u_X) = u_X \\ HDL \leftarrow f^L_{HDL}(x, u_{C1}) = x \oplus u_{C1} \\ LDL \leftarrow f^L_{LDL}(x, u_{C2}) = x \oplus u_{C2} \\ Y \leftarrow f^L_Y(hdl, ldl, u_Y) = (ldl \wedge \neg hdl) \oplus u_Y \end{cases} \tag{92}$$

$$P(\mathbf{U}_L) = \begin{cases} P(U_X = 1) = 0.5 \\ P(U_{C1} = 1) = 0.1 \\ P(U_{C2} = 1) = 0.1 \\ P(U_Y = 1) = 0.1 \end{cases} \tag{93}$$

It can be computed from $\mathcal{M}_L$ that a person is more likely to get heart disease if their diet consists of higher LDL levels and lower HDL levels, notably

$$P^{\mathcal{M}_L}(Y_{HDL=0,LDL=1} = 1) = 0.9, \tag{94}$$

$$P^{\mathcal{M}_L}(Y_{HDL=1,LDL=0} = 1) = 0.1. \tag{95}$$

Now, suppose a data scientist decides to abstract HDL and LDL together into a variable called "total cholesterol" (TC), defined as

$$TC = HDL + LDL. \tag{96}$$

This naturally leads to the choice of intervariable clusters

$$\mathbb{C} = \{\mathbf{C}_1 = \{X\}, \mathbf{C}_2 = \{HDL, LDL\}, \mathbf{C}_3 = \{Y\}\}, \tag{97}$$

and intravariable clusters

$$\mathbb{D}_{\mathbf{C}_2} = \begin{cases} tc_0 & = \{(HDL = 0, LDL = 0)\} \\ tc_1 & = \{(HDL = 0, LDL = 1), \\ & \quad (HDL = 1, LDL = 0)\} \\ tc_2 & = \{(HDL = 1, LDL = 1)\}. \end{cases} \tag{98}$$

For the other clusters, the variables remain the same. Let τ be the constructive abstraction function defined with this choice of $\mathbb{C}$ and $\mathbb{D}$ (i.e. $\tau_{\mathbf{C}_2}(hdl, ldl) = hdl + ldl$).

A violation of the AIC arises due to the grouping of values $(HDL = 0, LDL = 1)$ and $(HDL = 1, LDL = 0)$ into the same intravariable cluster. To witness, note that $\tau_{\mathbf{C}_1}(HDL = 0, LDL = 1) = \tau_{\mathbf{C}_2}(HDL = 1, LDL = 0) = (TC = 1)$. Consider two queries $Q_1 = P(Y_{HDL=0,LDL=1} = 1)$ and $Q_2 = P(Y_{HDL=1,LDL=0} = 1)$, and recall from Eqs. 94 and 95 that $Q_1^{\mathcal{M}_L} = 0.9$ and $Q_2^{\mathcal{M}_L} = 0.1$. However, since $\tau_{\mathbf{C}_1}(HDL = 0, LDL = 1) = \tau_{\mathbf{C}_2}(HDL = 1, LDL = 0) = (TC = 1)$, both Q_1 and Q_2 have the same high-level counterpart (i.e., $\tau(Q_1) = \tau(Q_2) = P(Y_{TC=1} = 1)$). No choice of $\mathcal{M}_H$ over $\mathbf{V}_H$ can be both Q_1-τ consistent and Q_2-τ consistent with $\mathcal{M}_L$ because $P^{\mathcal{M}_H}(Y_{TC=1} = 1)$ cannot both be equal to 0.9 and 0.1.

