GLORI: GRAPH-LOCAL ORDERING FROM RESIDUAL INSTABILITY ACROSS ENVIRONMENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

We study causal ordering from multi-environment data when an undirected graph skeleton is available but intervention targets are unknown. Our starting observation is that a stable causal mechanism can leave similar prediction errors across environments, whereas reverse prediction can change with the input distribution. We introduce GLORI (Graph-Local Ordering from Residual Instability). For each node, GLORI trains one predictor using all environments together, evaluates the same predictor separately in each environment, and measures changes in its environment-wise residual mean. The resulting node instability scores define a causal order. We establish the residual-instability contrast theoretically in a tractable one-root setting and then evaluate whether the same signal remains effective beyond this setting. Across 1,440 synthetic graph tasks spanning diverse graph structures and causal mechanisms, as well as real-world CausalBench data, GLORI consistently demonstrates strong causal-ordering performance while remaining computationally efficient relative to the compared methods.

1 INTRODUCTION

Distribution changes across environments can reveal causal asymmetries that are not identifiable from a single observational distribution. When the distribution of a cause changes while the mechanisms generating its effects remain unchanged, the conditional distributions associated with the causal direction can remain stable, whereas the corresponding reverse conditionals can change (Peters et al., 2016; Mooij et al., 2020; Huang et al., 2020; Perry et al., 2022). This principle provides an important source of information for causal discovery from multiple experimental conditions. Exploiting it in a multivariate graph, however, requires accounting for the local dependencies surrounding each variable.

Consider three variables connected by the undirected graph skeleton

$$X_1 - X_0 - X_2.$$

The skeleton indicates which variables are adjacent but contains no information about edge directions. For illustration, suppose that the underlying causal structure is

$$X_1 \longleftarrow X_0 \longrightarrow X_2,$$

and suppose that P_e, the data distribution in environment e, changes through $P_e(X_0)$, while the relationships from X_0 to X_1 and X_2 remain unchanged. Then the two conditionals in the causal direction satisfy

$$P_e(X_1 \mid X_0) = P(X_1 \mid X_0), \qquad P_e(X_2 \mid X_0) = P(X_2 \mid X_0).$$

Thus, a single model trained across all environments to predict X_1 or X_2 from X_0 encounters the same relationship in every environment. Its residuals should therefore remain stable across environments.

Predicting X_0 from X_1 and X_2 is different. By Bayes' rule,

$$P_e(X_0 \mid X_1, X_2) \propto P(X_1, X_2 \mid X_0) P_e(X_0).$$

The reverse conditional now depends explicitly on the changing distribution $P_e(X_0)$. A single model trained across all environments to predict X_0 from (X_1, X_2) therefore encounters a different

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.

Table 1: Prediction contrast in the three-node example. The forward conditionals yield stable residual behavior, whereas the reverse conditional varies with $P_e(X_0)$ and yields unstable residual behavior.

Prediction	Across environments	Residual behavior
$X_0 \mapsto X_1$	$P_e(X_1 \mid X_0) = P(X_1 \mid X_0)$	Stable
$X_0 \mapsto X_2$	$P_e(X_2 \mid X_0) = P(X_2 \mid X_0)$	
$(X_1, X_2) \mapsto X_0$	$P_e(X_0 \mid X_1, X_2)$ varies with $P_e(X_0)$	Unstable

relationship in each environment, causing its residuals to become unstable across environments. Table 1 summarizes this contrast.

The contrast requires only one prediction problem for each node. The two leaf nodes are predicted from their common neighbor X_0, while X_0 is predicted jointly from both X_1 and X_2. In this example, the residual for X_0 is more unstable across environments than those for X_1 and X_2. Placing variables with larger residual instability earlier therefore puts X_0 first and recovers

$$X_1 \longleftarrow X_0 \longrightarrow X_2.$$

We introduce **GLORI**, short for **Graph-Local Ordering from Residual Instability**. For each node, GLORI trains one predictor using all environments together, with all neighbors specified by the graph skeleton as inputs. It then evaluates the same predictor separately in each environment, summarizes the resulting residuals by their environment-wise mean, and measures how much these means differ across environments. Nodes are ordered from larger to smaller residual instability, and every supplied skeleton edge is directed from the node appearing earlier in this order to the node appearing later.

Beyond its simple construction, GLORI connects a local distributional signal to a globally consistent causal order. We establish this contrast theoretically in a tractable one-root linear model, where the ordering signal can be characterized exactly. We then evaluate whether the same local signal remains useful in more general settings, across 1,440 synthetic graph tasks with diverse graph structures and causal mechanisms and on real-world biological data. The results show that residual instability remains a practical ordering signal well beyond the structure covered by the theorem.

Our contributions are threefold:

- We propose GLORI, which measures how much the mean residual of each variable changes across environments.
- We theoretically analyze residual instability when each variable is predicted from all of its neighbors and prove population recovery for a linear-Gaussian graph with one root and multiple children.
- Beyond this theoretical setting, we evaluate GLORI across diverse synthetic graph structures and causal mechanisms and on real-world biological data, showing that the same residual-instability signal remains effective in more complex settings.

2 Related Work

Causal discovery across environments. Data collected from different environments can reveal causal information that is unavailable from a single observational distribution. Invariant Causal Prediction searches for predictors whose relationship with a chosen response remains stable across environments (Peters et al., 2016). Joint Causal Inference introduces environment labels as context variables and learns causal relationships over the combined system (Mooij et al., 2020). CD-NOD uses changes across domains or time to learn both causal connections and the mechanisms that vary (Huang et al., 2020). MSS searches for a causal graph that explains the observed distribution changes using as few changing mechanisms as possible (Perry et al., 2022). iSCAN compares multiple datasets to identify which variables have undergone mechanism changes (Chen et al., 2023). UT-IGSP considers experimental data in which the intervention targets are unknown and jointly estimates the targets and the causal structure (Squires et al., 2020); related interventional approaches

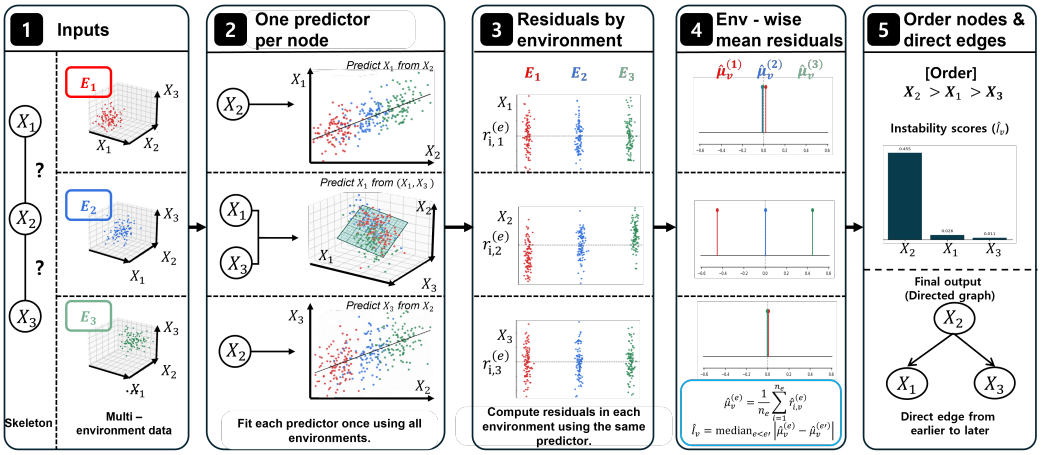

Figure 1: Overview of GLORI. Given a supplied undirected skeleton and observations from multiple environments, GLORI trains one predictor per node using all environments together, evaluates the same predictor separately in each environment, and computes environment-wise residual means. The median pairwise absolute difference between these means defines a node-level instability score. Nodes are ordered from larger to smaller instability, and each skeleton edge is directed from the earlier to the later node, yielding a fully directed acyclic graph over the supplied skeleton.

include differentiable causal discovery and neural unknown-target discovery (Brouillard et al., 2020; Lippe et al., 2022). Although these methods differ in their assumptions and outputs, they share the principle that stable and changing relationships across environments provide information about causal structure. GLORI's score can also be read as a target-sensitive statistic, but it does not estimate intervention targets or feed them into a graph search; instead, it consumes the statistic directly as a node ordering over a supplied skeleton.

