Identifiability in Causal Abstractions: A Hierarchy of Criteria

Clément Yvernes[1] **Emilie Devijver**[1] **Marianne Clausel**[2] **Eric Gaussier**[1]

[1]Univ Grenoble Alpes, CNRS, Grenoble INP, LIG
[2]Université de Lorraine, CNRS, IECL 54000 Nancy

Abstract

Identifying the effect of a treatment from observational data typically requires assuming a fully specified causal diagram. However, such diagrams are rarely known in practice, especially in complex or high-dimensional settings. To overcome this limitation, recent works have explored the use of causal abstractions—simplified representations that retain partial causal information. In this paper, we consider causal abstractions formalized as collections of causal diagrams, and focus on the identifiability of causal queries within such collections. We introduce and formalize several identifiability criteria under this setting. Our main contribution is to organize these criteria into a structured hierarchy, highlighting their relationships. This hierarchical view enables a clearer understanding of what can be identified under varying levels of causal knowledge. We illustrate our framework through examples from the literature and provide tools to reason about identifiability when full causal knowledge is unavailable.

1 INTRODUCTION

Knowing the effect of a treatment X on an outcome Y, encoded by an intervention $do(X)$ in the interventional distribution $P(Y|do(X))$, is crucial in many applications. However, performing interventions is often impractical due to ethical concerns, potential harm or prohibitive costs. In such cases, one can instead aim to identify do-free formulas that estimate the effects of interventions using only observational (non-experimental) data and a causal graph [Pearl, 2009]. Solving the identifiability problem typically involves establishing graphical criteria on causal diagrams under which the total effect is identifiable and providing a do-free formula for estimating it from observational data. However, specifying a causal diagram requires prior knowledge of the relationships between all observed variables, a requirement that is often unmet in real-world applications. This challenge is particularly acute in complex, high-dimensional settings, limiting the practical applicability of 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 some, but not necessarily all, of the underlying causal structure. Causal abstraction has been recently studied from multiple angles: its compositional properties [Rischel and Weichwald, 2021], the search for optimal abstractions [Zennaro et al., 2022], the use of macro-variables or variable groupings in place of micro-level variables [Parviainen and Kaski, 2017, Beckers and Halpern, 2018], and the quantification of information loss relative to the full causal model [Zennaro et al., 2023]. Abstractions have also been applied across a variety of contexts, including causal discovery [Wahl et al., 2024], query identification [Anand et al., 2023], reinforcement learning [Zennaro et al., 2024], as well as structured data settings such as time series [Assaad et al., 2022, 2024, Yvernes et al., 2025], neural networks [Geiger et al., 2021], mechanistic interpretability [Geiger et al., 2025] and hierarchical data [Weinstein and Blei, 2024].

In this work, we formalize a causal abstraction as a collection of causal diagrams, and focus on the problem of identifying a causal query within such a collection. While various notions of identifiability have been proposed in this context, there is currently no unified framework that connects them. Our goal is to clarify and relate these different notions by introducing a formal structure for reasoning about identifiability in abstracted causal settings. More precisely, our contributions are as follows:

- We formally define several notions of identifiability directly on collections of graphs;
- We analyze the relationships between these different notions;

Accepted for the 1st Workshop on Causal Abstractions and Representations (CAR) at UAI 2025 (CAR 2025).

- We illustrate each concept with examples and connections to existing work from the literature.

The remainder of the paper is structured as follows: Section 2 introduces the main concepts while Section 3 presents identifiability in a single graph; and Section 4 presents different notions of identifiability in a collection of graphs. Section 5 provides the link between those notions of identifiability, and Section 6 develops a conjecture. Lastly, Section 7 concludes the paper.

2 PRELIMINARIES

Graphs. A *(directed) mixed graph* $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ is a graph that contains two type of arrows, *directed arrows* ($\rightarrow$) and *dashed-bidirected arrows* ($\dashleftarrow\dashrightarrow$). Following the notations of Perković et al. [2018], an arrow ($\leftarrow\!\bullet$) represents either a directed arrow ($\leftarrow$) or a dashed-bidirected arrows ($\dashleftarrow\dashrightarrow$). We potentially distinguish arrows $\mathcal{E}$ into directed arrows $\mathcal{E}_D$ and dashed bidirected arrows $\mathcal{E}_B$. 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 a *causal path*, is a path in which all arrows are pointing 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 (i.e. $\bullet\!\rightarrow V \leftarrow\!\bullet$) in π. Otherwise, it is said to be a *non-collider* on π. A vertex V is said to be *active* on a path relative to $\mathbf{Z}$ if 1) V is a collider and V or any of its descendants are 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 $\mathcal{G}_{\overline{\mathbf{X}}\underline{\mathbf{Z}}}$ is the result of removing from a graph $\mathcal{G}$ edges with an arrowhead into $\mathbf{X}$ (e.g., $A \bullet\!\rightarrow \mathbf{X}$), and edges with a tail from $\mathbf{Z}$ (e.g., $A \leftarrow \mathbf{Z}$).

