CSALT: TRAINING-FREE CAUSAL SKELETON LEARNING WITH FROZEN TABULAR FOUNDATION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Learning causal skeletons from observational data is a fundamental yet challenging task in causal discovery. Constraint-based methods may lose power when data are scarce; continuous-optimization methods require per-dataset refitting; and pretrained causal discovery models require large-scale pretraining on simulations labeled with true graphs. Tabular foundation models offer another source of structural information for causal skeletons, but their attention gives only relative rankings and no statistical rule for determining edge existence. To bridge this gap, we introduce CSALT (Causal Skeleton Learning via Attention-Led Testing), which turns attention from a frozen tabular foundation model into the weights of a target-wise weighted Benjamini–Hochberg procedure. Permuting the target column on a disjoint split gives each candidate a permutation score, and an OR rule combines the target-wise rejection sets into an undirected skeleton. CSALT keeps all model parameters frozen and does not require additional pretraining with ground-truth graph supervision. We theoretically derive a sufficient condition involving attention ranking margins for excluding false edges with finitely many permutations, and further conditions for exact skeleton recovery. Empirical results on nonlinear synthetic data show that CSALT attains higher skeleton precision and F1 than all evaluated baselines. On two real datasets, our method achieves the highest F1 on UF and ties for the highest F1 on Sachs.

1 INTRODUCTION

Causal discovery aims to infer causal relationships among variables and thereby support scientific understanding and decision-making (Spirtes et al., 2000). When randomized experiments are constrained by cost, ethics, or feasibility, observational data are the primary source of evidence. A basic step towards the causal graph is to recover its skeleton: the undirected graph of direct causal adjacencies. An accurate skeleton is a prerequisite for causal graph construction, causal effect estimation, and decision analysis.

Despite steady progress, learning the skeleton from observational data remains difficult when samples are few, noise is strong, and mechanisms are complicated. Methods fitted to each dataset draw all evidence from the current sample, carry nothing across datasets, and rely on assumptions that may not match the data. Constraint-based methods such as PC and FCI may lose power on scarce or noisy data (Spirtes et al., 2000). Score-based methods rely on a chosen scoring criterion, and the search may stop at a local optimum (Chickering, 2002). Functional causal models require assumptions about functional form and noise (Shimizu et al., 2006). Continuous-optimization methods learn causal structure and mechanisms anew for each dataset (Zheng et al., 2020). Pretrained causal discovery models such as AVICI and CDFM amortize inference across datasets, but require large-scale graph-labeled synthetic pretraining; the resulting priors are shaped by the simulated training distribution, making transfer to new graph and mechanism regimes an empirical question (Lorch et al., 2022; Qiao et al., 2026). Structural knowledge thus comes either from a single sample or from graph supervision, leaving a gap between purely data-driven evidence and graph-supervised priors.

General-purpose tabular foundation models offer another source of structural cues. LimiX, for example, is pretrained on tabular tasks generated from structural causal models, without causal graphs

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: Comparison of representative causal discovery paradigms.

Method family	Per-dataset optimization	Graph-label pretraining	Structural information
Constraint-based (PC, FCI)	No	No	Independence tests on data
Functional (LiNGAM, ANM)	Yes	No	Mechanism assumptions
Score-based (GES)	Yes	No	Score-guided search
Continuous opt. (NOTEARS, GraN-DAG)	Yes	No	Graph + mechanism fitting
Pretrained (AVICI, TabCausal, CDFM)	No	Yes	Pretrained graph inference
CSALT	**No**	**No**	**Attention + permutation scores**

as training labels (Zhang et al., 2025). At inference time, attention indicates how strongly the model draws on each input when predicting a target. Under faithfulness, a neighbor of the target can provide predictive information that the other variables cannot replace, so attention is a plausible cue for adjacency. However, the attention scores are relative, their scale varies with graph size and sparsity, and indirect paths and common children can raise the scores of nonneighbors, including spouses.

We therefore introduce CSALT (Causal Skeleton Learning via Attention-Led Testing), which turns attention rankings into target-wise weighted multiple tests as shown in Figure 1. For each target, attention on one data split assigns the candidate weights that average to one in the weighted Benjamini–Hochberg (BH) procedure, and permuting the target column on a disjoint split yields a permutation score for each candidate. Because the splits are disjoint, one sampling fluctuation cannot inflate both the weight and the permutation score of the same candidate. Candidates whose weights fall below a threshold set by the testing level, the permutation count, and the number of rejections cannot be selected, however small their permutation scores; when true neighbors lead nonneighbors in attention by a large enough ranking margin, every nonneighbor weight falls below this threshold (Theorem 1). Among highly weighted candidates, the permutation score decides selection. An OR rule combines the target-wise rejection sets into an undirected skeleton. CSALT is training-free: it runs only forward passes of a standard pretrained model, without parameter updates, per-dataset graph optimization, or added graph-labeled pretraining. In summary, we make three contributions:

- We introduce CSALT, a training-free skeleton learning method that turns frozen-model attention into weighted-BH rejection thresholds for permutation scores from a disjoint split.
- We prove that a target selects no false edge when its ranking margin exceeds a threshold that grows logarithmically with the number of permutations, without requiring the permutation scores of nonneighbors such as spouses to be valid p-values. With a true-edge power condition, the target neighborhood is recovered exactly.
- We perform extensive evaluations on both synthetic and real-world data. CSALT achieves a higher skeleton F1 than all baselines evaluated on synthetic data and the highest or tied-highest F1 on real data. In addition, the ranking-margin condition holds for most synthetic targets, and ablations attribute most of the F1 gain to the attention weights.

2 Related Work

Causal discovery methods built on structural causal models (SCMs) (Pearl, 2009) fall into several families, as shown in Table 1 and a full discussion is provided in Appendix D. Constraint-based methods infer adjacencies through conditional independence tests (Spirtes et al., 2000; Zhang et al., 2011), score-based methods let a score function guide the search (Chickering, 2002), and functional causal models achieve identifiability through assumptions on the generating mechanisms (Shimizu et al., 2006; Hoyer et al., 2008). More recently, NOTEARS expressed acyclicity as a smooth constraint (Zheng et al., 2018) and inspired nonlinear extensions (Zheng et al., 2020; Lachapelle et al., 2020). Despite their empirical success, these methods fit each dataset from its own sample and carry distinct inductive biases; continuous optimization can exploit variance increasing along the causal order in simulated data (Reisach et al., 2021). In recent years, pretrained causal discovery models such as AVICI, TabCausal, and CDFM train graph predictors on simulated datasets paired with

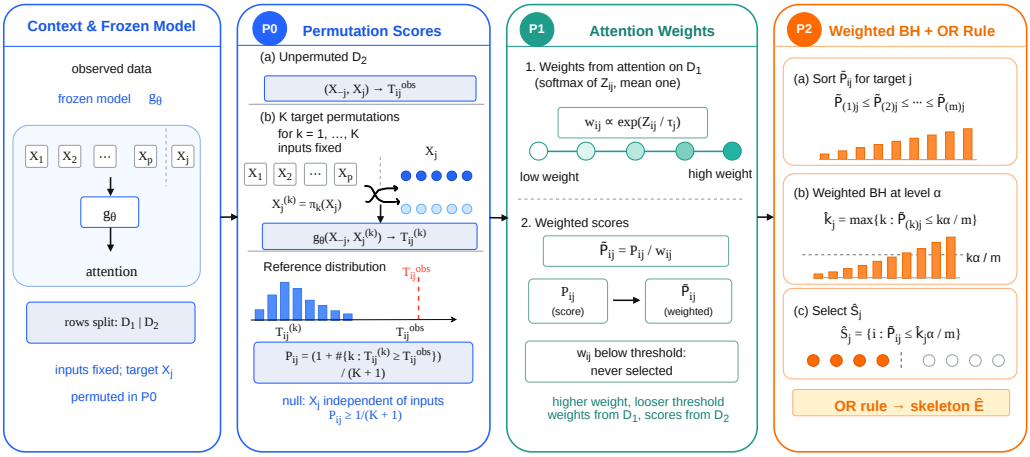


Figure 1: Overview of CSALT for target X_j and frozen model g_θ. P0: the target column of $\mathcal{D}_2$ is permuted K times with inputs fixed; the rank of T_{ij}^{obs} among $T_{ij}^{(1)}, \ldots, T_{ij}^{(K)}$ gives the permutation score P_{ij} for the null that the target is independent of all inputs. P1: attention on $\mathcal{D}_1$ gives symmetric scores Z_{ij} (Eq. (2)) and weights w_{ij} (Eq. (3)); the weighted scores are $\widetilde{P}_{ij} = P_{ij}/w_{ij}$. P2: weighted BH over $m = p - 1$ candidates gives $\widehat{S}_j$; the OR rule assembles the skeleton $\widehat{E}$.

ground-truth graphs, infer structure in a single forward pass (Lorch et al., 2022; Li et al., 2026; Qiao et al., 2026). Their learned structural prior consequently depends on the simulation distribution.

The methods discussed above obtain structural information either from the current dataset or from graph-supervised synthetic pretraining. Tabular foundation models offer a third source. For example, TabPFN and LimiX are pretrained on SCM-generated prediction tasks without graph labels (Hollmann et al., 2025; Zhang et al., 2025), and Swelam et al. (2025) found that a learned graph-supervised decoder can recover adjacencies from frozen TabPFN embeddings. By contrast, CSALT learns skeleton from a frozen tabular model without training a graph decoder, thereby reusing the knowledge from a broad pretraining that makes these models accurate on small datasets (Hollmann et al., 2025). Edges are still decided by tests on the current data, and the cost stays low: CSALT fits no model to a new dataset and needs no graph-labeled pretraining.

3 Problem Setup

A structural causal model (SCM) (Pearl, 2009; Peters et al., 2017) consists of a directed acyclic graph $G^\star = (\mathcal{V}, E)$ and structural equations. Node $j \in \mathcal{V} = \{1, \ldots, p\}$ represents the observed variable X_j, and an edge $(i, j) \in E$ represents a direct causal effect of X_i on X_j:

$$X_j := f_j(\mathbf{PA}_j, U_j), \quad j = 1, \ldots, p, \tag{1}$$

where $\mathbf{PA}_j$ denotes the parents of X_j in $G^\star$, f_j is its causal mechanism, and the noise terms $\{U_j\}_{j=1,\ldots,p}$ are mutually independent. Hence the joint distribution P_X satisfies the *Markov condition* with respect to $G^\star$, that is, $p(x_1, \ldots, x_p) = \prod_{j=1}^{p} p(x_j \mid \mathbf{pa}_j)$ (Spirtes et al., 2000). The skeleton $\mathrm{skel}(G^\star)$ is the undirected graph on $\mathcal{V}$ in which i and j are adjacent if and only if $G^\star$ contains $i \to j$ or $j \to i$; $S^\star \in \{0, 1\}^{p \times p}$ denotes its adjacency matrix. We aim to learn $S^\star$ from n i.i.d. observations $X \in \mathbb{R}^{n \times p}$. Under the Markov condition, causal sufficiency, and faithfulness, the skeleton is identifiable from the observational distribution (Spirtes et al., 2000).

4 CSALT

CSALT weights permutation tests of candidate edges by attention from a frozen tabular foundation model and assembles the selected edges into an undirected skeleton. An illustration is shown in Figure 1, and the full procedure is given in Algorithm 1. We randomly split the observations into

disjoint sets $\mathcal{D}_1$ and $\mathcal{D}_2$, using $\mathcal{D}_1$ only for the attention weights and $\mathcal{D}_2$ only for the permutation scores. Because the splits are disjoint, the weights and permutation scores are independent, so a chance fluctuation in one split cannot inflate both the weight and the permutation score of the same candidate edge.

4.1 Attention Weights

For each target X_j, CSALT predicts X_j from the other $p-1$ variables with the frozen tabular foundation model g_θ on $\mathcal{D}_1$. The score $a_{j\leftarrow i}(\mathcal{D}_1)$ is the between-feature attention from X_j to input X_i, averaged over the samples and attention heads of a fixed set $\mathcal{L}$ of intermediate and later layers (Appendix B). Because the skeleton is undirected, we average the two prediction directions for each pair $\{i,j\}$ to obtain the symmetric score

$$Z_{ij} = \frac{1}{2}\left\{a_{j\leftarrow i}(\mathcal{D}_1) + a_{i\leftarrow j}(\mathcal{D}_1)\right\}, \qquad Z_{ij} = Z_{ji}. \tag{2}$$

For each target j with $m = p-1$ candidates, the mean-one softmax weights are

$$w_{ij} = \frac{\exp(Z_{ij}/\tau_j)}{m^{-1}\sum_{k\neq j}\exp(Z_{kj}/\tau_j)}, \qquad \frac{1}{m}\sum_{i\neq j} w_{ij} = 1, \tag{3}$$

where $\tau_j = c_\tau \max\{\text{median}_{k\neq j} Z_{kj},\, \varepsilon_\tau\}$ is a target-adaptive temperature, $c_\tau > 0$ is a temperature multiplier, and $\varepsilon_\tau \in (0,1)$ is a numerical floor (10^{-10} in our implementation).

