Causal Order: The Key to Leveraging Imperfect Experts in Causal Inference

Aniket Vashishtha
UIUC
vashisthaaniket07@gmail.com

Abbavaram Gowtham Reddy
IIT Hyderabad, India
cs19resch11002@iith.ac.in

Abhinav Kumar
MIT
akumar03@mit.edu

Saketh Bachu
IIT Hyderabad, India
sakethvnit@gmail.com

Vineeth N Balasubramanian
IIT Hyderabad, India
vineethnb@iith.ac.in

Amit Sharma
Microsoft Research, India
amshar@microsoft.com

Abstract

Large Language Models (LLMs) have recently been used to infer causal graphs, often by repeatedly applying a pairwise prompt that asks about causal relationship of each variable pair. However, we identify a key limitation of using graphs as the output interface for the domain knowledge provided by LLMs and other imperfect experts. Even perfect experts cannot distinguish between direct and indirect edges given a pairwise prompt, leading to unnecessary errors. Instead, we propose *causal order* as a more stable output interface that experts should be evaluated on. Causal order is also a useful structure by itself; we show, both theoretically and empirically, that causal order better correlates with effect estimation error than commonly used graph metrics. We propose a triplet-based prompting method that considers three variables at a time rather than a pair of variables. For both LLMs and human annotators as experts, the proposed triplet method leads to more accurate causal order with significantly fewer cycles. We also show how the estimated causal order can be used to reduce error in downstream discovery and effect inference.

1 Introduction

Based on evidence that LLMs' domain knowledge, even if imperfect, can be used to decide the direction of causal relationship between a pair of variables Kıcıman et al. [2023], Willig et al. [2022], LLMs have been used to infer the entire causal graph for a problem by invoking the same pairwise prompt—does A cause B?—multiple times for different pairs of variables Long et al. [2022], Antonucci et al. [2023], Kıcıman et al. [2023], Cohrs et al. [2023]. Others have used the graph or edges obtained from LLMs as a prior Takayama et al. [2024] or constraint Long et al. [2023], Khatibi et al. [2024], Ban et al. [2023a] for causal discovery algorithms and shown that the input graph from LLMs can improve accuracy of graph discovery.

However, there is a key limitation of using graphs as the *output interface* for the domain knowledge provided by LLMs and other imperfect experts. Obtaining the complete graph requires distinguishing between direct and indirect effects among variables. Given only a pair of variables, it is impossible to decide whether an edge exists or it is mediated by another variable, even for a perfect human expert; because the existence of an edge depends on which other variables are considered to be a part of the node set. For example, consider the true data-generating process, *lung cancer* $\rightarrow$ *doctor visit* $\rightarrow$ *positive Xray*. If an expert is asked whether there should be a direct causal edge from *lung cancer* to *positive Xray*, they would answer "Yes" (indeed, such an edge exists in the BNLearn *Cancer* dataset [Scutari and Denis, 2014]). However, if they are told that the set of observed variables additionally includes *doctor visit*, then the correct answer would be to not create a direct edge between

NeurIPS 2024 Workshop on Causality and Large Models (CaLM).

lung cancer and *positive Xray*, but rather create edges mediated through *doctor visit*. In large graphs, keeping track of the different variables that can affect a given pairwise decision can be cumbersome.

Instead, we propose *causal order* as a more stable interface between experts' domain knowledge and downstream causal algorithms. Causal order is defined as the topological ordering over graph variables. Formally, we show that for an (optimal) perfect expert that is given only a subset of observed variables at a time, the predicted causal graph can be incorrect but the predicted causal order is always correct. As a result, the standard practice of obtaining a causal graph from *imperfect* experts such as LLMs and crowd-sourced human annotators (using pairwise questions) may include many errors in the inferred edges, leading to further errors in the downstream causal algorithms; which can be alleviated using a causal order. Moreover, order is a more stable causal construct since it does not depend on the existence of other variables, thus being more generalizable to new scenarios.

While the causal order is a simpler structure than the full graph, it is useful by itself, aiding downstream tasks like effect inference and graph discovery. For example, correct causal order is sufficient for identifying a valid backdoor set for any pair of treatment and outcome variables. Moreover, a causal order-based metric, topological divergence (D_{top}), correlates better with the effect estimation accuracy than commonly used graph metrics such as structural hamming distance (*SHD*). Specifically, $D_{top} = 0$ if and only if the causal order provides a valid backdoor adjustment set. In contrast, there exist predicted graphs where the identified backdoor set is accurate (and topological divergence is zero), but their SHD can be arbitrarily high. In addition to effect inference, causal order can also be used to improve the accuracy of graph discovery algorithms. To this end, we provide simple algorithms for using causal order to improve existing causal discovery algorithms.

Hence, given the ambiguity in graph edges obtained from experts and the utility of causal order for downstream tasks, we posit that causal order is more suitable for evaluating the quality of a causal structure provided by experts than commonly used metrics on edge accuracy. In practice, however, obtaining causal order from experts is still a challenge because we need to account for *imperfect experts* such as crowd-sourced human annotators and LLMs. Using the standard method Kıcıman et al. [2023], Long et al. [2022] of iterating with a pairwise prompt/question over a set of variables, while a perfect expert would always predict the correct causal order, we find that using LLMs as experts leads to many cycles. To reduce the number of cycles from LLM output, we propose a novel *triplet-based* prompting strategy for obtaining causal order. Rather than asking questions about a pair of variables, the triplet prompt asks about the causal relationship between three variables at once. Since each variable pair occurs in more than one triplet, this results in multiple, possibly conflicting predictions for each pair ($A \to B, B \to A$, or no edge), which are then aggregated based on majority vote. We theoretically show that given an imperfect expert with an error ϵ on each prediction, using the triplet-based prompt results in an error less than ϵ, which is less than the error of the pairwise prompt. For practical usage, we use the variance in votes for each pair to motivate a final edge removal step that ensures that no cycles are present in the final output. Comprehensive experiments using both LLMs and human annotators as imperfect experts demonstrate our contributions.

2 Causal Order: A Stable Output Interface for Experts' Domain Knowledge

Let $\mathcal{G}(\mathbf{X}, \mathbf{E})$ be a causal DAG with variables $\mathbf{X} = X_1, \ldots, X_n$ and directed edges $\mathbf{E}$ among them. An edge $X_i \to X_j \in \mathbf{E}$ indicates a direct causal influence of X_i on X_j. The parents of X_i, denoted $pa(X_i)$, are the set $X_k \mid X_k \to X_i$, and its descendants, $de(X_i)$, are the set $X_k \mid X_k \leftarrow \cdots \leftarrow X_i$. If X_k is a descendant of X_i but not directly connected, X_i has an indirect effect on X_k. The causal (topological) order is then defined as a sequence (or ordered permutation) π of variables $\mathbf{X}$ is a topological order iff for each edge $X_i \to X_j \in \mathbf{E}$, $\pi_i < \pi_j$.

To evaluate the quality of a causal order relative to a ground-truth graph, we use the topological divergence metric $D_{\text{top}}(\hat{\pi}, A)$, defined as $D_{top}(\hat{\pi}, A) = \sum_{i=1}^{n} \sum_{j:\hat{\pi}_i > \hat{\pi}_j} A_{ij}$ where $A_{ij} = 1$ if there is a directed edge from node i to j and 0 otherwise. This metric counts the edges that cannot be recovered due to the estimated topological order $\hat{\pi}$. While a perfect expert provides an accurate causal order, the causal graph may still be inaccurate because determining an edge between two variables requires knowledge of potential mediators. Our approach formalizes the idea that an LLM's pairwise answers can reflect the ancestor-descendant relationship between variables, demonstrating that causal order remains invariant to the presence of unobserved variables. Detailed proofs are provided in Appendix E.

Proposition 2.1. *Consider a Perfect Expert that infers causal relationships based on its (optimal) domain knowledge. It is given only a subset of observed variables* $\mathbf{V}$ *to predict the causal structure. Assuming that the Perfect Expert assumes causal sufficiency (the variables not given do not exist in the system), the predicted causal structure by the Expert might be wrong whereas the predicted causal order remains correct.*

These observations suggest using causal order as the output structure for experts' domain knowledge instead of a graph. It facilitates local engagement and objective evaluation via topological divergence, which is zero for a perfect expert, unlike Structural Hamming Distance (SHD) (see § 5 for examples).

3 Downstream Utility of Causal Order: Discovery and Effect Inference

While the causal order is a more stable measure of experts' knowledge than the full graph, a natural question is whether it is a useful measure by itself too. We now show the utility of causal order for effect estimation and causal discovery, which are also supported strongly by our experimental results in Sec 5. Specifically, we show that although backdoor adjustment is defined with respect to the entire graph, causal order is sufficient to find a valid backdoor set. Moreover, the effect estimation error correlates more with topological divergence than it does with SHD. Causal order is also useful as a prior or constraint to increase accuracy of discovery algorithms.

Proposition 3.1. *[Pearl, 2009, Cinelli et al., 2022] Under the no latent confounding assumption, for a pair of treatment and target variables* (X_i, X_j) *in a DAG* $\mathcal{G}$, $\mathbf{Z} = \{X_k | \pi_k < \pi_i\}$ *is a valid adjustment set relative to* (X_i, X_j) *for any topological order* π *of* $\mathcal{G}$.

Proofs are provided in Appendix § E. Propn D.1 states, in simple words, that all variables that precede the treatment variable in a topological order π of $\mathcal{G}$ constitute a valid adjustment set.

Note that the set $\mathbf{Z}$ may contain variables that are not necessary to adjust for (e.g., ancestors of only treatment or target variables). For statistical efficiency purposes, ancestors of the target variable are helpful for precise effect estimation, whereas ancestors of treatment variable can be harmful [Cinelli et al., 2022]. However, asymptotically, as the number of data points increases, such variables do not have any impact on the estimation. We now show that D_{top} is an optimal metric to minimize for effect estimation. That is, D_{top} being 0 for a topological order is equivalent to obtaining the correct backdoor adjustment set using Proposition D.1. And if $D_{top} \neq 0$, there exists some treatment-target pair whose backdoor set is not correctly identified.

Proposition 3.2. *For an estimated topological order* $\hat{\pi}$ *and a true topological order* π *of a causal DAG* $\mathcal{G}$ *with the corresponding adjacency matrix* A, $D_{top}(\hat{\pi}, A) = 0$ *iff* $\mathbf{Z} = \{X_k | \hat{\pi}_k < \hat{\pi}_i\}$ *is a valid adjustment set relative to* (X_i, X_j), $\forall \pi_i < \pi_j$.

Comparison with SHD. In comparison, the widely used metric for evaluating graphs, Structural Hamming Distance (SHD), does not share such properties. Given a true causal DAG $\mathcal{G}$ and an estimated causal DAG $\hat{\mathcal{G}}$, SHD counts the number of missing, falsely detected, and falsely directed edges in $\hat{\mathcal{G}}$. Formally, D_{top} acts as a lower-bound on SHD [Rolland et al., 2022]. However, SHD is not an ideal metric for evaluating downstream effect estimation accuracy. Specifically, we show that SHD can be high even when $D_{top} = 0$ and a valid backdoor set can be inferred. This result is of significance since most estimated graphs (including those that are LLM-generated [Ban et al., 2023b, Long et al., 2023]) are evaluated using SHD.

Definition 3.1. ***Level Order****. Given a causal DAG* $\mathcal{G}(\mathbf{X}, \mathbf{E})$*, its level order refers to a systematic assignment of levels to variables. This assignment begins with assigning level* 0 *to the set of variables* $\{X_i | pa(X_i) = \emptyset\}$*. Subsequently, each of the remaining variables is assigned a level* i *such that all variables within a given level* i *have a directed path of length* i *from one/more variables in level 0.*

Proposition 3.3. *In a causal DAG* $\mathcal{G}$ *with* N *levels in the level order of variables (level* i *contains* n_i *variables),* $\exists\, \hat{\mathcal{G}}$ *s.t.* $SHD(\hat{\mathcal{G}}, \mathcal{G}) \geq \sum_{i=1}^{N-1} (n_i \times \sum_{j=i+1}^{N} n_j) - |\mathbf{E}|$ *and* $D_{top}(\hat{\pi}, A) = 0\, \forall \hat{\pi}$ *of* $\hat{\mathcal{G}}$.

Fig 1 shows the results of a simple study that highlights the limitations of SHD in the context of our work.

Given a fixed number of nodes, we sample a graph at random as the 'ground truth' and then consider all graph orientations of the same size (number of nodes) such that $D_{top} = 0$ with respect to (w.r.t) the ground truth graph. For this set of graphs, we compute SHD w.r.t the ground truth graph. Notice the variance in SHD, despite D_{top} being 0. For graphs with six nodes, SHD can vary from 0 to 14 even as $D_{top} = 0$ and backdoor set validity stays the same. The same pattern translates to effect estimation error (shown in Table A3). For the same SHD, increasing D_{top} leads to an increase in estimation error; but if D_{top} is the same, increasing SHD has no effect on effect estimation error. The correlation of D_{top} with effect estimation is also shown in Appendix G with additional results.

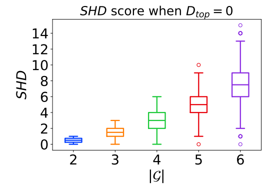


Figure 1: Variability of SHD for various graph sizes with $D_{top} = 0$ within each graph.

Topological order can improve accuracy of graph discovery algorithms. Constraints implied by the topological order can be used to reduce the search space for discovery algorithms. For instance, if $X_i \prec X_j$ in the order, then X_i cannot be a descendant of X_j in the corresponding causal graph.

4 Obtaining a Causal Order from Imperfect Experts

While we assumed a perfect expert so far, obtaining causal order from an imperfect expert requires the design of effective querying strategies. One naive way is to provide the expert with all nodes in the graph and ask for the whole causal order in one go. But this might lead to information overload for an expert and thus a suboptimal result. For e.g., if we use LLMs as an expert, they might hallucinate with a large context size.

Prop 2.1 show that an expert predicts correct causal order even when given a limited set of observed nodes. Hence, we resort to querying the order by giving partial information (*local query*) to the expert. Going beyond existing *pairwise querying* strategies, we propose a novel *triplet-based* query strategy that aids in mitigating cycles in the obtained causal graph and has provably better performance (see Prop E.2).

Existing Efforts: Pairwise Queries. A natural way to elicit causal order from experts, also adopted by recent work on inferring graph edges [Kıcıman et al., 2023, Ban et al., 2023b, Long et al., 2022], is to ask about each pair of variables and aggregate the results. Beyond a basic prompt, we also study augmented pairwise strategies with additional contextual information. We study four types of pairwise query strategies, which are briefly summarized below and described in Appendix § I: **Basic query:** This is the simplest technique. We directly ask the expert to identify the causal direction between a given pair of variables [Kıcıman et al., 2023]. **Iterative Context:** Here we provide the previously oriented pairs as context in the query while iteratively prompting for next pair. **Markov Blanket Context:** Providing previously oriented pairs can become prohibitive for large graphs. As a variable is independent of other nodes given its Markov Blanket [Pearl, 2009], we include the Markov Blanket of the node pairs as additional context in the query.

Chain-of-Thought (+In-context Learning): Building on previous findings regarding in-context examples in LLM prompts for various tasks [Brown et al., 2020], we present examples of the ordering task (node pairs and their correct causal ordering) before posing the question about the given nodes. Refer to Appendix § I for implementation details.