Structural Causal Models. A *Structural Causal Model (SCM)* $\mathcal{M}$ is 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 an ADMG – $\mathcal{G}_{\mathcal{M}} := (\mathcal{V}, \mathcal{E} = (\mathcal{E}_D, \mathcal{E}_B))$, known as a *causal diagram*, that encodes the structural relations among $\mathcal{V} \cup \mathcal{U}$, where every $V_i \in \mathcal{V}$ 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 $V_i \dashleftarrow\dashrightarrow V_j$.

Intervention. In a given SCM $\mathcal{M} = \langle \mathcal{U}, \mathcal{V}, \mathcal{F}, P(\mathcal{U}) \rangle$, let $\mathbf{V} = (V_1, \ldots, V_n)$ be a subset 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 $P_{\mathcal{M}}(\mathcal{V} \mid \text{do}(\mathcal{V} = \mathbf{v}))$ or $P_{\mathcal{M}}(\mathcal{V} \mid \text{do}(\mathbf{v}))$ for short. Let $\mathbf{X}$, $\mathbf{Y}$, and $\mathbf{Z}$ be pairwise distinct subsets of $\mathcal{V}$.

Causal Query and Estimand. Let $\mathcal{V}$ denote the set of observable variables and let $\mathbf{X}, \mathbf{Y}, \mathbf{Z} \subseteq \mathcal{V}$ be pairwise disjoint. A *causal query* is any expression of the form $P(\mathbf{y} \mid \text{do}(\mathbf{x}), \mathbf{z})$. A *causal estimand* is any function $\mathcal{F}$ mapping distributions over $\mathcal{V}$ to $\mathbb{R}^+$. Given an SCM $\mathcal{M} = \langle \mathcal{U}, \mathcal{V}, \mathcal{F}, P(\mathcal{U}) \rangle$, $P_{\mathcal{M}}(\mathcal{V})$ its observational distribution and a causal query $P(\mathbf{y} \mid \text{do}(\mathbf{x}), \mathbf{z})$, a causal estimand $\mathcal{F}$ is said to be *valid* in $\mathcal{M}$ for $P(\mathbf{y} \mid \text{do}(\mathbf{x}), \mathbf{z})$ if it satisfies $\mathcal{F}(P_{\mathcal{M}}(\mathcal{V})) = P(\mathbf{y} \mid \text{do}(\mathbf{x}), \mathbf{z})$.

3 IDENTIFIABILITY IN A SINGLE GRAPH

Different structural causal models (SCMs) can induce the same causal diagram. Consequently, a causal diagram represents a collection of SCMs. A causal query $P(\mathbf{y} \mid \text{do}(\mathbf{x}), \mathbf{z})$ is said to be *identifiable* if it can be uniquely determined from the observational distribution.

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 pair-*

[1] By convention, a vertex is a descendant and an ancestor of itself.

wise disjoint. The causal query $P(\mathbf{y} \mid do(\mathbf{x}), \mathbf{z})$ is identifiable *in $\mathcal{G}$ if there exists a causal estimand valid for $P(\mathbf{y} \mid do(\mathbf{x}), \mathbf{z})$ in every SCM $\mathcal{M}$ satisfying $\mathcal{G}_{\mathcal{M}} = \mathcal{G}$. In this case, we also say that the causal estimand is valid for $P(\mathbf{y} \mid do(\mathbf{x}), \mathbf{z})$ in $\mathcal{G}$.*

Pearl's do-calculus provides a set of three rules for transforming interventional distributions into observational ones. We restate these rules below.

Theorem 1 (Do-Calculus [Pearl, 2009])**.** *Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be a causal diagram, and let $\mathbf{X}, \mathbf{Y}, \mathbf{Z}, \mathbf{W} \subseteq \mathcal{V}$ be pairwise disjoint. For any SCM $\mathcal{M}$ inducing $\mathcal{G}$, denote its interventional density by $P_{\mathcal{M}}$. Then:*

R1. Insertion/deletion of observations.

$$P_{\mathcal{M}}(\mathbf{y} \mid do(\mathbf{w}),\ \mathbf{x},\ \mathbf{z}) = P_{\mathcal{M}}(\mathbf{y} \mid do(\mathbf{w}),\ \mathbf{z})$$

if $\mathbf{Y} \perp\!\!\!\perp_{\mathcal{G}_{\overline{\mathbf{W}}}} \mathbf{X} \mid (\mathbf{W} \cup \mathbf{Z})$.

