EXACT SEARCH, UNCERTAIN ARROWS: AUDITING LEARNED SCORE TABLES

Anonymous authors
Paper under double-blind review

ABSTRACT

A learned score can prefer the wrong causal graph even when search finds its exact optimum. To understand this mismatch, we introduce an audit framework that separates the ability to find high-scoring graphs from the ability of the score to favor accurate structures. We also examine which arrow directions the predictor favors. Our framework uses tables that store one score per variable and candidate parent set. We project these scores so that all graphs in a Markov equivalence class receive the same total score, then use exact search to find a highest-scoring graph. The component removed by this projection compares how well opposite edge directions fit the data. Better predictive fit alone does not establish the causal direction. We build 180 full ten-variable TabPFN tables and compute their exact optima, enabling repeated search comparisons without further predictor queries. On 70 tables, policy-gradient search finds a graph closer to the true structure than the highest-scoring graph, despite receiving a lower score. BOSS-style search does so on 51. These reversals occur mainly on scale-free graphs, where the highest-scoring equivalence classes often direct edges that the true class leaves undirected. This orientation error persists from 500 to 2000 observations. On Erdős-Rényi graphs, the highest-scoring graph is usually closer to the true structure than the graphs returned by heuristic search. For larger graphs, we find the highest-scoring graph within a restricted search space or using a table with estimated entries. Together, these tables provide a reusable benchmark for assessing search methods and identifying when higher scores fail to yield more accurate causal structures.

1 INTRODUCTION

When a learned score returns the wrong causal graph, should we improve the search algorithm or the score itself? Graph accuracy alone cannot answer this question. Local-score methods predict each variable from its candidate parents, the variables with arrows into it (Hollmann et al., 2025; Sypniewski et al., 2025; Gajewski & Olko, 2026); a table stores one score per variable and parent set. Once this table is fixed, search can change without querying the predictor again: the exact optimum shows how much score the search missed, and comparison with the true graph, when known, shows whether that score rewards accurate graphs (Liu et al., 2012; Wienöbst et al., 2026).

Graphs that imply the same conditional independences form a *Markov equivalence class*, represented by a completed partially directed acyclic graph (CPDAG) that leaves a direction undetermined when members of the class disagree. We measure class error by CPDAG structural Hamming distance (CPDAG-SHD), the number of vertex pairs whose edge state (absent, undirected, or either direction) differs from the true CPDAG.

We first *project* the learned scores so equivalent graphs tie, then find a highest-scoring graph with a *certificate of optimality*, a proof that no allowed graph scores higher. The removed component, the *residual*, records direction preferences within a class; it compares predictive fit, not causal direction.

On 180 full ten-variable TabPFN tables, closing a search method's score gap has opposite effects on the two graph families (Figure 1). The exact optimum has lower mean class error on Erdős–Rényi (ER) graphs, but higher mean error on scale-free (SF) graphs, where its remaining errors are mainly in orientation. Policy-gradient search by proximal policy optimization (PPO) and a search modeled on best order score search (BOSS-style) return strictly lower-scoring, more accurate

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.

classes on 70 and 51 tables, respectively. These counterexamples rule out incomplete optimization as the explanation for the corresponding score misrankings; neither heuristic improves pooled mean accuracy. The audit then asks whether the discrepancy survives fixing the skeleton, changing the objective, or examining nearly tied classes. Section 4.1 reports those checks and their limits.

Contributions. We construct 180 full learned-score tables with raw and projected cells, bootstrap variances, exact projected optima and generating DAGs. They extend precomputed-score benchmarks for exact solvers (Malone et al., 2018): a new search can be compared on both missed score and class error without further predictor queries. We also test whether changing how parent counts are rewarded can repair a ranking. No finite additions depending only on target and parent count can make a true-class DAG globally co-optimal on $101/180$ raw tables; the remaining 79 admit truth-dependent constructions, not a learned repair. On 91 tables, even strictly increasing local-score transformations specific to each target and parent count cannot make a true-class DAG optimal. A matched-domain control isolates how cells unusable by a graph can influence its projected score. Finally, the residual compares predictive direction preferences; its diagnostics validate no arrows, and the held-out guarantee concerns only predictive-independence nulls.

2 SETUP

The data matrix has n rows and d columns, indexed by $V = \{1, \dots, d\}$; the finite-sample statements of Section 3.4 assume independent and identically distributed (i.i.d.) rows from a distribution P^o. A causal reading assumes a distribution Markov and faithful to a directed acyclic graph (DAG) G^o, with no unobserved common causes or relevant feedback. Optimization certificates need none of these assumptions.

Write $\mathrm{pa}_G(i)$ for the parents of i and $|G|$ for the edge count. Two DAGs are Markov equivalent exactly when they have the same skeleton (edges without directions) and the same unshielded colliders $x \to z \leftarrow y$ with x, y nonadjacent (Verma & Pearl, 1990). Write $\mathcal{C}(G)$ for the equivalence class of G; its CPDAG directs the arrows shared by every member and leaves the others undirected.

From cells to graph scores. A local score table contains a finite number $t(i, S)$ for each target i and candidate parent set $S \subseteq V \setminus \{i\}$. A full table has $d\,2^{d-1}$ cells. Each graph selects one cell per target and pays a penalty $\lambda > 0$ per edge:

$$s_t(G) := \sum_{i \in V} t(i, \mathrm{pa}_G(i)) - \lambda |G|. \tag{1}$$

Learning the cells. Split the rows equally into a context D_{tr}, which fits the predictor, and an evaluation half D_{est}, which measures its fit; let $M := |D_{\mathrm{est}}|$ and let m index those rows. For a cell (i, S), the predictor sees only context columns $S \cup \{i\}$. Write $q(x_i \mid x_S; D_{\mathrm{tr}})$ for its scoring function. Nonfinite log scores receive a fixed finite floor (Appendix B); $\log q$ denotes the floored score. A density interpretation additionally requires q to be positive and integrate to one. The learned table is

$$\hat{l}(i, S) := \sum_{m \in D_{\mathrm{est}}} \log q\big(x_i^{(m)} \mid x_S^{(m)}; D_{\mathrm{tr}}\big). \tag{2}$$

We retain row terms and their sum and use the BIC-motivated edge penalty $\lambda = \frac{1}{2} \log M$. Scoring on D_{est} removes an added parent's in-sample gain, but its held-out gain remains noisy. At finite M, irrelevant or weakly relevant parents can still raise the score. The penalty sets the required gain at BIC's rate; Section 4.1 halves and doubles it. An empty parent set gives the no-parent score (for TabPFN, with a fixed-seed auxiliary noise feature, not a fitted marginal).

We compare three predictors through this interface. *TabPFN v2*, a transformer pretrained on synthetic tabular tasks, returns histogram densities (Hollmann et al., 2025); for each target the bins are shared across parent sets, so binning does not by itself create score differences between graphs (unless the floor intervenes). *NIG* is Gaussian linear regression with a normal-inverse-gamma prior and Student-t posterior predictive, correctly specified only for our linear Gaussian mechanisms; its cell priors are assigned separately and its scores need not tie equivalent graphs. *Score-PFN* is a 4.48M-parameter network trained on ER graphs with 3–12 nodes and the benchmark's mechanism families. It clips tails into extreme bins, so its exponentiated score is not a normalized density on $\mathbb{R}$,

and it is a partial prior-alignment control, with a graph-family shift on SF and a size shift at $d = 20$ (Appendix B).

Class scores and direction preferences. For two variables, the true density can be factored in either order: $p^o(x)p^o(y \mid x) = p^o(x, y) = p^o(y)p^o(x \mid y)$, so the two log scores agree on every row. Separate predictive queries see different context columns and need not be conditionals of one joint predictor. For a common conditioning set S, the gap uses four cells:

$$\Delta_t(x, y, S) := [t(x, S) + t(y, S \cup \{x\})] - [t(y, S) + t(x, S \cup \{y\})], \tag{3}$$

where $x \neq y$ and $S \subseteq V \setminus \{x, y\}$; the first bracket adds x before y, the second y before x.

The difference Δ_t also governs equivalence in larger graphs. An edge $x \to y$ is *covered* when x has the parents S and y exactly $S \cup \{x\}$; reversing it stays within the Markov class, leaves the penalty unchanged and changes the score by $-\Delta_t(x, y, S)$. Any two equivalent DAGs are connected by such reversals (Chickering, 1995), so a full table respects all class ties exactly when every such Δ_t is zero. The four queried distributions are *compatible* if they are marginals and conditionals of one joint density for the pair given S, which zeroes the four-cell difference on every row.

The oracle table $l^*(i, S)$ sums the true conditional log density over the evaluation rows. When these values are finite, the density chain rule gives $\Delta_{l^*} = 0$ row by row. The population table $l^o(i, S) := \mathbb{E}_{P^o} \log p^o(X_i \mid X_S)$ has the same property when its conditional entropies are finite. These are reference tables, not convergence claims about learned predictors.

3 What a table gives

3.1 A certificate of optimality

The *domain* is the set of DAGs that the available parent sets allow. For a full table, subset dynamic programming (DP) returns a maximizer over all DAGs on V (Koivisto & Sood, 2004; Silander & Myllymäki, 2006). Its post-query cost is $O(d^2 2^d)$ (Appendix A).

Larger tables use *screening*: for each target, retain the r candidate parents with largest gains $t(i, \{j\}) - t(i, \emptyset)$ and query their subsets up to an in-degree cap c; all singletons remain available (we use $r = 8$, $c = 3$ at $d = 20$). DP remains exact on this restricted domain; an integer linear program (ILP) can instead certify its optimum (Cussens, 2011), and we use ILP at larger d. A timeout can return a candidate graph and an upper bound, but a remaining gap means there is no certificate; screening omits parent sets and does not estimate their scores.

