
A Characterization of Latent Variable Causal Models Consistent with Observational Data

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. In this work, we study the equivalence class of structural causal models (SCMs) with latent variables sharing the same set of conditional independencies, as represented by a partial ancestral graph (PAG). Specifically, we provide a new characterization of SCM-induced causal diagrams in this equivalence class using differentiable parameters. Building on this characterization, we introduce a new model class called PAG-constrained neural causal models (\mathcal{P} -NCMs) which parameterize this equivalence class of SCMs. We prove that \mathcal{P} -NCMs are expressive enough to represent counterfactual distributions induced by any SCM in the class, while remaining consistent with the structural constraints shared across all its members. Finally, we demonstrate how the differentiable parameterization of this model class enables causal inference under Markov equivalence by reducing counterfactual partial identification to an optimization problem over \mathcal{P} -NCMs. We establish the theoretical soundness of this approach and validate its performance on simulations.

1 Introduction

Despite recent advances in artificial intelligence and large language models, the ability to represent, learn, and reason about causal relationships remains a fundamental challenge [18, 19]. As models grow in size, complexity, and data scale, it is still unclear how to enable them to reason reliably as causal agents [20]. This limitation is particularly evident in tasks that require reasoning beyond observed correlations, such as counterfactual reasoning and decision-making under uncertainty [23, 8]. This challenge echoes Nancy Cartwright’s well-known dictum, “no causes in, no causes out” [5], highlighting that meaningful causal inference requires explicit assumptions about the underlying data-generating process. Understanding these assumptions is therefore essential for determining both the scope and the limits of the inferences a model can support.

To make causal assumptions explicit, we must understand how they arise and how they are represented. The true causal mechanism underlying any phenomenon can be formalized as a *Structural Causal Model (SCM)* [18], which induces distributions across three layers of the ladder of causation—observational, interventional, and counterfactual, collectively known as the *Pearl Causal Hierarchy (PCH)* [4]. An SCM also induces a graphical representation, the *causal diagram*, which encodes structural constraints linking these quantities. The compatibility between the PCH distributions and the causal diagram is formalized through *graphical causal models*, which provide a compact encoding of causal assumptions and serve as an essential component of the causal inference pipeline, as illustrated in Fig. 1. When the true underlying SCM is unknown, these causal assumptions play a central role in bridging the gap between the available data and the target causal queries.

* Equal contribution.

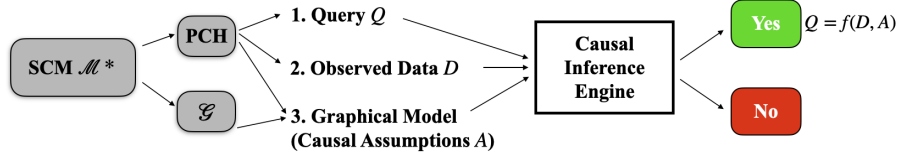


Figure 1: Unobserved SCM and the causal inference engine. The engine takes as input a query, a model, and datasets, and returns whether the query is computable from the assumptions and data.

In many real-world applications across science and machine learning, however, only observational data are available, while even the causal diagram is unknown. A large body of work therefore focuses on learning causal structure from data prior to performing inference [24, 19]. In general, without strong 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) [24, 28].

Given a learned PAG, a natural goal is to use the structural invariances it encodes to enable causal inference beyond the observational level [27, 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 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 [27, 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 differentiable 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, and translate this characterization into equivalent differentiable constraints over the weighted adjacency matrices of these diagrams.
2. We introduce PAG-constrained neural causal models (\mathcal{P} -NCMs), which incorporate these differentiable constraints to parameterize the space of SCMs in the MEC of a PAG. We prove that this model class has universal expressiveness over categorical counterfactual distributions induced by SCMs in the MEC.
3. 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 prove the theoretical soundness of this approach and empirically validate its performance on simulated causal bounding tasks.

2 Preliminaries

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].

An SCM \mathcal{M} induces a *causal diagram* \mathcal{G} where \mathbf{V} is the set of vertices, and there is a directed edge ($V_i \rightarrow V_j$) for every $V_i, V_j \in \mathbf{V}$ if V_i appears as an argument of f_j , and a bidirected edge ($V_i \leftrightarrow V_j$) 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: BN for \mathcal{L}_1 , CBN for \mathcal{L}_2 , and CTFBN for \mathcal{L}_3 [17, 4, 7]. In particular, CBNs admit equivalent global and local characterizations, where the graphical structure can be understood through truncated factorization, modularity, and local causal invariance conditions [3].

ADMGs, MAGs, and PAGs. A *directed mixed graph* (DMG) is a graph whose edges may be either directed (\rightarrow) or bidirected (\leftrightarrow). An *acyclic directed mixed graph* (ADMG) is a DMG with no directed cycles. In this paper, we consider ADMGs allowing bows (i.e., directed and bidirected edges between the same variables), which coincide with causal diagrams induced by SCMs. We will use both terms interchangeably. 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. A *maximal ancestral graph* (MAG) encodes a set of causal diagrams over the same observed variables that entail identical conditional independence and ancestral relations [22]. 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 [24, 28]. PAGs can be learned using the FCI algorithm from conditional independence information from an observation distribution, or from an SCM-induced causal diagram [24, 28, 27]. In this work, we assume access to the true PAG compatible with the causal diagram.

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, 26]. 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}_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 PAGs

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 [21, 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 differentiable constraints over the weighted adjacency matrices of the causal diagrams.

3.1 PAG MEC Signatures and Structural Characterization

Given a PAG \mathcal{P} , we define its MEC signature in terms of ordered triples, which coincides with the invariants shared by all MAGs in the equivalence class [6].

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

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

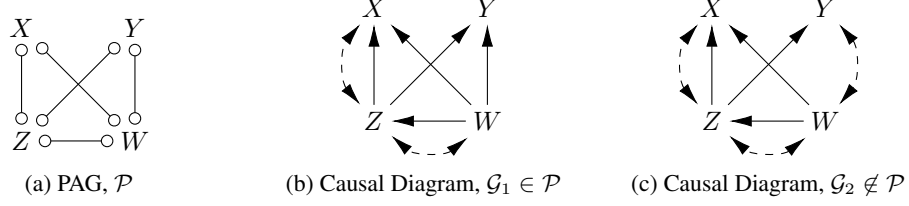


Figure 2: A PAG \mathcal{P} (a) and two causal diagrams induced by different mask assignments of the \mathcal{P} -NCM. The causal diagram in (b) is consistent with the MEC of \mathcal{P} , whereas the causal diagram in (c) falls outside the MEC due to the inducing path $\langle X, Z, W, Y \rangle$.

1. $a * - * b * - * c$ is a collider (resp. non-collider) in \mathcal{P} , and
2. $\exists q : \langle q, a, b \rangle \in \mathfrak{C}_{j < i}$ and $\langle q, a, c \rangle \in \mathfrak{D}_{k < i}$.

Definition 2 (PAG MEC Signature). *The MEC signature of a PAG \mathcal{P} is the triplet $\langle \mathfrak{S}, \mathfrak{C}, \mathfrak{D} \rangle$, where \mathfrak{S} is the (undirected) skeleton of \mathcal{P} , and \mathfrak{C} and \mathfrak{D} are the sets of ordered collider and ordered non-collider triples (Def. 1) shared by all MAGs $M \in \mathcal{P}$.*

Lemma 1. [MAG – PAG MEC Connection] *Given a PAG \mathcal{P} and a MAG $M \in \mathcal{P}$, the MEC signature output from the MAG-to-MEC algorithm ([6, Alg. 1]) is identical when applied to either \mathcal{P} or M .*

Example 1 (PAG MEC Signature). *Consider the PAG \mathcal{P} in Fig. 2(a). Its MEC signature will have $\mathfrak{S} = \{\langle X, Z \rangle, \langle X, W \rangle, \langle Z, W \rangle, \langle Z, Y \rangle, \langle W, Y \rangle\}$, $\mathfrak{C} = \emptyset$, and $\mathfrak{D} = \{\langle X, Z, Y \rangle, \langle X, W, Y \rangle\}$. ■*

The MEC signature provides a complete set of structural invariants that characterize the equivalence class. Using this signature, we characterize the set of causal diagrams in the MEC of a given PAG.

Theorem 1. [Characterization of Causal Diagrams Represented by a PAG] *A causal diagram \mathcal{G} lies in the MEC represented by a PAG \mathcal{P} if and only if the following conditions hold:*

1. **Acyclicity:** \mathcal{G} contains no directed cycles;
2. **Skeleton:** Variables adjacent in \mathcal{P} if and only if connected by an inducing path in \mathcal{G} ;
3. **Collider constraints:** Every triple in \mathfrak{C} that appears in \mathcal{G} is a collider in \mathcal{G} ;
4. **Non-collider constraints:** Every triple in \mathfrak{D} that appears in \mathcal{G} is a non-collider in \mathcal{G} .

3.2 Constraints on Weighted Adjacency Matrices

The characterization above provides a direct way to describe the space of SCMs compatible with a PAG through causal diagrams they induce. However, the resulting space over causal diagrams is discrete and therefore difficult to explore efficiently during inference. Fortunately, each condition in the characterization admits an equivalent differentiable parameterization over weighted adjacency matrices $\langle \mathbf{m}^D, \mathbf{m}^B \rangle$, defined as follows:

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

With the definition above, constraints in Thm. 1 can be translated into constraints over the weighted adjacency matrices. In particular, we formulate the non-collider and inducing-path constraints in this representation and establish their soundness.

Definition 4 (Non-collider Constraint). *Given a PAG \mathcal{P} with MEC signature $\langle \mathfrak{S}, \mathfrak{C}, \mathfrak{D} \rangle$ and a causal diagram \mathcal{G} with weighted adjacency matrices $(\mathbf{m}^D, \mathbf{m}^B)$, the non-collider constraint in \mathcal{G} with respect to \mathcal{P} is defined as*

$$nc^{\mathcal{P}}(\mathbf{m}^D, \mathbf{m}^B, \mathfrak{D}) := \langle \mathbf{D}, \mathbf{H} \rangle, \quad (1)$$

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 1. [Correctness of Non-collider Constraint] *Given a PAG \mathcal{P} with MEC signature $\langle \mathfrak{S}, \mathfrak{C}, \mathfrak{D} \rangle$ and a causal diagram \mathcal{G} with weighted adjacency matrices $(\mathbf{m}^D, \mathbf{m}^B)$, \mathcal{G} satisfies the non-collider constraints with respect to \mathcal{P} in Thm. 1 if and only if $nc^{\mathcal{P}}(\mathbf{m}^D, \mathbf{m}^B, \mathfrak{D}) = 0$.*

Example 2 (Non-collider constraint). Consider the \mathcal{P} in Fig. 2(a) with MEC signature from Example 1, and the causal diagram \mathcal{G}_1 in Fig. 2(b). The non-collider constraint in \mathcal{G}_1 with respect to \mathcal{P} will have $D_{ijk} = 0$ except $D_{123} = 1$, $D_{234} = 1$ and

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

which gives $nc^{\mathcal{P}} = \langle \mathbf{D}, \mathbf{H} \rangle = D_{123}H_{123} + D_{234}H_{234} = 1 \cdot 0 + 1 \cdot 0 = 0$, and implies \mathcal{G}_1 satisfies the non-collider constraint with respect to \mathcal{P} in Thm. 1. ■

Definition 5 (Inducing path constraint). Given a PAG \mathcal{P} with MEC signature $\langle \mathfrak{S}, \mathfrak{C}, \mathfrak{D} \rangle$ and a causal diagram \mathcal{G} with weighted adjacency matrices $(\mathbf{m}^D, \mathbf{m}^B)$, with \mathbf{m}^D nilpotent, the inducing path constraint in \mathcal{G} with respect to \mathcal{P} is defined as

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

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 $\mathbf{P}_{ij} := \log\left(1 + \sum_{k=1}^K [(\mathbf{W}^{ij})^k]_{ij}\right)$, where $\mathbf{P}_{ij} > 0$ 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 $\mathbf{W}^{ij} = \mathbf{S} \circ (\boldsymbol{\alpha}^{ij}(\boldsymbol{\alpha}^{ij})^\top) \circ \text{Tail}^{ij} \circ \text{Head}^{ij}$, via components defined as follows, where I is the identity and $\mathbf{1}$ is the all-ones vector:

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

Proposition 2. [Correctness of inducing path constraint] Given a \mathcal{P} with MEC signature $\langle \mathfrak{S}, \mathfrak{C}, \mathfrak{D} \rangle$ and a causal diagram \mathcal{G} with weighted adjacency matrices $(\mathbf{m}^D, \mathbf{m}^B)$, \mathcal{G} satisfies the skeleton constraints with respect to \mathcal{P} in Thm. 1 if and only if $p^{\mathcal{P}}(\mathbf{m}^D, \mathbf{m}^B, \mathfrak{S}) = 0$.