R2. Action/observation exchange.

$$P_{\mathcal{M}}(\mathbf{y} \mid do(\mathbf{w}),\ do(\mathbf{x}),\ \mathbf{z}) = P_{\mathcal{M}}(\mathbf{y} \mid do(\mathbf{w}),\ \mathbf{x},\ \mathbf{z})$$

if $\mathbf{Y} \perp\!\!\!\perp_{\mathcal{G}_{\overline{\mathbf{W}},\underline{\mathbf{X}}}} \mathbf{X} \mid (\mathbf{W} \cup \mathbf{Z})$.

R3. Insertion/deletion of actions.

$$P_{\mathcal{M}}(\mathbf{y} \mid do(\mathbf{w}),\ do(\mathbf{x}),\ \mathbf{z}) = P_{\mathcal{M}}(\mathbf{y} \mid do(\mathbf{w}),\ \mathbf{z})$$

if $\mathbf{Y} \perp\!\!\!\perp_{\mathcal{G}_{\overline{\mathbf{W}},\overline{\mathbf{X}(\mathbf{Z})}}} \mathbf{X} \mid (\mathbf{W} \cup \mathbf{Z})$*, where* $\mathbf{X}(\mathbf{Z}) := \mathbf{X} \backslash Anc_{\mathcal{G}_{\overline{\mathbf{W}}}}(\mathbf{Z})$.

A *proof of do-calculus* is a finite sequence of applications of the do-calculus rules, possibly combined with standard probability manipulations, that rewrites a causal query $P(\mathbf{y} \mid \mathrm{do}(\mathbf{x}), \mathbf{z})$ into a do-free formula $\mathcal{F}(P(\mathcal{V}))$. Such a proof is said to *hold in* a causal diagram $\mathcal{G}$ if all the d-separation conditions required by each rule application are satisfied in $\mathcal{G}$. In this case, the resulting expression $\mathcal{F}$ is a valid causal estimand for the query and is referred to as an *identification formula*.

Do-calculus is both sound and complete for causal diagrams [Shpitser and Pearl, 2006]: if a query is identifiable in a causal diagram $\mathcal{G}$, then there exists a proof of do-calculus that holds in $\mathcal{G}$. This proof yields a valid identification formula for the query in $\mathcal{G}$. Conversely, if the query is not identifiable in $\mathcal{G}$, then no such a proof exists.

4 IDENTIFIABILITY IN A COLLECTION OF GRAPHS.

Specifying a full causal diagram demands detailed knowledge of all causal relationships among observed variables—an assumption rarely satisfied 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.

To our knowledge, each causal abstraction determines a collection C of causal diagrams on $\mathcal{V}$ that satisfy its specified constraints. We study identifiability of a causal query within such a class: broadly, a query is *identifiable* in C if the same estimand can be derived for every graph $G \in C$. 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.

4.1 EPISTEMIC DIFFERENCES

We now extend Pearl's notion of identifiability from a single causal diagram to an entire class of diagrams. Let C be a set of causal diagrams on the variable set $\mathcal{V}$. Intuitively, a causal query is identifiable in C if it admits the same estimand for every graph in C. We formalize this below.

Definition 2 (Identifiability through Graphs (IG))**.** *Let C be a class of causal diagrams over $\mathcal{V}$, and let $\mathbf{X}, \mathbf{Y}, \mathbf{Z} \subseteq \mathcal{V}$ be pairwise disjoint. The causal query $P(\mathbf{y} \mid do(\mathbf{x}), \mathbf{z})$ is* identifiable through graphs *in C if there exists a causal estimand valid for $P(\mathbf{y} \mid do(\mathbf{x}), \mathbf{z})$ in every SCM $\mathcal{M}$ satisfying $\mathcal{G}_{\mathcal{M}} \in C$.*

Remark that it means that the causal query $P(\mathbf{y} \mid \mathrm{do}(\mathbf{x}), \mathbf{z})$ is *identifiable through graphs* in C if there exists a causal estimand valid for $P(\mathbf{y} \mid \mathrm{do}(\mathbf{x}), \mathbf{z})$ in all causal diagram $\mathcal{G} \in C$ from Definition 1.

For instance, Jaber et al. [2022] provide a complete characterization of identifiability through graphs in partial ancestral graphs.