We call a maximizer returned with a certificate the *certified optimum*; on a full table with DP it is the exact optimum. Search comparisons require the *same table, domain and edge penalty*. If a heuristic returns a strictly lower-scoring graph closer to the truth, the score ranks the two graphs incorrectly; if they tie, choosing the worse graph may be a tie-breaking issue. The certificate alone does not identify the source of score error. For example, if the true graph is $X \to Y \to Z$ but the certified optimum is $X \to Y \leftarrow Z$ at score 10.0 while the chain scores 9.7, a heuristic returning the chain is causally closer (CPDAG-SHD 0 against 2) despite its 0.3-nat gap.

3.2 One score per equivalence class

Projection changes the learned table as little as possible so that equivalent graphs tie. Its purpose is not accuracy: it gives a well-defined class score, so that a certificate over DAGs is one over classes, and isolates the directional part (Section 3.3). For $d = 2$, if the forward cells $(X;\ Y \mid X)$ score 1 and 3 while the reverse cells $(Y;\ X \mid Y)$ score 1 and 1, the two graphs score 4 and 2 before their common edge penalty. Moving 0.5 from each forward cell to each reverse cell makes both score 3, with the smallest total squared change; at $d = 2$ the signed unweighted adjustments are $\pm\Delta/4$.

For larger tables, let $\mathcal{E} := \{t : \Delta_t(x, y, S) = 0 \text{ for every } x, y, S\}$ be the linear space that respects all class ties. Every table in it is described by a *potential*, one number $L(T)$ per subset $T \subseteq V$, each cell recording how much this number changes when its target is added (Appendix A).

Theorem 1 (Potential characterization; Cussens et al., 2017, Lemma 5)**.** *A full table t belongs to $\mathcal{E}$ exactly when a potential $L : 2^V \to \mathbb{R}$ satisfies $L(\emptyset) = 0$ and $t(i, S) = L(S \cup \{i\}) - L(S)$ for every*

cell. This potential is unique, and $\dim \mathcal{E} = 2^d - 1$. *Every table in* $\mathcal{E}$ *ties each Markov class, so its full-domain argmax is a union of classes;* l^o *and* l^* *lie in* $\mathcal{E}$.

We project $\hat{l}$ onto $\mathcal{E}$ by minimizing weighted squared cell changes, $\Pi\hat{l} := \arg\min_{t\in\mathcal{E}} \|\hat{l} - t\|$, where the squared norm sums squared cell differences with positive, finite weights (unit weights give the unweighted case). Each weight is the reciprocal of a bootstrap cell variance plus a positive ridge, $w_{i,S} = 1/(\widehat{\mathrm{Var}}_{\mathrm{boot}}[\hat{l}(i,S)] + \tau)$, $\tau > 0$; the identity below uses these same fitted weights.

Proposition 2 (Projection error)**.** *The projection exists and is unique, and for every* $t \in \mathcal{E}$, $\|\hat{l} - t\|^2 = \|\hat{l} - \Pi\hat{l}\|^2 + \|\Pi\hat{l} - t\|^2$*; in particular* $\|\Pi\hat{l} - l^*\| \le \|\hat{l} - l^*\|$*, for sampling error and bias alike.*

This reduces table error in the chosen norm, for the fitted weights; it need not reduce graph error. The removed part, the *residual* $r := \hat{l} - \Pi\hat{l}$, carries every four-cell difference $\Delta_{\hat{l}}$ because the projected component has none. The two parts are orthogonal in the chosen norm but need not be statistically independent. Projection changes stored sums; it does not replace the four queried predictors with a coherent joint predictor. Summing the residual over the cells of a DAG G gives its *residual charge* $\kappa(G) := \sum_i r(i, \mathrm{pa}_G(i))$.

Proposition 3 (What projection removes from a class)**.** *For every DAG* G, $s_{\hat{l}}(G) = s_{\Pi\hat{l}}(G) + \kappa(G)$*; hence for every Markov class* $\mathcal{C}$*, whose members share one projected score* $s_{\Pi\hat{l}}(\mathcal{C})$, $\max_{G\in\mathcal{C}} s_{\hat{l}}(G) = s_{\Pi\hat{l}}(\mathcal{C}) + \beta(\mathcal{C})$ *with* $\beta(\mathcal{C}) := \max_{G\in\mathcal{C}} \kappa(G)$*. If a maximizer of* $s_{\hat{l}}$ *lies in* $\mathcal{C}_{\mathrm{raw}}$ *and one of* $s_{\Pi\hat{l}}$ *in* $\mathcal{C}_\Pi$*, then* $0 \le s_{\Pi\hat{l}}(\mathcal{C}_\Pi) - s_{\Pi\hat{l}}(\mathcal{C}_{\mathrm{raw}}) \le \beta(\mathcal{C}_{\mathrm{raw}}) - \beta(\mathcal{C}_\Pi)$.

So the raw table ranks classes by projected score plus a bonus β, the best charge among their members, and projection drops it; within a class, charges differ only by signed sums of four-cell differences (proof in Appendix A). Whether dropping it helps is empirical: on the 180 tables the two optima share a class on only 11 datasets; the raw optimum is closer on SF, with no consistent difference on ER (Section 4.1). On screened domains, we fit potential differences only on retained cells, using their incident subsets as potential coordinates. Downward closure preserves ties between admissible equivalent DAGs; excluded members of the returned CPDAG receive no certificate (Appendix A).

3.3 What the residual measures

Fix an undirected edge x–y in the CPDAG of a projected-table optimum and let S be the union of its endpoints' already-directed parents; the edge and S are estimated, and S need not be a valid causal adjustment set. Each direction builds a predictive reconstruction of the pair in two steps: the forward direction predicts x from S, then y from x and S; the reverse direction swaps these steps. For values a, b, s of x, y, S, the two products are $J_{x\to y}(a, b \mid s) := q(a \mid s)q(b \mid a, s)$ and $J_{y\to x}(a, b \mid s) := q(b \mid s)q(a \mid b, s)$. Define $\delta_m := \log J_{x\to y} - \log J_{y\to x}$ on evaluation row m; then $\sum_m \delta_m = \Delta_{\hat{l}}(x, y, S)$, and a positive row difference means the forward product assigned that row a higher score. For TabPFN's $S = \emptyset$ cells, which use the auxiliary noise input, the statements below hold row by row, with independent but not identically distributed row terms (Appendix B).

Write $P^o_{xy|X_S}$ for the true conditional law of the pair and $D_{\mathrm{KL}}(P\|Q)$ for the Kullback–Leibler divergence from the true law P to a predictive law Q (smaller means better fit).

Proposition 4 (Compatibility and predictive fit)**.** *Fix the context and a triple* (x, y, S). *Assume positive, normalized queried densities. The relevant true and queried log densities must have finite expected absolute values under* P^o*. The four queries are compatible exactly when the two products agree almost everywhere, or equivalently when* $\delta = 0$ *almost everywhere. Averaging over an independent row from* P^o *gives*

$$\mathbb{E}[\delta \mid D_{\mathrm{tr}}] = \mathbb{E}_{X_S}\Big[D_{\mathrm{KL}}(P^o_{xy|X_S}\|J_{y\to x}) - D_{\mathrm{KL}}(P^o_{xy|X_S}\|J_{x\to y})\Big]. \tag{4}$$

If the queries are compatible for every triple, the table ties every class almost surely and has no residual orientation signal. Incompatible queries can still have zero mean residual.

Equation 4 follows because each product's expected log density is the truth's minus its divergence from the truth, and the true-density terms cancel. A positive mean residual therefore favors the better-fitting reconstruction; it does not by itself identify the causal direction. For the numerical floor and score-PFN's tail clipping, which need not preserve normalization, only the log-ratio algebra survives (Appendix A).

Identifiability requires mechanism assumptions, such as an additive-noise or linear non-Gaussian model, and an estimator that favors the causal reconstruction. Model misspecification can reverse that preference (Schultheiss & Bühlmann, 2023); equal fit of both reconstructions leaves no mean residual to test.

3.4 Direction diagnostics

We use two diagnostics, both with the limitations above. For $K = 10$ disjoint blocks with separately fitted contexts, let z_k be block k's residual mean. The statistic $T_K = \sqrt{K}\,\bar{z}/s_z$ uses their sample mean $\bar{z}$ and sample standard deviation s_z, with a t_{K-1} reference. It selects edges and conditioning sets on the same data, so its calibration and false discovery rate (FDR) are unproved; Benjamini–Hochberg is a descriptive threshold.

A held-out check fixes the predictor and selects the class, conditioning sets and projection weights without its test rows. The forward probability integral transform (PIT) is $u_y = \hat{F}(y \mid x, S)$, the predicted conditional CDF evaluated at the observed target; the reverse swaps x and y. A kernel permutation test checks whether each PIT is independent of its covariates. The edge null is that both independence statements hold; directional p-values give $p_{\text{edge}} = \min\{1, 2\min(p_{\rightarrow}, p_{\leftarrow})\}$. Under i.i.d. rows and this independent selection, the edge p-value is super-uniform under its null; Benjamini–Yekutieli controls FDR over each dataset's selected edge nulls. This controls predictive-independence discoveries, not wrong arrows. A heuristic proposes the direction with the larger p-value, abstaining on ties. Appendix E gives the construction and proof. Non-rejection never validates an arrow.

