Causal Discovery Under Hard Selection Bias: A New Robust Score-Matching Approach

Yiwen Qiu
ISTA
Am Campus 1, Klosterneuburg, 3400
Yiwen.Qiu@ista.ac.at

Francesco Montagna
ISTA
Am Campus 1, Klosterneuburg, 3400
Francesco.Montagna@ista.ac.at

Shimeng Huang
ISTA
Am Campus 1, Klosterneuburg, 3400
Shimeng.Huang@ista.ac.at

Francesco Locatello
ISTA
Am Campus 1, Klosterneuburg, 3400
Francesco.Locatello@ista.ac.at

Abstract

Selection bias is a difficult yet widespread problem in causal discovery, occurring whenever non-random selection processes lead to data that is not representative of the underlying populations. Due to its practical importance, related works have been proposed to address the problem under local [Versteeg et al., 2022] or interventional settings [Dai et al., 2025a], but general algorithms for observational data remain elusive. In this paper, we focus on truncated data (hard selection), and provide a general result for Additive Noise Models (ANM). We show that score-based methods are a surprisingly suitable candidate due to a special property of the score function ($\nabla \log p(x)$): *deterministic truncation preserves both the score* of the joint distribution as well as the Jacobian of the score. While established score-matching algorithms still fail in most settings, we actively leverage this observation to propose a new method for causal discovery that is robust under truncated data with ANMs. Overall, our approach offers a general, robust solution that is agnostic to external information about the selection process and still achieves comparable performance to the state-of-the-art approaches when selection is not present.

1 Introduction

Identifying causal relationships from observational data is a fundamental problem in many scientific disciplines. However, selection bias can significantly distort the observed data distribution, leading to incorrect inference about causal relationships [Heckman, 1979, Hernán et al., 2004]. Selection bias occurs when the data collected is not representative of the underlying population due to a non-random selection process, e.g., in epidemiological studies where only patients with certain traits (socio-economic or health factors) are recorded [Swanson, 2012, van Alten et al., 2024], or in molecular biology where only transcripts in cells that are highly expressed are measured and included in the study [Dwyer and Pleiss, 2023].

Some methods have been developed to address this problem in causal discovery, but they often rely on access to interventional data [Dai et al., 2025a, Luo et al., 2025], or only focus on a local structure (with three or four variables) [Gregory, 1995, Versteeg et al., 2022]. The seminal work FCI [Spirtes et al., 2000] and its variants [Colombo et al., 2012] recover the causal graph under latent variables and selection bias up to a Markov equivalence class without further interpretation. Recent studies based on Functional Causal Models (FCMs) by Zhang et al. [2016] have shown that if we know the location of the selection variable, i.e., if it is only outcome-dependent, then we can identify the full

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

causal graph under certain assumptions. Beyond these specific, partial solutions, a general approach for selection bias in causal discovery for observational data is missing.

In this paper, we investigate the case of *hard selection bias*: the densities of (some or all) observed variables are *truncated*, meaning that under selection, we do not observe data on certain parts of the (true) support at all. We begin by showing a surprising yet simple result: regardless of the location of selection, *causal asymmetry persists under hard selection bias within the score function of the data ($\nabla \log p(x)$)*. This introduces an opportunity, as the score function has been a central component in a rich line of work on causal discovery [Rolland et al., 2022, Montagna et al., 2023, 2025]. Unfortunately, we find that while the SCORE [Rolland et al., 2022] algorithm for non-linear additive Gaussian noise is robust to hard selection bias, the more general extensions [Montagna et al., 2025] are not. To this end, we propose a novel score-based method for causal discovery, and to our knowledge, it is the first general causal discovery method that is robust to hard selection, without relying on any external information about the selection process. Our contributions are as follows:

Contribution We evaluate the existing score function method—SCORE [Rolland et al., 2022], DAS [Montagna et al., 2023], NoGAM [Montagna et al., 2025]— and show that they can be surprisingly robust to hard selection bias (Section 3). However, failures persist: to tackle these cases, we propose a new method (Section 4) based on the conditional rank of the Hessian log density ($\nabla^2 \log(x)$), which is fully robust to hard selection bias under all Additive Noise Model (ANM) settings, with the only exception being the linear Gaussian case (see Table 1 for a summary). We provide a practical estimation algorithm for our method in Section 5, and demonstrate its effectiveness through extensive synthetic and real-world [Pinna et al., 2010] experiments in Section 6.

ANM Assumption	LiNGAM	CAM-UV	SCORE	NoGAM	CHess (Ours)
Nonlinear Gaussian	✗	✓	✓	✓	✓
Linear non-Gaussian	✓	✗	✗	✓	✓
Nonlinear non-Gaussian	✗	✓	✗	✓	✓
Hard Selection (Truncation)	✗	✗	✓*	✗	✓

Table 1: Identifiability guarantees of causal discovery methods across ANM assumptions. ✓ denotes full DAG identifiability; ✗ denotes no guarantee. LiNGAM [Shimizu et al., 2006, 2011] only deals with linear non-Gaussian ANMs, SCORE [Rolland et al., 2022] requires nonlinear mechanisms with Gaussian noise. NoGAM [Montagna et al., 2025] extends SCORE to arbitrary noise distributions. CAM-UV [Maeda and Shimizu, 2021] additionally handles latent confounders under nonlinear separable (CAM) assumptions. SCORE can handle latent hard selection bias in both Gaussian and non-Gaussian cases, which is not discussed in the original paper but is one of the surprising findings in this work (✓*). Our method (CHess) is the only method that is resilient to hard selection under all the above ANMs. Linear Gaussian models are known to be not identifiable beyond the Markov equivalence class [Peters and Bühlmann, 2014] thus omitted in the table.

2 Preliminaries and Problem Setup

Problem Setting: Hard Selection and What We Know about it

We consider a set of random variables $\mathbf{X} = \{X_1, \dots, X_d\}$ generated by a restricted additive noise model (ANM):

$$X_i := f_i(X_{\mathrm{PA}_i}) + N_i, \qquad i = 1, \dots, d, \tag{1}$$

where f_i are C^2 causal mechanisms, N_i are mutually independent noise terms with strictly positive densities, and the model satisfies the restricted ANM identifiability condition [Bühlmann et al., 2014]. The associated causal graph $\mathcal{G}$ is a DAG.

In the remainder of the paper, we assume the Markov and faithfulness conditions for the DAG $\mathcal{G}$ and the distribution $p(\mathbf{X})$ [Spirtes et al., 2000]. We say two nodes X_i and X_j are *d-separated* by a set of nodes $\mathbf{Z}$ in $\mathcal{G}$ if every path between them is blocked by $\mathbf{Z}$, and a path is blocked by conditioning on $\mathbf{Z}$ if the path contains a non-collider node in $\mathbf{Z}$ or a collider node that is not in $\mathbf{Z}$ and has no descendant in $\mathbf{Z}$. For a full definition, please refer to the Appendix.

Hard Selection (Truncation) We observe biased data from $p(\mathbf{X} \mid S = 1)$, where $p(\mathbf{X} \mid S = 1) \neq p(\mathbf{X})$. The selection mechanism is unknown and can be dependent on an arbitrary subset of $\mathbf{X}$. Here we focus on the hard selection (truncation) case, which is the scenario where the selection mechanism is deterministic and only data points in a subset $\mathcal{B}$ of the support of $\mathbf{X}$ are observed, i.e. $p(S = 1 \mid \mathbf{X} = x) = \mathbb{1}(x \in \mathcal{B})$ for a set $\mathcal{B} \subseteq \mathbb{R}^d$ with $\mathrm{vol}(\mathcal{B}) > 0$.

Causal Discovery under Selection Bias Given the biased observational data $\mathcal{D} = \{x^{(i)}\}_{i=1}^N$ from $p(\mathbf{X} \mid S = 1)$, the goal is to recover the causal graph $\mathcal{G}$ under the original distribution $p(\mathbf{X})$. This is a challenging problem that has been studied in the causal discovery literature in local causal discovery [Versteeg et al., 2022], interventional settings [Dai et al., 2025a, Luo et al., 2025], latent variable settings [Dai et al., 2025b], etc. However, to the best of our knowledge, *no method has been developed in the general setting* to tackle selection bias for causal discovery only with observational data. For a comprehensive review of the related work, please refer to Appendix C.

Given the generality of the problem, we focus on the following question in this paper:

> *Can we find properties of the original distribution $p(\mathbf{X})$ preserved under hard selection, and how can we leverage them for robust causal discovery?*

We take the first step to answer this question by revisiting the score function and score-matching based causal discovery methods, and show its potentials.

3 The Blessing of Score Functions

In this section, we look at the score function (Section 3.1) and review score-matching for causal discovery (Section 3.2). Then in Section 3.3, we present the advantage of using the score directly for causal discovery under truncation, and an example showing a failure case, which motivates our proposed method in the next section (Section 4).

3.1 Score Preservation

Consider the score of the selected distribution $\nabla_X \log p(x \mid s)$. The selected density satisfies

$$\log p(x \mid S = 1) = \log p(x) + \underbrace{\log \mathbb{1}(x \in \mathcal{B})}_{=0 \text{ for all } x \in \mathrm{int}(\mathcal{B})} \underbrace{- \log p(S = 1)}_{\text{constant}}, \quad \forall x \in \mathbb{R}^d,$$

because of Bayes' rule. Take the derivative, for $x \in \mathrm{int}(\mathcal{B})$,

$$\boxed{\underbrace{\nabla_X \log p(x \mid S = 1)}_{\text{(the score on the biased data)}} = \underbrace{\nabla_X \log p(x)}_{\text{(the score on the true data)}}} \tag{2}$$

The score is exactly preserved at every observed data point.

Remark 1 (Score preservation under hard selection)**.** *This is an important and nontrivial observation that we leverage in this paper for developing a novel score-based method for causal discovery under hard selection. The same argument also applies to the second-order score $\nabla_X^2 \log p(x)$, which is also preserved under hard selection, i.e. $\nabla_X^2 \log p(x \mid S = 1) = \nabla_X^2 \log p(x)$ (in the interior of $\mathcal{B}$). This aligns with the local geometrical nature of the score, which characterizes the curvature of the log-density pointwise, and which is preserved under truncation.*

Remark 2 (Boundary effect)**.** *One might notice that the score is only preserved in the interior of $\mathcal{B}$, and not on its boundary, where we have an extra term $\delta(x)$ in the score due to the derivative of the indicator function. This boundary effect can be problematic for score estimation: [Rolland et al., 2022, Montagna et al., 2025] show that we can use Stein identity [Stein, 1972] for score estimation, which relies on an vanishing boundary condition on $p(x)$, i.e. $h(x)p(x) \to 0$ as $x \to \infty$, for any testing function $h(x)$. However, under truncation, the selected density $p(x \mid s)$ does not satisfy this condition. In practice, we propose to use the diffusion score matching method [Sanchez et al., 2023], which is more robust to boundary effects. We will discuss this in more detail in Section 4.*

3.2 Revisiting Score-Matching

We provide a review of the score-matching based causal discovery methods. The score component for node X_i decomposes as:

$$\partial_{X_i} \log p(x) = s_{n_i}(N_i) - \sum_{k \in X_{\mathrm{CH}_i}} \frac{\partial f_k}{\partial x_i}(x_{\mathrm{PA}_k}) \cdot s_{n_k}(N_k), \tag{3}$$

where $s_{n_i}(N_i) := (\log p_{N_i})'$ is the score function of the i-th noise distribution and X_{CH_i} denotes the children of X_i in $\mathcal{G}$.

Sink Identification For a sink node X_l (with $X_{\mathrm{CH}_l} = \emptyset$):

$$\partial_{X_l} \log p(x) = s_{n_l}(N_l). \tag{4}$$

Under *Gaussian* noise, Rolland et al. [2022] propose the variance test (Proposition 1) as sink node identification method criterion:

Proposition 1 (Gaussian case: Variance test [Rolland et al., 2022])**.** *Let* $\mathbf{X}$ *be generated by a restricted ANM* (1) *with mutually independent Gaussian noise terms.*

$$\mathrm{Var}\left[\partial_{X_i} s_{x_i}(\mathbf{X})\right] = 0 \Longleftrightarrow X_i \textit{ is a sink node, } \forall i = 1, \ldots, d. \tag{5}$$

For a sink node, $s_{x_l}(x_l) = -\frac{x_l - f_l(\mathrm{PA}_l(x_l))}{\sigma_l^2}$, and $s'_{x_l}(x_l) = c$. Therefore, $\mathrm{Var}_X(s'_{n_l}(N_l)) = 0$.

Extending beyond Gaussian noise, Montagna et al. [2025] construct the regression residual test (Proposition 2) as sink-node identification method criterion:

Proposition 2 (Non-Gaussain case: Residual test [Montagna et al., 2025])**.** *Let* $\mathbf{X}$ *be generated by a restricted ANM* (1) *with mutually independent non-Gaussian noise terms. For any* $i \in [d]$*,*

$$\mathbb{E}\left[\left(\partial_{X_i} s_i(\mathbf{X}) - \mathbb{E}\left[\partial_{X_i} s_i(\mathbf{X}) \mid R_i\right]\right)^2\right] = 0 \Longleftrightarrow X_i \textit{ is a sink node.} \tag{6}$$

For a sink node, $R_l = X_l - \mathbb{E}[X_l \mid X_{\setminus l}]$, which equals the centered noise $N_l - \mathbb{E}(N_l)$ for a true sink node. This means for a sink node (without selection bias), the score of X_l can be perfectly predicted from the residual R_l.

Edge (neighbor) Identification The Markov equivalence graph can be constructed by reading the conditional independence of pairs of random variables as shown by [Spantini et al., 2018, Montagna et al., 2023], in Proposition 3.

Proposition 3 ([Spantini et al., 2018, Montagna et al., 2023])**.** *Let* $\mathbf{X}$ *be generated by a restricted ANM* (1) *with mutually independent noise terms. For any* $i, j \in [d]$*,*

$$\partial^2_{X_i X_j} \log p(x) = 0 \Longleftrightarrow X_i \perp\!\!\!\perp X_j \mid \mathbf{X}_{\setminus\{i,j\}} \textit{in } \mathcal{G}. \tag{7}$$

3.3 When Score-Matching is Surprisingly Resilient and When it is Not

How does score-matching behave under hard selection? In this subsection, we show that for sink-node identification under truncation, Proposition 1 and 3 still hold (modified as Proposition 4 and 8), but Proposition 2 fails in general. Unlike other causal discovery methods that rely on conditional independence tests that are susceptible to hidden selection variables, the score-matching based methods are surprisingly resilient to hard selection bias. However, the residual test relies not only on the score, but also on the structure underlying the density, which can be distorted by selection bias. This motivates us to leverage the score directly for causal discovery, without relying on the residual or any graphical assumptions.

The first fold: Variance test still holds Under nonlinear ANM and Gaussian noise assumptions, the result in Proposition 1 can be directly extended to truncated data, as stated in Proposition 4.

Proposition 4 (Gaussian case: Variance test under truncation)**.** *Let* $\mathbf{X}$ *be generated by a restricted ANM* (1) *with mutually independent Gaussian noise terms, and let* $S = \mathbb{1}(x \in \mathcal{B})$ *denote a deterministic selection for a truncation region* $\mathcal{B}$*.*

$$\mathrm{Var}\left[\partial_{X_i} s_i(\mathbf{X} \mid S = 1)\right] = 0 \Longleftrightarrow X_i \textit{ is a sink node, } \forall i = 1, \ldots, d. \tag{8}$$

This implies that with data from $p(\mathbf{X} \mid S = 1)$, we can still identify the sink nodes by checking the variance of the Hessian log-density, which is an additional outcome not realized before. The proof directly relies on the score preservation property in (2).

Similarly, the result from Proposition 3 naturally extend to testing $\partial^2_{X_i X_j} \log p(x \mid S = 1)$ for skeleton discovery, as shown in Appendix B.

The second fold: Residual test fail Under hard selection, the selected (observed) residual is:

$$R_l^{\text{sel}} = X_l - \mathbb{E}[X_l \mid X_{\setminus l}, S = 1].$$

If the selection is only a function of a single node, then the biased residual R_l^{sel} is still a function of X_l. This is a special case where the residual test still holds, which is elaborated in Appendix B. But in general, when we do not know whether the selection is only a function on a single node or not, the residual test fails as in Example 1.