Causal ordering. Causal discovery can also be simplified by first estimating an order of the variables and then selecting the connections that are consistent with that order. SCORE constructs such an order by repeatedly identifying variables with no children under nonlinear additive Gaussian noise models (Rolland et al., 2022). NoGAM extends this approach to additive models with more general noise distributions by combining score estimation with regression residuals (Montagna et al., 2023). DiffAN uses a diffusion model to estimate the required score information and similarly builds an order through iterative variable removal (Sanchez et al., 2023). These methods obtain a causal order from a single observational distribution and typically use an additional pruning step to recover the final graph.

Residual asymmetry and graph skeletons. Residuals provide another source of information about causal direction. In an additive noise model, predicting an effect from its cause can leave a residual that is independent of the cause, whereas prediction in the reverse direction generally does not (Hoyer et al., 2008; Peters et al., 2014). Related approaches distinguish directions by comparing regression errors (Blöbaum et al., 2018). Classical constraint-based methods instead use conditional independence relations to first determine which variables are adjacent and then direct the connections supported by the resulting equivalence class (Colombo & Maathuis, 2014). Their output may remain partially directed when a single observational distribution does not contain enough information to determine every direction.

GLORI combines these ideas by training one predictor for each variable using all environments together, with its neighbors in the supplied graph skeleton as inputs, and measuring how its mean residual changes across environments.

3 PROBLEM AND METHOD

Figure 1 provides an overview of GLORI, from per-node prediction across environments and residual-instability scoring to node ordering and edge direction assignment.

3.1 PROBLEM SETTING

Let $\mathcal{E} = \{1, \ldots, K\}$ denote a set of $K \geq 2$ environments. In each environment $e \in \mathcal{E}$, we observe

$$\mathcal{D}^{(e)} = \{\mathbf{x}_i^{(e)}\}_{i=1}^{n_e}, \qquad \mathbf{x}_i^{(e)} \in \mathbb{R}^d, \qquad \mathbf{x}_i^{(e)} \overset{\text{i.i.d.}}{\sim} P_e. \tag{1}$$

Here, $\mathcal{D}^{(e)}$ is the dataset collected in environment e, n_e is its number of observations and may differ across environments, and $\mathbf{x}_i^{(e)}$ is the i-th observation containing the values of all d variables. The distribution P_e denotes the joint distribution of these variables in environment e. The same variables $\mathbf{X} = (X_0, \ldots, X_{d-1})$ are observed in every environment.

We assume a shared directed acyclic causal graph across environments and observe the environment labels, but not which variables or mechanisms change. We are given its undirected skeleton

$$G = (V, E), \tag{2}$$

where G represents the supplied adjacency structure, $V = \{0, \ldots, d-1\}$ is the set of observed variables, and $\{u, v\} \in E$ indicates that X_u and X_v are adjacent without specifying the edge direction. The skeleton specifies that the two variables are connected but does not specify whether the direction is $X_u \to X_v$ or $X_v \to X_u$.

For each node $v \in V$, its neighborhood in G is

$$\mathcal{N}_G(v) = \{u \in V : \{u, v\} \in E\}. \tag{3}$$

For example, the three-variable skeleton from Section 1 is

$$X_1 - X_0 - X_2, \tag{4}$$

which corresponds to

$$V = \{0, 1, 2\}, \qquad E = \big\{\{0, 1\}, \{0, 2\}\big\}. \tag{5}$$

Its neighborhoods are

$$\mathcal{N}_G(0) = \{1, 2\}, \qquad \mathcal{N}_G(1) = \{0\}, \qquad \mathcal{N}_G(2) = \{0\}. \tag{6}$$

Thus, X_0 is predicted from (X_1, X_2), whereas X_1 and X_2 are each predicted from X_0.

Our objective is to estimate a causal order

$$\widehat{\pi} = (\pi_1, \ldots, \pi_d), \tag{7}$$

where $\pi_a \in V$ denotes the node placed at position a. In a correct causal order, causes appear before their effects. The estimated order is then used to direct every edge in the supplied skeleton.

3.2 PER-NODE PREDICTION ACROSS ENVIRONMENTS

Before fitting, each variable is standardized once using its mean and standard deviation computed over all environments. The same standardization is then used in every environment.

For every non-isolated node v, GLORI predicts X_v from its neighboring variables $\mathbf{X}_{\mathcal{N}_G(v)}$. Since the supplied skeleton is undirected, the neighborhood may contain both causes and effects of X_v. GLORI uses the entire neighborhood without assigning causal roles before prediction. The predictor for each node is trained using samples from all environments:

$$\widehat{f}_v = \arg\min_{f \in \mathcal{F}} \left\{ \sum_{e \in \mathcal{E}} \sum_{i=1}^{n_e} \left[x_{i,v}^{(e)} - f\left(\mathbf{x}_{i,\mathcal{N}_G(v)}^{(e)}\right) \right]^2 + \lambda \Omega(f) \right\}, \tag{8}$$

where $\mathcal{F}$ denotes the class of affine linear prediction models, $\Omega(f)$ is the squared ℓ_2 norm of the non-intercept coefficients, and $\lambda \geq 0$ controls the regularization strength. Each observation receives equal weight in Equation (8), so environments with more observations contribute proportionally more to the fit. The predictor $\widehat{f}_v$ is fitted once using samples from all environments and then kept fixed. Further implementation details are provided in Appendix B.

3.3 RESIDUAL INSTABILITY

Applying the same predictor in environment e gives the residual random variable

$$R_v^{(e)} = X_v^{(e)} - \widehat{f}_v\left(\mathbf{X}_{\mathcal{N}_G(v)}^{(e)}\right). \tag{9}$$

Because the same predictor $\widehat{f}_v$ is used in every environment, changes in its residuals reflect how differently this predictor explains X_v across environments.

For the observed samples, we compute

$$r_{i,v}^{(e)} = x_{i,v}^{(e)} - \widehat{f}_v\left(\mathbf{x}_{i,\mathcal{N}_G(v)}^{(e)}\right), \tag{10}$$

and summarize the residuals in each environment by their mean,

$$\widehat{\mu}_v^{(e)} = \frac{1}{n_e}\sum_{i=1}^{n_e} r_{i,v}^{(e)}. \tag{11}$$

Thus, $\widehat{\mu}_v^{(e)}$ measures the average signed prediction error of the predictor for node v in environment e.

