CAUSAL STRUCTURE SHIFT DETECTION VIA ORTHOGONAL RESIDUAL MOMENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Causal relationships may appear or disappear across environments, and capturing them is essential for understanding the structure of the underlying system. Most existing methods can identify which causal mechanisms shift under a common graph, but they cannot attribute a detected shift to a change in the causal structure. In this work, we consider nonlinear additive noise models sharing a causal order and establish conditions under which observed data uniquely identify the causal structure shifts, even when the order is not unique. We also provide conditions that preserve this guarantee under structure shifts consistent with a common order. Building on these identifiability results, we propose STORM, a two-stage algorithm that first estimates a common order by repeatedly removing a node with no children in any environment and then detects shifts by testing whether omitting a candidate parent changes the target node's conditional mean. To reduce the impact of regression estimation errors on both stages, we derive orthogonal scores whose expectations are insensitive to the errors to first order. Experiments on synthetic and real-world datasets showed the effectiveness of STORM in various scenarios.

1 INTRODUCTION AND RELATED WORK

This paper focuses on *causal structure shifts*, where direct causal relationships appear or disappear across environments. For example, in cellular signaling networks, an edge represents a direct influence of one protein on the activity of another. The same downstream protein can be activated by different sets of upstream proteins in response to inflammatory signaling and ultraviolet exposure (Brancho et al., 2003). Identifying these differences can guide interventions that change the protein's response to one stimulus while preserving its response to another. Mapping these relationships experimentally requires perturbing proteins across conditions, but cost and available biological material limit the scale of these experiments (Gabor et al., 2021). Detecting links gained or lost from observational data therefore provides a focused set of candidates for experimental validation. To detect causal structure shifts, a natural starting point is to ask whether the joint distribution varies across environments. However, from the perspective of a structural causal model (SCM), differences in structural functions or exogenous noise distributions can change the joint distribution even under a fixed causal structure (Rothenhäusler et al., 2015; Mameche et al., 2022; Montagna, 2026). In light of this ambiguity, we factorize the joint distribution in each environment according to its causal graph into a product of *causal mechanisms*—the conditional distributions of the nodes given their parents (Ghassami et al., 2018; Schölkopf et al., 2021; Perry et al., 2022)—and examine distribution shifts at the level of individual causal mechanisms. Under this factorization, a causal structure shift corresponds to a change in the parent set of a mechanism.

A substantial body of work uses changes in causal mechanisms across heterogeneous data to learn causal models (Huang et al., 2020; Perry et al., 2022; Mameche et al., 2022; Hägele et al., 2023; Mameche et al., 2025; Xu et al., 2025) or causal representations (Lachapelle et al., 2022; Squires et al., 2023; Buchholz et al., 2023; Acartürk et al., 2024; Reizinger et al., 2025; Alimisis & Diou, 2026). These methods hold a single causal structure fixed across environments and identify which mechanisms shift. However, they do not identify whether causal mechanism shifts stem from changes in a causal structure. DCI (Wang et al., 2018) detects differences in structural coefficients, including changes that preserve a causal structure, but it supports only linear Gaussian models. For nonlinear additive noise models (ANMs) sharing a causal order, iSCAN (Chen et al., 2023) utilizes derivatives of the log-density to identify mechanism shifts and then estimates parent-set differences at the selected

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

nodes. In nonparametric settings, conditional means are often easier to estimate than derivatives of the log-density (Strobl & Visweswaran, 2021; Wibisono et al., 2024). This motivates the development of a regression-based approach to detecting causal structure shifts.

On the theoretical side, Peters et al. (2014) establish graph identifiability from a single environment distribution. Chen et al. (2023) addresses multiple environments, but its guarantees rely on Gaussian noise. We then seek sufficient conditions on structural functions and noise distributions to recover a valid common order and identify causal structure shifts, while permitting non-Gaussian noise. We also investigate when these identifiability results are preserved under causal structure shifts.

Contributions. We present STORM, a two-stage algorithm for causal **ST**ructure shift detection via **O**rthogonal **R**esidual **M**oments. We consider samples from multiple environments, each governed by a nonlinear ANM with a common causal order, while parent sets, structural functions, and noise distributions may vary across environments (Assumption 1). STORM first estimates a common order by repeatedly removing a node with no children in any environment. Using each node's predecessors in this order as candidate parents, it identifies parent sets in each environment by testing whether each candidate parent changes the conditional mean of a target node, and compares these sets to detect structure shifts. Since regression estimation errors can lead to incorrect decisions, we construct scores based on Neyman orthogonality (Chernozhukov et al., 2018; 2022) whose expectations are insensitive to first-order estimation errors. Specifically, our work makes the following contributions:

- **Identifiability Guarantees (§3.2).** We establish sufficient conditions under which the distributions in the population uniquely determine causal structure shifts. Under conditions on the structural functions and exogenous noises, checking residual independence in every environment identifies a valid common order, even when no single environment satisfies the conditions for every edge (Theorem 1). We also derive the identifiability of the parent set given any valid common order, implying that the identified structure shifts are invariant to the choice of order (Proposition 1).
- **Preservation of Identifiability under Structure Shifts (§3.2).** Causal structure shifts can alter the conditional input distributions involved in the restricted-ANM identifiability conditions of Peters et al. (2014). Therefore, these conditions need to be verified for the shifted distributions. We establish a model class in which identification is preserved under causal structure shifts compatible with a common order, without specifying in advance which edges change (Theorem 2).
- **Estimation with Orthogonal Scores (§3.3).** In the first stage, to find a node with no children in any environment, residual independence is assessed through an exact criterion based on the characteristic functions, with a correction that reduces the impact of regression error (Lemma 1). In the second stage, whether each candidate parent changes the conditional mean of a node is tested following a projected-covariance score (Lundborg et al., 2024) (Lemma 2).
- **Experimental Verification (§4).** Synthetic experiments showed that STORM outperformed the baselines across various evaluation metrics, with ablation results supporting the benefit of the orthogonal residual moments. On mass-cytometry data, STORM identified biologically plausible shifts associated with MEK inhibition. In the light-tunnel data, it reported programmed edge deletions under the structural intervention and no candidate shifts under the noise-only control.

2 Problem Setup

Notations. We define $[n] = \{1, \dots, n\}$ for $n \in \mathbb{N}$. Let $G = (V, \mathcal{E})$ denote a *directed acyclic graph* (DAG), where $V = [d]$ is a node set and $\mathcal{E}$ is an edge set, and let π denote a *causal order*. For a graph G, let $\mathrm{pa}(j)$ be the parent set of node j, and $\mathrm{Top}(G)$ be the set of causal orders. For a causal order π, we write $k \prec_\pi j$ when k precedes j in π, and define the predecessor set of node j by $P_\pi(j) \coloneqq \{\ell \in V : \ell \prec_\pi j\}$. A *structural causal model* (SCM) is denoted by $M = (G, f, \mathbb{P}_N)$, where $f = (f_j)_{j \in V}$ is the collection of *structural functions* and $\mathbb{P}_N$ is the joint distribution of its *exogenous noise* variables. We write $G(M)$ for its graph and $\mathbb{P}_X^M$ for its induced observational distribution, where $X \sim \mathbb{P}_X^M$. In environment e, we write $X^e = (X_1^e, \dots, X_d^e)$ for the random vector and $\mathbb{E}_e$ for the expectation under its distribution. The DAG in environment e and its parent sets are written by $G^e = (V, \mathcal{E}^e)$ and $\mathrm{pa}_e(j)$, respectively. Let $\mathcal{M}$ denote the model space on V, and the model in environment e is $M^e \in \mathcal{M}$ with $G(M^e) = G^e$. Let $o : \mathcal{M} \to \mathcal{M}$ denote a *model transformation*, and $\mathcal{O}$ denote a class of such operations. A detailed list of notation and the relevant definitions are given in Table 2 and Appendix B.

Causal models across environments. We consider nonlinear ANMs over the same d nodes in E environments. In environment e, any model $M^e \in \mathcal{M}$ has the form

$$X_j^e = f_j^e\big(X^e_{\mathrm{pa}_e(j)}\big) + N_j^e, \qquad j \in V, \tag{1}$$

where we assume mutually independent noises, and, without loss of generality, mean-zero noises. We further require models in $\mathcal{M}$ to satisfy causal sufficiency (Assumption 5) and causal minimality (Assumption 4). The Markov property follows from an SCM with mutually independent noise variables, and faithfulness is not required (see the preliminaries in Appendix B.1 for details). Our parameter of interest is the causal structure shift set, defined by:

Definition 1 (Causal structure shift). *For $j \in V$ and distinct environments $e_1, e_2 \in [E]$, the following symmetric difference defines the causal structure shift set of node j.*

$$S_j^{e_1,e_2} \coloneqq \mathrm{pa}_{e_1}(j) \triangle \mathrm{pa}_{e_2}(j), \tag{2}$$

where $\triangle$ is a symmetric difference. An edge $k \to j$ shifts between e_1 and e_2 if and only if $k \in S_j^{e_1,e_2}$.

We next specify how the causal models may vary across environments. We allow the parent sets, structural functions, and noise distributions to vary, unlike settings with a fixed causal structure (Rothenhäusler et al., 2015; Mameche et al., 2022; Montagna, 2026) or the sparse mechanism shift hypothesis (Schölkopf et al., 2021; Hägele et al., 2023; Xu et al., 2025). Specifically, although the sparse mechanism shift framework introduced by Perry et al. (2022) restricts changes to a subset of mechanisms while keeping the parent sets of an underlying DAG G fixed, we relax this restriction by allowing some or all incoming edges to be removed, so that $\mathrm{pa}_e(j) \subseteq \mathrm{pa}(j)$. Under this relaxation, parent sets may vary across environments while the corresponding DAGs are assumed to share a common causal order. The following assumption states this structural requirement.

Assumption 1 (Union graph acyclicity). *$G_\cup = (V, \mathcal{E}_\cup)$ is acyclic, where $\mathcal{E}_\cup \coloneqq \bigcup_{e=1}^E \mathcal{E}^e$.*

Assumption 1 guarantees the existence of a common order across environments, while allowing their edge sets to differ. Under this assumption, we have $\Pi_0 \coloneqq \bigcap_{e=1}^E \mathrm{Top}(G^e) = \mathrm{Top}(G_\cup) \neq \varnothing$. We refer to any $\pi \in \Pi_0$ as a *valid common order*. We impose this requirement on model transformations by restricting $\mathcal{O}$ to *structure-shift operations* satisfying

$$\mathrm{Top}\big(G(M)\big) \cap \mathrm{Top}\big(G(o(M))\big) \neq \varnothing, \qquad M \in \mathcal{M}. \tag{3}$$

The operations mean allowed model changes, but our estimand remains the causal structure shift set in Definition 1. A valid common order $\pi \in \Pi_0$ serves as an intermediate target for estimating $S_j^{e_1,e_2}$. For every environment, we have $\mathrm{pa}_e(j) \subseteq P_\pi(j)$, and these predecessors contain no descendants of j. This gives a predecessor set across environments and reduces parent identification to conditional-mean comparisons under the conditions established in §3.2. Any valid common order yields the same parent sets and hence the same structure shifts under those conditions (Proposition 1).

3 Proposed Method: STORM

3.1 Overview of Methodology

With access to the environment distributions, we now describe an overall procedure that first recovers a common causal order and then uses it to identify the parent sets in each environment.

Common-order recovery. Although Assumption 1 guarantees the existence of a common causal order, we must recover one from the observational distributions. A node with no children in any environment can be placed last in a common causal order. We refer to such a node as a *common sink*. At each step, this definition applies to the graphs induced by the remaining node set $R \subseteq V$. For $j \in R$ and $e \in [E]$, define the conditional-mean residual

$$\varepsilon^e_{R,j} \coloneqq X_j^e - \mathbb{E}_e\Big[X_j^e \mid X^e_{R\setminus\{j\}}\Big]. \tag{4}$$

Initially, the remaining node set R is set to the node set V. For a common sink j, the conditioning set $R \setminus \{j\}$ contains all its parents and none of its descendants. Therefore, its conditional-mean residual equals N_j^e and is independent of $X^e_{R\setminus\{j\}}$ in every environment. Removing such a node leaves the

structural equations of the remaining nodes unchanged, so the same reasoning applies as common sinks are successively removed. At each step, select a node $j \in R$ satisfying $\varepsilon^e_{R,j} \perp\!\!\!\perp X^e_{R\setminus\{j\}}$ for every $e \in [E]$ and remove it from R. Repeat until no nodes remain, then reverse the removal sequence to obtain an order π.

Structure shift detection. Each node j in the estimated predecessor set $P_\pi(j)$ specifies the candidate parents. For the common order π, an environment $e \in [E]$, and $k \in P_\pi(j)$, define the *conditional-mean contrast* and its squared L_2 norm by

$$\delta^{e,\pi}_{j,k} := \mathbb{E}_e\left[X^e_j \mid X^e_{P_\pi(j)}\right] - \mathbb{E}_e\left[X^e_j \mid X^e_{P_\pi(j)\setminus\{k\}}\right], \tag{5}$$

$$\Delta^{e,\pi}_{j,k} := \mathbb{E}_e\left[\left(\delta^{e,\pi}_{j,k}\right)^2\right]. \tag{6}$$

The contrast compares the conditional mean using all predecessors with the mean after omitting k. Its squared norm $\Delta^{e,\pi}_{j,k}$ indicates the edge strength of $k \to j$ and is determined by the observational distribution for a fixed order. We retain a candidate k in environment e precisely when $\Delta^{e,\pi}_{j,k} > 0$, then we take symmetric differences of the retained sets across environments.

3.2 Identifiability Guarantees

Next, we provide sufficient conditions under which the population distributions $(\mathbb{P}^{M^e}_X)^E_{e=1}$ determine a valid common order and the structure shift sets. We then identify a model class for which the guarantee is preserved by the allowed structure-shift operations.

Identifiability of common causal order. To justify the procedure for common-order recovery, we must also show that every remaining node that is not a common sink fails the residual independence criterion in at least one environment. We establish this converse under the following conditions on the structural functions and noise distributions. We write C^1 for continuously differentiable functions and $PC^1(\mathbb{R})$ for continuous functions that are C^1 outside a finite set.

Assumption 2 (Regularity)**.** *Each structural function f^e_j is bounded and continuously differentiable. Each exogenous noise N^e_j has a strictly positive, log-concave density $q^e_j \in PC^1(\mathbb{R})$.*

Log-concavity restricts distribution shape while permitting non-Gaussian noise (Saumard & Wellner, 2014), and piecewise smoothness allows for distributions with a kink, such as Laplace noise. The following parent noise condition allows each edge to use a different environment. Figure 1 illustrates this flexibility.

Assumption 3 (Edgewise noise condition)**.** *For any edge $j \to k \in \mathcal{E}_\cup$, there is an environment e with $j \to k \in \mathcal{E}^e$ such that $\log q^e_j$ has a kink or q^e_j has a super-exponential tail in at least one direction.*

Under these conditions, residual independence in every environment characterizes common sinks and justifies the recovery procedure in §3.1.

Theorem 1 (Identifiability of common order)**.** *Under the ANM in Eq.* (1) *and Assumptions 1-4, at every remaining set R reached by removing common sinks, a node j is a common sink if and only if $\varepsilon^e_{R,j} \perp\!\!\!\perp X^e_{R\setminus\{j\}}$ for all $e \in [E]$. Successively removing these nodes and reversing the removal sequence returns a valid common order $\pi \in \Pi_0$.*

Proof sketch. The key step is to exclude an independent regression residual at a node that still has a child. We explain how the function and noise conditions yield this exclusion. Fix a remaining set R reached by removing common sinks, a remaining edge $j \to k$, and an environment e supplied by Assumption 3. Writing u for a value of $X^e_{R\setminus\{j\}}$, for every $x \in \mathbb{R}$, the conditional density factors as

$$p^e(x \mid u) = q^e_j(x - f^e_j(u_{\mathrm{pa}_e(j)}))K^e_u(x), \qquad 0 < L^e(u) \le K^e_u(x) \le H^e(u) < \infty \tag{7}$$

where K^e_u is the product of the remaining children's noise-density factors, divided by the normalizing constant. For fixed u, bounded structural functions keep the arguments of these densities in compact intervals. Their positivity and continuity therefore give the bounds in Eq. (7).

If the regression residual were independent of the other remaining variables, all these conditional densities would be translates of a single residual density. Let $a^e(u)$ denote the difference between the conditional mean of X^e_j given u and the structural function $f^e_j(u_{\mathrm{pa}_e(j)})$. If $a^e(u)$ differed across

conditioning values, equating the densities at a common residual value would require the factors K_u^e to offset a nonzero shift of the noise density. With a super-exponential tail, the absolute log-density ratio for such a shift is unbounded, contradicting the positive bounds in Eq. (7). For a noise density with a kink, comparing the numbers and locations of kinks likewise forces $a^e(u)$ to be constant, with log-concavity preventing cancelation of derivative jumps. In either case, the conditional mean equals the structural function up to a constant; the mean-zero noise assumption makes this constant zero. Thus, the supposedly independent residual is N_j^e itself. Independence from all other remaining variables would then make a remaining child's mechanism constant in its jth argument, contradicting causal minimality. This argument rules out each remaining nonsink in one environment containing a qualifying outgoing edge. Removing a common sink leaves the remaining mechanisms and noises unchanged, so every remaining edge retains such an environment. This is why different environments can supply the conditions needed at different edges in common-order recovery. □

Identifiability of causal structure shifts. For each candidate of parents $k \in P_\pi(j)$, the contrast in Eq. (5) measures how omitting node k changes the conditional mean of X_j^e. When π is a valid common order, the predecessor set of each node contains all its parents and none of its descendants in each environment. Then, if node k is not a parent of node j in environment e, omitting it leaves the conditional means equal to the structural function, so $\Delta_{j,k}^{e,\pi} = 0$. The following proposition reveals that any true parent has positive edge strength, completing the identification of the parent sets.

Proposition 1 (Identifiability of parent set)**.** *Under the ANM in Eq.* (1) *and Assumptions 1-4, suppose that each environment satisfies the intersection property. Then, for every $e \in [E]$, $j \in V$, and $\pi \in \Pi_0$,*

$$\mathrm{pa}_e(j) = \left\{ k \in P_\pi(j) : \Delta_{j,k}^{e,\pi} > 0 \right\}. \tag{8}$$

Taking the symmetric difference of the identified parent sets yields causal structure shift set $S_j^{e_1,e_2}$. This set has exactly candidates k for which $k \to j$ has zero edge strength in one environment and positive edge strength in the other. Different valid common orders may change the predecessor sets and numerical edge strengths, but identify the same parent sets and hence the same structure shifts. Thus, Theorem 1 and Proposition 1 identify the target even when the common order is not unique.

Example: complementary environments. Consider the SCMs in Figure 1. The models share the causal order $\pi = (Z, X, Y)$ and satisfy Assumptions 2 and 4. The Laplace log-density has a kink, whereas the logistic log-density is smooth and its density has exponential tails. Thus, in environment 1, the parent noise condition holds for $Z \to X$ and $Z \to Y$, but not for $X \to Y$. In environment 2, it holds for $X \to Y$, but not for $Z \to X$, while $Z \to Y$ is absent. Therefore, neither environment meets the condition for all of its present edges. Together, however, they satisfy Assumption 3: every edge in the union graph has an environment in which it is present and its parent noise meets the condition. By Theorem 1, the procedure recovers $\pi = (Z, X, Y)$ by removing the common sinks in the order Y, X, Z. For this order, the positivity of the joint densities ensures that Proposition 1 applies, which implies $\Delta_{Y,Z}^{1,\pi} > 0$ and $\Delta_{Y,Z}^{2,\pi} = 0$. Comparing the identified parent sets yields $S_Y^{1,2} = \{Z\}$ and $S_Z^{1,2} = S_X^{1,2} = \varnothing$. Thus, despite changes in the noise distributions of Z and X, the identified structure shift is exactly the deletion of $Z \to Y$.

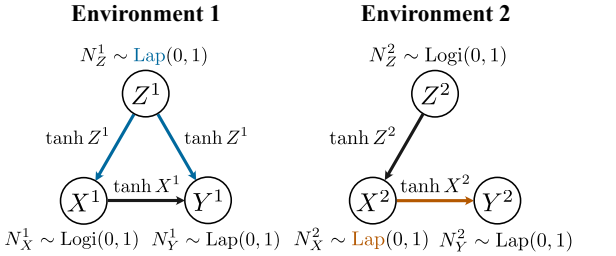


Figure 1: Complementary environments. Each variable equals its noise plus the sum of the functions labeling its incoming arrows. Colored arrows mark edges whose parent noise satisfies the kink-or-tail condition in Assumption 3, whereas black arrows denote those that do not. Different environments supply this condition for different edges of the union graph.

