

Causal Effect Identifiability in Causal Abstractions: A Strict Hierarchy of Criteria

Clément Yvernes

Univ Grenoble Alpes, CNRS, Grenoble INP, LIG

clement.yvernes@univ-grenoble-alpes.fr

Marianne Clausel

Université de Lorraine, CNRS, CRAN, 54000 Nancy

marianne.clausel@univ-lorraine.fr

Emilie Devijver

Univ Grenoble Alpes, CNRS, Grenoble INP, LIG

emilie.devijver@univ-grenoble-alpes.fr

Eric Gaussier

Univ Grenoble Alpes, CNRS, Grenoble INP, LIG

eric.gaussier@imag.fr

Abstract

Identifying causal effects from observational data typically requires a fully specified causal graph, an assumption that is often unrealistic in complex or high-dimensional settings. Causal abstractions address this limitation by retaining only partial causal information, thereby defining equivalence classes of compatible causal graphs. In this paper, we study causal effect identifiability directly at the level of these graph collections, rather than at the level of individual causal diagrams. We formalize several notions of identifiability in this setting and organize them into a hierarchy that clarifies their relationships and differences. In particular, we establish that a causal effect may be identified over a collection of graphs without admitting a single causal derivation that holds uniformly across all compatible graphs. These results help reveal gaps and clarify ambiguities in the existing literature.

1 Introduction

Knowing the effect of a treatment $\mathbf{X}$ on an outcome $\mathbf{Y}$, represented by the interventional distribution $P(\mathbf{Y} \mid do(\mathbf{X}))$, is central to many applications. Yet, performing interventions is often impractical due to ethical concerns, potential harm, or prohibitive costs. In such cases, causal effect identification aims to determine whether interventional quantities can be expressed solely in terms of observational data (Pearl, 2009). This identifiability problem is typically addressed by deriving graphical criteria on causal diagrams and corresponding do-free formulas. However, these approaches generally assume that the underlying causal diagram is fully specified, requiring prior knowledge of the relationships between all observed variables. In practice, such knowledge is rarely available, especially in complex or high-dimensional settings, which substantially limits the applicability of standard causal inference methods.

A more relaxed assumption than having access to a fully specified causal graph is to rely on a causal abstraction, a representation that captures partial causal information about the underlying system. Rather than fixing a single graph, such abstractions induce classes of compatible causal diagrams and allow reasoning at a higher level of description. However, each compatible causal diagram may induce different answers to the same interventional query. As a result, identifiability is no longer a unique notion, but depends on how one aggregates or quantifies validity across the collection of graphs. While several such notions have been proposed, they are usually developed independently and lack a common framework for comparison.

Causal abstractions have been studied from multiple complementary perspectives. On the one hand, several works focus on their theoretical properties, including compositionality (Rischel & Weichwald, 2021), the computation of optimal abstractions (Zennaro et al., 2022), the use of macro-variables or grouped variables as substitutes for micro-level variables (Parviainen & Kaski, 2017; Beckers & Halpern, 2018), and the quantification of information loss relative to the underlying causal model (Zennaro et al., 2023). On the other hand, causal abstractions have also been used as a tool in a variety of downstream settings, including causal discovery (Wahl et al., 2024), query identifiability on clusters of variable (Anand et al., 2023; Yvernes et al., 2025c), reinforcement learning (Zennaro et al., 2024), time series analysis (Assaad et al., 2022; 2024; Yvernes et al., 2025b), neural network analysis (Geiger et al., 2021), mechanistic interpretability (Geiger et al., 2025; Méloux et al., 2025), and hierarchical modeling (Weinstein & Blei, 2026). Despite this broad range of studies, existing notions of identifiability under causal abstractions remain largely developed in isolation, making them difficult to compare within a unified framework.

We address this gap by unifying several notions of identifiability over collections of causal diagrams. These notions form a strict hierarchy, allowing us to compare their strength and determine when one strictly subsumes another. This hierarchy can assess, for a given abstraction, whether existing theory misses identifiable queries or fully captures the identifiable ones. More precisely, our contributions are as follows:

- We unify several notions of identifiability over collections of causal graphs and establish a strict hierarchy between them, enabling systematic comparison of their relative strength,
- In particular, we show that
 - We establish that adjustment-based identifiability (ICA) does not imply uniform do-calculus identifiability (ICD), complementing the known result that ICD does not imply ICA.
 - We prove that graphical identifiability (IG) strictly subsumes ICD, resolving the conjecture of Yvernes et al. (2025a).

The remainder of the paper is organized as follows. Section 2 provides the necessary background, Section 3 introduces identifiability notions for collections of causal diagrams, and Section 4 establishes their strict hierarchy. Section 5 interprets this hierarchy and shows that, under a mild assumption, identifiability over a collection is determined by its maximal elements, allowing the collection to be reduced without loss of identifiability. Section 6 concludes, and the Appendix contains all the proofs.

2 Preliminaries

2.1 Foundations of Causal Modeling

Graphs. A *mixed graph* $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ is a graph that contains two type of arrows, *directed arrows* ($\rightarrow$) and *dashed-bidirected arrows* ($\leftrightarrow$). For a mixed graph $\mathcal{G}$, if $X \rightarrow Y$, then X is a *parent* of Y and Y is a *child* of X . A *path* is a sequence of distinct vertices such that there exists an arrow in $\mathcal{G}$ connecting each element to its successor. A *directed path* (or *causal path*) is a path consisting only of directed edges ($\rightarrow$), such that all arrows are oriented towards the last vertex. If there is a directed path from X to Y , then X is an *ancestor* of Y , and Y is a *descendant* of X .¹ The sets of parents, children, ancestors and descendants of X in $\mathcal{G}$ are denoted by $\text{Pa}(X, \mathcal{G})$, $\text{Ch}(X, \mathcal{G})$, $\text{Anc}(X, \mathcal{G})$ and $\text{Desc}(X, \mathcal{G})$ respectively. For two disjoint subsets $\mathbf{X}, \mathbf{Y} \subseteq \mathcal{V}$, a *path* from $\mathbf{X}$ to $\mathbf{Y}$ is a path from some $X \in \mathbf{X}$ to some $Y \in \mathbf{Y}$. A path from $\mathbf{X}$ to $\mathbf{Y}$ is *proper* if only its first node is in $\mathbf{X}$. A *directed cycle* is a circular list of distinct vertices in which each vertex is a parent of its successor. When a (directed) mixed graph does not contain any directed cycle, then this graph is called an *Acyclic Directed Mixed Graph (ADMG)*. When an ADMG does not contain any dashed-bidirected arrow, it is called a *Directed Acyclic Graph (DAG)*. A graph $\mathcal{G}_1 = (\mathcal{V}_1, \mathcal{E}_1)$ is said to be a subgraph of another graph $\mathcal{G}_2 = (\mathcal{V}_2, \mathcal{E}_2)$ if $\mathcal{V}_1 \subseteq \mathcal{V}_2$ and $\mathcal{E}_1 \subseteq \mathcal{E}_2$. In this case, we write $\mathcal{G}_1 \subseteq \mathcal{G}_2$, which defines a partial order over graphs.

On a path π , a vertex V is said to be a *collider* on π if it has two incoming arrows in π (i.e., $\rightarrow V \leftarrow$, or $\rightarrow V \leftrightarrow$, or $\leftrightarrow V \leftarrow$ or $\leftrightarrow V \leftrightarrow$). Otherwise, it is said to be a *non-collider* on π . A vertex V is said to be

¹Note that a vertex is a descendant and an ancestor of itself, since for all variable V the path containing V alone is a directed path from V to V .

active on a path relative to $\mathbf{Z}$ if 1) V is a collider and one of its descendants is in $\mathbf{Z}$ or 2) V is a non-collider and is not in $\mathbf{Z}$. A path π is said to be *active* given (or conditioned on) $\mathbf{Z}$ if every vertex on π is active relative to $\mathbf{Z}$. Otherwise, π is said to be *inactive* given $\mathbf{Z}$. Given a graph $\mathcal{G}$, the sets $\mathbf{X}$ and $\mathbf{Y}$ are said to be *d-separated* by $\mathbf{Z}$ if every path between $\mathbf{X}$ and $\mathbf{Y}$ is inactive given $\mathbf{Z}$. We denote this by $\mathbf{X} \perp\!\!\!\perp_{\mathcal{G}} \mathbf{Y} \mid \mathbf{Z}$. Otherwise, $\mathbf{X}$ and $\mathbf{Y}$ are said to be *d-connected* given $\mathbf{Z}$, which we denote by $\mathbf{X} \not\perp\!\!\!\perp_{\mathcal{G}} \mathbf{Y} \mid \mathbf{Z}$.

The mutilated graph $\mathcal{G}_{\overline{\mathbf{XZ}}}$ is obtained from $\mathcal{G}$ by removing all edges with an arrowhead into $\mathbf{X}$ (i.e., $A \rightarrow \mathbf{X}$ or $A \leftrightarrow \mathbf{X}$, for any vertex A) and all edges with a tail out of $\mathbf{Z}$ (i.e., $\mathbf{Z} \rightarrow A$, for any vertex A).

Structural Causal Models. A *Structural Causal Model (SCM)* $\mathcal{M}$ consists of a 4-tuple $\langle \mathcal{U}, \mathcal{V}, \mathcal{F}_{\mathcal{M}}, P(\mathcal{U}) \rangle$, where $\mathcal{U}$ is a set of exogenous (latent) mutually independent variables and $\mathcal{V}$ is a set of endogenous (measured) variables. $\mathcal{F}_{\mathcal{M}}$ is a collection of functions $\{f_i\}_{i=1}^{|\mathcal{V}|}$ such that each endogenous variable $V_i \in \mathcal{V}$ is a function $f_i \in \mathcal{F}_{\mathcal{M}}$ of $\mathcal{U}_i \cup \text{Pa}(V_i)$, where $\mathcal{U}_i \subseteq \mathcal{U}$ and $\text{Pa}(V_i) \subseteq \mathcal{V} \setminus \{V_i\}$. The uncertainty is encoded through a probability distribution over the exogenous variables, $P(\mathcal{U})$. The resulting joint distribution over $\mathcal{V}$, obtained by marginalizing out $\mathcal{U}$, is called the *observational distribution* and is denoted $P_{\mathcal{M}}(\mathcal{V})$. Each SCM $\mathcal{M}$ induces a mixed graph $\mathcal{G}_{\mathcal{M}} := (\mathcal{V}, \mathcal{E} = (\mathcal{E}_D, \mathcal{E}_B))$, known as a *causal diagram*, that encodes the structural relations among $\mathcal{V} \cup \mathcal{U}$, where every endogenous variable $V_i \in \mathcal{V}$ is a vertex. $\mathcal{G}_{\mathcal{M}}$ contains directed edges which connect each variable $V_i \in \mathcal{V}$ to its parents $V_j \in \text{Pa}(V_i)$ as $(V_j \rightarrow V_i)$, and bidirected edges between variables $V_i, V_j \in \mathcal{V}$ that share a common exogenous parent, i.e., such that $\mathcal{U}_i \cap \mathcal{U}_j \neq \emptyset$, denoted $V_i \leftrightarrow V_j$. Throughout this paper, we restrict attention to acyclic causal models, meaning that the associated causal diagrams are ADMGs.

Interventions. In a given SCM $\mathcal{M} = \langle \mathcal{U}, \mathcal{V}, \mathcal{F}, P(\mathcal{U}) \rangle$, let $\mathbf{V} = (V_1, \dots, V_n)$ be a subset of $\mathcal{V}$ and $\mathbf{v}$ be a valuation of $\mathcal{V}$. An intervention $\text{do}(\mathbf{V} = \mathbf{v})$ replaces each structural function f_i for $V_i \in \mathbf{V}$ by the constant assignment $V_i := v_i$, which in the causal diagram $\mathcal{G}_{\mathcal{M}}$ corresponds to removing all incoming edges into each V_i . The resulting joint distribution over $\mathcal{V}$, obtained by marginalizing out $\mathcal{U}$, is called the *interventional distribution* and is denoted $P_{\mathcal{M}}(\mathcal{V} \mid \text{do}(\mathbf{V} = \mathbf{v}))$ or $P_{\mathcal{M}}(\mathcal{V} \mid \text{do}(\mathbf{v}))$ for short. For any pairwise disjoint subsets $\mathbf{Y}, \mathbf{X}, \mathbf{W} \subseteq \mathcal{V}$, with $\mathbf{Y} \neq \emptyset$ and corresponding valuations $\mathbf{y}, \mathbf{x}, \mathbf{w}$, we will denote $P(\mathbf{Y} = \mathbf{y} \mid \text{do}(\mathbf{X} = \mathbf{x}), \mathbf{X} = \mathbf{w})$ by $P(\mathbf{y} \mid \text{do}(\mathbf{x}), \mathbf{w})$.

Compatibility A probability distribution P is said to be *compatible* with an ADMG $\mathcal{G}$ if there exists a SCM inducing $\mathcal{G}$ whose observational distribution is P . In particular, any distribution compatible with $\mathcal{G}$ is Markovian with respect to $\mathcal{G}$ (Proposition 6.31 in Peters et al. (2017)), that is, it satisfies the global Markov property, i.e., every d-separation relation in $\mathcal{G}$ implies the corresponding conditional independence in P .

Causal Queries & Identification Functionals. Let $\mathcal{V}$ denote the set of observable variables and let $\mathbf{X}, \mathbf{Y}, \mathbf{Z} \subseteq \mathcal{V}$ be pairwise disjoint, with $\mathbf{Y} \neq \emptyset$. A *causal query* is an expression of the form $P(\mathbf{y} \mid \text{do}(\mathbf{x}), \mathbf{z})$. An *identification functional* is a map ϕ from joint distributions over $\mathcal{V}$ to $\mathbb{R}^+$. Given an SCM $\mathcal{M}$, ϕ is *valid* for $P(\mathbf{y} \mid \text{do}(\mathbf{x}), \mathbf{z})$ if $\phi(P_{\mathcal{M}}(\mathcal{V})) = P_{\mathcal{M}}(\mathbf{y} \mid \text{do}(\mathbf{x}), \mathbf{z})$.