Example 3 (Inducing path constraint). Consider again the \mathcal{P} in Fig. 2(a) from Example 5 with MEC signature from Example 1, and the causal diagram \mathcal{G}_1 in Fig. 2(b). Setting $\lambda = 1 - e^{-5}$, the inducing path constraint in \mathcal{G}_1 with respect to \mathcal{P} will have

$$\mathbf{P} \approx \begin{bmatrix} 0 & 1.3829 & 2.3856 & 0 \\ 1.3829 & 0 & 1.0986 & 1.3829 \\ 2.3856 & 1.0986 & 0 & 1.3829 \\ 0 & 1.3829 & 1.3829 & 0 \end{bmatrix} \quad \mathfrak{S}^- = \begin{bmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix}. \quad (4)$$

The entries of \mathfrak{S}^- equal to 1 correspond precisely to pairs of variables that are non-adjacent in \mathcal{P} . Thus, it suffices to check the corresponding entries in \mathbf{P} and ensure they are zero to indicate absence of a forbidden inducing path. In this case, \mathbf{P} is zero at all positions where \mathfrak{S}^- equals 1, and the constraint evaluates to $p^{\mathcal{P}}(\mathbf{m}^D, \mathbf{m}^B, \mathfrak{S}) = \langle \mathbf{P}, \mathfrak{S}^- \rangle = 0$. Thus, \mathcal{G}_1 satisfies the inducing path constraint with respect to \mathcal{P} in Thm. 1.

For the causal diagram \mathcal{G}_2 in Fig. 2(c), its inducing path constraint with respect to \mathcal{P} will have

$$\mathbf{P} \approx \begin{bmatrix} 0 & 1.3829 & 2.3856 & 0.6864 \\ 1.3829 & 0 & 1.0986 & 1.0986 \\ 2.3856 & 1.0986 & 0 & 1.3829 \\ 0.6864 & 1.0986 & 1.3829 & 0 \end{bmatrix} \quad \mathfrak{S}^- = \begin{bmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix}. \quad (5)$$

In this case, \mathbf{P} is non-zero at positions where \mathfrak{S}^- equals 1, indicating a forbidden inducing path between X and Y . The constraint evaluates to $p^{\mathcal{P}}(\mathbf{m}^D, \mathbf{m}^B, \mathfrak{S}) = \langle \mathbf{P}, \mathfrak{S}^- \rangle = 1.3728$, which implies that \mathcal{G}_2 does not satisfy the inducing path constraint with respect to \mathcal{P} in Thm. 1. ■

Algorithm 1 PAG to U

```

1: Input: a PAG  $\mathcal{P}$  over  $\mathbf{V}$ , and its MEC signature  $\langle \mathfrak{S}, \mathfrak{C}, \mathfrak{D} \rangle$ 
2: Output: a set of exogenous variables  $\mathbf{U}$  for the PAG-NCM
3: Initialize  $\mathbf{U} \leftarrow \emptyset$ 
4: Let  $\mathcal{P}'$  be the graph obtained from  $\mathcal{P}$  by removing all visible edges
5: for each  $V \in \mathbf{V}$  do
6:   Add  $U_V$  to  $\mathbf{U}$ 
7: end for
8: for  $k = 2$  to  $|\mathbf{V}|$  do
9:    $n_k \leftarrow 0$ 
10:  for each  $S \subseteq \mathbf{V}$  such that  $|S| = k$  and  $S$  is a clique in  $\mathcal{P}'$  do
11:    if there is no ordered non-collider triple  $\langle A, B, C \rangle \in \mathfrak{D}$  with  $\{A, B, C\} \subseteq S$  then
12:      Add  $U_S$  to  $\mathbf{U}$ 
13:       $n_k \leftarrow n_k + 1$ 
14:    end if
15:  end for
16:  if  $n_k = 0$  then
17:    break
18:  end if
19: end for
20: Return  $\mathbf{U}$ 

```

4 PAG-Constrained Neural Causal Models

The differentiable parameterization of constraints from the previous section can be incorporated into causal models as additional parameters to account for structural uncertainty among SCMs within the same PAG MEC. In this section, we define this new model class and show that it is sufficiently expressive to match the counterfactual distributions induced by any SCM in the MEC.

Definition 6 (\mathcal{P} -Constrained Neural Causal Model (\mathcal{P} -NCM)). *Let \mathcal{P} be a PAG over variables $\mathbf{V} = \{V_1, \dots, V_n\}$ 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;
- $\mathbf{U} = \{U_S : S \subseteq \mathbf{V} \text{ is a possible bidirected clique in } \mathcal{P}\}$ is a set of exogenous variables constructed via Alg. 1;
- $\mathcal{F} = \{f_i : V_i \in \mathbf{V}\}$ is a collection of structural functions, where

$$V_i \leftarrow f_i(\mathbf{m}_{:,i}^D \odot \mathbf{V}, \mathbf{m}_{:,i}^C \odot \mathbf{U}; \theta_i).$$

Here, \odot denotes the Hadamard product, and each f_i is a feedforward neural network parameterized by θ_i ;

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

The parameters are constrained as follows:

- Directed mask: $m_{ij}^D = \begin{cases} 1, & \text{if } V_i \rightarrow V_j \in \mathcal{P} \text{ or } V_i \circ \rightarrow V_j \in \mathcal{P}, \\ 0, & \text{if } i = j, V_i \leftarrow *V_j \in \mathcal{P}, \text{ or } V_i \not\sim V_j \text{ in } \mathcal{P}, \\ 1 - m_{ji}^D, & \text{if } V_i \circ - \circ V_j \in \mathcal{P}; \end{cases}$
- Bidirected mask: $m_{ij}^B = \begin{cases} 1, & \text{if } V_i \leftrightarrow V_j \in \mathcal{P}, \\ 0, & \text{if } i = j, V_i \xrightarrow{v} V_j \in \mathcal{P}, V_j \xrightarrow{v} V_i \in \mathcal{P}, \text{ or } V_i \not\sim V_j \text{ in } \mathcal{P}; \end{cases}$
- Maximal bidirected clique mask: $m_{S_i}^C = \begin{cases} 0, & \text{if } V_i \notin S, \\ 1, & \text{if } S = \{V_i\}, \\ \prod_{V_j, V_k \in S, j \neq k} m_{jk}^B, & \text{otherwise;} \end{cases}$
- Acyclicity constraint: $h(\mathbf{m}^D) = 0$, where h is a sound differentiable 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 fixed entries in the adjacency matrices correspond to the presence or absence of edges in the causal diagram implied by the PAG. The maximal bidirected clique matrix \mathbf{m}^C is constructed from the bidirected adjacency matrix \mathbf{m}^B and encodes the presence of shared latent confounders among variables within each maximal bidirected clique. The final three constraints correspond to the conditions in Thm. 1. In particular, the acyclicity constraint can be enforced using standard formulations such as NOTEARS. The collider constraints do not appear explicitly in the definition, as they are already enforced by the fixed edge mask parameters specified by the PAG.

Example 4 (\mathcal{P} -NCM). *Consider the PAG \mathcal{P} in Fig. 2(a). The \mathcal{P} -NCM \mathcal{N} will have $\mathbf{V} = \{X, Z, W, Y\}$ and $\mathbf{U} = \{U_X, U_Z, U_W, U_Y, U_{XZ}, U_{XW}, U_{ZW}, U_{WY}, U_{ZY}, U_{XZW}, U_{ZWY}\}$ from Alg. 1. Its mask parameters will be instantiated as*

$$\mathbf{m}^D = \begin{bmatrix} 0 & m_{12}^D & m_{13}^D & 0 \\ 1 - m_{12}^D & 0 & m_{23}^D & m_{24}^D \\ 1 - m_{13}^D & 1 - m_{23}^D & 0 & m_{34}^D \\ 0 & 1 - m_{24}^D & 1 - m_{34}^D & 0 \end{bmatrix} \quad \mathbf{m}^C = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ m_{12}^B & m_{12}^B & 0 & 0 \\ m_{13}^B & 0 & m_{13}^B & 0 \\ 0 & m_{23}^B & m_{23}^B & 0 \\ 0 & 0 & m_{34}^B & m_{34}^B \\ 0 & 0 & 0 & 0 \\ m_{12}^B m_{13}^B m_{23}^B & m_{12}^B m_{13}^B m_{23}^B & m_{12}^B m_{13}^B m_{23}^B & 0 \\ 0 & m_{23}^B m_{24}^B m_{34}^B & m_{23}^B m_{24}^B m_{34}^B & m_{23}^B m_{24}^B m_{34}^B \end{bmatrix} \quad \mathbf{m}^B = \begin{bmatrix} 0 & m_{12}^B & m_{13}^B & 0 \\ m_{12}^B & 0 & m_{23}^B & m_{24}^B \\ m_{13}^B & m_{23}^B & 0 & m_{34}^B \\ 0 & m_{24}^B & m_{34}^B & 0 \end{bmatrix} \quad (6)$$

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. 6. ■

A \mathcal{P} -NCM induces a causal diagram following the construction procedure below.

Definition 7 (\mathcal{P} -NCM Causal Diagram). *Given a \mathcal{P} -NCM $\mathcal{N}(\theta, \mathbf{m}^D, \mathbf{m}^B, \mathbf{m}^C) = \langle \mathbf{V}, \mathbf{U}, P(\mathbf{U}), \mathcal{F} \rangle$, 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 $V_i \rightarrow V_j$ whenever $\mathbf{m}_{ij}^D > 0$; and (3) add a bidirected edge $V_i \leftrightarrow V_j$ whenever $\mathbf{m}_{ij}^B > 0$.*

Example 5 (\mathcal{P} -NCM Causal Diagram). *Consider again the PAG \mathcal{P} in Fig. 2(a) and its \mathcal{P} -NCM \mathcal{N} as instantiated in Example 4. If the free mask parameters in Eq. (6) are assigned with values shown below, it will induce the causal diagram shown in Fig. 2(b), following Def. 7.*

$$\mathbf{m}^D = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix} \quad \mathbf{m}^B = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \quad (\mathbf{m}^C)^\top = \begin{bmatrix} 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \quad (7)$$

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

Definition 8 (\mathcal{P} -NCM \mathcal{L}_3 Valuation). *Given a \mathcal{P} -NCM $\mathcal{N}(\theta, \mathbf{m}^D, \mathbf{m}^B, \mathbf{m}^C) = \langle \mathbf{V}, \mathbf{U}, P(\mathbf{U}), \mathcal{F} \rangle$, for any subsets $\mathbf{Y}, \mathbf{Z}, \dots, \mathbf{X}, \mathbf{W} \subseteq \mathbf{V}$ and values $\mathbf{x}, \mathbf{y}, \mathbf{z}, \mathbf{w}$, the induced counterfactual distribution is*

$$\mathbf{P}^{\mathcal{N}, \mathbf{m}^D, \mathbf{m}^C}(\mathbf{y}_{\mathbf{x}}, \dots, \mathbf{z}_{\mathbf{w}}) = \sum_{\mathbf{u}} \mathbf{1}[\mathbf{Y}_{\mathbf{x}}(\mathbf{m}_{\cdot, \mathbf{Y}}^C \odot \mathbf{u}) = \mathbf{y}, \dots, \mathbf{Z}_{\mathbf{w}}(\mathbf{m}_{\cdot, \mathbf{Z}}^C \odot \mathbf{u}) = \mathbf{z}]P(\mathbf{u}) \quad (8)$$

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 3. [\mathcal{P} -NCM – SCM Equivalence] *Given a \mathcal{P} -NCM $\mathcal{N}(\theta, \mathbf{m}^D, \mathbf{m}^B, \mathbf{m}^C) = \langle \mathbf{V}, \mathbf{U}, P(\mathbf{U}), \mathcal{F} \rangle$ with causal diagram \mathcal{G} (Def. 7) and counterfactual distributions $\mathbf{P}^{\mathcal{L}_3}(\mathcal{N})$ (Def. 8). Then, there exists an SCM \mathcal{M} inducing the same causal diagram \mathcal{G} with $\mathbf{P}^{\mathcal{L}_3}(\mathcal{M}) = \mathbf{P}^{\mathcal{L}_3}(\mathcal{N})$.*

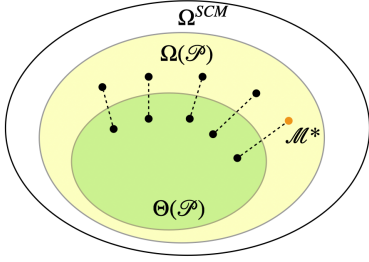


Figure 3: \mathcal{P} -NCM Expressiveness: each SCM $\mathcal{M}^* \in \Omega(\mathcal{P})$ has a \mathcal{P} -NCM $\in \Theta(\mathcal{P})$ matching on $\mathbf{P}^{\mathcal{L}_3}$.

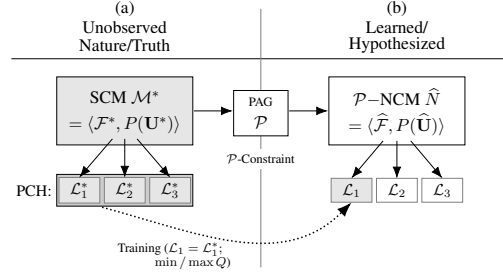


Figure 4: Left: true SCM $\mathcal{M}^* \in \Omega(\mathcal{P})$ that induces PCH's three layers. Right: a \mathcal{P} -NCM \mathcal{N} constrained by inductive bias in \mathcal{P} and matching \mathcal{M}^* on $\mathcal{L}_1 = P(\mathbf{V})$ while minimizing or maximizing a target query Q through training.