Preservation of identifiability under structure shifts. Theorem 1 identifies a common order for a given collection of environments. We next ask when this guarantee remains available as the models undergo structure shifts, without specifying in advance which edges change. The common-order requirement in Eq. (3) restricts the graphs, but does not by itself preserve conditions on structural functions and noise distributions. To obtain a model class that permits any node to become a parent, we impose the noise-shape condition at every node, strengthening the edgewise requirement of

Assumption 3. The following theorem then allows replacements of structural functions and noise distributions that retain the stated conditions.

Theorem 2 (Preservation of identifiability under structure shifts)**.** *Let $\mathcal{M}_\star \subseteq \mathcal{M}$ consist of models satisfying Assumption 2 and the noise-shape condition of Assumption 3 at every node, and suppose these conditions are preserved under structure-shift operations. Then*

$$o(M) \in \mathcal{M}_\star \qquad \text{for every } M \in \mathcal{M}_\star \text{ and } o \in \mathcal{O}. \tag{9}$$

$G(M)$ and $G(o(M))$ are identifiable from their observational distributions. Also, $\mathbb{P}_X^M$ and $\mathbb{P}_X^{o(M)}$ identify a valid common order and the structure shift sets in Eq. (2)*. These conclusions require no additional conditions on the induced conditional input distributions.*

The conditions on the replacement functions and noise distributions make Eq. (9) hold; Theorem 1 and Proposition 1 then identify the shifted graphs and their differences. More generally, we can choose any number of models in $\mathcal{M}_\star$ whose parent sets share an order. The union graph is acyclic, and the noise-shape condition holds at every node in every environment. Therefore, our identifiability results guarantee the recovery of a valid common order and the identification of all parent sets across environments. Thus, the guarantee permits edge additions and deletions compatible with a common order, together with changes in structural functions and noise distributions within the stated classes.

Comparison with restricted ANMs. Peters et al. (2014) also justify residual-independence sink removal. Their general restricted-ANM condition combines the structural function and child noise with conditional input distributions through a smooth differential criterion. A causal structure shift can change these input distributions even for a surviving mechanism whose structural function and noise distribution stay fixed. Hence, applying that criterion to the shifted model requires its new conditional input distributions. Our proof derives the residual-independence directly from the stated function and noise-shape conditions, including piecewise smooth densities such as the Laplace density.

3.3 Estimation Procedure with Orthogonal Scores

Both stages of §3.1 rely on conditional means, whose estimation errors can imitate the dependence used to recover the common order and parent sets. STORM addresses this problem with orthogonal scores, whose expectation is insensitive to first-order estimation errors. We first describe the construction and interpretation of the scores for both stages, then give their finite-sample implementation.

Common-order recovery. Theorem 1 identifies common sinks by utilizing the independence of the conditional-mean residual in Eq. (4) and its regressors. We represent this independence through an exact criterion based on the characteristic functions. Let $m^e_{R,j} = \mathbb{E}[X^e_j \mid X^e_{R\setminus\{j\}}]$. For frequencies $t \in \mathbb{R}$ and $\omega \in \mathbb{R}^{|R|-1}$, we define two types of characteristic functions $\phi^e_{R,j}(t) = \mathbb{E}[\exp(\mathrm{i}t\varepsilon^e_{R,j})]$ and $\rho^e_{R,j}(\omega) = \mathbb{E}[\exp(\mathrm{i}\omega^\top X^e_{R\setminus\{j\}})]$, where $\mathrm{i}^2 = -1$. Multiplying the centered complex exponential function measures how far they deviate from independence, but estimating $m^e_{R,j}$ also changes the residual feature. Specifically, we use the following:

$$\chi^e_{R,j}(t,\omega) \coloneqq \{\exp(\mathrm{i}t\varepsilon^e_{R,j}) - \phi^e_{R,j}(t) - \mathrm{i}t\phi^e_{R,j}(t)\varepsilon^e_{R,j}\}\{\exp(\mathrm{i}\omega^\top X^e_{R\setminus\{j\}}) - \rho^e_{R,j}(\omega)\}. \tag{10}$$

At residual independence, the conditional expectation of the derivative of $\exp(\mathrm{i}t\varepsilon^e_{R,j})$ with respect to $\varepsilon^e_{R,j}$ is $\mathbb{E}_e[\mathrm{i}t\exp(\mathrm{i}t\varepsilon^e_{R,j}) \mid X^e_{R\setminus\{j\}}] = \mathrm{i}t\phi^e_{R,j}(t)$. Therefore, subtracting $\mathrm{i}t\phi^e_{R,j}(t)\varepsilon^e_{R,j}$ cancels the first-order effect of estimating $m^e_{R,j}$ when the residual is independent of its regressors. The following lemma shows that this correction preserves the characterization of residual independence.

Lemma 1 (Characteristic residual moments)**.** *Under the integrability conditions required to differentiate under the expectation, the score in Eq.* (10) *has expectation*

$$M^e_{R,j}(t,\omega) \coloneqq \mathbb{E}_e[\chi^e_{R,j}(t,\omega)] = \mathbb{E}_e[\exp\{\mathrm{i}(t\varepsilon^e_{R,j} + \omega^\top X^e_{R\setminus\{j\}})\}] - \phi^e_{R,j}(t)\rho^e_{R,j}(\omega).$$

This moment vanishes for every (t,ω) if and only if $\varepsilon^e_{R,j} \perp\!\!\!\perp X^e_{R\setminus\{j\}}$. At residual independence, its real and imaginary components are Neyman-orthogonal with respect to $(m^e_{R,j}, \phi^e_{R,j}(t), \rho^e_{R,j}(\omega))$.

Structure shift detection. Given a valid common order π, Proposition 1 reduces parent identification to deciding which conditional-mean contrasts are nonzero. Directly squaring a fitted contrast can turn regression error into an apparent contribution even for a nonparent. Following the projected-covariance principle of Lundborg et al. (2024), we instead learn a direction approximating the

contrast and evaluate its association with the response on a separate sample. This separates learning a direction that can reveal a parent from assessing the evidence along that direction. Let $g^e_{j,k}$ be a fixed square-integrable function of $X^e_{P_\pi(j)}$, called the *direction*. To isolate the contribution beyond the other predecessors, center both the response and the direction conditionally on $X^e_{P_\pi(j)\setminus\{k\}}$. With $\mu^e_{j,k} = \mathbb{E}_e[X^e_j \mid X^e_{P_\pi(j)\setminus\{k\}}]$ and $\nu^e_{j,k} = \mathbb{E}_e[g^e_{j,k} \mid X^e_{P_\pi(j)\setminus\{k\}}]$, the score $\theta^e_{j,k}$ is defined as follows:

$$\psi^e_{j,k} \coloneqq (X^e_j - \mu^e_{j,k})(g^e_{j,k} - \nu^e_{j,k}), \qquad \theta^e_{j,k} \coloneqq \mathbb{E}_e[\psi^e_{j,k}]. \tag{11}$$

The following lemma relates the population score to the conditional-mean contrast and establishes orthogonality with respect to the two centering functions. Specifically, this result shows that estimation errors in the two centering functions affect the population score only through the expectation of their product, allowing flexible regression fits under suitable product-rate conditions.

Lemma 2 (Projected-covariance identity and orthogonality)**.** *For any fixed square-integrable direction $g^e_{j,k}$, $\theta^e_{j,k} = \mathbb{E}_e[\delta^e_{j,k} g^e_{j,k}]$. In particular, $\delta^e_{j,k} = 0$ implies $\theta^e_{j,k} = 0$, while $g^e_{j,k} = \delta^e_{j,k}$ gives $\theta^e_{j,k} = \Delta^e_{j,k}$. The moment is Neyman-orthogonal with respect to $(\mu^e_{j,k}, \nu^e_{j,k})$.*

Estimation from finite samples. Lemma 2 and Proposition 1 establish that causal structure shifts are identified by whether the population score is zero or nonzero across environments. In practice, estimating population scores from finite samples requires learning several conditional regression functions that serve as nuisance functions. Orthogonality reduces sensitivity to estimation errors in these nuisance functions. To further improve robustness, we employ a cross-fitting strategy (Chernozhukov et al., 2018) (see Algorithms 1 and 2). We also approximate the criterion in Lemma 1 on a dictionary $\mathcal{D}_{n,R} = \{(t_b, \omega_b)\}_{b=1}^B$ of frequency pairs at each remaining set R. Specifically, we aggregate them by their root mean square (RMS) over the B pairs and select a node minimizing

$$\widehat{D}_{n,R}(j) \coloneqq \max_{e\in[E]} \left\{ \frac{1}{B} \sum_{b=1}^{B} \left| \widehat{M}^e_{R,j}(t_b, \omega_b) \right|^2 \right\}^{1/2}. \tag{12}$$

The maximum across environments retains evidence against a candidate even when only one environment reveals its dependence. We remove the minimizer, repeat on the remaining nodes, and reverse the removal sequence to obtain an estimated order $\widehat{\pi}$. For parent-set estimation, we use $\widehat{\pi}$ to align candidate edges across environments and split each environment sample into a direction-learning sample and a support-evaluation sample. On the direction-learning sample, we fit $\widehat{g}^e_{j,k}$ to approximate the contrast in Eq. (5) under $\widehat{\pi}$. Holding this direction fixed, we cross-fit the two centering regressions for $(\mu^e_{j,k}, \nu^e_{j,k})$ on the support-evaluation sample and average the held-out scores in Eq. (11). We studentize these score means and use the $(1-\alpha)$ quantile of a Gaussian multiplier maximum across candidate edges and environments as the common one-sided threshold, where $\alpha \in (0,1)$ is the significance level (Appendix C.2).

4 Experiments

4.1 Synthetic Datasets

Setup. We generated samples from nonlinear ANMs on $d \in \{10, 20\}$ variables in two environments, with $n \in \{300, 600, 900, 1200\}$ observations per environment. We used Erdős–Rényi (ER) (Erdös & Rényi, 1960) and scale-free (SF) (Barabasi & Albert, 1999) graphs with edge factors $r \in \{1, 2\}$. Each structural function was a neural network with one hidden layer of 16 units and $\tanh$ activation functions. To introduce causal structure shifts, we selected four nonroot nodes in the first environment and deleted up to two incoming edges per selected node in the second, while the resulting graphs shared a causal order. We used Gaussian, exponential, Laplace, and Gumbel noise. The exponential setting was used to evaluate empirical performance beyond our noise assumptions.

We compared STORM with iSCAN (Chen et al., 2023), ORION (Mian et al., 2023), SCORE (Rolland et al., 2022), DCI (Wang et al., 2018), and MG-LiNGAM (Shimizu, 2012). For SCORE and ORION, we estimated causal structure shifts by taking the directed symmetric difference of their estimated causal graphs in each environment. To validate the contribution of the order-stage correction, we prepared an ablated variant *w/o correction*, which removes the correction term from Eq. (10) while retaining cross-fitting and the same parent-set estimation procedure. We used kernel ridge regression

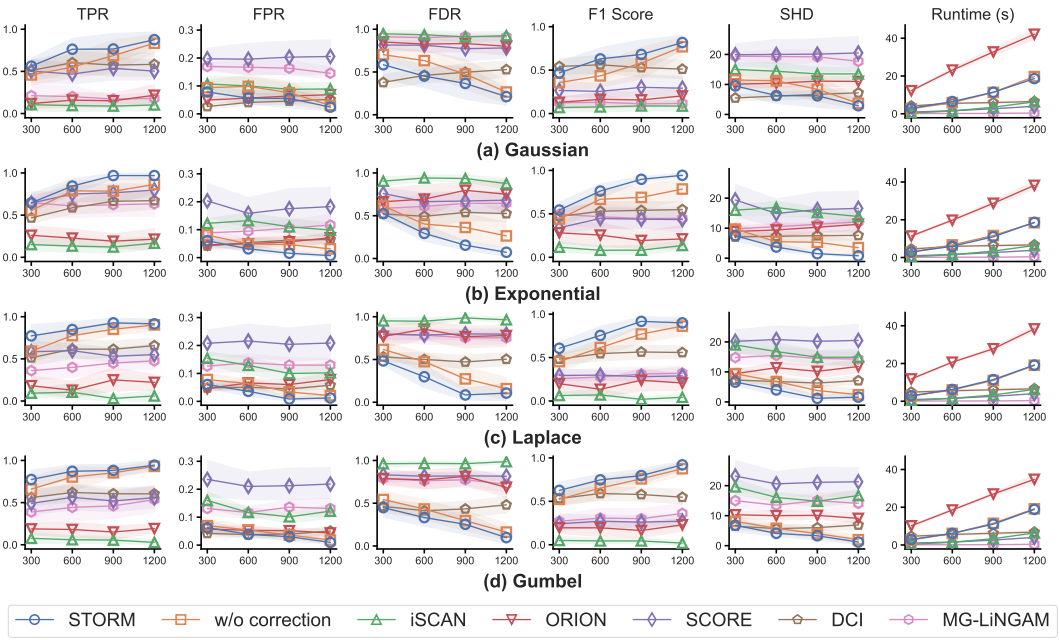


Figure 2: Structure shift detection accuracy and runtime on ER2 graphs with $d = 10$ variables. Curves show means over ten random seeds, with shaded pointwise 95% percentile confidence intervals.

with an RBF kernel for conditional means. Order recovery used five-fold cross-fitting, and parent-set estimation used two-fold cross-fitting on the support-evaluation sample. The support threshold was calibrated using 1999 Gaussian multiplier draws at a nominal level of $\alpha = 0.05$, with the plus-one correction (Phipson & Smyth, 2010) applied to the Monte Carlo p-values. All synthetic experiments were conducted with 10 random seeds. Detailed experimental settings are in Appendix E.

Structure shift detection. Figure 2 shows structure shift detection accuracy and computational resources on synthetic data generated from ER2 graphs. Additional results for other graphs are in Appendix F.1. For STORM, overall performance generally improved with larger sample sizes across the four noise distributions. DCI was competitive with our algorithm at the small sample size, but improved less thereafter because additional data cannot remove the mismatch between its linear regressions and the nonlinear generating mechanisms. In contrast, iSCAN detected few true shifts, possibly because errors in estimating log-density Hessians obscured the signals used to identify shifted mechanisms. We can also see that the full method detects structure shifts more accurately than *w/o correction*, meaning that the correction generally improves the accuracy of causal structure shift detection. By removing the first-order effect of regression estimation error, the correction may stabilize order recovery and reduce errors propagated to parent-set comparison, which is consistent with the analysis in Lemma 1. The smaller gap at the largest sample size is consistent with improved regression estimates reducing the need for the order-stage correction.

4.2 Real-World Datasets

We qualitatively evaluated the real-world results using biological evidence for single-cell signaling data from mass and flow cytometry and known interventions in a light-tunnel apparatus. Results on flow-cytometry data are provided in Appendix F.5.

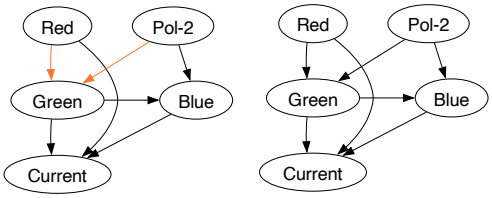
Mass cytometry. We evaluated the performance of STORM on the mass-cytometry dataset obtained from the Single Cell Signaling in Breast Cancer (SCSBrC) DREAM challenge (Tognetti et al., 2021; Gabor et al., 2021). This dataset contains 31 phosphorylation markers in breast cancer BT-20 cells from biological replicate A, where protein signals were measured 9 minutes after stimulation with epidermal growth factor (EGF) and serum. We compared 9,925 cells without inhibitor pretreatment (EGF condition) against 8,360 cells pretreated with the selective MEK inhibitor CI-1040 (iMEK condition). The estimated cellular signaling networks shared 135 edges and yielded 96 structure shifts.

Table 1: Selected estimated edges in BT-20 cells.

Inferred relationship	EGF	iMEK
$\{\text{MEK}, \text{ERK}, \text{p90RSK}\} \to \text{4EBP1}$	✓	-
$\{\text{AKT(S473)}, \text{AKT(T308)}\} \to \text{4EBP1}$	✓	✓
$\text{ERK} \to \text{p90RSK}$	✓	✓

Table 1 highlights selected estimated shifts around 4EBP1: the MEK, ERK, and p90RSK inputs were detected only in EGF, whereas inputs from both AKT phosphorylation sites were detected in both conditions. These results are consistent with previous perturbation experiments (Romeo et al., 2013; She et al., 2010). The mean 4EBP1 signal changed little, whereas ERK and p90RSK phosphorylation decreased significantly. The shared AKT $\to$ 4EBP1 edges suggest continued AKT-associated regulation as a candidate explanation for the maintained 4EBP1 level, motivating a test with additional PI3K/AKT inhibition. Meanwhile, ERK $\to$ p90RSK remained detected in both conditions despite the large phosphorylation decreases. This direction agrees with experiments showing that disruption of ERK docking impairs EGF-induced RSK activation (Roux et al., 2003). Together, these results illustrate how STORM identifies candidate input changes that are inconspicuous in a downstream mean, while retaining relationships despite large phosphorylation decreases. Further experimental results and biological discussions are provided in Appendix F.4.

Light tunnel. We also evaluated STORM on the light-tunnel apparatus from Gamella et al. (2025). The tunnel contains a controllable RGB light source and a camera at opposite ends. Two rotatable linear polarizers are positioned between them. Each reference and intervention environment contained 5,000 samples on five variables: the programmed RGB settings (Red, Green, and Blue), the second polarizer's angle setting (Pol-2), and the measured LED current (Current). The *green intervention* deletes Red $\to$ Green and Pol-2 $\to$ Green from the input generating mechanism and changes Green's exogenous noise distribution. The *red intervention* changes the exogenous noise distribution of the root variable Red while preserving the causal structure. The reader can find details in the original paper. Figure 3 shows that the detected differences matched the programmed structural intervention while the noise-only control produced no structure shifts.



(a) green intervention (b) red intervention

Figure 3: Estimated structure shifts from light tunnel data. Orange edges mean existence only in the reference environment.

5 Conclusion

This paper studied *causal structure shifts* in nonlinear ANMs with a common causal order. Under the stated assumptions, we established the identifiability of a valid common order and parent sets in each environment, even when the order is not unique. We also specified a model class preserved by order-compatible shifts without separate conditions on the induced conditional input distributions. Our algorithm, STORM, estimates a common causal order and compares parent sets using orthogonal scores, sample splitting, and cross-fitting. Synthetic experiments generally showed higher F1 than the compared baselines and supported the contribution of the order-stage correction. In the light-tunnel experiment, detected shifts matched the programmed edge deletions under the structural intervention, whereas no candidates were reported under the noise-only control. The mass-cytometry analysis yielded biologically plausible candidates for targeted experimental investigation.

Limitations. Our approach has some limitations that may be worth addressing in the future. First, this work assumes no unobserved confounders and union graph acyclicity. Then, extending to latent confounding and feedback (Bongers et al., 2021) would broaden its applicability to partially observed signaling systems, where feedback can mediate responses to inhibition (Turke et al., 2012). Second, post-selection family-wise error control for edge detection is not established for our implementation. Differences in statistical significance do not by themselves establish differences in the underlying quantities (Gelman & Stern, 2006). In STORM, unequal detection power can yield a candidate shift even when an edge is present in both environments. Finite-sample theory accounting for common-order recovery, regression error, and edge strength would clarify the conditions and sample sizes needed for reliable structure shift detection.

AI USE STATEMENT

In this work, we used generative AI tools to formulate mathematical claims, provide critical ingredients for proving mathematical claims, assist in the writing of proofs, design or provide feedback on research methodology or experiments, implement methods, assist with translation, support qualitative and thematic data analysis, interpret results. We have not used generative AI tools to generate synthetic data sets, help develop theoretical models or conceptual frameworks, propose or refine hypotheses, clean and reformat dataset. Additionally, we used generative AI tools for brainstorming, sourcing/searching for information, editing a research paper to improve readability, identifying relevant literature, formatting references. We have reviewed all AI-assisted work. For example, LLM-generated code was read and tested carefully and LLM-provided information was verified using primary sources by the first author. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Emre Acartürk, Burak Varı cı, Karthikeyan Shanmugam, and Ali Tajer. Sample complexity of interventional causal representation learning. In A Globerson, L Mackey, D Belgrave, A Fan, U Paquet, J Tomczak, and C Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 39350–39385. Curran Associates, Inc., 2024.

Panagiotis Alimisis and Christos Diou. Learning robust intervention representations with delta embeddings. In *The Fourteenth International Conference on Learning Representations*, 2026.

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

Norman P Archer and Shouhong Wang. Application of the back propagation neural network algorithm with monotonicity constraints for two-group classification problems. *Decision Sciences*, 24(1): 60–75, 1993.

Mona Azadkia and Sourav Chatterjee. A simple measure of conditional dependence. *The Annals of Statistics*, 49(6):3070–3102, 2021.

A L Barabasi and R Albert. Emergence of scaling in random networks. *Science (New York, N.Y.)*, 286(5439):509–512, 15 October 1999.

Stephan Bongers, Patrick Forré, Jonas Peters, and Joris M Mooij. Foundations of structural causal models with cycles and latent variables. *The Annals of Statistics*, 49(5):2885–2915, 2021.

