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

Aniket Vashishtha[1], Abbavaram Gowtham Reddy[2], Abhinav Kumar[3], Saketh Bachu[2], Vineeth N Balasubramanian[2], Amit Sharma[1]

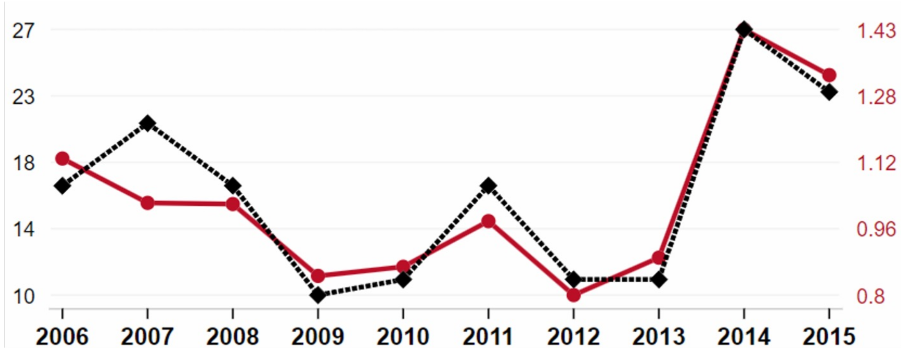
Microsoft Research India[1], IIT Hyderabad[2], MIT[3]

భారతీయ సాంకేతిక విజ్ఞాన సంస్థ హైదరాబాద్
भारतीय प्रौद्योगिकी संस्थान हैदराबाद
Indian Institute of Technology Hyderabad

Massachusetts Institute of Technology

Challenges in Causal Graph Discovery

- Learning the graph is a fundamental task
- All downstream causal tasks depend on the graph



Differentiating causation from correlation through observational data requires **domain knowledge**, in addition to discovery algorithms. [Tu et al. 2019, Huang et al. 2021, Kaiser & Sipos 2022]

Idea: Use LLMs' world knowledge to infer graph edges

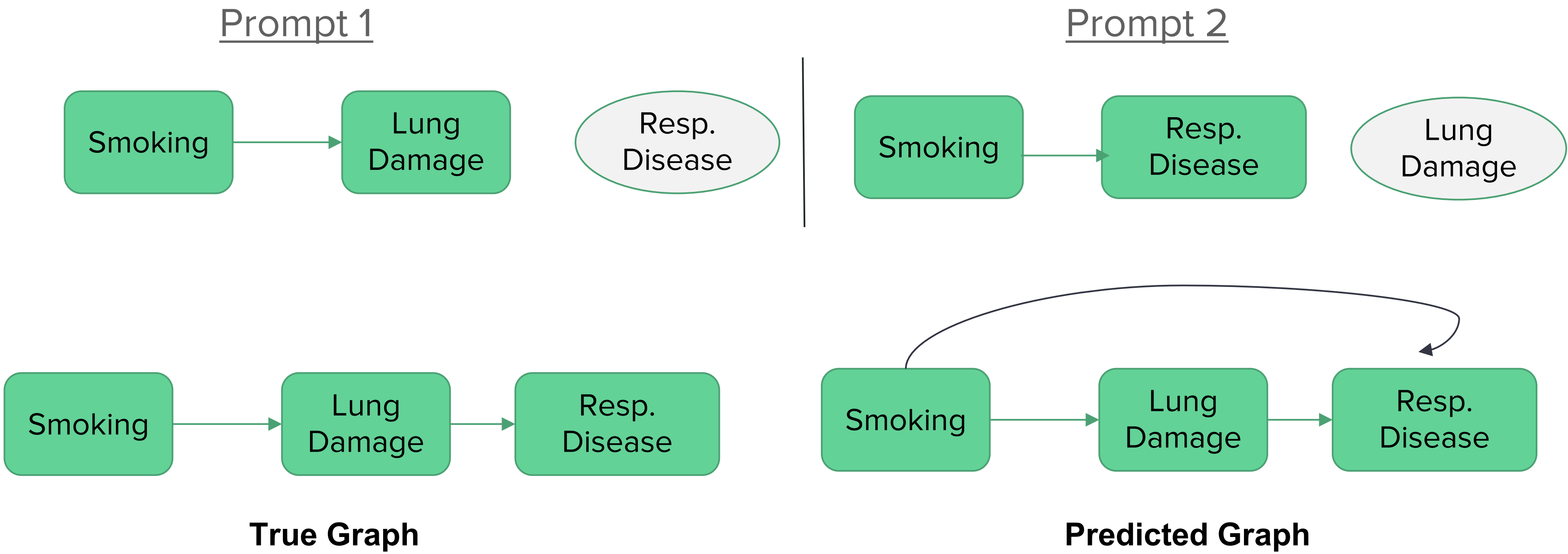
- **Pairwise Prompt:** A popular way to infer causal edge in a graph [Antonucci et al. 2023, Cohrs et al. 2023, Kiciman et al. 2023, Long et al. 2023, Willig et al. 2022].
  - Given a pair of variables, ask LLM to determine direction and existence of an edge.
- **Iterate** over all pairs to build a causal graph.

