Proceedings of Machine Learning Research 323:1–26, 2026 5th Conference on Causal Learning and Reasoning

Retrieving Classes of Causal Orders with Inconsistent Knowledge Bases

Federico Baldo FEDERICO.BALDO@INSERM.FR
Simon Ferreira SIMON.FERREIRA@INSERM.FR
Charles K. Assaad CHARLES.ASSAAD@INSERM.FR
Sorbonne Université, INSERM,
Institut Pierre Louis d'Epidémiologie
et de Santé Publique,
F75012, Paris, France

Editors: Bijan Mazaheri and Niels Richard Hansen

Abstract

Traditional causal discovery methods often depend on strong, untestable assumptions, making them unreliable in real-world applications. In this context, Large Language Models (LLMs) have emerged as a promising alternative for extracting causal knowledge from text-based metadata, effectively consolidating domain expertise. However, LLMs are prone to hallucinations, necessitating strategies that account for these limitations. One effective approach is to use a consistency measure as a proxy of reliability. Moreover, LLMs do not clearly distinguish direct from indirect causal relationships, complicating the discovery of causal Directed Acyclic Graphs (DAGs), which are often sparse. This ambiguity is evident in the way informal sentences are formulated in various domains. For this reason, focusing on causal orders provides a more practical and direct task for LLMs. We propose a new method for deriving abstractions of causal orders that maximizes a consistency score obtained from an LLM. Our approach begins by computing pairwise consistency scores between variables, from which we construct a semi-complete partially directed graph that consolidates these scores into an abstraction. Using this structure, we identify both a maximally oriented partially directed acyclic graph and an optimal set of acyclic tournaments that maximize consistency across all configurations. We further demonstrate how both the abstraction and the class of causal orders can be used to estimate causal effects. We evaluate our method on a wide set of causal DAGs extracted from scientific literature in epidemiology and public health. Our results show that the proposed approach can effectively recover the correct causal order, providing a reliable and practical LLM-assisted causal framework.

Keywords: Causal Inference, Information Retrieval, Large Language Models

1. Introduction

Traditional causal discovery algorithms rely on observational (or interventional) data to uncover causal relationships. To do so, they often make strong assumptions (Spirtes et al., 2001; Glymour et al., 2019; Peters et al., 2017; Assaad et al., 2022), such as causal sufficiency and faithfulness. The recent rise in popularity of Large Language Models (LLMs) offers a new tool to discover causal models (Long et al., 2023a,b; Cohrs et al., 2023; Vashishtha et al., 2025; Kiciman et al., 2024). Unlike traditional causal discovery methods, LLM-aided approaches operate on textual data, leveraging pre-collected knowledge encoded in their training data. However, LLMs have been frequently observed to provide ambiguous, inconsistent, or prompt-sensitive answers when queried about causal relationships (Zečević et al., 2023). Moreover, LLMs have limited capabilities when it comes to

© 2026 F. Baldo, S. Ferreira & C. K. Assaad.

uncovering new knowledge or simply generating novel ideas, which undermines their applicability in the discovery of causal structures (Si et al., 2025)

Most importantly, as noted in Vashishtha et al. (2025), in natural language, direct and indirect causes are often conflated, making them difficult to distinguish. This ambiguity is evident in informal statements across various domains, including philosophy, medicine, and epidemiology. For instance, we commonly assert that a sedentary lifestyle causes type 2 diabetes, when in fact this link is fully mediated by obesity (Li et al., 2022). More generally, causal relations are frequently verbally expressed as a simple relationship: "X causes Y" or "X affects Y" or "X prevents Y", etc. This oversimplification obscures the complex web of direct and indirect influences, including immediate "parents" and distant "ancestors" of a causal pathway. In this paper, we argue that *LLMs are not effective for discovering causal Directed Acyclic Graphs (DAGs), but rather allow us to retrieve abstractions of causal orders that can, under certain conditions, be used to estimate total effects*, or Average Treatment Effects (ATE).

In this context, we account for the inherent unreliability of LLMs, which often produce hallucinated or inconsistent responses. We propose to quantify the reliability of the LLM and use such a metric as a heuristic to identify abstractions of causal orders. Specifically, uncertainty in the LLM's responses is captured through self-consistency, defined as the stability of its answers when queried multiple times about pairwise causal relationships. By maximizing this consistency measure, we identify an abstraction in the form of a Semi-Complete Partially Directed Graph (PDG), which can be transformed into a dense Maximally oriented Partially Directed Acyclic Graph (MPDAG) (Perkovic, 2020). We show that such an MPDAG can be used to establish a simpler criterion for the identifiability of the total effect in settings with a single treatment. Additionally, based on the Semi-Complete PDG, we propose an exact method to derive all maximally consistent causal orders.

Unlike traditional causal discovery methods (Spirtes et al., 2001; Glymour et al., 2019; Peters et al., 2017; Assaad et al., 2022), our approach does not rely on faithfulness, or any parametric assumption, but only on acyclicity and causal sufficiency. We view the LLM as an inconsistent knowledge base (Zheng et al., 2024), rather than an imperfect expert (Long et al., 2023a), which can be used to rapidly access a large amount of information. This perspective reflects the skepticism toward the ability of LLMs to discover new causal relationships (Zečević et al., 2023; Si et al., 2025).

Contributions

- We show how to derive a dense MPDAG from the knowledge provided by an LLM and how it can be used to establish a simpler criterion for identifiability of the total effect.
- We provide an effective algorithm to find a class of causal orders maximally consistent with an LLM. Such a method is based on a top-down search strategy that does not require any parametric assumptions or faithfulness.

The remainder of the paper is organized as follows: in Section 2, we review relevant literature related to LLM-aided causal discovery; in Section 3, we provide some of the basic formal notions used in the paper; in Sections 4 and 5, we present the main contribution of the paper and details regarding the proposed method; Sections 6 and 7 conclude the paper with experimental results and final considerations. Further details regarding the experimental setting, the implementation of the algorithm, and the proofs of the propositions and theorems presented in the paper are provided in the Appendix.

2. Related Work

Causal Inference with Background Knowledge. The use of prior knowledge in causal discovery has been a long-standing research topic aimed at integrating domain-specific information to refine causal graphs. Meek (1995) proposed a set of rules, known as Meek rules, to orient edges in a Completed Partially Directed Acyclic Graph (CPDAG) based on prior knowledge. These rules allow us to refine the CPDAG by orienting edges while preserving the conditional independence encoded in the graph and the acyclicity constraint. More recently, Maathuis and Colombo (2015); Perkovic et al. (2017); Perković et al. (2017); Perkovic (2020); Venkateswaran and Perković (2024) propose a generalization of identifiability by adjustment to abstraction of causal graphs, such as Partially Directed Acyclic Graphs (PDAGs), CPDAGs, and MPDAGs.

LLMs in Causal Discovery. In causal discovery, LLMs have often been referred to as *imperfect experts* (Long et al., 2023a; Vashishtha et al., 2025), since they are trained on vast amounts of textual data, including scientific literature. In this context, assuming we have a textual description of a set of variables —for instance, provided by a human expert —numerous works attempted to use LLMs to uncover causal DAGs (Kiciman et al., 2024; Long et al., 2023a; Cohrs et al., 2023; Vashishtha et al., 2025; Jiralerspong et al., 2024). However, LLMs tend to provide unreliable replies; most notably, they can hallucinate. Concerns regarding their capability to effectively reason have been raised, as they might be just capturing verbal patterns without actually learning the underlying reasoning (Zečević et al., 2023). To tackle this problem, most of the approaches proposed until now quantify the reliability of these models. As presented in Cohrs et al. (2025), there are two primary approaches to compute uncertainties associated with the answers provided by LLMs: the first is to use the probabilities associated with the response tokens, i.e., the likelihood of the generated answer; the second is to evaluate the *consistency*, i.e., the self-coherence, of the LLM output when queried multiple times. In Cohrs et al. (2023), authors propose an LLM-informed variant of the PC algorithm, where conditional independencies are detected by the language model. In Long et al. (2023a), a pairwise prompt strategy is proposed to complete the orientation of edges in a CPDAG. Each edge is associated with an uncertainty —based on the probabilities assigned to the response tokens —, then the Markov Equivalence Class (MEC) is refined through a Bayesian optimization process. Lastly, in Jiralerspong et al. (2024), the authors focus on an optimized query strategy to the LLMs aimed at reducing computational complexity.

LLMs and Causal Orders. In Vashishtha et al. (2025), the authors propose a method to estimate causal orders using LLMs. As previously stated, the intuition is that LLMs can be more effective in identifying causal orders rather than causal DAGs, given the inherent ambiguity of causal relationships in natural language. The method proposed estimates the topological order and identifies the total effect using the backdoor criterion (Pearl, 2009). To recover the causal order, the LLM is asked to provide a DAG for every possible triplet of variables, which are then aggregated to obtain a single causal order.

In this paper, we argue that identifying abstractions of causal orders rather than full causal DAGs presents a more direct task for LLMs. However, unlike Vashishtha et al. (2025), our approach focuses on: 1) using a pairwise prompt strategy to compute the self-consistency of the LLM, 2) identifying an abstraction of the maximally consistent causal orders, namely a fully connected Semi-Complete PDG, and 3) refining the abstraction to obtain a dense MPDAG and a class of causal orders maximally consistent with the LLM. This design explicitly allows the inherent incongruity of the

LLM to be reflected in the resulting causal representations, rather than forcing a single, potentially unreliable, structure.

3. Background

Causal Graphs and Causal Orders A causal DAG, $\mathcal{G} = (\mathbb{V}, \mathbb{E})$, is defined by a set of nodes $\mathbb{V}$ and directed edges $\mathbb{E}$, where an edge $X_i \to X_j$ denotes a direct causal effect of X_i on X_j (Pearl, 2009; Spirtes et al., 2001). In this context, the presence of confounding variables can lead to biased estimates of the causal effect. To address this issue, assuming that all confounders are observed, we can use the backdoor criterion, which identifies sets of variables that, when controlled for, allow us to estimate the causal effect of X_i on X_j without biasing the estimate (Pearl, 2009). The backdoor criterion is always satisfied if the conditioning set $\mathbb{Z}$ contains all the parents of X_i. Given a causal DAG, we can define its causal order as follows:

Definition 1 (Causal Order) *Suppose a causal DAG $\mathcal{G}$. A causal order compatible with $\mathcal{G}$ is a bijective mapping $\pi : \mathbb{V} \mapsto \{1, \ldots, d\}$ such that if X_j is a descendant of X_i in $\mathcal{G}$, $X_i \succ_{\mathcal{G}} X_j$, then $\pi(X_i) < \pi(X_j)$, $\forall X_i, X_j \in \mathbb{V}$.*

Note that multiple causal orders can be compatible with the same causal DAG, as a single DAG may admit several valid topological orderings of its nodes. A graph that admits a unique causal order does not permit ambiguity in the direction of the causal relationships between any pair of nodes. We refer to these graphs as acyclic tournaments.

Definition 2 (Acyclic tournament) *An acyclic tournament is a DAG with one direct edge between every pair of distinct nodes.*

Acyclic tournaments provide a graphical representation that fully encodes a unique causal order. Specifically, the direction of the edge between any two nodes directly reflects their relative position in the causal order. Crucially, the backdoor criterion is always satisfied for causal orders by conditioning on all the parents of the treatment, assuming causal sufficiency (Pearl, 2009; Vashishtha et al., 2025). In the case of tournaments, the set of parents coincides with the set of predecessors in the causal order, eliminating the need to enumerate all paths between treatment and the outcome to identify a valid backdoor set.

Maximally Oriented Partially Directed Graphs (MPDAGs) In general, when causal sufficiency holds, and only observational data are available, if we do not make any parametric assumption, the best one can recover is a CPDAG (Spirtes et al., 2001). A CPDAG is a partially directed graph that represents the equivalence class of all DAGs that are Markov equivalent to each other, meaning they encode the same conditional independences. In the presence of background knowledge — for instance, provided by an expert — we can refine a CPDAG into an MPDAG (Perkovic et al., 2017) by incorporating the additional edge orientations and applying the Meek rules to propagate their implications (Meek, 1995). In the scope of this paper, we will mainly focus on the second Meek rule (R2), which enforces acyclicity in the graph, i.e., if orienting an undirected edge $X_i - X_j$ as $X_i \to X_j$ creates a directed cycle, then the edge must be oriented as $X_j \leftarrow X_i$.

The backdoor criterion cannot be directly applied to these abstractions, which include undirected edges; to this end, the generalized backdoor criterion (Maathuis and Colombo, 2015; Perkovic et al., 2017) allows us to identify causal effects in CPDAGs.

Definition 3 (Generalized Backdoor Criterion) *Let $\mathbb{X}_i$ and $\mathbb{X}_j$ be two disjoint sets of variables in a CPDAG $\mathcal{C}$. A set of variables $\mathbb{Z}$ satisfies the generalized backdoor criterion relative to $(\mathbb{X}_i, \mathbb{X}_j)$ if:*

1. *$\mathbb{Z}$ contains no possible descendants of $\mathbb{X}_i$ in any DAG consistent with $\mathcal{C}$;*
2. *Every path between $\mathbb{X}_i$ and $\mathbb{X}_j$ in $\mathcal{C}$ that contains an edge pointing into $\mathbb{X}_i$ is blocked by $\mathbb{Z}$.*

If $\mathbb{Z}$ satisfies this criterion, then the causal effect of $\mathbb{X}_i$ on $\mathbb{X}_j$ is identifiable and given by:

$$P(\mathbb{X}_j \mid do(\mathbb{X}_i = x)) = \sum_{z \in \mathbb{Z}} P(\mathbb{X}_j \mid \mathbb{X}_i = x, \mathbb{Z} = z) P(\mathbb{Z} = z). \tag{1}$$

The generalized backdoor criterion naturally extends to MPDAGs, which are enriched CPDAGs where all conditional independencies are preserved, and edge orientations are refined based on additional knowledge. Whenever a backdoor set can be identified for an MPDAG, the total effect is identifiable and can be estimated using the standard adjustment formula (1).

