Identifiability Is Not Enough: A Causal Discovery Case Study with r-Partite Graphs

Anonymous authors
Paper under double-blind review

Abstract

Observational causal discovery methods have grown rapidly in popularity over recent years, with methodological success evidenced by controlled simulation experiments. However, there have been limited successes in causal discovery under real world data conditions. This suggests a gap between theory and practice. In this work, we investigate this mismatch by studying the behavior of causal discovery methods in cases where causal graphs are provably identifiable and recoverable by all studied algorithms. To contextualize performance, we introduce a simple baseline benchmark that isolates performance gained from respecting the problem's structural topology and basic statistical tests. This baseline establishes an empirical performance standard. Our experiments show that across various r-partite settings, existing methods do not consistently surpass this baseline. We conclude that these inconsistent performance gains demonstrate that existing causal theory is insufficient to guarantee strong empirical performance. Consequently, our baseline benchmark provides a necessary reality check for evaluating data-driven causal discovery.

1 Introduction

The study of causality, the pursuit of understanding the fundamental relationships between events or variables, traces its origin to classical antiquity (Charlton, 1970). In recent decades, these causal relationships have been mathematically formalized through Directed Acyclic Graphs (DAGs), where nodes denote variables and directed edges indicate the presence of causal influence (Pearl, 2009). Modern observational causal discovery methods aim to detect the presence of true relationships from data (Spirtes et al., 2000). Various theories for detecting these relationships have produced a range of distinct approaches.

These approaches can be broadly categorized into four main streams. **Constraint-based** approaches, such as PC (Spirtes & Glymour, 1991), rely on conditional independence tests to infer edges and their orientations. **Score-based** methods, exemplified by GES (Chickering, 2002), instead search for the graph that maximizes a scoring criterion like BIC (Bayesian Information Criterion), yielding the Markov equivalence class of a DAG. **Functional form-based** approaches, including DirectLiNGAM (Shimizu et al., 2011), exploit explicit functional assumptions (e.g., linearity) to distinguish causes from effects. **Gradient-based** methods, beginning with NOTEARS (Zheng et al., 2018), recast structure learning as a continuous optimization problem with an acyclicity constraint.

Despite this algorithmic diversity, general DAG structures remain unidentifiable from purely observational data (Spirtes et al., 2000). This prevents causal discovery methods from producing predictions converging to a unique DAG; they must instead predict a Markov Equivalence Class (MEC),[1] in which the orientation of certain edges remains underdetermined.

While some methods such as PC only offer identifiability up to MEC; others make use of additional, restrictive and often untestable assumptions to obtain stronger identifiability guarantees. Examples include LiNGAM, which requires the data to be linear with independent non-Gaussian errors (Shimizu et al., 2006); and additive

[1] Two DAGs are Markov equivalent if and only if they have (i) the same skeleton (i.e., the same adjacencies between nodes), and (ii) the same set of v-structures (induced subgraphs $X \to Z \leftarrow Y$ where X and Y are non-adjacent). (Pearl & Verma, 1995)

noise models (ANM), which model each variable as a deterministic function of parents plus independent noise (Hoyer et al., 2008). Such assumptions are justified only by their mathematical convenience. As such, it remains a question whether these identifiability claims retain their validity in real-world scenarios.

On the other hand, many practical domains naturally exhibit known structural properties, such as temporal settings (Peters et al., 2017, pp.197-198). Beyond improving model identifiability, enforcing these types of structural partitions has recently emerged as a powerful tool for enhancing algorithmic scalability. Specifically, Leiner et al. (2023) introduce a strategy that temporally separates nodes into sets to alleviate the computational burden of structure learning. Although this approach relies on black-box models that may limit interpretability, it nevertheless demonstrates the scalability benefits of node splitting. This is particularly beneficial as discovering the correct structure of a DAG from data remains a computationally challenging problem and the number of candidate DAGs grows superexponentially with the number of nodes in the graph (Chickering, 2002; Robinson, 2006). More general frameworks that implement structural constraints include tiered background knowledge (Andrews et al., 2020; Bang & Didelez, 2023), cluster-DAGs (Ruiz de Vargas et al., 2025), typing assumptions (Brouillard et al., 2022), and order-based causal discovery for multistage processes (Ma et al., 2026). Andrews et al. (2020) and Bang & Didelez (2023) utilize tiered background knowledge, which organizes variables into ordered groups where causal effects can only flow from earlier tiers to later ones, thereby significantly reducing the space of Markov equivalent graphs. Building on the idea of group-level constraints, Ruiz de Vargas et al. (2025) introduce cluster-DAGs, a framework that leverages known relationships between sets of variables to provide powerful priors for structure learning. Additionally, Brouillard et al. (2022), offers a categorical approach to identification: if a DAG can be partitioned into k-colors (types), model identifiability substantially improves in high-dimensional or complex domains. Finally, Ma et al. (2026), incorporates process knowledge in the causal discovery framework to preserve the sequential and hierarchical structure of multistage data.

Despite these theoretical insights into identifiability and the integration of background knowledge, a significant disconnect persists between algorithmic sophistication and empirical reliability. While simulations under idealized conditions suggest that causal discovery is a maturing field, these successes rarely translate to real-world datasets where modeling assumptions are frequently violated (Brouillard et al., 2024). Compounding this concern, Reisach et al. (2021) demonstrate that the remarkable benchmark performance of several continuous structure learning methods, including NOTEARS, can be largely attributed to an artifact of simulation: marginal variance tends to increase along the causal order in generically sampled additive noise models. This discrepancy raises a fundamental question: is identifiability enough to guarantee strong method performance?

In this paper, we interrogate this discrepancy by evaluating state-of-the-art causal discovery methods on provably identifiable, straightforward structures. In particular, we explore block-directed bipartite and more general r-partite DAGs. We test their ability to recover fundamental topologies that should, in theory, be easily resolved, rather than overwhelming algorithms with complex DAG structures. To establish a rigorous point of comparison, we introduce a simple topological baseline that exploits basic structural constraints. Our contributions are as follows:

1. We introduce a class of causal discovery benchmark problems with theoretical identifiability guarantees that are recoverable using only standard assumptions.
2. An evaluation tool derived from a set of baseline algorithms that serves as an empirical performance benchmark for existing methods.
3. Experimental evidence for a range of popular causal discovery methods that demonstrate sub-optimal empirical performance.

We structure the article as follows. Section 2 introduces the notation, formal assumptions and theoretical foundation of our work. In particular, we focus on graph identifiability, the core property that ensures a DAG is recoverable from the observed distribution. In Section 3, we present a set of synthetic experiments on tabular data, designed to benchmark existing causal discovery methods on a diverse range of datasets that correspond to r-partite DAGs. These experiments test the methods under idealized conditions, such as

having the Data Generating Process (DGP) satisfy a particular method's assumptions (for example, linearity in the case of methods such as NOTEARS (Zheng et al., 2018)); and, under more general conditions, where although all standard causal discovery assumptions as stated in Section 2.2 hold, any extra assumptions may be violated. Importantly, we emphasize that all r-partite DAGs used to generate these tabular datasets are all identifiable, as detailed in Section 2.3. We discover that method performance is in general insufficient and, in many cases, unable to consistently outperform the baseline algorithms. We then proceed to benchmark causal discovery methods on realistically generated temporal data in Section 4. To account for this time-series setting, we employ temporal causal discovery methods. The resulting empirical findings closely mirror those of the synthetic tabular experiments, confirming that sophisticated temporal models similarly struggle to outperform our structural baselines. Finally, Section 5 describes limitations and Section 6 summarizes our thoughts and conclusions. We conclude that existing methods' inconsistent and performance gains over the baselines indicate that existing causal theory is insufficient to guarantee strong empirical performance, both with synthetic data and realistically simulated datasets. Our code is publicly available online.[2]

2 Theory

2.1 Definitions

We formalise the data-generating process (DGP) via a Structural Causal Model (SCM) to establish the necessary identifiability guarantees. Let $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ be a DAG where the vertex set, $\mathbf{V}$, is partitioned into disjoint sets and the edge set is denoted as $\mathbf{E}$. We define the SCM, $\mathcal{M}$, by the set of structural equations $\mathcal{S}$ and the joint probability distribution over the exogeneous noise variables $\mathbf{U}$, $P(\mathbf{U})$. That is, $\mathcal{M} = \langle \mathcal{S}, P(\mathbf{U}) \rangle$ such that for every endogenous variable $X_i \in \mathcal{V}$, the structural equation is defined as:

$$X_i := f_i(Pa(X_i), U_i), \tag{1}$$

where $Pa(X_i)$ denotes the set of parent nodes of X_i, and U_i represents the mutually independent exogenous noise variables.

Definition 1 (Bipartite graph)**.** *A bipartite graph, $\mathcal{G} = (\mathbf{V}, \mathbf{E})$, consists of nodes, $\mathbf{V}$, and edges, $\mathbf{E}$, such that $\mathbf{V}$ may be partitioned into two disjoint subsets $\mathcal{S}_1$ and $\mathcal{S}_2$, where every edge in $\mathbf{E}$ has one endpoint in $\mathcal{S}_1$ and another in $\mathcal{S}_2$ (Fouss et al., 2016). For ease of notation, when discussing the bipartite graphs we will refer to $\mathcal{S}_1$ and $\mathcal{S}_2$ as $\mathcal{A}$ and $\mathcal{B}$, respectively.*

Definition 2 (r-Partite Graph)**.** *An r-partite (multipartite) graph, $\mathcal{G} = (\mathbf{V}, \mathbf{E})$, consists of nodes $\mathbf{V}$ and edges $\mathbf{E}$ such that $\mathbf{V}$ may be partitioned into r disjoint subsets $\mathbf{V} = (\mathcal{S}_1, \ldots, \mathcal{S}_r)$, where every edge in $\mathbf{E}$ has one endpoint in $\mathcal{S}_i$ and another in $\mathcal{S}_j$ and $i \neq j$ (Zdravkovska, 2018).*

Definition 3 (Block-Directed Bipartite DAG)**.** *We define a block-directed bipartite DAG as a bipartite graph, $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ with $\mathbf{V} = (\mathcal{A}, \mathcal{B})$, with directed edges such that causality flows exclusively from nodes in $\mathcal{A}$ to nodes in $\mathcal{B}$. $\mathcal{A}$ is denoted as the set of source nodes and $\mathcal{B}$ represents the set of sink nodes. An example of a block-directed bipartite DAG is shown in Figure 1a.*

Definition 4 (Block-directed r-Partite DAG)**.** *We define a block-directed r-partite DAG as an r-partite graph, $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ with $\mathbf{V} = (\mathcal{S}_1, \ldots, \mathcal{S}_r)$, where we can establish a causal ordering such that there can only be causation from a node in an earlier partition to a node in a later partition. An example of a block-directed r-parite DAG is shown in Figure 1b.*

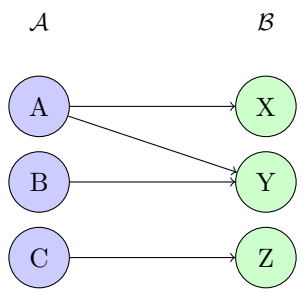
2.2 Formal Assumptions and Structural Causal Framework

To bridge these graph definitions with data distributions, we maintain the following formal assumptions (Peters et al., 2017):

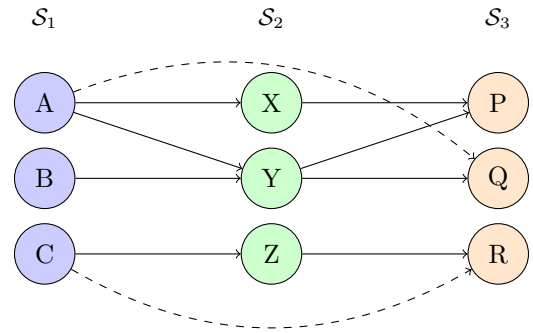
Assumption 1 (Causal Markov Property)**.** *The joint distribution $P(\mathbf{X})$ factors according to the graph $\mathcal{G}$. Formally, d-separation in the graph implies independence in the probability distribution:*

$$(X \perp_{\mathcal{G}} Y \mid Z) \implies (X \perp_P Y \mid Z), \tag{2}$$

[2]https://anonymous.4open.science/r/r-dag-paper-5C41



(a) A block-directed bipartite DAG ($r = 2$). Nodes in $\mathcal{A}$ (blue) are source nodes and nodes in $\mathcal{B}$ (green) are sink nodes. All edges flow exclusively from $\mathcal{A}$ to $\mathcal{B}$.



(b) A block-directed r-partite DAG with $r = 3$. Edges between adjacent partitions (solid) represent lag-1 effects. Dashed edges represent lag-2 effects spanning from $\mathcal{S}_1$ directly to $\mathcal{S}_3$. All edges respect the causal ordering across partitions.

Figure 1: Examples of block-directed partition DAG structures. In both cases, causality flows strictly from earlier partitions to later ones with no within-partition edges. The bipartite DAG **(a)** is the special case of the r-partite DAG **(b)** with $r = 2$.

where $\perp_{\mathcal{G}}$ denotes the d-separation criterion in the graph topology (Pearl, 2009), and $\perp_P$ denotes probabilistic conditional independence in the data.

Assumption 2 (Causal Sufficiency)**.** *There are no unmeasured common causes (latent confounders) affecting any pair of variables (X_i, X_j). Mathematically, this is represented by the mutual independence of exogenous variables: $U_i \perp U_j$ for all $i \neq j$.*

Assumption 3 (Faithfulness)**.** *All conditional independence relations in $P(\mathbf{X})$ are entailed by d-separation in $\mathcal{G}$. Formally, $X_i \perp_P X_j \mid \mathbf{Z} \implies X_i \perp_{\mathcal{G}} X_j \mid \mathbf{Z}$.*

Assumption 4 (Minimality)**.** *Any proper subgraph of $\mathcal{G}$ does not satisfy the Causal Markov Property with respect to the joint distribution $P(\mathbf{X})$.*

2.3 Identifiability

Identifiability of DAGs holds when each observational distribution corresponds to a unique graph whose structure encodes exactly its conditional independencies. Ensuring indentifiability is crucial for causal discovery, as without this property, recovering the complete and correct underlying structure would be impossible (Pearl, 2009).

Identifiability of DAGs from observational data alone is not possible in the general case; additional restrictions, such as DGP assumptions, are needed to distinguish between structurally distinct but observationally equivalent graphs (e.g., distinguishing a chain from a fork). This includes all methods that we examine; for example, the PC algorithm can identify a graph up to a Markov equivalence class (Spirtes & Glymour, 1991). However, due to the structure of bipartite and r-partite graphs, one only requires the standard causal discovery assumptions defined in Section 2.2 to recover a DAG. This structural setting ensures that algorithms restricted to standard causal assumptions can, in theory, recover the unique ground-truth DAG. The corresponding proofs of identifiability are provided in Appendix B.

Theorem 1 (Identifiability of Block-Directed Bipartite DAGs)**.** *Given causal sufficiency, faithfulness, the Markov property and knowledge of direction of causality in bipartition; a block-directed bipartite DAG is identifiable from independent and identically distributed (iid) observational data from its conditional independence relations.*

We note that the bipartite assumption is strong enough that faithfulness need not be assumed, as long as minimality still holds.

Theorem 2 (Identifiability of Block-Directed r-Partite DAGs)**.** *Given causal sufficiency, faithfulness, the Markov property, and knowledge of the causal ordering of partitions; a block-directed r-partite DAG is identifiable from independent and identically distributed (iid) observational data from its conditional independence relations.*

3 Case Study with Synthetic Tabular Experiments

3.1 Bipartite Context

Table 1: Benchmark design dimensions and settings.

Dimension	Values	Description
Node count	10, 15, 20	Small to medium-sized networks
Proportion of source to sink nodes	$20\%, 50\%, 80\%$	20% = sink-heavy; 50% = balanced; 80% = source-heavy
Global density	$20\%, 50\%, 80\%$	Overall density of edges: sparse, medium, dense
Local density	`balanced`, `source_dom`, `sink_dom`	Vertex degree distribution

To bridge theoretical identifiability guarantees with practical empirical validation, we construct a benchmark encompassing a diverse array of structural configurations. Evaluating algorithmic performance across diverse regimes requires our benchmark to generate a comprehensive range of bipartite graph topologies. This structural diversity ensures a rigorous and equitable comparison that avoids biasing results toward methods tailored to specific configurations.

The benchmark design ensures coverage of both global properties (e.g., overall density) and local structural features (e.g., local density); details are summarized in Table 1. In particular, we assess local density or vertex degree distribution (VDD) where some graphs will have a node with higher degree than others, where the degree of the vertex is the number of edges incident to it (Diestel, 2024). In such cases we have modeled the graph to have a dominant source (or sink) node that is three times more likely to have incident edges than all other source (or sink) nodes. We further include a `balanced` setting where every source or sink node is equally likely to have incident edges.

These benchmark settings are constructed towards capturing a wide range of real-world scenarios. **Node count** reflects problem scale; for instance, Tennant et al. (2021) review DAG usage for confounder detection in applied health research, finding an average of 12 nodes across DAGs of varying sizes. The **proportion of source to sink nodes** distinguishes sink-heavy cases, such as many patient features predicting a single disease, from balanced regimes like patient–lab test interactions or user–item recommendations, and source–heavy cases where one intervention (e.g., a vaccination campaign) influences many outcomes (Fisker et al., 2015). **Global density** represents the degree of interconnectedness: sparse structures arise in genomics where most gene–phenotype links are absent, medium density mirrors epidemiological data with overlapping risk factors, and dense networks resemble hospital sensors or electronic health records where many variables correlate (Boll et al., 2024). Finally, **local density** regimes capture source–dominant settings such as policies affecting many outcomes, sink–dominant settings where chronic diseases are shaped by diverse causes, and balanced cases reflecting uniform interaction patterns like gene–phenotype studies or recommender systems (Köhler et al., 2014).

Algorithm 1 Bipartite Baseline

Require: Bipartite sets $\mathcal{A}$ (sources), $\mathcal{B}$ (sinks)
Ensure: Bipartite DAG G

1: Initialize G: add edges $X_i^{(\mathcal{A})} \to X_j^{(\mathcal{B})}$ for all $X_i^{(\mathcal{A})} \in \mathcal{A}, X_j^{(\mathcal{B})} \in \mathcal{B}$.
2: **for all** edges $X_i^{(\mathcal{A})} \to X_j^{(\mathcal{B})}$ **do**
3: Test marginal independence of $X_i^{(\mathcal{A})}$ and $X_j^{(\mathcal{B})}$.
4: **if** $X_i^{(\mathcal{A})} \perp X_j^{(\mathcal{B})}$ **then**
5: Remove edge $X_i^{(\mathcal{A})} \to X_j^{(\mathcal{B})}$.
6: **end if**
7: **end for**
8: **return** G.

3.1.1 Bipartite Baseline

Motivated by the considered structural constraints, we adapt the PC algorithm to establish a simple baseline method. Our baseline takes advantage of the distinct bipartite setting. Our goals in introducing this baseline are to evaluate a simple method tailored specifically to the bipartite setting and to establish an empirical sanity check for contextualizing existing algorithms.

The procedure consists of three steps: (1) start with a fully connected block-directed bipartite DAG (every source node linked causally to every sink node); (2) perform marginal independence tests to prune edges by establishing causal independence, under assumptions mentioned in Section 2.2, and (3) output the final block-directed bipartite DAG. The pseudocode for this algorithm can be found in Algorithm 1. Note that the bipartite baseline is agnostic of any specific marginal independence tests (line 3 in Algorithm 1). This choice is left to the user.

By exploiting the bipartite structure of the graph, the baseline restricts the number of required marginal independence tests, substantially mitigating computational overhead. This efficiency gain enables the incorporation of more computationally intensive independence measures; beyond partial correlations or χ^2 tests, such as the Hilbert–Schmidt Independence Criterion (HSIC), which would otherwise be prohibitive in large, unconstrained settings. For the synthetic bipartite experiments we utilized an HSIC independence tests with significance level 0.025.[3]

Importantly, we can establish the correctness of the bipartite baseline algorithm in finding the unique ground truth, following Proposition 3 in Ankan & Textor (2025) (refer to Appendix B.2 for the formal proof).