You are a helpful assistant to a neuropathic pain diagnosis expert. Which cause-and-effect relationship is more likely?

A. Left T6 Radiculopathy causes DLS T5-T6.
B. DLS T5-T6 causes Left T6 Radiculopathy.
C. No causal relationship exists.

Let's work this out in a step by step way to be sure that we have the right answer. Then provide your final answer within the tags <Answer>A/B</Answer>.

Fundamental Problem: Cannot distinguish direct and indirect effects using pairwise prompting



Fundamental Problem: Cannot distinguish direct and indirect effects using pairwise prompting, **even for human experts**

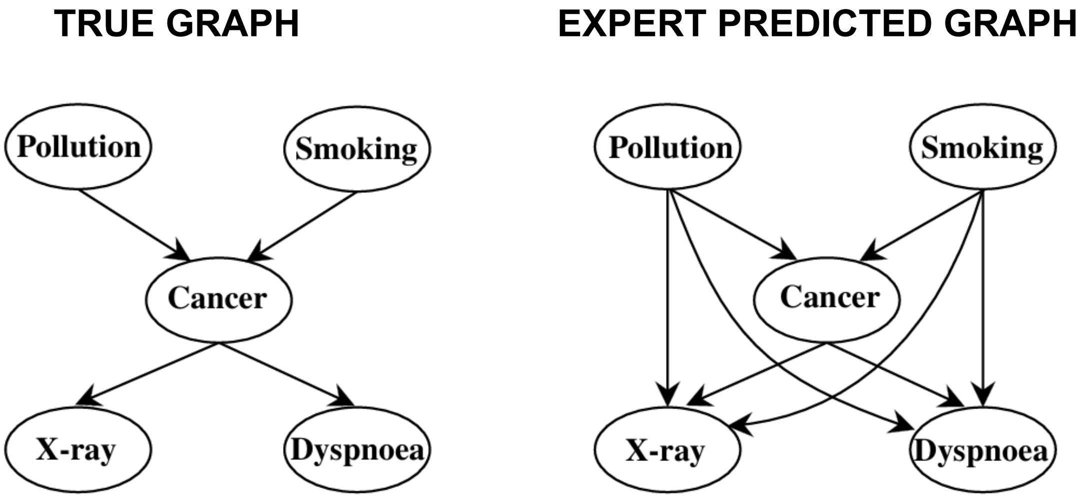
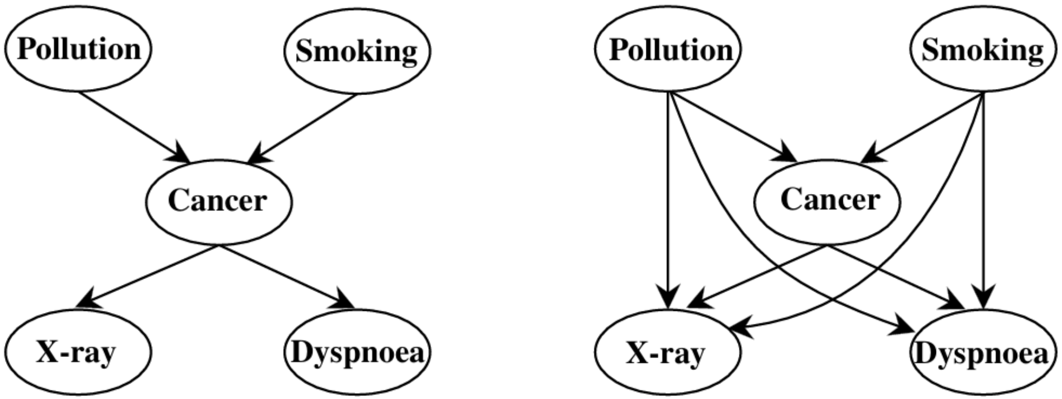