Proposed Triplet-based Queries. As we shall see, while pairwise prompts are conceptually simple, they are prone to yielding higher number of cycles in the graph since they decide about each edge separately. Taking inspiration from the PC algorithm that employs constraints over three variables, we now describe a prompting technique based on iterating over all possible triplets given a set of nodes. Once the LLM has provided subgraphs for each triplet, we determine causal order between a pair by aggregating over all triplet LLM answers where the pair was included. The algorithm has the following steps after generating all possible triplets from a set of nodes. **1)** Generate subgraphs over all triplets through LLMs by prompting them to orient the three causal edges for each triplet. **2)** Merge the resultant structure between any two nodes by aggregating the number of LLM answers for the three orientations: (A $\rightarrow$ B; B $\rightarrow$ A; No connection) and choosing the majority answer. If there's a tie in edge orientation, GPT-4 is used with a CoT prompt to make the final decision. Then the causal order is extracted from the graph. Our triplet queries additionally use in-context examples and the chain-of-thought strategy from the pairwise setup. An example query is shown in Table A24. Now, we analyze the triplet strategy for its impact on predicting incorrect edges. We begin by defining (imperfect) ϵ-experts (inspired from Long et al. [2023]):

Definition 4.1 (ϵ-Experts)**.** *Given two nodes A and B of a graph and three options of the causal relationship between them: (i) $A \rightarrow B$, (ii) $A \leftarrow B$, and (iii) no edge between A and B (denoted as $[c_1, c_2, c_3]$), an expert $\mathcal{E}$ queried for the causal relationship between A and B is said to be an ϵ-expert (denoted as $\mathcal{E}_\epsilon$) if the probability of making an error in the prediction of the causal relationship between A and B is ϵ, where $\epsilon \in (0, 1)$.*

Referring E.2 and E.2, a querying strategy using triplets will have error probability $< \epsilon$ over determining the causal relationship between A and B than a pairwise strategy (proof in Appendix E). Still, some cycles may be produced, we hence use a cycle removal algorithm inspired from [Zheng et al., 2018] as the last stage of our triplet method. We leverage triplet pipeline votes to establish a probability distribution over edge orientations. We use this to compute entropy for each edge, removing those with higher entropy (lower confidence). To minimize D_{top}, we prune edges with entropy below the mean of all entropies.

5 Experiments and Results

We performed a comprehensive suite of experiments across multiple benchmark datasets from the BNLearn repository [Scutari and Denis, 2014]: Earthquake, Cancer, Survey, Asia, Asia modified (Asia-M), and Child, to validate the usefulness of the proposed framework. Asia-M is derived from Asia by removing the node *either* since it is not a node with a semantic meaning (see Appendix§ I for details). We also used a medium-sized subset graph (see Appendix Fig A8) from the Neuropathic dataset [Tu et al., 2019] used for pain diagnosis to explicitly study memorization in the context of our experiments.We considered two types of imperfect experts in our empirical studies: LLMs and human annotators.

Our human studies used 15 annotators, each with undergrad-level training in STEM but no formal experience in causality. Each annotator was randomly allotted a graph for both pairwise and triplet query strategies while ensuring no annotator got the same graph to query with both strategies. For the triplet method, to get an estimate of the upper bound of human performance in such studies, we used a ground truth-based oracle (proxy for a human domain expert) to resolve conflicts in the triplet. For LLM-based expert assessment, GPT-3.5 turbo was used as the imperfect expert for causal order, and GPT-4 for tie-breaking. We first present the results of using our triplet-based approach over other existing pairwise query strategies. Subsequently, we present the results of using the causal order obtained from imperfect experts to downstream tasks including causal discovery and effect inference.

Dataset	Metric	Pairwise (CoT)	Triplet
Using LLM			
Earthquake	D_{top}	**0**	**0**
	SHD	**4**	**4**
	Cycles	**0**	**0**
	IN/TN	**0/5**	**0/5**
Asia-M	D_{top}	-	**1**
	SHD	13	**11**
	Cycles	1	**0**
	IN/TN	**0/7**	**0/7**
Child	D_{top}	-	**1**
	SHD	138	**28**
	Cycles	»500	**0**
	IN/TN	**0/20**	10/20
Using Human Annotators			
Earthquake	D_{top}	**0**	**0**
	SHD	4.67	**1.67**
	Cycles	**0**	**0**
	IN	**0**	0.33
Asia-M	D_{top}	-	**1.33**
	SHD	11.67	**11.33**
	Cycles	3	**0**
	IN	**0**	**0**

Table 1: *(Top)* Results using LLM *(Bottom)* Results using human annotators (mean value across annotators reported). Performance of triplet method vs best performing pairwise query strategy (Chain of Thought) on multiple benchmark datasets across diff metrics: D_{top}, SHD, (Num of) Cycles, IN (Isolated Nodes), TN (Total Nodes). When num of cycles>0, $\hat{\pi}$ cannot be computed, hence D_{top} is given by '-'. Triplet consistently outperforms the pairwise strategy across metrics & datasets, esp by significant amounts on larger graphs like *Child*. Refer A5 for results on other graphs

Triplet vs Pairwise Query Strategies. Table A5 and 1 presents the experimental results of a direct comparison between triplet and pairwise query strategies on multiple benchmark datasets on different metrics: D_{top}, SHD, (Num of) Cycles, IN (Isolated Nodes), TN (Total Nodes), using LLMs (*top*) and human annotators (*bottom*). Using GPT-3.5 as the imperfect expert, the triplet query strategy outperforms the best performing pairwise approach (CoT) consistently across datasets (see Appendix G.3 for full pairwise results), the difference is even more pronounced, especially in the number of cycles. Results on Neuropathic dataset presented in appendix G.4 further corroborates this claim. Note that the triplet method prioritizes precision over coverage; it hence obtains the lowest D_{top} and zero cycles on Child but with isolated nodes as 10 out of 20. We believe this is an appropriate tradeoff for the output of an expert, i.e. its output is aimed to be correct but it may not cover all nodes. Similarly, even with human annotators, graphs like *Survey* and *Asia-M* result in cycles when queried pairwise. However, no cycle formations were observed across annotators when they were queried to orient causal graphs using the triplet strategy. Also, the triplet strategy

	Dataset	PC	SCORE	ICA LiNGAM	Direct LiNGAM	NOTEARS	CaMML	Ours (PC+LLM)	Ours (CaMML+LLM)	Ours (PC+Human)	Ours (CaMML+Human)
$N = 250$	Earthquake	0.16±0.28	4.00±0.00	3.20±0.39	3.00±0.00	1.80±0.74	2.00±0.00	**0.00±0.00**	**0.00±0.00**	**0.00±0.00**	1.00±0.00
	Cancer	**0.00±0.00**	3.00±0.00	4.00±0.00	3.60±0.48	2.00±0.00	2.00±0.00	**0.00±0.00**	**0.00±0.00**	**0.00±0.00**	**0.00±0.00**
	Survey	0.50±0.00	3.00±0.00	6.00±0.00	6.00±0.00	3.20±0.39	3.33±0.94	**0.00±0.00**	3.33±0.94	**0.00±0.00**	**0.00±0.00**
	Asia	2.00±0.59	5.00±0.00	6.20±0.74	7.00±0.00	4.00±0.00	1.85±0.58	1.00±0.00	**0.97±0.62**	N/A	N/A
	Asia-M	1.50±0.00	5.00±0.00	7.60±0.48	6.20±1.16	3.40±0.48	**1.00±0.00**	**1.00±0.00**	1.71±0.45	**1.00±0.00**	2.00±0.00
	Child	5.75±0.00	8.80±2.70	12.8±0.97	13.0±0.63	15.0±1.09	**3.00±0.00**	4.00±0.00	3.53±0.45	N/A	N/A
	Neuropathic	4.00±0.00	6.00±0.00	13.0±6.16	10.0±0.00	9.00±0.00	10.4±1.95	**3.00±0.00**	5.00±0.00	N/A	N/A
$N = 10000$	Earthquake	**0.00±0.00**	4.00±0.00	3.00±0.00	3.00±0.00	1.00±0.00	0.40±0.48	**0.00±0.00**	**0.00±0.00**	**0.00±0.00**	**0.00±0.00**
	Cancer	2.00±0.00	3.00±0.00	3.00±0.00	3.00±0.00	2.00±0.00	0.60±0.80	2.00±0.00	**0.00±0.00**	2.00±0.00	**0.00±0.00**
	Survey	2.00±0.00	4.00±0.00	5.00±0.00	5.00±0.00	3.00±0.00	3.60±1.35	2.00±0.00	1.83±0.00	2.00±0.00	**0.00±0.00**
	Asia	1.5±0.00	4.00±0.00	6.00±0.00	4.40±1.35	3.00±0.00	1.40±0.48	**0.00±0.00**	0.34±0.47	N/A	N/A
	Asia-M	1.00±0.00	4.00±0.00	8.00±0.00	4.80±0.39	3.00±0.00	2.00±0.00	**0.00±0.00**	**0.00±0.00**	**0.00±0.00**	3.00±0.00
	Child	6.00±3.04	3.00±0.00	12.2±1.46	11.6±0.48	14.4±0.48	2.80±0.84	5.00±2.64	**1.00±0.00**	N/A	N/A
	Neuropathic	10.00±0.00	6.00±0.00	**1.00±0.00**	10.0±0.00	10.0±0.00	3.00±0.00	10.00±0.00	**1.00±0.00**	N/A	N/A

Table 2: Comparison with causal discovery methods, showing mean and std dev of D_{top} over 3 runs. (For the Neuropathic subgraph (1k samples), PC Algorithm returns cyclic graphs in the MEC). Human experiments not conducted for Child (due to feasibility issues), hence rows marked as N/A. Refer full table A10

Dataset	Metric: ϵ_{ACE} (Treatment, Target)	PC	SCORE	ICA LiNGAM	Direct LiNGAM	NOTEARS	CaMML	Ours (PC+LLM)	Ours (CaMML+LLM)
Earthquake	(JohnCalls,alarm)	$\mathbf{0.00 \pm 0.00}$	0.85 ± 0.02	0.63 ± 0.10	0.63 ± 0.10	0.21 ± 0.12	0.08 ± 0.03	$\mathbf{0.00 \pm 0.00}$	$\mathbf{0.00 \pm 0.00}$
Cancer	(dyspnoea,cancer)	0.20 ± 0.01	0.30 ± 0.00	0.30 ± 0.01	0.30 ± 0.01	0.18 ± 0.02	0.06 ± 0.00	0.30 ± 0.00	$\mathbf{0.00 \pm 0.00}$
Survey	(T,E)	0.02 ± 0.00	0.04 ± 0.00	0.05 ± 0.01	0.05 ± 0.01	0.03 ± 0.00	0.03 ± 0.00	0.02 ± 0.01	$\mathbf{0.01 \pm 0.01}$
Asia	(smoke,dyspnoea)	0.10 ± 0.00	0.09 ± 0.00	0.27 ± 0.03	0.27 ± 0.04	0.14 ± 0.01	0.05 ± 0.00	0.02 ± 0.00	$\mathbf{0.00 \pm 0.00}$
Child	(Lung Parench, Lowerbody O2)	0.22 ± 0.01	0.02 ± 0.00	0.52 ± 0.00	0.52 ± 0.00	0.52 ± 0.07	0.01 ± 0.00	0.22 ± 0.00	$\mathbf{0.00 \pm 0.00}$

Table 3: Comparison of causal effect inference with existing methods, showing mean and std dev of error in Average Causal Effect (ϵ_{ACE}) of a variable on another, over 3 runs.

showed consistently low D_{top} across all human outputs. Quality of the causal graph curated improves when using the triplet strategy, and expectedly leads to better D_{top} results than the pairwise strategy.

Performance on Downstream Applications: Causal Discovery. Table 2 presents the D_{top} results of using the causal order obtained from expert output (both using LLMs and humans) to assist causal discovery methods. We compare our overall approach using triplet queries with well-known causal discovery methods: PC [Spirtes et al., 2000], SCORE [Rolland et al., 2022], ICA-LiNGAM [Shimizu et al., 2006], Direct-LiNGAM [Shimizu et al., 2011], NOTEARS [Zheng et al., 2018], and Causal discovery via minimum message length (CaMML) [Wallace et al., 1996] across five different sample sizes: $250, 500, 1000, 5000, 10000$. Among the discovery algorithms, we find that PC and CaMML perform the best, with the lowest D_{top} across all datasets. We hence studied 4 variants of our approach: PC+Human, CaMML+Human, PC+LLM, and CaMML+LLM.

The results show that using our approach improves D_{top} across our experiments consistently. Specifically, the improvement (reduction) in D_{top} when using our approach is larger at lower sample sizes. This indicates that obtaining causal order from imperfect experts like humans and LLMs can help with causal discovery in limited sample settings. CaMML+Human/LLM yields benefits even at higher sample sizes. At a sample size of 10,000, CaMML's D_{top} for Child and Asia surpasses CaMML+LLM by three and fivefold respectively. In specific datasets like *Survey* where the variables are better understood by humans, incorporating human priors to CaMML leads to consistently zero D_{top}, outperforming LLM output. Overall, these results show that expert output can significantly improve the accuracy of existing causal discovery algorithms.

Performance on Downstream Applications: Causal Effect Inference. Table 3 presents results of an experimental study that uses the causal order obtained from our LLMs to compute average causal effect (ACE). Taking the output graph from the methods discussed above, we extract and use the backdoor set from them along with the data, to implement causal effect estimation using the dowhy package. We report the error in ACE ϵ_{ACE}. The obtained causal order shows unanimous improvement in performance across the studies, especially when using the causal order from CaMML+LLM.

6 Concluding Discussion

We presented causal order as a suitable output interface to elicit causal knowledge from imperfect experts like LLMs and human annotators. Compared to the full graph, we showed that causal order is a more stable quantity to elicit from imperfect experts since it avoids making a distinction between direct and indirect effects. We also proposed a novel triplet-based method to query experts for obtaining the causal order. Empirical results show the benefits of triplet method in avoiding cycles in the causal order and how causal order helps in improving accuracy of downstream discovery and effect inference tasks.

References

Alessandro Antonucci, Gregorio Piqué, and Marco Zaffalon. Zero-shot causal graph extrapolation from text via llms. *arXiv preprint arXiv:2312.14670*, 2023.

Taiyu Ban, Lyuzhou Chen, Derui Lyu, Xiangyu Wang, and Huanhuan Chen. Causal structure learning supervised by large language model. *arXiv preprint arXiv:2311.11689*, 2023a.

Taiyu Ban, Lyvzhou Chen, Xiangyu Wang, and Huanhuan Chen. From query tools to causal architects: Harnessing large language models for advanced causal discovery from data. *arXiv preprint arXiv:2306.16902*, 2023b.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. *CoRR*, abs/2005.14165, 2020.

Carlos Cinelli, Andrew Forney, and Judea Pearl. A crash course in good and bad controls. *Sociological Methods & Research*, page 00491241221099552, 2022.

Kai-Hendrik Cohrs, Emiliano Diaz, Vasileios Sitokonstantinou, Gherardo Varando, and Gustau Camps-Valls. Large language models for constrained-based causal discovery. In *AAAI 2024 Workshop on"Are Large Language Models Simply Causal Parrots?"*, 2023.

Anthony C. Constantinou, Zhigao Guo, and Neville K. Kitson. The impact of prior knowledge on causal structure learning. *Knowledge and Information Systems*, 65(8):3385–3434, 2023.

Uzma Hasan and Md Osman Gani. Kcrl: A prior knowledge based causal discovery framework with reinforcement learning. In *Proceedings of the 7th Machine Learning for Healthcare Conference*, 2022.

David Heckerman and Dan Geiger. Learning bayesian networks: a unification for discrete and gaussian domains. *arXiv preprint arXiv:1302.4957*, 2013.

Thomas Jiralerspong, Xiaoyin Chen, Yash More, Vedant Shah, and Yoshua Bengio. Efficient causal graph discovery using large language models. *arXiv preprint arXiv:2402.01207*, 2024.

Elahe Khatibi, Mahyar Abbasian, Zhongqi Yang, Iman Azimi, and Amir M Rahmani. Alcm: Autonomous llm-augmented causal discovery framework. *arXiv preprint arXiv:2405.01744*, 2024.

Emre Kıcıman, Robert Ness, Amit Sharma, and Chenhao Tan. Causal reasoning and large language models: Opening a new frontier for causality. *arXiv preprint arXiv:2305.00050*, 2023.

Stephanie Long, Tibor Schuster, and Alexandre Piché. Can large language models build causal graphs? In *NeurIPS 2022 Workshop on Causality for Real-world Impact*, 2022.

Stephanie Long, Alexandre Piché, Valentina Zantedeschi, Tibor Schuster, and Alexandre Drouin. Causal discovery with language models as imperfect experts. In *ICML 2023 Workshop on Structured Probabilistic Inference & Generative Modeling*, 2023.

Christopher Meek. Causal inference and causal explanation with background knowledge. In *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence*, pages 403–410, 1995.

Narmada Naik, Ayush Khandelwal, Mohit Joshi, Madhusudan Atre, Hollis Wright, Kavya Kannan, Scott Hill, Giridhar Mamidipudi, Ganapati Srinivasa, Carlo Bifulco, et al. Applying large language models for causal structure learning in non small cell lung cancer. *arXiv preprint arXiv:2311.07191*, 2023.