Corollary 1. *Under causal sufficiency, faithfulness, the Markov property and knowledge of direction of causality in bipartition, the bipartite baseline (Algorithm 1) converges to the ground truth DAG in the infinite data limit.*

3.2 r-Partite Context

Paralleling the bipartite framework, the experimental design for r-partite graphs evaluates a comprehensive suite of topologies across three key dimensions, summarized in Table 2. This ensures robust evidence on algorithmic performance across diverse regimes, enabling fair comparison between methods and avoiding bias towards any specific r-partite topology.

The **number of partitions** r governs the depth of the causal hierarchy. The value $r = 2$ reduces to the bipartite setting, while larger values (e.g. $r = 5$) reflect richer multi-stage processes such as longitudinal clinical studies, where variables are measured across successive time points (Caruana et al., 2015). For example, Müller et al. (2021) use three time points for studying neurodegeneration with MRI. The **maximum lag** determines the maximal topological distance over which an inter-partition causal effect can occur. A

[3]We adopt a slightly more conservative threshold than the Fisher standard of 0.05 (Fisher, 1970), following the findings in Kalisch & Bühlman (2007).

lag of one restricts edges to adjacent partitions, reflecting Markov-like temporal assumptions common in econometrics (Hamilton, 1989), whereas larger lags permit long-range dependencies, as observed in gene regulatory networks where upstream regulators affect downstream phenotypes after several intermediate steps (Dasika et al., 2003), who include examples with a maximum lag of two, for instance. Finally, **global density** controls the overall connectivity of the graph: our sparse settings (0.1) reflect systems where most inter-partition links are absent, as in genomics where few gene–phenotype associations are active (Köhler et al., 2014); medium densities $(0.3, 0.5)$ mirror epidemiological settings with overlapping risk factors; and dense settings (0.8) resemble sensor networks or electronic health records where many variables interact (Boll et al., 2024).

Table 2: Benchmark design dimensions and settings for r-partite graphs.

Dimension	Values	Description
Number of partitions	$2, 3, 5$	Few to moderately many ordered tiers
Maximum lag	$1, 2, 3$	Maximum partition distance over which edges are permitted
Global density	$10\%, 30\%, 50\%, 80\%$	Overall edge density: sparse, low-medium, medium, dense

3.2.1 r-Partite Baseline

We also introduce a baseline for the general r-partite setting that directly exploits the known partition structure and causal ordering. Mirroring the bipartite setup, we introduce this baseline to analyze a simple, structure-aware approach and to provide a critical performance standard for contextualizing existing unconstrained methods.

The procedure consists of three steps: (1) initialize a fully connected block-directed r-partite DAG $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ with $\mathbf{V} = (\mathcal{S}_1, \ldots, \mathcal{S}_r)$, respecting the known causal ordering of partitions and the maximum lag ℓ, so that edges are added from partition $\mathcal{S}_k$ to partition $\mathcal{S}_l$ only if $k < l$ and $l - k \leq \ell$; (2) for each candidate edge $X_i^{(k)} \to X_j^{(l)}$ with $X_i^{(k)} \in \mathcal{S}_k$ and $X_j^{(l)} \in \mathcal{S}_l$, perform a conditional independence test of $X_i^{(k)}$ and $X_j^{(l)}$ given all nodes in partitions strictly between $\mathcal{S}_k$ and $\mathcal{S}_l$, and in all partitions preceding $\mathcal{S}_k$; and (3) remove any edge for which independence is not rejected, and return the resulting $\mathcal{G}$. The pseudocode is given in Algorithm 2. Note that this baseline is agnostic to the choice of independence test (line 4 in Algorithm 2), which is left to the user. For our synthetic r-partite experiments, we utilized conditional independence Fisher-z tests with a significance level of 0.025.

The correctness of the r-partite baseline follows directly from Proposition 3 in Ankan & Textor (2025) (refer to Appendix B.2 for the formal proof).

Corollary 2. *Under causal sufficiency, faithfulness, the Markov property, and knowledge of the causal ordering of partitions, the r-partite baseline as described in Algorithm 2 converges to the ground truth DAG in the infinite data limit.*

Algorithm 2 r-Partite Baseline

Require: Partitions $\mathcal{S}_1, \ldots, \mathcal{S}_r$ (causally ordered), maximum lag ℓ
Ensure: Block-directed r-partite DAG $\mathcal{G} = (\mathbf{V}, \mathbf{E})$

```
Initialize G: add edges X_i^(k) → X_j^(l) for all X_i^(k) ∈ S_k, X_j^(l) ∈ S_l where k < l and l − k ≤ ℓ.
for all edges X_i^(k) → X_j^(l) with X_i^(k) ∈ S_k, X_j^(l) ∈ S_l do
    Construct conditioning set S = (⋃_{p=0}^{k−1} S_p) ∪ (⋃_{p=k+1}^{l−1} S_p).
    Test conditional independence of X_i^(k) and X_j^(l) given S.
    if X_i^(k) ⊥ X_j^(l) | S then
        Remove edge X_i^(k) → X_j^(l).
    end if
end for
return G.
```

3.3 Data Generating Processes

We ground our empirical analysis by evaluating the candidate models under a wide range of DGPs. Thus, we generated synthetic independent and iid data from the following distributions: linear with exponential noise (equation 3), linear with Gumbel noise (equation 4), quadratic with Gaussian noise (equation 5), Gaussian Process (GP) with Gaussian noise (equation 6, where $k_{\mathrm{RBF}}(x, x') = \exp\Big(-\frac{\|x-x'\|^2}{2\ell^2}\Big)$ is the radial basis function kernel and the inputs, x and x' are the parents to the node being simulated) and a combination setting where nodes were randomly assigned across the four DGP categories. Formally, our DGPs are defined

$$X_j = \sum_{i \in \mathrm{Pa}(j)} w_{ij} X_i + \varepsilon_j, \quad \varepsilon_j \sim \mathrm{Exp}(\lambda), \quad \text{with density } p(\varepsilon_j) = \lambda e^{-\lambda \varepsilon_j}, \quad \varepsilon_j \geq 0, \tag{3}$$

$$X_j = \sum_{i \in \mathrm{Pa}(j)} w_{ij} X_i + \varepsilon_j, \quad \varepsilon_j \sim \mathrm{Gumbel}(\mu, \beta), \tag{4}$$

$$X_j = \sum_{i \in \mathrm{Pa}(j)} a_i X_i + \sum_{i \in \mathrm{Pa}(j)} b_i X_i^2 + \sum_{i<k,\, i,k \in \mathrm{Pa}(j)} c_{ik} X_i X_k + \varepsilon_j, \quad \varepsilon_j \sim \mathcal{N}(0, \sigma^2), \tag{5}$$

$$X_j \sim \mathcal{GP}\Big(0,\ k_{\mathrm{RBF}}(x, x')\Big) + \varepsilon_j, \quad \varepsilon_j \sim \mathcal{N}(0, \sigma^2). \tag{6}$$

The linear settings were designed to satisfy the strict functional modeling assumptions of linearity for LiNGAM and NOTEARS (Shimizu et al., 2006; Zheng et al., 2018). The remaining DGPs serve to test how well the models perform when such functional assumptions are violated.

To ensure statistical robustness, we repeated each simulation ten times. Because our bipartite analysis (Figure 2c) demonstrates that performance is agnostic to linear noise types but severely degraded by quadratic non-linearities, we narrow the focus of the r-partite experiments. Specifically, we evaluate the methods on a single representative linear baseline (exponential noise) alongside the most challenging non-linear configuration (quadratic with Gaussian noise).

3.4 Benchmarked Methods

Towards a comprehensive benchmark, we test algorithms from the main causal discovery paradigms. We select PC, GES, DirectLiNGAM and NOTEARS as representatives for constrained-based, score-based, functional form-based and gradient-based methods respectively.

Every method tested was run under four regimes: **default** settings; after **post hoc pruning**, removing edges incompatible with bipartite or r-partite structure; with explicit **prior knowledge** of forbidden edges supplied; and, finally, with **both prior knowledge and pruning**. We denote any variant involving prior knowledge or pruning as a variant that incorporates structural knowledge into the method.

All algorithms included in this study intrinsically supported the input of prior knowledge of forbidden edges, with the exception of NOTEARS. To address this disparity, we modified the NOTEARS implementation to incorporate prior knowledge of forbidden edges (refer to Appendix D.1 for modification details).

3.5 Evaluation Metrics

We compare predictions from all model variants against the ground truth, our proposed baseline, and three structural controls: a full, an empty, and a randomly generated block-directed bipartite DAG.

The performance of each method is measured using standard metrics (Zheng et al., 2018): **Structural Hamming Distance (SHD)**, which measures the total number of changes required to transform an estimated graph structure into the true structure, counting edge additions, deletions, or reversals; **(arrow) precision**, which measures the proportion of correctly oriented edges among all predicted edge directions; **(arrow) recall**, which represents the proportion of true edge orientations that were correctly detected in the estimation; and, wall-clock **time** measured in seconds to run the method.[4] Full mathematical definitions of the metrics are provided in Appendix C. We focus our discussion on SHD, following standard practice in causal discovery with tabular data (see e.g., Zheng et al. (2018); Machlanski et al. (2024)). We report all other metrics in Appendix F.

Standard metrics such as SHD are context dependent, as SHD reports the absolute number of total changes required to obtain the correct DAG regardless of DAG size. Thus, to contextualize the performance of the existing methods we use our bipartite and r-partite baselines. For an easier comparison with our baselines, we report the differences in SHD between tested methods and the baselines: $\Delta\text{SHD} = \text{SHD}_{\text{method}} - \text{SHD}_{\text{baseline}}$.

Performance variations are analyzed for every combination of parameters (e.g. node count, node ratio, global density, local density, etc.). To isolate the effect of a single parameter (e.g., node count), we marginalize over the remaining parameters, reporting the means and standard deviations of the causal discovery methods' performance for the parameter of interest.

3.6 Results

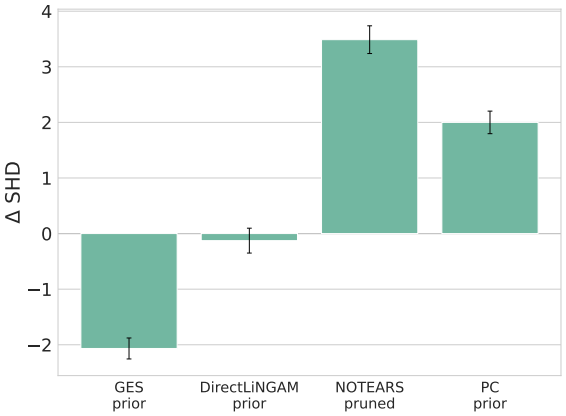
We next present empirical results for the methods introduced in Section 3.4, evaluated on the synthetic datasets described in Section 3.3. The complete set of results is available in Appendix F.

3.6.1 Bipartite Results

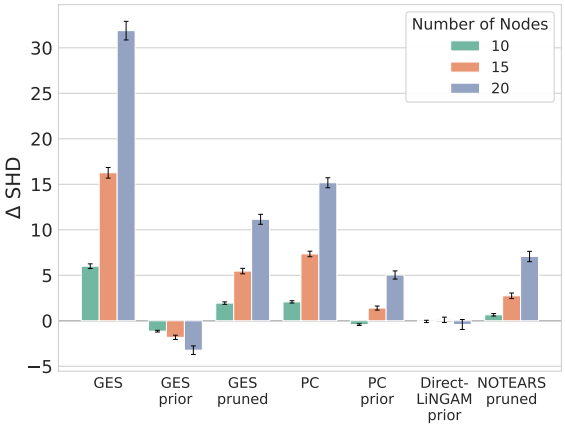
Figure 2a summarizes aggregate performance in the bipartite setting relative to the baseline. The bipartite baseline achieves SHD 8.5 ± 0.2 (refer to Appendix F.1, Table 5). The only two methods that surpass this structure-aware baseline are GES with prior knowledge, which achieves the largest aggregate SHD improvement (ΔSHD -2.1), followed by DirectLiNGAM with prior (ΔSHD -0.1). All other methods (NOTEARS, PC, and their variants) trail by a substantial margin. Based on these observations, we now further explore the question: *under which conditions can existing methods surpass our baseline.*

Number of nodes Scalability follows a consistent pattern (refer to Figure 2b). The baseline performance, as expected, degrades with increasing number of nodes: SHD 3.2 at 10 nodes, SHD 7.4 at 15 nodes and SHD 14.8 at 20 nodes (refer to Appendix F.1, Tables 6, 7 and 8). Notably, all methods degrade with scale, but only those that incorporate structural knowledge show slower degradation comparable to that of the baseline. In particular, methods with the prior knowledge variant, with the exception of NOTEARS, have the slowest degradation. For example, GES with prior obtains ΔSHDs $-1.1, -1.8$ and -3.2 at 10, 15 and 20 nodes respectively; compared to GES default with ΔSHDs $6.0, 16.3$ and 31.9. The pruned variant of

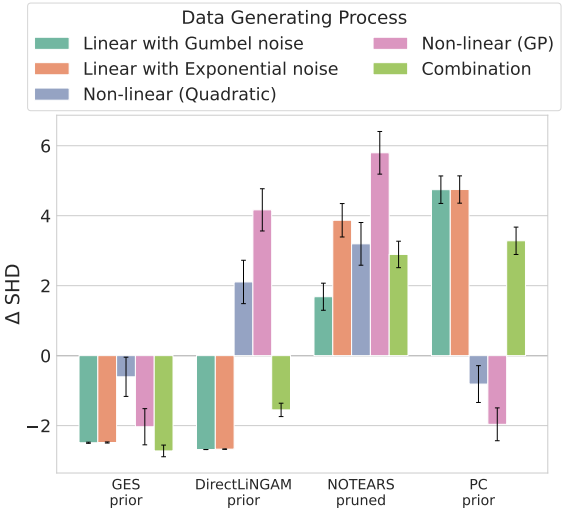
[4]The computational set up used was composed of 60 AMD EPYC™ 7003 Series Processors.



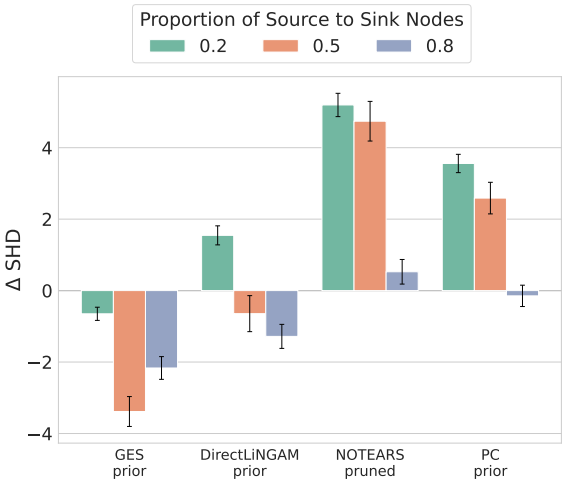
(a) Aggregate ΔSHD across all benchmark configurations. The best variant is selected for each causal discovery method (refer to Subsection 3.4 for variant descriptions).



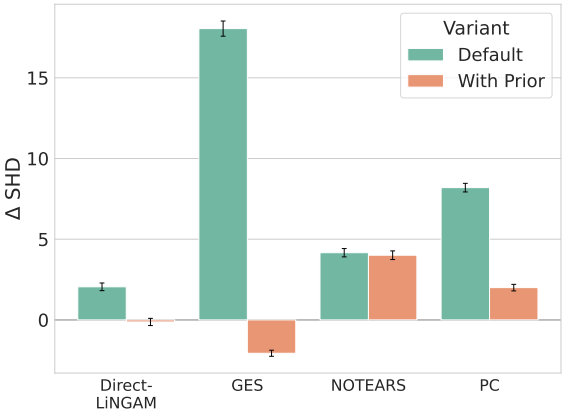
(b) ΔSHD grouped by number of nodes.



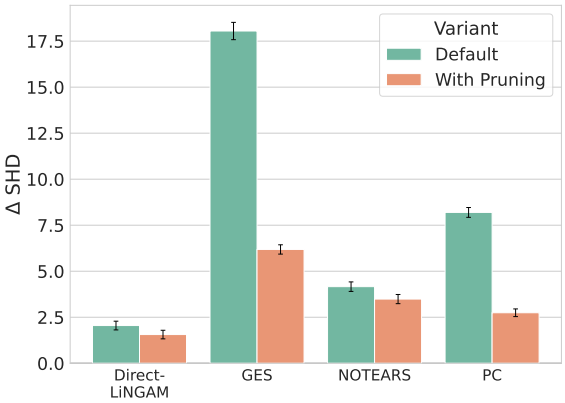
(c) ΔSHD by DGP.



(d) ΔSHD by proportion of source to sink nodes.



(e) Effect of prior knowledge on ΔSHD.



(f) Effect of post-hoc pruning on ΔSHD.

Figure 2: Summary of bipartite causal discovery results. ΔSHD denotes $\text{SHD}_{\text{method}} - \text{SHD}_{\text{baseline}}$ (lower is better). Error bars denote one standard error.

methods, although not optimal, also taper off degradation. Using the same example, GES pruned obtains ΔSHDs $1.9, 5.5$ and 11.2 at 10, 15 and 20 nodes. This suggests that leveraging structural constraints slows down degradation of methods with scale.

DGP The DGP breakdown (Figure 2c) outlines a clear trend in behavior. On linear DGPs, where functional assumptions are satisfied, DirectLiNGAM achieves near-perfect recovery (ΔSHD ≈ -2.7) and GES prior achieves ΔSHD ≈ -2.5; both far outperforming the baseline. However, on nonlinear DGPs, this advantage largely vanishes. Under the quadratic setting, all results for the methods cluster close to the results obtained by empty graph: GES prior ΔSHD -0.6 and DirectLiNGAM prior ΔSHD 2.1, compared to the empty graph at ΔSHD 4.4 and the baseline. Under Gaussian Process data, GES prior (ΔSHD -2.0) maintains a modest advantage over the baseline, but DirectLiNGAM prior (ΔSHD 4.2) no longer surpasses the expected minimum performance offered by the baseline. The noise type within linear settings (exponential vs Gumbel) had negligible effect. This pattern reveals that general methods only reliably outperform the baseline when their built-in functional assumptions are met; structural awareness alone, without assumption alignment, is insufficient for meaningful gains.

Proportion of source to sink nodes The proportion of source to sink nodes had more impact. Although balanced (50% source nodes) configurations were most challenging for all methods (refer to Appendix F.1, Tables 9, 10 and 11), attributable to the higher maximum possible edge count under balanced partitions; sink-heavy (20% source nodes) configurations were where existing methods struggled most to exceed the baseline (refer to Figure 2d).

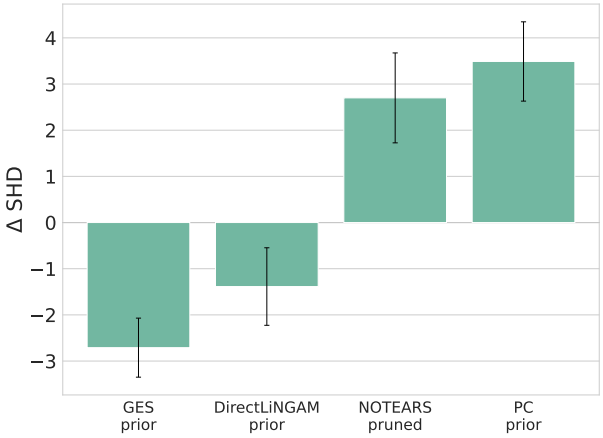
Structural knowledge The effect of the incorporation of structural knowledge, both in the form of prior knowledge and pruning, was positive across methods (Figures 2e and 2f). Nonetheless, the effectiveness of incorporating structural knowledge varied amongst methods. GES followed by PC, obtained the greatest benefit from both prior knowledge and pruning. For example, GES' overall performance improved from ΔSHD 18.0 (the worst overall SHD) to ΔSHD 6.2 with pruning, and to ΔSHD -2.1 (best performance) with the addition of prior knowledge. Although the effect of structural knowledge was milder in DirectLiNGAM and NOTEARS, it was never detrimental to any of the methods. These results highlight how using known structural knowledge of a problem, can reliably improve the performance of the causal discovery methods.