All probability expressions and equalities in this paper are understood on their natural domains of definition.

2.2 Identifiability in a Single Graph

Different SCMs may induce the same causal diagram which represents an equivalence class of SCMs. A causal query $P(\mathbf{y} \mid \text{do}(\mathbf{x}), \mathbf{z})$ is said to be *identifiable* in a graph $\mathcal{G}$ if all SCMs that induce $\mathcal{G}$ and the same observational distribution agree on the value of the query. Definition 1 formalizes this notion.

Definition 1 (Identifiability in a Single Graph). *Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be a causal diagram and let $\mathbf{X}, \mathbf{Y}, \mathbf{Z} \subseteq \mathcal{V}$ be pairwise disjoint, with $\mathbf{Y} \neq \emptyset$. The causal query $P(\mathbf{y} \mid \text{do}(\mathbf{x}), \mathbf{z})$ is identifiable in $\mathcal{G}$ if there exists an identification functional ϕ valid for $P(\mathbf{y} \mid \text{do}(\mathbf{x}), \mathbf{z})$ in every SCM $\mathcal{M}$ satisfying $\mathcal{G}_{\mathcal{M}} = \mathcal{G}$. In this case, ϕ is said to be valid for $P(\mathbf{y} \mid \text{do}(\mathbf{x}), \mathbf{z})$ in $\mathcal{G}$.*

Definition 1 formalizes identifiability as the existence of a valid identification functional for a given causal query. It is equivalent to Definition 3.2.3 in Pearl (2009), rewritten from a functional perspective. The following example illustrates this notion on a simple causal graph that we shall use later as a running example.

Example 1. Consider the DAG $\mathcal{G} := Z \overset{\curvearrowright}{\rightarrow} X \rightarrow Y$ and the query $P(y \mid \text{do}(x))$. Then the identification functional $\phi : P(\mathcal{V}) \mapsto \sum_z P(y \mid x, z)P(z)$ is valid for $P(y \mid \text{do}(x))$ in $\mathcal{G}$. The derivation of this functional will be given later.

However, this definition is purely semantic: it does not directly provide a procedure for determining whether a query is identifiable, nor for constructing an identification functional when one exists. Addressing these questions requires systematic tools for transforming causal queries into valid identification functional. Pearl's do-calculus provides such a framework through three transformation rules that enable reasoning about interventions directly from the graphical structure of the causal diagram. We restate these rules in Theorem 1.

Theorem 1 (Do-calculus rules, Pearl (2009)). *Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be an ADMG, and let $\mathbf{X}, \mathbf{Y}, \mathbf{Z}, \mathbf{W}$ be pairwise disjoint subsets of $\mathcal{V}$, with $\mathbf{Y} \neq \emptyset$. Then the following rules hold for all distributions compatible with $\mathcal{G}$:*

R1 *insertion/deletion of observations:*

$$\mathbf{Y} \perp\!\!\!\perp_{\mathcal{G}_{\overline{\mathbf{X}}}} \mathbf{Z} \mid \mathbf{X}, \mathbf{W} \Rightarrow P(\mathbf{y} \mid \text{do}(\mathbf{x}), \mathbf{z}, \mathbf{w}) = P(\mathbf{y} \mid \text{do}(\mathbf{x}), \mathbf{w}),$$

R2 *action/observation exchange:*

$$\mathbf{Y} \perp\!\!\!\perp_{\mathcal{G}_{\overline{\mathbf{X}}}\mathbf{Z}} \mathbf{Z} \mid \mathbf{X}, \mathbf{W} \Rightarrow P(\mathbf{y} \mid \text{do}(\mathbf{x}), \mathbf{z}, \mathbf{w}) = P(\mathbf{y} \mid \text{do}(\mathbf{x}), \mathbf{z}, \mathbf{w}),$$

R3 *insertion/deletion of actions :*

$$\mathbf{Y} \perp\!\!\!\perp_{\mathcal{G}_{\overline{\mathbf{X}}}\mathbf{Z}(\mathbf{W})} \mathbf{Z} \mid \mathbf{X}, \mathbf{W} \Rightarrow P(\mathbf{y} \mid \text{do}(\mathbf{x}), \mathbf{z}, \mathbf{w}) = P(\mathbf{y} \mid \text{do}(\mathbf{x}), \mathbf{w}),$$

where $\mathbf{Z}(\mathbf{W})$ is the set of $\mathbf{Z}$ -nodes non-ancestors of $\mathbf{W}$ in $\mathcal{G}_{\overline{\mathbf{X}}}$.

The do-calculus proceeds by repeatedly applying its rules to transform observational and interventional probabilities of the form $P(\mathbf{y} \mid \text{do}(\mathbf{x}), \mathbf{w})$. Starting from a causal quantity of interest, each application replaces the current expression with an equivalent one, with the validity of the transformation guaranteed by a graphical condition in the causal graph. Combining applications of do-calculus rules with standard probability manipulations, such as Bayes' rule and marginalization, enables causal reasoning.

In this paper, a *causal derivation* is any finite sequence of applications of do-calculus rules, possibly combined with standard probability manipulations, that transforms a causal quantity into another one. Such a derivation is said to *hold in* a causal diagram $\mathcal{G}$ if every d-separation condition required by the applied do-calculus rules is satisfied in $\mathcal{G}$. Standard probability manipulations are always interpreted within their usual domain of validity, and do not affect the notion of validity of the derivation. Given a set of variables $\mathcal{V}$, a derivation is said to be *on* $\mathcal{V}$ if every variable appearing in any expression throughout the derivation belongs to $\mathcal{V}$.

When a causal derivation rewrites a causal query $P(\mathbf{y} \mid \text{do}(\mathbf{x}), \mathbf{z})$ into a do-free expression $\phi(P(\mathcal{V}))$, $\phi(P(\mathcal{V}))$ depends only on the observational distribution $P(\mathcal{V})$ and ϕ is a valid identification functional for the query. In this case, ϕ is called an *identification formula* and the causal derivation is said to *identify* $P(\mathbf{y} \mid \text{do}(\mathbf{x}), \mathbf{z})$. Example 2 illustrates a causal derivation identifying a causal query in a given graph.

Example 2. Consider the DAG $\mathcal{G} := Z \overset{\curvearrowright}{\rightarrow} X \rightarrow Y$. Then $P(y \mid \text{do}(x))$ is identified by the following causal derivation:

$$\begin{aligned} P(y \mid \text{do}(x)) &= \sum_z P(y, z \mid \text{do}(x)) && \text{(by marginalisation)} \\ &= \sum_z P(y \mid \text{do}(x), z)P(z \mid \text{do}(x)) && \text{(by chain rule)} \\ &= \sum_z P(y \mid \text{do}(x), z)P(z) && \text{(by R3)} \\ &= \sum_z P(y \mid x, z)P(z) && \text{(by R2)} \end{aligned}$$

In this case, we recover Example 1 and an example of backdoor adjustment (Pearl, 2009).

Do-calculus is sound and complete for causal identification in causal diagrams (Shpitser & Pearl, 2006). Consequently, a causal query is identifiable in $\mathcal{G}$ if and only if there exists a causal derivation that rewrites the query into a do-free expression and holds in $\mathcal{G}$. Whenever such a derivation exists, it yields an identification formula for the query in $\mathcal{G}$. This principle underlies the ID algorithm (Shpitser & Pearl, 2006), which systematically constructs such a derivation whenever the query is identifiable.

In practice, identifiability can also be established through more direct *graphical criteria*, namely conditions expressed solely in terms of graphical properties that guarantee the existence of a valid identification functional. A standard example is the backdoor criterion (Pearl, 2009).

Definition 2 (Backdoor criterion (Pearl, 2009)). *Let X and Y be two variables in a causal graph. A set of variables $\mathbf{Z}$ satisfies the backdoor criterion with respect to (X, Y) if:*

1. *No element of $\mathbf{Z}$ is a descendant of X , and*
2. *$\mathbf{Z}$ blocks all paths between X and Y that contain an arrow into X .*

Theorem 2 (Backdoor adjustment (Pearl, 2009)). *Let X and Y be two variables in a causal graph. If a set of variables $\mathbf{Z}$ satisfies the backdoor criterion with respect to (X, Y) , then the causal effect of X on Y is identifiable and is given by*

$$P(y \mid \text{do}(x)) = \sum_{\mathbf{z}} P(y \mid x, \mathbf{z}) P(\mathbf{z}).$$

When no set satisfying the backdoor criterion exists, the causal effect may still be identifiable through an alternative graphical criterion. The frontdoor criterion (Pearl, 2009) provides a classical example of such an identification strategy. Rather than adjusting directly for variables that block backdoor paths from X to Y , it exploits variables lying on the directed paths from X to Y to identify the effect indirectly.

Definition 3 (Frontdoor criterion (Pearl, 2009)). *Let X and Y be two variables in a causal graph. A set of variables $\mathbf{Z}$ satisfies the frontdoor criterion with respect to (X, Y) if:*

1. *$\mathbf{Z}$ intercepts all directed paths from X to Y ;*
2. *there is no unblocked backdoor path from X to Y ; and*
3. *X blocks all paths from $\mathbf{Z}$ to Y that contain an arrow into any variable in $\mathbf{Z}$.*

Theorem 3 (Frontdoor adjustment (Pearl, 2009)). *Let X and Y be two variables in a causal graph. If a set of variables $\mathbf{Z}$ satisfies the frontdoor criterion with respect to (X, Y) and $P(x, \mathbf{z}) > 0$, then the causal effect of X on Y is identifiable and is given by*

$$P(y \mid \text{do}(x)) = \sum_{\mathbf{z}} P(\mathbf{z} \mid x) \sum_{x'} P(y \mid \mathbf{z}, x') P(x').$$

The backdoor criterion also admits a more general formulation. We therefore recall the generalized adjustment criterion of Perković et al. (2015), which characterizes valid adjustment sets; see also der Zander et al. (2014).

Definition 4 (Generalized Adjustment Criterion). *Let $\mathbf{X}, \mathbf{Y}$ and $\mathbf{Z}$ be pairwise distinct sets of variables in a causal diagram $\mathcal{G}$. $\mathbf{Z}$ is said to satisfy the adjustment criterion relative to $\mathbf{X}$ and $\mathbf{Y}$ in $\mathcal{G}$ if:*

1. *$\text{Forb}(\mathbf{X}, \mathbf{Y}, \mathcal{G}) \cap \mathbf{Z} = \emptyset$; and*
2. *$\mathbf{Z}$ blocks all proper non-causal paths from $\mathbf{X}$ to $\mathbf{Y}$ in $\mathcal{G}$.*

where the forbidden set $\text{Forb}(\mathbf{X}, \mathbf{Y}, \mathcal{G})$ is the set of all descendants of any $W \notin \mathbf{X}$ which lies on a proper causal path from $\mathbf{X}$ to $\mathbf{Y}$.

Theorem 4 (Adjustment(Perković et al., 2015)). *Let $\mathbf{X}, \mathbf{Y}$ and $\mathbf{Z}$ be pairwise distinct sets of variables in a causal diagram $\mathcal{G}$. Then the following propositions are equivalent:*

- $\mathbf{Z}$ satisfies the adjustment criterion relative to $\mathbf{X}$ and $\mathbf{Y}$ in $\mathcal{G}$
- For all P Markovian with respect to $\mathcal{G}$, we have

$$P(\mathbf{y} \mid \text{do}(\mathbf{x})) = \begin{cases} P(\mathbf{y} \mid \mathbf{x}) & \text{if } \mathbf{Z} = \emptyset, \\ \sum_{\mathbf{z}} P(\mathbf{y} \mid \mathbf{x}, \mathbf{z}) P(\mathbf{z}) & \text{otherwise.} \end{cases}$$

3 Identifiability in a Collection of Graphs.

When a single causal diagram is considered, do-calculus provides a complete characterization of causal effect identification: a query is identifiable if and only if it can be transformed into a do-free expression via a valid causal derivation in the graph. However, specifying a full causal diagram demands detailed knowledge of all causal relationships among observed variables, an assumption that is often unrealistic in practice. To mitigate this challenge, many recent works have adopted *causal abstractions*: simplified or coarse-grained models that capture only a subset of causal dependencies.

All the causal abstractions we are aware of determine a collection $\mathcal{C}$ of causal diagrams on $\mathcal{V}$ that satisfy their specified constraints. We study identifiability of a causal query within such a class: broadly, a query is *identifiable* in $\mathcal{C}$ if the same identification functional can be derived for every graph $G \in \mathcal{C}$. In practice, however, this notion can be ambiguous. The literature has proposed multiple identification strategies under causal abstractions, giving rise to several distinct definitions of identifiability in this setting.

3.1 Identifiability Under Different Knowledge Conditions

We extend Pearl’s notion of identifiability from a single causal diagram to classes of diagrams. In the classical setting, identifiability is a uniqueness property over a fixed causal graph: a causal query is identifiable if all SCMs inducing the same graph and observational distribution agree on its value. When moving to classes of graphs, this requirement naturally lifts to uniform validity over all SCMs whose graph belongs to the class. Definition 5 formalizes this extension.

Definition 5 (Identifiability through Graphs (IG)). *Let $\mathcal{C}$ be a class of causal diagrams over $\mathcal{V}$, and let $\mathbf{X}, \mathbf{Y}, \mathbf{Z} \subseteq \mathcal{V}$ be pairwise disjoint with $\mathbf{Y} \neq \emptyset$. The causal query $P(\mathbf{y} \mid \text{do}(\mathbf{x}), \mathbf{z})$ is identifiable through graphs in $\mathcal{C}$ if there exists an identification functional valid for $P(\mathbf{y} \mid \text{do}(\mathbf{x}), \mathbf{z})$ in every SCM $\mathcal{M}$ satisfying $\mathcal{G}_{\mathcal{M}} \in \mathcal{C}$.*

Example 3. *Consider the graphs $\mathcal{G}_1 := Z \leftarrow X \rightarrow Y$ and $\mathcal{G}_2 := X \leftarrow Z \rightarrow Y$. Then $P(y \mid \text{do}(x))$ is identifiable through graphs in $\mathcal{C} := \{\mathcal{G}_1, \mathcal{G}_2\}$. Details are provided in the proof of Proposition 4.*

This notion is called graphical because identifiability is required to hold uniformly over all SCMs whose causal graph lies in $\mathcal{C}$, regardless of their observational distributions. For instance, Jaber et al. (2022) provide a complete characterization of identifiability through graphs in Partial Ancestral Graphs (PAGs), which represent Markov equivalence classes $\mathcal{C}$ of ADMGs and encode the ancestral relations that are invariant across the class.

