PREQUENTIAL EVIDENCE PRUNING: INFORMATION-THEORETIC EDGE SELECTION FOR ORDERING-BASED CAUSAL DISCOVERY

Anonymous authors
Paper under double-blind review

ABSTRACT

Ordering-based causal discovery reduces structure learning to parent selection under a candidate order, yet its pruning stage remains the primary bottleneck: widely used procedures rely on marginal, additivity-constrained tests and tuned thresholds, which fail to capture non-additive interactions and compromise reproducibility. We introduce *Prequential Evidence Pruning* (PEP), a framework that reframes pruning as a local cost-benefit analysis grounded in information theory. For each candidate edge, PEP computes a prequential (out-of-fold) log-evidence gain by evaluating the child's predictive density in the context of its current co-parents, and retains the edge only when this gain exceeds a computed Minimum Description Length (MDL) code-length penalty that adapts to sample size, the number of admissible parents, and the set size. Theoretically, the population target of the evidence gain equals conditional mutual information (CMI); the statistic is stable under bounded log-loss regret of the predictive component; and prequential scoring yields finite-sample concentration. Empirically, instantiating PEP with a pre-trained tabular model that provides calibrated, zero-shot predictive densities yields consistent improvements across diverse ordering backbones and datasets, including stress tests under misspecification. PEP thus replaces fragile heuristics with a principled, auditable rule, elevating the pruning stage of ordering-based discovery from marginal testing to context-aware evidence maximization.

1 INTRODUCTION

Causal discovery from observational data is fundamental to mechanistic understanding across science and engineering (Sachs et al., 2005; Van Koten & Gray, 2006; Hicks et al., 1980), yet exhaustive search over directed acyclic graphs (DAGs) is super-exponential and therefore intractable without strong inductive biases (Bongers et al., 2021). Ordering-based methods address this computational challenge by first estimating a topological order and then pruning forward edges (Teyssier & Koller, 2012; Bühlmann et al., 2014; Peters et al., 2014; Rolland et al., 2022; Montagna et al., 2023c;b; Sanchez et al., 2023; Xu et al., 2024). This two-stage paradigm has seen significant advances in the ordering step. In contrast, the pruning step remains the practical bottleneck: widely used Causal Additive Model (CAM) (Bühlmann et al., 2014) pruning evaluates each candidate parent *marginally* under additivity constraints and makes pruning decisions via fixed thresholds, which can obscure non-additive interactions among co-parents and induce unstable behavior across datasets. We illustrate this core challenge, which motivates our work, in Figure 1.

We propose *Prequential Evidence Pruning (PEP)*, a principled framework that reframes pruning as a localized cost–benefit analysis grounded in information theory. For a candidate edge $i \to j$ evaluated with its current co-parents $S \setminus \{i\}$, PEP quantifies a *prequential* (out-of-fold) log-evidence gain—the improvement in predictive log-likelihood of the child when conditioning on X_i in addition to $X_{S\setminus\{i\}}$. Computing evidence strictly out of sample mitigates in-sample optimism and underpins finite-sample stability. The decision rule is MDL-based: the edge is retained only when the data-compression benefit exceeds a *computed* code-length penalty that accounts for the identity of the added parent, the change in set size, and a fixed overhead (Grünwald, 2007). This transforms a sequence of ad-hoc tests into a single, auditable principle that preserves the search efficiency of ordering while directly addressing the pruning failure modes that constrain current pipelines.

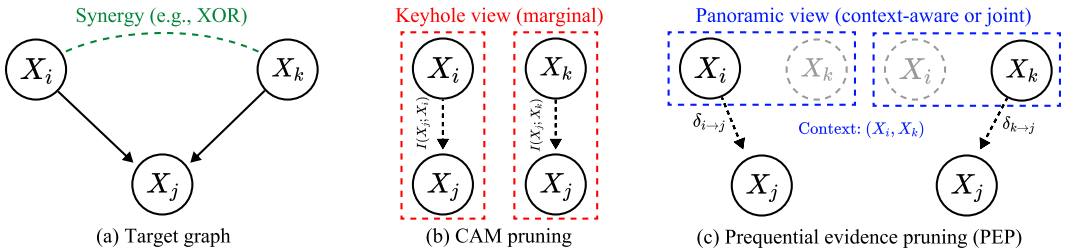

Figure 1: A conceptual illustration of our pruning framework. (a) The target graph depicts parents X_i and X_k having a synergistic effect on their child X_j. (b) In contrast, CAM pruning adopts a *keyhole view*, evaluating each parent in isolation. This approach fails to capture synergies when the marginal signal is null (e.g., $I(X_j; X_i) \approx 0$). (c) Our PEP framework addresses this limitation by adopting a *panoramic view*, which evaluates each parent (X_i) in the context of its co-parents (X_k) to compute an evidence gain ($\delta_{i \to j}$) that captures the interaction. For mathematical examples, see Appendix D.