By allowing different instantiations of the mask parameters, a \mathcal{P} -NCM can induce different causal diagrams within the MEC represented by \mathcal{P} . Importantly, the resulting model class, $\Theta(\mathcal{P})$, is at least as expressive as the set of SCMs inducing causal diagrams in the MEC, $\Omega(\mathcal{P})$. In particular, every counterfactual distribution induced by an SCM in $\Omega(\mathcal{P})$ can also be represented by a \mathcal{P} -NCM in $\Theta(\mathcal{P})$, as illustrated in Fig. 3 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 $\mathcal{N}(\theta, \mathbf{m}^D, \mathbf{m}^B, \mathbf{m}^C)$ 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 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 9 (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 $P(\mathbf{y}_x, \dots, \mathbf{z}_w)$ is defined as, respectively, the minimum and maximum of the following optimization problem:

$$\begin{aligned} \min / \max_{\mathcal{M} \in \Omega(\mathcal{P})} P^{\mathcal{M}}(\mathbf{y}_x, \dots, \mathbf{z}_w) \\ \text{s.t. } P^{\mathcal{M}}(\mathbf{V}) = P(\mathbf{V}) \end{aligned} \quad (9)$$

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 formulate equivalently using these models, as defined below.

Definition 10 (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 interventional probability $P(\mathbf{y}_x)$ is defined as, respectively, the minimum and maximum of the following optimization problem:

$$\begin{aligned} \min / \max_{\mathcal{N} \in \Theta(\mathcal{P})} P^{\mathcal{N}(\theta, \mathbf{m}^D, \mathbf{m}^B, \mathbf{m}^C)}(\mathbf{y}_x) \\ \text{s.t. } P^{\mathcal{N}(\theta, \mathbf{m}^D, \mathbf{m}^B, \mathbf{m}^C)}(\mathbf{V}) = P(\mathbf{V}), \end{aligned} \quad (10)$$

where $\Theta(\mathcal{P})$ is the set of all \mathcal{P} -NCMs, with mask parameters satisfying constraints in Def. 6.

The key idea is illustrated in Fig.4: 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. 2.

Theorem 3. [Correctness of PAG Neural Partial ID] The true optimal counterfactual bounds (Def. 9) and the optimal neural counterfactual bounds (Def. 10) coincide.

Algorithm 2 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:** $P^{\hat{\mathcal{N}}(\Theta_{min})}(\mathbf{y}_*|\mathbf{x}_*)$, $P^{\hat{\mathcal{N}}(\Theta_{max})}(\mathbf{y}_*|\mathbf{x}_*)$
 - 3: $\hat{\mathcal{N}}(\Theta) \leftarrow \mathcal{P}\text{-NCM}(\mathcal{P}, \mathbf{V})$ with parameters $\Theta = (\theta, \mathbf{m}^D, \mathbf{m}^B, \mathbf{m}^C)$ from Def. 6.
 - 4: $\Theta_{min} \leftarrow \arg \min_{\Theta} P^{\hat{\mathcal{N}}(\theta)}(\mathbf{y}_*|\mathbf{x}_*)$ s.t. $P^{\hat{\mathcal{N}}(\Theta)}(\mathbf{V}) = P(\mathbf{V})$.
 - 5: $\Theta_{max} \leftarrow \arg \max_{\Theta} P^{\hat{\mathcal{N}}(\theta)}(\mathbf{y}_*|\mathbf{x}_*)$ s.t. $P^{\hat{\mathcal{N}}(\Theta)}(\mathbf{V}) = P(\mathbf{V})$.
 - 6: **Return** $P^{\hat{\mathcal{N}}(\Theta_{min})}(\mathbf{y}_*|\mathbf{x}_*)$, $P^{\hat{\mathcal{N}}(\Theta_{max})}(\mathbf{y}_*|\mathbf{x}_*)$
-

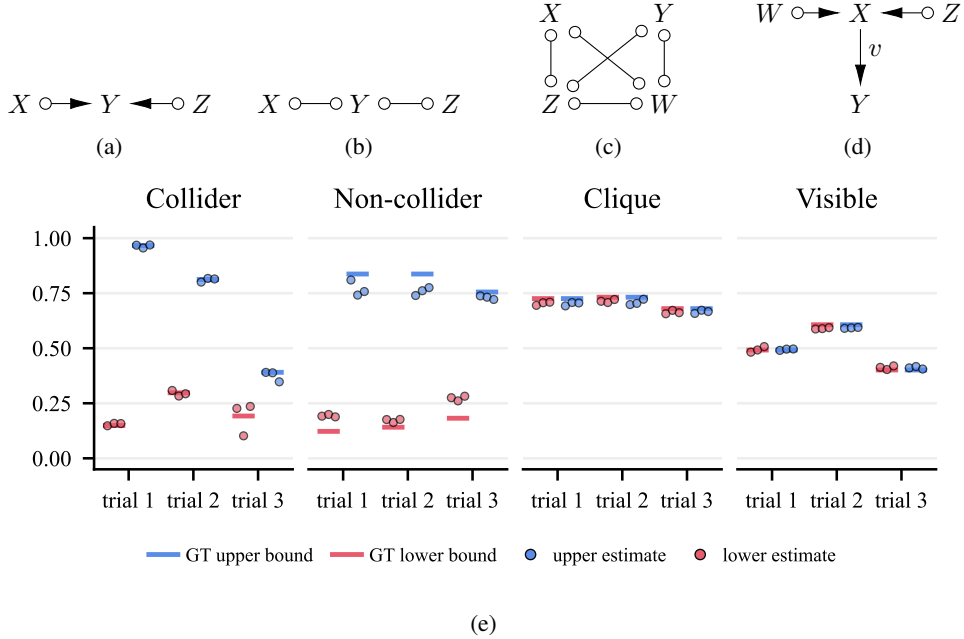


Figure 5: Experimental performance of neural partial identification with \mathcal{P} -NCMs (Alg. 2), demonstrating accurate recovery of ground-truth bounds across all tested PAGs.

5.1 Experiments

We test the counterfactual partial ID based on Alg. 2 using PAGs shown in Fig. 5. In practice, the parameters of the \mathcal{P} -NCM are learned by optimizing a penalized objective consisting of the observational likelihood, a query objective, and penalty terms for constraints in Def. 6 (see Appendix D for details). The target query is $P(y | do(x))$ for PAGs with a collider (Fig. 5(a)), a non-collider (Fig. 5(b)), and a visible edge (Fig. 5(d)), and $P(y | do(x), w, z)$ for the PAG with circle cliques (Fig. 5(c)). The query is identifiable in (c) and (d), while partially identifiable in (a) and (b). In all cases, the algorithm recovers the ground-truth bounds of the query with minimal error.

6 Discussions

In this paper, we introduced a characterization of SCM-induced causal diagrams in the MEC represented by a PAG (Thm. 1), and derived two differentiable constraints from this characterization (Def. 4, Def. 5). These constructions lead to the definition of \mathcal{P} -NCMs (Def. 6), a new class of neural causal models that compactly encode the structural constraints shared across SCMs in the equivalence class. We established the expressiveness of this model class by showing that it can represent counterfactual distributions induced by any SCM in the MEC (Thm. 2), and demonstrated its usefulness for counterfactual partial identification through both theoretical guarantees (Thm. 3) and experiments on simulated data. Several directions remain for future work. First, improving the scalability of the framework is an important challenge. In particular, it would be valuable to empirically evaluate how well the current construction scales to large graphs, given the potentially

exponential number of latent variables in \mathbf{U} (Alg. 1). In addition, the inducing-path constraints may admit more efficient formulations. Second, our framework currently focuses on PAGs derived from observational data. An important extension is to generalize the approach to equivalence classes refined by interventional distributions or other forms of causal knowledge.

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] Nancy Cartwright. *Nature’s Capacities and Their Measurement*. Clarendon Press, Oxford, 1989.
- [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] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.
- [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] Drago Plecko, Patrik Okanovic, Shreyas Havaldar, Torsten Hoefler, and Elias Bareinboim. Epidemiology of large language models: A benchmark for observational distribution knowledge, 2025.

- [21] Thomas Richardson. Markov Properties for Acyclic Directed Mixed Graphs. *Scandinavian Journal of Statistics*, 30(1):145–157, 2003.
- [22] Thomas Richardson and Peter Spirtes. Ancestral Graph Markov Models. *The Annals of Statistics*, 30(4):962–1030, 2002.
- [23] Bernhard Scholkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward Causal Representation Learning. *Proceedings of the IEEE*, 109(5):612–634, May 2021.
- [24] P Spirtes, C N Glymour, and R Scheines. *Causation, Prediction, and Search*. MIT Press, Cambridge, MA, 2nd edition, 2000.
- [25] Jin Tian and Judea Pearl. A General identification condition for causal effects. In *Proceedings of the 18th AAAI Conference on Artificial Intelligence*, 2002.
- [26] 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.
- [27] Jiji Zhang. Causal Reasoning with Ancestral Graphs. *Journal of Machine Learning Research*, 9(47):1437–1474, 2008.
- [28] 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.
- [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.

Appendices

Contents

| | |
|---|-----------|
| A Background and Definitions | 12 |
| A.1 Graph Definitions | 12 |
| A.2 SCMs and Casual Diagrams | 13 |
| A.3 MAG and PAGs | 14 |
| A.4 Partial Identification using DAG-Constrained Neural Causal Models | 15 |
| B Useful Lemmas | 16 |
| B.1 Characterization of Mixed Graphs in PAG MEC | 16 |
| B.1.1 Lemmas on PAG to MAG | 16 |
| B.1.2 Lemmas on MAG to ADMG | 18 |
| B.1.3 Lemmas on PAG to ADMG | 20 |
| B.2 Details on PAG-NCM Masking Parameters | 22 |
| C Proofs | 26 |
| D Experiments | 29 |
| D.1 Experimental Setup | 29 |
| D.1.1 A Relaxed Training Objective | 29 |
| Fitting the observational distribution. | 29 |
| Optimizing the causal query. | 29 |
| D.1.2 Reinforcement learning over the space of masks | 29 |
| D.2 Additional Experimental Results | 30 |

A Background and Definitions

A.1 Graph Definitions

Definition 11 (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 (\rightarrow) and bi-directed edges (\leftrightarrow).*

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

Definition 12 (Confounded Component [25]). *Let C_1, C_2, \dots, C_k be a partition over the set of variables V , where C_i is said to be a confounded component (for short, C-component) of the ADMG \mathcal{G} if for every $V_i, V_j \in C_i$ there exists a path made entirely of bidirected edges between V_i and V_j in \mathcal{G} and C_i is maximal.*

Definition 13 (Augmented Parents). *Let $<$ be a topological order over the variables V_1, \dots, 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, \dots, V_i , and let $C(V_i)$ 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 $C(V_i)$ that comes before V_i in topological order:*

$$Pa_i^+ = \cup_{j|V_j \in T_i} Pa_j \setminus \{V_i\} \tag{11}$$

where $T_i = \{X \in C(V_i) : X \leq V_i\}$.

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 14 (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} :*

$$P(\mathbf{V}) = \prod_i P(v_i | pa_i^{\mathbf{X}^+}) \quad (12)$$

A.2 SCMs and Casual Diagrams

Definition 15 (Structural Casual 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 $\{V_1, V_2, \dots, V_n\}$ 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 $\{f_1, f_2, \dots, f_n\}$ 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 $i = 1, \dots, n$, each $f_i \in \mathcal{F}$ is such that

$$v_i \leftarrow f_i(pa_i, u_i), \quad (13)$$

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 16 (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 $V_i \rightarrow V_j$ for every $V_i, V_j \in \mathbf{V}$ if V_i appears as an argument of f_j
- (3) Add a bidirected edge $V_i \leftrightarrow V_j$ 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 17 (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*

$$\mathcal{M}_{\mathbf{x}} = \langle \mathbf{U}, \mathbf{V}, \mathcal{F}_{\mathbf{x}}, P(\mathbf{U}) \rangle, \quad (14)$$

where

$$\mathcal{F}_{\mathbf{x}} = \{f_i : V_i \notin \mathbf{X}\} \cup \{\mathbf{X} \leftarrow \mathbf{x}\}. \quad (15)$$

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

Definition 18 (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 19 (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 $\mathbf{Y}, \mathbf{Z}, \dots, \mathbf{X}, \mathbf{W} \subseteq \mathbf{V}$, the three layers of distributions are given by:*

- \mathcal{L}_1 (Observational):

$$\mathbf{P}^{\mathcal{M}}(y) = \sum_{\mathbf{u}} \mathbf{1}[\mathbf{Y}(\mathbf{u}) = y]P(\mathbf{u}) \quad (16)$$

- \mathcal{L}_2 (Interventional):

$$\mathbf{P}^{\mathcal{M}}(\mathbf{y}_{\mathbf{x}}) = \sum_{\mathbf{u}} \mathbf{1}[\mathbf{Y}_{\mathbf{x}}(\mathbf{u}) = \mathbf{y}]P(\mathbf{u}) \quad (17)$$

- \mathcal{L}_3 (Counterfactual):

$$\mathbf{P}^{\mathcal{M}}(\mathbf{y}_{\mathbf{x}}, \dots, \mathbf{z}_{\mathbf{w}}) = \sum_{\mathbf{u}} \mathbf{1}[\mathbf{Y}_{\mathbf{x}}(\mathbf{u}) = \mathbf{y}, \dots, \mathbf{Z}_{\mathbf{w}}(\mathbf{u}) = \mathbf{z}]P(\mathbf{u}) \quad (18)$$

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 20 (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 21 (Inducing Path). ¹ *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 $X \rightarrow Y$ in the MAG.*
2. *Out of Y and into X : which makes $Y \in An(X)$ and induces $Y \rightarrow X$ in the MAG.*
3. *Into both X and Y :*
 - (a) *If $X \in An(Y)$, it induces $X \rightarrow Y$ in the MAG;*
 - (b) *If $Y \in An(X)$, it induces $Y \rightarrow X$ in the MAG;*
 - (c) *If $X \notin An(Y)$ and $Y \notin An(X)$, it induces $X \leftrightarrow Y$ in the MAG.*

Definition 22 (Discriminating Path). *A path $p = \langle X, \dots, W, V, Y \rangle$ 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 23 (Visibility (Def. 8 [27])). *Given a MAG M , a directed edge $A \rightarrow B$ 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 ². Otherwise $A \rightarrow B$ is said to be invisible.*

Definition 24 (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;*

¹Single edges are inducing paths too.