4. Inconsistent Knowledge Base

The aim of this paper is to identify factual knowledge embedded in the LLMs from commonsense knowledge and established literature regarding cause-effect relationships in a specific domain. We assume that these relationships do not directly indicate the presence of an edge in a causal DAG, but rather a causal ordering of the variables involved. This knowledge is then used to identify abstractions of causal orders maximally compatible with the information provided by the LLM. In this context, the LLM acts as a knowledge base (Zheng et al., 2024) that: 1) has access to a large body of knowledge and 2) may provide incorrect responses, e.g., hallucinated responses. Measuring the reliability of an LLM's response is a well-established practice, which aims at quantifying the trustworthiness of the information provided by the model (Cohrs et al., 2025). In this paper, we rely on self-consistency as a proxy of uncertainty —which has been shown to outperform other metrics, such as entropy, confidence elicitation, and token probabilities (Manakul et al., 2023; Savage et al., 2024).

The goal is to define a consistency matrix encapsulating a measure of self-consistency of the LLM over pairwise causal relationships, which can be leveraged to identify maximally consistent causal orders. Following the approach adopted in Long et al. (2023a); Kadavath et al. (2022), we assume that we have a set of variables $X_1, ..., X_d$ associated with a set of textual descriptions, $\mu_1, ..., \mu_d$. The consistency is the degree of agreement of the LLM when queried multiple times about the causal relationship between two variables, X_i and X_j, with semantically equivalent queries. The queries are generated by the LLM itself, which is asked to rephrase an initial sentence, such as *"Is μ_i a cause of μ_j?"*, into a set of semantically equivalent questions. To reduce the number of incorrect responses, we restrict the possible answers to a `Yes` or a `No` —more details in Appendix D. The consistency score is then computed as the proportion of positive responses to the prompts. Specifically, when an LLM is queried n times, the consistency score for $X_i \succ X_j$ is calculated as:

$$C_{X_i \succ X_j} = \frac{1}{n} \sum_{k=1}^{n} r_k,$$

where r_k is the response of the LLM to the k-th query, and $r_k = 1$ if the response is `Yes`, $r_k = 0$ if `No`. Note that $C_{X_i \succ X_j}$ and $C_{X_j \succ X_i}$ are computed independently of each other, i.e., there is no complementarity between the two. As shown in Figure 1 (a), we can derive a consistency matrix, C, from the LLM's responses over a set of variables. We formalize notions of consistency for a knowledge base as follows:

Definition 4 (Consistent Knowledge Base) *Given a causal DAG, $\mathcal{G}$, a knowledge base is consistent if, for every pair of variables X_i and X_j, such that $X_i \succ_{\mathcal{G}} X_j$ the consistency score $C_{X_i \succ X_j} \geq C_{X_j \succ X_i}$.*

Definition 5 (Strictly Consistent Knowledge Base) *Given a causal DAG, $\mathcal{G}$, a knowledge base is strictly consistent if, for every pair of variables X_i and X_j, such that $X_i \succ_{\mathcal{G}} X_j$ the consistency score $C_{X_i \succ X_j} > C_{X_j \succ X_i}$.*

5. Retrieving Maximally Consistent Abstractions of Causal Orders

The consistency matrix can be leveraged to construct a maximally consistent graph. Specifically, we can build a partially directed graph, $\mathcal{S}$, such that: a directed edge $X_i \to X_j$ is included in $\mathcal{S}$ if $C_{X_i \succ X_j} > C_{X_j \succ X_i}$, and an undirected edge $X_i - X_j$ is included if $C_{X_i \succ X_j} = C_{X_j \succ X_i}$. In Figure 1, we illustrate how the consistency matrix can be used to derive such graph —in bold the consistencies that determine the orientation and the type of the edges. Note that $\mathcal{S}$ can contain directed cycles, since there is no explicit mechanism preventing them. This can be seen in Figure 1 (b), where the directed edges $W \to V$, $V \to Y$, and $Y \to W$ form a directed cycle. We refer to a graph defined in this fashion as *semi-complete partially directed graph*.

Definition 6 (Semi-Complete Partially Directed Graph) *A semi-complete partially directed graph, or semi-complete PDG, is a dense graph in which every pair of distinct nodes is connected by exactly one edge, which may be either directed or undirected.*

5.1. Maximally Consistent Semi-Complete Partially Directed Graphs

The maximally consistent semi-complete PDG, $\mathcal{S}$, holds specific properties based on the consistency of the knowledge base used to construct it. Indeed, we can show that $\mathcal{S}$ is a dense MPDAG if the LLM is consistent (Definition 4).

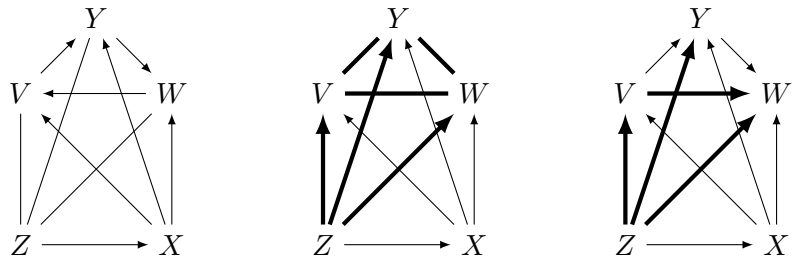
Proposition 7 *The maximally consistent semi-complete PDG, $\mathcal{S}$, does not contain directed cycles if the consistency matrix is provided by a consistent knowledge base.*

Proposition 8 *The maximally consistent semi-complete PDG, $\mathcal{S}$, obtained from a consistent knowledge base and by applying R2 of the Meek rules, is an MPDAG.*

Note that a dense MPDAG, i.e., with an edge between every pair of variables, has no v-structures, but only shielded colliders; this implies that we only require R2 for maximal orientation. Unfortunately, we have no guarantees on the consistency of the LLM; thus, the semi-complete PDG may contain directed cycles. In this case, we can still derive an MPDAG from $\mathcal{S}$ by unorienting directed cycles and enforcing acyclicity.

	Y	W	V	X	Z
Y	-	**0.6**	0.2	0.3	**1.0**
W	0.3	-	**0.7**	0.4	**0.4**
V	**0.7**	0.5	-	0.5	**0.6**
X	0.4	**0.9**	**0.9**	-	0.7
Z	**1.0**	**0.4**	**0.6**	**0.8**	-

(a)



(b) (c) (d)

Figure 1: (a) Consistency Matrix. (b) Maximally consistent Semi-Complete PDG. (c) MPDAG derived from the Semi-Complete PDG. (d) Maximally consistent acyclic tournament. In bold edges modified with respect to the Semi-Complete PDG.

Proposition 9 *Assuming a semi-complete PDG, $\mathcal{S}$, containing directed cycles, $C_1, \ldots, C_m$, we can derive an MPDAG from $\mathcal{S}$ by:*

1. *Unorienting every edge in a cycle C_k, i.e., for every pair of nodes X_i and X_j, such that $X_i \rightarrow X_j$ is part of a cycle C_k, we substitute the directed edge with an undirected one, $X_i - X_j$.*

2. *Maximally orient undirected edges applying R2 of the Meek rules.*

An application of Proposition 9 is illustrated in Figure 1 (c), where the directed cycle involving nodes Y, W, and V is unoriented, and then R2 is applied to orient additional edges. Note also that the one-to-one correspondence between tournaments and orders does not hold for MPDAGs, which rather map to a set of partial orders.

Reasoning with Dense Maximally Oriented Partially Directed Graphs. As per Maathuis and Colombo (2015); Perković et al. (2017), we can use MPDAGs to identify the total effect through the generalized backdoor criterion. We show that a dense MPDAG, obtained from the maximally consistent Semi-Complete PDG, as per Proposition 7, 8, and 9, allows us to establish a simpler identifiability criterion for univariate treatments:

Proposition 10 *Given a dense MPDAG and the treatment variable X_i, the total effect is identifiable, using the adjustment formula (1), if and only if there are no undirected edges related to X_i.*

This implies that it suffices to examine whether any undirected edges are incident on the treatment variable X_i to determine if the total effect can be identified via adjustment. Note that some nodes in $\mathcal{S}$ may be *order invariant*, meaning that they are connected to every other node with undirected edges. These nodes do not provide any information about the causal order and prevent identifiability.

Corollary 11 *Given a dense MPDAG, if it contains an order invariant node, X_i, any causal effect between X_i and every other variable is not identifiable.*

The key assumption underlying the construction of a dense MPDAG from the consistency matrix is that the LLM provides reliable information about pairwise causal relationships. While this assumption may not always hold, Proposition 9 allows us to handle inconsistencies in the knowledge base by unorienting directed cycles. Alternatively, one could interpret cycles within $\mathcal{S}$ as potential errors, which can be addressed by applying corrective procedures that enforce acyclicity.

5.2. Maximally Consistent Causal Orders

Orienting all edges while enforcing acyclicity and maximizing consistency over the semi-complete PDG, $\mathcal{S}$, allows us to identify a class of maximally consistent acyclic tournaments (MCAT). Each acyclic tournament corresponds to a unique causal order, and vice versa (Section 3). To establish a connection between semi-complete PDGs and acyclic tournaments, we introduce the notion of compatibility:

Definition 12 (Compatible Acyclic Tournament) *Given a semi-complete PDG, $\mathcal{S}$, an acyclic tournament, $\mathcal{T}$, is said to be compatible with $\mathcal{S}$ if:*

1. *$\mathcal{T}$ is obtained by reversing certain directed edges of $\mathcal{S}$ to produce an acyclic directed graph, while leaving all other edges unchanged.*
2. *$\mathcal{T}$ contains an acyclic orientation of all undirected edges.*

We can then identify a causal order that maximally aligns with the knowledge provided by the LLM by finding all MCATs compatible with the semi-complete PDG.

Definition 13 (Maximally consistent acyclic tournament) *Given a semi-complete PDG, $\mathcal{S}$, an acyclic tournament, $\mathcal{T}$, compatible with $\mathcal{S}$, is said to be maximally consistent if it maximizes the consistency score relative to all other acyclic tournaments compatible with $\mathcal{S}$.*

In this context, if $\mathcal{S}$ contains directed cycles or undirected edges, we need to obtain an acyclic transformation. Finding such a transformation is a well-studied problem in graph theory, known as the Feedback Arc Set (FAS) problem (Karp, 1972). In its general form, the FAS problem consists of identifying a minimal set of edges whose removal eliminates all cycles in a directed graph. The same set can be used to obtain an acyclic graph by reversing the edges instead of removing them (Barthélemy et al., 1995). In our setting, we aim to find the acyclic transformation of $\mathcal{S}$ that maximizes the consistency score. To this end, we consider a weighted version of the FAS, seeking the minimal set of edges whose reversal achieves the highest score. Instead of considering consistency scores as edges' weights, we define a cost score, $B_{i\to j}$, for each edge, $X_i \to X_j$, in $\mathcal{S}$:

$$B_{i\to j} = \xi - C_{X_i \succ X_j} + C_{X_j \succ X_i}.$$

where ξ is the total consistency score of $\mathcal{S}$, calculated as the sum of the consistency scores of all edges. The cost score $B_{i\to j}$ represents the cost of reversing the edge $X_i \to X_j$ in $\mathcal{S}$. The goal is to find the solution to the FAS that maximizes the total cost score, which is equivalent to maximizing

the total consistency score of the resulting acyclic tournament. Thus, we can identify the set of edges, A, whose reversal leads to an MCAT by solving the following optimization problem:

$$A = arg \max_{\hat{A} \in \mathcal{F}_{\mathcal{S}}} \sum_{X_i \succ X_j \in \hat{A}} B_{i \to j} \tag{2}$$

where $\mathcal{F}_{\mathcal{S}}$ is the set of all FAS of $\mathcal{S}$. The second step consists of orienting the undirected edges in $\mathcal{S}$. Since undirected edges have equal consistency scores in both directions, we can consider all acyclic orientations of the undirected edges as equally consistent —for more details, see Appendix B. In Figure 1 (d), we illustrate an MCAT compatible with the semi-complete PDG in Figure 1 (b).

Finding All Maximally Consistent Acyclic Tournaments. The FAS maximizing the total cost score and compatible with $\mathcal{S}$ is not necessarily unique. We can find all such sets by iteratively solving the weighted FAS while excluding previously found solutions. Let A be an optimal solution to (2), meaning that there is an MCAT, $\mathcal{T} = (V, E_{\mathcal{T}})$, such that $E_{\mathcal{T}} = (E_{\mathcal{S}} \cap \mathbb{A}) \cup \mathbb{A}^T$. We want to find all the acyclic tournaments $\mathcal{T}' = (V, E'_{\mathcal{T}})$ such that $E'_{\mathcal{T}} = (E_{\mathcal{S}} \cap \mathbb{A}') \cup \mathbb{A}'^T$ and $\mathbb{A}' \neq \mathbb{A}$, such that A' is a solution to (2). This can be obtained by excluding subsets of edges of the optimal solution, $\mathbb{F} \in \mathcal{P}(\mathbb{A})$, from the search space, i.e., we will search through all solutions to the FAS where the edges in $\mathbb{F}$ are prohibited from being reversed. If the exclusion of $\mathbb{F}$ leads to a suboptimal solution —non-maximal —then any other solution containing $\mathbb{F}$ will be suboptimal. This allows us to prune the search space efficiently. We denote this method as **M**aximal Weighted **A**cyclic **T**ournaments **S**earch, or MATS —more details about the procedure can be found in Appendix B.

Theorem 14 *The MATS algorithm is sound, complete, and terminates.*

Proposition 15 *If the consistency matrix is provided by a consistent knowledge base, the MATS algorithm is guaranteed to return a class of acyclic tournaments containing the true causal order.*

Proposition 16 *If the consistency matrix is provided by a strictly consistent knowledge base, the MATS algorithm is guaranteed to return the true causal order.*

Reasoning with Classes of Causal Orders. If a treatment variable X_i is part of a cycle in the Semi-Complete PDG, the total effect is not identifiable by deriving the associated MPDAG (Proposition 9 and 10). In this case, we can estimate the total effect using the class of maximally consistent causal orders. Given an estimated causal order, $\hat{\pi}$, we can use the predecessors of the treatment, X_i, to estimate the total effect over another variable, X_j, assuming that $\hat{\pi}(X_i) > \hat{\pi}(X_j)$ and causal sufficiency (Pearl, 2009; Vashishtha et al., 2025). This allows for efficient identification of the adjustment set without enumerating all backdoor paths from X_i to X_j. Moreover, we do not require the correct causal order to identify the total effect as long as all the predecessors are correctly ordered with respect to the treatment variable. We can generalize this to every maximally consistent causal order in the recovered class. In this context, two or more orders might share the same predecessors, e.g, if $\hat{\pi}(\mathbb{Z}) > \hat{\pi}(X_i) > \hat{\pi}(\mathbb{M}) > \hat{\pi}(X_j)$ and $\hat{\pi}(\mathbb{Z}) > \hat{\pi}(X_i) > \hat{\pi}(\mathbb{M}') > \hat{\pi}(X_j)$, meaning that the total effect is the same for both orders. This implies that we can efficiently estimate the total effect once for multiple orders sharing the same backdoor set. Ultimately, the total effect for an MPDAG is uniquely identifiable, whereas for a class of causal orders, we might have multiple estimations.