This holds true fundamentally on the SCM-level as well. Note that a τ-abstraction with this choice of τ cannot exist for $\mathcal{M}_L$ for similar reasons. Specifically, note that

$$Y_{L[HDL=0,LDL=1]}(U_Y = 0) = 1, \tag{99}$$
$$Y_{L[HDL=1,LDL=0]}(U_Y = 0) = 0, \tag{100}$$

but $Y_{H[TC=1]}(\tau_{\mathbf{U}}(U_Y = 0))$ cannot both be equal to 0 and 1. This violates Eq. 17, implying that no such τ-abstraction can exist. ■

Below, we give an example of an SCM projection followed by a partial SCM projection for comparison.

Example 7. For concreteness, consider a setting in which different insurance companies (Z) offer various insurance plans (X), which affect whether an insurance claim is approved (Y). For simplicity, suppose there are two insurance companies (z_1 and z_2) which offer three different insurance plans (x_1, x_2, and x_3), and the claim is either approved ($Y = 1$) or not approved ($Y = 0$). Suppose the true model $\mathcal{M}^* = \mathcal{M}_L$ is described as follows.

$$\mathcal{M}_L = \begin{cases} \mathbf{U}_L & = \{U_Z, U_X^1, U_X^2, U_Y^1, U_Y^2, U_Y^3\} \\ \mathbf{V}_L & = \{Z, X, Y\} \\ \mathcal{F}_L & = \begin{cases} f_Z^L(u_Z) & = u_Z \\ f_X^L(z, u_X^1, u_X^2) & = \begin{cases} u_X^1 & z = z_1 \\ u_X^2 & z = z_2 \end{cases} \\ f_Y^L(x, u_Y^1, u_Y^2, u_Y^3) & = \begin{cases} u_Y^1 & x = x_1 \\ u_Y^2 & x = x_2 \\ u_Y^3 & x = x_3 \end{cases} \end{cases} \\ P(\mathbf{U}_L) & = \begin{cases} P(U_Z = z_1) = P(U_Z = z_2) = 0.5 \\ P(U_X^1 = x_1) = 0.4, P(U_X^1 = x_2) = 0.1, P(U_X^1 = x_3) = 0.5 \\ P(U_X^2 = x_1) = 0.1, P(U_X^2 = x_2) = 0.4, P(U_X^2 = x_3) = 0.5 \\ P(U_Y^1 = 1) = 0.9, P(U_Y^2 = 1) = 0.1, P(U_Y^3 = 1) = 0.9 \end{cases} \end{cases} \tag{101}$$

The interpretation of the model is as follows: Insurance plans x_1 and x_3 are very effective, with 0.9 probability of claim acceptance, while x_2 is very ineffective at only 0.1 probability. Insurance company z_1 is more reputable than z_2 and is more likely to offer plan x_1 over x_2, while company z_2 prefers to offer plan x_2 over x_1.

A data scientist may be interested in studying which insurance company (z_1 or z_2) is the better company for getting claims approved. In this case, the specific plan X that is being used may not be relevant. One may wish to instead study only the set of variables $\{Z, Y\}$, excluding X from the set. In other words, the SCM of interest is the *SCM projection* of $\mathcal{M}_L$ to the variable set $\mathbf{V}_H = \{Z, Y\}$. The SCM projection of $\mathcal{M}_L$ over $\mathbf{V}_H$ is quite straightforward to specify.

$$\mathcal{M}_H = \begin{cases} \mathbf{U}_H & = \mathbf{U}_L \\ \mathbf{V}_H & = \{Z, Y\} \\ \mathcal{F}_H & = \begin{cases} f_Z^H(u_Z) & = u_Z \\ f_Y^H(z, u_X^1, u_X^2, u_Y^1, u_Y^2, u_Y^3) & = \begin{cases} u_Y^1 & f_X^L(z, u_X^1, u_X^2) = x_1 \\ u_Y^2 & f_X^L(z, u_X^1, u_X^2) = x_2 \\ u_Y^3 & f_X^L(z, u_X^1, u_X^2) = x_3 \end{cases} \end{cases} \\ P(\mathbf{U}_L) & = P(\mathbf{U}_H) \end{cases} \tag{102}$$

With X excluded from the model, the functionality of X is projected into the function of its child, Y. Hence, the natural construction of the SCM projection $\mathcal{M}_H$ is simply the same as the construction of $\mathcal{M}_L$, but with f_Y computing X internally using f_X (comparing Eq. 3 with Eq. 102). It is not difficult to verify that computations of values of Z and Y under any choice of $\mathbf{U}_L$ remains the same in both models. Consequently, the induced PCH distributions are also the same, and $\mathcal{M}_H$ can be viewed simply as $\mathcal{M}_L$ but ignoring X.