We now refine this perspective by incorporating additional observational information. In particular, we consider the case where the true observational distribution $P^*(\mathcal{V})$ is known. This induces a further restriction on the admissible SCMs, yielding a smaller class defined by both graphical and distributional constraints.

Definition 6 (Identifiability through Graphs knowing P^* (IGP)). *Let $\mathcal{C}$ be a class of causal diagrams over $\mathcal{V}$, and let $\mathbf{X}, \mathbf{Y}, \mathbf{Z} \subseteq \mathcal{V}$ be pairwise disjoint with $\mathbf{Y} \neq \emptyset$. Let $P^*(\mathcal{V})$ denote the true data density. The causal query $P(\mathbf{y} \mid \text{do}(\mathbf{x}), \mathbf{z})$ is identifiable through graphs knowing P^* in $\mathcal{C}$ if there exists an identification functional valid for $P(\mathbf{y} \mid \text{do}(\mathbf{x}), \mathbf{z})$ in every SCM $\mathcal{M}$ satisfying $\mathcal{G}_{\mathcal{M}} \in \mathcal{C}$ and $P_{\mathcal{M}}(\mathcal{V}) = P^*(\mathcal{V})$.*

The distinction between IG and IGP lies in the information available to the analyst. IG assumes access only to graphical knowledge, and requires the query to be identifiable uniformly over all SCMs whose causal graph

belongs to a given class. IGP additionally assumes knowledge of the true observational distribution, thereby imposing both graphical and distributional constraints, but still allowing for a collection of SCMs. In this sense, IGP corresponds to a strictly more informed setting, in which identifiability is no longer based on graphical information alone.

Here, IGP should be viewed as a representative example of a broader family of extensions in which identifiability is defined under additional non-graphical information. Such side information may include structural assumptions, such as linearity, additive noise, or restrictions on the functional form of causal mechanisms, as well as distributional assumptions. While other forms of side information could be considered, we focus on IGP as an illustrative example to demonstrate how the framework accommodates such extensions.

From a computational perspective, both notions remain difficult to verify. IG would require validity over all graphs belonging to $\mathcal{C}$, while IGP depends on exact knowledge of $P^*(\mathcal{V})$, which is usually unavailable in practice. These limitations motivate our focus on graphical characterizations and sufficient conditions for identifiability in $\mathcal{C}$.

3.2 Identifiability Under Different Methodological Conditions

IG provides a natural extension of identifiability to collections of diagrams, without requiring structural assumptions or additional constraints on the observational distribution. A key distinction between notions of identifiability in graphs therefore arises in how identifiability is established over a collection of causal diagrams. As discussed in Section 2.2, methods developed for a single graph can often be extended to a collection by requiring that the same identification strategy applies uniformly across all compatible graphs. This leads to weaker, yet more tractable, notions of identifiability, trading some degree of generality for computational and conceptual simplicity.

One natural approach consists in applying do-calculus uniformly over the entire collection. More precisely, the objective is to derive a single causal derivation that holds simultaneously in every causal diagram of the class.

Definition 7 (Identifiability by Common Derivation (ICD)). *Let $\mathcal{C}$ be a class of causal diagrams over the variable set $\mathcal{V}$, and let $\mathbf{X}, \mathbf{Y}, \mathbf{Z} \subseteq \mathcal{V}$ be pairwise disjoint with $\mathbf{Y} \neq \emptyset$. The causal query $P(\mathbf{y} \mid \text{do}(\mathbf{x}), \mathbf{z})$ is identifiable by common derivation in $\mathcal{C}$ if there exists a causal derivation that holds in every $\mathcal{G} \in \mathcal{C}$ and transforms the query into an expression involving only observational distributions.*

Moreover, we restate the notion of an *atomically complete calculus*, introduced in Jaber et al. (2022), for a class $\mathcal{C}$. Such a calculus consists of three atomic rules ($\mathcal{A}1, \mathcal{A}2, \mathcal{A}3$), which are counterparts of the three do-calculus rules ($\mathcal{R}1, \mathcal{R}2, \mathcal{R}3$). Each atomic rule specifies, based only on the structure of the class $\mathcal{C}$, when the corresponding do-calculus rule application holds uniformly over $\mathcal{C}$. The rules are required to satisfy the following two properties, for each i :

1. **(Soundness)** If the atomic rule $\mathcal{A}i$ applies to $\mathcal{C}$, then the corresponding rule $\mathcal{R}i$ holds in every graph in $\mathcal{C}$.
2. **(Completeness)** If the atomic rule $\mathcal{A}i$ does *not* apply to $\mathcal{C}$, then there exists at least one graph $G \in \mathcal{C}$ in which the corresponding do-calculus rule $\mathcal{R}i$ does not hold.

Thus, the atomic rules characterize exactly which do-calculus transformations hold uniformly over the class: soundness guarantees that every applicable atomic rule holds in every graph in $\mathcal{C}$, while completeness guarantees that a transformation cannot hold uniformly over $\mathcal{C}$ when its corresponding atomic rule does not apply. Consequently, any derivation obtained within an atomically complete calculus immediately yields a causal derivation that holds uniformly over the entire class, thereby establishing ICD for the causal query in $\mathcal{C}$. An important example is given by Partial Ancestral Graphs (PAGs), which represent Markov equivalence classes $\mathcal{C}$ of ADMGs and encode the ancestral relations that are invariant across the class. For this setting, ICD was established in Jaber et al. (2022). Other examples of atomically complete calculi are given in Anand et al. (2023); Yvernes et al. (2025c); Ferreira & Assaad (2025).

The main distinction between identifiability through graphs and identifiability by common derivation lies in the permissible proof techniques. In the latter, one must provide a single causal derivation that holds in all

compatible graphs. By contrast, IG permits the use of different graphical identification procedures for different graphs in the class, provided that all resulting procedures ultimately yield the same valid identification functional.

Beyond do-calculus, graphical identification criteria developed for single graphs can also be extended uniformly to collections of graphs. This leads to notions such as *identifiability by common backdoor* (ICB), *identifiability by common frontdoor* (ICF) or *identifiability by common adjustment* (ICA) where one requires that the corresponding graphical criterion is satisfied in every graph of the class $\mathcal{C}$. More generally, this gives rise to a broad family of graphical identifiability notions based on uniform graphical conditions. For instance, Example 4 illustrates a class in which a causal query is identifiable by common backdoor adjustment.

Example 4 (ICB). *Let us consider the following class of causal diagrams $\mathcal{C}$ as follows:*

$$\mathcal{C} := \left\{ \begin{array}{c} V \xrightarrow{\quad} Y \\ \uparrow \quad \downarrow \\ U \xrightarrow{\quad} Z \rightarrow X \rightarrow Y, \quad Y \leftarrow X \leftarrow Z \\ \uparrow \quad \downarrow \\ U \xrightarrow{\quad} Z \end{array} \right\}$$

In both graphs of $\mathcal{C}$, $\{Z\}$ is a valid backdoor set. As a result, any causal query $P(y \mid \text{do}(x))$ is identified by the formula $P(y \mid \text{do}(x)) = \sum_z P(y \mid x, z) P(z)$.

As an example, identifiability by common backdoor has been studied in the context of Summary Causal Graphs (SCGs), a causal abstraction that summarizes causal relationships between time series by abstracting away the delays between causes and their effects, in Assaad et al. (2024). This was later extended into a complete criterion for identifiability by common adjustment by Yvernes et al. (2025b). In a different framework, Perković et al. (2018) established a sound and complete criterion for identifiability by common adjustment in Markov equivalence classes of ancestral graphs.

4 The Strict Hierarchy of Identifiability in Collections of Graphs

Section 3 introduced several notions of identifiability for collections of causal diagrams. These notions are not independent, however, but instead form a hierarchy: some forms of identifiability imply others in all cases. This hierarchy is strict, with the implication relations and counterexamples establishing its strictness summarized in Figure 1. We first establish the implication relations in Section 4.1, and then show that their converses fail in general and discuss the resulting cases of incomparability in Section 4.2.

4.1 The Hierarchy of Identifiability

We now establish the implications underlying the hierarchy. We first consider the relationship between IG and IGP. Since IGP corresponds to IG with additional knowledge, we know that IG always implies IGP, as formalized by Proposition 1.

Proposition 1 ($IG \Rightarrow IGP$). *Let $\mathcal{C}$ be a class of causal diagrams over the variable set $\mathcal{V}$, let P^* be a distribution over $\mathcal{V}$, and let $\mathbf{X}, \mathbf{Y}, \mathbf{Z} \subseteq \mathcal{V}$ be pairwise disjoint with $\mathbf{Y} \neq \emptyset$. If the causal query $P(\mathbf{y} \mid \text{do}(\mathbf{x}), \mathbf{z})$ is identifiable through graphs in $\mathcal{C}$, then it is identifiable through graphs knowing P^* in $\mathcal{C}$.*

Proof. Assume that IG holds for $P(\mathbf{y} \mid \text{do}(\mathbf{x}), \mathbf{z})$ over $\mathcal{C}$. Then there exists an identification functional ϕ such that for every SCM $\mathcal{M}$ satisfying $\mathcal{G}_{\mathcal{M}} \in \mathcal{C}$, $P_{\mathcal{M}}(\mathbf{y} \mid \text{do}(\mathbf{x}), \mathbf{z}) = \phi(P_{\mathcal{M}})$. Since $\{\mathcal{M} \mid \mathcal{G}_{\mathcal{M}} \in \mathcal{C}, P_{\mathcal{M}} = P^*\} \subseteq \{\mathcal{M} \mid \mathcal{G}_{\mathcal{M}} \in \mathcal{C}\}$, ϕ is valid for all SCM $\mathcal{M}$ such that $\mathcal{G}_{\mathcal{M}} \in \mathcal{C}$ and such that $P_{\mathcal{M}} = P^*$. Hence IGP holds. $\square$

Similarly, because ICD requires a single causal derivation to hold simultaneously across all causal diagrams in the class, ICD implies IG, as stated by Proposition 2.

Proposition 2 ($ICD \Rightarrow IG$). *Let $\mathcal{C}$ be a class of causal diagrams over the variable set $\mathcal{V}$, and let $\mathbf{X}, \mathbf{Y}, \mathbf{Z} \subseteq \mathcal{V}$ be pairwise disjoint with $\mathbf{Y} \neq \emptyset$. If the causal query $P(\mathbf{y} \mid \text{do}(\mathbf{x}), \mathbf{z})$ is identifiable by common derivation in $\mathcal{C}$, then it is identifiable through graphs in $\mathcal{C}$.*

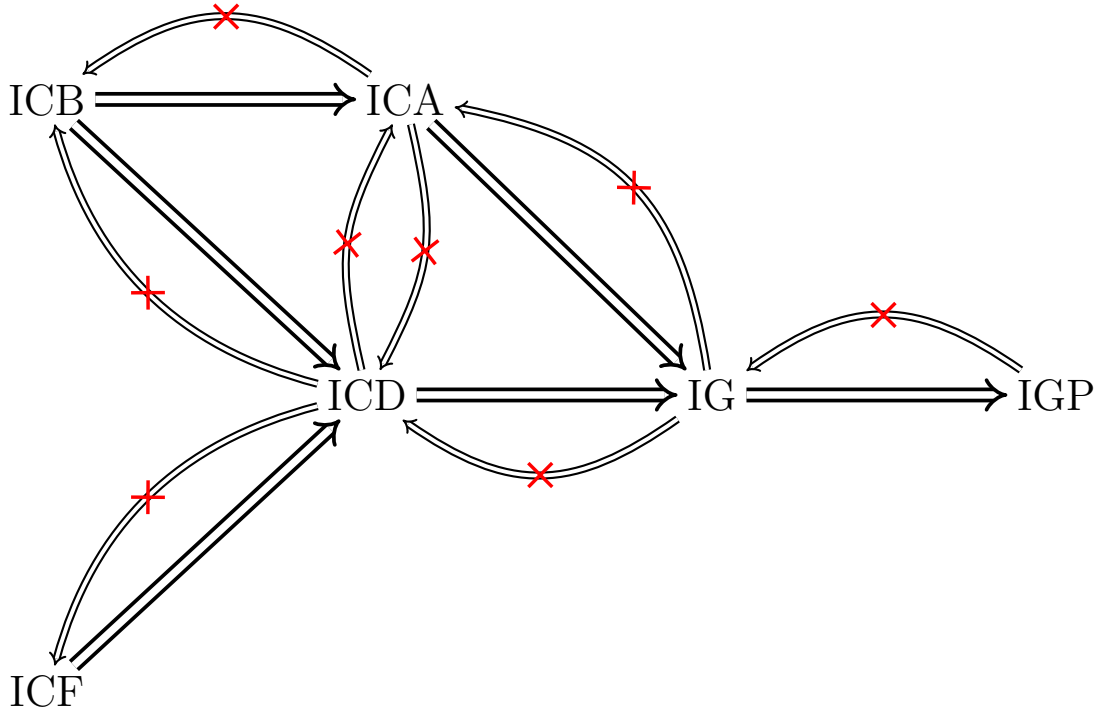


Figure 1: Hierarchy between the notions of identifiability. An arrow indicates that the implication holds in every case. A red cross on an arrow indicates that the implication fails in at least one case.