To convert evidence into a decision, PEP compares $\delta_{i \to j}(q; S)$ against a computable Minimum Description Length (MDL) (Grünwald, 2007) penalty that prices the *order-aware combinatorics* of adding one parent. The per-sample gate $\tau_j^{\mathrm{MDL}}(S, i)$ encodes the identity of the added parent among the admissible predecessors and the change in set cardinality (Eq. (3)–Eq. (4)), yielding an explicit, sample-size aware acceptance threshold, obviating the need for a user-tuned significance level. This design penalizes search-space complexity rather than parametric dimension and is therefore compatible with amortized or nonparametric predictive components. Our framework is model-class agnostic and requires only a predictive component that outputs proper, calibrated conditional densities. In experiments we instantiate this component with a single pre-trained tabular foundation model (Hollmann et al., 2025b) that provides zero-shot, well-calibrated predictive densities for mixed data types, allowing the empirical study to focus on the contribution of the principle rather than on model-specific engineering.

Contributions. (1) A prequential, context-aware edge statistic is introduced, which measures the out-of-sample predictive gain of a parent conditioned on its co-parents to capture synergistic and non-additive interactions. (2) A decision gate based on the MDL principle is developed, replacing user-tuned significance thresholds with a computed, adaptive penalty that enhances the robustness of pruning decisions. (3) A modular, plug-in pruning framework (PEP) is presented, which improves diverse ordering-based backbones by directly addressing their pruning shortcomings. (4) We provide theoretical guarantees for stability and extensive experiments on synthetic and real-world data, demonstrating that our framework offers significant improvements over state-of-the-art baselines.

2 Related Work

Ordering-based Causal Discovery. Ordering-based approaches circumvent the super-exponential DAG search by first estimating a topological order and then pruning edges consistent with that order. Early work such as CAM (Bühlmann et al., 2014) and RESIT (Peters et al., 2014) pioneered this two-stage paradigm. A recent line, initiated by SCORE (Rolland et al., 2022), identifies leaves via properties of the score function and has given rise to several effective variants, including NoGAM (Montagna et al., 2023c), DAS (Montagna et al., 2023b), DiffAN (Sanchez et al., 2023), and CaPS (Xu et al., 2024). Despite this progress on the ordering step, most pipelines still employ CAM-style, additivity-constrained post-processing for pruning, which evaluates candidates marginally and fails to account for synergistic (non-additive) interactions among parents. We address this under-explored bottleneck: our PEP module performs joint, context-aware evaluation via a prequential log-evidence gain and utilizes a computed MDL penalty in place of tuned thresholds, integrating with diverse ordering backbones without changing their ordering criteria. See Appendix E for additional related work in causal discovery.

Information-Theoretic Approaches in Causal Discovery. Information theory has been foundational to causal discovery along two primary lines. Constraint-based procedures (e.g., PC (Spirtes & Glymour, 1991)) rely on statistical tests for conditional independence, using estimators of conditional

mutual information (CMI) with user-specified significance levels. In contrast, score-based methods (e.g., GES (Chickering, 2002)) optimize a global objective that balances model fit and complexity, often with an MDL-derived penalty like BIC. Our framework, PEP, synthesizes these two traditions: it uses an information-theoretic evidence statistic (the prequential log-evidence gain) to quantify dependence in context, and compares this against a computed MDL code-length penalty to make local edge decisions. This approach retains the semantic appeal of CMI while inheriting MDL's parsimony, yet avoids tuned thresholds and global parametric assumptions, remaining applicable with nonparametric or amortized predictors (see §3 for definitions and guarantees).

Positioning relative to prior paradigms. Constraint-based pipelines adjudicate edges via hypothesis tests for (surrogates of) conditional mutual information with user-chosen significance levels, whereas global score-based pipelines optimize in-sample objectives with parametric penalties, and practical post-ordering modules often employ cross-validated gains with tuned thresholds. PEP differs along three axes: (i) *evidence semantics* via a prequential, context-aware edge score that targets CMI at the oracle and concentrates under cross-fitting; (ii) *gate construction* via a computable MDL penalty that prices the *order-restricted combinatorics* of adding one parent rather than parametric dimension; and (iii) *scope* in its applicability to amortized or nonparametric predictors without global likelihood optimization. A broader discussion of related paradigms, including continuous optimizations and Bayesian structure learning, is provided in Appendix E.

3 The Prequential Evidence Pruning (PEP) Framework

We consider i.i.d. observations $X = (X_1, \dots, X_d) \sim p$ that are Markov to an unknown DAG $G^\star$. Given a topological order π, the pruning problem is formulated as determining, for each node j, which forward candidates in $\mathrm{Pred}_\pi(j)$ to include. PEP addresses this decision locally by combining a *prequential, context-aware* evidence statistic with a *computed* Minimum Description Length (MDL) gate, while preserving the computational advantages of ordering-based pipelines.