4 Experiments

Design. The main study uses ER and SF DAGs with $2d$ edges and neural-network mechanisms, $d \in \{10, 20\}$, $n \in \{500, 1000, 2000\}$, and 30 datasets per setting. The 180 datasets at $d = 10$ give full TabPFN tables. SF graphs grow by preferential attachment: hubs are causes, a convention varied in the NIG controls. We report mean CPDAG-SHD and standard error across datasets; edges are not independent replicates. The context/evaluation split is 50:50, the edge penalty is $\lambda = \frac{1}{2}\log M$, and 100 bootstrap resamples of evaluation rows estimate projection weights. Additional audits split datasets at each d into 108 exploration and 72 held-out tables (36 per family), distinct from evaluation rows. Their tests were specified before inspecting these held-out contrasts, but after aggregate patterns were visible (Appendix C).

External baselines are PC-stable, GES and BOSS with BIC, CAM, SCORE, linear and MLP DAGMA, and released AVICI, SEA, BCNP and Arrow checkpoints (Andrews et al., 2023; Zheng et al., 2024; Bühlmann et al., 2014; Rolland et al., 2022; Bello et al., 2022; Lorch et al., 2022; Wu et al., 2025; Dhir et al., 2025; Thompson et al., 2026). Their preprocessing follows the respective methods' protocols; none is tuned on these datasets. Standardization can affect causal-discovery benchmarks (Reisach et al., 2021). BCNP is scored by expected CPDAG-SHD over posterior samples. These end-to-end methods differ from searches on our fixed table. The latter use our PPO policy and a BOSS-style relocation search, with the same table, domain and penalty as the exact solver. They are not reproductions of the published ACD sampler or TabPFN-GES pipelines (Sypniewski et al., 2025; Gajewski & Olko, 2026).

4.1 Exact scores can rank causal structures incorrectly

Our PPO search returns the best of 256 final-policy samples, on average 8.9 ± 0.6 nats below the exact optimum at $d = 10$. Across the 180 tables, it nevertheless returns a closer class on 70 and a farther one on 89; BOSS-style does so on 51 and 47. These counts require a score gap greater than $10^{-6} + 10^{-9}|s^*|$, with s^* the optimum's score. No distinct optimal class ties within this tolerance on any of the 180 tables (Appendix C). Neither search has lower pooled mean SHD (ΔSHD 0.63 ± 0.50 for PPO, 0.17 ± 0.39 for BOSS-style).

SF accounts for 52 PPO and 35 BOSS-style reversals, versus 18 and 16 on ER. SF post hoc sign tests are nonsignificant after Holm correction: PPO $p = 0.098$, BOSS-style $p = 0.41$. On ER, lower

Table 1: CPDAG-SHD at $n = 1000$ (mean $\pm$ s.e. over 30 NN datasets per column; lower is better). "Ours" is each estimator's projected-table optimum: full-table DP at $d = 10$, screened ILP with $r = 8, c = 3$ at $d = 20$. BCNP averages over posterior graph samples. Bold: column best; underlined: best external baseline.

Method	ER(10,20)	SF(10,20)	ER(20,40)	SF(20,40)
Ours (TabPFN)	5.3±0.6	11.4±0.8	20.1±0.9	26.0±1.1
Ours (score-PFN)	**4.2±0.7**	10.8±0.9	19.1±0.9	24.4±1.3
Ours (NIG)	11.3±0.8	16.7±0.8	27.8±0.9	36.4±1.5
PC (Fisher-z)	15.3±0.5	18.7±0.5	26.4±0.7	36.5±0.8
GES (BIC)	15.7±0.9	19.0±0.7	26.8±1.6	41.0±1.7
BOSS (BIC)	11.5±0.9	14.9±0.7	16.4±0.9	27.1±1.1
CAM	10.0±0.9	10.5±1.0	**15.4±1.0**	**20.1±1.5**
SCORE	8.6±0.8	**10.1±1.0**	24.3±1.3	21.7±0.9
DAGMA-linear	14.0±0.6	13.5±0.6	27.4±0.6	28.8±0.9
DAGMA-MLP	10.9±0.7	11.5±0.7	23.0±0.9	25.0±1.0
AVICI	14.8±0.5	14.6±0.6	29.8±0.7	30.2±0.8
SEA	11.5±0.4	13.2±0.5	19.1±0.6	26.7±0.7
Arrow	16.1±1.0	16.5±1.1	28.4±0.9	34.2±1.4
BCNP	22.0±0.2	21.2±0.2	60.3±0.8	58.8±0.9

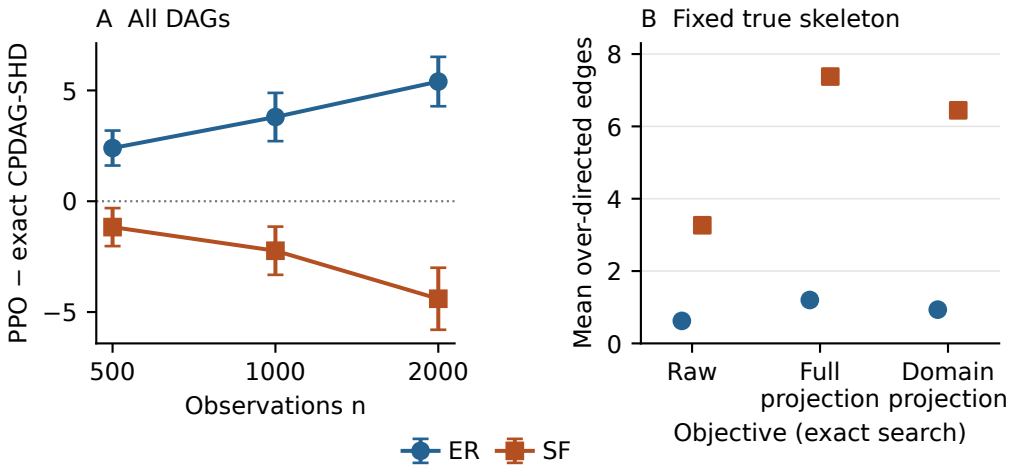

Figure 1: Separate search error from the effect of projection. TabPFN, $d = 10$. (A) All DAGs: paired PPO-minus-exact CPDAG-SHD on full projected tables, mean $\pm$ one s.e. over 30 NN datasets per family and sample size; positive values favor exact search. (B) Fixed true skeleton: mean over-directed edges for exact optima of three objectives on all 90 tables per family, pooling sample sizes (descriptive; uncertainty not displayed). Full projection fits all cells; domain projection fits only cells usable on that skeleton, with the original weights. Every certificate concerns its stated objective and search domain, not causal accuracy.

mean SHD gives a strict majority of wins only against PPO; Appendix Table 2 includes ties and sample-size breakdowns.

Changing the objective in full search. Adjacency errors add or omit an edge; orientation errors disagree on its state when both CPDAGs contain it. On SF, the projected optimum's adjacency errors fall with n, but its orientation errors remain 8.2, 8.0 and 8.9 at $n = 500$, 1000 and 2000. It directs edges that the true class leaves undirected; projected score-PFN and NIG show the same pattern. With search still over all DAGs, the raw optimum has orientation errors 5.6, 4.6 and 4.2. It is closer than the projected optimum on 55 SF tables, equal on 8 and farther on 27 (Holm-adjusted $p = 0.021$). Halving or doubling the penalty, using unit projection weights, and removing a shared nonlinear parent-set-size profile leave substantial over-direction (Appendix C). Projection can remove a residual charge that offsets class-ranking error (Proposition 3); this does not establish causal information in the charge. Raw-table PPO still returns $57/180$ lower-score, closer graphs, including $39/90$ on SF (descriptive post hoc control).

Keeping the original projection fixed, we optimize raw and projected scores over all acyclic orientations of each original optimum's skeleton with the original penalty. Both paths start with the raw objective and its optimum's skeleton, and end with the projected objective and its optimum's skeleton. On 90 SF tables, mean CPDAG-SHD follows $8.44 \to 10.73 \to 11.67$ when changing objective then skeleton, versus $8.44 \to 9.00 \to 11.67$ in the reverse order. The corresponding paths on 90 ER tables are $5.41 \to 5.76 \to 5.54$ and $5.41 \to 5.06 \to 5.54$. Thus SF deterioration persists on average when either selected skeleton is fixed. These post hoc paths describe fixed tables, not causal fractions or stability under new scoring contexts.

Holding the true skeleton fixed. We next optimize all three objectives over the same acyclic orientations of each true skeleton (Figure 1B). Restricting search alone retains the influence of all cells used to fit the full projection. Refitting only cells usable by these orientations, with their original weights and ridge, reverses the full-projected winner's strict ordering against truth on $39/90$ SF and $26/90$ ER tables. These fixed-pair reversals do not count recovered true classes: after reoptimization, the SF optimum still outranks truth on $63/90$ tables, versus $73/90$ before. Mean false colliders remain 1.76 versus 1.80 and increase on $16/90$ tables. This post hoc audit uses all 180 stored tables and the true skeleton; it is a sensitivity diagnostic, not a proposed recovery method (Appendix C). In the earlier 36 held-out SF subset, raw/full-projected optima over-direct 3.14/7.00 edges and add 1.03/1.69 false colliders per table. That collider decrease did not pass the registered Holm correction (Appendix C).

Which parent-count corrections can repair the ranking? Oracle indegree priors already constrain parent counts (Eggeling et al., 2019). We test their limit across the whole true class.