4.2 Permutation Scores

For target X_j, let $T_{ij}^{\text{obs}} = a_{j\leftarrow i}(\mathcal{D}_2)$ be the attention assigned to X_i on the unpermuted $\mathcal{D}_2$. Keeping all input columns fixed, we permute only the target column K times independently and obtain $T_{ij}^{(1)}, \ldots, T_{ij}^{(K)}$. Permuting the target breaks its association with all inputs, so these values form a reference distribution for T_{ij}^{obs} under the null that X_j is independent of all inputs. We use the plus-one-corrected right-tail permutation score (Phipson & Smyth, 2010):

$$P_{ij} = \frac{1 + \sum_{r=1}^{K} \mathbb{I}\{T_{ij}^{(r)} \geq T_{ij}^{\text{obs}}\}}{K+1}, \qquad P_{ij} \geq \frac{1}{K+1}. \tag{4}$$

This global null differs from the skeleton null of nonadjacency: a spouse of X_j is nonadjacent but can be predictive given their common child, so P_{ij} need not be super-uniform (i.e., satisfy $\Pr(P_{ij} \leq t) \leq t$ for all $t \in [0,1]$) for nonneighbors. Hence the false discovery rate (FDR) guarantee does not directly apply to skeleton edges; Section 5 instead uses the ranking margin and the floor of P_{ij} to give conditions for excluding false edges.

4.3 Weighted Edge Selection and Skeleton Assembly

For target j, order the weighted scores $\widetilde{P}_{ij} = P_{ij}/w_{ij}$ as $\widetilde{P}_{(1)j} \leq \cdots \leq \widetilde{P}_{(m)j}$. At level α, CSALT applies weighted BH (Benjamini & Hochberg, 1995; Genovese et al., 2006):

$$\widehat{k}_j = \max\left(\left\{k : \widetilde{P}_{(k)j} \leq \frac{k\alpha}{m}\right\} \cup \{0\}\right), \qquad \widehat{S}_j = \left\{i : \widetilde{P}_{ij} \leq \frac{\widehat{k}_j \alpha}{m}\right\}. \tag{5}$$

The ratio P_{ij}/w_{ij} adjusts the rejection threshold: a larger w_{ij} lowers the weighted score and makes selection more likely. As in neighborhood selection (Meinshausen & Bühlmann, 2006), the OR rule gives the final undirected skeleton:

$$\{i,j\} \in \widehat{E} \quad \Longleftrightarrow \quad i \in \widehat{S}_j \text{ or } j \in \widehat{S}_i. \tag{6}$$

4.4 Computational Cost

Extracting weights on $\mathcal{D}_1$ takes p forward passes; each pass gives the target's attention to all inputs, so these p passes cover both prediction directions. Attention on the unpermuted and K permuted versions of $\mathcal{D}_2$ takes another $(K+1)p$ passes, for $(K+2)p$ single passes in total. With batches of B permutations, this becomes $p[2 + \lceil K/B \rceil]$ batched passes; the main experiments use $K = 100$ and $B = 8$, giving $15p$ batched passes (i.e., $102p$ single passes). Appendix C reports runtimes.

Algorithm 1 Skeleton learning with CSALT

Require: Data X, frozen model g_θ, K, α
Ensure: Skeleton $\widehat{E}$
1: Randomly split X into $\mathcal{D}_1$ and $\mathcal{D}_2$; compute all Z_{ij} from attention on $\mathcal{D}_1$ (Eq. (2))
2: **for** $j = 1, \dots, p$ **do**
3: Compute weights w_{ij} (Eq. (3))
4: Compute T_{ij}^{obs} on $\mathcal{D}_2$ and P_{ij} from K target-column permutations (Eq. (4))
5: Apply weighted BH to P_{ij}/w_{ij} to obtain $\widehat{S}_j$ (Eq. (5))
6: Return $\widehat{E}$ by the OR rule (Eq. (6))

5 Sufficient Conditions for Skeleton Recovery

For target j, let $N_j = \{i : (i, j) \in \text{skel}(G^\star)\}$ be its skeleton neighborhood, $\mathcal{H}_{0,j} = \{i \neq j : i \notin N_j\}$ the set of nonneighbors, $V_j = |\widehat{S}_j \setminus N_j|$ the number of false rejections, $R_j = |\widehat{S}_j|$ the number of rejections, and $R_j^+ \in [1, m]$ a constant cap on R_j used in the analysis. Let $Z_{j,\text{skel},(s)}$ be the s-th largest value in $\{Z_{kj} : k \in N_j\}$ and define the ranking margin

$$G_j(s) = Z_{j,\text{skel},(s)} - \max_{i \in \mathcal{H}_{0,j}} Z_{ij}, \qquad 1 \leq s \leq |N_j|. \tag{7}$$

When $\mathcal{H}_{0,j}$ is empty, $G_j(s) = +\infty$ by convention. The margin depends on the true skeleton, so it serves only for analysis and for diagnostics on SCMs with known graphs (Section 6.4). Define the event

$$\mathcal{E}_j(s, R_j^+) = \left\{ R_j \leq R_j^+, \;\; G_j(s) > \tau_j \log \frac{\alpha R_j^+ (K+1)}{s} \right\}. \tag{8}$$

Theorem 1 (False-edge exclusion and neighborhood recovery with finite K)**.** *Consider* CSALT *with two disjoint data splits and a fixed number K of permutations. Fix a target j with $|N_j| \geq 1$, some $1 \leq s \leq |N_j|$, and $R_j^+ \in [1, m]$. (i) On $\mathcal{E}_j(s, R_j^+)$, $V_j = 0$; equivalently, $\widehat{S}_j \subseteq N_j$. (ii) If, in addition, every true neighbor $k \in N_j$ satisfies the true-edge power condition*

$$P_{kj} \leq w_{kj} \frac{\alpha |N_j|}{m}, \tag{9}$$

then $\widehat{S}_j = N_j$.

Proof sketch (full proofs in Appendix A). (i) The margin $G_j(s)$ and at least s true-edge terms in the softmax denominator together give $w_{ij} < m/[\alpha R_j^+ (K+1)]$ for every nonneighbor i. If such an i were rejected, weighted-BH self-consistency would require $P_{ij} \leq w_{ij} \alpha R_j / m$, which with $R_j \leq R_j^+$ gives $P_{ij} < 1/(K+1)$, contradicting the Phipson–Smyth floor in Eq. (4). The argument never uses the distribution of P_{ij}, so it needs no super-uniformity and also covers spouses (Section 4). (ii) The condition places all $|N_j|$ true edges below the weighted-BH threshold at $k = |N_j|$, so they are all selected, and part (i) rules out any other rejection (Proposition A.9). □

Because $V_j = 0$ on $\mathcal{E}_j$,

$$\text{FDR}_j^{\text{skel}} := \mathbb{E}\left[\frac{V_j}{R_j \vee 1} \right] \leq \Pr(V_j > 0) \leq \Pr(\mathcal{E}_j^c). \tag{10}$$

Corollary 1 (Complete skeleton recovery)**.** *If isolated targets have no rejections and, for every nonisolated target j, $\mathcal{E}_j(s_j, R_j^+)$ holds for some $s_j \in \{1, \dots, |N_j|\}$ and $R_j^+ \in [1, m]$ (which may vary across targets) while Eq. (9) holds for all $k \in N_j$, then $\widehat{E} = \text{skel}(G^\star)$.*

6 Experiments

We evaluate CSALT on nonlinear synthetic SCMs, where the true skeleton is known, and on two real datasets. Section 6.2 compares skeleton recovery with ten baselines, Section 6.3 isolates the contribution of each component, and Section 6.4 examines whether the conditions of Section 5 account for the observed behavior. Appendix C reports complete results.

6.1 Experimental Setup

Datasets. The main synthetic benchmark comprises 50 nonlinear SCMs on Erdős–Rényi (Erdős & Rényi, 1959) or scale-free graphs. Edge functions are finite-order Fourier series, and mechanisms are additive noise models (ANMs) with heteroscedastic Gaussian noise or post-nonlinear models (PNLs) with Laplace noise and a cubic transform. Sample size n, number of variables p, mean in-degree $\bar{d}$, noise standard deviation σ, and Fourier order M are drawn independently for each SCM from the priors in Appendix B. Each SCM yields one dataset on which each method runs once, and error bars are standard errors across SCMs. For the sample-size and noise experiments, we independently draw 50 SCMs from the same priors at each level, fixing only $n \in \{200, 500, 1000, 2000, 5000, 10000\}$ or $\sigma \in \{0.1, 0.3, 0.5, 1, 2, 5\}$. For real data, Sachs (Sachs et al., 2005) measures a protein signaling network in human immune cells, with 11 variables and 17 edges in the reference skeleton, and UF, from the industrial-process causal discovery benchmark CIPCaD-Bench (Menegozzo et al., 2022), records ultra-processed food production, with 17 variables and 83 edges.

Baselines. We compare against ten methods from five paradigms: the constraint-based method PC (Spirtes et al., 2000) with the kernel conditional independence test KCI (Zhang et al., 2011) or Fisher's z test (PC-KCI and PC-FisherZ); the score-based greedy equivalence search (GES) (Chickering, 2002); the functional causal model DirectLiNGAM (Shimizu et al., 2011); the continuous-optimization methods NOTEARS-MLP (Zheng et al., 2020), GraN-DAG (Lachapelle et al., 2020), and DAG-GNN (Yu et al., 2019); and the pretrained causal discovery models AVICI (Lorch et al., 2022), CDFM (Qiao et al., 2026), and TabCausal (Li et al., 2026). The sample-size and noise experiments add PC-nonlinear, which uses KCI for at most 30 variables and the approximate kernel test RCIT (Strobl et al., 2019) otherwise. PC-KCI and PC-nonlinear run on random subsamples of at most 1,000 observations. Baseline settings appear in Appendix B.

Evaluation metrics. We report edge-wise skeleton F1, precision, and recall (higher is better). Precision corresponds to false-edge exclusion in Theorem 1, and recall to the true-edge power condition in Eq. (9); F1, their harmonic mean, gives the overall comparison. On synthetic data we also report the mean target-wise false discovery proportion (FDP; lower is better): the fraction of false edges among each target's neighbors in the assembled skeleton, averaged over targets within each SCM and then across SCMs.

CSALT settings. CSALT reads attention from LimiX (Zhang et al., 2025) (the default, layers 5–11) or TabPFN v2.5 (Grinsztajn et al., 2025), with identical downstream edge selection. Unless stated otherwise, $\mathcal{D}_1$ and $\mathcal{D}_2$ are equal halves of the observations, and we fix the permutation count $K = 100$, the nominal weighted-BH level $\alpha = 0.10$, and the temperature multiplier $c_\tau = 0.5$. Attention on $\mathcal{D}_1$ uses at most 10,000 observations; permutation scores on $\mathcal{D}_2$ use at most 500 observations when $p \leq 15$ and at most 300 otherwise. None of these hyperparameters was tuned: the layer range was chosen based on the more discriminative attention in intermediate and later layers documented in Figure 5, and the other values were set in advance and are shared by all synthetic and real datasets. Appendix B gives implementation details, Appendix C gives full results, the choice of permutation count, alternative weight functions, and tail estimation for permutation scores.

6.2 Main Results

Higher precision at competitive recall. Table 2 reports skeleton recovery on the synthetic benchmark. With LimiX, CSALT attains the highest F1, together with higher precision and recall than every external baseline. The gain comes mainly from precision. Baselines either select many edges at low precision, as CDFM does, or keep precision high while recovering only about a third of the true edges, as TabCausal and GraN-DAG do; CSALT avoids this trade-off. Paired results in Appendix Table 4 support the advantage on most individual SCMs: against the runner-up PC-KCI, CSALT has higher F1 on 42 of the 49 SCMs where both methods return results.

Robustness to limited samples and strong noise. At every sample size and noise level in Figure 2, CSALT has higher mean F1 than both PC baselines, including at the smallest n and the largest σ. In paired comparisons, it beats PC-nonlinear on about nine in ten SCMs along each axis (263

Table 2: Skeleton recovery on the synthetic benchmark.

Method	F1 ↑	Prec ↑	Rec ↑
CSALT (LimiX)	**0.6346**	0.8963	**0.5275**
CSALT (TabPFN v2.5)	0.4720	**0.9174**	0.3329
PC-KCI	0.5391	0.6949	0.4613
PC-FisherZ	0.4312	0.4769	0.4203
GES	0.4118	0.5609	0.3466
DirectLiNGAM	0.3441	0.5405	0.2710
NOTEARS-MLP	0.2981	0.5751	0.2183
GraN-DAG	0.4535	0.7793	0.3587
DAG-GNN	0.2611	0.5849	0.1911
CDFM	0.4399	0.4681	0.5128
TabCausal	0.3871	0.7897	0.3049
AVICI	0.2311	0.6455	0.1788

Macro-averages across 50 SCMs, omitting SCMs without results (one each for PC-KCI, PC-FisherZ, and AVICI; eight for GES). Paired per-SCM comparisons are in Appendix Table 4.

Table 3: Skeleton F1 on real datasets.