Proof. Assume that ICD holds for $P(\mathbf{y} \mid \text{do}(\mathbf{x}), \mathbf{z})$ over $\mathcal{C}$. By definition, there exists a single do-calculus derivation that holds for every graph in $\mathcal{C}$ and produces an identification formula $\mathcal{F}(P)$. Consequently, for every SCM $\mathcal{M}$ such that $\mathcal{G}_{\mathcal{M}} \in \mathcal{C}$, we have $P_{\mathcal{M}}(\mathbf{y} \mid \text{do}(\mathbf{x}), \mathbf{z}) = \mathcal{F}(P_{\mathcal{M}})$. Therefore, the identification functional $(P_{\mathcal{M}} \mapsto \mathcal{F}(P_{\mathcal{M}}))$ is valid for all $\mathcal{M}$ such that $\mathcal{G}_{\mathcal{M}} \in \mathcal{C}$. Hence IG holds. $\square$

Similarly, methods that consist of applying a graphical criterion uniformly across all graphs in the collection (i.e., ICB, ICA, and ICF) imply either IG or ICD. Specifically, ICB and ICF imply ICD, while ICA implies IG. Theorem 5 summarizes all implication relations between the identifiability notions introduced above.

Theorem 5 (Hierarchy of Identifiability Notions). *Let $\mathcal{C}$ be a class of causal diagrams over $\mathcal{V}$, and let $\mathbf{X}, \mathbf{Y}, \mathbf{Z} \subseteq \mathcal{V}$ be pairwise disjoint with $\mathbf{Y} \neq \emptyset$. The implications between identifiability notions summarized in Figure 1 are correct.*

Proof Sketch. The implications $IG \Rightarrow IGP$ and $ICD \Rightarrow IG$ follow directly from Propositions 1 and 2, respectively. The implications $ICB \Rightarrow ICD$ and $ICF \Rightarrow ICD$ follow from the fact that backdoor and frontdoor adjustment admit a common do-calculus derivation (Pearl, 2009), which can therefore be applied uniformly across all graphs in $\mathcal{C}$. Finally, $ICB \Rightarrow ICA$ follows from the fact that the backdoor criterion is a special case of Definition 4, while $ICA \Rightarrow IG$ follows from Theorem 4. Full details are given in Appendix A.1. $\square$

4.2 Strictness of the Hierarchy

The hierarchy presented in Theorem 5 is strict. More specifically, the following result holds.

Proposition 3 (Strictness of the Hierarchy). *None of the converse implications in Figure 1 holds in general, meaning that there are counterexamples to each of them. Moreover, ICA and ICD are incomparable, meaning that there are counterexamples in both directions.*

We establish Proposition 3 by considering the relevant separations individually. Most of them follow from existing counterexamples in the single-graph setting or from straightforward constructions. The separation between IG and ICD, however, requires a substantially more involved construction and constitutes the main technical contribution of this section. We first establish the simpler separations and then turn to this more challenging case. The formal proof, including all technical details, is provided in Appendix A.2.

4.2.1 Straightforward Separations

We first establish the separations that follow directly from known results in the single-graph setting or from simple constructions. In the single-graph setting, the back-door criterion, the front-door criterion, and generalized adjustment are not complete for causal effect identification, whereas do-calculus is complete (Pearl, 2009; Perković et al., 2015). These limitations immediately carry over to collections of causal diagrams, since the single-graph setting is recovered by considering collections consisting of a single diagram. Consequently, the corresponding converse implications in Figure 1 do not hold in general. The formal constructions are given in Appendix A.2; they can also be obtained directly from the standard counterexamples for the incompleteness of these criteria.

We next consider the converse implication $IGP \Rightarrow IG$. Since IGP incorporates additional information about the data-generating distribution, it can identify effects that are not identifiable from the graphical structure alone.

Example 5 (Failure of $IGP \Rightarrow IG$). *Consider the collection of causal diagrams $\mathcal{C} := \{X \rightarrow Y; X \leftarrow Y\}$. Suppose that the observed variables X and Y are independent. Then $P(y \mid \text{do}(x))$ is identifiable from $\mathcal{C}$ knowing the data density P^* , although it is not identifiable from $\mathcal{C}$ alone. Thus, IGP does not imply IG. All details are provided in Appendix A.2.1.*

The example is deliberately simple, as our definitions allow the observed distribution P^* to provide information beyond the conditional independences encoded by the graph. More elaborate instances of the same phenomenon can be obtained by imposing additional distributional constraints. In that direction, Tikka et al. (2019) show that context-specific independence relations can provide information beyond the conditional independences encoded by a causal graph, and can turn an otherwise non-identifiable causal effect into an identifiable one. Thus, this separation between graphical and distributional information already occurs in the single-graph setting. The example above illustrates the same underlying mechanism in its simplest form: information contained in P^* can rule out causal models that remain compatible with the graphical information alone.

4.2.2 ICA, ICD and IG

Historically, the relationship between ICD and IG has been an open question, particularly in the context of causal abstractions such as ancestral graphs PAGs. Zhang (2008) first extended do-calculus to ancestral graphs, providing a framework for identifiability by common derivation within such classes. However, their work left unresolved whether IG could be strictly more general than ICD. This gap was later addressed by Jaber et al. (2022), who proved that in the specific case of PAGs, ICD and IG are in fact equivalent. Their result established that, for PAGs, the existence of a common causal derivation is both necessary and sufficient for identifiability through graphs.

Yet, this equivalence does not hold in full generality. In this section, we demonstrate that IG and ICD are distinct notions. We begin by showing the failure of the implication from ICA to ICD, thereby illustrating that this graphical criterion does not guarantee a common derivation. To formalize this separation, we introduce derivation graphs from Yvernes et al. (2026), as a tool to characterize the structure of possible causal derivations.

Definition 8 (Derivation Graph (Yvernes et al., 2026)). *Let $\mathcal{G} = (\mathcal{V}_{\mathcal{G}}, \mathcal{E}_{\mathcal{G}})$ be an ADMG. Its derivation graph is defined as the undirected graph $\mathcal{D}[\mathcal{G}] = (\mathcal{V}_{\mathcal{D}}, \mathcal{E}_{\mathcal{D}})$, where:*

- *The set of vertices $\mathcal{V}_{\mathcal{D}}$ consists of all expressions of the form $P(\mathbf{y} \mid \text{do}(\mathbf{x}), \mathbf{w})$, with pairwise disjoint subsets $(\mathbf{Y}, \mathbf{X}, \mathbf{W}) \subseteq \mathcal{V}_{\mathcal{G}}$, with $\mathbf{Y} \neq \emptyset$.*

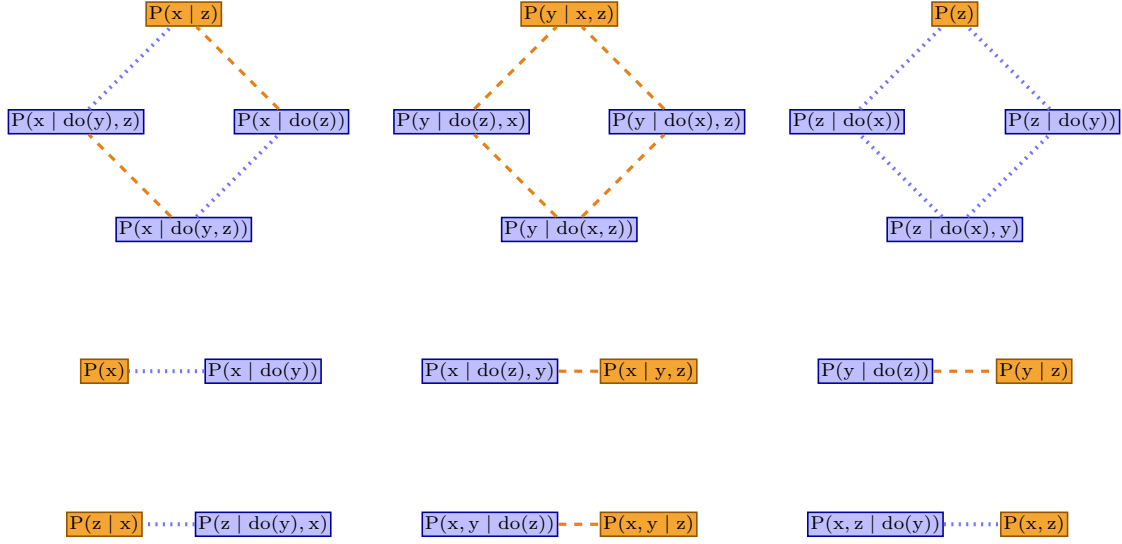


Figure 2: Derivation graph of the DAG from Example 2. Blue nodes denote interventional queries, while orange nodes denote do-free queries. Edges of the derivation graph represent valid applications of the do-calculus rules in the DAG: gray solid lines for Rule $\mathcal{R}1$, orange dashed lines for Rule $\mathcal{R}2$, and blue dotted lines for Rule $\mathcal{R}3$. Only vertices that are non-isolated are displayed.

- There is an edge $(Q_1, Q_2) \in \mathcal{E}_{\mathcal{D}}$ if Q_2 can be obtained from Q_1 by applying a single do-calculus rule that is valid in $\mathcal{G}$.

The derivation graph takes the causal diagram as its starting point and characterizes the do-calculus rules that hold in it. Figure 2 provides an example of a derivation graph, illustrating how the graphical structure determines which transformations are permitted.

We now reverse this perspective: rather than starting from a graph and deriving the rules that hold, we start from a given causal derivation and ask which graphical conditions are required for each of its individual steps to hold. This shift in perspective allows us to characterize the graphical conditions associated with each step of a derivation and leads to the notion of a *step graph*.

Definition 9 (Step Graph of a Derivation). *Let D be a causal derivation. Let $\mathcal{V}$ be the set of all variables represented in D . Its step graph is defined as the following undirected graph $\mathcal{S}[D] = (\mathcal{V}_{\mathcal{S}}, \mathcal{E}_{\mathcal{S}})$, where:*

- The set of vertices $\mathcal{V}_{\mathcal{D}}$ consists of all expressions of the form $P(\mathbf{y} \mid \text{do}(\mathbf{x}), \mathbf{w})$, with pairwise disjoint subsets $(\mathbf{Y}, \mathbf{X}, \mathbf{W}) \subseteq \mathcal{V}$, with $\mathbf{Y} \neq \emptyset$, that appear in D .
- There is an edge $(Q_1, Q_2) \in \mathcal{E}_{\mathcal{S}}$ if Q_1 is rewritten into Q_2 by applying a single do-calculus rule in D .

Example 6. Consider the derivation from Example 2. This derivation involves the variables X, Y , and Z , and uses $\mathcal{R}3$ to obtain $P(z \mid \text{do}(x)) = P(z)$ and $\mathcal{R}2$ to obtain $P(y \mid \text{do}(x), z) = P(y \mid x, z)$. As a result, the nodes of its step graph consist of all causal queries that can be written using X, Y, Z , and its edges are $P(z \mid \text{do}(x)) - P(z)$ and $P(y \mid \text{do}(x), z) - P(y \mid x, z)$.

With these definitions, derivation graphs can be used to characterize the set of valid derivations in any graph $G \in \mathcal{C}$. Conversely, step graphs can be used to characterize the set of graphs in which a given derivation is valid. This duality is formalized in the following theorem.

Theorem 6. Let $\mathcal{V}$ be a set of variables, $\mathcal{G}$ be an ADMG on $\mathcal{V}$, and D be a derivation on $\mathcal{V}$. The following holds:

$$D \text{ holds in } \mathcal{G} \quad \text{if and only if} \quad \mathcal{S}[D] \subseteq \mathcal{D}[\mathcal{G}].$$

Proof. By construction, $\mathcal{S}[D]$ contains all expressions and do-calculus rules required for the derivation D . If D holds in $\mathcal{G}$, then every rule in D is valid in $\mathcal{G}$, hence all edges of $\mathcal{S}[D]$ are present in $\mathcal{D}[\mathcal{G}]$, implying $\mathcal{S}[D] \subseteq \mathcal{D}[\mathcal{G}]$. Conversely, if $\mathcal{S}[D] \subseteq \mathcal{D}[\mathcal{G}]$, then all rules in D are valid in $\mathcal{G}$, hence D holds in $\mathcal{G}$. $\square$

Example 7. Let $\mathcal{G}$ be the DAG considered in Example 2, with derivation graph $\mathcal{D}[\mathcal{G}]$ shown in Figure 2. Let D be the derivation from the same example. D holds in $\mathcal{G}$, and its step graph $\mathcal{S}[D]$, constructed in Example 6, satisfies

$$\mathcal{S}[D] \subseteq \mathcal{D}[\mathcal{G}].$$

Conversely, consider a derivation D' that applies $\mathcal{R}3$ to obtain $P(x \mid \text{do}(z)) = P(x)$. This rule is not valid in $\mathcal{G}$. Consequently the edge $P(x \mid \text{do}(z)) - P(x)$ is present in the step graph but not in the derivation graph. Hence, D' does not hold in $\mathcal{G}$ and

$$\mathcal{S}[D'] \not\subseteq \mathcal{D}[\mathcal{G}].$$

Building on this theorem, we demonstrate the failure of the implication from ICA to ICD.