Graph	PC	GES	NOTEARS	BOSS	GRaSP	GOBNILP	Triplet	MATS (Our)	
								MPDAG	π
Asia	$3.66^{\pm 2.47}$	$1.75^{\pm 1.20}$	$4.08^{\pm 0.92}$	$1.50^{\pm 0.87}$	$1.50^{\pm 0.87}$	$4.31^{\pm 1.20}$	$2.75^{\pm 3.03}$	$2.87^{\pm 1.81}$	$1.22^{\pm 1.03}$
Cancer	$0.00^{\pm 0.00}$	$0.00^{\pm 0.00}$	$0.83^{\pm 0.12}$	$0.00^{\pm 0.00}$	$0.00^{\pm 0.00}$	$2.00^{\pm 0.86}$	$0.16^{\pm 0.37}$	$3.53^{\pm 1.86}$	$0.00^{\pm 0.00}$
Climate	$0.14^{\pm 0.35}$	$1.67^{\pm 3.13}$	$1.83^{\pm 0.17}$	$0.00^{\pm 0.00}$	$0.83^{\pm 0.50}$	$4.42^{\pm 1.21}$	$2.00^{\pm 0.00}$	$0.60^{\pm 0.80}$	$0.40^{\pm 0.80}$
Covid 1	$0.00^{\pm 0.00}$	$0.00^{\pm 0.00}$	$0.25^{\pm 0.08}$	$0.00^{\pm 0.00}$	$0.17^{\pm 0.14}$	$1.00^{\pm 0.74}$	$1.00^{\pm 0.0}$	$0.55^{\pm 0.78}$	$0.00^{\pm 0.00}$
Covid 2	$0.50^{\pm 0.50}$	$0.50^{\pm 0.50}$	$1.08^{\pm 0.49}$	$0.50^{\pm 0.50}$	$0.50^{\pm 0.50}$	$1.67^{\pm 0.88}$	$1.66^{\pm 0.47}$	$0.00^{\pm 0.00}$	$0.00^{\pm 0.00}$
Covid 3	$0.50^{\pm 0.50}$	$1.62^{\pm 1.41}$	$0.75^{\pm 0.42}$	$1.25^{\pm 0.42}$	$1.17^{\pm 0.75}$	$2.64^{\pm 1.25}$	$0.00^{\pm 0.00}$	$1.00^{\pm 0.50}$	$1.00^{\pm 0.50}$
Covid 4	$1.00^{\pm 0.73}$	$1.03^{\pm 0.72}$	$2.25^{\pm 0.67}$	$1.00^{\pm 0.71}$	$1.00^{\pm 0.71}$	$3.49^{\pm 1.49}$	$0.83^{\pm 0.68}$	$0.00^{\pm 0.00}$	$0.00^{\pm 0.00}$
Genetic	$0.00^{\pm 0.00}$	$0.00^{\pm 0.00}$	$0.50^{\pm 0.00}$	$0.00^{\pm 0.00}$	$0.25^{\pm 0.25}$	$2.75^{\pm 0.93}$	$0.00^{\pm 0.00}$	$1.10^{\pm 0.70}$	$1.20^{\pm 0.40}$
MSU	$1.29^{\pm 0.70}$	$1.56^{\pm 1.07}$	$1.33^{\pm 0.17}$	$1.00^{\pm 0.00}$	$1.00^{\pm 0.00}$	$3.00^{\pm 1.32}$	$1.33^{\pm 0.74}$	$3.22^{\pm 1.73}$	$2.00^{\pm 0.58}$
Neighbor	$5.65^{\pm 3.12}$	$0.00^{\pm 0.00}$	$2.33^{\pm 0.12}$	$0.88^{\pm 0.14}$	$1.21^{\pm 0.14}$	$4.30^{\pm 1.59}$	$3.00^{\pm 0.00}$	$5.96^{\pm 2.64}$	$4.89^{\pm 2.42}$
Sachs	$9.83^{\pm 3.17}$	$10.02^{\pm 3.12}$	$8.77^{\pm 1.73}$	$7.38^{\pm 1.97}$	$8.42^{\pm 1.61}$	$9.31^{\pm 2.57}$	$10.33^{\pm 9.19}$	$7.40^{\pm 2.04}$	$5.50^{\pm 0.50}$
Supermarket	$6.38^{\pm 2.63}$	$9.11^{\pm 3.64}$	$4.31^{\pm 1.30}$	$5.12^{\pm 1.19}$	$4.21^{\pm 0.88}$	$6.49^{\pm 1.83}$	$5.0^{\pm 2.23}$	$6.45^{\pm 2.58}$	$1.56^{\pm 1.26}$

Table 1: $\mathcal{D}_{top}$ (↓) of the estimated causal orders. Best results are highlighted in gray.

Graph	PC	GES	NOTEARS	BOSS	GRaSP	GOBNILP	Triplet	MATS (Our)	
								MPDAG	π
Asia	$8.37^{\pm 2.27}$	$6.62^{\pm 0.99}$	$6.06^{\pm 0.36}$	$7.44^{\pm 0.29}$	$7.25^{\pm 0.08}$	$6.94^{\pm 0.37}$	$2.75^{\pm 3.03}$	$2.87^{\pm 1.81}$	$1.22^{\pm 1.03}$
Cancer	$3.67^{\pm 0.47}$	$3.00^{\pm 0.00}$	$2.62^{\pm 0.49}$	$3.92^{\pm 0.08}$	$3.92^{\pm 0.08}$	$3.25^{\pm 0.63}$	$0.16^{\pm 0.37}$	$3.53^{\pm 1.86}$	$0.00^{\pm 0.00}$
Climate	$8.00^{\pm 2.52}$	$8.00^{\pm 0.00}$	$5.98^{\pm 0.43}$	$7.92^{\pm 0.08}$	$7.67^{\pm 0.00}$	$8.00^{\pm 0.00}$	$2.00^{\pm 0.00}$	$0.60^{\pm 0.80}$	$0.40^{\pm 0.80}$
Covid 1	$0.00^{\pm 0.00}$	$1.00^{\pm 0.00}$	$2.00^{\pm 0.00}$	$2.00^{\pm 0.00}$	$2.00^{\pm 0.00}$	$1.54^{\pm 0.37}$	$1.00^{\pm 0.0}$	$0.55^{\pm 0.78}$	$0.00^{\pm 0.00}$
Covid 2	$1.50^{\pm 0.50}$	$1.00^{\pm 0.00}$	$2.75^{\pm 0.20}$	$3.00^{\pm 0.00}$	$3.00^{\pm 0.00}$	$2.25^{\pm 0.59}$	$1.66^{\pm 0.47}$	$0.00^{\pm 0.00}$	$0.00^{\pm 0.00}$
Covid 3	$2.00^{\pm 0.00}$	$2.00^{\pm 0.00}$	$4.46^{\pm 0.34}$	$5.00^{\pm 0.00}$	$5.00^{\pm 0.00}$	$4.45^{\pm 0.45}$	$0.00^{\pm 0.00}$	$1.00^{\pm 0.50}$	$1.00^{\pm 0.50}$
Covid 4	$3.17^{\pm 1.67}$	$4.00^{\pm 0.00}$	$3.82^{\pm 0.31}$	$5.50^{\pm 0.14}$	$5.50^{\pm 0.14}$	$5.67^{\pm 0.33}$	$0.83^{\pm 0.68}$	$0.00^{\pm 0.00}$	$0.00^{\pm 0.00}$
Genetic	$4.00^{\pm 0.00}$	$6.00^{\pm 1.26}$	$2.90^{\pm 0.73}$	$4.92^{\pm 0.08}$	$4.92^{\pm 0.08}$	$3.65^{\pm 0.73}$	$0.00^{\pm 0.00}$	$1.10^{\pm 0.70}$	$1.20^{\pm 0.40}$
MSU	$1.50^{\pm 0.50}$	$3.40^{\pm 0.80}$	$4.74^{\pm 0.71}$	$5.92^{\pm 0.08}$	$5.92^{\pm 0.08}$	$4.22^{\pm 0.91}$	$1.33^{\pm 0.74}$	$3.22^{\pm 1.73}$	$2.00^{\pm 0.58}$
Neighbor	$6.67^{\pm 1.25}$	$7.20^{\pm 1.60}$	$5.65^{\pm 0.54}$	$7.75^{\pm 0.08}$	$7.75^{\pm 0.08}$	$7.00^{\pm 0.58}$	$3.00^{\pm 0.00}$	$5.96^{\pm 2.64}$	$4.89^{\pm 2.42}$
Sachs	$7.75^{\pm 1.20}$	$10.02^{\pm 3.12}$	$13.50^{\pm 0.99}$	$15.15^{\pm 0.61}$	$14.95^{\pm 0.76}$	$15.00^{\pm 1.08}$	$10.33^{\pm 9.19}$	$7.40^{\pm 2.04}$	$5.50^{\pm 0.50}$
Supermarket	$12.33^{\pm 2.36}$	$12.50^{\pm 0.50}$	$8.08^{\pm 0.58}$	$10.92^{\pm 0.08}$	$10.83^{\pm 0.17}$	$11.50^{\pm 0.50}$	$5.0^{\pm 2.23}$	$6.45^{\pm 2.58}$	$1.56^{\pm 1.26}$

Table 2: $\mathcal{D}_{top}$ (↓) of the estimated causal orders with non-linear synthetic data. Best results are highlighted in gray.

6. Experimental Results

The code[1] is designed to be compatible with OpenAI API and open-source LLMs available through the Ollama platform. Results presented in Table 1 are obtained using `gpt-4.1-nano`[2], a fast and lightweight version of `gpt`. Additional results relative to `mistral:7b`[3] and `llama3.1:8b`[4] are available in Table 10 of Appendix C. To minimize hallucinations and ensure a higher degree of consistency, all the experiments relative to LLMs have been conducted with *temperature* set to 0.1 —more details in Appendix C. Graphs are implemented using `igraph`, a `C++` library that offers an implementation of `ExactFAS` based on an integer programming formulation, which is guaranteed to yield an optimal result (Baharev et al., 2021) —more details in Appendix B. Note that finding a FAS is NP-complete, even though for larger graphs, one could also resort to heuristic methods (Eades et al., 1993), also available on `igraph`.

1. https://github.com/Federic0Bald0/MATS
2. https://platform.openai.com/docs/models/gpt-4.1-nano
3. https://ollama.com/library/mistral
4. https://ollama.com/library/llama3.1

Baselines. Concerning LLM-aided approaches, we compare to a state-of-the-art method for recovering causal orders, proposed in Vashishtha et al. (2025), which we refer to as the triplet method, or Triplet[5]. Moreover, we evaluate our approach against more traditional causal discovery methods. In particular, we consider the PC algorithm with Fisher's conditional independence test in the linear case and KCI for the non-linear one, GES[6], NOTEARS[7], GRaSP (Lam et al., 2022), BOSS (Andrews et al., 2023), and GOBNILP (Cussens, 2020), using 1000 data samples. Finally, we experimented with a hybrid approach that orients edges in the skeleton recovered by PC, using the orders obtained from MATS and Triplet —results in Appendix C.

Graphs & Datasets We tested MATS and the baselines on 12 causal graphs. Among these, 3 are well-known causal DAGs included in the `bnlearn` library, namely Asia (Lauritzen and Spiegelhalter, 2018), Cancer (Korb and Nicholson, 2004), and Sachs (Sachs et al., 2005). Additionally, we tested 9 causal graphs sourced from scientific literature in the fields of epidemiology and public health. These causal DAGs include Covid 1, Covid 2, Covid 3 (Griffith et al., 2020), Covid 4 (Glemain et al., 2025), Genetic (Palmer et al., 2012), MSU (Piccininni et al., 2023), Neighborhood (Chaix et al., 2009), Climate (Barrero Guevara et al., 2025), and Supermarket (Chaix et al., 2012). The datasets used for data-driven causal discovery, i.e., for PC, GES, and NOTEARS, and total effect estimation, are generated based on the true causal DAG. We rely on both linear and non-linear synthetic data, which is meaningful since text-driven methods do not require parametric assumptions —more details on the graphs and the synthetic datasets in Appendix B and D.

Evaluation Traditional metrics fall short in evaluating the error over the estimated causal orders. Indeed, there can be multiple causal orders that are consistent with the same causal DAG (Section 3). To this end, we rely on a metric proposed in Rolland et al. (2022) that measures how well an estimated ordering respects the true causal graph, $\mathcal{G}$. A correct causal order respects the ordering constraints of $\mathcal{G}$, meaning that if $X_i \succ_{\mathcal{G}} X_j$, then X_i must precede X_j in the estimated order $\hat{\pi}$. The metric is then defined as follows:

$$\mathcal{D}_{top}(\hat{\pi}, \mathcal{G}) = \sum_{(X_i, X_j) \in \mathcal{C}} \mathbf{1}[\hat{\pi}(X_i) > \hat{\pi}(X_j)] \quad \text{where} \quad \mathcal{C} = \{(X_i, X_j) | X_i \succ_{\mathcal{G}} X_j\}$$

where $\mathbf{1}$ is an indicator function measuring the number of violated constraints in $\mathcal{C}$. We have that $\mathcal{D}_{top}(\hat{\pi}, \mathcal{G}) = 0$ when $\hat{\pi}$ is a correct causal order for $\mathcal{G}$. Evaluating the estimated MCAT using more traditional metrics, such as SHD, does not capture the quality of the estimated orders, since the graph is dense. For completeness, we provide a comparison using SHD in Appendix C. Concerning causal effect estimation, we measure the absolute error between the estimated and true total effect, ϵ_{ATE}.