Deborah Brancho, Nobuyuki Tanaka, Anja Jaeschke, Juan-Jose Ventura, Nyaya Kelkar, Yoshinori Tanaka, Masanao Kyuuma, Toshikazu Takeshita, Richard A Flavell, and Roger J Davis. Mechanism of p38 MAP kinase activation in vivo. *Genes & Development*, 17(16):1969–1978, 15 August 2003.

Simon Buchholz, Goutham Rajendran, Elan Rosenfeld, Bryon Aragam, Bernhard Schölkopf, and Pradeep Ravikumar. Learning linear causal representations from interventions under general nonlinear mixing. In A Oh, T Naumann, A Globerson, K Saenko, M Hardt, and S Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 45419–45462. Curran Associates, Inc., 2023.

Tianyu Chen, Kevin Bello, Bryon Aragam, and Pradeep Ravikumar. iSCAN: Identifying causal mechanism shifts among nonlinear additive noise models. In *Advances in Neural Information Processing Systems*, volume 36, pp. 44671–44706. Curran Associates, Inc., 2023.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, February 2018.

Victor Chernozhukov, Juan Carlos Escanciano, Hidehiko Ichimura, Whitney K Newey, and James M Robins. Locally robust semiparametric estimation. *Econometrica*, 90(4):1501–1535, 2022.

Max Chickering. A transformational characterization of equivalent bayesian network structures. In *In Proceedings of Eleventh Conference on Uncertainty in Artificial Intelligence*, January 1995.

Barry J Collins, Maria Deak, J Simon C Arthur, Laura J Armit, and Dario R Alessi. In vivo role of the PIF-binding docking site of PDK1 defined by knock-in mutation. *The EMBO Journal*, 22(16): 4202–4211, 15 August 2003.

P Erdös and A Rényi. On the evolution of random graphs. *Publications of the Mathematical Institute of the Hungarian Academy of Sciences*, 5:17–61, 1960.

Attila Gabor, Marco Tognetti, Alice Driessen, Jovan Tanevski, Baosen Guo, Wencai Cao, He Shen, Thomas Yu, Verena Chung, Single Cell Signaling in Breast Cancer DREAM Consortium members, Bernd Bodenmiller, and Julio Saez-Rodriguez. Cell-to-cell and type-to-type heterogeneity of signaling networks: insights from the crowd. *Molecular Systems Biology*, 17(10):e10402, October 2021.

Juan L Gamella, Jonas Peters, and Peter Bühlmann. Causal chambers as a real-world physical testbed for AI methodology. *Nature Machine Intelligence*, 7(1):107–118, 1 January 2025.

Andrew Gelman and Hal Stern. The difference between "significant" and "not significant" is not itself statistically significant. *The American Statistician*, 60(4):328–331, 2006.

Amiremad Ghassami, Negar Kiyavash, Biwei Huang, and Kun Zhang. Multi-domain causal structure learning in linear systems. In *Advances in Neural Information Processing Systems*, volume 31, pp. 6266–6276, December 2018.

Biwei Huang, Kun Zhang, Jiji Zhang, J Ramsey, Ruben Sanchez-Romero, C Glymour, and B Scholkopf. Causal discovery from heterogeneous/NOnstationary data. *Journal of machine learning research*, 21(89):89:1–89:53, 2020.

Alexander Hägele, Jonas Rothfuss, Lars Lorch, Vignesh Ram Somnath, Bernhard Schölkopf, and Andreas Krause. BaCaDI: Bayesian causal discovery with unknown interventions. In Francisco Ruiz, Jennifer Dy, and Jan-Willem van de Meent (eds.), *International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pp. 1411–1436. PMLR, 2023.

Sebastien Lachapelle, Pau Rodriguez, Yash Sharma, Katie E Everett, Rémi L E Priol, Alexandre Lacoste, and Simon Lacoste-Julien. Disentanglement via mechanism sparsity regularization: A new principle for nonlinear ICA. In Bernhard Schölkopf, Caroline Uhler, and Kun Zhang (eds.), *Proceedings of the First Conference on Causal Learning and Reasoning*, volume 177 of *Proceedings of Machine Learning Research*, pp. 428–484. PMLR, 2022.

Sébastien Lachapelle, Philippe Brouillard, Tristan Deleu, and Simon Lacoste-Julien. Gradient-based neural DAG learning. In *International Conference on Learning Representations*, 2020.

Anton Rask Lundborg, Ilmun Kim, Rajen D Shah, and Richard J Samworth. The projected covariance measure for assumption-lean variable significance testing. *The Annals of Statistics*, 52(6):2851–2878, 2024.

Sarah Mameche, David Kaltenpoth, and Jilles Vreeken. Discovering invariant and changing mechanisms from data. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '22, pp. 1242–1252, New York, NY, USA, 14 August 2022. Association for Computing Machinery.

Sarah Mameche, Lénaïg Cornanguer, Urmi Ninad, and Jilles Vreeken. SpaceTime: Causal discovery from non-stationary time series. In *AAAI conference on artificial intelligence*, 17 January 2025.

Osman Mian, Michael Kamp, and Jilles Vreeken. Information-theoretic causal discovery and intervention detection over multiple environments. In Brian Williams, Yiling Chen, and Jennifer Neville (eds.), *Proceedings of the 37th AAAI Conference on Artificial Intelligence*, volume 37, pp. 9171–9179, Washington, DC, USA, June 2023. Association for the Advancement of Artificial Intelligence (AAAI).

Olga K Mirzoeva, Debopriya Das, Laura M Heiser, Sanchita Bhattacharya, Doris Siwak, Rina Gendelman, Nora Bayani, Nicholas J Wang, Richard M Neve, Yinghui Guan, Zhi Hu, Zachary Knight, Heidi S Feiler, Philippe Gascard, Bahram Parvin, Paul T Spellman, Kevan M Shokat, Andrew J Wyrobek, Mina J Bissell, Frank McCormick, Wen-Lin Kuo, Gordon B Mills, Joe W

Gray, and W Michael Korn. Basal subtype and MAPK/ERK kinase (MEK)-phosphoinositide 3-kinase feedback signaling determine susceptibility of breast cancer cells to MEK inhibition. *Cancer Research*, 69(2):565–572, 15 January 2009.

Francesco Montagna. On the identifiability of causal graphs with the invariance principle. In *International Conference on Learning Representations*, volume 2026, pp. 13837–13868, 2026.

Judea Pearl. *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. Morgan Kaufmann, San Mateo, CA, 1988.

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

Ronan Perry, Julius von Kügelgen, and B Scholkopf. Causal discovery in heterogeneous environments under the sparse mechanism shift hypothesis. In *Advances in Neural Information Processing Systems*, volume 35, pp. 10904–10917, 4 June 2022.

Jonas Peters, Joris Mooij, Dominik Janzing, and Bernhard Schölkopf. Causal discovery with continuous additive noise models. *Journal of Machine Learning Research*, 2014.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of Causal Inference: Foundations and Learning Algorithms*. MIT Press, 2017. ISBN 9780262037310. URL https://mitpress.mit.edu/9780262037310/elements-of-causal-inference/.

Belinda Phipson and Gordon K Smyth. Permutation P-values should never be zero: Calculating exact P-values when permutations are randomly drawn. *Statistical Applications in Genetics and Molecular Biology*, 9(1), 2010.

Patrik Reizinger, Siyuan Guo, Ferenc Huszar, Bernhard Schölkopf, and Wieland Brendel. Identifiable exchangeable mechanisms for causal structure and representation learning. In Y Yue, A Garg, N Peng, F Sha, and R Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 62196–62223, 2025.

J Rissanen. Modeling by shortest data description. *Automatica: the journal of IFAC, the International Federation of Automatic Control*, 14(5):465–471, September 1978.

Paul Rolland, V Cevher, Matthäus Kleindessner, Chris Russel, B Scholkopf, D Janzing, and Francesco Locatello. Score matching enables causal discovery of nonlinear additive noise models. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162, pp. 18741–18753. PMLR, 8 March 2022.

Y Romeo, J Moreau, P-J Zindy, M Saba-El-Leil, G Lavoie, F Dandachi, M Baptissart, K L B Borden, S Meloche, and P P Roux. RSK regulates activated BRAF signalling to mTORC1 and promotes melanoma growth. *Oncogene*, 32(24):2917–2926, 1 June 2013.

Dominik Rothenhäusler, Christina Heinze, Jonas Peters, and Nicolai Meinshausen. BACKSHIFT: Learning causal cyclic graphs from unknown shift interventions. In C Cortes, N Lawrence, D Lee, M Sugiyama, and R Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015.

Philippe P Roux, Stephanie A Richards, and John Blenis. Phosphorylation of p90 ribosomal S6 kinase (RSK) regulates extracellular signal-regulated kinase docking and RSK activity. *Molecular and Cellular Biology*, 23(14):4796–4804, July 2003.

Philippe P Roux, David Shahbazian, Hieu Vu, Marina K Holz, Michael S Cohen, Jack Taunton, Nahum Sonenberg, and John Blenis. RAS/ERK signaling promotes site-specific ribosomal protein S6 phosphorylation via RSK and stimulates cap-dependent translation. *The Journal of Biological Chemistry*, 282(19):14056–14064, 11 May 2007.

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.

Adrien Saumard and Jon A Wellner. Log-concavity and strong log-concavity: A review. *Statistics Surveys*, 8:45–114, 2014.

Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021.

Qing-Bai She, Ensar Halilovic, Qing Ye, Wei Zhen, Senji Shirasawa, Takehiko Sasazuki, David B Solit, and Neal Rosen. 4E-BP1 is a key effector of the oncogenic activation of the AKT and ERK signaling pathways that integrates their function in tumors. *Cancer Cell*, 18(1):39–51, 13 July 2010.

Shohei Shimizu. Joint estimation of linear non-gaussian acyclic models. *Neurocomputing*, 81: 104–107, April 2012.

Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, prediction, and search*. Lecture Notes in Statistics. Springer-Verlag, New York, 1st edition, 4 March 1993.

Chandler Squires, Anna Seigal, Salil S Bhate, and Caroline Uhler. Linear causal disentanglement via interventions. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 32540–32560. PMLR, 2023.

Eric V Strobl and Shyam Visweswaran. Dirac delta regression: Conditional density estimation with clinical trials. In Thuc Duy Le, Jiuyong Li, Greg Cooper, Sofia Triantafyllou, Elias Bareinboim, Huan Liu, and Negar Kiyavash (eds.), *Proceedings of The KDD'21 Workshop on Causal Discovery*, volume 150 of *Proceedings of Machine Learning Research*, pp. 78–125. PMLR, 15 August 2021.

Marco Tognetti, Attila Gabor, Mi Yang, Valentina Cappelletti, Jonas Windhager, Oscar M Rueda, Konstantina Charmpi, Elham Esmaeilishirazifard, Alejandra Bruna, Natalie de Souza, Carlos Caldas, Andreas Beyer, Paola Picotti, Julio Saez-Rodriguez, and Bernd Bodenmiller. Deciphering the signaling network of breast cancer improves drug sensitivity prediction. *Cell Systems*, 12(5): 401–418.e12, 19 May 2021.

Alexa B Turke, Youngchul Song, Carlotta Costa, Rebecca Cook, Carlos L Arteaga, John M Asara, and Jeffrey A Engelman. MEK inhibition leads to PI3K/AKT activation by relieving a negative feedback on ERBB receptors. *Cancer Research*, 72(13):3228–3237, 1 July 2012.

Thomas S Verma and Judea Pearl. Equivalence and synthesis of causal models. In *Proceedings of the Sixth Conference on Uncertainty in Artificial Intelligence*, UAI '90, pp. 220–227, Cambridge, MA, USA, 1990. Association for Uncertainty in Artificial Intelligence.

Yuhao Wang, Chandler Squires, Anastasiya Belyaeva, and Caroline Uhler. Direct estimation of differences in causal graphs. In S Bengio, H Wallach, H Larochelle, K Grauman, N Cesa-Bianchi, and R Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

Andre Wibisono, Yihong Wu, and Kaylee Yingxi Yang. Optimal score estimation via empirical bayes smoothing. In Shipra Agrawal and Aaron Roth (eds.), *Proceedings of Thirty Seventh Conference on Learning Theory*, volume 247 of *Proceedings of Machine Learning Research*, pp. 4958–4991. PMLR, 2024.

Sascha Xu, Sarah Mameche, and Jilles Vreeken. Information-theoretic causal discovery in topological order. In Yingzhen Li, Stephan Mandt, Shipra Agrawal, and Emtiyaz Khan (eds.), *Proceedings of The 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pp. 2008–2016. PMLR, 2025.

APPENDIX

A Notation and Terminology **15**

B Preliminaries **16**

B.1 Causal Graphical Models 16

B.2 Distributional Properties 19

C Estimation Procedure **20**

C.1 Nuisance Estimation via Cross-Fitting 20

C.2 Multiplier Maximum for Parent-Set Estimation Stage 21

D Proofs **21**

D.1 Proof of Theorem 1 21

D.2 Proof of Lemma 1 25

D.3 Proof of Proposition 1 27

D.4 Proof of Lemma 2 27

D.5 Proof of Theorem 2 27

E Experimental Setup **28**

E.1 Data Generating Process 28

E.2 Baselines 29

E.3 Implementation Details 29

E.4 Evaluation Metrics 30

F Additional Results **31**

F.1 Structure Shift Detection 31

F.2 Common-Order Recovery 35

F.3 Regression Learner Sensitivity 35

F.4 Further Results and Discussion on Mass-Cytometry Data 37

F.5 Results on Flow-Cytometry Data 39

A NOTATION AND TERMINOLOGY

Number and Indices	
n	Number of observations per environment
d	Number of variables
E	Number of environments
e	Index of an environment, $e \in [E]$
j, k	Indices of a target and a candidate parent, $j, k \in V$
r	Index of an observation in the support-evaluation sample, $r \in \mathcal{I}_s^e$
n_s^e	Number of observations in the support-evaluation sample, $n_s^e = \lvert\mathcal{I}_s^e\rvert$
Random Variables	
X^e	Vector of observed variables in environment e, i.e., $X^e = (X_1^e, \dots, X_d^e)$
N_j^e	Additive noise of X_j^e
$\varepsilon_{R,j}^e$	Conditional-mean residual, i.e., $\varepsilon_{R,j}^e = X_j^e - m_{R,j}^e$
$\delta_{j,k}^{e,\pi}$	Conditional-mean contrast for X_j^e, defined in Eq. (5)
$\Delta_{j,k}^{e,\pi}$	Population edge strength, i.e., $\Delta_{j,k}^{e,\pi} = \mathbb{E}_e[(\delta_{j,k}^{e,\pi})^2]$
ξ_r^e	Gaussian multiplier for observation r in environment e
Estimated Quantities	
$\widehat{\pi}$	Estimated common order
$\widehat{M}_{R,j}^e(t,\omega)$	Estimate of the characteristic-residual moment $M_{R,j}^e(t,\omega)$
$\widehat{g}_{j,k}^e$	Learned full-minus-reduced regression direction, fitted on the direction-learning sample and fixed across support-evaluation folds
$\widehat{\mu}_{j,k}^e, \widehat{\nu}_{j,k}^e$	Fold-specific centering regressions for the response and learned direction $\widehat{g}_{j,k}^e$, respectively
$\widehat{\psi}_{j,k,r}^e$	Cross-fitted projected-covariance score for observation r
$\bar{\psi}_{j,k}^e$	Sample mean of $\widehat{\psi}_{j,k,r}^e$
$\widehat{\sigma}_{j,k}^e$	Sample standard deviation of $\widehat{\psi}_{j,k,r}^e$
Functions and Operators	
M	Structural causal model, i.e., $M = (G, f, \mathbb{P}_N)$
$\mathcal{M}$	Space of structural causal model M
$\mathbb{P}_X^M$	Observational distribution induced by M
$\mathbb{P}_N$	Joint distribution of the exogenous noises in M
$\mathbb{E}_e$	Expectation in environment e
f	Collection of structural functions, i.e., $f = (f_j)_{j \in V}$
f_j^e	Structural function of node j in environment e
$\mathrm{pa}_e(j)$	Parent set of node j in environment e
$\mathrm{Top}(G)$	Set of common orders of G
q_j^e	Probability density of the exogenous noise N_j^e in environment e
o	Structure-shift operation, i.e., $o : \mathcal{M} \to \mathcal{M}$
$\mathcal{O}$	Class of order-compatible operations
$P_\pi(j)$	Predecessor set of j, i.e., $P_\pi(j) = \{\ell \in V : \ell \prec_\pi j\}$
$m_{R,j}^e$	Conditional mean, i.e., $m_{R,j}^e = \mathbb{E}_e[X_j^e \mid X_{R\setminus\{j\}}^e]$
$\phi_{R,j}^e(t)$	Residual characteristic function, i.e., $\phi_{R,j}^e(t) = \mathbb{E}_e[\exp(\mathrm{i}t\varepsilon_{R,j}^e)]$
$\rho_{R,j}^e(\omega)$	Regressor characteristic function, i.e., $\rho_{R,j}^e(\omega) = \mathbb{E}_e[\exp(\mathrm{i}\omega^\top X_{R\setminus\{j\}}^e)]$
$\mu_{j,k}^e$	Reduced conditional mean, i.e., $\mu_{j,k}^e = \mathbb{E}_e[X_j^e \mid X_{P_\pi(j)\setminus\{k\}}^e]$
$\nu_{j,k}^e$	Conditional mean for the direction $g_{j,k}^e$, i.e, $\nu_{j,k}^e = \mathbb{E}_e[g_{j,k}^e \mid X_{P_\pi(j)\setminus\{k\}}^e]$
Common-Order Recovery	
G^e	Causal DAG in environment e, i.e., $G^e = (V, \mathcal{E}^e)$
$G_\cup$	Union graph, i.e., $G^\cup = (V, \bigcup_{e=1}^E \mathcal{E}^e)$

Common-Order Recovery (continued)	
Π_0	Set of valid common orders, i.e., $\Pi_0 = \bigcap_{e=1}^{E} \mathrm{Top}(G^e)$
R	Set of remaining nodes during sink removal
$\mathcal{D}_{n,R}$	Shared dictionary of residual and regressor frequencies (t, ω)
$\chi^e_{R,j}(t,\omega)$	Corrected characteristic-residual score, defined in Eq. (10)
$M^e_{R,j}(t,\omega)$	Population characteristic-residual moment $\mathbb{E}_e[\chi^e_{R,j}(t,\omega)]$
Structure Shift Detection	
$S_j^{e_1,e_2}$	Causal structure shift set $\mathrm{pa}_{e_1}(j) \triangle \mathrm{pa}_{e_2}(j)$
$g^e_{j,k}$	Fixed square-integrable direction, a function of $X^e_{P_\pi(j)}$
$\psi^e_{j,k}$	Projected-covariance score, i.e., $\psi^e_{j,k} = (X^e_j - \mu^e_{j,k})(g^e_{j,k} - \nu^e_{j,k})$
$\theta^e_{j,k}$	Population score mean, i.e., $\theta^e_{j,k} = \mathbb{E}_e[\psi^e_{j,k}] = \mathbb{E}_e[\delta^{e,\pi}_{j,k} g^e_{j,k}]$
$\widehat{T}^e_{j,k}$	Studentized projected-covariance statistic $\sqrt{n^e_s}\,\bar{\psi}^e_{j,k}/\widehat{\sigma}^e_{j,k}$
$T^*_{\mathrm{str,max}}$	Gaussian multiplier maximum, calculated using Eq. (20)
$\widehat{c}_{\mathrm{str},1-\alpha}$	Conditional $(1-\alpha)$ quantile of $T^*_{\mathrm{str,max}}$

B PRELIMINARIES

B.1 CAUSAL GRAPHICAL MODELS

We retain the notation of §2: $G = (V, \mathcal{E}_G)$ is a DAG on the observed variables $X_V = (X_v)_{v \in V}$. The terminology below is standard in causal graphical models (Pearl, 2009).

Graph basics. Two distinct nodes are *adjacent* if an edge joins them, irrespective of its direction. A path is a sequence of distinct nodes $(v_0, \ldots, v_m)$ in which consecutive nodes are adjacent; its length is m. It is a *directed path* if $v_{r-1} \to v_r$ for every r; we write $v_0 \rightsquigarrow v_m$ only for a directed path of positive length. Closing a directed path with $v_m \to v_0$ produces a directed cycle, which a DAG by definition cannot contain. For $v \in V$, we use the following graph relations:

- The **parents** of v are $\mathrm{pa}_G(v) = \{u \in V : u \to v \in \mathcal{E}_G\}$.
- The **children** of v are $\mathrm{ch}_G(v) = \{u \in V : v \to u \in \mathcal{E}_G\}$.
- The **proper ancestors** of v are $\mathrm{an}_G(v) = \{u \in V : u \rightsquigarrow v\}$.
- The **proper descendants** of v are $\mathrm{de}_G(v) = \{u \in V : v \rightsquigarrow u\}$.
- The **nodes adjacent to** v are $\mathrm{adj}_G(v) = \mathrm{pa}_G(v) \cup \mathrm{ch}_G(v)$.
- The **non-descendants** of v are $\mathrm{nd}_G(v) = V \setminus \big(\{v\} \cup \mathrm{de}_G(v)\big)$.