Proposition 4 (Failure of $\text{ICA} \Rightarrow \text{ICD}$). *There exist a causal query $P(\mathbf{y} \mid \text{do}(\mathbf{x}), \mathbf{z})$ and a class of causal graphs $\mathcal{C}$ such that $P(\mathbf{y} \mid \text{do}(\mathbf{x}), \mathbf{z})$ is identifiable by common adjustment in $\mathcal{C}$, but not identifiable by common derivation in $\mathcal{C}$.*

Proof Sketch. See Appendix A.3 for a detailed proof. We construct a counterexample in which ICA holds but ICD does not. Consider the graphs $\mathcal{G}_1 := Z \leftarrow X \rightarrow Y$ and $\mathcal{G}_2 := X \leftarrow Z \rightarrow Y$, and the class $\mathcal{C} := \{\mathcal{G}_1, \mathcal{G}_2\}$. We have the following properties:

- $\{Z\}$ satisfies the adjustment criterion in both $\mathcal{G}_1$ and $\mathcal{G}_2$, hence ICA holds in $\mathcal{C}$ for $P(y \mid \text{do}(x))$.
- Consider the graph $\mathcal{G}_3 := X \rightarrow Z \rightarrow Y$. One can verify that:

$$\mathcal{D}[\mathcal{G}_1] \cap \mathcal{D}[\mathcal{G}_2] \subseteq \mathcal{D}[\mathcal{G}_3].$$

Assume, by contradiction, that $P(y \mid \text{do}(x))$ is ICD in $\mathcal{C}$. Then, there exists a derivation D that identifies $P(y \mid \text{do}(x))$ in all $\mathcal{G} \in \mathcal{C}$. Hence:

$$\mathcal{S}[D] \subseteq \mathcal{D}[\mathcal{G}_1] \cap \mathcal{D}[\mathcal{G}_2] \subseteq \mathcal{D}[\mathcal{G}_3].$$

Therefore, by Theorem 6, D holds in $\mathcal{G}_3$, hence $P(y \mid \text{do}(x))$ is ICD, thus IG (by Proposition 2), in $\{\mathcal{G}_1, \mathcal{G}_2, \mathcal{G}_3\}$. However, since $\{Z\}$ does not satisfy the adjustment criterion in $\mathcal{G}_3$, by Theorem 4, there exists a density P^* Markovian with respect to $\mathcal{G}_3$ such that:

$$P^*(y \mid \text{do}(x)) \neq \sum_z P^*(y \mid z, x) P^*(z).$$

The construction of P^* yields an SCM $\mathcal{M}_3$ with $P_{\mathcal{M}_3} = P^*$ and $\mathcal{G}_{\mathcal{M}_3} = \mathcal{G}_3$. Moreover, P^* is also Markovian with respect to $\mathcal{G}_2$, hence by Proposition 7.1 from Peters et al. (2017), there exists an SCM $\mathcal{M}_2$ with $P_{\mathcal{M}_2} = P^*$ and $\mathcal{G}_{\mathcal{M}_2} = \mathcal{G}_2$. In $\mathcal{G}_2$, since $\{Z\}$ satisfies the adjustment criterion, we have:

$$P^*(y \mid \text{do}(x)) = \sum_z P^*(y \mid z, x) P^*(z).$$

This contradicts the previous inequality. Therefore, $P(y \mid \text{do}(x))$ cannot be ICD in $\mathcal{C}$.

Hence, $\text{ICA} \Rightarrow \text{ICD}$ fails. $\square$

The failure of $\text{ICA} \Rightarrow \text{ICD}$ also has direct implications for the relationship between IG and ICD. Since IG includes ICA as a special case, the non-implication of ICA to ICD extends directly to IG. This is formalized in Corollary 1.

Corollary 1 (Failure of $\text{IG} \Rightarrow \text{ICD}$). *There exists a class of causal diagrams $\mathcal{C}$ and a causal query $P(\mathbf{y} \mid \text{do}(\mathbf{x}), \mathbf{z})$ such that $P(\mathbf{y} \mid \text{do}(\mathbf{x}), \mathbf{z})$ is identifiable through graphs in $\mathcal{C}$, but not identifiable by common derivation in $\mathcal{C}$.*

Proof. Let $\mathcal{C}$ be the class constructed in Proposition 4. Since ICA holds in $\mathcal{C}$, there exists a single identification functional (e.g., the adjustment formula) valid for all SCMs with graphs in $\mathcal{C}$. Thus, IG holds by Definition 5. However, as shown in Proposition 4, ICD fails in $\mathcal{C}$. This establishes the corollary. $\square$

The results above have implications for the use of atomically complete calculi (e.g., Yvernes et al. (2025c)) as tools for identifying causal queries in causal abstractions. While an atomically complete calculus guarantees ICD (and thus IG, since ICD implies IG), our findings show that atomic completeness is stronger than necessary for IG. Indeed, IG may hold even when no atomically complete calculus exists, because atomic completeness requires derivations to be uniform across all graphs represented by a causal abstraction, whereas IG allows graph-specific derivations. This suggests that characterizing IG may require identification tools that are more flexible than atomically complete calculi.

5 Discussion

5.1 Interpretation of the Hierarchy

Beyond its formal implications, the hierarchy provides a useful lens for interpreting existing results on causal effect identifiability in abstractions. It makes it possible to distinguish precisely which notion is characterized by a given approach and to identify potential gaps between existing guarantees and stronger notions of identifiability. In particular, a claim of completeness should always be interpreted with respect to a specific level of the hierarchy: completeness for ICD, for instance, does not imply completeness for IG. The atomically complete calculus established in Yvernes et al. (2025c) solves ICD for the causal abstractions and causal queries considered in this work, while IG remains open. Thus, the hierarchy provides a framework for locating existing results and assessing which stronger notions remain uncharacterized.

From this perspective, IG can be viewed as a natural target when causal effects are to be identified from a purely graphical abstraction. Among the notions considered here, IG is the least restrictive, as it requires only graphical identifiability and does not impose a specific form of identification strategy. When the available information is purely graphical, establishing IG guarantees that no effect identifiable from this information is missed. Thus, IG provides a natural completeness target for purely graphical causal abstractions. This makes IG particularly useful for characterizing the limits of what can be inferred from the graphical abstraction alone.

Conversely, if a causal effect is not IG under the available graphical abstraction, then identifying it requires information beyond that encoded by the abstraction. Such additional information may take the form of a refinement of the graphical abstraction or of non-graphical assumptions or information about the underlying causal model. In this sense, failure of IG does not merely indicate that a particular identification method is insufficient; it indicates that the available graphical information alone is insufficient to identify the effect.

5.2 Maximal Elements Is All You Need

Identifiability notions defined over a collection of causal diagrams can become difficult to apply when the collection $\mathcal{C}$ is large. A useful reduction is that, for the notions considered in this paper, it is sufficient to analyze only the maximal elements of $\mathcal{C}$ under graph inclusion. The key reason for this reduction is monotonicity under edge deletion: if an independence property holds in a graph, then it also holds in any graph obtained from it by deleting edges.

We make the following assumption throughout this section.

Assumption 1 (Existence of Maximal Elements). *For every graph $\mathcal{G} \in \mathcal{C}$, there exists a maximal element $\mathcal{G}^{\max} \in \mathcal{C}$ such that $\mathcal{G}$ is contained in $\mathcal{G}^{\max}$.*

This assumption is automatic when $\mathcal{C}$ is finite: starting from any $\mathcal{G} \in \mathcal{C}$, repeatedly moving to a strict supergraph in $\mathcal{C}$ must eventually terminate at a maximal element containing $\mathcal{G}$. In particular, when $\mathcal{V}$ is finite, there are only finitely many causal diagrams over $\mathcal{V}$, and hence $\mathcal{C}$ is finite. Finiteness of $\mathcal{C}$, however, is not necessary. The assumption also holds in some infinite settings, such as causal abstractions for time series,

including Extended Summary Causal Graphs (ESCGs) Assaad et al. (2022) and Summary Causal Graphs (SCGs) Yvernes et al. (2025b).

Lemma 1 formalizes this monotonicity.

Lemma 1 (Monotonicity under Graph Inclusion). *Let $\mathcal{V}$ be the set of observable variables, and let $\mathcal{G}_1$ and $\mathcal{G}_2$ be two causal diagrams on $\mathcal{V}$ such that $\mathcal{G}_2$ is a subgraph of $\mathcal{G}_1$. Let $\mathbf{X}, \mathbf{Y}, \mathbf{Z} \subseteq \mathcal{V}$ be pairwise disjoint with $\mathbf{Y} \neq \emptyset$. Then:*

1. *Any identification functional valid in $\mathcal{G}_1$ for $P(\mathbf{y} \mid \text{do}(\mathbf{x}), \mathbf{z})$ is also valid in $\mathcal{G}_2$.*
2. *Let $P^*(\mathcal{V})$ be an observational distribution. Any identification functional valid for $P(\mathbf{y} \mid \text{do}(\mathbf{x}), \mathbf{z})$ in all SCMs $\mathcal{M}_1$ inducing $\mathcal{G}_1$ and satisfying $P_{\mathcal{M}_1}(\mathcal{V}) = P^*(\mathcal{V})$ is also valid in all SCMs $\mathcal{M}_2$ inducing $\mathcal{G}_2$ and satisfying $P_{\mathcal{M}_2}(\mathcal{V}) = P^*(\mathcal{V})$.*
3. *Any causal derivation that holds in $\mathcal{G}_1$ also holds in $\mathcal{G}_2$.*
4. *Any graphical criterion satisfied in $\mathcal{G}_1$ is also satisfied in $\mathcal{G}_2$.*

Proof sketch. All four statements follow from the fact that $\mathcal{G}_2$ is obtained from $\mathcal{G}_1$ by removing edges.

For items (1) and (2), any SCM inducing $\mathcal{G}_2$ can be extended to an SCM inducing $\mathcal{G}_1$ by adding the missing parent arguments to the structural equations while keeping the functional dependence unchanged on the original parents. This extension preserves both the observational and interventional distributions. In particular, it preserves any fixed observational distribution $P^*(\mathcal{V})$. Items (3) and (4) follow from the fact that every graphical independence that holds in $\mathcal{G}_1$ also holds in $\mathcal{G}_2$.

The full proof is given in Appendix B. □

Lemma 1 implies that the identifiability notions considered in this paper are determined by the maximal graphs in $\mathcal{C}$. This is formalized in the following theorem.

Theorem 7 (Reduction to Maximal Elements). *Let $\mathcal{C}$ be a class of causal diagrams over a set of variables $\mathcal{V}$ satisfying Assumption 1, and let $\mathbf{X}, \mathbf{Y}, \mathbf{Z} \subseteq \mathcal{V}$ be pairwise disjoint with $\mathbf{Y} \neq \emptyset$. Let $\mathcal{C}^{\max}$ denote the subcollection of maximal elements of $\mathcal{C}$ under graph inclusion. Consider any notion of identifiability discussed above (namely, IG, IGP, ICD, ICB, ICF, or ICA). Then the following statements are equivalent:*

1. *$P(\mathbf{y} \mid \text{do}(\mathbf{x}), \mathbf{z})$ is identifiable in $\mathcal{C}$.*
2. *$P(\mathbf{y} \mid \text{do}(\mathbf{x}), \mathbf{z})$ is identifiable in $\mathcal{C}^{\max}$.*

Proof. The implication (1) $\Rightarrow$ (2) is immediate since $\mathcal{C}^{\max} \subseteq \mathcal{C}$ by Assumption 1.

For (2) $\Rightarrow$ (1), let $\mathcal{G} \in \mathcal{C}$. By Assumption 1, there exists $\mathcal{G}^{\max} \in \mathcal{C}^{\max}$ such that $\mathcal{G}$ is a subgraph of $\mathcal{G}^{\max}$.

We distinguish the different notions of identifiability considered above:

- If identifiability in $\mathcal{C}^{\max}$ is established through a common identification functional, then this functional is valid in $\mathcal{G}^{\max}$. By Lemma 1, it is also valid in $\mathcal{G}$.
- If identifiability in $\mathcal{C}^{\max}$ is established through a common identification functional knowing P^* , then this functional is valid for all SCMs inducing $\mathcal{G}^{\max}$ and satisfying $P_{\mathcal{M}}(\mathcal{V}) = P^*(\mathcal{V})$. By Lemma 1, it is also valid for all SCMs inducing $\mathcal{G}$ and satisfying $P_{\mathcal{M}}(\mathcal{V}) = P^*(\mathcal{V})$.
- If identifiability in $\mathcal{C}^{\max}$ is established through a common causal derivation, then this derivation holds in $\mathcal{G}^{\max}$. Again by Lemma 1, it also holds in $\mathcal{G}$.
- If identifiability in $\mathcal{C}^{\max}$ is established through a common graphical criterion, then this criterion is satisfied in $\mathcal{G}^{\max}$. By Lemma 1, it is therefore also satisfied in $\mathcal{G}$.

Hence, the query is identifiable in $\mathcal{C}$. □

Theorem 7 shows that every notion of identifiability becomes easier to assess when $\mathcal{C}$ has few maximal elements. In particular, if $\mathcal{C}$ admits a greatest element under graph inclusion, then identifiability in $\mathcal{C}$ reduces to identifiability in a single graph.

Corollary 2. *If $\mathcal{C}$ contains a greatest element under graph inclusion, then IG is equivalent to ICD.*

Proof. Let $\mathcal{G}^*$ denote the greatest element of $\mathcal{C}$. Then $\mathcal{C}^{\max} = \{\mathcal{G}^*\}$. By Theorem 7, IG in $\mathcal{C}$ is equivalent to IG in $\mathcal{C}^{\max}$, and thus to IG in the single graph $\mathcal{G}^*$. On a single causal diagram, soundness and completeness of do-calculus imply that IG in $\mathcal{G}^*$ is equivalent to ICD in $\{\mathcal{G}^*\}$. Finally, applying again Theorem 7 yields that ICD in $\mathcal{G}^*$ is equivalent to ICD in $\mathcal{C}$. Hence IG and ICD are equivalent in $\mathcal{C}$. □

This situation arises, for example, in cluster DAGs (Anand et al., 2023), where the partition admissibility assumption ensures the existence of a unique maximal element under inclusion. When this assumption is relaxed, as in (Yvernes et al., 2025c), Corollary 2 no longer applies. In that setting, ICD has been established, whereas IG remains open.

More generally, causal abstractions that induce only a small number of maximal graphs appear favorable from the standpoint of tractable identifiability analysis.

6 Conclusion

In this work, we proposed a unified framework for studying causal identifiability in causal abstractions represented as collections of causal graphs. Within this setting, we formalized several notions of identifiability, namely IG, ICD, IGP, ICA, ICB, and ICF, providing a common language for comparing previously disparate definitions.

We established the hierarchy relating these notions (Figure 1) and characterized its strictness through explicit counterexamples. These results delineate the relative strength of the different notions and show that several natural implications fail in general.

To analyze these relationships, we used derivation graphs and step graphs as methodological tools for reasoning about causal identifiability across collections of graphs. These constructions provide a fine-grained representation of causal derivations that hold and help explain why existing approaches, including atomically complete calculi, do not necessarily capture all identifiable queries in causal abstractions.