Results. Table 1, 2, and 3 report evaluation results for causal order recovery in the linear case, the non-linear case, and total effect estimation, respectively. Both tables show the mean and standard deviation over 5 runs for each method, using different random seeds. For methods that return abstractions (CPDAGs produced by PC, GES, BOSS, GRaSP, and GOBNILP; MPDAGs; or classes of causal orders) rather than a single causal order, the reported error is computed by averaging over all admissible orders across configurations. In the case of NOTEARS, which returns a single DAG,

5. We implemented this method based on the paper and discussions with the authors.
6. https://github.com/py-why/causal-learn
7. https://github.com/xunzheng/notears

Graph	PC	GES	NOTEARS	BOSS	GRaSP	GOBNILP	Triplet	MATS (Our)	
								MPDAG	Causal Orders
Asia	$0.26^{\pm 0.18}$	$0.26^{\pm 0.18}$	$0.16^{\pm 0.13}$	$0.16^{\pm 0.12}$	$0.17^{\pm 0.13}$	$0.19^{\pm 0.09}$	$0.00^{\pm 0.00}$	-	$0.00^{\pm 0.00}$
Cancer	$0.03^{\pm 0.01}$	$0.03^{\pm 0.01}$	$0.01^{\pm 0.00}$	$0.01^{\pm 0.00}$	$0.01^{\pm 0.00}$	$0.14^{\pm 0.08}$	$0.00^{\pm 0.00}$	-	$0.00^{\pm 0.00}$
Climate	$0.03^{\pm 0.03}$	$0.03^{\pm 0.03}$	$0.03^{\pm 0.01}$	$0.01^{\pm 0.00}$	$0.08^{\pm 0.06}$	$0.20^{\pm 0.11}$	$0.01^{\pm 0.00}$	-	$0.14^{\pm 0.24}$
Covid 1	$0.03^{\pm 0.01}$	$0.03^{\pm 0.01}$	$0.01^{\pm 0.00}$	$0.01^{\pm 0.00}$	$0.01^{\pm 0.00}$	$0.01^{\pm 0.00}$	$0.01^{\pm 0.00}$	$0.00^{\pm 0.00}$	$0.00^{\pm 0.00}$
Covid 2	$0.52^{\pm 0.24}$	$0.52^{\pm 0.24}$	$0.01^{\pm 0.00}$	$0.01^{\pm 0.00}$	$0.01^{\pm 0.00}$	$0.01^{\pm 0.00}$	$0.01^{\pm 0.00}$	$0.00^{\pm 0.00}$	$0.00^{\pm 0.00}$
Covid 3	$0.03^{\pm 0.02}$	$0.13^{\pm 0.23}$	$0.02^{\pm 0.01}$	$0.01^{\pm 0.00}$	$0.08^{\pm 0.06}$	$0.17^{\pm 0.15}$	$0.00^{\pm 0.00}$	$0.03^{\pm 0.06}$	$0.03^{\pm 0.06}$
Covid 4	$0.03^{\pm 0.01}$	$0.05^{\pm 0.04}$	$0.14^{\pm 0.02}$	$0.07^{\pm 0.08}$	$0.07^{\pm 0.08}$	$0.12^{\pm 0.05}$	$0.13^{\pm 0.15}$	$0.01^{\pm 0.01}$	$0.00^{\pm 0.00}$
Genetic	$0.02^{\pm 0.01}$	$0.02^{\pm 0.01}$	$0.01^{\pm 0.00}$	$0.01^{\pm 0.00}$	$0.01^{\pm 0.00}$	$0.01^{\pm 0.00}$	$0.00^{\pm 0.00}$	-	$0.00^{\pm 0.00}$
MSU	$0.19^{\pm 0.14}$	$0.19^{\pm 0.14}$	$0.26^{\pm 0.00}$	$0.26^{\pm 0.00}$	$0.26^{\pm 0.00}$	$0.41^{\pm 0.12}$	$0.25^{\pm 0.14}$	-	$0.25^{\pm 0.14}$
Neighbor	$0.45^{\pm 0.37}$	$0.26^{\pm 0.17}$	$0.17^{\pm 0.00}$	$0.18^{\pm 0.00}$	$0.14^{\pm 0.00}$	$0.25^{\pm 0.07}$	$0.28^{\pm 0.17}$	-	$0.16^{\pm 0.07}$
Sachs	$0.08^{\pm 0.08}$	$0.15^{\pm 0.10}$	$0.23^{\pm 0.14}$	$0.18^{\pm 0.12}$	$0.21^{\pm 0.02}$	$0.18^{\pm 0.08}$	$0.04^{\pm 0.05}$	-	$0.11^{\pm 0.07}$
Supermarket	$0.40^{\pm 0.35}$	$0.47^{\pm 0.52}$	$0.43^{\pm 0.15}$	$0.37^{\pm 0.07}$	$0.42^{\pm 0.09}$	$0.45^{\pm 0.14}$	$0.30^{\pm 0.31}$	-	$0.46^{\pm 0.25}$

Table 3: ϵ_{ATE} ($\downarrow$) of total effect estimation. Best results are highlighted in gray. If non identifiable the value is set to '-'.

we first recover the corresponding CPDAG and evaluate the error over that, so as not to penalize it for returning a single representative of its equivalence class —results relative to the recovered DAG are provided in Appendix C. Additionally, this evaluation may penalize methods that return large equivalence classes with high variance. To provide a more complete picture, additional intraclass estimation results are reported in Appendix C. The results in Table 1, for PC, GES, NOTEARS, BOSS, GRaSP, and GOBNILP, as well as for total effect estimation, are obtained using linearly generated data, a setting that is favorable to traditional causal discovery methods. Indeed, as shown in Table 2, their performance deteriorates substantially when data are generated from a non-linear model. In contrast, text-based methods such as MATS and Triplet do not rely on parametric assumptions. In the linear setting, MATS achieves the lowest $\mathcal{D}_{top}$ error in 7 out of the 12 graphs, while in the non-linear setting it outperforms all other methods in 8 out of the 12 graphs. Moreover, in 4 graphs, MATS recovers a class containing exclusively correct causal orders, both in the linear and non-linear case; in two additional cases, the resulting MPDAG is an acyclic tournament and therefore requires no further edge orientation. Total effects are estimated using the adjustment formula in (1), implemented via linear regression. Adjustment sets for data-driven methods are identified on the estimated graph using the generalized backdoor criterion (Definition 3). The resulting estimation errors are reported in Table 3. Overall, we observe a clear correlation between the topological distance $\mathcal{D}_{top}$ and the absolute error in total effect estimation, ϵ_{ATE}. Finally, we note that in the majority of cases, the estimated MPDAG does not permit identification of the total effect.

7. Discussion & Conclusion

The method described in this paper offers an approach leveraging LLMs as knowledge bases for retrieving abstractions of causal orders. This approach relies solely on textual descriptions of the variables and does not require faithfulness, parametric assumptions, or observational data. A key insight behind the method is that natural language often leaves causal mechanisms implicit. This implies that some causal relationships may not be directly stated in the text, yet the causal order remains intact. We compared our methods with traditional approaches, such as PC, GES, NOTEARS, BOSS, GRaSP, and GOBNILP, as well as an LLM-aided method recovering causal order (Vashishtha et al.,

2025). The results show that our approach can provide an accurate estimation most of the time, outperforming other methods.

Background Knowledge. Note that the proposed method is compatible with the use of background knowledge —for instance, orientations provided by human experts or temporal priority. In particular, we can force orderings among variables by assigning specific values in the consistency matrix. For instance, if we know that X_i is a predecessor of X_j: we can set $C_{X_i \succ X_j} = 1$, being maximally consistent, while $C_{X_j \succ X_i} = -\infty$, effectively removing the edge $X_j \to X_i$ from every maximally consistent graph.

Limitations. The MATS algorithm has several limitations that should be taken into account. Indeed, the accuracy of the estimation strongly relies on the consistency of the LLM. In the presence of an inconsistent knowledge base, we cannot guarantee the correctness of the class of orders retrieved. Moreover, the computational complexity of the method can increase significantly with larger graphs. This is primarily due to the computation of the consistency matrix, which has a quadratic complexity in relation to the number of nodes, and the ExactFAS algorithm, which is an NP-hard problem. Note, however, that MATS can be made more efficient by introducing heuristic solvers for the FAS problem.

Future Work. Extensions of this work will focus on reducing computational complexity and enhancing the reliability of the LLM. The use of chain-of-thought based LLMs (Wei et al., 2022) could potentially improve the accuracy of the method. Moreover, in scenarios where a collection of documents relevant to a specific domain is available, we can enhance the reliability of the knowledge base by using Retrieval-Augmented Generation (RAG), proposed by Lewis et al. (2020). This approach grounds replies generated by the LLM on a specific corpus of documents.

Acknowledgments

This work was supported by the CIPHOD project (ANR-23-CPJ1-0212-01).

References

Bryan Andrews, Joseph Ramsey, Ruben Sanchez Romero, Jazmin Camchong, and Erich Kummerfeld. Fast scalable and accurate discovery of dags using the best order score search and grow shrink trees. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 63945–63956. Curran Associates, Inc., 2023.

Charles K. Assaad, Emilie Devijver, and Eric Gaussier. Survey and evaluation of causal discovery methods for time series. *J. Artif. Intell. Res.*, 73:767–819, 2022.

Ali Baharev, Hermann Schichl, Arnold Neumaier, and Tobias Achterberg. An exact method for the minimum feedback arc set problem. *ACM J. Exp. Algorithmics*, 26, April 2021. ISSN 1084-6654. doi: 10.1145/3446429.

Laura Andrea Barrero Guevara, Sarah C. Kramer, Tobias Kurth, and Matthieu Domenech de Cellès. Causal inference concepts can guide research into the effects of climate on infectious

diseases. *Nature Ecology & Evolution*, 9(2):349–363, Feb 2025. ISSN 2397-334X. doi: 10.1038/s41559-024-02594-3.

Jean-Pierre Barthélemy, Olivier Hudry, Garth Isaak, Fred S. Roberts, and Barry Tesman. The reversing number of a diagraph. *Discrete Applied Mathematics*, 60(1):39–76, 1995. ISSN 0166-218X. doi: https://doi.org/10.1016/0166-218X(94)00042-C.

Basile Chaix, Cinira Leal, and David Evans. Neighborhood-level confounding in epidemiologic studies unavoidable challenges, uncertain solutions. *Epidemiology (Cambridge, Mass.)*, 21:124–7, 11 2009. doi: 10.1097/EDE.0b013e3181c04e70.

Basile Chaix, Kathy Bean, Mark Daniel, Shannon Zenk, Yan Kestens, Hélène Charreire, Cinira Leal, Frédérique Thomas, Noëlla Karusisi, Christiane Weber, Jean-Michel Oppert, Chantal Simon, Juan Merlo, and Bruce Pannier. Associations of supermarket characteristics with weight status and body fat: A multilevel analysis of individuals within supermarkets (record study). *PloS one*, 7:e32908, 04 2012. doi: 10.1371/journal.pone.0032908.

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

Kai-Hendrik Cohrs, Emiliano Diaz, Vasileios Sitokonstantinou, Gherardo Varando, and Gustau Camps-Valls. Large language models for causal hypothesis generation in science. *Machine Learning: Science and Technology*, 6(1):013001, January 2025. doi: 10.1088/2632-2153/ada47f. Publisher: IOP Publishing.

James Cussens. Gobnilp: Learning bayesian network structure with integer programming. In Manfred Jaeger and Thomas Dyhre Nielsen, editors, *Proceedings of the 10th International Conference on Probabilistic Graphical Models*, volume 138 of *Proceedings of Machine Learning Research*, pages 605–608. PMLR, 23–25 Sep 2020.

Peter Eades, Xuemin Lin, and W. F. Smyth. A fast and effective heuristic for the feedback arc set problem. *Inf. Process. Lett.*, 47(6):319–323, October 1993. ISSN 0020-0190. doi: 10.1016/0020-0190(93)90079-O.

Benjamin Glemain, Charles Assaad, Walid Ghosn, Paul Moulaire, Xavier de Lamballerie, Marie Zins, Gianluca Severi, Mathilde Touvier, Jean-François Deleuze, Nathanaël Lapidus, Fabrice Carrat, Pierre-Yves Ancel, Marie-Aline Charles, Sofiane Kab, Adeline Renuy, Stephane LeGot, Celine Ribet, Mireille Pellicer, Emmanuel Wiernik, Marcel Goldberg, Fanny Artaud, Pascale Gerbouin-Rérolle, Mélody Enguix, Camille Laplanche, Roselyn Gomes-Rima, Lyan Hoang, Emmanuelle Correia, Alpha Amadou Barry, Nadège Senina, Julien Allegre, Fabien Szabo de Edelenyi, Nathalie Druesne-Pecollo, Younes Esseddik, Serge Hercberg, Mélanie Deschasaux, Valérie Benhammou, Anass Ritmi, Laetitia Marchand, Cecile Zaros, Elodie Lordmi, Adriana Candea, Sophie de Visme, Thierry Simeon, Xavier Thierry, Bertrand Geay, Marie-Noelle Dufourg, Karen Milcent, Delphine Rahib, Nathalie Lydie, Clovis Lusivika-Nzinga, Gregory Pannetier, Nathanael Lapidus, Isabelle Goderel, Céline Dorival, Jérôme Nicol, Olivier Robineau, Cindy Lai, Liza Belhadji, Hélène Esperou, Sandrine Couffin-Cadiergues, Jean-Marie Gagliolo, Hélène Blanché, Jean-Marc Sébaoun, Jean-Christophe Beaudoin, Laetitia Gressin, Valérie Morel,

Ouissam Ouili, Laetitia Ninove, Stéphane Priet, Paola Mariela Saba Villarroel, Toscane Fourié, Souand Mohamed Ali, Abdenour Amroun, Morgan Seston, Nazli Ayhan, Boris Pastorino, and SAPRIS-SERO study group. Revisiting the link between covid-19 incidence and infection fatality rate during the first pandemic wave. *Scientific Reports*, 15(1):15638, May 2025. ISSN 2045-2322. doi: 10.1038/s41598-025-99078-6.

Clark Glymour, Kun Zhang, and Peter Spirtes. Review of causal discovery methods based on graphical models. *Frontiers in Genetics*, 10, 2019. ISSN 1664-8021. doi: 10.3389/fgene.2019.00524.