²This 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. 23.

Definition 25 (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 26 (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:*

$$\text{Rule 1 } P(\mathbf{y}|\text{do}(\mathbf{x}), \mathbf{z}, \mathbf{w}) = P(\mathbf{y}|\text{do}(\mathbf{x}), \mathbf{w}) \quad \text{if } (\mathbf{Y} \perp\!\!\!\perp \mathbf{Z} | \mathbf{X}, \mathbf{W})_{\mathcal{P}_{\overline{\mathbf{X}}}}. \quad (19)$$

$$\text{Rule 2 } P(\mathbf{y}|\text{do}(\mathbf{x}), \text{do}(\mathbf{z}), \mathbf{w}) = P(\mathbf{y}|\text{do}(\mathbf{x}), \mathbf{z}, \mathbf{w}) \quad \text{if } (\mathbf{Y} \perp\!\!\!\perp \mathbf{Z} | \mathbf{X}, \mathbf{W})_{\mathcal{P}_{\overline{\mathbf{XZ}}}}. \quad (20)$$

$$\text{Rule 3 } P(\mathbf{y}|\text{do}(\mathbf{x}), \text{do}(\mathbf{z}), \mathbf{w}) = P(\mathbf{y}|\text{do}(\mathbf{x}), \mathbf{w}) \quad \text{if } (\mathbf{Y} \perp\!\!\!\perp \mathbf{Z} | \mathbf{X}, \mathbf{W})_{\mathcal{P}_{\overline{\mathbf{XZ}(\mathbf{W})}}}, \quad (21)$$

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

Definition 27 (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*

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

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 [26, 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 28 (Optimal Counterfactual Bound [29, Def. 2.1]). *For a causal diagram \mathcal{G} and distributions $\{P(\mathbf{V}_{\mathbf{z}}) \mid \mathbf{z} \in \mathbb{Z}\}$, the optimal bound $[l, r]$ over a counterfactual probability $P(\mathbf{y}_{\mathbf{x}}, \dots, \mathbf{z}_{\mathbf{w}})$ is defined as, respectively, the minimum and maximum of the following optimization problem:*

$$\begin{aligned} & \min / \max_{\mathcal{M} \in \Omega(\mathcal{G})} P^{\mathcal{M}}(\mathbf{y}_{\mathbf{x}}, \dots, \mathbf{z}_{\mathbf{w}}) \\ & \text{s.t. } P^{\mathcal{M}}(\mathbf{V}_{\mathbf{z}}) = P(\mathbf{V}_{\mathbf{z}}) \forall \mathbf{z} \in \mathbb{Z} \end{aligned}$$

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

Proposition 4 (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 [26, 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 $\mathcal{M} \in \Omega(\mathcal{G})$, there exists a \mathcal{G} -NCM \mathcal{N} such that $P^{\mathcal{M}}(\mathbf{y}_{\mathbf{x}}, \dots, \mathbf{z}_{\mathbf{w}}) = P^{\mathcal{N}}(\mathbf{y}_{\mathbf{x}}, \dots, \mathbf{z}_{\mathbf{w}})$.

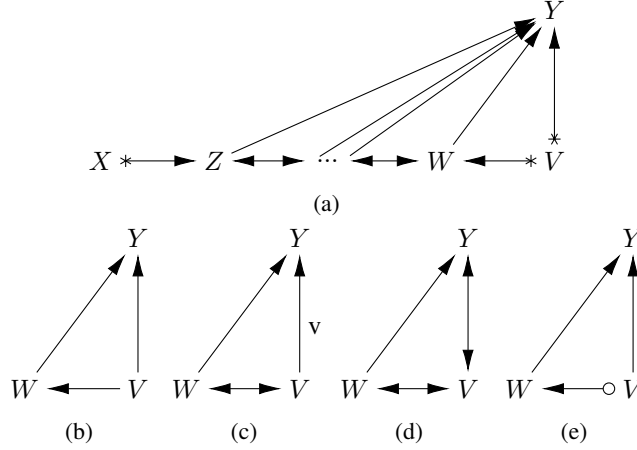


Figure 6: A discriminating path between X and Y for V (a), and possible configurations of edge orientations at V (b-d). $X \perp\!\!\!\perp Y|Z, W, V, \dots$ and $X \not\perp\!\!\!\perp Y|Z, W, \dots$ in (b) and (c), while $X \not\perp\!\!\!\perp Y|Z, W, V, \dots$ and $X \perp\!\!\!\perp Y|Z, W, \dots$ in (d). Or, V is a definite non-collider in (b) and (c), while it is a definite collider in (d). The edges out of V cannot both be confounded in (b). In a PAG, the possible configurations of edge orientations at V can also be (e) if $X \perp\!\!\!\perp Y|Z, W, V, \dots$ and $X \not\perp\!\!\!\perp Y|Z, W, \dots$.

By construction of RelationalNeuralID [26, Alg. 1], it must be that $P^{\mathcal{N}}(\mathbf{y}_x, \dots, \mathbf{z}_w) \in [\hat{l}, \hat{r}]$. Next, to see that $[\hat{l}, \hat{r}] \subseteq [l, r]$: by construction of [26, Alg. 1], there exist \mathcal{G} -NCMs $\mathcal{N}_{\hat{l}}$ and $\mathcal{N}_{\hat{r}}$ such that $P^{\mathcal{N}_{\hat{l}}}(\mathbf{y}_x, \dots, \mathbf{z}_w) = \hat{l}$ and $P^{\mathcal{N}_{\hat{r}}}(\mathbf{y}_x, \dots, \mathbf{z}_w) = \hat{r}$. By [26, 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. 28) and we are done. \square

B Useful Lemmas

B.1 Characterization of Mixed Graphs in PAG MEC

B.1.1 Lemmas on PAG to MAG

Lemma 2 (MEC of MAGs 1 [24]). *Two MAGs are Markov equivalent if and only if*

1. *They have the same skeleton;*
2. *They have the same unshielded colliders;*
3. *If a path p is a discriminating path for a vertex V in both MAGs, then V is a collider on p in one MAG iff it is a collider on p in the other MAG.*

However, a discriminating path for a given triple may not be present in all graphs within a Markov equivalence class [1]. There is a sub-class of discriminating paths and associated triples (those “with order”) that are always present, and sufficient for Markov equivalence.

Definition 29 (Ordered Colliders and Non-Colliders in MAGs [6, Def. 2]). *Let \mathcal{C}_i (resp. \mathcal{D}_i), with $i \geq 0$, be the set of set of collider (resp. non-collider) triples with order i in a MAG M , defined recursively as:*

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

Algorithm 3 PAG to MEC

```
1: Input: a PAG  $\mathcal{P}$  over  $\mathbf{V}$ 
2: Output: MEC  $\langle \mathfrak{S}, \mathfrak{C}, \mathfrak{D} \rangle$  for all MAGs in  $\mathcal{P}$ 
3: Phase 1: initialize, process unshielded triples
4:  $\mathfrak{S} \leftarrow \text{Skeleton}(\mathcal{P})$ 
5:  $\mathfrak{C}_0, \mathfrak{D}_0 \leftarrow$  unshielded colliders (non-colliders)  $\langle x, z, y \rangle \in \mathcal{P}$ 
6: for all  $\langle x, z, y \rangle \in \mathfrak{D}_0$  do
7:   if  $\exists q : \langle x, z, q \rangle \in \mathfrak{C}_0$  and  $q, y$  adjacent in  $\mathcal{P}$  then
8:      $\mathfrak{L} \leftarrow \langle z, q, y \rangle$ 
9:   end if
10: end for
11: Phase 2: process candidate triples until no more left
12: repeat
13:    $\langle x, z, y \rangle \leftarrow \mathfrak{L}$ 
14:   if  $x * \rightarrow z \leftarrow * y$  in  $\mathcal{P}$  then
15:     add  $\langle x, z, y \rangle$  to  $\mathfrak{C}$ 
16:      $\forall q : \langle x, z, q \rangle \in \mathfrak{D}$  and  $q, y$  adjacent in  $\mathcal{P} : \mathfrak{L} \leftarrow \langle z, y, q \rangle$ 
17:   else if  $x * - * z - * y$  or  $x * - z * - * y$  in  $\mathcal{P}$  then
18:     add  $\langle x, z, y \rangle$  to  $\mathfrak{D}$ 
19:      $\forall q : \langle x, z, q \rangle \in \mathfrak{C}$  and  $q, y$  adjacent in  $\mathcal{P} : \mathfrak{L} \leftarrow \langle z, q, y \rangle$ 
20:   end if
21: until  $\mathfrak{L}$  is empty
22: Return  $\langle \mathfrak{S}, \mathfrak{C}, \mathfrak{D} \rangle$ 
```

Definition 30 (MAG MEC Signature [6, Def. 3]). *The MEC signature of a MAG M , is defined as the triplet $\langle \mathfrak{S}, \mathfrak{C}, \mathfrak{D} \rangle$, with \mathfrak{S} the (undirected) skeleton of M , and \mathfrak{C} and \mathfrak{D} the corresponding lists of ordered collider and non-collider triples from Def. 29.*

Lemma 3 (Markov Equivalence of MAGs using Ordered Triples [6, Cor. 3]). *Two MAGs are Markov equivalent if and only if they have identical MEC signatures.*

In fact, these MEC signatures are also readable from the PAG representing the MAGs in the equivalence class, by adapting the algorithm of extracting MEC from MAGs [6, Alg. 1] to taking a PAG as the input (Alg. 3). It keeps the same polynomial efficiency: Consider graphs over n nodes with e edges and max. node degree d . For sparse graphs with $d \leq k$ we have $e = O(n)$, whereas in general we can have $e = O(n^2)$.

Lemma 4 (Soundness of Algo. 3). *Given a PAG \mathcal{P} , Algorithm 3 outputs the correct MEC representation for all MAGs in the equivalence class represented by \mathcal{P} .*

Proof. The proof proceeds by showing that running the algorithm on any MAG $M \in [\mathcal{P}]$ yields the same output as running it directly on \mathcal{P} .

Consider two parallel executions of the algorithm: one with input M and the other with input \mathcal{P} .

Phase 1. By definition, \mathcal{P} and M share the same skeleton and the same set of unshielded collider and non-collider triples. Hence, the initial sets \mathfrak{C}_0 and \mathfrak{D}_0 coincide in both executions. Since the list \mathfrak{L} is initialized from \mathfrak{C}_0 and the skeleton, it is also identical.

Phase 2. We prove by induction on the iterations that both executions remain identical.

Assume that after k iterations, the sets \mathfrak{C} , \mathfrak{D} , \mathfrak{S} , and the list \mathfrak{L} coincide. Consider the $(k + 1)$ -th iteration. The same triple is removed from \mathfrak{L} in both runs.

We distinguish two cases:

- **Collider case.** If the triple is a collider in M , then it is invariant across all MAGs in $[\mathcal{P}]$. Therefore, both incident edges have arrowheads into the middle node in every MAG, and these arrowheads are invariant edge marks. By completeness of PAG edge marks, the same arrowheads appear in \mathcal{P} , so the triple is also identified as a collider in the PAG-based execution.

Algorithm 4 ADMG to MAG

```
1: Input: an ADMG  $\mathcal{G}$  over  $\mathbf{V}$ 
2: Output: a MAG  $M$  over  $\mathbf{V}$ , representing the same CIs over  $\mathbf{V}$ 
3: Initialize  $M$  with nodes in  $\mathbf{V}$ 
4: for each pair of variables  $A, B \in \mathbf{V}$  do
5:   if there is an inducing path between  $A$  and  $B$  in  $\mathcal{G}$  then
6:     if  $A \in \text{An}(B)$  in  $\mathcal{G}$  then
7:       Add edge  $A \rightarrow B$  to  $M$ 
8:     else if  $B \in \text{An}(A)$  in  $\mathcal{G}$  then
9:       Add edge  $B \rightarrow A$  to  $M$ 
10:    else
11:      Add edge  $A \leftrightarrow B$  to  $M$ 
12:    end if
13:  end if
14: end for
15: Return  $M$ 
```

- **Non-collider case.** If the triple is a non-collider in M , then by the ordered-triple characterization of the MEC, this status is also invariant. In the PAG, this invariance is reflected by a tail mark at the middle node (from the definition of core PAG and Algorithm 2 of [6]). As a result, the conditions of the algorithm identify the triple as a non-collider in the PAG-based execution.