Proposition 5 (Parent-count corrections)**.** *For a full raw table of finite real scores, some true-class DAG can become globally co-optimal for some choice of finite target-specific additions $h_i(|\mathrm{pa}_G(i)|)$ if and only if some true-class DAG maximizes raw score among all DAGs sharing its labeled indegree vector.*

Proof. Use unpenalized raw cells; h_i absorbs the edge penalty. Within a vector all corrections cancel, proving necessity. For sufficiency, reward each node for matching the chosen DAG's count by a bonus larger than the sum of local score ranges. Same-vector graphs cannot beat it; any different vector loses a bonus exceeding its possible raw-score advantage. □

Complete enumeration by covered reversals (Chickering, 1995) excludes all corrections on $101/180$ tables ($51/90$ ER, $50/90$ SF) and constructs them on the other 79. Exclusions fall from 50 to 31 to 20 of 60 as $n = 500, 1000, 2000$. We check all 4634 true members using exact dyadic representations of the stored cells; these are deterministic gaps, not uncertainty estimates. The truth-dependent corrections establish existence, not estimable calibration or unique recovery.

The obstruction is stronger on $91/180$ tables (47 ER, 44 SF): every true-class DAG has a strictly better raw-score competitor obtained by replacing one parent at one target. No strictly increasing transformation of local scores, chosen separately for each target and parent count, can make any true-class DAG globally co-optimal, even with arbitrary count additions. Only one cell changes, so its strict ordering survives every such transformation (Appendix A). The remaining tables are uncertified for this wider family. These post hoc witnesses change the skeleton. They do not explain over-direction or cover parent-identity corrections, transformations introducing ties, refitting or reprojection.

An optimal class can have small reversal margins. For an arrow in the optimal CPDAG, an exact forced-opposite query measures the score lost by the best DAG with that arrow reversed. This margin is an exact loss in the chosen objective, not an arrow confidence level. On 36 held-out SF tables, undirecting arrows with margin at most λ reduces mean SHD from 11.9 to 6.5 (paired gain 5.39, 95% CI $[3.89, 6.89]$). The gain exceeds random undirection of the same number of arrows by 3.75. On ER the same threshold increases SHD from 6.0 to 7.2. These outputs need not be valid CPDAGs; we flag small margins rather than recommend undirection. On six additional ER tables, 200 joint row resamples retain the projected optimal class only 11–132 times per table, refitting projection with fixed context and weights (Appendix C). The SF results concern stored predictor realizations; stability under new contexts or score-generation seeds at $d = 10$ remains untested.

A separate held-out audit at $d = 20$ forces each admissible pair into its three DAG states: absent or either direction. Among the certified classes found by these queries, higher scores rank classes more accurately on average for all three estimators and both families, yet on 71 of 72 TabPFN tables a closer class is found below the optimum; the smallest such gap per table has median 0.58 nats (0.19λ). This is a statement about the queried classes, not all possible neighbors. The full-table SF/ER reversal contrast at $d = 10$ also does not generalize unchanged: at $d = 20$, BOSS-style is closer about as often as farther in both families (Appendix C).

4.2 What the controls explain, and what remains unknown

New synthetic data test whether over-direction is specific to learned scores. For NIG the SF excess replicates under NN mechanisms and is descriptively at least as large under correctly specified LinGauss mechanisms at $n \leq 2000$. The registered hypothesis that misspecification enlarges it was not confirmed. Matching ER and SF on true-class composition descriptively recovers a share 0.87 of the gap, but the registered equivalence test was not passed (Appendix D).

From $n = 2000$ to $20\,000$, over-direction falls under LinGauss from 7.22 to 3.83 edges per table, versus 7.82 and 8.08 under NN. The difference in decline is 3.65 (95% CI $[1.67, 5.63]$). For NN at $n = 20\,000$, a mean share 0.753 of NIG's false colliders on the true skeleton also occurs in the population Gaussian-regression optimum, computed from the covariance matrix (share 0.430 at $n = 2000$). Appendix Lemma 6 constructs dependent but uncorrelated children, each correlated with their parent: the Gaussian criterion then prefers a collider to the true fork. Its explanation need not transfer to TabPFN: overlap between their false colliders was not confirmed. A prospectively specified covariance-matched three-node pilot gives a boundary case (Appendix D.2). Across 24 full tables, raw and projected TabPFN retain the true-density oracle's fixed fork-versus-collider preference, whereas NIG favors the collider on all 12 nonlinear/nearby-fork tables. This engineered covariance trap does not reproduce TabPFN's error; the cause of TabPFN's errors on the SF benchmark remains unresolved.

Predictive preference does not validate direction. The orientation study uses six-variable ER graphs with six edges and $n = 1000$, with NN, LiNGAM and LinGauss mechanisms (30 datasets each). A proposed arrow is wrong if absent from the generating DAG or oriented oppositely; counts precede acyclicity filtering and Meek propagation (Meek, 1995). The block diagnostic decides $8/65$ NN, $19/61$ LiNGAM and $1/61$ LinGauss edges, none wrongly, too few to establish an error rate. Summed across datasets, held-out BY rejects $3/68$ NN edge nulls and $5/64$ LiNGAM edge nulls; respectively $1/3$ and $2/5$ proposed arrows are wrong. For misspecified NIG on NN, $28/30$ are wrong. A rejection detects predictive misfit, not causal direction.

On 16 wind-tunnel walks with known wiring (Gamella et al., 2025), all 19 block-diagnostic decisions are wrong. Temporal dependence violates the i.i.d. assumptions. Marginal entropy also predicts residual preferences in some controls, so agreement with the generating direction need not isolate causal information (Appendix C).

4.3 Beyond full tables

Screening can exclude accurate graphs before scoring begins. We quantify this restriction by minimizing adjacency error over each of the 540 stored $d = 20$ domains, using an oracle that rewards true adjacencies and penalizes false ones. For TabPFN the mean achievable minimum is 1.96 on ER and 2.61 on SF, against 13.66 and 13.28 for its score optimum. Every one of the 540 score optima has adjacency error above its domain minimum. Thus restricted parent sets do not account for all adjacency mistakes. This post hoc audit concerns skeleton error, not the minimum CPDAG-SHD or the predictor's behavior on omitted cells (Appendix C).

Tree imputation sums queried cells along each subset's fixed-order path from $\emptyset$ to define L, then fills cells by $L(S \cup \{i\}) - L(S)$ (Appendix A). Its certificate concerns these imputed scores, not the estimator's unqueried values. At $d = 30, c = 3$ it queries 31,930 of 122,700 capped cells. On five ER datasets at $n = 2000$, all five imputed and all five screened tables obtain certified optima, though not every solver attempt finishes. Imputation yields SHD 37.6 ± 3.3, against 29.4 ± 2.9 for screening and 22.6 for BOSS (BIC). The result illustrates different optimization targets, not a scaling guarantee.

5 Related work

Exact solvers, benchmarked on precomputed scores (Malone et al., 2018), separate search from score failure for classical scores (Liu et al., 2012). FLOP and exact search can beat the true graph's BIC (Wienöbst et al., 2026). For learned scores, five-node enumeration removes search error under TabPFN (Sypniewski et al., 2025) and MLP likelihoods, where Olko et al. (2025) attribute misranking to violated λ-strong faithfulness. TabPFN also drives nonparametric GES (Gajewski & Olko, 2026).

The potential characterization (Cussens et al., 2017) and least-squares projection onto set-function gradients (Stern & Tettenhorst, 2019) are classical. PIT/covariate independence is part of the local coverage null of Zhao et al. (2021); our held-out check tests this predictive property. Its FDR guarantee concerns predictive-independence nulls. Appendix E states the assumptions of the separate block diagnostic. Beyond five-node enumeration, our full and certified screened tables support two linked audits. Comparing exact optima and nearby classes tests how raw and projected scores rank structures. Comparing the residual charges of equivalent DAGs tests preferences within a class. Forced-opposite score losses differ from Bayesian edge posteriors, which average over graphs (Koivisto & Sood, 2004).

6 Conclusion and limitations

To benchmark a new search, keep the table, parent domain and penalty fixed, and report both its gap to the exact optimum and CPDAG-SHD against the known truth. For transformations of stored cells or changes to the penalty, recompute the optimum; no new predictor queries are needed. Compare raw and projected objectives, and specify which cells fit the projection: restricting search alone does not remove other cells' influence. Small orientation margins flag barely preferred arrows; undirecting them helps on SF but hurts on ER.

False colliders persist on the true SF skeleton, and the screened-domain audit finds substantial adjacency error beyond the best feasible skeleton. The controls explain part of NIG's behavior, but the TabPFN and score-PFN error mechanisms remain unknown. Predictive fit and the residual do not establish causal direction; the held-out check controls FDR for predictive-independence nulls, not arrows.

Full tables stop at small d. Screening and imputation change the optimization target. Temporal dependence limits the wind-tunnel interpretation. In structural intervention distance, most reversals hold only for the classes' best members (Appendix C).

Reproducibility statement

Appendix A gives proofs and exact-search recurrences. Appendix B specifies graph and data generation, scoring models, search domains, baseline settings and evaluation metrics. Appendices C and D describe the additional analyses, control experiments and statistical tests; Appendix E details the direction checks.

AI assistance disclosure

AI tools were used to assist with translation into English and to polish the writing, as well as for retrieval and discovery of related work. The authors remain responsible for all claims, code and results.

References

Bryan Andrews, Joseph Ramsey, Rubén Sánchez-Romero, Jazmin Camchong, and Erich Kummerfeld. Fast scalable and accurate discovery of DAGs using the best order score search and grow-shrink trees, 2023.