Figure 1: **Cancer dataset** (Scutari & Denis, 2014):

Key Insight: Graph is not the right output interface for experts' knowledge

- Experts such as LLMs and humans can only convey ancestral constraints [Ban et al. 2023]
- We propose Causal Order as the output interface of experts' knowledge



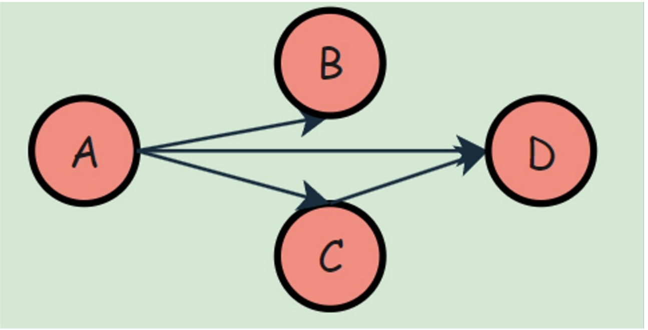
Unique Causal Order:

Pollution < Cancer < Dyspnoea

Figure 1: **Cancer dataset** (Scutari & Denis, 2014):

Causal Order: Definition

Causal order is a partial order relation over a set of entities or events, satisfying asymmetry and transitivity, representing the **temporal precedence of cause-and-effect relationships**.



Causal Order

$$\mathbf{A \prec \{B, C\} \prec D}$$

Graphs from experts can have significant errors, even when the expert is "Perfect"!

- **Perfect Expert:** Outputs an edge if a directed path exists between two nodes

Theorem: Consider a procedure to estimate a graph by using pairwise queries to a Perfect Expert.
SHD of the estimated graph can have significant error, whereas causal order is always correct.

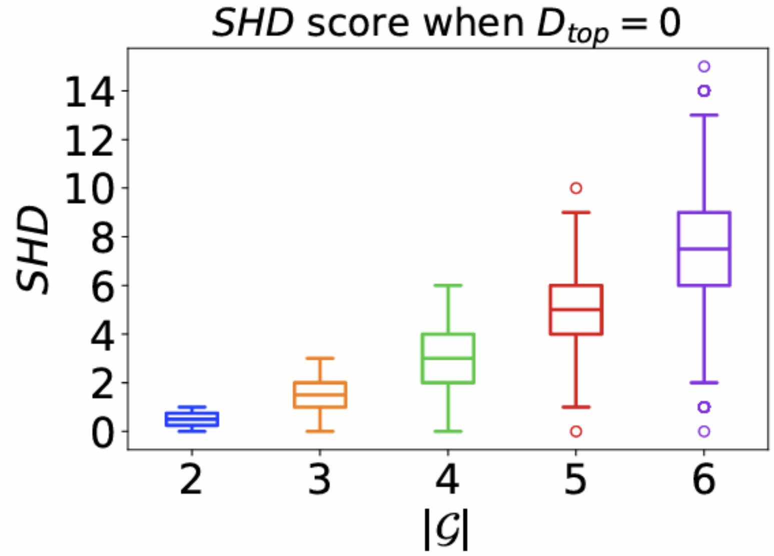


Figure 3: Variability of SHD for various graph sizes with $D_{top} = 0$ within each graph.

