Improving constraint-based discovery with robust propagation and LLM priors

Ruiqi Lyu[*]
Carnegie Mellon University

Alistair Turcan[*]
Carnegie Mellon University

Martin Jinye Zhang[†]
Carnegie Mellon University

Bryan Wilder[†]
Carnegie Mellon University

Abstract

Constraint-based causal discovery recovers causal DAG structure from conditional independence (CI) relations. Classical methods such as PC orient v-structures first, then propagate edge directions from these seeds, relying on accurate CI tests and rich separating-set searches. In practice, these conditions often fail, causing cascading orientation errors. Recent work uses large language models (LLMs) as experts to augment edge orientation when standard assumptions fail, but often treats LLM outputs as reliable or assumes stable error behavior, despite hallucinations and instability. We propose MosaCD, a constraint-based framework that robustly combines CI tests with LLM inference to obtain high-confidence orientation seeds, then propagates them with Seeded Propagation Rules (SPR), which mitigate the fragility of collider-first orientation. We prove oracle-level soundness for SPR under explicit assumptions on the skeleton, separating-set record, and seed orientations, and give a stylized finite-sample analysis showing why prioritizing non-collider evidence can reduce orientation errors. Across 13 real-world benchmarks, MosaCD and SPR achieve substantially higher accuracy than existing methods through more reliable seeds and more robust propagation.

1 Introduction

Causal discovery aims to recover cause-effect graphs over variables. A prominent family, originating with PC [1], is constraint-based discovery: CI tests first remove edges, then orientation rules direct the remaining ones. Orientation starts from seed edges induced by v-structures and iteratively propagates directions. These methods are widely used for their guarantees, flexibility across data types, and interpretable outputs [1]. However, errors can accumulate at every step. Theory assumes CI tests perfectly distinguish dependence from independence across conditioning subsets [1, 2], but finite-sample CI tests are noisy and exhaustive subset search is infeasible. Orientation can amplify these errors: seed directions depend on fine-grained conclusions from separating sets found during skeleton search. For an unshielded triple $X - Z - Y$, collider orientation requires the recorded minimal separating sets for (X, Y) to consistently exclude Z; if this record is incomplete or corrupted by finite-sample CI errors, a false collider can be introduced and propagated.

Recent advances in LLMs create new opportunities for seeding edge orientations. Despite their imperfections, LLMs encode broad knowledge useful for inferring pairwise causal relationships [3, 4]. Prior work prompts LLMs with causal queries, e.g., "Does A cause B?" [3, 5] or "Is A conditionally independent of B given $C, D, \ldots$?" [6], or uses LLM outputs as causal-order priors [4] or enforced

[*]Equal first-author contribution.
[†]Equal last-author contribution.

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

constraints [7]. However, most LLM-based causal discovery keeps the LLM and algorithm separate: LLM outputs are fed into existing algorithms as priors [7], or standard algorithms like PC are run first and the LLM is used for post hoc orientation [4, 8]. In both cases, the discovery algorithm remains unchanged.

In this work, we study how causal discovery algorithms can directly integrate external domain knowledge. We focus on two challenges: (i) generating useful orientation information without human experts, and (ii) propagating it to recover causal relations in the DAG. We propose MosaCD, a constraint-based causal discovery algorithm that combines CI tests with LLM-based inference to obtain high-quality orientation seeds, then propagates them using Seeded Propagation Rules (SPR), a new set of orientation rules that leverage reliable initial orientations unavailable to classical PC-style algorithms. For robust LLM seeding, we use randomized prompting to reduce variance and mitigate hallucinations. Theoretically and empirically, we show that with high-quality seeds, SPR's non-collider-first strategy yields more robust graph recovery than classical collider-first constraint-based methods. Across 13 causal discovery benchmarks, MosaCD achieves powerful constraint-based performance through more reliable seeding and propagation.

Our contributions are:

1. We develop MosaCD, a constraint-based causal discovery algorithm that robustly integrates domain knowledge by combining a reliable LLM-based seeding strategy with a tailored non-collider-first orientation strategy, termed Seeded Propagation Rules (SPR).
2. We provide oracle-level soundness guarantees for SPR under explicit assumptions, and a stylized analysis showing why non-collider-first propagation is less exposed to finite-sample CI errors than collider-first propagation.
3. We conduct extensive empirical evaluations across 13 real-world benchmark datasets spanning diverse settings, demonstrating consistent performance improvements from both the LLM-based seeding step and the proposed propagation strategy.

2 Related Work

Constraint-based causal discovery. Constraint-based causal discovery infers edges and orientations from CI relations and logical rules. PC [1] removes edges via CI tests and orients the remainder with logical rules, assuming no hidden confounders. FCI [9] extends PC to latent confounders and selection bias, returning PAGs (partial ancestral graphs) rather than CPDAGs. PC-stable [10] removes PC's order dependence by fixing adjacency sets at each conditioning-set size. Conservative PC (CPC) [11] reduces false colliders by requiring unanimity across observed separating sets before orienting an edge, while still assuming those sets are comprehensive. Post-hoc consistency methods revisit CI tests to reconcile a PDAG with skeleton evidence. PC-max [12] avoids contradictions by emphasizing conditioning sets with the most significant p-values. These methods still rely on v-structures to seed orientations, then propagate the rest of the graph by avoiding new v-structures.

Incorporating domain knowledge into constraint-based causal discovery. Constraint-based methods can naturally incorporate domain constraints or priors. Tiered orders and path constraints can prune orientations after PC or its variants, then be closed under orientation rules [13]. However, this still inherits the v-structure-first bias because domain knowledge is applied after initial collider orientations. Beyond this paradigm, Hyttinen et al. [14] encode tested (in)dependences and prior knowledge as logical constraints and minimize the total weight of violations, but the problem is NP-hard in the worst case and degrades on dense graphs. Claassen and Heskes [15] assign Bayesian reliabilities to (in)dependence claims and process them in decreasing reliability, returning one model with confidence tags per decision, but require enumerating parts of the Markov equivalence class, which is computationally expensive. Other causal discovery paradigms integrate domain knowledge differently, such as through a global optimization objective.

LLMs for causal discovery. Recent work shows that LLMs can provide useful domain knowledge for causal discovery, but it remains difficult to distinguish accurate causal information from plausible errors. Some methods directly query LLMs for causal constraints or full graphs, e.g., asking "Does A cause B?" [3, 5, 16] or "Is A conditionally independent of B given $\{C, D, \ldots\}$?" [6]. Others use LLM outputs as prior selectors [4, 17–19] or discovery constraints [20]. While effective in some settings, many treat direct LLM outputs as reliable or assume stable error behavior, without explicitly

addressing hallucination and positional bias. In contrast, MosaCD uses customized querying and aggregation to identify and filter unreliable LLM responses before causal discovery. Appendix R discusses related LLM methods and assumptions in detail.

Hallucination detection in LLMs. LLMs often fail to acknowledge uncertainty: when unsure on multiple-choice questions, they may select the first option rather than "I don't know." Empirical studies report positional and presentation biases and reluctance to admit uncertainty; shuffle-and-vote mitigations help but do not eliminate the need for calibration [21, 22]. Hallucination-detection methods often require direct model access [23] or fine-tuning [24]. Cheaper prompt-only methods typically rely on self-reflection [25] or LLM-generated confidence scores [26], requiring the model to reason about its own errors. Causal discovery methods that extract LLM priors would benefit from scalable, prompt-only hallucination filtering that requires no calibration.

3 Preliminaries

Notations. Let $G = (V, E)$ be the (unknown) ground-truth DAG, where V is the set of observed variables and E is the edge set. For disjoint $X, Y, S \subseteq V$, write $X \perp Y \mid S$ if S d-separates X and Y in G, and $X \not\perp Y \mid S$ otherwise. We use $X - Y$ for an undirected edge, $X \to Y$ for a directed edge, and $X \to \cdots \to Y$ for a directed path.

We use curly braces $\{X, Y, \ldots\}$ to denote an unordered node set. A triple $X - Z - Y$ is *unshielded* if X and Y are non-adjacent but both are adjacent to Z. A partially directed graph (PDAG) is a graph whose edges may be directed or undirected; throughout, an acyclic PDAG means one with no directed cycle.

PC-style skeleton search (PC, CPC, PC-stable). The procedure [1] starts from the complete undirected graph K_V and removes an unordered edge between X and Y whenever a conditional independence (CI) test accepts $X \perp Y \mid S$ for some $S \subseteq V \setminus \{X, Y\}$. The resulting undirected graph is $\widehat{Skel} = (V, \hat{E})$. For each removed edge $\{X, Y\} \notin \hat{E}$, we maintain a recorded minimal separating-set collection $\Sigma_{\min}(X, Y) = \Sigma_{\min}(Y, X) \subseteq 2^{V \setminus \{X,Y\}}$. Operationally, $\Sigma_{\min}(X, Y)$ contains accepted separating sets of minimum cardinality among the sets tested for (X, Y); equivalently, accepted supersets of already accepted separating sets are not used for orientation. Each recorded set S is stored with its CI-test p-value $p_{X,Y,S}$; when a score is needed, write $p_{\max}(X, Y) = \max_{S \in \Sigma_{\min}(X,Y)} p_{X,Y,S}$. PC and PC-stable typically retain one such set per removed edge, whereas CPC records multiple sets at the first separating level.

Intuition of MosaCD and SPR. Given a skeleton, PC-style algorithms need seed orientations before further propagation, e.g., via Meek's rules [13]. They obtain these seeds from unshielded colliders: for each unshielded triple $X - Z - Y$, if $Z \notin S$ for every $S \in \Sigma_{\min}(X, Y)$, they orient $X \to Z \leftarrow Y$. Conversely, if $Z \in S$ for every $S \in \Sigma_{\min}(X, Y)$, the recorded minimal separators support treating Z as a non-collider; if both separator types appear, conservative procedures leave the triple unresolved. With an oracle and complete minimal-separator records, collider and non-collider decisions follow the same separator criterion. In finite samples, however, CI tests can falsely accept independence, and level-wise skeleton search tests many sets excluding Z. A false accepted separator excluding Z can create spurious collider evidence, which is especially harmful as collider orientations seed propagation. SPR instead propagates first from external seed orientations and prioritizes non-collider evidence from $\Sigma_{\min}(X, Y)$ before applying collider orientations. This lets external domain knowledge resolve directions in non-collider configurations that PC alone cannot distinguish.

4 Method

MosaCD takes as input a dataset $\mathcal{D}$, variables V, and known edge orientations $\mathcal{O}$, by default generated by an LLM reasoning over variables and CI tests. Its core SPR stage outputs a PDAG; an optional vote-completion step orients remaining reversible edges into a DAG for evaluation.The oracle guarantee in Section 5.1 concerns only the SPR PDAG, not this optional completion. Stage-wise results appear in Figure 8 (Appendix L.1). MosaCD has four steps. First, it builds an undirected skeleton using a constraint-based method, e.g., PC, CPC, or PC-stable, producing G_{skel} and recorded minimal separating sets $\Sigma_{\min}$ with CI p-values. Second, it applies domain knowledge $\mathcal{O}$ to seed skeleton edges; $\mathcal{O}$ can be LLM-generated from variable names, descriptions, and $\Sigma_{\min}(X, Y)$, with shuffled

answer order and repeated queries to reduce positional bias and hallucination. Third, SPR propagates orientations across the seeded graph. For unshielded triples $X - Z - Y$, SPR uses $\Sigma_{\min}(X, Y)$ to prioritize *non-collider* evidence, where Z appears in every recorded minimal separator, before collider evidence, where Z appears in none. SPR then resolves some remaining undirected edges by choosing the orientation with the fewest conflicts with the recorded CI evidence $\Sigma_{\min}$, leaving ties undirected. Finally, optional vote completion orients leftover undirected edges using a topological order from Step 2 LLM votes.

Step 1: Skeleton search. MosaCD initializes a skeleton G_{skel} and a recorded minimal separating-set collection $\Sigma_{\min}$ from dataset $\mathcal{D}$ using a constraint-based algorithm such as PC, CPC, or PC-stable. The PDAG P is initialized as G_{skel}. For each removed edge $\{X, Y\}$, $\Sigma_{\min}(X, Y)$ contains one or more minimum-cardinality accepted separating sets, together with the corresponding CI p-values. PC and PC-stable usually contribute a single recorded separator, while CPC may contribute several. All SPR orientation decisions below use $\Sigma_{\min}(X, Y)$.

Step 2: Domain knowledge orientation seeding. Given known orientations $\mathcal{O}$ in the skeleton, orient all such edges in the skeleton following this knowledge.

2.1 **(Default) Generate domain knowledge from LLMs** If seed orientations $\mathcal{O}$ are unavailable, we generate high-confidence seeds for undirected edges by querying an LLM. For each undirected edge $X - Y$, we provide variable names, descriptions, and $\Sigma_{\min}(X, Y)$ with CI p-values. To reduce positional bias and hallucination, we randomize candidate answer order, e.g., "$X \to Y$ or $Y \to X$" and its reverse, and repeat each query five times. We retain edges with consistent answers, defined as the majority vote under both answer orders. We discard any proposed seed $X \to Y$ that (i) contradicts separator evidence in $\Sigma_{\min}$ at any unshielded triple or (ii) creates a directed cycle in P.

Step 3: Apply SPR to propagate edges from seeds.

1. **Iterative orientation propagation.** SPR first rejects the candidate if it would create a directed cycle, then repeats the following steps until convergence.

   1.1 **Unsupervised acyclic propagation:** Before CI evidence is certified, apply the directed-path acyclicity rule: if $X - Y$ and $X \to \cdots \to Y$ in P, then orient $X \to Y$.

   1.2 **CI-supervised propagation:** For each unshielded and partially ordered triple $X \to Z - Y$, sort in descending order by $\max p$ in $\Sigma_{\min}(X, Y)$, since larger p-values provide stronger evidence for conditional independence. Orient $Z \to Y$ if Z appears in every recorded minimal separator in $\Sigma_{\min}(X, Y)$, and orient $Y \to Z$ if Z appears in none. Prioritizing by $\max p$ ensures that orientations with stronger CI support are applied first.

   1.3 **Collider orientation:** For each unshielded and unordered triple $X - Z - Y$, again sort by descending $\max p$ in $\Sigma_{\min}(X, Y)$. Orient $X \to Z \leftarrow Y$ if Z appears in no recorded minimal separator in $\Sigma_{\min}(X, Y)$.

   1.4 **Certified Meek closure:** Apply a Meek rule only when every unshielded collider/non-collider fact needed by its premise has been certified by $\Sigma_{\min}$. Reject orientations that would create a directed cycle.