Example 1 (Failure case)**.** Consider a sink node X_l with parent X_{PA_l}. If the selection mechanism depends on both X_l and some $X_k \in X_{\text{PA}_l} \subseteq X_{\setminus l}$ (e.g., $S = \mathbb{1}(|X_l + X_k| \leq c)$), conditioning on $S = 1$ induces dependence between N_l and X_{PA_l}, so $R_l^{\text{sel}} = X_l - \mathbb{E}[X_l \mid X_{\setminus l}, S = 1] \neq X_l - \mathbb{E}[X_l \mid X_{\setminus l}] = R_l = N_l - \mathbb{E}N_l$. Note that the score $\partial_{X_l} \log p(x) = s_l(N_l)$ is preserved, but the residual R_l^{sel} is a biased proxy for N_l, and the residual test incorrectly rejects the true sink.

How do we surpass the issue? Our proposal: exploit the score directly for causal asymmetry. The residual test fails because it depends on the global structure of the density, while the score is a local property of the density at each data point, and is preserved under truncation (2). Therefore, we need to exploit this property of the score directly for a novel method for causal discovery. Since we have seen that Proposition 3 remains true under selection, the remaining challenge is to identify the sink nodes given the skeleton.

4 Causal Discovery Method Robust to Hard Selection

We present our main result in three-fold: we first introduce the intuition of the method in Section 4.1, then in Section 4.2 we give the main theoretical result under linear ANM assumptions (with non-Gaussian noise) in Proposition 6 and the more general result for non-linear ANM in Proposition 9.

4.1 Intuition and Definitions

As shown in the example in Section 3.3, the residual test [Montagna et al., 2025] fails under hard-selected data in general (in non-Gaussian noise cases).

Intuition: Asymmetry of Number of Independent Noise Sources. The key intuition is that the score of a sink node has exactly *one* independent noise term that contributes to it, after conditioning on all its neighbors[1], while a non-sink node has ≥ 2 independent noise terms. Consider the graph in Figure 1, the path between a non-sink node and the noise of its child is opened by conditioning on its child, while the argument does not apply to a sink node that does not have child terms.

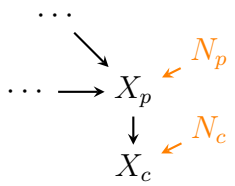


Figure 1: In the subgraph, the sink node X_c is only d-connected to its own noise N_c after conditioning on X_{PA_c}, while X_p is d-connected to N_p and N_c after conditioning on $X_{\text{PA}_p} \cup X_{\text{CH}_p}$.

Algebraic Connection: To translate the number of independent noises into an algebraic criterion, we consider the score function $s(x) := \nabla \log p(x)$. The main observation is that if we take the gradient of the score on node X_i w.r.t all noise terms[2], we obtain a vector that has contributions from multiple source directions. The conditional rank (conditioning on all neighbors of X_i) of the variance of such a vector's outer product indicates the number of independently varying factors, e.g., for a sink node the rank is 1 (for non-Gaussian noise), which we use as an identification rule to distinguish a sink node from a non-sink node.

[1] In practice (Algorithm 1), we perform skeleton discovery first to identify the neighbors of each node.
[2] Could be understood as taking the gradient w.r.t. all nodes with a change of basis

In light of the intuition and algebraic connection above, we formally define the Conditional Hessian (M_i^{cond}) outer product matrices in Definition 1.

Definition 1 (Conditional Hessian outer product matrix)**.** For each node X_i, define the i-th Hessian row vector evaluated at x:

$$H_i(x) := \left(\frac{\partial^2 \log p(x)}{\partial X_i \partial X_1}, \ \ldots, \ \frac{\partial^2 \log p(x)}{\partial X_i \partial X_d} \right) \in \mathbb{R}^d. \tag{9}$$

We define the *conditional gradient outer product matrix* as:

$$M_i^{\text{cond}}(z) := \text{Var}\big(H_i(\mathbf{X}) \,\big|\, X_{\text{NB}_i} = z\big) \in \mathbb{R}^{d \times d}, \tag{10}$$

where X_{NB_i} denotes the set of neighbors of X_i.

Hessian Row Structure: sinks vs. Non-sinks The Hessian row H_i is the gradient of the score component $s_i(x) = \partial_{X_i} \log p(x)$. By the chain rule applied to (3):

$$H_i(x) = \underbrace{s_i'(N_i) \cdot \nabla_x N_i}_{①} - \sum_{k \in X_{\text{CH}_i}} \Big[\underbrace{s_k'(N_k) \cdot \frac{\partial f_k}{\partial x_i} \cdot \nabla_x N_k}_{②\text{child noise direction}} + \underbrace{s_k(N_k) \cdot \nabla_x \left(\frac{\partial f_k}{\partial x_i} \right)}_{③\text{mechanism curvature}} \Big], \tag{11}$$

where $N_i = x_i - f_i(x_{\text{PA}_i})$ is the noise term at node X_i, so $\nabla_x N_i = e_i - \nabla_x f_i(x_{\text{PA}_i})$ and e_i is the i-th standard basis vector. For a sink node, the whole summation term over all children vanishes, so H_i only has one noise term direction contributed by N_i:

$$H_l(x) = s_l'(N_l) \cdot \nabla_x N_l \quad (\text{only term } ① \text{ remains}). \tag{12}$$

Such direction $\nabla_x N_i$ is a constant vector if the mechanism is linear, or a varying vector function of x.

In the following section, we will formalize the intuition for the linear ANM case first: under linearity, ③ vanishes and non-sink node H_i has $|\,\text{CH}_i\,| + 1$ noise term directions contributed by N_i in (11), while sink node H_l only has one noise term direction contributed by N_l in (12).

4.2 Main Results: from Linear to Nonlinear ANM

First, we present the result under linear ANM with non-Gaussian noise, which clearly illustrates the intuition of the method and the algebraic connection between the number of independent noise terms and the rank of M_i^{cond}. Then we present the more general result for nonlinear ANM with both Gaussian and non-Gaussian noise.

For a linear ANM, the decomposition (11) simplifies to a combination of noise term directions (only terms ① and ② remain, and $\nabla N_j, j \in \{i\} \cup \text{CH}_i$ is a constant vector). The rank of M_i^{cond} then directly counts the number of independent noise terms that contribute to the variance of H_i after conditioning on the neighbors of X_i: conditioning on the parents of X_i blocks the contribution of the noise terms from the upstream (parents), while conditioning on the children of X_i still keeps the contribution of the noise terms of the children.

Proposition 5 (Connection between Conditional Rank and Independent Noise Terms in Linear ANM)**.** *Under a linear non-Gaussian ANM, the rank of $M_i^{\text{cond}}(z)$ equals the number of independent noise terms contributes to s_i by conditioning on X_{NB_i} for generic z[a]:*

$$\text{rank}\big(M_i^{\text{cond}}(z)\big) \ = \ |\mathcal{N}_i|,$$

where the noise index set $\mathcal{N}_i \ := \ \{i\} \cup \text{CH}_i$.

[a] By genericity, we exclude the degenerate cases where the graph structure is not faithfully reflected in the rank property, similar to the cases where standard faithfulness is violated. See Remark. 5

Proof Sketch. The proof consists of two main steps: (1) express M_i^{cond} as outer products $\mathbf{v}_i \mathbf{v}_i^\top$ weighted by the conditional covariates of the noise terms; (2) show that the vectors $\mathbf{v}_i$ has full column rank (linearly independent noise terms), so the rank of M_i^{cond} equals the number of noise terms.

To illustrate the intuition of Proposition 5, we present the following example of a chain structure, where the sink node has only one independent noise term while the non-sink node has two independent noise terms after conditioning on its neighbors.

Example 2 (Number of Independent Terms in Chain)**.** Consider a chain $X_1 \to X_2 \to X_3$, where $X_1 = N_1$, $X_2 = aX_1 + N_2$, $X_3 = bX_2 + N_3$ with non-Gaussian noise.

Illustration of Example 2. For non-sink X_1 and X_2, both have two independent noise terms that contribute to their score function after conditioning on their neighbors, thus the rank of the conditional outer product matrix M_1^{cond} and M_2^{cond} (Eqn (10)) is 2, while M_3^{cond} has rank 1. The reason is that, for source node X_1, conditioning on its neighbor (child in this case) $X_2 = z_2$ induces dependency of N_2 on N_1, thus leading to the direction of N_1 and a non-affine transformation of N_1, while for a sink node X_3, conditioning on its neighbor (parent in this case) will only leave a single direction term contributed by N_3. A formal proof is provided in Appendix A.2 □

Proposition 6 (Conditional Rank Test for Sink Identification (Linear))**.** *Let* $\mathbf{X}$ *be generated by a restricted* linear *ANM* (1) *with independent noise terms. Then for generic* z*:*

(a) ***Sink.*** *If* X_l *is a sink (i.e.,* $\text{CH}_l = \emptyset$*), then* $\text{rank}\big(M_l^{\text{cond}}(z)\big) = 1$.

(b) ***Non-sink.*** *If* X_i *is not a sink (i.e.,* $|\,\text{CH}_i\,| \geq 1$*), then* $\text{rank}\big(M_i^{\text{cond}}(z)\big) \geq 2$.

Therefore, a node X_i *is a sink if and only if* $\text{rank}(M_i^{\text{cond}}(z)) = 1$.

Proof sketch. This is a direct consequence of Proposition 5 and the fact that the rank of M_i^{cond} is invariant to conditioning on X_{NB_i} (since the noise terms are independent of X_{NB_i}, so the variance of $s_j'(N_j)$ does not change after conditioning). □

More generally, for the nonlinear ANM, *the rank gap between sink and non-sink nodes still exists*, because the decomposition (11) contains an additional mechanism curvature term, and this term is generally linearly independent of the noise term directions. Due to space limit, we present the informal Proposition 7 and the full proposition under nonlinear ANM in the Appendix (Proposition 9).

Proposition 7 (Informal Conditional Rank Test for Sink Identification (Nonlinear))**.** *Let* $\mathbf{X}$ *be generated by a restricted* nonlinear *ANM* (1) *with independent noise terms. Then the rank gap of* $\text{rank}(M_i^{\text{cond}}(z))$ *for generic* z *between sink and non-sink node still exists and is at least one.*

For a summary of the rank values for sink vs. non-sink nodes across different ANM settings, see Table 2. In particular, the rank gap between sink and non-sink nodes exists (thus sink nodes are identifiable) in all settings except for the linear Gaussian case, where all nodes have rank 0.

Setting	Sink rank	Non-sink rank	Gap
Linear Gaussian	0	0	0 (not identifiable)
Linear non-Gaussian	1	≥ 2	≥ 1
Nonlinear Gaussian	0	≥ 1	≥ 1
Nonlinear non-Gaussian	1	≥ 2	≥ 1

Table 2: Rank of $M_i^{\text{cond}}(z)$ for sink vs. non-sink nodes across different ANM settings.

Remark 3 (Practical Usage of Proposition 6 and Table 2)**.** *Without knowledge about the type of ANM and type of noise one is dealing with, a* $\text{rank}(M_i^{cond}) = 1$ *result could either imply a sink node in a linear non-Gaussian model, or a non-sink node in a nonlinear Gaussian model. In practice, for either case, we do not bother to distinguish the model / noise types, but only take the node that has the minimal rank to be the current sink node (and we perform this procedure iteratively), i.e., as long as a gap exists.*

Discussion of Unconditional Hessian Rank Test Why do we need the conditioning on the neighbors? Can we just use the unconditional Hessian outer product matrix $M_i = \mathbb{E}[H_i(\mathbf{X})H_i(\mathbf{X})^\top]$ to identify sink nodes? The short answer is yes, but it requires strong structural constraints. In Appendix B.3, we introduce the *Unconditional Rank Test*. However, it requires stronger structural assumptions (involving the number of $|\,\mathrm{PA}_i\,|$ and $|\,\mathrm{CH}_i\,|$) since the noise coming from the parents is not blocked by conditioning on the neighbors anymore. On the other hand, the Conditional Hessian Rank Test proposed in Proposition 6 is free from the structural assumptions and thus more practical.

5 Estimation

In this section, we present a practical estimation procedure for causal discovery based on the conditional Hessian rank test. This procedure involve 4 main steps:(1) estimation of Hessian log-density ($H_j(x^{(i)})$), (2) use the estimated Hessian $H_j(x^{(i)})$ to estimate the conditional variance matrix $\widehat{M}_j^{\mathrm{cond}}$, (3) compute the rank of $\widehat{M}_j^{\mathrm{cond}}$ to identify sinks, and (4) . The overall algorithm is provided in Algorithm 1 and more details on the estimation procedure in Appendix E.2.

Score and Hessian Estimation Estimation of the score function (and its Jacobian) is a well-studied problem in statistics and machine learning, e.g., the Stein estimator [Stein, 1972, Rolland et al., 2022, Montagna et al., 2025], the kernel density estimator (KDE) [Sasaki et al., 2016], and most recently a diffusion model based estimator [Sanchez et al., 2023]. In this paper, we apply the commonly used Stein-estimator, and also a diffusion model based estimator, because it does not rely on a strictly positive density $p(x)$ while both Stein and KDE estimators suffer when $p(x) \approx 0$ (See Remark 2).

Remark 4. *(Practical Estimation of M_i^{cond}) In practice, we use a residual estimator $\widehat{M}_i^{cond} := \frac{1}{n}\sum_{t=1}^{n} r_i(x^{(t)})r_i(x^{(t)})^\top$ as a proxy, where $r_i(x) := H_i(x) - \hat{m}_i(x_{NB_i})$ is the residual of using $\hat{m}_i(\cdot)$ to estimate H_i. The residual estimator $\widehat{M}_i^{cond}$ shares a similar spirit with the estimation method in NoGAM [Montagna et al., 2025], which also uses the residual of a regression to estimate the noise. Further details are discussed in the Appendix. B.4*

6 Experiments

Given the proposed method, we conduct experiments to evaluate its performance on both synthetic (Section 6.1) and real-world (Section 6.2) datasets. The goal is to show that (1) for general causal discovery, the proposed method performs competitively with state-of-the-art baselines, and (2) in the presence of hard selection bias, the proposed method outperforms all baselines that do not explicitly account for selection bias.

6.1 Synthetic Experiments

Data generation. We generate Erdős–Rényi DAGs with $d \in \{5, 10\}$ nodes and target average degree 2.0. We consider three SEM regimes: (1)Linear non-Gaussian (Linear-nG). (2)Nonlinear Gaussian (MLP-G): $X_i = f_i(\mathrm{Pa}(X_i)) + \epsilon_i$, with $\epsilon_i \sim \mathcal{N}(0, \sigma_i^2)$ and f_i a randomly initialised 1-hidden-layer MLP. (3)Nonlinear non-Gaussian (MLP-nG). Same MLP mechanism as MLP-G but with Gumbel or log-normal noise as in Linear-nG.

Selection mechanism. We add K latent selection variables $S_1, \ldots, S_K$. We simulate the hardest case of selection bias where the selection mechanism depends on *all* observed variables. Each S_k is a random linear combination $S_k = \sum_{j=1}^{d} \alpha_{kj} X_j$ with coefficients $\alpha_{kj} \sim \mathcal{U}(-1, 1)$, and we keep the samples where $S_k \geq q_k$ for all k simultaneously, where q_k is the $1-\rho$ empirical quantile, ρ being the retaining fraction. After applying the selection mechanism, we keep sample sizes $n = 1000$ (from an initial pool of 10^5 samples) to avoid the interference of different sample sizes on the performance.

Baselines. We compare against constraint-based methods PC and FCI [Spirtes et al., 2000, 2013], score-based methods GES [Chickering, 2003], permutation-based method GRaSP [Lam et al., 2022], and LiNGAM-family estimators ICA-LiNGAM [Shimizu et al., 2006] and DirectLiNGAM [Shimizu et al., 2011], together with score-matching baselines SCORE [Rolland et al., 2022], DAS [Montagna et al., 2023], NoGAM [Montagna et al., 2025], CAM [Bühlmann et al., 2014], and DiffAN [Sanchez

et al., 2023]. We report F1 of the arrow on the estimated DAG with part of the baselines due to space limit[3], with one selection variables as a function of all observed variables. We also include sensitivity analysis with varying selection fraction $\rho \in \{1.0, 0.1, 0.05\}$, selection shape {`halfspace`, `annulus`, `two boxes` } in Appendix F.

Results. In the absence of selection bias ($\rho = 1.0$), the proposed method performs competitively with the baselines across all SEM regimes, and outperforms all baselines when hard selection bias is present (in the linear-nG regime, the LiNGAM baselines sometimes outperform even when selection bias is present, which is discussed in Appendix F).