So, what’s the solution?

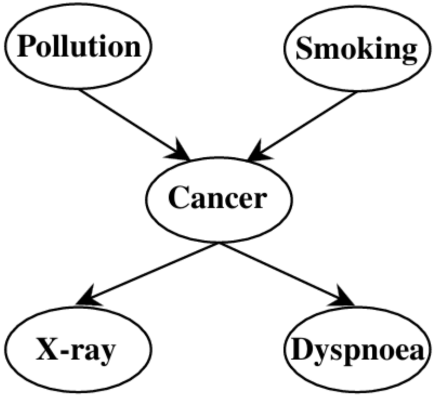
Use causal order.

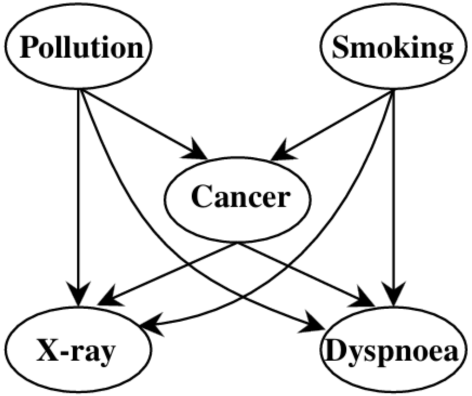
Causal order is a simpler construct, but still useful

For **causal effect estimation**, causal order is necessary and sufficient for estimating a valid backdoor set (under no unobserved confounding).

[consider nodes before treatment in causal order]

For **obtaining the graph**, we can use causal order as a prior or constraint for graph discovery algorithms.



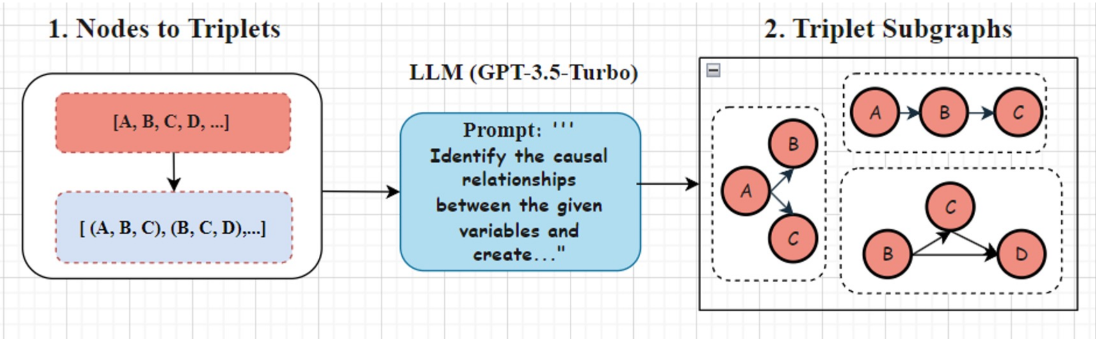


Causal Order:
{Smoking, Pollution} < Cancer < X-Ray <Dyspnoea

Causal Effect of Cancer on Dyspnoea:
Backdoor set: {Smoking, Pollution}

How to estimate causal order? **Triplet Method**

- Inspired by PC Algorithm, group nodes into **sets of three**
  - Boosts accuracy in identifying **direct and indirect effects**
- Utilize LLMs to orient edges in each triplet group
  - For each variable pair, obtain (n-2) edge predictions
- For each variable pair, decide final edge through **majority voting**
- Extract final **causal order** as **domain prior**