Method	Sachs	UF
CSALT (LimiX)	**0.6667**	**0.7166**
CSALT (TabPFN v2.5)	0.6400	0.0430
PC-KCI	**0.6667**	0.2698
PC-FisherZ	0.6400	0.3817
GES	0.6400	0.6044
DirectLiNGAM	0.5600	0.5521
NOTEARS-MLP	0.5200	0.4147
GraN-DAG	0.5969	0.1569
DAG-GNN	0.5896	0.1477
CDFM	0.6154	0.6136
TabCausal	0.5600	0.4000
AVICI	0.5000	0.4967

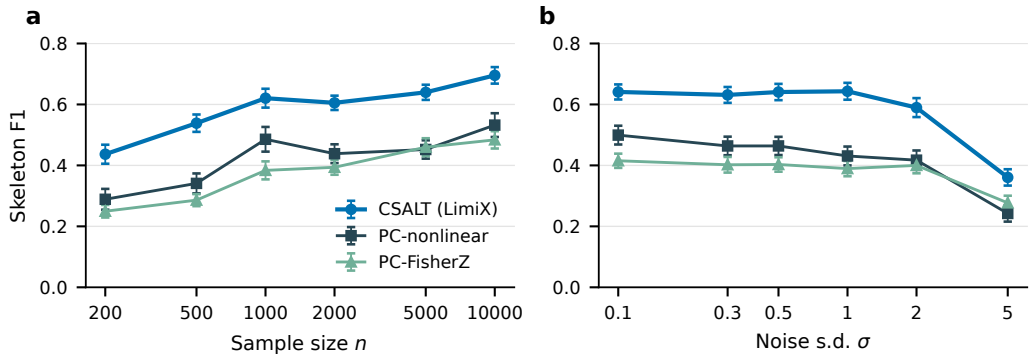


Figure 2: Skeleton F1 versus sample size (a) and noise standard deviation (b), on the SCMs with results for all three methods (CSALT, PC-nonlinear, and PC-FisherZ; 50, 50, 50, 50, 49, and 46 along the sample-size axis; 50, 48, 50, 49, 49, and 50 along the noise axis). Missing cases are PC-FisherZ timeouts. PC-nonlinear uses a random 1,000-observation subsample when $n > 1000$. Error bars are standard errors across SCMs. Each level contains different SCMs, so differences between levels do not isolate the effect of n or σ.

of 295 and 261 of 296). A controlled experiment that varies one factor at a time gives the same ranking across sample sizes (Appendix C). Strong noise lowers recall far more than precision. At $\sigma = 5$, CSALT keeps precision above 0.97 and its errors are mostly missed edges, consistent with weighted BH concentrating rejections on a few high-weight candidates with strong evidence.

Transfer to real systems. With unchanged settings, the LimiX-backed CSALT obtains F1 of 0.7166 on UF, about 0.10 above CDFM, and ties PC-KCI at 0.6667 on Sachs (Table 3). UF has about 4.9 edges per variable, exceeding the mean in-degree range of the synthetic benchmark prior. The backbone comparison shows that transfer depends on the attention representation; the real-data advantage reported here is established for the LimiX-backed configuration.

6.3 Ablation Studies

Attention weights drive the gain. Figure 3a replaces one component at a time. The two variants that remove LimiX attention lose the most F1. Uniform weights, which reduce CSALT to unweighted BH on the same permutation scores, also give the lowest precision. Random-forest importance, used for both the weights and the permutation statistic, lowers F1 further, suggesting that attention from a model pretrained across many tasks carries more adjacency information than importance scores fitted to one dataset. Replacing the permutation null and weighted BH with a

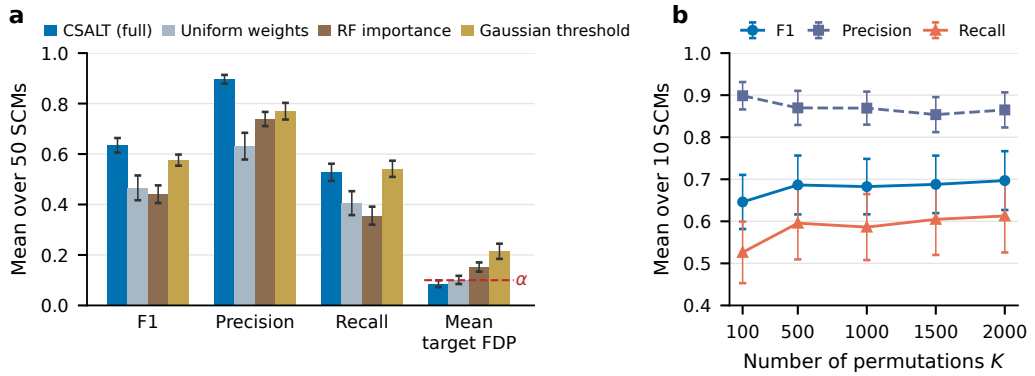


Figure 3: Ablations and permutation count. (a) F1, precision, recall, and mean target-wise FDP of full CSALT and three variants on 50 synthetic SCMs; the red dashed line marks the nominal weighted-BH level $\alpha = 0.10$. (b) F1, precision, and recall on 10 synthetic SCMs as the permutation count K increases from 100 to 2,000. Error bars are standard errors across SCMs; Appendix C examines alternative weight functions.

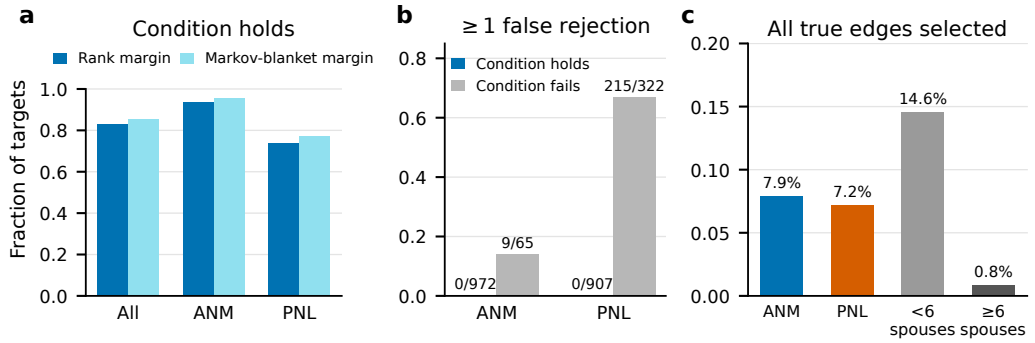


Figure 4: Diagnostics of the theoretical conditions (50 synthetic SCMs; 2,323 targets). (a) Fraction of targets satisfying the ranking-margin and weaker Markov-blanket margin conditions (Appendix A.7), overall and by mechanism. (b) Fraction of targets with at least one false rejection when the condition holds or fails, with target counts above the bars. (c) Fraction of nonisolated targets with all true edges selected, by mechanism and spouse count.

target-wise Gaussian threshold costs less F1 but more than doubles the mean target-wise FDP, so the testing stage is what keeps false edges in check.

Backbone choice and permutation resolution. The two backbones have similar precision on synthetic SCMs, whereas LimiX yields substantially higher recall (Table 2). Increasing K raises recall and slightly lowers precision (Figure 3b), as finer permutation resolution allows more true edges to cross the selection threshold. Adaptive K and tail estimation improve recall at lower cost; the finite-permutation guarantee does not apply to the tail-estimation variant because it replaces the empirical score floor. Softmax weighting outperforms the rank-, quantile-, and sigmoid-based alternatives (Appendix C).

6.4 Empirical Assessment of the Recovery Conditions

The ranking-margin condition covers most targets. Figure 4 summarizes target-wise checks of the conditions of Theorem 1. The ranking-margin condition holds for 82.9% of the 2,266 nonisolated targets, and, as the theorem guarantees, every target with a false rejection violates it. The condition

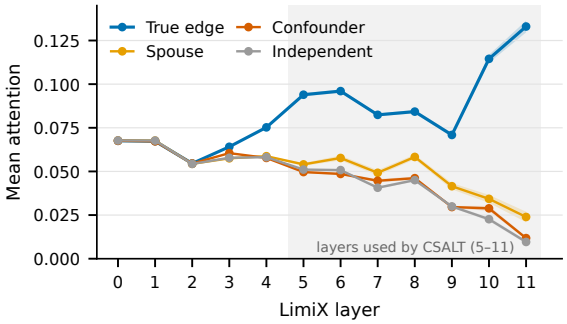


Figure 5: Mean LimiX attention by relationship type across layers (10 synthetic SCMs with 15 variables each). Shading shows standard errors from resampling SCMs, and the gray region marks the default layers 5–11.

holds less often under PNL than under ANM mechanisms, and PNL SCMs account for 215 of the 224 targets with a false rejection.

Attention separates neighbors in intermediate and later layers. The layer-wise analysis in Figure 5 helps explain why the margin often holds. Early layers barely distinguish relationship types; from layer 4 onward, attention to true neighbors rises above that to other variables, and at layer 11 it exceeds attention to independent pairs by more than an order of magnitude. Spouses receive more attention than independent pairs but far less than true neighbors, so the margin can hold even when spouses are present; the default layer range 5–11 lies in this regime.

Precision and true-edge recovery. The ranking-margin condition accounts for the observed false-edge exclusion, while full target-neighborhood recovery is more demanding at the default $K = 100$. Full-neighborhood recall is achieved for 7.5% of nonisolated targets, with lower rates among targets having more neighbors or spouses (Figure 4c). The permutation-count experiment in Figure 3b identifies finer score resolution as one route to improved recall.

7 Conclusion and Discussion

CSALT combines attention from a frozen tabular foundation model with permutation-based edge tests on a disjoint data split. Attention prioritizes candidate edges, while the tests ground edge selection in evidence from the new dataset; no parameter updates or graph supervision are required. We establish two target-wise sufficient conditions for exact neighborhood recovery: a ranking margin that excludes false edges with finitely many permutations, and a true-edge power condition that selects genuine neighbors. Together, they imply complete skeleton recovery under the stated conditions. On nonlinear synthetic SCMs, CSALT achieves the highest skeleton F1 among the evaluated baselines, leading the external methods in both precision and recall and retaining its F1 advantage across the tested sample sizes and noise levels. Without retuning, it also attains the highest F1 on UF and ties for the highest on Sachs. Ablations highlight the contributions of attention weighting and permutation-based testing.

Improving attention separation and allocating permutation effort adaptively offer direct routes to higher recall. Extending the framework beyond the causally sufficient observational setting is an important direction for future work.

AI USE STATEMENT

We used generative AI tools to assist with manuscript polishing, theoretical derivations, and proof checking. The authors reviewed the resulting text and proofs and take responsibility for the entire paper.

REPRODUCIBILITY STATEMENT

Section 4 specifies the full algorithm and its computational cost under the default settings, Section 5 states the theoretical conditions, and Appendix A provides the proofs. Section 6 and Appendix B describe data generation, baselines, metrics, hyperparameters, and random seeds; Appendix C gives supplementary results.

REFERENCES

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300, 1995. doi: 10.1111/j.2517-6161.1995.tb02031.x. URL https://doi.org/10.1111/j.2517-6161.1995.tb02031.x.

Yikang Chen, Zhengkang Guan, Haoyuan Qian, Xingxuan Zhang, Peng Cui, Yi Yang, Fei Wu, and Kun Kuang. DAG-FM: A foundation model for causal discovery under heterogeneous causal mechanisms, 2026. URL https://arxiv.org/abs/2607.11510.

David Maxwell Chickering. Optimal structure identification with greedy search. *Journal of Machine Learning Research*, 3:507–554, 2002. URL https://jmlr.org/papers/v3/chickering02b.html.

Diego Colombo, Marloes H. Maathuis, Markus Kalisch, and Thomas S. Richardson. Learning high-dimensional directed acyclic graphs with latent and selection variables. *The Annals of Statistics*, 40(1):294–321, 2012. doi: 10.1214/11-AOS940. URL https://doi.org/10.1214/11-AOS940.

Paul Erdős and Alfréd Rényi. On random graphs I. *Publicationes Mathematicae Debrecen*, 6(3–4): 290–297, 1959. doi: 10.5486/PMD.1959.6.3-4.12. URL https://publi.math.unideb.hu/paper/2769.

Tomas Geffner, Javier Antorán, Adam Foster, Wenbo Gong, Chao Ma, Emre Kıcıman, Amit Sharma, Angus Lamb, Martin Kukla, Nick Pawlowski, Agrin Hilmkil, Joel Jennings, Meyer Scetbon, Miltiadis Allamanis, and Cheng Zhang. Deep end-to-end causal inference. *Transactions on Machine Learning Research*, 2024. URL https://openreview.net/forum?id=e6sqttxEGX.

Christopher R. Genovese, Kathryn Roeder, and Larry Wasserman. False discovery control with p-value weighting. *Biometrika*, 93(3):509–524, 2006. doi: 10.1093/biomet/93.3.509. URL https://doi.org/10.1093/biomet/93.3.509.

Léo Grinsztajn, Klemens Flöge, Oscar Key, Felix Birkel, Philipp Jund, Brendan Roof, Benjamin Jäger, Dominik Safaric, Simone Alessi, Adrian Hayler, Mihir Manium, Rosen Yu, Felix Jablonski, Shi Bin Hoo, Anurag Garg, Jake Robertson, Magnus Bühler, Vladyslav Moroshan, Lennart Purucker, Clara Cornu, Lilly Charlotte Wehrhahn, Alessandro Bonetto, Bernhard Schölkopf, Sauraj Gambhir, Noah Hollmann, and Frank Hutter. TabPFN-2.5: Advancing the state of the art in tabular foundation models, 2025. URL https://arxiv.org/abs/2511.08667.

Noah Hollmann, Samuel Müller, Katharina Eggensperger, and Frank Hutter. TabPFN: A transformer that solves small tabular classification problems in a second. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=cp5PvcI6w8_.

Noah Hollmann, Samuel Müller, Lennart Purucker, Arjun Krishnakumar, Max Körfer, Shi Bin Hoo, Robin Tibor Schirrmeister, and Frank Hutter. Accurate predictions on small data with a tabular foundation model. *Nature*, 637(8045):319–326, 2025. doi: 10.1038/s41586-024-08328-6. URL https://www.nature.com/articles/s41586-024-08328-6.