We define the residual-instability score of node v as the median absolute difference in its mean residual across pairs of environments:

$$\widehat{I}_v = \underset{1\le e<e'\le K}{\text{median}} \left|\widehat{\mu}_v^{(e)} - \widehat{\mu}_v^{(e')}\right|. \tag{12}$$

A large $\widehat{I}_v$ indicates that the same predictor has different systematic prediction errors across environments, whereas a small $\widehat{I}_v$ indicates that its mean residual remains stable. For an isolated node, its position in the order does not affect any edge direction because it has no incident edges. In the implementation, we use its cross-environment marginal-mean variation only to determine its position in the total order.

3.4 CAUSAL ORDERING AND EDGE ASSIGNMENT

GLORI converts the node-wise instability scores into a total order by placing nodes with larger residual instability earlier:

$$\widehat{I}_{\pi_1} \ge \widehat{I}_{\pi_2} \ge \cdots \ge \widehat{I}_{\pi_d}. \tag{13}$$

GLORI places nodes with larger $\widehat{I}_v$ earlier, motivated by Theorem 1, where the shifted root has larger residual instability than its stable children.

Given the estimated order $\widehat{\pi}$, GLORI directs each edge in the supplied skeleton from the node appearing earlier in the order to the node appearing later. The resulting directed graph is

$$\widehat{G} = (V, \widehat{A}), \tag{14}$$

where

$$\widehat{A} = \{(\pi_r, \pi_s) \mid 1 \le r < s \le d,\ \{\pi_r, \pi_s\} \in E\}. \tag{15}$$

The ordering only determines the directions of the supplied edges. Since every directed edge follows the same total order, $\widehat{G}$ is acyclic by construction.

One-root population recovery. We first establish the residual-instability contrast in a tractable local setting. Consider a one-root causal graph

$$X_0 \longrightarrow X_j, \qquad j = 1, \ldots, m,$$

where each child is generated linearly from X_0 with independent, zero-mean, environment-invariant noise. The environments differ through the mean of X_0, while the root-to-child mechanisms remain unchanged. Let $\mu_e = \mathbb{E}[X_0^{(e)}]$ denote the root mean in environment e, and let I_v denote the population counterpart of the residual-instability score $\widehat{I}_v$.

Theorem 1 (One-root population recovery) *If* $\Delta_\mu = \mathrm{median}_{e<e'} |\mu_e - \mu_{e'}| > 0$*, then*

$$I_0 > 0, \qquad I_j = 0 \quad \textit{for every } j = 1, \dots, m. \tag{16}$$

Consequently, GLORI places X_0 *before every child and recovers all root-to-child directions in the supplied skeleton.*

Theorem 1 explains the residual-instability contrast in a one-root setting. The GLORI method itself applies to general graph structures, and the experiments in Section 4 evaluate whether the same residual-instability signal remains effective in more complex multi-node graphs.

Computational cost. GLORI fits at most one neighborhood predictor per non-isolated node. Let

$$N = \sum_{e=1}^{K} n_e, \qquad k = \max_{v \in V} |\mathcal{N}_G(v)|, \tag{17}$$

where N is the total number of observations and k is the maximum degree of the supplied skeleton. Let $C_{\mathrm{fit}}(N, p)$ and $C_{\mathrm{eval}}(N, p)$ denote the costs of fitting and evaluating the chosen prediction model on N observations with p input variables. Fitting and evaluating all neighborhood predictors therefore costs

$$O\left(d\, C_{\mathrm{fit}}(N, k) + d\, C_{\mathrm{eval}}(N, k)\right). \tag{18}$$

Computing residual summaries across all nodes costs $O(Nd)$, while pairwise comparisons across the K environments cost $O(K^2 d)$. Sorting the nodes costs $O(d \log d)$, and directing the supplied skeleton adds $O(|E|)$, giving the total cost

$$O\left(d\, C_{\mathrm{fit}}(N, k) + d\, C_{\mathrm{eval}}(N, k) + Nd + K^2 d + d \log d + |E|\right). \tag{19}$$

In practice, the computational cost is dominated by fitting and evaluating the neighborhood prediction models.

4 Experiments

4.1 Experimental setup

We evaluate GLORI on synthetic multi-environment causal graphs and on the K562 and RPE1 perturbation datasets from CausalBench (Chevalley et al., 2025). For synthetic experiments, all estimated orders are evaluated on the same supplied ground-truth skeleton, so the comparison isolates edge orientation. Performance is measured by graph-averaged edge-direction accuracy with 95% confidence intervals from 10,000 bootstrap resamples over graph tasks. For CausalBench, we report W1 and precision.

We compare against MSS–Soft (Perry et al., 2022), SCORE (Rolland et al., 2022), NoGAM (Montagna et al., 2023), and iSCAN (Chen et al., 2023). MSS–Soft and iSCAN are multi-environment methods. SCORE and NoGAM are observational causal-ordering methods, which we apply to the pooled samples.

The full-data setting uses 500 observations from each of 10 environments, whereas the equal-sample setting uses the same stratified 10 observations per environment for all methods. Runtime is reported in seconds per graph for synthetic experiments and seconds per evaluation instance for real-data experiments. All experiments were run on a single NVIDIA RTX PRO 6000 Blackwell Edition GPU.

4.2 Main results

4.2.1 Synthetic results

Table 2 reports the main synthetic results. Unlike the one-root structure used for the theoretical analysis, these experiments cover general multi-node graphs with diverse structures and causal mechanisms, testing whether the residual-instability signal remains useful beyond the setting covered by Theorem 1. With full data, GLORI achieves the highest overall accuracy at 0.704 ± 0.007, outperforming MSS–Soft, iSCAN, and SCORE, with consistent gains across all mechanism classes.

Table 2: Synthetic results for edge-direction accuracy, reported as mean ± 95% bootstrap CI half-width with runtime in seconds per graph. Best results are shown in bold and second-best results are underlined. In Panel A, × marks the official NoGAM implementation exceeding available memory in the full-data run.

Method	Overall	Linear	Tanh	Binary	Time (s)
Panel A: full-data comparison					
Multi-environment methods					
GLORI	$\mathbf{0.704 \pm 0.007}$	$\mathbf{0.790 \pm 0.009}$	$\mathbf{0.702 \pm 0.011}$	$\mathbf{0.619 \pm 0.011}$	$\mathbf{0.108 \pm 0.003}$
MSS–Soft	$\underline{0.545 \pm 0.006}$	$\underline{0.558 \pm 0.010}$	$\underline{0.553 \pm 0.010}$	$\underline{0.523 \pm 0.010}$	$\underline{1.583 \pm 0.056}$
iSCAN	0.448 ± 0.008	0.360 ± 0.011	0.547 ± 0.012	0.438 ± 0.011	9.181 ± 0.362
Observational causal-ordering methods					
SCORE	0.433 ± 0.007	0.377 ± 0.011	0.488 ± 0.011	0.434 ± 0.011	6.419 ± 0.224
NoGAM	×	×	×	×	×
Panel B: equal-sample comparison					
Multi-environment methods					
GLORI	$\mathbf{0.586 \pm 0.008}$	$\mathbf{0.697 \pm 0.011}$	$\underline{0.548 \pm 0.011}$	$\mathbf{0.513 \pm 0.012}$	$\mathbf{0.082 \pm 0.002}$
MSS–Soft	$\underline{0.502 \pm 0.006}$	$\underline{0.509 \pm 0.010}$	0.499 ± 0.010	$\underline{0.497 \pm 0.009}$	3.235 ± 0.160
iSCAN	0.480 ± 0.008	0.471 ± 0.013	0.537 ± 0.012	0.433 ± 0.012	0.197 ± 0.007
Observational causal-ordering methods					
SCORE	0.447 ± 0.007	0.429 ± 0.012	0.511 ± 0.011	0.401 ± 0.012	$\underline{0.107 \pm 0.004}$
NoGAM	0.480 ± 0.008	0.401 ± 0.012	$\mathbf{0.552 \pm 0.012}$	0.486 ± 0.012	4.263 ± 0.188

Table 3: Real-data results on the K562 and RPE1 CausalBench datasets. Higher W1 and Precision indicate better performance. Values are reported as mean ± 95% bootstrap CI half-width. Best results are shown in bold and second-best results are underlined among the learned methods; Random is included as a diagnostic baseline.