Triplet Approach to Infer Accurate Causal Order

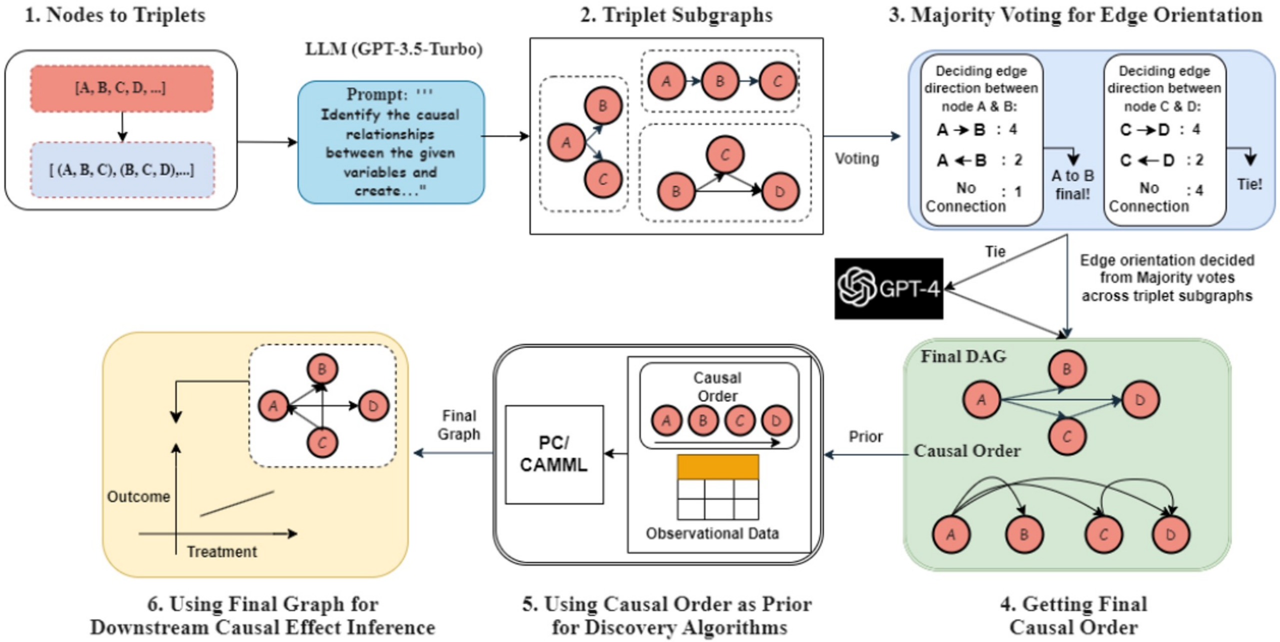


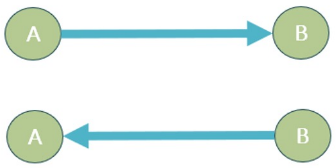
Figure 1: The *LLM-augmented* causal inference process based on inferring causal order. We propose a triplet-based prompting technique to infer all three-variable subgraphs and aggregate them using majority voting to produce a causal order. The causal order (optionally combined with discovery algorithms like PC or CaMML) can then be used to identify a valid back-door adjustment set. Ties in causal order are broken using GPT-4.

Evaluation: BnLearn and 3 recent real-world datasets

- Evaluation conducted on benchmark datasets from **BNLearn** repository across different scales of nodes
- Lesser known **Neuropathic pain diagnosis, Covid-19 and Alzheimers dataset.**

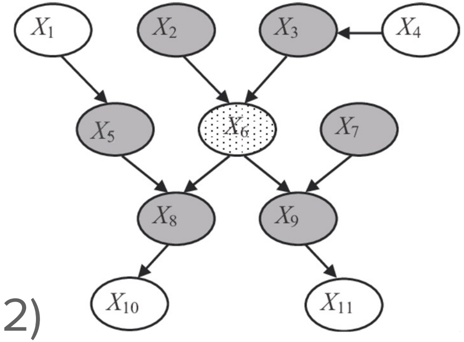
Dataset	Number of Nodes	Number of Edges	Description (used as a context)
Asia	8	8	Model the possible respiratory problems someone can have who has recently visited Asia and is experiencing shortness of breath
Cancer	5	4	Model the relation between various variables responsible for causing Cancer and its 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

Baseline methods based on pairwise prompt



Pairwise Base: Orienting edges between a given pair of nodes.