Gareth J. Griffith, Tim T. Morris, Matthew J Tudball, Annie Herbert, Giulia Mancano, Lindsey Pike, Gemma C. Sharp, Jonathan A. C. Sterne, Tom M. Palmer, George Davey Smith, Kate Tilling, Luisa Zuccolo, Neil Martin Davies, and Gibran Hemani. Collider bias undermines our understanding of covid-19 disease risk and severity. *Nature Communications*, 11, 2020.

Thomas Jiralerspong, Xiaoyin Chen, Yash More, Vedant Shah, and Yoshua Bengio. Efficient causal graph discovery using large language models. In *ICLR 2024 Workshop: How Far Are We From AGI*, 2024.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, Deep Ganguli, Danny Hernandez, Josh Jacobson, Jackson Kernion, Shauna Kravec, Liane Lovitt, Kamal Ndousse, Catherine Olsson, Sam Ringer, Dario Amodei, Tom Brown, Jack Clark, Nicholas Joseph, Ben Mann, Sam McCandlish, Chris Olah, and Jared Kaplan. Language models (mostly) know what they know, 2022.

Richard M Karp. Reducibility among combinatorial problems. In *50 Years of Integer Programming 1958-2008: from the Early Years to the State-of-the-Art*, pages 219–241. Springer, 1972.

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*, 2024. ISSN 2835-8856. Featured Certification.

Kevin B. Korb and Ann E. Nicholson. Bayesian artificial intelligence. 2004.

Wai-Yin Lam, Bryan Andrews, and Joseph Ramsey. Greedy relaxations of the sparsest permutation algorithm. In James Cussens and Kun Zhang, editors, *Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence*, volume 180 of *Proceedings of Machine Learning Research*, pages 1052–1062. PMLR, 01–05 Aug 2022.

S. L. Lauritzen and D. J. Spiegelhalter. Local computations with probabilities on graphical structures and their application to expert systems. *Journal of the Royal Statistical Society: Series B (Methodological)*, 50(2):157–194, 12 2018. ISSN 0035-9246. doi: 10.1111/j.2517-6161.1988.tb01721.x.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks. In H. Larochelle,

M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc., 2020.

Dan-dan Li, Yang Yang, Zi-yi Gao, Li-hua Zhao, Xue Yang, Feng Xu, Chao Yu, Xiu-lin Zhang, Xue-Qin Wang, Li-hua Wang, and Jian-Bin Su. Sedentary lifestyle and body composition in type 2 diabetes. *Diabetology & Metabolic Syndrome*, 14, 01 2022. doi: 10.1186/s13098-021-00778-6.

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

Stephanie Long, Tibor Schuster, and Alexandre Piché. Can large language models build causal graphs? *arXiv preprint arXiv:2303.05279*, 2023b.

Marloes H. Maathuis and Diego Colombo. A generalized back-door criterion. *The Annals of Statistics*, 43(3):1060 – 1088, 2015. doi: 10.1214/14-AOS1295.

Potsawee Manakul, Adian Liusie, and Mark Gales. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9004–9017, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.557.

Christopher Meek. Causal inference and causal explanation with background knowledge. In *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, UAI'95, page 403–410, San Francisco, CA, USA, 1995. Morgan Kaufmann Publishers Inc. ISBN 1558603859.

Tom M. Palmer, Deborah A. Lawlor, Roger M. Harbord, Nuala A. Sheehan, Jonathan H. Tobias, Nicholas John Timpson, George Davey Smith, and Jonathan A. C. Sterne. Using multiple genetic variants as instrumental variables for modifiable risk factors. *Statistical Methods in Medical Research*, 21:223 – 242, 2012.

Judea Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, Cambridge, 2 edition, 2009. doi: https://doi.org/10.1017/CBO9780511803161.

Emilija Perkovic. Identifying causal effects in maximally oriented partially directed acyclic graphs. 124:530–539, 03–06 Aug 2020.

Emilija Perković, Markus Kalisch, and Maloes H Maathuis. Interpreting and using cpdags with background knowledge. *Association for Uncertainty in Artificial Intelligence (UAI)*, 2017.

Emilija Perkovic, Johannes Textor, Markus Kalisch, and Marloes H. Maathuis. Complete graphical characterization and construction of adjustment sets in markov equivalence classes of ancestral graphs. *J. Mach. Learn. Res.*, 18(1):8132–8193, January 2017. ISSN 1532-4435.

Jonas Peters, Dominik Janzing, and Bernhard Schlkopf. *Elements of Causal Inference: Foundations and Learning Algorithms*. The MIT Press, 2017. ISBN 0262037319.

Marco Piccininni, Tobias Kurth, Heinrich J Audebert, and Jessica L Rohmann. The effect of mobile stroke unit care on functional outcomes: an application of the front-door formula. *Epidemiology*, 34(5):712–720, 2023.

Paul Rolland, Volkan Cevher, Matthäus Kleindessner, Chris Russell, Dominik Janzing, Bernhard Schölkopf, and Francesco Locatello. Score matching enables causal discovery of nonlinear additive noise models. 162:18741–18753, 17–23 Jul 2022.

Karen Sachs, Omar D. Perez, Dana Pe'er, Douglas A. Lauffenburger, and Garry P. Nolan. Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, 308:523 – 529, 2005.

Thomas Savage, John Wang, Robert Gallo, Abdessalem Boukil, Vishwesh Patel, Seyed Amir Ahmad Safavi-Naini, Ali Soroush, and Jonathan H Chen. Large language model uncertainty measurement and calibration for medical diagnosis and treatment. *medRxiv*, pages 2024–06, 2024.

Chenglei Si, Diyi Yang, and Tatsunori Hashimoto. Can LLMs generate novel research ideas? a large-scale human study with 100+ NLP researchers. In *The Thirteenth International Conference on Learning Representations*, 2025.

Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, prediction, and search*. MIT press, 2001.

Aniket Vashishtha, Abbavaram Gowtham Reddy, Abhinav Kumar, Saketh Bachu, Vineeth N. Balasubramanian, and Amit Sharma. Causal order: The key to leveraging imperfect experts in causal inference. In *The Thirteenth International Conference on Learning Representations*, 2025.

Aparajithan Venkateswaran and Emilija Perković. Towards complete causal explanation with expert knowledge, 2024.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed H. Chi, Quoc V Le, and Denny Zhou. Chain of thought prompting elicits reasoning in large language models. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022.

Matej Zečević, Moritz Willig, Devendra Singh Dhami, and Kristian Kersting. Causal parrots: Large language models may talk causality but are not causal. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856.

Danna Zheng, Mirella Lapata, and Jeff Z. Pan. Large language models as reliable knowledge bases? *CoRR*, abs/2407.13578, 2024.

Appendix

The appendix is organized as follows:

- Appendix A, contains the proofs of the propositions and theorems presented in the main text.
- Appendix B, contains details regarding the implementation of the MATS algorithm.
- Appendix C, contains additional results.
- Appendix D, contains details on the graphs used in the experiments and the prompts used to query the LLM.

Appendix A. Proofs

Proof [Proposition 7] Given a DAG $\mathcal{G} = (\mathbb{V}, \mathbb{E})$, a consistent knowledge base induces a consistency matrix W such that: for every pair of variables X_i and X_j, if $X_i \succ_{\mathcal{G}} X_j$ then $C_{X_i \succ X_j} \geq C_{X_j \succ X_i}$. If we build a semi-complete PDG $\mathcal{S}$ from the consistency matrix W, it follows that:

- If $C_{X_i \succ X_j} > C_{X_j \succ X_i}$, then we orient the edge as $X_i \to X_j$; the ensemble of edges oriented in this way does not contain directed cycles since it is a directed reflection of the DAG $\mathcal{G}$.
- If $C_{X_i \succ X_j} = C_{X_j \succ X_i}$, then we orient the edge as $X_i - X_j$; this does not introduce any directed cycle.

■

Proof [Proposition 8] The proof follows from Proposition 7. If the expert is consistent, the maximally consistent semi-complete PDG $\mathcal{S}$ does not contain directed cycles, meaning that it is a dense Partially Directed Acyclic Graph (PDAG). Additionally, since the graph is fully connected, there cannot be unshielded colliders. Thus, we can use R2 exclusively to maximize the orientation of the PDAG. ■

Proof [Proposition 9] Substituting directed edges in a cycle with undirected ones transforms the Semi-Complete PDG into a dense PDAG. Applying the Meek rules, in particular R2, to the PDAG, we get a maximally oriented PDAG (Proposition 8). ■

Proof [Proposition 10] Since the MPDAG is dense, between every pair of variables there is an edge, either directed or undirected. If the treatment and the outcome are connected by an undirected edge, it means that the total effect is not identifiable by adjustment. If the treatment and the outcome are connected by a directed edge, we can reduce the problem of identifiability to triplets of variables. If any of the subgraphs containing the treatment, the outcome, and any other variables are of the form in Figure 2, we can see that we cannot distinguish between a collider and a non-collider. Thus, the total effect is not identifiable. ■

Proof [Corollary 11] The proof follows from Proposition 10. ■

Proof [Theorem 14] We define the following notation in reference to Algorithm 1:

- `maxScore` is the maximal consistency score; thus, given a maximally consistent acyclic tournament $\mathcal{T} = (\mathbb{V}, \mathbb{E}_{\mathcal{T}})$:

$$\texttt{maxScore} = \sum_{(i,j) \in \mathbb{E}_{\mathcal{T}}} W[i,j]$$

- The method `ExactFAS` returns a FAS of a directed graph $\mathcal{G}$, which is a set of edges that can be reversed to make $\mathcal{G}$ acyclic and of maximal weight. If applied to a maximally consistent semi-complete partially directed graph $\mathcal{S}$, it returns a set $\mathbb{A}_i$ for $\mathcal{S}$ with respect to the consistency matrix W. If $\mathbb{A}_i$ is optimal, meaning that leads to a maximally consistent acyclic tournament, then it holds that:

$$\mathcal{T}_i = (\mathbb{V}, (\mathbb{E}_{\mathcal{S}} \setminus \mathbb{A}_i) \cup \mathbb{A}_i^T)$$

where $\mathbb{A}_i^T$ is the transpose of $\mathbb{A}_i$.

$$\texttt{score}(\mathcal{T}_i) = \texttt{maxScore}$$

Soundness. *Every tournament $\mathcal{T}$ in* `Results` *is a maximally consistent acyclic tournament.*

- **Initialization:** The algorithm starts by computing the maximally consistent semi-complete partially directed graph $\mathcal{S}$. Then, it computes a FAS $\mathbb{A}_0$ of $\mathcal{S}$ with respect to the consistency matrix W. By definition of `ExactFAS`, a FAS $\mathbb{A}_0$ is a set of edges that can be reversed to transform $\mathcal{S}$ into a maximally weighted acyclic tournament, $\mathcal{T}^0$; the tournament is then added to `Results` and the maximal consistency score `maxScore` is computed.

- **Iteration:** At every iteration, we compute a FAS of the semi-complete partially directed graph $\mathcal{S}$ with respect to the consistency matrix W' in which some edges have been excluded from the solution space. a FAS A is computed by `ExactFAS`$(\mathbb{E}_{\mathcal{S}}, W')$. The acyclic tournament $\mathcal{T}_A$ and its score are then computed. If the score of $\mathcal{T}_A$ is equal to `maxScore`, it is maximally consistent, thus it is added to `Results`.

Completeness. *Every maximally consistent acyclic tournament $\mathcal{T}$ is in* `Results`.

The solution space is explored by exclusion. The idea is that any maximally consistent solution is unique, thus removing from the solution space subsets of optimal FAS will force `ExactFAS` to search for other maximally consistent solutions. This is achieved by iterating over the power set of the optimal FAS, contained in `maximalFAS`. To exclude a set of edges $\mathbb{F}$ from the admissible solutions, we set the cost of reversing edges in $\mathbb{F}$ to $-\infty$, making them suboptimal by construction. So we can build W' as a copy of W where $W[j,i] = -\infty$ for every edge $(i,j) \in \mathbb{F}$. If there is another maximally consistent solution, A', such that $\mathbb{F} \not\subset \mathbb{A}'$, then it will be found in the next iteration of the algorithm by `ExactFAS`$(\mathcal{S}, W')$. We define $\mathbb{F}$ as the union of subsets of optimal solutions, $\mathbb{A}_i \in$ `maximalFAS`.

$$\mathbb{F} = \mathbb{F}_0 \cup \mathbb{F}_1 \cup \ldots \cup \mathbb{F}_k$$

where $\mathbb{F}_i \subset \mathbb{A}_i$. Given an undiscovered set $\mathbb{F}$ of forbidden edges:

- $\mathbb{F} = \emptyset$, which corresponds to the case where the algorithm has not yet explored any subset of the optimal FAS. The algorithm will compute the first FAS $\mathbb{A}_0$ and add it to `Results`.

- $\mathbb{F} \neq \emptyset$, meaning that the algorithm has already explored some subsets of the optimal FAS. In this case, the algorithm will compute a FAS $\mathbb{A}_i$ that is either: 1) optimal, meaning that it is a maximally consistent acyclic tournament, or 2) suboptimal, meaning that it is not a maximally consistent acyclic tournament. In the first case, $\mathbb{A}_i$ will be added to `Results`; moreover, we will add to the `Queue` the combinatorial union between $\mathbb{F}$ and all the subsets of $\mathbb{A}_i$, which represents a tighter constraint on the admissible solutions. Eventually, this will lead to the discovery of a new maximally consistent acyclic tournament.

Termination. *The MATS algorithm terminates.*

MATS systematically explores the power set of the optimal FAS, which is finite. Thus, the algorithm will eventually terminate. Additionally, the algorithm adopts a caching mechanism to avoid exploring the same subsets multiple times. Also, any set leading to a suboptimal solution allows us to identify sets of edges that can be excluded from the search space. Indeed, any set containing a subset of edges that leads to a suboptimal solution will not be explored.