2. **Least-conflict orientation.** SPR can optionally resolve remaining undirected edges by choosing the direction least conflicting with recorded conditional independences in $\Sigma_{\min}$. For each undirected pair $X - Y$ in random order, we consider $X \to Y$ and $Y \to X$. For each candidate, we close the graph under the orientation rules and count local separator-triple contradictions in $\Sigma_{\min}$; candidates creating a directed cycle are rejected. We choose the lower-conflict direction and leave $X - Y$ undirected on ties. This finite-sample step is inactive in the oracle setting of Theorem 5.1: rule closure has already exhausted all sound compelled orientations, so least-conflict adds none. *Example.* After SPR Step 1, suppose $U \to Y$, $V \to Y$, $W \to Y$, $X - Y$ is undirected, and the recorded CI evidence includes $\{Y\} \in \Sigma_{\min}(X, U)$, $\{Y\} \in \Sigma_{\min}(X, V)$, and $\emptyset \in \Sigma_{\min}(X, W)$. Then $X \to Y$ creates the colliders $X \to Y \leftarrow U$ and $X \to Y \leftarrow V$ (2 conflicts), whereas $Y \to X$ makes $W \to Y \to X$ induce $X \not\perp W$ despite $\emptyset \in \Sigma_{\min}(X, W)$ (1 conflict); by the least-conflict step, we choose $Y \to X$.

(Optional) Step 4: Final orientation via votes. If edges remain undirected, we use the Step 2 LLM votes to complete the DAG, analogous to Vashishtha et al. [4], but with shuffled answer orders to reduce positional bias. We aggregate votes into a weighted directed graph and remove the weakest edges, defined by net directional support, to break cycles. We then derive a topological order and orient any remaining undirected edges accordingly, producing a final DAG for evaluation. This optional step is outside the oracle CPDAG/PDAG guarantee in Section 5.1 and is used only when a fully oriented DAG is required.

5 Theoretical Analysis

We start by analyzing the oracle behavior of SPR. The goal of this result is narrower than full finite-sample correctness: it shows that, when the skeleton, recorded minimal separators, and seed orientations are correct and mutually compatible, SPR's rule-closure stage does not introduce orientations that violate these constraints. We distinguish the ordinary CPDAG from the PDAG obtained after adding background orientation constraints, since external seed orientations may select a subset of the Markov equivalence class. We assume causal sufficiency, Markovness, faithfulness, a perfect CI oracle, and a seed oracle that returns only orientations compatible with the target background-knowledge class, though it may abstain. We then examine departures from these assumptions, particularly the noisiness in CI tests that motivates SPR. To complement our empirical results, we provide a theoretical analysis in a stylized model, demonstrating that orienting non-colliders first (as SPR does) yields fewer errors than orienting colliders first (as in PC algorithms).

5.1 Correctness

We distinguish two objects. First, the CPDAG of a DAG G is the unique PDAG representing the Markov equivalence class $[G]$: it has the same skeleton and unshielded colliders as G, directed edges are those compelled in every DAG in $[G]$, and undirected edges are reversible within $[G]$ [27, Theorem 4.1]. Second, when background orientation constraints are provided, let $\mathcal{K}$ denote the set of accepted seed orientations and define $[G]_{\mathcal{K}} = \{H \in [G] : H \text{ satisfies every orientation in } \mathcal{K}\}$. When $[G]_{\mathcal{K}} \neq \emptyset$, the $\mathcal{K}$-compatible essential graph is the maximally oriented PDAG representing this restricted class: an edge is directed iff it has the same orientation in every DAG in $[G]_{\mathcal{K}}$, and is left undirected otherwise. For $\mathcal{K} = \emptyset$, this reduces to the ordinary CPDAG.

Theorem 5.1. *Let G be a causally sufficient DAG over V, and let $\mathcal{K}$ be a set of background orientation constraints on edges in the true skeleton. Assume: (i)* ***Markov and faithfulness:*** *the observational distribution is Markov and faithful to G; (ii)* ***Perfect CI oracle and skeleton consistency:*** *skeleton search returns $\widehat{Skel} = \mathrm{Skel}(G)$ and the recorded minimal separator collection $\Sigma_{\min}$ is correct; (iii)* ***Compatible seeds:*** *the seed set $E_{\text{seed}} = \mathcal{K}$ is acyclic and compatible with at least one DAG in $[G]$, i.e., $[G]_{\mathcal{K}} \neq \emptyset$; (iv)* ***Sound propagation rules:*** *SPR Step 1 applies only acyclicity-preserving orientation rules and CI-guarded collider/non-collider implications derived from $\Sigma_{\min}$. Then every orientation made by SPR Step 1 is sound with respect to the restricted equivalence class $[G]_{\mathcal{K}}$: after each orientation, at least one DAG in $[G]_{\mathcal{K}}$ remains consistent with the current PDAG, the recorded minimal separators, and the seed constraints. The resulting PDAG is an acyclic PDAG with skeleton $\mathrm{Skel}(G)$ satisfying the recorded separator constraints and the seed constraints $\mathcal{K}$. If $\mathcal{K} = \emptyset$, then under the same oracle and faithfulness assumptions, SPR Step 1 reduces to the usual PC-style collider orientation followed by orientation-rule closure, and returns the ordinary CPDAG of G; in this oracle-faithful setting its closure coincides with PC, PC-stable, and CPC.*

See Appendix M for the proof.

Remark 5.2. Theorem 5.1 is an oracle-level statement about the SPR rule-closure stage. It does not assume that LLM-provided seeds are always correct in finite samples, and it does not cover the optional vote-completion step that orients the remaining reversible edges into a DAG. If retained seeds are incorrect, exact recovery of the true CPDAG cannot be guaranteed in general; our finite-sample robustness to such errors is evaluated empirically in Figure 2 and Appendix K.

5.2 Prioritizing non-collider over collider identification improves accuracy

For an unshielded triple $X - Z - Y$, PC-style collider orientation checks whether Z is absent from the recorded minimal separating sets $\Sigma_{\min}(X, Y)$. SPR instead first uses non-collider evidence, namely

cases where Z appears in every recorded minimal separator in $\Sigma_{\min}(X, Y)$, before applying collider orientations. With a complete and correct minimal-separator record, collider and non-collider status are both determined by $\Sigma_{\min}(X, Y)$. However, under noisy CI tests with finite samples, collider-first procedures have more opportunities to accept a spurious separator that excludes Z, thereby creating an erroneous early collider seed. We analyze this effect in a stylized level-wise search model.

Level-wise search and error events. Let $\ell = |C|$ denote the conditioning-set size. The search proceeds by levels $\ell = 0, 1, 2, \ldots$, testing $X \perp Y \mid C$ over all $C \subseteq V \setminus \{X, Y\}$ with $|C| = \ell$. At level ℓ, if any candidate C is accepted as a sepset, the search stops; we declare Z a collider if $Z \notin C$, and a non-collider if $Z \in C$. Accordingly, a collider error occurs if Z is a non-collider but $Z \notin C$, and a non-collider error occurs if Z is a collider but $Z \in C$. We measure their relative frequency via

$$\mathcal{R}_\ell := \frac{\Pr(\text{collider error at level } \ell)}{\Pr(\text{non-collider error at level } \ell)}. \tag{1}$$

At level ℓ, PC uses the **first** accepted sepset C for (X, Y). PC-stable has the same collider/non-collider decision as PC, but with adjacency sets frozen within level ℓ (order-invariant), so $\mathcal{R}_\ell$ matches PC. At the first level ℓ where independence holds, CPC gathers all minimal sepsets for (X, Y) and orient the collider iff Z is in **none**, treat as non-collider iff Z is in **all**, otherwise leave the triple unoriented.

We compute $\mathcal{R}_\ell$ for the PC (the same as that of PC-stable) and CPC rules (denoted $\mathcal{R}_\ell^{\mathrm{PC}}$ and $\mathcal{R}_\ell^{\mathrm{CPC}}$). PC-stable has the same $\mathcal{R}_\ell$ as PC, since it only removes within-level order dependence. Whenever $\mathcal{R}_\ell > 1$ for a given rule set, tests of non-colliders (prioritized by SPR) will have a lower error rate than test of colliders (prioritized by existing algorithms). We make the following assumptions to make theoretical analysis tractable, noting they are not required for running SPR in practice.

Assumption 5.3. (Simple CI test model) Conditional independence (CI) tests act independently across candidates/levels given truth labels, with false positive rate α and false negative rate β that do not vary with ℓ or C. Thus, a true sepset rejects dependence with probability $1 - \beta$, while a non-sepset does so with probability α.

Assumption 5.4. (Z controls the $X - Y$ path) (a) All $X - Y$ paths of length at most $2\ell + 1$ pass through Z; (b) whether the $X - Y$ path is open is determined fully by whether Z is conditioned on.

Similar to related work [11, 4], these assumptions are deliberately simplified to make analysis tractable. Here, they let us illustrate the core dynamic that non-collider evidence is less error-prone by isolating the impact of the single node Z (intuitively, Assumption 5.4 describes a locally sparse graph without redundant $X - Y$ paths) and imposing a single set of parameters describing the performance of the CI test. They should not be read as conditions required for running MosaCD or as a complete model of finite-sample CI testing. Assumption 5.3 replaces heterogeneous and correlated CI-test behavior with a homogeneous independent-error model in order to isolate the orientation-order effect. Assumption 5.4 is a local structural simplification rather than a claim that all real graphs have a single $X - Y$ path. Across the 13 datasets, the local condition holds for 71% of relevant triples on average. Performance remains strong even when the condition holds less often; for example, it holds for 42% of relevant triples in Child, where MosaCD still achieves the best F1 score (Table 1).

In this model, we obtain exact analytical expressions for R_ℓ^{PC} and R_ℓ^{CPC}, derived and shown in the appendix. These expressions involve a number of combinatorial quantities, but in the asymptotic regime where the error rates of the CI test are small relative to the graph size (a necessary condition for the algorithm to not be overwhelmed with errors), we can further simplify and show that R_ℓ^{PC} and R_ℓ^{CPC} must be strictly above 1.

Theorem 5.5. *Let $M = |V \setminus \{X, Y\}|$. Suppose that $\alpha, \beta = o\left(\frac{1}{M}\right)$ and $\ell = \Theta(1)$. For M sufficiently large compared to ℓ, the error ratios satisfy*

$$R_\ell^{\mathrm{CPC}} = \beta^{\binom{M-1}{\ell-1} - \binom{M-1}{\ell}} \cdot \frac{M - \ell}{\ell} + o\left(\frac{1}{M}\right) > 1$$

$$R_\ell^{\mathrm{PC}} = \frac{n(n+1)}{m(m+1)}(1 + o(1)) = \left(\frac{M - \ell}{\ell}\right)^2 (1 + o(1)) > 1,$$

where $m = \binom{M-1}{\ell-1}, n = \binom{M-1}{\ell}$. Thus, for both algorithms, the error rate among colliders will be higher than noncolliders.

See Appendix N for the proof. In this stylized noisy regime, the intuition is that there are more candidate subsets that do *not* contain Z than subsets that do contain Z, so collider-first strategies have more opportunities to make a mistake at a reached level. SPR flips the ordering, starting with candidates that do contain Z, since this set is smaller and the total number of mistakes made in early stages will be limited. Numerical experiments (Appendix O) also show substantially lower FPRs under the non-collider first strategy across three standard skeleton learners, consistent with the theory.

6 Experimental results

		F1/SHD						
Group	Method	Cancer (5 nodes)	Asia (8 nodes)	Covid Resp. (11 nodes)	ALZ (11 nodes)	Child (20 nodes)	Insurance (27 nodes)	Water (32 nodes)
Non-LLM	PC	0.50/2	0.67/3	0.31/18	0.49/13	0.70/11	0.62/28	0.45/53
	Meek	0.50/2	0.67/3	0.32/**16**	0.53/**11**	**0.78/6**	**0.70/19**	**0.57/38**
	Shapley-PC	**1.00/0**	0.53/4	0.28/26	0.31/25	0.67/19	0.67/28	0.47/53
	NOTEARS	0.00/4	0.57/5	0.45/**16**	0.07/19	0.48/17	0.15/50	0.29/54
	DAG-GNN	0.00/4	0.27/9	**0.46**/25	**0.68**/16	0.24/23	0.20/60	0.26/55
	GES	0.50/2	0.12/9	0.33/18	0.42/25	0.17/63	0.26/105	0.29/71
	HC	0.50/2	0.12/9	0.33/18	0.53/17	0.33/51	0.33/95	0.36/60
	MMHC	**1.00/0**	**0.93/1**	0.32/19	0.07/19	0.60/12	0.61/23	0.29/71
	SPR (Ours)	0.50/2	0.67/3	0.31/18	0.47/12	0.76/7	0.67/22	0.45/51
Oracle	SPR + one oracle edge	0.60/1	0.71/2	0.31/18	0.52/11	0.79/7	0.74/17	0.51/46
LLM	ILS-CSL*	0.50/2	**0.93/1**	0.31/18	0.49/13	0.83/6	0.70/25	0.60/38
	SCP*	0.50/2	0.67/3	0.39/15	0.53/11	0.78/6	0.68/21	0.57/38
	Jiralerspong*	0.57/3	0.55/5	0.25/18	0.48/15	0.10/38	0.21/52	0.00/71
	Vashishtha*	0.75/1	**0.93/1**	0.65/11	0.48/16	0.67/9	0.70/21	0.65/34
	Causal Disco*	0.57/2	0.86/2	**0.72/6**	0.68/10	0.56/13	0.34/41	**0.87/15**
	LLM-CD*	**1.00/0**	0.80/2	0.39/15	0.49/13	0.40/30	0.29/59	0.24/81
	Memorization*	0.80/2	0.78/4	0.55/14	0.55/12	0.32/26	0.35/47	0.53/42
	MosaCD* (Ours)	**1.00/0**	**0.93/1**	0.65/11	**0.73/8**	**0.90/2**	**0.87/12**	0.59/36

Group	Method	Mildew (35 nodes)	Alarm (37 nodes)	Hailfinder (56 nodes)	Covid. Comp. (63 nodes)	Hepar2 (70 nodes)	Win95pts (76 nodes)
Non-LLM	PC	0.63/23	0.85/12	0.38/105	0.23/152	0.36/100	0.59/68
	Meek	0.69/**17**	**0.90/8**	0.40/97	0.28/**134**	0.39/87	0.64/58
	Shapley-PC	**0.75**/19	0.84/14	0.38/105	0.18/174	0.44/89	0.65/70
	NOTEARS	0.25/44	0.25/40	0.27/**58**	0.13/144	0.06/119	0.14/111
	DAG-GNN	0.43/50	0.53/35	0.13/77	0.07/648	0.13/119	0.37/93
	GES	0.23/152	0.26/101	0.33/150	**0.29**/302	0.45/122	0.51/122
	HC	0.24/140	0.30/92	0.31/162	**0.29**/300	0.47/119	0.46/133
	MMHC	0.52/27	0.68/16	**0.59**/72	0.17/156	**0.67/55**	**0.68/50**
	SPR (Ours)	0.64/21	0.87/10	0.40/100	0.22/151	0.37/91	0.61/63
Oracle	SPR + one oracle edge	0.86/10	0.87/10	0.43/95	0.23/148	0.48/80	0.73/50
LLM	ILS-CSL*	0.89/10	0.85/12	0.44/97	0.00/138	0.54/78	0.69/54
	SCP*	0.69/17	0.87/8	0.39/99	0.27/129	0.38/88	0.63/56
	Jiralerspong*	0.07/50	0.09/80	0.10/72	0.03/145	0.16/137	0.03/135
	Vashishtha*	0.78/13	0.72/15	**0.60**/77	0.27/131	0.66/59	0.74/46
	Causal Disco*	0.61/24	0.51/27	0.33/**53**	**0.39**/120	**0.81/36**	0.55/69
	LLM-CD*	0.24/52	0.30/59	0.22/92	0.27/129	0.32/117	0.17/160
	Memorization*	0.24/59	0.29/46	0.17/79	0.17/156	0.12/147	0.00/182
	MosaCD* (Ours)	**0.90/8**	**0.93/6**	0.49/90	**0.39/118**	0.72/52	**0.81/37**

Table 1: **Benchmark evaluation.** Each entry reports F1/SHD for each dataset and method (nodes in brackets). Best F1 and SHD is **bold** and second-best is underlined within the LLM and non-LLM groups. Oracle rows receive one ground-truth edge and are separate diagnostics, excluded from non-oracle rankings.