- **Iterative Context**: Previously oriented pairs added as context for next orientation



- **Markov Blanket**: Markov blanket of node pairs being evaluated as context
- **Chain of Thought Reasoning:** Applying Wei et al.'s (2022) approach, aligning LLMs in-context using real-world entity pairs with causal orientation and reasoning

CoT enhances pairwise accuracy, but Triplet prompt yields **significantly more accurate causal order.**

Triplet prompt avoids cycles.

Accuracy difference is higher for larger graphs.

Dataset	Metric	Pairwise (Base)	Pairwise (CoT)	Triplet
Using LLM				
Earthquake	D_{top}	**0**	**0**	**0**
	SHD	7	**4**	**4**
	Cycles	**0**	**0**	**0**
	IN/TN	**0/5**	**0/5**	**0/5**
Survey	D_{top}	3	1	**0**
	SHD	12	**9**	**9**
	Cycles	**0**	**0**	**0**
	IN/TN	**0/6**	2/6	**0/6**
Cancer	D_{top}	**0**	-	1
	SHD	6	-	**6**
	Cycles	0	-	**0**
	IN/TN	**0/5**	-	**0/5**
Asia-M	D_{top}	-	-	**1**
	SHD	15	13	**11**
	Cycles	7	1	**0**
	IN/TN	**0/7**	**0/7**	**0/7**
Child	D_{top}	-	-	**1**
	SHD	177	138	**28**
	Cycles	»3k	»500	**0**
	IN/TN	**0/20**	**0/20**	**0/20**
Covid	D_{top}	-	**0**	**0**
	SHD	41	**27**	30
	Cycles	»1000	**0**	**0**
	IN/TN	**0/20**	**0/20**	**0/20**
Alzheimers	D_{top}	-	6	**4**
	SHD	42	**26**	28
	Cycles	684	**0**	**0**
	IN/TN	**0/20**	**0/20**	**0/20**
Neuropathic	D_{top}	-	-	**3**
	SHD	212	64	**24**
	Cycles	»5k	5	**0**
	IN/TN	**0/22**	**0/22**	13/22

CoT enhances pairwise accuracy, but Triplet prompt yields **significantly more accurate causal order**

Triplet prompt obtains better accuracy even with smaller models like Phi-3 and Llama-3, *compared to pairwise with GPT-4.*

Dataset	Metric	Pairwise GPT-4	Triplet Phi-3	Triplet Llama3
Asia	D_{top}	1	**0**	2
	SHD	18	**13**	17
	Cycles	**0**	**0**	**0**
	IN/TN	**0/5**	1/5	**0/5**
Alzheimers	D_{top}	-	7	**5**
	Cycles	1	**0**	**0**
	IN/TN	**0/11**	**0/11**	1/11
Child	D_{top}	-	17	**12**
	SHD	148	**69**	129
	Cycles	»10k	**0**	**0**
	IN/TN	**0/20**	**0/20**	**0/20**

Table 2: Comparison of GPT-4 Pairwise Base with Phi-3/Llama3 using the Triplet method, showing how smaller models outperform GPT-4 by producing cycle-free graphs. This underscores the importance of the triplet strategy, regardless of the expert model used.

Causal Discovery: Triplet order output enhances accuracy of discovery algorithms, esp. in data-constrained settings

	Dataset	PC	SCORE	ICA LINGAM	Direct LINGAM	NOTEARS	CaMML	Ours (PC+LLM)	Ours (CaMML+LLM)
$N = 250$	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**
	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**
	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**	1.00±0.21
	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	0.00±1.00	**0.00±0.00**
	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**	1.21±0.30
	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
	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

Causal Effect Inference: Triplet order + graph discovery enhances accuracy of backdoor estimation algorithms.