Rodney T O'Donnell, Ann E Nicholson, Bin Han, Kevin B Korb, Md Jahangir Alam, and Lucas R Hope. Causal discovery with prior information. In *AI 2006: Advances in Artificial Intelligence: 19th Australian Joint Conference on Artificial Intelligence, Hobart, Australia, December 4-8, 2006. Proceedings 19*, pages 1162–1167. Springer, 2006.

Judea Pearl. *Causality*. Cambridge university press, 2009.

Jonas Peters and Peter Bühlmann. Structural intervention distance for evaluating causal graphs. *Neural computation*, 27(3):771–799, 2015.

Paul Rolland, Volkan Cevher, Matthäus Kleindessner, Chris Russell, Dominik Janzing, Bernhard Schölkopf, and Francesco Locatello. Score matching enables causal discovery of nonlinear additive noise models. In *ICML*, 2022.

M. Scutari and J.B. Denis. *Bayesian Networks: With Examples in R*. Chapman & Hall/CRC Texts in Statistical Science. Taylor & Francis, 2014.

Shohei Shimizu, Patrik O Hoyer, Aapo Hyvärinen, Antti Kerminen, and Michael Jordan. A linear non-gaussian acyclic model for causal discovery. *JMLR*, 7(10), 2006.

Shohei Shimizu, Takanori Inazumi, Yasuhiro Sogawa, Aapo Hyvarinen, Yoshinobu Kawahara, Takashi Washio, Patrik O Hoyer, Kenneth Bollen, and Patrik Hoyer. Directlingam: A direct method for learning a linear non-gaussian structural equation model. *JMLR*, 12(Apr):1225–1248, 2011.

Peter Spirtes, Clark N Glymour, and Richard Scheines. *Causation, prediction, and search*. MIT press, 2000.

Masayuki Takayama, Tadahisa Okuda, Thong Pham, Tatsuyoshi Ikenoue, Shingo Fukuma, Shohei Shimizu, and Akiyoshi Sannai. Integrating large language models in causal discovery: A statistical causal approach. *arXiv preprint arXiv:2402.01454*, 2024.

Takeshi Teshima and Masashi Sugiyama. Incorporating causal graphical prior knowledge into predictive modeling via simple data augmentation. In *UAI*, pages 86–96. PMLR, 2021.

Ruibo Tu, Kun Zhang, Bo Bertilson, Hedvig Kjellstrom, and Cheng Zhang. Neuropathic pain diagnosis simulator for causal discovery algorithm evaluation. *NeurIPS*, 32, 2019.

Chris Wallace, Kevin B Korb, and Honghua Dai. Causal discovery via mml. In *ICML*, volume 96, pages 516–524, 1996.

Moritz Willig, Matej Zečević, Devendra Singh Dhami, and Kristian Kersting. Probing for correlations of causal facts: Large language models and causality. 2022.

Yuzhe Zhang, Yipeng Zhang, Yidong Gan, Lina Yao, and Chen Wang. Causal graph discovery with retrieval-augmented generation based large language models. *arXiv preprint arXiv:2402.15301*, 2024.

Xun Zheng, Bryon Aragam, Pradeep K Ravikumar, and Eric P Xing. Dags with no tears: Continuous optimization for structure learning. *NeurIPS*, 31, 2018.

Appendix

In this appendix, we include the following additional information, which we could not include in the main paper due to space constraints:

- Appendix A: Related Work
- Appendix B: Illustration of Triplet Query Strategy
- Appendix C: Using Causal Order with Different Classes of Discovery Methods
- Appendix D: Relation of Causal Order with Backdoor Adjustment Set
- Appendix E: Proofs of propositions
- Appendix F: Algorithms to integrate causal order into existing discovery methods
- Appendix G: Additional results, including LLMs used in post-processing for graph discovery and a discussion of triplet vs pairwise query strategies
- Appendix H: More details and examples of our query strategies
- Appendix I: Causal graphs used in our experiments of the datasets

A Related Work

Domain Expertise-aided Causal Discovery. Prior knowledge has been used in causal discovery literature [Hasan and Gani, 2022, Constantinou et al., 2023, Heckerman and Geiger, 2013, Teshima and Sugiyama, 2021, O'Donnell et al., 2006, Wallace et al., 1996]. These methods rely on domain expert opinions and documented knowledge from randomized controlled trials (RCT). Various priors have been studied in literature, such as *edge existence*, *forbidden edge* Meek [1995], *ancestral constraints* [Constantinou et al., 2023, Ban et al., 2023b]. Recent advancements in LLMs have led to more attention on how LLMs may act as imperfect experts and provide causal knowledge based on metadata such as variable names [Kıcıman et al., 2023, Ban et al., 2023b, Long et al., 2023, Willig et al., 2022]. Early methods [Kıcıman et al., 2023, Willig et al., 2022, Long et al., 2022] rely on LLMs to predict the complete causal structure, which is evaluated using metrics for full graph structure such as Structural Hamming Distance (SHD). Recent methods however use LLM's output to improve accuracy of graph discovery algorithms. The key idea is that LLM can provide information about edges in the graph, which can then be added as a prior or constraint Long et al. [2023], Jiralerspong et al. [2024] to reduce the search space for a causal discovery algorithm. For example, [Long et al., 2023] use LLMs to improve output of a constraint-based algorithm for full graph discovery by orienting undirected edges in the CPDAG. Most of these works, however, depend on obtaining correct edge information from LLMs and evaluate LLMs' quality by full graph metrics Naik et al. [2023], Zhang et al. [2024] such as SHD Kıcıman et al. [2023], Long et al. [2023]. Instead, we argue that LLMs (or even humans) are incapable of providing edge information given a pair (or subset) of variables. Hence, *causal order* may be a more appropriate causal structure to elicit from experts. For the same reason, the quality of an imperfect expert's output for such tasks are better evaluated on the accuracy of causal order, rather than the full graph structure.

LLM-based Prompting Strategies. Existing LLM-based algorithms for graph discovery [Kıcıman et al., 2023, Long et al., 2022, Ban et al., 2023b, Antonucci et al., 2023] use a pairwise prompt, essentially asking "does A cause B?" with varying levels of prompt complexity. Going beyond this line of work, we propose a triplet-based prompt that provides more accurate answers through aggregation and provides an uncertainty score for each edge to aid in cycle removal. As a result, our triplet-based prompt may be of independent interest for causal tasks.

B Illustration of our Triplet Query Strategy

We present an intuitive illustration of our overall triplet querying framework to obtain causal order from imperfect experts in Fig A1 below.

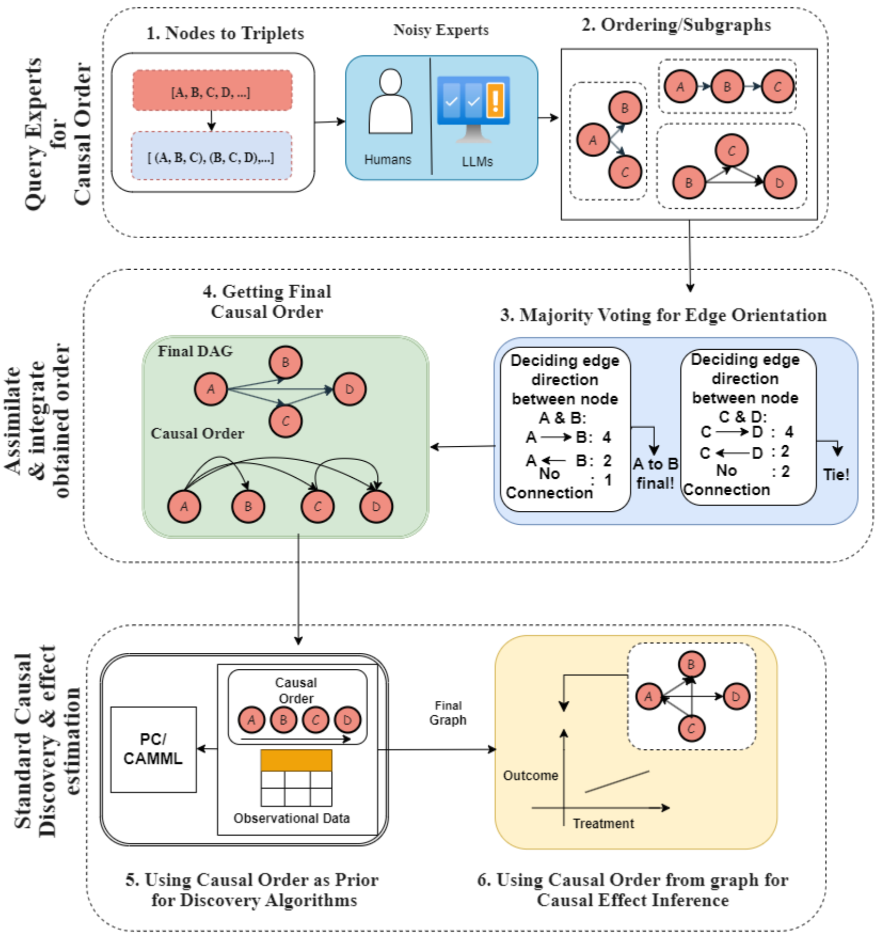


Figure A1: **Leveraging Causal Order from Imperfect Experts.** Our triplet-based querying method infers all three-variable subgraphs from imperfect experts and aggregates them (using majority voting) to produce a causal order. Ties in causal order are broken using a high-cost expert. Expert-generated causal order is integrated with discovery algorithms, before estimating causal effect.

C Using Causal Order with Different Classes of Discovery Methods

Using causal order with Constraint-Based Discovery Methods. Constraint-based causal discovery algorithms usually return a Completed Partially Directed Acyclic Graph (CPDAG), from which a Markov equivalence class of graphs can be obtained. However, not all edges in a CPDAG are oriented. Given a CPDAG from a constraint-based algorithm like PC [Spirtes et al., 2000], we use the causal order $\hat{\pi}$ obtained from experts to orient the undirected edges, similar to the algorithm from Meek [1995]. Iterating over undirected edges, we first check if the nodes of that edge occur in $\hat{\pi}$. If yes, we orient the edge according to $\hat{\pi}$. Since it is possible that the causal order obtained from querying experts may not include some nodes (isolated nodes), if either (or both) nodes of the undirected edge are not in $\hat{\pi}$, we query a superior expert (e.g. oracle) (see Sec 4) to finalize a direction between the pair. Algorithm 1 outlines the specific steps for this integration.

D Relation of Causal Order with Backdoor Adjustment Set

Correct topological order is necessary and sufficient for finding a valid backdoor set. Average causal effect [Pearl, 2009] (ACE) of a variable X_i on a variable X_j is defined as: $ACE_{X_i}^{X_j} = \mathbb{E}[X_j|do(X_i = x_i)] - \mathbb{E}[X_j|do(X_i = x_i^*)]$, where X_i is called the *treatment*, X_j is called the *target*, and $do(X_i = x_i)$ denotes an external intervention to the variable X_i with the value x_i. If a set of variables $\mathbf{Z}$ satisfies the backdoor criterion (Defn. E.1) relative to (X_i, X_j), $\mathbb{E}[X_j|do(X_i = x_i)]$ can be computed as: $\mathbb{E}[X_j|do(X_i = x_i)] = \mathbb{E}_{\mathbf{z}\sim\mathbf{Z}}\mathbb{E}[X_j|X_i = x_i, \mathbf{Z} = \mathbf{z}]$ (Thm. 3.3.2 of [Pearl, 2009]). To ensure that all variables in $\mathbf{Z}$ are observed, we assume there are no unobserved variables in the underlying causal graph. We now first show that a correct causal order is sufficient for identifying a backdoor set.

Proposition D.1. *[Pearl, 2009, Cinelli et al., 2022] Under the no latent confounding assumption, for a pair of treatment and target variables (X_i, X_j) in a DAG $\mathcal{G}$, $\mathbf{Z} = \{X_k|\pi_k < \pi_i\}$ is a valid adjustment set relative to (X_i, X_j) for any topological order π of $\mathcal{G}$.*

Using causal order with score-based discovery methods. Score-based methods like CaMML Wallace et al. [1996] allow the specification of prior constraints which are respected while obtaining the complete graph. We hence utilize the causal order $\hat{\pi}$ obtained from experts as a level order (Defn 3.1) prior to such methods. We handle any cycles in the expert's output by assigning all nodes in a cycle to the same level. Unlike a similar LLM-prior approach by Ban et al. [2023b], where the output of LLM and a score-based method are combined using an ancestral constraint as a prior, ours is a sequential approach where the score-based algorithm starts with the provided order-based constraint. This approach also allows us to provide a prior probability to control the influence of prior on the discovery method. Algorithm 2 outlines the specific steps for this integration.

E Proofs of Propositions

To estimate $\mathbb{E}[X_j|do(X_i = x_i)]$ from observational data, the *backdoor adjustment* formula is used.

Definition E.1. ***Backdoor Adjustment*** *[Pearl, 2009]. Given a DAG $\mathcal{G}$, a set of variables $\mathbf{Z}$ satisfies the backdoor criterion relative to a pair of treatment and target variables (X_i, X_j) if (i) no variable in $\mathbf{Z}$ is a descendant of X_i; and (ii) $\mathbf{Z}$ blocks every path between X_i and X_j that contains an arrow into X_i.*

Proposition D.1. *[Pearl, 2009, Cinelli et al., 2022] Under the no latent confounding assumption, for a pair of treatment and target variables (X_i, X_j) in a DAG $\mathcal{G}$, $\mathbf{Z} = \{X_k|\pi_k < \pi_i\}$ is a valid adjustment set relative to (X_i, X_j) for any topological order π of $\mathcal{G}$.*

Proof. Before starting the proof, we define a confounding variable. A confounder is a variable that should be casually associated with both the treatment and the target variables and is not on the causal pathway between treatment and target. An unmeasured common cause can also be a source of confounding the treatment $\rightarrow$ target relationship. Coming to the proof, we need to show that the set $\mathbf{Z} = \{X_k|\pi_k < \pi_i\}$ satisfies the conditions (i) and (ii) in Defn E.1. For any variable X_k such that $\pi_k < \pi_i$, we have $X_k \notin de(X_i)$ and hence the condition (i) is satisfied. Additionally, for each $X_k \in pa(X_i)$ we have $\pi_k < \pi_i$ and hence $pa(X_i) \subseteq \mathbf{Z}$. Since $pa(X_i)$ blocks all paths from X_i to X_j that contains an arrow into X_i [Peters and Bühlmann, 2015], $\mathbf{Z}$ satisfies condition (ii). □

Proposition E.1. *Given two nodes A and B of an underlying causal graph, access to an ϵ-expert $\mathcal{E}_\epsilon$ that doesn't produce any cycles in the predicted causal graph (see Assm E.1 for formal statement) and Assm E.2, let $C \neq A \neq B$ be any other node in the graph. If $\mathcal{E}_\epsilon$ predicts causal relationship between all pairs of nodes sequentially, the marginalized probability that $\mathcal{E}_\epsilon$ makes an error in predicting the causal relationship between A and B, after it has already predicted the causal relationships between (C, A) and (C, B), is less than ϵ, where marginalization is over all possible causal graphs that can be formed between A, B and C, with each of such graphs being equally likely.*

Definition E.2 (ϵ-Experts). *Given two nodes A and B of a graph and three options of the causal relationship between them: (i) $A \rightarrow B$, (ii) $A \leftarrow B$, and (iii) no edge between A and B (denoted as $[c_1, c_2, c_3]$), an expert $\mathcal{E}$ queried for the causal relationship between A and B is said to be an ϵ-expert (denoted as $\mathcal{E}_\epsilon$) if the probability of making an error in the prediction of the causal relationship between A and B is ϵ, where $\epsilon \in (0, 1)$.*

Proposition E.2. *Given two nodes A and B of an underlying causal graph, access to an ϵ-expert $\mathcal{E}_\epsilon$ that doesn't produce any cycles in the predicted causal graph (see Assm E.1 for formal statement) and Assm E.2, let $C \neq A \neq B$ be any other node in the graph. If $\mathcal{E}_\epsilon$ predicts causal relationship between all pairs of nodes sequentially, the marginalized probability that $\mathcal{E}_\epsilon$ makes an error in predicting the causal relationship between A and B, after it has already predicted the causal relationships between (C, A) and (C, B), is less than ϵ, where marginalization is over all possible causal graphs that can be formed between A, B and C, with each of such graphs being equally likely.*

Proposition 3.2. *For an estimated topological order $\hat{\pi}$ and a true topological order π of a causal DAG $\mathcal{G}$ with the corresponding adjacency matrix A, $D_{top}(\hat{\pi}, A) = 0$ iff $\mathbf{Z} = \{X_k | \hat{\pi}_k < \hat{\pi}_i\}$ is a valid adjustment set relative to (X_i, X_j), $\forall \pi_i < \pi_j$.*

Proof. The statement of proposition is of the form $A \iff B$ with A being "$D_{top}(\hat{\pi}, A) = 0$" and B being "$\mathbf{Z} = \{X_k | \hat{\pi}_k < \hat{\pi}_i\}$ is a valid adjustment set relative to (X_i, X_j), $\forall i, j$". We prove $A \iff B$ by proving (i) $A \implies B$ and (ii) $B \implies A$.