Since the sets \mathcal{C} , \mathcal{D} , and \mathcal{S} are identical up to this point, the newly added triples to \mathcal{L} are also identical in both executions.

By induction, the two executions produce identical outputs. Therefore, the algorithm applied to \mathcal{P} yields the same MEC representation as when applied to any MAG in $[\mathcal{P}]$, proving soundness. \square

With the results above, we formulate a characterization of MAGs that fall within the equivalence class represented by a PAG.

Lemma 5 (A Characterization of MAGs Represented by a PAG). *Let \mathcal{P} be a PAG with MEC signature $(\mathcal{S}, \mathcal{C}, \mathcal{D})$. A MAG M lies in the MEC represented by \mathcal{P} if and only if the following conditions hold:*

1. **Skeleton / Maximality:** For any pair of vertices X, Y , X and Y are adjacent in \mathcal{P} if and only if they are adjacent or connected by an inducing path in M ;
2. **Ancestrality:** M contains no directed cycles or almost directed cycles;
3. **Collider constraints:** Every ordered collider triple in \mathcal{C} is a collider in M ;
4. **Non-collider constraints:** Every ordered non-collider triple in \mathcal{D} is a non-collider in M .

Proof. Conditions (1) and (2) ensure that M is a maximal ancestral graph (MAG). Condition (1) guarantees maximality via the absence of inducing paths between non-adjacent vertices, while (2) enforces ancestrality.

Conditions (3) and (4) ensure that M shares the same ordered collider and non-collider triples as the MEC represented by \mathcal{P} . By the ordered-triple characterization of MAG equivalence classes (Lem. 3), they imply that M is Markov equivalent to every MAG in the class together with Condition (1). \square

Moreover, invariant edge marks in \mathcal{P} are automatically respected by M whenever Conditions (1) to (4) of Lem. 5 are satisfied, since such marks correspond to edge features shared by all MAGs in the MEC. Although not required, these invariant marks may be enforced explicitly to simplify the construction or enumeration of MAGs represented by \mathcal{P} .

B.1.2 Lemmas on MAG to ADMG

The following facts follow from the construction of MAG from ADMG using Algo. 4.

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

Fact 2. If $A \rightarrow B$ 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 $A \in \text{An}(B)$.

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 $A \notin \text{An}(B)$ and $B \notin \text{An}(A)$.

Lemma 6 (Contrapositive of Lemma 9 [27]). If $A \rightarrow B$ 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 .

In other words, the visible edge must be induced by an inducing path that is out of A and into B .

Lemma 7 (Characterizing ADMGs Represented by a MAG). Let M be a MAG and \mathcal{G} an ADMG. Then \mathcal{G} belongs to the MEC represented by M if and only if:

1. For all distinct X, Y , X and Y are adjacent in M if and only if they are adjacent or connected by an inducing path in \mathcal{G} ;
2. For all X, Y , X is an ancestor of Y in M if and only if X is an ancestor of Y in \mathcal{G} .

Proof. This lemma follows directly from the construction of the MAG from an ADMG in Algo. 4. \square

In addition, the collider and non-collider properties of ordered triples in MAGs must be preserved in ADMGs in the MEC.

Lemma 8. Let M be a MAG with MEC signature $\langle \mathfrak{S}, \mathfrak{C}, \mathfrak{D} \rangle$, and let \mathcal{G} be an ADMG that induces M . Then for any triple $\langle a, b, c \rangle$ that is present in \mathcal{G} :

- If $\langle a, b, c \rangle \in \mathfrak{C}$, then it is a collider in \mathcal{G} ;
- If $\langle a, b, c \rangle \in \mathfrak{D}$, then it is a non-collider in \mathcal{G} .

Proof. We prove each statement separately.

Collider case. Let $\langle a, b, c \rangle \in \mathfrak{C}$. Then in M , both edges incident to b have arrowheads at b , i.e., $a^* \rightarrow b \leftarrow^* c$. Since \mathcal{G} induces M , any edge between a and b (and between b and c) in \mathcal{G} cannot introduce an arrowhead away from b without creating an inducing path inconsistent with M (cf. Fact 1). Thus, all such edges in \mathcal{G} must also have arrowheads into b , and $\langle a, b, c \rangle$ is a collider in \mathcal{G} .

Non-collider case. Let $\langle a, b, c \rangle \in \mathfrak{D}$. Then there exists a discriminating path

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

for b in M , certifying that $\langle a, b, c \rangle$ is a non-collider.

Since \mathcal{G} induces M , the path p is also induced in \mathcal{G} . Suppose for contradiction that $\langle a, b, c \rangle$ is a collider in \mathcal{G} , i.e., $a^* \rightarrow b \leftarrow^* c$. We consider two cases:

- **No inducing path out of b to a or c .** Then the induced MAG M' from \mathcal{G} must contain arrowheads into b on both sides, making $\langle a, b, c \rangle$ a collider in M' , contradicting $M' = M$.
- **An inducing path exists out of b to a or c .** Then this path, together with p and the paths formed the collider at b , creates an inducing path between x and c , contradicting the fact that p is a discriminating path for b .

Both cases lead to contradictions. Therefore, $\langle a, b, c \rangle$ must be a non-collider in \mathcal{G} . \square

Lemma 9. Let \mathcal{G} be an ADMG and let M be its induced MAG. Let \mathcal{G}' be the ADMG obtained by adding any subset of edges in M that are not already present in \mathcal{G} . Then the induced MAG of \mathcal{G}' is also M .

Algorithm 5 \mathcal{G} to $\mathcal{G}[M]$

```
1: Input: an ADMG  $\mathcal{G}$  over  $\mathbf{V}$ 
2: Output: an ADMG  $\mathcal{G}[M]$  with edges in the MAG  $M$  of  $\mathcal{G}$  added to  $\mathcal{G}$ .
3:  $M \leftarrow \text{ADMG to MAG}(\mathcal{G})$ 
4: Initialize  $\mathcal{G}[M]$  to be  $\mathcal{G}$ 
5: for each edge  $e \in M$  do
6:   if  $e \notin \mathcal{G}[M]$  then
7:     Add edge  $e$  to  $\mathcal{G}[M]$ 
8:   end if
9: end for
10: Return  $\mathcal{G}[M]$ 
```

Proof. Since M and \mathcal{G} share the same ancestral relations, adding edges from M to \mathcal{G} cannot introduce or remove any ancestral relations. Thus, it suffices to show that adding any edge $e \in M \setminus \mathcal{G}$ to \mathcal{G} does not create any new inducing paths between pairs of vertices that were not already connected by an inducing path in \mathcal{G} . Since \mathcal{G}' is obtained by adding a subset of such edges, the result follows by iterating this argument.

Let $e \in M \setminus \mathcal{G}$. We consider two cases.

Case 1: $e = A \rightarrow B$. By Fact 2, there exists an inducing path p from A to B that is into B , and $A \in \text{An}(B)$ in \mathcal{G} .

Suppose adding $A \rightarrow B$ creates a new inducing path between A and some vertex C that was not previously connected to A by an inducing path. Such a path must begin with the edge $A \rightarrow B$, followed by a subpath p' from B to C that is into B .

By the definition of inducing paths, every node on p' (except C) is a collider and lies in $\text{An}(A)$ or $\text{An}(C)$. Since $A \in \text{An}(B)$, to avoid a directed cycle we must have $B \in \text{An}(C)$.

Now consider the original inducing path p from A to B . All colliders on p lie in $\text{An}(A)$ or $\text{An}(B)$. By transitivity of ancestry, any node in $\text{An}(B)$ also lies in $\text{An}(C)$. Thus, concatenating p with p' yields an inducing path from A to C already present in \mathcal{G} , a contradiction.

Case 2: $e = A \leftrightarrow B$. By Fact 3, there exists an inducing path p between A and B that is into both endpoints, with $A \notin \text{An}(B)$ and $B \notin \text{An}(A)$.

Suppose adding $A \leftrightarrow B$ creates a new inducing path between A and some vertex C . Without loss of generality, assume the path begins with $A \leftrightarrow B$, followed by a subpath p' from B to C that is into B .

Again, all intermediate nodes on p' are colliders and lie in $\text{An}(A)$ or $\text{An}(C)$. Since $B \notin \text{An}(A)$, we must have $B \in \text{An}(C)$.

On the original inducing path p , all colliders lie in $\text{An}(A)$ or $\text{An}(B)$. By transitivity, any node in $\text{An}(B)$ also lies in $\text{An}(C)$. Thus, concatenating p with p' yields an inducing path from A to C already present in \mathcal{G} , a contradiction.

In both cases, adding edges from M to \mathcal{G} does not introduce new inducing paths. Hence the set of inducing paths remains unchanged, and the induced MAG of \mathcal{G}' coincides with M . \square

This lemma implies that for any ADMG \mathcal{G} represented by M that does not share the same skeleton as M , there exists a supergraph $\mathcal{G}' \supset \mathcal{G}$ that is also represented by M and has the same skeleton as M . Such a \mathcal{G}' can be obtained by running Alg. 5. By Prop. 4, it therefore suffices to bound causal queries over the subclass of ADMGs whose skeleton matches that of M .

Moreover, the lemma implies that if an ADMG contains an almost directed cycle, one can add a directed edge between the corresponding pair of vertices without leaving the equivalence class. In the induced MAG, such an edge is necessarily invisible.

B.1.3 Lemmas on PAG to ADMG

As compared to MAGs, the space of ADMGs are further relaxed to allow some additional features:

| PAG | MAG | ADMG (with bows) |
|-------------------------|---|--|
| $a \circ - \circ b$ | $a \rightarrow b, b \rightarrow a, a \leftrightarrow b$ | Forbidden: None Required: Inducing path between a and b |
| $a \circ \rightarrow b$ | $a \rightarrow b, a \leftrightarrow b$ | Forbidden: Inducing path out of b Required: Inducing path into b |
| $a \leftarrow \circ b$ | $b \rightarrow a, a \leftrightarrow b$ | Forbidden: Inducing path out of a Required: Inducing path into a |
| $a \leftrightarrow b$ | $a \leftrightarrow b$ | Forbidden: Directed path between a and b Required: Inducing path into a and b |
| $a \xrightarrow{v} b$ | $a \xrightarrow{v} b$ | Forbidden: Inducing path into a Required: Inducing path into b |

Table 1: Mapping of edge marks from PAGs to MAGs and ADMGs (with bows).

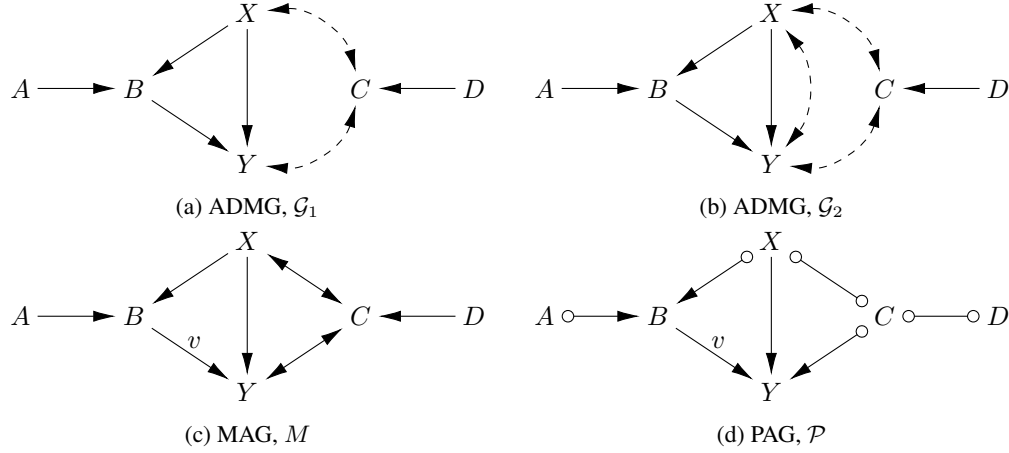


Figure 7: Two ADMGs (a,b) inducing the same MAG (c) and (d), but have different collider properties for definite non-collider triple (C, X, Y) : non-collider in (a) and collider in (b).

1. bows
2. almost directed cycles

With the edge orientations in the ADMC following Table 1, the following facts help ensure that the oriented ADMGs stay within the MEC of the PAG.

Fact 4. Let \mathcal{P} be a PAG and \mathcal{G} an ADMG consistent with \mathcal{P} . Then invariant edge marks in \mathcal{P} restrict the edges in \mathcal{G} as follows:

1. If the edge between a and b has an arrowhead at b in \mathcal{P} , then $b \rightarrow a \notin \mathcal{G}$.
2. If the edge between a and b is visible in \mathcal{P} , then $a \leftrightarrow b \notin \mathcal{G}$.

Fact 5. Let (X, Z, Y) be a definite collider triple in a PAG \mathcal{P} . For any ADMG \mathcal{G} consistent with \mathcal{P} in which X and Y are both adjacent to Z , the triple (X, Z, Y) is a collider in \mathcal{G} .