The key distinction between identifiability in a single graph and identifiability through graphs lies in the set of SCMs under consideration. In the single-graph setting, we restrict attention to those SCMs whose causal diagram is exactly $\mathcal{G}$. By contrast, in the class-of-graphs setting, we consider the union of the SCM classes induced by every graph in C. One may further restrict this union by incorporating additional domain knowledge.

As an illustrative example, suppose we know the true observational density $P^{\star}(\mathcal{V})$ (e.g. corresponding to an infinite sample size). We then restrict our attention to those SCMs whose observational distribution satisfies $P_{\mathcal{M}}(\mathcal{V}) = P^{\star}(\mathcal{V})$. Under this constraint, we say a causal query is *identifiable through Graphs knowing $P^{\star}(\mathcal{V})$* (IGP) in the class C.

Definition 3 (Identifiability through Graphs knowing $P^{\star}$ (IGP))**.** *Let C be a class of causal diagrams over $\mathcal{V}$, and let $\mathbf{X}, \mathbf{Y}, \mathbf{Z} \subseteq \mathcal{V}$ be pairwise disjoint. Let $P^{\star}(\mathcal{V})$ denote the true data density. The causal query $P(\mathbf{y} \mid do(\mathbf{x}), \mathbf{z})$ is* identifiable through graphs knowing $P^{\star}$ *in C if there exists a causal estimand valid for $P(\mathbf{y} \mid do(\mathbf{x}), \mathbf{z})$ in every SCM $\mathcal{M}$ satisfying $\mathcal{G}_{\mathcal{M}} \in C$ and $P_{\mathcal{M}}(\mathcal{V}) = P^{\star}(\mathcal{V})$.*

Despite their theoretical appeal, both definitions encounter serious practical limitations. Identifiability through graphs knowing $P^\star(\mathcal{V})$ (IGP) requires knowing the true observational density—effectively demanding an infinite sample. IG can, in principle, be checked by enumerating every $G \in \mathcal{C}$ and verifying that all derived estimands coincide; however, $\mathcal{C}$ may be extremely large or even infinite, rendering this procedure infeasible. To overcome these obstacles, we propose to distinguish between different notions of identifiability based on the proof techniques used to establish IG in $\mathcal{C}$.

4.2 METHODOLOGICAL DIFFERENCES

IG represents the strongest notion of graphical identifiability one can hope to achieve without making structural assumption or additional assumptions about the data distribution. In general, IG is a highly challenging problem. Hence, simpler notions of identifiability are proposed here.

To establish identifiability of a causal query in a collection $\mathcal{C}$, one possible approach is to apply do-calculus uniformly across all graphs in $\mathcal{C}$. Specifically, this entails finding a single proof that holds in all causal diagram of the collection.

Definition 4 (Identifiability by Common Do-Calculus (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. The causal query $P(\mathbf{y} \mid do(\mathbf{x}), \mathbf{z})$ is* identifiable by common do-calculus *in $\mathcal{C}$ if there exists a proof of do-calculus 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 $\mathcal{C}$.[2] Such a calculus consists of three atomic rules $(A1, A2, A3)$, which are counterparts of the three do-calculus rules $(R1, R2, R3)$, and for each i:

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

Consequently, any derivation constructed in an atomically complete calculus immediately yields a common do-calculus proof, and hence establishes ICD for the underlying class of diagrams. An example of a class where ICD has been achieved via this approach is that of Markov equivalence class of PAGs [Jaber et al., 2022].

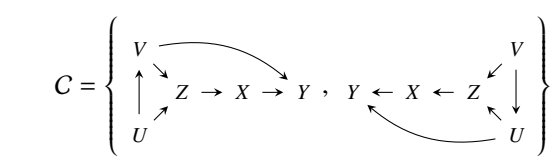
The main distinction between identifiability through graphs and identifiability by common do-calculus lies in the permissible proof techniques. In the latter, one must provide a single proof of do-calculus that holds in all compatible graphs. In contrast, identifiability through graphs allows one to apply any graphical identification procedure on each compatible graph individually, so long as every such procedure yields the same, valid causal estimand.

Sometimes, it is also possible to identify an effect on an isolated graph using a *graphical criterion* i.e. a set of graphical independence conditions that the diagram must satisfy in order for the criterion to apply. For instance, Pearl [2009] mentions the backdoor criterion and the frontdoor criterion. This leads to concepts such as *identifiability by common backdoor (ICB)* or *identifiability by common frontdoor (ICF)*, where one verifies that the graphical criterion, e.g., the backdoor criterion, applies to all graphs in $\mathcal{C}$. This defines a multitude of different graphical identifiability notions. The union of all them gives Definition 5.