Table 3: F1 (Arrow) for Linear-nG (gumbel), $d = 5$ and $d = 10$

	$d = 5$, gumbel			$d = 10$, gumbel		
Fraction kept	$\rho = 1.0$	$\rho = 0.1$	$\rho = 0.05$	$\rho = 1.0$	$\rho = 0.1$	$\rho = 0.05$
SCORE	0.57 (±0.25)	0.37 (±0.17)	0.41 (±0.14)	0.65 (±0.12)	0.46 (±0.11)	0.45 (±0.10)
DAS	0.57 (±0.25)	0.37 (±0.17)	0.41 (±0.14)	0.66 (±0.11)	0.48 (±0.12)	0.49 (±0.10)
NoGAM	0.81 (±0.21)	0.40 (±0.16)	0.40 (±0.14)	0.78 (±0.16)	0.47 (±0.14)	0.45 (±0.11)
CAM	0.31 (±0.20)	0.18 (±0.14)	0.19 (±0.12)	0.23 (±0.08)	0.22 (±0.11)	0.19 (±0.12)
DirectLiNGAM	**0.99** (±0.03)	0.64 (±0.14)	0.65 (±0.16)	**0.98** (±0.03)	0.56 (±0.05)	0.54 (±0.07)
DiffAN	0.40 (±0.28)	0.31 (±0.24)	0.29 (±0.23)	0.33 (±0.15)	0.25 (±0.16)	0.25 (±0.14)
CHess (Stein)	0.90 (±0.14)	**0.76** (±0.11)	0.70 (±0.16)	0.72 (±0.19)	**0.59** (±0.14)	0.51 (±0.09)
CHess (DiffAN)	0.81 (±0.25)	0.71 (±0.17)	**0.72** (±0.11)	0.64 (±0.15)	0.49 (±0.15)	**0.64** (±0.09)

6.2 Real-world experiments

Dataset. We use the protein-signalling dataset of Sachs et al. [2005], loaded through the `cdt` package, which yields 7466 samples over $d = 11$ proteins and a verified ground-truth DAG. To probe robustness to selection bias in severe cases, we apply the selection mechanism as in Section 6.1 with $K = 1, \rho = 0.1$. After selection, there are 373 samples, so we apply random downsampling to $n = 373$ samples for the non-selection bias setting $\rho = 1.0$ to avoid the interference of different sample sizes with pruning parameter for a fair comparison.

Results. In both cases with and without selection bias, our method outperforms the baselines. Because Sachs essentially provide perturbation data and we aggregate all the data, the result should be interpreted with care. In addition, we provide results on separate sub-datasets in Appendix F.4, with varying selection strengths and shapes. Overall, these experiments consistently demonstrate that the proposed method achieves competitive performance compared with other baselines.

Table 4: F1 (Arrow) on Sachs Dataset for $\rho = 1$ (downasmpled) and $\rho = 0.1$

Method	$\rho = 1$ ($n = 373$)	$\rho = 0.1$
SCORE	0.28 (±0.05)	0.29 (±0.05)
DAS	0.19 (±0.10)	0.23 (±0.08)
NoGAM	0.29 (±0.04)	0.30 (±0.06)
CAM	0.40 (±0.08)	0.28 (±0.09)
PC	0.25 (±0.09)	0.23 (±0.07)
FCI	0.13 (±0.04)	0.14 (±0.11)
GES	0.25 (±0.06)	0.22 (±0.08)
GRaSP	0.28 (±0.07)	0.22 (±0.10)
ICA-LiNGAM	0.27 (±0.08)	0.30 (±0.07)
DirectLiNGAM	0.30 (±0.07)	0.33 (±0.09)
DiffAN	0.30 (±0.08)	0.19 (±0.05)
CHess (Stein)	**0.42** (±0.09)	0.32 (±0.12)
CHess (DiffAN)	0.41 (±0.10)	**0.38** (±0.12)

7 Conclusion

In this paper, we have presented a novel approach to address the challenges of hard selection bias in causal discovery. Our method leverages the strengths of the score function, which is preserved under hard selection, allowing us to effectively identify causal relationships even in

[3]For details of the evaluation metrics (SHD, F1 adjacency, order accuracy, FPR), and results on all baselines, see Appendix F.

the presence of selection bias, while all baseline methods suffer under selection. We have demonstrated the effectiveness of our approach through experiments on both synthetic and real-world datasets.

Limitations and Future Work. First, our method assumes hard selection bias, which may not always be the case in real-world scenarios. Future work could explore extensions of our method to handle soft (non-deterministic) selection bias. Additionally, although we provide the theoretical result that the score function is preserved under hard selection, in practice, we found that it is hard to estimate when the truncation is severe. Therefore, further research is needed to develop more robust estimation techniques for the score-matching based causal discovery.

8 Acknowledgments

This research was funded in whole or in part by the Austrian Science Fund (FWF) [10.55776/COE12], ERC grant (Causal-AIM) [PR.A0184.1.F.1087.A]. FM acknowledges the Chan Zuckerberg Initiative for financial support.

References

Elias Bareinboim and Judea Pearl. Controlling selection bias in causal inference. In *Artificial Intelligence and Statistics*, pages 100–108. PMLR, 2012.

Elias Bareinboim and Jin Tian. Recovering causal effects from selection bias. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29, 2015.

Elias Bareinboim, Jin Tian, and Judea Pearl. Recovering from selection bias in causal and statistical inference. In *Probabilistic and causal inference: The works of Judea Pearl*, pages 433–450, 2014.

Peter Bühlmann, Jonas Peters, and Jan Ernest. CAM: Causal additive models, high-dimensional order search and penalized regression. *Ann. Statist.*, 42(6), December 2014. ISSN 0090-5364. doi: 10.1214/14-AOS1260.

David Maxwell Chickering. Optimal structure identification with greedy search. *J. Mach. Learn. Res.*, 3(null):507–554, mar 2003. ISSN 1532-4435. doi: 10.1162/153244303321897717. URL https://doi.org/10.1162/153244303321897717.

Diego Colombo, Marloes H Maathuis, Markus Kalisch, and Thomas S Richardson. Learning high-dimensional directed acyclic graphs with latent and selection variables. *The Annals of Statistics*, pages 294–321, 2012.

Juan D. Correa, Jin Tian, and Elias Bareinboim. Identification of Causal Effects in the Presence of Selection Bias. *AAAI*, 33(01):2744–2751, July 2019. ISSN 2374-3468, 2159-5399. doi: 10.1609/aaai.v33i01.33012744.

Haoyue Dai, Ignavier Ng, Jianle Sun, Zeyu Tang, Gongxu Luo, Xinshuai Dong, Peter Spirtes, and Kun Zhang. When Selection Meets Intervention: Additional Complexities in Causal Discovery, March 2025a.

Haoyue Dai, Yiwen Qiu, Ignavier Ng, Xinshuai Dong, Peter Spirtes, and Kun Zhang. Latent Variable Causal Discovery under Selection Bias, December 2025b.

Zachary W Dwyer and Jeffrey A Pleiss. The problem of selection bias in studies of pre-mrna splicing. *nature communications*, 14(1):1966, 2023.

Cooper Gregory. Causal Discovery from Data in the Presence of Selection Bias, 1995.

James J Heckman. *Sample selection bias as a specification error (with an application to the estimation of labor supply functions)*, volume 172. National Bureau of Economic Research Cambridge, MA, 1977.

James J. Heckman. Sample Selection Bias as a Specification Error. *Econometrica*, 47(1):153–161, 1979. ISSN 0012-9682. doi: 10.2307/1912352.

Miguel A Hernán, Sonia Hernández-Díaz, and James M Robins. A structural approach to selection bias. *Epidemiology*, pages 615–625, 2004.

Amin Jaber, Jiji Zhang, and Elias Bareinboim. Causal identification under markov equivalence: Completeness results. In *International Conference on Machine Learning*, pages 2981–2989. PMLR, 2019.

Wai-Yin Lam, Bryan Andrews, and Joseph Ramsey. Greedy relaxations of the sparsest permutation algorithm. In *Uncertainty in Artificial Intelligence*, pages 1052–1062. PMLR, 2022.

Gongxu Luo, Haoyue Dai, Loka Li, Chengqian Gao, Boyang Sun, and Kun Zhang. Gene regulatory network inference in the presence of selection bias and latent confounders. *arXiv preprint arXiv:2501.10124*, 2025.

Takashi Nicholas Maeda and Shohei Shimizu. Causal additive models with unobserved variables. In *Uncertainty in Artificial Intelligence*, pages 97–106. PMLR, 2021.

Francesco Montagna, Nicoletta Noceti, Lorenzo Rosasco, Kun Zhang, and Francesco Locatello. Scalable Causal Discovery with Score Matching. In *Proceedings of the Second Conference on Causal Learning and Reasoning*, pages 752–771. PMLR, August 2023.

Francesco Montagna, Philipp M. Faller, Patrick Bloebaum, Elke Kirschbaum, and Francesco Locatello. Score matching through the roof: Linear, nonlinear, and latent variables causal discovery, March 2025.

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

Jonas Peters and Peter Bühlmann. Identifiability of gaussian structural equation models with equal error variances. *Biometrika*, 101(1):219–228, 2014.

Andrea Pinna, Nicola Soranzo, and Alberto De La Fuente. From knockouts to networks: establishing direct cause-effect relationships through graph analysis. *PloS one*, 5(10):e12912, 2010.

Joseph Ramsey and Bryan Andrews. Fask with interventional knowledge recovers edges from the sachs model. arxiv. org preprint. *arXiv preprint arXiv:1805.03108*, 2018.

James M Robins, Miguel Angel Hernan, and Babette Brumback. Marginal structural models and causal inference in epidemiology, 2000.

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

Karen Sachs, Omar Perez, Dana Pe'er, Douglas A Lauffenburger, and Garry P Nolan. Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, 308(5721):523–529, 2005.

Pedro Sanchez, Xiao Liu, Alison Q. O'Neil, and Sotirios A. Tsaftaris. Diffusion Models for Causal Discovery via Topological Ordering, June 2023.

Hiroaki Sasaki, Yung-Kyun Noh, Gang Niu, and Masashi Sugiyama. Direct density derivative estimation. *Neural computation*, 28(6):1101–1140, 2016.

Ashka Shah, Arvind Ramanathan, Valerie Hayot-Sasson, and Rick Stevens. Causal discovery and optimal experimental design for genome-scale biological network recovery. In *Proceedings of the Platform for Advanced Scientific Computing Conference*, pages 1–11, 2023.

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.

Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32, 2019.

Alessio Spantini, Daniele Bigoni, and Youssef Marzouk. Inference via low-dimensional couplings. *Journal of Machine Learning Research*, 19(66):1–71, 2018.

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

Peter L. Spirtes, Christopher Meek, and Thomas S. Richardson. Causal Inference in the Presence of Latent Variables and Selection Bias, February 2013.

Charles Stein. A bound for the error in the normal approximation to the distribution of a sum of dependent random variables. In *Proceedings of the sixth Berkeley symposium on mathematical statistics and probability, volume 2: Probability theory*, volume 6, pages 583–603. University of California Press, 1972.

James M. Swanson. The UK Biobank and selection bias. *The Lancet*, 380(9837):110, July 2012. ISSN 0140-6736, 1474-547X. doi: 10.1016/S0140-6736(12)61179-9.

Sjoerd van Alten, Benjamin W Domingue, Jessica Faul, Titus Galama, and Andries T Marees. Reweighting UK Biobank corrects for pervasive selection bias due to volunteering. *Int J Epidemiol*, 53(3):dyae054, May 2024. ISSN 0300-5771. doi: 10.1093/ije/dyae054.

Benito Van der Zander, Maciej Liskiewicz, and Johannes Textor. Constructing separators and adjustment sets in ancestral graphs. In *CI@ UAI*, pages 11–24, 2014.

Philip Versteeg, Joris Mooij, and Cheng Zhang. Local Constraint-Based Causal Discovery under Selection Bias. In *Proceedings of the First Conference on Causal Learning and Reasoning*, pages 840–860. PMLR, June 2022.

Kun Zhang, Jiji Zhang, Biwei Huang, Bernhard Scholkopf, and Clark Glymour. On the Identifiability and Estimation of Functional Causal Models in the Presence of Outcome-Dependent Selection. In *Uncertainty in Artificial Intelligence (UAI)*, 2016.

Yujia Zheng, Biwei Huang, Wei Chen, Joseph Ramsey, Mingming Gong, Ruichu Cai, Shohei Shimizu, Peter Spirtes, and Kun Zhang. Causal-learn: Causal discovery in python. *Journal of Machine Learning Research*, 25(60):1–8, 2024a.

Yujia Zheng, Zeyu Tang, Yiwen Qiu, Bernhard Schölkopf, and Kun Zhang. Detecting and Identifying Selection Structure in Sequential Data, June 2024b.

A Examples and Definitions

A.1 Definitions

Definition 2. (Markov Condition (Spirtes et al. [2000], Pearl [2009])) Given a DAG $\mathcal{G}$ and distribution P over the variable set $\mathbf{X}$, every variable X in $\mathbf{X}$ is probabilistically independent of its non-descendants given its parents in $\mathcal{G}$.

Definition 3. (Faithfulness Assumption (Spirtes et al. [2000], Pearl [2009])) Given a DAG $\mathcal{G}$ and distribution P over the variable set $\mathbf{X}$, there are no independencies between variables that are not entailed by the Markov Condition.

A.2 Illustrative Example: Linear Chain $X_1 \to X_2 \to X_3$

Example 2 (Number of Independent Terms in Chain)**.** Consider a chain $X_1 \to X_2 \to X_3$, where $X_1 = N_1$, $X_2 = aX_1 + N_2$, $X_3 = bX_2 + N_3$ with non-Gaussian noise.

The noise gradients are

$$\nabla N_1 = (1, 0, 0), \qquad \nabla N_2 = (-a, 1, 0), \qquad \nabla N_3 = (0, -b, 1).$$

Non-sink X_1 has $\operatorname{rank}(M_1^{\text{cond}}) = 2$: $\text{NB}_1 = \{X_2\}$. $H_1(x) = (s'_{n_1}(N_1) + a^2 s'_{n_2}(N_2),\ -a\, s'_{n_2}(N_2),\ 0)$. Conditioning on $X_2 = z_2$ gives $N_2 = z_2 - aN_1$:

$$H_1 \mid z_2 = \big(\underbrace{s'_{n_1}(N_1) + a^2 s'_{n_2}(z_2 - aN_1)}_{q_1(N_1)},\ \underbrace{-a\, s'_{n_2}(z_2 - aN_1)}_{q_2(N_1)},\ 0\big).$$

In general, q_1 and q_2 are distinct nonlinear functions of N_1, i.e. $q_1(N_1)$ is not an affine transformation of $q_2(N_1)$ under non-Gaussian noise assumption ($s'_{n_1}(t) \neq \text{const}$). Then the conditional matrix $M_1^{\text{cond}}(z_2)$ has the form

$$M_1^{\text{cond}}(z_2) = \begin{pmatrix} \operatorname{Var}(q_1) & \operatorname{Cov}(q_1, q_2) & 0 \\ \operatorname{Cov}(q_1, q_2) & \operatorname{Var}(q_2) & 0 \\ 0 & 0 & 0 \end{pmatrix}.$$

i.e. the components of $H_1|z_2$ span a 2-dimensional function space, thus $\operatorname{rank}(M_1^{\text{cond}}) = 2$.

Non-sink X_2 has $\operatorname{rank}(M_2^{\text{cond}}) = 2$: $\text{NB}_2 = \{X_1, X_3\}$. Conditioning on $(X_1, X_3) = (z_1, z_3)$ gives $N_3 = z_3 - b(az_1 + N_2) = c - wN_2$:

$$H_2 \mid (z_1, z_3) = s'_{n_2}(N_2) \cdot (-a, 1, 0) + b\, s'_{n_3}(c - wN_2) \cdot (0, b, -1).$$

Components involve $s'_{n_2}(N_2)$ and $s'_{n_3}(c - wN_2)$ which are generally independent functions. Following the same argument as for X_1, $M_2^{\text{cond}}(z_1, z_3)$ also has $\operatorname{rank}(M_2^{\text{cond}}) = 2$.