We use 13 benchmark datasets: 10 from BNLearn [28]—Cancer, Asia, Child, Insurance, Water, Mildew, Alarm, Hailfinder, Hepar2, and Win95pts—and 3 from Srivastava et al. [29]: Covid Resp., ALZ, and Covid. Comp., which were released after the GPT-4o-mini training cutoff used in our experiments and were designed to reduce benchmark memorization. To reduce memorization, we replace original variable names with human-readable versions less easily linked to the datasets, e.g., "HrglssDrtnAftrPrnt" becomes "Hourglass Duration Too Long." The datasets span 5–76 nodes and

include real and simulated graphs with categorical and continuous data. For simulated datasets, we generate 20,000 samples each. We evaluate true edge orientation recovery using F1 score and structural Hamming distance (SHD). Additional 211-node MACH and 283-node BEELINE evaluations are reported separately in Appendix L.3.

Baselines. We use three skeleton methods—PC [1], PC-stable [10], and CPC [11]—all compatible with our orientation strategies. Given a skeleton, we evaluate 15 baselines: (i) **PC**, using the corresponding skeleton method's standard orientation procedure; (ii) **Meek** [13], applying Meek's rules after PC; (iii) **Shapley-PC** [30], using Shapley-value feature importance; (iv) **ILS-CSL** [31], an LLM method incorporating statistical knowledge; (v) **SCP** [6], an LLM orientation method; (vi) **Jiralerspong** [16], a pairwise LLM method prompting on two nodes; (vii) **Vashishtha** [4], a pairwise LLM causal-ordering method given the full graph; (viii) **Causal Disco** [18], using LLM information as a noise-robust prior; (ix) **LLM-CD** [19], also using noisy LLM priors; (x–xiv) non-LLM methods, **NOTEARS** [32], **DAG-GNN** [33], **GES** [34], **Hill climbing (HC)** [35], and **MMHC** [36], implemented in gCastle [37] and run independently of the skeleton, except MMHC, initialized with the same skeleton as MosaCD; and (xv) **Memorization**, which prompts the LLM to return a causal graph from all variables. These cover state-of-the-art LLM and non-LLM constraint-based, score-based, and continuous-optimization methods. All methods access the dataset; LLM methods also receive identical variable names, metadata, and GPT-4o-mini backbone (Appendix L.6). We compare **MosaCD**, the full pipeline, and **SPR**, propagation without domain knowledge. We separately report **SPR + one oracle edge**, propagation with one random ground-truth orientation averaged over 100 replicates, as an oracle diagnostic. Matched oracle-edge controls appear in Appendix L.5. For fairness, LLM and non-LLM methods are compared only within groups. See Appendix P for details.

6.1 Benchmarking experiments

We applied MosaCD, SPR, and baselines to 13 benchmark datasets. Table 1 reports results using PC skeletons; PC-Stable and CPC results appear in Appendix B. We draw two main conclusions. First, MosaCD broadly outperforms LLM baselines, achieving the best performance on 9 of 13 datasets and the best overall results: average F1 of 0.76 versus 0.66 for the next-best method, Vashishtha, and average SHD of 29 versus 32 for the next-best method, Causal Disco. This includes the three datasets released after the LLM training cutoff. Second, the oracle diagnostic SPR + one oracle edge improves over Meek on 7 datasets and ties on at least one metric on 2 more; this comparison includes extra ground-truth information, so matched-access results are reported separately in Appendix L.5. Adding just one edge can sharply improve SPR by reducing reliance on collider orientations; for example, in Mildew, F1 increases by 0.22. Consistent with our theory, PC skeleton orientations identify undirected edges with 92.2% accuracy, while collider orientations are correct only 73.1% of the time. MosaCD similarly outperforms baselines with PC-Stable and CPC skeletons (Appendix B).

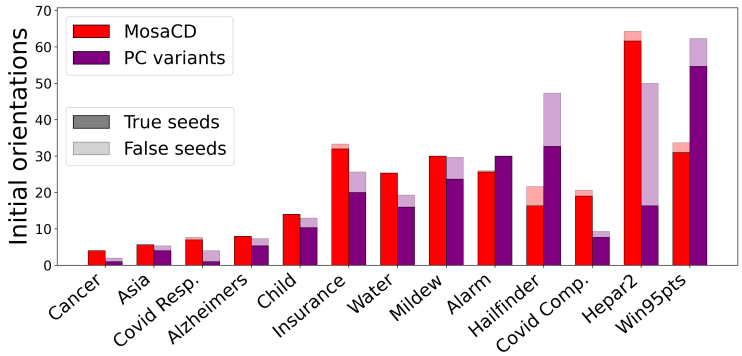


Figure 1: **Accuracy of orientation seeds.** Number of true and false directions found by MosaCD LLM-based orientation seeding (Step 2) and standard PC v-structure orientation, averaged for each dataset across PC, PC-stable, and CPC skeletons.

We further evaluated the effectiveness of MosaCD's LLM-based orientation seeding (Step 2) compared to the standard PC procedure (orienting v-structures). Results are reported in Figure 1, averaged across PC, PC-Stable, and CPC skeletons. First, MosaCD's seeding procedure identified substantially

more true directions (avg MosaCD vs. PC true seed ratio 2.15) and markedly fewer false directions (avg 5.0% vs. 27.1%) than PC across the 13 datasets. Second, in the 4 largest datasets where the memorization baseline underperformed, MosaCD remained robust: it did not introduce excessive false seeds compared to PC (average 11.2% vs. 32.1% across such datasets), likely due to the hallucination filtering procedure. Full results, cost analysis, and additional ablations appear in Appendices C–I.

6.2 Secondary analyses and ablation studies

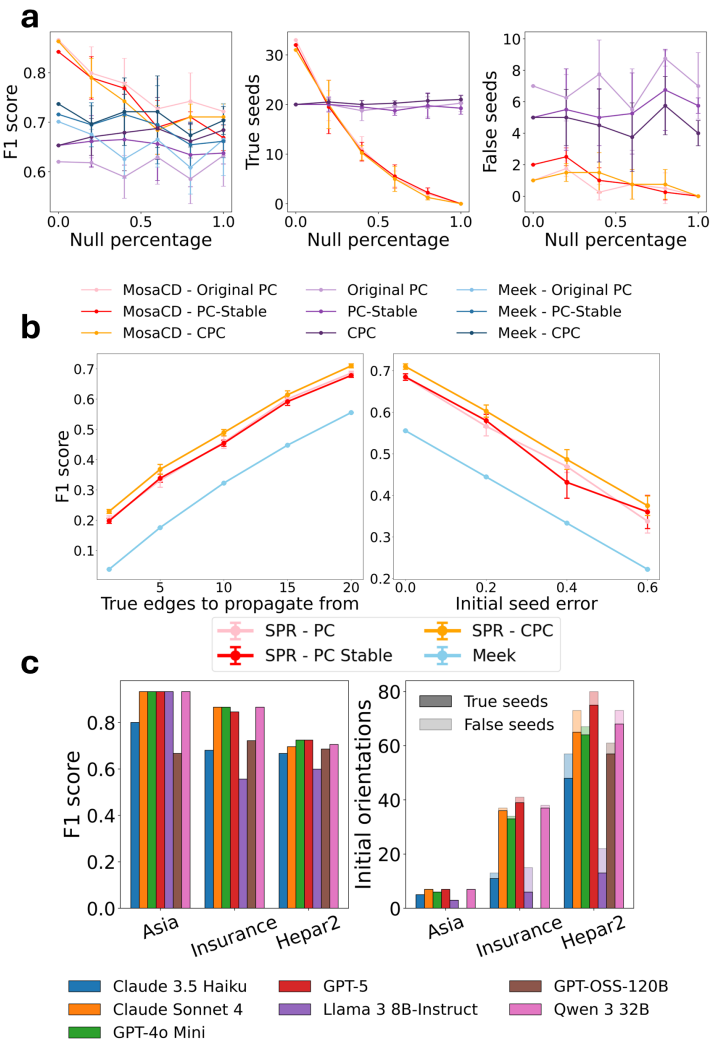


Figure 2: **Secondary analysis. (a)** F1 score, true seeds, and false seeds for MosaCD, PC, and Meek as uninformative variable descriptions increase, using PC, PC-Stable, and CPC skeletons on Insurance. **(b)** F1 score for SPR and Meek on Insurance across PC, PC-Stable, and CPC skeletons: left, varying true seeds with no false seeds; right, varying false-seed proportion with 20 total seeds. **(c)** F1 scores and LLM correctness across seven LLMs on Asia, Insurance, and Hepar2 using PC skeletons.

First, we assessed the robustness of MosaCD's LLM inference by replacing a proportion of variable names and descriptions in the "Insurance" dataset with uninformative ones. Results are reported in Figure 2a. While MosaCD's performance and the number of true orientation seeds declined expectedly as the proportion of uninformative variables increased, it consistently generated only a small number of false seeds and remained competitive or superior overall to the baseline, demonstrating robustness and consistent with SPR's theory. Experiments in Asia and Hepar2 yield similar results (Appendix J). Graph-independent aliases (Figure 9) and post-cutoff comparisons appear in Appendix L.2.

Second, we evaluated the robustness of SPR through ablation studies on the "Insurance" dataset. We vary the amount and accuracy of external domain knowledge to propagate from, comparing against Meek's propagation rules starting from the same seeds. Results are reported in Figure 2b. SPR consistently outperformed Meek both when varying the number of true seeds (with false seeds fixed at 0) and when varying the proportion of false seeds (with total seeds fixed at 20), demonstrating its effectiveness in utilizing such information. We repeat this analysis in the Asia and Hepar2 datasets and achieve similar results (Appendix K).

Third, we varied the LLM backbone in MosaCD across efficient models (Claude-3.5-Haiku, GPT-4o-mini), frontier reasoning models (GPT-5, Claude-Sonnet-4), and open-source models (GPT-oss-120b, Llama 3 8B-instruct, Qwen 3 32B) on 3 representative datasets: Asia, Insurance, and Hepar2. As shown in Figure 2c, MosaCD remains consistent across LLMs, except for slightly weaker performance with GPT-oss-120b and Llama 3 8B-instruct. A seven-sample-size study with learned-skeleton ceilings appears in Figure 10 (Appendix L.4).

References

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

[2] Markus Kalisch and Peter Bühlman. Estimating high-dimensional directed acyclic graphs with the pc-algorithm. *Journal of Machine Learning Research*, 8(3), 2007.

[3] Emre Kiciman, Robert Ness, Amit Sharma, and Chenhao Tan. Causal reasoning and large language models: Opening a new frontier for causality. *Transactions on Machine Learning Research*, 2023.

[4] Aniket Vashishtha, Gowtham Reddy Abbavaram, Abhinav Kumar, Saketh Bachu, Vineeth N Balasubramanian, and Amit Sharma. Causal order: The key to leveraging imperfect experts in causal inference. 2025.

[5] Aniket Vashishtha, Abbavaram Gowtham Reddy, Abhinav Kumar, Saketh Bachu, Vineeth N Balasubramanian, and Amit Sharma. Causal inference using llm-guided discovery. *arXiv preprint arXiv:2310.15117*, 2023.

[6] Kai-Hendrik Cohrs, Gherardo Varando, Emiliano Diaz, Vasileios Sitokonstantinou, and Gustau Camps-Valls. Large language models for constrained-based causal discovery. *arXiv preprint arXiv:2406.07378*, 2024.

[7] Uzma Hasan and Md Osman Gani. Optimizing data-driven causal discovery using knowledge-guided search. *arXiv preprint arXiv:2304.05493*, 2023.

[8] 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.

[9] Peter L Spirtes, Christopher Meek, and Thomas S Richardson. Causal inference in the presence of latent variables and selection bias. *arXiv preprint arXiv:1302.4983*, 2013.

[10] Diego Colombo, Marloes H Maathuis, et al. Order-independent constraint-based causal structure learning. *J. Mach. Learn. Res.*, 15(1):3741–3782, 2014.

[11] Joseph Ramsey, Jiji Zhang, and Peter L Spirtes. Adjacency-faithfulness and conservative causal inference. *arXiv preprint arXiv:1206.6843*, 2012.

[12] Joseph Ramsey. Improving accuracy and scalability of the pc algorithm by maximizing p-value. *arXiv preprint arXiv:1610.00378*, 2016.

[13] Christopher Meek. Causal inference and causal explanation with background knowledge. *arXiv preprint arXiv:1302.4972*, 2013.

[14] Antti Hyttinen, Frederick Eberhardt, and Matti Järvisalo. Constraint-based causal discovery: Conflict resolution with answer set programming. In *UAI*, pages 340–349, 2014.

[15] Tom Claassen and Tom Heskes. A bayesian approach to constraint based causal inference. *arXiv preprint arXiv:1210.4866*, 2012.

[16] Thomas Jiralerspong, Xiaoyin Chen, Yash More, Vedant Shah, and Yoshua Bengio. Efficient causal graph discovery using large language models, 2024. URL https://arxiv.org/abs/2402.01207.