Definition 5. *(Identifiability by Common Graphical Criterion (ICGC)) 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. The causal query $P(\mathbf{y} \mid do(\mathbf{x}), \mathbf{z})$ is* identifiable by common graphical criterion *in $\mathcal{C}$ if there exists a graphical criterion satisfied in all graphs in $\mathcal{C}$.*

Example 1 illustrates a class in which a query is identifiable by common backdoor.

Example 1. *Let us consider the following class of causal diagram $\mathcal{C}$ as follows:*



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

Identifiability by common backdoor in summary causal graphs has been studied in Assaad et al. [2024], and was later extended into a complete criterion for identifiability by common adjustment by Yvernes et al. [2025]. 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.

5 PROPERTIES OF IDENTIFIABILITY NOTIONS IN COLLECTIONS OF GRAPHS

5.1 A PROPERTY VALID FOR ALL GRAPHICAL NOTIONS

All notions of identifiability relying solely on the collection of graphs may be difficult to apply in practice because the

[2] An example of an atomically complete calculus that turns out to be complete is given in Anand et al. [2023].

collection C can be very large. However, it is possible to reduce the size of C without affecting the identifiability of the collection. Since edges in a graph represent dependencies, removing an edge preserves the graphical independences present in the original graph. This intuition is formalized in Lemma 1.

Lemma 1. *Let $\mathcal{V}$ denotes the set of observable variables. Let $\mathcal{G}_1$ and $\mathcal{G}_2$ be two causal diagrams on $\mathcal{V}$ and let $\mathbf{X}, \mathbf{Y}, \mathbf{Z} \subseteq \mathcal{V}$ be pairwise disjoint. If $\mathcal{G}_2$ is a subgraph of $\mathcal{G}_1$, then the following propositions hold:*

1. *Any causal estimand valid in $\mathcal{G}_1$ for $P(\mathbf{y} \mid do(\mathbf{x}), \mathbf{z})$ is valid in $\mathcal{G}_2$ for $P(\mathbf{y} \mid do(\mathbf{x}), \mathbf{z})$.*
2. *Any proof of do-calculus that holds in $\mathcal{G}_1$ also holds in $\mathcal{G}_2$.*
3. *Any graphical criterion satisfied in $\mathcal{G}_1$ is satisfied in $\mathcal{G}_2$.*

Proof. Let us prove the different points:

1. Let $\mathcal{F}_1$ be a causal estimand 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 $P_{\mathcal{M}_1}(\mathbf{y} \mid \text{do}(\mathbf{x}), \mathbf{z}) = P_{\mathcal{M}_2}(\mathbf{y} \mid \text{do}(\mathbf{x}), \mathbf{z})$:

   Indeed, for all $V_i \in \mathcal{V}$, we know that $\text{Pa}(V_i, \mathcal{G}_2) \subseteq \text{Pa}(V_i, \mathcal{G}_1)$. For any variable A, let $\mathcal{X}_A$ denote its valuation space. We construct $\mathcal{M}_1$ by keeping the same exogenous variables $\mathcal{U}$, endogenous variables $\mathcal{V}$, and exogenous distribution $P(\mathcal{U})$ as in $\mathcal{M}_2$, and by defining new structural functions $f_i^{(1)} \in \mathcal{F}_{\mathcal{M}_1}$ for each $V_i \in \mathcal{V}$ as:

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

   where $f_i^{(2)} \in \mathcal{F}_{\mathcal{M}_2}$ is the corresponding structural function in $\mathcal{M}_2$. That is, $f_i^{(1)}$ coincides with $f_i^{(2)}$ and ignores the arguments in $\text{Pa}(V_i, \mathcal{G}_1) \backslash \text{Pa}(V_i, \mathcal{G}_2)$. As a result, $\mathcal{M}_1$ and $\mathcal{M}_2$ induce the same observational and interventional distributions, while $\mathcal{M}_1$ respects the structural dependencies encoded in $\mathcal{G}_1$.

   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. 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, the proof of do-calculus also holds in $\mathcal{G}_2$.
3. 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 graphical criterion satisfied in $\mathcal{G}_1$ is satisfied in $\mathcal{G}_2$.

□

Lemma 1 implies that, to determine identifiability through graphs in the collection C, it is sufficient to consider only the maximal graphs in C (with respect to subgraph inclusion). This is formalized in Theorem 2.