Sink X_3 has $\operatorname{rank}(M_3^{\text{cond}}) = 1$: $\text{NB}_3 = \{X_2\}$. $H_3(x) = s'_{n_3}(N_3) \cdot (0, -b, 1)$. Therefore:

$$M_3^{\text{cond}}(z_2) = \operatorname{Var}(s'_{n_3}) \begin{pmatrix} 0 & 0 & 0 \\ 0 & b^2 & -b \\ 0 & -b & 1 \end{pmatrix}.$$

$$\text{so, } \operatorname{rank}(M_3^{\text{cond}}) = 1$$

Therefore, the sink X_3 is uniquely identified as the only node with $\operatorname{rank}(M_i^{\text{cond}}) = 1, i \in \{1, 2, 3\}$.

B Proofs

B.1 Preservation of Skeleton Discovery

Proposition 8. *Let* $\mathbf{X}$ *be generated by a restricted ANM* (1) *with mutually independent noise terms and let* $S = \mathbb{1}(x \in \mathcal{B})$ *denote a deterministic selection for a truncation region* $\mathcal{B}$*. For any* $i, j \in [d]$*,*

$$\partial^2_{X_i X_j} \log p(x|\ S = 1) = 0 \iff X_i \perp\!\!\!\perp X_j \mid \mathbf{X}_{\setminus\{i,j\}} \text{ in } \mathcal{G}. \tag{13}$$

The proof of Proposition 8 directly follows of the preservation of score function under hard selection (truncation).

B.2 Proofs of Conditional Hessian Rank Characterization

Note that, throughout this subsection, for linear ANM, $X_i = \sum_{p \in \text{PA}_i} a_{ip} X_p + N_i$, i.e. $\mathbf{X} = B\mathbf{X} + \mathbf{N}$ with $B_{kp} = a_{kp}$, where $a_{kp} \neq 0$ if and only if $p \in \text{PA}_k$. Since $\mathcal{G}$ is acyclic, $I - B$ is lower-triangular (after permutation). We also assume $N_1, \ldots, N_d$ are mutually independent and non-Gaussian, with positive densities p_{N_k}, and we define $s_k := (\log p_{N_k})'$ and $s'_k := (\log p_{N_k})''$. For a strictly positive C^2 density, s'_k is constant if and only if $\log p_{N_k}$ is quadratic, i.e. if and only if N_k is Gaussian.

Remark 5 (Rank faithfulness)**.** *For a non-sink* X_i*, let* $\mathcal{N}_i = \{i\} \cup \text{CH}_i$*, and let* D_i *be the set of* $\beta \in \mathbb{R}^{|\,\text{CH}_i\,|}$ *for which the functions*

$$t \mapsto 1, \qquad t \mapsto s'_i(t), \qquad t \mapsto s'_k(\beta_k - a_{ki}t) \quad (k \in \text{CH}_i)$$

are linearly dependent on $\mathbb{R}$*. The model is* rank-faithful *at* X_i *if* D_i *is Lebesgue-null.*

Proposition 5 (Connection between Conditional Rank and Independent Noise Terms in Linear ANM)**.** *Under a linear non-Gaussian ANM, the rank of $M_i^{\mathrm{cond}}(z)$ equals the number of independent noise terms contributes to s_i by conditioning on X_{NB_i} for generic z[a]:*

$$\mathrm{rank}\big(M_i^{\mathrm{cond}}(z)\big) \;=\; |\mathcal{N}_i|,$$

where the noise index set $\mathcal{N}_i \;:=\; \{i\} \cup \mathrm{CH}_i$.

[a] By genericity, we exclude the degenerate cases where the graph structure is not faithfully reflected in the rank property, similar to the cases where standard faithfulness is violated. See Remark. 5

Proof. Let $\mathbf{X} = (X_1, \ldots, X_d)$ be generated by a linear additive noise model

$$X_i \;=\; \sum_{p \in \mathrm{PA}_i} a_{ip}\, X_p \;+\; N_i, \qquad i = 1, \ldots, d,$$

where $N_1, \ldots, N_d$ are mutually independent, non-Gaussian, with strictly positive smooth densities p_{N_k}. Let $s_k(n) := (\log p_{N_k})'(n)$ and $s_k'(n) := (\log p_{N_k})''(n)$, and assume $\mathrm{Var}(s_k'(N_k)) > 0$ for all k (equivalent to non-Gaussianity). Here, "generic" refers to rank faithfulness at X_i (Remark 5).

Compute Hessian row H_i. Differentiating the score function with respect to $\mathbf{x}$ gives $\nabla_{\mathbf{x}}^2 \log p(\mathbf{x}) = \sum_k s_k'(N_k)\, \mathbf{v}_k \mathbf{v}_k^\top$, whose i-th row is

$$H_i(\mathbf{x}) \;=\; s_i'(N_i)\, \mathbf{v}_i \;-\; \sum_{k \in \mathrm{CH}_i} a_{ki}\, s_k'(N_k)\, \mathbf{v}_k, \tag{14}$$

where we define $\mathbf{v}_k := \nabla_{\mathbf{x}} N_k = \mathbf{e}_k - \sum_{\ell \in \mathrm{PA}_k} a_{k\ell} \mathbf{e}_\ell$ (a fixed vector in $\mathbb{R}^d$ because of linearity). Thus H_i depends on exactly the noise indices $\{i\} \cup \mathrm{CH}_i$.

By rewriting the coefficients as $c_{ij} := 1$ if $j = i$, and $c_{ij} := -a_{ji}$ if $j \in \mathrm{CH}_i$, we have:

$$H_i(x) = \sum_{j \in \mathcal{N}_i} c_{ij} s_j'(N_j)\, v_j = \tilde{V}_i g_i(x),$$

where $\tilde{V}_i \;=\; [c_{ij} v_j]_{j \in \mathcal{N}_i} \;\in\; \mathbb{R}^{d \times |\mathcal{N}_i|}, g_i \;=\; \big(s_j'(N_j)\big)_{j \in \mathcal{N}_i}$ and $\tilde{V}_i$ is deterministic, constant in x, independent of $\mathcal{B}$.

So the target matrix can be expressed as:

$$M_i^{\mathrm{cond}}(z) = \mathrm{Var}\left(H_i \mid X_{\mathrm{NB}_i} = z\right) = \tilde{V}_i \Sigma_i(z) \tilde{V}_i^\top, \tag{15}$$

where $\Sigma_i(z) := \mathrm{Cov}\left(g_i \mid X_{\mathrm{NB}_i} = z\right) \in \mathbb{R}^{|\mathcal{N}_i| \times |\mathcal{N}_i|}$.

For the selected data, it is

$$M_{i,\mathrm{sel}}^{\mathrm{cond}}(z) = \tilde{V}_i \Sigma_i^{\mathrm{sel}}(z) \tilde{V}_i^\top, \tag{16}$$

and $\Sigma_i^{\mathrm{sel}}(z) := \mathrm{Cov}\left(g_i \mid X_{\mathrm{NB}_i} = z,\ S = 1\right) \in \mathbb{R}^{|\mathcal{N}_i| \times |\mathcal{N}_i|}$.

Note that in (16), selection acts only inside $\Sigma_i^{\mathrm{sel}}(z)$, and the eigenvalues of $M_{i,\mathrm{sel}}^{\mathrm{cond}}(z)$ do differ from the untruncated ones, but it does not influence the rank for generic z.

Next we prove that $\mathrm{rank}\, M_i^{\mathrm{cond}}(z) = \mathrm{rank}\, \Sigma_i(z)$ because $\tilde{V}_i$ has full column rank. The vectors $\mathbf{v}_j^\top$, $j \in \mathcal{N}_i$, are distinct rows of the invertible matrix $I - B$, so they are linearly independent. Moreover $c_{ij} \neq 0$ for all $j \in \mathcal{N}_i$, because $a_{ji} \neq 0$ for $j \in \mathrm{CH}_i$. Hence the columns $c_{ij}\mathbf{v}_j$ of $\tilde{V}_i$ are linearly independent, i.e. $\tilde{V}_i$ has full column rank $|\mathcal{N}_i|$. Writing $\Sigma_i(z) = LL^\top$, we get $\mathrm{rank}(\tilde{V}_i \Sigma_i(z) \tilde{V}_i^\top) = \mathrm{rank}(\tilde{V}_i L) = \mathrm{rank}\, L = \mathrm{rank}\, \Sigma_i(z)$. The same holds for $\Sigma_i^{\mathrm{sel}}(z)$.

□

Proposition 6 (Conditional Rank Test for Sink Identification (Linear))**.** *Let* $\mathbf{X}$ *be generated by a restricted* linear *ANM* (1) *with independent noise terms. Then for generic* z*:*

(a) ***Sink.*** *If* X_l *is a sink (i.e.,* $\mathrm{CH}_l = \emptyset$*), then* $\operatorname{rank}\big(M_l^{\mathrm{cond}}(z)\big) = 1$*.*

(b) ***Non-sink.*** *If* X_i *is not a sink (i.e.,* $|\,\mathrm{CH}_i\,| \geq 1$*), then* $\operatorname{rank}\big(M_i^{\mathrm{cond}}(z)\big) \geq 2$*.*

Therefore, a node X_i *is a sink if and only if* $\operatorname{rank}(M_i^{\mathrm{cond}}(z)) = 1$*.*

Proof. Assume the model is rank-faithful (Remark 5) at every non-sink. By Proposition 5, for generic z under the truncated distribution, $\operatorname{rank} M_i^{\mathrm{cond}}(z) = |\mathcal{N}_i| = 1 + |\,\mathrm{CH}_i\,|$. (a) If X_l is a sink, $\mathrm{CH}_l = \emptyset$, so $\operatorname{rank} M_l^{\mathrm{cond}}(z) = 1$. (b) If X_i is not a sink, $|\,\mathrm{CH}_i\,| \geq 1$, so $\operatorname{rank} M_i^{\mathrm{cond}}(z) \geq 2$. Hence X_i is a sink if and only if $\operatorname{rank} M_i^{\mathrm{cond}}(z) = 1$. □

Proposition 9 (Conditional Rank Test for Sink Identification (Nonlinear))**.** *Let* $\mathbf{X}$ *be generated by a restricted* nonlinear *ANM* (1) *with independent noise terms. Then:*

(a) ***Sink.*** *If* X_l *is a sink (i.e.,* $\mathrm{CH}_l = \emptyset$*), then:*

$$\operatorname{rank}\big(M_l^{\mathrm{cond}}\big) = \begin{cases} 1, & \text{if } N_i \text{ is non-Gaussian}, \forall i \\ 0, & \text{if } N_i \text{ is Gaussian}, \forall i \end{cases}. \tag{17}$$

(b) ***Non-sink.*** *If* X_i *is not a sink (i.e.,* $|\,\mathrm{CH}_i\,| \geq 1$*), then:*

$$\operatorname{rank}\big(M_i^{\mathrm{cond}}\big) \geq \begin{cases} 2, & \text{if } N_i \text{ is non-Gaussian}, \forall i \\ 1, & \text{if } N_i \text{ is Gaussian}, \forall i \end{cases} \tag{18}$$

Therefore, a node X_i *is a sink if and only if* $\operatorname{rank}(M_i^{\mathrm{cond}}) = 1$ *when the noise is non-Gaussian, and* $\operatorname{rank}(M_i^{\mathrm{cond}}) = 0$ *when the noise is Gaussian.*

Proof. **Part (a): Sink.** Let X_l be a sink, so $X_{\mathrm{CH}_l} = \emptyset$ and $X_{\mathrm{NB}_l} = X_{\mathrm{PA}_l}$. By (11), the Hessian row reduces to

$$H_l(x) = s'_{n_l}(N_l) \cdot \nabla N_l, \quad \text{where } \nabla N_l = e_l - \nabla_x f_l(x_{\mathrm{PA}_l}). \tag{19}$$

There are two terms: $s'_l(N_l)$ and ∇N_l in this equation. For ∇N_l, conditioning on $X_{\mathrm{NB}_l} = X_{\mathrm{PA}_l} = z$ fixes ∇N_l to a constant vector

$$c_l(z) := e_l - \nabla_x f_l(z),$$

where e_i is the i-th standard basis vector. The exogenous noise N_l remains free (not determined by the conditioning). Thus the only remaining random quantity is the scalar $s'_{n_l}(N_l)$, and

$$H_l \mid X_{\mathrm{PA}_l} \;=\; s'_{n_l}(N_l) \cdot c_l(z). \quad (X_{\mathrm{PA}_l} = z)$$

The conditional covariance is

$$M_l^{\mathrm{cond}}(z) = \operatorname{Var}\big(s'_{n_l}(N_l)\big) \cdot c_l(z)\, c_l(z)^\top.$$

When noise is non-Gaussian, s'_{n_l} is non-constant, so $\operatorname{Var}(s'_{n_l}(N_l)) > 0$, and $c_l(z) \neq 0$ (as its l-th entry is 1). Therefore $\operatorname{rank}(M_l^{\mathrm{cond}}(z)) = 1$.

When noise is Gaussian, $s'_{n_l}(N_l) = -1/\sigma^2$ is constant, so $\operatorname{Var}(s'_{n_l}(N_l)) = 0$ and $\operatorname{rank}(M_l^{\mathrm{cond}}(z)) = 0$.

Part (b): Non-sink. Let X_i be a non-sink with at least one child $X_k \in \mathrm{CH}_i$. We condition on $X_{\mathrm{NB}_i} = z$, which partitions as $z = (z_{\mathrm{PA}_i}, z_{\mathrm{CH}_i}, z_{\mathrm{SP}_i})$, where SP_i collects the other parents of the children of X_i.

Remind that the Hessian row for a non-sink i is computed as:

$$H_i(x) = s'_i(N_i) \cdot \nabla N_i \;-\; \sum_{k \in \mathrm{CH}_i} \Big[\underbrace{s'_k(N_k) \cdot \frac{\partial f_k}{\partial x_i} \cdot \nabla N_k}_{\text{child noise direction}} + \underbrace{s_k(N_k) \cdot \nabla\left(\frac{\partial f_k}{\partial x_i}\right)}_{\text{mechanism curvature}} \Big], \tag{20}$$

First term in (20): conditioning on $X_{\mathrm{PA}_i} = z_{\mathrm{PA}_i}$ fixes

$$\nabla N_i = e_i - \nabla_x f_i(z_{\mathrm{PA}_i}) =: c_i \quad \text{(constant vector)},$$

Child terms in (20): For child X_k, the structural equation reads $X_k = f_k(X_i, X_{\mathrm{PA}_k \setminus \{i\}}) + N_k$. Under the conditioning $X_k = z_k$ and $X_{\mathrm{PA}_k \setminus \{i\}} = z_{\mathrm{PA}_k \setminus \{i\}}$, the child noise becomes a function of N_i:

$$N_k = z_k - f_k\big(f_i(z_{\mathrm{PA}_i}) + N_i,\ z_{\mathrm{PA}_k \setminus \{i\}}\big) \coloneqq q_k(N_i;\, z). \tag{21}$$

Overall, substituting into (20) and using that all ∇N_k, $\partial f_k / \partial x_i$, and $\nabla(\partial f_k / \partial x_i)$ are deterministic functions of (N_i, z) (they depend on $x_i = f_i(z_{\mathrm{PA}_i}) + N_i$, which the conditioning does not fix), the conditional Hessian row has the form

$$H_i \mid X_{\mathrm{NB}_i} = z \;=\; s_i'(N_i) \cdot c_i - \sum_{k \in \mathrm{CH}_i} \underbrace{s_k'\big(q_k(N_i; z)\big) \cdot \alpha_k(N_i; z) \cdot d_k(N_i; z)}_{\text{child noise direction}} + \underbrace{R(N_i; z)}_{\text{mech curvature}}, \tag{22}$$

where $\alpha_k(N_i; z) \coloneqq \partial f_k / \partial x_i$, and $d_k(N_i; z) \coloneqq \nabla N_k$ are deterministic functions of N_i (when f_k) is linear, α_k is constant and $R(N_i; z)$ collects the mechanism curvature terms (also deterministic functions of N_i through q_k).

Finally, consider the number of independent functions. The conditional Hessian row (22) is a vector-valued function of the single scalar N_i. Its components include at least two functionally distinct terms:

- $s_i'(N_i)$, arising from the own-noise contribution, and
- $s_k'\big(q_k(N_i; z)\big) \cdot \alpha_k(N_i; z) \cdot d_k(N_i; z)$, arising from the k-th child noise.