Kevin Bello, Bryon Aragam, and Pradeep Ravikumar. DAGMA: Learning DAGs via M-matrices and a log-determinant acyclicity characterization, 2022.

Yoav Benjamini and Daniel Yekutieli. The control of the false discovery rate in multiple testing under dependency, 2001.

Peter Bühlmann, Jonas Peters, and Jan Ernest. CAM: Causal additive models, high-dimensional order search and penalized regression, 2014.

David Maxwell Chickering. A transformational characterization of equivalent Bayesian network structures, 1995.

James Cussens. Bayesian network learning with cutting planes, 2011.

James Cussens, David Haws, and Milan Studený. Polyhedral aspects of score equivalence in Bayesian network structure learning, 2017.

Anish Dhir, Matthew Ashman, James Requeima, and Mark van der Wilk. A meta-learning approach to Bayesian causal discovery, 2025.

Ralf Eggeling, Jussi Viinikka, Aleksis Vuoksenmaa, and Mikko Koivisto. On structure priors for learning Bayesian networks, 2019.

Mateusz Gajewski and Mateusz Olko. Nonparametric greedy equivalence search with prior-fitted networks, 2026.

Juan L. Gamella, Jonas Peters, and Peter Bühlmann. Causal chambers as a real-world physical testbed for AI methodology, 2025.

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, 2025.

Mikko Koivisto and Kismat Sood. Exact Bayesian structure discovery in Bayesian networks, 2004.

Zhifa Liu, Brandon Malone, and Changhe Yuan. Empirical evaluation of scoring functions for Bayesian network model selection, 2012.

Lars Lorch, Scott Sussex, Jonas Rothfuss, Andreas Krause, and Bernhard Schölkopf. Amortized inference for causal structure learning, 2022.

Brandon Malone, Kustaa Kangas, Matti Järvisalo, Mikko Koivisto, and Petri Myllymäki. Empirical hardness of finding optimal Bayesian network structures: Algorithm selection and runtime prediction, 2018.

Christopher Meek. Causal inference and causal explanation with background knowledge, 1995.

Mateusz Olko, Mateusz Gajewski, Joanna Wojciechowska, Mikołaj Morzy, Piotr Sankowski, and Piotr Miłoś. Since faithfulness fails: The performance limits of neural causal discovery, 2025.

Jonas Peters and Peter Bühlmann. Structural intervention distance for evaluating causal graphs, 2015.

Alexander G. Reisach, Christof Seiler, and Sebastian Weichwald. Beware of the simulated DAG! causal discovery benchmarks may be easy to game, 2021.

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

Christoph Schultheiss and Peter Bühlmann. On the pitfalls of Gaussian likelihood scoring for causal discovery, 2023.

Tomi Silander and Petri Myllymäki. A simple approach for finding the globally optimal Bayesian network structure, 2006.

Ari Stern and Alexander Tettenhorst. Hodge decomposition and the Shapley value of a cooperative game, 2019.

Mateusz Sypniewski, Mateusz Olko, Mateusz Gajewski, and Piotr Miłoś. Amortized causal discovery with prior-fitted networks, 2025.

Ryan Thompson, He Zhao, Daniel M. Steinberg, and Edwin V. Bonilla. Arrow: A foundation model for causal discovery, 2026.

Thomas S. Verma and Judea Pearl. Equivalence and synthesis of causal models, 1990.

Marcel Wienöbst, Leonard Henckel, and Sebastian Weichwald. Embracing discrete search: A reasonable approach to causal structure learning, 2026.

Menghua Wu, Yujia Bao, Regina Barzilay, and Tommi S. Jaakkola. Sample, estimate, aggregate: A recipe for causal discovery foundation models, 2025.

David Zhao, Niccolò Dalmasso, Rafael Izbicki, and Ann B. Lee. Diagnostics for conditional density models and Bayesian inference algorithms, 2021.

Yujia Zheng, Biwei Huang, Wei Chen, Joseph Ramsey, Mingming Gong, Ruichu Cai, Shohei Shimizu, Peter Spirtes, and Kun Zhang. Causal-learn: Causal discovery in python, 2024.

A Proofs and optimization

Potential characterization (Theorem 1). If $t(i, S) = L(S \cup \{i\}) - L(S)$, the two routes from S to $S \cup \{x, y\}$ telescope to the same difference, so $\Delta_t = 0$. Conversely, for an ordering $(i_1, \ldots, i_k)$ of a set T, define $L(T) = \sum_{j=1}^{k} t(i_j, \{i_1, \ldots, i_{j-1}\})$ and $L(\emptyset) = 0$. An adjacent swap changes this sum by one covered asymmetry. If every asymmetry vanishes, any two orderings have the same sum, since adjacent swaps connect them. Adding i last gives $L(S \cup \{i\}) - L(S) = t(i, S)$. Induction on $|T|$ gives uniqueness. Thus the gradient map $\nabla : L \mapsto t$ is an injective linear map from $2^d - 1$ free potential values onto $\mathcal{E}$. Equivalent DAGs are connected by covered reversals (Chickering, 1995); each changes their score by a zero asymmetry and leaves the edge penalty unchanged. Hence each class ties. The density chain rule gives the same identity for l^* row by row and for l^o in expectation. The dimension includes the d empty-parent scores; subtracting target-specific constants gives the normalized dimension $2^d - d - 1$ in Cussens et al. (2017).

Projection (Proposition 2). Let W be diagonal with the positive finite fitted cell weights. The image of ∇ is a finite-dimensional linear subspace. The weighted least-squares problem has one projected table, and with $L(\emptyset) = 0$ its full-table potential is unique. The normal equations are $\nabla^\top W \nabla L = \nabla^\top W \hat{l}$; they also specify the weighted implementation. The residual $r = \hat{l} - \Pi\hat{l}$ is W-orthogonal to $\mathcal{E}$. For $t \in \mathcal{E}$, expand $\hat{l} - t = r + (\Pi\hat{l} - t)$: the cross term vanishes, proving the identity and the oracle bound. The argument conditions on the fitted weights.

Class charge (Proposition 3). Summing $\hat{l} = \Pi\hat{l} + r$ over a DAG's cells proves the first identity; the edge penalties cancel. The projected sum is constant within a Markov class, so maximizing the raw score within that class adds $\beta(\mathcal{C}) = \max_{G \in \mathcal{C}} \kappa(G)$. Projected optimality gives $s_{\Pi\hat{l}}(\mathcal{C}_\Pi) \geq s_{\Pi\hat{l}}(\mathcal{C}_{\text{raw}})$. Raw optimality gives $s_{\Pi\hat{l}}(\mathcal{C}_{\text{raw}}) + \beta(\mathcal{C}_{\text{raw}}) \geq s_{\Pi\hat{l}}(\mathcal{C}_\Pi) + \beta(\mathcal{C}_\Pi)$. Rearranging proves the bound; it does not separate systematic preferences from maximization over class members.

Compatibility and fit (Proposition 4). For fixed s, positive normalized queried conditionals make both products $J_{x \to y}$ and $J_{y \to x}$ normalized joint densities. If the four queries come from one joint density, its two factorizations agree. Conversely, equality of the products defines one joint density; integrating either factorization recovers its queried marginal, and division by the positive marginal recovers its queried conditional. This proves compatibility. Integrability permits subtracting the two KL expressions: their common true-log-density term cancels and leaves $\mathbb{E} \log(J_{x \to y} / J_{y \to x})$. Averaging over X_S proves Eq. 4. Compatibility of every queried triple zeroes each row difference and every covered asymmetry, so projection removes nothing. The converse from zero mean is false because positive and negative row differences can cancel. If flooring or tail clipping destroys normalization, only the log-ratio algebra is asserted.

Strictly increasing local calibrations. For each target i and parent count k, let $\phi_{i,k} : \mathbb{R} \to \mathbb{R}$ be strictly increasing and $h_i(k)$ finite. Consider

$$F(G) = \sum_i \big[\phi_{i,|\mathrm{pa}_G(i)|}(t(i, \mathrm{pa}_G(i))) + h_i(|\mathrm{pa}_G(i)|)\big].$$

An ordinary edge penalty is absorbed by h_i. Suppose every true-class DAG G has a feasible competitor H differing only in one target's parent set, with the same count and a strictly larger raw cell. All unchanged terms and the count addition cancel in $F(H) - F(G)$; strict monotonicity makes the remaining difference positive. Thus, for every choice of these functions, every true member is strictly beaten. The functions may depend on the table and even on the truth. This includes target-specific positive rescaling, but not edge-specific penalties or density transformations with parent-dependent normalization.

We enumerate each complete true class and test every single-parent replacement, rejecting cycles. Exact dyadic cell differences certify the sufficient condition on $91/180$ tables; absence of a witness does not prove that a monotone repair exists. Whole-graph score improvement with several changed targets would not suffice: different transformations can reverse a sum of positive and negative changes. Non-strict transformations can introduce ties, and reprojection can change the local ordering. No uniform positive transformed gap or stability across new predictor contexts is claimed.

Restricted domains. The screened domains are downward closed separately for every target: retaining (i, S) retains (i, S') for $S' \subseteq S$. Regard each cell as an oriented edge $S \to S \cup \{i\}$ of a subset graph. Every cycle in this graph is a sum of complete square faces. To see this, take a divergence-free edge flow and a largest set T incident to its support. At least two support edges enter T; downward closure supplies the two lower edges of their square. Subtracting a multiple of its circulation cancels one upper edge. Repeat to eliminate all support at this height, then descend. At height one no nonzero cycle remains. Zero circulation on all present squares therefore defines a potential on each connected component, which extends to all subsets. Conversely, every gradient has zero circulation. The implemented least-squares projection on the present face constraints thus ties any equivalent admissible DAGs. Potentials can have component-wise constants, but the projected table is unique. A returned CPDAG may include DAGs outside the restricted domain; no certificate is assigned to them. Without downward closure, a face-free hexagon gives a counterexample to the argument.