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. Curran Associates, Inc., 2008. URL https://proceedings.neurips.cc/paper/2008/hash/f7664060cc52bc6f3d620bcedc94a4b6-Abstract.html.

Sarthak Jain and Byron C. Wallace. Attention is not explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 3543–3556, Minneapolis, Minnesota, 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1357. URL https://aclanthology.org/N19-1357/.

Markus Kalisch and Peter Bühlmann. Estimating high-dimensional directed acyclic graphs with the PC-algorithm. *Journal of Machine Learning Research*, 8(22):613–636, 2007. URL https://jmlr.org/papers/v8/kalisch07a.html.

Theo A. Knijnenburg, Lodewyk F. A. Wessels, Marcel J. T. Reinders, and Ilya Shmulevich. Fewer permutations, more accurate p-values. *Bioinformatics*, 25(12):i161–i168, 2009. doi: 10.1093/bioinformatics/btp211. URL https://doi.org/10.1093/bioinformatics/btp211.

Sébastien Lachapelle, Philippe Brouillard, Tristan Deleu, and Simon Lacoste-Julien. Gradient-based neural DAG learning. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=ryl6nX398r.

Zi-Rong Li, Si-Yang Liu, Tian-Zuo Wang, and Han-Jia Ye. TabCausal: Pretraining across causal environments for tabular causal discovery, 2026. URL https://arxiv.org/abs/2605.31156.

Lars Lorch, Scott Sussex, Jonas Rothfuss, Andreas Krause, and Bernhard Schölkopf. Amortized inference for causal structure learning. In *Advances in Neural Information Processing Systems*, volume 35, pp. 13104–13118. Curran Associates, Inc., 2022. doi: 10.52202/068431-0952. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/54f7125dee9b8b3dc798bb9a082b09e2-Abstract.html.

Nicolai Meinshausen and Peter Bühlmann. High-dimensional graphs and variable selection with the Lasso. *The Annals of Statistics*, 34(3):1436–1462, 2006. doi: 10.1214/009053606000000281. URL https://doi.org/10.1214/009053606000000281.

Giovanni Menegozzo, Diego Dall'Alba, and Paolo Fiorini. CIPCaD-Bench: Continuous industrial process datasets for benchmarking causal discovery methods. In *2022 IEEE 18th International Conference on Automation Science and Engineering (CASE)*, pp. 2124–2131. IEEE, 2022. doi: 10.1109/CASE49997.2022.9926420. URL https://doi.org/10.1109/CASE49997.2022.9926420.

Samuel Müller, Noah Hollmann, Sebastian Pineda Arango, Josif Grabocka, and Frank Hutter. Transformers can do Bayesian inference. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=KSugKcbNf9.

Ignavier Ng, AmirEmad Ghassami, and Kun Zhang. On the role of sparsity and DAG constraints for learning linear DAGs. In *Advances in Neural Information Processing Systems*, volume 33, pp. 17943–17954. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/hash/d04d42cdf14579cd294e5079e0745411-Abstract.html.

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

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

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):Article 39, 2010. doi: 10.2202/1544-6115.1585. URL https://doi.org/10.2202/1544-6115.1585.

Jie Qiao, Ruichu Cai, Zijian Li, Weilin Chen, Pengfei Hua, Boyan Xu, Zhengming Chen, Zhifeng Hao, and Peng Cui. CDFM: Towards a general-purpose causal discovery foundation model, 2026. URL https://arxiv.org/abs/2607.11508.

Jingang Qu, David Holzmüller, Gaël Varoquaux, and Marine Le Morvan. TabICL: A tabular foundation model for in-context learning on large data. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 50817–50847. PMLR, 2025. URL https://proceedings.mlr.press/v267/qu25d.html.

Alexander Reisach, Christof Seiler, and Sebastian Weichwald. Beware of the simulated DAG! causal discovery benchmarks may be easy to game. In *Advances in Neural Information Processing Systems*, volume 34, pp. 27772–27784. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/hash/e987eff4a7c7b7e580d659feb6f60c1a-Abstract.html.

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. doi: 10.1126/science.1105809. URL https://doi.org/10.1126/science.1105809.

Shohei Shimizu, Patrik O. Hoyer, Aapo Hyvärinen, and Antti Kerminen. A linear Non-Gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7(72):2003–2030, 2006. URL https://www.jmlr.org/papers/v7/shimizu06a.html.

Shohei Shimizu, Takanori Inazumi, Yasuhiro Sogawa, Aapo Hyvärinen, Yoshinobu Kawahara, Takashi Washio, Patrik O. Hoyer, and Kenneth Bollen. DirectLiNGAM: A direct method for learning a linear Non-Gaussian structural equation model. *Journal of Machine Learning Research*, 12(33):1225–1248, 2011. URL https://jmlr.org/papers/v12/shimizu11a.html.

Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, Prediction, and Search*. MIT Press, Cambridge, Massachusetts, 2 edition, 2000. ISBN 9780262194402. URL https://mitpress.mit.edu/9780262527927/causation-prediction-and-search/.

Eric V. Strobl, Kun Zhang, and Shyam Visweswaran. Approximate kernel-based conditional independence tests for fast non-parametric causal discovery. *Journal of Causal Inference*, 7 (1):20180017, 2019. doi: 10.1515/jci-2018-0017. URL https://doi.org/10.1515/jci-2018-0017.

Omar Swelam, Lennart Purucker, Jake Robertson, Hanne Raum, Joschka Boedecker, and Frank Hutter. Does TabPFN understand causal structures?, 2025. URL https://arxiv.org/abs/2511.07236.

Ryan Thompson, He Zhao, Daniel M. Steinberg, and Edwin V. Bonilla. Arrow: A foundation model for causal discovery, 2026. URL https://arxiv.org/abs/2605.07204.

Sarah Wiegreffe and Yuval Pinter. Attention is not not explanation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pp. 11–20. Association for Computational Linguistics, 2019. doi: 10.18653/v1/D19-1002. URL https://aclanthology.org/D19-1002/.

Menghua Wu, Yujia Bao, Regina Barzilay, and Tommi S. Jaakkola. Sample, estimate, aggregate: A recipe for causal discovery foundation models. *Transactions on Machine Learning Research*, 2025. URL https://openreview.net/forum?id=h434zx5SX0.

Yue Yu, Jie Chen, Tian Gao, and Mo Yu. DAG-GNN: DAG structure learning with graph neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 7154–7163. PMLR, 2019. URL https://proceedings.mlr.press/v97/yu19a.html.

Kun Zhang and Aapo Hyvärinen. On the identifiability of the post-nonlinear causal model. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, pp. 647–655. AUAI Press, 2009. URL https://dl.acm.org/doi/10.5555/1795114.1795190.

Kun Zhang, Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. Kernel-based conditional independence test and application in causal discovery. In *Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence*, pp. 804–813. AUAI Press, 2011. URL https://dl.acm.org/doi/10.5555/3020548.3020641.

Xingxuan Zhang, Gang Ren, Han Yu, Hao Yuan, Hui Wang, Jiansheng Li, Jiayun Wu, Lang Mo, Li Mao, Mingchao Hao, Ningbo Dai, Renzhe Xu, Shuyang Li, Tianyang Zhang, Yue He, Yuanrui Wang, Yunjia Zhang, Zijing Xu, Dongzhe Li, Fang Gao, Hao Zou, Jiandong Liu, Jiashuo Liu, Jiawei Xu, Kaijie Cheng, Kehan Li, Linjun Zhou, Qing Li, Shaohua Fan, Xiaoyu Lin, Xinyan Han, Xuanyue Li, Yan Lu, Yuan Xue, Yuanyuan Jiang, Zimu Wang, Zhenlei Wang, and Peng Cui. LimiX: Unleashing structured-data modeling capability for generalist intelligence, 2025. URL https://arxiv.org/abs/2509.03505.

Xun Zheng, Bryon Aragam, Pradeep K. Ravikumar, and Eric P. Xing. DAGs with NO TEARS: Continuous optimization for structure learning. In *Advances in Neural Information Processing Systems*, volume 31, pp. 9472–9483. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper/2018/hash/e347c51419ffb23ca3fd5050202f9c3d-Abstract.html.

Xun Zheng, Chen Dan, Bryon Aragam, Pradeep Ravikumar, and Eric Xing. Learning sparse nonparametric DAGs. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pp. 3414–3425. PMLR, 2020. URL https://proceedings.mlr.press/v108/zheng20a.html.

Max Zhu, Martino Mansoldo, Ching-Hao Wang, and Stefan Groha. A causal foundation model for structure and outcome prediction, 2026. URL https://arxiv.org/abs/2606.26467.

APPENDIX

A PROOFS AND EXTENDED CONDITIONS

This appendix proves the results in Section 5 and gives probability statements, a Markov-blanket inclusion condition, and properties of the skeleton after OR assembly. Except in Appendix A.8, all results concern each target's rejection set $\widehat{S}_j$ before OR assembly.

A.1 NOTATION AND ANALYSIS SETTING

Let $\mathcal{V} = \{1, \dots, p\}$ and $m = p - 1$. For target j, let N_j be its skeleton neighborhood, $d_j = |N_j|$, $\mathcal{H}_{0,j} = \mathcal{V} \setminus (\{j\} \cup N_j)$ the set of nonneighbors, $R_j = |\widehat{S}_j|$, and $V_j = |\widehat{S}_j \setminus N_j|$. The target-wise false discovery proportion and false discovery rate are

$$\mathrm{FDP}_j^{\mathrm{skel}} = \frac{V_j}{R_j \vee 1}, \qquad \mathrm{FDR}_j^{\mathrm{skel}} = \mathbb{E}\big[\mathrm{FDP}_j^{\mathrm{skel}}\big], \tag{11}$$

where the expectation is over the observations, data split, and permutations, with the pretrained model parameters held fixed. Let $\mathrm{Pa}(j)$ and $\mathrm{Ch}(j)$ denote the parents and children of j, respectively, and define

$$\mathrm{MB}(j) = \Big(\mathrm{Pa}(j) \cup \mathrm{Ch}(j) \cup \{i : \exists\, c \in \mathrm{Ch}(j),\ i \in \mathrm{Pa}(c)\}\Big) \setminus \{j\}, \qquad \overline{\mathrm{MB}}_j = \mathcal{V} \setminus \big(\{j\} \cup \mathrm{MB}(j)\big).$$

Thus $\mathcal{H}_{0,j}$ is the disjoint union of the spouse set $\mathrm{MB}(j) \setminus N_j$ and $\overline{\mathrm{MB}}_j$. A spouse shares a child with j without being adjacent to it. We set the maximum over an empty set to $-\infty$, so $G_j(s) = +\infty$ when $\mathcal{H}_{0,j} = \varnothing$; in this case, $V_j = 0$ always.

We analyze the algorithm in Section 4 under the following conventions. (i) Z_{ij}, τ_j, and w_{ij} are computed only from $\mathcal{D}_1$ and the frozen pretrained model according to Eqs. (2)–(3). Since $c_\tau > 0$ and $\varepsilon_\tau \in (0, 1)$, we have $\tau_j > 0$, $w_{ij} > 0$, and $\sum_{i \neq j} w_{ij} = m$. (ii) For target j, permutation scores use T_{ij}^{obs} and the K_j permutation statistics actually obtained on $\mathcal{D}_2$, in the form of Eq. (4): $P_{ij} = \big\{1 + \sum_{r=1}^{K_j} \mathbb{I}(T_{ij}^{(r)} \geq T_{ij}^{\mathrm{obs}})\big\}/(K_j + 1)$, where $1 \leq K_j \leq K$ and $K_j = K$ when all permutation statistics are obtained. (iii) Each target applies weighted BH at level $\alpha \in (0, 1)$ to all m candidates as in Eq. (5), with no restriction on the number of rejections and no additional filtering before or after assembly. (iv) $p \geq 2$ (hence $m \geq 1$), both $\mathcal{D}_1$ and $\mathcal{D}_2$ are nonempty, and $K \geq 1$.

Except for Appendix A.6, the results are either deterministic implications for every realization or probability statements obtained directly from those implications. They require neither independence of $\mathcal{D}_1$ and $\mathcal{D}_2$ nor super-uniformity of permutation scores under the skeleton null, $\Pr(P_{ij} \leq t) \leq t$. The probability results in Appendix A.6 additionally assume concentration of attention scores, as stated there.

A.2 BASIC LEMMAS

Lemma A.1 (Finite-permutation floor). *For any $i \neq j$, $P_{ij} \geq 1/(K_j + 1) \geq 1/(K + 1)$.*

Proof. The numerator in Eq. (4) is at least 1 (Phipson & Smyth, 2010), while the denominator satisfies $K_j + 1 \leq K + 1$. □

Lemma A.2 (Weighted-BH self-consistency). *$R_j = \widehat{k}_j$, and every $i \in \widehat{S}_j$ satisfies*

$$P_{ij} \leq w_{ij} \frac{\alpha R_j}{m}. \tag{12}$$

Proof. Because $0 < w_{ij} \leq m$ and Lemma A.1 holds, every weighted score $\widetilde{P}_{ij} = P_{ij}/w_{ij}$ is strictly positive. If $\widehat{k}_j = 0$, the selection threshold in Eq. (5) is zero, no candidate is selected, and $R_j = 0$. If $\widehat{k}_j \geq 1$, then $\widetilde{P}_{(\widehat{k}_j)j} \leq \widehat{k}_j \alpha/m$ by definition, so at least the smallest $\widehat{k}_j$ weighted scores do not exceed the threshold and $R_j \geq \widehat{k}_j$. If $R_j > \widehat{k}_j$, at least $\widehat{k}_j + 1$ scores are at most $\widehat{k}_j \alpha/m$,

giving $\widetilde{P}_{(\widehat{k}_j+1)j} \le \widehat{k}_j\alpha/m < (\widehat{k}_j+1)\alpha/m$, contradicting the maximality of $\widehat{k}_j$. Hence $R_j = \widehat{k}_j$. For $i \in \widehat{S}_j$, $\widetilde{P}_{ij} \le \widehat{k}_j\alpha/m = R_j\alpha/m$; multiplying by $w_{ij} > 0$ yields Eq. (12). □

Lemma A.3 (Low-weight candidates cannot be selected)**.** *Fix $R_j^+ \in [1, m]$. On the event $\{R_j \le R_j^+\}$, if candidate i satisfies*

$$w_{ij} < \frac{m}{\alpha R_j^+(K+1)}, \tag{13}$$

then $i \notin \widehat{S}_j$.