For the conditional rank to equal 1 (degerate case), every component of $H_i|z$ must be an affine function of a single "base function" $\psi(N_i)$. Thus a rank-one degeneracy requires an exact identity coupling s_i', $s_k' \circ q_k$, $s_k \circ q_k$ and the derivatives of f_k.

For non-Gaussian noise, the noise score functions s_i' and s_k' are non-constant. The composition $s_k' \circ q_k$ introduces an additional nonlinearity through the mechanism f_k. For generic mechanism parameters and noise distributions, the affine relation fails on a set of full Lebesgue measure in the parameter and z-space under the restricted ANM assumption. Therefore, for Lebesgue-almost every z, the components of $H_i|z$ span at least a 2-dimensional function space, and $\mathrm{rank}(M_i^{\mathrm{cond}}(z)) \geq 2$.

For Gaussian noise, s_i' and s_k' are constant, but $H_i \mid z$ still varies with N_i through α_k, d_k and R, which is non-constant unless f_k is affine in x_i. Therefore, $\mathrm{rank}(M_i^{\mathrm{cond}}(z)) \geq 1$ for almost every z.

Under truncation, the same argument applies to the selected probability distribution (assuming that given z and $S = 1$, N_i has a density on a set of positive measure). □

Remark 6 (Neighbor VS. Markov Blanket)**.** *As in Algorithm 1,* NB_i *is the neighbourhood of* X_i *in the undirected graph recovered with Proposition 8. Because* $\partial^2_{X_i X_j} \log p \not\equiv 0$ *if and only if* $X_i \not\perp\!\!\!\perp X_j \mid \mathbf{X}_{\setminus\{i,j\}}$*, this graph is the moral graph of* $\mathcal{G}$*, so*

$$\mathrm{NB}_i = \mathrm{MB}_i := \mathrm{PA}_i \cup \mathrm{CH}_i \cup \mathrm{SP}_i, \qquad \mathrm{SP}_i := \textstyle\bigcup_{k \in \mathrm{CH}_i} \mathrm{PA}_k \setminus \{i\}.$$

For a sink X_l, $\mathrm{MB}_l = \mathrm{PA}_l = \mathrm{CH}_l$*, so all the arguments for sink nodes do not change.*

For a non-sink node, if we condition on $\mathrm{PA}_i \cup \mathrm{CH}_i$ *only, consider the cases when some* $k \in \mathrm{CH}_i$ *has a parent* $p \notin \mathrm{PA}_i \cup \mathrm{CH}_i \cup \{i\}$ *that is not conditioned on. This means that we leave more free variables in the conditional hessian term, therefore the rank arguments for part (b) also still hold.*

B.3 Proofs of Unconditional Hessian Rank Test

The following method introduces the rank test of unconditional Hessian outer product matrix M_i for sink identification, but no conditioning on the neighbors means that we are not blocking the contribution of noise terms from the upstream (parents) of X_i, thus the rank of M_i is determined by the number of parents and children of X_i instead of just the number of children. A particular challenge is to distinguish a sink node with m parents from a source node with the same number of m children, as in Example 3.

Example 3 (Unconditional Rank Test Failure)**.** Consider a chain $X_1 \to X_2 \to X_3$, where the unconditional rank of M_1 is 2, but the rank of M_3 is also 2. Therefore, we cannot distinguish the sink node X_3 from the non-sink node X_1 by just looking at the rank of M_i.

More generally, the following result (Proposition 10) characterizes the rank of M_i for sink vs. non-sink nodes under both linear and nonlinear ANM.

Proposition 10 (Unconditional Rank)**.** *Let* $\mathbf{X}$ *be generated by a restricted ANM* (1) *with mutually independent non-Gaussian noise terms (i.e.,* s_i' *is non-constant for each* i*).*

(a) ***sink:*** *If* X_l *is a sink with* $|\,\mathrm{PA}_l\,|$ *parents, then for linear mechanisms,* $\mathrm{rank}(M_l) = 1$*. For nonlinear mechanisms,* $\mathrm{rank}(M_l) \le |\,\mathrm{PA}_l\,| + 1$*.*

(b) ***Non-sink:*** *If* X_i *has* $|\,\mathrm{PA}_i\,|$ *parents and* $|\,\mathrm{CH}_i\,| \ge 1$ *children, then for linear mechanisms,* $\mathrm{rank}(M_i) \ge 2$*. For nonlinear mechanisms,* $\mathrm{rank}(M_i) \le |\,\mathrm{PA}_i\,| + |\,\mathrm{CH}_i\,| + 1$ *and generically* $\mathrm{rank}(M_i) > |\,\mathrm{PA}_i\,| + 1$*.*

If X_l *is a sink node, it is identifiable if and only if for every non-sink node* X_i*,* $|\,\mathrm{PA}_i\,| + |\,\mathrm{CH}_i\,| - |\,\mathrm{PA}_l\,| \ge 1$*, (i.e. there exists a rank gap between a sink node and all non-sink nodes).*

Proof sketch. For a sink X_l: the decomposition (11) reduces to $H_l(x) = s_l'(N_l) \cdot \nabla N_l$. If f_l is linear, then ∇N_l is a constant vector c_l, and every $H_l(x)$ is parallel to c_l. Thus $M_l = \mathbb{E}[s_l'^2] \cdot c_l c_l^\top$, which has rank 1. If f_l is nonlinear, ∇N_l varies with x^{PA_l} and spans at most $m+1$ dimensions (the entries corresponding to $\{l\} \cup \mathrm{PA}_l$).

For a non-sink X_i with child X_k: H_i contains an additional term with coefficient $s_k'(N_k)$ (independent of $s_i'(N_i)$ since $N_i \perp\!\!\!\perp N_k$) multiplying the direction ∇N_k. Since ∇N_k has a non-zero entry at position $k \notin \{i\} \cup \mathrm{PA}_i$, it is linearly independent of ∇N_i. The two independent random coefficients times two linearly independent directions produce a matrix of rank ≥ 2 (linear case), or of rank at least $\mathrm{rank}(M_l) + 1$ (nonlinear case).

Under non-Gaussianity, s_i' is non-constant, preventing the random coefficients from collapsing to scalars (which would happen for Gaussian noise where $s'(N) = -1/\sigma^2$ is constant).

Under Gaussianity but with non-linear mechanisms, s_i' is constant, but the mechanism curvature term $\nabla\left(\frac{\partial f_k}{\partial x_j}\right)$ is generally non-zero and linearly independent of ∇N_i, so the rank gap still exists. □

Table 5: Rank of M_j for sink vs. non-sink nodes across different ANM settings.

Setting	**sink rank**	**Non-sink rank**	**Gap**
Linear + Gaussian	1	1	0 (not identifiable)
Linear + non-Gaussian	1	$\ge 1 + \lvert\,\mathrm{CH}_l\,\rvert$	$\lvert\,\mathrm{CH}_j\,\rvert \ge 1$
Nonlinear + Gaussian	$\le \lvert\,\mathrm{PA}_l\,\rvert + 1$	$\le \lvert\,\mathrm{PA}_l\,\rvert + \lvert\,\mathrm{CH}_l\,\rvert + 1$	$\lvert\,\mathrm{PA}_j\,\rvert + \lvert\,\mathrm{CH}_j\,\rvert - \lvert\,\mathrm{PA}_l\,\rvert$
Nonlinear + non-Gaussian	$\le \lvert\,\mathrm{PA}_l\,\rvert + 1$	$\le \lvert\,\mathrm{PA}_l\,\rvert + \lvert\,\mathrm{CH}_l\,\rvert + 1$	$\lvert\,\mathrm{PA}_j\,\rvert + \lvert\,\mathrm{CH}_j\,\rvert - \lvert\,\mathrm{PA}_l\,\rvert$

Remark 7 (Comparing Conditional and Unconditional Rank Test)**.** *The unconditional rank test in Proposition 10 imposes stronger conditions on the graph structure, since it requires that* $|\,\mathrm{PA}_j\,| + |\,\mathrm{CH}_j\,| - |\,\mathrm{PA}_l\,| \ge 1$ *for all non-sink nodes* X_i *for sink-node identification. A tree-like structure (where a non-sink node has more than one child and a sink node has at most one parent) satisfies this condition, but more general graphs may not. In contrast, the conditional rank test in Proposition 6 and 9 does not require such a degree constraint, since conditioning on the neighbors blocks the contribution of the parents, thus only counting the number of noise terms that directly feed into the score of* X_i*.*

B.4 Proofs of Consistent Estimator of $M_i^{\mathbf{cond}}$

Conditional Variance Matrix Estimation In Algorithm 2, we directly estimate $M_i^{\mathrm{cond}}(z)$ via a local kernel-weighted covariance estimator. In practice, when the ANM is linear non-Gaussian or

nonlinear Gaussian, we find that the residual-based estimator is more robust to estimation errors in the Hessian, as a direct estimator of M_i^{cond} relies on selecting the number of conditional neighbors x_{NB_i} and the kernel bandwidth, thus suffers more from estimation errors.

Specifically, we define the test statistic as $\widehat{M}_i^{\text{cond}} := \frac{1}{n}\sum_{t=1}^{n} r_i(x^{(t)})r_i(x^{(t)})^\top$ as a proxy, where $r_i(x) := H_i(x) - \hat{m}_i(x_{\text{NB}_i})$ is the residual of using $\hat{m}_i(\cdot)$ to estimate H_i, i.e. $\hat{m}_i(\cdot)$ is a regressor of H_i on X_{NB_i} approximating $m_i := \mathbb{E}[H_i \mid X_{\text{NB}_i}]$.

Proposition 11 (Consistent Estimator of M_i^{cond})**.** *Under a linear non-Gaussian or nonlinear Gaussian ANM, $\widehat{M}_i^{\text{cond}}$ is a consistent estimator of M_i^{cond} if $\hat{m}_i$ is a consistent estimator of m_i.*

Proof. The proof of Proposition 11 relies on the law of total variance: $\mathbb{E}M_i^{\text{cond}} = \text{Var}(H_i) - \text{Var}(m_i(X_{\text{NB}_i}))$. We first show that $M_i^{\text{cond}} = \mathbb{E}[r_i r_i^\top]$, and then by the continuous mapping theorem, If $\hat{m}_i \to m_i$ in L^2 as $n \to \infty$, then $\widehat{M}_i^{\text{cond}} \to M_i^{\text{cond}}$.

$$\begin{aligned}\mathbb{E}[r_i r_i^\top] &= \mathbb{E}(H_i(x) - \hat{m}_i)(H_i(x) - \hat{m}_i)^\top \\ &= \mathbb{E}[H_i H_i^\top] - \mathbb{E}[\hat{m}_i H_i^\top] - \mathbb{E}[H_i \hat{m}_i^\top] + \mathbb{E}\hat{m}_i\hat{m}_i^\top\end{aligned}$$

By the towering property, we can expand $H_i = \mathbb{E}[H_i \mid X_{\text{NB}_i}] = m_i$, so

$$\begin{aligned}\mathbb{E}[r_i r_i^\top] &= \mathbb{E}[H_i H_i^\top] - \mathbb{E}[\mathbb{E}[H_i \mid X_{\text{NB}_i}]H_i^\top] - \mathbb{E}[H_i \mathbb{E}[H_i \mid X_{\text{NB}_i}]^\top] + \mathbb{E}\hat{m}_i\hat{m}_i^\top \\ &= \mathbb{E}[H_i H_i^\top] - \mathbb{E}[\hat{m}_i\hat{m}_i^\top] - \mathbb{E}[\hat{m}_i\hat{m}_i^\top] + \mathbb{E}[\hat{m}_i\hat{m}_i^\top] \\ &= \mathbb{E}[H_i H_i^\top] - \mathbb{E}[\hat{m}_i\hat{m}_i^\top].\end{aligned}$$

On the other hand,

$$\begin{aligned}\mathbb{E}M_i^{\text{cond}} &= \text{Var}(H_i) - \text{Var}(m_i(X_{\text{NB}_i})) \\ &= \mathbb{E}[H_i H_i^\top] - \mathbb{E}[\hat{m}_i\hat{m}_i^\top] - \mathbb{E}[\hat{m}_i\hat{m}_i^\top] + \mathbb{E}[\hat{m}_i\hat{m}_i^\top] \\ &= \mathbb{E}[H_i H_i^\top] - \mathbb{E}[\hat{m}_i\hat{m}_i^\top] \\ &= \mathbb{E}[r_i r_i^\top]\end{aligned}$$

When the ANM is linear non-Gaussian or nonlinear Gaussian, the local and averaged conditional matrix coincide, i.e. $M_i^{\text{cond}}(z) = \mathbb{E}_z M_i^{\text{cond}}$ for generic z. As is in the proof of Proposition 10, for a sink node l, the geometrical factorization of the target matrix $M_l^{\text{cond}}(z) = \text{Var}\left(s'_{n_l}(N_l)\right) \cdot c_l(z)c_l(z)^\top$, where $c_l(z) = e_k - \nabla_x f_l(z)$. In the linear case, $c_l(z)$ is a constant vector that does not rotate with z. Therefore the aggregation $\mathbb{E}_z M_l^{\text{cond}}(z)$ is an aggregation of rank 1 matrices that has the same eigenvector, therefore its value is preserved after pooling for linear ANM. In the nonlinear-Gaussian case, $\text{Var}\left(s'_{n_l}\right) = 0$ thus gives 0 rank of the matrix, and the rotation of $c_l(z)$ also does not matter.

□

C Related Works

Causal Discovery and Inference under Selection Bias There are limited works on causal discovery under selection bias, and they either focus on local causal structure learning [Versteeg et al., 2022], or assuming access to interventional data [Dai et al., 2025a, Luo et al., 2025], or assuming sequential data [Zheng et al., 2024b], or performing discovery at latent variable level [Dai et al., 2025b] with linear Gaussian assumption. In contrast, we focus on learning the full causal structure among observed variables under selection bias, without requiring interventional or structural assumptions on the selection mechanism, and allowing for nonlinear and non-Gaussian mechanisms.

Causal Inference under Selection Bias Selection bias is a well-known problem in causal inference, and there are various methods to address it, such as [Bareinboim et al., 2014, Bareinboim and Pearl, 2012, Bareinboim and Tian, 2015, Correa et al., 2019], which characterizes conditions under which causal effects remain identifiable despite selection bias. Subsequent works extended this approach by developing testable implications of selection mechanisms [Correa et al., 2019] and providing

adjustment criteria. Additionally, Van der Zander et al. [2014], Jaber et al. [2019] developed causal effect identification results under Markov equivalence classes (MEC) which takes into account the potential presence of latent selections. Selection bias has been studied extensively in economics Heckman [1977] and epidemiology Robins et al. [2000]. Selection bias has been studied extensively in economics Heckman [1977] and epidemiology Robins et al. [2000]. However, these works primarily focus on *correcting* the bias rather than causal structure learning, which is the focus of our work.

D Algorithm

D.1 Main Algorithm

The overall algorithm for causal discovery via the Conditional Hessian Rank Test (CHess) is outlined in Algorithm 1.

Algorithm 1 Causal discovery via Conditional HessianRank Test (CHess)

Require: Data $\{x^{(i)}\}_{i=1}^{n}$ (potentially from truncated distribution $P(\mathbf{X} \mid S = 1)$)
Ensure: Causal DAG $\mathcal{G}$

Estimate Hessian: Compute $\nabla^2 \log p(x^{(i)})$ at each data point (e.g., via Stein estimator or kernel score matching on the truncated data).
$\mathbf{X} \leftarrow \{X_1, \ldots, X_d\}$
while $\mathbf{X} \neq \emptyset$ **do**
    MEC recovery: For all pairs $X_i, X_j \in \mathbf{X}$, test $\frac{\partial^2 \log p|x}{\partial X_i \partial X_j} = 0$ for conditional independence [Montagna et al., 2023]. Recover skeleton via Proposition 8.
    Sink identification: For each $X_j \in \mathbf{X}$, find its neighbor NB_j and estimate its conditional hessian matrix rank r_j via Algorithm 2.
    $X_l \leftarrow \arg\min_j r_j$ ▷ Smallest estimated rank = sink
    Orient edges: Orient all edges adjacent to X_l as incoming.
    $\mathbf{X} \leftarrow \mathbf{X} \setminus \{X_l\}$; update Hessian estimates for the reduced variable set.
    DAS pruning: a final pruning step that is robust to hard selection [Montagna et al., 2023].
end while
return $\mathcal{G}$

D.2 Algorithm for Conditional Hessian Matrix Estimation

The algorithms used for estimating M_i^{cond} and sink node identification. We present two alternative algorithms for estimating the conditional Hessian matrix rank: (1) Algorithm 2 is a conditional rank estimator based on local averaging, used for nonlinear non-Gaussian ANMs and (2) Algorithm 3 is a residual rank estimator based on regression, used for linear non-Gaussian ANMs and nonlinear Gaussian ANMs.