■

Proof [Proposition 15] Following Proposition 7, the maximally consistent semi-complete PDG, $\mathcal{S}$, does not contain directed cycles. Moreover, since the consistency matrix is provided by a consistent

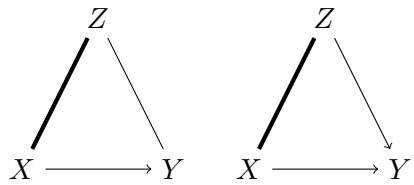


Figure 2: Non-identifiable subgraphs. In bold, the undirected edges are related to the treatment variable.

knowledge base, it holds that for every pair of variables X_i and X_j, if $X_i \succ X_j$ then $C_{X_i \succ X_j} > C_{X_j \succ X_i}$. The MATS algorithm will find all acyclic orientations of the undirected edges in $\mathcal{S}$ that do not introduce new cycles. Among these, at least one will be the true causal order. Note that in the limit case where every edge is undirected, MATS will return the list of all acyclic tournaments. ■

Proof [Proposition 16] The proof follows from Proposition 10. A strictly consistent expert guarantees that for every pair of variables X_i and X_j, such that $X_i \succ X_j$, it holds that $C_{X_i \succ X_j} > C_{X_j \succ X_i}$. Meaning that the maximally consistent semi-complete partially directed graph $\mathcal{S}$ is an acyclic tournament, since there cannot be undirected edges. ■

Appendix B. Implementation Details

Undirected Edges. To reduce computational time, undirected edges are handled separately. Before applying Algorithm 1, all undirected edges are removed from the graph $\mathcal{S}$. After generating a class of acyclic tournaments, each undirected edge is reintroduced as a directed one, oriented in one of its possible directions, as long as it doesn't create new cycles. Since undirected edges contribute equally to consistency, their orientation does not change the consistency score of the maximal acyclic tournament, but reduces the size of the SCCs processed by `ExactFAS`.

Data Generation. For each node x in the DAG,

$$x = f_x(Parents(x)) + \epsilon_x,$$

where ϵ_x is sampled from a Gaussian distribution, $Parents(x)$ is the set of parents of x in the graph; f_x in the linear case is of the form $f_x = \sum_{i=1}^{k} x_i m_i$, whereas in the non-linear one is a function randomly picked from $\{sin(.), cos(.), square(.)\}$.

Appendix C. Additional Results

This section provides further experimental results and analysis that complement those presented in the main text.

- Figure 3 shows the results of MATS using `gpt-4.1-nano` with different temperatures. We can see that a lower temperature leads to a lower $\mathcal{D}_{top}$. Indeed, a lower temperature leads to more deterministic outputs, reducing hallucinations.

Algorithm 1 Maximally Weighted Acyclic Tournaments Search (MATS)

Input: $\mathbb{V}$, variables; W, weights

Output: $Results$, a set of maximally weighted acyclic tournaments

$\mathcal{S} \leftarrow$ MAXIMALLYWEIGHTEDGRAPH$(\mathbb{V}, W)$
$(\mathbb{V}, \mathbb{E}_{\mathcal{S}}) \leftarrow \mathcal{S}$
$\mathbb{A} \leftarrow$ EXACTFAS$(\mathbb{E}_{\mathcal{S}}, W)$ {Find feedback arc set}
$\mathbb{E}_{\mathcal{T}} \leftarrow (\mathbb{E}_{\mathcal{S}} \setminus \mathbb{A}) \cup \mathbb{A}^T$
$maxScore \leftarrow \sum_{(i,j)\in\mathbb{E}_{\mathcal{T}}} W[i,j]$
$Queue \leftarrow \mathcal{P}(\mathbb{A})$ {Power set of $\mathbb{A}$}
$maximalFAS \leftarrow \{\mathbb{A}\}$
$Results \leftarrow \{\mathbb{E}_{\mathcal{T}}\}$
$Cache \leftarrow \emptyset$
while $Queue \neq \emptyset$ **do**
  $\mathbb{F} \leftarrow$ POP$(Queue)$
  $Cache \leftarrow Cache \cup \mathbb{F}$
  $W' \leftarrow$ COPY(W)
  for $(i,j) \in \mathbb{F}$ **do**
    $W'[j,i] \leftarrow -\infty$
  end for
  $\mathbb{A} \leftarrow$ EXACTFAS$(\mathbb{E}_{\mathcal{S}}, W')$
  $\mathbb{E}_{\mathcal{T}} \leftarrow (\mathbb{E}_{\mathcal{S}} \setminus \mathbb{A}) \cup \mathbb{A}^T$
  $newScore \leftarrow \sum_{(i,j)\in\mathbb{E}_{\mathcal{T}}} W[i,j]$
  if $newScore = maxScore$ **and** $\mathbb{A} \notin maximalFAS$ **then**
    $maximalFAS \leftarrow maximalFAS \cup \{\mathbb{A}\}$
    $Results \leftarrow Results \cup \{\mathbb{E}_{\mathcal{T}}\}$
    for $\mathbb{F}' \in \mathcal{P}(\mathbb{A})$ **do**
      if $\mathbb{F}' \cup \mathbb{F} \notin Cache$ **and** $\mathbb{F}' \cup \mathbb{F} \notin Queue$ **then**
        PUSH(Queue, $\mathbb{F}' \cup \mathbb{F}$)
      end if
    end for
  else
    for $\mathbb{F}' \in Queue$ **do**
      if $\mathbb{F} \subset \mathbb{F}'$ **then**
        REMOVE(Queue, $\mathbb{F}'$)
      end if
      $Cache \leftarrow Cache \cup \{\mathbb{F}'\}$
    end for
  end if
end while
return $Results$

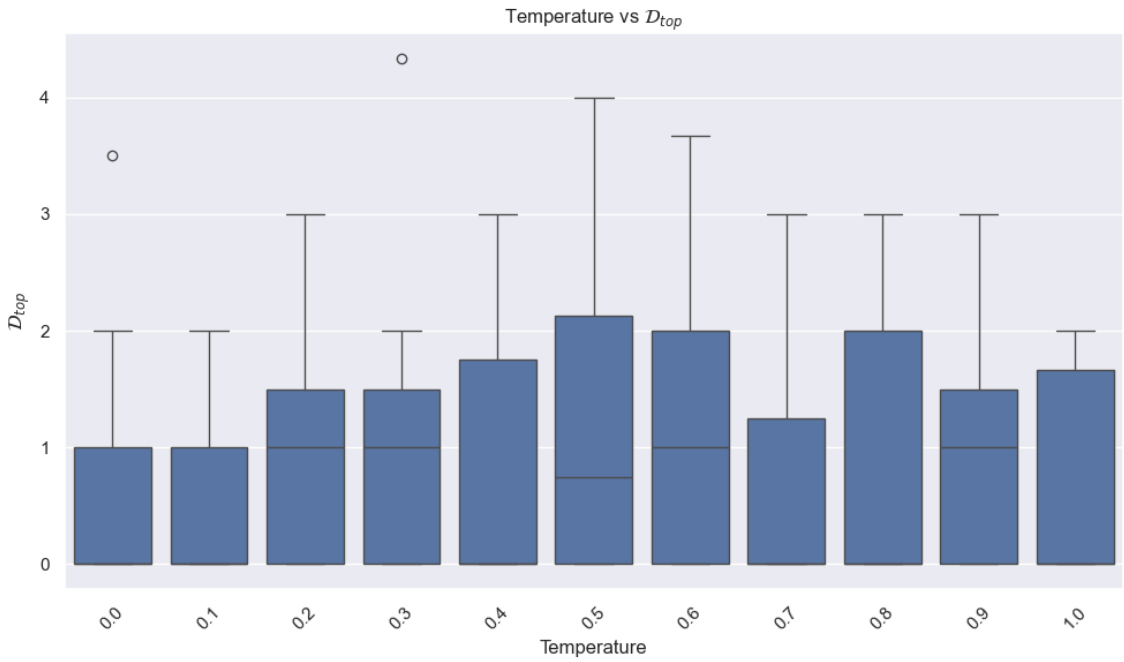


Figure 3: Boxplot intraclass $\mathcal{D}_{top}$ of MATS using `gpt-4.1-nano` with different temperatures. Error is bounded to values higher than 0. Note that when the median is not visible signifies it is equal to 0.

- Tables 4 complements results in Table 1 by providing SHD results on linear data. As argued in the main text, MATS provides a fully connected DAG, which leads to higher SHD values compared to data-driven baselines that return sparser graphs.
- Tables 5 provides the SHD associated with non-linear data for data-driven methods. We can see how MATS does not consistently outperform other methods, mainly due to the fact that we are recovering tournaments.
- Tables 6 and 7 present results of the hybridized approach using PC to estimate the skeleton, while using the approximated orders from MATS and Triplets to orient the edges. In this context, we can see a general improvement of the error associated with the estimated DAG.
- Table 8 and 9 compare MATS to NOTEARS. In this context, we display both the error associated with the DAG discovered by NOTEARS and the relative CPDAG.
- Table 10 showcases the best results obtained for intraclass $\mathcal{D}_{top}$ using different LLMs to compute the consistency matrix. We can see that generally, `gpt-4.1-nano` and `mistral` perform similarly, followed by `llama3.1`. Most importantly, in many instances, MATS recovers only correct orders.

Graph	PC	GES	NOTEARS	BOSS	GRaSP	GOBNILP	Triplet	MATS (Our)	
								MPDAG	π
Asia	$4.39^{\pm 2.34}$	$3.00^{\pm 1.73}$	$5.83^{\pm 1.84}$	$3.00^{\pm 1.73}$	$3.00^{\pm 1.73}$	$14.50^{\pm 2.41}$	$19.75^{\pm 3.90}$	$24.00^{\pm 2.25}$	$22.00^{\pm 1.63}$
Cancer	$0.00^{\pm 0.00}$	$0.00^{\pm 0.00}$	$1.00^{\pm 0.24}$	$0.00^{\pm 0.00}$	$0.00^{\pm 0.00}$	$7.50^{\pm 1.72}$	$3.50^{\pm 0.76}$	$9.73^{\pm 1.61}$	$6.00^{\pm 0.00}$
Climate	$0.44^{\pm 0.68}$	$1.56^{\pm 2.95}$	$2.00^{\pm 0.33}$	$0.00^{\pm 0.00}$	$2.50^{\pm 1.01}$	$12.83^{\pm 2.33}$	$24.00^{\pm 0.00}$	$21.20^{\pm 1.60}$	$20.80^{\pm 1.60}$
Covid 1	$0.00^{\pm 0.00}$	$0.00^{\pm 0.00}$	$0.33^{\pm 0.17}$	$0.00^{\pm 0.00}$	$0.50^{\pm 0.27}$	$2.33^{\pm 1.49}$	$3.00^{\pm 0.00}$	$1.00^{\pm 0.00}$	$1.00^{\pm 0.00}$
Covid 2	$1.00^{\pm 1.00}$	$1.00^{\pm 1.00}$	$1.67^{\pm 0.97}$	$1.00^{\pm 1.00}$	$1.00^{\pm 1.00}$	$3.67^{\pm 1.76}$	$6.33^{\pm 0.94}$	$3.00^{\pm 0.00}$	$3.00^{\pm 0.00}$
Covid 3	$1.00^{\pm 1.00}$	$3.25^{\pm 2.82}$	$1.17^{\pm 0.83}$	$1.00^{\pm 1.00}$	$2.33^{\pm 1.51}$	$5.33^{\pm 2.28}$	$1.67^{\pm 1.49}$	$3.00^{\pm 1.00}$	$3.00^{\pm 1.00}$
Covid 4	$2.00^{\pm 1.35}$	$2.14^{\pm 1.57}$	$3.17^{\pm 1.10}$	$2.00^{\pm 1.41}$	$2.00^{\pm 1.41}$	$6.50^{\pm 2.24}$	$5.50^{\pm 1.12}$	$4.00^{\pm 0.00}$	$4.00^{\pm 0.00}$
Genetic	$0.12^{\pm 0.33}$	$0.00^{\pm 0.00}$	$0.50^{\pm 0.00}$	$0.00^{\pm 0.00}$	$2.00^{\pm 0.50}$	$8.50^{\pm 1.87}$	$10.00^{\pm 0.00}$	$12.20^{\pm 1.40}$	$12.40^{\pm 0.80}$
MSU	$2.44^{\pm 0.96}$	$3.33^{\pm 2.54}$	$2.17^{\pm 0.33}$	$2.00^{\pm 0.00}$	$2.00^{\pm 0.00}$	$6.17^{\pm 2.59}$	$5.50^{\pm 0.76}$	$9.76^{\pm 2.99}$	$8.00^{\pm 1.15}$
Neighbor	$5.59^{\pm 3.49}$	$1.00^{\pm 0.00}$	$3.67^{\pm 0.29}$	$1.00^{\pm 0.00}$	$4.83^{\pm 0.17}$	$12.17^{\pm 2.53}$	$13.00^{\pm 0.00}$	$13.57^{\pm 3.38}$	$10.71^{\pm 1.28}$
Sachs	$16.85^{\pm 3.73}$	$15.26^{\pm 3.77}$	$13.67^{\pm 2.97}$	$14.33^{\pm 3.92}$	$16.00^{\pm 3.17}$	$24.73^{\pm 4.26}$	$21.83^{\pm 2.11}$	$50.59^{\pm 2.48}$	$49.00^{\pm 1.00}$
Supermarket	$8.04^{\pm 2.56}$	$13.25^{\pm 6.44}$	$8.67^{\pm 2.04}$	$11.83^{\pm 2.29}$	$9.83^{\pm 1.65}$	$17.43^{\pm 3.29}$	$16.50^{\pm 1.12}$	$18.84^{\pm 3.70}$	$11.59^{\pm 2.14}$

Table 4: SHD (↓) of the estimated causal orders with linear synthetic data. Best results are highlighted in gray.

