Characterizing Markov-Equivalent Causal Models with Latent Variables for Neural Partial Identification

Hongshuo Yang* **Adiba Ejaz*** **Yushu Pan*** **Elias Bareinboim**
Causal Artificial Intelligence Lab
Columbia University
{yhs,adiba.ejaz,yushupan,eb}@cs.columbia.edu

Abstract

Causal reasoning from observational data becomes significantly challenging when the underlying causal structure is only partially known, particularly in the presence of latent confounding. In this work, we study the space of structural causal models (SCMs) whose induced causal diagrams belong to a Markov equivalence class (MEC) represented by a partial ancestral graph (PAG). Specifically, we characterize SCM-induced causal diagrams in this MEC through two counterfactually expressive classes, *skeleton-matched diagrams* (SMDs) and *directed-backboned diagrams* (DBDs), and then translate these graphical characterizations into constraints over adjacency masks. Building on these results, we introduce a new model class called *PAG-constrained neural causal models* ($\mathcal{P}$-NCMs) which incorporate these constraints to parameterize a restricted class of SCMs that preserves the full counterfactual expressiveness of the MEC. Finally, we formulate counterfactual partial identification under Markov equivalence as an optimization problem over $\mathcal{P}$-NCMs, establish its theoretical correctness, and empirically evaluate its performance on simulated causal bounding tasks.

1 Introduction

Causal inference requires assumptions about the underlying data-generating process to reason beyond observational associations [18]. These assumptions can be understood as structural invariances imposed by an underlying *Structural Causal Model* (SCM) on the distributions it induces across the three layers of the *Pearl Causal Hierarchy* (PCH). In particular, *causal diagrams* and their associated graphical models provide a compact representation of these assumptions by relating graphical features to distributional properties across the PCH [4, 7, 25]. In many real-world applications across science and machine learning, however, only observational data are available, while the causal diagram is unknown. A large body of work therefore focuses on learning causal structure from data prior to performing inference [22, 19]. In general, without additional assumptions, only a *Markov equivalence class* (MEC) of causal diagrams can be recovered from observational data. In the presence of latent confounding, such equivalence classes are commonly represented by *partial ancestral graphs* (PAGs) [22, 27].

Given a learned PAG, a natural goal is to use the structural invariances it encodes to enable causal inference beyond the observational level [26, 10]. However, due to the structural uncertainty inherent in PAGs, causal inference becomes significantly more challenging, as its validity must hold across all SCMs in the corresponding MEC. Direct enumeration of these models is infeasible because of the combinatorial size of the MEC space. Existing approaches attempt to mitigate this challenge by reducing the search space using graphical properties of the PAG. For example, Hyttinen et al. propose an algorithm that iteratively prunes the set of candidate graphs [9], while Malinsky et al. leverage

*Equal contribution.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

the generalized back-door criterion to restrict attention to admissible queries [16]. Another line of work extends do-calculus to PAGs, known as PAG calculus, to enable interventional inference at the level of the MEC without explicit enumeration [26, 11, 12, 14, 15, 13, 10]. Despite these advances, existing methods are largely limited to the interventional layer and do not provide an expressive representation of SCMs in the MEC that supports more general inference tasks, particularly at the counterfactual level, without resorting to enumeration.

When the causal diagram is known, a special class of SCMs, known as *neural causal models (NCMs)* can be constructed to encode the structural constraints of the diagram while providing expressive parameterizations capable of matching counterfactual distributions across all SCMs inducing that diagram. The parameterization of NCMs makes it possible to formulate inference tasks as continuous optimization problems that support counterfactual reasoning while fitting observational data. However, when only a PAG is available, this construction breaks down due to the uncertainty in variable dependencies. In this paper, we generalize this expressive parameterization to account for structural uncertainty in SCMs compatible with a PAG. Our main contributions are as follows:

1. We characterize the space of SCM-induced causal diagrams in the MEC of a PAG through two classes: *skeleton-matched diagrams* (SMDs) and *directed-backboned diagrams* (DBDs), and show that they preserve the full counterfactual expressiveness of SCMs in the MEC.
2. We translate these graphical characterizations into equivalent constraints over directed and bidirected adjacency masks, providing a parameterization of compatible causal diagrams amenable to optimization.
3. We introduce PAG-constrained neural causal models ($\mathcal{P}$-NCMs), which incorporate these constraints to parameterize a counterfactually expressive class of SCMs in the MEC.
4. We demonstrate how $\mathcal{P}$-NCMs enable causal inference under structural uncertainty by formulating counterfactual partial identification in PAGs as an optimization problem over this model class. We establish the theoretical soundness of this approach and empirically validate its performance on simulated causal bounding tasks.

2 Preliminaries

In this section, we introduce the key definitions and background results used throughout the paper. Detailed definitions are provided in Appendix A.

Structural Causal Models. We use *Structural Causal Models* (SCM) as the underlying semantical framework [18, 2]. An SCM $\mathcal{M}$ is a 4-tuple $\langle \mathbf{V}, \mathbf{U}, \mathcal{F}, P(\mathbf{u}) \rangle$, where $\mathbf{U}$ is a set of exogenous (latent) variables; $\mathbf{V}$ is a set of endogenous (observable) variables; $\mathcal{F}$ is a collection of functions such that each variable $V_i \in \mathbf{V}$ is determined by a function $f_i \in \mathcal{F}$. Each f_i is a mapping from a set of exogenous variables $U_i \subseteq \mathbf{U}$ and a set of endogenous variables $\mathbf{Pa}_i \subseteq \mathbf{V}$ to the domain of V_i. Uncertainty is encoded through a probability distribution over the exogenous variables, $P(\mathbf{u})$ [4]. Throughout this paper, we assume that domains of endogenous variables $\mathbf{V}$ are discrete and finite.

An SCM $\mathcal{M}$ induces a *causal diagram* $\mathcal{G}$ where $\mathbf{V}$ is the set of vertices, and there is a directed edge ($V_i \longrightarrow V_j$) for every $V_i, V_j \in \mathbf{V}$ if V_i appears as an argument of f_j, and a bidirected edge ($V_i \dashleftarrow\dashrightarrow V_j$) for every $V_i, V_j \in \mathbf{V}$ if functions f_i, f_j share some common $U \in \mathbf{U}$ as an argument, or the corresponding $U_i, U_j \in \mathbf{U}$ are correlated [4]. We assume the underlying model is recursive, i.e. it has no cyclic dependencies among the variables. An SCM $\mathcal{M}$ also induces all quantities within the *Pearl Causal Hierarchy* (PCH): *observational* ($\mathcal{L}_1$), *interventional* ($\mathcal{L}_2$) and *counterfactual* ($\mathcal{L}_3$). There is also a family of graphical causal models capturing the compatibility relationships between the PCH and causal diagrams: Bayesian networks (BNs) for $\mathcal{L}_1$, causal Bayesian networks (CBNs) for $\mathcal{L}_2$, and counterfactual Bayesian networks (CTFBNs) for $\mathcal{L}_3$ [17, 3, 4, 7].

MAGs, PAGs, and Graphical Notions. A *maximal ancestral graph* (MAG) encodes a set of causal diagrams over the same observed variables that entail identical conditional independence and ancestral relations [21]. A *partial ancestral graph* (PAG) represents a Markov equivalence class (MEC) of MAGs: it shares the same adjacencies as all MAGs in the class and displays precisely the invariant edge marks [22, 27]. PAGs can be learned using the FCI algorithm from conditional independence information from an observation distribution, or from an SCM-induced causal diagram [22, 27, 26]. In this work, we assume access to the true PAG compatible with the causal diagram.

In a graph, a *triple* $\langle X, Y, Z\rangle$ consists of three distinct variables such that X is adjacent to Y and Y is adjacent to Z. In a PAG, a triple is a *collider* if both incident edges have arrowheads at the middle node. It is a *(definite) non-collider* if at least one incident edge has a tail at the middle node, or if both incident edges have circle marks at the middle node and the two endpoint nodes are nonadjacent. In a causal diagram or MAG, an *inducing path* is either a single edge or a collider path whose non-endpoint nodes are ancestors of at least one of the endpoints. Given a graph, the *ordinary skeleton* of the graph is the undirected graph obtained by replacing every edge between two vertices with an undirected edge, regardless of its type or orientation. Two graphs are called *skeleton-matching* if their ordinary skeletons are identical.

Neural Causal Models. *Neural causal models* (NCMs) are a class of SCMs in which the structural functions are parameterized by neural networks and the exogenous variables follow fixed distributions [4, 24]. Prior work assumes a known causal diagram $\mathcal{G}$, and trains a $\mathcal{G}$-constrained NCM ($\mathcal{G}$-NCM) to match observational data and answer causal queries via direct evaluation or sampling. NCMs are proved to solve identification and estimation tasks across all three layers of the PCH.

Notations. We denote variables by capital letters, X, and values by small letters, x. Bold letters, $\mathbf{X}$ represent a set of variables and $\mathbf{x}$ a set of values. The domain of a variable X is denoted by $Val(X)$ or $Dom(X)$. $\mathbf{Y}_*$ or $\mathbf{Y}_\mathbf{x}$ denotes sets of counterfactual variables. We assume the domain of every variable is finite. We adopt standard graph-theoretic terminology, such as parent/child, ancestor/descendant, directed paths, and cycles, across all graph classes considered in this work.

3 Characterizing Causal Diagrams in PAG MEC

A PAG is typically interpreted as representing an equivalence class of MAGs, while each MAG in turn represents an equivalence class of causal diagrams. Existing characterizations of the equivalence class associated with a PAG therefore focus on invariant features of MAGs within its MEC [20, 1, 6]. However, unlike causal diagrams, neither PAGs nor MAGs admit a direct connection to SCMs, due to the uncertainty in functional relationships and latent confounding they encode. To address this gap, we provide a direct characterization of SCM-induced causal diagrams in the MEC of a PAG, bypassing the intermediate MAG representation. We then translate the conditions of this characterization into constraints over the adjacency masks of the causal diagrams.

3.1 PAG MEC Signatures and Graphical Characterization

First, we recall the ordered-triple characterization of MAG Markov equivalence introduced by [6]. In particular, [6, Def. 2] recursively characterize the invariant collider and non-collider triples of a MAG, and use these together with the skeleton to represent its MEC [6, Def. 3]. Importantly, the same MAG-to-MEC procedure can be applied directly to a PAG [6, Alg. 1]. We therefore apply the ordered-triple construction directly to a given PAG and refer to the resulting invariant features as its *MEC signature*.

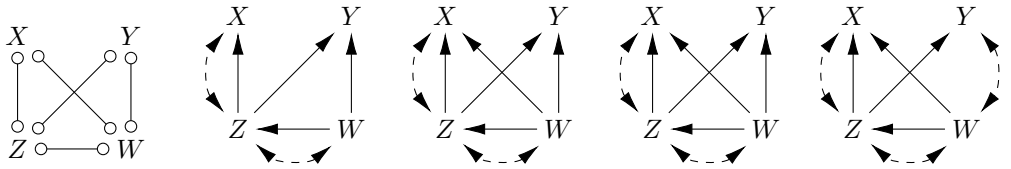
Definition 1 (Ordered Colliders and Non-Colliders in PAGs)**.** *Following [6, Def. 2], let $\mathfrak{C}_i$ (resp. $\mathfrak{D}_i$), with $i \geq 0$, be the set of collider (resp. definite non-collider) triples with order i in a PAG $\mathcal{P}$, defined recursively as:*

- *A triple $\langle a, b, c\rangle \in \mathfrak{C}_0$ (resp. $\mathfrak{D}_0$), if $a *\!\!-\!\!* b *\!\!-\!\!* c$ is an unshielded collider (resp. definite non-collider) in $\mathcal{P}$.*
- *A triple $\langle a, b, c\rangle \in \mathfrak{C}_i$ (resp. $\mathfrak{D}_i$), with $i \geq 1$, if $\langle a, b, c\rangle \notin \mathfrak{C}_{j<i}$ (resp. $\mathfrak{D}_{j<i}$), and*
  1. *$a *\!\!-\!\!* b *\!\!-\!\!* c$ is a collider (resp. definite non-collider) in $\mathcal{P}$, and*
  2. *$\exists q : \langle q, a, b\rangle \in \mathfrak{C}_{j<i}$ and $\langle q, a, c\rangle \in \mathfrak{D}_{k<i}$.*

Example 1 (Ordered Triples)**.** *Consider the PAG $\mathcal{P}$ in Fig. 1(a). Applying the definition above, its ordered collider and non-collider sets are $\mathfrak{C}_0 = \emptyset, \mathfrak{D}_0 = \{\langle X, Z, Y\rangle, \langle X, W, Y\rangle\}$, and there are no ordered triples of higher order.* ■

Definition 2 (PAG MEC Signature)**.** *Following the MEC representation of [6, Def. 3], The MEC signature of a PAG $\mathcal{P}$ is the triplet $\langle \mathfrak{S}, \mathfrak{C}, \mathfrak{D}\rangle$, where $\mathfrak{S}$ is the ordinary skeleton of $\mathcal{P}$, and $\mathfrak{C} = \bigcup_{i\geq 0} \mathfrak{C}_i$ and $\mathfrak{D} = \bigcup_{i\geq 0} \mathfrak{D}_i$ are its sets of ordered collider and ordered non-collider triples, respectively.*

The PAG MEC signature thus provides a semi-local characterization of the MAGs in an equivalence class through their skeleton and ordered triples. Specifically, the skeleton captures the global



(a) PAG, $\mathcal{P}$ (b) $\mathcal{G}_1 \in \mathcal{P}$ (c) SMD, $\mathcal{G}_2 \in \mathcal{P}$ (d) DBD, $\mathcal{G}_3 \in \mathcal{P}$ (e) $\mathcal{G}_4 \notin \mathcal{P}$

Figure 1: A PAG $\mathcal{P}$ (a) and representative causal diagrams. Diagrams (b)–(d) belong to its MEC: (b) is not skeleton-matched, (c) is skeleton-matched, and (d) is directed-backboned. Diagram (e) lies outside the MEC because the inducing path $\langle X, Z, W, Y \rangle$ violates the projected adjacency condition.

connectivity of the graph, while the ordered-triple constraints can be verified locally. However, this characterization does not transfer directly to causal diagrams because inducing paths can create a mismatch between their ordinary skeletons and the PAG skeleton. In particular, an adjacency in the PAG may be induced by a path in a causal diagram rather than represented by an actual edge. This creates two difficulties: (i) the ordinary skeleton of the causal diagram may differ from that of the PAG, and (ii) an ordered triple in the PAG may fail to appear as an actual triple in the causal diagram.

Example 2 (Triple Absent from Causal Diagram)**.** *Consider the PAG $\mathcal{P}$ in Fig. 1(a) and causal diagram $\mathcal{G}_1$ in its MEC, shown in Fig. 1(b). The triple $\langle X, W, Y \rangle$ belongs to $\mathfrak{D}_0$ in the MEC signature of $\mathcal{P}$. However, as X and W are not adjacent in $\mathcal{G}_1$, this triple is absent from $\mathcal{G}_1$. Thus, its collider or non-collider status cannot be determined directly from the local edge orientations in $\mathcal{G}_1$.* ■

To address this issue, we restrict attention to *skeleton-matched causal diagrams* (SMDs), which serve as an explicit-adjacency normal form: every adjacency encoded by the PAG is represented by an actual edge rather than only implicitly through an inducing path. This restriction does not sacrifice representational power, since every non-skeleton-matched causal diagram is a subgraph of some SMD (Lemma 7 in Appendix C).

Proposition 1. *[Characterization of Skeleton-Matched Causal Diagrams Represented by a PAG]*

Let $\mathcal{P}$ be a PAG with MEC signature $\langle \mathfrak{S}, \mathfrak{C}, \mathfrak{D} \rangle$. A graph $\mathcal{G}$ is a skeleton-matched causal diagram represented by $\mathcal{P}$ if and only if the following conditions hold:

1. ***Ordinary Adjacency:*** *Variables adjacent in $\mathcal{P}$ are adjacent in $\mathcal{G}$;*
2. ***Projected Adjacency:*** *Variables not adjacent in $\mathcal{P}$ are not connected by inducing paths in $\mathcal{G}$;*
3. ***Acyclicity:*** *$\mathcal{G}$ contains no directed cycles;*
4. ***Collider Triples:*** *Every triple in $\mathfrak{C}$ is a collider in $\mathcal{G}$;*
5. ***Non-collider Triples:*** *Every triple in $\mathfrak{D}$ is a non-collider in $\mathcal{G}$.*

Here, a triple $\langle X, Y, Z \rangle$ is a collider *in $\mathcal{G}$ if there exists an edge between X and Y and an edge between Y and Z that both have arrowheads at Y. Otherwise, the triple is a* non-collider.

Conditions (1) and (2) together ensures that the ordinary skeleton of the causal diagram matches that of the PAG, while Condition (2) additionally rules out non-trivial inducing paths that would create extra adjacency between variables that are non-adjacent in the PAG.

Given a PAG, there exists a set of special representative MAGs, called *loyal equivalent graphs* (LEGs), that contain the fewest bidirected edges among all MAGs in the equivalence class [28]. These LEGs motivate a directed backbone for a subset of SMDs, where the ordinary adjacency condition is satisfied by enforcing a directed adjacency between variables that are not connected by a bidirected edge in the PAG. We refer to this subset of causal diagrams in the MEC of the PAG as *directed-backboned causal diagrams* (DBDs), formally defined below.

Proposition 2. *[Characterization of Directed-Backboned Causal Diagrams Represented by a PAG]*

Let $\mathcal{P}$ be a PAG with MEC signature $\langle \mathfrak{S}, \mathfrak{C}, \mathfrak{D} \rangle$. A graph $\mathcal{G}$ is a directed-backboned causal diagram represented by $\mathcal{P}$ if and only if:

1. ***Directed Backbone:*** *Every adjacency in $\mathcal{P}$ that is not bidirected is represented by a directed edge or bow in $\mathcal{G}$, and every bidirected adjacency in $\mathcal{P}$ is represented by a bidirected edge in $\mathcal{G}$;*
2. *$\mathcal{G}$ satisfies Conditions (2–5) in Proposition 1.*

Comparing the two propositions, the directed-backbone condition strengthens the ordinary adjacency condition by requiring specific edge types rather than mere adjacency. Since the remaining conditions are shared, the corresponding spaces of SCMs satisfy $\Omega^{DBD}(\mathcal{P}) \subseteq \Omega^{SMD}(\mathcal{P}) \subseteq \Omega(\mathcal{P})$, where these spaces consist of SCMs inducing DBDs, SMDs, and arbitrary causal diagrams in the MEC of $\mathcal{P}$, respectively. Despite this restriction, DBDs retain the full expressiveness of the MEC, as formalized by the following theorem.