Theorem 2. *Let C be a class of causal diagrams over a set of variables $\mathcal{V}$, and let $\mathbf{X}, \mathbf{Y}, \mathbf{Z} \subseteq \mathcal{V}$ be pairwise disjoint. Let $C^{\max}$ denote the subcollection of maximal elements of C under graph inclusion. Consider any graphical notion of identifiability as described above (i.e. IG, ICD, ICGC, ICB, ICF). Then the following propositions are equivalent:*

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

Proof. To treat all identifiability notions uniformly, we define an *identification object* as either a causal estimand, a do-calculus proof, or a graphical criterion, and say it *applies* to $\mathcal{G}$ if it is valid in $\mathcal{G}$, holds in $\mathcal{G}$, or is satisfied in $\mathcal{G}$, respectively. Let us prove the two implications:

(1) ⇒ (2): Suppose $P(\mathbf{y} \mid \text{do}(\mathbf{x}), \mathbf{z})$ is identifiable in C. $C^{\max}$ is a subcollection of C. Hence, the query is identifiable in $C^{\max}$.

(2) ⇒ (1): Suppose now that $P(\mathbf{y} \mid \text{do}(\mathbf{x}), \mathbf{z})$ is identifiable in $C^{\max}$. By definition, there exists an identification object $\theta^{\max}$ applying to every $\mathcal{G}^{\max} \in C^{\max}$. Let $\mathcal{G}$ be any causal diagram in C. Then, by definition of maximality, there exists $\mathcal{G}^{\max} \in C^{\max}$ such that $\mathcal{G}$ is a subgraph of $\mathcal{G}^{\max}$. By Lemma 1, $\theta^{\max}$ also applies to $\mathcal{G}$. Then the query is identifiable in C. □

Theorem 2 shows that identifying a causal effect becomes computationally easier when the subcollection of maximal elements under graph inclusion is small. As formalized in Corollary 1, if this subcollection reduces to a singleton, then identifiability in the entire class C reduces to identifiability in a single graph.

Corollary 1. *If C contains a greatest element under graphical inclusion, then IG is equivalent to ICD.*

This situation arises, for example, in Cluster DAGs [Anand et al., 2023], where the partition admissibility assumption ensures that there is a unique maximal element under inclusion. Designing causal abstractions that induce few maximal elements therefore appears to be a good practice when aiming for tractable identifiability conditions.

5.2 LINKS BETWEEN THESE CONCEPTS

This section explores the fundamental relationships between the different notions of identifiability introduced earlier: Theorem 3 details the different implications and Figure 1

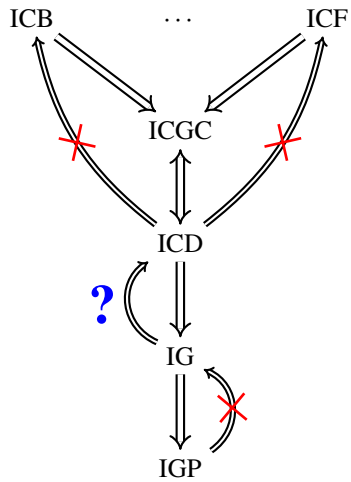


Figure 1: Relations between the notions of identifiability. An arrow indicates that the implication holds in every case. A red cross on an arrow indicates that the implication fails in at least one case. A blue question mark on an arrow indicates that the validity of the implication is still open.

illustrates all the logical implications that relate the various notions of identifiability discussed in this work.

Theorem 3. *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. We have the following properties:*

- *If $P(\mathbf{y} \mid do(\mathbf{x}), \mathbf{z})$ is identifiable by a common specific graphical criterion in $\mathcal{C}$ (e.g. ICB or ICF), then it is ICGC in $\mathcal{C}$.*
- ***ICGC $\Leftrightarrow$ ICD:*** *$P(\mathbf{y} \mid do(\mathbf{x}), \mathbf{z})$ is ICGC in $\mathcal{C}$ if and only if it is ICD in $\mathcal{C}$.*
- ***ICD $\Rightarrow$ IG:*** *if $P(\mathbf{y} \mid do(\mathbf{x}), \mathbf{z})$ is ICD in $\mathcal{C}$, then it is IG in $\mathcal{C}$.*
- ***IG $\Rightarrow$ IGP:*** *if $P(\mathbf{y} \mid do(\mathbf{x}), \mathbf{z})$ is IG in $\mathcal{C}$, then it is IGP in $\mathcal{C}$.*

Proof. Let us prove the implications:

- Identifibility by a common specific (e.g backdoor or frontdoor) criterion is a special case of ICGC.
- **ICGC $\Rightarrow$ ICD:** any graphical criterion on a single graph can be established by a proof of do-calculus. Therefore, if a query is identifiable by a common graphical criterion in $\mathcal{C}$, it is also identifiable by common do-calculus in $\mathcal{C}$.
- **ICD $\Rightarrow$ ICGC:** if a query is identifiable by common do-calculus in $\mathcal{C}$, then there exists a single do-calculus proof that holds in every $\mathcal{G} \in \mathcal{C}$. Since each rule in this proof depends solely on graphical independence conditions, the conjunction of all these conditions defines a common graphical criterion satisfied by every $\mathcal{G} \in \mathcal{C}$.
- **ICD $\Rightarrow$ IG:** the causal query is identifiable by common do-calculus. Therefore, there exists a proof of do-calculus that transforms the causal query $P(\mathbf{y} \mid \text{do}(\mathbf{x}), \mathbf{z})$ into a do-free formula $\mathcal{F}(P(\mathcal{V}))$, where $\mathcal{F}$ is a valid identification formula for every $\mathcal{G} \in \mathcal{C}$. Therefore, $P(\mathbf{y} \mid \text{do}(\mathbf{x}), \mathbf{z})$ is identifiable through graphs in $\mathcal{C}$.
- **IG $\Rightarrow$ IGP:** let $\mathcal{C}$ be a collection of graphs in which the causal query $P(\mathbf{y} \mid \text{do}(\mathbf{x}), \mathbf{z})$ is identifiable through graphs. By definition, the underlying class of SCMs in Definition 3 is strictly smaller. Thus, IGP holds in $\mathcal{C}$ for this query.

□

In a single graph, neither the backdoor criterion nor the frontdoor criterion is complete for identification. Consequently, ICB and ICF cannot be complete for ICD. Moreover, no existing graphical criterion is complete in a single graph, when consider causal diagrams (ADMGs). This suggests that no standard graphical criterion is likely to be complete for identifiability via common do-calculus.

To establish IGP, one can use knowledge on the true density of data. As a result, graphical knowledge may not be sufficient to establish IGP. That's why IG is not complete for IGP as shown in Example 2.

Example 2. *Let us consider the collection of causal diagrams $\mathcal{C} := \{X \to Y, X \leftarrow Y\}$ and let us consider the causal query $P(Y = 1 \mid do(X = 1))$.*

$P(Y = 1 \mid do(X = 1))$ ***is not identifiable through graphs in $\mathcal{C}$.*** *Indeed, let us consider the SCM $\mathcal{M}_1$ defined by $X \sim \mathcal{B}\left(\frac{1}{2}\right)$ and $Y = X \otimes U$ where $U \sim \mathcal{B}\left(\frac{1}{10}\right)$ and $\otimes$ denotes the XOR operator. Moreover, let us consider the SCM $\mathcal{M}_2$ defined by $Y \sim \mathcal{B}\left(\frac{1}{2}\right)$ and $X = Y \otimes U$ where $U \sim \mathcal{B}\left(\frac{1}{10}\right)$. 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, $P_{\mathcal{M}_1}(Y = 1 \mid do(X = 1)) = \frac{9}{10}$ and $P_{\mathcal{M}_2}(Y = 1 \mid do(X = 1)) = \frac{1}{2}$. Therefore, $P(Y = 1 \mid do(X = 1))$ is not identifiable through graphs in $\mathcal{C}$.

$P(Y = 1 \mid do(X = 1))$ ***may be identifiable through graphs knowing $P^\star(X, Y)$ in $\mathcal{C}$.*** *Indeed, let us consider the case where X and Y are independent in the true data distribution $P^\star(X, Y)$. Let $\mathcal{M}_1$ be an SCM inducing $X \to Y$ such that $P_{\mathcal{M}_1}(X, Y) = P^\star(X, Y)$. By the do-calculus, we know that $P_{\mathcal{M}_1}(Y = 1 \mid do(X = 1)) = P_{\mathcal{M}_1}(Y = 1 \mid X = 1) = P^\star(Y = 1)$. Similarly, let $\mathcal{M}_2$ be an SCM inducing $X \leftarrow Y$ such that $P_{\mathcal{M}_2}(X, Y) = P^\star(X, Y)$. By the do-calculus, we know that $P_{\mathcal{M}_2}(Y = 1 | do(X = 1)) = P^\star(Y = 1)$. Therefore the causal estimand $\mathcal{F} := P(X, Y) \mapsto P(Y = 1)$ is a valid causal estimand in all SCM $\mathcal{M}$ satisfying $\mathcal{G}_{\mathcal{M}} \in \mathcal{C}$, $P_M(X, Y) =$*

$P^\star(X, Y)$, and $P(Y = 1 \mid do(X = 1))$ is identifiable through graphs knowing $P^\star(X, Y)$ in $\mathcal{C}$.