Graph	PC	GES	NOTEARS	BOSS	GRaSP	GOBNILP	Triplet	MATS (Our)	
								MPDAG	Causal Orders
Asia	$8.80^{\pm 1.78}$	$8.75^{\pm 1.48}$	15.83 ± 0.33	$10.00^{\pm 0.53}$	$9.50^{\pm 0.17}$	$8.67^{\pm 0.77}$	$19.75^{\pm 3.90}$	$24.00^{\pm 2.25}$	$22.00^{\pm 1.63}$
Cancer	$2.67^{\pm 0.47}$	$4.00^{\pm 0.00}$	6.83 ± 1.16	$4.00^{\pm 0.17}$	$4.00^{\pm 0.17}$	$5.00^{\pm 1.00}$	$3.50^{\pm 0.76}$	$9.73^{\pm 1.61}$	$6.00^{\pm 0.00}$
Climate	$7.67^{\pm 0.47}$	$8.00^{\pm 0.00}$	13.33 ± 0.40	$8.50^{\pm 0.17}$	$8.33^{\pm 0.00}$	$8.00^{\pm 0.00}$	$24.00^{\pm 0.00}$	$21.20^{\pm 1.60}$	$20.80^{\pm 1.60}$
Covid 1	$0.00^{\pm 0.00}$	$2.00^{\pm 0.00}$	2.67 ± 0.00	$2.00^{\pm 0.00}$	$2.00^{\pm 0.00}$	$2.17^{\pm 0.64}$	$3.00^{\pm 0.00}$	$1.00^{\pm 0.00}$	$1.00^{\pm 0.00}$
Covid 2	$3.00^{\pm 1.00}$	$3.00^{\pm 1.00}$	4.50 ± 0.40	$3.17^{\pm 0.00}$	$3.17^{\pm 0.00}$	$3.25^{\pm 1.18}$	$6.33^{\pm 0.94}$	$3.00^{\pm 0.00}$	$3.00^{\pm 0.00}$
Covid 3	$3.00^{\pm 0.00}$	$5.00^{\pm 0.00}$	5.17 ± 0.57	$5.00^{\pm 0.00}$	$5.00^{\pm 0.00}$	$5.20^{\pm 0.77}$	$1.67^{\pm 1.49}$	$3.00^{\pm 1.00}$	$3.00^{\pm 1.00}$
Covid 4	$6.00^{\pm 1.29}$	$6.00^{\pm 0.00}$	7.50 ± 0.44	$5.83^{\pm 0.27}$	$5.83^{\pm 0.27}$	$6.00^{\pm 0.67}$	$5.50^{\pm 1.12}$	$4.00^{\pm 0.00}$	$4.00^{\pm 0.00}$
Genetic	$1.00^{\pm 0.00}$	$5.00^{\pm 0.63}$	9.50 ± 0.94	$5.00^{\pm 0.17}$	$5.00^{\pm 0.17}$	$5.80^{\pm 1.46}$	$10.00^{\pm 0.00}$	$12.20^{\pm 1.40}$	$12.40^{\pm 0.80}$
MSU	$2.50^{\pm 0.50}$	$6.00^{\pm 0.63}$	7.67 ± 1.23	$6.00^{\pm 0.17}$	$6.00^{\pm 0.17}$	$6.20^{\pm 1.66}$	$5.50^{\pm 0.76}$	$9.76^{\pm 2.99}$	$8.00^{\pm 1.15}$
Neighbor	$6.00^{\pm 0.82}$	$8.00^{\pm 0.63}$	9.83 ± 0.89	$8.17^{\pm 0.17}$	$8.17^{\pm 0.17}$	$9.00^{\pm 1.15}$	$13.00^{\pm 0.00}$	$13.57^{\pm 3.38}$	$10.71^{\pm 1.28}$
Sachs	$18.75^{\pm 1.56}$	$17.00^{\pm 0.63}$	21.83 ± 1.21	$18.83^{\pm 1.16}$	$18.33^{\pm 1.35}$	$17.00^{\pm 2.16}$	$21.83^{\pm 2.11}$	$50.59^{\pm 2.48}$	$49.00^{\pm 1.00}$
Supermarket	$11.00^{\pm 0.00}$	$13.50^{\pm 1.50}$	15.20 ± 0.95	$12.00^{\pm 0.17}$	$11.50^{\pm 0.33}$	$12.00^{\pm 1.00}$	$16.50^{\pm 1.12}$	$18.84^{\pm 3.70}$	$11.59^{\pm 2.14}$

Table 5: SHD (↓) of the estimated causal orders with non-linear synthetic data. Best results are highlighted in gray.

Graph	PC	GES	NOTEARS	BOSS	GRaSP	GOBNILP	Triplet + PC	MATS + PC (Ours)
Asia	$3.66^{\pm 2.47}$	$1.75^{\pm 1.20}$	$4.08^{\pm 0.92}$	$1.50^{\pm 0.87}$	$1.50^{\pm 0.87}$	$4.31^{\pm 1.20}$	$2.00^{\pm 1.73}$	$1.00^{\pm 0.82}$
Cancer	$0.00^{\pm 0.00}$	$0.00^{\pm 0.00}$	$0.83^{\pm 0.12}$	$0.00^{\pm 0.00}$	$0.00^{\pm 0.00}$	$2.00^{\pm 0.86}$	$0.17^{\pm 0.37}$	$0.00^{\pm 0.00}$
Climate	$0.14^{\pm 0.35}$	$1.67^{\pm 3.13}$	$1.83^{\pm 0.17}$	$0.00^{\pm 0.00}$	$0.83^{\pm 0.50}$	$4.42^{\pm 1.21}$	$2.00^{\pm 0.00}$	$0.50^{\pm 0.87}$
Covid 1	$0.00^{\pm 0.00}$	$0.00^{\pm 0.00}$	$0.25^{\pm 0.08}$	$0.00^{\pm 0.00}$	$0.17^{\pm 0.14}$	$1.00^{\pm 0.74}$	$1.00^{\pm 0.00}$	$0.00^{\pm 0.00}$
Covid 2	$0.50^{\pm 0.50}$	$0.50^{\pm 0.50}$	$1.08^{\pm 0.49}$	$0.50^{\pm 0.50}$	$0.50^{\pm 0.50}$	$1.67^{\pm 0.88}$	$1.67^{\pm 0.47}$	$0.00^{\pm 0.00}$
Covid 3	$0.50^{\pm 0.50}$	$1.62^{\pm 1.41}$	$0.75^{\pm 0.42}$	$1.25^{\pm 0.42}$	$1.17^{\pm 0.75}$	$2.64^{\pm 1.25}$	$0.00^{\pm 0.00}$	$0.88^{\pm 0.60}$
Covid 4	$1.00^{\pm 0.73}$	$1.03^{\pm 0.72}$	$2.25^{\pm 0.67}$	$1.00^{\pm 0.71}$	$1.00^{\pm 0.71}$	$3.49^{\pm 1.49}$	$0.83^{\pm 0.37}$	$0.17^{\pm 0.37}$
Genetic	$0.00^{\pm 0.00}$	$0.00^{\pm 0.00}$	$0.50^{\pm 0.00}$	$0.00^{\pm 0.00}$	$0.25^{\pm 0.25}$	$2.75^{\pm 0.93}$	$0.00^{\pm 0.00}$	$1.25^{\pm 0.43}$
MSU	$1.29^{\pm 0.70}$	$1.56^{\pm 1.07}$	$1.33^{\pm 0.17}$	$1.00^{\pm 0.00}$	$1.00^{\pm 0.00}$	$3.00^{\pm 1.32}$	$1.33^{\pm 0.75}$	$2.00^{\pm 0.58}$
Neighbor	$5.65^{\pm 3.12}$	$0.00^{\pm 0.00}$	$2.33^{\pm 0.12}$	$0.88^{\pm 0.14}$	$1.21^{\pm 0.14}$	$4.30^{\pm 1.59}$	$3.83^{\pm 0.37}$	$3.00^{\pm 1.41}$
Sachs	$9.83^{\pm 3.17}$	$10.02^{\pm 3.12}$	$8.77^{\pm 1.73}$	$7.38^{\pm 1.97}$	$8.42^{\pm 1.61}$	$9.31^{\pm 2.57}$	$5.83^{\pm 2.73}$	$5.00^{\pm 1.22}$
Supermarket	$6.38^{\pm 2.63}$	$9.11^{\pm 3.64}$	$4.31^{\pm 1.30}$	$5.12^{\pm 1.19}$	$4.21^{\pm 0.88}$	$6.49^{\pm 1.83}$	$6.67^{\pm 0.47}$	$1.50^{\pm 1.32}$

Table 6: $\mathcal{D}_{top}$ (↓) of the estimated causal graphs (obtained hybridizing MATS and Triplets with PC) with linear synthetic data. Best results are highlighted in gray.

Graph	PC	GES	NOTEARS	BOSS	GRaSP	GOBNILP	Triplet + PC	MATS + PC (Our)
Asia	$4.39^{\pm 2.34}$	$3.00^{\pm 1.73}$	$5.83^{\pm 1.84}$	$3.00^{\pm 1.73}$	$3.00^{\pm 1.73}$	$14.50^{\pm 2.41}$	$2.83^{\pm 0.69}$	$2.22^{\pm 1.81}$
Cancer	$0.00^{\pm 0.00}$	$0.00^{\pm 0.00}$	$1.00^{\pm 0.24}$	$0.00^{\pm 0.00}$	$0.00^{\pm 0.00}$	$7.50^{\pm 1.72}$	$4.20^{\pm 0.40}$	$0.33^{\pm 0.47}$
Climate	$0.44^{\pm 0.68}$	$1.56^{\pm 2.95}$	$2.00^{\pm 0.33}$	$0.00^{\pm 0.00}$	$2.50^{\pm 1.01}$	$12.83^{\pm 2.33}$	$2.00^{\pm 0.00}$	$1.25^{\pm 1.64}$
Covid 1	$0.00^{\pm 0.00}$	$0.00^{\pm 0.00}$	$0.33^{\pm 0.17}$	$0.00^{\pm 0.00}$	$0.50^{\pm 0.27}$	$2.33^{\pm 1.49}$	$3.50^{\pm 1.12}$	$0.00^{\pm 0.00}$
Covid 2	$1.00^{\pm 1.00}$	$1.00^{\pm 1.00}$	$1.67^{\pm 0.97}$	$1.00^{\pm 1.00}$	$1.00^{\pm 1.00}$	$3.67^{\pm 1.76}$	$1.00^{\pm 1.83}$	$0.17^{\pm 0.37}$
Covid 3	$1.00^{\pm 1.00}$	$3.25^{\pm 2.82}$	$1.17^{\pm 0.83}$	$1.00^{\pm 1.00}$	$2.33^{\pm 1.51}$	$5.33^{\pm 2.28}$	$4.33^{\pm 2.36}$	$1.88^{\pm 1.05}$
Covid 4	$2.00^{\pm 1.35}$	$2.14^{\pm 1.57}$	$3.17^{\pm 1.10}$	$2.00^{\pm 1.41}$	$2.00^{\pm 1.41}$	$6.50^{\pm 2.24}$	$0.83^{\pm 0.37}$	$0.67^{\pm 0.75}$
Genetic	$0.12^{\pm 0.33}$	$0.00^{\pm 0.00}$	$0.50^{\pm 0.00}$	$0.00^{\pm 0.00}$	$2.00^{\pm 0.50}$	$8.50^{\pm 1.87}$	$0.17^{\pm 0.37}$	$2.75^{\pm 0.83}$
MSU	$2.44^{\pm 0.96}$	$3.33^{\pm 2.54}$	$2.17^{\pm 0.33}$	$2.00^{\pm 0.00}$	$2.00^{\pm 0.00}$	$6.17^{\pm 2.59}$	$4.33^{\pm 0.75}$	$4.33^{\pm 1.25}$
Neighbor	$5.59^{\pm 3.49}$	$1.00^{\pm 0.00}$	$3.67^{\pm 0.29}$	$1.00^{\pm 0.00}$	$4.83^{\pm 0.17}$	$12.17^{\pm 2.53}$	$6.33^{\pm 0.94}$	$4.00^{\pm 1.20}$
Sachs	$16.85^{\pm 3.73}$	$15.26^{\pm 3.77}$	$13.67^{\pm 2.97}$	$14.33^{\pm 3.92}$	$16.00^{\pm 3.17}$	$24.73^{\pm 4.26}$	$13.67^{\pm 1.89}$	$11.75^{\pm 1.92}$
Supermarket	$8.04^{\pm 2.56}$	$13.25^{\pm 6.44}$	$8.67^{\pm 2.04}$	$11.83^{\pm 2.29}$	$9.83^{\pm 1.65}$	$17.43^{\pm 3.29}$	$8.67^{\pm 1.70}$	$3.47^{\pm 2.12}$

Table 7: SHD (↓) of the estimated causal graphs (obtained hybridizing MATS and Triplets with PC) with linear synthetic data. Best results are highlighted in gray.

Graph	NOTEARS				MATS (Our)	
	Linear		Non-Linear		MPDAG	π
	DAG	CPDAG	DAG	CPDAG		
Asia	$0.75^{\pm 0.97}$	$4.08^{\pm 0.92}$	$7.40^{\pm 1.50}$	$6.06^{\pm 0.36}$	$2.87^{\pm 1.81}$	$1.22^{\pm 1.03}$
Cancer	$1.20^{\pm 0.98}$	$0.83^{\pm 0.12}$	$4.20^{\pm 0.75}$	$2.62^{\pm 0.49}$	$3.53^{\pm 1.86}$	$0.00^{\pm 0.00}$
Climate	$3.38^{\pm 2.69}$	$1.83^{\pm 0.17}$	$8.33^{\pm 2.49}$	$5.98^{\pm 0.43}$	$0.60^{\pm 0.80}$	$0.40^{\pm 0.80}$
Covid 1	$0.00^{\pm 0.00}$	$0.25^{\pm 0.08}$	$1.33^{\pm 0.47}$	$2.00^{\pm 0.00}$	$0.55^{\pm 0.78}$	$0.00^{\pm 0.00}$
Covid 2	$0.20^{\pm 0.40}$	$1.08^{\pm 0.49}$	$1.00^{\pm 0.82}$	$2.75^{\pm 0.20}$	$0.00^{\pm 0.00}$	$0.00^{\pm 0.00}$
Covid 3	$0.12^{\pm 0.33}$	$0.75^{\pm 0.42}$	$2.00^{\pm 0.00}$	$4.46^{\pm 0.34}$	$1.00^{\pm 0.50}$	$1.00^{\pm 0.50}$
Covid 4	$0.50^{\pm 0.71}$	$2.25^{\pm 0.67}$	$3.33^{\pm 1.25}$	$3.82^{\pm 0.31}$	$0.00^{\pm 0.00}$	$0.00^{\pm 0.00}$
Genetic	$1.00^{\pm 1.73}$	$0.50^{\pm 0.00}$	$3.00^{\pm 1.41}$	$2.90^{\pm 0.73}$	$1.10^{\pm 0.70}$	$1.20^{\pm 0.40}$
MSU	$0.00^{\pm 0.00}$	$1.33^{\pm 0.17}$	$2.67^{\pm 1.89}$	$4.74^{\pm 0.71}$	$3.22^{\pm 1.73}$	$2.00^{\pm 0.58}$
Neighbor	$1.62^{\pm 1.58}$	$2.33^{\pm 0.12}$	$5.33^{\pm 2.05}$	$5.65^{\pm 0.54}$	$5.96^{\pm 2.64}$	$4.89^{\pm 2.42}$
Sachs	$3.62^{\pm 2.78}$	$8.77^{\pm 1.73}$	$7.67^{\pm 0.94}$	$13.50^{\pm 0.99}$	$7.40^{\pm 2.04}$	$5.50^{\pm 0.50}$
Supermarket	$2.00^{\pm 2.40}$	$4.31^{\pm 1.30}$	$9.00^{\pm 3.74}$	$8.08^{\pm 0.58}$	$6.45^{\pm 2.58}$	$1.56^{\pm 1.26}$

Table 8: $\mathcal{D}_{top}$ (↓) comparing MATS to NOTEARS using linear and non-linear data. We display both the $\mathcal{D}_{top}$ associated with the DAG found by NOTEARS and the relative CPDAG. Best results are highlighted in gray.