Because a DAG has no self-loops and $\rightsquigarrow$ denotes a path of positive length, none of these six sets contains v itself. A node with no parents is a *source* (often called a root), while a node with no children is a *sink* (or leaf). A topological order is a bijection $\pi : V \to [d]$ such that $\pi(i) < \pi(j)$ whenever $i \to j \in \mathcal{E}_G$; every DAG admits at least one such order.

Let i, k, j be three consecutive distinct nodes on a path in a DAG. Relative to this path, the local orientation at the middle node k falls into exactly one of the following three configurations, where the two consistently oriented cases are grouped as a chain:

- A **chain structure** is $i \to k \to j$, and k is a non-collider on the path. For a causal query from the upstream endpoint to the downstream endpoint, k is a **mediator** on the displayed directed path. More generally, a mediator is an internal node on a directed path from a specified treatment to a specified outcome.
- A **fork structure** is $i \leftarrow k \to j$, and k is a non-collider and a direct common cause of i and j. Relative to a causal query concerning the effect of either endpoint on the other, k is commonly called a **confounder** for that query.
- A **collider structure** is $i \to k \leftarrow j$, and k is a **collider** on the path and a direct common effect of i and j. The triple is a **v-structure**, or unshielded collider, if and only if i and j are not adjacent.

Thus, along any fixed path in a DAG, each internal node is locally classified by one of these three configurations. This is not a global partition of the vertices: the same node can be a collider on one path and a non-collider on another. Moreover, the terms mediator and confounder are relative to a specified causal query.

Definition 2 (D-separation (Pearl, 1988)). *Let $A, B, S \subseteq V$ be pairwise disjoint. A path between A and B is* active *(or d-connecting) given S precisely when*

(i) every non-collider internal node on the path lies outside S; and
(ii) every collider c on the path satisfies $c \in S$ or $\mathrm{de}_G(c) \cap S \neq \varnothing$.

Otherwise the path is blocked. The sets A and B are d-separated *by S, written $A \perp_G B \mid S$, if every path between them is blocked; otherwise they are d-connected.*

Hence conditioning on k blocks the displayed chain or fork path. The displayed collider path is blocked when neither k nor any descendant of k is conditioned on, but conditioning on either can open that path. These statements concern the status of a particular path; without further assumptions they do not by themselves assert marginal or conditional dependence.

Connection between graphs and distributions. D-separation is a property of the graph, whereas conditional independence is a property of a probability distribution. The Markov property connects the two by translating graphical separations into conditional independences. In our ANMs, it follows from the mutually independent exogenous noises in §2; it is not a separate assumption.

Definition 3 (Markov property (Pearl, 2009)). *Writing $X_A = (X_v)_{v \in A}$, a distribution $\mathbb{P}$ is Markov to G if, for every pairwise disjoint $A, B, S \subseteq V$, graphical separation implies probabilistic conditional independence:*

$$A \perp_G B \mid S \quad \Longrightarrow \quad X_A \perp\!\!\!\perp X_B \mid X_S. \tag{13}$$

In particular, the local Markov property states

$$X_v \perp\!\!\!\perp X_{\mathrm{nd}_G(v) \setminus \mathrm{pa}_G(v)} \mid X_{\mathrm{pa}_G(v)}. \tag{14}$$

The local statement says that, given its parents, a variable is independent of its remaining nondescendants. For identification, we also require that the graph contain no redundant edges.

Assumption 4 (Causal minimality (Peters et al., 2014)). *For every $M \in \mathcal{M}$, the induced distribution $\mathbb{P}_X^M$ is not Markov with respect to any proper subgraph of $G(M)$ on the same vertex set.*

The stronger property of *faithfulness* requires every conditional independence in $\mathbb{P}$ to follow from d-separation in G (Spirtes et al., 1993; Peters et al., 2014). We do not impose this converse of Eq. (13): causal minimality rules out redundant edges but permits additional conditional independences. The parent-set identification proof in Appendix D.3 uses the Markov property and causal minimality together with the intersection property stated below.

Latent confounders. The preceding properties describe the relation between a graph and a distribution on the variables under consideration. A separate issue is whether unobserved variables are common causes of the observed variables. To explain how common causes can contribute to an association, consider a causal query concerning the effect of X_i on X_j. A *back-door path* is a path from i to j whose first edge points into i. If such a path is active under the chosen conditioning set, it provides a route for noncausal association. The fork $i \leftarrow c \rightarrow j$ introduced above is the elementary case. More generally, a common cause of i and j is a node in $\mathrm{an}_G(i) \cap \mathrm{an}_G(j)$. Because the term "confounder" is query-dependent, we use common causes and open back-door paths as the primitive notions. In an underlying causal DAG that may also contain latent nodes, an unobserved common cause of observed variables is often called a *latent confounder*. Since our graphs are defined on observed variables, we specify the exclusion of such shared latent causes through causal sufficiency.

Assumption 5 (Causal sufficiency (Spirtes et al., 1993)). *No unobserved variable is a common cause of two or more observed variables.*

Causal sufficiency excludes such shared latent causes; it does not require every latent variable in the data-generating system to be observed. Observed common causes and the back-door paths through them may still be present.

Combining conditional independences. Reasoning with conditional independence also requires relating statements that use different conditioning sets. The intersection property specifies when two

such statements imply conditional independence from the union of two variable sets, given a smaller conditioning set.

Definition 4 (Intersection property)**.** *A distribution satisfies the intersection property if, for every four pairwise disjoint index sets $A, B, C, D \subseteq V$,*

$$\left.\begin{aligned} X_A \perp\!\!\!\perp X_B \,|\, X_{C\cup D}, \\ X_A \perp\!\!\!\perp X_C \,|\, X_{B\cup D} \end{aligned}\right\} \quad \Longrightarrow \quad X_A \perp\!\!\!\perp X_{B\cup C} \,|\, X_D. \tag{15}$$

Strictly positive joint densities with respect to a product measure are sufficient for this property (Peters et al., 2014). Positivity only on the support does not suffice in general. In Appendix D.3, this property is used with the local Markov property and causal minimality to show that a child remains conditionally dependent on each parent when conditioning on its other parents and additional nondescendants of the child.

Causal Bayesian networks. A Bayesian network represents a joint distribution using a DAG and node-specific conditional distributions. Writing p for a probability density or mass function, the joint distribution factorizes as

$$p(x_V) = \prod_{v\in V} p_v\big(x_v \,|\, x_{\mathrm{pa}_G(v)}\big).$$

A *causal Bayesian network* (CBN) gives this factorization a causal interpretation through autonomous mechanisms (Pearl, 2009).

Definition 5 (Causal mechanism)**.** *In a CBN, the causal mechanism of node v is the local conditional distribution specified by the model for X_v given its parents, represented by $p_v(\,\cdot \,|\, x_{\mathrm{pa}_G(v)})$ for each parent configuration $x_{\mathrm{pa}_G(v)}$. Its autonomy means that an intervention replaces the mechanisms of its target nodes while preserving those of all other nodes.*

In particular, a perfect intervention $\mathrm{do}(X_I = a_I)$, for $I \subseteq V$, fixes X_I to a_I and yields the *truncated factorization*

$$p\big(x_{V\setminus I} \,|\, \mathrm{do}(X_I = a_I)\big) = \prod_{v\in V\setminus I} p_v\big(x_v \,|\, x_{\mathrm{pa}_G(v)}\big)\Big|_{x_I = a_I}.$$

The factors for nonintervened nodes are the same as in the observational distribution. A local conditional distribution can change while its parent set remains fixed. A node's marginal distribution can also change because of changes in upstream mechanisms, even when its own mechanism remains unchanged.

Structural causal models. An acyclic *structural causal model* (SCM) on G specifies structural assignments

$$X_v \coloneqq g_v\big(X_{\mathrm{pa}_G(v)}, N_v\big), \qquad v \in V,$$

together with a joint distribution $\mathbb{P}_N$ of the exogenous variables $N_V = (N_v)_{v\in V}$ (Pearl, 2009; Peters et al., 2017). The X_v are *endogenous* variables determined within the model, while the N_v are *exogenous* inputs whose causes are not specified within the model. Acyclicity allows the assignments to be evaluated uniquely in a topological order. Let h_g denote the resulting solution map from exogenous to endogenous variables, so that

$$X_V = h_g(N_V).$$

For measurable structural functions, the induced observational distribution $\mathbb{P}$ is the pushforward of $\mathbb{P}_N$ under h_g:

$$\mathbb{P} = (h_g)_{\#}\mathbb{P}_N, \qquad \mathbb{P}(A) = \mathbb{P}_N\big(\{n : h_g(n) \in A\}\big)$$

for every measurable set A. For a fixed SCM, this distribution is unique because recursive substitution determines all endogenous variables from the noises, as in the proof of Proposition 6.3 (Entailed distributions) in Peters et al. (2017).

If the exogenous variables are mutually independent, this distribution is Markov to G. Under this independence condition, the mechanism in Definition 5 is the distribution of $g_v(x_{\mathrm{pa}_G(v)}, N_v)$ at each fixed parent configuration, with N_v drawn from its marginal distribution $\mathbb{P}_{N_v}$. Thus, the structural function g_v and noise distribution $\mathbb{P}_{N_v}$ determine the local conditional distribution p_v; different such pairs can induce the same conditional distribution (Pearl, 2009). A perfect intervention $\mathrm{do}(X_v = x)$

replaces the assignment for X_v by $X_v := x$, leaving all other assignments and $\mathbb{P}_N$ unchanged; graphically, it removes the incoming edges of v. Under the same independence condition, this replacement yields the truncated factorization of the corresponding CBN (Pearl, 2009).

Conversely, if $\mathbb{P}$ has a density with respect to Lebesgue measure and is Markov to G, there exists an SCM on G with mutually independent noises that induces $\mathbb{P}$ (Peters et al., 2017). The proof in Peters et al. (2017) constructs the structural functions using conditional generalized inverse distribution functions and independent uniform noises. This representation is generally nonunique, and the construction does not impose an additive-noise form.

Example 1 (Nonlinear additive-noise model)**.** *An additive-noise model (ANM) is the special case $g_v(x, n) = f_v(x) + n$ with mutually independent noises. For the chain $1 \to 2 \to 3$, consider nonlinear functions f_2, f_3 and the assignments*

$$X_1 = N_1, \qquad X_2 = f_2(X_1) + N_2, \qquad X_3 = f_3(X_2) + N_3.$$

The solution map is

$$h_g(n_1, n_2, n_3) = \begin{pmatrix} n_1 \\ f_2(n_1) + n_2 \\ f_3\big(f_2(n_1) + n_2\big) + n_3 \end{pmatrix}.$$

The third component shows how local structural functions compose and propagate upstream noise. If the noises have densities p_{N_v}, the corresponding CBN mechanisms are

$$p_v\big(x_v \,|\, x_{\mathrm{pa}_G(v)}\big) = p_{N_v}\big(x_v - f_v(x_{\mathrm{pa}_G(v)})\big),$$

where $f_1 \equiv 0$ for the source node in this example.

Markov equivalence classes. Let $\mathcal{I}(G)$ collect all d-separation statements entailed by G. Two DAGs are *Markov equivalent* when their sets $\mathcal{I}(G)$ coincide. Equivalently, they have the same skeleton and the same v-structures (Verma & Pearl, 1990). Their Markov equivalence class (MEC) is represented uniquely by a completed partially directed acyclic graph (CPDAG): an edge is directed when its orientation is shared by every DAG in the MEC (*compelled*), and is undirected when both orientations occur among MEC members (*reversible*) (Andersson et al., 1997). Its undirected edges cannot in general be oriented independently: a DAG in the MEC must preserve all directed edges of the CPDAG and orient the remaining edges without creating a directed cycle or a new v-structure. Thus, on the three-node path skeleton with nonadjacent endpoints, the two directed chains and the fork are Markov equivalent, whereas the collider belongs to a different MEC.

Finally, an edge $i \to j$ is *covered* when $\mathrm{pa}_G(j) = \mathrm{pa}_G(i) \cup \{i\}$. Reversing a covered edge preserves Markov equivalence, and any two DAGs in the same MEC are connected by a sequence of covered edge reversals (Chickering, 1995).

B.2 Distributional Properties

This subsection collects the density properties used in Assumptions 2 and 3. Let q be a strictly positive density on $\mathbb{R}$ and write $\ell = \log q$.

Log-concavity and piecewise smoothness. The density q is *log-concave* if ℓ is concave. Recall that $PC^1(\mathbb{R})$ denotes continuous functions that are C^1 outside a finite, possibly empty, set, as used in Assumption 2. A finite-valued concave log-density has finite left and right derivatives at every point. A *kink* is a point c where $\ell'(c+) < \ell'(c-)$, with $-$ and $+$ denoting the left and right derivatives. Thus, a concave log-density in $PC^1(\mathbb{R})$ is C^1 except at finitely many kinks, and it may have no kinks at all.

Definition 6 (Super-exponential tail)**.** *A strictly positive density q on $\mathbb{R}$ has a* super-exponential tail *in direction $s \in \{-1, 1\}$ if*

$$\lim_{t \to \infty} \frac{-\log q(st)}{t} = \infty. \tag{16}$$

Equivalently, $q(st) = o(e^{-at})$ as $t \to \infty$ for every $a > 0$. The choices $s = 1$ and $s = -1$ specify the right and left tails, respectively.

For a concave log-density $\ell = \log q$ that is C^1 outside a finite set $\mathcal{B}$ of kinks, having at least one super-exponential tail is equivalent to

$$\sup_{x \in \mathbb{R} \setminus \mathcal{B}} |\ell'(x)| = \infty. \tag{17}$$

Indeed, concavity makes the one-sided derivatives nonincreasing, and their limits at the tails agree with the corresponding limiting secant slopes. Thus, an unbounded derivative gives Eq. (16) at one tail, whereas a bounded derivative makes ℓ globally Lipschitz and rules it out at both tails.

C Estimation Procedure

This appendix specifies full-sample cross-fitting for common-order recovery and the multiplier threshold used for the projected-covariance scores in the parent-set estimation stage of §3.3.

C.1 Nuisance Estimation via Cross-Fitting

For common-order recovery, we use the full sample from each environment. We first pool these samples to compute a mean and standard deviation for each variable, and denote the standardized observations by z_r^e. Each environment sample is then divided into K folds, with the same partition used for every candidate and sink removal step. The standardization statistics are shared across folds: held-out observations contribute to these statistics, but are excluded from the regression fits used to evaluate their fold.

At the start of each sink removal step, we draw a new frequency dictionary for the current remaining set R and use it for all candidates, environments, and folds. For each candidate $j \in R$ and environment e, we estimate the conditional mean $m^e_{R,j}$ in Eq. (4) using all observations outside the held-out fold. Let T denote the indices of these out-of-fold units and $\widehat{m}$ the fitted regression function. We form residuals $\widehat{\varepsilon}_r = z^e_{r,j} - \widehat{m}(z^e_{r,R\setminus\{j\}})$ and estimate the characteristic functions of the residuals and predictors by averaging their complex exponential features over T. For a frequency pair (t_b, ω_b), these estimates are

$$\widehat{\phi}_b = \frac{1}{|T|} \sum_{r \in T} \exp(\mathrm{i} t_b \widehat{\varepsilon}_r), \qquad \widehat{\rho}_b = \frac{1}{|T|} \sum_{r \in T} \exp\Big(\mathrm{i} \omega_b^\top z^e_{r,R\setminus\{j\}}\Big). \tag{18}$$

To keep the notation readable, we leave the dependence of these estimates on R, j, e, and the fold implicit. We then evaluate the score in Eq. (10) on the held-out observations using the fitted regression and characteristic functions. Repeating this step for all folds and averaging the score contributions over all observations in environment e gives $\widehat{M}^e_{R,j}(t, \omega)$, which enters the discrepancy in Eq. (12). The additional residual scaling used in our experiments is described in Appendix E.3.

For parent-set estimation, we return to the original observations and split each environment sample into a direction-learning sample and a support-evaluation sample. Standardization for this stage uses means and standard deviations computed from the pooled direction-learning samples, with the same coordinatewise transformation applied to both parts. For each candidate edge $k \to j$ in environment e, we use the direction-learning sample to fit $\widehat{g}^e_{j,k}$ and keep this function fixed throughout the subsequent cross-fitting steps. As in Eq. (11), we leave the dependence on the estimated order $\widehat{\pi}$ implicit in the notation.

Cross-fitting is carried out within the support-evaluation sample. For each held-out fold, we use the remaining observations in that sample to regress the response and the values of $\widehat{g}^e_{j,k}$ separately on the predecessors of j other than k. These two regressions yield $\widehat{\mu}^e_{j,k}$ and $\widehat{\nu}^e_{j,k}$, respectively; the latter estimates the conditional mean of the direction function learned on the direction-learning sample. Substituting these estimates and $\widehat{g}^e_{j,k}$ into Eq. (11), we obtain $\widehat{\psi}^e_{j,k,r}$ by evaluating the score at each held-out observation r. Within an environment, we use the same direction-learning sample and the same support-evaluation folds for every candidate edge.

Algorithms 1 and 2 give the complete procedures, starting from the original observations $\{x^e_r\}_{r=1}^{n_e}$. In both algorithms, $z^e_{A,P}$ denotes the standardized data restricted to rows A and coordinates P. The operation Fit uses the regression learner specified for each stage, taking predictors and responses as inputs and returning the training-response mean when the predictor set is empty. In Algorithm 1,

$u^e_{r,b}$ is the held-out score contribution for the current R and j, which are omitted from its indices. In Algorithm 2, we abbreviate $\widehat{g}^e_{j,k}(z^e_{r,P})$ as $\widehat{g}^e_{r,k}$, leaving j implicit, and use $\widehat{g}^e_{T,k}$ for the collection of these values over $r \in T$.

C.2 Multiplier Maximum for Parent-Set Estimation Stage

Given the estimated candidate order $\widehat{\pi}$ returned by the full-sample order-recovery stage, we consider each candidate edge $k \to j$ with $k \in P_{\widehat{\pi}}(j)$ in each environment $e \in [E]$. Let $\mathcal{I}^e_s$ index the n^e_s observations in the support-evaluation sample. Let $\widehat{\psi}^e_{j,k,r}$ denote the cross-fitted projected-covariance score for observation r, and let $\bar{\psi}^e_{j,k}$ and $\widehat{\sigma}^e_{j,k}$ denote its sample mean and standard deviation, respectively. In each environment, candidate edges whose standard deviations fall below a prespecified floor are *variance-ineligible*; the others are *variance-eligible* and form the variance-eligible family $\widehat{\mathcal{R}}$.

For each $(e, j, k) \in \widehat{\mathcal{R}}$, define

$$\widehat{T}^e_{j,k} \coloneqq \frac{\sqrt{n^e_s}\,\bar{\psi}^e_{j,k}}{\widehat{\sigma}^e_{j,k}}. \tag{19}$$

One Gaussian multiplier draw yields

$$T^*_{\mathrm{str,max}} \coloneqq \max_{(e,j,k)\in\widehat{\mathcal{R}}} \frac{1}{\sqrt{n^e_s}\,\widehat{\sigma}^e_{j,k}} \sum_{r\in\mathcal{I}^e_s} \xi^e_r\{\widehat{\psi}^e_{j,k,r} - \bar{\psi}^e_{j,k}\}. \tag{20}$$

Observation-level multipliers are shared across all cells in an environment. Let $\widehat{c}_{\mathrm{str},1-\alpha}$ be the conditional $(1-\alpha)$ quantile of Eq. (20). A variance-eligible candidate edge $k \to j$ in environment e is *detected nonzero* when

$$\widehat{T}^e_{j,k} > \widehat{c}_{\mathrm{str},1-\alpha}. \tag{21}$$

D Proofs

D.1 Proof of Theorem 1

Theorem 1 (Identifiability of common order)**.** *Under the ANM in Eq.* (1) *and Assumptions 1-4, at every remaining set R reached by removing common sinks, a node j is a common sink if and only if $\varepsilon^e_{R,j} \perp\!\!\!\perp X^e_{R\setminus\{j\}}$ for all $e \in [E]$. Successively removing these nodes and reversing the removal sequence returns a valid common order $\pi \in \Pi_0$.*

Proof. A set R reached by removing common sinks is ancestral in $G_\cup$ and hence in every environment-specific DAG. Its variables retain their original parents, structural equations, and mutually independent noises. Their density in each environment is the product of the corresponding noise-density factors: integrating out the removed sinks successively contributes a factor of one. This density is strictly positive and continuous on $\mathbb{R}^{|R|}$. If j is a common sink, every variable in $X^e_{R\setminus\{j\}}$ is a nondescendant of j, and the conditioning set contains every parent. Subtracting the conditional mean thus leaves $\varepsilon^e_{R,j} = N^e_j \perp\!\!\!\perp X^e_{R\setminus\{j\}}$ in each environment.