Theorem 1. *[Expressiveness of MEC SMDs and DBDs] For every SCM $\mathcal{M}$ inducing a causal diagram $\mathcal{G}$ in the MEC of a PAG $\mathcal{P}$, there exist SCMs $\mathcal{M}^{SMD}$ and $\mathcal{M}^{DBD}$ inducing causal diagrams $\mathcal{G}^{SMD}$ and $\mathcal{G}^{DBD}$ in the same MEC, respectively, such that $\mathbf{P}^{\mathcal{L}_3}(\mathcal{M}) = \mathbf{P}^{\mathcal{L}_3}(\mathcal{M}^{SMD}) = \mathbf{P}^{\mathcal{L}_3}(\mathcal{M}^{DBD})$.*

Example 3 (SMD vs. DBD)**.** *Consider the PAG $\mathcal{P}$ in Fig. 1(a) and two causal diagrams $\mathcal{G}_2$ and $\mathcal{G}_3$ in its MEC, shown in Fig. 1(c) and (d), respectively. Both match the ordinary skeleton of $\mathcal{P}$, but differ in how their adjacencies are realized. In $\mathcal{G}_3$, every non-bidirected adjacency in $\mathcal{P}$ is realized by a directed edge or a bow, satisfying the directed-backbone condition in Prop. 2. In contrast, $\mathcal{G}_2$ violates this condition because the adjacency $Z \circ\!\!-\!\!\circ W$ is not realized by a directed edge. Thus, $\mathcal{G}_2$ is an SMD but not a DBD, whereas $\mathcal{G}_3$ is both.* ■

3.2 Adjacency Mask Constraints

The characterizations above provide a direct way to describe the space of SCMs compatible with a PAG through causal diagrams they induce. Specifically, each condition in the characterizations admits an equivalent constraint formulation over the adjacency masks $\langle \mathbf{m}^D, \mathbf{m}^B \rangle$.

Definition 3 (Causal Diagram Adjacency Masks)**.** *Given a causal diagram $\mathcal{G}$ over endogenous variables $\mathbf{V}$ with $|\mathbf{V}| = n$, its adjacency masks $\mathbf{m}^D \in [0,1]^{n\times n}, \mathbf{m}^B \in [0,1]^{n\times n}$ are defined as: $m^D_{ij} > 0 \iff V_i \to V_j \in \mathcal{G}$ and $m^B_{ij} > 0 \iff V_i \leftrightarrow V_j \in \mathcal{G}$.*

All conditions in Propositions 1 and 2 can be expressed as flexible constraints over the mask parameters. However, since the PAG itself provides sufficient structural information to fix certain mask entries, we focus directly on the conditions in Proposition 2 and use a fixed-mask construction to encode part of the required graphical structure by design.

Definition 4 (PAG-Induced Adjacency Mask Parameterization)**.** *Given a PAG $\mathcal{P}$ over variables $\mathbf{V}$, the PAG-induced parameterization of the adjacency masks $\langle \mathbf{m}^D, \mathbf{m}^B \rangle$ is defined as:*

- ***Directed adjacency:*** $m^D_{ij} = \begin{cases} 1, & \text{if } V_i \to V_j \in \mathcal{P} \text{ or } V_i \circ\!\!\to V_j \in \mathcal{P}, \\ 0, & \text{if } i = j,\ V_i \leftarrow\!\!* V_j \in \mathcal{P},\ \text{or } V_i \not\to V_j \text{ in } \mathcal{P}, \\ 1 - m^D_{ji}, & \text{if } V_i \circ\!\!-\!\!\circ V_j \in \mathcal{P}; \end{cases}$
- ***Bidirected adjacency:*** $m^B_{ij} = \begin{cases} 1, & \text{if } V_i \leftrightarrow V_j \in \mathcal{P}, \\ 0, & \text{if } i = j,\ V_i \overset{v}{\to} V_j \in \mathcal{P},\ V_j \overset{v}{\to} V_i \in \mathcal{P},\ \text{or } V_i \not\to V_j \text{ in } \mathcal{P}. \end{cases}$

The mask parameters are classified into three types: *fixed*, *free*, and *derived*. Fixed entries encode edge presence, absence, and orientations imposed by the PAG; free entries are trainable during optimization; and derived entries are determined by the free parameters to enforce adjacency constraints.

Example 4 (PAG-Induced Adjacency Mask Parameterization)**.** *Consider the PAG $\mathcal{P}$ in Fig. 1(a). Its induced mask parameters will be instantiated as*

$$\mathbf{m}^D = \begin{bmatrix} 0 & m^D_{12} & m^D_{13} & 0 \\ 1-m^D_{12} & 0 & m^D_{23} & m^D_{24} \\ 1-m^D_{13} & 1-m^D_{23} & 0 & m^D_{34} \\ 0 & 1-m^D_{24} & 1-m^D_{34} & 0 \end{bmatrix} \qquad \mathbf{m}^B = \begin{bmatrix} 0 & m^B_{12} & m^B_{13} & 0 \\ m^B_{12} & 0 & m^B_{23} & m^B_{24} \\ m^B_{13} & m^B_{23} & 0 & m^B_{34} \\ 0 & m^B_{24} & m^B_{34} & 0 \end{bmatrix} \tag{1}$$

where gray entries 0 and 1 denote fixed parameters, blue entries such as m^D_{12} denote free parameters, and red entries such as $1 - m^D_{12}$ denote derived (constrained) parameters. In other words, only the free parameters are independently trainable, subject to other constraints over the masks. ■

Proposition 3. *[Satisfaction of Directed-Backbone and Collider Conditions] Let $\mathcal{P}$ be a PAG and $\mathcal{G}$ a causal diagram whose adjacency masks $\langle \mathbf{m}^D, \mathbf{m}^B \rangle$ follow the PAG-induced parameterization in Definition 4. Then $\mathcal{G}$ satisfies the directed-backbone and collider conditions in Proposition 2.*

With the directed-backbone and collider conditions satisfied by construction, we next formulate the non-collider and projected adjacency conditions as equality constraints over the PAG MEC signature and the adjacency masks.

Definition 5 (Non-Collider Triples Constraint)**.** *Given a PAG $\mathcal{P}$ with MEC signature $\langle \mathfrak{S}, \mathfrak{C}, \mathfrak{D} \rangle$ and a causal diagram $\mathcal{G}$ with adjacency masks $(\mathbf{m}^D, \mathbf{m}^B)$, the non-collider constraint in $\mathcal{G}$ with respect to $\mathcal{P}$ is defined as*

$$nc^{\mathcal{P}}(\mathbf{m}^D, \mathbf{m}^B, \mathfrak{D}) := \langle \mathbf{D}, \mathbf{H} \rangle, \tag{2}$$

where $\mathbf{W} := \mathbf{m}^D + \mathbf{m}^B - \mathbf{m}^D \odot \mathbf{m}^B$ encodes the presence of arrowheads, $\mathbf{H} \in \mathbb{R}^{n \times n \times n}$ is defined by $H_{ijk} := W_{ij} W_{kj}$, and $\mathbf{D} \in \{0,1\}^{n \times n \times n}$ is defined by $D_{ijk} = 1 \iff \langle V_i, V_j, V_k \rangle \in \mathfrak{D}$.

Proposition 4. *[Correctness of Non-Collider Triples Constraint] Given a PAG $\mathcal{P}$ with MEC signature $\langle \mathfrak{S}, \mathfrak{C}, \mathfrak{D} \rangle$ and a causal diagram $\mathcal{G}$ with adjacency masks $(\mathbf{m}^D, \mathbf{m}^B)$, $\mathcal{G}$ satisfies the non-collider triples condition in Propositions 1 and 2 if and only if $nc^{\mathcal{P}}(\mathbf{m}^D, \mathbf{m}^B, \mathfrak{D}) = 0$.*

Example 5 (Non-Collider Triples Constraint)**.** *Consider the PAG $\mathcal{P}$ in Fig. 1(a), with ordered triples given in Example 1, and the causal diagram $\mathcal{G}_3$ in Fig. 1(d) with variable ordering $(X, Z, W, Y) = (1, 2, 3, 4)$. To evaluate the non-collider constraint in Prop. 1 and 2, the tensor $\mathbf{D}$ has exactly two nonzero entries, $D_{124} = D_{134} = 1$, while the slices of $\mathbf{H}$ are given by*

$$\mathbf{H}_{:,1,:} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \mathbf{H}_{:,2,:} = \begin{bmatrix} 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \mathbf{H}_{:,3,:} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \mathbf{H}_{:,4,:} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

Since $H_{124} = H_{134} = 0$, the non-collider penalty gives $nc^{\mathcal{P}} = \langle \mathbf{D}, \mathbf{H} \rangle = D_{124} H_{124} + D_{134} H_{134} = 1 \cdot 0 + 1 \cdot 0 = 0$. Hence, $\mathcal{G}_3$ satisfies the non-collider constraint with respect to $\mathcal{P}$ in Prop. 1 and 2. ■

Definition 6 (Projected Adjacency Constraint)**.** *Given a PAG $\mathcal{P}$ with MEC signature $\langle \mathfrak{S}, \mathfrak{C}, \mathfrak{D} \rangle$ and a causal diagram $\mathcal{G}$ with adjacency masks $(\mathbf{m}^D, \mathbf{m}^B)$, with $\mathbf{m}^D$ nilpotent, the inducing path constraint in $\mathcal{G}$ with respect to $\mathcal{P}$ is defined as*

$$p^{\mathcal{P}}(\mathbf{m}^D, \mathbf{m}^B, \mathfrak{S}) = \langle \mathbf{P}, \mathfrak{S}^- \rangle, \tag{3}$$

where $\mathfrak{S}^- := \mathbf{1}\mathbf{1}^\top - \mathfrak{S} - I$ denotes the complement of the skeleton, i.e., pairs that must not admit an inducing path. $\mathbf{P}$ encodes inducing-path connectivity, with $\mathbf{P}_{ij} := \log\Big(1 + \sum_{k=1}^{K} \big[(\mathbf{W}^{ij})^k\big]_{ij}\Big)$, where $K \geq n-1$, and $\mathbf{P}_{ij} > 0$ if and only if there exists an inducing path between V_i and V_j. For each pair (i, j), the step matrix $\mathbf{W}^{ij} \in \mathbb{R}^{n \times n}$ is defined as $\mathbf{W}^{ij} = \mathbf{S} \odot \big(\boldsymbol{\alpha}^{ij} (\boldsymbol{\alpha}^{ij})^\top\big) \odot \mathrm{Tail}^{ij} \odot \mathrm{Head}^{ij}$, via components defined as follows, where I is the identity and $\mathbf{1}$ is the all-ones vector:

- *Ancestry:* $T := e^{\mathbf{m}^D} - I, \qquad \mathrm{Anc} := (1 - e^{-\lambda T}) \odot (1 - I) + I, \lambda > 0.$
- *Adjacency:* $\mathbf{S} := 1 - (1 - \mathbf{m}^D) \odot (1 - (\mathbf{m}^D)^\top) \odot (1 - \mathbf{m}^B)$.
- *Arrowheads:* $\mathrm{ArrAt} := 1 - (1 - \mathbf{m}^D) \odot (1 - \mathbf{m}^B)$.
- *Pairwise anchor vector:* $\boldsymbol{\alpha}^{ij} = \mathbf{1} - (\mathbf{1} - e_i) \odot (\mathbf{1} - e_j) \odot (\mathbf{1} - \mathrm{Anc}_{:i}) \odot (\mathbf{1} - \mathrm{Anc}_{:j})$.
- *Endpoint indicator:* $\boldsymbol{\eta}^{ij} := e_i + e_j$.
- *Endpoint masks:* $E^{ij}_{\mathrm{col}} := \mathbf{1}(\boldsymbol{\eta}^{ij})^\top, \qquad E^{ij}_{\mathrm{row}} := \boldsymbol{\eta}^{ij} \mathbf{1}^\top$.
- *Edgemarks:* $\mathrm{Head}^{ij} := E^{ij}_{\mathrm{col}} + (1 - E^{ij}_{\mathrm{col}}) \odot \mathrm{ArrAt}, \mathrm{Tail}^{ij} := E^{ij}_{\mathrm{row}} + (1 - E^{ij}_{\mathrm{row}}) \odot \mathrm{ArrAt}^\top$.

Proposition 5. *[Correctness of Projected Adjacency Constraint] Given a $\mathcal{P}$ with MEC signature $\langle \mathfrak{S}, \mathfrak{C}, \mathfrak{D} \rangle$ and a causal diagram $\mathcal{G}$ with adjacency masks $(\mathbf{m}^D, \mathbf{m}^B)$, $\mathcal{G}$ satisfies the projected adjacency condition in Propositions 1 and 2 if and only if $p^{\mathcal{P}}(\mathbf{m}^D, \mathbf{m}^B, \mathfrak{S}) = 0$.*

The projected adjacency constraint accounts for trivial inducing paths consisting of a single edge. Under the PAG-induced adjacency mask parameterization in Def. 4, such paths are already excluded by fixing the adjacency masks of PAG-nonadjacent pairs to zero. We nevertheless retain the general formulation to accommodate settings without this fixed parameterization. For a PAG with $|\mathbf{V}| = n$ and a maximum path length of k, the computational complexity is $O(kn^3)$ per endpoint pair, yielding a worst-case complexity of $O(kn^5)$ over all non-adjacent pairs. In practice, the cost can be substantially lower for sparse PAGs, where fewer endpoint pairs and candidate paths need to be considered.

Example 6 (Projected Adjacency Constraint)**.** *Consider the PAG $\mathcal{P}$ in Fig. 1(a) and the causal diagrams $\mathcal{G}_3$ and $\mathcal{G}_4$ in Fig. 1(d) and (e), respectively. Setting $\lambda = 1 - e^{-5}$, we illustrate how the*

projected adjacency constraint distinguishes these two diagrams. Since (X, Y) is the only non-adjacent pair in $\mathcal{P}$, the non-adjacency matrix $\mathfrak{S}^-$ has exactly two nonzero entries, $\mathfrak{S}^-_{14} = \mathfrak{S}^-_{41} = 1$. The corresponding projected adjacency matrices are

$$\mathbf{P}(\mathcal{G}_3) \approx \begin{bmatrix} 0 & 1.3829 & 2.3856 & 0 \\ 1.3829 & 0 & 1.0986 & 1.3829 \\ 2.3856 & 1.0986 & 0 & 1.3829 \\ 0 & 1.3829 & 1.3829 & 0 \end{bmatrix}, \quad \mathbf{P}(\mathcal{G}_4) \approx \begin{bmatrix} 0 & 1.3829 & 2.3856 & 0.6864 \\ 1.3829 & 0 & 1.0986 & 1.0986 \\ 2.3856 & 1.0986 & 0 & 1.3829 \\ 0.6864 & 1.0986 & 1.3829 & 0 \end{bmatrix}.$$

For $\mathcal{G}_3$, the projected adjacency penalty evaluates to $p^{\mathcal{P}}(\mathbf{m}^D, \mathbf{m}^B, \mathfrak{S}) = \langle \mathbf{P}(\mathcal{G}_3), \mathfrak{S}^- \rangle = P_{14} + P_{41} = 0$. Thus, $\mathcal{G}_3$ satisfies the projected adjacency constraint in Props. 1 and 2. In contrast, $\mathcal{G}_4$ contains the inducing path $\langle X, Z, W, Y \rangle$ between the non-adjacent pair (X, Y), yielding $p^{\mathcal{P}}(\mathbf{m}^D, \mathbf{m}^B, \mathfrak{S}) = \langle \mathbf{P}(\mathcal{G}_4), \mathfrak{S}^- \rangle = P_{14} + P_{41} \approx 1.3728 > 0$. Hence, $\mathcal{G}_4$ violates the projected adjacency constraint and lies outside the MEC represented by $\mathcal{P}$. ■

4 PAG-Constrained Neural Causal Models

The adjacency mask constraints developed in the previous section can be incorporated into causal models to capture structural uncertainty within a PAG MEC. In this section, we introduce *$\mathcal{P}$-constrained neural causal models* ($\mathcal{P}$-NCMs) and establish their expressiveness in representing the counterfactual distributions induced by any SCM in the MEC. We begin by defining the admissible exogenous variables and their associated masks, which allow for a shared latent confounder for each bidirected clique to preserve counterfactual expressiveness.

Definition 7 (PAG-Induced Exogenous Variables)**.** *Let $\mathcal{P}$ be a PAG over $\mathbf{V}$ with MEC signature $\langle \mathfrak{S}, \mathfrak{C}, \mathfrak{D} \rangle$, and let $\mathcal{P}'$ be obtained from $\mathcal{P}$ by removing all visible edges. Define*

$$\mathcal{Q}(\mathcal{P}) := \big\{\{V\} : V \in \mathbf{V}\big\} \cup \left\{ S \subseteq \mathbf{V} : \begin{array}{l} |S| \geq 2,\ S \text{ is a clique in } \mathcal{P}', \\ \nexists\, \langle A, B, C \rangle \in \mathfrak{D} \text{ such that } \{A, B, C\} \subseteq S \end{array} \right\}. \tag{4}$$

The set of exogenous variables induced by $\mathcal{P}$ is $\mathbf{U}(\mathcal{P}) := \{U_S : S \in \mathcal{Q}(\mathcal{P})\}$.

Definition 8 (Bidirected Clique Mask)**.** *Let $\mathcal{Q}(\mathcal{P})$ be the admissible subsets in Definition 7. Given $\mathbf{m}^B$, the bidirected clique mask $\mathbf{m}^C \in [0,1]^{|\mathcal{Q}(\mathcal{P})| \times n}$ is defined by*

$$m^C_{Si} = \mathbb{I}[V_i \in S] \prod_{\substack{V_j, V_k \in S \\ j<k}} m^B_{jk}, \quad S \in \mathcal{Q}(\mathcal{P}), \tag{5}$$

where an empty product is defined as 1.

Example 7 (PAG-Induced Exogenous Variables and Bidirected Clique Mask)**.** *Consider the PAG $\mathcal{P}$ in Fig. 1(a). The set of exogenous variables induced by $\mathcal{P}$ with Def. 7 will be $\mathbf{U} = \{U_X, U_Z, U_W, U_Y, U_{XZ}, U_{XW}, U_{ZW}, U_{WY}, U_{ZY}, U_{XZW}, U_{ZWY}\}$. The bidirected clique mask is then fully derived from $\mathbf{U}$ and $\mathbf{m}^B$ from Example 4. Refer to Appendix B for details of $\mathbf{m}^C$.* ■

Def. 7 introduces a sufficiently rich set of exogenous variables to capture shared latent confounding within bidirected cliques, which is essential for the expressiveness of NCMs [24]. For a PAG with $|\mathbf{V}| = n$ and maximal bidirected clique size k, the number of exogenous variables scales as $O(n^k)$ for fixed k. However, this complexity can be substantially reduced for sparse PAGs, where the number of admissible bidirected cliques is much smaller.