Proof. If $i \in \widehat{S}_j$, Lemma A.2 and $R_j \le R_j^+$ give $P_{ij} \le w_{ij}\alpha R_j/m \le w_{ij}\alpha R_j^+/m < 1/(K+1)$, contradicting Lemma A.1. □

Lemma A.4 (Upper bound on softmax weights)**.** *Suppose $a, b \in \mathbb{R}$ and $1 \le s \le m$. If at least s indices $k \ne j$ satisfy $Z_{kj} \ge a$ and candidate i satisfies $Z_{ij} \le b$, then*

$$w_{ij} \le \frac{m}{s}\exp\Big(-\frac{a-b}{\tau_j}\Big). \tag{14}$$

Proof. Since $\tau_j > 0$, the map $x \mapsto \exp(x/\tau_j)$ is increasing. The denominator in Eq. (3) is $m^{-1}\sum_{k\ne j}\exp(Z_{kj}/\tau_j)$, a sum of positive terms. Retaining only the at least s terms with $Z_{kj} \ge a$ bounds it below by $(s/m)\exp(a/\tau_j)$; the numerator is at most $\exp(b/\tau_j)$. Dividing gives the result. □

Remark A.5 (Two effects of the permutation count). Lemmas A.1 and A.2 imply that selecting candidate i requires $w_{ij} \ge m/[\alpha R_j(K+1)] \ge 1/[\alpha(K+1)]$. A candidate with weight below $1/[\alpha(K+1)]$ therefore cannot be selected, however small its permutation score. With $\alpha = 0.10$ and $K = 100$, this floor is $1/10.1$. Moreover, because $w_{ij} \le m$, obtaining k rejections requires $\alpha k(K+1) \ge 1$. Increasing K lowers the threshold on the right-hand side of Eq. (13): true edges can more readily cross a rejection threshold, but the weight condition for excluding false edges becomes stricter. In the definition of $\mathcal{E}_j$, the margin threshold $\tau_j \log[\alpha R_j^+(K+1)/s]$ grows logarithmically with K.

A.3 Proof of False-Edge Exclusion with Finitely Many Permutations

Proof of Theorem 1. (i) If $\mathcal{H}_{0,j} = \varnothing$, part (i) is immediate. Otherwise, let $a_j = Z_{j,\mathrm{skel},(s)}$ and $b_j = \max_{i\in\mathcal{H}_{0,j}} Z_{ij}$, so $G_j(s) = a_j - b_j$. By the definition of the s-th largest value, at least s candidates in N_j have symmetric scores at least a_j, while every $i \in \mathcal{H}_{0,j}$ satisfies $Z_{ij} \le b_j$. Lemma A.4 gives

$$w_{ij} \le \frac{m}{s}\exp\Big\{-\frac{G_j(s)}{\tau_j}\Big\}, \qquad i \in \mathcal{H}_{0,j}. \tag{15}$$

Let $c_j = \alpha R_j^+(K+1)/s > 0$. On $\mathcal{E}_j(s, R_j^+)$, $G_j(s) > \tau_j \log c_j$; because $\tau_j > 0$, $\exp\{-G_j(s)/\tau_j\} < 1/c_j$. Substituting into Eq. (15) gives

$$w_{ij} < \frac{m}{s\,c_j} = \frac{m}{\alpha R_j^+(K+1)}, \qquad i \in \mathcal{H}_{0,j}.$$

The event $\mathcal{E}_j(s, R_j^+)$ also implies $R_j \le R_j^+$. By Lemma A.3, no $i \in \mathcal{H}_{0,j}$ is selected. Hence $\widehat{S}_j \subseteq N_j$ and $V_j = 0$. The argument applies whether $c_j \ge 1$ or $c_j < 1$.

(ii) Let $d_j = |N_j|$. If Eq. (9) holds for every $k \in N_j$, all d_j true-edge weighted scores satisfy $\widetilde{P}_{kj} \le \alpha d_j/m$. Hence $\widetilde{P}_{(d_j)j} \le \alpha d_j/m$ and $\widehat{k}_j \ge d_j$, so the weighted-BH selection threshold is at least $\alpha d_j/m$ and $N_j \subseteq \widehat{S}_j$. On $\mathcal{E}_j(s, R_j^+)$, part (i) gives $\widehat{S}_j \subseteq N_j$; therefore $\widehat{S}_j = N_j$. □

This proof gives the pathwise inclusion $\mathcal{E}_j(s, R_j^+) \subseteq \{V_j = 0\}$ for every realization. It uses only Lemmas A.1, A.2, and A.4, without assumptions on the distribution of permutation scores for non-neighbors. Because the inclusion holds pathwise, s and R_j^+ may be selected for a particular realization.

Corollary A.6 (Realization-dependent parameters). *Let $d_j \geq 1$ and define*

$$\mathcal{E}_j^* = \{R_j = 0\} \cup \Big\{R_j \geq 1,\ \exists\, s \in \{1, \ldots, d_j\} :\ G_j(s) > \tau_j \log \frac{\alpha R_j (K+1)}{s}\Big\}.$$

On $\mathcal{E}_j^*$, $V_j = 0$. For any fixed $s \in \{1, \ldots, d_j\}$ and $R_j^+ \in [1, m]$, $\mathcal{E}_j(s, R_j^+) \subseteq \mathcal{E}_j^*$.

Proof. If $R_j = 0$, then $V_j = 0$. Suppose a realization has $R_j \geq 1$ and some s satisfies $G_j(s) > \tau_j \log[\alpha R_j(K+1)/s]$. Taking the constant R_j^+ to equal this realized $R_j \in [1, m]$ puts the realization in $\mathcal{E}_j(s, R_j^+)$, and Theorem 1 gives $V_j = 0$. Conversely, consider a realization in $\mathcal{E}_j(s, R_j^+)$. If $R_j = 0$, it belongs to $\mathcal{E}_j^*$. If $R_j \geq 1$, then $1 \leq R_j \leq R_j^+$, and monotonicity of the logarithm with $\tau_j > 0$ yields $\tau_j \log[\alpha R_j(K+1)/s] \leq \tau_j \log[\alpha R_j^+(K+1)/s] < G_j(s)$, so it also belongs to $\mathcal{E}_j^*$. □

The ranking-margin diagnostic in Section 6.4 uses $R_j^+ = \max(R_j, 1)$ and selects the most favorable $s \in \{1, \ldots, d_j\}$; its counted event is contained in $\mathcal{E}_j^*$.

Remark A.7 (Spouse candidates and super-uniformity). The standard weighted-BH FDR proof requires super-uniform p-values under null hypotheses, together with conditions on dependence among p-values (Benjamini & Hochberg, 1995; Genovese et al., 2006). The reference distribution in Eq. (4) comes from permuting the target column and corresponds to the global null that X_j is independent of all inputs. The skeleton null $i \notin N_j$ says only that i and j are nonadjacent; the two nulls differ for spouse candidates. For a linear Gaussian collider, let $X_c = aX_i + bX_j + \varepsilon_c$, where $X_i, X_j \sim N(0,1)$ and $\varepsilon_c \sim N(0, \sigma_c^2)$ are mutually independent, with $ab \neq 0$. Although i and j are nonadjacent, applying the Gaussian conditional-variance formula, and using $X_c - aX_i = bX_j + \varepsilon_c$ once X_i is given, yields

$$\mathrm{Var}(X_j \mid X_c) = \frac{a^2 + \sigma_c^2}{a^2 + b^2 + \sigma_c^2}, \qquad \mathrm{Var}(X_j \mid X_c, X_i) = \frac{\sigma_c^2}{b^2 + \sigma_c^2},$$

and their difference is $a^2 b^2 / [(a^2 + b^2 + \sigma_c^2)(b^2 + \sigma_c^2)] > 0$. Thus, even given the common child, the spouse can still reduce the mean squared error of predicting X_j. Because the model inputs contain both X_c and X_i, attention to X_i can be systematically higher on unpermuted than permuted data, so P_{ij} cannot be assumed super-uniform under the skeleton null. Theorem 1 does not require this property: a spouse, like any other nonneighbor, cannot be selected if its weight is below the threshold in Eq. (13).

A.4 An Upper Bound on FDR

Corollary A.8 (Target-wise FDR). *Under the conditions of Theorem 1, Eq. (10) holds. More generally, for any fixed $s \in \{1, \ldots, d_j\}$ and $R_j^+ \in [1, m]$,*

$$\mathrm{FDR}_j^{\mathrm{skel}} \leq \Pr(V_j > 0) \leq \Pr\{(\mathcal{E}_j^*)^c\} \leq \Pr\{\mathcal{E}_j(s, R_j^+)^c\}. \tag{16}$$

Proof. Since $V_j \leq R_j$, we have $0 \leq V_j/(R_j \vee 1) \leq \mathbb{I}\{V_j > 0\}$. Taking expectations gives the first inequality. Corollary A.6 gives $\{V_j > 0\} \subseteq (\mathcal{E}_j^*)^c \subseteq \mathcal{E}_j(s, R_j^+)^c$; taking probabilities gives the remaining inequalities. Equation (10) is the special case stated in the main text. □

This bound first controls the probability of at least one false rejection and then uses it to control FDR, as in family-wise error rate arguments. It does not rely on calibrated p-values under the skeleton null. If a separate argument establishes $\Pr\{\mathcal{E}_j(s, R_j^+)^c\} \leq \alpha'$, then $\mathrm{FDR}_j^{\mathrm{skel}} \leq \alpha'$. Taking $\alpha' = \alpha$ requires an additional condition; it does not follow from weighted BH alone.

A.5 Including True Edges and Recovering a Neighborhood Exactly

For $d_j \geq 1$, let the event that Eq. (9) holds for every true edge be

$$\mathcal{C}_j = \Big\{P_{kj} \leq w_{kj} \frac{\alpha d_j}{m} \text{ for all } k \in N_j\Big\}.$$

Proposition A.9 (Exact neighborhood recovery). *Suppose $d_j \geq 1$. (i) On $\mathcal{C}_j$, $N_j \subseteq \widehat{S}_j$. (ii) $\{\widehat{S}_j = N_j\} = \{V_j = 0\} \cap \mathcal{C}_j$; in this case, $R_j = \widehat{k}_j = d_j$. (iii) For any $s \in \{1, \ldots, d_j\}$ and $R_j^+ \in [1, m]$, $\widehat{S}_j = N_j$ on $\mathcal{E}_j(s, R_j^+) \cap \mathcal{C}_j$.*

Proof. (i) On $\mathcal{C}_j$, all d_j candidates in N_j have weighted scores at most $\alpha d_j/m$. Hence the d_j-th smallest weighted score satisfies $\widetilde{P}_{(d_j)j} \leq \alpha d_j/m$, and the definition of $\widehat{k}_j$ gives $\widehat{k}_j \geq d_j$. The selection threshold $\widehat{k}_j \alpha/m \geq \alpha d_j/m$ is no smaller than the weighted score of any $k \in N_j$, so $N_j \subseteq \widehat{S}_j$.

(ii) If $V_j = 0$ and $\mathcal{C}_j$ holds, part (i) gives $N_j \subseteq \widehat{S}_j \subseteq N_j$. Conversely, if $\widehat{S}_j = N_j$, then $V_j = 0$ and Lemma A.2 gives $\widehat{k}_j = R_j = d_j$. Each $k \in N_j = \widehat{S}_j$ satisfies $\widetilde{P}_{kj} \leq \widehat{k}_j \alpha/m = \alpha d_j/m$, so $\mathcal{C}_j$ holds.

(iii) This is exactly Theorem 1(ii). □

Part (ii) shows that exact neighborhood recovery is equivalent to two conditions: no false rejections and satisfaction of Eq. (9) by all true edges. If a false rejection occurs, the additional rejections can raise the selection threshold, so selecting all true edges no longer requires that equation. The event $\mathcal{C}_j$ can be rewritten as

$$1 + \#\big\{r \leq K_j : T_{kj}^{(r)} \geq T_{kj}^{\mathrm{obs}}\big\} \leq (K_j + 1)\, w_{kj} \frac{\alpha d_j}{m}, \qquad k \in N_j.$$

The left-hand side is at least 1, so $\mathcal{C}_j$ requires every true edge to have weight $w_{kj} \geq m/[\alpha d_j(K_j+1)]$ and its unpermuted attention on $\mathcal{D}_2$ to be rarely exceeded by permuted attention. Thus, the condition depends on both weights from $\mathcal{D}_1$ and permutation scores from $\mathcal{D}_2$.