Fact 6. Let \mathcal{P} be a PAG and (X, Z, Y) an ordered non-collider triple in \mathcal{P} . Then, for any ADMG \mathcal{G} consistent with \mathcal{P} in which X and Y are both adjacent to Z , the triple (X, Z, Y) is a non-collider in \mathcal{G} .

Note that we do not enforce all definite non-collider triples in the PAG to remain non-colliders in the ADMG. In particular, when the tail mark in such a triple arises from an invisible edge, allowing bows may introduce an arrowhead at that endpoint, and turning the triple into a collider, with an example shown in Fig. 7. Therefore, permitting bows in the ADMG requires relaxing some of the definite non-collider constraints that are invariant across MAGs in the MEC.

B.2 Details on PAG-NCM Masking Parameters

Soundness of Enforcing PAG Adjacency

Lemma 10 ($\circ - \circ$). *Let \mathcal{P} be a PAG and let $A \circ - \circ B$ be an edge in \mathcal{P} . Let \mathcal{G} be an ADMG inducing a MAG $M \in \mathcal{P}$ with $A * - * B \in \mathcal{G}$. Then there exists a direction $\sigma \in \{A \rightarrow B, B \rightarrow A\}$ such that the graph $\mathcal{G}' = \mathcal{G} \cup \{\sigma\}$ induces a MAG in \mathcal{P} .*

The set of admissible directions is determined by the ancestral relations in \mathcal{G} : it may be $\{A \rightarrow B\}$, $\{B \rightarrow A\}$, or both.

Proof. Since \mathcal{G} induces a MAG in \mathcal{P} , it satisfies all conditions in Thm. 1. We show that after adding the directed edge, the resulting graph \mathcal{G}' continues to satisfy these conditions. The proof proceeds by verifying that each requirement is preserved; the arguments are identical across all cases, differing only in the direction of the added edge.

(i) Acyclicity. Adding $A \rightarrow B$ (or $B \rightarrow A$) does not introduce directed cycles. Depending on the edge between A and B in the induced MAG M , we have:

- If $A \rightarrow B$ in M , then $B \notin An(A)$ in \mathcal{G} , so adding $A \rightarrow B$ preserves acyclicity.
- If $A \leftarrow B$ in M , the argument is symmetric.
- If $A \leftrightarrow B$ in M , then neither $A \in An(B)$ nor $B \in An(A)$ in \mathcal{G} , so adding either direction preserves acyclicity.

(ii) Skeleton. We show that no new inducing path is created between any pair of non-adjacent vertices. Without loss of generality, we assume the edge added has direction $A \rightarrow B$. Suppose, for contradiction, that a new inducing path is created in \mathcal{G}' between A and some node C not adjacent to A . Such a path must use the newly added edge $A \rightarrow B$, so it has the form $A \rightarrow B \rightsquigarrow C$. Let p denote the subpath from B to C . By definition of inducing paths, all internal nodes on p , including B and excluding C , are colliders and lie in $An(A) \cup An(C)$. But then replacing $A \rightarrow B$ with $A \leftrightarrow B$ yields an inducing path between A and C in \mathcal{G} , contradicting non-adjacency of A and C . Hence no new inducing path is created.

(iii) Collider and non-collider constraints. Adding the directed edge does not change the collider or non-collider status of any triple involving A and B . Since $A \leftrightarrow B \in \mathcal{G}$, both endpoints already have arrowheads, and the additional directed edge does not introduce new arrowhead configurations. Thus all ordered collider and non-collider triples required by Thm. 1 remain satisfied.

All conditions in Thm. 1 remain satisfied after adding the directed edge. Therefore \mathcal{G}' induces a MAG in the MEC represented by \mathcal{P} . \square

Lemma 11 ($\circ \rightarrow$). *Let \mathcal{P} be a PAG and let $A \circ \rightarrow B$ be an edge in \mathcal{P} . Let \mathcal{G} be an ADMG inducing a MAG $M \in \mathcal{P}$ with $A \leftrightarrow B \in \mathcal{G}$. Then the graph $\mathcal{G}' = \mathcal{G} \cup \{A \rightarrow B\}$ induces a MAG in \mathcal{P} .*

Proof. Since \mathcal{G} induces a MAG in \mathcal{P} , it satisfies all conditions in Thm. 1. We verify that these conditions are preserved after adding the edge $A \rightarrow B$.

(i) Acyclicity. Since $A \circ \rightarrow B$ in \mathcal{P} , the arrowhead at B is invariant, implying that $B \notin An(A)$ in every MAG in \mathcal{P} , and hence in \mathcal{G} . Therefore, adding $A \rightarrow B$ does not create a directed cycle.

(ii) Skeleton / maximality. The skeleton is unchanged. As in Lem. 10, adding $A \rightarrow B$ does not create any new inducing paths between non-adjacent vertices, since any such path would already exist via the edge $A \leftrightarrow B$ in \mathcal{G} .

(iii) Collider and non-collider constraints. Adding $A \rightarrow B$ does not alter the ordered collider or non-collider status of any triple. In particular, since $A \leftrightarrow B \in \mathcal{G}$, both endpoints already carry arrowheads, and the additional directed edge does not introduce any new configurations violating the constraints in Thm. 1.

All conditions of Thm. 1 remain satisfied. Hence \mathcal{G}' induces a MAG in the MEC represented by \mathcal{P} . \square

Algorithm 6 \mathcal{G} to $\mathcal{G}[\mathcal{P}]$

```

1: Input: an ADMG  $\mathcal{G}$  over  $\mathbf{V}$ 
2: Output: an ADMG  $\mathcal{G}'$ 
3:  $\mathcal{G}' \leftarrow \mathcal{G}[M]$  (Alg. 5)
4: for each bidirected edge  $V_i \leftrightarrow V_j \in \mathcal{G}'$  such that  $V_i \rightarrow V_j \notin \mathcal{G}'$  and  $V_j \rightarrow V_i \notin \mathcal{G}'$  do
5:   if  $V_i \circ \rightarrow V_j \in \mathcal{P}$  then
6:     Add edge  $V_i \rightarrow V_j$  to  $\mathcal{G}'$ 
7:   else if  $V_j \circ \rightarrow V_i \in \mathcal{P}$  then
8:     Add edge  $V_j \rightarrow V_i$  to  $\mathcal{G}'$ 
9:   else if  $V_i \circ - \circ V_j \in \mathcal{P}$  then
10:    Add either edge  $V_i \rightarrow V_j$  or  $V_j \rightarrow V_i$  to  $\mathcal{G}'$  such that no directed cycles are introduced
11:   end if
12: end for
13: Return  $\mathcal{G}'$ 

```

The two lemmas above implies that adding directed edges for edges with circles marks in the PAG does not bring the causal diagram out of the MEC.

Corollary 2. *Given a causal diagram $\mathcal{G} \in \mathcal{P}$, the causal diagram \mathcal{G}' output from Alg. 6 is in the MEC of \mathcal{P} .*

Proof. This follows immediately from Lem. 10 and Lem. 11. □

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 31 (\mathcal{P} -NCM to Standard SCM). *Let $\mathcal{N}(\theta, \mathbf{m}^D, \mathbf{m}^B, \mathbf{m}^C) = \langle \mathbf{V}, \mathbf{U}, P(\mathbf{U}), \mathcal{F} \rangle$ 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

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

- Construct $f'_i \in \mathcal{F}'$ by fixing the masks to their assigned values and restricting inputs to the active parents:

$$f'_i(\mathbf{pa}_i, \mathbf{u}_i) := f_i(\mathbf{m}_{\mathbf{Pa}_i, i}^D \odot \mathbf{pa}_i, \mathbf{m}_{\mathbf{U}_i, i}^C \odot \mathbf{u}_i).$$

Proposition 5 (\mathcal{P} -NCM–SCM \mathcal{L}_3 Equivalence). *The SCM \mathcal{M} constructed from a \mathcal{P} -NCM via Def. 31 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 $\mathbf{m}_{ji}^D > 0$. By Def. 7, these are precisely the variables that induce directed edges $V_j \rightarrow V_i$ in \mathcal{G} . Since directed edges in \mathcal{G}' are also determined by parent relationships (Def. 16), it follows that \mathcal{G} and \mathcal{G}' share the same directed edges.

For bidirected edges, \mathcal{M} inherits exogenous parents U_S with $\mathbf{m}_{U_S, i}^C > 0$ for each V_i . Each such U_S is a shared latent parent of all variables in S , which induces bidirected edges between every pair of variables in S in \mathcal{G}' (Def. 16). By definition,

$$\mathbf{m}_{U_S, i}^C = \prod_{\substack{V_j, V_k \in S \\ j \neq k}} \mathbf{m}_{jk}^B,$$

which implies $\mathbf{m}_{jk}^B > 0$ for all distinct $V_j, V_k \in S$. By Def. 7, this induces bidirected edges between all such pairs in \mathcal{G} as well.

Therefore, $\mathcal{G} = \mathcal{G}'$.

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}}, \dots, \mathbf{z}_{\mathbf{W}}) \\
&= \sum_{\mathbf{u}} \mathbf{1}[\mathbf{Y}_{\mathbf{X}}(\mathbf{m}_{\cdot, \mathbf{Y}}^C \odot \mathbf{u}) = \mathbf{y}, \dots, \mathbf{Z}_{\mathbf{W}}(\mathbf{m}_{\cdot, \mathbf{Y}}^C \odot \mathbf{u}) = \mathbf{z}] P(\mathbf{u}) \\
&= \sum_{\mathbf{u} \in \text{dom}(\mathbf{U})} \left(\prod_{V_i \in \mathbf{Y} \setminus \mathbf{X}} \mathbf{1}[f_i(\mathbf{m}_{\prec, i}^D \odot \mathbf{v}_{\prec, i}, \mathbf{m}_{\cdot, i}^C \odot \mathbf{u}) = v_i] \right) \left(\prod_{V_i \in \mathbf{X}} \mathbf{1}[V_i = x_i] \right) \\
&\quad \left(\prod_{V_j \in \mathbf{Z} \setminus \mathbf{W}} \mathbf{1}[f_j(\mathbf{m}_{\prec, j}^D \odot \mathbf{v}_{\prec, j}, \mathbf{m}_{\cdot, j}^C \odot \mathbf{u}) = v_j] \right) \left(\prod_{V_j \in \mathbf{W}} \mathbf{1}[V_j = w_j] \right) P(\mathbf{u}) \quad (\text{By Def. 8}) \\
&= \sum_{\mathbf{u} \in \text{dom}(\mathbf{U})} \left(\prod_{V_i \in \mathbf{Y} \setminus \mathbf{X}} \mathbf{1}[f_i(\mathbf{m}_{\mathbf{pa}_i}^D \odot \mathbf{pa}_i, \mathbf{m}_{\mathbf{U}_i}^C \odot \mathbf{u}_i) = v_i] \right) \left(\prod_{V_i \in \mathbf{X}} \mathbf{1}[V_i = x_i] \right) \\
&\quad \left(\prod_{V_j \in \mathbf{Z} \setminus \mathbf{W}} \mathbf{1}[f_j(\mathbf{m}_{\mathbf{pa}_j}^D \odot \mathbf{pa}_j, \mathbf{m}_{\mathbf{U}_j}^C \odot \mathbf{u}_j) = v_j] \right) \left(\prod_{V_j \in \mathbf{W}} \mathbf{1}[V_j = w_j] \right) P(\mathbf{u}) \\
&\quad \quad \quad (0 \text{ entries in } \mathbf{m}^D, \mathbf{m}^C \text{ do not affect } \odot) \\
&= \sum_{\mathbf{u} \in \text{dom}(\mathbf{U})} \left(\prod_{V_i \in \mathbf{Y} \setminus \mathbf{X}} \mathbf{1}[f'_i(\mathbf{pa}_i, \mathbf{u}_i) = v_i] \right) \left(\prod_{V_i \in \mathbf{X}} \mathbf{1}[V_i = x_i] \right) \\
&\quad \left(\prod_{V_j \in \mathbf{Z} \setminus \mathbf{W}} \mathbf{1}[f'_j(\mathbf{pa}_j, \mathbf{u}_j) = v_j] \right) \left(\prod_{V_j \in \mathbf{W}} \mathbf{1}[V_j = w_j] \right) P(\mathbf{u}) \quad (\text{By Def. 31}) \\
&= \mathbf{P}^{\mathcal{M}}(\mathbf{y}_{\mathbf{X}}, \dots, \mathbf{z}_{\mathbf{W}}) \quad (\text{By Def. 19})
\end{aligned}$$

□

For every SCM \mathcal{M} inducing a causal diagram $\mathcal{G} \in \mathcal{P}$, there exists a \mathcal{P} -NCM \mathcal{N} whose induced causal diagram \mathcal{G}' satisfies $\mathcal{G}' \supseteq \mathcal{G}$ and $\mathcal{G}' \in \mathcal{P}$, and shares the same counterfactual distributions $\mathbf{P}^{\mathcal{L}^3}$ as \mathcal{M} .

First, we show that for every SCM \mathcal{M} inducing a causal diagram $\mathcal{G} \in \mathcal{P}$, there is an SCM \mathcal{M}' inducing \mathcal{G}' , the causal diagram obtained from Alg. 6 with \mathcal{G} as input, and shares the same counterfactual distributions $\mathbf{P}^{\mathcal{L}^3}$ as \mathcal{M} .