(i) Proof of $A \implies B$: If $D_{top}(\hat{\pi}, A) = 0$, for all pairs of nodes (X_i, X_j), we have $\hat{\pi}_i < \hat{\pi}_j$ whenever $\pi_i < \pi_j$. That is, causal order in estimated graph is same that of the causal order in true graph. Hence, from Propn D.1, $\mathbf{Z} = \{X_k | \hat{\pi}_k < \hat{\pi}_i\}$ is a valid adjustment set relative to (X_i, X_j), $\forall i, j$.

(ii) Proof of $B \implies A$: we prove the logical equivalent form of $B \implies A$ i.e., $\neg A \implies \neg B$, the *contrapositive* of $B \implies A$. To this end, assume $D_{top}(\hat{\pi}, A) \neq 0$, then there will be at least one edge $X_i \to X_j$ that cannot be oriented correctly due to the estimated topological order $\hat{\pi}$. i.e., $\hat{\pi}_j < \hat{\pi}_i$ but $\pi_j > \pi_i$. Hence, to find the causal effect of X_i on X_l; $l \neq j$, X_j is included in the back-door adjustment set $\mathbf{Z}$ relative to (X_i, X_l). Adding X_j to $\mathbf{Z}$ renders $\mathbf{Z}$ an invalid adjustment set because it violates the condition (i) of Defn E.1. □

Proposition 2.1. *Consider a Perfect Expert that infers causal relationships based on its (optimal) domain knowledge. It is given only a subset of observed variables $\mathbf{V}$ to predict the causal structure. Assuming that the Perfect Expert assumes causal sufficiency (the variables not given do not exist in the system), the predicted causal structure by the Expert might be wrong whereas the predicted causal order remains correct.*

Proof. Let $\boldsymbol{V} = \{V_1, \dots, V_n\}$ be the set of observed nodes and let $\boldsymbol{U} = \{U_1, \dots, U_m\}$ be the set of unobserved variables. Let V_1 and V_2 be two observed variables and U_1 be an unobserved node (for the expert). Also, let the underlying true causal graph have edges $V_1 \to U_1 \to V_2$ but no direct edges between V_1 and V_2. Thus, node V_1 comes before V_2 in the true causal order. Now if we ask any (optimal) expert that assumes causal sufficiency (i.e there is no unobserved variable) to predict the causal relationship between them, it will predict there is an edge between V_1 and V_2 since V_1 has an indirect causal effect on V_2 in the underlying true causal graph, thus making an error in causal structure prediction. However, the predicted causal order between node V_1 and V_2 (i.e. V_1 comes before V_2) is still unaffected by this unobserved variable U_1. The same argument can be extended to other subsets of nodes in the graph with more complicated sub-structures thus completing the proof. □

Corollary E.1. *If an (optimal) algorithm is given only a subset of observed variables to predict the causal structure, the predicted causal structure might be wrong whereas the predicted causal order remains correct.*

Proof. The proof of this corollary follows the proof of Prop 2.1 (see Proof E). Let $\boldsymbol{V} = \{V_1, \dots, V_n\}$ be the set of nodes present in the graph and let $\boldsymbol{U} = \{U_1, \dots, U_m\}$ be the set of unobserved variables. Let $\boldsymbol{V}_S$ be the subset of the variable given to the algorithm and $\boldsymbol{V}_R = \boldsymbol{V} \setminus \boldsymbol{V}_S$ be the rest of the observed variable. Now define $\boldsymbol{V}_S$ to be the observed variable for this proof and $\boldsymbol{V}_R \cup \boldsymbol{U}$ to be the unobserved variable for this proof and rerun the same arguments the proof of Prop 2.1 (see Proof E). □

Assumption E.1 (DAG Acyclicity)**.** *Given that ϵ-expert $\mathcal{E}_\epsilon$ is used to predict a causal graph between a set of nodes, the predicted causal graph is acyclic.*

Remark E.2. *For ease of exposition, we define the ϵ-expert to have error probability exactly equal to ϵ; this could however be generalized to have error probability at most ϵ.*

Assumption E.2 (Error Distribution and Probability Renormalization)**.** *Let $[c_1, c_2, c_3]$ be the three choices for a causal relationship between node A and B (see Def E.2). Let $P(c_1)$, $P(c_2)$ and $P(c_3)$ be the probability of selecting the corresponding three choices by the ϵ-expert $\mathcal{E}_\epsilon$. We assume that the probability for the two wrong options are equally likely i.e. equal to $\epsilon/2$. If any constraint $\mathcal{T}$ renders some of the choices as not possible i.e. $P(c_j|\mathcal{T}) = 0$ for some $j \in \{1, 2, 3\}$, then $\mathcal{E}_\epsilon$ renormalizes the posterior probability over the other choices i.e $P(c_i|\mathcal{T}) = \frac{P(c_i)}{\sum_{j, P(c_j|\mathcal{T})\neq 0} P(c_j)}$ where the denominator is summed over j s.t. $P(c_j|\mathcal{T}) \neq 0$.*

Proposition E.2. *Given two nodes A and B of an underlying causal graph, access to an ϵ-expert $\mathcal{E}_\epsilon$ that doesn't produce any cycles in the predicted causal graph (see Assm E.1 for formal statement) and Assm E.2, let $C \neq A \neq B$ be any other node in the graph. If $\mathcal{E}_\epsilon$ predicts causal relationship between all pairs of nodes sequentially, the marginalized probability that $\mathcal{E}_\epsilon$ makes an error in predicting the causal relationship between A and B, after it has already predicted the causal relationships between (C, A) and (C, B), is less than ϵ, where marginalization is over all possible causal graphs that can be formed between A, B and C, with each of such graphs being equally likely.*

Proof. Without any additional constraint, ϵ-expert ($\mathcal{E}_\epsilon$) has "ϵ" probability of making incorrect prediction. But in presence of additional constraint, e.g. DAG constraint (see Assm E.1), the probability of error changes and is given by the following lemma:

Lemma E.3. *Suppose we have two nodes A and B and three possible choices $[c_1, c_2, c_3]$ for causal relationship between them i.e $A \to B$, $B \to A$ or no edge between them (not in any particular order). Without loss of generality, let c_3 be the ground truth causal relationship between node A and B. Thus, without any additional constraint, let the probability assigned to each of the three choices by ϵ-expert ($\mathcal{E}_\epsilon$) is $P(c_1) = \epsilon_1, P(c_2) = \epsilon_2$ and $P(c_3) = 1 - \epsilon_1 - \epsilon_2$ respectively where $\epsilon = \epsilon_1 + \epsilon_2$. If due to additional constraint (e.g. acyclicity Assm E.1), one of the incorrect choice gets discarded, say c_1, then the new probability of selecting the wrong choice (c_2 given by ϵ') is always less than ϵ. However if the correct/ground truth choice is discarded due to this additional constraint the new probability of selecting the wrong choice (c_1 or c_2) is 1. In case, no options are discarded the new probability of choosing the wrong choice remains same i.e ϵ as before.*

Proof. For the case when the correct/ground truth choice i.e c_3 is discarded due to some constraint, the only left out choices are wrong choices i.e. c_1 and $c2$. Thus the probability of making error in selecting the correct choice is 1. Next, for the case when one of the incorrect choice (here c_1 w.l.o.g) is discarded, we are left with one incorrect (c_2) and one correct choice (c_3). From Assm E.2 once a particular option is discarded, the ϵ-expert renormalizes the probability proportional to their initial probability. Thus the new probability ($\tilde{P}(c_2)$) of choosing wrong option c_2 is:

$$\tilde{P}(c_2) = \frac{\epsilon_2}{1 - \epsilon_1 - \epsilon_2 + \epsilon_2} = \frac{\epsilon_2}{1 - \epsilon_1} = \frac{\epsilon/2}{1 - \epsilon/2} = \frac{\epsilon}{2 - \epsilon} \tag{1}$$

where $\epsilon_1 = \epsilon_2 = \epsilon/2$ from Assm E.2. Next, we can show that $\tilde{P}(c_2) < \epsilon$ completing our proof. To have $\tilde{P}(c_2) < \epsilon$ we need:

$$\begin{aligned} \tilde{P}(c_2) =& \frac{\epsilon_2}{1 - \epsilon_1} < \epsilon = \epsilon_1 + \epsilon_2 \\ \implies& \epsilon_2 < \epsilon_1 + \epsilon_2 - \epsilon_1^2 - \epsilon_1\epsilon_2 \\ \implies& \epsilon_1(\epsilon_1 + \epsilon_2 - 1) < 0 \end{aligned} \tag{2}$$

which is always true since from Assm E.2 we have $\epsilon_1 > 0$, $\epsilon_2 > 0$ and $1 - \epsilon_1 - \epsilon_2 > 0$. □

Now, give any three nodes A, B and C, Table A1 summarizes all possible *partially completed* graph (henceforth partial graph) possible between those nodes. Each partially-completed DAG in Table A1 generated more DAG based on the orientation of the node A and B. Specifically, each of the partial graph $1, 2, 3, 4, 5, 7$ and 9 generated three graphs ($A \to B$, $B \to A$ or no edge between A and B) and partial graph 6 and 8 will give two DAG (one option is not possible to maintain acyclicity constraint). Thus overall we have 25 possible graphs. Our next goal is to show that the marginal probability of choosing the wrong causal relationship for node (A, B) when oriented last among is less than ϵ, where marginalization is over all the causal graph depicted in Table A1 (assuming all graphs are equally likely). The expert $\mathcal{E}_\epsilon$ finds the causal relationship sequentially for all the pairs in $\{(C,A),(C,B),(A,B)\}$. We are interested in the case when $\mathcal{E}_\epsilon$ finds the causal relationship for pair (A,B) in the end. Let F,S,T (called first,second and third) be three binary random variable and the value 0 represent whether the causal relationship discovered by $\mathcal{E}_\epsilon$ for first, second or last/third pair respectively is incorrect and 1 represent it is correct. So the probability of error when finding the causal relationship between node A and B when oriented last/third (denoted by $P(T)$) is given by:

$$\begin{aligned} P(T=0) &= \sum_{G\in\mathcal{G}} \sum_{S,T\in\{0,1\}\times\{0,1\}} P(G)P(F,S|G)P(T=0|F,S,G) \\ &= \frac{1}{25}\cdot \sum_{G\in\mathcal{G}} \sum_{S,T\in\{0,1\}\times\{0,1\}} P(F,S|G)P(T=0|F,S,G) \end{aligned} \tag{3}$$

where $\mathcal{G}$ denotes the set of graphs generated by orienting the causal relationship between A and B for all *partial* graphs in Table A1, $|\mathcal{G}|=25$ and all the graphs are equally likely, different configuration of (F,S) shows whether the causal relationship between first two pairs (C,A) and (C,B) are correct or not. When orienting the first two pair of nodes i.e (C,A) and (C,B) there is no DAG constraint thus we have:

$$P(F,S) = \begin{cases} \epsilon^2 & \text{when } \; S=0,T=0 \\ \epsilon(1-\epsilon) & \text{when } \; S=0,T=1 \\ \epsilon(1-\epsilon) & \text{when } \; S=1,T=0 \\ (1-\epsilon)^2 & \text{when } \; S=1,T=1 \end{cases} \tag{4}$$

Now based on the graph $G\in\mathcal{G}$ and the setting of S,T, $P(T=0|F,S,G)$ takes different values. Suppose that the causal relationship between the first two pairs (C,A) and (C,B) are already predicted by the expert. We observe that the DAG acyclcity constraint (Assm E.2) will only change the probability of error for orienting nodes (A,B) ($P(T=0|F,S,G)$ given by Lemma E.3) when the predicted causal graphs is either $B\rightarrow C\rightarrow A$ or $A\rightarrow C\rightarrow B$ after orienting (C,A) and (C,B). For all the other predictions of (C,A) and (C,B), they don't enforce any acyclicity constant for finding the causal relationship between (A,B), thus, $P(T=0|F,S,G)=\epsilon$ (from Lemma E.3). Table A2 summarizes of error probability for all the partial graphs in Table A1 ($P(F,S|G)$ and $P(T|F,S,G)$). The first column shows different partial graphs from Table A1. The second column then shows different causal relationships that are possible between the nodes A and B for a particular partial graph. Given one *true orientation* between node A and B we get a final ground truth graph. Thus the third column shows the probability of prediction of structure $A\leftarrow C\rightarrow B$ for a particular true graph and the fourth column shows the probability of making an error in predicting the third causal relationship i.e between (A,B) given the first and second pair (C,A) and $((C,B))$ is already predicted. Similarly, the fifth and sixth columns show the same thing for the predicted structure $A\rightarrow C\rightarrow B$ for each of the ground truth graphs. The partial-graph number $4,7,8$ is not depicted in the table but the entries for 4th graph is the same as 2nd, 7th is the same as 3rd and 8th is same as 6th due to symmetry in the partial-structure. The value of $\epsilon^{'}=\frac{\epsilon}{2-\epsilon}$ in 4th and 6th column is given by renormalized probability Eq. 1 in Lemma E.3. Substituting the values from Table A2 in Eq. 3 and using the value $P(T|F,S,G)=\epsilon$ for the rest of the predicted structure not mentioned in the Table A2 we get:

$$\begin{aligned} P(T=0) = \frac{1}{25}\cdot \Bigg\{ & 2*\frac{\epsilon^2}{4}\Big[2\epsilon^{'}+1\Big] + \Big[1-2*\frac{\epsilon^2}{4}\Big]\epsilon \\ & + \Bigg(\Big[\frac{\epsilon^2}{4}+\frac{\epsilon(1-\epsilon)}{2}\Big]\Big[2\epsilon^{'}+1\Big] + \Big[1-\frac{\epsilon^2}{4}-\frac{\epsilon(1-\epsilon)}{2}\Big]\epsilon\Bigg)*4 \\ & + \Bigg(2*\frac{\epsilon(1-\epsilon)}{2}\Big[2\epsilon^{'}+1\Big] + \Big[1-2*\frac{\epsilon(1-\epsilon)}{2}\Big]\epsilon\Bigg)*2 \\ & + \Bigg((1-\epsilon)^2\Big[2\epsilon^{'}\Big] + \frac{\epsilon^2}{4}\Big[\epsilon^{'}+1\Big] + \Big[1-(1-\epsilon)^2-\frac{\epsilon^2}{4}\Big]\epsilon\Bigg)*2\Bigg\} \\ = \frac{1}{25}\cdot &\Bigg\{\frac{\epsilon(3\epsilon^2-30\epsilon+52)}{4-2\epsilon}\Bigg\} \end{aligned} \tag{5}$$

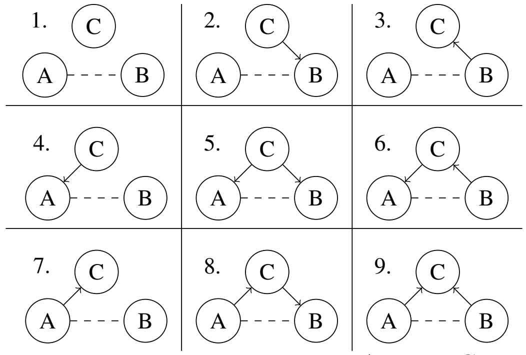


Table A1: All possible causal graph between three variables A, B and C. The dashed arrow represented undecided causal relationship between node A and B. So, the dashed arrow can take one of three choices $A \rightarrow B$, $A \leftarrow B$ or no edge between A and B. To ensure that the graph is acyclic, some of the graphs above might not allow all three choice for causal relationship between node A and B. Hence the causal-graph $1, 2, 3, 4$ and 7 each have three possible graphs and $5, 6, 8$ and 9 each have two possible graphs based on the valid choice of causal relationship between A and B that preserves acyclicity constraint. So overall there are 25 possible different causal graph between three variables A, B and C.