Definition 9 ($\mathcal{P}$-Constrained Neural Causal Model ($\mathcal{P}$-NCM))**.** *Let $\mathcal{P}$ be a PAG over variables $\mathbf{V} = \{V_1, \ldots, V_n\}$ with MEC signature $\langle \mathfrak{S}, \mathfrak{C}, \mathfrak{D} \rangle$. A $\mathcal{P}$-constrained neural causal model ($\mathcal{P}$-NCM) $\mathcal{N}(\boldsymbol{\theta}, \mathbf{m}^D, \mathbf{m}^B, \mathbf{m}^C)$ is an SCM $\langle \mathbf{V}, \mathbf{U}, P(\mathbf{U}), \mathcal{F} \rangle$ with parameters $\boldsymbol{\theta} = \{\theta_i : V_i \in \mathbf{V}\}$, $\mathbf{m}^D \in [0,1]^{n \times n}$, $\mathbf{m}^B \in [0,1]^{n \times n}$, $\mathbf{m}^C \in [0,1]^{|\mathbf{U}| \times n}$, defined as follows:*

- *$\mathbf{V}$ is a set of discrete endogenous variables;*
- *$\mathbf{U}$ is a set of exogenous variables constructed via Def. 7;*
- *$\mathcal{F} = \{f_i : V_i \in \mathbf{V}\}$ is a collection of structural functions, where*

$$V_i \leftarrow f_i\big(\mathbf{V}, \mathbf{U}; \mathbf{m}^D_{:,i}, \mathbf{m}^C_{:,i}, \theta_i\big).$$

Each f_i is a neural network parameterized by θ_i, satisfying:

(i) Mask invariance: f_i *is invariant to inputs with zero masks.*
(ii) Expressiveness: f_i *can represent any function over the inputs with positive masks.*

- $P(\mathbf{U})$ *is such that each* $U \in \mathbf{U}$ *is independently distributed as* $\text{Uniform}(0, 1)$.

The parameters are constrained as follows:

- $\mathbf{m}^D, \mathbf{m}^B$ *follow the PAG-induced parameterization in Definition 4;*
- $\mathbf{m}^C$ *is the bidirected clique mask constructed from Def. 8 using* $\mathbf{U}$ *and* $\mathbf{m}^B$*;*
- *Acyclicity constraint:* $h(\mathbf{m}^D) = 0$*, where* h *is a sound acyclicity constraint;*
- *Non-collider constraint:* $nc^{\mathcal{P}}(\mathbf{m}^D, \mathbf{m}^B, \mathfrak{D}) = 0$*;*
- *Inducing-path constraint:* $p^{\mathcal{P}}(\mathbf{m}^D, \mathbf{m}^B, \mathfrak{S}) = 0$.

The constraints on the adjacency mask parameters are equivalent to the graphical conditions in Prop. 2, ensuring that the resulting masks induce a DBD in the MEC represented by $\mathcal{P}$. In particular, acyclicity can be enforced using standard formulations such as NOTEARS [30]. The causal diagram induced by a $\mathcal{P}$-NCM is constructed as follows.

Definition 10 ($\mathcal{P}$-NCM Causal Diagram)**.** *Given a* $\mathcal{P}$*-NCM* $\mathcal{N}(\boldsymbol{\theta}, \mathbf{m}^D, \mathbf{m}^B, \mathbf{m}^C) = \langle \mathbf{V}, \mathbf{U}, P(\mathbf{U}), \mathcal{F} \rangle$*, its induced causal diagram* $\mathcal{G}$ *is constructed by (1) adding a vertex for each endogenous variable in* $\mathbf{V}$*; (2) add a directed edge* $V_i \to V_j$ *whenever* $\mathbf{m}^D_{ij} > 0$*; and (3) add a bidirected edge* $V_i \leftrightarrow V_j$ *whenever* $\mathbf{m}^B_{ij} > 0$.

Example 8 ($\mathcal{P}$-NCM Causal Diagram)**.** *Consider the PAG* $\mathcal{P}$ *in Fig. 1(a) and the corresponding* $\mathcal{P}$*-NCM* $\mathcal{N}$ *with the mask parameterization given in Example 4. Assigning the free mask parameters in Eq.* (1) *yields*

$$\mathbf{m}^D = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix} \qquad \mathbf{m}^B = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

The resulting clique mask $\mathbf{m}^C$ *has nonzero rows only for the singleton sets and the pairs* $\{X, Z\}$ *and* $\{Z, W\}$*. By Def. 10, these masks induce the causal diagram* $\mathcal{G}_3$ *in Fig. 1(d).* ■

A $\mathcal{P}$-NCM also induces a collection of counterfactual distributions, defined below.

Definition 11 ($\mathcal{P}$-NCM $\mathcal{L}_3$ Valuation)**.** *Given a* $\mathcal{P}$*-NCM* $\mathcal{N}(\boldsymbol{\theta}, \mathbf{m}^D, \mathbf{m}^B, \mathbf{m}^C) = \langle \mathbf{V}, \mathbf{U}, P(\mathbf{U}), \mathcal{F} \rangle$*, for any subsets* $\mathbf{Y}, \mathbf{Z}, ..., \mathbf{X}, \mathbf{W} \subseteq \mathbf{V}$ *and values* $\mathbf{x}, \mathbf{y}, \mathbf{z}, \mathbf{w}$*, the induced counterfactual distribution is*

$$\mathbf{P}^{\mathcal{N}, \mathbf{m}^D, \mathbf{m}^C}(\mathbf{y}_{\mathbf{x}}, ..., \mathbf{z}_{\mathbf{w}}) = \int \mathbf{1}[\mathbf{Y}_{\mathbf{x}}(\mathbf{u}; \mathbf{m}^{\mathbf{C}}_{:,\mathbf{Y}}) = \mathbf{y}, ..., \mathbf{Z}_{\mathbf{w}}(\mathbf{u}; \mathbf{m}^{\mathbf{C}}_{:,\mathbf{Z}}) = \mathbf{z}] dP(\mathbf{u}) \tag{6}$$

where $\mathbf{Y}_{\mathbf{x}}$ *and* $\mathbf{Z}_{\mathbf{w}}$ *are obtained by evaluating the structural equations* $\{f_i\}_{i=1}^n$ *in the corresponding submodels, with each* $X_i \in \mathbf{X}$ *replaced by* x_i *and each* $W_i \in \mathbf{W}$ *replaced by* w_i*. The evaluation proceeds in a topological order consistent with* $\mathbf{m}^D$*. We denote the collection of all such counterfactual distributions induced by* $\mathcal{N}$ *as* $\mathbf{P}^{\mathcal{L}_3}(\mathcal{N})$.

The following proposition shows that any $\mathcal{P}$-NCM is equivalent to an SCM in terms of both the induced causal diagram and the collection of counterfactual distributions.

Proposition 6. *[*$\mathcal{P}$*-NCM – SCM Equivalence] Given a* $\mathcal{P}$*-NCM* $\mathcal{N}(\boldsymbol{\theta}, \mathbf{m}^D, \mathbf{m}^B, \mathbf{m}^C) = \langle \mathbf{V}, \mathbf{U}, P(\mathbf{U}), \mathcal{F} \rangle$ *with DBD* $\mathcal{G}$ *(Def. 10) and counterfactual distributions* $\mathbf{P}^{\mathcal{L}_3}(\mathcal{N})$ *(Def. 11), there exists an SCM* $\mathcal{M}$ *inducing the same causal diagram* $\mathcal{G}$ *with* $\mathbf{P}^{\mathcal{L}_3}(\mathcal{M}) = \mathbf{P}^{\mathcal{L}_3}(\mathcal{N})$.

By varying the adjacency mask parameters, a $\mathcal{P}$-NCM can induce different causal diagrams within the MEC represented by $\mathcal{P}$. Importantly, the resulting model class $\Theta(\mathcal{P})$ preserves the full counterfactual expressiveness of the SCM space $\Omega(\mathcal{P})$ associated with this MEC. Specifically, for every SCM in $\Omega(\mathcal{P})$, there exists a $\mathcal{P}$-NCM in $\Theta(\mathcal{P})$ that induces the same counterfactual distributions, as illustrated in Fig. 2 and formalized below.

Theorem 2. *[Expressiveness of* $\mathcal{P}$*-NCM] For every SCM* $\mathcal{M}$ *inducing a causal diagram* $\mathcal{G}$ *in the MEC of a PAG* $\mathcal{P}$*, there exists a* $\mathcal{P}$*-NCM* $\mathcal{N}(\boldsymbol{\theta}, \mathbf{m}^D, \mathbf{m}^B, \mathbf{m}^C)$ *such that* $\mathbf{P}^{\mathcal{L}_3}(\mathcal{N}) = \mathbf{P}^{\mathcal{L}_3}(\mathcal{M})$.

5 Neural Partial Identification with $\mathcal{P}$-NCM

In this section, we demonstrate the usefulness of $\mathcal{P}$-NCMs for causal inference under structural uncertainty by applying them to the problem of partial identification given a PAG. The goal of partial

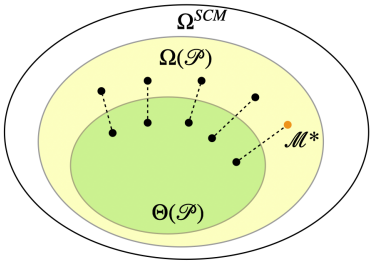


Figure 2: $\mathcal{P}$-NCM Expressiveness: each SCM $\mathcal{M}^* \in \Omega(\mathcal{P})$ has a $\mathcal{P}$-NCM $\in \Theta(\mathcal{P})$ matching on $\mathbf{P}^{\mathcal{L}_3}$.

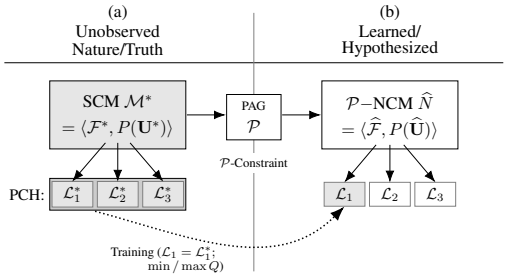


Figure 3: Left: true SCM $\mathcal{M}^* \in \Omega(\mathcal{P})$ that induces PCH's three layers. Right: a $\mathcal{P}$-NCM $\mathcal{N}$ constrained by inductive bias in $\mathcal{P}$ and matching $\mathcal{M}^*$ on $\mathcal{L}_1 = P(\mathbf{V})$ while minimizing or maximizing a target query Q through training.

Algorithm 1 PAG Neural Counterfactual Bounding

1: **Input:** a query $Q = P(\mathbf{y}_*|\mathbf{x}_*)$, observed datasets $P(\mathbf{V})$, a PAG $\mathcal{P}$ over $\mathbf{V}$
2: **Output:** $P^{\hat{\mathcal{N}}(\mathbf{\Theta}_{min})}(\mathbf{y}_*|\mathbf{x}_*)$, $P^{\hat{\mathcal{N}}(\mathbf{\Theta}_{max})}(\mathbf{y}_*|\mathbf{x}_*)$
3: $\hat{\mathcal{N}}(\mathbf{\Theta}) \leftarrow \mathcal{P}$-NCM$(\mathcal{P}, \mathbf{V})$ with parameters $\mathbf{\Theta} = (\boldsymbol{\theta}, \mathbf{m}^D, \mathbf{m}^B, \mathbf{m}^C)$ from Def. 9.
4: $\mathbf{\Theta}_{min} \leftarrow \arg\min_{\mathbf{\Theta}} P^{\hat{\mathcal{N}}(\boldsymbol{\theta})}(\mathbf{y}_*|\mathbf{x}_*)$ s.t. $P^{\hat{\mathcal{N}}(\mathbf{\Theta})}(\mathbf{V}) = P(\mathbf{V})$.
5: $\mathbf{\Theta}_{max} \leftarrow \arg\max_{\mathbf{\Theta}} P^{\hat{\mathcal{N}}(\mathbf{\Theta})}(\mathbf{y}_*|\mathbf{x}_*)$ s.t. $P^{\hat{\mathcal{N}}(\mathbf{\Theta})}(\mathbf{V}) = P(\mathbf{V})$.
6: **Return** $P^{\hat{\mathcal{N}}(\mathbf{\Theta}_{min})}(\mathbf{y}_*|\mathbf{x}_*)$, $P^{\hat{\mathcal{N}}(\mathbf{\Theta}_{max})}(\mathbf{y}_*|\mathbf{x}_*)$

identification in this setting is to compute bounds on a counterfactual query over all SCMs whose induced causal diagrams lie in the MEC represented by the PAG. To this end, we first extend the notion of optimal counterfactual bounds from causal diagrams [29, Def. 2.1] to PAGs.

Definition 12 (Optimal Counterfactual Bound (PAG))**.** *For a PAG $\mathcal{P}$ and an observational distribution $P(\mathbf{V})$ faithful to $\mathcal{P}$, the optimal bound $[l, r]$ over a counterfactual probability $P(\mathbf{y}_\mathbf{x}, ..., \mathbf{z}_\mathbf{w})$ is defined as, respectively, the minimum and maximum of the following optimization problem:*

$$\min / \max_{\mathcal{M} \in \Omega(\mathcal{P})} P^{\mathcal{M}}(\mathbf{y}_\mathbf{x}, \ldots, \mathbf{z}_\mathbf{w}) \quad \textit{s.t.} \quad P^{\mathcal{M}}(\mathbf{V}) = P(\mathbf{V}) \tag{7}$$

where $\Omega(\mathcal{P})$ is the set of all SCMs inducing a graph $\mathcal{G}$ in the MEC represented by $\mathcal{P}$.

Given the expressiveness of $\mathcal{P}$-NCMs in encoding counterfactual distributions from SCMs in $\Omega(\mathcal{P})$, the partial identification problem can be formulated equivalently using these models, as defined below.

Definition 13 (Optimal Neural Counterfactual Bound (PAG))**.** *For a PAG $\mathcal{P}$ and an observational distribution $P(\mathbf{V})$ faithful to $\mathcal{P}$, the optimal neural bound $[l, r]$ over a counterfactual probability $P(\mathbf{y}_\mathbf{x}, ..., \mathbf{z}_\mathbf{w})$ is defined as, respectively, the minimum and maximum of the following optimization problem:*

$$\min / \max_{\mathcal{N} \in \Theta(\mathcal{P})} \quad P^{\mathcal{N}(\boldsymbol{\theta}, \mathbf{m}^D, \mathbf{m}^B, \mathbf{m}^C)}(\mathbf{y}_\mathbf{x}, ..., \mathbf{z}_\mathbf{w}) \quad \textit{s.t.} \quad P^{\mathcal{N}(\boldsymbol{\theta}, \mathbf{m}^D, \mathbf{m}^B, \mathbf{m}^C)}(\mathbf{V}) = P(\mathbf{V}), \tag{8}$$

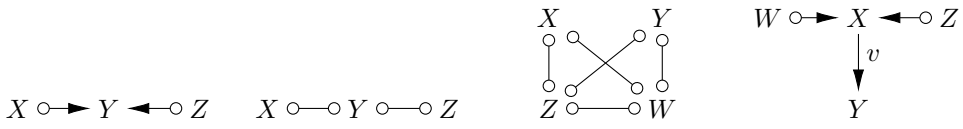
where $\Theta(\mathcal{P})$ is the set of all $\mathcal{P}$-NCMs, with mask parameters satisfying constraints in Def. 9.

The key idea is illustrated in Fig.3: we search for $\mathcal{P}$-NCMs $\in \Theta(\mathcal{P})$ matching the observational data from a true, unobserved SCM $\mathcal{M}^* \in \Omega(\mathcal{P})$, while minimizing and maximizing a target query. This optimization problem can equivalently be expressed as Alg. 1.

Theorem 3. *[Correctness of PAG Neural Partial ID] The true optimal counterfactual bounds (Def. 12) and the optimal neural counterfactual bounds (Def. 13) coincide.*

5.1 Experiments

We evaluate counterfactual partial identification using Alg. 1 on the PAGs shown in Fig. 4. In practice, the $\mathcal{P}$-NCM parameters are learned by optimizing an objective combining the observational negative



(a) Collider (b) Non-Collider (c) Diamond (d) Visible-Edge

Figure 4: PAGs tested in simulated partial identification experiments.

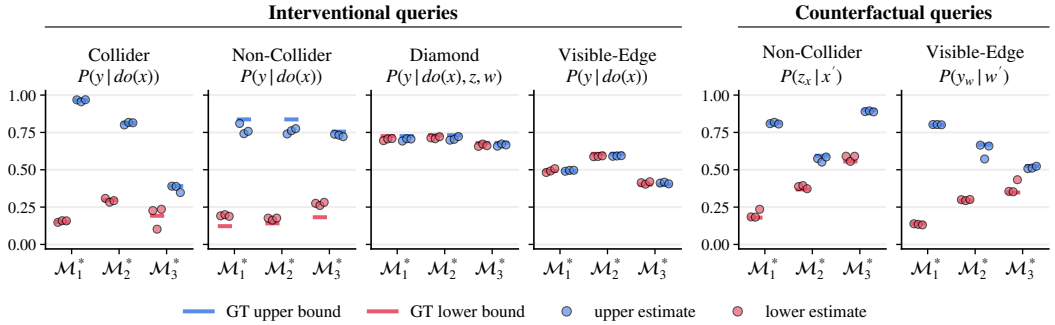


Figure 5: Estimated vs. ground-truth (GT) bounds for interventional (left) and counterfactual (right) queries on the PAGs in Fig. 4. $\mathcal{M}_1^*$–$\mathcal{M}_3^*$ denote three ground-truth SCMs for each PAG. Each dot represents one of three independent runs minimizing (lower bound) or maximizing (upper bound) the target query.

log-likelihood, the target query, and penalties enforcing the constraints in Def. 9 (see Appendix E for details).

Interventional Queries: We consider $P(y \mid do(x))$ for the collider, non-collider, and visible-edge PAGs in Fig. 4(a), (b), and (d), respectively, and $P(y \mid do(x), w, z)$ for the diamond PAG in Fig. 4(c). The queries are identifiable in (c) and (d), and partially identifiable in (a) and (b).

Counterfactual Queries: We consider $P(z_x \mid x')$ for the non-collider PAG in Fig. 4(b) and $P(y_w \mid w')$ for the visible-edge PAG in Fig. 4(d), both of which are partially identifiable. For evaluation, we set $x = w = 1$ and $x' = w' = 0$, yielding the queries $P(Z_{X=1} = 1 \mid X = 0)$ and $P(Y_{W=1} = 1 \mid W = 0)$, respectively.

Across all settings, the algorithm recovers the ground-truth bounds with minimal error, as illustrated in Fig. 5.

6 Discussions

In this paper, we characterized SCM-induced causal diagrams in the MEC represented by a PAG through two classes, skeleton-matched diagrams (SMDs) and directed-backboned diagrams (DBDs) (Props. 1 and 2), and proved their counterfactual expressiveness (Thm. 1). We then translated these graphical characterizations into constraints over adjacency masks (Defs. 4, 5, and 6) that are amenable for optimization. Building on these results, we introduced $\mathcal{P}$-NCMs (Def. 9), a class of neural causal models that encode structural uncertainty within the MEC while preserving its full counterfactual expressiveness (Thm. 2). We further established the theoretical correctness of counterfactual partial identification over this model class (Thm. 3) and demonstrated its effectiveness through experiments on simulated causal bounding tasks.