Corollary A.10 (Probability of exact recovery). *If, for fixed $s \in \{1, \ldots, d_j\}$ and $R_j^+ \in [1, m]$, $\Pr\{\mathcal{E}_j(s, R_j^+)^c\} \leq \delta_j^F$ and $\Pr(\mathcal{C}_j^c) \leq \delta_j^P$, then*

$$\Pr(\widehat{S}_j = N_j) \geq 1 - \delta_j^F - \delta_j^P. \tag{17}$$

Proof. By Proposition A.9(ii) and Corollary A.6, $\{\widehat{S}_j \neq N_j\} = \{V_j > 0\} \cup \mathcal{C}_j^c \subseteq \mathcal{E}_j(s, R_j^+)^c \cup \mathcal{C}_j^c$. Apply the union bound. □

Here δ_j^F bounds failure of the entire event $\mathcal{E}_j(s, R_j^+)$, including both $R_j \leq R_j^+$ and the ranking-margin condition. When $R_j^+ = m$, the former holds automatically.

A.6 Probability of the Ranking-Margin Event and Attention Concentration

Theorem 1 states what follows when $\mathcal{E}_j(s, R_j^+)$ holds. A numerical value for δ_j^F requires control of three random quantities in this event: the ranking margin $G_j(s)$, the temperature τ_j, and the rejection count R_j. We give a sufficient condition based on concentration inequalities; its assumptions do not follow from the algorithm itself.

Let $N_1 \geq 1$ be the number of observations from $\mathcal{D}_1$ actually used to compute symmetric scores, and let $\Theta_{ij} = \mathbb{E}[Z_{ij}]$ be their expectation at this N_1. Define the population ranking margin

$$\Gamma_j(s) = \Theta_{j,\mathrm{skel},(s)} - \max_{i \in \mathcal{H}_{0,j}} \Theta_{ij},$$

where $\Theta_{j,\mathrm{skel},(s)}$ is the s-th largest value in $\{\Theta_{kj} : k \in N_j\}$. In this subsection, assume $d_j \geq 1$ and $\mathcal{H}_{0,j} \neq \varnothing$.

Assumption A.11 (Concentration of symmetric scores). *For a fixed target j and sample size N_1, suppose there is $\sigma_Z = \sigma_{Z,j}(N_1) > 0$ such that for every $i \neq j$ and $t \geq 0$,*

$$\Pr\big(|Z_{ij} - \Theta_{ij}| > t\big) \leq 2 \exp\Big(-\frac{N_1 t^2}{2\sigma_Z^2}\Big). \tag{18}$$

Remark A.12 (Scope of Assumption A.11). Because the backbone processes observations jointly, bounded attention scores alone do not imply the N_1-dependent concentration in Assumption A.11. The probability bound below is conditional on that assumption; its constants may depend on the target and N_1. A rate uniform across sample sizes would require corresponding uniform constants.

Lemma A.13 (Concentration of margins and temperature). *For $r > 0$, let $\mathcal{A}_j(r) = \{\max_{i \neq j} |Z_{ij} - \Theta_{ij}| \leq r\}$. Under Assumption A.11,*

$$\Pr\{\mathcal{A}_j(r)^c\} \leq 2m \exp\Big(-\frac{N_1 r^2}{2\sigma_Z^2}\Big). \tag{19}$$

On $\mathcal{A}_j(r)$, for every $1 \leq s \leq d_j$,

$$|G_j(s) - \Gamma_j(s)| \leq 2r, \qquad c_\tau \max\{\mu_j - r, \varepsilon_\tau\} \leq \tau_j \leq c_\tau \max\{\mu_j + r, \varepsilon_\tau\}, \tag{20}$$

where $\mu_j = \mathrm{median}_{k \neq j} \Theta_{kj}$.

Proof. Equation (19) follows by applying the union bound to Eq. (18) over the m indices $i \neq j$. Suppose real sequences $x_1, \ldots, x_n$ and $y_1, \ldots, y_n$ satisfy $\max_k |x_k - y_k| \leq r$, and let $x_{(q)}$ and $y_{(q)}$ be their q-th largest values. At least q entries of y are at least $y_{(q)}$, and the corresponding x_k satisfy $x_k \geq y_k - r \geq y_{(q)} - r$. Thus $x_{(q)} \geq y_{(q)} - r$; interchanging x and y yields $|x_{(q)} - y_{(q)}| \leq r$. Apply this result to the s-th largest values of $\{Z_{kj}\}_{k \in N_j}$ and $\{\Theta_{kj}\}_{k \in N_j}$ and to the maxima over $\mathcal{H}_{0,j}$; subtracting gives $|G_j(s) - \Gamma_j(s)| \leq 2r$. A median is an order statistic or the mean of two adjacent order statistics, so $|\mathrm{median}_{k \neq j} Z_{kj} - \mu_j| \leq r$. The temperature bounds follow from the definition of τ_j and monotonicity of $x \mapsto c_\tau \max\{x, \varepsilon_\tau\}$. □

Proposition A.14 (Sources of failure probability for the ranking-margin event). *Fix $1 \leq s \leq d_j$ and $R_j^+ \in [1, m]$, and suppose $c_j = \alpha R_j^+ (K+1)/s \geq 1$. Under Assumption A.11, choose $\delta_j^Z \in (0,1)$ and*

$$r_j = \sigma_Z \sqrt{\frac{2\log(2m/\delta_j^Z)}{N_1}}.$$

Suppose a constant $\bar{\tau}_j$ satisfies $\Pr(\tau_j > \bar{\tau}_j) \leq \delta_j^\tau$ and $\Pr(R_j > R_j^+) \leq \delta_j^R$. If

$$\Gamma_j(s) > \bar{\tau}_j \log c_j + 2r_j, \tag{21}$$

then

$$\Pr\{\mathcal{E}_j(s, R_j^+)^c\} \leq \delta_j^Z + \delta_j^\tau + \delta_j^R, \tag{22}$$

and consequently $\mathrm{FDR}_j^{\mathrm{skel}} \leq \delta_j^Z + \delta_j^\tau + \delta_j^R$. If also $\Pr(\mathcal{C}_j^c) \leq \delta_j^P$, then $\Pr(\widehat{S}_j = N_j) \geq 1 - \delta_j^Z - \delta_j^\tau - \delta_j^R - \delta_j^P$.

Proof. By the choice of r_j and Eq. (19), $\Pr\{\mathcal{A}_j(r_j)^c\} \leq \delta_j^Z$. On $\mathcal{A}_j(r_j) \cap \{\tau_j \leq \bar{\tau}_j\} \cap \{R_j \leq R_j^+\}$, Lemma A.13 and Eq. (21) yield

$$G_j(s) \geq \Gamma_j(s) - 2r_j > \bar{\tau}_j \log c_j \geq \tau_j \log c_j,$$

where the last step uses $\log c_j \geq 0$ and $\tau_j \leq \bar{\tau}_j$. Thus this intersection is contained in $\mathcal{E}_j(s, R_j^+)$. Applying the union bound to its complement gives Eq. (22). The other statements follow from Corollaries A.8 and A.10. □

Remark A.15 (Choosing the failure probabilities). Taking $R_j^+ = m$ sets $\delta_j^R = 0$ at the cost of increasing c_j to $\alpha m(K+1)/s$; if $\alpha(K+1) \geq 1$, then $c_j \geq 1$ automatically. Lemma A.13 implies that with $\bar{\tau}_j = c_\tau \max\{\mu_j + r_j, \varepsilon_\tau\}$, $\mathcal{A}_j(r_j) \subseteq \{\tau_j \leq \bar{\tau}_j\}$. The temperature event can then be removed from the intersection in the proof, reducing the bound in Eq. (22) to $\delta_j^Z + \delta_j^R$. Combining these choices, whenever

$$\Gamma_j(s) > c_\tau \max\{\mu_j + r_j, \varepsilon_\tau\} \log \frac{\alpha m(K+1)}{s} + 2r_j,$$

we have $\Pr\{\mathcal{E}_j(s, m)^c\} \leq \delta_j^Z$. The attention-scale bound $\bar{\tau}_j = c_\tau \max\{\mu_j + r_j, \varepsilon_\tau\}$ is the useful choice here; the generic bound $\tau_j \leq c_\tau$ is too loose under the default settings. For $c_j < 1$, use the lower temperature bound in Eq. (20) because $\log c_j < 0$.

Condition (21) requires population-level attention separation of true neighbors from all nonneighbors, including spouses. Section 6.4 evaluates the corresponding ranking behavior empirically; true-edge power is addressed separately through $\mathcal{C}_j$.

A.7 Markov-Blanket Inclusion

When spouses shrink $G_j(s)$, a weaker condition requires true neighbors to be separated only from variables outside the Markov blanket. Define

$$G_j^{\mathrm{MB}}(s) = Z_{j,\mathrm{skel},(s)} - \max_{i \in \overline{\mathrm{MB}}_j} Z_{ij},$$

$$\mathcal{E}_j^{\mathrm{MB}}(s, R_j^+) = \left\{ R_j \le R_j^+,\ G_j^{\mathrm{MB}}(s) > \tau_j \log \frac{\alpha R_j^+ (K+1)}{s} \right\}.$$

Because $\overline{\mathrm{MB}}_j \subseteq \mathcal{H}_{0,j}$, we always have $G_j^{\mathrm{MB}}(s) \ge G_j(s)$ and hence $\mathcal{E}_j(s, R_j^+) \subseteq \mathcal{E}_j^{\mathrm{MB}}(s, R_j^+)$.

Proposition A.16 (Markov-blanket inclusion)**.** *Let $d_j \ge 1$, $1 \le s \le d_j$, and $R_j^+ \in [1, m]$. On $\mathcal{E}_j^{\mathrm{MB}}(s, R_j^+)$, $\widehat{S}_j \subseteq \mathrm{MB}(j)$; on $\mathcal{E}_j^{\mathrm{MB}}(s, R_j^+) \cap \mathcal{C}_j$, $N_j \subseteq \widehat{S}_j \subseteq \mathrm{MB}(j)$.*

Proof. If $\overline{\mathrm{MB}}_j = \varnothing$, then $\mathcal{V} \setminus \{j\} = \mathrm{MB}(j)$ and the first claim is immediate. Otherwise, replace b_j in the proof of Theorem 1 with $\max_{i \in \overline{\mathrm{MB}}_j} Z_{ij}$. The same argument shows that every $i \in \overline{\mathrm{MB}}_j$ satisfies Eq. (13) and is excluded by Lemma A.3. Thus $\widehat{S}_j \subseteq \mathcal{V} \setminus (\{j\} \cup \overline{\mathrm{MB}}_j) = \mathrm{MB}(j)$. The second claim follows with Proposition A.9(i). □

In an SCM without latent variables and with mutually independent exogenous noise, the Markov-blanket property (Pearl, 2009) implies that variables in $\overline{\mathrm{MB}}_j$ are conditionally independent of X_j given all other variables and provide no additional information for predicting it, whereas spouses may provide information (Remark A.7). This motivates reporting $G_j^{\mathrm{MB}}(s)$ as a weaker condition. The proposition itself depends only on the computed symmetric scores; it does not assume that attention reflects this conditional independence.

A.8 Isolated Targets and the OR-Assembled Skeleton

Proposition A.17 (Isolated targets)**.** *If $N_j = \varnothing$, then $V_j = R_j$, $\mathrm{FDP}_j^{\mathrm{skel}} = \mathbb{I}\{R_j > 0\}$, and $\mathrm{FDR}_j^{\mathrm{skel}} = \Pr(R_j > 0)$.*

Proof. When $N_j = \varnothing$, every selected candidate is a false rejection, so $V_j = R_j$ and $V_j/(R_j \vee 1)$ is 1 when $R_j > 0$ and 0 when $R_j = 0$. □

For an isolated target, all candidates are nulls and the ranking margin is undefined. Controlling $\Pr(R_j > 0)$ requires a separate condition on the lower tail of its permutation scores. Weights alone cannot exclude all candidates: at least one mean-one weight is at least 1, while the threshold in Eq. (13) with $R_j^+ = m$ is $1/[\alpha(K+1)] < 1$ when $\alpha(K+1) > 1$, as in the main-text setting.

Proposition A.18 (OR-assembled skeleton)**.** *Let $E_{\mathrm{skel}}^\star = \{\{i, j\} : i \in N_j\}$ be the true skeleton edge set, and let $\widehat{E}$ be defined by Eq. (6). (i) $\widehat{E} \subseteq E_{\mathrm{skel}}^\star$ if and only if $V_j = 0$ for every $j \in \mathcal{V}$. (ii) $\widehat{E} = E_{\mathrm{skel}}^\star$ if and only if condition (i) holds and every $\{i, j\} \in E_{\mathrm{skel}}^\star$ has $i \in \widehat{S}_j$ or $j \in \widehat{S}_i$. (iii) If $\widehat{S}_j = N_j$ for each nonisolated target and $\widehat{S}_j = \varnothing$ for each isolated target, then $\widehat{E} = E_{\mathrm{skel}}^\star$.*

Proof. The skeleton is undirected, so $i \in N_j$ if and only if $j \in N_i$, equivalently $\{i, j\} \in E_{\mathrm{skel}}^\star$. (i) A pair $\{i, j\}$ belongs to $\widehat{E} \setminus E_{\mathrm{skel}}^\star$ if and only if $i \in \widehat{S}_j \setminus N_j$ or $j \in \widehat{S}_i \setminus N_i$. Thus $\widehat{E}$ has no false edges exactly when every $\widehat{S}_j \subseteq N_j$. (ii) By Eq. (6), $E_{\mathrm{skel}}^\star \subseteq \widehat{E}$ is precisely the stated condition. (iii) Under the premise, $\widehat{S}_j = N_j$ for all j, so both (i) and (ii) hold. □

Part (iii), together with Proposition A.9(iii), gives Corollary 1. By part (ii), the OR rule needs only one endpoint to select each true edge, so part (iii) is stronger than necessary for exact skeleton recovery. Excluding false edges, however, requires every target to have no false rejections.