3.6.2 r-Partite Results

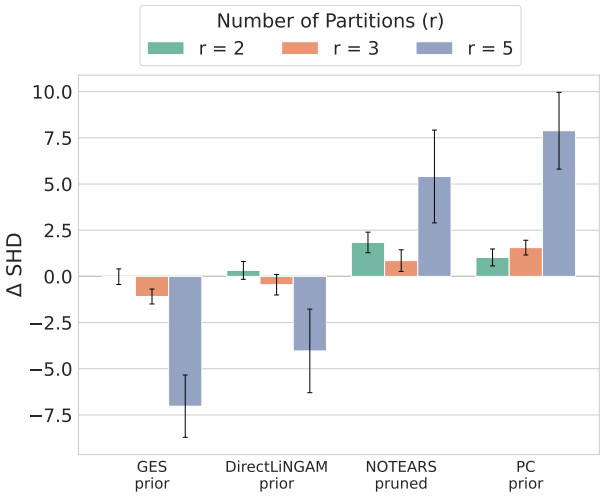
The r-partite results replicate the bipartite findings. In aggregate (Figure 3a), the r-partite baseline (SHD 10.3 ± 0.7, refer to Appendix F.2, Table 23) is only surpassed by GES prior (ΔSHD -2.7), DirectLiNGAM prior (ΔSHD -1.4) and DirectLiNGAM pruned (ΔSHD -0.1). We now study the conditions under which existing method performance surpasses our baseline.

Number of partitions The most informative dimension is the number of partitions r (Figure 3b). As expected, as the number of partitions increases, the performance of every method degrades. Nonetheless, GES prior and DirectLiNGAM prior are the only methods whose performance appears to degrade at a lower rate than that of the baseline. At $r = 2$, GES prior (ΔSHD -0.0) and DirectLiNGAM prior (ΔSHD 0.3) are both comparable to the baseline. This result is consistent with the bipartite findings, since $r = 2$ reduces to the bipartite case. At $r = 3$, we start to see a noticeable benefit in running existing causal discovery methods: GES prior (ΔSHD -1.1) and DirectLiNGAM prior (ΔSHD -0.0) as opposed to the baseline. At $r = 5$, this benefit is even larger: GES prior (ΔSHD -7.0), DirectLiNGAM prior (ΔSHD -4.0) versus the baseline. This finding suggests that as the number of partitions increases, existing method performance of GES and DirectLiNGAM surpasses the baseline as long as structural knowledge is provided.

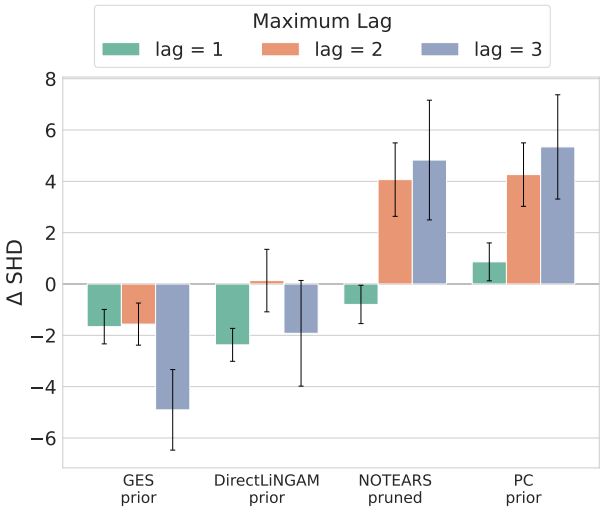
Maximum lag Maximum lag shows a similar pattern to the number of partitions (Figure 3c): method performance degrades with increased maximum lag. At lag $\ell = 1$, GES prior (ΔSHD -1.7) and DirectLiNGAM prior (ΔSHD -1.0) are comparable with the baseline. At $\ell = 3$, we observe that the degradation in performance is less for GES prior (ΔSHD -4.9) and DirectLiNGAM prior (ΔSHD -1.9) than that of the baseline.



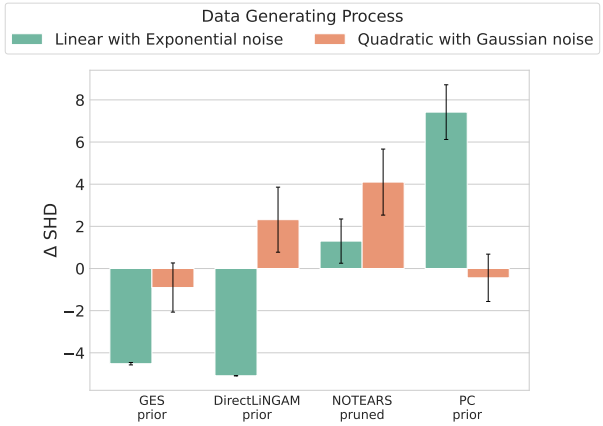
(a) Aggregate ΔSHD across all r-partite benchmark configurations. The best variant is selected for each method (refer to Subsection 3.4 for variant descriptions).



(b) ΔSHD by number of partitions $r \in \{2, 3, 5\}$.



(c) ΔSHD by maximum lag $\ell \in \{1, 2, 3\}$.



(d) ΔSHD by DGP.

Figure 3: Summary of r-partite causal discovery results across four dimensions. ΔSHD denotes $\text{SHD}_{\text{method}} - \text{SHD}_{\text{baseline}}$ (lower is better). Error bars denote one standard error.

Notably, the only methods that show a slower rate of performance degradation than the baseline are: GES prior, DirectLiNGAM prior and DirectLiNGAM pruned.

DGP The DGP pattern replicates the bipartite finding almost exactly (Figure 3d). On linear data, DirectLiNGAM achieves perfect recovery and GES prior near-perfect (ΔSHD −4.5). This is a clear gap under assumption-aligned conditions. On quadratic data, all methods cluster tightly: GES prior ΔSHD −0.9, and DirectLiNGAM prior ΔSHD 2.3. When functional assumptions are violated, the existing general methods do not obtain a statistically significant increase in performance compared to the baseline.

Taken together, the r-partite results tell a more differentiated story than the bipartite case. When the graph is small and shallow ($r = 2$, $\ell = 1$), the baseline is an adequate substitute for general methods. As the causal hierarchy grows deeper and edges span more and more distant partitions, GES prior, DirectLiNGAM prior and DirectLiNGAM pruned obtain a statistically significant increase in performance. The practical implication is that the relative value of the baseline depends on the complexity of the partition structure: it

provides a strong (empirical) lower bound for simple cases, but more sophisticated methods are warranted as r and ℓ increase, provided their assumptions are satisfied and that they incorporate structural knowledge.

4 Case Study with Realistically-Generated Time Series Data

The synthetic experiments in the previous case study are useful to understand the behavior of causal discovery methods in fully identifiable settings. However, there is limitation in that the datasets studied lack some of the complexities of the real world, such as time-delayed causation. Thus, to complement the synthetic experiments, we present a case study on realistically-generated multivariate time series: CausalTime (Cheng et al., 2023). The study consists of three datasets: Medical (20 nodes), AQI (36 nodes), and Traffic (20 nodes). Full details on the datasets may be found in Appendix E.

Time series causal discovery can be framed as an r-partite problem under the assumption of no instantaneous effects (no variable can cause another measured at the same time step). Concretely, variables measured at each time step $T, T{+}1, \ldots, T{+}(r{-}1)$ form the partitions $\mathcal{S}_1, \mathcal{S}_2, \ldots, \mathcal{S}_r$ respectively. The maximum lag ℓ corresponds to the maximum number of time steps over which a direct causal effect can propagate.

In practice, many systems naturally exhibit lagged causal dependencies, as seen in neuroscience (Butler et al., 2023) and Earth sciences (Runge et al., 2019). Thus, evaluating bipartite and r-partite formulations on such data provides insight into the efficacy of simple structural priors. Moreover, the realistic nature of these datasets allows us to test these structural priors even when some assumptions, such as no instantaneous effects, may fail. We study this problem using the r-partite formulation with maximum lags of one and three, corresponding to the bipartite ($r = 2$) and the r-partite ($r = 4$) cases, respectively.

In these experiments, we use F1 scores as our primary metric to align our presentation with recent works in time series causal discovery (Pamfil et al., 2020; Nauta et al., 2019; Sanchez et al., 2025).

4.1 Bipartite Formulation: Maximum Lag of one

We start with a bipartite ($r = 2$) formulation of the time-series problem.

Preprocessing for the baseline. The original dataset contains one observation per variable per time step. To cast the time series into a bipartite format, we follow two standard assumptions in causal time-series discovery: (i) the process is stationary, and (ii) causal effects propagate with a maximum lag of one. Under these assumptions, we construct two variable sets: the first contains all variables at time T, and the second contains the same variables at time $T{+}1$. Edges are permitted only from the T-set to the $T{+}1$-set, resulting in a bipartite dataset on which the bipartite baseline can be applied directly.

Time-series causal discovery methods. We compare the bipartite baseline (using HSIC, Spearman correlation and Fisher for the independence tests) with several state-of-the-art (SotA) time-series causal discovery methods. These include the following temporal approaches: PCMCI (and PCMCI+) (Runge, 2020), VARLiNGAM (Hyvärinen et al., 2010), DYNOTEARS (Pamfil et al., 2020), TCDF (Nauta et al., 2019), TiMINo (Peters et al., 2013) and DOTS (Sanchez et al., 2025). We incorporate the results reported by Sanchez et al. (2025), denoted by the superscript "*", and append the bipartite baseline performance using the same data processing and evaluation metric (F1 score) that was employed in their work. Further details can be found in Appendix E. Overall, this setting allows us to assess how a purely bipartite formulation, which ignores explicit temporal modeling, performs relative to methods designed specifically for time-indexed data.

Results. The results, reported in Table 3, show that even the strongest time-series SotA methods fail to surpass the bipartite baseline on the Medical dataset. In the other cases (AQI and Traffic) SotA performance remains comparable to that of the baseline despite the fact that the baseline was not explicitly designed for time-series data. Similarly, no SotA method is able to outperform bipartite baseline on average, which highlights an important point: when the true system can be modeled with a bipartite formulation, exploiting this structural prior can yield interpretable and robust solutions even before deploying more targeted

Table 3: F1 scores (higher is better) of representative time-series methods and the bipartite baseline obtained on CausalTime datasets. Best per column are in **bold**. Avg. denotes average F1 across all three datasets.

Category	Method	Medical ↑	AQI ↑	Traffic ↑	Avg. ↑
Time-series	DOTS*	0.548	0.457	0.349	0.451
	DYNOTEARS*	0.107	0.010	0.315	0.144
	PCMCI/PCMCI+*	0.427	0.382	0.372	0.393
	TCDF*	0.286	0.218	0.000	0.168
	TiMINo*	0.553	0.429	0.340	0.441
	VARLiNGAM*	0.457	**0.464**	**0.391**	0.437
Bipartite baseline	HSIC	0.504	0.394	0.271	0.390
	Spearman	**0.662**	0.395	0.362	**0.473**
	Fisher-z	0.590	0.429	0.357	0.459

methods. We also note that the choice of the independence test can play an important role as evidenced by performance differences between the bipartite baseline with HSIC and Spearman tests.

4.2 r-Partite Formulation: Maximum Lag of three

For greater modeling flexibility, we also test a larger maximum lag of three, which corresponds to a more general r-partite formulation with $r = 4$. In other words, we allow for variables to have a causal effect on others with any lag that is less than four.

Preprocessing for the baseline. The original dataset contains one observation per variable per time step. To cast the time series into an r-partite format with $r = 4$, we follow two standard assumptions in causal time-series discovery: (i) the process is stationary, and (ii) causal effects propagate with a maximum lag of three time steps. Under these assumptions, we construct four variable sets corresponding to variables at times T, $T{+}1$, $T{+}2$, and $T{+}3$, forming partitions $\mathcal{S}_1, \mathcal{S}_2, \mathcal{S}_3$, and $\mathcal{S}_4$ respectively. Edges are permitted only from earlier partitions to later ones, with no within-partition edges, resulting in a block-directed 4-partite dataset on which the r-partite baseline can be applied directly.

Time-series causal discovery methods. We compare the r-partite baseline (using Fisher-z for the independence tests) with the same SotA time-series causal discovery methods mentioned in Section 4.1. We perform the same processing steps with the evaluation metrics found in Sanchez et al. (2025), with the exception of using a maximum lag of three for all methods (for fair comparison with the baseline). Further details can be found in Appendix E. Similarly to Section 4.1, this setting allows us to assess how a purely r-partite formulation, ignoring explicit temporal modeling, performs relative to methods designed specifically for time-indexed data.

Table 4: F1 scores (higher is better) of representative time-series methods and the r-partite baseline obtained on CausalTime datasets (all with a maximum lag of three). Best per column are in **bold**. Avg. denotes average F1 across all three datasets.

Category	Method	Medical ↑	AQI ↑	Traffic ↑	Avg. ↑
Time-series	DOTS	0.538	0.436	0.354	0.443
	DYNOTEARS	0.107	0.010	0.293	0.137
	PCMCI	0.462	0.376	**0.391**	0.409
	TCDF	0.286	0.218	0.000	0.168
	TiMINo	**0.553**	0.429	0.340	0.441
	VARLINGAM	0.454	**0.499**	0.381	0.445
r-partite baseline	Fisher-z	**0.553**	0.429	0.383	**0.455**

Results. The results, summarised in Table 4, show that TiMINo is the only method that is able to match the baseline's performance (F1 score of 0.553) on the Medical dataset. Additionally, no SotA method is able to surpass the baseline's performance on average (0.455). Competitive performance is obtained by VARLiNGAM (0.445) and DOTS (0.443). On the AQI dataset, VARLiNGAM achieves the strongest result (0.499), while on Traffic, PCMCI leads (0.391) narrowly outperforming the r-partite baseline (0.383). Notably, DYNOTEARS and TCDF perform poorly across a majority of datasets, which may be due to a violation of method assumptions. For instance, DYNOTEARS assumes linearity, which may not hold here.

These results are particularly striking given that the r-partite baseline exploits only a partition structure and a marginal independence test. The achievement of the highest average F1 score against methods specifically designed for time-series data reinforces a central finding of this paper: *when the structural prior is correctly specified, a simple baseline that exploits structural knowledge is sufficient to match or exceed methods designed for the temporal domain.* The $r = 4$, $\ell = 3$ formulation appears well-suited to the causal dynamics present in these datasets, suggesting that the lag-3 assumption captures the dominant temporal dependencies without overfitting to spurious long-range associations.

Discussion. When testing a method on a real dataset, there is a possibility of assumptions being violated. However, the strong empirical performance of the bipartite baseline on the CausalTime datasets, suggests that the r-partite structural prior remains a robust inductive bias for many systems and it suggests that simple r-partite formulations may serve as an effective first step before deploying more sophisticated, or computationally expensive temporal methods.

5 Limitations

The primary limitation of this work concerns the scope of the structural settings considered. Our benchmark and theoretical guarantees are restricted to block-directed bipartite and r-partite DAGs, a specific and well-defined class of identifiable structures. While this focus is deliberate, enabling rigorous identifiability guarantees and controlled comparison, it means our findings do not directly generalize to arbitrary DAG topologies. Real-world causal systems may exhibit within-partition edges, cycles at the macro level, or hierarchical structures that do not decompose neatly into ordered partitions. Extending the baseline strategy and identifiability arguments to richer graph families, such as DAGs with partial tiered structure or cluster level constraints, remains an important direction for future work.

A second limitation concerns the synthetic nature of the benchmark. Although we designed the benchmark to span a wide range of topological configurations, density regimes, and data generating processes, synthetic datasets are generated under controlled assumptions that may not fully reflect the complexity and noise characteristics of real-world data. The CausalTime case study partially addresses this by demonstrating that our findings extend to realistic multivariate time-series data, but a broader empirical validation across domains (including genomics, epidemiology, and economics) would further strengthen the conclusions. We view such validation as a natural next step.

These limitations are inherent to any work that prioritizes theoretical grounding and controlled evaluation over broad empirical coverage. We argue that establishing a rigorous minimum performance standard in simple, provably identifiable settings is a necessary first step before making claims about performance in more complex analogs.

6 Conclusion

Causal discovery has seen remarkable theoretical progress over the past decade. Yet, translating this progress into reliable practical tools remains an open challenge. In this work, we have investigated this gap in two ways. Firstly, we evaluated established causal discovery methods on synthetic tabular datasets based on DAGs that corresponded to provably identifiable, structurally simple graphs (block-directed bipartite and r-partite DAGs) where correct recovery should, in theory, be tractable using only standard assumptions. Secondly, we compared SotA temporal causal discovery methods on a set of realistically generated datasets,

where identifiability may be violated, to understand to what extent r-partite structures can be beneficial in time-series causal discovery.

Our central finding is that existing methods do not consistently surpass the simple topological baselines that exploit only the known partition structure via marginal or conditional independence tests. This holds across a wide range of benchmark configurations: varying graph size, density, node proportions, maximum lag, and data generating process. Most importantly, the only methods that do outperform the baselines are those that, as the baselines, are given partial structural knowledge of the problem in the form of either prior knowledge or pruning. It is further corroborated by the CausalTime case study, where the bipartite and r-partite baselines achieve the highest average F1 score despite being agnostic to the temporal structure of the data. We argue that if sophisticated methods cannot reliably outperform a structurally-aware but otherwise trivial baseline in the simplest recoverable settings, their theoretical guarantees offer limited practical assurance.

We argue that this finding has a concrete methodological consequence. Simple baselines that exploit known structural priors should be established as standard practice before deploying more complex methods. A method that cannot outperform such a baseline in a provably identifiable setting provides little benefit for its extra modeling assumptions. Establishing minimum performance standards of this kind, analogous to sanity checks in supervised learning, is a necessary step toward making causal discovery a more reliable empirical discipline.

More broadly, our results suggest that structural priors are a largely underexploited resource in causal discovery. Rather than designing ever more expressive models under increasingly restrictive functional assumptions, the field may benefit from investing in methods that are explicitly designed to leverage partial structural knowledge; whether in the form of tiered orderings, cluster constraints, or partition structure. The strong performance of structurally-aware baselines, even on realistic data, suggests that this is a tractable and practically valuable direction.

Finally, our benchmarks, baselines, and identifiability results introduced in this work aim to contribute to ongoing efforts to close the gap between causal discovery theory and practice (Reisach et al., 2021; Montagna et al., 2023; Gideon et al., 2025; Faltenbacher et al., 2026), including the incorporation of prior knowledge to improve method scalability (Schubert et al., 2026).

References

Bryan Andrews, Peter Spirtes, and Gregory F Cooper. On the completeness of causal discovery in the presence of latent confounding with tiered background knowledge. In *International Conference on Artificial Intelligence and Statistics*, pp. 4002–4011. PMLR, 2020.

Ankur Ankan and Johannes Textor. Expert-in-the-loop causal discovery: Iterative model refinement using expert knowledge. In *The 41st Conference on Uncertainty in Artificial Intelligence*, 2025.

Charles K Assaad, Emilie Devijver, and Eric Gaussier. Survey and evaluation of causal discovery methods for time series. *Journal of Artificial Intelligence Research*, 73:767–819, 2022.

Christine W Bang and Vanessa Didelez. Do we become wiser with time? on causal equivalence with tiered background knowledge. In *Uncertainty in Artificial Intelligence*, pp. 119–129. PMLR, 2023.

Heloísa Oss Boll, Ali Amirahmadi, Mirfarid Musavian Ghazani, Wagner Ourique de Morais, Edison Pignaton de Freitas, Amira Soliman, Farzaneh Etminani, Stefan Byttner, and Mariana Recamonde-Mendoza. Graph neural networks for clinical risk prediction based on electronic health records: A survey. *J. Biomed. Informatics*, 151:104616, 2024.

Philippe Brouillard, Perouz Taslakian, Alexandre Lacoste, Sébastien Lachapelle, and Alexandre Drouin. Typing assumptions improve identification in causal discovery. In *Conference on Causal Learning and Reasoning*, pp. 162–177. PMLR, 2022.

Philippe Brouillard, Chandler Squires, Jonas Wahl, Konrad P Kording, Karen Sachs, Alexandre Drouin, and Dhanya Sridhar. The landscape of causal discovery data: Grounding causal discovery in real-world applications. *arXiv preprint arXiv:2412.01953*, 2024.

Kurt Butler, Duncan Cleveland, Charles B Mikell, Sima Mofakham, Yuri B Saalmann, and Petar M Djurić. An approach to learning the hierarchical organization of the frontal lobe. In *2023 31st European Signal Processing Conference (EUSIPCO)*, pp. 1410–1414. IEEE, 2023.