	K562		RPE1		
Method	W1 ↑	Precision ↑	W1 ↑	Precision ↑	Time (s)
GLORI	$\mathbf{0.111 \pm 0.022}$	$\mathbf{0.348 \pm 0.076}$	$\mathbf{0.221 \pm 0.040}$	$\mathbf{0.506 \pm 0.078}$	$\mathbf{0.548 \pm 0.012}$
MSS–Soft	$\underline{0.096 \pm 0.014}$	$\mathbf{0.348 \pm 0.073}$	0.115 ± 0.018	0.373 ± 0.073	103.476 ± 1.473
SCORE	0.094 ± 0.013	$\mathbf{0.348 \pm 0.076}$	$\underline{0.132 \pm 0.025}$	$\underline{0.404 \pm 0.073}$	$\underline{1.628 \pm 0.037}$
NoGAM	0.088 ± 0.012	$\underline{0.323 \pm 0.073}$	0.109 ± 0.018	0.349 ± 0.073	14.436 ± 1.294
iSCAN	0.087 ± 0.011	0.316 ± 0.073	0.109 ± 0.017	0.380 ± 0.073	2.153 ± 0.127
Random	0.094 ± 0.014	0.316 ± 0.073	0.178 ± 0.037	0.416 ± 0.072	< 0.001

The task-paired bootstrap difference against the closest multi-environment baseline, MSS–Soft, is $+0.159$ with 95% interval $[+0.151, +0.168]$. Panel B compares all methods under the same equal-sample budget of 10 observations in each of 10 environments. The gains are largest on the linear mechanisms, where GLORI reaches 0.790 ± 0.009 in the full-data setting and 0.697 ± 0.011 in the equal-sample setting. Under the equal-sample budget, GLORI also gives the best overall accuracy and the best average accuracy over continuous mechanisms, while MSS–Soft is comparable on the binary mechanism. GLORI is also substantially faster than the main competitors, requiring 0.108 seconds per graph with full data and 0.082 seconds per graph in the equal-sample setting.

4.2.2 Real-data results

For each predicted edge $u \to v$, CausalBench compares the distribution of v under perturbation of u with its control distribution. We report

$$\mathrm{W1} = \frac{1}{|\widehat{A}|} \sum_{(u,v)\in\widehat{A}} W_1\left(P_v^{(u)}, P_v^{(\mathrm{ctrl})}\right), \qquad \mathrm{Precision} = \frac{1}{|\widehat{A}|} \sum_{(u,v)\in\widehat{A}} \mathbf{1}\{p_{uv} < \alpha\}. \tag{20}$$

where $\widehat{A}$ denotes the set of predicted directed edges, $P_v^{(u)}$ and $P_v^{(\mathrm{ctrl})}$ denote the distributions of v under perturbation of u and under control, respectively, and p_{uv} is the Mann–Whitney U-test p-value comparing the two distributions. α is the significance level specified in Appendix E.

Table 3 reports the real-data results on CausalBench. GLORI achieves the highest W1 on both K562 and RPE1, ties for the highest precision on K562, and achieves the highest precision on RPE1. These results indicate that the residual-instability signal remains informative when applied to biological perturbation data, despite the substantial differences from the synthetic settings used in our controlled evaluation. The consistent W1 gains across both datasets further suggest that the proposed ordering

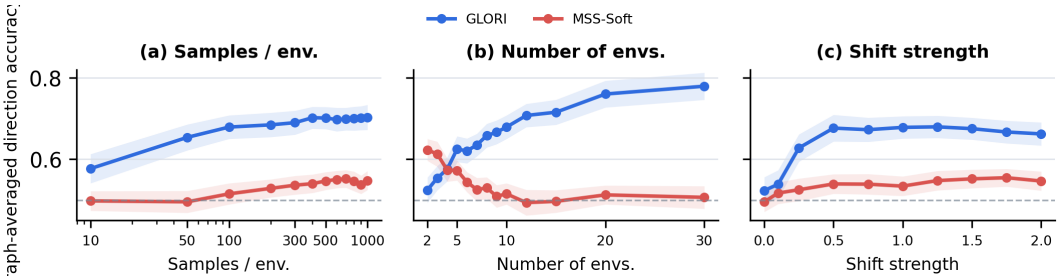

Figure 2: Operating-condition sensitivity on synthetic data. Each point averages results over the same set of underlying graphs across conditions while varying (a) samples per environment with $K = 10$, (b) the number of environments with $n_e = 100$, and (c) shift strength.

Table 4: Compact synthetic analyses. Panel A reports dimension and runtime; × marks NoGAM exceeding available memory. Panel B reports sensitivity to predictor choice and Ridge regularization at $K = 10, n_e = 100$. Entries are mean ± 95% CI half-width. Extra Trees denotes an Extremely Randomized Trees regressor. Best results are shown in bold and second-best results are underlined. lower runtime is better.

Method	Overall	$d = 10$	$d = 30$	$d = 50$	Time (s)
Panel A: graph dimension					
GLORI	**0.704 ± 0.007**	**0.716 ± 0.016**	**0.700 ± 0.010**	**0.696 ± 0.010**	**0.108 ± 0.003**
MSS–Soft	0.545 ± 0.006	0.541 ± 0.013	0.547 ± 0.008	0.545 ± 0.007	1.583 ± 0.056
SCORE	0.433 ± 0.007	0.469 ± 0.015	0.416 ± 0.010	0.414 ± 0.009	6.419 ± 0.224
NoGAM	×	×	×	×	×
iSCAN	0.448 ± 0.008	0.464 ± 0.017	0.445 ± 0.011	0.435 ± 0.011	9.181 ± 0.362
Panel B: predictor and regularization sensitivity					
Predictor	Overall	Linear	Tanh	Binary	Time (s)
Ridge ($\lambda = 0.01$)	**0.668 ± 0.030**	0.738 ± 0.056	**0.678 ± 0.038**	0.587 ± 0.043	**0.075 ± 0.008**
Ridge ($\lambda = 0.1$)	0.666 ± 0.030	0.738 ± 0.056	**0.678 ± 0.038**	0.583 ± 0.047	0.076 ± 0.008
Ridge ($\lambda = 1.0$)	0.667 ± 0.030	**0.739 ± 0.056**	**0.678 ± 0.038**	0.583 ± 0.047	**0.075 ± 0.008**
Extra Trees	0.521 ± 0.027	0.484 ± 0.038	0.495 ± 0.048	0.584 ± 0.045	1.482 ± 0.161
MLP	0.661 ± 0.029	0.723 ± 0.048	0.668 ± 0.038	**0.592 ± 0.052**	2.713 ± 0.321

criterion can recover useful directional structure under a fixed supplied skeleton. GLORI is also substantially faster than all external baselines in this real-data evaluation.