Exact search and imputation. For an allowed parent family $\mathcal{A}_i$, set

$$B_i(U) = \max_{S \subseteq U,\ S \in \mathcal{A}_i} \{t(i, S) - \lambda|S|\}, \qquad C(\emptyset) = 0, \quad C(U) = \max_{i \in U} \{C(U \setminus \{i\}) + B_i(U \setminus \{i\})\}.$$

Every DAG has a sink i. Removing it bounds the rest by $C(U \setminus \{i\})$ and its cell by B_i. Conversely, appending a maximizing sink to an optimal smaller DAG attains the recurrence without creating a cycle. This proves exactness for full and restricted domains. Subset-max preprocessing and this recurrence use $O(d^2 2^d)$ operations.

For tree imputation with cap c, fix an order of variables and let $i(T)$ be the last element of a nonempty set T, $|T| \le c+1$. Query only $t(i(T), T \setminus \{i(T)\})$ and recursively define $L(T) = L(T \setminus \{i(T)\}) + t(i(T), T \setminus \{i(T)\})$, with $L(\emptyset) = 0$. Fill each capped cell by $L(S \cup \{i\}) - L(S)$. The recursion reproduces every queried cell and makes all present squares close. Its query count is $\sum_{k=1}^{c+1} \binom{d}{k}$, versus $d \sum_{k=0}^{c} \binom{d-1}{k}$ full capped cells. It does not reproduce unqueried predictor values in general and can depend on the chosen order. Search certifies this imputed table alone.

B Experimental design and estimators

Graphs and mechanisms. ER DAGs draw $2d$ edges uniformly among forward pairs of a random order. SF graphs add nodes in a random arrival order, drawing $\mathrm{round}(2d/(d-1))$ earlier parents (all if fewer) without replacement with weights degree plus one; uniform forward pairs complete the requested $2d$ edges. Hubs are therefore causes. For $k \ge 1$ parents, NN mechanisms are $f_i(z) = \sum_{h=1}^{16} w_h \tanh(a_h^\top z + b_h)$, with independent $a_h \sim N(0, (1.5^2/k)I)$, $b_h \sim N(0, 0.5^2)$ and $w_h \sim N(0, 0.375^2)$. Roots have $f_i = 0$; independent Gaussian noises have standard deviations

uniform on $[0.4, 0.8]$. LiNGAM and LinGauss use linear mechanisms with uniform and Gaussian noise, respectively. The main study uses 30 datasets per graph family, dimension and sample size. Baselines receive the same matrix.

Scoring details. TabPFN uses package 2.2.1 and a fixed v2 regressor checkpoint, with two ensemble members for structure and four for the primary orientation study. Histogram borders are shared across a target's parent-set queries and checked by the adapter. With fixed bin shapes, every cell has a target-specific shape contribution; a DAG uses each target once, so this contribution cancels in score differences. Re-estimating variance weights after a change need not preserve that invariance. For empty parent sets, the adapter supplies a fixed-seed Gaussian noise feature because the interface requires an input. This is a numerical no-parent score, not a fitted marginal. Auxiliary noise can make row terms independent but not identically distributed; the synthetic split-PIT tests use nonempty conditioning sets. Nonfinite row log-densities are replaced by $\log(2^{-126})$; finite values below this number are retained. Flooring occurred in 48 of 360 sweep tables. It changes the table and need not preserve density normalization. Predictors fit preprocessing on context rows only.

NIG regresses the target on $(1, x_S)$ with $\beta \mid \sigma^2 \sim N(0, 10\sigma^2 I)$, including the intercept, and $\sigma^2 \sim \mathrm{InvGamma}(2, 1)$. Its Student-$t$ posterior predictive is evaluated in the target's original units. These independently assigned cell priors need not yield score equivalence. The linear Gaussian family is correctly specified under LinGauss, although the finite-context predictive still has estimation error.

Score-PFN is a 4.48M-parameter pre-norm transformer (8 layers, width 256, four heads) with a 1000-bin head on $[-6, 6]$ in context-standardized units. Its prior draws 3–12 node ER graphs with $m = \min(\mathrm{round}(dU), d(d-1)/2)$, $U \sim \mathrm{Unif}[1, 2]$, the three mechanism families uniformly, and a log-uniform row count from 200 to 2000. With probabilities $1/2, 1/4, 1/4$, parent sets include each other variable with probability $1/2$, draw a uniform size in $\{0, \ldots, \min(3, d-1)\}$ and then a uniform subset of that size, or contain every true parent plus each remaining variable with probability 0.3. Training uses AdamW, learning rate 10^{-4} with cosine decay, batch 32 and 250,000 steps in bfloat16. The original checkpoint is reused without retraining. Tail values map to extreme bins, so the exponentiated score has constant tails and is not a normalized density. Its nominal histogram CDF, clipped to $[10^{-6}, 1 - 10^{-6}]$, is not the CDF of that score.

Domains, comparisons and metrics. Screening uses raw singleton gains, keeps subsets of the top r parents up to cap c, and also keeps every out-of-candidate singleton. At $d = 20$, $r = 8, c = 3$. Weights use the reciprocal of bootstrap variance plus a positive ridge $\tau = \max(10^{-8}, 10^{-3}\,\mathrm{median}(v))$, where v are the cell variances; the structure study uses 100 shared row-count bootstrap draws, the estimated-class orientation study 200, and the split-PIT selection 20. These weights stay fixed during search. Our PPO trains a Plackett–Luce order vector and independent Bernoulli edges given that order for 76,800 graph evaluations, then returns the best of 256 final-policy samples. Excluding those samples, the saved training-best graphs have lower/equal/higher CPDAG-SHD than the exact optimum on $41/19/30$ SF and $18/17/55$ ER tables (90 per family); these counts are descriptive. BOSS-style uses order relocation and is distinct from external BOSS with BIC. SHD is also computed when PC returns an invalid CPDAG. We check solver status and independently recompute objectives, using the main text's tolerance and identical domains.

Baseline configurations. PC-stable uses Fisher-z at 0.05; GES uses Gaussian BIC in `pcalg` without turning; external BOSS uses BIC in causal-learn. BCNP subsamples at most 1000 rows, standardizes and pads to 20 variables, and averages CPDAG-SHD over posterior DAG samples; its checkpoint adapter is checked against the reference implementation. DAGMA-MLP uses 20 000/40 000 warm-up/final iterations per subproblem, below its defaults of 50 000/80 000; the larger budget was not tested. Synthetic PC-KCI was excluded in full after repeated six-hour timeouts, rather than averaging completed cases. Pretrained models are not adapted to our mechanisms; DAG outputs are converted to CPDAGs.

C Additional audit results and their interpretation

Objective controls. These post hoc controls do not uniquely justify the original choices. On 90 full $d = 10$ SF TabPFN tables (30 per n), exact search over all DAGs yields mean over-directed edge

Table 2: Search on all 180 full $d = 10$ TabPFN tables: counts of closer and farther classes at strictly lower score, with mean search-minus-optimum SHD (and s.e.). The retained optimum and each search use the same projected table, domain and penalty. Score ties are included in Equal.

Graph	n	Search	N	Better	Worse	Equal	Score ties	ΔSHD
ER	500	PPO	30	9	16	5	5	2.40±0.79
ER	500	BOSS-style	30	10	10	10	10	0.77±0.75
ER	1000	PPO	30	6	21	3	2	3.80±1.09
ER	1000	BOSS-style	30	3	7	20	19	1.77±0.92
ER	2000	PPO	30	3	23	4	3	5.40±1.11
ER	2000	BOSS-style	30	3	9	18	18	2.17±0.94
SF	500	PPO	30	14	11	5	1	−1.17±0.86
SF	500	BOSS-style	30	11	8	11	10	−0.30±0.75
SF	1000	PPO	30	17	11	2	2	−2.23±1.09
SF	1000	BOSS-style	30	9	9	12	11	−0.70±0.70
SF	2000	PPO	30	21	7	2	0	−4.40±1.40
SF	2000	BOSS-style	30	15	4	11	9	−2.70±1.36
ER	All	PPO	90	18	60	12	10	3.87±0.59
ER	All	BOSS-style	90	16	26	48	47	1.57±0.50
SF	All	PPO	90	52	29	9	3	−2.60±0.66
SF	All	BOSS-style	90	35	21	34	30	−1.23±0.57
Pooled	All	PPO	180	70	89	21	13	0.63±0.50
Pooled	All	BOSS-style	180	51	47	82	77	0.17±0.39

counts $7.5/8.2/7.1/7.3$ for the original penalty, $\lambda/2$, 2λ, and unit projection weights; nonlinear-profile subtraction gives 7.37. The profile accumulates mean gains from true non-descendant, non-parent additions to true parent sets on same-n ER tables, leaving the scored ER table out; we subtract its residual after fitting a line against parent count, since a linear size term changes the penalty. Raw is closer/equal/farther than the projected optimum on $34/22/34$ ER datasets.