E More Details on the Experiments

E.1 Datasets

Synthetic Dataset The binary adjacency $\mathbf{B}_{\text{bin}}$ is sampled by drawing $\lfloor d \cdot \text{degree}/2 \rfloor$ edges uniformly over all possible indices. For **linear SEM**, we generate the weights $\mathbf{B}_{ij}$ uniformly from $[-0.8, -0.2] \cup [0.2, 0.8]$, and per-node noise variances are drawn i.i.d. from $\mathcal{U}(1, 3)$. For each configuration we sample $n = 1000$ observations and average over 20 seeds. And the data are generated as: $X_i = \mathbf{B}_{\cdot i}^{\top} \mathbf{X} + \epsilon_i$, with ϵ_i drawn from a Gumbel or log-normal distribution. For **nonlinear** SEM, we use a one-layer MLP with 64 hidden units and a sigmoid activation: $f_i(\mathbf{x}) = \sigma(\mathbf{x} W_1) W_2$, where the entries of $W_1 \in \mathbb{R}^{|\mathrm{Pa}(X_i)| \times 64}$ and $W_2 \in \mathbb{R}^{64}$ are drawn i.i.d. from $\mathcal{U}(0.5, 2)$ with each entry's sign flipped independently with probability $1/2$.

Sachs [Sachs et al., 2005] Dataset The Sachs data [Sachs et al., 2005] is a single-cell biological dataset in order to estimate protein interactions. It consists of nine files, each with one interventional

Algorithm 2 Conditional Rank Estimation: Kernel-Weighted Local Covariance

Require: Data $\{x^{(i)}\}_{i=1}^n$; neighborhood sets $\{\mathrm{NB}_j\}_{j=1}^d$ from MEC recovery; number of reference points L; bandwidth h.
Ensure: Identified sink node X_l.

for each node $X_j \in V$ **do**
  Select L reference points $\{z^{(\ell)}\}_{\ell=1}^L$ in the support of NB_j ▷ e.g., K-means centroids
  for $\ell = 1, \dots, L$ **do**
    Compute weights: $w_i^{(\ell)} \leftarrow K_h(x_{\mathrm{NB}_j}^{(i)} - z^{(\ell)}), \quad K_h(u) = \exp\!\Big(-\frac{\|u\|^2}{2h^2}\Big), \quad i = 1, \dots, n$
    Compute $\bar{H}_j^{(l)} = \frac{\sum_i w_i^{(l)} H_j(x^{(i)})}{\sum_i w_i^{(l)}}$.
    Compute $M_j^{\mathrm{cond}}(z^{(l)}) = \frac{\sum_i w_i^{(l)} \left(H_j(x^{(i)}) - \bar{H}_j^{(l)}\right)\left(H_j(x^{(i)}) - \bar{H}_j^{(l)}\right)^\top}{\sum_i w_i^{(l)}}$.
    Compute eigenvalues $\hat{\lambda}_1^{(\ell)} \geq \hat{\lambda}_2^{(\ell)} \geq \cdots \geq \hat{\lambda}_d^{(\ell)}$
    $r_j^{(\ell)} \leftarrow (\sum_k \hat{\lambda}_k^{(\ell)})^2 / \sum_k (\hat{\lambda}_k^{(\ell)})^2$ ▷ effective rank at $z^{(\ell)}$
  end for
  $\hat{r}_j \leftarrow \mathrm{median}(r_j^{(1)}, \dots, r_j^{(L)})$ ▷ aggregate over reference points
end for
$X_l \leftarrow \arg\min_j \hat{r}_j$ ▷ smallest conditional effective rank = sink
return X_l

Algorithm 3 Residual Rank Estimator

Require: Data $\{x^{(i)}\}_{i=1}^n$; neighborhood sets $\{\mathrm{NB}_j\}_{j=1}^d$ from MEC recovery
Ensure: Identified sink node X_l

for each candidate node X_j **do**
  for $i = 1, \dots, n$ **do**
    Compute $H_j(x^{(i)}) \leftarrow \nabla_x \, \partial_{x_j} \log \hat{p}(x^{(i)}) \in \mathbb{R}^d$ ▷ estimate Hessian rows
  end for
  Fit $\hat{m}_j \leftarrow \arg\min_f \frac{1}{n} \sum_{i=1}^n \left\| H_j(x^{(i)}) - f(x_{\mathrm{NB}_j}^{(i)}) \right\|^2$ ▷ estimate conditional mean via regression
  for $i = 1, \dots, n$ **do**
    $e_j^{(i)} \leftarrow H_j(x^{(i)}) - \hat{m}_j(x_{\mathrm{NB}_j}^{(i)}) \in \mathbb{R}^d$ ▷ compute residuals
  end for
  $\widehat{M}_j^{\mathrm{cond}} \leftarrow \frac{1}{n} \sum_{i=1}^n e_j^{(i)} \left(e_j^{(i)}\right)^\top \in \mathbb{R}^{d \times d}$ ▷ form conditional outer-product matrix
  Compute eigenvalues $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \cdots \geq \hat{\lambda}_d \geq 0$ of $\widehat{M}_j^{\mathrm{cond}}$
  $\widehat{r}_j \leftarrow \mathrm{Rank}\left(\widehat{M}_j^{\mathrm{cond}}\right)$ ▷ effective rank
end for
return $X_l \leftarrow \arg\min_j \widehat{r}_j$

context and the same set of variables $raf, mek, plc, pip2, pip3, erk, akt, pka, pkc, p38, jnk$, representing the measurements of different biochemical concentrations. There are between 700 and 900 data points in each file, for a total of 7466 data points. The data is continuous and very skewed, shifting far away from our assumptions of Gaussian distributions. Following Ramsey and Andrews [2018], we apply log transformation to the data.

DREAM4 Dataset We consider the in-silico Size-10 sub-challenge of DREAM4[4] [Pinna et al., 2010, Shah et al., 2023], which provides five ground-truth gene regulatory networks of $d = 10$ genes each.

[4] https://gnw.sourceforge.net/resources/DREAM4%20in%20silico%20challenge.pdf

E.2 Estimation Method

Diffusion model-based score estimation [Sanchez et al., 2023] sidesteps the boundary problem by learning the score $\nabla_x \log p_\sigma(x)$ via denoising score matching with noise level σ. We use diffusion probabilistic models (DPMs) to learn to reverse a diffusion process, by starting with noise and recovery data distribution to be estimated. A DPM is trained with the denoising score matching objective:

$$\theta^* = \arg\min_\theta \mathbb{E}_{x_0,t,\epsilon}\left[\lambda(t)\|\epsilon_\theta(x_t,t) - \epsilon\|_2^2\right]$$

where $x_t = \sqrt{\alpha_t}x_0 + \sqrt{1-\alpha_t}\epsilon$, with $\epsilon \sim \mathcal{N}(0, I)$. The key result [Song and Ermon, 2019] is that the trained noise predictor ϵ_θ approximates the score of the data distribution: $\epsilon_\theta(x,t) \approx \nabla_x \log p(x)$. Once ϵ_θ is trained, the (i,j)-th entry of the Hessian is obtained by a single backpropagation pass through the network: $H_{i,j}(\log p(\mathbf{x})) \approx \nabla_{i,j}\epsilon_\theta(\mathbf{x},t)$, i.e. differentiate the i-th output of ϵ_θ with respect to the j-th input.

E.3 Evaluation metrics

The output of PC [Spirtes et al., 2000] and GES [Chickering, 2003] are complete partially directed acyclic grap (CPDAGs), while the output of FCI [Spirtes et al., 2013] is a partial acyclic graph (PAG). We convert the output of the baseline algorithms according to Table. 6.

True	Estimated	Adjacency	Arrowhead
$a \leftarrow b$	$a \leftarrow b$	TP	TP, TN
	$a \rightarrow b$	TP	FP, FN
	a — b	TP	FN
	$a \quad b$	FN	FN
$a \quad b$	$a \leftarrow b$	FP	FP
	$a \rightarrow b$	FP	FP
	a — b	FP	\
	$a \quad b$	TN	\

Table 6: Metrics used for evaluation. Precision=TP/(TP+FP), Recall=TP/(TP+FN).

E.4 Hyperparameters

We have the following hyperparameters for our method, which are set to the same values across all experiments unless otherwise specified:

1. MEC threshold: relative off-diagonal cutoff for the MEC skeleton recovery step. An off-diagonal entry of the RMS Hessian matrix is treated as a skeleton edge if it exceeds MEC threshold. Set to 0.1 in all experiments.
2. Rank test: method for estimating the numerical rank of M_j^{cond}. `chi2` uses an eigenvalue-based χ^2 test (EigRankTest) at significance level. Set to `chi2` in all scripts.
3. Rank test significance level: α for the χ^2 rank test above. Set to 0.05 in all scripts.
4. KRR ridge: the ridge regularization parameter λ for the kernel ridge regressor used by residual estimator of M_j^{cond}. Controls the bias–variance tradeoff of the residual score estimate. Set to 10^{-5} in all scripts.
5. CAM pruning threshold: p-value cutoff for the final CAM-style edge pruning step. Each candidate edge is tested by a GAM regression; edges whose coefficient p-value exceeds this threshold are removed. Set to 0.001 for DREAM4, 0.01 for Sachs.

E.5 Baseline Implementations

For PC, GES, FCI, GRaSP, LiNGAM, we use the implementations in the `causallearn` Python package [Zheng et al., 2024a] (`https://github.com/py-why/causal-learn`), with the default hyperparameters. For CAM, SCORE, DAS, and NoGAM, we use the implementations in `dodiscover` (`https://www.pywhy.org/dodiscover/dev/index.html`), with the default hyperparameters.

F More Experimental Results

F.1 All Baselines in Linear non-Gaussian ANM

Results. When hard selection bisa is not present, LiNGAM-based methods achieves the best performance, which is not a surprising result. However, when selection bias becomes severe, our conditional Hessian rank-based method starts to outperform. We also underlines the second-best performance.

Table 7: $d = 5$, noise=lognormal: F1 Metrics

Method	F1 (Adj) ↑			F1 (Arrow) ↑		
	$K = 0$	$K = 1$	$K = 2$	$K = 0$	$K = 1$	$K = 2$
PC	**0.96**(0.00)	0.76(0.02)	0.70(0.02)	\	\	\
FCI	**0.96**(0.00)	0.76(0.02)	0.70(0.02)	\	\	\
Exact	0.70(0.02)	0.65(0.02)	0.66(0.01)	0.29(0.04)	0.31(0.02)	0.34(0.06)
GES	0.96(0.00)	0.78(0.02)	0.72(0.02)	\	\	\
GRaSP	0.89(0.01)	**0.79**(0.03)	0.71(0.02)	\	\	\
ICA-LiNGAM	0.96(0.00)	0.74(0.02)	0.72(0.01)	**0.95**(0.01)	0.59(0.07)	0.50(0.08)
Direct-LiNGAM	0.96(0.00)	**0.79**(0.02)	0.74(0.01)	**0.95**(0.01)	**0.76**(0.02)	0.68(0.04)
CAM	0.78(0.01)	0.71(0.01)	0.72(0.01)	0.12(0.03)	0.07(0.01)	0.13(0.02)
SCORE	0.91(0.01)	0.71(0.01)	0.68(0.02)	0.89(0.01)	0.53(0.04)	0.47(0.06)
DAS	0.91(0.01)	0.71(0.01)	0.68(0.02)	0.89(0.01)	0.53(0.04)	0.47(0.06)
NoGAM	0.91(0.01)	0.74(0.01)	0.69(0.02)	0.83(0.03)	0.54(0.05)	0.49(0.05)
DiffAN	0.79(0.01)	0.74(0.01)	0.71(0.01)	0.26(0.05)	0.30(0.03)	0.23(0.05)
CHess (Stein)	0.94(0.01)	0.76(0.01)	0.73(0.01)	0.91(0.02)	0.70(0.03)	**0.70**(0.02)
CHess (DiffAN)	0.93(0.01)	0.76(0.01)	**0.74**(0.02)	0.88(0.03)	0.72(0.02)	**0.70**(0.03)

F.2 All Baselines in MLP non-Gaussian ANM

Results. When hard selection bias is not present, the performance of all methods are similar (F1 around $0.7 - 0.8$). However, when selection bias becomes severe, our conditional Hessian rank-based method starts to outperform while other methods' performance drops significantly.

Table 8: $d = 5$, noise=lognormal: SHD Metrics

Method	SHD ↓ $K = 0$	$K = 1$	$K = 2$
PC	2.81(1.85)	4.81(3.26)	5.11(2.25)
FCI	4.48(0.92)	5.96(0.85)	6.33(2.07)
Exact	6.11(2.84)	6.33(2.59)	6.22(4.02)
GES	2.78(1.58)	4.63(3.79)	5.70(3.54)
GRaSP	3.44(1.80)	4.00(2.74)	5.15(2.13)
ICA-LiNGAM	**0.41**(0.39)	4.00(4.59)	5.04(7.07)
Direct-LiNGAM	**0.41**(0.39)	**2.63**(2.16)	3.63(3.20)
CAM	7.15(3.83)	8.30(1.34)	7.43(1.65)
SCORE	1.11(1.14)	4.43(3.25)	5.03(5.37)
DAS	1.11(1.14)	4.43(3.25)	5.03(5.37)
NoGAM	1.44(2.32)	4.33(4.22)	5.03(3.97)
DiffAN	6.04(4.63)	6.26(3.08)	6.78(3.95)
CHess (Stein)	<u>0.81</u>(1.19)	<u>3.17</u>(2.34)	<u>3.53</u>(1.92)
CHess (DiffAN)	1.11(2.32)	<u>3.03</u>(2.10)	**3.47**(2.85)

Table 9: $d = 10$, noise=lognormal: F1 Metrics

Method	F1 (Adj) ↑ $K = 0$	$K = 1$	$K = 2$	F1 (Arrow) ↑ $K = 0$	$K = 1$	$K = 2$
PC	0.90(0.01)	0.76(0.01)	0.73(0.01)	0.59(0.04)	\	\
FCI	0.90(0.01)	0.76(0.01)	0.73(0.01)	0.55(0.03)	\	\
Exact	0.53(0.00)	0.50(0.00)	0.46(0.00)	0.23(0.01)	0.25(0.01)	0.21(0.01)
GES	0.89(0.00)	0.70(0.01)	0.64(0.01)	0.58(0.04)	0.49(0.02)	0.44(0.02)
GRaSP	0.88(0.01)	**0.80**(0.01)	**0.75**(0.02)	0.55(0.05)	\	\
ICA-LiNGAM	<u>0.93</u>(0.00)	0.66(0.01)	0.58(0.01)	<u>0.92</u>(0.00)	0.51(0.05)	0.40(0.04)
Direct-LiNGAM	**0.93**(0.00)	0.72(0.01)	0.62(0.01)	**0.93**(0.00)	**0.71**(0.01)	**0.60**(0.02)
CAM	0.72(0.01)	0.61(0.00)	0.57(0.01)	0.09(0.01)	0.09(0.01)	0.09(0.01)
SCORE	0.80(0.01)	0.65(0.01)	0.58(0.01)	0.75(0.03)	0.52(0.02)	0.40(0.03)
DAS	0.82(0.01)	0.66(0.01)	0.60(0.02)	<u>0.77</u>(0.02)	0.56(0.02)	0.46(0.03)
NoGAM	0.81(0.01)	0.63(0.01)	0.58(0.01)	0.67(0.04)	0.46(0.03)	0.42(0.03)
DiffAN	0.70(0.01)	0.61(0.01)	0.57(0.01)	0.22(0.02)	0.21(0.02)	0.17(0.01)
CHess (Stein)	0.80(0.01)	0.65(0.01)	0.58(0.01)	0.68(0.03)	<u>0.57</u>(0.02)	0.48(0.02)
CHess (DiffAN)	0.81(0.01)	0.65(0.01)	0.58(0.01)	0.70(0.02)	<u>0.58</u>(0.02)	<u>0.51</u>(0.01)