Now we want to show that the error probability for the third pair (A, B) given by the above equation is less than ϵ. For that, we need:

$$\begin{gathered} \frac{1}{25} \cdot \left\{ \frac{\epsilon(3\epsilon^2 - 30\epsilon + 52)}{4 - 2\epsilon} \right\} < \epsilon \\ 3\epsilon^2 + 20\epsilon - 48 < 0 \end{gathered} \tag{6}$$

The above inequality is always satisfied since $\epsilon \in (0, 1)$ and $3\epsilon^2 + 20\epsilon - 48$ is always less than 0 in the allowed range of ϵ since the roots of the quadratic equation are $-10/3 - 2\sqrt{61}/3 = -8.5$ and $-10/3 + 2\sqrt{61}/3 = 1.87$. Thus $P(T = 0) < \epsilon$ for all values of $\epsilon \in (0, 1)$ completing our proof.

$\square$

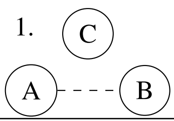
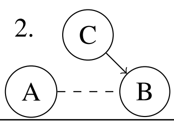
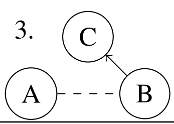
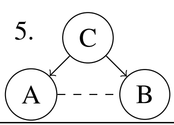
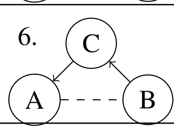
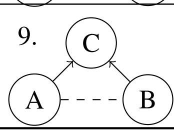
Partial True Graph	True Orientation (A, B)	Predicted Orientation in first two steps (F, S)			
		$A \leftarrow C \leftarrow B$		$A \rightarrow C \rightarrow B$	
		$P(F, S \mid G)$	$P(T \mid F, S, G)$	$P(F, S \mid G)$	$P(T \mid F, S, G)$
1. 	no edge $A \rightarrow B$ $A \leftarrow B$	$\left(\frac{\epsilon}{2}\right)^2$	$\epsilon^{'}$ 1 $\epsilon^{'}$	$\left(\frac{\epsilon}{2}\right)^2$	$\epsilon^{'}$ $\epsilon^{'}$ 1
2. 	no edge $A \rightarrow B$ $A \leftarrow B$	$\left(\frac{\epsilon}{2}\right)^2$	$\epsilon^{'}$ 1 $\epsilon^{'}$	$\left(\frac{\epsilon}{2}\right)(1-\epsilon)$	$\epsilon^{'}$ $\epsilon^{'}$ 1
3. 	no edge $A \rightarrow B$ $A \leftarrow B$	$\left(\frac{\epsilon}{2}\right)(1-\epsilon)$	$\epsilon^{'}$ 1 $\epsilon^{'}$	$\left(\frac{\epsilon}{2}\right)^2$	$\epsilon^{'}$ $\epsilon^{'}$ 1
5. 	no edge $A \rightarrow B$ $A \leftarrow B$	$\left(\frac{\epsilon}{2}\right)(1-\epsilon)$	$\epsilon^{'}$ 1 $\epsilon^{'}$	$\left(\frac{\epsilon}{2}\right)(1-\epsilon)$	$\epsilon^{'}$ $\epsilon^{'}$ 1
6. 	no edge $A \leftarrow B$	$(1-\epsilon)^2$	$\epsilon^{'}$ $\epsilon^{'}$	$\left(\frac{\epsilon}{2}\right)^2$	$\epsilon^{'}$ 1
9. 	no edge $A \rightarrow B$ $A \leftarrow B$	$\left(\frac{\epsilon}{2}\right)(1-\epsilon)$	$\epsilon^{'}$ 1 $\epsilon^{'}$	$\left(\frac{\epsilon}{2}\right)(1-\epsilon)$	$\epsilon^{'}$ $\epsilon^{'}$ 1

Table A2: Summary of Error Probability for all the partial graphs in Table A1 ($P(F, S|G)$ and $P(T|F, S, G)$): The first column shows different partial graphs from Table A1. The second column then shows different causal relationships that are possible between the nodes A and B for a particular partial graph. Given one *true orientation* between node A and B we get a final ground truth graph. Now we observed in the proof of Proposition E.2 (see Proof E), that the error probability for the prediction of causal relationship for the pair (A, B) will only change when the ϵ-expert predicts the the structure $A \leftarrow C \rightarrow B$ or $A \rightarrow C \rightarrow B$ for the pair of nodes (C, A) and (C, B) for any ground truth graph. For the rest of the possible predictions of a pair of nodes (C, A) and (C, B) in any ground truth graph, the error probability for (A, B) remains ϵ (see Lemma E.3). Thus the third column shows the probability of prediction of structure $A \leftarrow C \rightarrow B$ for a particular true graph and the fourth column shows the probability of making an error in predicting the third causal relationship i.e between (A, B) given the first and second pair (C, A) and $((C, B))$ is already predicted. Similarly, the fifth and sixth columns show the same thing for the predicted structure $A \rightarrow C \rightarrow B$ for each of the ground truth graphs. The partial-graph number $4, 7, 8$ is not depicted in the table but the entries for 4th graph is the same as 2nd, 7th is the same as 3rd and 8th is same as 6th due to symmetry in the partial-structure. The value of $\epsilon^{'} = \frac{\epsilon}{2-\epsilon}$ in 4th and 6th column is given by renormalized probability Eq. 1 in Lemma E.3.

Algorithm 1 Integrating $\hat{\pi}$ in constraint-based methods

Input: Noisy expert topological ordering $\hat{\pi}$, Expert $\mathcal{E}$, CPDAG $\hat{\mathcal{G}}$
Output: Estimated topological order $\hat{\pi}_{\text{final}}$ of $\{X_1, \dots, X_n\}$.
for $(i - j) \in$ undirected-edges($\hat{\mathcal{G}}$) **do**
    If both nodes i and j are in $\hat{\pi}$ and if $\hat{\pi}_i < \hat{\pi}_j$, orient $(i - j)$ as $(i \rightarrow j)$ in $\hat{\mathcal{G}}$.
    Otherwise, use expert $\mathcal{E}$ to orient the edge.
end for
$\hat{\pi}_{\text{final}} =$ topological ordering of $\hat{\mathcal{G}}$
return $\hat{\pi}_{\text{final}}$

Algorithm 2 Integrating $\hat{\pi}$ in score-based methods

Input: Dataset $\mathcal{D}$, Variables $\{X_1, \dots, X_n\}$, Expert $\mathcal{E}$, Score-based method $\mathcal{S}$, *Prior* probability p.
Output: Estimated topological order $\hat{\pi}_{\text{final}}$ of $\{X_1, \dots, X_n\}$.
$\hat{\mathcal{G}} = \mathcal{E}(X_1, \dots, X_n)$
L = level order of $\hat{\mathcal{G}}$
for cycle $C \in \hat{\mathcal{G}}$ **do**
    for node $\in C$ **do**
        L(node) $=$ min(level(c) $\forall c \in C$)
    end for
end for
$\hat{\mathcal{G}} = \mathcal{S}(\mathcal{D}, L, p)$ `//L is provided as prior`
$\hat{\pi}_{\text{final}} =$ topological ordering of $\hat{\mathcal{G}}$
return $\hat{\pi}_{\text{final}}$

F Algorithms for Integrating Causal Order in Existing Discovery Methods

In continuation to the discussion in Sec 3, the algorithms for integrating causal order into existing constraint-based and score-based discovery methods are summarized in Algorithms 1 and 2 respectively.

G Additional Results

G.1 D_{top} vs SHD: Better Measure of Effect Estimation Error

As discussed in Sec 3 of the main paper, we show herein that D_{top} has a strong correlation with effect estimation error and hence is a valid metric for effect inference.

Specifically, we study how the error in causal effect, ϵ_{ACE}, changes as values of the metric D_{top} change. Fig A2 shows results on Asia and Child datasets. When D_{top} is zero, ϵ_{ACE} is also zero; and ϵ_{ACE} of a feature on other features increases as D_{top} increases. We use the DoWhy library for evaluating causal effects given a set of backdoor set of variables (we take number of samples $N = 10000$). Figure A2 shows a comparison of various methods w.r.t ϵ_{ACE} vs D_{top}. Results show that ϵ_{ACE} increases as D_{top} increases, aligning with theoretical observations. Table A3 shows that for the same SHD, increasing D_{top} leads to an increase in estimation error; but if D_{top} is the same, increasing SHD has no effect on effect estimation error.

Cancer			
$D_{top} = 0$		$SHD = 2$	
SHD	ϵ_{ACE}	D_{top}	ϵ_{ACE}
0	0.00	0	0.00
2	0.00	1	0.25
4	0.00	2	0.50
Asia			
$D_{top} = 0$		$SHD = 3$	
SHD	ϵ_{ACE}	D_{top}	ϵ_{ACE}
0	0.00	1	0.14
6	0.00	2	0.22
10	0.00	3	0.57
Survey			
$D_{top} = 0$		$SHD = 2$	
SHD	ϵ_{ACE}	D_{top}	ϵ_{ACE}
0	0.00	0	0.00
2	0.00	1	0.25
4	0.03	2	0.50

Table A3: ϵ_{ACE} vs SHD given D_{top} (& D_{top} given SHD)

G.2 LLMs used in post processing for graph discovery

We conducted some experiments where we utilised discovery algorithms like PC for creating skeletons of the graph and employed LLMs for orienting the undirected edges. The idea was to utilise LLMs ability to correctly estimate the causal direction while leveraging PC algorithm's ability to give a skeleton which could be oriented in a post processing setup. We saw that LLM ended up giving improved results as compared to PC alone.

G.3 Triplet vs Pairwise Query Strategies

In continuation to the discussion in Sec 5 of the main paper, we include Tables A4 for more details. The pairwise strategy also shows flaws when LLMs are used as noisy experts. In many cases, pairwise querying yields cycles due to which D_{top} cannot be computed. In particular, for the Child dataset

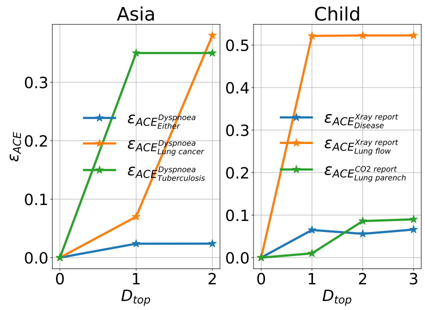


Figure A2: D_{top} vs. ϵ_{ACE}. ϵ_{ACE} increases as D_{top} increases, aligning with theoretical observations.

with 20 nodes, pairwise querying of LLMs yields an extremely high number of cycles (see Table A4). LLM output tends to overconnect, resulting in high SHD. Overall, among the prompting strategies, the chain of thought prompt performs the best: it has the lowest number of cycles for Child and Neuropathic datasets. This indicates that in-context examples and chain-of-thought reasoning help to increase the accuracy of causal order output, but other contextual cues do not matter.

Dataset	$\mathbf{D_{top}}$	**SHD**	**IN/TN**	**Cycles**
Base Prompt				
Earthquake	0	7	0/5	0
Cancer	0	6	0/5	0
Survey	3	12	0/6	0
Asia	-	21	0/8	1
Asia-M	-	15	0/7	7
Child	-	177	0/20	>>3k
Neuropathic	-	212	0/22	>>5k
All Directed Edges				
Earthquake	1	9	0/5	0
Cancer	1	7	0/5	0
Survey	2	11	0/6	0
Asia	-	21	0/8	6
Asia-M	0	13	0/7	0
Child	-	139	0/20	>>300
Neuropathic	-	194	0/22	>>1k
Markov Blanket				
Earthquake	0	8	0/5	0
Cancer	0	6	0/5	0
Survey	3	12	0/6	0
Asia	-	21	0/8	1
Asia-M	0	14	0/7	0
Child	-	167	0/20	>>400
Neuropathic	-	204	0/22	>>4k

Table A4: Comparison of various querying strategies for only LLM-based setups, providing different contextual cues in each setup about the graph. IN: Isolated Nodes, TN:Total Nodes.

Dataset	Metric	Pairwise (CoT)	Triplet
Using LLM			
Earthquake	D_{top}	**0**	**0**
	SHD	**4**	**4**
	Cycles	**0**	**0**
	IN/TN	**0/5**	**0/5**
Survey	D_{top}	1	**0**
	SHD	**9**	**9**
	Cycles	**0**	**0**
	IN/TN	2/6	**0/6**
Cancer	D_{top}	-	**1**
	SHD	-	**6**
	Cycles	-	**0**
	IN/TN	-	**0/5**
Asia	D_{top}	-	1
	SHD	18	14
	Cycles	1	0
	IN/TN	0/8	0/8
Asia-M	D_{top}	-	**1**
	SHD	13	**11**
	Cycles	1	**0**
	IN/TN	**0/7**	**0/7**
Child	D_{top}	-	**1**
	SHD	138	**28**
	Cycles	»500	**0**
	IN/TN	**0/20**	10/20
Using Human Annotators			
Earthquake	D_{top}	**0**	**0**
	SHD	4.67	**1.67**
	Cycles	**0**	**0**
	IN	**0**	0.33
Survey	D_{top}	-	**0**
	SHD	6.33	**3.67**
	Cycles	0.67	**0**
	IN	0.67	**0**
Cancer	D_{top}	**0**	**0**
	SHD	4.33	**3.67**
	Cycles	**0**	**0**
	IN	0.67	**0**
Asia-M	D_{top}	-	**1.33**
	SHD	11.67	**11.33**
	Cycles	3	**0**
	IN	**0**	**0**

Table A5: *(Top)* Results using LLM *(Bottom)* Results using human annotators (mean value across annotators reported). Performance of triplet method vs best performing pairwise query strategy (Chain of Thought) on multiple benchmark datasets across diff metrics: D_{top}, SHD, (Num of) Cycles, IN (Isolated Nodes), TN (Total Nodes). When num of cycles>0, $\hat{\pi}$ cannot be computed, hence D_{top} is given by '-'. Triplet consistently outperforms the pairwise strategy across metrics & datasets, esp by significant amounts on larger graphs like *Child*.

Dataset	$\mathbf{D_{top}}$	SHD	IN/TN	Cycles
Chain of Thought				
Earthquake	0	4	0/5	0
Survey	1	9	2/6	0
Asia	-	18	0/8	1
Asia-M	-	13	0/7	1
Child	-	138	0/20	$>>$500
Neuropathic	-	64	0/22	5
Triplet Query				
Earthquake	0	4	0/5	0
Cancer	1	6	0/5	0
Survey	0	9	0/6	0
Asia	1	14	0/8	0
Asia-M	1	11	0/7	0
Child	-	138	0/20	391
Child (+ Cycle Remover)	1	28	10/20	0
Neuropathic	-	151	0/22	772
Neuropathic(+ Cycle remover)	3	24	13/20	0

Table A6: Triplet query output using variable names with their descriptions (Cancer not included since CoT prompt has examples from this graph). IN: Isolated Nodes, TN:Total Nodes. Since calculating total number of cycles in a DAG is computationally challenging (NP Hard), we find a lower bound of cycles present in each graph based on total k lenght cycles in each setting, where k=5. If k is scaled up, the number of such unique cycles in the LLM output will also scale significantly. Lower bound helps us make a comparison with number of cycles in outputs like in Triplet strategy, where numbers are comparatively smaller and can be calculated easily.

Graph	Sample size	Before LLM prior	After LLM Prior
	250	18	16
	500	16	15
Child	1000	14	13
	5000	13.5	12
	10000	9.66	6
	250	3.83	3
	500	3.6	3
Earthquake	1000	3.6	3
	5000	1.16	0.66
	10000	0	0
	250	1	0
	500	3.83	3.83
Cancer	1000	2.6	2.6
	5000	2.3	2.3
	10000	2	2
	250	7.5	7
	500	6	5
Asia	1000	7	7
	5000	2	1
	10000	2	1
	250	4.5	4
	500	4	4
Asia-M	1000	5.5	5
	5000	4	4
	10000	4	4
	250	27	26
	500	31	29
Neuro	1000	41	40
	5000	55	53

Table A7: Comparison of SHD Values Before and After Incorporating LLM Priors Using the PC Algorithm Across Various Graphs

Finally, the triplet prompt provides the most accurate causal order. For small-scale graphs, it produces no cycles and consistently produces minimal D_{top} (ranging from 0 to 1) while also producing no isolated nodes. Even for medium-size graphs like Child and Neuropathic, the LLM output includes significantly fewer cycles than the pairwise strategy, which were removed leading to a significant and accurate causal order used further as prior. That said, we do see that isolated nodes in the output increase after cycles are removed for medium graphs (all graphs are connected, so outputting an isolated node is an error). Considering LLMs as virtual experts, this indicates that there are some nodes on which the LLM expert cannot determine the causal order. This is still a better tradeoff than providing the wrong causal order, which can confuse downstream algorithms. Overall, we conclude that the triplet query strategy provides the most robust causal order predictions. Additional results showing the error introduced by the LLM with respect to a ground truth order are shown in two different settings in Tables A11 and A12.