Lemma 12. *Given an SCM $\mathcal{M} = \langle \mathbf{U}, \mathbf{V}, \mathcal{F}, P(\mathbf{U}) \rangle$, with causal diagram \mathcal{G} , there exists an SCM $\mathcal{M}' = \langle \mathbf{U}', \mathbf{V}', \mathcal{F}', P(\mathbf{U}') \rangle$, with causal diagram \mathcal{G}' , the causal diagram obtained from Alg. 6 with \mathcal{G} as input, and shares the same counterfactual distributions $\mathbf{P}^{\mathcal{L}^3}$ as \mathcal{M} .*

Proof. \mathcal{M}' can be constructed as follows:

- Set $\mathbf{V}' = \mathbf{V}$.
- Set $\mathbf{U}' = \mathbf{U} \cup \mathbf{U}''$, where $U'_{ij} \in \mathbf{U}''$ if $V_i \leftrightarrow V_j \in \mathcal{G}' \setminus \mathcal{G}$.
- Set $P(\mathbf{U}') = P(\mathbf{U})P(\mathbf{U}'')$ with each $U' \in \mathbf{U}'' \sim \text{Unif}(0, 1)$.
- For each $V_i \in \mathbf{V}$, set

$$f'_i(\mathbf{pa}'_i, \mathbf{u}'_i) = f_i(\mathbf{pa}_i, \mathbf{u}_i).$$

First, given that $\mathcal{G} \subseteq \mathcal{G}'$, for each $V_i \in \mathbf{V}$, $\mathbf{Pa}_i \subseteq \mathbf{Pa}'_i$ and $\mathbf{U}_i \subseteq \mathbf{U}'_i$.

Therefore, 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 $\mathbf{v} \in \text{dom}(\mathbf{V})$,

$$\begin{aligned}
& \mathbf{P}^{\mathcal{M}'}(\mathbf{y}_{\mathbf{x}}, \dots, \mathbf{z}_{\mathbf{w}}) \\
&= \sum_{\mathbf{u}'} \mathbf{1}[\mathbf{Y}_{\mathbf{x}}(\mathbf{u}') = \mathbf{y}, \dots, \mathbf{Z}_{\mathbf{w}}(\mathbf{u}') = \mathbf{z}] P(\mathbf{u}') \\
&= \sum_{\mathbf{u}'} \left(\prod_{V_i \in \mathbf{Y} \setminus \mathbf{X}} \mathbf{1}[f'_i(\mathbf{pa}'_i, \mathbf{u}'_i) = v_i] \right) \left(\prod_{V_i \in \mathbf{X}} \mathbf{1}[V_i = x_i] \right) \\
&\quad \left(\prod_{V_j \in \mathbf{Z} \setminus \mathbf{W}} \mathbf{1}[f'_j(\mathbf{pa}'_j, \mathbf{u}'_j) = v_j] \right) \left(\prod_{V_j \in \mathbf{W}} \mathbf{1}[V_j = w_j] \right) P(\mathbf{u}') \quad (\text{By Def. 19}) \\
&= \sum_{\mathbf{u}} \left(\prod_{V_i \in \mathbf{Y} \setminus \mathbf{X}} \mathbf{1}[f_i(\mathbf{pa}_i, \mathbf{u}_i) = v_i] \right) \left(\prod_{V_i \in \mathbf{X}} \mathbf{1}[V_i = x_i] \right) \\
&\quad \left(\prod_{V_j \in \mathbf{Z} \setminus \mathbf{W}} \mathbf{1}[f_j(\mathbf{pa}_j, \mathbf{u}_j) = v_j] \right) \left(\prod_{V_j \in \mathbf{W}} \mathbf{1}[V_j = w_j] \right) P(\mathbf{u}) \sum_{\mathbf{u}''} P(\mathbf{u}'') \\
&\hspace{15em} (\text{By definition of } \mathcal{F}' \text{ and } P(\mathbf{U}')) \\
&= \mathbf{P}^{\mathcal{M}}(\mathbf{y}_{\mathbf{x}}, \dots, \mathbf{z}_{\mathbf{w}}) \quad (\text{By Def. 19})
\end{aligned}$$

□

Definition 32 (\mathcal{G} -Constrained Neural Causal Model (\mathcal{G} -NCM) [26, Def. 2]). *Given a causal diagram \mathcal{G} , a \mathcal{G} -constrained Neural Causal Model (for short, \mathcal{G} -NCM) $\widehat{M}(\boldsymbol{\theta})$ over variables \mathbf{V} with parameters $\boldsymbol{\theta} = \{\theta_{V_i} : V_i \in \mathbf{V}\}$ is an SCM $\langle \widehat{\mathbf{U}}, \mathbf{V}, \widehat{\mathcal{F}}, \widehat{P}(\widehat{\mathbf{U}}) \rangle$ such that $\widehat{\mathbf{U}} = \{\widehat{U}_{\mathbf{C}} : \mathbf{C} \in \mathbb{C}(\mathcal{G})\}$, where $\mathbb{C}(\mathcal{G})$ is the set of all maximal cliques over bidirected edges of \mathcal{G} , and $D_{\widehat{U}_{\mathbf{C}}} = [0, 1]$ for all $\widehat{U}_{\mathbf{C}} \in \widehat{\mathbf{U}}$; $\widehat{\mathcal{F}} = \{\widehat{f}_{V_i} : V_i \in \mathbf{V}\}$, where each \widehat{f}_{V_i} is a feedforward neural network parameterized by $\theta_{V_i} \in \boldsymbol{\theta}$ mapping values of $\mathbf{U}_i \cup \mathbf{Pa}_i$ to values of V_i for $\mathbf{U}_i = \{\widehat{U}_{\mathbf{C}} : \widehat{U}_{\mathbf{C}} \in \widehat{\mathbf{U}} \text{ s.t. } V_i \in \mathbf{C}\}$ and $\mathbf{Pa}_i V_i = \mathbf{Pa}_{\mathcal{G}}(V_i)$; $\widehat{P}(\widehat{\mathbf{U}})$ is defined s.t. $\widehat{U}_{\mathbf{C}} \sim \text{Unif}(0, 1)$ for each $\widehat{U}_{\mathbf{C}} \in \widehat{\mathbf{U}}$.*

Lemma 13 (\mathcal{L}_3 - \mathcal{G} Expressiveness [26, Thm. 3]). *For any SCM \mathcal{M}^* that induces causal diagram \mathcal{G} , there exists a \mathcal{G} -NCM $\widehat{M}(\boldsymbol{\theta}) = \langle \widehat{\mathbf{U}}, \mathbf{V}, \widehat{\mathcal{F}}, \widehat{P}(\widehat{\mathbf{U}}) \rangle$ s.t. \widehat{M} is \mathcal{L}_3 -consistent w.r.t. \mathcal{M}^* .*

Lemma 14. *Given a causal diagram $\mathcal{G} \in \mathcal{P}$, and \mathcal{G}' the output from Alg. 6 with \mathcal{G} as input, let \mathcal{G}' -NCM be the \mathcal{G}' -constrained NCM. there exists a \mathcal{P} -NCM equivalent to the \mathcal{G}' -NCM.*

Proof. First, we show that there exists a fixed mask parameter assignment $\mathbf{m}^D, \mathbf{m}^B$ that induces \mathcal{G}' .

By construction of \mathcal{P} -NCM (Def. 6, the mask values set to 0 match exactly the ones missing from \mathcal{G}' , because if this not the case, \mathcal{G}' will not induce the same skeleton as \mathcal{P} and fall outside the MEC of \mathcal{P} by Thm. 1.

For directed edge masks in \mathbf{m}^D fixed to 1, they correspond to edges in the following two cases:

- $V_i \rightarrow V_j \in \mathcal{P}$: This implies that $V_i \rightarrow V_j$ in all MAGs in the MEC, and they must also appear in \mathcal{G}' by Alg. 6.
- $V_i \circ \rightarrow V_j \in \mathcal{P}$: This implies that either $V_i \rightarrow V_j$ or $V_i \leftrightarrow V_j$ must be in the MAG. In either case, $V_i \rightarrow V_j$ will be added to \mathcal{G}' by Alg. 6.

For circle edges like $V_i \circ - \circ V_j \in \mathcal{P}$, it implies that V_i and V_j are adjacent in the MAG, and they would also induce a directed edge in \mathcal{G}' . With acyclicity constraint, $m_{ij}^D + m_{ji}^D = 1$ as only one of the two masks are set to 1. They satisfy the mask constraint $M_{ij}^D = 1 - m_{ji}^D$.

For bidirected edges masks fixed to 1, it only happens when $V_i \leftrightarrow V_j \in \mathcal{P}$. This implies that $V_i \leftrightarrow V_j$ in all MAGs in the MEC, and they must also appear in \mathcal{G}' by Alg. 6.

The rest of the mask values are free variables which can be assigned to match edges in \mathcal{G} , with $m_{i,j}^D = 1 \iff V_i \rightarrow V_j \in \mathcal{G}'$ and $m_{i,j}^B = 1 \iff V_i \leftrightarrow V_j \in \mathcal{G}'$.

Therefore, the masks fixed using the above procedure will induce \mathcal{G}' as the causal diagram. More importantly, the mask values are binary where $m_{i,j}^D = 1 \iff V_i \rightarrow V_j \in \mathcal{G}'$ and $m_{i,j}^B = 1 \iff V_i \leftrightarrow V_j \in \mathcal{G}'$.

Next, we show that for each \mathcal{G}' -NCM $\widehat{M}(\theta) = \langle \widehat{\mathbf{U}}, \mathbf{V}, \widehat{\mathcal{F}}, \widehat{P}(\widehat{\mathbf{U}}) \rangle$, there is a \mathcal{P} -NCM $\mathcal{N}(\theta', \mathbf{m}^D, \mathbf{m}^B, \mathbf{m}^C) = \langle \mathbf{V}', \mathbf{U}', P(\mathbf{U}'), \mathcal{F}' \rangle$ with the above masks that induces the same $\mathbf{P}^{\mathcal{L}_3}$ distributions as $\widehat{M}(\theta)$.

First, we note that $\mathbf{V} = \mathbf{V}'$, and $\widehat{\mathbf{U}} \subseteq \mathbf{U}'$ as \mathbf{U}' contains one U for each possible bidirected clique, while $\widehat{\mathbf{U}} \subseteq \mathbf{U}'$ only contains one U for each maximal clique in \mathcal{G}' .

We define

- $P(\mathbf{U}') = P(\widehat{\mathbf{U}})P(\mathbf{U}' \setminus \widehat{\mathbf{U}})$ where variables in $\mathbf{U}' \setminus \widehat{\mathbf{U}}$ also follow $Unif(0, 1)$.
- $\theta' = \theta \cup \omega$, where ω contains all other parameters of the \mathcal{P} -NCM.
- For each variable $V_i \in \mathbf{V}$, define

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

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 $\mathbf{v} \in \text{dom}(\mathbf{V})$,

$$\begin{aligned} & \mathbf{P}^{\mathcal{N}(\mathbf{m}^D, \mathbf{m}^B, \mathbf{m}^C, \theta')}(\mathbf{y}_{\mathbf{x}}, \dots, \mathbf{z}_{\mathbf{w}}) \\ &= \sum_{\mathbf{u}'} \mathbf{1} \left[\mathbf{Y}_{\mathbf{x}}^{\theta'}(\mathbf{m}_{:,Y}^C \odot \mathbf{u}') = \mathbf{y}, \dots, \mathbf{Z}_{\mathbf{w}}^{\theta'}(\mathbf{m}_{:,Z}^C \odot \mathbf{u}') = \mathbf{z} \right] P(\mathbf{u}') \\ &= \sum_{\mathbf{u}'} \left(\prod_{V_i \in \mathbf{Y} \setminus \mathbf{X}} \mathbf{1} \left[f'_i(\mathbf{m}_{:,i}^D \odot \mathbf{v}', \mathbf{m}_{:,i}^C \odot \mathbf{u}') = v_i \right] \right) \left(\prod_{V_i \in \mathbf{X}} \mathbf{1} [V_i = x_i] \right) \\ & \quad \cdot \left(\prod_{V_j \in \mathbf{Z} \setminus \mathbf{W}} \mathbf{1} \left[f'_j(\mathbf{m}_{:,j}^D \odot \mathbf{v}', \mathbf{m}_{:,j}^C \odot \mathbf{u}') = v_j \right] \right) \left(\prod_{V_j \in \mathbf{W}} \mathbf{1} [V_j = w_j] \right) P(\mathbf{u}') \\ & \hspace{15em} \text{(By Def. 8)} \\ &= \sum_{\mathbf{u}} \left(\prod_{V_i \in \mathbf{Y} \setminus \mathbf{X}} \mathbf{1} \left[f_i^{\theta}(\mathbf{p}a_i, \mathbf{u}_i) = v_i \right] \right) \left(\prod_{V_i \in \mathbf{X}} \mathbf{1} [V_i = x_i] \right) \\ & \quad \cdot \left(\prod_{V_j \in \mathbf{Z} \setminus \mathbf{W}} \mathbf{1} \left[f_j^{\theta}(\mathbf{p}a_j, \mathbf{u}_j) = v_j \right] \right) \left(\prod_{V_j \in \mathbf{W}} \mathbf{1} [V_j = w_j] \right) P(\mathbf{u}) \sum_{\mathbf{u}' \setminus \mathbf{u}} P(\mathbf{u}' \setminus \mathbf{u}) \\ & \hspace{15em} \text{(By definition of } \mathcal{F}' \text{ and } P(\mathbf{U}')) \\ &= \mathbf{P}^{\mathcal{M}}(\mathbf{y}_{\mathbf{x}}, \dots, \mathbf{z}_{\mathbf{w}}). \hspace{10em} \text{(By Def. 19)} \end{aligned}$$

□

C Proofs

Lemma 1. [MAG – PAG MEC Connection] Given a PAG \mathcal{P} and a MAG $M \in \mathcal{P}$, the MEC signature output from the MAG-to-MEC algorithm ([6, Alg. 1]) is identical when applied to either \mathcal{P} or M .