4.3 Ablation Studies

Operating-condition sensitivity. Figure 2 evaluates GLORI under three operating conditions using paired synthetic graphs. We vary the number of samples per environment, the number of environments, and the magnitude of distribution shifts across environments. GLORI consistently improves over MSS–Soft as within-environment sample size increases, becomes especially strong when more environments are available, and remains robust across a wide range of shift magnitudes.

Sources of the ordering signal. Figure 3 compares three node-level cross-environment scores: the raw marginal-mean score, the residual-scale score, and GLORI's residual-mean score. Using the notation from Section 3.3, all three use the same median pairwise aggregation across environments:

$$\underset{e<e'}{\text{median}} \begin{cases} \text{Raw marginal mean:} & \left|\widehat{\mu}_{X,v}^{(e)} - \widehat{\mu}_{X,v}^{(e')}\right|, \quad \widehat{\mu}_{X,v}^{(e)} = \frac{1}{n_e}\sum_{i=1}^{n_e} x_{i,v}^{(e)}, \\ \text{Residual scale:} & \left|\widehat{\sigma}_{R,v}^{(e)} - \widehat{\sigma}_{R,v}^{(e')}\right|, \quad \widehat{\sigma}_{R,v}^{(e)} = \sqrt{\frac{1}{n_e}\sum_{i=1}^{n_e}\left(r_{i,v}^{(e)} - \widehat{\mu}_v^{(e)}\right)^2}, \\ \text{Residual mean (GLORI):} & \left|\widehat{\mu}_v^{(e)} - \widehat{\mu}_v^{(e')}\right|. \end{cases}$$

Raw marginal-mean changes tend to propagate to downstream nodes, making them less informative about causal position, while residual-scale changes show only a weak relationship with graph depth. In contrast, GLORI's residual-mean score exhibits a clearer decrease from roots toward leaves. This comparison indicates that the useful ordering signal is not simply explained by marginal distribution

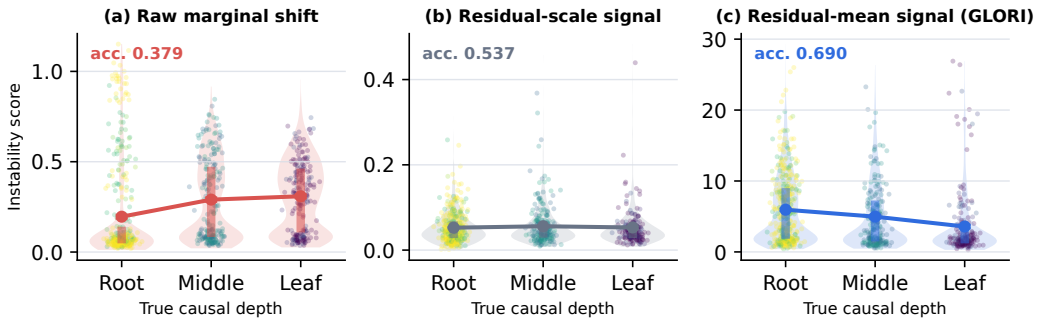

Figure 3: Diagnostic comparison of node-level shift signals on synthetic graphs. Each panel shows the distribution of instability scores across true causal-depth buckets. Raw marginal shifts increase toward downstream nodes, residual-scale changes are weakly aligned with depth, and GLORI's residual-mean signal gives a strong root-to-leaf ordering signal. This diagnostic uses five fixed graph seeds per base condition.

shifts or changes in residual magnitude. Instead, systematic changes in the signed residual mean provide a more informative contrast for distinguishing earlier from later nodes across environments.

5 Limitations

GLORI requires a supplied skeleton, and the population recovery guarantee is established for a one-root causal model. Limited distributional change across environments, few environments, small samples, and missing skeleton adjacencies can weaken the contrast. The residual-mean signal can also be affected by finite-sample estimation, particularly when environmental variation is weak. Its strength depends on how distribution changes interact with local causal mechanisms and the supplied graph structure.

6 Conclusion

We presented GLORI, which estimates a causal order from multi-environment data given an undirected skeleton by training one predictor per node using all environments together and ordering nodes by residual-mean instability. It directs all supplied edges without graph search or a decision threshold. The key idea is that a stable causal relationship can leave similar prediction errors across environments, while prediction in the reverse direction can become unstable when the input distribution changes. We established this contrast in a one-root causal model and found that the same residual-instability signal remains effective across more complex synthetic graphs and real biological perturbation data. These results suggest that simple changes in environment-wise prediction residuals can provide a practical signal for causal ordering when graph adjacency information is already available.

Reproducibility Statement

Appendix A provides the complete theoretical assumptions and proofs. Appendices B–E document the implementation and experimental details, including preprocessing, benchmark settings, comparator configurations, and hyperparameters. The supplementary material provides the code and scripts needed to reproduce the experiments.

AI Use Statement

Generative AI tools were used only for writing assistance and code debugging, with all scientific content, implementations, experiments, and reported results verified by the authors.

REFERENCES

Patrick Blöbaum, Dominik Janzing, Takashi Washio, Shohei Shimizu, and Bernhard Schölkopf. Cause-effect inference by comparing regression errors. In *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, 2018.

Philippe Brouillard, Sébastien Lachapelle, Alexandre Lacoste, Simon Lacoste-Julien, and Alexandre Drouin. Differentiable causal discovery from interventional data. In *Advances in Neural Information Processing Systems*, volume 33, pp. 21865–21877, 2020.

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*, 2023.

Mathieu Chevalley, Yusuf H. Roohani, Arash Mehrjou, Jure Leskovec, and Patrick Schwab. A large-scale benchmark for network inference from single-cell perturbation data. *Communications Biology*, 8:412, 2025.

Diego Colombo and Marloes H. Maathuis. Order-independent constraint-based causal structure learning. *Journal of Machine Learning Research*, 15:3921–3962, 2014.

Patrik O. Hoyer, Dominik Janzing, Joris M. Mooij, Jonas Peters, and Bernhard Schölkopf. Nonlinear causal discovery with additive noise models. In *Advances in Neural Information Processing Systems*, volume 21, pp. 689–696, 2008.

Biwei Huang, Kun Zhang, Jiji Zhang, Joseph Ramsey, Ruben Sanchez-Romero, Clark Glymour, and Bernhard Schölkopf. Causal discovery from heterogeneous/nonstationary data. *Journal of Machine Learning Research*, 21(89):1–53, 2020.

Phillip Lippe, Taco Cohen, and Efstratios Gavves. ENCO: Efficient neural causal discovery without acyclicity constraints. In *International Conference on Learning Representations*, 2022. URL `https://openreview.net/forum?id=eYciPrLuUhG`.

Francesco Montagna, Nicoletta Noceti, Lorenzo Rosasco, Kun Zhang, and Francesco Locatello. Causal discovery with score matching on additive models with arbitrary noise. In *Proceedings of the Second Conference on Causal Learning and Reasoning*, 2023.