G.4 Neuropathic Results

Studying Memorization: Results on Neuropathic Dataset. The datasets we considered from BNLearn are popular ones. To study if our framework's performance is limited to popular datasets where there could have been memorization, we consider another dataset, a medium-sized subset graph from the Neuropathic dataset [Tu et al., 2019] used for pain diagnosis. This dataset is relatively less popular and harder to access on the web. The results shown in Tables A14 and A13 show similar trends as before – that obtaining causal order using triplet

Graph	Sample size	Before LLM prior	After LLM Prior
Child	250	18	16
	500	16	15
	1000	14	13
	5000	13.5	12
	10000	9.66	6
Earthquake	250	2	0
	500	1	0
	1000	1	0
	5000	1	0
	10000	1	0
Cancer	250	4	2
	500	3	4
	1000	3	2
	5000	2	0
	10000	3	0
Asia	250	4	4
	500	4	2
	1000	2	2
	5000	2	2
	10000	2	2
Asia-M	250	6	3
	500	2	1
	1000	2	2
	5000	2	1
	10000	2	1
Neuro	250	27	26
	500	31	29
	1000	41	40
	5000	55	53

Table A8: Comparison of SHD Values Before and After Incorporating LLM Priors Using the CamML Algorithm Across Various Graphs

1000 samples				
Context	Base prompt	Past iteration orientations	Markov Blanket	PC (Avg. over MEC)
D_{top}	8.0	5.3	6.6	9.61
SHD	14.33	12.66	14.0	17.0
10000 samples				
D_{top}	6.33	9.66	6.0	7.67
SHD	9.0	13.33	8.33	12.0

Table A9: PC + LLM results where LLM is used to orient the undirected edges of the skeleton PC returns over different data sample sizes. We show how LLMs can be used in a post processing setup for edge orientation besides having the capability of acting as a strong prior for different discovery algorithms.

method leads to low D_{top}, and then using this for causal discovery improves the performance of both PC and CaMML algorithms.

	PC	**SCORE**	**ICA LiNGAM**	**Direct LiNGAM**	**NOTEARS**	**CaMML**	**Ours (PC+LLM)**	**Ours (CaMML+LLM)**
$N = 250$	4.00±0.00	6.00±0.00	13.0±6.16	10.0±0.00	9.00±0.00	10.4±1.95	**3.00±0.00**	5.00±0.00
$N = 10000$	10.00±0.00	6.00±0.00	**1.00±0.00**	10.0±0.00	10.0±0.00	3.00±0.00	10.00±0.00	**1.00±0.00**

Table A13: Performance on causal discovery for the *Neuropathic* dataset subgraph (1k samples), showing mean and std dev of D_{top} over 3 runs.

	Dataset	PC	SCORE	ICA LiNGAM	Direct LiNGAM	NOTEARS	CaMML	Ours (PC+LLM)	Ours (CaMML+LLM)	Ours (PC+Human)	Ours (CaMML+Human)
$N = 250$	Earthquake	0.16±0.28	4.00±0.00	3.20±0.39	3.00±0.00	1.80±0.74	2.00±0.00	**0.00±0.00**	**0.00±0.00**	**0.00±0.00**	1.00±0.00
	Cancer	**0.00±0.00**	3.00±0.00	4.00±0.00	3.60±0.48	2.00±0.00	2.00±0.00	**0.00±0.00**	**0.00±0.00**	**0.00±0.00**	**0.00±0.00**
	Survey	0.50±0.00	3.00±0.00	6.00±0.00	6.00±0.00	3.20±0.39	3.33±0.94	**0.00±0.00**	3.33±0.94	**0.00±0.00**	**0.00±0.00**
	Asia	2.00±0.59	5.00±0.00	6.20±0.74	7.00±0.00	4.00±0.00	1.85±0.58	1.00±0.00	**0.97±0.62**	N/A	N/A
	Asia-M	1.50±0.00	5.00±0.00	7.60±0.48	6.20±1.16	3.40±0.48	**1.00±0.00**	**1.00±0.00**	1.71±0.45	**1.00±0.00**	2.00±0.00
	Child	5.75±0.00	8.80±2.70	12.8±0.97	13.0±0.63	15.0±1.09	**3.00±0.00**	4.00±0.00	3.53±0.45	N/A	N/A
	Neuropathic	4.00±0.00	6.00±0.00	13.0±6.16	10.0±0.00	9.00±0.00	10.4±1.95	**3.00±0.00**	5.00±0.00	N/A	N/A
$N = 500$	Earthquake	0.75±0.25	4.00±0.00	3.20±0.39	3.40±0.48	1.20±0.40	**0.00±0.00**	**0.00±0.00**	**0.00±0.00**	**0.00±0.00**	**0.00±0.00**
	Cancer	0.16±0.28	3.00±0.00	3.40±0.48	3.00±0.00	2.00±0.00	1.00±0.00	0.33±0.57	1.00±0.00	**0.00±0.00**	**0.00±0.00**
	Survey	1.25±0.00	4.00±0.00	6.00±0.0	6.00±0.00	3.40±0.48	3.39±0.08	1.00±0.00	3.33±0.94	1.00±0.00	**0.00±0.00**
	Asia	3.06±0.00	5.00±0.00	5.60±0.48	7.00±0.00	3.20±0.39	3.81±0.39	1.00±0.00	**0.97±0.62**	N/A	N/A
	Asia-M	2.00±0.00	6.00±0.00	7.60±0.48	5.00±0.00	3.80±0.39	2.00±0.00	1.00±0.00	**0.17±0.45**	1.33±0.57	3.00±0.00
	Child	8.09±0.00	6.20±1.32	12.2±0.74	10.6±1.35	15.4±0.48	**2.00±0.00**	5.00±1.73	**2.00±0.00**	N/A	N/A
	Neuropathic	7.50±0.00	6.00±0.00	9.00±1.41	13.0±0.00	11.0±0.00	**5.32±0.57**	8.00±0.00	7.49±0.64	N/A	N/A
$N = 1000$	Earthquake	1.33±0.57	4.00±0.00	3.00±0.00	3.00±0.00	1.00±0.00	**0.00±0.00**	0.66±0.57	**0.00±0.00**	0.66±0.57	**0.00±0.00**
	Cancer	1.33±0.57	3.00±0.00	3.00±0.00	3.00±0.00	2.00±0.00	1.60±0.48	1.33±0.57	**0.00±0.00**	1.33±0.57	**0.00±0.00**
	Survey	1.00±0.00	4.00±0.00	5.80±0.39	5.40±0.48	3.20±0.39	2.71±0.27	1.00±0.00	2.83±0.00	1.00±0.00	**0.00±0.00**
	Asia	5.00±0.00	4.00±0.00	6.20±0.74	6.60±0.48	3.40±0.48	1.75±0.43	5.00±0.00	**0.97±0.62**	N/A	N/A
	Asia-M	1.50±0.00	4.00±0.00	8.00±0.00	5.20±0.39	3.40±0.48	2.04±0.51	1.00±0.00	**0.65±0.47**	1.33±0.57	3.00±0.00
	Child	8.25±0.00	3.80±0.74	12.2±1.72	11.8±0.74	15.2±0.97	**2.00±0.00**	7.00±0.00	**2.00±0.40**	N/A	N/A
	Neuropathic	-	6.00±0.00	**4.00±0.81**	12.0±0.00	12.0±0.00	5.54±0.75	-	10.1±2.12	N/A	N/A
$N = 5000$	Earthquake	0.50±0.86	4.00±0.00	2.80±0.39	3.00±0.00	1.00±0.00	0.80±0.97	**0.00±0.00**	**0.00±0.00**	**0.00±0.00**	**0.00±0.00**
	Cancer	1.33±0.57	3.00±0.00	3.00±0.00	3.00±0.00	2.00±0.00	2.00±0.00	1.33±0.57	**0.00±0.00**	1.33±0.57	**0.00±0.00**
	Survey	2.00±0.00	4.00±0.00	5.00±0.00	5.00±0.00	3.00±0.00	3.33±0.69	2.00±0.00	2.60±0.00	2.00±0.00	**0.00±0.00**
	Asia	1.00±0.00	4.00±0.00	6.60±0.79	4.40±1.35	3.40±0.48	1.75±0.43	**0.00±0.00**	0.97±0.62	N/A	N/A
	Asia-M	2.00±0.00	4.00±0.00	7.60±0.48	4.60±0.48	3.20±0.39	1.68±0.46	2.00±0.00	**0.00±0.00**	2.00±0.00	2.00±0.00
	Child	8.25±0.00	3.00±0.00	12.6±0.79	10.8±1.72	14.2±0.40	**3.00±0.00**	7.00±0.00	**3.00±0.00**	N/A	N/A
	Neuropathic	8.62±0.00	6.00±0.00	9.33±0.94	10.0±0.00	10.0±0.00	4.20±0.96	9.00±0.00	**1.23±0.42**	N/A	N/A
$N = 10000$	Earthquake	**0.00±0.00**	4.00±0.00	3.00±0.00	3.00±0.00	1.00±0.00	0.40±0.48	**0.00±0.00**	**0.00±0.00**	**0.00±0.00**	**0.00±0.00**
	Cancer	2.00±0.00	3.00±0.00	3.00±0.00	3.00±0.00	2.00±0.00	0.60±0.80	2.00±0.00	**0.00±0.00**	2.00±0.00	**0.00±0.00**
	Survey	2.00±0.00	4.00±0.00	5.00±0.00	5.00±0.00	3.00±0.00	3.60±1.35	2.00±0.00	1.83±0.00	2.00±0.00	**0.00±0.00**
	Asia	1.5±0.00	4.00±0.00	6.00±0.00	4.40±1.35	3.00±0.00	1.40±0.48	**0.00±0.00**	0.34±0.47	N/A	N/A
	Asia-M	1.00±0.00	4.00±0.00	8.00±0.00	4.80±0.39	3.00±0.00	2.00±0.00	**0.00±0.00**	**0.00±0.00**	**0.00±0.00**	3.00±0.00
	Child	6.00±3.04	3.00±0.00	12.2±1.46	11.6±0.48	14.4±0.48	2.80±0.84	5.00±2.64	**1.00±0.00**	N/A	N/A
	Neuropathic	10.00±0.00	6.00±0.00	**1.00±0.00**	10.0±0.00	10.0±0.00	3.00±0.00	10.00±0.00	**1.00±0.00**	N/A	N/A

Table A10: Comparison with causal discovery methods, showing mean and std dev of D_{top} over 3 runs. (For the Neuropathic subgraph (1k samples), PC Algorithm returns cyclic graphs in the MEC). Human experiments not conducted for Child (due to feasibility issues), hence rows marked as N/A.

Dataset	Samples	LLM	Ground Truth	PC (Average over MEC)
Asia	250	1.00±0.00	0.00±0.00	2.00±0.00
	1000	3.00±0.00	2.00±0.00	3.00±0.00
	10000	3.00±0.00	3.00±0.00	3.00±0.00
Child	250	5.00±0.00	5.00±0.00	6.50±0.00
	1000	6.00±0.00	6.00±0.00	8.43±0.00
	10000	9.00±0.00	9.00±0.00	9.75±0.00

Table A11: Comparing D_{top} of final graph using LLM order vs Ground truth order as prior to PC algorithm for Child and Asia graph, averaged over 4 runs

Dataset	Samples	$\epsilon_{ATE}(S_1)$	$\epsilon_{ATE}(S_2)$	$\epsilon_{ATE}(S_3)$	Δ_{12}	Δ_{13}
Asia	250	0.70±0.40	0.70±0.39	0.69±0.39	0.00±0.00	0.00±0.00
	500	0.64±0.39	0.64±0.39	0.64±0.38	0.00±0.00	0.00±0.00
	1000	0.59±0.32	0.59±0.32	0.59±0.32	0.00±0.00	0.00±0.00
	5000	0.59±0.30	0.59±0.30	0.59±0.29	0.00±0.00	0.00±0.00
	10000	0.49±0.00	0.49±0.00	0.49±0.00	0.00±0.00	0.00±0.00

Table A12: Results on Asia dataset. Here we test the difference in the estimated causal effect of *lung* on *dyspnoea* when the causal effect is estimated using the backdoor set S_1 = *{smoke}* vs. the causal effect estimated when all variables in two topological orders as backdoor sets: S_2 = {*asia, smoke*}, S_2= {*asia, tub, smoke*}. Δ_{12}, Δ_{13} refers to the absolute difference between the pairs $\epsilon_{ATE}(S_1), \epsilon_{ATE}(S_2)$ and $\epsilon_{ATE}(S_1), \epsilon_{ATE}(S_3)$ respectively. From the last two columns, we observe that using the variables that come before the treatment node in a topological order as a backdoor set does not result in the deviation of causal effects from the ground truth effects.

Dataset	Number of Nodes	Number of Edges	Description (used as a context)
Asia	8	8	Model the possible respiratory problems someone can have who has recently visited Asia and is experiencing shortness of breath
Cancer	5	4	Model the relation between various variables responsible for causing Cancer and its possible outcomes
Earthquake	5	5	Model factors influencing the probability of a burglary
Survey	6	6	Model a hypothetical survey whose aim is to investigate the usage patterns of different means of transport
Child	20	25	Model congenital heart disease in babies
Neuropathic Pain Diagnosis (subgraph)	22	25	For neuropathic pain diagnosis

Table A15: Overview of datasets used

Metric	Pairwise	Triplet
D_{top}	-	**3**
SHD	64	**24**
Cycles	5	**0**
IN/TN	**0/22**	13/20

Table A14: Results using LLM for triplet vs pairwise query strategies on the *Neuropathic* dataset subgraph (metrics/values follow same procedure as Table A5)

H Cost Estimation Analysis

While the number of total API calls to LLM for triplet-based prompt is higher than the standard pairwise setup, our method still ensures scalability by optimizing majority of API calls to cheaper and lower version model (GPT-3.5-Turbo) while ensuring better performance. Our triplet pipeline further optimizes on performance by utilizing multiple context switch (varying third node across all triplets) for orienting each pair of nodes, which further leads to higher confidence and accuracy in its final orientation. Moreover, strategically deploying advanced language models such as GPT-4 only for conflict resolution enhances effectiveness, streamlines costs and enables scalability. Our analysis also suggests that an estimated cost of using a 100 node graph for Pairwise orientation using GPT-4 will lead to a total cost of 574 USD, whereas using our triplet strategy which leverages GPT-4 calls while optimizing performance of weaker models like GPT-3.5-turbo, will cost an estimated total of 55 USD. Pairwise analysis of a 100 node graph using GPT-3.5-turbo on the other hand will lead to an estimated cost of 10 USD. Therefore while our proposed triplet pipeline uses more calls, but is able to use GPT-4 more optimally because of inbuilt error correction thus leading to cost optimization while improving performance significantly over pairwise orientation using GPT-3.5-turbo.

I Causal Graphs used in Experiments

Figures A3-A7 show the causal graphs and details we considered from BNLearn repository [Scutari and Denis, 2014].