[17] Alex Havrilla, David Alvarez-Melis, and Nicolo Fusi. Igda: Interactive graph discovery through large language model agents. *arXiv preprint arXiv:2502.17189*, 2025.

[18] Stephanie Long, Alexandre Piché, Valentina Zantedeschi, Tibor Schuster, and Alexandre Drouin. Causal discovery with language models as imperfect experts. *URL https://arxiv.org/abs/2307.02390*, 2023.

[19] Taiyu Ban, Lyuzhou Chen, Derui Lyu, Xiangyu Wang, Qinrui Zhu, Qiang Tu, and Huanhuan Chen. Integrating large language model for improved causal discovery. *IEEE Transactions on Artificial Intelligence*, 2025.

[20] 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.

[21] Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Qi Liu, Tianyu Liu, and Zhifang Sui. Large language models are not fair evaluators. *arXiv preprint arXiv:2305.17926*, 2023.

[22] Pouya Pezeshkpour and Estevam Hruschka. Large language models sensitivity to the order of options in multiple-choice questions. *arXiv preprint arXiv:2308.11483*, 2023.

[23] Sebastian Farquhar, Jannik Kossen, Lukas Kuhn, et al. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630, 2024. doi: 10.1038/s41586-024-07421-0. URL https://doi.org/10.1038/s41586-024-07421-0.

[24] Hanning Zhang, Shizhe Diao, Yong Lin, Yi R. Fung, Qing Lian, Xingyao Wang, Yangyi Chen, Heng Ji, and Tong Zhang. R-tuning: Instructing large language models to say 'i don't know', 2024. URL https://arxiv.org/abs/2311.09677.

[25] Potsawee Manakul, Adian Liusie, and Mark J. F. Gales. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models, 2023. URL https://arxiv.org/abs/2303.08896.

[26] Xinran Zhao, Hongming Zhang, Xiaoman Pan, Wenlin Yao, Dong Yu, Tongshuang Wu, and Jianshu Chen. Fact-and-reflection (far) improves confidence calibration of large language models, 2024. URL https://arxiv.org/abs/2402.17124.

[27] Steen A Andersson, David Madigan, and Michael D Perlman. A characterization of markov equivalence classes for acyclic digraphs. *The Annals of Statistics*, 25(2):505–541, 1997.

[28] Marco Scutari and Jean-Baptiste Denis. *Bayesian Networks: With Examples in R*. Chapman & Hall/CRC Texts in Statistical Science, Taylor & Francis, 2014. ISBN 978-1-4665-7378-5.

[29] Ashutosh Srivastava, Lokesh Nagalapatti, Gautam Jajoo, Aniket Vashishtha, Parameswari Krishnamurthy, and Amit Sharma. Realizing llms' causal potential requires science-grounded, novel benchmarks. *arXiv preprint arXiv:2510.16530*, 2025.

[30] Fabrizio Russo and Francesca Toni. Shapley-pc: Constraint-based causal structure learning with shapley values. *arXiv preprint arXiv:2312.11582*, 2023.

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

[32] Xun Zheng, Bryon Aragam, Pradeep Ravikumar, and Eric P. Xing. Dags with no tears: Continuous optimization for structure learning, 2018. URL https://arxiv.org/abs/1803.01422.

[33] Yue Yu, Jie Chen, Tian Gao, and Mo Yu. Dag-gnn: Dag structure learning with graph neural networks. In *International conference on machine learning*, pages 7154–7163. PMLR, 2019.

[34] David Maxwell Chickering. Optimal structure identification with greedy search. *Journal of machine learning research*, 3(Nov):507–554, 2002.

[35] David Heckerman, Dan Geiger, and David M Chickering. Learning bayesian networks: The combination of knowledge and statistical data. *Machine learning*, 20(3):197–243, 1995.

[36] Ioannis Tsamardinos, Laura E Brown, and Constantin F Aliferis. The max-min hill-climbing bayesian network structure learning algorithm. *Machine learning*, 65(1):31–78, 2006.

[37] Keli Zhang, Shengyu Zhu, Marcus Kalander, Ignavier Ng, Junjian Ye, Zhitang Chen, and Lujia Pan. gcastle: A python toolbox for causal discovery. *arXiv preprint arXiv:2111.15155*, 2021.

[38] Katharine Sink and Tom Brikowski. MACH: A multi-attribute catchment hydrometeorological dataset. *Scientific Data*, 13:816, 2026. doi: 10.1038/s41597-026-07162-x. URL `https://www.nature.com/articles/s41597-026-07162-x`.

[39] Aditya Pratapa, Amogh P. Jalihal, Jeffrey N. Law, Aditya Bharadwaj, and T. M. Murali. Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data. *Nature Methods*, 17:147–154, 2020. doi: 10.1038/s41592-019-0690-6. URL `https://www.nature.com/articles/s41592-019-0690-6`.

A Limitations

MosaCD inherits several limitations from constraint-based causal discovery. Its theoretical guarantees rely on standard assumptions such as causal sufficiency, Markovness, faithfulness, an accurate skeleton, correct separating-set records, and seed orientations that are compatible with the target equivalence class. In finite samples, violations of these assumptions can lead to incorrect skeleton edges, missing edges, or erroneous separating sets; while SPR is designed to reduce the impact of fragile collider-first propagation, it cannot fully correct errors introduced earlier in skeleton discovery.

The method also depends on the quality of external orientation information. Although our randomized prompting and filtering reduce positional bias and hallucinated LLM seeds, LLM-provided domain knowledge may still be wrong, sensitive to variable descriptions, affected by memorization, or inconsistent across model versions. In domains where variable names are uninformative, mechanisms are highly specialized, or causal relations are counterintuitive, MosaCD may provide fewer useful seeds and fall back closer to standard constraint-based behavior.

Our empirical evaluation covers 13 benchmark datasets with up to 76 nodes, additional 211-node MACH and 283-node BEELINE tasks, and a larger runtime study, but this does not exhaust the range of real-world causal discovery settings. In particular, we do not evaluate latent confounding, selection bias, feedback cycles, interventions, time-varying systems, or very high-dimensional graphs. Scaling also remains partly limited by the skeleton-search step and by the number of LLM queries, although the latter is easily parallelized. Finally, the optional vote-completion step produces a fully oriented DAG for evaluation, but it is not covered by the oracle SPR guarantee and should be interpreted as a pragmatic completion heuristic rather than a theoretically guaranteed recovery step.

B Results with different PC skeletons

	F1/SHD						
	Cancer (5)	Asia (8)	Covid Resp. (11)	Alzheimers (11)	Child (20)	Insurance (27)	Water (32)
PC-Stable	0.50/2	0.67/3	0.32/17	0.44/14	0.70/11	0.63/28	0.48/51
CPC	0.50/2	0.75/4	0.32/17	0.47/14	0.90/3	0.63/28	0.47/52
PC-Stable + Meek	0.50/2	0.67/3	0.40/14	0.48/12	0.78/6	0.69/20	0.58/39
CPC + Meek	0.50/2	0.93/1	0.40/14	0.55/11	0.90/3	0.72/19	0.57/40
PC-Stable + MosaCD	1.00/0	0.93/1	0.65/11	0.73/8	0.90/3	0.84/13	0.65/11
CPC + MosaCD	1.00/0	0.93/1	0.69/10	0.73/8	0.94/2	0.86/12	0.61/36

	Mildew (35)	Alarm (37)	Hailfinder (56)	Covid. Comp. (63)	Hepar2 (70)	Win95pts (76)
PC-Stable	0.69/21	0.85/12	0.38/101	0.18/146	0.40/92	0.64/64
CPC	0.64/23	0.85/12	0.38/108	0.17/150	0.42/89	0.64/64
PC-Stable + Meek	0.73/16	0.90/8	0.40/98	0.27/123	0.43/82	0.69/53
CPC + Meek	0.71/17	0.90/8	0.39/99	0.27/124	0.44/83	0.69/53
PC-Stable + MosaCD	0.87/10	0.93/5	0.56/81	0.39/118	0.72/53	0.74/42
CPC + MosaCD	0.87/10	0.91/6	0.47/91	0.40/117	0.73/52	0.79/37

Table 2: Benchmark evaluation. Each entry reports F1/SHD for each dataset and method (nodes in brackets).

C Orientation correctness by skeleton

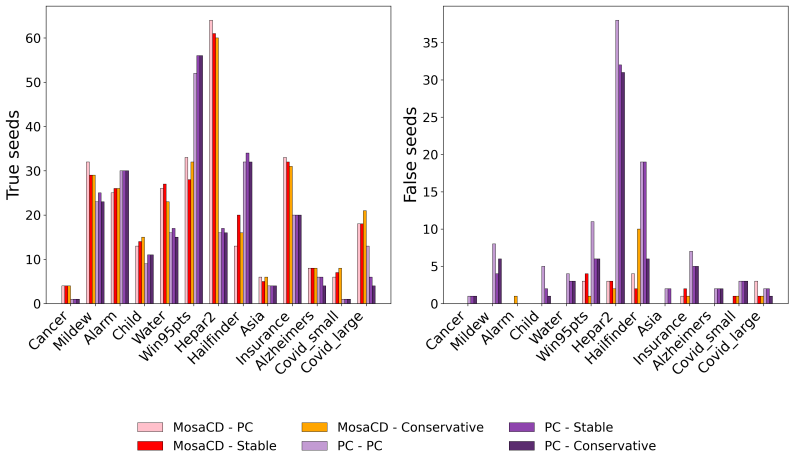


Figure 3: **Initial orientation correctness.** The average number of true and false seeds discovered by MosaCD and PC variants in each dataset, aggregated across PC, PC-stable, and CPC.

D Positional bias

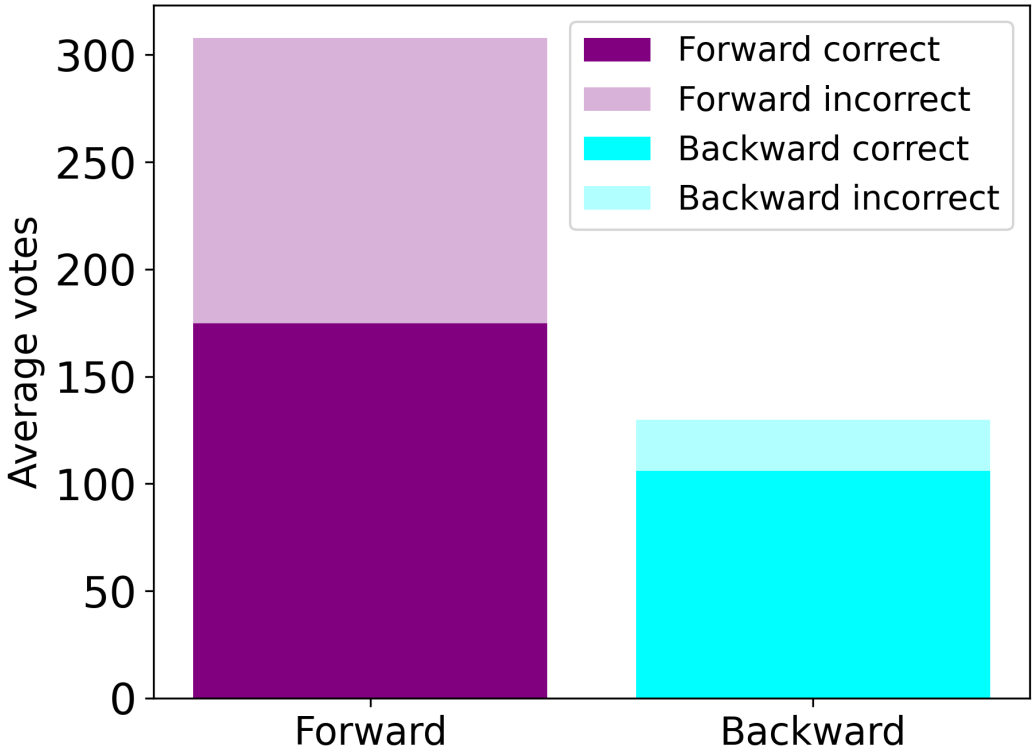


Figure 4: **Average count of total votes in each direction for a given prompt.** Forward appears before backward, and thus is chosen more frequently by the LLM. Shaded bars indicate votes in the wrong direction. Of note, I don't know was the first option, which was never selected.

E Runtime

Here we retain the original stage-timing measurements in seconds for skeleton search, LLM voting, and propagation. These measurements are separate from the comparative resource analyses in Appendix L.6; accumulated query time and parallel wall-clock time should not be interchanged. The original larger runtime case, Andes (223 nodes), required 15 minutes of skeleton searching and 120 minutes of LLM voting. Reported end-to-end timings and API costs for the additional MACH and BEELINE tasks appear in Appendix L.3.

	Asia (8 nodes)			**Insurance** (27 nodes)			**Hepar2** (70 nodes)		
	PC	PC-Stable	CPC	PC	PC-Stable	CPC	PC	PC-Stable	CPC
Skeleton	0.23	0.50	0.55	50.32	76.51	75.89	187.05	313.21	264.50
LLM	256.82	232.13	264.14	1501.94	1397.03	1434.65	2492.51	2453.28	2260.18
Propagation	0.00	0.00	0.00	0.01	0.01	0.01	0.03	0.03	0.03

Table 3: **Runtime analysis.** Runtime (seconds) of each stage of MosaCD for each dataset and PC variant.

F Ablating least-conflict orientation

Dataset	MosaCD	Randomized conflict
Cancer	1.00	1.00
Asia	0.93	0.93
Covid Resp.	0.65	0.52
Alzheimer's	0.73	0.73
Child	0.90	0.86
Insurance	0.87	0.87
Water	0.59	0.59
Mildew	0.90	0.90
Alarm	0.93	0.93
Hailfinder	0.49	0.47
Covid. Comp.	0.39	0.39
Hepar2	0.72	0.72
Win95pts	0.81	0.77

Table 4: F1 scores for the conflict-resolution ablation comparing MosaCD with a randomized conflict tie-breaker.

G Ablating variable information

Dataset	Names + descriptions	Names only
Cancer	1.00	1.00
Asia	0.93	0.93
Covid Resp.	0.65	0.65
Alzheimer's	0.73	0.73
Child	0.90	0.83
Insurance	0.87	0.87
Water	0.59	0.57
Mildew	0.90	0.89
Alarm	0.93	0.93
Hailfinder	0.49	0.44
Covid. Comp.	0.39	0.39
Hepar2	0.72	0.72
Win95pts	0.81	0.74

Table 5: F1 scores for the variable-description ablation, when MosaCD sees variable names and descriptions, or variable names only.