Several directions remain for future work. First, improving the scalability of the framework is an important challenge, particularly given the potentially exponential number of admissible exogenous variables (Def. 7) and the computational complexity of the inducing-path constraints. Developing more efficient parameterizations and constraint formulations could enable applications to larger graphs. Second, while our framework focuses on PAGs representing equivalence classes derived from observational data, a natural extension is to accommodate equivalence classes further refined by interventional distributions or other forms of causal knowledge.

Acknowledgments and Disclosure of Funding

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

References

[1] R. Ayesha Ali, Thomas S. Richardson, and Peter Spirtes. Markov equivalence for ancestral graphs. *The Annals of Statistics*, 37(5B), October 2009. arXiv:0908.3605 [math].

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

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

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

[5] Alexis Bellot. Towards bounding causal effects under markov equivalence. In *Proceedings of the Fortieth Conference on Uncertainty in Artificial Intelligence*, volume 244, pages 308–332. PMLR, 2024.

[6] Tom Claassen and Ioan G. Bucur. Greedy equivalence search in the presence of latent confounders. In *Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence*, pages 443–452. PMLR, August 2022.

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

[8] Adiba Ejaz, Yushu Pan, Hongshuo Yang, and Elias Bareinboim. Neural causal models under markov equivalence. In *The Fortieth Annual Conference on Neural Information Processing Systems*, 2026.

[9] Antti Hyttinen, Frederick Eberhardt, and Matti Jarvisalo. Do-calculus when the True Graph Is Unknown.

[10] Amin Jaber, Adele Ribeiro, Jiji Zhang, and Elias Bareinboim. Causal identification under markov equivalence: Calculus, algorithm, and completeness. In *Proceedings of the 36th Annual Conference on Neural Information Processing Systems*, 2022.

[11] Amin Jaber, Jiji Zhang, and Elias Bareinboim. Causal identification under markov equivalence. In *Proceedings of the 34th Conference on Uncertainty in Artificial Intelligence*, 2018.

[12] Amin Jaber, Jiji Zhang, and Elias Bareinboim. A graphical criterion for effect identification in equivalence classes of causal diagrams. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, 2018.

[13] Amin Jaber, Jiji Zhang, and Elias Bareinboim. Causal identification under markov equivalence: Completeness results. In *Proceedings of the 36th International Conference on Machine Learning*, Proceedings of Machine Learning Research, 2019.

[14] Amin Jaber, Jiji Zhang, and Elias Bareinboim. Identification of conditional causal effects under markov equivalence. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

[15] Amin Jaber, Jiji Zhang, and Elias Bareinboim. On causal identification under markov equivalence. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, 2019.

[16] Daniel Malinsky and Peter Spirtes. Estimating Causal Effects with Ancestral Graph Markov Models. *JMLR workshop and conference proceedings*, 52:299–309, August 2016.

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

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

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

[20] Thomas Richardson. Markov Properties for Acyclic Directed Mixed Graphs. *Scandinavian Journal of Statistics*, 30(1):145–157, 2003.

[21] Thomas Richardson and Peter Spirtes. Ancestral Graph Markov Models. *The Annals of Statistics*, 30(4):962–1030, 2002.

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

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

[24] Kevin Muyuan Xia, Yushu Pan, and Elias Bareinboim. Neural causal models for counterfactual identification and estimation. In *The Eleventh International Conference on Learning Representations*, 2023.

[25] Hongshuo Yang and Elias Bareinboim. A hierarchy of graphical models for counterfactual inferences. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

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

[27] Jiji Zhang. On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias. *Artificial Intelligence*, 172(16-17):1873–1896, November 2008.

[28] Jiji Zhang and Peter L. Spirtes. A transformational characterization of markov equivalence for directed acyclic graphs with latent variables. *arXiv preprint arXiv:1207.1419*, 2012.

[29] Junzhe Zhang, Jin Tian, and Elias Bareinboim. Partial counterfactual identification from observational and experimental data. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 26548–26558. PMLR, 17–23 Jul 2022.

[30] Xun Zheng, Bryon Aragam, Pradeep K Ravikumar, and Eric Xing. DAGs with NO TEARS: Continuous Optimization for Structure Learning. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

Appendices

Contents

A Background and Definitions **13**

A.1 Graph Definitions 13

A.2 SCMs and Causal Diagrams 14

A.3 MAG and PAGs 15

A.4 Partial Identification using DAG-Constrained Neural Causal Models 16

B Further Details **18**

B.1 Examples for definitions in the main text 18

B.2 Relationship among different space of causal diagrams 18

B.3 Constraints for SMDs without fixed mask parameterization 18

B.4 Sensitivity to PAG Misspecification and Constraint Relaxation 19

C Supporting Lemmas **22**

C.1 Lemmas over graphs 22

C.2 Lemmas over SCMs 24

D Proofs **28**

E Experiments **33**

E.1 Experimental Setup 33

E.2 Additional Experimental Results 34

E.2.1 Ground-truth bounds and experimental errors 34

E.2.2 Ablation - Partial ID without PAG Constraints 34

A Background and Definitions

A.1 Graph Definitions

Definition 14 (Acyclic Directed Mixed Graph)**.** *An acyclic directed mixed graph (ADMG) is a vertex-edge graph without any directed cycles that may contain two kinds of edges: directed edges ($\longrightarrow$) and bi-directed edges ($\longleftrightarrow$).*

Standard kinship notations for variable relationships apply in ADMGs: parents (Pa), children (Ch), descendants (De), ancestors (An).

Definition 15 (Confounded Component [23])**.** *Let $\mathbf{C}_1, \mathbf{C}_2, ...\mathbf{C}_k$ be a partition over the set of variables $\mathbf{V}$, where $\mathbf{C}_i$ is said to be a confounded component (for short, C-component) of the ADMG $\mathcal{G}$ if for every $V_i, V_j \in \mathbf{C}_i$ there exists a path made entirely of bidirected edges between V_i and V_j in $\mathcal{G}$ and $\mathbf{C}_i$ is maximal.*

Definition 16 (Augmented Parents)**.** *Let $<$ be a topological order over the variables $V_1, ..., V_n$ in the ADMG $\mathcal{G}$, let $\mathcal{G}(V_i)$ be the subgraph of $\mathcal{G}$ consists only of variables in $V_1, ..., V_i$, and let $\mathbf{C}(V_i)$ be the C-component of V_i in $\mathcal{G}(V_i)$. The augmented parents of V_i, denoted as Pa_i^+, is the union of parents of all variables in $\mathbf{C}(V_i)$ that comes before V_i in topological order:*

$$Pa_i^+ = \cup_{j|V_j \in \mathbf{T_i}} Pa_j \backslash \{V_i\} \tag{9}$$

where $\mathbf{T}_i = \{X \in \mathbf{C}(V_i) : X \leq V_i\}$.

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

Definition 17 (Semi-Markov Relative to [4]). *A probability $P(\mathbf{V})$ is said to be semi-Markov relative to an ADMG $\mathcal{G}$ if for any topological order $<$ of $\mathcal{G}$:*

$$P(\mathbf{V}) = \prod_i P(v_i|pa_i^+) \tag{10}$$

A.2 SCMs and Causal Diagrams

Definition 18 (Structural Causal Model (SCM) [4]). *A structural causal model $\mathcal{M}$ is a 4-tuple $\langle \mathbf{U}, \mathbf{V}, \mathcal{F}, P(\mathbf{U}) \rangle$, where*

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

  $$v_i \leftarrow f_i(pa_i, u_i), \tag{11}$$

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

The causal diagram is an ADMG induced by an underlying SCM.

Definition 19 (Causal Diagram [4]). *Consider an SCM $\mathcal{M} = \langle \mathbf{U}, \mathbf{V}, \mathcal{F}, P(\mathbf{u}) \rangle$. Then an ADMG $\mathcal{G}$ is a causal diagram of $\mathcal{M}$ if constructed as follows:*

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

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

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

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

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

Definition 20 (Submodel — "Interventional SCM" [18]). *Let $\mathcal{M}$ be a structural causal model, $\mathbf{X}$ a set of variables in $\mathbf{V}$, and $\mathbf{x}$ a particular realization of $\mathbf{X}$. A submodel $\mathcal{M}_{\mathbf{x}}$ of $\mathcal{M}$ is the causal model*

$$\mathcal{M}_{\mathbf{x}} = \langle \mathbf{U}, \mathbf{V}, \mathcal{F}_{\mathbf{x}}, P(\mathbf{U}) \rangle, \tag{12}$$

where

$$\mathcal{F}_{\mathbf{x}} = \{f_i : V_i \notin \mathbf{X}\} \cup \{\mathbf{X} \leftarrow \mathbf{x}\}. \tag{13}$$

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

Definition 21 (Potential Outcomes [18]). *Let $\mathbf{X}$ and $\mathbf{Y}$ be two sets of variables in $\mathbf{V}$, and $\mathbf{u}$ be a unit. The potential outcome $\mathbf{Y}_{\mathbf{x}}(\mathbf{u})$ is defined as the solution for $\mathbf{Y}$ of the set of equations $\mathcal{F}_{\mathbf{x}}$ with respect to SCM $\mathcal{M}$ (or, $\mathbf{Y}_{\mathcal{M}_{\mathbf{x}}}(\mathbf{u})$). That is, $\mathbf{Y}_{\mathbf{x}}(\mathbf{u}) \triangleq \mathbf{Y}_{\mathcal{M}_{\mathbf{x}}}(\mathbf{u})$.*

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

Definition 22 (Pearl Causal Hierarchy (PCH) ([4])). *An SCM $\mathcal{M} = \langle \mathbf{U}, \mathbf{V}, \mathcal{F}, P(\mathbf{u}) \rangle$ induces three layers of probability distributions that form the Pearl Causal Hierarchy. For any $\mathbf{Y}, \mathbf{Z}, ..., \mathbf{X}, \mathbf{W} \subseteq \mathbf{V}$, the three layers of distributions are given by:*

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

$$\mathbf{P}^{\mathcal{M}}(\mathbf{y}) = \sum_{\mathbf{u}} \mathbf{1}[\mathbf{Y}(\mathbf{u}) = \mathbf{y}]P(\mathbf{u}) \tag{14}$$

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

$$\mathbf{P}^{\mathcal{M}}(\mathbf{y_x}) = \sum_{\mathbf{u}} \mathbf{1}[\mathbf{Y_x}(\mathbf{u}) = \mathbf{y}]P(\mathbf{u}) \tag{15}$$

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

$$\mathbf{P}^{\mathcal{M}}(\mathbf{y_x}, ..., \mathbf{z_w}) = \sum_{\mathbf{u}} \mathbf{1}[\mathbf{Y_x}(\mathbf{u}) = \mathbf{y}, ..., \mathbf{Z_w}(\mathbf{u}) = \mathbf{z}]P(\mathbf{u}) \tag{16}$$

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

PCH specifies both the symbolic representation and the valuations of each probabilistic quantity given an underlying SCM.

A.3 MAG and PAGs

In an ADMG $\mathcal{G}$, an *almost directed cycle* occurs when $Y \leftrightarrow X$ is in $\mathcal{G}$ and $X \in An_{\mathcal{G}}(Y)$. An *inducing path relative to* $\mathbf{L}$ is a path on which every vertex not in $\mathbf{L}$ (except for the endpoints) is a collider on the path and every collider is an ancestor of an endpoint of the path.

Definition 23 (MAG)**.** *An ADMG is called a maximal ancestral graph (MAG) if*

1. *the graph does not contain any directed or almost directed cycles (ancestral); and*
2. *there is no inducing path between any two non-adjacent vertices (maximal).*

Definition 24 (Inducing Path)**.** [2] *A path p between X and Y is called an inducing path if every non-endpoint vertex is a collider and every collider is an ancestor of either X or Y.*

There are three types of inducing paths between X and Y:

1. Out of X and into Y: which makes $X \in An(Y)$ and induces $X \to Y$ in the MAG.
2. Out of Y and into X: which makes $Y \in An(X)$ and induces $Y \to X$ in the MAG.
3. Into both X and Y:
   (a) If $X \in An(Y)$, it induces $X \to Y$ in the MAG;
   (b) If $Y \in An(X)$, it induces $Y \to X$ in the MAG;
   (c) If $X \notin An(Y)$ and $Y \notin An(X)$, it induces $X \leftrightarrow Y$ in the MAG.

Definition 25 (Discriminating Path)**.** *A path $p = \langle X, ..., W, V, Y \rangle$ is a discriminating path for V if*

- *p includes at least three edges; and*
- *V is a non-endpoint on p, and is adjacent to Y on p; and*
- *X is not adjacent to Y, and every vertex between X and V is a collider on p and is a parent of Y.*

Definition 26 (Visibility (Def. 8 [26]))**.** *Given a MAG M, a directed edge $A \to B$ in M is visible if there is a vertex C not adjacent to B, such that either there is an edge between C and A that is into A, or there is a collider path between C and A that is into A and every vertex on the path is a parent of B* [3]*. Otherwise $A \to B$ is said to be invisible.*

Definition 27 (PAG)**.** *Let $[M]$ be the Markov equivalence class of an arbitrary MAG M. The partial ancestral graph (PAG) for $[M]$, $\mathcal{P}_{[M]}$, is a partial mixed graph such that*

1. *$\mathcal{P}_{[M]}$ has the same adjacencies as M (and any member of $[M]$) does;*

[2] Single edges are inducing paths too.

[3] This path is a discriminating path for A, where A is a definite non-collider.

2. *A mark of arrowhead is in $\mathcal{P}_{[M]}$ if and only if it is shared by all MAGs in $[M]$;and*

3. *A mark of tail is in $\mathcal{P}_{[M]}$ if and only if it is shared by all MAGs in $[M]$.*

The definition of visibility applies in PAGs, such that a directed edge in a PAG is called *definitely visible* if it satisfies the condition for visibility in Def. 26.

Definition 28 (Definite Colliders & Non-Colliders in PAGs)**.** *Let $\langle A, B, C\rangle$ be any consecutive triple along a path P in PAG $\mathcal{P}$. B is a collider on p if both edges are into B (i.e., $A* \rightarrow B \leftarrow *C$). B is a definite non-collider on p if one of the edges is out of B ($A \leftarrow B * - * C$ or $A * - * B \rightarrow C$), or both edges have circle marks at B and there is no edge between A and C (i.e., $A * - \circ B \circ - * C$, where A and B are not adjacent). Otherwise, B has a non-definite status along p.*

Definition 29 (Manipulations of PAGs)**.** *Given a PAG $\mathcal{P}$ and a set of variables $\mathbf{X}$ therein,*

- *the $\mathbf{X}$-lower-manipulation of $\mathcal{P}$ deletes all those edges that are definitely visible in $\mathcal{P}$ and are out of variables in $\mathbf{X}$, replaces all those edges that are out of variables in $\mathbf{X}$ but are not definitely visible in $\mathcal{P}$ with bi-directed edges, and otherwise keeps $\mathcal{P}$ as it is. The resulting graph is denoted as $\mathcal{P}_{\underline{\mathbf{X}}}$.*
- *the $\mathbf{X}$-upper-manipulation of $\mathcal{P}$ deletes all those edges in $\mathcal{P}$ that are into variables in $\mathbf{X}$, and otherwise keeps $\mathcal{P}$ as it is. The resulting graph is denoted as $\mathcal{P}_{\overline{\mathbf{X}}}$.*

Theorem 4 (Do-Calculus in PAGs [10])**.** *Given a PAG $\mathcal{P}$ over $\mathbf{V}$ and joint sets $\mathbf{X}, \mathbf{Y}, \mathbf{Z}, \mathbf{W} \subseteq \mathbf{V}$. For all ADMGs represented by $\mathcal{P}$, their CBNs satisfy the consequent of the following three rules:*

$$\textbf{\textit{Rule 1}}\ P(\mathbf{y}|do(\mathbf{x}), \mathbf{z}, \mathbf{w}) = P(\mathbf{y}|do(\mathbf{x}), \mathbf{w}) \qquad \text{if } (\mathbf{Y} \perp\!\!\!\perp \mathbf{Z}|\mathbf{X}, \mathbf{W})_{\mathcal{P}_{\overline{\mathbf{X}}}}. \tag{17}$$

$$\textbf{\textit{Rule 2}}\ P(\mathbf{y}|do(\mathbf{x}), do(\mathbf{z}), \mathbf{w}) = P(\mathbf{y}|do(\mathbf{x}), \mathbf{z}, \mathbf{w}) \qquad \text{if } (\mathbf{Y} \perp\!\!\!\perp \mathbf{Z}|\mathbf{X}, \mathbf{W})_{\mathcal{P}_{\overline{\mathbf{X}}\underline{\mathbf{Z}}}}. \tag{18}$$

$$\textbf{\textit{Rule 3}}\ P(\mathbf{y}|do(\mathbf{x}), do(\mathbf{z}), \mathbf{w}) = P(\mathbf{y}|do(\mathbf{x}), \mathbf{w}) \qquad \text{if } (\mathbf{Y} \perp\!\!\!\perp \mathbf{Z}|\mathbf{X}, \mathbf{W})_{\mathcal{P}_{\overline{\mathbf{X}\mathbf{Z}(\mathbf{W})}}}, \tag{19}$$

where $\mathbf{Z}(\mathbf{W}) := \mathbf{Z} \backslash PossAn(\mathbf{W})_{\mathcal{P}_{\mathbf{V}\backslash\mathbf{X}}}$.

Definition 30 (Acyclicity Constraint - NO TEARS)**.** *Given a masked $\mathcal{P}$-NCM with $|\mathbf{V}| = n$ and fixed mask value assignments $\mathbf{m}^D$, its induced causal diagram is acyclic if and only if*

$$h(\mathbf{m}^D) = tr(e^{\mathbf{m}^D \odot \mathbf{m}^D}) - n = 0,$$

where $\odot$ is the Hadamard product and e^A is the matrix exponential of A.

A.4 Partial Identification using DAG-Constrained Neural Causal Models

Theorem 5 (NCM Expressiveness [24, Thm. 2])**.** *For any SCM $\mathcal{M}$ inducing a causal diagram $\mathcal{G}$, there exists a $\mathcal{G}$-NCM $\mathcal{N}$ such that $\mathcal{M}$ and $\mathcal{N}$ are $\mathcal{L}_3$ consistent.*

Let $\mathbb{Z}$ be a finite set of realizations $\mathbf{z}_i$ of subsets of variables $\mathbf{Z}_i \subseteq \mathbf{V}$.

Definition 31 (Optimal Counterfactual Bound [29, Def. 2.1])**.** *For a causal diagram $\mathcal{G}$ and distributions $\{P(\mathbf{V_z}) \mid \mathbf{z} \in \mathbb{Z}\}$, the optimal bound $[l, r]$ over a counterfactual probability $P(\mathbf{y_x}, ..., \mathbf{z_w})$ is defined as, respectively, the minimum and maximum of the following optimization problem:*

$$\min / \max_{\mathcal{M} \in \Omega(\mathcal{G})} P^{\mathcal{M}}(\mathbf{y_x}, ..., \mathbf{z_w})$$
$$\text{s.t. } P^{\mathcal{M}}(\mathbf{V_z}) = P(\mathbf{V_z}) \forall \mathbf{z} \in \mathbb{Z}$$

where $\Omega(\mathcal{G})$ is the set of all SCMs satisfying inducing the graph $\mathcal{G}$.