Conversely, let j be a non-sink of $G_\cup[R]$. Choose a remaining edge $j \to k$ and a qualifying environment e from Assumption 3. Let $\mathcal{B}^e_v$ be the finite set of kinks of $\log q^e_v$. The conditional density of X^e_j given $X^e_{R\setminus\{j\}} = u$ is

$$\begin{aligned} p^e(x \mid u) &= q^e_j\big(x - f^e_j(u_{\mathrm{pa}_e(j)})\big)\, K^e_u(x), \\ K^e_u(x) &= \frac{1}{Z^e(u)} \prod_{v\in \mathrm{ch}_{G^e}(j)\cap R} q^e_v\big(u_v - f^e_v(x, u_{\mathrm{pa}_e(v)\setminus\{j\}})\big)\,, \end{aligned} \tag{22}$$

where $Z^e(u)$ normalizes the density. For fixed u, boundedness of the structural functions keeps the arguments of each child density in a compact interval. Positivity and continuity bound the unnormalized child-density product above and away from zero. Integrating against q^e_j gives $0 < Z^e(u) < \infty$, and hence

$$0 < L^e(u) \le K^e_u(x) \le H^e(u) < \infty \quad \text{for every } x \in \mathbb{R}. \tag{23}$$

Algorithm 1 STORM: Cross-fitting for common-order recovery

Input: (a) Original samples $\{x_r^e\}_{r=1}^{n_e}$ for $e \in [E]$
(b) Number of folds K
(c) Regression learner
(d) Frequency-dictionary rule with B pairs
Output: Estimated common order $\widehat{\pi}$.

$(a, s) \leftarrow \textsc{MeanSD}(\{x_r^e : e \in [E],\ r \in [n_e]\})$
$z_r^e \leftarrow (x_r^e - a)/s$ for every e, r
$(I_1^e, \ldots, I_K^e) \leftarrow \textsc{Partition}([n_e], K)$ for every e
$R \leftarrow V$; $L \leftarrow [\,]$
while $|R| > 1$ **do**
  $\mathcal{D}_{n,R} = \{(t_b, \omega_b)\}_{b=1}^{B} \leftarrow \textsc{DrawFrequencies}(|R| - 1, B)$
  for all $j \in R$ **do**
    $P \leftarrow R \setminus \{j\}$
    for $e = 1, \ldots, E$ **do**
      for $\ell = 1, \ldots, K$ **do**
        $I \leftarrow I_\ell^e$ ▷ `In-fold units`
        $T \leftarrow [n_e] \setminus I$ ▷ `Out-of-fold units`
        $\widehat{m} \leftarrow \textsc{Fit}(z_{T,P}^e, z_{T,j}^e)$
        $\widehat{\varepsilon}_r \leftarrow z_{r,j}^e - \widehat{m}(z_{r,P}^e)$ for all $r \in [n_e]$ ▷ Eq. (4)
        for $b = 1, \ldots, B$ **do**
          $\widehat{\phi}_b \leftarrow |T|^{-1} \sum_{r \in T} \exp(\mathrm{i} t_b \widehat{\varepsilon}_r)$ ▷ Eq. (18)
          $\widehat{\rho}_b \leftarrow |T|^{-1} \sum_{r \in T} \exp(\mathrm{i} \omega_b^\top z_{r,P}^e)$ ▷ Eq. (18)
          $u_{r,b}^e \leftarrow \{\exp(\mathrm{i} t_b \widehat{\varepsilon}_r) - \widehat{\phi}_b - \mathrm{i} t_b \widehat{\phi}_b \widehat{\varepsilon}_r\}\{\exp(\mathrm{i} \omega_b^\top z_{r,P}^e) - \widehat{\rho}_b\}$ ▷ Eq. (10)
        end for
      end for
      $\widehat{M}_{R,j}^e(t_b, \omega_b) \leftarrow n_e^{-1} \sum_{r=1}^{n_e} u_{r,b}^e$ for $b \in [B]$
    end for
    $\widehat{D}_{n,R}(j) \leftarrow \max_{e \in [E]} \{B^{-1} \sum_{b=1}^{B} |\widehat{M}_{R,j}^e(t_b, \omega_b)|^2\}^{1/2}$ ▷ Eq. (12)
  end for
  $j_R \leftarrow \min \arg\min_{j \in R} \widehat{D}_{n,R}(j)$ ▷ `Estimated common sink`
  $L \leftarrow \textsc{Append}(L, j_R)$; $R \leftarrow R \setminus \{j_R\}$
end while
$L \leftarrow \textsc{Append}(L,$ the remaining node in $R)$
return $\widehat{\pi} \leftarrow \textsc{Reverse}(L)$

Algorithm 2 STORM: Direction learning and cross-fitting for parent-set estimation

Input: (a) Original samples $\{x_r^e\}_{r=1}^{n_e}$ for $e \in [E]$
(b) Common order $\widehat{\pi}$
(c) Regression learners
(d) Number of folds K_{supp}
(e) Multiplier count B_{mult}
(f) Significance level α

Output: Candidate-edge labels and candidate shift sets $\widehat{S}_j^{e_1,e_2}$.

$(\mathcal{I}_d^e, \mathcal{I}_s^e) \leftarrow$ SPLITDIRECTIONSUPPORT$([n_e])$ for every e
$(a, s) \leftarrow$ MEANSD$(\{x_r^e : e \in [E],\ r \in \mathcal{I}_d^e\})$
$z_r^e \leftarrow (x_r^e - a)/s$ for every e, r
$(H_1^e, \ldots, H_{K_{\text{supp}}}^e) \leftarrow$ PARTITION$(\mathcal{I}_s^e, K_{\text{supp}})$ for every e
for $e = 1, \ldots, E$ **do**
  for all $j \in V$ with $P_{\widehat{\pi}}(j) \neq \varnothing$ **do**
    $P \leftarrow P_{\widehat{\pi}}(j)$ ▷ `Parent candidates`
    $\widehat{f}^e \leftarrow$ FIT$(z^e_{\mathcal{I}_d^e, P}, z^e_{\mathcal{I}_d^e, j})$
    for all $k \in P$ **do**
      $\widehat{h}^e \leftarrow$ FIT$(z^e_{\mathcal{I}_d^e, P\setminus\{k\}}, z^e_{\mathcal{I}_d^e, j})$
      $\widehat{g}^e_{r,k} \leftarrow \widehat{f}^e(z^e_{r,P}) - \widehat{h}^e(z^e_{r,P\setminus\{k\}})$ for $r \in \mathcal{I}_s^e$ ▷ Eq. (5)
      for $\ell = 1, \ldots, K_{\text{supp}}$ **do**
        $I \leftarrow H_\ell^e$; $T \leftarrow \mathcal{I}_s^e \setminus I$
        $\widehat{\mu}_k^e \leftarrow$ FIT$(z^e_{T,P\setminus\{k\}}, z^e_{T,j})$
        $\widehat{\nu}_k^e \leftarrow$ FIT$(z^e_{T,P\setminus\{k\}}, \widehat{g}^e_{T,k})$
        $\widehat{\psi}^e_{j,k,r} \leftarrow \{z^e_{r,j} - \widehat{\mu}_k^e(z^e_{r,P\setminus\{k\}})\}\{\widehat{g}^e_{r,k} - \widehat{\nu}_k^e(z^e_{r,P\setminus\{k\}})\}$ ▷ Eq. (11)
      end for
      $(\bar{\psi}^e_{j,k}, \widehat{\sigma}^e_{j,k}) \leftarrow$ MEANSD$(\{\widehat{\psi}^e_{j,k,r} : r \in \mathcal{I}_s^e\})$
    end for
  end for
end for
$\widehat{\mathcal{R}} \leftarrow \{(e, j, k) : k \in P_{\widehat{\pi}}(j),\ \widehat{\sigma}^e_{j,k} \geq \sigma_{\min}\}$; $\widehat{c}_{\text{str},1-\alpha} \leftarrow +\infty$
if $\widehat{\mathcal{R}} \neq \varnothing$ **then**
  for $b = 1, \ldots, B_{\text{mult}}$ **do**
    Draw $\xi_r^{e,(b)} \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$ for $e \in [E], r \in \mathcal{I}_s^e$
    $T^{*(b)}_{\text{str,max}} \leftarrow \max_{(e,j,k)\in\widehat{\mathcal{R}}} \dfrac{\sum_{r\in\mathcal{I}_s^e} \xi_r^{e,(b)}(\widehat{\psi}^e_{j,k,r} - \bar{\psi}^e_{j,k})}{\sqrt{|\mathcal{I}_s^e|}\,\widehat{\sigma}^e_{j,k}}$ ▷ Eq. (20)
  end for
  $\widehat{c}_{\text{str},1-\alpha} \leftarrow$ QUANTILE$(\{T^{*(b)}_{\text{str,max}}\}_{b=1}^{B_{\text{mult}}}, 1-\alpha)$
end if
$\text{label}_{e,j,k} \leftarrow$ unresolved for all candidate-edge cells; $\widehat{\text{pa}}_e(j) \leftarrow \varnothing$ for all e, j
for all $(e, j, k) \in \widehat{\mathcal{R}}$ **do**
  $\widehat{T}^e_{j,k} \leftarrow \sqrt{|\mathcal{I}_s^e|}\,\bar{\psi}^e_{j,k}/\widehat{\sigma}^e_{j,k}$ ▷ Eq. (19)
  if $\widehat{T}^e_{j,k} > \widehat{c}_{\text{str},1-\alpha}$ **then** ▷ Eq. (21)
    $\text{label}_{e,j,k} \leftarrow$ detected nonzero; $\widehat{\text{pa}}_e(j) \leftarrow \widehat{\text{pa}}_e(j) \cup \{k\}$
  end if
end for
$\widehat{S}_j^{e_1,e_2} \leftarrow \widehat{\text{pa}}_{e_1}(j) \triangle \widehat{\text{pa}}_{e_2}(j)$ for all $j, e_1 < e_2$ ▷ Eq. (2)
return label and $\{\widehat{S}_j^{e_1,e_2}\}$

Suppose for contradiction that $\varepsilon^e_{R,j} \perp\!\!\!\perp X^e_{R\setminus\{j\}}$. The conditional densities must then be translates of a common residual density h^e:

$$p^e(x \mid u) = h^e\Big(x - \mathbb{E}_e[X^e_j \mid X^e_{R\setminus\{j\}} = u]\Big) \text{ for almost every } u \text{ and almost every } x. \tag{24}$$

Identifying the translation. First, suppose that q^e_j has a super-exponential tail. By Appendix B.2, $(\log q^e_j)'$ is unbounded outside $\mathcal{B}^e_j$, and concavity makes its one-sided derivative diverge at a tail. For any $c \neq 0$, integrating that derivative over an interval of length $|c|$ gives

$$\sup_{x\in\mathbb{R}} |\log q^e_j(x+c) - \log q^e_j(x)| = \infty. \tag{25}$$

Fix a value u_0 at which Eq. (24) holds. Evaluating the two conditional densities at the same residual value r yields

$$\frac{q^e_j\Big(r + \mathbb{E}_e[X^e_j \mid X^e_{R\setminus\{j\}} = u] - f^e_j(u_{\mathrm{pa}_e(j)})\Big)}{q^e_j\Big(r + \mathbb{E}_e[X^e_j \mid X^e_{R\setminus\{j\}} = u_0] - f^e_j((u_0)_{\mathrm{pa}_e(j)})\Big)} = \frac{K^e_{u_0}\Big(r + \mathbb{E}_e[X^e_j \mid X^e_{R\setminus\{j\}} = u_0]\Big)}{K^e_u\Big(r + \mathbb{E}_e[X^e_j \mid X^e_{R\setminus\{j\}} = u]\Big)}.$$

By Eq. (23), the right-hand side is bounded above and bounded away from zero, uniformly in r. The density ratio is continuous, so these bounds extend from almost every r to every r. Equation (25) forces

$$\mathbb{E}_e[X^e_j \mid X^e_{R\setminus\{j\}} = u] - f^e_j(u_{\mathrm{pa}_e(j)}) = \mathbb{E}_e[X^e_j \mid X^e_{R\setminus\{j\}} = u_0] - f^e_j((u_0)_{\mathrm{pa}_e(j)})$$

for almost every u.

It remains to treat $\mathcal{B}^e_j \neq \varnothing$ when the derivative is bounded. A finite-valued concave log-density has finite one-sided derivatives at every point. At each actual nondifferentiability point $s \in \mathcal{B}^e_v$, their difference is strictly negative:

$$(\log q^e_v)'(s+) - (\log q^e_v)'(s-) < 0.$$

If $u_v - f^e_v(x_0, u_{\mathrm{pa}_e(v)\setminus\{j\}}) = s$ and $\partial_j f^e_v(x_0, u_{\mathrm{pa}_e(v)\setminus\{j\}}) \neq 0$, the child term $\log q^e_v(u_v - f^e_v(x, u_{\mathrm{pa}_e(v)\setminus\{j\}}))$ has derivative jump

$$\big|\partial_j f^e_v(x_0, u_{\mathrm{pa}_e(v)\setminus\{j\}})\big| \left\{(\log q^e_v)'(s+) - (\log q^e_v)'(s-)\right\} < 0 \tag{26}$$

at x_0. At a preimage of s where this partial derivative is zero, the composition is differentiable: $\log q^e_v$ is locally Lipschitz and the change in f^e_v is $o(|x - x_0|)$. Consequently, all derivative jumps in the conditional log-density have the same sign and cannot cancel.

Suppose a remaining child $v \in \mathrm{ch}_{G^e}(j) \cap R$ has $\mathcal{B}^e_v \neq \varnothing$. On the nonempty open set where $u_w > \sup f^e_w + \max \mathcal{B}^e_w$ for every child $w \in \mathrm{ch}_{G^e}(j) \cap R$ with $\mathcal{B}^e_w \neq \varnothing$, all child terms are C^1 as functions of x. The conditional log-density then has exactly $|\mathcal{B}^e_j|$ nondifferentiability points, namely $f^e_j(u_{\mathrm{pa}_e(j)}) + \mathcal{B}^e_j$. Causal minimality implies that f^e_v cannot be constant in its jth argument everywhere, since otherwise deleting $j \to v$ would preserve the generated distribution. Because f^e_v is C^1, there is a slice with $\partial_j f^e_v \neq 0$ on an open interval. Fix the other coordinates and choose x_0 in this interval outside the finite set $f^e_j(u_{\mathrm{pa}_e(j)}) + \mathcal{B}^e_j$. For any $s \in \mathcal{B}^e_v$, set $u_v = f^e_v(x_0, u_{\mathrm{pa}_e(v)\setminus\{j\}}) + s$. This choice does not change $f^e_j(u_{\mathrm{pa}_e(j)})$ or $u_{\mathrm{pa}_e(v)\setminus\{j\}}$, since a child is neither a parent of j nor its own parent. The implicit function theorem and continuity of f^e_j give an open neighborhood of conditioning values with an additional derivative jump separate from $f^e_j(u_{\mathrm{pa}_e(j)}) + \mathcal{B}^e_j$. All $|\mathcal{B}^e_j|$ original jumps remain, and Eq. (26) prevents cancellation by other children. Thus, on this second open set, the conditional log-density has at least $|\mathcal{B}^e_j| + 1$ nondifferentiability points. Both open sets have positive probability because $X^e_{R\setminus\{j\}}$ has a strictly positive density. Choosing conditional densities from the full-measure set in Eq. (24), their positive continuous versions must agree up to translation everywhere. Translation preserves the number of nondifferentiability points, contradicting the two counts.

Therefore, under the assumed residual independence, every remaining child $v \in \mathrm{ch}_{G^e}(j) \cap R$ has $\mathcal{B}^e_v = \varnothing$. The nondifferentiability set of the conditional log-density is then exactly $f^e_j(u_{\mathrm{pa}_e(j)}) + \mathcal{B}^e_j$ for every u. After translating by $\mathbb{E}_e[X^e_j \mid X^e_{R\setminus\{j\}} = u]$ as in Eq. (24), the sets

$$\mathcal{B}^e_j - \mathbb{E}_e[X^e_j \mid X^e_{R\setminus\{j\}} = u] + f^e_j(u_{\mathrm{pa}_e(j)})$$

must coincide for almost every u. Since $\mathcal{B}_j^e$ is finite and nonempty, equality of their minima forces $\mathbb{E}_e[X_j^e \mid X_{R\setminus\{j\}}^e = u] - f_j^e(u_{\mathrm{pa}_e(j)})$ to be constant. This proves in both cases that

$$\mathbb{E}_e[X_j^e \mid X_{R\setminus\{j\}}^e] = f_j^e(X_{\mathrm{pa}_e(j)}^e) + c_0^e \quad \text{almost surely.}$$

Contradiction with causal minimality. The tower property and the mean-zero structural noise give

$$\mathbb{E}_e\Big[\mathbb{E}_e[X_j^e \mid X_{R\setminus\{j\}}^e] \mid X_{\mathrm{pa}_e(j)}^e\Big] = \mathbb{E}_e[X_j^e \mid X_{\mathrm{pa}_e(j)}^e] = f_j^e(X_{\mathrm{pa}_e(j)}^e),$$

so $c_0^e = 0$. The supposedly independent residual is therefore N_j^e, and $N_j^e \perp\!\!\!\perp X_{R\setminus\{j\}}^e$. For a remaining child $k \in \mathrm{ch}_{G^e}(j) \cap R$, independence from $X_{R\setminus\{j\}}^e$ implies

$$N_j^e \perp\!\!\!\perp \Big(X_k^e, X_{\mathrm{pa}_e(j)\cup(\mathrm{pa}_e(k)\setminus\{j\})}^e\Big).$$

Also,

$$N_k^e \perp\!\!\!\perp \Big(N_j^e, X_{\mathrm{pa}_e(j)\cup(\mathrm{pa}_e(k)\setminus\{j\})}^e\Big)$$

by the acyclic structural equations, since every variable in $X_{\mathrm{pa}_e(j)\cup(\mathrm{pa}_e(k)\setminus\{j\})}^e$ is a nondescendant of k. Hence

$$\begin{aligned} f_k^e\Big(f_j^e(X_{\mathrm{pa}_e(j)}^e) + N_j^e, X_{\mathrm{pa}_e(k)\setminus\{j\}}^e\Big) &= \mathbb{E}_e\Big[X_k^e \mid N_j^e, X_{\mathrm{pa}_e(j)\cup(\mathrm{pa}_e(k)\setminus\{j\})}^e\Big] \\ &= \mathbb{E}_e\Big[X_k^e \mid X_{\mathrm{pa}_e(j)\cup(\mathrm{pa}_e(k)\setminus\{j\})}^e\Big]. \end{aligned}$$

The right-hand side does not depend on N_j^e. Since $N_j^e \perp\!\!\!\perp X_{\mathrm{pa}_e(j)\cup(\mathrm{pa}_e(k)\setminus\{j\})}^e$ and q_j^e is positive on all of $\mathbb{R}$, continuity of f_k^e makes its dependence on the jth argument constant for almost every value of the other parents. Removing $j \to k$ would therefore preserve the generated distribution, contradicting causal minimality. We have shown that every remaining non-sink has dependent residuals in at least one environment.

All properties used above persist after removing a common sink: the remaining functions and noise densities are unchanged, and each remaining union edge retains its qualifying environment. Every nonempty finite DAG has a sink, so the procedure can continue until all nodes have been removed. Reversing the sequence gives a topological order of $G_\cup$, which is a valid common order by the definition of Π_0. □

D.2 Proof of Lemma 1

Lemma 1 (Characteristic residual moments)**.** *Under the integrability conditions required to differentiate under the expectation, the score in Eq.* (10) *has expectation*

$$M_{R,j}^e(t,\omega) := \mathbb{E}_e[\chi_{R,j}^e(t,\omega)] = \mathbb{E}_e[\exp\{\mathrm{i}(t\varepsilon_{R,j}^e + \omega^\top X_{R\setminus\{j\}}^e)\}] - \phi_{R,j}^e(t)\rho_{R,j}^e(\omega).$$

This moment vanishes for every (t,ω) if and only if $\varepsilon_{R,j}^e \perp\!\!\!\perp X_{R\setminus\{j\}}^e$. At residual independence, its real and imaginary components are Neyman-orthogonal with respect to $(m_{R,j}^e, \phi_{R,j}^e(t), \rho_{R,j}^e(\omega))$.