H Ablating logprob-based confidence

Dataset	MosaCD	Logprobs
Cancer	1.00	1.00
Asia	0.93	0.93
Covid Resp.	0.65	0.32
Alzheimer's	0.73	0.53
Child	0.90	0.56
Insurance	0.87	0.61
Water	0.59	0.53
Mildew	0.90	0.30
Alarm	0.93	0.33
Hailfinder	0.49	0.33
Covid. Comp.	0.39	0.27
Hepar2	0.72	0.54
Win95pts	0.81	0.63

Table 6: F1 scores for the logprob-based confidence ablation, comparing MosaCD with a variant using logprobs instead of randomized shuffling and voting.

I Varying sample size

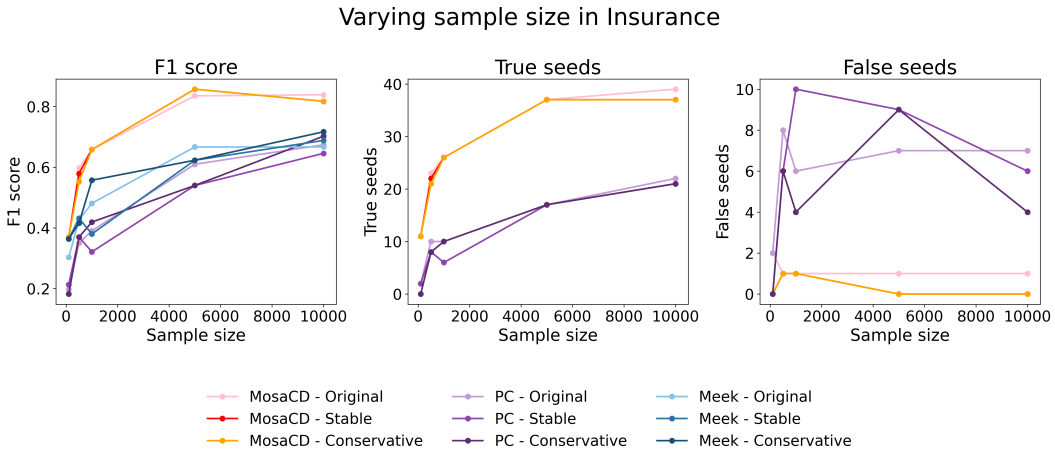


Figure 5: **Varying sample size.** Results showing F1 score, true seed counts, and false seed counts for MosaCD, PC, and Meek's rules for sample size from 100 to 10000.

J Additional null simulations

Dataset	PC	Meek	MosaCD (0% null)	MosaCD (40% null)	MosaCD (100% null)
Asia	0.67	0.67	0.93	0.81	0.67
Hepar2	0.36	0.39	0.72	0.60	0.42

Table 7: F1 scores for the null information experiment on Asia and Hepar2.

K Propagation ablations

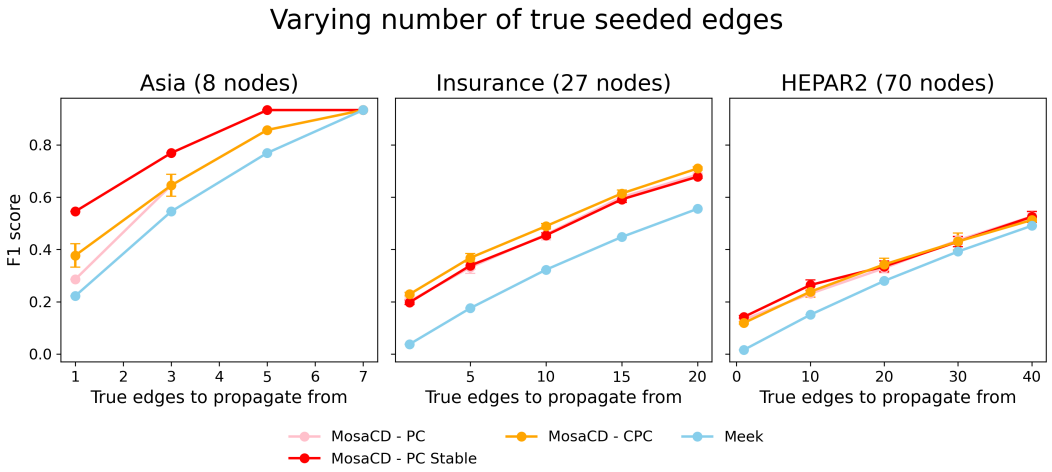


Figure 6: **Experiments varying number of true seeds.** F1 score for SPR (using PC, PC-Stable, and CPC skeletons) and Meek.

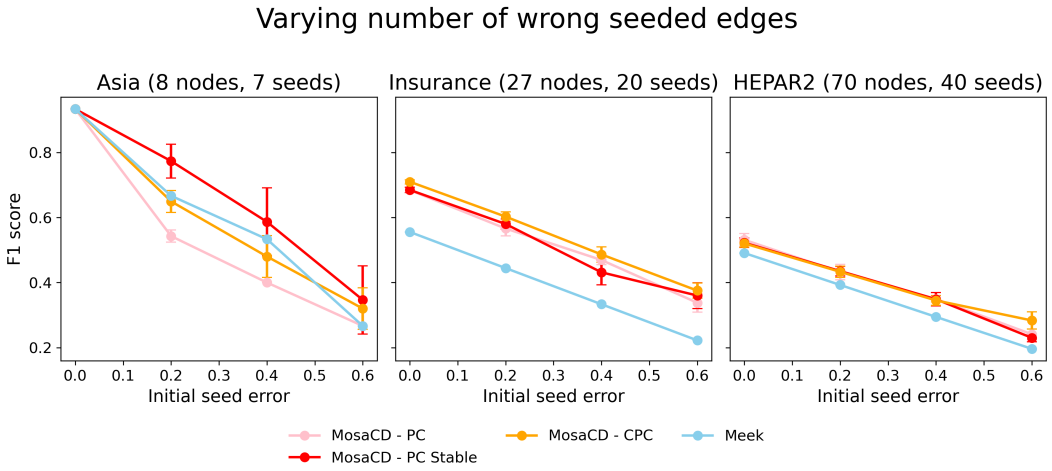


Figure 7: **Experiments varying number of false seeds.** F1 score for SPR (using PC, PC-Stable, and CPC skeletons) and Meek.

L Additional experiments and reproducibility details

L.1 Separating propagation from optional completion

Table 8 and Figure 8 distinguish accepted LLM seeds, SPR closure, least-conflict orientation, and optional vote completion. Holding the accepted seeds fixed, SPR increases reported F1 from .75 to .93 on Asia, .78 to .85 on Insurance, and .66 to .69 on Hepar2. Optional completion then adds .00, .02, and .03, respectively. These empirical gains do not extend the oracle guarantee of Theorem 5.1 to incorrect seeds or to completed DAGs. Observational CI information identifies an equivalence class under the stated assumptions; accepted external seeds restrict that class, and completion chooses residual directions for full-DAG evaluation.

Stage	Asia	Insurance	Hepar2
Skeleton	.38	.36	.32
+ cross-order consistency	.75	.78	.67
+ CI/cycle filtering	.75	.78	.66
+ SPR closure	.93	.85	.69
+ least-conflict	.93	.85	.69
+ optional vote completion	.93	.87	.72

Table 8: **Stage-wise reported F1.** Each row is the output after the named stage on Asia, Insurance, or Hepar2. The SPR and least-conflict rows precede optional full-DAG completion.

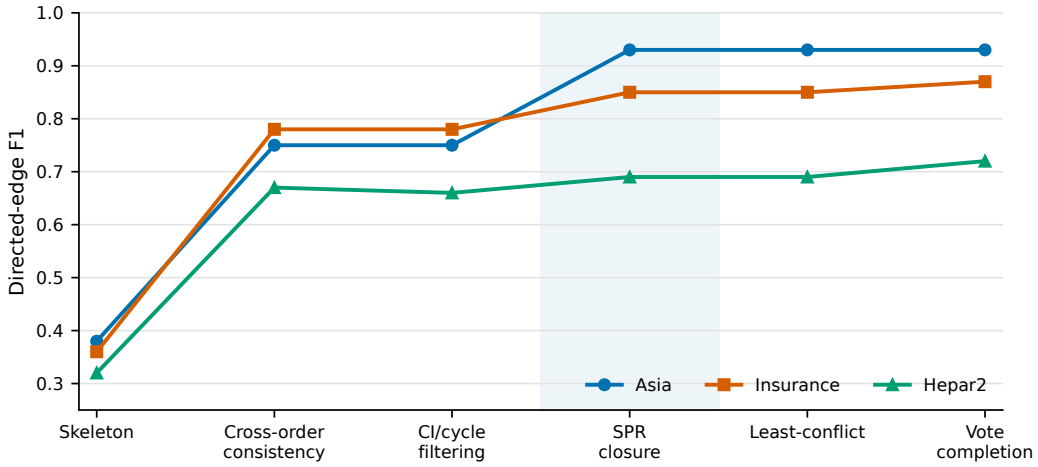


Figure 8: **Contribution of propagation and completion.** Reported F1 at successive pipeline stages for Asia (circles), Insurance (squares), and Hepar2 (triangles), using the numerical values in Table 8. The shaded band marks SPR closure. The comparison after CI/cycle filtering holds the accepted seed set fixed; least-conflict orientation does not change F1 in these three cases. Lines connect stage measurements and do not imply a continuous or monotone process.

L.2 Graph-independent aliases and benchmark familiarity

We replace names with graph-independent animal, color, or atom aliases, using five replicates of each alias type on Asia, Insurance, and Hepar2 and applying the same intervention to every LLM baseline. Table 9 and Figure 9 report the original-label and alias results. MosaCD falls toward the PC/Meek/SPR range when useful semantics are removed. The null-information experiment in Appendix J is a separate diagnostic and is not pooled with these alias measurements.

The three post-cutoff datasets in the original benchmark are Covid Resp., ALZ, and Covid Comp. MACH adds a fourth post-cutoff evaluation. Table 10 compares these tasks with whole-graph LLM controls and PC-family methods.

Method	Asia	Insurance	Hepar2
MosaCD	.93 → .67	.87 → .65	.72 → .36
Memorization	.71 → .22	.27 → .05	.11 → .05
SCP	.67 → .67	.68 → .60	.36 → .25
Vashishtha	.93 → .54	.77 → .30	.70 → .25
Jiralerspong	.46 → .22	.16 → .05	.16 → .05
LLM-CD	.94 → .62	.22 → .22	.18 → .16
Causal Disco	.86 → .46	.65 → .22	.92 → .25
ILS-CSL	.67 → .62	.76 → .58	.51 → .36
PC / Meek / SPR	.58 / .67 / .67	.53 / .60 / .65	.26 / .36 / .37

Table 9: **Reported F1 with graph-independent aliases.** Each arrow gives original-label F1 followed by random-alias F1. The final row gives the statistical controls reported for this experiment. The alias measurements are separate from the main benchmark and null-information runs; their baseline values should not be interchanged. Five replicates per alias type were performed.

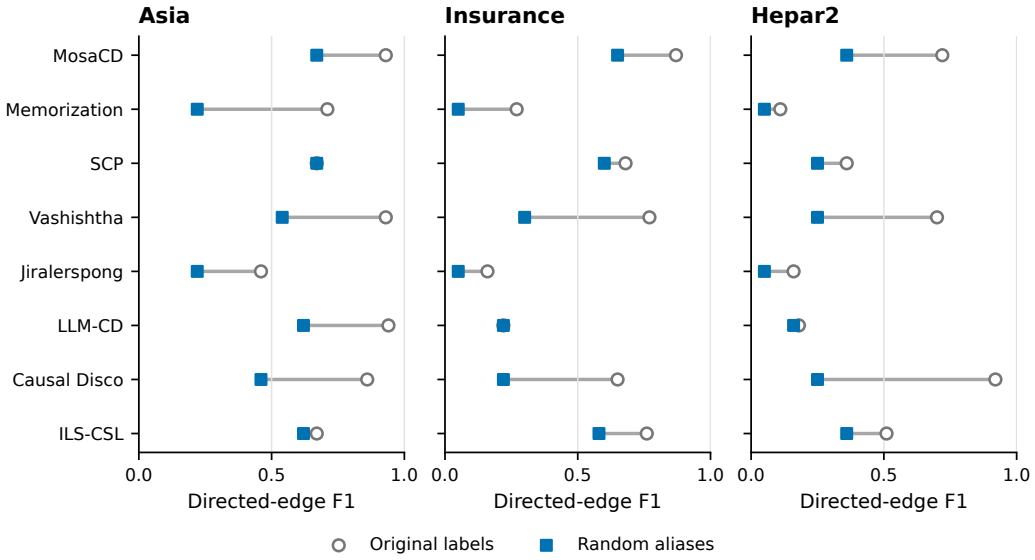


Figure 9: **Effect of removing informative variable names.** Open circles denote original-label F1 and blue squares denote random-alias F1 for each method, using Table 9. Each line connects the two reported aggregate values for one method and dataset. Aliases are readable animal, color, or atom labels assigned independently of the graph. All methods receive the same naming intervention. This figure quantifies changes in F1, not false-positive counts or a causal attribution of the change to memorization.

Method	Covid Resp.	ALZ	Covid Comp.	MACH
MosaCD	.65	.73	.39	.40
Vashishtha	.65	.48	.27	.32
Memorization	.55	.55	.17	.22
Best PC-family	.32	.53	.28	.25

Table 10: **Reported F1 on post-cutoff tasks.** The first three tasks belong to the original 13-dataset comparison; MACH is an additional evaluation. "Best PC-family" denotes the best reported PC, Meek, or SPR result for each task.

L.3 Additional evaluations at 211 and 283 nodes

We evaluate a 211-node streamflow task derived from MACH [38] and a 283-node mHSC-GM gene-regulatory task derived from BEELINE [39]. Table 11 gives the reported F1, end-to-end runtime, and API cost for the methods evaluated. MosaCD improves over the best tested PC-family F1 from .25 to .40 on MACH and .06 to .18 on BEELINE.

	MACH (211 nodes)			BEELINE (283 nodes)		
Method	F1	Time (s)	USD	F1	Time (s)	USD
MosaCD	.40	530	.70	.18	761	1.18
Memorization	.22	25	.01	.08	33	.01
Vashishtha	.32	987	.13	.05	2074	.25
SCP	.29	85	.01	.06	28	.01
PC	.25	501	—	.06	690	—
Meek	.24	505	—	.06	705	—
SPR	.25	503	—	.06	711	—

Table 11: **Additional larger-graph evaluations.** Time is reported end-to-end runtime in seconds; USD is the reported LLM API cost, with a dash for methods without an LLM API cost. Only the methods shown were reported in these evaluations. These measurements are separate from the original stage-timing study in Appendix E and the small-graph resource table below.