Proof. The result follows immediately from Lem. 4. □

Theorem 1. [Characterization of Causal Diagrams Represented by a PAG] A causal diagram \mathcal{G} lies in the MEC represented by a PAG \mathcal{P} if and only if the following conditions hold:

1. **Acyclicity:** \mathcal{G} contains no directed cycles;
2. **Skeleton:** Variables adjacent in \mathcal{P} if and only if connected by an inducing path in \mathcal{G} ;
3. **Collider constraints:** Every triple in \mathfrak{C} that appears in \mathcal{G} is a collider in \mathcal{G} ;
4. **Non-collider constraints:** Every triple in \mathfrak{D} that appears in \mathcal{G} is a non-collider in \mathcal{G} .

Proof. Since the transformation from a causal diagram to its associated MAG is fixed, membership of a causal diagram in the MEC represented by \mathcal{P} can be checked through its induced MAG. Thus, if violating any of the above constraints yields an induced MAG that violates one of the MEC conditions in Lem. 3, then these constraints characterize exactly the causal diagrams whose induced MAG lies in the MEC of \mathcal{P} .

First, we show that if all conditions are satisfied, the graph \mathcal{G} is a causal diagram in \mathcal{P} . Condition (1) ensures that \mathcal{G} is acyclic and hence a causal diagram. Condition (2) ensures that \mathcal{G} induces a MAG with skeleton \mathfrak{S} . By Lem. 8, conditions (3) and (4) ensure that the induced MAG has the same sets \mathfrak{C} and \mathfrak{D} as \mathcal{P} . Therefore, the induced MAG has the same MEC signature as \mathcal{P} , and thus lies in the MEC of \mathcal{P} by Lem. 3.

Second, we show that if any condition is violated, then \mathcal{G} is either not a causal diagram or does not lie in the MEC of \mathcal{P} . If condition (1) is violated, then \mathcal{G} is not acyclic and hence not a causal diagram. If condition (2) is violated, the MAG induced by \mathcal{G} has a skeleton $\mathfrak{S}' \neq \mathfrak{S}$ by Algo. 4 and therefore lies outside the MEC of \mathcal{P} by Lem. 3. Similarly, if conditions (3) or (4) are violated, the induced MAG has $\mathfrak{C}' \neq \mathfrak{C}$ or $\mathfrak{D}' \neq \mathfrak{D}$, and thus lies outside the MEC of \mathcal{P} by the same lemma. \square

Proposition 1. [Correctness of Non-collider Constraint] Given a PAG \mathcal{P} with MEC signature $\langle \mathfrak{S}, \mathfrak{C}, \mathfrak{D} \rangle$ and a causal diagram \mathcal{G} with weighted adjacency matrices $(\mathbf{m}^D, \mathbf{m}^B)$, \mathcal{G} satisfies the non-collider constraints with respect to \mathcal{P} in Thm. 1 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 Thm. 1 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 $\mathbf{m}_{ij}^D > 0$ or $\mathbf{m}_{ij}^B > 0$.

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 constraint in Thm. 1. Therefore, the constraint is satisfied if and only if

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

\square

Proposition 2. [Correctness of inducing path constraint] Given a \mathcal{P} with MEC signature $\langle \mathfrak{S}, \mathfrak{C}, \mathfrak{D} \rangle$ and a causal diagram \mathcal{G} with weighted adjacency matrices $(\mathbf{m}^D, \mathbf{m}^B)$, \mathcal{G} satisfies the skeleton constraints with respect to \mathcal{P} in Thm. 1 if and only if $p^{\mathcal{P}}(\mathbf{m}^D, \mathbf{m}^B, \mathfrak{S}) = 0$.

Proof. All matrices in Def. 5 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

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

Next, by construction from the weighted adjacency matrices,

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

and

$\text{ArrAt}_{uv} > 0 \iff V_u \rightarrow V_v \text{ or } V_u \leftrightarrow V_v$ (i.e., the edge between V_u and V_v 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

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

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

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

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

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

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,

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

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 $K \geq n - 1$,

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

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

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

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} . \square

Proposition 3. [*\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 causal diagram \mathcal{G} (Def. 7) and counterfactual distributions $\mathbf{P}^{\mathcal{L}_3}(\mathcal{N})$ (Def. 8). Then, 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 Prop. 5. \square

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 $\mathbf{P}^{\mathcal{L}_3}(\mathcal{N}) = \mathbf{P}^{\mathcal{L}_3}(\mathcal{M})$.

Proof. By Lem. 12, for any SCM \mathcal{M} with causal diagram $\mathcal{G} \in \mathcal{P}$, there exists an SCM \mathcal{M}' whose induced causal diagram \mathcal{G}' satisfies $\mathcal{G}' \supseteq \mathcal{G}$ and induces the same counterfactual distributions $\mathbf{P}^{\mathcal{L}_3}$ as \mathcal{M} .

By Lem. 13, 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. 14, there exists a \mathcal{P} -NCM \mathcal{N} that induces the same $\mathbf{P}^{\mathcal{L}_3}$ as $\widehat{\mathcal{M}}$. \square

Theorem 3. [Correctness of PAG Neural Partial ID] The true optimal counterfactual bounds (Def. 9) and the optimal neural counterfactual bounds (Def. 10) coincide.

Proof. By Prop. 5, every \mathcal{P} -NCM is equivalent to an SCM on counterfactual valuations that induces the same causal diagram. Therefore, optimizing over \mathcal{P} -NCMs is equivalent to optimizing over SCMs whose induced causal diagrams are representable by such models. In particular, this procedure does not leave the space of SCMs.

Acyclicity are enforced using standard constraints such as NOTEARS or DAGMA, which are known to be correct. Moreover, all collider constraints in \mathcal{P} are satisfied by construction, as the corresponding masks are fixed to one (Def. 6). In addition, Prop. 1 and Prop. 2 establish the correctness of the non-collider and inducing-path constraints. Consequently, the causal diagram \mathcal{G} induced by any feasible \mathcal{P} -NCM satisfies Thm. 1 and thus lies in the MEC of \mathcal{P} . This restricts the search space to SCMs whose causal diagrams \mathcal{G} belong to \mathcal{P} , i.e., $\Omega(\mathcal{P})$, ensuring that any obtained bound is valid.

Conversely, by the expressiveness of \mathcal{P} -NCMs (Thm. 2), for every SCM inducing a causal diagram in \mathcal{P} , there exists a \mathcal{P} -NCM that induces the same counterfactual distributions $\mathbf{P}^{\mathcal{L}_3}$. In particular, there exist \mathcal{P} -NCMs that match the SCMs attaining the lower and upper bounds. Therefore, the computed bounds $[l, r]$ coincide with the true optimal counterfactual bounds. \square

D Experiments

D.1 Experimental Setup

D.1.1 A Relaxed Training Objective

In this section, we illustrate a practical implementation to solve the challenging, non-convex optimization problem of neural partial identification in \mathcal{P} -NCMs.

Fitting the observational distribution. First, while the optimization is formulated in terms of the population $P(\mathbf{v})$, in practice we have a finite dataset of i.i.d. samples $\mathcal{D} = \{\mathbf{v}_1, \dots, \mathbf{v}_m\}$. Let $\hat{P}(\mathbf{v})$ be the empirical distribution and $\mathbb{D}(\cdot|\cdot)$ be some divergence function between empirical distributions, e.g., KL-divergence, negative log-likelihood, or max-mean discrepancy.

Optimizing the causal query. Given a PAG \mathcal{P} , we learn the parameters of \mathcal{P} -NCMs as follows. By design, some masking parameters are fixed in Def. 6. Then, extending the objective of standard neural causal models to our case, to maximize/minimize a query $P(\mathbf{y}_*|\mathbf{x}_*)$, we minimize the loss

$$\mathcal{L}(\theta, \mathbf{m}) = \mathbb{D}(\hat{P}^{\mathcal{N}(\theta, \mathbf{m})}(\mathbf{V}^n) || \hat{P}(\mathbf{V})) + \lambda_h \cdot h(\mathbf{m}^D) + \lambda_{nc} \cdot nc(\mathbf{m}^D, \mathbf{m}^B) + \lambda_p \cdot p(\mathbf{m}^D, \mathbf{m}^B) \pm \mathbb{D}(\hat{P}^{\mathcal{N}(\theta, \mathbf{m})}(\mathbf{y}_*^{(n)} | \mathbf{x}_*) || \hat{P}(\mathbf{y}_* | \mathbf{x}_*))$$

where $\hat{P}(\mathbf{y}_* | \mathbf{x}_*)$ comprises our desired empirical distribution, e.g., a sample with all $\mathbf{y} = 1$ if we want to maximize $P(\mathbf{y} = 1 | do(\mathbf{x}))$.

We implement an attention mechanism that respects the non-parent invariance of Def. 6. At a high level, the architecture takes the exogenous variables \mathbf{U} and the endogenous variables \mathbf{V} as token inputs and produces predictions for all endogenous \mathbf{V} as token outputs. At each position $i \leq n$, the input is the exogenous draw U_i and the corresponding output is the endogenous prediction V_i ; in this sense the attention block serves as the structural function \hat{f}_i , mapping U_i to V_i . In Def. 6, for a variable V_i , we have

$$V_i \leftarrow f_i(\mathbf{m}_{:,i}^D \odot \mathbf{V}, \mathbf{m}_{:,i}^C \odot \mathbf{U}; \theta_i). \quad (22)$$

To guarantee that f_i does not depend on masked values v_j in \mathbf{v}_i where $m_{ji} = 0$, we inject the $\mathbf{m}_{:,i}^D$ and $\mathbf{m}_{:,i}^C$ as attention mask for gating the non-parent variables causing V_i .

D.1.2 Reinforcement learning over the space of masks

The differentiable-mask formulation searches over a continuous relaxation of adjacency matrices, and generates samples without hard thresholding of these masks. While this is desirable for gradient-based

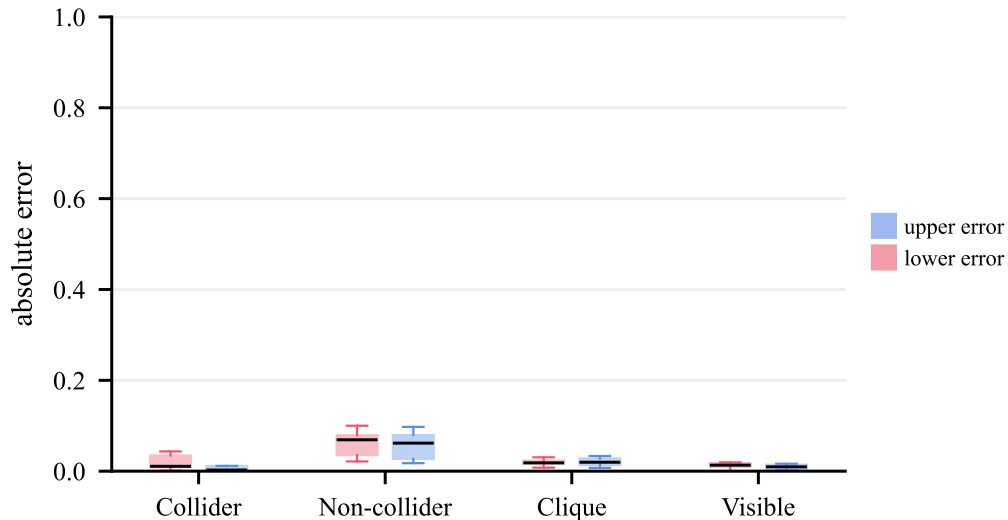


Figure 8: Error plot for experiments of neural partial identification for PAGs in Fig. 5.

optimization, it may come at a cost of accuracy. We next present an alternative search strategy over masks that leverages thresholding in the attention-based architecture.

At a high-level, we indirectly sample from the space of masks by sampling from the space of topological orderings over n nodes. We parameterize a distribution over orders using Plackett–Luce scores: a higher score for a variable V_i makes it more likely to appear earlier in the sampled topological order. Given a sampled order π , we construct a hard acyclic mask $m(\pi)$ and evaluate the \mathcal{P} -NCM under that mask. Since all oriented reversible edges point forward in π , acyclicity is guaranteed by construction; only the collider constraints remain to be enforced. We then optimize over the space of these ordering probabilities and attention parameters with a reward dependent on data fit and collider constraints.

D.2 Additional Experimental Results

Experimental errors are plotted in Fig. 8.