Proof. We derive the auxiliary score by removing the first-order sensitivity of the centered characteristic function product

$$\{\exp\big(\mathrm{i}t\varepsilon_{R,j}^e\big) - \phi_{R,j}^e(t)\}\{\exp\Big(\mathrm{i}\omega^\top X_{R\setminus\{j\}}^e\Big) - \rho_{R,j}^e(\omega)\}$$

to estimation of the conditional mean $m_{R,j}^e$. Let a be a perturbation direction that is a function of $X_{R\setminus\{j\}}^e$, and perturb the conditional mean to $m_{R,j}^e + \eta a$ for a real parameter η. The residual then becomes $\varepsilon_{R,j}^e - \eta a$. Keeping the distribution, $\phi_{R,j}^e(t)$, and $\rho_{R,j}^e(\omega)$ fixed at their true values, the derivative of the product's expectation at $\eta = 0$ is

$$\begin{aligned} &-\mathbb{E}_e\Big[\mathrm{i}t\exp\big(\mathrm{i}t\varepsilon_{R,j}^e\big)a\{\exp\Big(\mathrm{i}\omega^\top X_{R\setminus\{j\}}^e\Big) - \rho_{R,j}^e(\omega)\}\Big] \\ &\quad = -\mathbb{E}_e\Big[\mathbb{E}_e[\mathrm{i}t\exp\big(\mathrm{i}t\varepsilon_{R,j}^e\big) \mid X_{R\setminus\{j\}}^e]a\Big\{\exp\Big(\mathrm{i}\omega^\top X_{R\setminus\{j\}}^e\Big) - \rho_{R,j}^e(\omega)\Big\}\Big]. \end{aligned}$$

The equality follows by conditioning on $X^e_{R\setminus\{j\}}$, since both a and the centered covariate feature are functions of these variables. This derivative need not be zero. The conditional coefficient in this expression suggests subtracting the residual correction

$$\mathbb{E}_e[\mathrm{i}t\exp\big(\mathrm{i}t\varepsilon^e_{R,j}\big)\mid X^e_{R\setminus\{j\}}]\varepsilon^e_{R,j}\{\exp\Big(\mathrm{i}\omega^\top X^e_{R\setminus\{j\}}\Big)-\rho^e_{R,j}(\omega)\}$$

from the original product. Holding the conditional coefficient fixed, the expectation of this correction has exactly the same derivative as the original moment, because the derivative of the residual is $-a$. Subtracting it therefore cancels the first-order effect for every such direction a. This gives the auxiliary score

$$\begin{aligned}\chi^{e,\perp}_{R,j}(t,\omega) \coloneqq \Big\{\exp\big(\mathrm{i}t\varepsilon^e_{R,j}\big)-\phi^e_{R,j}(t)-\mathbb{E}_e\Big[\mathrm{i}t\exp\big(\mathrm{i}t\varepsilon^e_{R,j}\big)\mid X^e_{R\setminus\{j\}}\Big]\varepsilon^e_{R,j}\Big\}\\ \times\Big\{\exp\Big(\mathrm{i}\omega^\top X^e_{R\setminus\{j\}}\Big)-\rho^e_{R,j}(\omega)\Big\}.\end{aligned} \tag{27}$$

Because $\mathbb{E}_e[\varepsilon^e_{R,j}\mid X^e_{R\setminus\{j\}}]=0$ and the conditional expectation in the correction term is a function of the conditioning variables,

$$\begin{aligned}\mathbb{E}_e[\chi^{e,\perp}_{R,j}(t,\omega)] &= \mathbb{E}_e\Big[\{\exp\big(\mathrm{i}t\varepsilon^e_{R,j}\big)-\phi^e_{R,j}(t)\}\{\exp\Big(\mathrm{i}\omega^\top X^e_{R\setminus\{j\}}\Big)-\rho^e_{R,j}(\omega)\}\Big]\\ &= \mathbb{E}_e\Big[\exp\{\mathrm{i}(t\varepsilon^e_{R,j}+\omega^\top X^e_{R\setminus\{j\}})\}\Big]-\phi^e_{R,j}(t)\rho^e_{R,j}(\omega).\end{aligned}$$

The same calculation holds with $\mathbb{E}_e[\mathrm{i}t\exp\big(\mathrm{i}t\varepsilon^e_{R,j}\big)\mid X^e_{R\setminus\{j\}}]$ replaced by $\mathrm{i}t\phi^e_{R,j}(t)$. Factorization and uniqueness of characteristic functions give the independence equivalence.

To verify orthogonality jointly in all nuisance quantities, consider perturbation directions (a,b,d,r) for

$$\Big(m^e_{R,j},\phi^e_{R,j}(t),\mathbb{E}_e[\mathrm{i}t\exp\big(\mathrm{i}t\varepsilon^e_{R,j}\big)\mid X^e_{R\setminus\{j\}}],\rho^e_{R,j}(\omega)\Big),$$

where a and d are functions of $X^e_{R\setminus\{j\}}$ and b,r are constants. The Gateaux derivative of the corrected moment at the truth is

$$\begin{aligned}\mathbb{E}_e\Big[\Big\{\Big(\mathbb{E}_e[\mathrm{i}t\exp\big(\mathrm{i}t\varepsilon^e_{R,j}\big)\mid X^e_{R\setminus\{j\}}]-\mathrm{i}t\exp\big(\mathrm{i}t\varepsilon^e_{R,j}\big)\Big)a-b-d\varepsilon^e_{R,j}\Big\}\\ \times\Big\{\exp\Big(\mathrm{i}\omega^\top X^e_{R\setminus\{j\}}\Big)-\rho^e_{R,j}(\omega)\Big\}\Big]\\ -r\,\mathbb{E}_e\Big[\exp\big(\mathrm{i}t\varepsilon^e_{R,j}\big)-\phi^e_{R,j}(t)-\mathbb{E}_e[\mathrm{i}t\exp\big(\mathrm{i}t\varepsilon^e_{R,j}\big)\mid X^e_{R\setminus\{j\}}]\varepsilon^e_{R,j}\Big]=0.\end{aligned}$$

The a and d terms vanish by conditioning on $X^e_{R\setminus\{j\}}$; the b term vanishes because $\exp\Big(\mathrm{i}\omega^\top X^e_{R\setminus\{j\}}\Big)-\rho^e_{R,j}(\omega)$ is centered; and the r term vanishes because

$$\mathbb{E}_e[\exp\big(\mathrm{i}t\varepsilon^e_{R,j}\big)]=\phi^e_{R,j}(t),\qquad \mathbb{E}_e\Big[\mathbb{E}_e[\mathrm{i}t\exp\big(\mathrm{i}t\varepsilon^e_{R,j}\big)\mid X^e_{R\setminus\{j\}}]\varepsilon^e_{R,j}\Big]=0.$$

Thus, the real and imaginary parts of the auxiliary moment are Neyman-orthogonal with respect to

$$\Big(m^e_{R,j},\phi^e_{R,j}(t),\mathbb{E}_e[\mathrm{i}t\exp\big(\mathrm{i}t\varepsilon^e_{R,j}\big)\mid X^e_{R\setminus\{j\}}],\rho^e_{R,j}(\omega)\Big)$$

at their true values, without requiring residual independence.

At residual independence, the conditional coefficient simplifies to

$$\mathbb{E}_e[\mathrm{i}t\exp\big(\mathrm{i}t\varepsilon^e_{R,j}\big)\mid X^e_{R\setminus\{j\}}]=\mathrm{i}t\phi^e_{R,j}(t),$$

so the auxiliary score reduces to Eq. (10), whose regression derivative therefore also vanishes. The additional derivative through $\phi^e_{R,j}(t)$ is proportional to

$$\mathbb{E}_e\Big[\varepsilon^e_{R,j}\{\exp\Big(\mathrm{i}\omega^\top X^e_{R\setminus\{j\}}\Big)-\rho^e_{R,j}(\omega)\}\Big]=0,$$

proving the implemented score's null orthogonality. □

D.3 Proof of Proposition 1

Proposition 1 (Identifiability of parent set)**.** *Under the ANM in Eq.* (1) *and Assumptions 1-4, suppose that each environment satisfies the intersection property. Then, for every $e \in [E]$, $j \in V$, and $\pi \in \Pi_0$,*

$$\mathrm{pa}_e(j) = \left\{ k \in P_\pi(j) : \Delta^{e,\pi}_{j,k} > 0 \right\}. \tag{8}$$

Proof. Fix an environment e, a valid common order $\pi \in \Pi_0$, and a node j, and suppress the environment index. Write $P = P_\pi(j)$, and fix $k \in P$ with $S = P \setminus \{k\}$. The set P contains every parent of j and consists only of its nondescendants. Hence $N_j \perp\!\!\!\perp X_P$, and, writing $F_j = f_j(X_{\mathrm{pa}(j)})$, the ANM gives

$$\mathbb{E}[X_j \mid X_P] = F_j, \qquad \mathbb{E}[X_j \mid X_S] = \mathbb{E}[F_j \mid X_S].$$

Consequently,

$$\Delta^{e,\pi}_{j,k} = \mathbb{E}\big[\{F_j - \mathbb{E}[F_j \mid X_S]\}^2\big] = \mathbb{E}[\mathrm{Var}(F_j \mid X_S)]. \tag{28}$$

If $k \notin \mathrm{pa}(j)$, then F_j is a function of X_S, so $\Delta^{e,\pi}_{j,k} = 0$.

Conversely, let $k \in \mathrm{pa}(j)$ and suppose that $\Delta^{e,\pi}_{j,k} = 0$. Then $F_j = h(X_S)$ almost surely for a measurable function h. Since $X_j = h(X_S) + N_j$ with $N_j \perp\!\!\!\perp X_P$, its conditional distribution given X_P depends only on X_S. Thus, $X_j \perp\!\!\!\perp X_k \mid X_S$. Set $A = \mathrm{pa}(j) \setminus \{k\}$ and $U = P \setminus \mathrm{pa}(j)$, so that $S = A \cup U$. The local Markov property and the preceding conditional independence give, respectively,

$$X_j \perp\!\!\!\perp X_U \mid X_{A\cup\{k\}}, \qquad X_j \perp\!\!\!\perp X_k \mid X_{A\cup U}.$$

The intersection property in Eq. (15) implies $X_j \perp\!\!\!\perp X_{U\cup\{k\}} \mid X_A$, and hence $X_j \perp\!\!\!\perp X_k \mid X_A$. The conditional distribution of X_j given its parents therefore depends only on X_A. Its factor in the DAG factorization can omit X_k, leaving the distribution Markov after deleting $k \to j$. This contradicts causal minimality in Assumption 4; this is the intersection argument of Peters et al. (2014, Lemma 38). Therefore every parent has $\Delta^{e,\pi}_{j,k} > 0$, proving Eq. (8). The argument applies to every valid common order and environment, so comparing the identified parent sets gives Eq. (2). □

D.4 Proof of Lemma 2

Lemma 2 (Projected-covariance identity and orthogonality)**.** *For any fixed square-integrable direction $g^e_{j,k}$, $\theta^e_{j,k} = \mathbb{E}_e[\delta^e_{j,k} g^e_{j,k}]$. In particular, $\delta^e_{j,k} = 0$ implies $\theta^e_{j,k} = 0$, while $g^e_{j,k} = \delta^e_{j,k}$ gives $\theta^e_{j,k} = \Delta^e_{j,k}$. The moment is Neyman-orthogonal with respect to $(\mu^e_{j,k}, \nu^e_{j,k})$.*

Proof. Fix the direction $g^e_{j,k}$ throughout and use the fixed-order notation of §3.3. Conditioning first on all predecessors gives

$$\theta^e_{j,k} = \mathbb{E}_e[\{\mathbb{E}_e[X^e_j \mid X^e_{P_\pi(j)}] - \mu^e_{j,k}\}(g^e_{j,k} - \nu^e_{j,k})] = \mathbb{E}_e[\delta^e_{j,k}(g^e_{j,k} - \nu^e_{j,k})] = \mathbb{E}_e[\delta^e_{j,k} g^e_{j,k}],$$

where the last equality follows from $\mathbb{E}_e[\delta^e_{j,k} \mid X^e_{P_\pi(j)\setminus\{k\}}] = 0$. The null identity and the case $g^e_{j,k} = \delta^e_{j,k}$ follow immediately. For square-integrable functions $\widetilde{\mu}^e_{j,k}$ and $\widetilde{\nu}^e_{j,k}$ of $X^e_{P_\pi(j)\setminus\{k\}}$, expand the score around the true centering functions. The two linear terms vanish after conditioning on these variables, leaving

$$\mathbb{E}_e[(X^e_j - \widetilde{\mu}^e_{j,k})(g^e_{j,k} - \widetilde{\nu}^e_{j,k})] - \theta^e_{j,k} = \mathbb{E}_e[(\mu^e_{j,k} - \widetilde{\mu}^e_{j,k})(\nu^e_{j,k} - \widetilde{\nu}^e_{j,k})]. \tag{29}$$

This product remainder establishes Neyman orthogonality with respect to both centering functions. □

D.5 Proof of Theorem 2

Theorem 2 (Preservation of identifiability under structure shifts)**.** *Let $\mathcal{M}_\star \subseteq \mathcal{M}$ consist of models satisfying Assumption 2 and the noise-shape condition of Assumption 3 at every node, and suppose these conditions are preserved under structure-shift operations. Then*

$$o(M) \in \mathcal{M}_\star \qquad \text{for every } M \in \mathcal{M}_\star \text{ and } o \in \mathcal{O}. \tag{9}$$

$G(M)$ and $G(o(M))$ are identifiable from their observational distributions. Also, $\mathbb{P}^M_X$ and $\mathbb{P}^{o(M)}_X$ identify a valid common order and the structure shift sets in Eq. (2)*. These conclusions require no additional conditions on the induced conditional input distributions.*

Proof. Since $o : \mathcal{M} \to \mathcal{M}$, the transformed model remains a causally minimal ANM with independent, mean-zero noises. The specified replacements retain the function and density restrictions defining $\mathcal{M}_\star$, proving Eq. (9). Every model in $\mathcal{M}_\star$ satisfies Assumptions 2 and 3 for $E = 1$. Applying Theorem 1 and Proposition 1 therefore identifies its graph from its observational distribution; positive noise densities supply the intersection property. For the pair $(M, o(M))$, Eq. (3) gives an acyclic union graph, and the noise condition holds in every environment containing an edge. The same two results identify a valid common order and both parent sets, whose symmetric differences give the structure shift sets. □

E Experimental Setup

E.1 Data Generating Process

We generated two environments with n observations each from nonlinear ANMs on d variables. Let $A^e_{k,j} = \mathbf{1}\{k \in \mathrm{pa}_e(j)\}$ denote the adjacency matrix in environment $e \in \{1, 2\}$, with rows indexing parents and columns indexing children. The two environments shared mechanism parameters and a topological order, and differed through deletions of incoming edges.

Graphs and causal structure shifts. We used two types of random graph generators, Erdős–Rényi (ER) (Erdös & Rényi, 1960) and scale-free (SF) (Barabasi & Albert, 1999), with an edge factor r controlling sparsity. For ER graphs, we sampled an undirected graph uniformly with exactly rd edges. For SF graphs, we used the undirected Barabási preferential-attachment generator with r attachments per new node. Thus, r has different finite-sample edge-count interpretations in the two families. We oriented each edge from the smaller to the larger node index and then randomly relabeled all nodes. For SF graphs, this orientation used the node indices assigned by the growth process before relabeling. The resulting graph defines A^1 and a known topological order π_0.

We resampled the graph until it contained at least four nonroot nodes, then chose four of these nodes uniformly without replacement. Let $\mathcal{S}$ denote this set and $q_j = |\mathrm{pa}_1(j)|$. For each $j \in \mathcal{S}$, we removed $\min(2, q_j)$ parents, chosen uniformly without replacement, to obtain A^2. All other edges were retained, so $\mathcal{E}^2 \subseteq \mathcal{E}^1$ and the union graph is acyclic. Each generated pair therefore contained between four and eight deleted edges, affecting exactly four target nodes. The ground-truth causal structure shift matrix is

$$D_{k,j} \coloneqq \mathbf{1}\{A^1_{k,j} \neq A^2_{k,j}\}, \qquad k \neq j. \tag{30}$$

Nonlinear mechanisms with parent interactions. For each node, we used a feedforward neural network with one hidden layer of width $H = 16$ and $\tanh$ activation. The adjacency mask removes non-parent inputs by setting their input weights to zero, as in the neural DAG models of Lachapelle et al. (2020). The structural equation is

$$X^e_j = a_j \sum_{\ell=1}^{H} v_{j,\ell} \left[\tanh\left(b_{j,\ell} + \sum_{k=1}^{d} W_{j,\ell,k} A^e_{k,j} X^e_k \right) - \tanh(b_{j,\ell}) \right] + \sigma Z^e_j. \tag{31}$$

Here $a_j \sim \mathrm{Unif}[1.5, 2.5]$ and $b_{j,\ell} \sim \mathrm{Unif}[-0.5, 0.5]$. We imposed monotonicity in each parent through weight-sign constraints (Archer & Wang, 1993). For our experiments, we sampled the weights as follows, with $\kappa_j = \max(q_j, 1)$:

$$\begin{aligned} W_{j,\ell,k} &= s_{j,k} |G_{j,\ell,k}|, & G_{j,\ell,k} &\sim \mathcal{N}(0, 4/\kappa_j), \\ v_{j,\ell} &= \frac{|Q_{j,\ell}|}{\sum_{h=1}^{H} |Q_{j,h}|}, & Q_{j,\ell} &\sim \mathcal{N}(0, 1/H), \end{aligned} \tag{32}$$

with independent equiprobable signs $s_{j,k} \in \{-1, 1\}$, one per parent–target pair shared across hidden units. The positive output weights sum to one. Holding the other inputs fixed, the deterministic mechanism is increasing in parent X^e_k when $s_{j,k} = 1$ and decreasing when $s_{j,k} = -1$. The nonlinear transformation of weighted sums permits interactions between parents. All weights, biases, and amplitudes were drawn once and shared across environments, and only the adjacency mask changed. Input weights retained their normalization by the original indegree after edge deletion. Subtracting $\tanh(b_{j,\ell})$ makes the deterministic contribution zero for a node with no parents.

Table 3: Synthetic experiment settings.

Parameter	Value
Dimension d	$10, 20$
Graph family	ER, SF
Edge factor r	$1, 2$
Observations n per environment	$300, 600, 900, 1200$
Noise distribution	Gaussian, exponential, Laplace, Gumbel
Noise standard deviation σ	0.7
Number of shifted targets	4
Maximum deleted parents per target	2

Noise distributions. Innovations are independent across observations, nodes, and environments. We used four centered, unit-variance families:

$$Z = \begin{cases} U, & U \sim \mathcal{N}(0,1), \\ U - 1, & U \sim \mathrm{Exp}(1), \\ U/\sqrt{2}, & U \sim \mathrm{Laplace}(0,1), \\ (U - \gamma_{\mathrm{E}})\sqrt{6}/\pi, & U \sim \mathrm{Gumbel}(0,1), \end{cases} \tag{33}$$

where the exponential parameter is the rate, the Laplace and Gumbel parameters are location and scale, and γ_{E} is the Euler–Mascheroni constant. Centering and scaling used population moments, rather than empirical moments of the generated samples. We set $\sigma = 0.7$, giving noise variance 0.49 in both environments. The noise distribution and scale were identical across environments, and changes were introduced only by edge deletions. Lastly, Table 3 summarizes the synthetic experiment settings.

E.2 Baselines

- **iSCAN (Chen et al., 2023).** It identifies shifted nodes using log-density Hessians and estimates shifted edges using FOCI (Azadkia & Chatterjee, 2021). We used the official implementation[1] and followed the parameter setting in the original paper.
- **ORION (Mian et al., 2023).** It utilizes greedy search with a minimum description-length (MDL) score (Rissanen, 1978) to discover causal networks and intervention targets across multiple environments. We followed default parameter setting for the authors' implementation[2].
- **SCORE (Rolland et al., 2022).** It uses score matching to estimate DAG structures under nonlinear additive Gaussian noise models. We used the authors' implementation[3], which we partially implemented in Python based on the R version. We followed its default settings.
- **DCI (Wang et al., 2018).** It estimates the skeleton of the difference-DAG using invariance tests of regression coefficients and orients some edges using invariance tests of regression residual variances. We followed default parameter setting for the `causaldag` package[4].
- **MG-LiNGAM (Shimizu, 2012).** It estimates a common causal ordering and the connection strengths of each group under linear models with independent non-Gaussian external influences. We used `MultiGroupDirectLiNGAM` from the `lingam` package[5]. We pruned the adjacency matrix of each group with adaptive lasso.

E.3 Implementation Details

Reproducibility. Our source code is available at this repository: `https://anonymous.4open.science/r/STORM-EE22`. The repository also includes instructions for obtaining the real-world datasets used in our experiments.

[1] `https://github.com/kevinsbello/iscan`
[2] `https://github.com/osman-mian/Orion`
[3] `https://github.com/paulrolland1307/SCORE`
[4] `https://github.com/uhlerlab/causaldag`
[5] `https://github.com/cdt15/lingam`

Computing infrastructure. All experiments were conducted on a MacBook Pro equipped with an Apple M2 Max processor and 64 GB of unified memory, running macOS (version 26.4).

Regression learners. For STORM, we used kernel ridge regression for common-order recovery, direction learning, and both centering regressions in parent-set estimation, with penalty 1 and the RBF kernel $K(z, z') = \exp\{-\|z - z'\|_2^2/(2p)\}$ for p predictors. Each kernel regression standardizes predictors and centers responses using its training rows, then restores the response mean at prediction time. Regressions with no predictors return the training response mean. Regression and numerical-library thread counts were set to one.