L.4 Finite-sample degradation and the skeleton ceiling

We vary sample size from 100 to 10,000 in Insurance and Hepar2, thereby jointly changing the quality of the learned skeleton and separating-set record. The skeleton ceiling is the maximum directed-edge F1 attainable if every retained true skeleton edge is oriented correctly. MosaCD exceeds Meek at every reported sample size in both datasets, while remaining bounded by the learned skeleton.

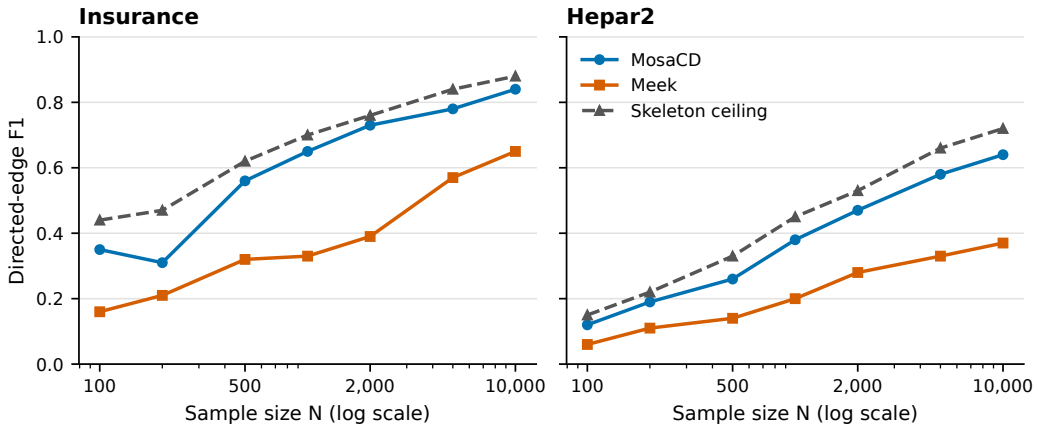


Figure 10: **Finite-sample performance relative to the learned skeleton.** Reported F1 for MosaCD (blue circles), Meek (orange squares), and the skeleton ceiling (dashed gray triangles) on Insurance and Hepar2 at $N \in \{100, 200, 500, 1000, 2000, 5000, 10000\}$. The horizontal axis is logarithmic. The ceiling conditions on the learned skeleton and assumes correct directions for its retained true edges. Lines connect the reported sample sizes. Exact values appear in Table 12.

L.5 A matched one-oracle-edge diagnostic

To separate propagation from access to additional ground-truth information, we supply the identical sampled correct edge to each compatible non-LLM method. Table 13 reports aggregate F1 across the original 13 datasets. SPR with one oracle edge reaches .57, compared with .54 for PC and Meek with the same supplied edge.

	Insurance			Hepar2		
N	MosaCD	Meek	Ceiling	MosaCD	Meek	Ceiling
100	.35	.16	.44	.12	.06	.15
200	.31	.21	.47	.19	.11	.22
500	.56	.32	.62	.26	.14	.33
1,000	.65	.33	.70	.38	.20	.45
2,000	.73	.39	.76	.47	.28	.53
5,000	.78	.57	.84	.58	.33	.66
10,000	.84	.65	.88	.64	.37	.72

Table 12: **Reported sample-size sensitivity.** Entries are directed-edge F1. "Ceiling" is the maximum attainable F1 conditional on the learned skeleton, as defined in the text.

Method	Baseline F1	With one oracle edge
PC	.48	.54
Meek	.52	.54
NOTEARS	.23	.28
GES	.38	.40
SPR	.53	.57

Table 13: **Matched access to one correct orientation.** The same sampled ground-truth edge is supplied to every compatible method. Values are the reported aggregate F1 across 13 datasets.

L.6 Comparative computational resources and baseline interfaces

All LLM methods use the same backbone and variable metadata, but otherwise follow their published interfaces. MosaCD's CI context, two-order voting, filtering, and SPR are components of its algorithm, rather than additions made to every baseline. The comparison therefore does not match prompt budgets or CI access across methods. Tables 14 and 15 give the reported small-graph measurements.

	Asia		Insurance		Hepar2	
Method	Time (s)	USD	Time (s)	USD	Time (s)	USD
Memorization	1.80	.01	4.00	.01	12.60	.01
SCP	.20	.00	52.90	.01	247.50	.01
Vashishtha	18.60	.01	186.70	.01	507.00	.02
Jiralerspong	16.60	.01	72.50	.01	82.10	.01
LLM-CD	55.10	.01	812.40	.01	1291.90	.01
Causal Disco	15.30	.01	106.80	.01	270.70	.03
ILS-CSL	95.60	.01	620.60	.05	1346.30	.11
MosaCD	10.60	.01	56.20	.08	250.00	.15

Table 14: **Reported runtime and API cost for LLM methods.** Times and costs are transcribed from the comparative small-graph analysis. API costs are reported to two decimal places. These are parallelized measurements across all queries and thus differ from the sequential stage timings in Appendix E.

Method	Asia (s)	Insurance (s)	Hepar2 (s)
PC	.18	46.08	237.52
Meek	.18	46.08	237.53
SPR	.18	46.09	237.54
Shapley-PC	.43	347.84	$>$ 7200
NOTEARS	1.83	72.03	25.07
DAG-GNN	26.30	$>$ 7200	$>$ 7200
GES	.42	25.70	87.71
Hill climbing	.41	26.86	$>$ 7200
MMHC	101.30	$>$ 7200	$>$ 7200

Table 15: **Reported non-LLM wall-clock runtimes.** Entries are seconds for the three small-graph tasks. Values above 7,200 indicate runs exceeding two hours.

M Proofs for 5.1

Lemma M.1 (Collider/non-collider soundness w.r.t. recorded minimal separators)**.** *Consider an unshielded triple $X - Z - Y$ with X nonadjacent to Y. Under Markovness, faithfulness, and a perfect CI oracle, the recorded minimal separators satisfy:*

1. *if $X \to Z \leftarrow Y$ in the CPDAG, then $Z \notin S$ for every $S \in \Sigma_{\min}(X, Y)$;*
2. *if the unshielded triple is not a collider in G equivalently, in every DAG in $[G]$, then $Z \in S$ for every $S \in \Sigma_{\min}(X, Y)$.*

Thus the all-absent and all-present tests are sound when the minimal-separator record is correct. If the recorded minimal separators contain both sets that include Z and sets that exclude Z, SPR treats the triple as ambiguous and does not orient it using this rule.

Proof. This is the standard separator characterization of unshielded triples under faithfulness: for nonadjacent X and Y, the membership of the middle node Z in minimal separating sets determines whether the unshielded triple is a collider or a non-collider. The final statement follows from the definition of SPR's CI-guarded rules, which require unanimity over $\Sigma_{\min}(X, Y)$ and abstain under mixed evidence. □

Lemma M.2 (Soundness of CI-guarded rule closure)**.** *Start from the true skeleton after inserting an acyclic seed set $E_{\text{seed}} = \mathcal{K}$ compatible with $[G]_{\mathcal{K}}$. Under the assumptions of Theorem 5.1, every orientation added by SPR Step 1 preserves compatibility with $[G]_{\mathcal{K}}$ and preserves acyclicity.*

Proof. Step 3.1 first applies only the directed-path acyclicity rule: if $X - Y$ and $X \to \cdots \to Y$, then orient $X \to Y$, since the opposite orientation would create a directed cycle. Steps 3.2–3.3 are guarded by $\Sigma_{\min}(X, Y)$ on unshielded triples. By Lemma M.1, the all-present and all-absent separator tests are sound under the oracle assumptions and a correct recorded minimal-separator collection. Because candidate orientations that create directed cycles are rejected, each accepted orientation preserves acyclicity. By induction over the finite sequence of rule applications, all orientations added by SPR Step 1 remain compatible with the skeleton, $\Sigma_{\min}$, and $\mathcal{K}$. □

Proof of Theorem 5.1 By the perfect CI oracle, faithfulness, and skeleton consistency assumptions, the skeleton search returns $\mathrm{Skel}(G)$ and the recorded minimal separator collection $\Sigma_{\min}$ is correct. By seed compatibility, the initial PDAG after adding $E_{\text{seed}} = \mathcal{K}$ is acyclic and has at least one DAG extension in $[G]_{\mathcal{K}}$. Lemma M.2 shows that each orientation added by SPR Step 1 preserves compatibility with the restricted equivalence class and with the recorded separator constraints. Since the graph is finite and orientations are only added, Step 1 terminates. The returned graph is therefore an acyclic PDAG with the true skeleton satisfying $\Sigma_{\min}$ and $\mathcal{K}$.

When $\mathcal{K} = \emptyset$, the restricted class $[G]_{\mathcal{K}}$ equals the ordinary Markov equivalence class $[G]$. Under full faithfulness and a perfect CI oracle, the recorded minimal separators correctly identify the unshielded colliders, and exhausting the standard acyclicity/orientation closure yields the ordinary CPDAG. In this oracle-faithful setting, PC, PC-stable, and CPC have the same skeleton and collider information and hence the same closure. This final equivalence is not asserted under adjacency-faithfulness alone.

N Proofs for 5.2

Let $S_{Z,\ell}$ and $U_{Z,\ell}$ denote the numbers of true sepsets and true non-sepsets in incZ$_\ell$, and let $S_{\neg Z,\ell}$ and $U_{\neg Z,\ell}$ denote the same counts for notZ$_\ell$, and $S_\ell = S_{Z,\ell} + S_{\neg Z,\ell}$, $U_\ell = U_{Z,\ell} + U_{\neg Z,\ell}$.

Lemma N.1 (First-hit filtration across levels)**.** *Define event D "there is at least one independence hit at some level". Define F_ℓ as the event "no hit on levels $k < \ell$, and at least one hit on level ℓ". Under*

5.3 (independent CI tests, FPR $= \alpha$, FNR $= \beta$), we have

$$\Pr(D) = 1 - \prod_{\ell} \beta^{S_\ell}(1-\alpha)^{U_\ell}$$

$$\mathrm{PrevNoHit}_\ell := \Pr(\textit{no hit on } k < \ell) = \prod_{k<\ell} \beta^{S_k}(1-\alpha)^{U_k}$$

$$\Pr(F_\ell) = \mathrm{PrevNoHit}_\ell \cdot \left(1 - \beta^{S_\ell}(1-\alpha)^{U_\ell}\right)$$

where $\{F_\ell\}$ is a disjoint partition of D.

Proof. Under Assumption 5.3 each candidate on level k hits (declares independence) with probability $1-\beta$ if it is a true sepset and with probability α if not; candidates and levels are independent. Thus "no hit anywhere" has probability $\prod_\ell \beta^{S_\ell}(1-\alpha)^{U_\ell}$, giving $\Pr(D)$. The rest is by independence across levels and the definition of F_ℓ. □

Assumption N.2 (No short detours; Z-only control up to order ℓ)**.** Fix integers $\ell \geq 0$ and $L \geq 2\ell+1$. (a) (No short detours) Every $X \rightsquigarrow Y$ path of length $\leq L$ contains Z. (b) (Z-only control) For any conditioning set C with $|C| \leq \ell$, the segment $X - Z - Y$ is open/blocked iff $Z \notin C$/$Z \in C$ when Z is a non-collider, and blocked/open iff $Z \notin C$/$Z \in C$ when Z is a collider; in particular, conditioning on any other node (including descendants on the segment) cannot change the segment's status for $|C| \leq \ell$.

Lemma N.3 (Bucket counts and collider/non-collider separation)**.** *Assume Markov and Faithfulness and Assumption N.2 (with $2\ell+1 \leq L$), at level $\ell = 0$, if Z is a non-collider, $S_{\neg Z,0} = 0$, $U_{\neg Z,0} = 1$; if Z is a collider, $S_{\neg Z,0} = 1$, $U_{\neg Z,0} = 0$. In both cases, incZ$_0 = \varnothing$, hence $S_{Z,0} = U_{Z,0} = 0$.*

For all $\ell \geq 1$, if non-collider truth, $S_{Z,\ell} =$ incZ$_\ell$, $U_{Z,\ell} = 0$, $S_{\neg Z,\ell} = 0$, $U_{\neg Z,\ell} =$ notZ$_\ell$; if collider truth, $S_{Z,\ell} = 0$, $U_{Z,\ell} =$ incZ$_\ell$, $S_{\neg Z,\ell} =$ notZ$_\ell$, $U_{\neg Z,\ell} = 0$.

Proof. D-separation rules are that, a non-collider blocks a path iff it is in C; a collider blocks unless it or a descendant is in C.

When $\ell = 0$, for a non-collider chain $X - Z - Y$, the path is open unconditionally, so $\emptyset$ is a non-sepset; for a collider, it is blocked unconditionally, so $\emptyset$ is a sepset.

When $\ell \geq 1$: by Assumption N.2(a), all short $X - Y$ paths pass through Z; by Assumption N.2(b), with $|C| \leq \ell$ the local segment's status is controlled only by the inclusion of Z. So for a non-collider, if $Z \in C$ (bucket incZ$_\ell$), the local segment is blocked and, since every short path uses Z, all paths are blocked, thus sepset; if $Z \notin C$, the local segment is active and cannot be blocked by other vertices with $|C| \leq \ell$, thus non-sepset. For a collider, if $Z \in C$, the local segment is opened and cannot be re-blocked by $|C| \leq \ell$, thus non-sepset; if $Z \notin C$, the local segment remains blocked and no short detour exists, thus sepset. □

Lemma N.4 (Order-averaged first-hit factor within a level)**.** *Fix a level containing m true sepsets (each hits with prob. $a = 1-\beta$) and n non-sepsets (each hits with prob. $b = \alpha$). Under a uniformly random within-level permutation (independent of outcomes by Assumption 5.3), the average no-hit-from-predecessors factor for a fixed candidate equals*

$$I_{m,n}(a,b) = \int_0^1 (1-au)^m (1-bu)^n \, du = \sum_{i=0}^{m} \sum_{j=0}^{n} \binom{m}{i} \binom{n}{j} \frac{(-a)^i (-b)^j}{i+j+1}$$

with $I_{m,0} = \frac{1-(1-a)^{m+1}}{a(m+1)}$ and $I_{0,n} = \frac{1-(1-b)^{n+1}}{b(n+1)}$.

Proof. Let $N := m + n + 1$ be the number of candidates on the level including a fixed target k. Write $p_j \in \{a, b\}$ for the hit probability of candidate $j \neq k$ and, for a random permutation π, define $X_\pi = \prod_{j \prec_\pi k}(1-p_j)$, the product of "no-hit" factors over predecessors of k. We average X_π over all permutations.