Edward Joseph Caruana, Marius Roman, Jules Hernández-Sánchez, and Piergiorgio Solli. Longitudinal studies. *Journal of thoracic disease*, 7(11):E537, 2015.

William Charlton. *Aristotle's Physics: Books I and II*. Oxford University Press, 1970. (Original work written c. 350 BCE).

Yuxiao Cheng, Ziqian Wang, Tingxiong Xiao, Qin Zhong, Jinli Suo, and Kunlun He. Causaltime: Realistically generated time-series for benchmarking of causal discovery. *arXiv preprint arXiv:2310.01753*, 2023.

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

Madhukar S Dasika, Anshuman Gupta, and Costas D Maranas. A mixed integer linear programming (milp) framework for inferring time delay in gene regulatory networks. In *Biocomputing 2004*, pp. 474–485. World Scientific, 2003.

Reinhard Diestel. The basics. In *Graph Theory*, pp. 1–35. Springer, 2024.

Sofia Faltenbacher, Jonas Wahl, Rebecca Herman, and Jakob Runge. How pc-based methods err: Towards better reporting of assumption violations and small sample errors, 2026. URL `https://arxiv.org/abs/2502.14719`.

Ronald Aylmer Fisher. Statistical methods for research workers. In *Breakthroughs in statistics: Methodology and distribution*, pp. 66–70. Springer, 1970.

Ane B Fisker, Amabelia Rodrigues, Cesario Martins, Henrik Ravn, Stine Byberg, Sanne Thysen, Line Storgaard, Marie Pedersen, Manuel Fernandes, Christine S Benn, et al. Reduced all-cause child mortality after general measles vaccination campaign in rural guinea-bissau. *The Pediatric infectious disease journal*, 34(12):1369–1376, 2015.

François Fouss, Marco Saerens, and Masashi Shimbo. *Algorithms and models for network data and link analysis*. Cambridge University Press, 2016.

Stein Gideon, Shadaydeh Maha, Blunk Jan, Penzel Niklas, and Denzler Joachim. Causalrivers - scaling up benchmarking of causal discovery for real-world time-series. In *The Thirteenth International Conference on Learning Representations*, 2025. URL `https://openreview.net/forum?id=wmV4cIbgl6`.

James D Hamilton. A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica: Journal of the econometric society*, pp. 357–384, 1989.

Patrik Hoyer, Dominik Janzing, Joris M Mooij, Jonas Peters, and Bernhard Schölkopf. Nonlinear causal discovery with additive noise models. *Advances in neural information processing systems*, 21, 2008.

Aapo Hyvärinen, Kun Zhang, Shohei Shimizu, and Patrik O Hoyer. Estimation of a structural vector autoregression model using non-gaussianity. *Journal of Machine Learning Research*, 11(5), 2010.

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

Sebastian Köhler, Sandra C Doelken, Christopher J Mungall, Sebastian Bauer, Helen V Firth, Isabelle Bailleul-Forestier, Graeme CM Black, Daniel L Brown, Michael Brudno, Jeremy Campbell, et al. The human phenotype ontology project: linking molecular biology and disease through phenotype data. *Nucleic Acids Research*, 42(D1):D966–D974, 2014. doi: 10.1093/nar/gkt1026.

James Leiner, Brian Manzo, Aaditya Ramdas, and Wesley Tansey. Scalable causal structure learning via amortized conditional independence testing. *arXiv preprint arXiv:2310.16626*, 2023.

Eun-Yeol Ma, Junsub Jung, and Heeyoung Kim. Order-based causal discovery for multistage processes. *IEEE Transactions on Knowledge and Data Engineering*, 2026.

Damian Machlanski, Spyridon Samothrakis, and Paul S Clarke. Robustness of algorithms for causal structure learning to hyperparameter choice. In Francesco Locatello and Vanessa Didelez (eds.), *Proceedings of the Third Conference on Causal Learning and Reasoning*, volume 236 of *Proceedings of Machine Learning Research*, pp. 703–739. PMLR, 01–03 Apr 2024. URL `https://proceedings.mlr.press/v236/machlanski24a.html`.

Francesco Montagna, Atalanti Mastakouri, Elias Eulig, Nicoletta Noceti, Lorenzo Rosasco, Dominik Janzing, Bryon Aragam, and Francesco Locatello. Assumption violations in causal discovery and the robustness of score matching. *Advances in Neural Information Processing Systems*, 36:47339–47378, 2023.

Hans-Peter Müller, Anna Behler, G Bernhard Landwehrmeyer, Hans-Jürgen Huppertz, and Jan Kassubek. How to arrange follow-up time-intervals for longitudinal brain mri studies in neurodegenerative diseases. *Frontiers in Neuroscience*, 15:682812, 2021.

Meike Nauta, Doina Bucur, and Christin Seifert. Causal discovery with attention-based convolutional neural networks. *Machine Learning and Knowledge Extraction*, 1(1):19, 2019.

Roxana Pamfil, Nisara Sriwattanaworachai, Shaan Desai, Philip Pilgerstorfer, Konstantinos Georgatzis, Paul Beaumont, and Bryon Aragam. Dynotears: Structure learning from time-series data. In *International Conference on Artificial Intelligence and Statistics*, pp. 1595–1605. Pmlr, 2020.

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

Judea Pearl and Thomas S. Verma. A theory of inferred causation. In Dag Prawitz, Brian Skyrms, and Dag Westerståhl (eds.), *Logic, Methodology and Philosophy of Science IX*, volume 134 of *Studies in Logic and the Foundations of Mathematics*, pp. 789–811. Elsevier, 1995. doi: https://doi.org/10.1016/S0049-237X(06)80074-1. URL `https://www.sciencedirect.com/science/article/pii/S0049237X06800741`.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. Causal inference on time series using restricted structural equation models. *Advances in neural information processing systems*, 26, 2013.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: foundations and learning algorithms*. The MIT press, 2017.

Alexander Reisach, Christof Seiler, and Sebastian Weichwald. Beware of the simulated dag! causal discovery benchmarks may be easy to game. *Advances in Neural Information Processing Systems*, 34:27772–27784, 2021.

Robert W Robinson. Counting unlabeled acyclic digraphs. In *Combinatorial Mathematics V: Proceedings of the Fifth Australian Conference, Held at the Royal Melbourne Institute of Technology, August 24–26, 1976*, pp. 28–43. Springer, 2006.

Jan Marco Ruiz de Vargas, Kirtan Padh, and Niki Kilbertus. Cluster-dags as powerful background knowledge for causal discovery. *arXiv e-prints*, pp. arXiv–2512, 2025.

Jakob Runge. Discovering contemporaneous and lagged causal relations in autocorrelated nonlinear time series datasets. In *Conference on Uncertainty in Artificial Intelligence*, pp. 1388–1397. PMLR, 2020.

Jakob Runge, Sebastian Bathiany, Erik Bollt, Gustau Camps-Valls, Dim Coumou, Ethan Deyle, Clark Glymour, Marlene Kretschmer, Miguel D Mahecha, Jordi Muñoz-Marí, et al. Inferring causation from time series in earth system sciences. *Nature communications*, 10(1):2553, 2019.

Pedro Sanchez, Damian Machlanski, Steven McDonagh, and Sotirios A. Tsaftaris. Causal ordering for structure learning from time series. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL `https://openreview.net/forum?id=hWuTzqggSd`.

Mátyás Schubert, Theofanis Aslanidis, Tom Claassen, and Sara Magliacane. Integrating background knowledge for scalable causal discovery. *arXiv preprint arXiv:2607.10456*, 2026.

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

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

Peter Spirtes and Clark Glymour. An algorithm for fast recovery of sparse causal graphs. *Social science computer review*, 9(1):62–72, 1991.

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

Peter WG Tennant, Eleanor J Murray, Kellyn F Arnold, Laurie Berrie, Matthew P Fox, Sarah C Gadd, Wendy J Harrison, Claire Keeble, Lynsie R Ranker, Johannes Textor, et al. Use of directed acyclic graphs (dags) to identify confounders in applied health research: review and recommendations. *International journal of epidemiology*, 50(2):620–632, 2021.

N Zdravkovska. Crc standard mathematical tables and formulas. *CHOICE: Current Reviews for Academic Libraries*, 56(2):184–185, 2018.

Xun Zheng, Bryon Aragam, Pradeep K Ravikumar, and Eric P Xing. Dags with no tears: Continuous optimization for structure learning. *Advances in neural information processing systems*, 31, 2018.

A Appendix

B Proof of Results

B.1 Proofs of Identifiability Results

Proof. **of Theorem 1.** Let $\mathcal{A}$ and $\mathcal{B}$ be the sets of source and sink nodes in the bipartite graph, respectively.

Denote the elements of $\mathcal{A}$ and $\mathcal{B}$ as $X_i^{(\mathcal{A})}$ and $X_j^{(\mathcal{B})}$ respectively, where $i \in \{1, \ldots, n\}$ and $j \in \{1, \ldots, m\}$.

By the Markov property, if $X_i^{(\mathcal{A})} \in \mathcal{A}$ and $X_j^{(\mathcal{B})} \in \mathcal{B}$ are dependent variables conditioning on the empty set, then $X_i^{(\mathcal{A})} - X_j^{(\mathcal{B})}$. By the knowledge of directionality of causality, we have that $X_i^{(\mathcal{A})} \to X_j^{(\mathcal{B})}$.

By faithfulness, if $X_i^{(\mathcal{A})} \to X_j^{(\mathcal{B})}$, then $X_i^{(\mathcal{A})}$ and $X_j^{(\mathcal{B})}$ must be dependent variables when we condition on the empty set.

Thus, every edge, and hence the whole DAG, is identifiable.

□

Proof. **of Theorem 2.** Let $\mathcal{S}_1, \mathcal{S}_2, \ldots, \mathcal{S}_r$ be the r partitions of $\mathcal{G} = (\mathbf{V}, \mathbf{E})$, ordered such that causal influence flows from $\mathcal{S}_k$ to $\mathcal{S}_l$ only if $k < l$.

Denote the elements of partition $\mathcal{S}_k$ as $X_i^{(k)}$, where $i \in \{1, \ldots, n_k\}$ and n_k is the number of nodes in $\mathcal{S}_k$.

By the Markov property, if $X_i^{(k)} \in \mathcal{S}_k$ and $X_j^{(l)} \in \mathcal{S}_l$ are dependent conditioning on the empty set, then there exists an edge $X_i^{(k)} - X_j^{(l)}$ in the skeleton of $\mathcal{G}$. By knowledge of the causal ordering of partitions, since $k < l$, this edge must be oriented as $X_i^{(k)} \to X_j^{(l)}$.

By faithfulness, if $X_i^{(k)} \to X_j^{(l)}$, then $X_i^{(k)}$ and $X_j^{(l)}$ must be dependent when conditioning on the empty set. Since every edge in $\mathbf{E}$ connects nodes in distinct partitions and the causal ordering uniquely determines the direction of each such edge, every edge in $\mathcal{G}$ is identifiable. Hence, the entire DAG $\mathcal{G}$ is identifiable. □

B.2 Proofs of Correctness Results

Proof. **of Corollary 1.** By Theorem 1, there exists a unique ground truth $\mathcal{G}$ under these assumptions (since the combination of faithfulness and the Markov assumption implies minimality (Spirtes et al., 2000)).

Since the $\mathcal{G}$ must be a block-directed bipartite DAG, the full block-directed bipartite DAG is a supergraph $\mathcal{G}'$ of $\mathcal{G}$. Since there are no interconnections between elements of $\mathcal{A}$ or $\mathcal{B}$, two variables $X_i^{(\mathcal{A})} \in \mathcal{A}, X_j^{(\mathcal{B})} \in \mathcal{B}$ are d-separated by the empty set if and only if $X_i^{(\mathcal{A})} \perp X_j^{(\mathcal{B})}$. Therefore, the bipartite baseline follows the framework defined in Algorithm 3 in (Ankan & Textor, 2025), where Z is used as the conditioning set for all conditional independence tests. Thus, by Proposition 3 in (Ankan & Textor, 2025), the bipartite baseline correctly converges to $\mathcal{G}$.

□

Proof. **of Corollary 2.** By Theorem 2, there exists a unique ground truth $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ under these assumptions (since the combination of faithfulness and the Markov assumption implies minimality (Spirtes et al., 2000)).

Since $\mathcal{G}$ must be a block-directed r-partite DAG with $\mathbf{V} = (\mathcal{S}_1, \ldots, \mathcal{S}_r)$, the fully connected block-directed r-partite DAG (with maximum lag ℓ) is a supergraph $\mathcal{G}'$ of $\mathcal{G}$. Since there are no edges within any partition $\mathcal{S}_k$, and all edges in $\mathbf{E}$ respect the causal ordering, two variables $X_i^{(k)} \in \mathcal{S}_k$ and $X_j^{(l)} \in \mathcal{S}_l$ (with $k < l$) are d-separated given the conditioning set $\mathbf{S} = \left(\bigcup_{p=0}^{k-1} \mathcal{S}_p\right) \cup \left(\bigcup_{p=k+1}^{l-1} \mathcal{S}_p\right)$ if and only if $X_i^{(k)} \perp X_j^{(l)} \mid \mathbf{S}$. Therefore, the r-partite baseline follows the framework defined in Algorithm 3 in (Ankan & Textor, 2025), where $\mathbf{S}$ is used as the conditioning set for all conditional independence tests. Thus, by Proposition 3 in (Ankan & Textor, 2025), the r-partite baseline correctly converges to $\mathcal{G}$. □

C Mathematical Definitions of Evaluation Metrics

The evaluation metrics used are the following:

- *SHD (Structural Hamming Distance)*, which measures the total number of changes required to transform an estimated graph structure into the true structure, counting edge additions, deletions, or reversals,

$$\text{SHD} \;=\; \left|\hat{E} \setminus E^\star\right| \;+\; \left|E^\star \setminus \hat{E}\right| + \;2 \cdot \left|\{\, (i \to j) \in \hat{E} : (j \to i) \in E^\star \,\}\right|;$$

- *(arrow) precision*, which measures the proportion of correctly oriented edges among all predicted edge directions,

$$\frac{\text{TP}}{\text{TP} + \text{FP}};$$

- *(arrow) recall*, which represents the proportion of true edge orientations that were correctly detected in the estimation,

$$\frac{\text{TP}}{\text{TP} + \text{FN}};$$

and,

- wall-clock time taken in seconds to run the method,

where E^* is the true edge set, $\hat{E}$ is the estimated edge set, A^* is the true oriented edges, $\hat{A}$ is the estimated oriented edges and TP, FP and FN stand for true positive, false positive and false negative, respectively.

D Implementation of Causal Discovery Methods

The benchmarked methods described in Section 3.4, were implemented in the following Python packages:

- PC (default and prior variants): `castle`.
- GES (default and prior variants): `pgmpy`.
- DirectLiNGAM (default and prior variants): `lingam`.
- NOTEARS (default variant): `castle`. As NOTEARS does not natively incorporate prior knowledge we coded a modification of the algorithm based on the NOTEARS `castle` implementation. The detailed description of this modification can be found in Subsection D.1.

The pruned and prior pruned variants of these methods are simply the default and prior variants with post hoc pruning, respectively.

D.1 NOTEARS Modification

The NOTEARS algorithm ("Non-combinatorial Optimization via Trace Exponential and Augmented lagRangian for Structure learning") learns a DAG from observational data by turning the combinatorial acyclicity constraint into a smooth continuous one.

The objective function NOTEARS minimizes is:

$$\min_{W \in \mathbb{R}^{d \times d}} F(W) = \frac{1}{2n} \|X - XW\|_F^2 + \lambda \|W\|_1 \tag{7}$$

subject to the acyclicity constraint

$$h(W) = \operatorname{tr}\left(e^{W \circ W}\right) - d = 0, \tag{8}$$

where

- $X \in \mathbb{R}^{n \times d}$: data matrix with n samples and d variables,
- $W \in \mathbb{R}^{d \times d}$: weighted adjacency matrix (to be learned),
- $\| \cdot \|_F$: Frobenius norm,
- $\|W\|_1$: element-wise ℓ_1 penalty to promote sparsity,
- $W \circ W$: Hadamard (element-wise) square,
- $e^{\cdot}$: matrix exponential,
- $h(W) = 0$: condition that the learned graph is acyclic.

The original NOTEARS algorithm does not support the integration of expert or prior information, potentially altering predictions and impacting performance. We adapt NOTEARS to allow the incorporation of prior knowledge in the form of forbidden edges. Specifically, we modify the gradient-based optimization procedure by introducing a mask that encodes edges excluded by prior knowledge. At each iteration, this mask is applied element-wise to the adjacency matrix, ensuring that the corresponding entries remain zero throughout training. This guarantees that the learned graph respects the forbidden-edge constraints, while otherwise

following the standard NOTEARS optimization scheme. The pseudocode for this modification can be found in Algorithm 3.

Algorithm 3 Modified NOTEARS with Forbidden Edge Mask

Require: Data matrix $X \in \mathbb{R}^{n \times d}$, forbidden mask M
Ensure: Adjacency matrix W
1: Initialize W.
2: **while** not converged **do**
3: Compute gradient $\nabla L(W; X)$.
4: Update: $W \leftarrow W - \eta \nabla L(W; X)$.
5: Enforce: $W \leftarrow W \odot M$.
6: **end while**
7: **return** W.

E Details on the Time-Series Case Study

E.1 Datasets

To complement the synthetic experiments, we present a realistic-data case study on multivariate time series: CausalTime (Cheng et al., 2023). It consists of three datasets.

- **Medical**: containing 20 nodes representing the vital-sign and chart-event channels for 1000 MIMIC-IV ICU stays, discretized to a 2-hour temporal resolution.
- **Air Quality Index (AQI)**: containing 36 nodes representing monitoring stations in China for $PM_{2.5}$.
- **Traffic**: containing 20 nodes representing loop detectors for the San-Francisco Bay Area capturing average speed every 5 minutes.

Each dataset is processed according to (Sanchez et al., 2025) and has a total of 19679 samples. For further information on the treatment of the data, please refer to (Sanchez et al., 2025).

E.2 Performance Metric

The evaluation metric for the time-series case study is the F1 score on the *summary graph* (Assaad et al., 2022). The F1 score is a classification metric that measures trade-off between precision and recall by taking their harmonic mean, and is especially useful in settings with class imbalance.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \qquad \text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}.$$

The F1 score is defined as

$$\text{F1} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}.$$

Equivalently, substituting precision and recall gives

$$\text{F1} = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}},$$

where TP, FP, and FN denote the number of true positives, false positives, and false negatives, respectively.

E.3 Significance Levels for Conditional Independence Tests in Baselines

Bipartite baseline. The following significance levels (α) were used for the conditional independence tests:

- HSIC: $\alpha = 0.01$
- Spearman correlation: $\alpha = 0.5$
- Fisher z-test: $\alpha = 0.01$

r-partite baseline. For the r-partite baseline, the conditional independence tests were conducted using:

- Fisher z-test: $\alpha = 0.05$

F Full Results

F.1 Block-Directed Bipartite DAGs

Method	SHD	Precision	Recall	Time (s)
Bipartite baseline	8.460 ± 0.202	0.945 ± 0.003	0.665 ± 0.005	1.158 ± 0.010
DirectLiNGAM	10.510 ± 0.237	0.682 ± 0.007	0.546 ± 0.007	0.721 ± 0.006
DirectLiNGAM prior	8.333 ± 0.224	0.893 ± 0.005	0.655 ± 0.006	0.393 ± 0.004
DirectLiNGAM prior pruned	8.333 ± 0.224	0.893 ± 0.005	0.655 ± 0.006	0.393 ± 0.004
DirectLiNGAM pruned	10.020 ± 0.233	0.751 ± 0.007	0.546 ± 0.007	0.721 ± 0.006
Empty	21.675 ± 0.270	0.000 ± 0.000	0.000 ± 0.000	0.001 ± 0.000
Full	24.214 ± 0.287	0.473 ± 0.004	0.998 ± 0.001	0.001 ± 0.000
GES	26.509 ± 0.470	0.453 ± 0.004	0.479 ± 0.005	4.537 ± 0.076
GES prior	6.394 ± 0.189	0.965 ± 0.002	0.765 ± 0.005	1.341 ± 0.024
GES prior pruned	6.394 ± 0.189	0.965 ± 0.002	0.765 ± 0.005	1.341 ± 0.024
GES pruned	14.644 ± 0.253	0.859 ± 0.004	0.479 ± 0.005	4.537 ± 0.076
Notears	12.622 ± 0.258	0.790 ± 0.006	0.549 ± 0.006	5.347 ± 0.168
Notears prior	12.468 ± 0.269	0.671 ± 0.007	0.501 ± 0.007	0.252 ± 0.005
Notears prior pruned	12.468 ± 0.269	0.671 ± 0.007	0.501 ± 0.007	0.252 ± 0.005
Notears pruned	11.948 ± 0.249	0.824 ± 0.006	0.549 ± 0.006	5.347 ± 0.168
PC	16.655 ± 0.267	0.615 ± 0.004	0.615 ± 0.005	0.730 ± 0.036
PC prior	10.460 ± 0.203	0.919 ± 0.002	0.677 ± 0.004	0.356 ± 0.026
PC prior pruned	10.460 ± 0.203	0.919 ± 0.002	0.677 ± 0.004	0.356 ± 0.026
PC pruned	11.204 ± 0.209	0.916 ± 0.003	0.615 ± 0.005	0.730 ± 0.036
Random	22.927 ± 0.211	0.474 ± 0.004	0.498 ± 0.003	0.001 ± 0.000

Table 5: Total results for all configurations. We show the average value and plus/minus one standard error for each test result, obtained from a sample of 4050 repeated experiments.