Graph	NOTEARS				MATS (Our)	
	Linear		Non-Linear		MPDAG	π
	DAG	CPDAG	DAG	CPDAG		
Asia	$2.50^{\pm 1.58}$	$5.83^{\pm 1.84}$	$13.40^{\pm 1.96}$	$15.83^{\pm 0.33}$	$24.00^{\pm 2.25}$	$22.00^{\pm 1.63}$
Cancer	$0.50^{\pm 0.71}$	$1.00^{\pm 0.24}$	$6.40^{\pm 1.85}$	$6.83^{\pm 1.16}$	$9.73^{\pm 1.61}$	$6.00^{\pm 0.00}$
Climate	$1.75^{\pm 0.83}$	$2.00^{\pm 0.33}$	$14.33^{\pm 2.62}$	$13.33^{\pm 0.40}$	$21.20^{\pm 1.60}$	$20.80^{\pm 1.60}$
Covid 1	$0.12^{\pm 0.33}$	$0.33^{\pm 0.17}$	$3.00^{\pm 0.82}$	$2.67^{\pm 0.00}$	$1.00^{\pm 0.00}$	$1.00^{\pm 0.00}$
Covid 2	$0.88^{\pm 0.78}$	$1.67^{\pm 0.97}$	$2.67^{\pm 1.25}$	$4.50^{\pm 0.40}$	$3.00^{\pm 0.00}$	$3.00^{\pm 0.00}$
Covid 3	$0.50^{\pm 0.71}$	$1.17^{\pm 0.83}$	$4.67^{\pm 0.47}$	$5.17^{\pm 0.57}$	$3.00^{\pm 1.00}$	$3.00^{\pm 1.00}$
Covid 4	$1.88^{\pm 1.0}$	$3.17^{\pm 1.10}$	$7.67^{\pm 2.49}$	$7.50^{\pm 0.44}$	$4.00^{\pm 0.00}$	$4.00^{\pm 0.00}$
Genetic	$0.62^{\pm 0.86}$	$0.50^{\pm 0.00}$	$7.67^{\pm 2.49}$	$9.50^{\pm 0.94}$	$12.20^{\pm 1.40}$	$12.40^{\pm 0.80}$
MSU	$2.00^{\pm 0.71}$	$2.17^{\pm 0.33}$	$7.00^{\pm 4.24}$	$7.67^{\pm 1.23}$	$9.76^{\pm 2.99}$	$8.00^{\pm 1.15}$
Neighbor	$2.62^{\pm 1.32}$	$3.67^{\pm 0.29}$	$10.33^{\pm 2.05}$	$9.83^{\pm 0.89}$	$13.57^{\pm 3.38}$	$10.71^{\pm 1.28}$
Sachs	$7.12^{\pm 1.76}$	$13.67^{\pm 2.97}$	$24.33^{\pm 4.03}$	$21.83^{\pm 1.21}$	$50.59^{\pm 2.48}$	$49.00^{\pm 1.00}$
Supermarket	$4.62^{\pm 2.39}$	$8.67^{\pm 2.04}$	$15.67^{\pm 1.25}$	$15.20^{\pm 0.95}$	$18.84^{\pm 3.70}$	$11.59^{\pm 2.14}$

Table 9: SHD (↓) comparing MATS to NOTEARS using linear and non-linear data. We display both the $\mathcal{D}_{top}$ associated with the DAG found by NOTEARS and the relative CPDAG. Best results are highlighted in gray.

Graph	gpt-4.1-nano	mistral:7b	llama3.1
Asia	0.00 ± 0.00	0.00 ± 0.00	2.00 ± 0.00
Cancer	0.00 ± 0.00	0.00 ± 0.00	2.00 ± 0.00
Climate	0.00 ± 0.00	0.00 ± 0.00	3.00 ± 0.00
Covid 1	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
Covid 2	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
Covid 3	0.50 ± 0.50	0.50 ± 0.50	0.00 ± 0.00
Covid 4	0.00 ± 0.00	0.00 ± 0.00	1.00 ± 0.00
Genetic	1.00 ± 0.00	1.00 ± 0.00	0.00 ± 0.00
MSU	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00
Neighbor	1.00 ± 0.00	-	-
Sachs	0.50 ± 0.50	0.50 ± 0.50	13.00 ± 0.00
Supermarket	5.00 ± 0.00	5.00 ± 0.00	4.50 ± 0.00

Table 10: Best intraclass $\mathcal{D}_{top}$ ($\downarrow$) of MATS for causal orders using different LLMs. In gray best performing method. '-' indicates that the method did not return a valid causal order.

Graph	Nodes	Edges
Asia 6	8	8
Cancer 7	5	4
Climate 8	8	8
Covid 1 9	3	2
Covid 2 10	4	5
Covid 3 11	4	3
Covid 4 12	4	6
Genetic 13	6	5
MSU 14	5	6
Neighborhood 15	5	8
Sachs 16	11	17
Supermarket 17	7	12

Table 11: Causal DAGs used in the experiments.

Appendix D. Prompts & Graphs

Provide {*n_rephrase*} rephrased versions of the following sentence: {*sentence*}.
The rephrased sentences should preserve the semantic meaning, even if absurd, of the original one. The answers should be in the following format:
1. rephrased sentence 1
2. rephrased sentence 2
3. rephrased sentence 3

Figure 4: Prompt used to query the LLM for rephrasing.

You are an expert in the field of {*field*}.
The task is to provide causal relationships between variables. Keep your answers concise and to the point.
Does {var_i} cause {var_j}?
(A) Yes
(B) No
Answer:

Figure 5: Prompt used to query the LLM for causal relationships.

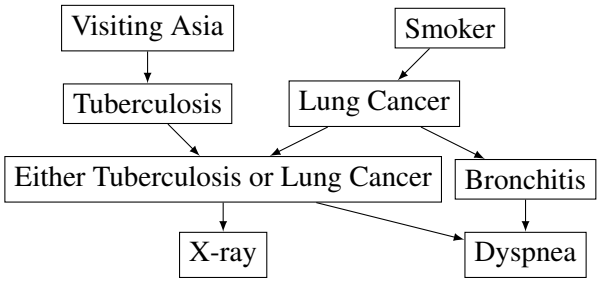


Figure 6: Asia causal DAG from `bnlearn`.

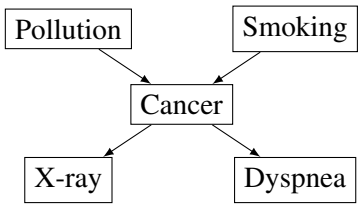


Figure 7: Cancer causal DAG from `bnlearn`.

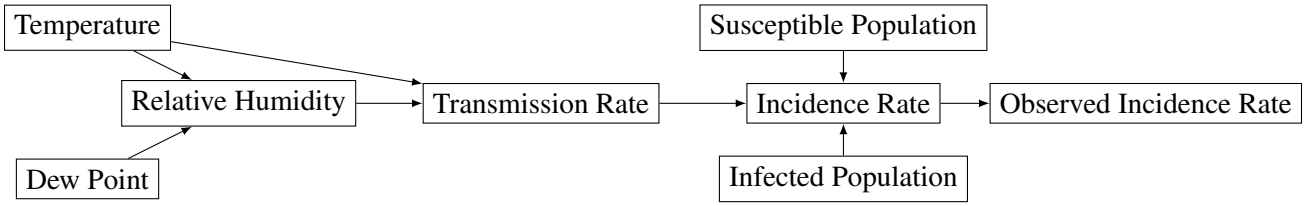


Figure 8: Climate causal DAG.

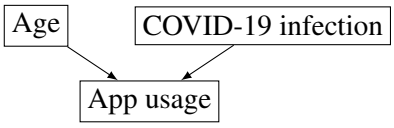


Figure 9: Covid 1 causal DAG.

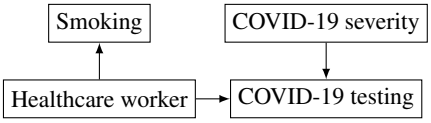


Figure 10: Covid 2 causal DAG.

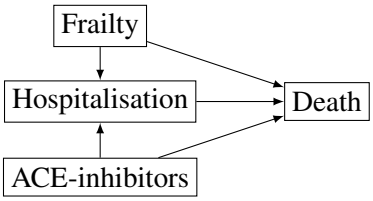


Figure 11: Covid 3 causal DAG.

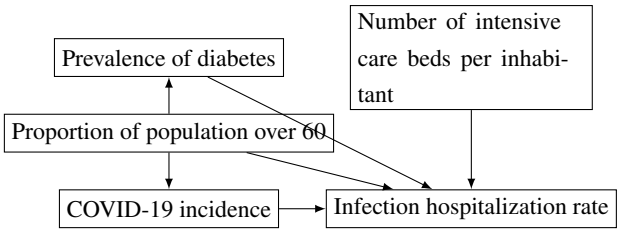


Figure 12: Covid 4 causal DAG.

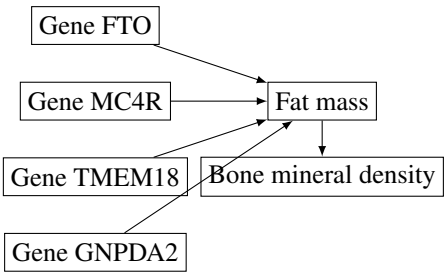


Figure 13: Genetic causal DAG.

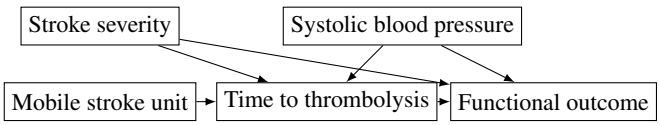


Figure 14: MSU causal DAG.

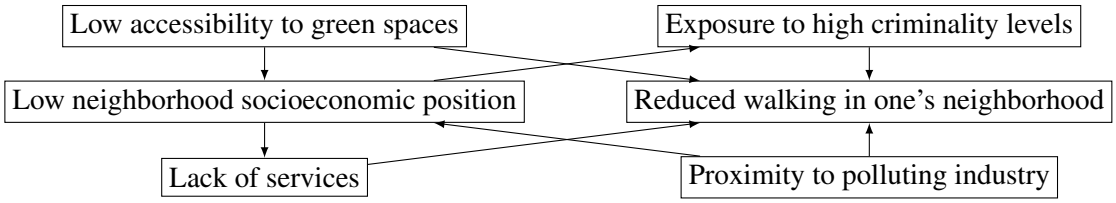


Figure 15: Neighborhood causal DAG.

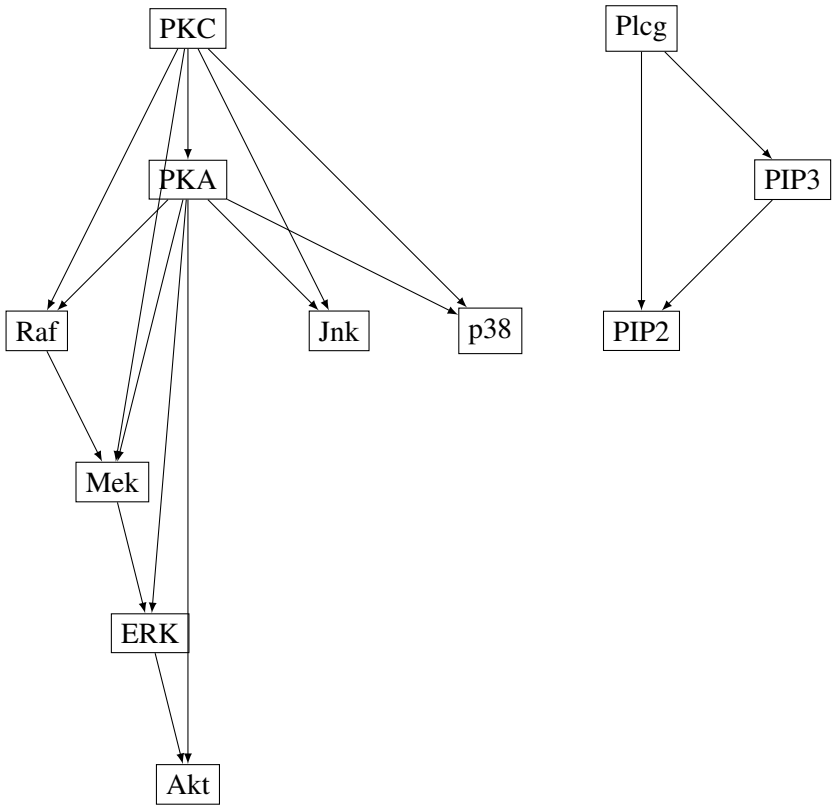


Figure 16: Sachs causal DAG from `bnlearn`.

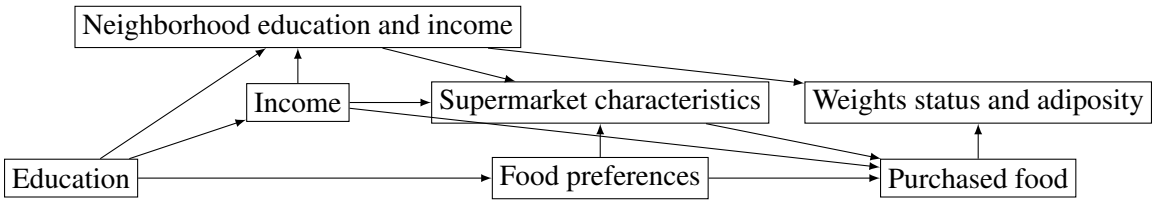


Figure 17: Supermarket causal DAG.