Corollary A.19 (Error control after skeleton assembly)**.** *For each nonisolated target, fix* $s_j \in \{1, \dots, d_j\}$ *and* $R_j^+ \in [1, m]$*. Let* $\mathrm{FDP}^\vee = |\widehat{E} \setminus E^\star_{\mathrm{skel}}|/(|\widehat{E}| \vee 1)$ *and* $\mathrm{FDR}^\vee = \mathbb{E}[\mathrm{FDP}^\vee]$*. Then*

$$\mathrm{FDR}^\vee \le \Pr\big(\exists\, j : V_j > 0\big) \le \sum_{j:d_j \ge 1} \Pr\{\mathcal{E}_j(s_j, R_j^+)^c\} + \sum_{j:d_j = 0} \Pr(R_j > 0), \tag{23}$$

$$\mathbb{E}\Big[\frac{1}{p}\sum_{j=1}^{p} \mathrm{FDP}_j^{\mathrm{skel}}\Big] \le \frac{1}{p}\Big[\sum_{j:d_j \ge 1} \Pr\{\mathcal{E}_j(s_j, R_j^+)^c\} + \sum_{j:d_j = 0} \Pr(R_j > 0)\Big]. \tag{24}$$

If, for every nonisolated target, $\Pr\{\mathcal{E}_j(s_j, R_j^+)^c\} \le \delta_j^F$ *and* $\Pr(\mathcal{C}_j^c) \le \delta_j^P$*, then*

$$\Pr\big(\widehat{E} = E^\star_{\mathrm{skel}}\big) \ge 1 - \sum_{j:d_j \ge 1} (\delta_j^F + \delta_j^P) - \sum_{j:d_j = 0} \Pr(R_j > 0). \tag{25}$$

Proof. Proposition A.18(i) gives $\mathrm{FDP}^\vee \le \mathbb{I}\{\widehat{E} \not\subseteq E^\star_{\mathrm{skel}}\} = \mathbb{I}\{\exists\, j : V_j > 0\}$. For nonisolated targets, $\{V_j > 0\} \subseteq \mathcal{E}_j(s_j, R_j^+)^c$ by Corollary A.6; for isolated targets, $\{V_j > 0\} = \{R_j > 0\}$. Taking expectations and applying the union bound gives Eq. (23). Equation (24) follows from linearity of expectation, Corollary A.8, and Proposition A.17. Propositions A.18(iii) and A.9(iii) give

$$\{\widehat{E} = E^\star_{\mathrm{skel}}\} \supseteq \bigcap_{j:d_j \ge 1} \{\mathcal{E}_j(s_j, R_j^+) \cap \mathcal{C}_j\} \cap \bigcap_{j:d_j = 0} \{R_j = 0\},$$

and applying the union bound to the complement of the right-hand event gives Eq. (25). □

Target-wise FDR bounds do not directly bound FDR for the assembled skeleton, because OR assembly changes both the false-edge count and the total edge count. Equation (23) follows instead from the stronger event that no target makes a false rejection. The quantity inside the expectation on the left of Eq. (24) averages each target's pre-assembly FDP within an SCM. Under this definition, the diagnostic in Section 6.4 yields 0.0702 when averaged within each SCM first, or 0.0776 when pooled directly across all targets. The main result of 0.0858 uses neighbors in the OR-assembled skeleton and does not correspond to that equation.

B Implementation Details

- **Backbones and attention aggregation.** For LimiX we use the publicly released pretrained LimiX-16M model (12 layers) and average attention over heads and observations in layers 5–11. For TabPFN v2.5 we use the officially released default tabpfn-v2.5-regressor model (18 layers), remove the 64 auxiliary tokens that the model adds, and average attention over observations and heads in layers 8–17.
- **Data split.** The observations are randomly split into disjoint sets with $\mathcal{D}_1 : \mathcal{D}_2 = 1 : 1$ and split seed 42. Attention is computed on at most 10,000 observations of $\mathcal{D}_1$, and permutation scores on at most 500 observations of $\mathcal{D}_2$ when $p \le 15$ and at most 300 otherwise. Feature scaling and label encoding are fitted separately within each split, without information from the other split.
- **Data generation.** Each synthetic SCM uses an additive noise model or a post-nonlinear model with equal probability and, also with equal probability, an Erdős–Rényi or a scale-free graph. The number of variables p and the number of observations n are drawn log-uniformly from $[5, 200]$ and $[500, 10000]$, respectively, and then rounded. Mean in-degree $\bar{d}$ is drawn uniformly from $[1, 4]$, Fourier order M uniformly from $\{3, \dots, 7\}$, heteroscedasticity strength (ANMs only) uniformly from $[0, 0.5]$, and noise standard deviation σ log-uniformly from $[0.3, 0.7]$. The target number of edges is $\lfloor p\bar{d} \rceil$; Erdős–Rényi graphs have exactly $\min\{\lfloor p\bar{d} \rceil, p(p-1)/2\}$ edges, and scale-free graphs have 0.45 to 1.11 times the target. The 50 SCMs of the main benchmark use seeds 700000–700049.
- **Hyperparameters.** $K = 100$ permutations, weighted-BH level $\alpha = 0.10$, temperature multiplier $c_\tau = 0.5$, and permutation batch size $B = 8$.

- **Baselines.** PC-FisherZ, PC-KCI, and PC-nonlinear use significance level 0.05 (causal-learn 0.1.4.4); the latter two randomly subsample 1,000 observations when the sample exceeds 1,000. GES uses the BIC score. DirectLiNGAM keeps edges whose adjacency coefficient exceeds 0.1 in absolute value. NOTEARS-MLP uses the gcastle 1.0.4 implementation with edge-weight threshold 0.3. GraN-DAG uses one hidden layer of 32 units, learning rate 10^{-3}, and CAM pruning after training. DAG-GNN uses hidden dimension 64, 100 outer iterations of 300 training epochs each, and edge threshold 0.1; each SCM runs in float32 precision in a separate process. AVICI (1.0.7) uses the official scm-v0 pretrained model and keeps edges with probability greater than 0.5. CDFM uses the official pretrained weights and its automatically calibrated threshold. TabCausal uses the tabcausal-base weights and keeps edges with probability at least 0.5. Apart from the subsampling in PC-KCI and PC-nonlinear, all methods use every observation of each SCM. On the main benchmark, each method has a runtime limit of 12 hours per SCM, and a timeout counts as a failure.
- **Compute environment.** Experiments ran on eight servers, each with eight NVIDIA GeForce RTX 3090 GPUs (24 GB). The main software versions are torch 2.7.0, causal-learn 0.1.4.4, gcastle 1.0.4, avici 1.0.7 (JAX 0.4.30), and tabpfn 6.4.1. Table 5 reports method runtimes.

C Complete Results and Additional Ablations

This appendix reports coverage and paired comparisons on the main benchmark, the two definitions of false-edge statistics, method runtimes, alternative weight functions, the choice of permutation count, tail estimation for permutation scores, a controlled one-factor experiment, the other axes of the 36-setting experiment, a replication with the current implementation, and directional information in attention.

Coverage and paired comparisons on the main benchmark. PC-KCI and PC-FisherZ each exceeded the 12-hour limit on one SCM, and GES on eight; these SCMs have 39–183 variables. On one SCM with 146 variables and 7,785 observations, AVICI exceeded the memory of a single 24 GB GPU and exhausted system memory when rerun on the CPU. Table 4 compares F1 per SCM on the SCMs where both the baseline and CSALT (LimiX) returned results. The mean F1 difference favors CSALT against every baseline and is smallest for PC-KCI (0.1007). GraN-DAG and CDFM have higher F1 than CSALT on nine and seven SCMs, respectively; every other baseline does so on at most six.

Table 4: Coverage and paired per-SCM comparisons on the main benchmark. *Successes* counts the SCMs on which the listed method returned a result. F1 values and differences are computed on the SCMs where both the listed method and CSALT (LimiX) returned results; F1 difference is CSALT (LimiX) minus the listed method, with the standard error across SCMs in parentheses. The last column counts SCMs on which CSALT (LimiX) has higher, equal, or lower F1.

Method	Successes	CSALT F1	Method F1	F1 diff.	Win / Tie / Loss
CSALT (TabPFN v2.5)	50	0.6346	0.4720	0.1626 (0.0206)	48 / 1 / 1
PC-KCI	49	0.6398	0.5391	0.1007 (0.0157)	42 / 1 / 6
PC-FisherZ	49	0.6406	0.4312	0.2093 (0.0150)	48 / 0 / 1
GES	42	0.6673	0.4118	0.2555 (0.0182)	42 / 0 / 0
DirectLiNGAM	50	0.6346	0.3441	0.2905 (0.0165)	50 / 0 / 0
NOTEARS-MLP	50	0.6346	0.2981	0.3365 (0.0195)	50 / 0 / 0
GraN-DAG	50	0.6346	0.4535	0.1811 (0.0322)	41 / 0 / 9
DAG-GNN	50	0.6346	0.2611	0.3735 (0.0219)	50 / 0 / 0
CDFM	50	0.6346	0.4399	0.1948 (0.0250)	43 / 0 / 7
TabCausal	50	0.6346	0.3871	0.2475 (0.0259)	47 / 0 / 3
AVICI	49	0.6339	0.2311	0.4028 (0.0268)	48 / 0 / 1

Two definitions of false-edge statistics. The diagnostics in Figure 4 use each target's weighted-BH rejection set before the OR rule is applied. Pooled over all 2,323 targets, the mean target-wise

FDP of these sets is 0.0776, and the fraction of targets with at least one false rejection is 0.0964. The mean target-wise FDP in the main results (0.0858) instead counts all neighbors of each target in the assembled skeleton and averages first within each SCM and then across SCMs. With this averaging order, the pre-assembly FDP is 0.0702.

Running time. Table 5 reports runtimes on the 360 SCMs of the controlled one-factor experiment described below, in which the methods in Table 5 were rerun on the same SCMs, with 5–200 variables and 200–10,000 observations. The time for CSALT includes attention extraction on both data splits, the permutations, and edge selection. The median time per SCM for CSALT (LimiX) is 35.7 seconds, below PC-KCI (372.6), PC-nonlinear (322.4), GraN-DAG (166.0), and DAG-GNN (448.2) but above PC-FisherZ, GES, DirectLiNGAM, NOTEARS-MLP, and AVICI. Mean runtimes are strongly affected by a few large SCMs.

Table 5: Runtime in seconds per SCM for each method over the 360 SCMs of the controlled one-factor experiment. GES timed out on ten runs; their times at termination are included.

Method	Median	Mean	Maximum
CSALT (LimiX)	35.7	227.8	5609.2
CSALT (TabPFN v2.5)	27.2	133.7	3152.1
PC-KCI	372.6	1622.2	37275.6
PC-nonlinear	322.4	486.1	3874.4
PC-FisherZ	0.2	8.8	355.6
GES	6.9	522.3	18388.1
DirectLiNGAM	0.6	45.1	2893.8
NOTEARS-MLP	10.4	101.1	3241.8
GraN-DAG	166.0	538.3	11081.8
DAG-GNN	448.2	694.5	8438.9
AVICI	9.9	19.1	258.1

Alternative weight functions. Table 6 compares seven mappings from attention to weights on the 50 main-benchmark SCMs. For each target the weights are normalized to mean one, and all other steps are held fixed. Softmax weights are given by Eq. (3); truncated softmax floors the softmax weights at 0.1 and renormalizes them. Rank and squared-rank weights use attention ranks and squared normalized ranks, respectively. Quantile weights assign 2 to candidates above the median attention and 0 otherwise. Sigmoid weights are $\mathrm{sigmoid}((Z_{ij} - \mathrm{median})/s)$, where s here denotes half the median absolute deviation of attention. Uniform weights recover unweighted BH. Softmax and truncated softmax reach precision near 0.90, compared with 0.71–0.72 for the other nonuniform weights, whose mean target-wise FDP is more than twice as large. Exponential weighting makes the weights of low-ranked candidates small enough, which is the property used in the proof of Theorem 1 (Lemma A.4). The softmax row and the main-text result (F1 0.6346) come from independent runs and differ in F1 by 0.0001.

Table 6: Skeleton recovery with different weight functions on the 50-SCM main benchmark. The last column is the mean target-wise FDP.