From a broader perspective, our results reveal a fundamental separation between graph-based and derivation-based notions of identifiability. This separation raises both methodological and computational questions. In particular, developing efficient procedures for deciding IG and ICD, and characterizing their computational complexity, remain open directions. More generally, the hierarchy provides a basis for understanding and designing causal abstractions with controlled identifiability properties.

References

- Tara V. Anand, Adele H. Ribeiro, Jin Tian, and Elias Bareinboim. Causal Effect Identification in Cluster DAGs. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(10):12172–12179, June 2023. ISSN 2374-3468. doi: 10.1609/aaai.v37i10.26435. URL <https://ojs.aaai.org/index.php/AAAI/article/view/26435>.
- Charles K. Assaad, Emilie Devijver, and Eric Gaussier. Discovery of extended summary graphs in time series. In James Cussens and Kun Zhang (eds.), *Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence*, volume 180 of *Proceedings of Machine Learning Research*, pp. 96–106. PMLR, 01–05 Aug 2022. URL <https://proceedings.mlr.press/v180/assaad22a.html>.
- Charles K. Assaad, Emilie Devijver, Eric Gaussier, Gregor Gössler, and Anouar Meynaoui. Identifiability of total effects from abstractions of time series causal graphs. In *Proceedings of the Fortieth Conference on Uncertainty in Artificial Intelligence*. JMLR.org, 2024.
- Sander Beckers and Joseph Y. Halpern. Abstracting causal models. *CoRR*, abs/1812.03789, 2018. URL <http://arxiv.org/abs/1812.03789>.
- Benito van der Zander, Maciej Liskiewicz, and Johannes Textor. Constructing separators and adjustment sets in ancestral graphs. In Nevin L. Zhang and Jin Tian (eds.), *Proceedings of the 30th Conference on Uncertainty in Artificial Intelligence*, volume R12 of *Proceedings of Machine Learning Research*, pp. 136–145. PMLR, 23–27 Jul 2014. URL <https://proceedings.mlr.press/r12/zander14a.html>. Reissued by PMLR on 04 October 2026.
- Simon Ferreira and Charles K. Assaad. Identifying macro conditional independencies and macro total effects in summary causal graphs with latent confounding. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(25):26787–26795, Apr. 2025. doi: 10.1609/aaai.v39i25.34882. URL <https://ojs.aaai.org/index.php/AAAI/article/view/34882>.
- Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. Causal abstractions of neural networks. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 9574–9586. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/4f5c422f4d49a5a807eda27434231040-Paper.pdf.
- Atticus Geiger, Duligur Ibeling, Amir Zur, Maheep Chaudhary, Sonakshi Chauhan, Jing Huang, Aryaman Arora, Zhengxuan Wu, Noah Goodman, Christopher Potts, and Thomas Icard. Causal abstraction: A theoretical foundation for mechanistic interpretability, 2025. URL <https://arxiv.org/abs/2301.04709>.
- Anna Guo, David Benkeser, and Razieh Nabi. Causal inference with the "napkin graph", 2025. URL <https://arxiv.org/abs/2512.19861>.
- Amin Jaber, Adele Ribeiro, Jiji Zhang, and Elias Bareinboim. Causal identification under markov equivalence: Calculus, algorithm, and completeness. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 3679–3690. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/17a9ab4190289f0e1504bbb98d1d111a-Paper-Conference.pdf.
- Maxime Méloux, Silviu Maniu, François Portet, and Maxime Peyrard. Everything, everywhere, all at once: Is mechanistic interpretability identifiable? In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 32900–32923, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/51157713f1f1f572a05647e1aef1b1ed-Paper-Conference.pdf.
- Pekka Parviainen and Samuel Kaski. Learning structures of bayesian networks for variable groups. *International Journal of Approximate Reasoning*, 88:110–127, 2017. ISSN 0888-613X. doi: <https://doi.org/10.1016/j.ijar.2017.05.006>.

-
- Judea Pearl. *Causality*. Cambridge University Press, Cambridge, 2 edition, 2009. ISBN 978-0-521-89560-6. doi: 10.1017/CBO9780511803161. URL <https://www.cambridge.org/core/books/causality/B0046844FAE10CBF274D4ACBDAEB5F5B>.
- Judea Pearl and Dana Mackenzie. *The Book of Why: The New Science of Cause and Effect*. Basic Books, New York, 2018. ISBN 978-0-465-09760-9.
- Emilija Perković, Johannes Textor, Markus Kalisch, and Marloes H. Maathuis. A complete generalized adjustment criterion. In *Proceedings of the Thirty-First Conference on Uncertainty in Artificial Intelligence*, UAI’15, pp. 682–691, Arlington, Virginia, USA, 2015. AUAI Press. ISBN 9780996643108.
- Emilija Perković, Johannes Textor, Markus Kalisch, and Marloes H. Maathuis. Complete graphical characterization and construction of adjustment sets in markov equivalence classes of ancestral graphs. *Journal of Machine Learning Research*, 18(220):1–62, 2018. URL <http://jmlr.org/papers/v18/16-319.html>.
- Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of Causal Inference - Foundations and Learning Algorithms*. Adaptive Computation and Machine Learning Series. The MIT Press, Cambridge, MA, USA, 2017.
- Eigil F. Rischel and Sebastian Weichwald. Compositional abstraction error and a category of causal models. In Cassio de Campos and Marloes H. Maathuis (eds.), *Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence*, volume 161 of *Proceedings of Machine Learning Research*, pp. 1013–1023. PMLR, 27–30 Jul 2021. URL <https://proceedings.mlr.press/v161/rischel21a.html>.
- Ilya Shpitser and Judea Pearl. Identification of joint interventional distributions in recursive semi-Markovian causal models. In *proceedings of the 21st national conference on Artificial intelligence - Volume 2*, pp. 1219–1226, Boston, Massachusetts, July 2006. AAAI Press. ISBN 978-1-57735-281-5.
- Santtu Tikka, Antti Hyttinen, and Juha Karvanen. Identifying causal effects via context-specific independence relations. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d’Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/d88518acbcc3d08d1f18da62f9bb26ec-Paper.pdf.
- Jonas Wahl, Urmi Ninad, and Jakob Runge. Foundations of causal discovery on groups of variables. *Journal of Causal Inference*, 12(1):20230041, 2024. doi: doi:10.1515/jci-2023-0041. URL <https://doi.org/10.1515/jci-2023-0041>.
- Eli N. Weinstein and David M. Blei. Hierarchical causal models. *Journal of Machine Learning Research*, 27(37):1–73, 2026. URL <http://jmlr.org/papers/v27/25-0899.html>.
- Clément Yvernes, Emilie Devijver, Marianne Clausel, and Eric Gaussier. Identifiability in causal abstractions: A hierarchy of criteria. CAR Workshop at the 41st Conference on Uncertainty in Artificial Intelligence (UAI), 2025a.
- Clément Yvernes, Emilie Devijver, and Eric Gaussier. Complete characterization for adjustment in summary causal graphs of time series. In Silvia Chiappa and Sara Magliacane (eds.), *Proceedings of the Forty-first Conference on Uncertainty in Artificial Intelligence*, volume 286 of *Proceedings of Machine Learning Research*, pp. 4844–4871. PMLR, 21–25 Jul 2025b. URL <https://proceedings.mlr.press/v286/yvernes25a.html>.
- Clément Yvernes, Emilie Devijver, Adèle H. Ribeiro, Marianne Clausel, and Eric Gaussier. Relaxing partition admissibility in cluster-dags: a causal calculus with arbitrary variable clustering. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, pp. 72492–72525. Curran Associates, Inc., 2025c. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/69078488008b859a347ed36bc6531453-Abstract-Conference.html.

Clément Yvernes, Emilie Devijver, Marianne Clausel, and Eric Gaussier. Unveiling the structure of do-calculus reasoning via derivation graphs. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*, pp. to appear, Seoul, South Korea, 2026. PMLR.

Fabio Massimo Zennaro, Paolo Turrini, and Theodoros Damoulas. Towards computing an optimal abstraction for structural causal models, 2022. URL <https://arxiv.org/abs/2208.00894>.

Fabio Massimo Zennaro, Paolo Turrini, and Theodoros Damoulas. Quantifying consistency and information loss for causal abstraction learning. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI '23*, 2023. ISBN 978-1-956792-03-4. doi: 10.24963/ijcai.2023/638. URL <https://doi.org/10.24963/ijcai.2023/638>.

Fabio Massimo Zennaro, Nicholas Bishop, Joel Dyer, Yorgos Felekis, Anisoara Calinescu, Michael Wooldridge, and Theodoros Damoulas. Causally abstracted multi-armed bandits. In *Proceedings of the Fortieth Conference on Uncertainty in Artificial Intelligence, UAI '24*. JMLR.org, 2024.

Jiji Zhang. Causal reasoning with ancestral graphs. *Journal of Machine Learning Research*, 9(47):1437–1474, 2008. URL <http://jmlr.org/papers/v9/zhang08a.html>.

A Proofs of Section 4

A.1 Proofs of Section 4.1

Theorem 5 (Hierarchy of Identifiability Notions). *Let $\mathcal{C}$ be a class of causal diagrams over $\mathcal{V}$, and let $\mathbf{X}, \mathbf{Y}, \mathbf{Z} \subseteq \mathcal{V}$ be pairwise disjoint with $\mathbf{Y} \neq \emptyset$. The implications between identifiability notions summarized in Figure 1 are correct.*

Proof. We prove each implication separately.

1. **ICB $\Rightarrow$ ICA.** Assume that ICB holds for $P(\mathbf{y} \mid \text{do}(\mathbf{x}))$ over $\mathcal{C}$. Then, there exists a set $\mathbf{Z}$ that satisfies the backdoor criterion relative to $\mathbf{X}$ and $\mathbf{Y}$ in all $\mathcal{G} \in \mathcal{C}$. Since the backdoor criterion is a special case of the adjustment criterion, $\mathbf{Z}$ also satisfies the adjustment criterion relative to $\mathbf{X}$ and $\mathbf{Y}$, in all $\mathcal{G} \in \mathcal{C}$. Therefore, ICA holds.
2. **ICB $\Rightarrow$ ICD.** Assume that ICB holds for $P(\mathbf{y} \mid \text{do}(\mathbf{x}))$ over $\mathcal{C}$, and let $\mathbf{Z}$ be the corresponding common backdoor adjustment set. For every $\mathcal{G} \in \mathcal{C}$, the backdoor criterion yields:

$$\begin{aligned}
 P(\mathbf{y} \mid \text{do}(\mathbf{x})) &= \sum_{\mathbf{z}} P(\mathbf{y}, \mathbf{z} \mid \text{do}(\mathbf{x})) && \text{(by marginalisation)} \\
 &= \sum_{\mathbf{z}} P(\mathbf{y} \mid \text{do}(\mathbf{x}), \mathbf{z}) P(\mathbf{z} \mid \text{do}(\mathbf{x})) && \text{(by chain rule)} \\
 &= \sum_{\mathbf{z}} P(\mathbf{y} \mid \text{do}(\mathbf{x}), \mathbf{z}) P(\mathbf{z}) && \text{(by } \mathcal{R}3) \\
 &= \sum_{\mathbf{z}} P(\mathbf{y} \mid \mathbf{x}, \mathbf{z}) P(\mathbf{z}) && \text{(by } \mathcal{R}2)
 \end{aligned}$$

where:

- the equality $P(\mathbf{z} \mid \text{do}(\mathbf{x})) = P(\mathbf{z})$ follows from Rule 3 of do-calculus since $\mathbf{Z}$ contains no descendant of $\mathbf{X}$;
- the equality $P(\mathbf{y} \mid \text{do}(\mathbf{x}), \mathbf{z}) = P(\mathbf{y} \mid \mathbf{x}, \mathbf{z})$ follows from Rule 2 of do-calculus since $\mathbf{Z}$ blocks all backdoor paths from $\mathbf{X}$ to $\mathbf{Y}$.

Since this derivation holds for every graph in $\mathcal{C}$, it is a common derivation. Hence ICD holds.

3. **ICF $\Rightarrow$ ICD.** Assume that ICF holds for $P(\mathbf{Y} \mid \text{do}(\mathbf{X}))$ over $\mathcal{C}$, and let $\mathbf{Z}$ be the corresponding common frontdoor set. For every $\mathcal{G} \in \mathcal{C}$, the frontdoor criterion yields:

$$\begin{aligned}
P(\mathbf{y} \mid \text{do}(x)) &= \sum_{\mathbf{z}} P(\mathbf{y}, \mathbf{z} \mid \text{do}(x)) && \text{(by marginalisation)} \\
&= \sum_{\mathbf{z}} P(\mathbf{z} \mid \text{do}(x)) P(\mathbf{y} \mid \mathbf{z}, \text{do}(x)) && \text{(by chain rule)} \\
&= \sum_{\mathbf{z}} P(\mathbf{z} \mid x) P(\mathbf{y} \mid \mathbf{z}, \text{do}(x)) && \text{(by } \mathcal{R}2) \\
&= \sum_{\mathbf{z}} P(\mathbf{z} \mid x) P(\mathbf{y} \mid \text{do}(\mathbf{z}), \text{do}(x)) && \text{(by } \mathcal{R}2) \\
&= \sum_{\mathbf{z}} P(\mathbf{z} \mid x) P(\mathbf{y} \mid \text{do}(\mathbf{z})) && \text{(by } \mathcal{R}3) \\
&= \sum_{\mathbf{z}} P(\mathbf{z} \mid x) \sum_{x'} P(\mathbf{y} \mid x', \mathbf{z}) P(x'), && \text{(by backdoor derivation)}
\end{aligned}$$

where:

- the equality $P(\mathbf{z} \mid \text{do}(x)) = P(\mathbf{z} \mid x)$ follows from Rule 2 of do-calculus since there is no unblocked path from X to Z , that contains an arrow into X ;
- the equality $P(\mathbf{y} \mid \mathbf{z}, \text{do}(x)) = P(\mathbf{y} \mid \text{do}(\mathbf{z}), \text{do}(x))$ follows from $\mathcal{R}2$ since all backdoor paths from $\mathbf{Z}$ to Y are blocked by X ;
- the equality $P(\mathbf{y} \mid \text{do}(\mathbf{z}), \text{do}(x)) = P(\mathbf{y} \mid \text{do}(\mathbf{z}))$ follows from $\mathcal{R}3$ since $\mathbf{Z}$ intercepts all directed paths from X to Y ;
- the equality $P(\mathbf{y} \mid \text{do}(\mathbf{z})) = \sum_{x'} P(\mathbf{y} \mid x', \mathbf{z}) P(x')$ follows from the backdoor criterion since all backdoor paths from $\mathbf{Z}$ to Y are blocked by X and since there is no collider-free backdoor path from X to Z , we know that X cannot be a descendant of $\mathbf{Z}$. The case (ICB $\Rightarrow$ ICD) gives us the corresponding derivation.