If r candidates precede k, then (i) r is uniform on $\{0, \dots, N-1\}$ with probability $1/N$; (ii) conditional on r, the predecessor set S is uniform over the $\binom{N-1}{r}$ subsets of $\{j \neq k\}$ of size r. Hence

$$\mathbb{E}_\pi[X_\pi] = \frac{1}{N} \sum_{r=0}^{N-1} \frac{1}{\binom{N-1}{r}} \sum_{\substack{S \subseteq \{j \neq k\} \\ |S|=r}} \prod_{j \in S} (1 - p_j).$$

Insert

$$\frac{1}{N\binom{N-1}{r}} = \frac{r!(N-1-r)!}{N!} = \int_0^1 u^r (1-u)^{N-1-r}\, du$$

and swap sum/integral (justified since the sums are finite), we have

$$\mathbb{E}_\pi[X_\pi] = \int_0^1 \left[\sum_{r=0}^{N-1} e_r\, u^r (1-u)^{N-1-r} \right] du$$

where $e_r = \sum_{|S|=r} \prod_{j \in S} (1 - p_j)$ are elementary symmetric sums of $\{1 - p_j\}_{j \neq k}$.

The generating function is

$$\sum_{r=0}^{N-1} e_r t^r = \prod_{j \neq k} \big(1 + (1 - p_j)t\big)$$

With $t = \frac{u}{1-u}$ and factoring $(1-u)^{N-1}$ we obtain

$$\sum_{r=0}^{N-1} e_r\, u^r (1-u)^{N-1-r} = \prod_{j \neq k} (1 - p_j u)$$

so

$$\mathbb{E}_\pi[X_\pi] = \int_0^1 \prod_{j \neq k} (1 - p_j u)\, du$$

There are m terms with $p_j = a$ and n with $p_j = b$, hence

$$I_{m,n}(a,b) = \int_0^1 (1-au)^m (1-bu)^n\, du$$

Expanding $(1-au)^m(1-bu)^n$ and integrating termwise yields the stated double sum. The special cases follow by taking $n = 0$ or $m = 0$.

For a specific candidate k with hit probability $p_k \in \{a, b\}$,

$$\Pr(k \text{ is first hit}) = p_k \cdot \mathbb{E}_\pi \left[\prod_{j \prec_\pi k} (1 - p_j) \right] = \begin{cases} a\, I_{S_\ell - 1, U_\ell}(a,b), & k \text{ a true sepset}, \\ b\, I_{S_\ell,\, U_\ell - 1}(a,b), & k \text{ a non-sepset}. \end{cases}$$

□

Lemma N.5 (Level-ℓ identification probabilities)**.** *Condition on the partition $\{F_\ell\}$ from Lemma N.1. Under Assumption 5.3,*

$$\Pr(E \mid D) = \sum_\ell \frac{\Pr(E \cap F_\ell)}{\Pr(D)} = \sum_\ell \frac{\mathrm{PrevNoHit}_\ell}{\Pr(D)} \cdot \Pr(E \textit{ via level } \ell \mid \textit{ level } \ell \textit{ has a hit}).$$

For CPC (bucket exclusivity within a level) and PC (first hit within a level),

$$\Pr_{\mathrm{CPC}}(\textit{identified as collider} \mid D)$$
$$= \frac{1}{\Pr(D)} \sum_{\ell} \mathrm{PrevNoHit}_{\ell} \underbrace{\beta^{S_{Z,\ell}}(1-\alpha)^{U_{Z,\ell}}}_{\textit{no } Z\textit{-hits at } \ell} \underbrace{\big(1-\beta^{S_{\neg Z,\ell}}(1-\alpha)^{U_{\neg Z,\ell}}\big)}_{\textit{some non-}Z\textit{ hit}},$$

$$\Pr_{\mathrm{CPC}}(Z \textit{ in all saved sepsets} \mid D)$$
$$= \frac{1}{\Pr(D)} \sum_{\ell} \mathrm{PrevNoHit}_{\ell} \beta^{S_{\neg Z,\ell}}(1-\alpha)^{U_{\neg Z,\ell}} \big(1-\beta^{S_{Z,\ell}}(1-\alpha)^{U_{Z,\ell}}\big),$$

$$\Pr_{\mathrm{PC}}(\textit{identified as collider} \mid D)$$
$$= \frac{1}{\Pr(D)} \sum_{\ell} \mathrm{PrevNoHit}_{\ell} \Big[S_{\neg Z,\ell}\,(1-\beta)\, I_{S_\ell-1,U_\ell}((1-\beta),\alpha)$$
$$+ U_{\neg Z,\ell}\,\alpha\, I_{S_\ell,U_\ell-1}((1-\beta),\alpha)\Big],$$

$$\Pr_{\mathrm{PC}}(Z \textit{ in saved sepsets} \mid D)$$
$$= \frac{1}{\Pr(D)} \sum_{\ell} \mathrm{PrevNoHit}_{\ell} \Big[S_{Z,\ell}\,(1-\beta)\, I_{S_\ell-1,U_\ell}((1-\beta),\alpha)$$
$$+ U_{Z,\ell}\,\alpha\, I_{S_\ell,U_\ell-1}((1-\beta),\alpha)\Big].$$

Proof. Condition on F_ℓ and apply CPC's exclusivity or PC's first-hit rule with the order-averaged factors from Lemma N.4. □

Proofs for Theorem 5.5 For CPC, under non-collider truth (sepsets in incZ$_\ell$), Lemma N.3 gives $S_{Z,\ell} = m$, $U_{Z,\ell} = 0$, $S_{\neg Z,\ell} = 0$, $U_{\neg Z,\ell} = n$. Plugging into Lemma N.5 yields $\Pr_{\mathrm{CPC}}(\text{collider at } \ell) = \mathrm{PrevNoHit}_\ell\, \beta^m\big[1-(1-\alpha)^n\big]$; under collider truth (sepsets in notZ$_\ell$), the symmetric expression is $\Pr_{\mathrm{CPC}}(Z \text{ in saved sepset at } \ell) = \mathrm{PrevNoHit}_\ell\, \beta^n\big[1-(1-\alpha)^m\big]$. Take the ratio to cancel $\mathrm{PrevNoHit}_\ell$. For CPC, with the same bucket counts and Lemma N.4, $\Pr_{\mathrm{PC}}(\text{collider at } \ell) = \mathrm{PrevNoHit}_\ell \cdot n \cdot b \cdot I_{m,\,n-1}(a,b)$, and $\Pr_{\mathrm{PC}}(Z \text{ in saved sepset at } \ell) = \mathrm{PrevNoHit}_\ell \cdot m \cdot b \cdot I_{n,\,m-1}(a,b)$. Divide to cancel $\mathrm{PrevNoHit}_\ell$ and b.

For Level-ℓ ($\ell \geq 1$), the wrong-orientation odds under Assumptions 5.3,5.4 can be obtained as

$$\mathcal{R}_\ell^{\mathrm{CPC}} = \beta^{\,m-n}\,\frac{1-(1-\alpha)^n}{1-(1-\alpha)^m}$$
$$\mathcal{R}_\ell^{\mathrm{PC}} = \frac{n}{m}\cdot\frac{I_{m,\,n-1}(1-\beta,\alpha)}{I_{n,\,m-1}(1-\beta,\alpha)}$$

where $I_{p,q}(a,b) = \int_0^1 (1-au)^p(1-bu)^q\,du$, $m = \text{incZ}_\ell$ and $n = \text{notZ}_\ell$.

Therefore, for early levels $\ell \geq 1$ and $\ell \ll \frac{M}{2}$:

$$\text{CPC: } \mathcal{R}_\ell^{\mathrm{CPC}} \approx \beta^{m-n}\frac{n}{m} = \beta^{\binom{M-1}{\ell-1}-\binom{M-1}{\ell}}\cdot\frac{M-\ell}{\ell} \quad (\alpha \text{ small})$$
$$\text{PC: } \mathcal{R}_\ell^{\mathrm{PC}} \approx \frac{n(n+1)}{m(m+1)}\left[1+\alpha\frac{(m-n)(m+n+1)}{(m+2)(n+2)}\right] \quad (\alpha,\beta \text{ small}),$$

Taking $\alpha, \beta = o\left(\frac{1}{M}\right)$ completes the proof for Theorem 5.5. As $\binom{M-1}{\ell}$ grows with ℓ up to $M/2$, $n > m$, $\beta^{m-n} > 1$ as $\beta < 1$, so $\mathcal{R}_\ell^{\mathrm{CPC}} > 1$. At zeroth order $\mathcal{R}_\ell^{\mathrm{PC}} \approx \dfrac{n(n+1)}{m(m+1)} \approx \big(\dfrac{M-\ell}{\ell}\big)^2$ when m, n are large, so $\mathcal{R}_\ell^{\mathrm{PC}} > 1$ for $\ell < M/2$.

Corollary N.6 (small-α approximation for CPC)**.** *For $\ell \geq 1$ and small α, as $1-(1-\alpha)^t = t\alpha + O(\alpha^2)$, we have*

$$\mathcal{R}_\ell^{\mathrm{CPC}} = \beta^{\,m-n}\,\frac{1-(1-\alpha)^n}{1-(1-\alpha)^m} \approx \beta^{\,m-n}\,\frac{n}{m} = \beta^{\binom{M-1}{\ell-1}-\binom{M-1}{\ell}}\cdot\frac{M-\ell}{\ell},$$

with relative error $O(\alpha)$.

Corollary N.7 (small-α and small-β approximation for PC)**.** *For $\ell \geq 1$ and small α, β,*

$$I_{m,n}(1-\beta,\alpha) \;\approx\; \frac{1}{m+1} \;+\; \frac{\beta}{m+1} \;-\; \frac{n\,\alpha}{(m+1)(m+2)} \;+\; O(\alpha^2, \alpha\beta, \beta^2).$$

Using the ratio expansion $(x+\delta_x)/(y+\delta_y) \approx (x/y)\,[1 + (\delta_x/x) - (\delta_y/y)]$, we have

$$\mathcal{R}_\ell^{\mathrm{PC}} \;\approx\; \frac{n(n+1)}{m(m+1)} \left[1 \;+\; \alpha\, \frac{(m-n)(m+n+1)}{(m+2)(n+2)}\right] \;+\; O(\alpha^2, \alpha\beta, \beta^2).$$

Zeroth order (ignore α, β) is

$$\mathcal{R}_\ell^{\mathrm{PC}} \approx \frac{n(n+1)}{m(m+1)} = \frac{M-\ell}{\ell} \cdot \frac{n+1}{m+1}.$$

and $\frac{n+1}{m+1}$ has no tidy closed form in M, ℓ. For large m, n, $\frac{n+1}{m+1} \approx \frac{n}{m}$, giving $\mathcal{R}_\ell^{\mathrm{PC}} \approx \left(\frac{M-\ell}{\ell}\right)^2$.

Proof.

$$I_{m,n}(1-\beta,\alpha) = \int_0^1 \big(1-(1-\beta)u\big)^m \big(1-\alpha u\big)^n \, du.$$

Write

$$\big(1-(1-\beta)u\big)^m = (1-u)^m \Big(1 + \frac{\beta u}{1-u}\Big)^m \approx (1-u)^m \Big(1 + m\frac{\beta u}{1-u}\Big),$$
$$(1-\alpha u)^n \approx 1 - n\alpha u,$$

keeping terms up to $O(\alpha, \beta)$ and dropping $O(\alpha^2, \alpha\beta, \beta^2)$ then multiplying (and ignoring the $\alpha\beta$ cross-term):

$$\big(1-(1-\beta)u\big)^m (1-\alpha u)^n \approx (1-u)^m + m\beta\, u(1-u)^{m-1} - n\alpha\, u(1-u)^m.$$

Integrate termwise, we have

$$I_{m,n}(1-\beta,\alpha) \;\approx\; \frac{1}{m+1} \;+\; \frac{\beta}{m+1} \;-\; \frac{n\,\alpha}{(m+1)(m+2)} \;+\; O(\alpha^2, \alpha\beta, \beta^2),$$

Since

$$\mathcal{R}_\ell^{\mathrm{PC}} = \frac{n}{m} \cdot \frac{I_{m,\,n-1}(1-\beta,\alpha)}{I_{n,\,m-1}(1-\beta,\alpha)}, \quad m = \mathrm{incZ}_\ell,\; n = \mathrm{notZ}_\ell.$$

Apply the expansion:

$$I_{m,\,n-1}(1-\beta,\alpha) \approx \frac{1}{m+1} + \frac{\beta}{m+1} - \frac{(n-1)\alpha}{(m+1)(m+2)},$$
$$I_{n,\,m-1}(1-\beta,\alpha) \approx \frac{1}{n+1} + \frac{\beta}{n+1} - \frac{(m-1)\alpha}{(n+1)(n+2)}.$$

Use the first-order ratio expansion $\frac{x+\delta_x}{y+\delta_y} \;\approx\; \frac{x}{y}\Big[1 + \frac{\delta_x}{x} - \frac{\delta_y}{y}\Big]$, where $x = \frac{1}{m+1}$, $y = \frac{1}{n+1}$. As $\frac{\delta_x}{x} = \beta - \frac{(n-1)\alpha}{m+2}$, $\frac{\delta_y}{y} = \beta - \frac{(m-1)\alpha}{n+2}$, the ratio is approximated by $\frac{n+1}{m+1}\left[1 + \alpha\left(\frac{m-1}{n+2} - \frac{n-1}{m+2}\right)\right]$.

Therefore

$$\mathcal{R}_\ell^{\mathrm{PC}} \approx \frac{n(n+1)}{m(m+1)} \left[1 + \alpha\, \frac{(m-n)(m+n+1)}{(m+2)(n+2)}\right] + O(\alpha^2, \alpha\beta, \beta^2).$$

At zeroth order (ignore α, β),

$$\mathcal{R}_\ell^{\mathrm{PC}} \approx \frac{n(n+1)}{m(m+1)} = \Big(\frac{n}{m}\Big)\Big(\frac{n+1}{m+1}\Big).$$

Since $n/m = \binom{M-1}{\ell} / \binom{M-1}{\ell-1} = \frac{M-\ell}{\ell}$,

$$\mathcal{R}_\ell^{\mathrm{PC}} \approx \frac{M-\ell}{\ell} \cdot \frac{n+1}{m+1},$$

and $\frac{n+1}{m+1}$ has no tidy closed form in M, ℓ. For large m, n, $\frac{n+1}{m+1} \approx \frac{n}{m}$, giving $\mathcal{R}_\ell^{\mathrm{PC}} \approx \left(\frac{M-\ell}{\ell}\right)^2$. $\square$

O Numerical experiments on non-collider / collider identification FPRs

We set max searching layer as $l = 3$, $\alpha = 0.05$ and $\beta = 0.1$, plugging in the number of nodes, arcs and average degrees of datasets, we can accordingly compute the expected FPRs as Table 16.