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	$\mathbf{0.00 \pm 0.00}$

Conclusion: Don’t use pairwise prompts to infer edges

- Pairwise prompts are the dominant way to infer edges in causal graph.
- **But they have a fundamental flaw, even when querying with Perfect Experts.**
- Causal Order is a simpler and more robust structure, that is useful for downstream tasks such as discovery and effect inference.
- To estimate causal order, we introduce a Triplet Prompting method, surpassing various pairwise prompting baselines.
    - Triplet output boosts causal discovery algorithms and causal effect accuracy.

Backup

Causal order correlates with effect inference errors, unlike SHD, unsuitable for evaluating noisy expert causal inference

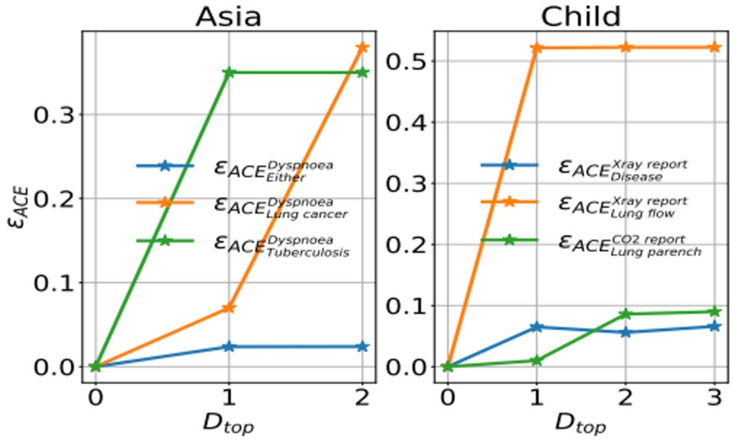
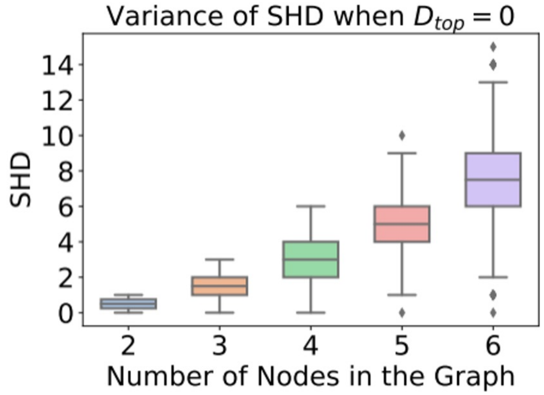


Figure 4. D_{top} vs. ϵ_{ACE}. ϵ_{ACE} increases as D_{top} increases, aligning with theoretical observations.

Cancer			
SHD vs. ϵ_{ACE} | $D_{top} = 0$		D_{top} vs. ϵ_{ACE} | $SHD = 2$	
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			
SHD vs. ϵ_{ACE} | $D_{top} = 0$		D_{top} vs. ϵ_{ACE} | $SHD = 3$	
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



How can Order based evaluation be done?

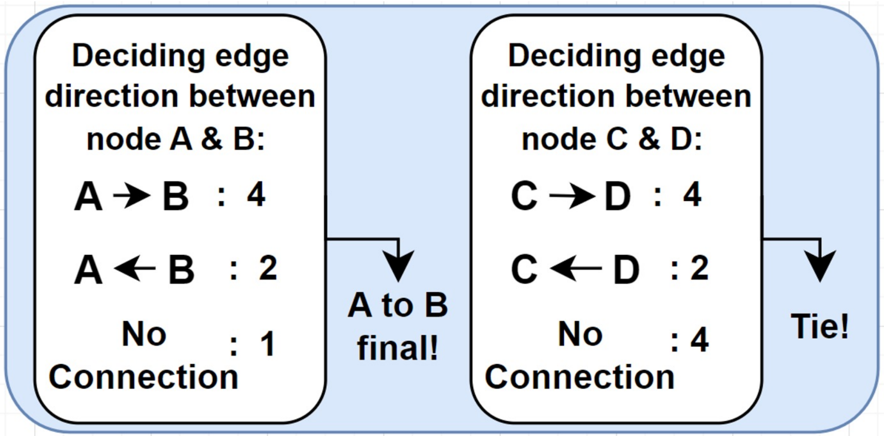
Topological Divergence (D_{Top}) helps compute the deviation in ordering of the predicted graph as compared to ground truth graph.