Since this derivation holds for every graph in $\mathcal{C}$, it is a common derivation. Hence ICD holds.

4. **ICA $\Rightarrow$ IG.** Assume that ICA holds for $P(\mathbf{y} \mid \text{do}(\mathbf{x}))$ over $\mathcal{C}$, and let $\mathbf{Z}$ be the corresponding common adjustment set.

By Theorem 4, for every graph $\mathcal{G} \in \mathcal{C}$, and for all probability P compatible with G , we have:

$$P(\mathbf{y} \mid \text{do}(\mathbf{x})) = \begin{cases} P(\mathbf{y} \mid \mathbf{x}) & \text{if } \mathbf{Z} = \emptyset, \\ \sum_{\mathbf{z}} P(\mathbf{y} \mid \mathbf{x}, \mathbf{z}) P(\mathbf{z}) & \text{otherwise.} \end{cases}$$

Therefore, the identification functional $P_{\mathcal{M}} \mapsto \begin{cases} P_{\mathcal{M}}(\mathbf{y} \mid \mathbf{x}) & \text{if } \mathbf{Z} = \emptyset, \\ \sum_{\mathbf{z}} P_{\mathcal{M}}(\mathbf{y} \mid \mathbf{x}, \mathbf{z}) P_{\mathcal{M}}(\mathbf{z}) & \text{otherwise.} \end{cases}$ is valid for all $\mathcal{M}$ such that $\mathcal{G}_{\mathcal{M}} \in \mathcal{C}$. Hence IG holds.

5. **ICD $\Rightarrow$ IG.** This implication is shown by Proposition 2.

6. **IG $\Rightarrow$ IGP.** This implication is shown by Proposition 1.

□

A.2 Proofs of Section 4.2

While Theorem 5 establishes a hierarchy of implications between these identifiability notions, Proposition 3 demonstrates that these implications are strict and that some notions are incomparable:

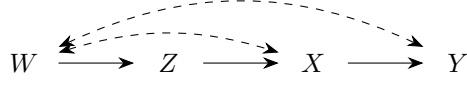


Figure 3: The Napkin Graph (Pearl & Mackenzie, 2018; Guo et al., 2025)

Proposition 3 (Strictness of the Hierarchy). *None of the converse implications in Figure 1 holds in general, meaning that there are counterexamples to each of them. Moreover, ICA and ICD are incomparable, meaning that there are counterexamples in both directions.*

Proof. We consider the converses of the implications presented in Theorem 5:

- **The failure of $ICA \Rightarrow ICB$** is given by the following counter example: Consider the graph $\mathcal{G} := Z \leftarrow X \rightarrow Y$. $\{Z\}$ satisfies the adjustment criterion relative to X and Y in G but does not satisfy the backdoor criterion since Z is a descendant of X . Hence, ICA holds for $P(y \mid \text{do}(x))$ in $\{\mathcal{G}\}$ but ICB does not hold for $P(y \mid \text{do}(x))$ in $\{\mathcal{G}\}$.
- **Failures of $ICD \Rightarrow ICA$, $IG \Rightarrow ICA$, $ICD \Rightarrow ICB$ and $ICD \Rightarrow ICF$ from the single graph setting.** We consider the collection $\mathcal{C}$ made of a single graph, the Napkin Graph presented in Figure 3 and we consider the query $P(y \mid \text{do}(x))$. We have the following properties:
 - As shown in (Pearl & Mackenzie, 2018; Guo et al., 2025), $P(y \mid \text{do}(x))$ is identifiable in the Napkin Graph. For completeness, identifiability is given by the following derivation:

$$\begin{aligned}
 P(y \mid \text{do}(x)) &= P(y \mid \text{do}(x), \text{do}(z)) && \text{(by } \mathcal{R}3) \\
 &= P(y \mid x, \text{do}(z)) && \text{(by } \mathcal{R}2) \\
 &= \frac{P(y, x \mid \text{do}(z))}{P(x \mid \text{do}(z))} && \text{(conditional probability)} \\
 &= \frac{\sum_w P(y, x \mid z, w)P(w)}{\sum_w P(x \mid z, w)P(w)}. && \text{(by backdoor derivation)}
 \end{aligned}$$

Hence, ICD and IG hold in $\mathcal{C}$ for $P(y \mid \text{do}(x))$.

- We will show that ICB , ICA and ICF do not hold in $\mathcal{C}$. We consider the cases separately:
 - * **ICB and ICA.** All candidate adjustment sets are $\{\emptyset, \{W\}, \{Z\}, \{W, Z\}\}$. However, none of them is an adjustment set. Indeed, conditioning on W or Z opens the backdoor path $X \leftrightarrow W \leftrightarrow Y$ and not conditioning on them lets the path $X \leftarrow Z \leftarrow W \leftrightarrow Y$ open. Hence, no common adjustment set exists and ICA does not hold. Since ICB is a special case of ICA , ICB does not hold either.
 - * **ICF.** All candidate frontdoor sets are $\{\emptyset, \{W\}, \{Z\}, \{W, Z\}\}$. However, none of them is an adjustment set since it is impossible to intercept $X \rightarrow Y$.

Hence, in this case, ICB , ICA , and ICF all fail to hold.

- **Incomparability of ICA and ICD :** The failure of $ICD \Rightarrow ICA$ has been established in the previous item and Proposition 4 establishes the failure of $ICA \Rightarrow ICD$.
- **The failure of $ICA \Rightarrow ICD$** is established by Proposition 4
- **The failure of $IG \Rightarrow ICD$** is established by Corollary 1
- **The failure of $IGP \Rightarrow IG$** is established by Proposition 5.

□

A.2.1 Proofs of Section 4.2.1

Example 5 (Failure of $IGP \Rightarrow IG$). Consider the collection of causal diagrams $\mathcal{C} := \{X \rightarrow Y; X \leftarrow Y\}$. Suppose that the observed variables X and Y are independent. Then $P(y \mid \text{do}(x))$ is identifiable from $\mathcal{C}$ knowing the data density P^* , although it is not identifiable from $\mathcal{C}$ alone. Thus, IGP does not imply IG . All details are provided in Appendix A.2.1.

Detailed Proof. We prove the failure of $IGP \Rightarrow IG$, by considering the collection of causal diagrams $\mathcal{C} := \{X \rightarrow Y, X \leftarrow Y\}$ and the causal query $P(Y = 1 \mid \text{do}(X = 1))$. We have the following properties:

- $P(Y = 1 \mid \text{do}(X = 1))$ is not identifiable through graphs in $\mathcal{C}$. Indeed, let us consider the SCM $\mathcal{M}_1$ defined by $X \sim \mathcal{B}(\frac{1}{2})$ and $Y = X \oplus U$ where $U \sim \mathcal{B}(\frac{1}{10})$ and $\oplus$ denotes the XOR operator. Moreover, let us consider the SCM $\mathcal{M}_2$ defined by $Y \sim \mathcal{B}(\frac{1}{2})$ and $X = Y \oplus U$ where $U \sim \mathcal{B}(\frac{1}{10})$. We have $\mathcal{G}_{\mathcal{M}_1} = X \rightarrow Y$ and $\mathcal{G}_{\mathcal{M}_2} = X \leftarrow Y$. Then, one can verify that:

$$P_{\mathcal{M}_1}(x, y) = \frac{9}{20} \cdot \mathbf{1}\{y = x\} + \frac{1}{20} \cdot \mathbf{1}\{y \neq x\} = P_{\mathcal{M}_2}(x, y)$$

However, $P_{\mathcal{M}_1}(Y = 1 \mid \text{do}(X = 1)) = \frac{9}{10}$ and $P_{\mathcal{M}_2}(Y = 1 \mid \text{do}(X = 1)) = \frac{1}{2}$. Therefore, $P(Y = 1 \mid \text{do}(X = 1))$ is not identifiable through graphs in $\mathcal{C}$.

- However, $P(Y = 1 \mid \text{do}(X = 1))$ may be identifiable through graphs knowing $P^*(X, Y)$ in $\mathcal{C}$. Indeed, let us consider the case where X and Y are independent in the true data distribution $P^*(X, Y)$. Let $\mathcal{M}_1$ be an SCM inducing $X \rightarrow Y$ such that $P_{\mathcal{M}_1}(X, Y) = P^*(X, Y)$. We have $P_{\mathcal{M}_1}(Y = 1 \mid \text{do}(X = 1)) = P_{\mathcal{M}_1}(Y = 1 \mid X = 1) = P^*(Y = 1 \mid X = 1) = P^*(Y = 1)$. Similarly, let $\mathcal{M}_2$ be an SCM inducing $X \leftarrow Y$ such that $P_{\mathcal{M}_2}(X, Y) = P^*(X, Y)$. We have $P_{\mathcal{M}_2}(Y = 1 \mid \text{do}(X = 1)) = P_{\mathcal{M}_2}(Y = 1) = P^*(Y = 1)$. Therefore the identification functional $\mathcal{F} := P(X, Y) \mapsto P(Y = 1)$ is valid in all SCM $\mathcal{M}$ satisfying $\mathcal{G}_{\mathcal{M}} \in \mathcal{C}$, $P_{\mathcal{M}}(X, Y) = P^*(X, Y)$, and $P(Y = 1 \mid \text{do}(X = 1))$ is identifiable through graphs knowing $P^*(X, Y)$ in $\mathcal{C}$.

Hence, in this case, IGP holds but IG does not hold. $\square$

A.3 Proofs of Section 4.2.2

Proposition 4 (Failure of $ICA \Rightarrow ICD$). There exist a causal query $P(\mathbf{y} \mid \text{do}(\mathbf{x}), \mathbf{z})$ and a class of causal graphs $\mathcal{C}$ such that $P(\mathbf{y} \mid \text{do}(\mathbf{x}), \mathbf{z})$ is identifiable by common adjustment in $\mathcal{C}$, but not identifiable by common derivation in $\mathcal{C}$.

Proof. Consider the graph $\mathcal{G}_1 := Z \leftarrow X \rightarrow Y$ and $\mathcal{G}_2 := X \leftarrow Z \rightarrow Y$, and the class $\mathcal{C} := \{\mathcal{G}_1, \mathcal{G}_2\}$. We have the following properties:

- $\{Z\}$ satisfies the adjustment criterion in both $\mathcal{G}_1$ and $\mathcal{G}_2$, hence ICA holds in $\mathcal{C}$ for $P(y \mid \text{do}(x))$.
- Consider the graph $\mathcal{G}_3 := X \rightarrow Z \rightarrow Y$. One can verify that:

$$\mathcal{D}[\mathcal{G}_1] \cap \mathcal{D}[\mathcal{G}_2] \subseteq \mathcal{D}[\mathcal{G}_3].$$

This inclusion was established by exhaustively enumerating all edges in $\mathcal{D}[\mathcal{G}_1]$ and $\mathcal{D}[\mathcal{G}_2]$, using the code of Yvernes et al. (2026). This enumeration produces the complete edge sets shown in Figures 4 and 5, representing $\mathcal{D}[\mathcal{G}_1] \cap \mathcal{D}[\mathcal{G}_2]$ and $\mathcal{D}[\mathcal{G}_3]$, respectively. Thus, the claimed inclusion is both computationally verified and directly checkable from the complete graphs displayed in the figures.

Assume, by contradiction, that $P(y \mid \text{do}(x))$ is ICD in $\mathcal{C}$. Then, there exists a derivation D that identifies $P(y \mid \text{do}(x))$ in all $\mathcal{G} \in \mathcal{C}$. Hence:

$$\mathcal{S}[D] \subseteq \mathcal{D}[\mathcal{G}_1] \cap \mathcal{D}[\mathcal{G}_2] \subseteq \mathcal{D}[\mathcal{G}_3].$$

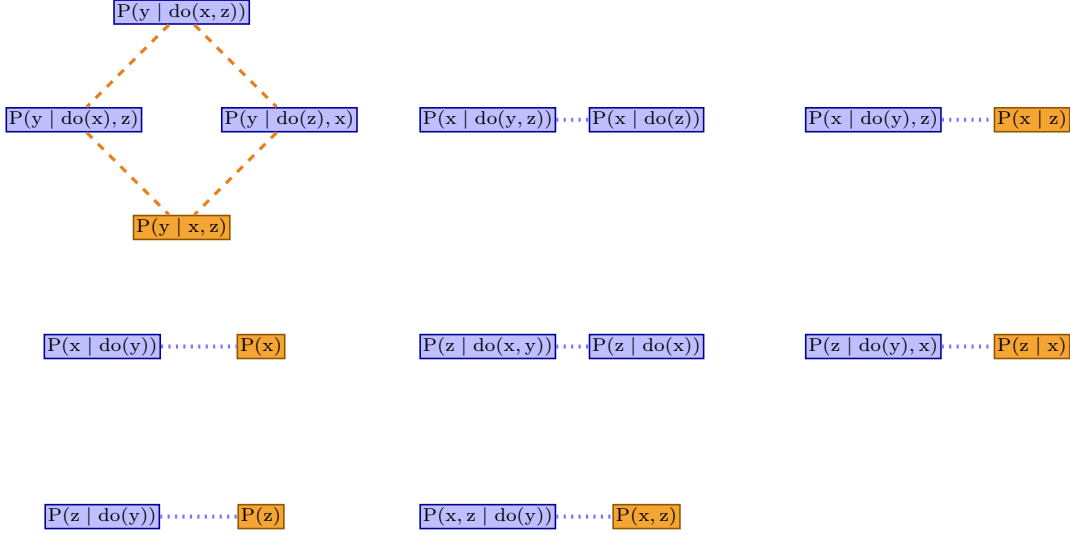


Figure 4: Intersection of the derivations graphs of $\mathcal{G}_1$ and $\mathcal{G}_2$. Blue nodes denote interventional queries, while orange nodes denote do-free queries. Edges represent applications of the do-calculus rules: gray solid lines for Rule $\mathcal{R}1$, orange dashed lines for Rule $\mathcal{R}2$, and blue dotted lines for Rule $\mathcal{R}3$. Only non-isolated queries are displayed.