One implication, specifically the relationship between IG and ICD, remains unresolved. This open question is the subject of a conjecture presented in Section 6.

The hierarchy established in Theorem 3 allows for a classification of the identifiability notions by increasing complexity. The more general a notion is, the more difficult it becomes to fully characterize it in a given collection of graphs.

6 IG VS ICD CONJECTURE

In this section we focus on two notions of identifiability in collections of causal diagrams: identifiability through graphs and identifiability by common do-calculus. Despite extensive study, the precise relationship between these notions remains unresolved. In particular, it is not known whether there exists a family of causal diagrams $\mathcal{C}$ and a causal query Q such that Q is IG in $\mathcal{C}$ but is not ICD in $\mathcal{C}$. We conjecture that such a class $\mathcal{C}$ indeed exists, even though no concrete example is yet available. Proving this conjecture would establish a strict separation between graphical identifiability and common do-calculus identifiability in causal abstractions. Conversely, if no such class exists, it would follow that any calculus which is atomically complete on $\mathcal{C}$ is in fact fully complete on $\mathcal{C}$, thereby unifying these two notions of identifiability.

How to prove the conjecture. To prove the conjecture, one may exhibit two causal diagrams $\mathcal{G}_1$ and $\mathcal{G}_2$, two do-calculus proofs $\mathcal{P}_1$ and $\mathcal{P}_2$, and a causal query $P(\mathbf{y} \mid \text{do}(\mathbf{x}), \mathbf{z})$ satisfying the following properties:

- $\mathcal{P}_1$ and $\mathcal{P}_2$ are respectively valid in $\mathcal{G}_1$ and $\mathcal{G}_2$.
- $\mathcal{P}_1$ and $\mathcal{P}_2$ yield the same identification formula for $P(\mathbf{y} \mid \text{do}(\mathbf{x}), \mathbf{z})$.
- There is *no* single do-calculus proof that is valid in both $\mathcal{G}_1$ and $\mathcal{G}_2$.

How to disprove the conjecture. To disprove the conjecture, one must show that for *every* class of causal diagrams $\mathcal{C}$ and *every* query $P(\mathbf{y} \mid \text{do}(\mathbf{x}), \mathbf{z})$, if the query is IG in $\mathcal{C}$, then it is also ICD in $\mathcal{C}$. Equivalently, one must provide a uniform procedure that takes as input any class $\mathcal{C}$ and a query IG in $\mathcal{C}$, and outputs a do-calculus proof that holds in all $\mathcal{G} \in \mathcal{C}$.

7 CONCLUSION

In this paper, we have introduced a unified framework for comparing the relative difficulty of various identifiability notions in causal abstractions, modelled as collections of causal diagrams. By arranging these notions into a hierarchy, we make explicit which criteria are strictly stronger than others. This organization allows one, for any given abstraction, to determine whether the strongest possible identifiability notion has been met, or if there remain stronger criteria yet to be achieved.

Looking ahead, it is essential to rigorously classify the algorithmic complexity associated with each identifiability notion, and to develop efficient procedures for verifying them. A pivotal open challenge is the IG versus ICD conjecture: determining whether there exists a class of causal diagrams in which a given causal query is ICD but not IG.

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, editors, *Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence*, volume 180 of *Proceedings of Machine Learning Research*, pages 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`.

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, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 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`.

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, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 3679–3690. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/17a9ab4190289f0e1504bbb98d1d111a-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. URL https://www.sciencedirect.com/science/article/pii/S0888613X17303134.

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.

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*, 18:1–62, May 2018. ISSN 1532-4435. doi: 10.3929/ethz-b-000278021. URL https://www.research-collection.ethz.ch/handle/20.500.11850/285660.

Eigil F. Rischel and Sebastian Weichwald. Compositional abstraction error and a category of causal models. In Cassio de Campos and Marloes H. Maathuis, editors, *Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence*, volume 161 of *Proceedings of Machine Learning Research*, pages 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*, pages 1219–1226, Boston, Massachusetts, July 2006. AAAI Press. ISBN 978-1-57735-281-5.

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, 2024. URL https://arxiv.org/abs/2401.05330.

Clément Yvernes, Emilie Devijver, and Eric Gaussier. Complete characterization for adjustment in summary causal graphs of timeseries. In *UAI2025*, 2025.

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.