Proposition 7 (Subgraph partial ID)**.** *Given two graphs $\mathcal{G}, \mathcal{G}'$ where $\mathcal{G}$ is a subgraph of $\mathcal{G}'$; query Q, and a set of distributions $\mathbb{P}$, the optimal counterfactual bounds of Q with respect to $\mathbb{P}$ and $\mathcal{G}$ are a subset of those with respect to $\mathbb{P}$ and $\mathcal{G}'$.*

Corollary 1 (Correctness of Neural Partial ID)**.** *Consider an SCM $\mathcal{M}^*$, causal diagram $\mathcal{G}$ and distributions $\mathbb{P}$ induced by $\mathcal{M}$, and a query Q. The bounds output by* NeuralID *[24, Alg. 1] given inputs $\mathcal{G}$, $\mathbb{P}$, and Q. Let l, r coincide with the optimal counterfactual bounds for Q from $\mathcal{G}$ and $\mathbb{P}$.*

Proof. Let $\hat{l}, \hat{r}$ be the bounds output by the $\mathcal{G}$-NCM procedure. To see that $[l, r] \subseteq [\hat{l}, \hat{r}]$, consider: for any SCM $|M \in \Omega(\mathcal{G})$, there exists a $\mathcal{G}$-NCM $\mathcal{N}$ such that $P^{\mathcal{M}}(\mathbf{y_x}, ..., \mathbf{z_w}) = P^{\mathcal{N}}(\mathbf{y_x}, ..., \mathbf{z_w})$.

By construction of RelationalNeuralID [24, Alg. 1], it must be that $P^{\mathcal{N}}(\mathbf{y_x}, ..., \mathbf{z_w}) \in [\hat{l}, \hat{r}]$. Next, to see that $[\hat{l}, \hat{r}] \subseteq [l, r]$: by construction of [24, Alg. 1], there exist $\mathcal{G}$-NCMs $\mathcal{N}_{\hat{l}}$ and $\mathcal{N}_{\hat{r}}$ such that $P^{\mathcal{N}_{\hat{l}}}(\mathbf{y_x}, ..., \mathbf{z_w}) = \hat{l}$ and $P^{\mathcal{N}_{\hat{r}}}(\mathbf{y_x}, ..., \mathbf{z_w}) = \hat{r}$. By [24, Thm. 1], $\mathcal{N}_{\hat{l}}$ and $\mathcal{N}_{\hat{r}}$ satisfy all $\mathcal{L}_3$ constraints encoded in $\mathcal{G}$; therefore, $\mathcal{N}_{\hat{l}}, \mathcal{N}_{\hat{r}} \in \Omega(\mathcal{G})$. This implies that $\hat{l}, \hat{r} \in [l, r]$ by definition of the optimal bounds (Def. 31) and we are done. □

B Further Details

B.1 Examples for definitions in the main text

The example below shows the details of the exogenous variables and mask parameterizations following the definitions in the main text.

Example 9 (PAG-Induced Exogenous Variables and Mask Parameterization)**.** *Consider the PAG $\mathcal{P}$ in Fig. 1(a) with $\mathbf{V} = \{X, Z, W, Y\}$.*

Its induced exogenous variables (Def. 7) are

$$\mathbf{U} = \{U_X, U_Z, U_W, U_Y, U_{XZ}, U_{XW}, U_{ZW}, U_{WY}, U_{ZY}, U_{XZW}, U_{ZWY}\}.$$

Its adjacency and bidirected clique mask parameters will be instantiated as

$$\mathbf{m}^D = \begin{bmatrix} 0 & m_{12}^D & m_{13}^D & 0 \\ 1-m_{12}^D & 0 & m_{23}^D & m_{24}^D \\ 1-m_{13}^D & 1-m_{23}^D & 0 & m_{34}^D \\ 0 & 1-m_{24}^D & 1-m_{34}^D & 0 \end{bmatrix} \quad \mathbf{m}^B = \begin{bmatrix} 0 & m_{12}^B & m_{13}^B & 0 \\ m_{12}^B & 0 & m_{23}^B & m_{24}^B \\ m_{13}^B & m_{23}^B & 0 & m_{34}^B \\ 0 & m_{24}^B & m_{34}^B & 0 \end{bmatrix}$$

$$\mathbf{m}^C = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ m_{12}^B & m_{12}^B & 0 & 0 \\ m_{13}^B & 0 & m_{13}^B & 0 \\ 0 & m_{23}^B & m_{23}^B & 0 \\ 0 & 0 & m_{34}^B & m_{34}^B \\ 0 & m_{24}^B & 0 & m_{24}^B \\ m_{12}^B m_{13}^B m_{23}^B & m_{12}^B m_{13}^B m_{23}^B & m_{12}^B m_{13}^B m_{23}^B & 0 \\ 0 & m_{23}^B m_{24}^B m_{34}^B & m_{23}^B m_{24}^B m_{34}^B & m_{23}^B m_{24}^B m_{34}^B \end{bmatrix}$$

where gray entries 0 and 1 denote fixed parameters, blue entries such as m_{12}^D denote free parameters, and red entries such as $1 - m_{12}^D$ denote constrained (derived) parameters. In other words, only the free parameters are independently trainable, subject to the constraints specified in Def. 9. ■

B.2 Relationship among different space of causal diagrams

Bellot [5] introduced a class of causal diagrams within the MEC represented by a PAG, termed *maximally bidirected diagrams* (MBDs), defined as follows.

Definition 32 (Maximally Bidirected Diagrams [5, Def. 8])**.** *A causal diagram G is said to be maximally bidirected if no additional bidirected edges can be added without violating a d-separation.*

The class of MBDs generally differs from that of DBDs within the same MEC. Bellot [5] further proposed a construction of MBDs starting from a loyal equivalent graph (LEG), yielding a more restricted class corresponding to the intersection of MBDs and DBDs.

In this section, we use examples to illustrate the relationships among these classes of causal diagrams within the MEC represented by a PAG.

PAG 1: Fig. 6 illustrates the relationships among the different classes of causal diagrams in the MEC represented by the PAG shown in the upper-right corner. Three observations are noteworthy: (1) DBDs form a subset of SMDs; (2) MBDs need not be SMDs; and (3) the intersection of DBDs and MBDs consists of causal diagrams obtained by adding a maximal set of bidirected edges to a loyal equivalent graph (LEG) in the MEC. Representative causal diagrams from these classes are shown in Fig. 7.

PAG 2: Fig. 8 illustrates the relationships among the classes of causal diagrams in the MEC represented by the PAG shown in the upper-right corner. The same relationships observed for PAG 1 hold here. Representative causal diagrams from these classes are shown in Fig. 9.

B.3 Constraints for SMDs without fixed mask parameterization

In this section, we prove a way to translate the ordinary adjacency and collider-triples conditions in Prop. 1 can be translated to flexible constraints over the adjacency masks rather than fixed parameterization.

Definition 33 (Ordinary Adjacency Constraint)**.** *Given a PAG $\mathcal{P}$ with MEC signature $\langle \mathfrak{S}, \mathfrak{C}, \mathfrak{D} \rangle$ and a causal diagram $\mathcal{G}$ with adjacency masks $(\mathbf{m}^D, \mathbf{m}^B)$, the ordinary adjacency constraint requires*

$$m_{ij}^D + m_{ji}^D + m_{ij}^B \geq \epsilon \mathfrak{S}_{ij}, \qquad \forall\, i < j, \tag{20}$$

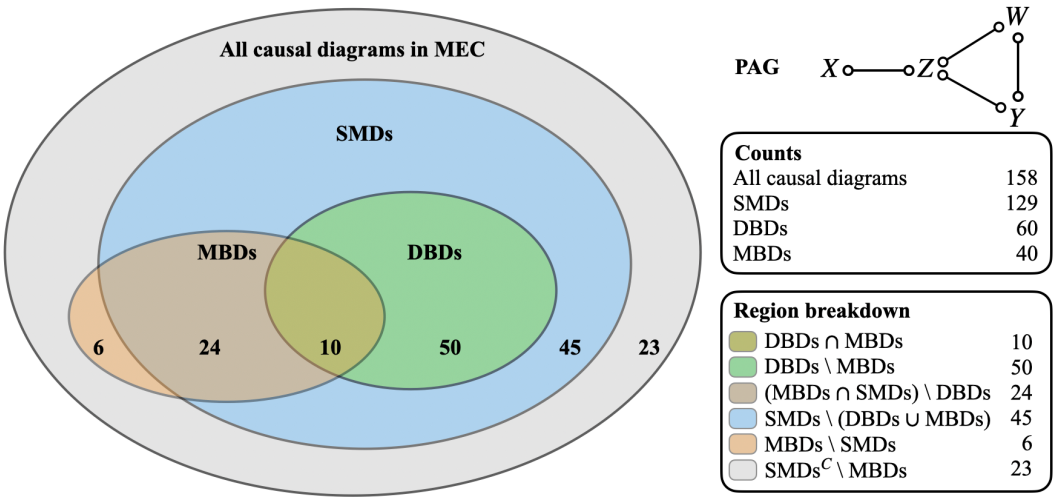


Figure 6: Illustration of the relationships among causal diagrams represented by the PAG $\mathcal{P}$. The full MEC contains 158 causal diagrams, of which 129 are skeleton-matched diagrams (SMDs), 60 are directed-backboned diagrams (DBDs), and 40 are maximally bidirected diagrams (MBDs). DBDs form a strict subset of SMDs, whereas MBDs intersect both the SMD class and its complement. The numbers indicate the sizes of the six disjoint regions for this PAG.

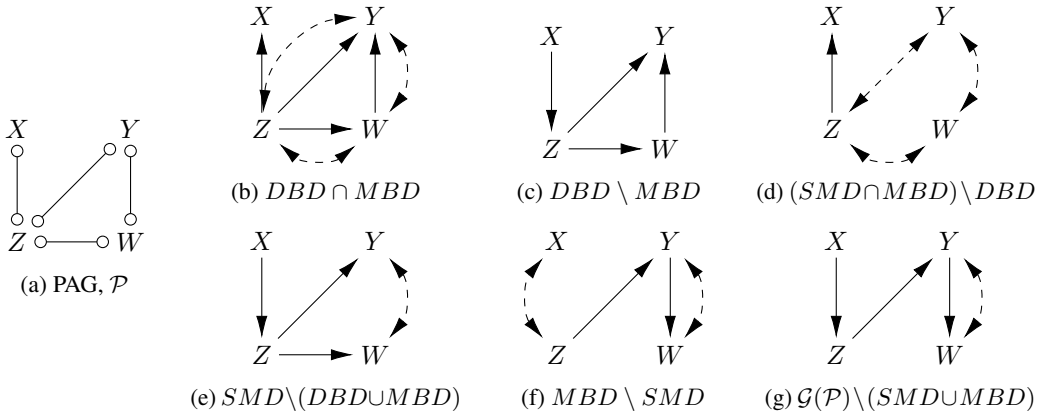


Figure 7: The PAG $\mathcal{P}$ (left) and representative causal diagrams from the six disjoint regions induced by the DBD, SMD, and MBD graph classes. The first four causal diagrams are skeleton-matched, whereas the last two are not: in both cases, the missing Z–W adjacency is recovered in the projected graph through the inducing path $\langle Z, Y, W \rangle$.

where $\epsilon > 0$ is a small positive threshold.

Definition 34 (Collider Triple Constraint)**.** *Given a PAG $\mathcal{P}$ with MEC signature $\langle \mathfrak{S}, \mathfrak{C}, \mathfrak{D} \rangle$ and a causal diagram $\mathcal{G}$ with adjacency masks $(\mathbf{m}^D, \mathbf{m}^B)$, define*

$$W_{ij} = m^D_{ij} + m^B_{ij} - m^D_{ij} m^B_{ij}, \qquad H_{ijk} = W_{ij} W_{kj}.$$

The collider triple constraint requires

$$H_{ijk} > 0, \qquad \forall \langle V_i, V_j, V_k \rangle \in \mathfrak{C}. \tag{21}$$

B.4 Sensitivity to PAG Misspecification and Constraint Relaxation

Throughout this paper, we assume that the true PAG is available to focus on searching within the correct MEC. In practice, however, PAGs are typically learned from finite observational data under assumptions such as causal Markov and faithfulness, and may therefore be misspecified. Understanding the effects of PAG misspecification, constraint relaxation, and optimization error is an

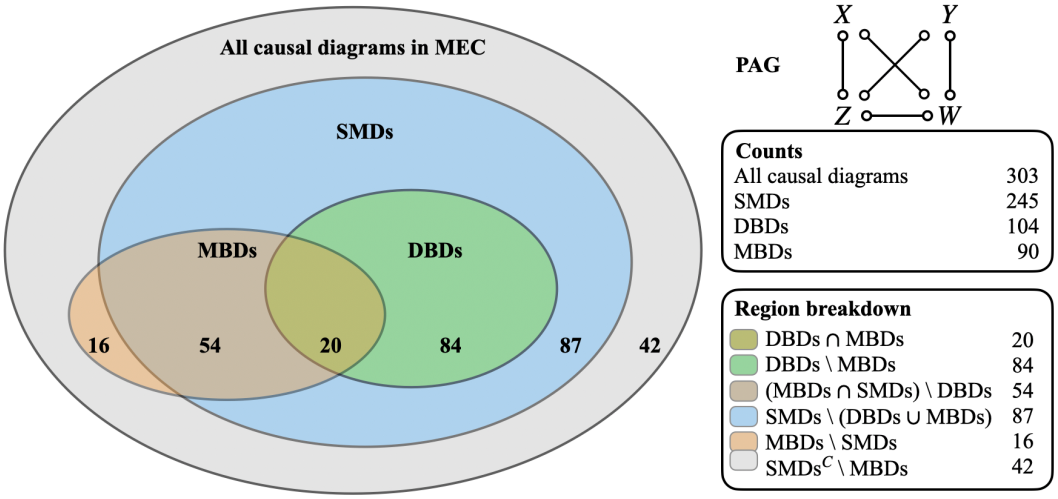


Figure 8: Illustration of the relationships among causal diagrams represented by the PAG $\mathcal{P}$. The full MEC contains 303 causal diagrams, of which 245 are skeleton-matched diagrams (SMDs), 104 are directed-backboned diagrams (DBDs), and 90 are maximally bidirected diagrams (MBDs). DBDs form a strict subset of SMDs, whereas MBDs intersect both the SMD class and its complement. The numbers indicate the sizes of the six disjoint regions for this PAG.

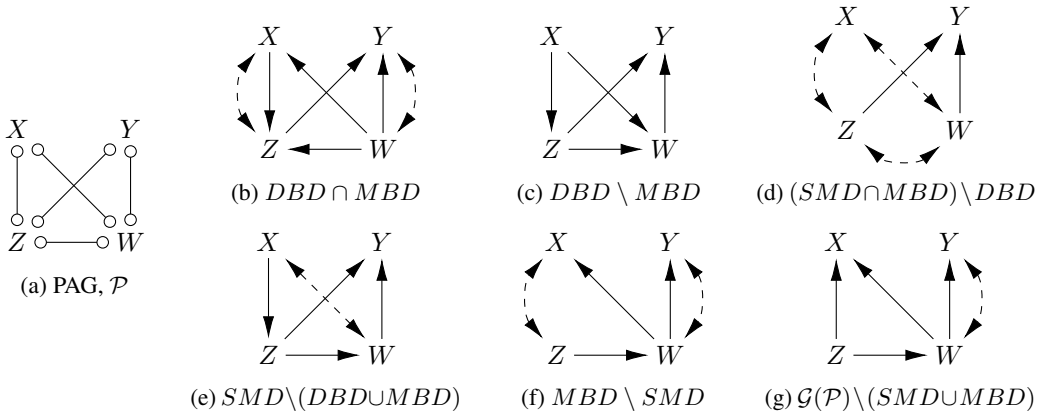


Figure 9: The PAG $\mathcal{P}$ (left) and representative causal diagrams from the six disjoint regions induced by the DBD, SMD, and MBD graph classes. The first four causal diagrams are skeleton-matched. The last two are not skeleton-matched: in both diagrams, the Z–Y adjacency is absent from the ordinary skeleton but recovered in the projected graph through the inducing path $\langle Z, W, Y \rangle$. None of the six diagrams admits an inducing path between the nonadjacent variables X and Y.

important direction for future work. Here, we provide a brief theoretical analysis of several controlled forms of misspecification to clarify the conditions under which the partial identification guarantees remain valid.

Over-specification by adding edges. Suppose the input PAG contains additional adjacencies relative to the true PAG. When these additions enlarge the class of compatible SCMs while preserving the original model class, the resulting partial identification bounds can only become wider. In particular, Lemma 8 establishes that suitable supergraphs can preserve the counterfactual distributions of the original SCMs, ensuring that their query values remain attainable in the enlarged class.

For example, consider the PAG in Fig. 4(d). Adding an edge between W and Z may admit additional causal structures excluded by the original PAG. Consequently, the query $P(y \mid do(x))$, identified as $P(y \mid x)$ under the original PAG, may become partially identified when alternative structures permit-

ting latent confounding between X and Y are admitted. Provided the enlarged model class retains the original compatible SCMs, the resulting bounds remain valid but potentially less informative.

Under-specification by removing edges. Suppose one or more true adjacencies are omitted from the input PAG. Under the faithfulness assumption, removing a true adjacency introduces conditional independence constraints that are not satisfied by the observational distribution $P(\mathbf{V})$. Consequently, the corresponding $\mathcal{P}$-NCM class cannot reproduce $P(\mathbf{V})$ exactly, and the resulting partial identification bounds are no longer guaranteed to be valid.

For example, consider again Fig. 4(d). Removing the edge between X and Y implies an m-separation between these variables given some conditioning set, contradicting the faithful observational distribution generated by the true causal structure containing $X \to Y$. Thus, no $\mathcal{P}$-NCM satisfying the misspecified structural constraints can exactly reproduce $P(\mathbf{V})$, and the theoretical correctness guarantee no longer applies.

Constraint relaxation and optimization error. Relaxing structural constraints may similarly enlarge the feasible model class, potentially yielding wider bounds when the original feasible class is preserved. In contrast, incomplete optimization or imperfect observational fitting may prevent the estimated bounds from reaching their true extrema, without guaranteeing that the resulting interval contains the correct bounds.

More general forms of misspecification require further analysis. These examples highlight that the correctness guarantees depend on the representational adequacy of the input PAG and the accurate enforcement of observational and structural constraints.