Table 16: Expected false positive rates (FPRs) during identifications.

network	**PC (colliders-first)**	**PC (nonc-first)**	**CPC (colliders-first)**	**CPC (nonc-first)**
asia	0.177849	0.000926	0.071491	$5. \times 10^{-8}$
alarm	0.583846	0.000159	0.128392	$5. \times 10^{-37}$
cancer	0.102895	0.001906	0.059310	$6. \times 10^{-5}$
child	0.399529	0.000309	0.105279	$5. \times 10^{-20}$
hailfinder	0.700605	0.000103	0.138733	$5. \times 10^{-56}$
hepar2	0.755133	0.000082	0.141944	$5. \times 10^{-70}$
insurance	0.488650	0.000222	0.117261	$5. \times 10^{-27}$
mildew	0.567216	0.000168	0.126597	$5. \times 10^{-35}$
water	0.540173	0.000185	0.123536	$5. \times 10^{-32}$
win95pts	0.773360	0.000075	0.142753	$5. \times 10^{-76}$

P Prompting Templates and Parsing Rule

P.1 Answer Tag Parsing

We extract the final choice using the following case-insensitive regular expression, which returns a single capital letter in {A,B,C,D,E}:

Listing 1: Regex for parsing the <Answer> tag.

```
_ans_re = re.compile(r"<\s*answer\s*>\s*([ABCDE])\s*<\s*/\s*answer\s*>
    ", re.I)
```

P.2 Prompt templates

Placeholders in braces are programmatically substituted (e.g., {u}, {v}, {data_desc}).

Listing 2: Full chain-of-thought selection template.

```
You are a senior researcher in causal discovery. We are studying the
    following dataset:

{data_desc}

The two target variables under review are {u} and {v}.

Conditional-independence tests mentioning these variables:

{ci_bullets}

Neighbour chain(s) that must normally remain non-collider:

{chains}

The nodes involved are described as below:

{node_desc}

Choose one explanation that best fits domain knowledge and/or decides
    a CI test is unreliable (avoid selecting D or E unless other
    options are strongly against common sense):

A. Undecided. We don't know enough to confidently pick a
    directionality.
B. Changing the state of {u} causally affects {v}, and {v} causally
    affects {u_theOther_2v}.
C. Changing the state of {v} causally affects {u}, and {u} causally
    affects {v_theOther_2u}.
D. Changing the state of {u} causally affects {v}, and {u_theOther_2v}
     also causally affects {v}, **violating corresponding CI tests**.
E. Changing the state of {v} causally affects {u}, and {v_theOther_2u}
     also causally affects {u}, **violating corresponding CI tests**.

Think step-by-step before selecting:
1. Mechanisms - What known causal pathways (biological, physical, etc
    .) support each direction?
2. Counterfactual test - What would happen if we intervened on one
    node? What would we expect?
3. Empirical check - Point to one key piece of information that favors
    /weakens a direction.
4. Comparison - Briefly weigh A vs B vs C vs D vs E and choose the
    most plausible.

Return exactly three lines:
1. Reasoning in support of one direction.
```

```
2. Reasoning against the weaker/less plausible direction.
3. Final choice:  <Answer>A/B/C/D/E</Answer>
```

P.3 Template When Only $v \to u$ Ancillary Edge Is Possible (_CHAIN_PROMPT_TMPL_None2u)

Listing 3: Restricted template (None2u).

```
You are a senior researcher in causal discovery. We are studying the
    following dataset:

{data_desc}

The two target variables under review are {u} and {v}.

Conditional-independence tests mentioning these variables:

{ci_bullets}

Neighbour chain(s) that must normally remain non-collider:

{chains}

The nodes involved are described as below:

{node_desc}

Choose one explanation that best fits domain knowledge and/or decides
    a CI test is unreliable (avoid selecting D unless other options
    are strongly against common sense):

A. Undecided. We don't know enough to confidently pick a
    directionality.
B. Changing the state of {u} causally affects {v}, and {v} causally
    affects {u_theOther_2v}.
C. Changing the state of {v} causally affects {u}.
D. Changing the state of {u} causally affects {v}, and {u_theOther_2v}
     also causally affects {v}, **violating corresponding CI tests**.

Think step-by-step before selecting:
1. Mechanisms - What known causal pathways (biological, physical, etc
    .) support each direction?
2. Counterfactual test - What would happen if we intervened on one
    node? What would we expect?
3. Empirical check - Point to one key piece of information that favors
    /weakens a direction.
4. Comparison - Briefly weigh A vs B vs C vs D and choose the most
    plausible.

Return exactly three lines:
1. Reasoning in support of one direction.
2. Reasoning against the weaker/less plausible direction.
3. Final choice:  <Answer>A/B/C/D</Answer>
```

P.4 Template When Only $u \to v$ Ancillary Edge Is Possible (_CHAIN_PROMPT_TMPL_None2v)

Listing 4: Restricted template (None2v).

```
You are a senior researcher in causal discovery. We are studying the
    following dataset:

{data_desc}

The two target variables under review are {u} and {v}.
```

```
Conditional-independence tests mentioning these variables:

{ci_bullets}

Neighbour chain(s) that must normally remain non-collider:

{chains}

The nodes involved are described as below:

{node_desc}

Choose one explanation that best fits domain knowledge and/or decides
    a CI test is unreliable (avoid selecting D unless other options
    are strongly against common sense):

A. Undecided. We don't know enough to confidently pick a
    directionality.
B. Changing the state of {u} causally affects {v}.
C. Changing the state of {v} causally affects {u}, and {u} causally
    affects {v_theOther_2u}.
D. Changing the state of {v} causally affects {u}, and {v_theOther_2u}
     also causally affects {u}, **violating corresponding CI tests**.

Think step-by-step before selecting:
1. Mechanisms - What known causal pathways (biological, physical, etc
    .) support each direction?
2. Counterfactual test - What would happen if we intervened on one
    node? What would we expect?
3. Empirical check - Point to one key piece of information that favors
    /weakens a direction.
4. Comparison - Briefly weigh A vs B vs C vs D and choose the most
    plausible.

Return exactly three lines:
1. Reasoning in support of one direction.
2. Reasoning against the weaker/less plausible direction.
3. Final choice:  <Answer>A/B/C/D</Answer>
```

P.5 Template Without CI/Neighbour Context (_CHAIN_PROMPT_TMPL_None)

Listing 5: Minimal template (None).

```
You are a senior researcher in causal discovery. We are studying the
    following dataset:

{data_desc}

The two target variables under review are {u} and {v}.

The nodes involved are described as below:

{node_desc}

Choose one explanation that best fits domain knowledge:

A. Undecided. We don't know enough to confidently pick a
    directionality.
B. Changing the state of {u} causally affects {v}.
C. Changing the state of {v} causally affects {u}.

Think step-by-step before selecting:
```

```
1. Mechanisms - What known causal pathways (biological, physical, etc
   .) support each direction?
2. Counterfactual test - What would happen if we intervened on one
   node? What would we expect?
3. Empirical check - Point to one key piece of information that favors
   /weakens a direction.
4. Comparison - Briefly weigh A vs B vs C and choose the most
   plausible.

Return exactly three lines:
1. Reasoning in support of one direction.
2. Reasoning against the weaker/less plausible direction.
3. Final choice:  <Answer>A/B/C</Answer>
```

Q Reasoning examples

Below are 2 examples of the LLM reasoning trace in support of a direction. The question asked is from Appendix P2.

Listing 6: Reasoning with large p-values

```
Reasoning in support of one direction.
B is supported by domain knowledge: 'either' (tuberculosis or lung
    cancer) can cause dyspnoea as a symptom. Intervening on 'either' (
    e.g., treating the disease) would alter dyspnoea, aligning with a
    causal direction from 'either' to dyspnoea. The neighbor chain
    being non-collider further supports this directionality.

Reasoning against the weaker/less plausible direction.
C and D contradict domain knowledge: dyspnoea is an effect, not a
    cause, of diseases like tuberculosis or lung cancer. The CI tests
    (p-values ~0.7-1) indicate no direct link between dyspnoea and the
     other variables given 'either', invalidating claims of dyspnoea
    influencing 'either' or its dependencies.

Final choice: <Answer>B</Answer>
```

Listing 7: Reasoning with small p-values

```
1. Reasoning in support of one direction.
Stroke Volume (SV) directly determines Cardiac Output (CO) via the
    relationship CO = SV x Heart Rate. CO, in turn, is a primary
    determinant of Blood Pressure (BP), as BP ~= CO x Peripheral
    Resistance. The CI tests confirm that once CO is known, SV and BP
    become independent (removing the Stroke Volume-Blood Pressure edge
    ), and Hypovolemia/Left-Ventricular Failure influence CO only via
    SV (removing their direct links to CO). This aligns with the
    physiological mechanism where SV -> CO -> BP.

2. Reasoning against the weaker/less plausible direction.
Option C reverses the causal relationship between CO and SV, which
    contradicts the physiological definition of CO as a product of SV
    and Heart Rate. Additionally, Stroke Volume influencing diagnostic
     variables like Hypovolemia is implausible, as these conditions
    are root causes (diagnostic variables have no predecessors).
    Options D and E imply violations of CI tests, which are explicitly
     rejected by the p-values, leaving B as the only direction
    consistent with both domain knowledge and statistical evidence.

3. Final choice: <Answer>B</Answer>
```

R Assumptions of related work

Vashishtha [4] treat the LLM as an imperfect expert queried for causal direction, and assume causal order (a topological order) is a more stable interface for expert knowledge ("we propose that causal order be used as a more stable output interface"), as opposed to directly prompting undirected edges for their orientations. Pairwise LLM querying produces inconsistencies ("pairwise prompts lead to multiple cycles"), so they add an explicit local consistency constraint by prompting on triplets and "instructing the LLM to avoid cycles within this triplet," then vote over redundant triplets; they also analyze an "imperfect expert with an error ϵ on each prediction" and assumes "the ϵ expert to have error probability exactly equal to ϵ".

Causal Disco [18] explicitly models the LLM as an imperfect expert oracle ($E(\mu_i, \mu_j)$) that consumes metadata (names/descriptions) and returns a hypothesized orientation for an edge. A core assumption of their setup is that the true Markov equivalence class is already available ("We assume that ($M_{G^\star}$) is known, e.g., that it has been obtained via some causal discovery algorithm"), and the LLM is used only to orient edges that are ambiguous within that MEC. They further assume a simple noise model for LLM mistakes: the expert's response depends only on the true orientation and is "incorrect with constant probability (ϵ)," which enables Bayesian reasoning over consistent orientations while controlling risk (they target keeping ($G^\star$) in the remaining set with probability at least ($1 - \eta$)).

LLM-CD [19] is another imperfect expert method. Concretely, they assume LLM reasoning should be restricted ("an accuracy-oriented prompting strategy restricts causal analysis to a reliable range"), because "LLM-based causal reasoning on all pairs. . . [is]" unreliable when mechanisms are unclear. Instead of edge constraints they translate LLM outputs into ancestral/path constraints: "We do not directly specify the existence of edges. . . [LLM] knowledge [is] qualitative and can represent indirect interactions." A key consideration of this approach is the reliance on a manually tuned confidence threshold for determining LLM accuracy, which as shown in their paper may vary between datasets.

Jiralerspong [16] assumes an LLM can answer causal queries from metadata alone and explicitly "utilize LLMs to respond to causal queries solely based on metadata. . . without accessing the numerical observations," analogizing this to how human experts build graphs from descriptions. Their main additional assumption for scalability is that full-graph discovery can be organized as a traversal: they construct the graph via BFS with only (O(n)) prompts by asking (i) root-node style questions ("Ask which variables are not caused by any other variables") and (ii) expansion questions ("ask which variables it causes"), inserting edges unless they "form a cycle" to enforce the DAG constraint.

ILS-CSL [31] assumes LLM judgments are most reliable when scoped to checking candidate edges suggested by data. Prior LLM-for-CSL approaches can yield "unreliable constraints from imperfect LLM inferences" and that full pairwise prompting of all nodes in the graph is costly; their key design choice is to "focus LLMs on verifying direct causal relationships already suggested by the data," i.e., use LLMs as an iterative supervisor of a data-driven CSL loop. They assume convergence can be defined operationally: "The iteration concludes when the LLM-based inferences and data-driven CSL align".

SCP [20] assumes LLMs can contribute background causal knowledge without needing data, and that this knowledge becomes more useful when conditioned on statistical signals from an initial causal discovery run. The LLM is "prompted with the results of the SCD" and "evaluates the probability of the causal relationships considering both the domain knowledge and the statistical characteristics suggested by SCD." They assume pairwise causal answers can be operationalized as a prior: the LLM "judges whether causal relationships exist between all pairs. . . with 'yes' or 'no'," and "probabilities of the responses. . . [are] transformed into the prior knowledge matrix. . . reapplied to SCD."

NeurIPS Paper Checklist

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

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

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

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

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

IMPORTANT, please:

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

1. **Claims**

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

   Answer: [Yes]

   Justification: Each claim is supported by an appropriate results section.

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: We have an explicit limitations section.

Guidelines:

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

3. **Theory assumptions and proofs**

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

   Answer: [Yes]

   Justification: Yes, this is a core contribution we make clear throughout the main text and appendix.

   Guidelines:

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

4. **Experimental result reproducibility**

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

   Answer: [Yes]

   Justification: We provide dataset details and documentation as well as our code for running our method.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.

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

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

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

   Answer: [Yes]

   Justification: We upload our code and provide data documentation.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).

- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

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

   Answer: [N/A]

   Justification: We do not have a train/test split setup here.

   Guidelines:

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

7. **Experiment statistical significance**

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

   Answer: [Yes]

   Justification: We report confidence intervals and results average across replicates.

   Guidelines:

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

8. **Experiments compute resources**

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

   Answer: [Yes]

   Justification: We have an explicit section talking about cost and runtime.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.

- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

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

   Answer: [Yes]

   Justification: We do not violate any ethical policies.

   Guidelines:

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

10. **Broader impacts**

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

    Answer: [Yes]

    Justification: We motivate the uses of causal discovery.

    Guidelines:

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

11. **Safeguards**

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

    Answer: [N/A]

    Justification: We do not believe there is any potential for misuse.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.

- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

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

    Answer: [Yes]

    Justification: All data and baselines are sufficiently cited.

    Guidelines:

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

13. **New assets**

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

    Answer: [N/A]

    Justification: We do not introduce any new assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

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

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

    Answer: [N/A]

    Justification: We do not do crowdsourcing.

Guidelines:

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

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

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

    Answer: [N/A]

    Justification: We do not need IRB approval.

    Guidelines:

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

16. **Declaration of LLM usage**

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

    Answer: [Yes]

    Justification: Yes, it is a core contribution of our paper and well documented throughout the main text, code, and appendix.

    Guidelines:

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