Joris M. Mooij, Sara Magliacane, and Tom Claassen. Joint causal inference from multiple contexts. *Journal of Machine Learning Research*, 21(99):1–108, 2020.

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

Jonas Peters, Joris M. Mooij, Dominik Janzing, and Bernhard Schölkopf. Causal discovery with continuous additive noise models. *Journal of Machine Learning Research*, 15(58):2009–2053, 2014.

Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. Causal inference by using invariant prediction: Identification and confidence intervals. *Journal of the Royal Statistical Society: Series B*, 78(5):947–1012, 2016.

Paul Rolland, Volkan Cevher, Matthäus Kleindessner, Chris Russell, Dominik Janzing, Bernhard Schölkopf, and Francesco Locatello. Score matching enables causal discovery of nonlinear additive noise models. In *Proceedings of the 39th International Conference on Machine Learning*, 2022.

Pedro Sanchez, Xiao Liu, Alison Q. O'Neil, and Sotirios A. Tsaftaris. Diffusion models for causal discovery via topological ordering. In *International Conference on Learning Representations*, 2023.

Chandler Squires, Yuhao Wang, and Caroline Uhler. Permutation-based causal structure learning with unknown intervention targets. In *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence*, volume 124 of *Proceedings of Machine Learning Research*, pp. 1039–1048, 2020.

A Theoretical Analysis

This appendix formalizes the residual-instability contrast described in Section 3. The analysis considers the one-root linear model used in the main text; predicting each child from the root leaves an environment-invariant mean residual, whereas predicting the root from its children leaves a systematic component of the root shift. This separation is sufficient for GLORI to recover all root-to-child directions in the supplied skeleton.

A.1 Population residual instability

For the population analysis, let $f_v^\star$ denote the population counterpart of the predictor $\widehat{f}_v$. Matching the pooled empirical fit in Equation (8), we define

$$f_v^\star = \arg\min_{f \in \mathcal{F}} \sum_{e \in \mathcal{E}} \frac{n_e}{N} \mathbb{E}_{P_e} \left[\left(X_v - f\left(\mathbf{X}_{\mathcal{N}_G(v)} \right) \right)^2 \right], \tag{21}$$

where $N = \sum_{e \in \mathcal{E}} n_e$. The theoretical analysis uses the unregularized population least-squares objective; the practical estimator uses Ridge regression for finite-sample stability.

Then

$$R_v^{(e)} = X_v^{(e)} - f_v^\star \left(\mathbf{X}^{(e)}_{\mathcal{N}_G(v)} \right). \tag{22}$$

Its environment-wise population mean and the corresponding population instability score are

$$\mu_v^{(e)} = \mathbb{E}_{P_e} \left[R_v^{(e)} \right], \qquad I_v = \operatorname*{median}_{e < e'} \left| \mu_v^{(e)} - \mu_v^{(e')} \right|. \tag{23}$$

Thus, $\mu_v^{(e)}$ and I_v are the population counterparts of $\widehat{\mu}_v^{(e)}$ and $\widehat{I}_v$ in Section 3.3.

Pooled standardization only rescales the population residual means by positive node-specific constants and therefore preserves the conclusion of Theorem 1, namely that the root is ordered before all of its children.

A.2 One-root linear-Gaussian model

Consider a supplied skeleton with one root X_0 connected to children $X_1, \ldots, X_m$ and no child-to-child edges. The underlying causal directions are $X_0 \to X_j$ for $j = 1, \ldots, m$. In environment e, let

$$X_0^{(e)} = \mu_e + \varepsilon_0^{(e)}, \tag{24}$$

$$X_j^{(e)} = a_j X_0^{(e)} + \varepsilon_j^{(e)}, \qquad j = 1, \ldots, m, \tag{25}$$

where $a_j \neq 0$. The noise variables are mutually independent, zero-mean, environment-invariant Gaussians with finite positive variances. The coefficients a_j are also shared across environments, so the environments differ only through the root mean μ_e.

The causal relationship from X_0 to each child is therefore unchanged across environments, while the marginal distribution of X_0 changes with μ_e. The supplied neighborhoods imply that GLORI predicts each child X_j from X_0 and predicts X_0 jointly from all of its children. Define the magnitude of the root-mean variation by

$$\Delta_\mu = \operatorname*{median}_{e < e'} \left| \mu_e - \mu_{e'} \right|. \tag{26}$$

The recovery result assumes $\Delta_\mu > 0$, so root-mean differences must occur across a sufficient fraction of environment pairs.

A.3 Residual stability for the children

Proposition 1 (Residual stability for the children) *Under Equations* (24)–(25)*, the population predictor of each child satisfies*

$$f_j^\star(X_0) = a_j X_0, \qquad \mu_j^{(e)} = 0 \text{ for every } e \in \mathcal{E}, \qquad I_j = 0. \tag{27}$$

Proof. Because $\varepsilon_j^{(e)}$ is independent of $X_0^{(e)}$, has mean zero, and has the same distribution in every environment, the relationship $f_j^\star(X_0) = a_j X_0$ minimizes the pooled population squared error. Hence

$$R_j^{(e)} = X_j^{(e)} - f_j^\star(X_0^{(e)}) = \varepsilon_j^{(e)},$$

so $\mu_j^{(e)} = \mathbb{E}[R_j^{(e)}] = 0$ in every environment. Every pairwise difference in Equation (23) is therefore zero, giving $I_j = 0$.

Interpretation. A shift in the root can change the marginal distribution of a child, but the predictor removes the component transmitted through the stable root-to-child relationship. What remains is environment-invariant noise. This is why a child can exhibit a large marginal shift while retaining a stable environment-wise mean residual.

A.4 Residual instability for the root

The reverse prediction problem is different because the children contain noisy information about the root and therefore do not determine it exactly.

Proposition 2 (Residual instability for the root) *Under Equations* (24)–(25)*, there exists a constant* $q > 0$ *such that, for every pair of environments* $e < e'$*,*

$$\left|\mu_0^{(e)} - \mu_0^{(e')}\right| = q\,|\mu_e - \mu_{e'}|\,. \tag{28}$$

Consequently,

$$I_0 = q\Delta_\mu > 0. \tag{29}$$

Proof. Because $f_0^\star$ is affine, write

$$f_0^\star(X_1, \ldots, X_m) = \beta_0 + \sum_{j=1}^{m} \beta_j X_j.$$

Using $X_j^{(e)} = a_j X_0^{(e)} + \varepsilon_j^{(e)}$, the root residual is

$$R_0^{(e)} = qX_0^{(e)} - \beta_0 - \sum_{j=1}^{m} \beta_j \varepsilon_j^{(e)}, \qquad q = 1 - \sum_{j=1}^{m} \beta_j a_j.$$

Under the weighted pooled population objective in Equation (21), the least-squares residual is orthogonal to its fitted value. Since $X_0 = f_0^\star + R_0$,

$$\mathrm{Cov}(R_0, X_0) = \mathrm{Var}(R_0) > 0.$$

The child noises are independent of X_0, so the residual expression also gives

$$\mathrm{Cov}(R_0, X_0) = q\ \mathrm{Var}(X_0).$$

Therefore

$$q = \frac{\mathrm{Var}(R_0)}{\mathrm{Var}(X_0)} > 0.$$

The strict inequality follows because independent child noise prevents the children from determining X_0 exactly.