Algorithm 2 ADMG to MAG

Input: an ADMG $\mathcal{G}$ over $\mathbf{V}$
Output: a MAG M over $\mathbf{V}$, representing the same CIs over $\mathbf{V}$
Initialize M with nodes in $\mathbf{V}$
for each pair of variables $A, B \in \mathbf{V}$ **do**
    if there is an inducing path between A and B in $\mathcal{G}$ **then**
        if $A \in \mathrm{An}(B)$ in $\mathcal{G}$ **then**
            Add edge $A \to B$ to M
        else if $B \in \mathrm{An}(A)$ in $\mathcal{G}$ **then**
            Add edge $B \to A$ to M
        else
            Add edge $A \leftrightarrow B$ to M
        end if
    end if
end for
Return M

C Supporting Lemmas

C.1 Lemmas over graphs

Fact 1. *If $A \to B$ or $A \leftrightarrow B$ is in the MAG, then for every ADMG represented by the MAG, there is no inducing path between A to B out of B.*

Fact 2. *If $A \to B$ is in the MAG, then for every ADMG represented by the MAG, there is an inducing path between A to B that is into B, and $A \in An(B)$.*

Fact 3. *If $A \leftrightarrow B$ is in the MAG, then for every ADMG represented by the MAG, there is an inducing path between A to B that is into both A and B, and $A \notin An(B)$ and $B \notin An(A)$.*

Lemma 1 (Contrapositive of Lemma 9 [26])**.** *If $A \to B$ is visible in the MAG, then for every ADMG represented by the MAG, there is no inducing path between A to B that is into A.*

Lemma 2 (Markov Equivalence of MAGs using Ordered Triples [6, Cor. 3])**.** *Two MAGs are Markov equivalent if and only if they have identical MEC signatures $\langle \mathfrak{S}, \mathfrak{C}, \mathfrak{D} \rangle$.*

Lemma 3 ([26])**.** *Let $\mathcal{P}$ be a PAG. Let M be the graph resulting from the following procedure applied to a $\mathcal{P}$:*

1. *replace all partially directed edges $\circ\!\!\to$ in $\mathcal{P}$ with directed edges $\to$, and*
2. *orient the subgraph of $\mathcal{P}$ consisting of all non-directed edges $\circ\!\!-\!\!\circ$ into a DAG with no unshielded colliders.*

Then M is in the Markov equivalence class of $\mathcal{P}$. Moreover, if X is a node in $\mathcal{P}$, then one can always find an orientation of (ii) that does not create any new edges into X.

Lemma 4. *Let $\mathcal{G}$ be a causal diagram and let M be the MAG induced by $\mathcal{G}$ sharing the same ordinary skeleton, with MEC signature $\langle \mathfrak{S}, \mathfrak{C}, \mathfrak{D} \rangle$. Then:*

- *every triple in $\mathfrak{C}$ is a collider in $\mathcal{G}$;*
- *every triple in $\mathfrak{D}$ is a non-collider in $\mathcal{G}$.*

Proof. **Collider case.** Let $\langle a, b, c \rangle$ be an ordered collider in M. Then both edges incident to b have arrowheads at b, i.e.,

$$a* \to b \leftarrow *c.$$

Since $\mathcal{G}$ induces M and the triple is present in $\mathcal{G}$, Fact 1 implies that the corresponding adjacencies in $\mathcal{G}$ must also admit arrowheads at b. Hence $\langle a, b, c \rangle$ is a collider in $\mathcal{G}$.

Non-collider case. Let $\langle a, b, c \rangle$ be an ordered non-collider in M.

If the triple has order 0, it is an unshielded non-collider. Suppose, for contradiction, that it is a collider in $\mathcal{G}$. Then the collider configuration at b, together with the inducing relation between $\mathcal{G}$ and

M, would either induce arrowheads at b on both corresponding MAG edges or create an inducing path between a and c. The former contradicts that the triple is a non-collider in M, while the latter contradicts that a and c are non-adjacent in the unshielded triple.

If the triple has order $i \geq 1$, there exists a discriminating path

$$p = \langle x, \ldots, a, b, c \rangle$$

for b in M certifying its non-collider status. Suppose, for contradiction, that $\langle a, b, c \rangle$ is a collider in $\mathcal{G}$. If no inducing path changes the endpoint mark at b, the induced MAG contains arrowheads into b from both sides, contradicting the non-collider status in M. Otherwise, the inducing path responsible for the changed endpoint mark can be combined with the discriminating-path structure to yield an inducing path between the non-adjacent endpoints of the discriminating path, again a contradiction.

Thus every ordered non-collider in M is a non-collider in $\mathcal{G}$.

□

Since a PAG and every MAG in the equivalence class it represents share the same MEC signature, Lemma 4 immediately yields the corresponding preservation result for PAGs.

Corollary 2 (PAG–Causal Diagram Ordered-Triple Preservation)**.** *Let $\mathcal{P}$ be a PAG with MEC signature $\langle \mathfrak{S}, \mathfrak{C}, \mathfrak{D} \rangle$, and let $\mathcal{G}$ be a skeleton-matched causal diagram represented by $\mathcal{P}$. Then:*

- *every triple in $\mathfrak{C}$ is a collider in $\mathcal{G}$;*
- *every triple in $\mathfrak{D}$ is a non-collider in $\mathcal{G}$.*

Lemma 5 (MAG Edge Augmentation)**.** *Let $\mathcal{G}$ be a causal diagram and let M be its induced MAG. Let $E' \subseteq E(M) \setminus E(\mathcal{G})$, and let $\mathcal{G}' = \mathcal{G} \cup E'$ be the graph obtained by adding the corresponding edges of M to $\mathcal{G}$. Then $\mathcal{G}'$ induces the same MAG M.*

Proof. It suffices to show that adding a single edge $e \in E(M) \setminus E(G)$ does not change the induced MAG. The result for any subset of edges then follows by iteration.

First, note that adding an edge of M cannot create new ancestral relations. Indeed, every directed edge $A \to B$ in M corresponds to an ancestral relation $A \in An(B)$ already present in G, while bidirected edges do not affect ancestry. Thus, acyclicity and ancestral relations are preserved.

Suppose, for contradiction, that adding e creates a new inducing path π between vertices that were not connected by an inducing path in G. Necessarily, π contains e.

Case 1: $e = A \to B$**.** Since every internal vertex on an inducing path must be a collider, A must be an endpoint of π. Without loss of generality, let the other endpoint be C. Since B is an internal collider on π, we have $B \in An(A) \cup An(C)$. But $A \in An(B)$ by Fact 2, and acyclicity rules out $B \in An(A)$. Hence $B \in An(C)$.

By Fact 2, there exists an inducing path p from A to B in G that is into B. Every internal vertex of p is a collider and an ancestor of A or B, and hence an ancestor of A or C. Replacing $A \to B$ in π with p therefore produces an inducing walk from A to C in G, a contradiction.

Case 2: $e = A \leftrightarrow B$**.** Let the endpoints of π be X and Y. Unlike Case 1, the edge e may occur anywhere along π.

By Fact 3, there exists an inducing path p between A and B in G that is into both endpoints.

Since A and B lie on the inducing path π, each is either an endpoint of π or an internal collider that is an ancestor of X or Y. Consequently, every internal vertex of p, being an ancestor of A or B, is also an ancestor of X or Y, or equals one of these endpoints.

Replacing $A \leftrightarrow B$ in π with p preserves the collider status at A and B whenever they are internal, because p has arrowheads at both endpoints. All newly introduced internal vertices are also colliders satisfying the required ancestral condition.

Thus, the replacement yields an inducing walk between X and Y in G. Removing repeated vertices from this walk yields an inducing path, contradicting the assumption.

In both cases, adding e introduces no new inducing-path adjacencies. Since the existing inducing paths and all ancestral relations are preserved, the induced MAG remains M. Repeating this argument for every edge in $E' \subseteq E(M) \setminus E(G)$ completes the proof. □

Lemma 6 (SMD-to-DBD Completion)**.** *Let $\mathcal{P}$ be a PAG and let $\mathcal{G}$ be an SMD represented by $\mathcal{P}$, inducing a MAG $M \in [\mathcal{P}]$. Then there exists a DBD $\mathcal{G}' \supseteq \mathcal{G}$ that induces a MAG $M' \in [\mathcal{P}]$. Moreover, M' can be chosen to be an LEG of M.*

Proof. The construction proceeds in two steps.

Step 1: Edge matching. Add to $\mathcal{G}$ every edge of M that is not already present in $\mathcal{G}$, and denote the resulting graph by $\widetilde{\mathcal{G}}$. By Lemma 5, $\widetilde{\mathcal{G}}$ still induces M.

Step 2: Directed-backbone completion. If $\widetilde{\mathcal{G}}$ is already a DBD, the claim follows. Otherwise, let M' be an LEG of M. By the transformational characterization of Zhang and Spirtes, M' can be obtained from M by replacing a sequence of admissible bidirected edges

$$A \leftrightarrow B$$

with directed edges

$$A \to B \quad \text{or} \quad A \leftarrow B,$$

while remaining in the same Markov equivalence class.

Add the corresponding directed edges to $\widetilde{\mathcal{G}}$. Since $\widetilde{\mathcal{G}}$ already contains the edges of M, these additions introduce precisely the ancestral relations encoded by the corresponding mark changes from M to M'. Denote the resulting causal diagram by $\mathcal{G}'$. Then $\mathcal{G}'$ induces M' and satisfies the directed-backbone requirement; hence $\mathcal{G}'$ is a DBD.

Finally, M' is Markov equivalent to M, so $M' \in [\mathcal{P}]$. Therefore, $\mathcal{G}'$ is a DBD represented by $\mathcal{P}$ and $\mathcal{G}' \supseteq \mathcal{G}$. □

Lemma 7 (MEC Supergraph Completion)**.** *For any causal diagram $\mathcal{G}$ in the MEC of a PAG $\mathcal{P}$, there exist a skeleton-matched causal diagram $\mathcal{G}^{SMD}$ and a directed-backboned causal diagram $\mathcal{G}^{DBD}$ in the same MEC such that*

$$\mathcal{G} \subseteq \mathcal{G}^{SMD} \subseteq \mathcal{G}^{DBD}.$$

Proof. Let M be the MAG induced by $\mathcal{G}$. By Lemma 5, adding all edges of M missing from $\mathcal{G}$ yields a supergraph $\mathcal{G}^{SMD} \supseteq \mathcal{G}$ in the same MEC. Since M shares the same skeleton as $\mathcal{P}$, the resulting causal diagram $\mathcal{G}^{SMD}$ is an SMD. By Lemma 6, there exists a DBD $\mathcal{G}^{DBD} \supseteq \mathcal{G}^{SMD}$ in the same MEC, establishing the result. □

C.2 Lemmas over SCMs

Lemma 8 (Subgraph-Supergraph Expressiveness)**.** *Let $\mathcal{G}$ and $\mathcal{G}'$ be two causal diagrams over the same set of observed variables $\mathbf{V}$, with $\mathcal{G} \subseteq \mathcal{G}'$. For every SCM $\mathcal{M}$ inducing $\mathcal{G}$, there exists an SCM $\mathcal{M}'$ inducing $\mathcal{G}'$ such that $\mathbf{P}^{\mathcal{L}_3}(\mathcal{M}') = \mathbf{P}^{\mathcal{L}_3}(\mathcal{M})$.*

Proof. Any SCM compatible with $\mathcal{G}$ can be represented over $\mathcal{G}'$ by assigning zero functional influence to the additional parents and latent dependencies (e.g., through terms of the form $0 \cdot V_i$). This preserves all counterfactual distributions, establishing the result. □

Connection between PAG-NCM and SCM

For every $\mathcal{P}$-NCM, there exists a corresponding SCM that preserves both the induced causal diagram and the counterfactual distributions $\mathbf{P}^{\mathcal{L}_3}$, as defined below.

Definition 35 ($\mathcal{P}$-NCM to Standard SCM)**.** *Let $\mathcal{N}(\boldsymbol{\theta}, \mathbf{m}^D, \mathbf{m}^B, \mathbf{m}^C) = \langle \mathbf{V}, \mathbf{U}, P(\mathbf{U}), \mathcal{F} \rangle$ be a $\mathcal{P}$-NCM. Its induced SCM $\mathcal{M} = \langle \mathbf{U}', \mathbf{V}', \mathcal{F}', P(\mathbf{U}') \rangle$ is constructed as follows:*

- *Let $\mathbf{V}' = \mathbf{V}$, $\mathbf{U}' = \mathbf{U}$, and $P(\mathbf{U}') = P(\mathbf{U})$.*
- *For each $V_i \in \mathbf{V}$, define the observed and latent parent sets as*

$$\mathbf{Pa}_i := \{V_j \in \mathbf{V} : \mathbf{m}^D_{ji} > 0\}, \mathbf{U}_i := \{U_S \in \mathbf{U} : \mathbf{m}^C_{Si} > 0\}.$$

- *Construct $f'_i \in \mathcal{F}'$ by restricting inputs to the observed and latent parent sets, such that:*

$$f'_i(\mathbf{pa}_i, \mathbf{u}_i; \theta_i) := f_i\big(\mathbf{v},\ \mathbf{u}; \mathbf{m}^D_{:,i}, \mathbf{m}^C_{:,i}, \theta_i\big).$$

Lemma 9 ($\mathcal{P}$-NCM–SCM $\mathcal{L}_3$ Equivalence)**.** *The SCM $\mathcal{M}$ constructed from a $\mathcal{P}$-NCM via Def .35 induces the same causal diagram $\mathcal{G}$ and counterfactual distributions $\mathbf{P}^{\mathcal{L}_3}$.*

Proof. Let $\mathcal{G}$ denote the causal diagram induced by the $\mathcal{P}$-NCM, and let $\mathcal{G}'$ denote the causal diagram induced by $\mathcal{M}$.

By construction, $\mathcal{M}$ inherits the same set of endogenous variables $\mathbf{V}$, which correspond to the nodes in both $\mathcal{G}$ and $\mathcal{G}'$.

For directed edges, $\mathcal{M}$ inherits as endogenous parents of each V_i exactly those variables V_j with $\mathbf{m}^D_{ji} > 0$. By Def. 10, these are precisely the variables that induce directed edges $V_j \to V_i$ in $\mathcal{G}$. Since directed edges in $\mathcal{G}'$ are also determined by parent relationships (Def. 19), it follows that $\mathcal{G}$ and $\mathcal{G}'$ share the same directed edges.

For bidirected edges, recall that the exogenous variables in the P-NCM are mutually independent. Hence, two endogenous variables V_i, V_j share a bidirected edge in G' if and only if they have a common exogenous parent U_S. We show that this is equivalent to $m^B_{ij} > 0$.

First, suppose that V_i and V_j share an exogenous parent U_S. By Def. 35, $m^C_{Si} > 0$ and $m^C_{Sj} > 0$. By the clique-mask construction in Def. 8, both variables belong to S, and

$$m^C_{Si} = \prod_{\substack{V_k, V_l \in S \\ k<l}} m^B_{kl} > 0.$$

In particular, $m^B_{ij} > 0$, so $V_i \leftrightarrow V_j$ is present in G.

Conversely, suppose $m^B_{ij} > 0$. By the PAG-induced parameterization in Def. 4, the pair $\{V_i, V_j\}$ is an admissible bidirected clique in Def. 7. Thus $U_{\{V_i,V_j\}} \in \mathbf{U}$, with

$$m^C_{\{V_i,V_j\},i} = m^C_{\{V_i,V_j\},j} = m^B_{ij} > 0.$$

Hence $U_{\{V_i,V_j\}}$ is a common exogenous parent of V_i and V_j in M, inducing $V_i \leftrightarrow V_j$ in G'.

Therefore, G and G' have identical bidirected edges. Together with the directed edge correspondence established above, this yields $G = G'$.

Fix a topological ordering $\prec$ of $\mathcal{G}$ consistent with $\mathbf{m}^D$. Consider any set of variables $\mathbf{X}, \mathbf{Z}, \mathbf{W}, \mathbf{Y} \subseteq \mathbf{V}$ and assignment $\mathbf{x} \in \text{dom}(\mathbf{X}), \mathbf{w} \in \text{dom}(\mathbf{W})$, etc. We have, for any value $\mathbf{v} \in \text{dom}(\mathbf{V})$,

$$
\begin{aligned}
&\mathbf{P}^{\mathcal{N},\mathbf{m}^D,\mathbf{m}^C}(\mathbf{y}_{\mathbf{x}},\ldots,\mathbf{z}_{\mathbf{w}}) \\
&= \int_{\mathbf{u}\in\mathrm{dom}(\mathbf{U})} \mathbf{1}[\mathbf{Y}_{\mathbf{x}}(\mathbf{u};\mathbf{m}^C)=\mathbf{y},\ldots,\mathbf{Z}_{\mathbf{w}}(\mathbf{u};\mathbf{m}^C)=\mathbf{z}]\,dP(\mathbf{u}) \\
&= \int_{\mathbf{u}\in\mathrm{dom}(\mathbf{U})} \left(\prod_{V_i\in\mathbf{Y}\setminus\mathbf{X}} \mathbf{1}\left[f_i\left(\{v_j : m^D_{ji}>0,\ j\prec i\},\{u_S : m^C_{Si}>0\};\theta_i\right)=v_i\right]\right)\left(\prod_{V_i\in\mathbf{X}}\mathbf{1}[V_i=x_i]\right) \\
&\qquad \left(\prod_{V_j\in\mathbf{Z}\setminus\mathbf{W}} \mathbf{1}\left[f_j\left(\{v_l : m^D_{lj}>0,\ l\prec j\},\{u_S : m^C_{Sj}>0\};\theta_j\right)=v_j\right]\right)\left(\prod_{V_j\in\mathbf{W}}\mathbf{1}[V_j=w_j]\right)dP(\mathbf{u}) \\
&\qquad\qquad\qquad

We define

- $P(\mathbf{U}') = P(\widehat{\mathbf{U}})P(\mathbf{U}'\backslash\widehat{\mathbf{U}})$ where variables in $\mathbf{U}'\backslash\widehat{\mathbf{U}}$ also follow $Unif(0,1)$.
- $\boldsymbol{\theta}'$ such that, for each variable $V_i \in \mathbf{V}$,

$$f_i'(\mathbf{v}', \mathbf{u}'; \mathbf{m}^D, \mathbf{m}^C, \boldsymbol{\theta}') = \widehat{f}_i(\mathbf{p}a_i, \mathbf{u}_i; \boldsymbol{\theta}).$$

  Such $\boldsymbol{\theta}'$ exists by expressiveness (Def. 9(ii)), since the unmasked inputs of f_i' are exactly $\mathbf{Pa}_i$ and the U_S with $V_i \in S$ and S a bidirected clique of $\mathcal{G}$, which include $\mathbf{U}_i$.