Weight function	F1	Precision	Recall	Mean target FDP
softmax (default)	**0.6345**	0.8959	0.5275	0.0859
truncated softmax	0.6338	**0.8975**	0.5266	**0.0850**
squared rank	0.6235	0.7222	**0.5783**	0.2121
quantile	0.5864	0.7094	0.5341	0.2030
sigmoid	0.5749	0.7172	0.5154	0.1868
rank	0.5689	0.7139	0.5134	0.1827
uniform	0.4660	0.6313	0.4054	0.1012

Choosing the permutation count. The main results use a fixed $K = 100$ to limit computation. Adaptive K first runs weighted BH with K_{base} permutations and then checks, for each target, whether finer permutation resolution could change the rejection set. For candidates with at most one permuted statistic no smaller than the unpermuted statistic, it replaces the permutation score with

the floor $1/(K_{\text{refine}}+1)$ and reruns weighted BH. If the number of rejections increases, the target uses K_{refine} permutations; otherwise it keeps the result obtained with K_{base}. We set $K_{\text{base}} = 100$. With $K_{\text{refine}} = 2000$ on 10 synthetic SCMs, 26.9% of targets per SCM on average (29.9% when pooled across all targets) switch to K_{refine}. This implies an average of about 610.5 permutations per target, about 30% of the number used with fixed $K = 2000$, and skeleton F1 is 0.6931 compared with 0.6969 for fixed $K = 2000$. Adaptive K was evaluated only on these 10 SCMs.

Tail estimation for permutation scores. With finitely many permutations, a permutation score cannot fall below $1/(K+1)$, so a true edge whose weighted rejection threshold lies below this floor cannot be selected even when its evidence is strong. Generalized Pareto distribution (GPD) tail estimation (Knijnenburg et al., 2009) keeps $K = 100$ and fits the extreme right tail of the permutation null distribution. For each target, the permutation statistics of the candidates are leave-one-out standardized and pooled into one reference distribution. Up to the 0.90 quantile of this distribution, tail probabilities are empirical proportions; beyond it, a fitted GPD gives probabilities below $1/(K+1)$. If fewer than 50 statistics are available, the method uses a Gaussian approximation; with fewer than 20 right-tail exceedances or a failed fit, it uses an empirical tail probability with a positive floor. Because this estimator removes the $1/(K+1)$ floor, the false-edge exclusion proof of Theorem 1 does not apply to this variant. On the synthetic benchmark, recall rises from 0.5275 to 0.5740, F1 from 0.6346 to 0.6715 (+0.0369), and mean target-wise FDP slightly, from 0.0858 to 0.0980. On UF, recall rises from 0.8072 to 0.8916 and F1 from 0.7166 to 0.7590 (+0.0424); one of the eight newly selected edges is false. Sachs results are unchanged. The estimator can be used when higher recall is needed.

Controlled one-factor experiment. The sample-size and noise experiments in the main text draw independent SCMs from the benchmark priors at each level, so the SCM composition changes between levels. The controlled one-factor experiment instead fixes the background parameters at $p = 15$, an Erdős–Rényi graph, an additive noise model, $n = 2000$, $\bar{d} = 2$, $M = 5$, heteroscedasticity strength 0.3, and $\sigma = 0.5$, and varies one parameter at a time, with 10 SCMs per level (360 SCMs in total). At $p = 15$, PC-nonlinear coincides with PC-KCI. As the sample size rises from 200 to 10,000, F1 rises from 0.6008 to 0.9127 for CSALT, from 0.5127 to 0.7001 for PC-KCI, and from 0.2998 to 0.5779 for PC-FisherZ. CSALT is highest at all six sample sizes; its F1 is close to 0.9 for $n \geq 1000$, where its curve flattens. At noise standard deviation 5, CSALT, PC-KCI, and PC-FisherZ have F1 of 0.3681, 0.3457, and 0.3199, respectively, and CSALT has precision 0.9875 and recall 0.2300.

Other axes of the 36-setting experiment. Figure 6 reports the four axes of the experiment behind Figure 2 that the main text does not show, with GraN-DAG, DAG-GNN, and DirectLiNGAM as additional baselines. CSALT has the highest mean F1 at 20 of the 24 levels, and its precision stays between 0.87 and 0.99 at every level except $M = 0$. From $M = 2$ upward, its lead grows with the Fourier order: at $M = 20$, its F1 is 0.59, against 0.40 for GraN-DAG and at most 0.33 for the other baselines. Across heteroscedasticity strengths, its F1 stays between 0.64 and 0.72, above every baseline. F1 falls with the number of variables for all methods except GraN-DAG, which changes little up to $p = 100$. CSALT drops from 0.90 at $p = 5$ to 0.45 at $p = 200$; GraN-DAG comes within 0.01 of it at $p = 50$ and is slightly ahead at $p = 100$.

Results across additional regimes. At $M = 0$, PC-FisherZ and DirectLiNGAM obtain F1 of 0.70 and 0.67, versus 0.62 for CSALT; at $M = 1$, PC-nonlinear obtains 0.80 versus 0.71. Both Fourier orders are below the main benchmark range. At $p = 100$, GraN-DAG obtains 0.50 versus 0.48 for CSALT. At nominal $\bar{d} = 16$, PC-FisherZ obtains 0.43 versus 0.38 on the 39 SCMs with results for all methods; CSALT obtains 0.32 across all 50 SCMs. In the latter two comparisons, CSALT has higher precision but lower recall. At nominal $\bar{d} = 16$, the realized mean in-degree is 9.3, and PC-FisherZ did not return results on 11 SCMs.

Replication with the current implementation. From each of the 36 settings of the experiment in Figures 2 and 6, we randomly selected 10 SCMs for which all methods returned results (360 in total) and reran CSALT on them with the current implementation, which processes eight permutations per batch; both runs use disjoint splits $\mathcal{D}_1$ and $\mathcal{D}_2$. The rerun-minus-original F1 difference has mean

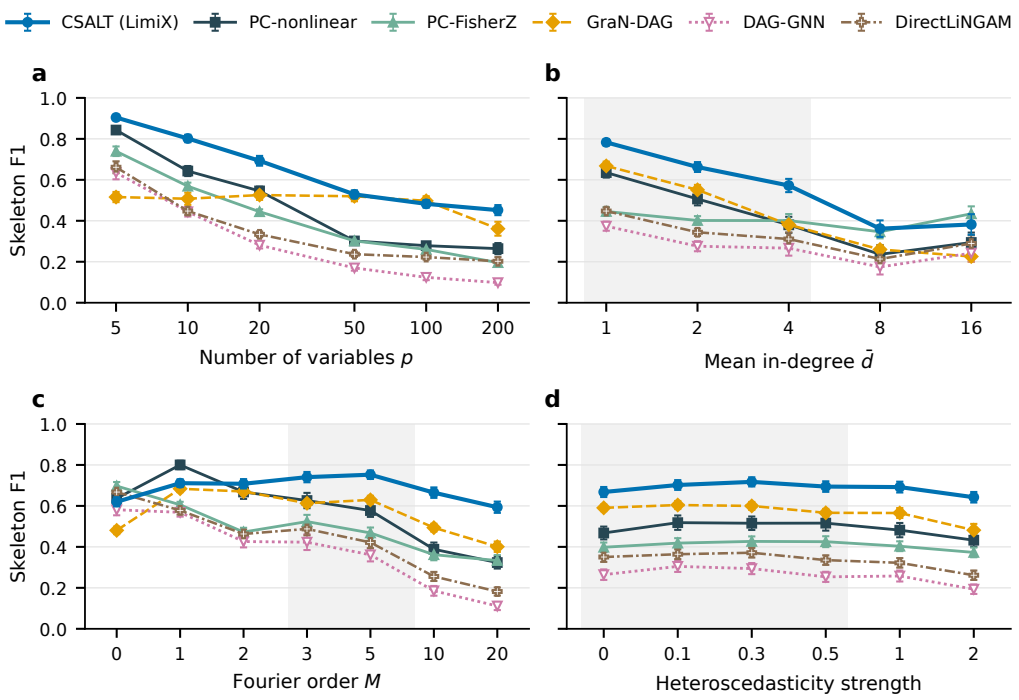


Figure 6: Skeleton F1 on the four axes of the 36-setting experiment that Figure 2 does not show: (a) number of variables p, (b) mean in-degree $\bar{d}$, (c) Fourier order M, and (d) heteroscedasticity strength. At each level, 50 SCMs are drawn with the swept parameter set to that level and the other parameters drawn from the benchmark priors (Appendix B). Shaded regions mark the prior ranges ($\bar{d} \in [1, 4]$, $M \in \{3, \ldots, 7\}$, strength in $[0, 0.5]$); every level in (a) lies inside the prior range $[5, 200]$ of p. In (c) and (d) all SCMs are ANMs, and $M = 0$ gives linear edge functions. In (b), the realized mean in-degree is lower than $\bar{d}$ because a DAG on p variables has at most $p(p-1)/2$ edges and the scale-free graphs in this panel have fewer than $\lfloor p\bar{d} \rfloor$ edges; over the plotted SCMs it averages 0.98, 1.90, 3.46, 6.65, and 9.33. Means are over the SCMs with results for all six methods (32 to 50 of 50 per level); the missing cases are PC-FisherZ and DirectLiNGAM runs that exceeded the 30-minute limit, mostly on SCMs with many variables or observations, and two PC-FisherZ runs that ran out of memory. PC-nonlinear uses a random 1,000-observation subsample when $n > 1000$. Levels in (c) and (d) are evenly spaced. Error bars are standard errors across SCMs.

0.0007 and standard deviation 0.0105, and the absolute value of the mean difference within each setting is at most 0.0061.

D Extended Related Work

Causal discovery fitted to each dataset. Constraint-based methods such as PC remove nonadjacent pairs through marginal and conditional independence tests and, under causal sufficiency, the Markov condition, and faithfulness, return a Markov equivalence class (Spirtes et al., 2000). In high-dimensional Gaussian settings, PC consistently recovers this class when the maximum graph degree is bounded and nonzero partial correlations are bounded below in absolute value (Kalisch & Bühlmann, 2007). FCI and RFCI extend the search to latent confounding and selection bias (Colombo et al., 2012). For Gaussian graphical models, neighborhood selection with the Lasso estimates the neighborhood of each node by a separate regression and combines the estimates by an AND or OR rule (Meinshausen & Bühlmann, 2006); CSALT assembles its target-wise selections in the same way. Kernel conditional independence tests and their random-feature approximations handle nonlinear relationships (Zhang et al., 2011; Strobl et al., 2019). Each edge deletion is traceable to a test, but the number of tests grows quickly with the number of variables and the size of conditioning sets, and finite-sample errors can propagate through deletion and orientation. Score-based methods

such as GES greedily add and remove edges in the space of equivalence classes (Chickering, 2002); results depend on the score, and the search can stop at a local optimum. Functional causal models identify directions through assumptions on functional form or noise (Shimizu et al., 2006; Hoyer et al., 2008; Zhang & Hyvärinen, 2009). Continuous-optimization methods express acyclicity as a smooth constraint (Zheng et al., 2018; Ng et al., 2020) and represent nonlinear conditional mechanisms with neural networks or flow models (Zheng et al., 2020; Yu et al., 2019; Lachapelle et al., 2020; Geffner et al., 2024); their objectives are typically nonconvex and must be optimized anew for each dataset. CSALT also uses the current data but requires neither conditional independence testing nor search or optimization over graphs.

Pretrained causal discovery models. These models are trained on many simulated tasks and applied directly to new datasets. AVICI trains a variational inference model to produce a posterior over causal structures (Lorch et al., 2022). More recent models broaden the scale and mechanistic diversity of the pretraining distribution (Wu et al., 2025; Thompson et al., 2026; Li et al., 2026; Chen et al., 2026; Qiao et al., 2026; Zhu et al., 2026). SEA runs classical methods on subsets of variables and uses a pretrained model to integrate these local outputs with statistics such as the inverse covariance into a structure estimate over all variables (Wu et al., 2025). Arrow predicts a DAG by separately predicting its undirected skeleton and its topological order (Thompson et al., 2026). DAG-FM identifies structure recursively through autoregressive stages that predict leaf nodes and parent nodes (Chen et al., 2026), and TabPFN-CFM predicts structure and outcomes jointly with one model (Zhu et al., 2026). TabCausal covers varied graph priors, functional mechanisms, noise types, dimensionalities, sample sizes, and intervention settings during pretraining (Li et al., 2026). CDFM is pretrained on large-scale synthetic structural causal models with diverse mechanisms and performs zero-shot structure inference on new datasets (Qiao et al., 2026). These models use ground-truth causal graphs as supervision, and their scope depends on the graph, mechanism, and noise families covered by the generator. Methods may also exploit patterns in simulated data that are unrelated to causal structure, such as marginal variance increasing along a topological order (Reisach et al., 2021). CSALT adds no graph-labeled pretraining, and the frozen model outputs no graph: its attention enters only as the weights and permutation statistics of candidate-edge tests on the current data.

Tabular foundation models and attention. A prior-data fitted network (PFN) trains one shared model on many synthetic supervised tasks and gives approximate posterior predictions from context examples at inference time (Müller et al., 2022). TabPFN is pretrained on numerous synthetic supervised tasks and predicts on new datasets without parameter updates (Hollmann et al., 2023; 2025; Grinsztajn et al., 2025); TabPFN v2.5 extends the supported sample size and number of features (Grinsztajn et al., 2025), and TabICL scales to larger sample sizes (Qu et al., 2025). LimiX is pretrained on tabular tasks generated from structural causal models, with objectives such as missing-value imputation but without causal graphs as labels (Zhang et al., 2025). Swelam et al. (2025) freeze the main TabPFN parameters and add learnable causal tokens and a structure decoder that recover adjacencies from hidden representations under additional supervision. The interpretation of attention remains debated: standard attention weights need not agree with other importance measures (Jain & Wallace, 2019), and their explanatory value depends on the question and the validation procedure (Wiegreffe & Pinter, 2019). CSALT reads the original attention of the frozen model, without training a decoder or treating attention as a direct edge decision. The ranking margin in Theorem 1 characterizes the required separation of true neighbors from nonneighbors.