Method	SHD	Precision	Recall	Time (s)
Bipartite baseline	3.199 ± 0.101	0.951 ± 0.006	0.681 ± 0.008	0.562 ± 0.004
DirectLiNGAM	4.340 ± 0.145	0.666 ± 0.012	0.545 ± 0.012	0.289 ± 0.002
DirectLiNGAM prior	3.125 ± 0.127	0.899 ± 0.008	0.685 ± 0.010	0.144 ± 0.001
DirectLiNGAM prior pruned	3.125 ± 0.127	0.899 ± 0.008	0.685 ± 0.010	0.144 ± 0.001
DirectLiNGAM pruned	4.116 ± 0.141	0.727 ± 0.012	0.545 ± 0.012	0.289 ± 0.002
Empty	9.048 ± 0.138	0.000 ± 0.000	0.000 ± 0.000	0.001 ± 0.000
Full	9.952 ± 0.141	0.476 ± 0.007	0.993 ± 0.002	0.001 ± 0.000
GES	9.195 ± 0.258	0.505 ± 0.009	0.518 ± 0.009	1.244 ± 0.018
GES prior	2.055 ± 0.094	0.967 ± 0.004	0.808 ± 0.008	0.430 ± 0.009
GES prior pruned	2.055 ± 0.094	0.967 ± 0.004	0.808 ± 0.008	0.430 ± 0.009
GES pruned	5.137 ± 0.135	0.869 ± 0.008	0.518 ± 0.009	1.244 ± 0.018
Notears	4.050 ± 0.137	0.811 ± 0.010	0.605 ± 0.011	0.949 ± 0.028
Notears prior	4.136 ± 0.145	0.724 ± 0.012	0.565 ± 0.012	0.042 ± 0.000
Notears prior pruned	4.136 ± 0.145	0.724 ± 0.012	0.565 ± 0.012	0.042 ± 0.000
Notears pruned	3.853 ± 0.135	0.835 ± 0.010	0.605 ± 0.011	0.949 ± 0.028
PC	5.274 ± 0.127	0.661 ± 0.007	0.702 ± 0.008	0.135 ± 0.005
PC prior	2.772 ± 0.090	0.923 ± 0.005	0.781 ± 0.007	0.092 ± 0.004
PC prior pruned	2.772 ± 0.090	0.923 ± 0.005	0.781 ± 0.007	0.092 ± 0.004
PC pruned	3.250 ± 0.098	0.909 ± 0.006	0.702 ± 0.008	0.135 ± 0.005
Random	9.444 ± 0.082	0.479 ± 0.007	0.497 ± 0.006	0.001 ± 0.000

Table 6: Results for networks with 10 nodes. We show the average value and plus/minus one standard error for each test result, obtained from a sample of 1350 repeated experiments.

Method	SHD	Precision	Recall	Time (s)
Bipartite baseline	7.358 ± 0.250	0.950 ± 0.006	0.677 ± 0.008	1.096 ± 0.008
DirectLiNGAM	9.409 ± 0.313	0.697 ± 0.011	0.554 ± 0.012	0.636 ± 0.002
DirectLiNGAM prior	7.460 ± 0.294	0.906 ± 0.008	0.656 ± 0.011	0.330 ± 0.002
DirectLiNGAM prior pruned	7.460 ± 0.294	0.906 ± 0.008	0.656 ± 0.011	0.330 ± 0.002
DirectLiNGAM pruned	9.001 ± 0.308	0.769 ± 0.011	0.554 ± 0.012	0.636 ± 0.002
Empty	19.963 ± 0.294	0.000 ± 0.000	0.000 ± 0.000	0.001 ± 0.000
Full	22.704 ± 0.308	0.469 ± 0.006	1.000 ± 0.000	0.001 ± 0.000
GES	23.621 ± 0.585	0.451 ± 0.007	0.483 ± 0.008	3.781 ± 0.062
GES prior	5.536 ± 0.234	0.967 ± 0.003	0.765 ± 0.009	1.137 ± 0.024
GES prior pruned	5.536 ± 0.234	0.967 ± 0.003	0.765 ± 0.009	1.137 ± 0.024
GES pruned	12.817 ± 0.302	0.874 ± 0.006	0.483 ± 0.008	3.781 ± 0.062
Notears	10.650 ± 0.312	0.815 ± 0.009	0.557 ± 0.011	4.802 ± 0.195
Notears prior	10.836 ± 0.345	0.686 ± 0.013	0.509 ± 0.012	0.282 ± 0.009
Notears prior pruned	10.836 ± 0.345	0.686 ± 0.013	0.509 ± 0.012	0.282 ± 0.009
Notears pruned	10.104 ± 0.306	0.850 ± 0.009	0.557 ± 0.011	4.802 ± 0.195
PC	14.700 ± 0.304	0.609 ± 0.006	0.617 ± 0.008	0.598 ± 0.040
PC prior	8.754 ± 0.224	0.922 ± 0.003	0.676 ± 0.007	0.309 ± 0.025
PC prior pruned	8.754 ± 0.224	0.922 ± 0.003	0.676 ± 0.007	0.309 ± 0.025
PC pruned	9.474 ± 0.231	0.926 ± 0.004	0.617 ± 0.008	0.598 ± 0.040
Random	21.379 ± 0.151	0.468 ± 0.007	0.499 ± 0.004	0.001 ± 0.000

Table 7: Results for networks with 15 nodes. We show the average value and plus/minus one standard error for each test result, obtained from a sample of 1350 repeated experiments.

Method	SHD	Precision	Recall	Time (s)
Bipartite baseline	14.824 ± 0.493	0.935 ± 0.006	0.637 ± 0.009	1.817 ± 0.013
DirectLiNGAM	17.780 ± 0.566	0.682 ± 0.011	0.537 ± 0.012	1.236 ± 0.003
DirectLiNGAM prior	14.415 ± 0.547	0.873 ± 0.009	0.623 ± 0.011	0.705 ± 0.003
DirectLiNGAM prior pruned	14.415 ± 0.547	0.873 ± 0.009	0.623 ± 0.011	0.705 ± 0.003
DirectLiNGAM pruned	16.944 ± 0.559	0.758 ± 0.011	0.537 ± 0.012	1.236 ± 0.003
Empty	36.015 ± 0.526	0.000 ± 0.000	0.000 ± 0.000	0.001 ± 0.000
Full	39.985 ± 0.540	0.474 ± 0.006	1.000 ± 0.000	0.001 ± 0.000
GES	46.712 ± 1.024	0.402 ± 0.006	0.435 ± 0.007	8.586 ± 0.163
GES prior	11.590 ± 0.471	0.960 ± 0.004	0.723 ± 0.009	2.454 ± 0.056
GES prior pruned	11.590 ± 0.471	0.960 ± 0.004	0.723 ± 0.009	2.454 ± 0.056
GES pruned	25.978 ± 0.548	0.834 ± 0.006	0.435 ± 0.007	8.586 ± 0.163
Notears	23.167 ± 0.586	0.742 ± 0.010	0.485 ± 0.010	10.289 ± 0.426
Notears prior	22.431 ± 0.618	0.605 ± 0.013	0.430 ± 0.012	0.433 ± 0.011
Notears prior pruned	22.431 ± 0.618	0.605 ± 0.013	0.430 ± 0.012	0.433 ± 0.011
Notears pruned	21.887 ± 0.569	0.786 ± 0.010	0.485 ± 0.010	10.289 ± 0.426
PC	29.991 ± 0.549	0.575 ± 0.006	0.527 ± 0.007	1.456 ± 0.098
PC prior	19.854 ± 0.449	0.913 ± 0.003	0.575 ± 0.007	0.666 ± 0.072
PC prior pruned	19.854 ± 0.449	0.913 ± 0.003	0.575 ± 0.007	0.666 ± 0.072
PC pruned	20.890 ± 0.459	0.913 ± 0.004	0.527 ± 0.007	1.456 ± 0.098
Random	37.959 ± 0.261	0.475 ± 0.006	0.499 ± 0.003	0.001 ± 0.000

Table 8: Results for networks with 20 nodes. We show the average value and plus/minus one standard error for each test result, obtained from a sample of 1350 repeated experiments.

Method	SHD	Precision	Recall	Time (s)
Bipartite baseline	4.162 ± 0.176	0.985 ± 0.003	0.795 ± 0.006	0.987 ± 0.013
DirectLiNGAM	9.133 ± 0.330	0.743 ± 0.009	0.551 ± 0.011	0.717 ± 0.011
DirectLiNGAM prior	5.708 ± 0.266	0.988 ± 0.003	0.728 ± 0.009	0.355 ± 0.007
DirectLiNGAM prior pruned	5.708 ± 0.266	0.988 ± 0.003	0.728 ± 0.009	0.355 ± 0.007
DirectLiNGAM pruned	8.304 ± 0.314	0.881 ± 0.009	0.551 ± 0.011	0.717 ± 0.011
Empty	18.319 ± 0.371	0.000 ± 0.000	0.000 ± 0.000	0.001 ± 0.000
Full	20.348 ± 0.385	0.472 ± 0.006	0.996 ± 0.002	0.001 ± 0.000
GES	21.530 ± 0.688	0.452 ± 0.007	0.560 ± 0.007	4.296 ± 0.108
GES prior	3.513 ± 0.184	0.979 ± 0.002	0.852 ± 0.006	0.595 ± 0.012
GES prior pruned	3.513 ± 0.184	0.979 ± 0.002	0.852 ± 0.006	0.595 ± 0.012
GES pruned	9.875 ± 0.300	0.930 ± 0.005	0.560 ± 0.007	4.296 ± 0.108
Notears	10.193 ± 0.348	0.904 ± 0.006	0.579 ± 0.010	7.882 ± 0.402
Notears prior	13.029 ± 0.411	0.614 ± 0.013	0.423 ± 0.012	0.242 ± 0.008
Notears prior pruned	13.029 ± 0.411	0.614 ± 0.013	0.423 ± 0.012	0.242 ± 0.008
Notears pruned	9.358 ± 0.326	0.963 ± 0.005	0.579 ± 0.010	7.882 ± 0.402
PC	17.745 ± 0.478	0.528 ± 0.006	0.638 ± 0.008	1.475 ± 0.103
PC prior	7.720 ± 0.256	0.952 ± 0.003	0.725 ± 0.007	0.833 ± 0.075
PC prior pruned	7.720 ± 0.256	0.952 ± 0.003	0.725 ± 0.007	0.833 ± 0.075
PC pruned	8.677 ± 0.264	0.945 ± 0.004	0.638 ± 0.008	1.475 ± 0.103
Random	19.332 ± 0.279	0.473 ± 0.007	0.498 ± 0.005	0.001 ± 0.000

Table 9: Results for networks with 20% of nodes designated as sources. We show the average value and plus/minus one standard error for each test result, obtained from a sample of 1350 repeated experiments.

Method	SHD	Precision	Recall	Time (s)
Bipartite baseline	12.103 ± 0.456	0.984 ± 0.003	0.626 ± 0.008	1.447 ± 0.020
DirectLiNGAM	13.741 ± 0.520	0.686 ± 0.012	0.535 ± 0.012	0.721 ± 0.011
DirectLiNGAM prior	11.458 ± 0.504	0.933 ± 0.007	0.630 ± 0.011	0.360 ± 0.005
DirectLiNGAM prior pruned	11.458 ± 0.504	0.933 ± 0.007	0.630 ± 0.011	0.360 ± 0.005
DirectLiNGAM pruned	13.313 ± 0.514	0.738 ± 0.012	0.535 ± 0.012	0.721 ± 0.011
Empty	28.259 ± 0.575	0.000 ± 0.000	0.000 ± 0.000	0.001 ± 0.000
Full	32.074 ± 0.609	0.469 ± 0.006	1.000 ± 0.000	0.001 ± 0.000
GES	32.795 ± 0.938	0.471 ± 0.007	0.459 ± 0.007	5.185 ± 0.153
GES prior	8.719 ± 0.418	0.977 ± 0.002	0.753 ± 0.008	1.669 ± 0.043
GES prior pruned	8.719 ± 0.418	0.977 ± 0.002	0.753 ± 0.008	1.669 ± 0.043
GES pruned	20.314 ± 0.548	0.859 ± 0.005	0.459 ± 0.007	5.185 ± 0.153
Notears	17.641 ± 0.573	0.793 ± 0.010	0.518 ± 0.011	5.931 ± 0.269
Notears prior	16.301 ± 0.585	0.714 ± 0.012	0.502 ± 0.012	0.258 ± 0.009
Notears prior pruned	16.301 ± 0.585	0.714 ± 0.012	0.502 ± 0.012	0.258 ± 0.009
Notears pruned	16.844 ± 0.555	0.821 ± 0.010	0.518 ± 0.011	5.931 ± 0.269
PC	19.801 ± 0.524	0.683 ± 0.005	0.591 ± 0.007	0.378 ± 0.013
PC prior	14.692 ± 0.442	0.921 ± 0.003	0.641 ± 0.007	0.178 ± 0.004
PC prior pruned	14.692 ± 0.442	0.921 ± 0.003	0.641 ± 0.007	0.178 ± 0.004
PC pruned	15.497 ± 0.456	0.925 ± 0.003	0.591 ± 0.007	0.378 ± 0.013
Random	30.080 ± 0.433	0.471 ± 0.007	0.500 ± 0.004	0.001 ± 0.000

Table 10: Results for networks with 50% of nodes designated as sources. We show the average value and plus/minus one standard error for each test result, obtained from a sample of 1350 repeated experiments.

Method	SHD	Precision	Recall	Time (s)
Bipartite baseline	9.116 ± 0.323	0.867 ± 0.009	0.573 ± 0.010	1.041 ± 0.014
DirectLiNGAM	8.656 ± 0.341	0.616 ± 0.013	0.551 ± 0.013	0.724 ± 0.011
DirectLiNGAM prior	7.834 ± 0.336	0.757 ± 0.012	0.606 ± 0.012	0.463 ± 0.007
DirectLiNGAM prior pruned	7.834 ± 0.336	0.757 ± 0.012	0.606 ± 0.012	0.463 ± 0.007
DirectLiNGAM pruned	8.444 ± 0.339	0.635 ± 0.013	0.551 ± 0.013	0.724 ± 0.011
Empty	18.448 ± 0.371	0.000 ± 0.000	0.000 ± 0.000	0.001 ± 0.000
Full	20.219 ± 0.394	0.479 ± 0.006	0.996 ± 0.002	0.001 ± 0.000
GES	25.203 ± 0.765	0.436 ± 0.009	0.417 ± 0.009	4.130 ± 0.125
GES prior	6.950 ± 0.318	0.939 ± 0.005	0.691 ± 0.010	1.757 ± 0.053
GES prior pruned	6.950 ± 0.318	0.939 ± 0.005	0.691 ± 0.010	1.757 ± 0.053
GES pruned	13.743 ± 0.378	0.787 ± 0.009	0.417 ± 0.009	4.130 ± 0.125
Notears	10.034 ± 0.349	0.672 ± 0.012	0.550 ± 0.012	2.226 ± 0.080
Notears prior	8.073 ± 0.337	0.686 ± 0.013	0.579 ± 0.013	0.257 ± 0.010
Notears prior pruned	8.073 ± 0.337	0.686 ± 0.013	0.579 ± 0.013	0.257 ± 0.010
Notears pruned	9.642 ± 0.344	0.686 ± 0.012	0.550 ± 0.012	2.226 ± 0.080
PC	12.419 ± 0.341	0.635 ± 0.007	0.616 ± 0.009	0.336 ± 0.024
PC prior	8.968 ± 0.299	0.885 ± 0.005	0.666 ± 0.009	0.056 ± 0.001
PC prior pruned	8.968 ± 0.299	0.885 ± 0.005	0.666 ± 0.009	0.056 ± 0.001
PC pruned	9.439 ± 0.305	0.877 ± 0.006	0.616 ± 0.009	0.336 ± 0.024
Random	19.370 ± 0.283	0.479 ± 0.007	0.497 ± 0.005	0.001 ± 0.000

Table 11: Results for networks with 80% of nodes designated as sources. We show the average value and plus/minus one standard error for each test result, obtained from a sample of 1350 repeated experiments.

Method	SHD	Precision	Recall	Time (s)
Bipartite baseline	2.285 ± 0.084	0.971 ± 0.004	0.757 ± 0.007	1.154 ± 0.016
DirectLiNGAM	4.327 ± 0.152	0.661 ± 0.012	0.544 ± 0.012	0.717 ± 0.011
DirectLiNGAM prior	2.748 ± 0.120	0.955 ± 0.006	0.717 ± 0.010	0.390 ± 0.007
DirectLiNGAM prior pruned	2.748 ± 0.120	0.955 ± 0.006	0.717 ± 0.010	0.390 ± 0.007
DirectLiNGAM pruned	4.083 ± 0.146	0.734 ± 0.012	0.544 ± 0.012	0.717 ± 0.011
Empty	8.974 ± 0.157	0.000 ± 0.000	0.000 ± 0.000	0.001 ± 0.000
Full	36.915 ± 0.577	0.195 ± 0.002	0.993 ± 0.002	0.001 ± 0.000
GES	7.330 ± 0.183	0.524 ± 0.007	0.591 ± 0.008	2.474 ± 0.050
GES prior	1.918 ± 0.086	0.948 ± 0.004	0.843 ± 0.007	0.820 ± 0.019
GES prior pruned	1.918 ± 0.086	0.948 ± 0.004	0.843 ± 0.007	0.820 ± 0.019
GES pruned	4.070 ± 0.104	0.902 ± 0.006	0.591 ± 0.008	2.474 ± 0.050
Notears	3.513 ± 0.137	0.873 ± 0.009	0.641 ± 0.011	2.023 ± 0.062
Notears prior	3.660 ± 0.144	0.813 ± 0.011	0.607 ± 0.012	0.272 ± 0.010
Notears prior pruned	3.660 ± 0.144	0.813 ± 0.011	0.607 ± 0.012	0.272 ± 0.010
Notears pruned	3.364 ± 0.134	0.892 ± 0.008	0.641 ± 0.011	2.023 ± 0.062
PC	6.650 ± 0.158	0.562 ± 0.007	0.741 ± 0.008	0.147 ± 0.005
PC prior	2.821 ± 0.093	0.849 ± 0.005	0.841 ± 0.006	0.098 ± 0.004
PC prior pruned	2.821 ± 0.093	0.849 ± 0.005	0.841 ± 0.006	0.098 ± 0.004
PC pruned	3.346 ± 0.103	0.846 ± 0.006	0.741 ± 0.008	0.147 ± 0.005
Random	22.790 ± 0.365	0.197 ± 0.003	0.499 ± 0.006	0.001 ± 0.000

Table 12: Results for networks with a 20% density. We show the average value and plus/minus one standard error for each test result, obtained from a sample of 1350 repeated experiments.