$$D_{top}(\hat{\pi}, A) = \sum_{i=1}^{n} \sum_{j:\hat{\pi}_i > \hat{\pi}_j} A_{ij}$$

- Recent studies in LLM based causal discovery employs **Structured Hamming Distance (SHD)** for edge analysis, capturing missing and falsely directed edges.

Understanding the Majority Voting System

- Iterate over all node pairs for orientation.
- Majority vote from triplets guide orientation.
- Expert intervention in case of tie-breaking conflicts among edge orientations.
- GPT-4 with CoT prompt used for Tie-Breaking



Best of Both Worlds - Merging LLMs and Discovery Algorithms

- LLMs exhibit unknown failure modes leading to inaccurate performance therefore making them unreliable.

  But how can their strengths be utilised while ensuring a principled approach to causal discovery?

- We propose pipelines for combining LLMs with causal discovery algorithms, by providing outputs of LLMs as priors. We focus on two types of discovery algorithms: **Constraint** and **Score** based methods.

Constraint Based Integration with LLMs

Causal order π from a triplet graph used for orienting undirected edges in Partial DAG.

To handle cases not covered by LLM output, we employ GPT-4 with CoT prompt for orientation.

Algorithm 1: Combining constraint based methods and experts to get $\hat{\pi}$ for a given set of variables.

1: **Input:** LLM topological ordering $\hat{\pi}$, Expert $\mathcal{E}_{GPT4}$, PC-CPDAG $\hat{\mathcal{G}}$
2: **Output:** Estimated topological order $\hat{\pi}_{\text{final}}$ of $\{X_1, \ldots, X_n\}$.
3: **for** $(i - j) \in$ undirected-edges$(\hat{\mathcal{G}})$ **do**
4: If both the node i and j are in $\hat{\pi}$ and if $\hat{\pi}_i < \hat{\pi}_j$, orient $(i - j)$ as $(i \rightarrow j)$ in $\hat{\mathcal{G}}$.
5: Otherwise, use the expert $\mathcal{E}_{GPT4}$ with CoT prompt to orient the edge $(i - j)$.
6: **end for**
7: $\hat{\pi}_{\text{final}} =$ topological ordering of $\hat{\mathcal{G}}$
8: return $\hat{\pi}$

Score Based Integration with LLMs

Level order of causal graph returned by LLM is used as prior for CamML

Algorithm 2: Combining score based methods and experts to get $\hat{\pi}$ for a given set of variables.

1: **Input:** $\mathcal{D}$, variables $\{X_1, \ldots, X_n\}$, Expert $\mathcal{E}$, Score based method $\mathcal{S}$, *Prior* probability p.
2: **Output:** Estimated topological order $\hat{\pi}$ of $\{X_1, \ldots, X_n\}$.
3: Step (I) $\hat{\mathcal{G}} = \mathcal{E}(X_1, \ldots, X_n)$
4: Step (II) *Prior* = level order traversal of $\hat{\mathcal{G}}$.
5: Step (II.I) If $\hat{\mathcal{G}}$ is cyclic, keep all the variables in a cycle at the same level in *Prior*.
6: Step (III) $\hat{\mathcal{G}} = \mathcal{S}(\mathcal{D}, \textit{Prior}, \textit{Prior probability} = p)$
7: Step (IV) $\hat{\pi} =$ topological ordering of $\hat{\mathcal{G}}$
8: return $\hat{\pi}$