■

Example 8. Continuing the insurance example in Ex. 7, suppose an important factor of consideration not shown in the model is that x_1 and x_2 are cheaper insurance plans, while x_3 is more expensive. A data scientist who is studying this model may choose to abstract the different plans away, categorizing them simply as "cheap" and "expensive" plans. Formally, they would study a set of higher-level variables $\mathbf{V}_H = \{Z_H, X_H, Y_H\}$, where $Z_H = Z$, $Y_H = Y$, and X_H has a domain $\mathcal{D}_{X_H} = \{x_C, x_E\}$ corresponding to cheap and expensive plans respectively. There exists an abstraction function $\tau : \mathcal{D}_{\mathbf{V}_L} \to \mathcal{D}_{\mathbf{V}_H}$ such that τ maps x_1 and x_2 to x_C and maps x_3 to x_E. We will use the notation Z and Y instead of Z_H and Y_H since the variables are the same. Note that in the new abstraction model $\mathcal{M}_H$, X is not removed entirely, but it is reduced down to only two possible values instead of three.

One possible method of accounting for this is as follows. First, redefine X into two parts, X^o and X^u, where X^o represents the observed portion of X and X^u represents the unobserved portion. X^o can simply be defined as $\tau(X)$. However, when $X^o = x_C$, it is ambiguous whether $X = x_1$ or x_2. Define X^u as a binary variable, where, whenever $X^o = x_C$, $X^u = 0$ represents $X = x_1$ while $X^u = 1$ represents $X = x_2$. X^u can be thought of as an indicator variable disambiguating any loss of information of X^o. Putting everything together, one can construct $\mathcal{M}_H$ as follows.

$$\mathcal{M}_H = \begin{cases} \mathbf{U}_H & = \mathbf{U}_L \cup \{X^u\} \\ \mathbf{V}_H & = \{Z, X_H, Y\} \\ \mathcal{F}_H & = \begin{cases} f_Z^H(u_Z) & = u_Z \\ f_X^H(z, u_X^1, u_X^2) & = \tau(f_X^L(z, u_X^1, u_X^2)) \\ f_Y^H(z, x^o, x^u, u_Y^1, u_Y^2, u_Y^3) & = \begin{cases} u_Y^1 & x^o = x_C, x^u = 0 \\ u_Y^2 & x^o = x_C, x^u = 1 \\ u_Y^3 & x^o = x_E \end{cases} \end{cases} \\ P(\mathbf{U}_H) & = P(U_L)P(X^o \mid \mathbf{U}_L) \\ & P(X^o = 0 \mid \mathbf{U}_L) = P(X = x_1 \mid X \in \{x_1, x_2\}) \\ & P(X^o = 1 \mid \mathbf{U}_L) = P(X = x_2 \mid X \in \{x_1, x_2\}) \end{cases} \tag{103}$$

Note that in this model, f_Y^H is trying to retain the same functionality as f_Y^L, but it is only given X_H as input instead of X. To disambiguate between $X = x_1$ and $X = x_2$, which both map to $X_H = x_C$, it utilizes the new exogenous variable X^u, whose probability is based on the probability of whether X is x_1 or x_2. In doing so, f_Y^H can mimic the functionality of f_Y^L in the sense that the lost information for X is partially projected into the exogenous space. ■