Search on raw scores. The raw-score control uses all 180 tables, the original domain and penalty, 600 batches of 128 training graphs, and the best of 256 final-policy samples. It makes no predictor queries. Each raw optimum is recomputed exactly; CPDAG-SHD evaluates the returned DAG's class although raw scores are not class-equivalent. PPO is closer/farther at strictly lower score on $18/31$ ER and $39/30$ SF tables. Score ties number $33/12$ and equal SHD at lower score $8/9$, respectively. Mean PPO-minus-optimum SHD is 1.31 ± 0.40 on ER and -1.08 ± 0.54 on SF (s.e.). The true-skeleton control instead optimizes each objective over all acyclic orientations of the same true skeleton, with a constant edge penalty. On 36 held-out SF tables, the paired one-sided false-collider comparison gives $p = 0.006$; it did not pass the original round-4 Holm correction.

Projection input on a fixed skeleton. For true skeleton H, let $A_H = \{(i, S) : S \subseteq N_H(i)\}$, the cells usable by its acyclic orientations. We refit the potential of Section 3.2 on these cells, retaining their full-table weights and ridge. All three objectives use the same orientations of H and edge penalty. For refitted table q_H, full-projection winner G_F and true DAG G_0, a reversal requires $s_{\Pi i}(G_F) - s_{\Pi i}(G_0) > \epsilon$ and $s_{q_H}(G_F) - s_{q_H}(G_0) < -\epsilon$. Here $\epsilon = 10^{-6} + 10^{-9}m$, where m is the largest absolute score among the two optima and their respective objectives at G_0. Independent solves and outside-class optima verify the numerical comparisons. The post hoc oracle diagnostic tests projection sensitivity, not new contexts or a predictor mechanism.

Checking optimal-class ties. We maximize over a union of constrained domains changing a skeleton adjacency or an unshielded-collider state. Every DAG outside the reference class satisfies at least one such constraint; none admits a member of that class. The best outside score therefore checks uniqueness of the optimal class. All 180 gaps exceed the declared tolerance; this is a numerical statement about the stored tables, not resampling stability.

Margins and nearby classes. For a directed optimal CPDAG edge, remove cells inconsistent with its opposite DAG direction and solve again. The loss is the exact margin for that query. Undirecting selected arrows produces a PDAG that need not be a valid CPDAG, so its SHD gain is a diagnostic and not a new causal estimator. Margin AUC distinguishes wrong from correct arrows on held-out SF tables (pooled-arrow descriptive AUC 0.89), but random undirection already recovers part of

the gain. Forced-pair queries at $d = 20$ retain only pairs for which all three states are admissible, deduplicate the resulting Markov classes, and compute Kendall τ_b between score and SHD within that set. The test uses datasets as units and requires negative mean τ_b for each estimator and family; it is an intersection–union test. Mean τ_b on SF/ER is $-0.281/-0.309$ for TabPFN, $-0.272/-0.286$ for score-PFN and $-0.201/-0.241$ for NIG; the intersection–union p-value is 5.5×10^{-11}. The t tests can be liberal under skewness; negative τ_b occurs on 71 of 72 held-out TabPFN tables. It measures a coarse ranking that can reflect adjacency changes, not discrimination among orientations alone. At $d = 20$, BOSS-style gives closer/farther TabPFN classes at strictly lower score on $35/35$ SF tables and $35/35$ ER tables.

Conditional row stability. Six additional ER/NN restriction-study datasets ($d = 10, n = 2000$) retain evaluation-row scores. For each, 200 multinomial vectors resample the 1000 rows jointly across cells and raw/projected branches. Each table is reprojected and optimized exactly, holding context, predictor, query randomness, penalty and original weights fixed. The original class is retained on 11–132 of 200 draws after projection, versus 1–14 before it. This is conditional class instability on six datasets, not 1200 independent samples or causal confidence. The cohort is neither SF nor the 180 benchmark tables; context and weight uncertainty remain untested.

Other limits of the interpretation. Within the raw optimum's class, agreement with truth on its true-skeleton undirected edges averages 0.84 across tied charge-maximizers, versus 0.61 uniformly across all members. These are dataset means on the $68/72$ held-out tables with eligible edges. A marginal-entropy rule recovers much of that gain on exploration data and all of it on ER, so the comparison does not isolate causal information from generator sortability. Structural intervention distance (SID; Peters & Bühlmann, 2015) depends on the class member chosen. Of the 70 PPO reversals in CPDAG-SHD, 59 persist using each class's minimum SID and 20 using its maximum SID. These counts are descriptive and do not compare a unique causal effect estimate.

Selection and uncertainty. The exploration/held-out split stratifies by family and sample size. Four stored-table rounds registered 34 hypotheses, tested 13 and confirmed ten; untested hypotheses counted as $p = 1$. Round k used Holm at $0.05/2^k$. Hypotheses followed visible aggregate patterns and earlier rounds, so these held-out estimates give no family-wise guarantee about a newly discovered phenomenon. Neither the $d = 20$ margin AUC threshold nor the shared parent-size explanation was confirmed. Separate post hoc sign tests are exact and two-sided, omit SHD ties (also score ties for searches), and use Holm over nine comparisons: PPO, BOSS-style or raw versus projected optima on ER, SF or pooled data.

The minimum error permitted by a screened domain. Let E be the true skeleton. Give each stored parent-set cell one unit per true adjacency and minus one per false adjacency. Its sum on a DAG G is $|E| - |E \triangle \mathrm{skel}(G)|$. Exact subset DP computes the minimum skeleton error on all 540 domains (90 per family and estimator). It agrees with 415 initial ILP certificates and resolves 125 timeouts within their bounds. Both solvers record their witnesses. The mean minima on ER/SF are 1.96/2.61 for TabPFN, 1.76/2.57 for score-PFN and 3.06/4.62 for NIG. Their mean score-optimum errors are respectively 13.66/13.28, 12.90/12.04 and 18.13/22.76. All 540 optima exceed their domain minimum. A feasible true skeleton need not admit the true Markov class.

D NEW DATA AND A SECOND-MOMENT EXPLANATION

Design and registration. All round-5/6 datasets have ten variables and 20 edges. Round 5 uses NN and LinGauss with ER, SF, reversed SF, randomly ordered SF skeletons, and class-matched ER, plus LiNGAM with ER and SF. Class matching rejects ER draws until their true class has an undirected-edge count within one of its paired SF count. There are 30 test replicates per mechanism, graph and $n \in \{500, 1000, 2000\}$, plus NN samples at 5000 and 20 000. Round 6 uses 60 new test replicates per NN/LinGauss SF cell at $n = 2000, 20\,000$. Pilot and test seeds are disjoint. Tables use the same NIG scoring, weighted projection and exact DP pipeline. True-skeleton optima use exact DP over acyclic orientations of that skeleton. Two tests instead reuse the earlier held-out SF datasets (regenerated NIG and stored TabPFN). The plan paired SF and reversed SF skeletons, but the generator drew them independently. One round-6 run timed out before computing its statistic and was rerun unchanged.

Hypotheses, directions, thresholds, exclusions and code were fixed before testing. Round 5 registered nine tests, with two withdrawn before testing; round 6 registered three. Holm levels were $0.05/2^5$ and $0.05/2^6$; withdrawn tests counted as $p = 1$. The unit is the dataset. Across all six rounds, 46 hypotheses were registered, 23 tested, 16 confirmed, seven not confirmed, and 23 withdrawn. The fresh data are distinct, but the hypotheses were motivated by earlier results; the two reused-split tests are less blind than those on fresh data.

Positive and negative findings. NIG over-directs a mean share 0.837 of eligible true-undirected pairs on SF versus 0.483 on ER under NN; pairs must be adjacent in both true and estimated graphs. The rates average $\mathrm{U}{\to}\mathrm{D}/T_u$ per table; T_u counts eligible true-undirected pairs. Excluding $T_u = 0$ leaves 90 SF and 81 ER tables out of 90 each. The registered misspecification contrasts compare the SF-minus-ER rate difference between mechanisms; they have the opposite sign to the prediction: -0.238 (95% CI $[-0.377, -0.099]$) for NN versus LinGauss and -0.177 ($[-0.320, -0.034]$) for NN versus LiNGAM. Neither is confirmed. The class-matched ER rate is 0.791 versus 0.837 on SF, recovering descriptively 0.87 of the gap; equivalence within 0.1 is not established (TOST $p = 0.054$). The correctly specified control therefore shows that SF over-direction at these sample sizes is not by itself proof of misspecification. With linear Gaussian mechanisms, dropping weak edges under the edge penalty can create false colliders, and this effect falls with increasing sample size.

The three round-6 tests distinguish this effect from nonlinear second-moment error. The fall with n is larger under LinGauss (main text). The Welch test is somewhat liberal under the right-skewed LinGauss cell at $n = 20\,000$; the prespecified stratified bootstrap-t check gives $p = 0.00033$, below its Holm threshold 3.9×10^{-4} with little margin. Over-direction decreases without vanishing. At $n = 20\,000$, more than half of NIG's true-skeleton false colliders under NN coincide with the joint population second-moment optimum; the mean share is 0.753 with 95% CI $[0.652, 0.854]$. This averages the per-table precision $|\mathrm{FC}_s \cap \mathrm{FC}_p|/|\mathrm{FC}_s|$ over 56 tables with nonempty sample-optimum collider sets FC_s (4 excluded); FC_p comes from the joint optimum over the whole true skeleton, not separate triple tests. Confirmation rests on the registered approximate one-sample t test of mean precision above $1/2$ ($p = 2.9 \times 10^{-6}$, Holm-adjusted 8.8×10^{-6}). The prespecified sign test ($p = 6.4 \times 10^{-5}$) concerns the median. A distribution-free Hoeffding bound does not pass Holm correction. The mean sensitivity $|\mathrm{FC}_s \cap \mathrm{FC}_p|/|\mathrm{FC}_p|$ is 0.834. At $n = 2000$ the share is only 0.430. The third test, of overlap with TabPFN's false colliders, is not confirmed. Covariances are approximated from 4×10^6 draws of each generating model; finite-context NIG is not identified with this population criterion. No asymptotic statement about TabPFN or score-PFN follows.