Then, for any set of variables $\mathbf{X}, \mathbf{Z}, \mathbf{W}, \mathbf{Y} \subseteq \mathbf{V}$ and assignment $\mathbf{x} \in \text{dom}(\mathbf{X}), \mathbf{w} \in \text{dom}(\mathbf{W})$, etc. We have, for any value $\mathbf{v} \in \text{dom}(\mathbf{V})$,

$$\begin{aligned}
&\mathbf{P}^{\mathcal{N}(\mathbf{m}^D, \mathbf{m}^B, \mathbf{m}^C, \boldsymbol{\theta}')}(\mathbf{y}_{\mathbf{x}}, \ldots, \mathbf{z}_{\mathbf{w}}) \\
&= \int_{\mathbf{u}'} \mathbf{1}\Big[\mathbf{Y}_{\mathbf{x}}^{\boldsymbol{\theta}'}(\mathbf{u}'; \mathbf{m}^C) = \mathbf{y}, \ldots, \mathbf{Z}_{\mathbf{w}}^{\boldsymbol{\theta}'}(\mathbf{u}'; \mathbf{m}^C) = \mathbf{z}\Big]\, dP(\mathbf{u}') \\
&= \int_{\mathbf{u}'} \left(\prod_{V_i \in \mathbf{Y}\backslash\mathbf{X}} \mathbf{1}\big[f_i'\left(\{v_l' : m_{li}^D > 0\}, \{u_S' : m_{Si}^C > 0\}; \theta_i'\right) = v_i\big]\right)\left(\prod_{V_i \in \mathbf{X}} \mathbf{1}[V_i = x_i]\right) \\
&\qquad \cdot \left(\prod_{V_j \in \mathbf{Z}\backslash\mathbf{W}} \mathbf{1}\big[f_j'\left(\{v_l' : m_{lj}^D > 0\}, \{u_S' : m_{Sj}^C > 0\}; \theta_j'\right) = v_j\big]\right)\left(\prod_{V_j \in \mathbf{W}} \mathbf{1}[V_j = w_j]\right) dP(\mathbf{u}') \\
&\qquad\qquad \text{(By Def. 11)} \\
&= \int_{\mathbf{u}} \left(\prod_{V_i \in \mathbf{Y}\backslash\mathbf{X}} \mathbf{1}[f_i(\mathbf{pa}_i, \mathbf{u}_i; \theta_i) = v_i]\right)\left(\prod_{V_i \in \mathbf{X}} \mathbf{1}[V_i = x_i]\right) \\
&\qquad \cdot \left(\prod_{V_j \in \mathbf{Z}\backslash\mathbf{W}} \mathbf{1}\big[f_j(\mathbf{pa}_j, \mathbf{u}_j; \theta_j) = v_j\big]\right)\left(\prod_{V_j \in \mathbf{W}} \mathbf{1}[V_j = w_j]\right) dP(\mathbf{u}) \int_{\mathbf{u}'\backslash\mathbf{u}} dP(\mathbf{u}' \setminus \mathbf{u}) \\
&\qquad\qquad \text{(By definition of } \mathcal{F}' \text{ and } P(\mathbf{U}')) \\
&= \mathbf{P}^{\mathcal{M}}(\mathbf{y}_{\mathbf{x}}, \ldots, \mathbf{z}_{\mathbf{w}}). \qquad \text{(By Def. 22)}
\end{aligned}$$

□

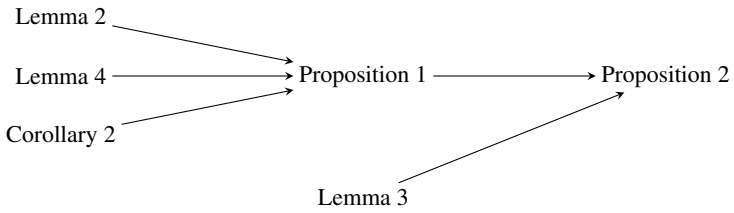


Figure 10: Roadmap to the MEC characterization results.

D Proofs

Proposition 1. *[Characterization of Skeleton-Matched Causal Diagrams Represented by a PAG]*

Let $\mathcal{P}$ be a PAG with MEC signature $\langle \mathfrak{S}, \mathfrak{C}, \mathfrak{D} \rangle$. A graph $\mathcal{G}$ is a skeleton-matched causal diagram represented by $\mathcal{P}$ if and only if the following conditions hold:

1. ***Ordinary Adjacency:*** *Variables adjacent in $\mathcal{P}$ are adjacent in $\mathcal{G}$;*
2. ***Projected Adjacency:*** *Variables not adjacent in $\mathcal{P}$ are not connected by inducing paths in $\mathcal{G}$;*
3. ***Acyclicity:*** *$\mathcal{G}$ contains no directed cycles;*
4. ***Collider Triples:*** *Every triple in $\mathfrak{C}$ is a collider in $\mathcal{G}$;*
5. ***Non-collider Triples:*** *Every triple in $\mathfrak{D}$ is a non-collider in $\mathcal{G}$.*

Here, a triple $\langle X, Y, Z \rangle$ is a collider *in $\mathcal{G}$ if there exists an edge between X and Y and an edge between Y and Z that both have arrowheads at Y. Otherwise, the triple is a* non-collider.

Proof. First, we show that each condition is necessary.

1. If Condition (1) fails, then some adjacency in $\mathcal{P}$ is absent from $\mathcal{G}$. Hence, $\mathcal{G}$ is not skeleton-matched to $\mathcal{P}$ by definition.
2. If Condition (2) fails, then some pair of variables that is non-adjacent in $\mathcal{P}$ is connected by an inducing path in $\mathcal{G}$. The corresponding variables are therefore adjacent in the MAG $M_{\mathcal{G}}$ induced by $\mathcal{G}$. Hence, $M_{\mathcal{G}}$ cannot belong to the MEC represented by $\mathcal{P}$ by Lemma 2. Since the MEC membership of $\mathcal{G}$ is determined by that of its induced MAG $M_{\mathcal{G}}$, $\mathcal{G}$ cannot be represented by $\mathcal{P}$.
3. If Condition (3) fails, then $\mathcal{G}$ contains a directed cycle and is therefore not a causal diagram by definition.
4. If Condition (4) fails, then some triple in $\mathfrak{C}$ is not a collider in $\mathcal{G}$. By Corollary 2, any skeleton-matched causal diagram represented by $\mathcal{P}$ must realize every triple in $\mathfrak{C}$ as a collider. Hence, $\mathcal{G}$ cannot be represented by $\mathcal{P}$.
5. If Condition (5) fails, then some triple in $\mathfrak{D}$ is a collider in $\mathcal{G}$. By Corollary 2, any skeleton-matched causal diagram represented by $\mathcal{P}$ must realize every triple in $\mathfrak{D}$ as a non-collider. Hence, $\mathcal{G}$ cannot be represented by $\mathcal{P}$.

Therefore, any skeleton-matched causal diagram represented by $\mathcal{P}$ must satisfy Conditions (1)–(5).

Next, we show the sufficiency of the conditions.

Conditions (1) and (2) ensure that $\mathcal{G}$ is a skeleton-matched causal diagram. Moreover, Condition (1) ensures that every adjacency in $\mathcal{P}$ is present in $\mathcal{G}$, and hence in $M_{\mathcal{G}}$, while Condition (2) excludes any adjacency in $M_{\mathcal{G}}$ between variables that are non-adjacent in $\mathcal{P}$. Therefore, the induced skeleton matches $\mathfrak{S}$.

It remains to show that $M_{\mathcal{G}}$ has the same ordered collider and non-collider sets as $\mathcal{P}$. By Conditions (4) and (5), every triple in $\mathfrak{C}$ is a collider in $\mathcal{G}$, while every triple in $\mathfrak{D}$ is a non-collider in $\mathcal{G}$. By Lemma 4, an ordered triple of $M_{\mathcal{G}}$ cannot have the opposite collider/non-collider status in $\mathcal{G}$. Hence, whenever the MAG-to-MEC procedure of [6, Alg. 1] considers a candidate ordered triple in $M_{\mathcal{G}}$, its classification agrees with the corresponding classification prescribed by $\mathfrak{C}$ and $\mathfrak{D}$.

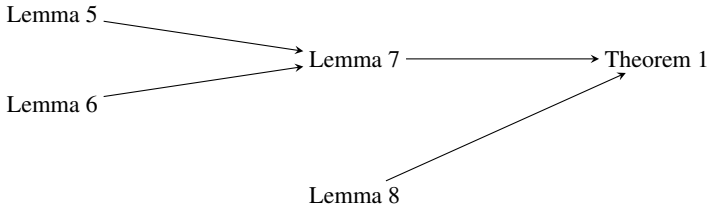


Figure 11: Roadmap to proving the expressiveness of MEC SMDs and DBDs.

Since $M_{\mathcal{G}}$ and $\mathcal{P}$ have the same skeleton, Alg. 1 begins with the same order-0 candidate triples. Recursively, once the previously identified collider and non-collider lists agree, the same higher-order candidates are generated and receive the same classifications. Consequently,

$$\mathfrak{C}(M_{\mathcal{G}}) = \mathfrak{C}, \qquad \mathfrak{D}(M_{\mathcal{G}}) = \mathfrak{D}.$$

Thus,

$$\langle \mathfrak{S}, \mathfrak{C}, \mathfrak{D} \rangle_{M_{\mathcal{G}}} = \langle \mathfrak{S}, \mathfrak{C}, \mathfrak{D} \rangle = \langle \mathfrak{S}, \mathfrak{C}, \mathfrak{D} \rangle_{\mathcal{P}}.$$

By Lemma 2, $M_{\mathcal{G}}$ belongs to the MEC represented by $\mathcal{P}$, and therefore $\mathcal{G}$ is represented by $\mathcal{P}$. □

Proposition 2. *[Characterization of Directed-Backboned Causal Diagrams Represented by a PAG]*

Let $\mathcal{P}$ be a PAG with MEC signature $\langle \mathfrak{S}, \mathfrak{C}, \mathfrak{D} \rangle$. A graph $\mathcal{G}$ is a directed-backboned causal diagram represented by $\mathcal{P}$ if and only if:

1. ***Directed Backbone:** Every adjacency in $\mathcal{P}$ that is not bidirected is represented by a directed edge or bow in $\mathcal{G}$, and every bidirected adjacency in $\mathcal{P}$ is represented by a bidirected edge in $\mathcal{G}$;*
2. *$\mathcal{G}$ satisfies Conditions (2–5) in Proposition 1.*

Proof. The only difference from Proposition 1 is Condition (1), where ordinary adjacency is strengthened to directed-backbone edge matching. The directed-backbone restriction is compatible with the MEC represented by $\mathcal{P}$. In particular, by Lem. 3, all partially directed edges $\circ\!\!\rightarrow$ in a PAG can be oriented as directed edges $\rightarrow$, while the remaining $\circ\!\!-\!\!\circ$ subgraph can be oriented into a DAG without unshielded colliders, such that the resulting MAG remains in the Markov equivalence class represented by $\mathcal{P}$. Thus, enforcing the directed backbone does not introduce additional restrictions on the MEC signature beyond those already captured by Conditions (2)–(5). The remainder follows directly from Proposition 1. □

Theorem 1. *[Expressiveness of MEC SMDs and DBDs] For every SCM $\mathcal{M}$ inducing a causal diagram $\mathcal{G}$ in the MEC of a PAG $\mathcal{P}$, there exist SCMs $\mathcal{M}^{SMD}$ and $\mathcal{M}^{DBD}$ inducing causal diagrams $\mathcal{G}^{SMD}$ and $\mathcal{G}^{DBD}$ in the same MEC, respectively, such that $\mathbf{P}^{\mathcal{L}_3}(\mathcal{M}) = \mathbf{P}^{\mathcal{L}_3}(\mathcal{M}^{SMD}) = \mathbf{P}^{\mathcal{L}_3}(\mathcal{M}^{DBD})$.*

Proof. By Lemma 7, there exist an SMD $\mathcal{G}^{SMD}$ and a DBD $\mathcal{G}^{DBD}$ in the same MEC such that $\mathcal{G} \subseteq \mathcal{G}^{SMD} \subseteq \mathcal{G}^{DBD}$. The result then follows directly from Lemma 8. □

Proposition 3. *[Satisfaction of Directed-Backbone and Collider Conditions] Let $\mathcal{P}$ be a PAG and $\mathcal{G}$ a causal diagram whose adjacency masks $\langle \mathbf{m}^D, \mathbf{m}^B \rangle$ follow the PAG-induced parameterization in Definition 4. Then $\mathcal{G}$ satisfies the directed-backbone and collider conditions in Proposition 2.*

Proof. Given $\mathcal{G}$ is a causal diagram, acyclic holds by definition.

Directed-backbone condition. The parameterization in Definition 4 directly enforces the edge structure required by Condition (1) of Proposition 2, following Lemma 3. In particular, every bidirected edge is enforced by setting $m^B_{ij} = 1$. Similarly, every directed or partially directed edge is enforced in the directed backbone by setting $m^D_{ij} = 1$. For an undetermined edge, the complementary parameterization $m^D_{ij} = 1 - m^D_{ji}$ ensures that exactly one of the two directed orientations is instantiated to comply with acyclicity, while retaining the flexibility permitted by the PAG. Hence, $\mathcal{G}$ satisfies the directed-backbone condition of Proposition 2.

Collider condition. The parameterization also preserves every arrowhead appearing in $\mathcal{P}$. In particular, if $\langle V_i, V_j, V_k \rangle \in \mathfrak{C}$, then the two PAG edges incident to V_j both contain arrowheads at

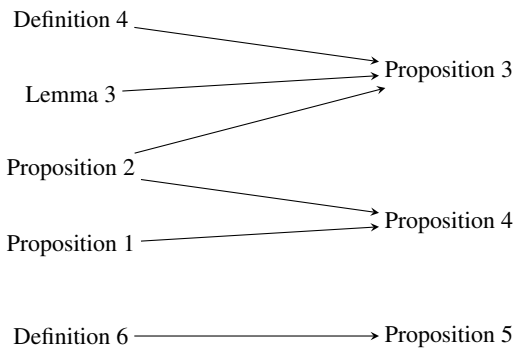


Figure 12: Roadmap to proving the correctness of the PAG-NCM mask parameterization and constraints.

V_j. By Definition 4, each such arrowhead is realized in $\mathcal{G}$ either by a directed edge into V_j or by a bidirected edge incident to V_j. Consequently, both edges of the corresponding triple in $\mathcal{G}$ have arrowheads at V_j, and $\langle V_i, V_j, V_k \rangle$ is a collider in $\mathcal{G}$. Therefore, $\mathcal{G}$ satisfies the collider condition of Proposition 2.

Finally, Definition 4 also sets both adjacency masks to zero for pairs that are non-adjacent in $\mathcal{P}$, thereby excluding certain edges between such pairs. This enforces the length-one component of the projected-adjacency condition. It does not, however, exclude longer inducing paths between non-adjacent variables, which must be ruled out separately. □

Proposition 4. *[Correctness of Non-Collider Triples Constraint] Given a PAG $\mathcal{P}$ with MEC signature $\langle \mathfrak{S}, \mathfrak{C}, \mathfrak{D} \rangle$ and a causal diagram $\mathcal{G}$ with adjacency masks $(\mathbf{m}^D, \mathbf{m}^B)$, $\mathcal{G}$ satisfies the non-collider triples condition in Propositions 1 and 2 if and only if $nc^{\mathcal{P}}(\mathbf{m}^D, \mathbf{m}^B, \mathfrak{D}) = 0$.*

Proof. First, each entry D_{ijk} of the tensor $\mathbf{D}$ indicates whether the ordered triple $\langle V_i, V_j, V_k \rangle$ belongs to $\mathfrak{D}$, i.e., whether the premise of the non-collider constraint in Prop. 1 and 2 is satisfied.

Each entry W_{ij} of the matrix $\mathbf{W}$ indicates whether there is an edge between V_i and V_j with an arrowhead at V_j. In particular, $W_{ij} > 0$ if and only if $\mathbf{m}^D_{ij} > 0$ or $\mathbf{m}^B_{ij} > 0$.

By construction, $H_{ijk} := W_{ij}W_{kj}$ indicates whether both edges $V_i * \!\!\rightarrow V_j$ and $V_k * \!\!\rightarrow V_j$ are present in $\mathcal{G}$, i.e., whether the triple $\langle V_i, V_j, V_k \rangle$ forms a collider at V_j. In particular, $H_{ijk} = 0$ if and only if the triple is not a collider.

Taking the inner product of $\mathbf{H}$ and $\mathbf{D}$, it is nonzero if and only if there exist indices i, j, k such that both $D_{ijk} > 0$ and $H_{ijk} > 0$, i.e., an ordered non-collider triple in $\mathfrak{D}$ is oriented as a collider in $\mathcal{G}$, violating the non-collider triples condition in the propositions. Therefore, the constraint is satisfied if and only if

$$nc^{\mathcal{P}}(\mathbf{m}^D, \mathbf{m}^B) := \langle \mathbf{D}, \mathbf{H} \rangle = 0.$$

□

Proposition 5. *[Correctness of Projected Adjacency Constraint] Given a $\mathcal{P}$ with MEC signature $\langle \mathfrak{S}, \mathfrak{C}, \mathfrak{D} \rangle$ and a causal diagram $\mathcal{G}$ with adjacency masks $(\mathbf{m}^D, \mathbf{m}^B)$, $\mathcal{G}$ satisfies the projected adjacency condition in Propositions 1 and 2 if and only if $p^{\mathcal{P}}(\mathbf{m}^D, \mathbf{m}^B, \mathfrak{S}) = 0$.*

Proof. All matrices in Def. 6 are entry-wise nonnegative. Since $\mathbf{m}^D$ is nilpotent, the matrix exponential satisfies

$$T := e^{\mathbf{m}^D} - I = \sum_{\ell=1}^{n-1} \frac{(\mathbf{m}^D)^\ell}{\ell!}.$$

Hence $T_{uv} > 0$ if and only if there exists a directed path $V_u \rightsquigarrow V_v$. Since the map $x \mapsto 1 - e^{-\lambda x}$ is strictly positive for $x > 0$ and zero at $x = 0$, we have

$$\mathrm{Anc}_{uv} > 0 \quad \Longleftrightarrow \quad V_u \in \mathrm{An}(V_v), \qquad u \neq v.$$

Next, by construction from the weighted adjacency matrices,

$$\mathbf{S}_{uv} > 0 \quad \iff \quad V_u \text{ and } V_v \text{ are adjacent,}$$

and

$$\mathrm{ArrAt}_{uv} > 0 \quad \iff \quad V_u \to V_v \text{ or } V_u \leftrightarrow V_v \text{ (i.e., the edge between } V_u \text{ and } V_v \text{ has an arrowhead at } V_v).$$

Fix a pair of variables (V_i, V_j), the anchor vector with respect to (V_i, V_j) satisfies

$$\alpha_u^{ij} > 0 \quad \iff \quad u = i \text{ or } u = j \text{ or } V_u \in \mathrm{An}(V_i) \cup \mathrm{An}(V_j).$$

Thus, the anchor condition coincides with the ancestry condition, which is automatically satisfied when the node coincides with the end nodes, and otherwise, must be ancestors of either end points if the node is a non-end one.

The endpoint indicators relax the arrowhead requirement at the two endpoints, where

$$\mathrm{Head}_{uv}^{ij} > 0$$

requires an arrowhead at V_v unless $v \in \{i, j\}$, and

$$\mathrm{Tail}_{uv}^{ij} > 0$$

requires an arrowhead at V_u unless $u \in \{i, j\}$.