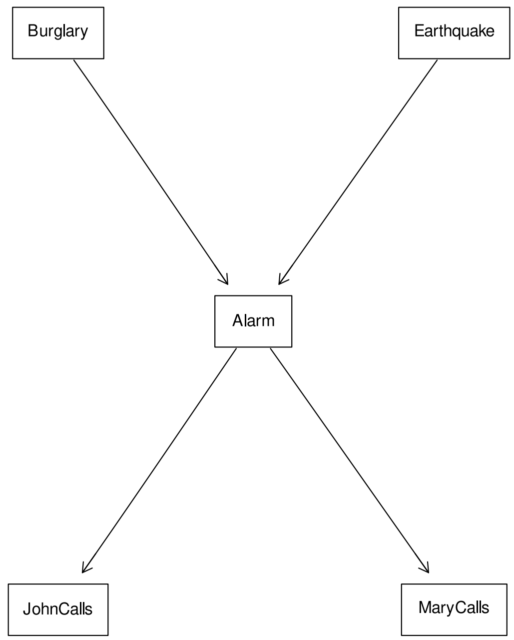


Figure A3: Earthquake Bayesian network. Abbreviations/Descriptions: Burglary: *burglar entering*, Earthquake: *earthquake hitting*, Alarm: *home alarm going off in a house*, JohnCalls: *first neighbor to call to inform the alarm sound*, Marycalls: *second neighbor to call to inform the alarm sound*.

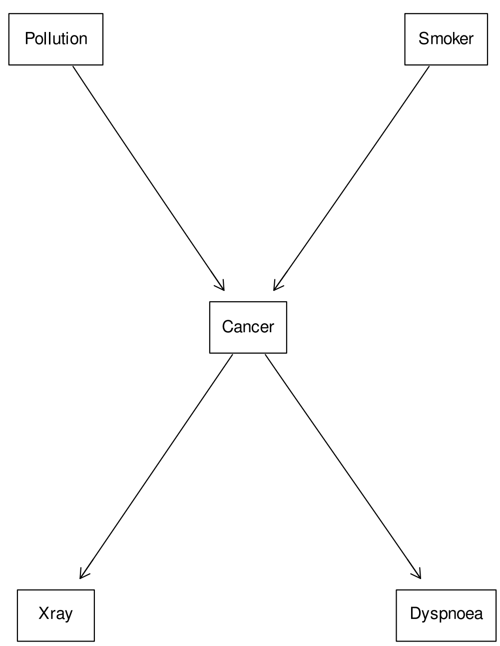


Figure A4: Cancer Bayesian network. Abbreviations/Descriptions: Pollution: *exposure to pollutants*, Smoker: *smoking habit*, Cancer: *Cancer*, Dyspnoea: *Dyspnoea*, Xray: *getting positive xray result*.

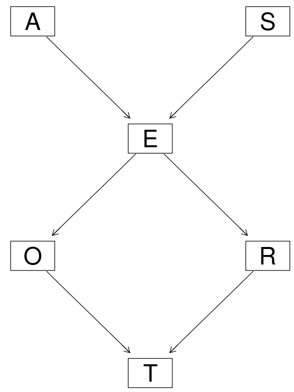


Figure A5: Survey Bayesian network. Abbreviations: A=*Age/Age of people using transport*, S=*Sex/male or female*, E=*Education/up to high school or university degree*, O=*Occupation/employee or self-employed*, R=*Residence/the size of the city the individual lives in, recorded as either small or big*, T=*Travel/the means of transport favoured by the individual*.

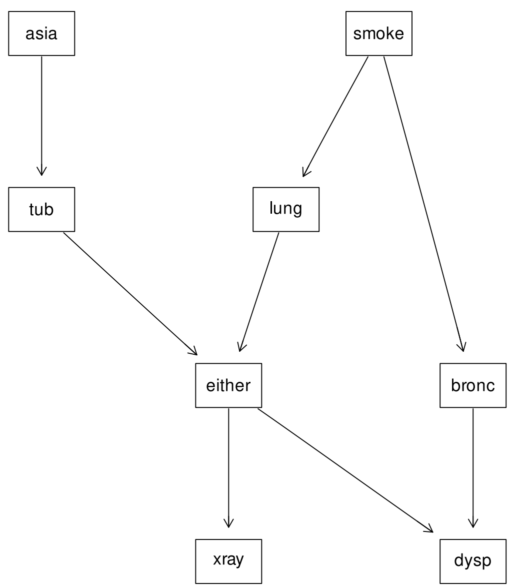


Figure A6: Asia Bayesian network. Abbreviations/Descriptions: asia=*visit to Asia/visiting Asian countries with high exposure to pollutants*, smoke=*smoking habit*, tub=*tuberculosis*, lung=*lung cancer*, either=*either tuberculosis or lung cancer*, bronc=*bronchitis*, dysp=*dyspnoea*, xray=*getting positve xray result*.

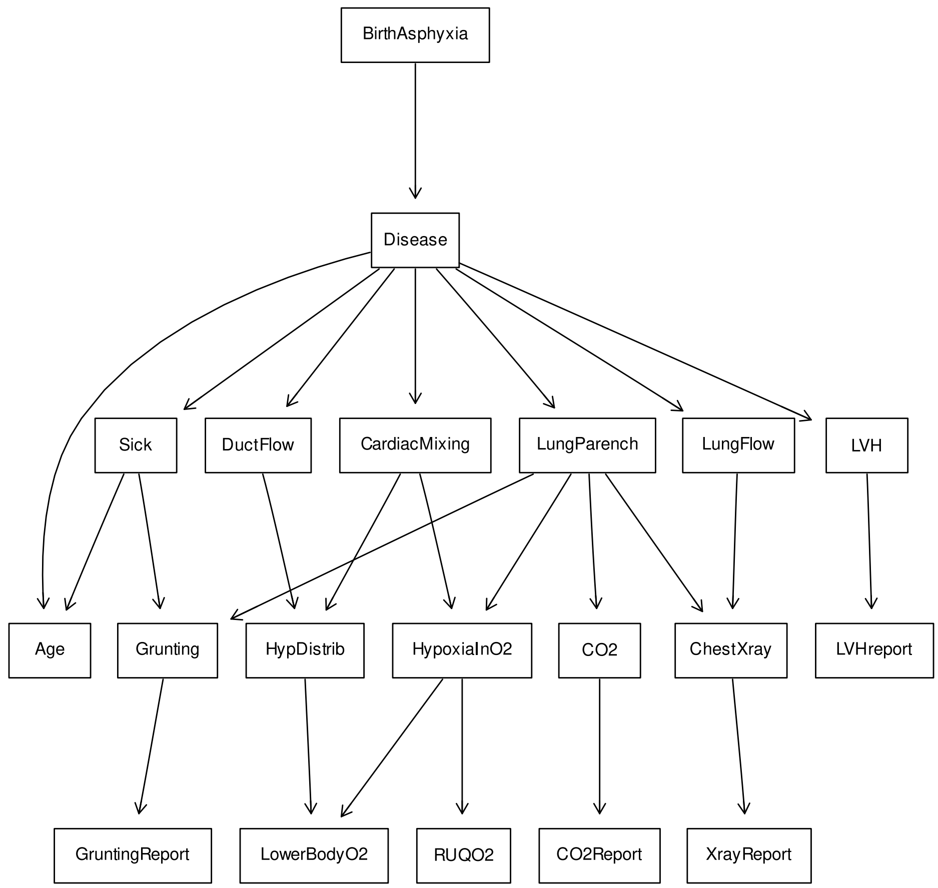


Figure A7: Child Bayesian network. Abbreviations: BirthAsphyxia: *Lack of oxygen to the blood during the infant's birth*, HypDistrib: *Low oxygen areas equally distributed around the body*, HypoxiaInO2: *Hypoxia when breathing oxygen*, CO2: *Level of carbon dioxide in the body*, ChestXray: *Having a chest x-ray*, Grunting: *Grunting in infants*, LVHreport: *Report of having left ventricular hypertrophy*, LowerBodyO2: *Level of oxygen in the lower body*, RUQO2: *Level of oxygen in the right upper quadricep muscle*, CO2Report: *A document reporting high levels of CO2 levels in blood*, XrayReport: *Report of having a chest x-ray*, Disease: *Presence of an illness*, GruntingReport: *Report of infant grunting*, Age: *Age of infant at disease presentation*, LVH: *Thickening of the left ventricle*, DuctFlow: *Blood flow across the ductus arteriosus*, CardiacMixing: *Mixing of oxygenated and deoxygenated blood*, LungParench: *The state of the blood vessels in the lungs*, LungFlow: *Low blood flow in the lungs*, Sick: *Presence of an illness*

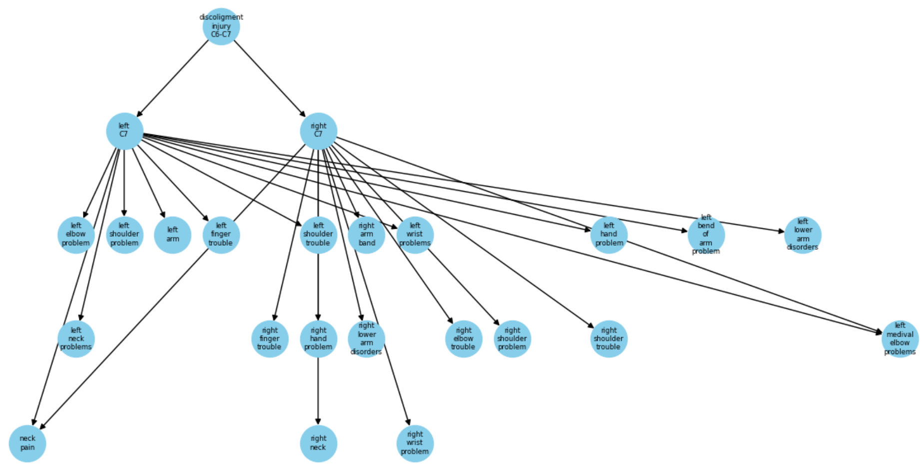


Figure A8: For Neuropathic dataset, we consider a sub-graph induced by one of the root nodes, containing the following 22 nodes and corresponding edges taken from `https://observablehq.com/@turuibo/the-complete-causal-graph-of-neuropathic-pain-diagnosis`: 'right C7', 'right elbow trouble', 'left shoulder trouble', 'left bend of arm problem', 'right shoulder trouble', 'right hand problem', 'left medival elbow problems', 'right finger trouble', 'left neck problems', 'left wrist problems', 'left shoulder problem', 'right neck', 'right wrist problem', 'right shoulder problem', 'discoligment injury C6 C7', 'left hand problem', 'left C7', 'right arm band', 'left lower arm disorders', 'neck pain', 'left finger trouble', 'left arm'. We did not use descriptions for the nodes of Neuropathic graph.

Question: For a causal graph used to model relationship of various factors and outcomes related to cancer with the following nodes: ['Pollution', 'Cancer', 'Smoker', 'Xray', 'Dyspnoea'],
Which cause-and-effect relationship is more likely between nodes 'smoker' and 'cancer'?

A. changing the state of node 'smoker' causally affects a change in another node 'cancer'.

B. changing the state of node 'cancer' causally affects a change in another node 'smoker'.

C. There is no causal relation between the nodes 'cancer' and 'smoker'.

Make sure to first provide a grounded reasoning for your answer and then provide the answer in the following format: ⟨Answer⟩A/B/C⟨/Answer⟩. It is very important that you output the final Causal graph within the tags like ⟨Answer⟩A/B/C⟨/Answer⟩ otherwise your answer will not be processed.

Answer: The causal effect of "smoker" directing to "cancer" is based on the strong evidence from epidemiological studies linking smoking to an increased risk of developing cancer. Smoking introduces harmful substances into the respiratory system, leading to cellular damage and mutation, which significantly raises the likelihood of cancer development in the lungs or respiratory tract, subsequently impacting the occurrence of respiratory problems like shortness of breath. Therefore answer is ⟨Answer⟩A⟨/Answer⟩

Question: For a causal graph used to model relationship of various factors and outcomes related to cancer with the following nodes: ['Pollution', 'Cancer', 'Smoker', 'Xray', 'Dyspnoea'], Which cause-and-effect relationship is more likely between nodes 'xray' and 'dyspnoea'?

A. changing the state of node 'xray' causally affects a change in another node 'dyspnoea'.

B. changing the state of node 'dyspnoea' causally affects a change in another node 'xray'.

C. There is no causal relation between the nodes 'xray' and 'dyspnoea'.

Make sure to first provide a grounded reasoning for your answer and then provide the answer in the following format: ⟨Answer⟩A/B/C⟨/Answer⟩. It is very important that you output the final Causal graph within the tags like ⟨Answer⟩A/B/C⟨/Answer⟩ otherwise your answer will not be processed.

Answer: Reasoning behind the lack of causal relation between X-ray and dyspnoea is that X-ray and dyspnoea are both effects of having cancer, but they do not directly cause or affect each other. X-ray is a diagnostic test that can help detect cancer in the lungs or other organs, while dyspnoea is a symptom of cancer that involves feeling short of breath. Therefore, X-ray and dyspnoea are not causally related, but they are both associated with cancer. Therefore answer is ⟨Answer⟩C⟨/Answer⟩

Question: For a causal graph used to model relationship of various factors and outcomes related to cancer with the following nodes: ['Pollution', 'Cancer', 'Smoker', 'Xray', 'Dyspnoea'], Which cause-and-effect relationship is more likely between nodes 'xray' and 'cancer'?

A. changing the state of node 'xray' causally affects a change in another node 'cancer'.

B. changing the state of node 'cancer' causally affects a change in another node 'xray'.

C. There is no causal relation between the nodes 'xray' and 'cancer'.

Make sure to first provide a grounded reasoning for your answer and then provide the answer in the following format: ⟨Answer⟩A/B/C⟨/Answer⟩. It is very important that you output the final Causal graph within the tags like ⟨Answer⟩A/B/C⟨/Answer⟩ otherwise your answer will not be processed.

Table A16: Chain of Thought Prompt

Answer: The causal effect of cancer on X-ray is that X-rays are often used to diagnose or detect cancer in different parts of the body, such as the bones, lungs, breasts, or kidneys123. X-rays use low doses of radiation to create pictures of the inside of the body and show the presence, size, and location of tumors. X-rays can also help monitor the response to treatment or check for signs of recurrence. Therefore, having cancer may increase the likelihood of getting an X-ray as part of the diagnostic process or follow-up care. Therefore answer is ⟨Answer⟩B⟨/Answer⟩

Question: For a causal graph used to model relationship of various factors and outcomes related to cancer with the following nodes: ['Pollution', 'Cancer', 'Smoker', 'Xray', 'Dyspnoea'], Which cause-and-effect relationship is more likely between nodes 'pollution' and 'cancer'?

A. changing the state of node 'pollution' causally affects a change in another node 'cancer'.

B. changing the state of node 'cancer' causally affects a change in another node 'pollution'.

C. There is no causal relation between the nodes 'pollution' and 'cancer'.

Make sure to first provide a grounded reasoning for your answer and then provide the answer in the following format: ⟨Answer⟩A/B/C⟨/Answer⟩. It is very important that you output the final Causal graph within the tags like ⟨Answer⟩A/B/C⟨/Answer⟩ otherwise your answer will not be processed.

Answer: The causal effect of pollution on cancer is that air pollution contains carcinogens (cancer-causing substances) that may be absorbed into the body when inhaled and damage the DNA of cells. Another possible reasoning is that particulate matter (tiny dust-like particles) in air pollution may cause physical damage to the cells in the lungs, leading to inflammation and oxidative stress and eventually cell mutations. A third possible reasoning is that air pollution may create an inflamed environment in the lungs that encourages the proliferation of cells with existing cancer-driving mutations. These are some of the hypotheses that researchers have proposed to explain how air pollution may cause cancer, but more studies are needed to confirm them. Therefore answer is ⟨Answer⟩A⟨/Answer⟩

Question: For a causal graph used to model relationship of various factors and outcomes related to cancer with the following nodes: ['Pollution', 'Cancer', 'Smoker', 'Xray', 'Dyspnoea'], Which cause-and-effect relationship is more likely between nodes 'pollution' and 'smoker'?

A. changing the state of node 'pollution' causally affects a change in another node 'smoker'.

B. changing the state of node 'smoker' causally affects a change in another node 'pollution'.

C. There is no causal relation between the nodes 'pollution' and 'smoker'.

Make sure to first provide a grounded reasoning for your answer and then provide the answer in the following format: ⟨Answer⟩A/B/C⟨/Answer⟩. It is very important that you output the final Causal graph within the tags like ⟨Answer⟩A/B/C⟨/Answer⟩ otherwise your answer will not be processed.

Answer: Reason behind the lack of causal relation between pollution and smoker is that pollution and smoking are both independent risk factors for respiratory problems, but they do not directly cause or affect each other. Pollution and smoking both contribute to air pollution, which can harm the health of people and the environment. However, pollution is mainly caused by human activities such as burning fossil fuels, deforestation, or industrial processes, while smoking is a personal choice that involves inhaling tobacco smoke. Therefore, pollution and smoker are not causally related, but they are both associated with respiratory problems. Therefore answer is ⟨Answer⟩C⟨/Answer⟩.