Common-order recovery. Each variable was centered and scaled using the pooled full samples from both environments, as described in Appendix C.1. Order recovery used five cross-fitting folds and a fresh Gaussian frequency dictionary at each sink removal step. With $p = |R| - 1$ predictors, frequencies satisfy $t_b \sim \mathcal{N}(0, 1)$ and $\omega_b \sim \mathcal{N}(0, I_p/p)$, independently for $b = 1, \ldots, B$. The dictionary was shared across candidates, environments, and folds within that round. We used $B = 256$ frequencies. Residuals in each fold were divided by their training-sample standard deviation, using denominator $n_{\text{train}} - 1$. The node with the smallest discrepancy in Eq. (12) was removed, breaking ties by node index. Fold assignments were reused across candidates and rounds, and equal-sized environments used the same index partitions.

Parent-set estimation. After common-order estimation, each environment sample was randomly split into a direction-learning sample containing $\lfloor n/2 \rfloor$ observations and a support-evaluation sample containing the remaining observations. Centering and scaling statistics were computed only from the pooled direction-learning samples and applied to both the direction-learning and support-evaluation samples. Equal-sized environments shared the split indices and support-evaluation fold assignments, and these assignments were reused across candidate edges. We calibrated the one-sided multiplier maximum in Appendix C.2 with 1999 Gaussian draws at $\alpha = 0.05$. Projected-covariance score standard deviations use denominator $n_{\text{eval}} - 1$, and cells with standard deviation below 10^{-6} remain unresolved. When at least one variance-eligible cell exists, the common threshold is the 1900th smallest multiplier maximum, with detection requiring a strictly larger statistic. If all cells are variance-ineligible, the threshold is $+\infty$ and no edge is detected. The final causal structure shift prediction was obtained by taking the symmetric difference of the detected parent sets across the two environments.

E.4 Evaluation Metrics

Structure shift detection. Let $\widehat{D}$ denote the predicted causal structure shift matrix. We compared it with the ground-truth causal structure shift matrix D in Eq. (30) over the common universe $\mathcal{U} = \{(k, j) \in V^2 : k \neq j\}$ of $d(d-1)$ directed pairs, including pairs that are incompatible with a method's estimated order. The positive class comprises pairs whose structure shifts between environments. The negative class includes both edges present in both environments and pairs absent in both. DCI represents an edge with unresolved orientation by positive entries in both directions, which were evaluated as separate directed predictions. The confusion counts are

$$\begin{aligned} \mathrm{TP} &= \sum_{(k,j)\in\mathcal{U}} D_{k,j}\widehat{D}_{k,j}, & \mathrm{FP} &= \sum_{(k,j)\in\mathcal{U}} (1 - D_{k,j})\widehat{D}_{k,j}, \\ \mathrm{FN} &= \sum_{(k,j)\in\mathcal{U}} D_{k,j}(1 - \widehat{D}_{k,j}), & \mathrm{TN} &= \sum_{(k,j)\in\mathcal{U}} (1 - D_{k,j})(1 - \widehat{D}_{k,j}). \end{aligned} \tag{34}$$

For each repetition, we computed the true-positive rate (TPR), false-positive rate (FPR), false discovery rate (FDR), F1 score, and structural Hamming distance (SHD):

$$\begin{gathered} \mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}, \qquad \mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}}, \qquad \mathrm{FDR} = \frac{\mathrm{FP}}{\max(\mathrm{TP} + \mathrm{FP}, 1)}, \\ \mathrm{F1} = \frac{2\mathrm{TP}}{2\mathrm{TP} + \mathrm{FP} + \mathrm{FN}}, \qquad \mathrm{SHD} = \mathrm{FP} + \mathrm{FN}. \end{gathered} \tag{35}$$

TPR measures the fraction of true causal structure shifts detected. FPR measures the fraction of unchanged directed pairs incorrectly declared shifted. FDR measures the fraction of detected shifts that are false positives. SHD is the unnormalized entrywise Hamming distance on the directed causal structure shift matrix, so reversing one changed edge incurs two errors: one false positive and one

false negative. Higher TPR and F1, along with lower FPR, FDR, and SHD, indicate more accurate causal structure shift detection.

Common-order recovery. For an estimated order $\widehat{\pi}$, write $\mathrm{pos}_{\widehat{\pi}}(j)$ for the position of node j. The topological order divergence counts edges in the union graph that point backward under the estimated order (Rolland et al., 2022), denoted by:

$$V_{\cup}(\widehat{\pi}) := \sum_{(k,j)\in\mathcal{E}_{\cup}} \mathbf{1}\{\mathrm{pos}_{\widehat{\pi}}(k) > \mathrm{pos}_{\widehat{\pi}}(j)\}. \tag{36}$$

Zero divergence indicates a valid common order, which need not be unique.

Aggregation across repetitions. The synthetic-results figures report arithmetic means of the displayed metrics across repetitions at each configuration, with shaded pointwise 95% percentile bootstrap confidence intervals for the mean from 1,000 resamples. Each metric was computed within a repetition before averaging. Thus, the plotted ratios average repetition-level rates rather than pooling confusion counts across repetitions. Repeated executions with the same resolved configuration were consolidated by retaining the latest completed run, so each configuration and seed contributed once.

F Additional Results

F.1 Structure Shift Detection

Figures 4-10 illustrate causal structure shift detection accuracy and computational costs across graph families, sparsity levels, noise distributions, and dimensions $d \in \{10, 20\}$. The experimental settings and aggregation follow Appendix E.

STORM generally achieved higher F1 than the baselines, including ORION, under both Gaussian and non-Gaussian noise. Larger samples generally improved F1 and SHD, with improvements extending to $d = 20$. DCI remained competitive on ER2 graphs, particularly under Gaussian noise at $d = 20$. On SF1 graphs, high TPR was accompanied by persistent false positives, limiting further F1 gains.

The order-stage correction improved F1 and SHD most consistently on ER2 and SF2 graphs, with less uniform gains on SF1. The full method and its ablation had similar runtimes and peak memory, indicating little computational overhead for the correction. Runtime increased with sample size and dimension, while STORM used substantially less peak memory than iSCAN.

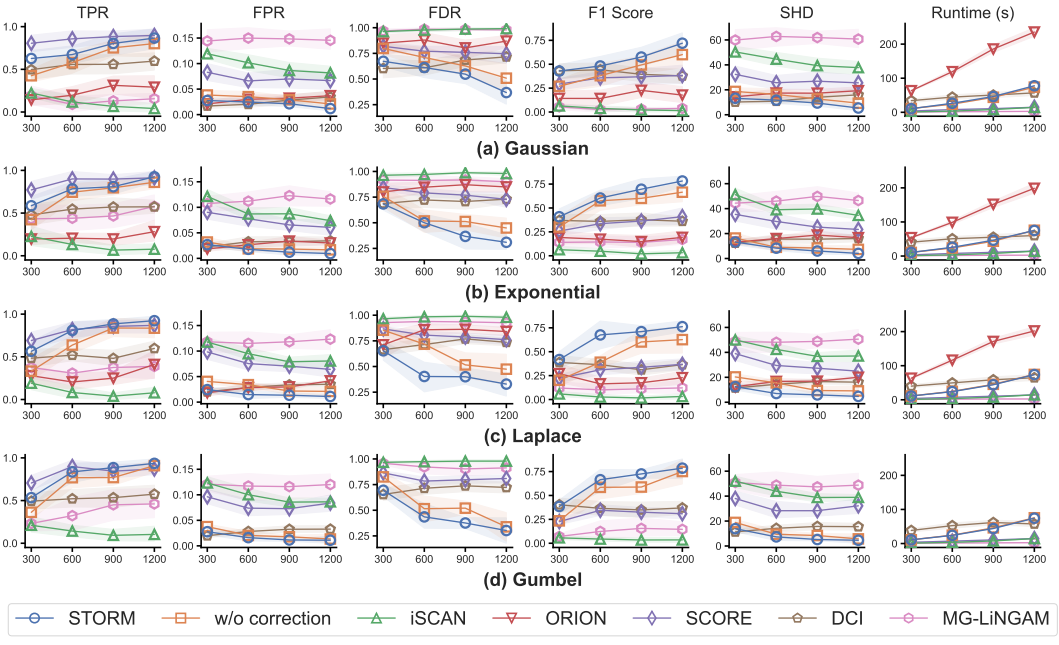


Figure 10: Causal structure shift detection on SF2 graphs with $d = 20$.

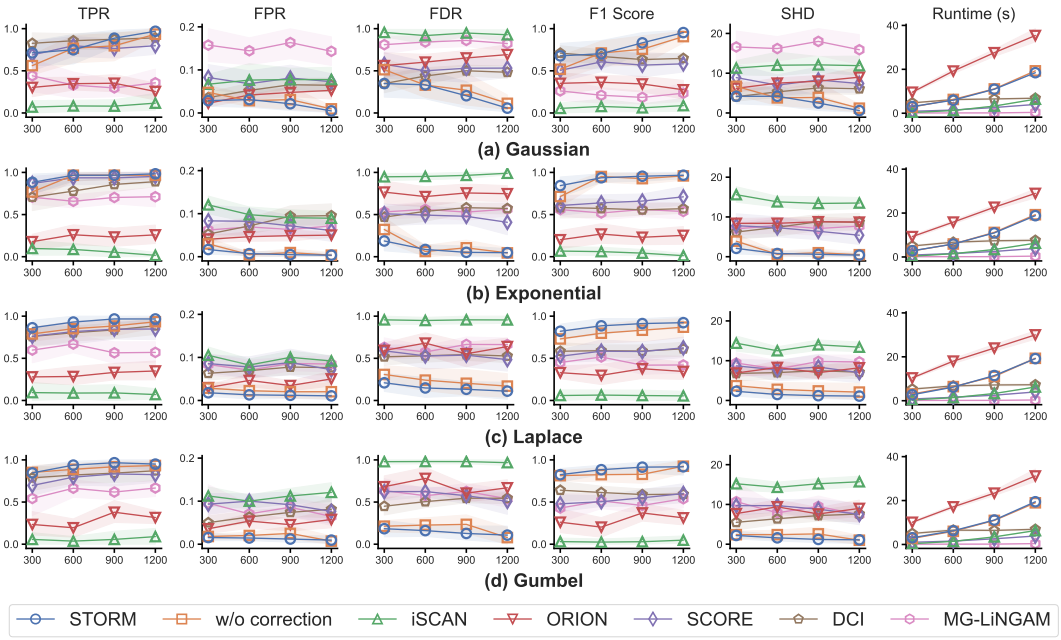


Figure 4: Causal structure shift detection on ER1 graphs with $d = 10$.

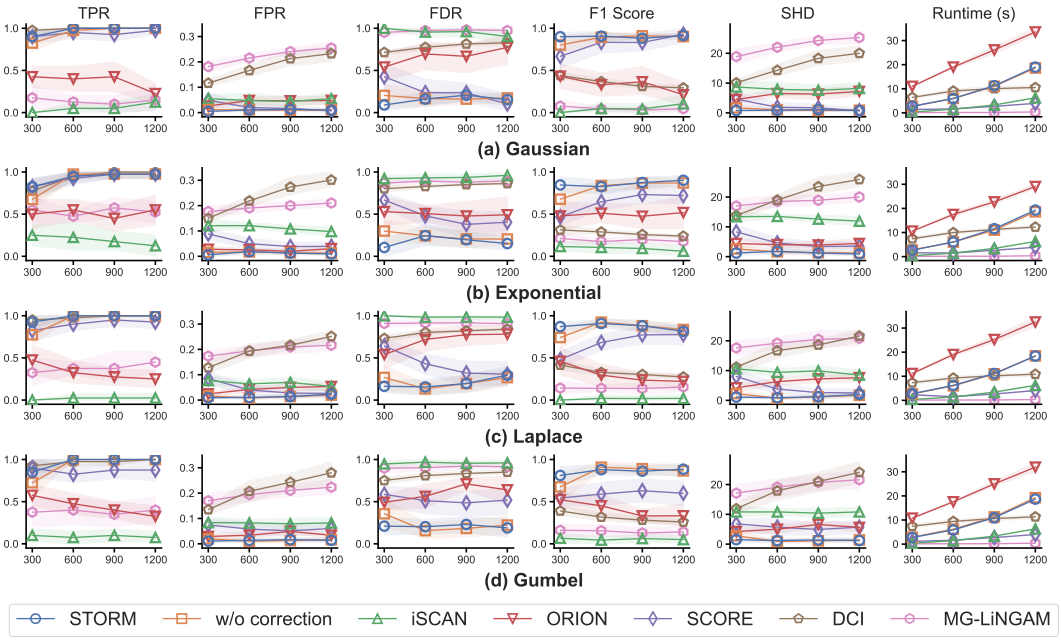


Figure 5: Causal structure shift detection on SF1 graphs with $d = 10$.

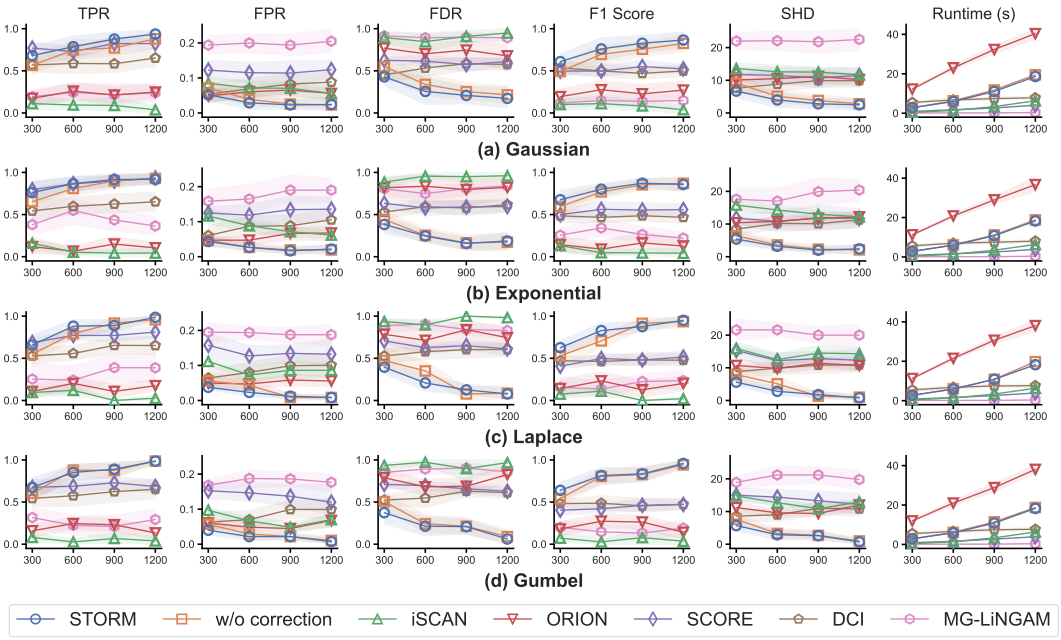


Figure 6: Causal structure shift detection on SF2 graphs with $d = 10$.

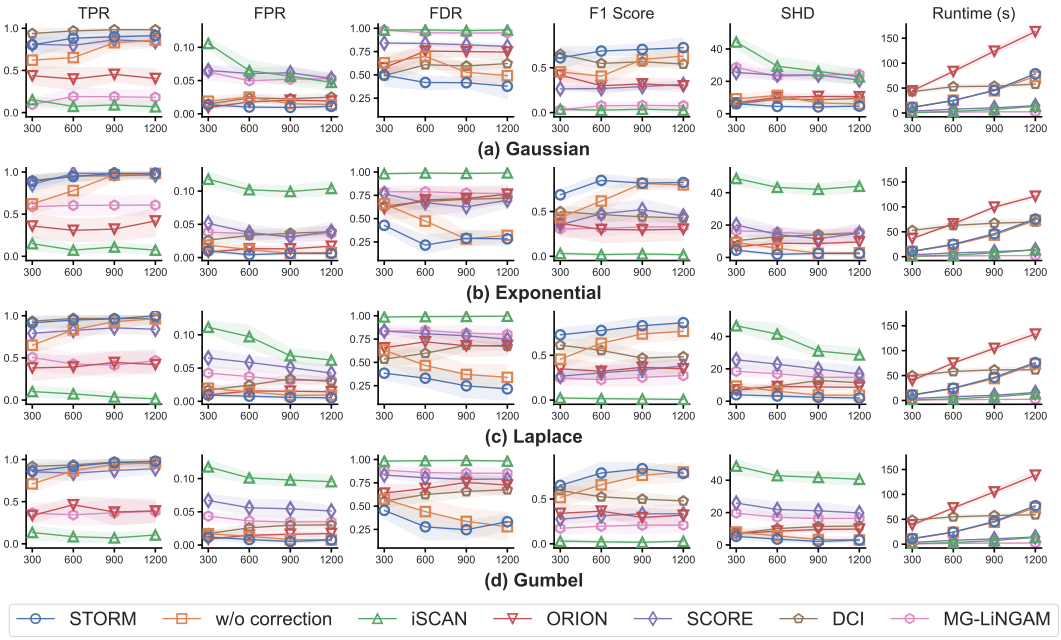


Figure 7: Causal structure shift detection on ER1 graphs with $d = 20$.

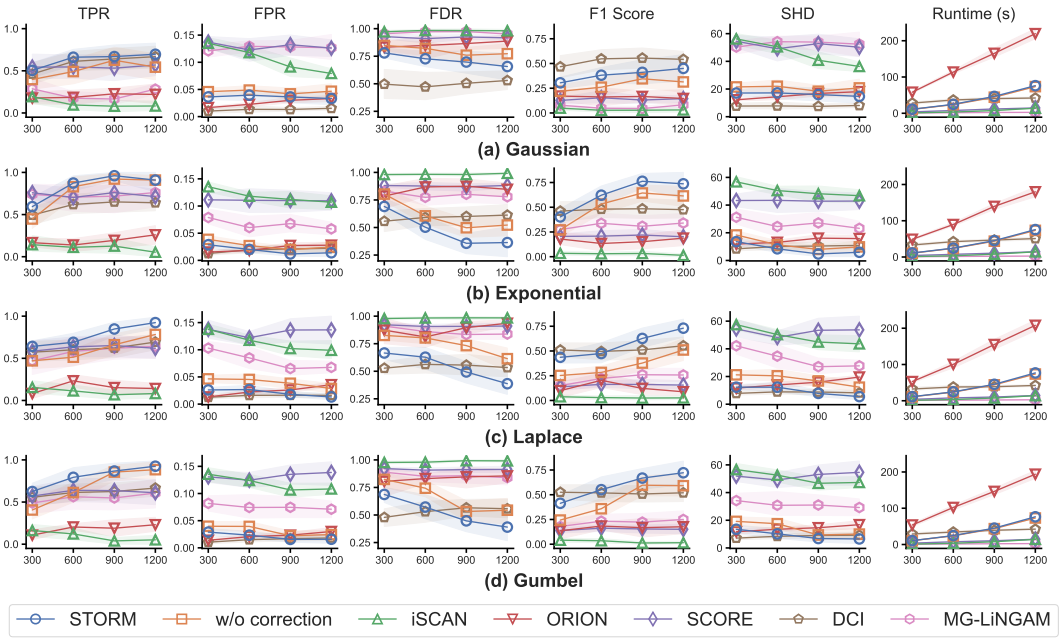


Figure 8: Causal structure shift detection on ER2 graphs with $d = 20$.

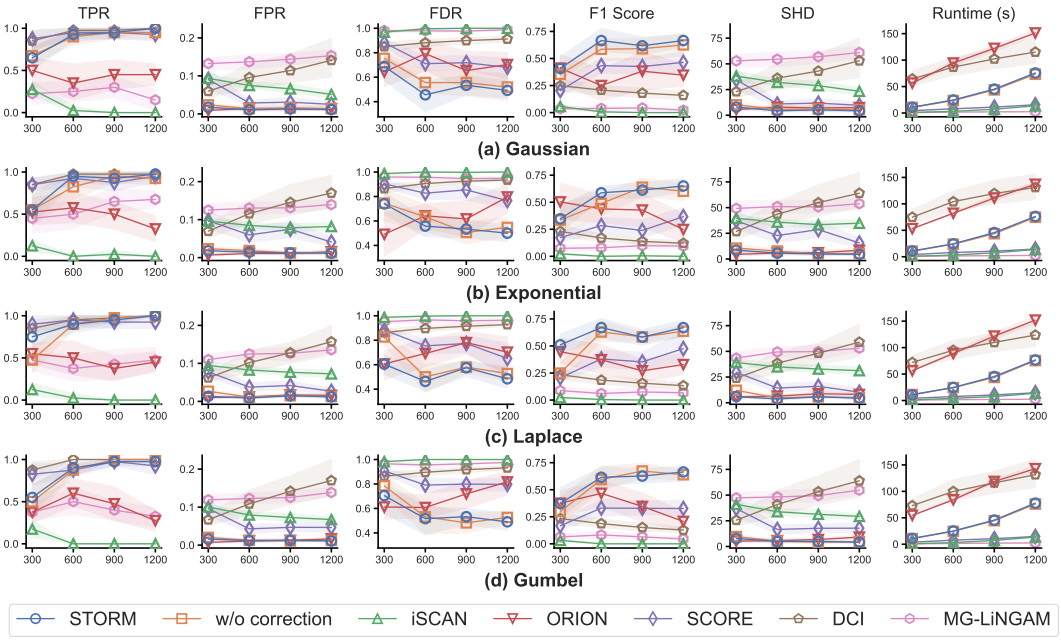


Figure 9: Causal structure shift detection on SF1 graphs with $d = 20$.