Table 10: $d = 10$, noise=lognormal: SHD Metrics

Method	SHD ↓ $K = 0$	$K = 1$	$K = 2$
PC	5.89(4.84)	<u>7.78</u>(6.10)	<u>9.85</u>(7.02)
FCI	9.74(1.97)	11.85(3.61)	12.85(5.31)
Exact	20.56(6.17)	21.74(11.16)	24.56(6.32)
GES	6.11(4.40)	10.33(7.19)	13.33(20.44)
GRaSP	5.96(5.07)	**7.22**(6.84)	**9.15**(15.16)
ICA-LiNGAM	<u>1.63</u>(1.94)	12.07(46.59)	16.37(44.60)
Direct-LiNGAM	**1.48**(1.66)	**7.22**(11.58)	12.15(31.31)
CAM	16.67(11.04)	20.40(8.31)	23.23(15.98)
SCORE	5.07(13.25)	10.67(9.29)	15.67(29.29)
DAS	4.22(6.47)	9.07(8.93)	12.13(15.72)
NoGAM	5.96(13.52)	11.93(15.06)	15.50(30.85)
DiffAN	15.56(19.36)	18.70(26.36)	21.12(21.49)
CHess (Stein)	5.93(13.48)	10.40(16.37)	14.40(18.04)
CHess (DiffAN)	5.52(9.81)	10.00(12.13)	13.93(18.20)

Table 11: $d = 5$, noise=gumbel: F1 Metrics

Method	F1 (Adj) ↑ $K = 0$	$K = 1$	$K = 2$	F1 (Arrow) ↑ $K = 0$	$K = 1$	$K = 2$
PC	<u>0.98</u>(0.00)	0.75(0.01)	0.67(0.01)	\	0.41(0.05)	\
FCI	<u>0.98</u>(0.00)	0.75(0.01)	0.66(0.02)	\	0.41(0.05)	\
Exact	0.76(0.01)	0.71(0.01)	0.68(0.01)	0.30(0.03)	0.31(0.03)	0.36(0.03)
GES	0.96(0.00)	0.71(0.01)	0.65(0.01)	\	\	\
GRaSP	0.95(0.00)	**0.76**(0.01)	0.66(0.02)	\	\	\
ICA-LiNGAM	**1.00**(0.00)	0.72(0.01)	0.67(0.00)	<u>0.99</u>(0.00)	0.41(0.05)	0.38(0.04)
Direct-LiNGAM	**1.00**(0.00)	0.71(0.01)	0.67(0.01)	**1.00**(0.00)	0.67(0.01)	0.51(0.03)
CAM	0.88(0.01)	0.72(0.01)	0.68(0.01)	0.15(0.02)	0.17(0.03)	0.23(0.02)
SCORE	0.90(0.01)	0.73(0.01)	0.69(0.01)	0.62(0.03)	0.41(0.04)	0.38(0.03)
DAS	0.90(0.01)	0.73(0.01)	0.69(0.01)	0.62(0.03)	0.41(0.04)	0.38(0.03)
NoGAM	0.94(0.01)	0.74(0.01)	<u>0.70</u>(0.01)	0.78(0.05)	0.49(0.03)	0.41(0.04)
DiffAN	0.87(0.01)	0.75(0.01)	**0.70**(0.01)	0.43(0.07)	0.30(0.04)	0.28(0.03)
CHess (Stein)	0.97(0.00)	<u>0.75</u>(0.01)	**0.70**(0.01)	0.94(0.02)	**0.69**(0.03)	**0.65**(0.02)
CHess (DiffAN)	0.95(0.01)	0.75(0.01)	**0.70**(0.01)	0.86(0.05)	<u>0.69</u>(0.02)	<u>0.64</u>(0.01)

Table 12: $d = 5$, noise=gumbel: SHD Metrics

Method	SHD ↓ $K = 0$	$K = 1$	$K = 2$
PC	2.37(1.94)	5.00(3.11)	5.93(1.55)
FCI	4.19(1.04)	6.63(1.64)	7.19(1.04)
Exact	5.74(1.75)	5.81(2.00)	5.81(1.85)
GES	2.59(2.32)	5.89(1.73)	6.33(1.85)
GRaSP	2.78(2.10)	4.85(2.50)	6.04(2.85)
ICA-LiNGAM	<u>0.07</u>(0.07)	5.70(3.39)	6.30(2.65)
Direct-LiNGAM	**0.00**(0.00)	4.00(1.48)	5.37(1.86)
CAM	5.52(2.03)	6.77(2.45)	7.00(2.33)
SCORE	2.70(2.13)	5.30(2.54)	6.07(2.60)
DAS	2.70(2.13)	5.30(2.54)	6.07(2.60)
NoGAM	1.59(3.13)	4.83(2.81)	5.73(2.93)
DiffAN	4.15(4.94)	5.78(2.62)	6.56(2.91)
CHess (Stein)	0.52(0.99)	**3.43**(2.58)	**4.10**(1.09)
CHess (DiffAN)	1.11(3.14)	<u>0.35</u>(2.38)	**4.10**(1.22)

Table 13: $d = 10$, noise=gumbel: F1 Metrics

Method	F1 (Adj) ↑ $K = 0$	$K = 1$	$K = 2$	F1 (Arrow) ↑ $K = 0$	$K = 1$	$K = 2$
PC	0.96(0.00)	0.77(0.01)	0.66(0.01)	\	0.54(0.02)	0.47(0.02)
FCI	0.96(0.00)	<u>0.77</u>(0.01)	<u>0.67</u>(0.01)	\	0.46(0.01)	0.37(0.01)
Exact	0.57(0.00)	0.53(0.00)	0.52(0.00)	0.19(0.01)	0.25(0.01)	0.27(0.02)
GES	0.96(0.00)	0.64(0.00)	0.55(0.01)	0.64(0.06)	0.44(0.02)	0.32(0.01)
GRaSP	<u>0.97</u>(0.00)	**0.81**(0.01)	**0.69**(0.01)	0.65(0.05)	0.45(0.05)	0.34(0.03)
ICA-LiNGAM	0.95(0.01)	0.57(0.01)	0.49(0.00)	<u>0.82</u>(0.06)	0.34(0.03)	0.21(0.02)
Direct-LiNGAM	**0.98**(0.00)	0.60(0.01)	0.52(0.00)	**0.98**(0.00)	**0.59**(0.01)	**0.50**(0.01)
CAM	0.86(0.00)	0.70(0.01)	0.61(0.01)	0.20(0.02)	0.21(0.01)	0.17(0.01)
SCORE	0.90(0.01)	0.70(0.01)	0.59(0.01)	0.63(0.03)	0.46(0.02)	0.34(0.01)
DAS	0.84(0.01)	0.66(0.01)	0.55(0.01)	0.63(0.03)	0.45(0.02)	0.35(0.02)
NoGAM	0.90(0.01)	0.69(0.01)	0.60(0.01)	0.73(0.03)	0.50(0.02)	0.41(0.01)
DiffAN	0.85(0.01)	0.68(0.01)	0.59(0.01)	0.37(0.03)	0.26(0.02)	0.22(0.01)
CHess (Stein)	0.92(0.01)	0.72(0.01)	0.60(0.01)	0.72(0.04)	<u>0.56</u>(0.03)	<u>0.48</u>(0.02)
CHess (DiffAN)	0.90(0.01)	0.70(0.01)	0.60(0.01)	0.66(0.02)	0.52(0.02)	0.40(0.02)

Table 14: $d = 10$, noise=gumbel: SHD Metrics

Method	SHD ↓ $K = 0$	$K = 1$	$K = 2$
PC	4.96(5.29)	<u>9.37</u>(6.60)	**12.26**(10.64)
FCI	9.30(2.43)	13.67(9.56)	16.15(10.13)
Exact	20.26(6.19)	21.48(7.95)	20.96(11.89)
GES	5.30(9.84)	14.78(9.14)	19.48(20.18)
GRaSP	4.81(5.56)	**8.96**(7.22)	<u>12.56</u>(11.73)
ICA-LiNGAM	<u>2.81</u>(17.34)	19.11(24.10)	25.11(21.14)
Direct-LiNGAM	**0.37**(0.46)	13.59(11.28)	18.48(19.81)
CAM	10.85(6.64)	14.93(9.40)	18.47(17.78)
SCORE	5.52(11.21)	11.47(8.05)	16.53(18.05)
DAS	5.67(10.74)	10.67(6.76)	14.80(11.76)
NoGAM	4.15(9.24)	11.03(9.77)	15.27(21.40)
DiffAN	9.15(10.94)	14.96(15.29)	18.85(17.39)
CHess (Stein)	4.00(11.04)	10.07(19.60)	14.17(20.47)
CHess (DiffAN)	4.85(6.13)	10.87(13.72)	15.80(24.23)

Table 15: $d = 5$, noise=lognormal, MLP: F1 Metrics

Method	F1 (Adj) ↑ $K = 0$	$K = 1$	$K = 2$	F1 (Arrow) ↑ $K = 0$	$K = 1$	$K = 2$
PC	0.79(0.02)	0.67(0.02)	0.64(0.02)	\	\	\
FCI	0.79(0.02)	0.67(0.02)	0.64(0.02)	\	\	\
Exact	0.62(0.02)	0.62(0.02)	0.65(0.01)	0.33(0.05)	0.32(0.03)	0.34(0.03)
GES	0.77(0.01)	0.67(0.03)	0.66(0.02)	\	\	\
GRaSP	0.68(0.04)	0.65(0.03)	0.62(0.02)	\	\	\
ICA-LiNGAM	0.79(0.01)	0.65(0.01)	0.65(0.01)	0.56(0.04)	0.38(0.03)	0.35(0.03)
Direct-LiNGAM	0.75(0.01)	0.67(0.02)	0.67(0.01)	0.50(0.05)	0.40(0.05)	0.45(0.04)
SCORE	0.86(0.01)	0.68(0.01)	0.70(0.01)	0.81(0.02)	0.57(0.02)	0.55(0.03)
DAS	0.86(0.01)	0.68(0.01)	0.70(0.01)	0.81(0.02)	0.57(0.02)	0.55(0.03)
CAM	0.86(0.02)	0.72(0.01)	**0.72**(0.01)	0.37(0.06)	0.28(0.03)	0.31(0.04)
NoGAM	**0.87**(0.01)	0.68(0.02)	0.70(0.01)	**0.83**(0.02)	0.57(0.02)	0.55(0.02)
DiffAN	0.86(0.01)	**0.72**(0.01)	0.71(0.01)	0.65(0.03)	0.47(0.03)	0.36(0.03)
CHess (Stein)	0.86(0.02)	0.69(0.01)	**0.72**(0.01)	0.77(0.03)	**0.64**(0.02)	**0.67**(0.02)
CHess (DiffAN)	0.85(0.02)	0.70(0.01)	0.71(0.01)	0.78(0.05)	0.62(0.03)	0.65(0.02)

Table 16: $d = 5$, noise=lognormal, MLP: SHD Metrics

Method	SHD ↓ $K = 0$	$K = 1$	$K = 2$
PC	3.77(1.85)	5.67(2.82)	5.83(2.47)
FCI	5.70(0.74)	6.73(1.60)	6.93(1.86)
Exact	6.23(2.91)	6.53(2.78)	6.33(1.89)
GES	4.27(2.46)	5.63(3.37)	6.50(2.98)
GRaSP	4.43(2.58)	5.20(2.49)	5.60(1.71)
ICA-LiNGAM	3.37(1.77)	6.00(2.20)	6.30(2.48)
Direct-LiNGAM	3.87(1.92)	5.70(3.48)	5.67(3.49)
SCORE	1.67(1.09)	4.57(1.78)	4.80(2.43)
DAS	1.67(1.09)	4.57(1.78)	4.80(2.43)
CAM	4.23(4.18)	6.43(2.58)	6.17(2.61)
NoGAM	**1.47**(1.05)	4.53(2.45)	4.70(2.08)
DiffAN	2.67(2.49)	5.07(2.20)	6.00(2.47)
CHess (Stein)	1.77(1.65)	**4.17**(2.54)	**3.90**(1.62)
CHess (DiffAN)	1.80(2.96)	**4.17**(2.21)	4.03(2.30)

F.3 DREAM4 experiments

Results. We report the SHD of all methods across 5 networks in Table 17. Due to the extremely small sample size (30 observations for each network), we compare all methods on the full data without selection. We see that our methods achieve the best performance for all networks, with the only exception being Network 5, where DAS [Montagna et al., 2023] achieves the best performance. This shows that our methods perform competitively on small-sized real-world datasets even without selection bias.

Table 17: SHD Results for $K = 0$ Across All Networks

Method	Network 1	Network 2	Network 3	Network 4	Network 5
SCORE	12.0	12.0	**12.0**	13.0	9.0
DAS	12.0	13.0	13.0	13.0	**8.0**
NoGAM	13.0	13.0	13.0	12.0	10.0
CAM	12.0	11.0	14.0	12.0	10.0
CHess (stein)	**11.0**	14.0	13.0	12.0	13.0
CHess (diffan)	12.0	**9.0**	**12.0**	**11.8**	10.4

F.4 Sachs experiments

We provide extra results for the Sachs dataset in Table 18-20 for F1 arrow and Table 21-23 for causal ordering accuracy on the aggregated perturbational dataset and in 24 for F1 arrow on the separated dataset. We vary the strength and shape of the selection bias, and apply random downsampling when $\rho = 1$ to the same sample size as when selection bias is present, to enable a fair comparison. The results are consistent with the main text, where our methods achieve the best performance.

Extra result for *aggregated* Sachs Dataset We combine all 9 experimental conditions into one dataset and score against the ground truth ($n = 7466$ before selection bias is applied), which is inline with the main text. We apply selection of different shapes (half-space, annulus, two-box) and different selection fraction $\rho \in \{1.0(\text{unselected}), 0.2, 0.1, 0.05\}$.

Table 18: F1 (Arrow), halfspace selection

Method	$\rho = 1$ $n = 1493$ / 747 / 373	$\rho = 0.2$ ($n = 1493$)	$\rho = 0.1$ ($n = 747$)	$\rho = 0.05$ ($n = 373$)
SCORE	0.21 (±0.04) / 0.26 (±0.06) / 0.28 (±0.05)	0.28 (±0.04)	0.29 (±0.05)	0.28 (±0.08)
DAS	0.17 (±0.06) / 0.18 (±0.10) / 0.19 (±0.10)	0.24 (±0.05)	0.23 (±0.08)	0.20 (±0.10)
NoGAM	0.23 (±0.02) / 0.26 (±0.01) / 0.29 (±0.04)	0.28 (±0.04)	0.30 (±0.06)	0.31 (±0.08)
CAM	**0.41** (±0.06) / 0.42 (±0.08) / 0.40 (±0.08)	0.23 (±0.08)	0.28 (±0.09)	0.27 (±0.10)
PC	0.27 (±0.10) / 0.26 (±0.10) / 0.25 (±0.09)	0.23 (±0.07)	0.23 (±0.07)	0.20 (±0.09)
FCI	0.16 (±0.05) / 0.16 (±0.06) / 0.13 (±0.04)	0.11 (±0.10)	0.14 (±0.11)	0.11 (±0.13)
GES	0.27 (±0.05) / 0.25 (±0.05) / 0.25 (±0.06)	0.26 (±0.08)	0.22 (±0.08)	0.23 (±0.09)
GRaSP	0.25 (±0.07) / 0.24 (±0.06) / 0.28 (±0.07)	0.23 (±0.11)	0.22 (±0.10)	0.28 (±0.14)
ICA-LiNGAM	0.23 (±0.07) / 0.23 (±0.06) / 0.27 (±0.08)	0.29 (±0.06)	0.30 (±0.07)	0.27 (±0.09)
DirectLiNGAM	0.30 (±0.06) / 0.28 (±0.06) / 0.30 (±0.07)	0.30 (±0.07)	0.33 (±0.09)	0.31 (±0.09)
DiffAN	0.34 (±0.06) / 0.36 (±0.09) / 0.30 (±0.08)	0.18 (±0.08)	0.19 (±0.05)	0.20 (±0.09)
CHess (Stein)	0.40 (±0.05) / **0.44** (±0.06) / **0.42** (±0.09)	0.34 (±0.12)	0.32 (±0.12)	**0.36** (±0.08)
CHess (DiffAN)	0.36 (±0.06) / 0.42 (±0.07) / 0.41 (±0.10)	**0.41** (±0.12)	**0.38** (±0.12)	0.36 (±0.10)