Table A17: Chain of Thought Prompt (continued..)

Question: For a causal graph used for modeling factors causing Coronary Heart Diseases with the following nodes: ['Family Disease', 'Gene', 'Smoking', 'Blood Pressure', 'Coronary Heart Disease', 'Headache'], Which cause-and-effect relationship is more likely between nodes 'Family Disease' and 'Gene'?

A. changing the state of node 'Family Disease' causally affects a change in another node 'Gene'.

B. changing the state of node 'Gene' causally affects a change in another node 'Family Disease'.

C. There is no causal relation between the nodes 'Family Disease' and 'Gene'.

Make sure to first provide a grounded reasoning for your answer and then provide the answer in the following format: ⟨Answer⟩A/B/C⟨/Answer⟩. It is very important that you output the final Causal graph within the tags like ⟨Answer⟩A/B/C⟨/Answer⟩ otherwise your answer will not be processed.

Answer: Reason behind the causal effect of family disease on gene is that family disease is a term that refers to diseases or health conditions that run in the family, meaning that they are influenced by genetic factors. Gene is a term that refers to the basic unit of heredity that carries information for a specific trait or function. Family disease can affect gene by altering the type or frequency of genes that are inherited by the offspring from their parents. For example, some family diseases are caused by deterministic genes, which are genes that guarantee the development of a disease if they are present in a person's genome. Other family diseases are influenced by risk genes, which are genes that increase the likelihood of developing a disease but do not guarantee it. Therefore, family disease can causally direct to gene by changing the genetic makeup of a person or a population. Therefore answer is ⟨Answer⟩A⟨/Answer⟩.

Question: For a causal graph used for modeling factors causing Coronary Heart Diseases with the following nodes: ['Family Disease', 'Gene', 'Smoking', 'Blood Pressure', 'Coronary Heart Disease', 'Headache'], Which cause-and-effect relationship is more likely between nodes 'Coronary Heart Disease' and 'Gene'?

A. changing the state of node 'Coronary Heart Disease' causally affects a change in another node 'Gene'.

B. changing the state of node 'Gene' causally affects a change in another node 'Coronary Heart Disease'.

C. There is no causal relation between the nodes 'Coronary Heart Disease' and 'Gene'.

Make sure to first provide a grounded reasoning for your answer and then provide the answer in the following format: ⟨Answer⟩A/B/C⟨/Answer⟩. It is very important that you output the final Causal graph within the tags like ⟨Answer⟩A/B/C⟨/Answer⟩ otherwise your answer will not be processed.

Answer: Possible reasoning behind the causal effect of gene on coronary heart disease is that gene is a term that refers to the basic unit of heredity that carries information for a specific trait or function. Gene can affect coronary heart disease by influencing the structure and function of the blood vessels, the metabolism and transport of lipids (fats) in the blood, the inflammation and clotting processes, or the response to environmental factors such as smoking or diet. For example, some genes code for proteins that regulate the cell cycle and growth of the cells that line the arteries, which can affect their susceptibility to damage or plaque formation. Other genes code for proteins that control the synthesis and clearance of cholesterol or other lipids, which can affect their levels and deposition in the arteries. Therefore, gene can causally direct to coronary heart disease by modifying the biological pathways that contribute to the development or progression of the disease. Therefore answer is ⟨Answer⟩B⟨/Answer⟩

Table A18: Chain of Thought Prompt (continued..)

Question: For a causal graph used for modeling factors causing Coronary Heart Diseases with the following nodes: [’Family Disease’, ’Gene’, ’Smoking’, ’Blood Pressure’, ’Coronary Heart Disease’, ’Headache’], Which cause-and-effect relationship is more likely between nodes ’Blood Pressure’ and ’Smoking’?

A. changing the state of node ’Blood Pressure’ causally affects a change in another node ’Smoking’.

B. changing the state of node ’Smoking’ causally affects a change in another node ’Blood Pressure’.

C. There is no causal relation between the nodes ’Blood Pressure’ and ’Smoking’.

Make sure to first provide a grounded reasoning for your answer and then provide the answer in the following format: ⟨Answer⟩A/B/C⟨/Answer⟩. It is very important that you output the final Causal graph within the tags like ⟨Answer⟩A/B/C⟨/Answer⟩ otherwise your answer will not be processed.

Answer: Possible reasoning behind the causal effect of smoking on blood pressure is that smoking is a habit that involves inhaling tobacco smoke, which contains nicotine and other harmful chemicals. Smoking can affect blood pressure by activating the sympathetic nervous system (SNS), which is the part of the nervous system that controls the body’s response to stress or danger. When the SNS is activated, it releases hormones such as adrenaline and noradrenaline, which cause the heart to beat faster and harder, and the blood vessels to constrict. This results in a temporary increase in blood pressure, which can last for 15 to 20 minutes after each cigarette. Therefore, smoking can causally direct to blood pressure by stimulating the SNS and increasing the cardiac output and vascular resistance. Therefore answer is ⟨Answer⟩B⟨/Answer⟩.

Question: For a causal graph used for modeling factors causing Coronary Heart Diseases with the following nodes: [’Family Disease’, ’Gene’, ’Smoking’, ’Blood Pressure’, ’Coronary Heart Disease’, ’Headache’], Which cause-and-effect relationship is more likely between nodes ’Headache’ and ’Smoking’?

A. changing the state of node ’Headache’ causally affects a change in another node ’Smoking’.

B. changing the state of node ’Smoking’ causally affects a change in another node ’Headache’.

C. There is no causal relation between the nodes ’Headache’ and ’Smoking’.

Make sure to first provide a grounded reasoning for your answer and then provide the answer in the following format: ⟨Answer⟩A/B/C⟨/Answer⟩. It is very important that you output the final Causal graph within the tags like ⟨Answer⟩A/B/C⟨/Answer⟩ otherwise your answer will not be processed.

Answer: One possible reasoning behind the lack of causal relation between headache and smoking is that headache and smoking are both associated with various health conditions, but they do not directly cause or affect each other12. Headache is a term that refers to pain or discomfort in the head, scalp, or neck, which can have many possible causes, such as stress, dehydration, infection, injury, or medication. Smoking is a habit that involves inhaling tobacco smoke, which contains nicotine and other harmful chemicals, which can increase the risk of diseases such as cancer, heart disease, stroke, and lung disease. Therefore, headache and smoking are not causally related, but they are both linked to different health problems. Therefore the answer is ⟨Answer⟩C⟨/Answer⟩

Table A19: Chain of Thought Prompt (continued..)

Question: For a causal graph used for modeling factors causing Coronary Heart Diseases with the following nodes: ['Family Disease', 'Gene', 'Smoking', 'Blood Pressure', 'Coronary Heart Disease', 'Headache'], Which cause-and-effect relationship is more likely between nodes 'Headache' and 'Smoking'?

A. changing the state of node 'Headache' causally affects a change in another node 'Smoking'.

B. changing the state of node 'Smoking' causally affects a change in another node 'Headache'.

C. There is no causal relation between the nodes 'Headache' and 'Smoking'.

Make sure to first provide a grounded reasoning for your answer and then provide the answer in the following format: ⟨Answer⟩A/B/C⟨/Answer⟩. It is very important that you output the final Causal graph within the tags like ⟨Answer⟩A/B/C⟨/Answer⟩ otherwise your answer will not be processed.

Answer: One possible reasoning behind the lack of causal relation between headache and smoking is that headache and smoking are both associated with various health conditions, but they do not directly cause or affect each other. Headache is a term that refers to pain or discomfort in the head, scalp, or neck, which can have many possible causes, such as stress, dehydration, infection, injury, or medication. Smoking is a habit that involves inhaling tobacco smoke, which contains nicotine and other harmful chemicals, which can increase the risk of diseases such as cancer, heart disease, stroke, and lung disease. Therefore, headache and smoking are not causally related, but they are both linked to different health problems. Therefore the answer is ⟨Answer⟩C⟨/Answer⟩

Question: For a causal graph used for modeling factors causing Coronary Heart Diseases with the following nodes: ['Family Disease', 'Gene', 'Smoking', 'Blood Pressure', 'Coronary Heart Disease', 'Headache'], Which cause-and-effect relationship is more likely between nodes 'Coronary Heart Disease' and 'Smoking'?

A. changing the state of node 'Smoking' causally affects a change in another node 'Coronary Heart Disease'.

B. changing the state of node 'Coronary Heart Disease' causally affects a change in another node 'Smoking'.

C. There is no causal relation between the nodes 'Coronary Heart Disease' and 'Smoking'.

Make sure to first provide a grounded reasoning for your answer and then provide the answer in the following format: ⟨Answer⟩A/B/C⟨/Answer⟩. It is very important that you output the final Causal graph within the tags like ⟨Answer⟩A/B/C⟨/Answer⟩ otherwise your answer will not be processed.

Answer: Possible reasoning behind the causal effect of smoking on coronary heart disease is smoking damages the heart and blood vessels by raising triglycerides, lowering HDL, increasing blood clotting, and impairing blood flow to the heart. This can lead to plaque buildup, heart attacks, and death. Therefore answer is ⟨Answer⟩A⟨/Answer⟩.

Question: For a causal graph used for context with the following nodes: nodes, Which cause-and-effect relationship is more likely between nodes X and Y?

A. changing the state of node X causally affects a change in another node Y.

B. changing the state of node Y causally affects a change in another node X.

C. There is no causal relation between the nodes X and Y.

Make sure to first provide a grounded reasoning for your answer and then provide the answer in the following format: ⟨Answer⟩A/B/C⟨/Answer⟩. It is very important that you output the final Causal graph within the tags like ⟨Answer⟩A/B/C⟨/Answer⟩ otherwise your answer will not be processed.

Table A20: Chain of Thought Queries (continued..)

Which cause-and-effect relationship is more likely?

A. changing the state of node which says X causally affects a change in another node which says Y.

B. changing the state of node which says Y causally affects a change in another node which says X.

C. There is no causal relationship between node X and Y.

Make sure to first output a factually grounded reasoning for your answer. X and Y are nodes of a Causal Graph. The causal graph is sparse and acyclic in nature. So option C could be chosen if there is some uncertainity about causal relationship between X and Y.

First give your reasoning and after that please make sure to provide your final answer within the tags ⟨Answer⟩A/B/C⟨/Answer⟩.
It is very important that you output your final answer between the tags like ⟨Answer⟩A/B/C⟨/Answer⟩ otherwise your response will not be processed.

Table A21: Base Queries

For the nodes X and Y which form an edge in a Causal Graph, you have to identify which cause-and-effect relationship is more likely between the nodes of the edge. This will be used to rearrange the nodes in the edge to create a directed edge which accounts for causal relation from one node to another in the edge.

A. changing the state of node X causally affects a change in another node Y.

B. changing the state of node Y causally affects a change in another node X.

C. There is no causal relation between the nodes X and Y.

You can also take the edges from the skeleton which have been rearranged to create a directed edge to account for causal relationship between the nodes: directed_edges.
Make sure to first output a factually grounded reasoning for your answer. First give your reasoning and after that please make sure to provide your final answer within the tags ⟨Answer⟩A/B/C⟨/Answer⟩.
It is very important that you output your final answer between the tags like ⟨Answer⟩A/B/C⟨/Answer⟩ otherwise your response will not be processed.

Table A22: Iterative orientation Queries

For the following undirected edge in a Causal Graph made of nodes X and Y, you have to identify which cause-and-effect relationship is more likely between the nodes of the edge. This will be used to rearrange the nodes in the edge to create a directed edge which accounts for causal relation from one node to another in the edge.

A. changing the state of node X causally affects a change in another node Y.

B. changing the state of node Y causally affects a change in another node X.

C. There is no causal relation between the nodes X and Y.

You can also take the other directed edges of nodes X: X_edges and Y: Y_edges of the Causal graph as context to redirect the edge to account for causal effect.
Make sure to first output a factually grounded reasoning for your answer. First give your reasoning and after that please make sure to provide your final answer within the tags ⟨Answer⟩A/B/C⟨/Answer⟩.
It is very important that you output your final answer between the tags like ⟨Answer⟩A/B/C⟨/Answer⟩ otherwise your response will not be processed.

Table A23: Markov Blanket Queries

*Identify the causal relationships between the given variables and create a directed acyclic graph to **{context}**. Make sure to give a reasoning for your answer and then output the directed graph in the form of a list of tuples, where each tuple is a directed edge. The desired output should be in the following form: [('A','B'), ('B','C')] where first tuple represents a directed edge from Node 'A' to Node 'B', second tuple represents a directed edge from Node 'B' to Node 'C'and so on.*

If a node should not form any causal relationship with other nodes, then you can add it as an isolated node of the graph by adding it seperately. For example, if 'C' should be an isolated node in a graph with nodes 'A', 'B', 'C', then the final DAG representation should be like [('A','B'), ('C')].
Use the description about the node provided with the nodes in brackets to form a better decision about the causal direction orientation between the nodes.

It is very important that you output the final Causal graph within the tags <Answer></Answer>otherwise your answer will not be processed.

Example:
Input: Nodes: ['A', 'B', 'C', 'D'];
Description of Nodes: [(description of Node A), (description of Node B), (description of Node C), (description of Node D)]
Output: <Answer>[('A','B'),('C','D')]</Answer>
Question:
*Input: Nodes: **{Triplet Nodes Input}***
*Description of Nodes: **{Description of Each Node from the Triplet}***
Output:

Table A24: The *triplet* query template, which includes a concise context of the graph, the input triplet nodes and their respective descriptions. As an example, for the Child graph, the context is *"to model congenital heart disease in babies"*, the three nodes may be *('HypoxiaInO2', 'Grunting', 'GruntingReport')*; and their node descriptions are *["hypoxia when breathing oxygen", "grunting in infants", "report of infant grunting"]* respectively.

Input: **('HypDistrib', 'LowerBodyO2')** Answer: Low oxygen areas equally distributed around the body can affect the level of oxygen in the lower body by reducing the amount of oxygen available for circulation. Therefore, the answer is ⟨Answer⟩A⟨/Answer⟩. Input: **('Disease', 'LungFlow')** Answer: Infant methemoglobinemia is a condition where the blood is unable to carry enough oxygen, which can lead to low blood flow in the lungs. Therefore, the answer is ⟨Answer⟩A⟨/Answer⟩. Input: **('CardiacMixing', 'HypDistrib')** Answer: Mixing of oxygenated and deoxygenated blood is a common cause of congenital heart disease in babies. This mixing can occur due to structural defects in the heart, such as a hole between the heart chambers or abnormal connections between the blood vessels. This mixing can lead to low oxygen areas equally distributed around the body. Therefore, the answer is ⟨Answer⟩B⟨/Answer⟩.

Table A25: Example LLM (GPT-3.5-turbo) outputs for estimating the causal direction between a given pair of nodes (with their descriptions) using the CoT Prompt (refer Table A16 in Appendix for the prompt). Given a pair (X,Y), Option A refers to X causing Y, option B refers to Y causing X, and option C refers to no causal relationship between the variables. The pairs shown are from the *Child* graph from BNLearn repository [Scutari and Denis, 2014], that is used for modeling Congenital Heart Disease in infants.

Input: **('Right C7','Discoligment injury C6-C7')** Answer: Discoligment injury C6-C7 can cause compression of the nerve roots that exit the spinal cord at the C7 level, which can lead to symptoms such as pain, numbness, and weakness in the right C7 dermatome. Therefore, the answer is ⟨Answer⟩B⟨/Answer⟩. Input: **('Right C7', 'Left C7')** Answer: Right C7 and left C7 are both parts of the cervical spine and are not known to directly influence each other. Therefore, the answer is ⟨Answer⟩C⟨/Answer⟩. Input: **('Right elbow trouble', 'Left shoulder trouble')** Answer: There is no direct causal relationship between right elbow trouble and left shoulder trouble. They may both be symptoms of a larger underlying condition, but they do not directly cause or affect each other. Therefore the answer is ⟨Answer⟩C⟨/Answer⟩.

Table A26: Example LLM (GPT-3.5-turbo) reasoning outputs for estimating causal directionality between different pairs of nodes using CoT queries (refer Table A16 for the query) for Neuropathic subgraph (used for pain diagnosis).