F.2 Common-Order Recovery

Figure 11 evaluates the estimated common causal order across the four graph families and noise distributions. The topological order divergence is the number of edges in the union graph that point backward under the estimated order, which provides good results when it is close to zero (Rolland et al., 2022). STORM generally reduced divergence as the sample size increased and had lower mean divergence than iSCAN throughout the displayed configurations. The correction reduced mean divergence relative to the ablation most clearly at smaller sample sizes and on ER2 graphs, while the curves often approach each other as more data become available. These results support the contribution of the order-stage correction to the structure shift detection improvements above.

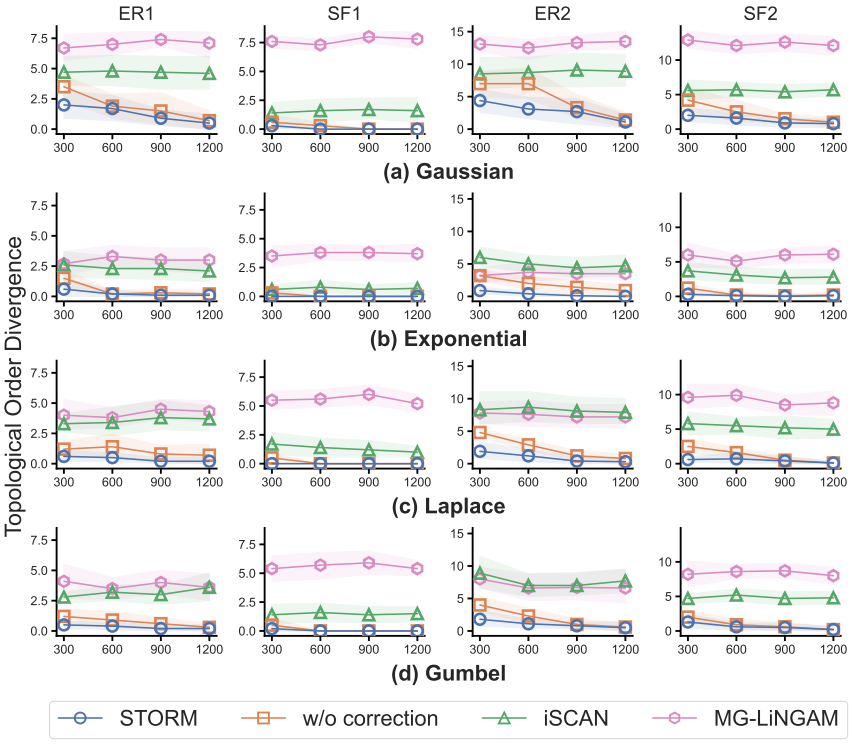


Figure 11: Common-order recovery on synthetic graphs with $d = 10$ variables (lower is better). Curves show means over ten random seeds, with shaded pointwise 95% percentile confidence intervals.

F.3 Regression Learner Sensitivity

Setup. We varied the regression learner used by STORM across RBF kernel ridge regression (KRR), XGBoost, ordinary least squares (OLS), spline ridge regression, and Extra Trees. KRR used the settings in Appendix E.3. Spline ridge used separate cubic B-spline expansions for each predictor, with eight quantile knots, constant extrapolation, and ridge penalty 1. XGBoost used 200 trees of maximum depth 2, learning rate 0.05, minimum child weight 10, row subsampling fraction 0.8, all predictors, and ℓ_2 regularization 1, with squared-error loss and histogram-based tree construction. Extra Trees used 50 trees, unrestricted depth, at least five observations per leaf, and all predictors at each split.

Results. Figures 12-15 report TPR, FPR, FDR, F1, SHD, and runtime, with aggregation as defined in Appendix E.4. The relative accuracy depended on the graph setting. On ER2 graphs at $n = 1200$, KRR attained the highest mean F1 among the five learners under each noise distribution, with values ranging from 0.82 to 0.94. On SF1 graphs at the same sample size, spline ridge and XGBoost achieved higher mean F1 than KRR under all four noise distributions, reaching 0.97–0.98 and 0.93–1.00, respectively. OLS showed a different pattern on SF1: increasing sample size raised its mean FPR under every noise distribution, and its mean FDR remained between 0.55 and 0.61 at $n = 1200$.

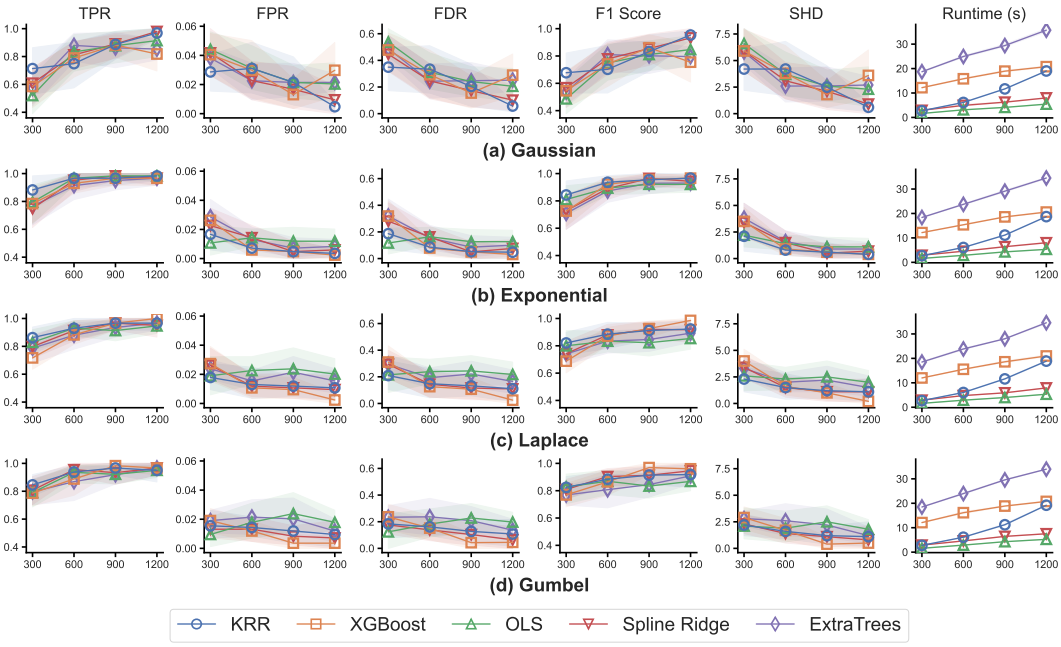


Figure 12: Regression learner sensitivity of STORM on ER1 graphs with $d = 10$. Curves show means over ten random seeds, with shaded pointwise 95% percentile bootstrap confidence intervals. The horizontal axis is the sample size per environment.

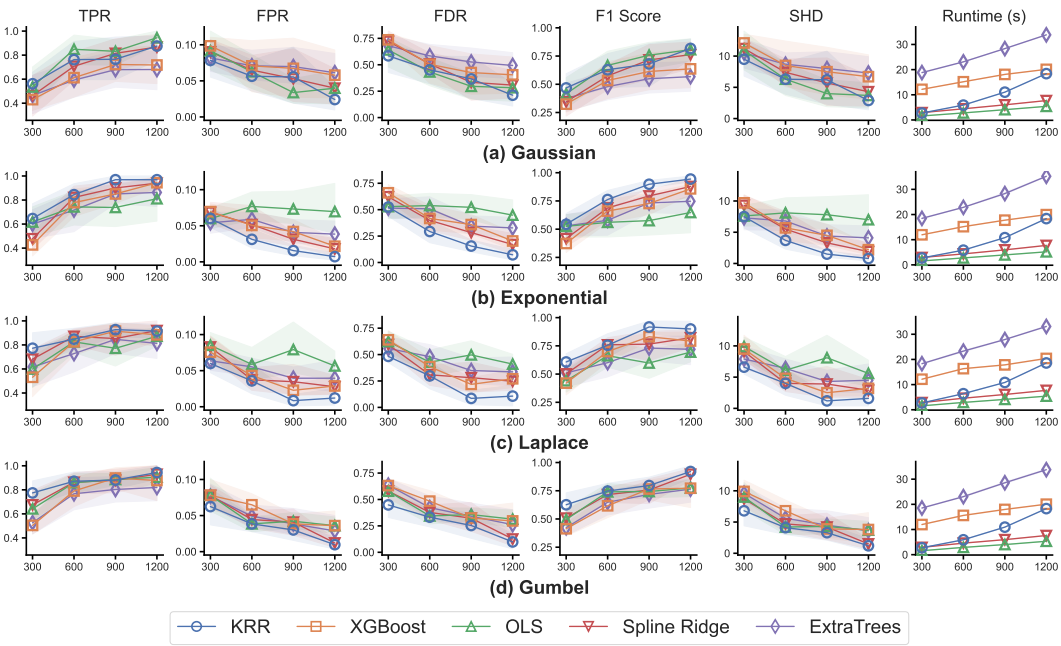


Figure 13: Regression learner sensitivity of STORM on ER2 graphs with $d = 10$.

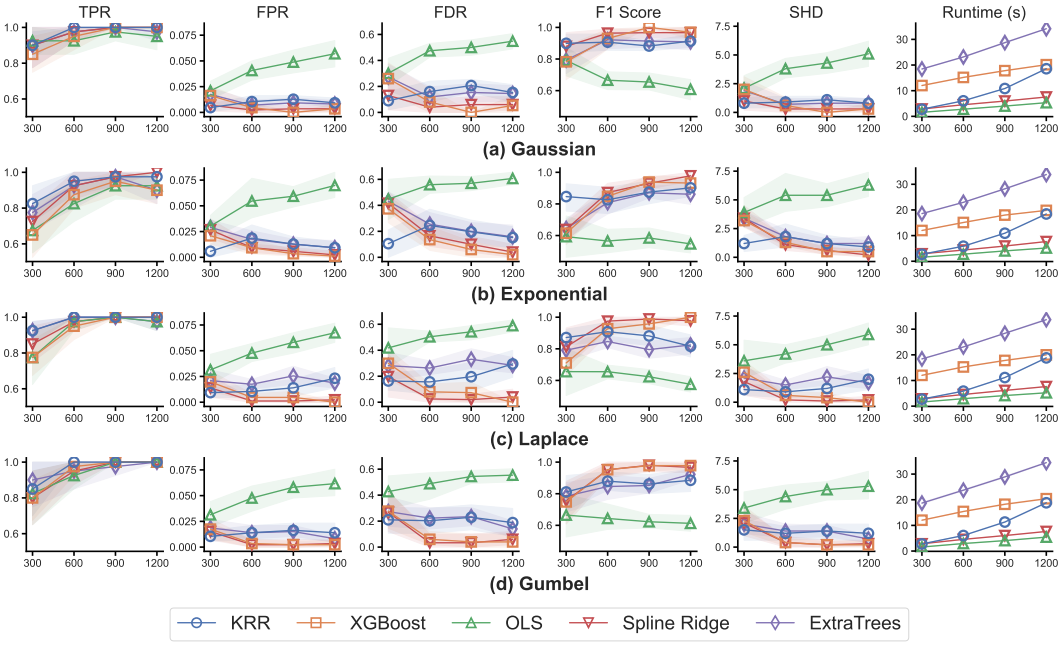


Figure 14: Regression learner sensitivity of STORM on SF1 graphs with $d = 10$.

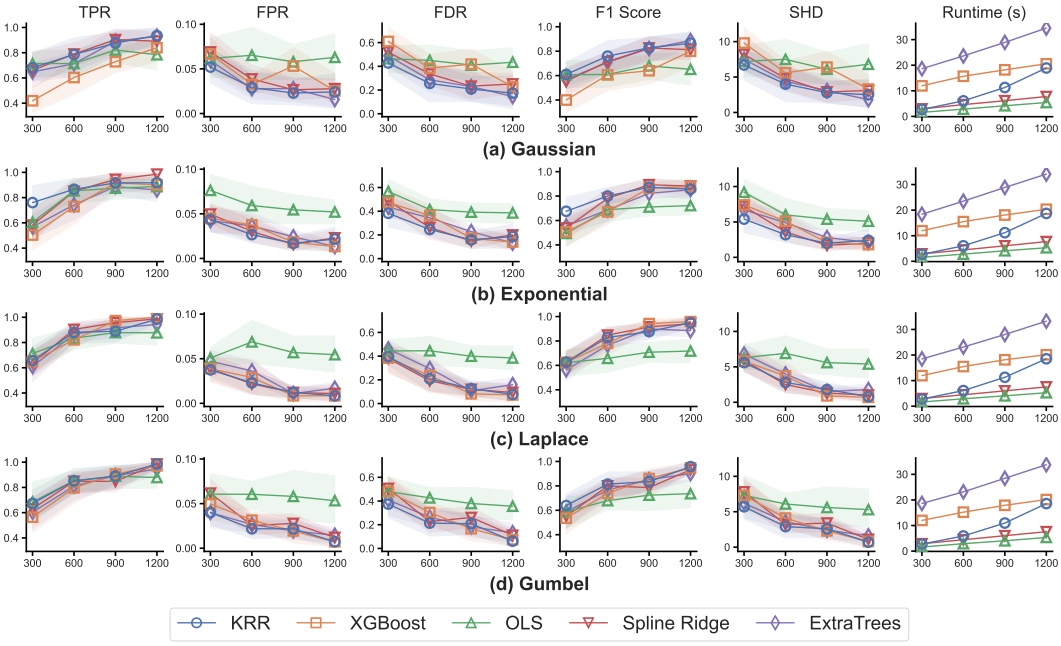


Figure 15: Regression learner sensitivity of STORM on SF2 graphs with $d = 10$.

F.4 Further Results and Discussion on Mass-Cytometry Data

For the BT-20 comparison in §4.2, both conditions used EGF (100 ng/ml) and fetal bovine serum (10%). In iMEK, cells were pretreated with CI-1040 (PD184352, 1.7 μM) for 15 minutes before stimulation. We used all cells in the two conditions and the author-provided compensated, arcsinh-

Table 4: Selected edges inferred by STORM in BT-20 cells.

Inferred relationship	EGF	iMEK
$\{\mathrm{MEK}, \mathrm{ERK}, \mathrm{p90RSK}\} \rightarrow \mathrm{4EBP1}$	✓	—
$\{\mathrm{AKT(S473)}, \mathrm{AKT(T308)}\} \rightarrow \mathrm{4EBP1}$	✓	✓
$\mathrm{HER2} \rightarrow \{\mathrm{AKT(S473)}, \mathrm{AKT(T308)}\}$	—	✓
$\mathrm{PDPK1} \rightarrow \mathrm{AKT(T308)}$	—	✓
$\mathrm{ERK} \rightarrow \mathrm{p90RSK}$	✓	✓
$\{\mathrm{ERK}, \mathrm{p90RSK}\} \rightarrow \mathrm{S6K}$	✓	✓
$\mathrm{AKT(S473)} \rightarrow \mathrm{S6}$	✓	✓
$\mathrm{ERK} \rightarrow \mathrm{S6}$	—	✓
$\mathrm{MKK3/6} \rightarrow \mathrm{p38}$	✓	✓
$\mathrm{MKK4} \rightarrow \mathrm{p38}$	—	✓

transformed, and batch-corrected measurements.[6] Protein names below denote their phosphorylation channels; AKT(S473) and AKT(T308) denote the two separately measured AKT sites.

4EBP1 regulation. MEK or RSK inhibition reduced 4EBP1 phosphorylation at T37/46, the sites measured here, in BRAF-mutant melanoma cells (Romeo et al., 2013). AKT/MEK inhibition and mutant-allele deletion experiments further showed that the dependence of 4EBP1 on either pathway changes with the available signaling inputs, and that single-pathway inhibition can leave its phosphorylation intact (She et al., 2010). In another breast cancer cell line, MDA-MB-231, CI-1040 suppressed ERK phosphorylation while largely preserving 4EBP1 T37/46 phosphorylation (Mirzoeva et al., 2009, Fig. 6B).

Upstream AKT signaling. Upstream of AKT, HER2 $\rightarrow$ AKT(S473), HER2 $\rightarrow$ AKT(T308), and PDPK1 $\rightarrow$ AKT(T308) were detected only in iMEK. Pathway-level support for the HER2–AKT relationships comes from evidence that MEK inhibition relieves negative feedback on ERBB receptors and promotes ERBB3-dependent AKT signaling in HER2-dependent cancer cells (Turke et al., 2012). The PDPK1 $\rightarrow$ AKT(T308) direction is consistent with the loss of IGF1-induced AKT T308 phosphorylation in mouse embryonic stem cells lacking PDPK1, also called PDK1 (Collins et al., 2003).

S6K and S6 regulation. ERK $\rightarrow$ S6K and p90RSK $\rightarrow$ S6K were detected in both conditions, consistent with the regulation of S6K T389 phosphorylation by MEK/RSK signaling (Romeo et al., 2013). A different response appeared at S6: ERK $\rightarrow$ S6 was detected only in iMEK, alongside a common AKT(S473) $\rightarrow$ S6 edge. The measured S6 sites, S235/236, receive input from both RSK and S6K, as established by kinase assays and RSK perturbations (Roux et al., 2007); AKT inhibition also reduced S6 phosphorylation in breast cancer cells (She et al., 2010). The condition-specific ERK $\rightarrow$ S6 detection motivates testing whether S6 remains responsive to residual ERK signaling under MEK inhibition, even as the ERK $\rightarrow$ 4EBP1 relationship is no longer detected.

p38 regulation. For p38, STORM retained MKK3/6 $\rightarrow$ p38 in both conditions and detected MKK4 $\rightarrow$ p38 only in iMEK. The coexistence of shared and condition-specific inputs is consistent with experiments showing that the upstream control of p38 depends on the stimulus: in mouse embryonic fibroblasts, combined loss of MKK3 and MKK6 severely impaired TNF-induced p38 activation, whereas MKK4 contributed to residual UV-induced p38 activation in those cells (Brancho et al., 2003). One possible interpretation is that MEK inhibition makes a normally redundant, stress-responsive MKK4 input more influential in p38 regulation. This would predict greater sensitivity of p38 phosphorylation to MKK4 depletion in iMEK than in EGF.

[6] Data were provided by the Bodenmiller laboratory through the SCSBrC DREAM challenge and accessed via Synapse (`https://www.synapse.org/Synapse:syn20564743`; BT-20 data: `syn20613595`, version 1).

F.5 Results on Flow-Cytometry Data

We also applied STORM to the flow-cytometry data[7] of Sachs et al. (2005), comprising 11 protein and phospholipid readouts in human primary naive $CD4^+$ T cells measured 15 minutes after stimulation. We compared all 853 cells stimulated with anti-CD3/CD28 (control) with all 810 cells receiving the same stimulation plus psitectorigenin, which perturbs phosphoinositide turnover. The estimator settings followed Appendix E.3, except that common-order recovery used $B = 4{,}096$ frequencies with $t_b \sim \mathcal{N}(0, 4)$ and $\omega_b \sim \mathcal{N}(0, 4I_p/p)$ for p predictors.

Figure 16 shows the complete union of the estimated graphs: 10 edges were detected in control and 7 under psitectorigenin, with 7 shared edges and 3 candidate structure shifts. The three condition-specific edges, Plcg $\to$ PIP2, Plcg $\to$ PIP3, and PIP2 $\to$ PIP3, were detected only in control; no edge was detected only under psitectorigenin. The detected differences were therefore confined to the Plcg–PIP2–PIP3 subnetwork, while relationships including Raf $\to$ Mek, PKA $\to$ Erk, and Erk $\to$ Akt were retained in both conditions. The localization is consistent with the intervention's involvement in phosphoinositide turnover.

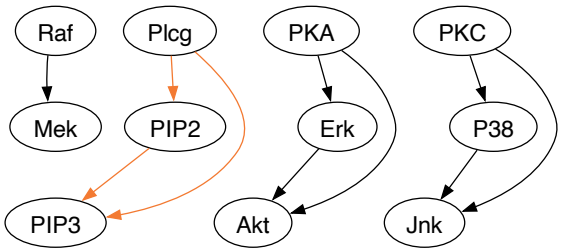


Figure 16: Estimated graphs for the flow-cytometry data. Orange edges were detected only under stimulation alone.

[7] Version 2: `https://doi.org/10.5281/zenodo.7681811`.