Therefore, by Theorem 6, D holds in $\mathcal{G}_3$, hence $P(y | \text{do}(x))$ is ICD, thus IG (by Proposition 2), in $\{\mathcal{G}_1, \mathcal{G}_2, \mathcal{G}_3\}$.

However, since $\{Z\}$ does not satisfy the adjustment criterion in $\mathcal{G}_3$, by Theorem 4, there exists a density P^* Markovian with respect to $\mathcal{G}_3$ such that:

$$P^*(y | \text{do}(x)) \neq \sum_z P^*(y | z, x)P^*(z).$$

The construction of P^* , given in the proof of Theorem 57 in Perković et al. (2018), yields an SCM $\mathcal{M}_3$ with $P_{\mathcal{M}_3} = P^*$ and $\mathcal{G}_{\mathcal{M}_3} = \mathcal{G}_3$. More precisely, $P_{\mathcal{M}_3}$ is a linear gaussian SCM, hence P^* as a density with respect to Lebesgue measure. Although P^* corresponds to a continuous distribution, we use the discrete notation adopted throughout the paper for simplicity. The arguments remain unchanged when expressed in terms of densities and integrals instead of probability masses and sums.

Moreover, since P^* is Markovian with respect to $\mathcal{G}_3$, thus by Bayes' rule we have:

$$P^*(x, y, z) = P^*(y | z)P^*(z | x)P^*(x) = P^*(y | z)P^*(x | z)P^*(z)$$

Hence P^* is also Markovian with respect to $\mathcal{G}_2$, hence by Proposition 7.1 from Peters et al. (2017), there exists an SCM $\mathcal{M}_2$ with $P_{\mathcal{M}_2} = P^*$ and $\mathcal{G}_{\mathcal{M}_2} = \mathcal{G}_2$.² In $\mathcal{G}_2$, since $\{Z\}$ satisfies the adjustment criterion, we have:

$$P^*(y | \text{do}(x)) = \sum_z P^*(y | z, x)P^*(z).$$

This contradicts the previous inequality. Therefore, $P(y | \text{do}(x))$ cannot be ICD in $\mathcal{C}$.

Hence, $ICA \Rightarrow ICD$ fails. □

²The fact that P^* as a density with respect to Lebesgue measure is needed here.

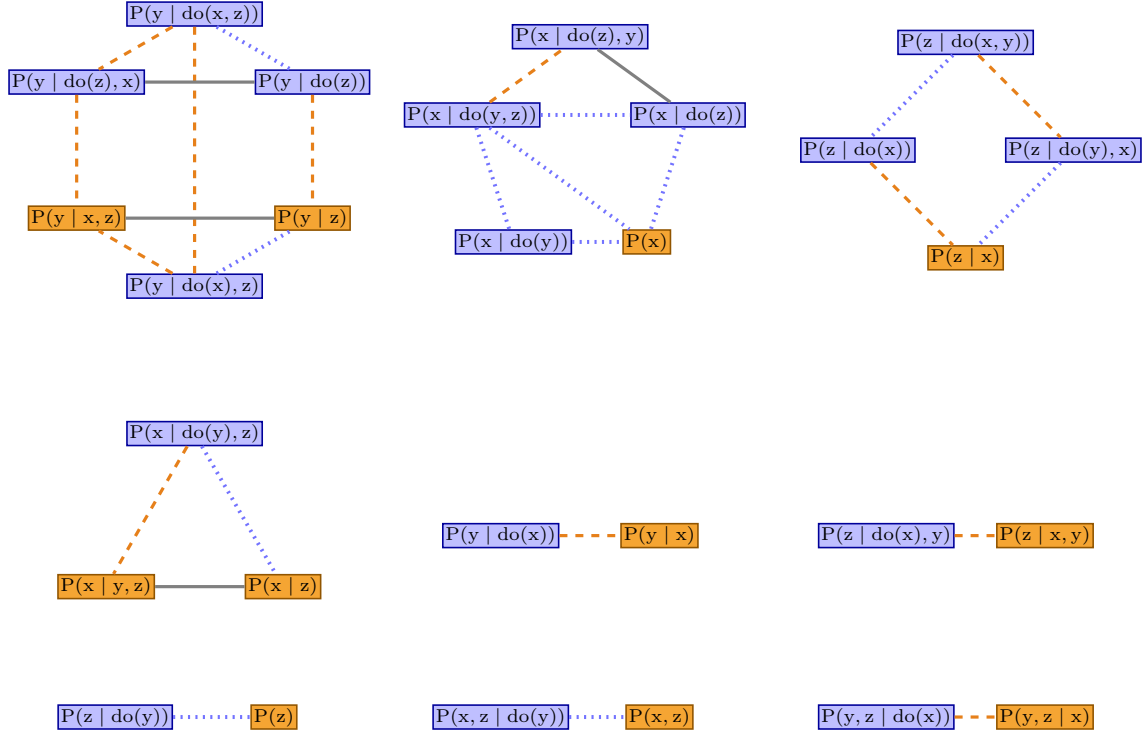


Figure 5: Derivation graph of $\mathcal{G}_3$. Blue nodes denote interventional queries, while orange nodes denote do-free queries. Edges represent applications of the do-calculus rules: gray solid lines for Rule $\mathcal{R}1$, orange dashed lines for Rule $\mathcal{R}2$, and blue dotted lines for Rule $\mathcal{R}3$. Only non isolated queries are displayed.

B Proofs of Section 5

Lemma 1 (Monotonicity under Graph Inclusion). *Let $\mathcal{V}$ be the set of observable variables, and let $\mathcal{G}_1$ and $\mathcal{G}_2$ be two causal diagrams on $\mathcal{V}$ such that $\mathcal{G}_2$ is a subgraph of $\mathcal{G}_1$. Let $\mathbf{X}, \mathbf{Y}, \mathbf{Z} \subseteq \mathcal{V}$ be pairwise disjoint with $\mathbf{Y} \neq \emptyset$. Then:*

1. *Any identification functional valid in $\mathcal{G}_1$ for $P(\mathbf{y} \mid \text{do}(\mathbf{x}), \mathbf{z})$ is also valid in $\mathcal{G}_2$.*
2. *Let $P^*(\mathcal{V})$ be an observational distribution. Any identification functional valid for $P(\mathbf{y} \mid \text{do}(\mathbf{x}), \mathbf{z})$ in all SCMs $\mathcal{M}_1$ inducing $\mathcal{G}_1$ and satisfying $P_{\mathcal{M}_1}(\mathcal{V}) = P^*(\mathcal{V})$ is also valid in all SCMs $\mathcal{M}_2$ inducing $\mathcal{G}_2$ and satisfying $P_{\mathcal{M}_2}(\mathcal{V}) = P^*(\mathcal{V})$.*
3. *Any causal derivation that holds in $\mathcal{G}_1$ also holds in $\mathcal{G}_2$.*
4. *Any graphical criterion satisfied in $\mathcal{G}_1$ is also satisfied in $\mathcal{G}_2$.*

Proof. Let us prove the different points:

1. Let $\mathcal{F}_1$ be an identification functional valid in $\mathcal{G}_1$ for $P(\mathbf{y} \mid \text{do}(\mathbf{x}), \mathbf{z})$. Let $\mathcal{M}_2$ be an SCM inducing $\mathcal{G}_2$. By adding missing arguments in the structural functions of $\mathcal{M}_2$, we can construct an SCM $\mathcal{M}_1$ inducing $\mathcal{G}_1$ such that $P_{\mathcal{M}_1}(\mathcal{V}) = P_{\mathcal{M}_2}(\mathcal{V})$ and $P_{\mathcal{M}_1}(\mathbf{y} \mid \text{do}(\mathbf{x}), \mathbf{z}) = P_{\mathcal{M}_2}(\mathbf{y} \mid \text{do}(\mathbf{x}), \mathbf{z})$:

Let $\mathcal{E}_B^+ := \mathcal{E}_B(\mathcal{G}_1) \setminus \mathcal{E}_B(\mathcal{G}_2)$ be the set of bidirected edges in $\mathcal{G}_1$ that are absent from $\mathcal{G}_2$. For each $V_i \leftrightarrow V_j \in \mathcal{E}_B^+$, introduce a fresh exogenous variable U_{ij} . Let $\mathcal{U}^+$ denote the set of these new variables and define the exogenous variables of $\mathcal{M}_1$ by $\mathcal{U}^{(1)} := \mathcal{U}^{(2)} \cup \mathcal{U}^+$. We retain the exogenous variables and their joint distribution from $\mathcal{M}_2$, and take each variable in $\mathcal{U}^+$ independently of them and of all other added variables. Thus, $\mathcal{M}_1$ has the same endogenous variables $\mathcal{V}$ and the same original exogenous

variables as $\mathcal{M}_2$, augmented by the new variables in $\mathcal{U}^+$. For each $V_i \in \mathcal{V}$, let $\mathcal{U}_i^+$ be the set of new exogenous variables corresponding to bidirected edges incident to V_i , and set $\mathcal{U}_i^{(1)} := \mathcal{U}_i^{(2)} \cup \mathcal{U}_i^+$.

Since $\mathcal{G}_2 \subseteq \mathcal{G}_1$, for all $V_i \in \mathcal{V}$ we have $\text{Pa}(V_i, \mathcal{G}_2) \subseteq \text{Pa}(V_i, \mathcal{G}_1)$. For any variable A , let $\mathcal{X}_A$ denote its valuation space. We define the structural functions $f_i^{(1)} \in \mathcal{F}_{\mathcal{M}_1}$ for each $V_i \in \mathcal{V}$ as:

$$f_i^{(1)} : \begin{cases} \mathcal{X}_{\text{Pa}(V_i, \mathcal{G}_1)} \times \mathcal{X}_{\mathcal{U}_i^{(2)}} \rightarrow \mathcal{X}_{V_i}, \\ (x_{\text{Pa}(V_i, \mathcal{G}_1)}, u_i^{(2)}) \mapsto f_i^{(2)}(x_{\text{Pa}(V_i, \mathcal{G}_2)}, u_i^{(2)}) \end{cases} \quad \text{if } \mathcal{U}_i^+ = \emptyset,$$

$$f_i^{(1)} : \begin{cases} \mathcal{X}_{\text{Pa}(V_i, \mathcal{G}_1)} \times \mathcal{X}_{\mathcal{U}_i^{(2)}} \times \mathcal{X}_{\mathcal{U}_i^+} \rightarrow \mathcal{X}_{V_i}, \\ (x_{\text{Pa}(V_i, \mathcal{G}_1)}, u_i^{(2)}, u_i^+) \mapsto f_i^{(2)}(x_{\text{Pa}(V_i, \mathcal{G}_2)}, u_i^{(2)}) \end{cases} \quad \text{if } \mathcal{U}_i^+ \neq \emptyset.$$

where $f_i^{(2)} \in \mathcal{F}_{\mathcal{M}_2}$ is the corresponding structural function in $\mathcal{M}_2$, and in the second case u_i^+ assigns values to the newly added exogenous variables in $\mathcal{U}_i^+$. In both cases, $f_i^{(1)}$ ignores the added directed-parent arguments in $\text{Pa}(V_i, \mathcal{G}_1) \setminus \text{Pa}(V_i, \mathcal{G}_2)$; in the second case, it also ignores u_i^+ . Each new U_{ij} is a shared exogenous parent of V_i and V_j , so it induces the required dashed bidirected edge. Since all added arguments are ignored, $\mathcal{M}_1$ and $\mathcal{M}_2$ induce the same observational and interventional distributions, while $\mathcal{M}_1$ induces $\mathcal{G}_1$.

Hence, $\mathcal{F}_1(P_{\mathcal{M}_2}(\mathcal{V})) = \mathcal{F}_1(P_{\mathcal{M}_1}(\mathcal{V})) = P_{\mathcal{M}_1}(\mathbf{y} \mid \text{do}(\mathbf{x}), \mathbf{z}) = P_{\mathcal{M}_2}(\mathbf{y} \mid \text{do}(\mathbf{x}), \mathbf{z})$. Therefore, $\mathcal{F}_1$ is valid in $\mathcal{G}_2$.

2. The construction above preserves the observational distribution and the interventional distribution. Hence, if $P_{\mathcal{M}_2}(\mathcal{V}) = P^*(\mathcal{V})$, then $P_{\mathcal{M}_1}(\mathcal{V}) = P^*(\mathcal{V})$ as well. identification functional valid for $P(\mathbf{y} \mid \text{do}(\mathbf{x}), \mathbf{z})$ in all SCMs $\mathcal{M}_1$ inducing $\mathcal{G}_1$ and satisfying $P_{\mathcal{M}_1}(\mathcal{V}) = P^*(\mathcal{V})$ is also valid in all SCMs $\mathcal{M}_2$ inducing $\mathcal{G}_2$ and satisfying $P_{\mathcal{M}_2}(\mathcal{V}) = P^*(\mathcal{V})$.
- 3 & 4. Since $\mathcal{G}_2$ is a subgraph of $\mathcal{G}_1$, every graphical independence that holds in $\mathcal{G}_1$ also holds in $\mathcal{G}_2$. Therefore, any causal derivation holding in $\mathcal{G}_1$, holds in $\mathcal{G}_2$ and any graphical criterion satisfied in $\mathcal{G}_1$ is satisfied in $\mathcal{G}_2$.

□