Consequently,

$$W_{uv}^{ij} > 0$$

if and only if the edge $V_u - V_v$ can be used as one step of a path from V_i to V_j such that every non-endpoint node incident to that edge receives an arrowhead (to form colliders along the path) and every non-endpoint node is an ancestor of V_i or V_j.

Because $\mathbf{W}^{ij}$ is entrywise nonnegative,

$$\left[(\mathbf{W}^{ij})^k\right]_{ij} > 0$$

if and only if there exists a length-k walk from V_i to V_j whose every edge satisfies the above local conditions. Removing repeated cycles from such a walk yields a simple path with the same local collider and ancestry properties. Therefore, for $K \geq n - 1$,

$$\sum_{k=1}^{K} \left[(\mathbf{W}^{ij})^k\right]_{ij} > 0$$

if and only if there exists a path between V_i and V_j whose non-end nodes are all colliders and ancestors of V_i or V_j, i.e., an inducing path. Since $\log(1 + x) > 0$ if and only if $x > 0$, this proves

$$\mathbf{P}_{ij} > 0 \quad \iff \quad \text{there exists an inducing path between } V_i \text{ and } V_j.$$

Finally, since both $\mathbf{P}$ and $\mathfrak{S}^-$ are entrywise nonnegative,

$$p^{\mathcal{P}}(\mathbf{m}^D, \mathbf{m}^B, \mathfrak{S}) = \langle \mathbf{P}, \mathfrak{S}^- \rangle = 0$$

if and only if $\mathbf{P}_{ij} = 0$ for every pair with $\mathfrak{S}_{ij}^- = 1$, equivalently for every pair absent from the skeleton $\mathfrak{S}$ (i.e., not adjacent in the PAG). Hence the constraint equals to zero exactly when no inducing path exists between any pair that is non-adjacent in $\mathfrak{S}$. □

Proposition 6. *[$\mathcal{P}$-NCM – SCM Equivalence] Given a $\mathcal{P}$-NCM $\mathcal{N}(\boldsymbol{\theta}, \mathbf{m}^D, \mathbf{m}^B, \mathbf{m}^C) = \langle \mathbf{V}, \mathbf{U}, P(\mathbf{U}), \mathcal{F} \rangle$ with DBD $\mathcal{G}$ (Def. 10) and counterfactual distributions $\mathbf{P}^{\mathcal{L}_3}(\mathcal{N})$ (Def. 11), there exists an SCM $\mathcal{M}$ inducing the same causal diagram $\mathcal{G}$ with $\mathbf{P}^{\mathcal{L}_3}(\mathcal{M}) = \mathbf{P}^{\mathcal{L}_3}(\mathcal{N})$.*

Proof. The result follows immediately from Lemma 9. □

Theorem 2. *[Expressiveness of $\mathcal{P}$-NCM] For every SCM $\mathcal{M}$ inducing a causal diagram $\mathcal{G}$ in the MEC of a PAG $\mathcal{P}$, there exists a $\mathcal{P}$-NCM $\mathcal{N}(\boldsymbol{\theta}, \mathbf{m}^D, \mathbf{m}^B, \mathbf{m}^C)$ such that $\mathbf{P}^{\mathcal{L}_3}(\mathcal{N}) = \mathbf{P}^{\mathcal{L}_3}(\mathcal{M})$.*

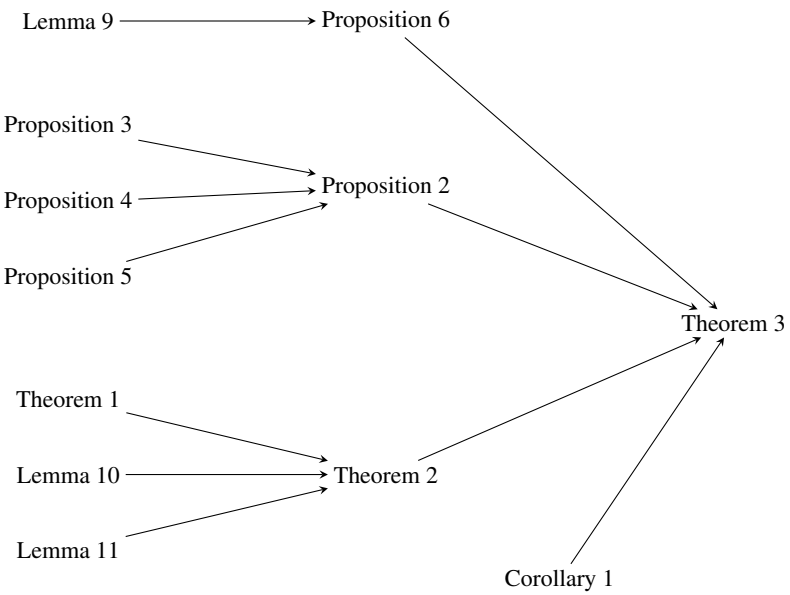


Figure 13: Roadmap to proving the expressiveness of $\mathcal{P}$-NCMs and the correctness of PAG partial identification.

Proof. By Theorem 1, there exists an SCM $\mathcal{M}'$ inducing a DBD $\mathcal{G}' \supseteq \mathcal{G}$ in the same MEC and the same counterfactual distributions $\mathbf{P}^{\mathcal{L}_3}$ as $\mathcal{M}$.

By Lem. 10, there exists a $\mathcal{G}'$-NCM $\widehat{\mathcal{M}}$ that induces the same counterfactual distributions $\mathbf{P}^{\mathcal{L}_3}$ as $\mathcal{M}'$.

Finally, by Lem. 11, there exists a $\mathcal{P}$-NCM $\mathcal{N}$ that induces the same $\mathbf{P}^{\mathcal{L}_3}$ as $\widehat{\mathcal{M}}$. □

Theorem 3. *[Correctness of PAG Neural Partial ID] The true optimal counterfactual bounds (Def. 12) and the optimal neural counterfactual bounds (Def. 13) coincide.*

Proof. By Prop. 6, every $\mathcal{P}$-NCM is equivalent to an SCM inducing the same causal diagram and counterfactual distributions. The constraints on the adjacency masks further ensure that the induced causal diagram belongs to the MEC of $\mathcal{P}$ (Propositions 3, 4, 5). Thus, the feasible distributions represented by $\mathcal{P}$-NCMs form a subset of those induced by SCMs in $\Omega(\mathcal{P})$, ensuring the validity of the computed bounds.

Conversely, by the expressiveness of $\mathcal{P}$-NCMs (Thm. 2), every SCM in $\Omega(\mathcal{P})$ admits a $\mathcal{P}$-NCM representation with identical $\mathcal{L}_3$ distributions. Consequently, the feasible $\mathcal{P}$-NCMs can reproduce the counterfactual distributions attaining the sharp lower and upper bounds.

The soundness of partial ID using NCMs (Corollary 1) further implies that the optimized bounds $[l, r]$ coincide with the sharp counterfactual bounds. □

E Experiments

E.1 Experimental Setup

We implement $\mathcal{P}$-NCMs following the masked-attention architecture and the relaxed training objective of [8], which learns NCMs over the Markov equivalence class of a CPDAG, and refer to that work for details. In brief, each variable V_i has an exogenous token, obtained by embedding a draw of its private exogenous variable $U_{\{V_i\}}$ with a variable-specific MLP, and a value token, obtained by embedding the value of V_i. The tokens pass through a masked attention block in which each mask entry m enters the attention logit additively as $\log m$, so a zero entry removes the corresponding input exactly, as required by mask invariance (Def. 9). A variable-specific output head maps the final exogenous token of V_i to a logit ℓ_i, and $V_i = \mathbb{I}[\varepsilon_i < \sigma(\ell_i)]$, where $\varepsilon_i \sim \mathrm{Uniform}(0,1)$ is treated as part of $U_{\{V_i\}}$, so that f_i is deterministic given $\mathbf{U}$. The observational likelihood is computed by enumerating $\mathrm{dom}(\mathbf{V})$, integrating out each ε_i analytically, and averaging over Monte Carlo draws of the remaining exogenous variables. Since a CPDAG has no latent confounders, we make the following changes for PAGs.

1. **Directed backbone.** [8] fix the directed edges of the CPDAG and learn the two orientations of each undirected edge as separate mask entries, without enforcing adjacency. Following Def. 4, we hard-code all mask entries fixed by $\mathcal{P}$ and parameterize each circle–circle adjacency $V_i \circ\!\!-\!\!\circ V_j$ by a single logit α_{ij}, with $m^D_{ij} = \sigma(\alpha_{ij})$ and $m^D_{ji} = 1 - \sigma(\alpha_{ij})$. Every such adjacency is thus represented by a directed edge in one of its two orientations, as required by the directed-backbone condition of Prop. 2.
2. **Latent confounding.** [8] use one exogenous variable per endogenous variable and forbid attention between exogenous tokens. Following Def. 7, we add one token for every shared exogenous variable U_S with $S \in \mathcal{Q}(\mathcal{P})$ and $|S| \geq 2$, obtained by embedding a draw of U_S. The exogenous token of V_i attends to the token of U_S with gate m^C_{Si} (Def. 8), where each free entry of $\mathbf{m}^B$ is a learnable $m^B_{ij} = \sigma(\beta_{ij})$, and the tokens of shared exogenous variables attend only to themselves. Hence f_i depends on U_S only if $m^C_{Si} > 0$, and two variables share exogenous input only through some U_S that contains both. The input sequence has $|\mathbf{U}(\mathcal{P})| + n$ tokens.
3. **Constraints in the objective.** The relaxed objective of [8] penalizes the acyclicity constraint $h(\mathbf{m})$ and the CPDAG non-collider constraint $c(\mathbf{m})$. We keep h on $\mathbf{m}^D$, replace c by the PAG non-collider constraint $nc^{\mathcal{P}}(\mathbf{m}^D, \mathbf{m}^B, \mathfrak{D})$ (Def. 5), which also counts arrowheads from bidirected edges, and add the projected adjacency constraint $p^{\mathcal{P}}(\mathbf{m}^D, \mathbf{m}^B, \mathfrak{S})$ (Def. 6), which rules out inducing paths between variables that are non-adjacent in $\mathcal{P}$. A CPDAG needs neither, since without latent variables there are no bidirected edges or inducing paths. Collider triples need no penalty, since Def. 4 satisfies them by construction (Prop. 3). To maximize $(+)$ or minimize $(-)$ a query $P(\mathbf{y}_* \mid \mathbf{x}_*)$, we minimize
$$\begin{aligned}\mathcal{L}(\boldsymbol{\theta}, \mathbf{m}) = \mathbb{D}\big(P^{\mathcal{N}(\boldsymbol{\theta},\mathbf{m})}(\mathbf{V}) \,\|\, \hat{P}(\mathbf{V})\big) + \lambda_h\, h(\mathbf{m}^D) + \lambda_{nc}\, nc^{\mathcal{P}}(\mathbf{m}^D, \mathbf{m}^B, \mathfrak{D}) \\ + \lambda_p\, p^{\mathcal{P}}(\mathbf{m}^D, \mathbf{m}^B, \mathfrak{S}) \pm \lambda_q\, \mathbb{D}\big(P^{\mathcal{N}(\boldsymbol{\theta},\mathbf{m})}(\mathbf{y}_* \mid \mathbf{x}_*) \,\|\, \hat{P}(\mathbf{y}_* \mid \mathbf{x}_*)\big),\end{aligned}$$
where $\hat{P}(\mathbf{V})$ is the empirical distribution of the data, $\mathbb{D}$ is a divergence (we use the negative log-likelihood), and $\hat{P}(\mathbf{y}_* \mid \mathbf{x}_*)$ is a synthetic target, e.g., all mass on $\mathbf{y} = 1$ when maximizing $P(\mathbf{y} = 1 \mid \mathbf{x}_*)$.
4. **Counterfactual queries.** [8] consider interventional queries only. We evaluate a counterfactual query $P(y_x \mid x')$ with the valuation of Def. 11, in which the factual and counterfactual worlds share the exogenous draw, i.e., $P(y_x \mid x') = \mathbb{E}_{\mathbf{U}\sim P(\mathbf{U}\mid x')}\big[\mathbb{I}[Y_x(\mathbf{U}) = y]\big]$. We estimate it by self-normalized importance sampling with proposal $P(\mathbf{U})$, weighting each draw by the probability of the factual evidence x'. Variables that are not descendants of X take the same value in both worlds, so they are enumerated jointly in the two worlds rather than integrated out separately in each.
5. **Search over masks.** As in [8], we also search over masks by sampling: a Plackett–Luce distribution over topological orders is trained with REINFORCE, using a reward that combines data fit and the query, while the network parameters are trained on the likelihood alone. Given a sampled order, each $\circ\!\!-\!\!\circ$ adjacency is oriented forward in the order, so every adjacency of $\mathcal{P}$ is kept and the mask is acyclic by construction, and each free entry of $\mathbf{m}^B$ is drawn as a Bernoulli variable.

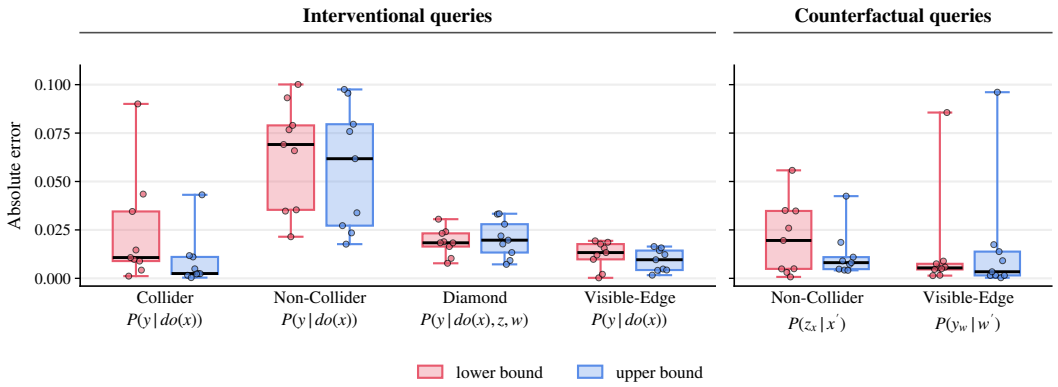


Figure 14: Absolute error between the estimated and ground-truth lower (red) and upper (blue) bounds in Fig. 5. Each box pools three ground-truth SCMs with three runs each; boxes show the median and interquartile range, whiskers the full range, and dots the individual runs.

E.2 Additional Experimental Results

E.2.1 Ground-truth bounds and experimental errors

Experimental errors are plotted in Fig. 14. The theoretical bounds for the interventional and counterfactual queries tested in the experiments are show in Table 1.

PAG	Query	Lower Bound	Upper Bound
Collider	$P(y\|do(x))$	$P(y,x)$	$P(y,x)+1-P(x)$
Non-collider	$P(y\|do(x))$	$P(y,x)$	$P(y,x)+1-P(x)$
Diamond	$P(y\|do(x),w,z)$	$P(y\|x,w,z)$	$P(y\|x,w,z)$
Visible	$P(y\|do(x))$	$P(y\|x)$	$P(y\|x)$
Non-collider	$P(z_x\|x')$	$\min(\min_y P(z\|y), P(z\|x'))$	$\max(\max_y P(z\|y), P(z\|x'))$
Visible	$P(y_w\|w')$	$\min(\min_x P(y\|x), P(y\|w'))$	$\max(\max_x P(y\|x), P(y\|w'))$

Table 1: Theoretical bounds on interventional and counterfactual queries for different PAGs.

E.2.2 Ablation - Partial ID without PAG Constraints

To assess the importance of PAG constraints for partial identification, we compare $\mathcal{P}$-NCM with a PAG-free NCM that fits the same observational data while enforcing only acyclicity, without PAG constraints or a structure score. Although the PAG-free NCM achieves comparable observational fit (KL divergence at most 0.002), its average bound estimation errors increase to 0.20–0.34 for three of the four PAGs (Fig. 15). The collider PAG is an exception, as its bounds coincide with those obtained over unrestricted causal diagrams. For example, on data generated from the visible-edge PAG, the PAG-free NCM learns $W \leftrightarrow Y$ and $W \rightarrow Y$, both of which are excluded by the PAG, while still fitting the observational distribution. These results demonstrate that observational fit alone is insufficient for accurate partial identification, highlighting the importance of incorporating PAG constraints.

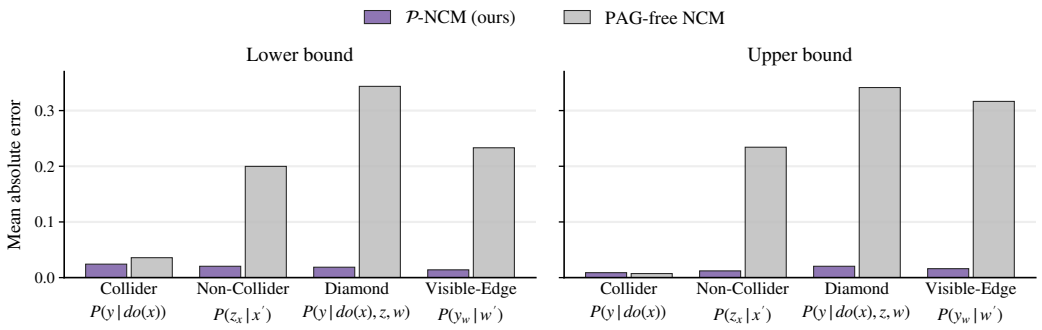


Figure 15: Bound estimation errors with and without PAG constraints. Mean absolute errors of the estimated lower (left) and upper (right) bounds, averaged over three ground-truth SCMs with three independent runs each.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes], [No], or [N/A].
- [N/A] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for [N/A]).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will also be asked to include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While [Yes] is generally preferable to [No], it is perfectly acceptable to answer [No] provided a proper justification is given (e.g., error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering [No] or [N/A] is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Yes, we clearly support our theory with assumptions and proofs and our empirical claims with clearly designed experiments.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

Justification: We discuss potential limitations of the work in terms of its untested empirical scalability on larger models, and suggest it to be a potential line of future work in the discussion

Guidelines:

- The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: Clear statements of theoretical results are provided in the main paper. Proofs are provided in the supplement.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

Justification: Experimental details are provided in the appendix, with the broader experimental setup provided in the main paper.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We will release the code for the experiments.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.

- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Yes, these details are briefly outlined in the main paper and provided in more specificity in the appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Yes, error bars are provided in the main paper; where limited in the main paper for legibility, they are provided in the appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Compute details are provided in the appendix.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: This paper comply with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We introduce a new parameterization for neural causal models for supporting causal inference. The results are fairly theoretical, without a direct path to particular negative applications.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: We do not use image generators or scraped datasets.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: We credit the implementations we use and build upon.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: Documentation will be provided alongside our code which we intend to release publicly.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.

- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: No crowdsourcing and research with human subjects

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: No crowdsourcing and research with human subjects

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: We do not use LLMs as a core method in this work.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.