Table 19: F1 (Arrow), annulus selection

Method	$\rho = 1$ $n = 1493$ / 747 / 373	$\rho = 0.2$ ($n = 1493$)	$\rho = 0.1$ ($n = 747$)	$\rho = 0.05$ ($n = 373$)
SCORE	0.21 (±0.04) / 0.26 (±0.06) / 0.28 (±0.05)	0.29 (±0.06)	0.32 (±0.07)	0.34 (±0.08)
DAS	0.17 (±0.06) / 0.18 (±0.10) / 0.19 (±0.10)	0.26 (±0.06)	0.29 (±0.09)	0.32 (±0.08)
NoGAM	0.23 (±0.02) / 0.26 (±0.01) / 0.29 (±0.04)	0.28 (±0.05)	0.31 (±0.06)	0.33 (±0.09)
CAM	**0.41** (±0.06) / 0.42 (±0.08) / 0.40 (±0.08)	0.30 (±0.08)	0.31 (±0.13)	0.26 (±0.10)
PC	0.27 (±0.10) / 0.26 (±0.10) / 0.25 (±0.09)	0.23 (±0.10)	0.25 (±0.07)	0.23 (±0.12)
FCI	0.16 (±0.05) / 0.16 (±0.06) / 0.13 (±0.04)	0.11 (±0.08)	0.15 (±0.08)	0.15 (±0.11)
GES	0.27 (±0.05) / 0.25 (±0.05) / 0.25 (±0.06)	0.22 (±0.05)	0.21 (±0.07)	0.20 (±0.10)
GRaSP	0.25 (±0.07) / 0.24 (±0.06) / 0.28 (±0.07)	0.17 (±0.07)	0.18 (±0.09)	0.19 (±0.08)
ICA-LiNGAM	0.23 (±0.07) / 0.23 (±0.06) / 0.27 (±0.08)	0.21 (±0.06)	0.26 (±0.06)	0.27 (±0.07)
DirectLiNGAM	0.30 (±0.06) / 0.28 (±0.06) / 0.30 (±0.07)	0.29 (±0.07)	0.27 (±0.07)	0.30 (±0.07)
DiffAN	0.34 (±0.06) / 0.36 (±0.09) / 0.30 (±0.08)	0.14 (±0.05)	0.20 (±0.06)	0.19 (±0.07)
CHess (Stein)	0.40 (±0.05) / **0.44** (±0.06) / **0.42** (±0.09)	**0.39** (±0.07)	**0.45** (±0.11)	**0.48** (±0.12)
CHess (DiffAN)	0.36 (±0.06) / 0.42 (±0.07) / 0.41 (±0.10)	0.36 (±0.08)	0.41 (±0.09)	0.37 (±0.14)

Table 20: F1 (Arrow), two boxes selection

Method	$\rho = 1$ $n = 1493$ / 747 / 373	$\rho = 0.2$ ($n = 1493$)	$\rho = 0.1$ ($n = 747$)	$\rho = 0.05$ ($n = 373$)
SCORE	0.21 (±0.04) / 0.26 (±0.06) / 0.28 (±0.05)	0.21 (±0.06)	0.24 (±0.06)	0.21 (±0.05)
DAS	0.17 (±0.06) / 0.18 (±0.10) / 0.19 (±0.10)	0.21 (±0.05)	0.21 (±0.07)	0.19 (±0.09)
NoGAM	0.23 (±0.02) / 0.26 (±0.01) / 0.29 (±0.04)	0.26 (±0.05)	0.27 (±0.04)	0.24 (±0.04)
CAM	**0.41** (±0.06) / 0.42 (±0.08) / 0.40 (±0.08)	0.31 (±0.09)	0.34 (±0.08)	0.31 (±0.09)
PC	0.27 (±0.10) / 0.26 (±0.10) / 0.25 (±0.09)	0.29 (±0.08)	0.27 (±0.07)	0.21 (±0.12)
FCI	0.16 (±0.05) / 0.16 (±0.06) / 0.13 (±0.04)	0.13 (±0.11)	0.18 (±0.07)	0.16 (±0.11)
GES	0.27 (±0.05) / 0.25 (±0.05) / 0.25 (±0.06)	0.25 (±0.06)	0.27 (±0.09)	0.28 (±0.09)
GRaSP	0.25 (±0.07) / 0.24 (±0.06) / 0.28 (±0.07)	0.26 (±0.09)	0.27 (±0.11)	0.25 (±0.13)
ICA-LiNGAM	0.23 (±0.07) / 0.23 (±0.06) / 0.27 (±0.08)	0.25 (±0.06)	0.28 (±0.06)	0.26 (±0.06)
DirectLiNGAM	0.30 (±0.06) / 0.28 (±0.06) / 0.30 (±0.07)	0.32 (±0.07)	0.35 (±0.08)	0.32 (±0.10)
DiffAN	0.34 (±0.06) / 0.36 (±0.09) / 0.30 (±0.08)	0.20 (±0.08)	0.23 (±0.09)	0.22 (±0.09)
CHess (Stein)	0.40 (±0.05) / **0.44** (±0.06) / **0.42** (±0.09)	0.32 (±0.10)	0.28 (±0.11)	0.29 (±0.11)
CHess (DiffAN)	0.36 (±0.06) / 0.42 (±0.07) / 0.41 (±0.10)	**0.37** (±0.10)	**0.40** (±0.14)	**0.39** (±0.12)

Table 21: Causal ordering accuracy $1 - D_{top}/|E|$ (0.5 = random order), halfspace selection

Method	$\rho = 1$ $n = 1493$ / 747 / 373	$\rho = 0.2$ ($n = 1493$)	$\rho = 0.1$ ($n = 747$)	$\rho = 0.05$ ($n = 373$)
SCORE	0.38 (±0.06) / 0.42 (±0.07) / 0.42 (±0.04)	0.52 (±0.08)	0.52 (±0.08)	0.52 (±0.09)
DAS	0.38 (±0.06) / 0.42 (±0.07) / 0.42 (±0.04)	0.52 (±0.08)	0.52 (±0.08)	0.52 (±0.09)
NoGAM	0.42 (±0.03) / 0.42 (±0.03) / 0.44 (±0.04)	0.53 (±0.06)	0.52 (±0.07)	0.52 (±0.06)
CAM	0.64 (±0.07) / 0.61 (±0.12) / 0.59 (±0.12)	0.40 (±0.12)	0.45 (±0.12)	0.43 (±0.14)
PC	- / - / -	-	-	-
FCI	- / - / -	-	-	-
GES	- / - / -	-	-	-
GRaSP	- / - / -	-	-	-
ICA-LiNGAM	0.46 (±0.12) / 0.44 (±0.10) / 0.48 (±0.13)	0.54 (±0.10)	0.56 (±0.10)	0.48 (±0.10)
DirectLiNGAM	0.60 (±0.11) / 0.58 (±0.07) / 0.61 (±0.12)	0.59 (±0.12)	**0.62** (±0.14)	**0.57** (±0.13)
DiffAN	0.58 (±0.09) / 0.58 (±0.11) / 0.56 (±0.11)	0.36 (±0.10)	0.38 (±0.07)	0.41 (±0.10)
CHess (Stein)	**0.73** (±0.04) / **0.75** (±0.05) / **0.71** (±0.07)	0.55 (±0.13)	0.48 (±0.15)	0.55 (±0.12)
CHess (DiffAN)	0.72 (±0.05) / 0.71 (±0.06) / **0.71** (±0.09)	**0.66** (±0.12)	0.60 (±0.16)	0.56 (±0.08)

Table 22: Causal ordering accuracy $1 - D_{top}/|E|$ (0.5 = random order), annulus selection

Method	$\rho = 1$ $n = 1493$ / 747 / 373	$\rho = 0.2$ ($n = 1493$)	$\rho = 0.1$ ($n = 747$)	$\rho = 0.05$ ($n = 373$)
SCORE	0.38 (±0.06) / 0.42 (±0.07) / 0.42 (±0.04)	0.45 (±0.07)	0.48 (±0.07)	0.49 (±0.06)
DAS	0.38 (±0.06) / 0.42 (±0.07) / 0.42 (±0.04)	0.45 (±0.07)	0.48 (±0.07)	0.49 (±0.06)
NoGAM	0.42 (±0.03) / 0.42 (±0.03) / 0.44 (±0.04)	0.46 (±0.04)	0.46 (±0.05)	0.47 (±0.08)
CAM	0.64 (±0.07) / 0.61 (±0.12) / 0.59 (±0.12)	0.49 (±0.14)	0.51 (±0.17)	0.45 (±0.19)
PC	- / - / -	-	-	-
FCI	- / - / -	-	-	-
GES	- / - / -	-	-	-
GRaSP	- / - / -	-	-	-
ICA-LiNGAM	0.46 (±0.12) / 0.44 (±0.10) / 0.48 (±0.13)	0.42 (±0.10)	0.49 (±0.09)	0.50 (±0.10)
DirectLiNGAM	0.60 (±0.11) / 0.58 (±0.07) / 0.61 (±0.12)	0.58 (±0.10)	0.53 (±0.10)	0.57 (±0.11)
DiffAN	0.58 (±0.09) / 0.58 (±0.11) / 0.56 (±0.11)	0.26 (±0.09)	0.34 (±0.13)	0.34 (±0.14)
CHess (Stein)	**0.73** (±0.04) / **0.75** (±0.05) / **0.71** (±0.07)	0.65 (±0.16)	**0.71** (±0.09)	**0.71** (±0.10)
CHess (DiffAN)	0.72 (±0.05) / 0.71 (±0.06) / **0.71** (±0.09)	**0.65** (±0.11)	**0.71** (±0.10)	0.65 (±0.12)

Table 23: Causal ordering accuracy $1 - D_{top}/|E|$ (0.5 = random order), two boxes selection

Method	$\rho = 1$ $n = 1493$ / 747 / 373	$\rho = 0.2$ ($n = 1493$)	$\rho = 0.1$ ($n = 747$)	$\rho = 0.05$ ($n = 373$)
SCORE	0.38 (±0.06) / 0.42 (±0.07) / 0.42 (±0.04)	0.39 (±0.06)	0.40 (±0.07)	0.40 (±0.07)
DAS	0.38 (±0.06) / 0.42 (±0.07) / 0.42 (±0.04)	0.39 (±0.06)	0.40 (±0.07)	0.40 (±0.07)
NoGAM	0.42 (±0.03) / 0.42 (±0.03) / 0.44 (±0.04)	0.45 (±0.05)	0.43 (±0.06)	0.41 (±0.06)
CAM	0.64 (±0.07) / 0.61 (±0.12) / 0.59 (±0.12)	0.51 (±0.14)	0.52 (±0.15)	0.51 (±0.13)
PC	- / - / -	-	-	-
FCI	- / - / -	-	-	-
GES	- / - / -	-	-	-
GRaSP	- / - / -	-	-	-
ICA-LiNGAM	0.46 (±0.12) / 0.44 (±0.10) / 0.48 (±0.13)	0.48 (±0.11)	0.51 (±0.11)	0.46 (±0.10)
DirectLiNGAM	0.60 (±0.11) / 0.58 (±0.07) / 0.61 (±0.12)	0.65 (±0.10)	0.64 (±0.11)	0.60 (±0.14)
DiffAN	0.58 (±0.09) / 0.58 (±0.11) / 0.56 (±0.11)	0.36 (±0.14)	0.37 (±0.12)	0.43 (±0.12)
CHess (Stein)	**0.73** (±0.04) / **0.75** (±0.05) / **0.71** (±0.07)	0.58 (±0.13)	0.55 (±0.14)	0.52 (±0.11)
CHess (DiffAN)	0.72 (±0.05) / 0.71 (±0.06) / **0.71** (±0.09)	**0.67** (±0.09)	**0.66** (±0.12)	**0.61** (±0.13)

Extra result for *separate* Sachs Dataset (per-condition) Each of the 9 experimental conditions is treated as its own dataset and scored against its own ground truth: a perfect intervention makes the perturbed protein independent of its parents, so that protein's incoming edges are cut. We apply selection of different shapes (half-space, annulus, two-box) to each of the per-condition subdataset, with 50% of the data remaining. Results show that the proposed CHess achieves the best performance when hard selection is present in the data. Overall CHess (Stein) dominates when selection is present, CHess (DiffAN) dominates in some of the 9 conditions (`psitect`, `aktinhib`,`pma`) as well, which is not shown in this averaged table.

Table 24: F1 arrow of per-condition Sachs dataset

Method	$\rho = 1.0$	$\rho = 0.5$		
	non-select	halfspace	annulus	two boxes
SCORE	**0.37** (±0.09)	**0.36** (±0.07)	0.37 (±0.08)	0.38 (±0.09)
DAS	0.31 (±0.08)	0.31 (±0.07)	0.31 (±0.08)	0.32 (±0.07)
NoGAM	0.35 (±0.07)	0.35 (±0.07)	0.35 (±0.07)	0.34 (±0.07)
CAM	0.28 (±0.11)	0.28 (±0.09)	0.28 (±0.08)	0.29 (±0.11)
PC	0.08 (±0.05)	0.16 (±0.02)	0.11 (±0.03)	0.15 (±0.03)
FCI	0.03 (±0.05)	0.09 (±0.03)	0.07 (±0.04)	0.12 (±0.06)
GES	0.12 (±0.03)	0.21 (±0.02)	0.18 (±0.02)	0.17 (±0.02)
GRaSP	0.07 (±0.04)	0.11 (±0.02)	0.11 (±0.05)	0.11 (±0.07)
ICA-LiNGAM	0.34 (±0.08)	0.31 (±0.05)	0.33 (±0.06)	0.34 (±0.06)
DirectLiNGAM	0.29 (±0.14)	0.29 (±0.07)	0.29 (±0.09)	0.28 (±0.10)
DiffAN	0.27 (±0.08)	0.26 (±0.07)	0.27 (±0.08)	0.27 (±0.08)
CHess (Stein)	0.36 (±0.13)	**0.36** (±0.11)	**0.37** (±0.13)	**0.39** (±0.12)
CHess (DiffAN)	0.34 (±0.06)	0.34 (±0.05)	0.34 (±0.07)	0.34 (±0.06)

G Hardware

All experiments were conducted on a server running Debian GNU/Linux 12, equipped with an Intel Xeon E5-2680 v4 CPU (2.40 GHz), 503 GB RAM, and ten NVIDIA GeForce GTX 1080 Ti GPUs (11 GB VRAM each), using Python 3.13.

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: The abstract and introduction accurately reflect the paper's contributions and scope, which is proposing a new score-matching based method for causal discovery.

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: The paper discuss the limitations of the work in the last Section.

   Guidelines:

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

3. **Theory assumptions and proofs**

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

   Answer: [Yes]

Justification: All the proofs are provided in the Appendix.

Guidelines:

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

4. **Experimental result reproducibility**

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

   Answer: [Yes]

   Justification: All the experimental details are in the appendix, and we will release the code as soon as possible.

   Guidelines:

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

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

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

Answer:[No]

Justification: We have provided the details of the experiments, access to the data in the appendix, but not yet released the code. We will do so as soon as possible.

Guidelines:

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

6. **Experimental setting/details**

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

   Answer: [Yes]

   Justification: We provided the experimental details in the appendix.

   Guidelines:

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

7. **Experiment statistical significance**

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

   Answer: [Yes]

   Justification: We reported the mean and variance of 20 seeds of repeated runs per experiment.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

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

   Answer: [Yes]

   Justification: We provide the details of the computation resources in the Appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

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

   Answer: [Yes]

   Justification: We conform in every respect, with the NeurIPS Code of Ethics

   Guidelines:

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

10. **Broader impacts**

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

    Answer: [N/A]

    Justification: This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, as selection bias is an important problem in causal inference and epidemiology, so our theory can have a vast range of positive applications. We do not find any risk that we feel must be specifically highlighted here.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

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

    Answer: [N/A]

    Justification: the paper poses no such risks.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

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

    Answer: [Yes]

    Justification: all the existing code packages or dataset used in the paper are cited.

    Guidelines:

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

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

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

    Answer: [N/A]

    Justification: the paper does not release new assets.

    Guidelines:

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

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

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

    Answer: [N/A]

    Justification: the paper does not involve crowdsourcing nor research934 with human subjects.

    Guidelines:

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

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

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

    Answer: [N/A]

    Justification: the paper does not involve crowdsourcing nor research951 with human subjects.

    Guidelines:

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

16. **Declaration of LLM usage**

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

Answer: [N/A]

Justification: the core method development in this research does not involve LLMs as any important, original, or non-standard components.

Guidelines:

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