Method	SHD	Precision	Recall	Time (s)
Bipartite baseline	8.533 ± 0.282	0.961 ± 0.005	0.644 ± 0.008	1.154 ± 0.017
DirectLiNGAM	10.873 ± 0.364	0.709 ± 0.011	0.551 ± 0.012	0.722 ± 0.011
DirectLiNGAM prior	8.585 ± 0.337	0.897 ± 0.008	0.637 ± 0.011	0.394 ± 0.007
DirectLiNGAM prior pruned	8.585 ± 0.337	0.897 ± 0.008	0.637 ± 0.011	0.394 ± 0.007
DirectLiNGAM pruned	10.329 ± 0.357	0.781 ± 0.011	0.551 ± 0.012	0.722 ± 0.011
Empty	22.541 ± 0.354	0.000 ± 0.000	0.000 ± 0.000	0.001 ± 0.000
Full	23.348 ± 0.375	0.494 ± 0.002	1.000 ± 0.000	0.001 ± 0.000
GES	25.682 ± 0.645	0.470 ± 0.007	0.498 ± 0.008	4.904 ± 0.128
GES prior	6.353 ± 0.271	0.979 ± 0.003	0.750 ± 0.009	1.423 ± 0.040
GES prior pruned	6.353 ± 0.271	0.979 ± 0.003	0.750 ± 0.009	1.423 ± 0.040
GES pruned	13.996 ± 0.335	0.881 ± 0.006	0.498 ± 0.008	4.904 ± 0.128
Notears	11.060 ± 0.347	0.806 ± 0.010	0.565 ± 0.011	4.905 ± 0.187
Notears prior	12.135 ± 0.387	0.672 ± 0.013	0.495 ± 0.012	0.254 ± 0.009
Notears prior pruned	12.135 ± 0.387	0.672 ± 0.013	0.495 ± 0.012	0.254 ± 0.009
Notears pruned	10.528 ± 0.344	0.835 ± 0.010	0.565 ± 0.011	4.905 ± 0.187
PC	15.516 ± 0.351	0.656 ± 0.006	0.617 ± 0.008	0.811 ± 0.078
PC prior	9.498 ± 0.260	0.944 ± 0.003	0.670 ± 0.007	0.579 ± 0.073
PC prior pruned	9.498 ± 0.260	0.944 ± 0.003	0.670 ± 0.007	0.579 ± 0.073
PC pruned	10.344 ± 0.271	0.939 ± 0.004	0.617 ± 0.008	0.811 ± 0.078
Random	23.018 ± 0.368	0.493 ± 0.003	0.501 ± 0.003	0.001 ± 0.000

Table 13: Results for networks with a 50% density. We show the average value and plus/minus one standard error for each test result, obtained from a sample of 1350 repeated experiments.

Method	SHD	Precision	Recall	Time (s)
Bipartite baseline	14.562 ± 0.474	0.903 ± 0.008	0.593 ± 0.009	1.168 ± 0.017
DirectLiNGAM	16.330 ± 0.546	0.675 ± 0.011	0.542 ± 0.012	0.723 ± 0.011
DirectLiNGAM prior	13.667 ± 0.527	0.826 ± 0.010	0.609 ± 0.012	0.394 ± 0.007
DirectLiNGAM prior pruned	13.667 ± 0.527	0.826 ± 0.010	0.609 ± 0.012	0.394 ± 0.007
DirectLiNGAM pruned	15.649 ± 0.540	0.738 ± 0.012	0.542 ± 0.012	0.723 ± 0.011
Empty	33.511 ± 0.530	0.000 ± 0.000	0.000 ± 0.000	0.001 ± 0.000
Full	12.378 ± 0.215	0.732 ± 0.002	1.000 ± 0.000	0.001 ± 0.000
GES	46.516 ± 0.984	0.365 ± 0.007	0.348 ± 0.007	6.233 ± 0.165
GES prior	10.910 ± 0.458	0.967 ± 0.004	0.703 ± 0.010	1.778 ± 0.055
GES prior pruned	10.910 ± 0.458	0.967 ± 0.004	0.703 ± 0.010	1.778 ± 0.055
GES pruned	25.866 ± 0.525	0.793 ± 0.008	0.348 ± 0.007	6.233 ± 0.165
Notears	23.295 ± 0.559	0.690 ± 0.011	0.442 ± 0.010	9.112 ± 0.442
Notears prior	21.608 ± 0.599	0.529 ± 0.014	0.402 ± 0.012	0.230 ± 0.008
Notears prior pruned	21.608 ± 0.599	0.529 ± 0.014	0.402 ± 0.012	0.230 ± 0.008
Notears pruned	21.951 ± 0.542	0.744 ± 0.011	0.442 ± 0.010	9.112 ± 0.442
PC	27.800 ± 0.570	0.628 ± 0.006	0.488 ± 0.007	1.232 ± 0.073
PC prior	19.061 ± 0.443	0.965 ± 0.002	0.521 ± 0.007	0.390 ± 0.020
PC prior pruned	19.061 ± 0.443	0.965 ± 0.002	0.521 ± 0.007	0.390 ± 0.020
PC pruned	19.924 ± 0.453	0.962 ± 0.003	0.488 ± 0.007	1.232 ± 0.073
Random	22.974 ± 0.365	0.732 ± 0.003	0.496 ± 0.003	0.001 ± 0.000

Table 14: Results for networks with a 80% density. We show the average value and plus/minus one standard error for each test result, obtained from a sample of 1350 repeated experiments.

Method	SHD	Precision	Recall	Time (s)
Bipartite baseline	8.962 ± 0.376	0.927 ± 0.007	0.656 ± 0.009	1.151 ± 0.016
DirectLiNGAM	11.004 ± 0.436	0.666 ± 0.012	0.541 ± 0.012	0.719 ± 0.011
DirectLiNGAM prior	8.824 ± 0.412	0.879 ± 0.009	0.645 ± 0.011	0.393 ± 0.007
DirectLiNGAM prior pruned	8.824 ± 0.412	0.879 ± 0.009	0.645 ± 0.011	0.393 ± 0.007
DirectLiNGAM pruned	10.541 ± 0.429	0.728 ± 0.012	0.541 ± 0.012	0.719 ± 0.011
Empty	22.663 ± 0.498	0.000 ± 0.000	0.000 ± 0.000	0.001 ± 0.000
Full	23.226 ± 0.513	0.496 ± 0.007	0.993 ± 0.002	0.001 ± 0.000
GES	27.601 ± 0.857	0.457 ± 0.008	0.469 ± 0.008	4.596 ± 0.134
GES prior	6.793 ± 0.349	0.961 ± 0.004	0.758 ± 0.009	1.365 ± 0.044
GES prior pruned	6.793 ± 0.349	0.961 ± 0.004	0.758 ± 0.009	1.365 ± 0.044
GES pruned	15.466 ± 0.464	0.852 ± 0.007	0.469 ± 0.008	4.596 ± 0.134
Notears	14.023 ± 0.490	0.765 ± 0.010	0.528 ± 0.011	6.112 ± 0.374
Notears prior	13.135 ± 0.497	0.666 ± 0.013	0.499 ± 0.012	0.239 ± 0.009
Notears prior pruned	13.135 ± 0.497	0.666 ± 0.013	0.499 ± 0.012	0.239 ± 0.009
Notears pruned	13.259 ± 0.470	0.802 ± 0.010	0.528 ± 0.011	6.112 ± 0.374
PC	17.667 ± 0.508	0.617 ± 0.007	0.604 ± 0.008	0.724 ± 0.062
PC prior	11.263 ± 0.381	0.916 ± 0.004	0.666 ± 0.008	0.217 ± 0.010
PC prior pruned	11.263 ± 0.381	0.916 ± 0.004	0.666 ± 0.008	0.217 ± 0.010
PC pruned	12.032 ± 0.391	0.913 ± 0.005	0.604 ± 0.008	0.724 ± 0.062
Random	22.943 ± 0.366	0.495 ± 0.007	0.493 ± 0.004	0.001 ± 0.000

Table 15: Results for networks with balanced vertex degrees. We show the average value and plus/minus one standard error for each test result, obtained from a sample of 1350 repeated experiments.

Method	SHD	Precision	Recall	Time (s)
Bipartite baseline	8.404 ± 0.335	0.959 ± 0.005	0.662 ± 0.008	1.156 ± 0.016
DirectLiNGAM	10.311 ± 0.403	0.687 ± 0.011	0.545 ± 0.012	0.722 ± 0.011
DirectLiNGAM prior	8.193 ± 0.378	0.900 ± 0.008	0.653 ± 0.011	0.392 ± 0.007
DirectLiNGAM prior pruned	8.193 ± 0.378	0.900 ± 0.008	0.653 ± 0.011	0.392 ± 0.007
DirectLiNGAM pruned	9.837 ± 0.395	0.756 ± 0.012	0.545 ± 0.012	0.722 ± 0.011
Empty	21.259 ± 0.458	0.000 ± 0.000	0.000 ± 0.000	0.001 ± 0.000
Full	24.630 ± 0.487	0.461 ± 0.006	1.000 ± 0.000	0.001 ± 0.000
GES	26.350 ± 0.827	0.454 ± 0.007	0.472 ± 0.008	4.521 ± 0.132
GES prior	6.305 ± 0.314	0.967 ± 0.003	0.760 ± 0.009	1.307 ± 0.040
GES prior pruned	6.305 ± 0.314	0.967 ± 0.003	0.760 ± 0.009	1.307 ± 0.040
GES pruned	14.573 ± 0.431	0.857 ± 0.007	0.472 ± 0.008	4.521 ± 0.132
Notears	12.167 ± 0.430	0.806 ± 0.010	0.557 ± 0.011	4.940 ± 0.256
Notears prior	11.959 ± 0.456	0.705 ± 0.012	0.520 ± 0.012	0.254 ± 0.009
Notears prior pruned	11.959 ± 0.456	0.705 ± 0.012	0.520 ± 0.012	0.254 ± 0.009
Notears pruned	11.524 ± 0.417	0.837 ± 0.010	0.557 ± 0.011	4.940 ± 0.256
PC	16.501 ± 0.452	0.614 ± 0.006	0.608 ± 0.008	0.537 ± 0.041
PC prior	10.456 ± 0.338	0.922 ± 0.004	0.668 ± 0.008	0.188 ± 0.008
PC prior pruned	10.456 ± 0.338	0.922 ± 0.004	0.668 ± 0.008	0.188 ± 0.008
PC pruned	11.144 ± 0.350	0.918 ± 0.005	0.608 ± 0.008	0.537 ± 0.041
Random	22.870 ± 0.366	0.465 ± 0.006	0.506 ± 0.004	0.001 ± 0.000

Table 16: Results for networks with sink dominant vertex degrees. We show the average value and plus/minus one standard error for each test result, obtained from a sample of 1350 repeated experiments.

Method	SHD	Precision	Recall	Time (s)
Bipartite baseline	8.015 ± 0.336	0.950 ± 0.006	0.676 ± 0.008	1.169 ± 0.017
DirectLiNGAM	10.213 ± 0.393	0.693 ± 0.011	0.551 ± 0.012	0.721 ± 0.011
DirectLiNGAM prior	7.983 ± 0.371	0.900 ± 0.008	0.665 ± 0.011	0.393 ± 0.007
DirectLiNGAM prior pruned	7.983 ± 0.371	0.900 ± 0.008	0.665 ± 0.011	0.393 ± 0.007
DirectLiNGAM pruned	9.682 ± 0.386	0.770 ± 0.011	0.551 ± 0.012	0.721 ± 0.011
Empty	21.104 ± 0.442	0.000 ± 0.000	0.000 ± 0.000	0.001 ± 0.000
Full	24.785 ± 0.491	0.462 ± 0.006	1.000 ± 0.000	0.001 ± 0.000
GES	25.577 ± 0.754	0.448 ± 0.007	0.495 ± 0.008	4.495 ± 0.126
GES prior	6.083 ± 0.316	0.966 ± 0.003	0.778 ± 0.008	1.350 ± 0.043
GES prior pruned	6.083 ± 0.316	0.966 ± 0.003	0.778 ± 0.008	1.350 ± 0.043
GES pruned	13.893 ± 0.415	0.868 ± 0.006	0.495 ± 0.008	4.495 ± 0.126
Notears	11.678 ± 0.415	0.798 ± 0.010	0.563 ± 0.011	4.988 ± 0.217
Notears prior	12.309 ± 0.440	0.643 ± 0.013	0.485 ± 0.012	0.263 ± 0.010
Notears prior pruned	12.309 ± 0.440	0.643 ± 0.013	0.485 ± 0.012	0.263 ± 0.010
Notears pruned	11.061 ± 0.404	0.832 ± 0.010	0.563 ± 0.011	4.988 ± 0.217
PC	15.796 ± 0.420	0.615 ± 0.006	0.633 ± 0.008	0.927 ± 0.079
PC prior	9.661 ± 0.332	0.920 ± 0.003	0.698 ± 0.007	0.661 ± 0.075
PC prior pruned	9.661 ± 0.332	0.920 ± 0.003	0.698 ± 0.007	0.661 ± 0.075
PC pruned	10.438 ± 0.340	0.916 ± 0.004	0.633 ± 0.008	0.927 ± 0.079
Random	22.969 ± 0.366	0.462 ± 0.006	0.496 ± 0.004	0.001 ± 0.000

Table 17: Results for networks with source dominant vertex degrees. We show the average value and plus/minus one standard error for each test result, obtained from a sample of 1350 repeated experiments.

Method	SHD	Precision	Recall	Time (s)
Bipartite baseline	2.673 ± 0.098	0.993 ± 0.003	0.874 ± 0.004	1.180 ± 0.022
DirectLiNGAM	0.000 ± 0.000	0.998 ± 0.002	0.998 ± 0.002	0.714 ± 0.014
DirectLiNGAM prior	0.000 ± 0.000	0.998 ± 0.002	0.998 ± 0.002	0.393 ± 0.009
DirectLiNGAM prior pruned	0.000 ± 0.000	0.998 ± 0.002	0.998 ± 0.002	0.393 ± 0.009
DirectLiNGAM pruned	0.000 ± 0.000	0.998 ± 0.002	0.998 ± 0.002	0.714 ± 0.014
Empty	21.675 ± 0.603	0.000 ± 0.000	0.000 ± 0.000	0.001 ± 0.000
Full	24.214 ± 0.643	0.473 ± 0.008	0.998 ± 0.002	0.001 ± 0.000
GES	32.721 ± 1.376	0.443 ± 0.011	0.605 ± 0.010	6.301 ± 0.210
GES prior	0.190 ± 0.015	0.983 ± 0.002	0.998 ± 0.002	1.856 ± 0.067
GES prior pruned	0.190 ± 0.015	0.983 ± 0.002	0.998 ± 0.002	1.856 ± 0.067
GES pruned	13.727 ± 0.604	0.813 ± 0.009	0.605 ± 0.010	6.301 ± 0.210
Notears	7.489 ± 0.540	0.940 ± 0.004	0.869 ± 0.007	6.233 ± 0.381
Notears prior	9.742 ± 0.673	0.784 ± 0.014	0.753 ± 0.015	0.280 ± 0.013
Notears prior pruned	9.742 ± 0.673	0.784 ± 0.014	0.753 ± 0.015	0.280 ± 0.013
Notears pruned	6.542 ± 0.478	0.968 ± 0.003	0.869 ± 0.007	6.233 ± 0.381
PC	12.888 ± 0.609	0.742 ± 0.007	0.805 ± 0.008	1.264 ± 0.115
PC prior	7.421 ± 0.390	0.947 ± 0.003	0.829 ± 0.007	0.568 ± 0.081
PC prior pruned	7.421 ± 0.390	0.947 ± 0.003	0.829 ± 0.007	0.568 ± 0.081
PC pruned	7.617 ± 0.395	0.952 ± 0.004	0.805 ± 0.008	1.264 ± 0.115
Random	22.933 ± 0.474	0.473 ± 0.009	0.492 ± 0.006	0.001 ± 0.000

Table 18: Results for datasets with the linear plus exponential noise DGP. We show the average value and plus/minus one standard error for each test result, obtained from a sample of 810 repeated experiments.

Method	SHD	Precision	Recall	Time (s)
Bipartite baseline	2.684 ± 0.098	0.991 ± 0.003	0.874 ± 0.004	1.181 ± 0.022
DirectLiNGAM	0.000 ± 0.000	0.998 ± 0.002	0.998 ± 0.002	0.728 ± 0.014
DirectLiNGAM prior	0.000 ± 0.000	0.998 ± 0.002	0.998 ± 0.002	0.393 ± 0.009
DirectLiNGAM prior pruned	0.000 ± 0.000	0.998 ± 0.002	0.998 ± 0.002	0.393 ± 0.009
DirectLiNGAM pruned	0.000 ± 0.000	0.998 ± 0.002	0.998 ± 0.002	0.728 ± 0.014
Empty	21.675 ± 0.603	0.000 ± 0.000	0.000 ± 0.000	0.001 ± 0.000
Full	24.214 ± 0.643	0.473 ± 0.008	0.998 ± 0.002	0.001 ± 0.000
GES	32.201 ± 1.341	0.430 ± 0.010	0.601 ± 0.010	6.332 ± 0.211
GES prior	0.194 ± 0.016	0.984 ± 0.002	0.998 ± 0.002	1.860 ± 0.068
GES prior pruned	0.194 ± 0.016	0.984 ± 0.002	0.998 ± 0.002	1.860 ± 0.068
GES pruned	13.409 ± 0.582	0.824 ± 0.008	0.601 ± 0.010	6.332 ± 0.211
Notears	4.812 ± 0.422	0.967 ± 0.003	0.924 ± 0.006	6.426 ± 0.366
Notears prior	2.475 ± 0.340	0.947 ± 0.008	0.937 ± 0.008	0.282 ± 0.013
Notears prior pruned	2.475 ± 0.340	0.947 ± 0.008	0.937 ± 0.008	0.282 ± 0.013
Notears pruned	4.368 ± 0.387	0.979 ± 0.003	0.924 ± 0.006	6.426 ± 0.366
PC	13.030 ± 0.605	0.734 ± 0.007	0.800 ± 0.008	1.318 ± 0.126
PC prior	7.427 ± 0.393	0.946 ± 0.003	0.830 ± 0.008	0.589 ± 0.091
PC prior pruned	7.427 ± 0.393	0.946 ± 0.003	0.830 ± 0.008	0.589 ± 0.091
PC pruned	7.716 ± 0.397	0.951 ± 0.004	0.800 ± 0.008	1.318 ± 0.126
Random	22.836 ± 0.470	0.478 ± 0.009	0.501 ± 0.006	0.001 ± 0.000

Table 19: Results for datasets with the linear plus Gumbel noise DGP. We show the average value and plus/minus one standard error for each test result, obtained from a sample of 810 repeated experiments.

Method	SHD	Precision	Recall	Time (s)
Bipartite baseline	17.265 ± 0.604	0.836 ± 0.013	0.321 ± 0.010	1.180 ± 0.022
DirectLiNGAM	21.010 ± 0.603	0.317 ± 0.014	0.062 ± 0.003	0.715 ± 0.014
DirectLiNGAM prior	19.372 ± 0.620	0.689 ± 0.016	0.205 ± 0.009	0.394 ± 0.009
DirectLiNGAM prior pruned	19.372 ± 0.620	0.689 ± 0.016	0.205 ± 0.009	0.394 ± 0.009
DirectLiNGAM pruned	20.773 ± 0.604	0.435 ± 0.017	0.062 ± 0.003	0.715 ± 0.014
Empty	21.675 ± 0.603	0.000 ± 0.000	0.000 ± 0.000	0.001 ± 0.000
Full	24.214 ± 0.643	0.473 ± 0.008	0.998 ± 0.002	0.001 ± 0.000
GES	21.960 ± 0.598	0.447 ± 0.009	0.242 ± 0.007	2.119 ± 0.052
GES prior	16.660 ± 0.560	0.921 ± 0.008	0.354 ± 0.009	0.530 ± 0.012
GES prior pruned	16.660 ± 0.560	0.921 ± 0.008	0.354 ± 0.009	0.530 ± 0.012
GES pruned	18.172 ± 0.571	0.870 ± 0.011	0.242 ± 0.007	2.119 ± 0.052
Notears	20.577 ± 0.611	0.519 ± 0.017	0.108 ± 0.006	1.296 ± 0.051
Notears prior	21.264 ± 0.607	0.296 ± 0.016	0.047 ± 0.004	0.235 ± 0.011
Notears prior pruned	21.264 ± 0.607	0.296 ± 0.016	0.047 ± 0.004	0.235 ± 0.011
Notears pruned	20.460 ± 0.610	0.553 ± 0.017	0.108 ± 0.006	1.296 ± 0.051
PC	23.223 ± 0.595	0.404 ± 0.008	0.289 ± 0.007	0.188 ± 0.015
PC prior	16.453 ± 0.526	0.853 ± 0.007	0.387 ± 0.009	0.107 ± 0.010
PC prior pruned	16.453 ± 0.526	0.853 ± 0.007	0.387 ± 0.009	0.107 ± 0.010
PC pruned	17.607 ± 0.537	0.829 ± 0.009	0.289 ± 0.007	0.188 ± 0.015
Random	22.935 ± 0.469	0.473 ± 0.009	0.502 ± 0.006	0.001 ± 0.000

Table 20: Results for datasets with the quadratic plus Gaussian noise DGP. We show the average value and plus/minus one standard error for each test result, obtained from a sample of 810 repeated experiments.