Taking expectations within environment e gives

$$\mu_0^{(e)} = q\mu_e - \beta_0.$$

Because the pooled least-squares predictor contains an intercept, its weighted population residual has zero mean. With $\overline{\mu} = \sum_{e\in\mathcal{E}}(n_e/N)\mu_e$, this gives $\beta_0 = q\overline{\mu}$ and therefore

$$\mu_0^{(e)} = q\,(\mu_e - \overline{\mu})\,.$$

Subtracting two environments yields Equation (28); taking the median over environment pairs gives $I_0 = q\Delta_\mu > 0$.

Interpretation. Each child is a noisy consequence of the root, so reverse prediction cannot completely remove an environmental shift in the root. The positive fraction that remains appears as a systematic change in the environment-wise mean residual.

Proof of Theorem 1. Proposition 1 gives $I_j = 0$ for every child, whereas Proposition 2 gives $I_0 = q\Delta_\mu > 0$. Hence $I_0 > I_j$ for every j. GLORI orders nodes from larger to smaller residual instability, so X_0 appears before every child. Because the supplied skeleton contains only root-to-child adjacencies, the edge assignment rule in Equation (15) directs every edge correctly.

A.5 Scope of the theoretical result

Theorem 1 establishes the residual-instability contrast motivating GLORI in a tractable population setting. It does not imply that residual instability is strictly monotone with causal depth in every directed acyclic graph.

Informative environmental variation. The theorem requires $\Delta_\mu > 0$. If the root mean does not vary across environments, all population scores in the one-root model are zero. Because GLORI uses the median across environment pairs, informative shifts must occur across a sufficient fraction of pairs to produce a positive score.

Invariant child mechanisms. The theorem assumes that the child mechanisms remain unchanged across environments. If a child mechanism changes, its environment-wise mean residual may also vary and its residual-instability score need not remain zero. The synthetic benchmark deliberately includes broader targeted shifts that can change intercepts, incoming coefficients, and noise scales, so those experiments test the signal outside the theorem's invariance assumptions.

Finite samples and broader graphs. The theorem is a population result based on exact least-squares prediction. With finite samples and Ridge regularization, estimation error can make stable nodes receive nonzero scores. In a general directed acyclic graph, a supplied neighborhood may also contain both parents and children, mixing forward and reverse prediction relationships. Broader graph structures are therefore evaluated empirically rather than covered by the recovery guarantee.

Role of Gaussianity. The theorem is stated for a linear-Gaussian model, but the core separation uses linear structural equations, independent zero-mean noise, and finite positive noise variances. The same population argument therefore extends to non-Gaussian noise satisfying these conditions.

Overall, the analysis provides population-level grounding for the residual-instability ordering signal in a concrete setting while keeping the scope of the recovery claim explicit.

B Algorithm and Implementation Details

This appendix records the complete GLORI procedure and the implementation details used in the experiments. We retain the notation from the main method and avoid introducing a second implementation-specific notation system.

B.1 Complete GLORI algorithm

Algorithm 1 summarizes the complete procedure.

Algorithm 1 GLORI

Require: Multi-environment datasets $\{\mathcal{D}^{(e)}\}_{e=1}^{K}$; supplied undirected skeleton $G = (V, E)$; Ridge regularization strength λ; tie-breaking seed.

Ensure: Estimated causal order $\widehat{\pi}$ and directed graph $\widehat{G} = (V, \widehat{A})$.

Pool the observations from all environments.
Standardize each variable using the pooled mean and standard deviation.
Generate a seed-fixed tie priority over V.
for each node $v \in V$ **do**
    Obtain the supplied neighborhood $\mathcal{N}_G(v)$.
    if $\mathcal{N}_G(v) = \varnothing$ **then**
        Compute $\widehat{I}_v$ from the environment-wise marginal means of X_v using the same median pairwise aggregation
        continue
    end if
    Fit the Ridge predictor $\widehat{f}_v$ using data from all environments.
    for each environment $e \in \mathcal{E}$ **do**
        Evaluate the same predictor $\widehat{f}_v$ in environment e.
        Compute the mean residual $\widehat{\mu}_v^{(e)}$.
    end for
    Compute $\widehat{I}_v$ using Equation (12).
end for
Sort nodes by decreasing $\widehat{I}_v$ to obtain $\widehat{\pi} = (\pi_1, \ldots, \pi_d)$.
Resolve exact score ties using the seed-fixed tie priority.
Direct each supplied edge from the earlier to the later node in $\widehat{\pi}$ using Equation (15).
return $\widehat{\pi}$ and $\widehat{G} = (V, \widehat{A})$.

B.2 PREPROCESSING AND NEIGHBORHOOD REGRESSION

Pooled standardization. Each variable is standardized once using its empirical mean and standard deviation over all $N = \sum_{e \in \mathcal{E}} n_e$ pooled observations. The same pooled transformation is then applied in every environment. If a variable has zero pooled variance, its scale is set to one. After this preprocessing, the sample values $x_{i,v}^{(e)}$ and neighborhood vectors $\mathbf{x}_{i,\mathcal{N}_G(v)}^{(e)}$ in the implementation refer to these standardized values.

Ridge predictor. For every non-isolated node, Equation (8) is instantiated as affine Ridge regression: $\mathcal{F}$ is the class of affine linear predictors, $\Omega(f)$ is the squared ℓ_2 norm of the non-intercept coefficients, and the intercept is not regularized. The main experiments use $\lambda = 0.1$ unless otherwise stated. Each observation receives equal weight, so environments with more samples contribute proportionally more to the pooled fit.

Residual evaluation. For the main Ridge predictor, $\widehat{f}_v$ is evaluated on the same observations used for fitting; no cross-fitting or validation split is used. Sample residuals, environment-wise residual means, and residual-instability scores are computed exactly as in Equations (10), (11), and (12).

B.3 AGGREGATION AND TIE HANDLING

Residuals are averaged with their signs before environments are compared, so $\widehat{I}_v$ measures changes in systematic prediction error rather than residual magnitude. Nodes are ordered by decreasing $\widehat{I}_v$ as in Equation (13). Exact score ties are resolved using a seed-fixed priority independent of the observed data. The ordering of isolated nodes does not affect any edge direction. Their implementation-level scores are used only to place them in the total order.

B.4 COMPUTATIONAL COMPLEXITY

Using the notation from the main text, let N be the total number of observations and k the maximum degree of the supplied skeleton. Fitting a Ridge predictor with at most k inputs by the normal equations costs $O(Nk^2 + k^3)$ per node, and evaluating it costs $O(Nk)$. The complete procedure is therefore bounded by

$$O\left(d\left(Nk^2 + k^3 + Nk\right) + Nd + K^2 d + d \log d + |E|\right). \tag{30}$$

C EXPERIMENTAL PROTOCOLS

This appendix records the fixed synthetic benchmark, evaluation, baseline, and compute settings used in Section 4. In each non-reference environment, a fraction of nodes is randomly selected for environment-specific changes to its local generating mechanism; the shift-strength parameter controls the magnitude of these changes. The main settings are summarized in Table 5.

Comparator implementations. Table 6 summarizes the implementation source and fixed input protocol used for each comparator. All estimated orders are evaluated on the same supplied skeleton.

D ADDITIONAL SYNTHETIC RESULTS

The main paper reports the primary synthetic comparisons and operating-condition analyses. Here we provide additional checks of the environmental signal, the GLORI score design, the location of environmental shifts, and errors in the supplied skeleton.