D. Experimental Details

In this section, we add further details to the experiments

D.1. Projected C-DAG Experiment

The first experiment tests the necessity of the projected C-DAGs in an estimation task where the AIC does not hold. The setting is described by three variables $\mathbf{V}_L = \{Z, X, Y\}$, and the low level model is described as

- Z is a 10-dimensional one-hot encoding ($\mathcal{D}_Z = \{0,1\}^{10}$) of a digit from 0-9, and it samples one uniformly at random.
- X is an MNIST image ($\mathbb{R}^{3\times32\times32}$) of the digit of Z. It is colored either red or blue and is shaded either light or dark. If the digit is odd, there is a 0.9 probability that the color will be red and 0.1 that it will be blue. The odds are flipped if the digit is even. Blue digits have a 0.7 probability of being light and 0.3 of being dark, and the odds are flipped for red digits.
- Y is a label ($\mathcal{D}_Y = \{0,1\}$) that predicts whether X is red ($Y = 1$) or blue ($Y = 0$), but it is incorrect with 0.1 probability.

On the high level, Z and Y remain the same, but $\tau(X) = X_H$, where X_H is a binary variable ($\mathcal{D}_{X_H} = \{0,1\}$) that represents whether X is light or dark.

The corresponding causal diagram $\mathcal{G}$ is shown in the l.h.s. of Fig. 3(a), which is also the C-DAG $\mathcal{G}_{\mathbb{C}}$. The r.h.s. shows the projected C-DAG $\mathcal{G}^{\dagger}_{\mathbb{C}}$, which is a result of X being an AIC violator.

The query being estimated is $P(Y_{X_H=1} = 1 \mid Z = 0)$, or the probability that Y predicts red under the intervention of forcing the image to be a light image, and conditioning on the digit being 0. The results are shown in Fig. 5. Three different GAN-NCMs (Xia et al., 2023) are trained. The first (red line) is a $\mathcal{G}$-NCM that is trained directly on the low-level data and attempts to estimate the low-level query without abstractions. The second (yellow line) is a $\mathcal{G}_{\mathbb{C}}$-NCM trained on the high-level data and is constrained by the C-DAG. The third (blue line) is similar to the second except it is a $\mathcal{G}^{\dagger}_{\mathbb{C}}$-NCM, constrained by the projected C-DAG. 95% confidence intervals of the errors across 10 trials are plotted in the figure.

D.2. Colored MNIST Sampling Experiment

The second experiment shows the ability of causal generative models to generate samples from causal queries involving high-dimensional images. The setting is described by three variables $\mathbf{V}_L = \{D, C, I\}$, and the low level model is described as

- D and C are 10-dimensional one-hot encodings ($\mathcal{D}_D = \mathcal{D}_C = \{0,1\}^{10}$ representing digits from 0-9 and colors from a spectrum respectively. Each digit is correlated with a color, a consequence of confounding. The correlated colors are shown on the right side of Fig. 6. A digit has a 0.9 probability of being its assigned color with a 0.1 probability of deviating.
- I is a corresponding MNIST digit ($\mathcal{D}_I = \mathbb{R}^{3\times64\times64}$) with color C and digit D.

The corresponding causal diagram is shown on the left side of Fig. 6. The results are shown in Fig. 4, demonstrating the ability for each of the methods on the left to sample images from the queries on the top. The non-causal approach simply trains a conditional GAN to sample image given digit. The RNCM (Xia & Bareinboim, 2024) maps images to a learned representation (i.e., τ is learned), which serves as the high-level space. However, due to AIC limitations, the dimensionality of X_H must remain high. When $\mathcal{D}_{X_H} \in \mathbb{R}^{16}$, the RNCM is able to sample the digits properly. However, when $\mathcal{D}_{X_H} \in \{0,1\}$, the RNCM is unable to get enough expressivity from the representation to perform the sampling. In contrast, the projected sampling approach, which trains a sampling model on top of the high-level model to sample from Eq. 7, is still able to reproduce the images despite the low-dimensional representation.