Method	SHD	Precision	Recall	Time (s)
Bipartite baseline	14.109 ± 0.586	0.915 ± 0.010	0.478 ± 0.011	1.124 ± 0.021
DirectLiNGAM	21.151 ± 0.603	0.268 ± 0.014	0.043 ± 0.003	0.720 ± 0.014
DirectLiNGAM prior	18.275 ± 0.604	0.788 ± 0.014	0.244 ± 0.008	0.389 ± 0.009
DirectLiNGAM prior pruned	18.275 ± 0.604	0.788 ± 0.014	0.244 ± 0.008	0.389 ± 0.009
DirectLiNGAM pruned	21.041 ± 0.604	0.363 ± 0.017	0.043 ± 0.003	0.720 ± 0.014
Empty	21.675 ± 0.603	0.000 ± 0.000	0.000 ± 0.000	0.001 ± 0.000
Full	24.214 ± 0.643	0.473 ± 0.008	0.998 ± 0.002	0.001 ± 0.000
GES	19.500 ± 0.588	0.518 ± 0.008	0.399 ± 0.009	2.700 ± 0.071
GES prior	12.077 ± 0.517	0.954 ± 0.005	0.575 ± 0.010	0.801 ± 0.020
GES prior pruned	12.077 ± 0.517	0.954 ± 0.005	0.575 ± 0.010	0.801 ± 0.020
GES pruned	15.079 ± 0.539	0.923 ± 0.008	0.399 ± 0.009	2.700 ± 0.071
Notears	19.915 ± 0.610	0.645 ± 0.017	0.139 ± 0.006	1.684 ± 0.060
Notears prior	21.077 ± 0.614	0.384 ± 0.017	0.072 ± 0.005	0.218 ± 0.011
Notears prior pruned	21.077 ± 0.614	0.384 ± 0.017	0.072 ± 0.005	0.218 ± 0.011
Notears pruned	19.906 ± 0.610	0.647 ± 0.017	0.139 ± 0.006	1.684 ± 0.060
PC	19.441 ± 0.553	0.525 ± 0.007	0.494 ± 0.008	0.376 ± 0.032
PC prior	12.146 ± 0.469	0.908 ± 0.005	0.598 ± 0.008	0.275 ± 0.028
PC prior pruned	12.146 ± 0.469	0.908 ± 0.005	0.598 ± 0.008	0.275 ± 0.028
PC pruned	13.526 ± 0.483	0.903 ± 0.006	0.494 ± 0.008	0.376 ± 0.032
Random	22.848 ± 0.472	0.478 ± 0.009	0.497 ± 0.006	0.001 ± 0.000

Table 21: Results for datasets with the Gaussian Process DGP. We show the average value and plus/minus one standard error for each test result, obtained from a sample of 810 repeated experiments.

Method	SHD	Precision	Recall	Time (s)
Bipartite baseline	5.570 ± 0.253	0.992 ± 0.003	0.777 ± 0.006	1.127 ± 0.021
DirectLiNGAM	10.388 ± 0.365	0.828 ± 0.007	0.628 ± 0.008	0.726 ± 0.014
DirectLiNGAM prior	4.020 ± 0.191	0.993 ± 0.003	0.830 ± 0.007	0.395 ± 0.009
DirectLiNGAM prior pruned	4.020 ± 0.191	0.993 ± 0.003	0.830 ± 0.007	0.395 ± 0.009
DirectLiNGAM pruned	8.288 ± 0.294	0.964 ± 0.005	0.628 ± 0.008	0.726 ± 0.014
Empty	21.675 ± 0.603	0.000 ± 0.000	0.000 ± 0.000	0.001 ± 0.000
Full	24.214 ± 0.643	0.473 ± 0.008	0.998 ± 0.002	0.001 ± 0.000
GES	26.164 ± 0.978	0.426 ± 0.008	0.547 ± 0.009	5.233 ± 0.164
GES prior	2.847 ± 0.167	0.981 ± 0.002	0.902 ± 0.005	1.655 ± 0.058
GES prior pruned	2.847 ± 0.167	0.981 ± 0.002	0.902 ± 0.005	1.655 ± 0.058
GES pruned	12.833 ± 0.504	0.865 ± 0.007	0.547 ± 0.009	5.233 ± 0.164
Notears	10.320 ± 0.438	0.875 ± 0.005	0.705 ± 0.008	11.094 ± 0.580
Notears prior	7.780 ± 0.378	0.945 ± 0.008	0.698 ± 0.010	0.246 ± 0.012
Notears prior pruned	7.780 ± 0.378	0.945 ± 0.008	0.698 ± 0.010	0.246 ± 0.012
Notears pruned	8.463 ± 0.378	0.970 ± 0.003	0.705 ± 0.008	11.094 ± 0.580
PC	14.694 ± 0.533	0.670 ± 0.007	0.688 ± 0.007	0.502 ± 0.036
PC prior	8.853 ± 0.391	0.942 ± 0.004	0.743 ± 0.007	0.240 ± 0.017
PC prior pruned	8.853 ± 0.391	0.942 ± 0.004	0.743 ± 0.007	0.240 ± 0.017
PC pruned	9.556 ± 0.401	0.944 ± 0.004	0.688 ± 0.007	0.502 ± 0.036
Random	23.084 ± 0.479	0.470 ± 0.009	0.500 ± 0.006	0.001 ± 0.000

Table 22: Results for datasets with the combination DGP. We show the average value and plus/minus one standard error for each test result, obtained from a sample of 810 repeated experiments.

F.2 Block-Directed r-Partite DAGs

Method	SHD	Precision	Recall	Time (s)
DirectLiNGAM	10.428 ± 0.877	0.635 ± 0.016	0.537 ± 0.018	0.724 ± 0.023
DirectLiNGAM prior	8.925 ± 0.840	0.941 ± 0.008	0.734 ± 0.014	0.270 ± 0.007
DirectLiNGAM prior pruned	8.925 ± 0.840	0.941 ± 0.008	0.734 ± 0.014	0.270 ± 0.007
DirectLiNGAM pruned	10.204 ± 0.870	0.729 ± 0.016	0.537 ± 0.018	0.724 ± 0.023
Empty	21.167 ± 1.113	0.000 ± 0.000	0.000 ± 0.000	0.001 ± 0.000
Full	99.389 ± 2.707	0.188 ± 0.006	1.000 ± 0.000	0.001 ± 0.000
GES	26.346 ± 1.573	0.446 ± 0.011	0.490 ± 0.012	4.853 ± 0.299
GES prior	7.601 ± 0.639	0.890 ± 0.006	0.836 ± 0.010	1.951 ± 0.115
GES prior pruned	7.601 ± 0.639	0.890 ± 0.006	0.836 ± 0.010	1.951 ± 0.115
GES pruned	17.292 ± 1.039	0.770 ± 0.011	0.490 ± 0.012	4.853 ± 0.299
Notears	13.893 ± 1.023	0.817 ± 0.012	0.589 ± 0.015	13.212 ± 1.543
Notears prior	15.839 ± 1.032	0.717 ± 0.015	0.509 ± 0.016	0.234 ± 0.015
Notears prior pruned	15.839 ± 1.032	0.717 ± 0.015	0.509 ± 0.016	0.234 ± 0.015
Notears pruned	13.011 ± 0.974	0.856 ± 0.012	0.589 ± 0.015	13.212 ± 1.543
PC	18.972 ± 1.065	0.549 ± 0.010	0.616 ± 0.012	0.393 ± 0.025
PC prior	13.800 ± 0.859	0.778 ± 0.008	0.762 ± 0.010	0.271 ± 0.017
PC prior pruned	13.800 ± 0.859	0.778 ± 0.008	0.762 ± 0.010	0.271 ± 0.017
PC pruned	15.144 ± 0.942	0.779 ± 0.009	0.616 ± 0.012	0.393 ± 0.025
r-partite baseline	10.311 ± 0.673	0.805 ± 0.007	0.844 ± 0.009	0.103 ± 0.002
Random	37.610 ± 1.191	0.292 ± 0.009	0.501 ± 0.009	0.001 ± 0.000

Table 23: Total results for all configurations. We show the average value and plus/minus one standard error for each test result, obtained from a sample of 720 repeated experiments.

Method	SHD	Precision	Recall	Time (s)
DirectLiNGAM	3.271 ± 0.483	0.602 ± 0.031	0.536 ± 0.031	0.198 ± 0.007
DirectLiNGAM prior	2.525 ± 0.482	0.917 ± 0.018	0.797 ± 0.023	0.097 ± 0.004
DirectLiNGAM prior pruned	2.525 ± 0.482	0.917 ± 0.018	0.797 ± 0.023	0.097 ± 0.004
DirectLiNGAM pruned	3.225 ± 0.483	0.637 ± 0.031	0.536 ± 0.031	0.198 ± 0.007
Empty	6.808 ± 0.615	0.000 ± 0.000	0.000 ± 0.000	0.001 ± 0.000
Full	30.358 ± 0.858	0.187 ± 0.008	1.000 ± 0.000	0.001 ± 0.000
GES	7.279 ± 0.971	0.535 ± 0.025	0.546 ± 0.026	0.810 ± 0.074
GES prior	2.183 ± 0.423	0.954 ± 0.011	0.865 ± 0.019	0.308 ± 0.026
GES prior pruned	2.183 ± 0.423	0.954 ± 0.011	0.865 ± 0.019	0.308 ± 0.026
GES pruned	4.833 ± 0.628	0.773 ± 0.026	0.546 ± 0.026	0.810 ± 0.074
Notears	4.263 ± 0.590	0.764 ± 0.027	0.610 ± 0.028	0.894 ± 0.132
Notears prior	3.829 ± 0.555	0.692 ± 0.030	0.591 ± 0.030	0.056 ± 0.003
Notears prior pruned	3.829 ± 0.555	0.692 ± 0.030	0.591 ± 0.030	0.056 ± 0.003
Notears pruned	4.037 ± 0.556	0.775 ± 0.027	0.610 ± 0.028	0.894 ± 0.132
PC	4.829 ± 0.553	0.664 ± 0.020	0.729 ± 0.023	0.088 ± 0.014
PC prior	3.225 ± 0.459	0.888 ± 0.013	0.846 ± 0.019	0.039 ± 0.003
PC prior pruned	3.225 ± 0.459	0.888 ± 0.013	0.846 ± 0.019	0.039 ± 0.003
PC pruned	3.625 ± 0.474	0.846 ± 0.018	0.729 ± 0.023	0.088 ± 0.014
r-partite baseline	2.200 ± 0.397	0.925 ± 0.012	0.872 ± 0.018	0.069 ± 0.001
Random	7.537 ± 0.325	0.351 ± 0.017	0.522 ± 0.021	0.001 ± 0.000

Table 24: Results for $r = 2$ partitions. We show the average value and plus/minus one standard error for each test result, obtained from a sample of 240 repeated experiments.

Method	SHD	Precision	Recall	Time (s)
DirectLiNGAM	6.062 ± 0.621	0.650 ± 0.028	0.549 ± 0.030	0.512 ± 0.016
DirectLiNGAM prior	4.763 ± 0.556	0.974 ± 0.009	0.703 ± 0.023	0.241 ± 0.006
DirectLiNGAM prior pruned	4.763 ± 0.556	0.974 ± 0.009	0.703 ± 0.023	0.241 ± 0.006
DirectLiNGAM pruned	5.896 ± 0.611	0.744 ± 0.028	0.549 ± 0.030	0.512 ± 0.016
Empty	12.517 ± 0.688	0.000 ± 0.000	0.000 ± 0.000	0.001 ± 0.000
Full	81.900 ± 2.575	0.182 ± 0.010	1.000 ± 0.000	0.001 ± 0.000
GES	13.154 ± 0.951	0.443 ± 0.016	0.494 ± 0.018	2.218 ± 0.096
GES prior	4.125 ± 0.405	0.883 ± 0.010	0.816 ± 0.017	0.946 ± 0.041
GES prior pruned	4.125 ± 0.405	0.883 ± 0.010	0.816 ± 0.017	0.946 ± 0.041
GES pruned	8.596 ± 0.596	0.809 ± 0.016	0.494 ± 0.018	2.218 ± 0.096
Notears	6.358 ± 0.598	0.878 ± 0.019	0.602 ± 0.026	2.843 ± 0.186
Notears prior	7.392 ± 0.614	0.767 ± 0.025	0.527 ± 0.027	0.238 ± 0.021
Notears prior pruned	7.392 ± 0.614	0.767 ± 0.025	0.527 ± 0.027	0.238 ± 0.021
Notears pruned	6.067 ± 0.586	0.912 ± 0.018	0.602 ± 0.026	2.843 ± 0.186
PC	10.125 ± 0.561	0.523 ± 0.016	0.629 ± 0.020	0.168 ± 0.009
PC prior	6.771 ± 0.398	0.740 ± 0.013	0.765 ± 0.016	0.118 ± 0.006
PC prior pruned	6.771 ± 0.398	0.740 ± 0.013	0.765 ± 0.016	0.118 ± 0.006
PC pruned	7.500 ± 0.480	0.748 ± 0.016	0.629 ± 0.020	0.168 ± 0.009
r-partite baseline	5.213 ± 0.372	0.775 ± 0.012	0.824 ± 0.016	0.090 ± 0.002
Random	27.892 ± 0.797	0.258 ± 0.015	0.480 ± 0.013	0.001 ± 0.000

Table 25: Results for $r = 3$ partitions. We show the average value and plus/minus one standard error for each test result, obtained from a sample of 240 repeated experiments.

Method	SHD	Precision	Recall	Time (s)
DirectLiNGAM	21.950 ± 2.339	0.653 ± 0.026	0.527 ± 0.031	1.461 ± 0.026
DirectLiNGAM prior	19.488 ± 2.261	0.932 ± 0.011	0.701 ± 0.023	0.471 ± 0.009
DirectLiNGAM prior pruned	19.488 ± 2.261	0.932 ± 0.011	0.701 ± 0.023	0.471 ± 0.009
DirectLiNGAM pruned	21.492 ± 2.327	0.805 ± 0.025	0.527 ± 0.031	1.461 ± 0.026
Empty	44.175 ± 2.635	0.000 ± 0.000	0.000 ± 0.000	0.001 ± 0.000
Full	185.908 ± 2.496	0.194 ± 0.010	1.000 ± 0.000	0.001 ± 0.000
GES	58.604 ± 3.724	0.360 ± 0.013	0.431 ± 0.015	11.531 ± 0.714
GES prior	16.496 ± 1.686	0.833 ± 0.011	0.828 ± 0.015	4.598 ± 0.269
GES prior pruned	16.496 ± 1.686	0.833 ± 0.011	0.828 ± 0.015	4.598 ± 0.269
GES pruned	38.446 ± 2.480	0.728 ± 0.013	0.431 ± 0.015	11.531 ± 0.714
Notears	31.058 ± 2.623	0.808 ± 0.018	0.555 ± 0.025	35.898 ± 4.266
Notears prior	36.296 ± 2.504	0.693 ± 0.023	0.409 ± 0.023	0.409 ± 0.037
Notears prior pruned	36.296 ± 2.504	0.693 ± 0.023	0.409 ± 0.023	0.409 ± 0.037
Notears pruned	28.929 ± 2.511	0.880 ± 0.016	0.555 ± 0.025	35.898 ± 4.266
PC	41.962 ± 2.496	0.460 ± 0.012	0.492 ± 0.017	0.923 ± 0.059
PC prior	31.404 ± 2.077	0.706 ± 0.010	0.677 ± 0.017	0.656 ± 0.039
PC prior pruned	31.404 ± 2.077	0.706 ± 0.010	0.677 ± 0.017	0.656 ± 0.039
PC pruned	34.308 ± 2.284	0.741 ± 0.011	0.492 ± 0.017	0.923 ± 0.059
r-partite baseline	23.521 ± 1.636	0.714 ± 0.011	0.836 ± 0.014	0.150 ± 0.003
Random	77.400 ± 1.119	0.266 ± 0.014	0.502 ± 0.007	0.001 ± 0.000

Table 26: Results for $r = 5$ partitions. We show the average value and plus/minus one standard error for each test result, obtained from a sample of 240 repeated experiments.

Method	SHD	Precision	Recall	Time (s)
DirectLiNGAM	5.737 ± 0.741	0.582 ± 0.030	0.528 ± 0.031	0.688 ± 0.035
DirectLiNGAM prior	4.358 ± 0.641	0.913 ± 0.015	0.785 ± 0.020	0.265 ± 0.011
DirectLiNGAM prior pruned	4.358 ± 0.641	0.913 ± 0.015	0.785 ± 0.020	0.265 ± 0.011
DirectLiNGAM pruned	5.325 ± 0.694	0.682 ± 0.030	0.528 ± 0.031	0.688 ± 0.035
Empty	11.325 ± 0.896	0.000 ± 0.000	0.000 ± 0.000	0.001 ± 0.000
Full	104.258 ± 4.806	0.127 ± 0.006	1.000 ± 0.000	0.001 ± 0.000
GES	15.000 ± 1.575	0.427 ± 0.020	0.503 ± 0.020	3.303 ± 0.286
GES prior	5.067 ± 0.671	0.842 ± 0.014	0.863 ± 0.016	1.454 ± 0.130
GES prior pruned	5.067 ± 0.671	0.842 ± 0.014	0.863 ± 0.016	1.454 ± 0.130
GES pruned	8.521 ± 0.834	0.748 ± 0.020	0.503 ± 0.020	3.303 ± 0.286
Notears	6.987 ± 0.903	0.786 ± 0.023	0.641 ± 0.025	12.375 ± 2.484
Notears prior	7.542 ± 0.811	0.718 ± 0.026	0.571 ± 0.027	0.194 ± 0.017
Notears prior pruned	7.542 ± 0.811	0.718 ± 0.026	0.571 ± 0.027	0.194 ± 0.017
Notears pruned	5.933 ± 0.747	0.821 ± 0.023	0.641 ± 0.025	12.375 ± 2.484
PC	11.387 ± 1.059	0.539 ± 0.019	0.677 ± 0.021	0.313 ± 0.037
PC prior	7.592 ± 0.738	0.719 ± 0.015	0.824 ± 0.016	0.218 ± 0.024
PC prior pruned	7.592 ± 0.738	0.719 ± 0.015	0.824 ± 0.016	0.218 ± 0.024
PC pruned	7.875 ± 0.782	0.733 ± 0.017	0.677 ± 0.021	0.313 ± 0.037
r-partite baseline	6.729 ± 0.808	0.759 ± 0.015	0.874 ± 0.015	0.102 ± 0.003
Random	37.612 ± 2.039	0.197 ± 0.013	0.485 ± 0.017	0.001 ± 0.000

Table 27: Results for a maximum lag of 1. We show the average value and plus/minus one standard error for each test result, obtained from a sample of 240 repeated experiments.