D.1 A NONLINEAR FORK CAN FAVOR A COLLIDER UNDER GAUSSIAN SCORES

For positive definite covariance Σ, define the population linear-regression residual variance $\sigma^2_{i|S} = \Sigma_{ii} - \Sigma_{iS}\Sigma_{SS}^{-1}\Sigma_{Si}$ and table $l^G(i, S) = -\frac{M}{2}\log(2\pi e\,\sigma^2_{i|S})$, with $M \geq 1$. Write s_G for Eq. 1 using l^G. The determinant identity $\sigma^2_{i|S} = \det \Sigma_{S\cup\{i\}} / \det \Sigma_S$ makes l^G score-equivalent. This criterion depends only on second moments, irrespective of the data distribution.

Lemma 6 (Second-moment error on a fork)**.** *For (X, H, Y) with finite second moments and positive definite covariance, let G_f be any member of the fork class on $X - H - Y$. For every edge penalty,*

$$s_G(X \to H \leftarrow Y) - s_G(G_f) = \frac{M}{2}\log\frac{1 - \rho^2_{XY}}{1 - \rho^2_{XY\cdot H}}.$$

Here $\rho_{XY\cdot H}$ is the correlation of the linear-regression residuals after regressing on H. The collider is strictly preferred exactly when $\rho^2_{XY\cdot H} > \rho^2_{XY}$. In a fork $X = f(H) + \varepsilon$, $Y = g(H) + \eta$, with $\operatorname{Var} H > 0$, finite second moments, $\mathbb{E}[\varepsilon \mid H] = \mathbb{E}[\eta \mid H] = \mathbb{E}[\varepsilon\eta \mid H] = 0$ and positive noise variances, an affine f or g precludes strict collider preference. Nevertheless, there are parameters of the paper's 16-unit $\tanh$ family with Gaussian H and independent Gaussian noises for which $\rho_{XY} = 0$ but $\rho_{XH}, \rho_{YH}, \rho_{XY\cdot H} \neq 0$. Then $N(0, \Sigma)$ is faithful to the collider and, for sufficiently large M with $\lambda = \frac{1}{2}\log M$, every maximizer over all three-node DAGs belongs to that collider class.

Proof. All orientations on the fixed skeleton pay the same penalty. Comparing the collider with $H \to X, H \to Y$ gives a ratio of products of regression residual variances. Substituting the de-

terminant identities cancels common terms and leaves $(1-\rho_{XY}^2)/(1-\rho_{XY\cdot H}^2)$ inside the score difference. Other fork-class members tie.

For the structural fork set $v = \operatorname{Var} H$, $a = \operatorname{Cov}(f(H), H)/v$, $b = \operatorname{Cov}(g(H), H)/v$, $\tilde{f} = f(H) - aH$, $\tilde{g} = g(H) - bH$, $c = \operatorname{Cov}(\tilde{f}, \tilde{g})$, $A = \operatorname{Var}\tilde{f} + \mathbb{E}\varepsilon^2$, $B = \operatorname{Var}\tilde{g} + \mathbb{E}\eta^2$. The conditional-mean assumptions cancel noise cross terms. The residual covariance matrix is $\left(\begin{smallmatrix} A & c \\ c & B \end{smallmatrix}\right)$, positive definite because it is a covariance matrix plus positive diagonal noise. Thus $\rho_{XY\cdot H} = c/\sqrt{AB}$ and $\rho_{XY} = (abv + c)/\sqrt{(a^2v + A)(b^2v + B)}$. If either mechanism is affine, $c = 0$, so the score difference is $\frac{M}{2}\log(1-\rho_{XY}^2) \le 0$.

For existence, take $H \sim N(0, 0.6^2)$. Let $o(h)$ be $\tanh h$ divided by its standard deviation, and $e(h)$ be $\tanh(h+1) + \tanh(1-h)$, centered and divided by its standard deviation under this law. Odd/even symmetry gives $\mathbb{E}o = \mathbb{E}e = \mathbb{E}[oe] = \mathbb{E}[He] = 0$, $\mathbb{E}o^2 = \mathbb{E}e^2 = 1$ and $t := \mathbb{E}[Ho] > 0$. Set $f = (o+e)/\sqrt{2}$, $g = (o-e)/\sqrt{2}$ and add independent Gaussian noises of standard deviation 0.6. Both functions use four $\tanh$ units, including a constant unit for centering; the other output weights can be zero. Hence they belong to the 16-unit function class, although this exact cancellation is not a typical random draw. Now $\operatorname{Cov}(X, Y) = 0$ and $\operatorname{Cov}(H, X) = \operatorname{Cov}(H, Y) = t/\sqrt{2} > 0$. The partial covariance is $-t^2/(2\operatorname{Var} H) \neq 0$.

The Gaussian law with this covariance has only the independence $X \perp Y$: the two nonzero hub correlations make every other partial correlation nonzero. Thus it is faithful to the collider. The collider and complete DAGs attain its population log-likelihood, but complete DAGs pay an extra edge penalty. Each of the finitely many other classes has a positive likelihood deficit, growing linearly in M and eventually exceeding any edge-count advantage of order $\log M$. □

D.2 A CONTROLLED BOUNDARY OF THE SECOND-MOMENT EXPLANATION

The construction above also separates the true-density and Gaussian criteria. The nonconstant analytic $f(H), g(H)$ have absolutely continuous marginals but lie jointly on a smooth curve, so are dependent. Independent Gaussian noise cannot make them independent: its nonzero characteristic functions cancel in the independence identity. Thus $I(X;Y) > 0$, while $X \perp Y \mid H$. Bounded signals and positive Gaussian noise ensure finite conditional entropies. The true-density collider-minus-fork score per row is $\log[p(X)p(Y)/p(X,Y)]$, with expectation $-I(X;Y) < 0$, opposite to the Gaussian preference.

The pilot fixes this fork, its covariance-matched Gaussian collider, a linear fork with signals $H/0.6$, and a nearby fork with signals $\sqrt{0.575}\,o \pm \sqrt{0.425}\,e$ (sibling covariance 0.15). Three independent contexts per condition and $n \in \{500, 1000\}$ give 24 full 12-cell tables, using TabPFN with two ensemble members and equal context/evaluation halves. Seeds and contrasts were fixed before queries; a timing-only rule admitted all tables. Projection uses 200 shared row resamples. The fixed contrast subtracts the mean score of the three fork orientations from the collider score; their edge penalties cancel. Global search uses $\lambda = \frac{1}{2}\log M$ and is checked against all 25 DAGs. It recovers truth on $24/24$ projected and $23/24$ raw tables; the raw exception adds an edge in a Gaussian control. True subset densities use Gaussian quadrature: every evaluation row agrees within 10^{-7} in log density between orders 1024 and 2048; independent adaptive integration agrees within the same tolerance on the checked support.

Per-context contrasts and row standard errors hold fitted weights fixed. These descriptive checks do not establish context stability, scaling/relabeling invariance or an internal TabPFN mechanism.

E DIRECTION CHECKS AND IMPLEMENTATION

Construction and guarantee. For the block statistic in Section 3.4, t_9 is exact only for independently fixed triples and i.i.d. nondegenerate Gaussian block means under a zero-mean null; estimated triples invalidate this argument. A rejected statistic proposes the direction of its sign. BH and BY use $q = 0.05$; the held-out check uses $B = 1999$ permutations.

For the held-out check, parent domains, predictors and preprocessing are fixed without the test rows. The other half of evaluation rows selects the class, conditioning sets and fitted projection weights.

Each test row is queried separately: neither TabPFN nor score-PFN allows query rows to attend to other query rows. For each direction, a kernel permutation independence test compares $u_y = \hat{F}(y \mid x, S)$ with (x, S). Under the directional null and conditional on context and selection, the test rows are exchangeable and independent PIT/covariate pairs remain equal in law after permuting one side. For B independent uniform permutations, use $p = (1 + \sum_{b=1}^{B} \mathbf{1}\{T_b \geq T_{\text{obs}}\})/(B+1)$. With N test rows, $T = N^{-2} \operatorname{tr}(K_u H K_z H)$ and $H = I - \mathbf{1}\mathbf{1}^\top/N$. Both kernels are $\exp(-\|a-b\|^2/(2m))$, with m the median positive squared pairwise distance in that input (or 1 if none); covariates are the input data coordinates. Large T indicates dependence. The permutation-rank p-value is super-uniform; a union bound gives the same property to $\min(1, 2\min(p_{\rightarrow}, p_{\leftarrow}))$ when both nulls hold. Benjamini–Yekutieli controls FDR across selected edge nulls under arbitrary dependence (Benjamini & Yekutieli, 2001), even for an incorrect estimated class. This is not wrong-arrow FDR: a valid predictor can satisfy both directional nulls, and tied directional p-values yield no arrow even if the edge null is rejected.

Graph postprocessing. The initial postprocessing applies signs by decreasing $|T_K|$, avoiding cycles and new colliders before Meek propagation. This need not admit a DAG extension in the selected class; enforcing one changes none of 34 recomputed outputs. Arrow errors are measured before postprocessing.