Environmental-signal controls. Table 7 tests whether GLORI's performance depends on environment-specific distributional information. Permuting environment labels or removing distribution shifts substantially reduces accuracy, while degree-only ordering remains near chance.

Residual-summary and aggregation ablation. Table 8 isolates the two choices used in the GLORI score. Signed residual means outperform raw marginal means and residual standard deviations, while median aggregation performs better than mean aggregation.

Shifted endpoint diagnostic. Table 9 groups true edges according to whether the source, target, both, or neither endpoint is directly shifted. GLORI performs best when the source is shifted, whereas shifting only the target can weaken or invert the residual-instability signal. This analysis highlights when the proposed ordering signal is most informative and identifies a concrete failure case.

Table 5: Main synthetic benchmark and implementation settings.

Component	Parameter	Setting
Graph	Model	Erdős–Rényi; Barabási–Albert
	Dimension	$d \in \{10, 30, 50\}$
	Average degree	$\bar{d} \in \{2, 4\}$
	Graph seeds	20 per condition
Mechanism	Type	Linear; tanh; binary
	Edge weight	$\lvert w \rvert \in [0.6, 1.4]$
	Noise scale	$[0.55, 0.85]$
Environment	Environments	$K = 10$
	Samples / environment	$n_e = 500$ (full); $n_e = 10$ (equal-sample)
	Shifted-node fraction	$\{0.1, 0.2\}$
	Shift strength	1.0
GLORI	Predictor	Ridge
	Regularization	$\lambda = 0.1$
Evaluation	Bootstrap resamples	10,000
Compute	GPU	NVIDIA RTX PRO 6000
	CPU	AMD Ryzen Threadripper 3970X
	Memory	125 GiB
	Python	3.12.4
	PyTorch	2.5.1+cu121
	CUDA	12.1

Table 6: Comparator implementations and fixed evaluation settings.

Method	Implementation	Input	Run detail
MSS–Soft	Official MSS code	Skeleton; environments	Soft objective
SCORE	Official SCORE code	Pooled data	CUDA
NoGAM	dodiscover	Pooled data	Full-data OOM
iSCAN	Official iSCAN code	Even/odd environments	CUDA patch

Table 7: Controls for the environmental ordering signal on the full-data synthetic benchmark. Entries are edge-direction accuracy, reported as mean ± 95% bootstrap CI half-width.

Variant	Overall	Linear	Tanh	Binary
Env. labels permuted	0.542 ± 0.008	0.642 ± 0.013	0.486 ± 0.012	0.498 ± 0.012
Distribution shifts removed	0.544 ± 0.008	0.656 ± 0.012	0.483 ± 0.012	0.493 ± 0.011
Degree-only	0.501 ± 0.007	0.506 ± 0.012	0.496 ± 0.012	0.499 ± 0.012

Table 8: Ablation of the residual summary and cross-environment aggregation on the full-data synthetic benchmark ($K = 10$, $n_e = 500$). Entries are mean ± 95% bootstrap CI half-width.

Residual summary	Aggregation	Overall	Linear	Tanh	Binary
Signed residual mean	**Median**	$\mathbf{0.704 \pm 0.007}$	$\mathbf{0.790 \pm 0.009}$	$\mathbf{0.702 \pm 0.011}$	$\mathbf{0.619 \pm 0.011}$
Raw marginal mean	Median	0.364 ± 0.008	0.346 ± 0.015	0.337 ± 0.013	0.411 ± 0.011
Residual standard deviation	Median	0.519 ± 0.007	0.550 ± 0.011	0.578 ± 0.011	0.430 ± 0.012
Signed residual mean	Mean	0.617 ± 0.007	0.704 ± 0.010	0.587 ± 0.011	0.559 ± 0.010

Robustness to skeleton errors. The main experiments use the ground-truth undirected skeleton to isolate edge orientation. Table 10 evaluates GLORI after randomly deleting true adjacencies or adding spurious adjacencies to the supplied skeleton. Adding extra edges has little effect on performance, whereas removing true edges leads to a clear drop in accuracy because informative neighborhood relationships are lost.

Table 9: Shift-target endpoint diagnostic on the full-data synthetic benchmark. Shift targets are used only for this analysis and are not provided to the methods. Entries are edge-direction accuracy with 95% bootstrap CI half-width; the last column reports the task-paired GLORI–MSS–Soft difference.

Shifted endpoint	Edges	GLORI	MSS–Soft	GLORI–MSS diff.
True source only	11,206	0.896 ± 0.009	0.631 ± 0.014	$+0.266$ [$+0.249$, $+0.283$]
True target only	11,402	0.462 ± 0.016	0.631 ± 0.014	-0.169 [-0.189, -0.149]
Both endpoints	35,127	0.714 ± 0.010	0.490 ± 0.008	$+0.224$ [$+0.212$, $+0.236$]
Neither endpoint	5,315	0.718 ± 0.019	0.569 ± 0.020	$+0.149$ [$+0.121$, $+0.176$]

Table 10: Robustness to errors in the supplied skeleton. Entries are edge-direction accuracy, reported as mean $\pm$ 95% bootstrap CI half-width. Best results are shown in bold and second-best results are underlined.

Skeleton	Overall	Linear	Tanh	Binary
Ground truth	$\mathbf{0.699 \pm 0.030}$	$\underline{0.809 \pm 0.041}$	0.677 ± 0.046	$\mathbf{0.612 \pm 0.044}$
10% edge deletion	0.605 ± 0.028	0.678 ± 0.045	0.601 ± 0.046	0.536 ± 0.043
20% edge deletion	0.515 ± 0.024	0.586 ± 0.032	0.522 ± 0.038	0.438 ± 0.039
10% edge addition	$\underline{0.698 \pm 0.030}$	0.806 ± 0.044	$\underline{0.678 \pm 0.046}$	$\underline{0.611 \pm 0.044}$
20% edge addition	$\mathbf{0.699 \pm 0.030}$	$\mathbf{0.810 \pm 0.042}$	$\mathbf{0.679 \pm 0.047}$	0.609 ± 0.042

E REAL-DATA PROTOCOL

The main text reports the headline CausalBench results. This appendix records the corresponding preprocessing and evaluation protocol.

Table 11: CausalBench preprocessing and evaluation protocol.

Setting	**K562**	**RPE1**
Variables	60	60
Observations	11,749	11,146
Environments	61	61
Gene selection	Top control-variance perturbable genes	
Minimum cells / perturbation	100	100
Skeleton construction	Top-3 absolute control-correlation union	
Skeleton edges	158	166
Normalization	10^4 + `log1p`	10^4 + `log1p`
Evaluation unit	Predicted directed edge	
Statistical test	Mann–Whitney U test	
Significance level	$\alpha = 0.05$	
Bootstrap resamples	10,000 edge-level	

CausalBench protocol. Table 11 summarizes the preprocessing and evaluation settings used for the K562 and RPE1 datasets. The same processed data and supplied skeleton are used for all compared methods. Headline W1, precision, and runtime results are reported in Table 3.

F SANITY CHECKS

We perform sanity checks to verify that the implementation behaves as expected. On hand-constructed graphs with known causal directions, GLORI recovers the expected edge orientations. We also verify that permuting the node indices and mapping them back to the original ordering leaves the results unchanged.