Method	SHD	Precision	Recall	Time (s)
DirectLiNGAM	9.417 ± 1.298	0.667 ± 0.027	0.551 ± 0.030	0.603 ± 0.029
DirectLiNGAM prior	7.767 ± 1.218	0.957 ± 0.011	0.745 ± 0.023	0.230 ± 0.009
DirectLiNGAM prior pruned	7.767 ± 1.218	0.957 ± 0.011	0.745 ± 0.023	0.230 ± 0.009
DirectLiNGAM pruned	9.267 ± 1.284	0.769 ± 0.027	0.551 ± 0.030	0.603 ± 0.029
Empty	19.383 ± 1.614	0.000 ± 0.000	0.000 ± 0.000	0.001 ± 0.000
Full	84.950 ± 3.757	0.190 ± 0.009	1.000 ± 0.000	0.001 ± 0.000
GES	22.646 ± 2.264	0.475 ± 0.019	0.533 ± 0.021	3.861 ± 0.332
GES prior	6.071 ± 0.820	0.898 ± 0.011	0.860 ± 0.016	1.558 ± 0.128
GES prior pruned	6.071 ± 0.820	0.898 ± 0.011	0.860 ± 0.016	1.558 ± 0.128
GES pruned	15.054 ± 1.522	0.796 ± 0.018	0.533 ± 0.021	3.861 ± 0.332
Notears	12.379 ± 1.513	0.840 ± 0.020	0.599 ± 0.026	11.809 ± 2.419
Notears prior	13.521 ± 1.408	0.744 ± 0.025	0.526 ± 0.028	0.264 ± 0.036
Notears prior pruned	13.521 ± 1.408	0.744 ± 0.025	0.526 ± 0.028	0.264 ± 0.036
Notears pruned	11.700 ± 1.431	0.883 ± 0.020	0.599 ± 0.026	11.809 ± 2.419
PC	17.133 ± 1.604	0.549 ± 0.017	0.635 ± 0.021	0.372 ± 0.045
PC prior	11.896 ± 1.237	0.799 ± 0.012	0.789 ± 0.017	0.259 ± 0.030
PC prior pruned	11.896 ± 1.237	0.799 ± 0.012	0.789 ± 0.017	0.259 ± 0.030
PC pruned	13.550 ± 1.388	0.791 ± 0.015	0.635 ± 0.021	0.372 ± 0.045
r-partite baseline	7.633 ± 0.774	0.819 ± 0.012	0.868 ± 0.014	0.100 ± 0.003
Random	31.783 ± 1.666	0.299 ± 0.014	0.516 ± 0.015	0.001 ± 0.000

Table 28: Results for a maximum lag of 2. We show the average value and plus/minus one standard error for each test result, obtained from a sample of 240 repeated experiments.

Method	SHD	Precision	Recall	Time (s)
DirectLiNGAM	16.129 ± 2.115	0.657 ± 0.028	0.533 ± 0.031	0.881 ± 0.049
DirectLiNGAM prior	14.650 ± 2.059	0.953 ± 0.013	0.672 ± 0.026	0.313 ± 0.014
DirectLiNGAM prior pruned	14.650 ± 2.059	0.953 ± 0.013	0.672 ± 0.026	0.313 ± 0.014
DirectLiNGAM pruned	16.021 ± 2.113	0.737 ± 0.028	0.533 ± 0.031	0.881 ± 0.049
Empty	32.792 ± 2.606	0.000 ± 0.000	0.000 ± 0.000	0.001 ± 0.000
Full	108.958 ± 5.250	0.246 ± 0.011	1.000 ± 0.000	0.001 ± 0.000
GES	41.392 ± 3.629	0.436 ± 0.018	0.435 ± 0.019	7.394 ± 0.758
GES prior	11.667 ± 1.569	0.929 ± 0.007	0.786 ± 0.019	2.840 ± 0.285
GES prior pruned	11.667 ± 1.569	0.929 ± 0.007	0.786 ± 0.019	2.840 ± 0.285
GES pruned	28.300 ± 2.424	0.767 ± 0.019	0.435 ± 0.019	7.394 ± 0.758
Notears	22.312 ± 2.415	0.824 ± 0.021	0.525 ± 0.027	15.451 ± 3.072
Notears prior	26.454 ± 2.487	0.690 ± 0.027	0.430 ± 0.027	0.245 ± 0.022
Notears prior pruned	26.454 ± 2.487	0.690 ± 0.027	0.430 ± 0.027	0.245 ± 0.022
Notears pruned	21.400 ± 2.333	0.863 ± 0.021	0.525 ± 0.027	15.451 ± 3.072
PC	28.396 ± 2.430	0.559 ± 0.016	0.537 ± 0.021	0.494 ± 0.045
PC prior	21.913 ± 2.032	0.815 ± 0.011	0.674 ± 0.019	0.336 ± 0.032
PC prior pruned	21.913 ± 2.032	0.815 ± 0.011	0.674 ± 0.019	0.336 ± 0.032
PC pruned	24.008 ± 2.215	0.812 ± 0.015	0.537 ± 0.021	0.494 ± 0.045
r-partite baseline	16.571 ± 1.608	0.837 ± 0.010	0.789 ± 0.018	0.106 ± 0.003
Random	43.433 ± 2.362	0.380 ± 0.018	0.504 ± 0.013	0.001 ± 0.000

Table 29: Results for a maximum lag of 3. We show the average value and plus/minus one standard error for each test result, obtained from a sample of 240 repeated experiments.

Method	SHD	Precision	Recall	Time (s)
DirectLiNGAM	2.144 ± 0.241	0.595 ± 0.035	0.546 ± 0.035	0.724 ± 0.037
DirectLiNGAM prior	0.761 ± 0.115	0.978 ± 0.011	0.858 ± 0.019	0.268 ± 0.013
DirectLiNGAM prior pruned	0.761 ± 0.115	0.978 ± 0.011	0.858 ± 0.019	0.268 ± 0.013
DirectLiNGAM pruned	2.033 ± 0.226	0.644 ± 0.036	0.546 ± 0.035	0.724 ± 0.037
Empty	4.400 ± 0.245	0.000 ± 0.000	0.000 ± 0.000	0.001 ± 0.000
Full	116.489 ± 5.629	0.050 ± 0.002	1.000 ± 0.000	0.001 ± 0.000
GES	3.472 ± 0.254	0.511 ± 0.023	0.627 ± 0.025	1.798 ± 0.100
GES prior	1.089 ± 0.120	0.897 ± 0.011	0.959 ± 0.009	0.747 ± 0.044
GES prior pruned	1.089 ± 0.120	0.897 ± 0.011	0.959 ± 0.009	0.747 ± 0.044
GES pruned	2.272 ± 0.180	0.770 ± 0.026	0.627 ± 0.025	1.798 ± 0.100
Notears	1.444 ± 0.173	0.865 ± 0.025	0.696 ± 0.028	1.187 ± 0.072
Notears prior	1.789 ± 0.205	0.796 ± 0.030	0.635 ± 0.033	0.215 ± 0.027
Notears prior pruned	1.789 ± 0.205	0.796 ± 0.030	0.635 ± 0.033	0.215 ± 0.027
Notears pruned	1.389 ± 0.162	0.878 ± 0.024	0.696 ± 0.028	1.187 ± 0.072
PC	5.300 ± 0.332	0.460 ± 0.022	0.741 ± 0.024	0.090 ± 0.005
PC prior	3.022 ± 0.228	0.691 ± 0.018	0.959 ± 0.009	0.067 ± 0.004
PC prior pruned	3.022 ± 0.228	0.691 ± 0.018	0.959 ± 0.009	0.067 ± 0.004
PC pruned	3.072 ± 0.219	0.645 ± 0.024	0.741 ± 0.024	0.090 ± 0.005
r-partite baseline	2.250 ± 0.184	0.763 ± 0.015	0.961 ± 0.008	0.103 ± 0.003
Random	37.878 ± 2.088	0.085 ± 0.006	0.536 ± 0.024	0.001 ± 0.000

Table 30: Results for networks with a 10% density. We show the average value and plus/minus one standard error for each test result, obtained from a sample of 180 repeated experiments.

Method	SHD	Precision	Recall	Time (s)
DirectLiNGAM	6.900 ± 1.076	0.662 ± 0.031	0.545 ± 0.035	0.693 ± 0.045
DirectLiNGAM prior	5.283 ± 0.966	0.927 ± 0.017	0.748 ± 0.026	0.265 ± 0.014
DirectLiNGAM prior pruned	5.283 ± 0.966	0.927 ± 0.017	0.748 ± 0.026	0.265 ± 0.014
DirectLiNGAM pruned	6.772 ± 1.070	0.760 ± 0.032	0.545 ± 0.035	0.693 ± 0.045
Empty	14.467 ± 1.364	0.000 ± 0.000	0.000 ± 0.000	0.001 ± 0.000
Full	103.311 ± 5.585	0.136 ± 0.005	1.000 ± 0.000	0.001 ± 0.000
GES	17.789 ± 2.520	0.490 ± 0.022	0.552 ± 0.022	4.269 ± 0.574
GES prior	4.978 ± 0.735	0.867 ± 0.015	0.853 ± 0.019	1.630 ± 0.177
GES prior pruned	4.978 ± 0.735	0.867 ± 0.015	0.853 ± 0.019	1.630 ± 0.177
GES pruned	11.394 ± 1.539	0.793 ± 0.021	0.552 ± 0.022	4.269 ± 0.574
Notears	6.978 ± 1.075	0.845 ± 0.025	0.633 ± 0.030	4.642 ± 0.595
Notears prior	10.083 ± 1.253	0.718 ± 0.030	0.538 ± 0.032	0.236 ± 0.024
Notears prior pruned	10.083 ± 1.253	0.718 ± 0.030	0.538 ± 0.032	0.236 ± 0.024
Notears pruned	6.733 ± 1.057	0.860 ± 0.024	0.633 ± 0.030	4.642 ± 0.595
PC	11.711 ± 1.203	0.587 ± 0.021	0.675 ± 0.023	0.238 ± 0.026
PC prior	7.939 ± 0.913	0.772 ± 0.015	0.816 ± 0.019	0.187 ± 0.021
PC prior pruned	7.939 ± 0.913	0.772 ± 0.015	0.816 ± 0.019	0.187 ± 0.021
PC pruned	8.922 ± 1.030	0.783 ± 0.017	0.675 ± 0.023	0.238 ± 0.026
r-partite baseline	7.361 ± 0.816	0.775 ± 0.016	0.857 ± 0.019	0.104 ± 0.003
Random	36.994 ± 2.477	0.209 ± 0.012	0.475 ± 0.018	0.001 ± 0.000

Table 31: Results for networks with a 30% density. We show the average value and plus/minus one standard error for each test result, obtained from a sample of 180 repeated experiments.

Method	SHD	Precision	Recall	Time (s)
DirectLiNGAM	9.889 ± 1.310	0.631 ± 0.032	0.536 ± 0.035	0.720 ± 0.050
DirectLiNGAM prior	8.633 ± 1.224	0.946 ± 0.012	0.717 ± 0.027	0.268 ± 0.014
DirectLiNGAM prior pruned	8.633 ± 1.224	0.946 ± 0.012	0.717 ± 0.027	0.268 ± 0.014
DirectLiNGAM pruned	9.633 ± 1.295	0.759 ± 0.032	0.536 ± 0.035	0.720 ± 0.050
Empty	20.144 ± 1.542	0.000 ± 0.000	0.000 ± 0.000	0.001 ± 0.000
Full	97.189 ± 5.857	0.185 ± 0.006	1.000 ± 0.000	0.001 ± 0.000
GES	26.739 ± 2.760	0.425 ± 0.021	0.472 ± 0.021	5.357 ± 0.602
GES prior	7.783 ± 0.992	0.872 ± 0.014	0.809 ± 0.020	1.928 ± 0.180
GES prior pruned	7.783 ± 0.992	0.872 ± 0.014	0.809 ± 0.020	1.928 ± 0.180
GES pruned	16.217 ± 1.505	0.753 ± 0.021	0.472 ± 0.021	5.357 ± 0.602
Notears	10.461 ± 1.341	0.853 ± 0.023	0.607 ± 0.031	13.261 ± 2.703
Notears prior	14.828 ± 1.460	0.745 ± 0.028	0.477 ± 0.030	0.270 ± 0.045
Notears prior pruned	14.828 ± 1.460	0.745 ± 0.028	0.477 ± 0.030	0.270 ± 0.045
Notears pruned	9.889 ± 1.281	0.884 ± 0.022	0.607 ± 0.031	13.261 ± 2.703
PC	17.183 ± 1.529	0.593 ± 0.018	0.621 ± 0.022	0.389 ± 0.039
PC prior	12.472 ± 1.174	0.787 ± 0.014	0.737 ± 0.019	0.300 ± 0.032
PC prior pruned	12.472 ± 1.174	0.787 ± 0.014	0.737 ± 0.019	0.300 ± 0.032
PC pruned	13.228 ± 1.259	0.818 ± 0.013	0.621 ± 0.022	0.389 ± 0.039
r-partite baseline	11.572 ± 1.148	0.791 ± 0.015	0.820 ± 0.019	0.100 ± 0.003
Random	35.806 ± 2.427	0.291 ± 0.011	0.503 ± 0.015	0.001 ± 0.000

Table 32: Results for networks with a 50% density. We show the average value and plus/minus one standard error for each test result, obtained from a sample of 180 repeated experiments.

Method	SHD	Precision	Recall	Time (s)
DirectLiNGAM	22.778 ± 2.849	0.652 ± 0.032	0.524 ± 0.036	0.759 ± 0.047
DirectLiNGAM prior	21.022 ± 2.761	0.913 ± 0.019	0.613 ± 0.032	0.277 ± 0.014
DirectLiNGAM prior pruned	21.022 ± 2.761	0.913 ± 0.019	0.613 ± 0.032	0.277 ± 0.014
DirectLiNGAM pruned	22.378 ± 2.834	0.753 ± 0.032	0.524 ± 0.036	0.759 ± 0.047
Empty	45.656 ± 3.230	0.000 ± 0.000	0.000 ± 0.000	0.001 ± 0.000
Full	80.567 ± 4.102	0.380 ± 0.011	1.000 ± 0.000	0.001 ± 0.000
GES	57.383 ± 4.122	0.358 ± 0.020	0.311 ± 0.018	7.987 ± 0.792
GES prior	16.556 ± 2.075	0.924 ± 0.011	0.724 ± 0.025	3.498 ± 0.353
GES prior pruned	16.556 ± 2.075	0.924 ± 0.011	0.724 ± 0.025	3.498 ± 0.353
GES pruned	39.283 ± 2.918	0.765 ± 0.021	0.311 ± 0.018	7.987 ± 0.792
Notears	36.689 ± 3.118	0.703 ± 0.026	0.419 ± 0.028	33.756 ± 5.198
Notears prior	36.656 ± 3.106	0.610 ± 0.032	0.384 ± 0.029	0.215 ± 0.019
Notears prior pruned	36.656 ± 3.106	0.610 ± 0.032	0.384 ± 0.029	0.215 ± 0.019
Notears pruned	34.033 ± 2.992	0.801 ± 0.026	0.419 ± 0.028	33.756 ± 5.198
PC	41.694 ± 3.173	0.556 ± 0.017	0.428 ± 0.022	0.854 ± 0.076
PC prior	31.767 ± 2.632	0.862 ± 0.010	0.538 ± 0.020	0.530 ± 0.048
PC prior pruned	31.767 ± 2.632	0.862 ± 0.010	0.538 ± 0.020	0.530 ± 0.048
PC pruned	35.356 ± 2.868	0.868 ± 0.011	0.428 ± 0.022	0.854 ± 0.076
r-partite baseline	20.061 ± 2.075	0.891 ± 0.011	0.737 ± 0.022	0.104 ± 0.003
Random	39.761 ± 2.520	0.582 ± 0.018	0.492 ± 0.009	0.001 ± 0.000

Table 33: Results for networks with a 80% density. We show the average value and plus/minus one standard error for each test result, obtained from a sample of 180 repeated experiments.

Method	SHD	Precision	Recall	Time (s)
DirectLiNGAM	0.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	0.725 ± 0.032
DirectLiNGAM prior	0.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	0.270 ± 0.010
DirectLiNGAM prior pruned	0.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	0.270 ± 0.010
DirectLiNGAM pruned	0.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	0.725 ± 0.032
Empty	21.167 ± 1.575	0.000 ± 0.000	0.000 ± 0.000	0.001 ± 0.000
Full	99.389 ± 3.831	0.188 ± 0.008	1.000 ± 0.000	0.001 ± 0.000
GES	31.381 ± 2.754	0.481 ± 0.017	0.594 ± 0.016	6.504 ± 0.556
GES prior	0.572 ± 0.057	0.958 ± 0.005	1.000 ± 0.000	2.370 ± 0.204
GES prior pruned	0.572 ± 0.057	0.958 ± 0.005	1.000 ± 0.000	2.370 ± 0.204
GES pruned	17.914 ± 1.609	0.772 ± 0.015	0.594 ± 0.016	6.504 ± 0.556
Notears	7.428 ± 1.190	0.958 ± 0.005	0.922 ± 0.010	19.238 ± 2.633
Notears prior	11.167 ± 1.292	0.922 ± 0.007	0.854 ± 0.013	0.255 ± 0.018
Notears prior pruned	11.167 ± 1.292	0.922 ± 0.007	0.854 ± 0.013	0.255 ± 0.018
Notears pruned	6.389 ± 1.049	0.984 ± 0.003	0.922 ± 0.010	19.238 ± 2.633
PC	16.506 ± 1.580	0.677 ± 0.012	0.773 ± 0.014	0.422 ± 0.033
PC prior	12.508 ± 1.297	0.819 ± 0.009	0.855 ± 0.012	0.299 ± 0.025
PC prior pruned	12.508 ± 1.297	0.819 ± 0.009	0.855 ± 0.012	0.299 ± 0.025
PC pruned	13.278 ± 1.362	0.836 ± 0.010	0.773 ± 0.014	0.422 ± 0.033
r-partite baseline	5.089 ± 0.552	0.856 ± 0.008	0.992 ± 0.002	0.102 ± 0.002
Random	37.842 ± 1.684	0.295 ± 0.013	0.501 ± 0.012	0.001 ± 0.000

Table 34: Results for datasets with the linear plus exponential noise DGP. We show the average value and plus/minus one standard error for each test result, obtained from a sample of 360 repeated experiments.

Method	SHD	Precision	Recall	Time (s)
DirectLiNGAM	20.856 ± 1.573	0.270 ± 0.018	0.075 ± 0.006	0.723 ± 0.032
DirectLiNGAM prior	17.850 ± 1.543	0.882 ± 0.015	0.468 ± 0.019	0.269 ± 0.010
DirectLiNGAM prior pruned	17.850 ± 1.543	0.882 ± 0.015	0.468 ± 0.019	0.269 ± 0.010
DirectLiNGAM pruned	20.408 ± 1.567	0.458 ± 0.026	0.075 ± 0.006	0.723 ± 0.032
Empty	21.167 ± 1.575	0.000 ± 0.000	0.000 ± 0.000	0.001 ± 0.000
Full	99.389 ± 3.831	0.188 ± 0.008	1.000 ± 0.000	0.001 ± 0.000
GES	21.311 ± 1.476	0.411 ± 0.014	0.386 ± 0.015	3.202 ± 0.185
GES prior	14.631 ± 1.166	0.822 ± 0.011	0.673 ± 0.016	1.532 ± 0.102
GES prior pruned	14.631 ± 1.166	0.822 ± 0.011	0.673 ± 0.016	1.532 ± 0.102
GES pruned	16.669 ± 1.316	0.769 ± 0.017	0.386 ± 0.015	3.202 ± 0.185
Notears	20.358 ± 1.594	0.675 ± 0.022	0.255 ± 0.014	7.186 ± 1.551
Notears prior	20.511 ± 1.573	0.513 ± 0.025	0.163 ± 0.014	0.213 ± 0.025
Notears prior pruned	20.511 ± 1.573	0.513 ± 0.025	0.163 ± 0.014	0.213 ± 0.025
Notears pruned	19.633 ± 1.566	0.727 ± 0.022	0.255 ± 0.014	7.186 ± 1.551
PC	21.439 ± 1.418	0.422 ± 0.013	0.460 ± 0.016	0.364 ± 0.037
PC prior	15.092 ± 1.123	0.737 ± 0.012	0.669 ± 0.015	0.242 ± 0.023
PC prior pruned	15.092 ± 1.123	0.737 ± 0.012	0.669 ± 0.015	0.242 ± 0.023
PC pruned	17.011 ± 1.296	0.721 ± 0.015	0.460 ± 0.016	0.364 ± 0.037
r-partite baseline	15.533 ± 1.165	0.754 ± 0.012	0.695 ± 0.015	0.103 ± 0.002
Random	37.378 ± 1.686	0.289 ± 0.013	0.502 ± 0.012	0.001 ± 0.000

Table 35: Results for datasets with the quadratic plus Gaussian noise DGP. We show the average value and plus/minus one standard error for each test result, obtained from a sample of 360 repeated experiments.