Prequential (cross-fitted) scoring. We partition $\{1, \dots, n\}$ into K folds $\{I_k\}_{k=1}^K$. For any held-out index $s \in I_k$, the $\log q_{j,S}(x_j^{(s)} \mid x_S^{(s)})$ is evaluated using a predictor trained only on I_k^c. This out-of-sample evaluation mitigates in-sample optimism and, *conditional on the fitted predictors*, renders per-sample contributions independent across s, a property that underpins the concentration results below.

3.1 Definition: the prequential log-evidence gain

For an edge $i \to j$ evaluated in context $S \subseteq \mathrm{Pred}_\pi(j)$ with $i \in S$, define the per-sample evidence

$$\delta_{i\to j}(q; S) = \frac{1}{n} \sum_{s=1}^{n} \Big\{ \log q_{j,S}\big(x_j^{(s)} \mid x_S^{(s)}\big) - \log q_{j,S\setminus\{i\}}\big(x_j^{(s)} \mid x_{S\setminus\{i\}}^{(s)}\big) \Big\}. \tag{1}$$

The statistic $\delta_{i\to j}$ quantifies the improvement in predictive log-likelihood (in *nats per sample*) resulting from the inclusion of X_i among X_j's parents conditioned on the other co-parents $S \setminus \{i\}$, thereby preserving non-additive interactions that marginal tests fail to capture.

3.2 Theoretical guarantees

We work under the following standing assumptions.

Assumption 1 (Data and regularity)**.** *(i)* $x^{(1)}, \dots, x^{(n)} \overset{\text{i.i.d.}}{\sim} p$*. (ii) For all* $S \subseteq \mathrm{Pred}_\pi(j)$*, the true conditional* $p(x_j \mid x_S)$ *and the predictor* $q_{j,S}(x_j \mid x_S)$ *have finite log-loss and variance. (iii) All likelihood terms are evaluated prequentially (out-of-fold). Unless stated otherwise, all logarithms are natural and code lengths are in nats.*

Theorem 1 (Population target equals CMI)**.** *With an ideal predictor* $q = p$*,*

$$\mathbb{E}\big[\delta_{i\to j}(p; S)\big] \;=\; I\big(X_j; X_i \,\big|\, X_{S\setminus\{i\}}\big)\,.$$

Proof sketch. Taking expectations in Eq. (1) with $q = p$ yields $-H(X_j \mid X_S) + H(X_j \mid X_{S\setminus\{i\}}) = I(X_j; X_i \mid X_{S\setminus\{i\}})$ by the chain rule. Full details are given in Appendix F.1. □

The statistic remains well-behaved with imperfect predictors; its deviation from the oracle target is controlled by the conditional log-loss regrets of the competing families.

Proposition 1 (Stability under log-loss regret). *Let* $r_S = \mathbb{E}[\log p(X_j \mid X_S) - \log q_{j,S}(X_j \mid X_S)] \geq 0$ *and define* $r_{S\setminus\{i\}}$ *analogously. Then*

$$\big|\mathbb{E}[\delta_{i\to j}(q;S)] - \mathbb{E}[\delta_{i\to j}(p;S)]\big| \;\leq\; r_S + r_{S\setminus\{i\}}.$$

Proof sketch. Add and subtract the oracle terms and rearrange; see Appendix F.2 for a Bregman-divergence formulation. □

To control finite-sample fluctuations, define per-sample differences $Z_s = \log q_{j,S}(X_j^{(s)} \mid X_S^{(s)}) - \log q_{j,S\setminus\{i\}}(X_j^{(s)} \mid X_{S\setminus\{i\}}^{(s)})$ and assume sub-exponential tails uniformly in s.

Theorem 2 (Concentration under prequential scoring). *Assume* $\{Z_s\}$ *are sub-exponential with parameters* (ν, b) *and are computed prequentially. Then, for any* $t > 0$*,*

$$\Pr\Big(\big|\delta_{i\to j}(q;S) - \mathbb{E}[\delta_{i\to j}(q;S)]\big| \geq t\Big) \;\leq\; 2\exp\Big(-c\,n\,\min\{t^2/\nu^2,\; t/b\}\Big),$$

for an absolute constant $c > 0$*.*

Proof sketch. Conditional on the fitted predictors, $\{Z_s\}$ are independent across s by cross-fitting; apply Bernstein's inequality and de-condition by the tower property. See Appendix F.3 for details and a uniform-over-edges extension. □

Moreover, if the sub-exponential parameters hold uniformly over forward candidates, a union bound yields a uniform tail bound over the edge set (Appendix E.3). This result has two immediate practical implications. First, in the absence of a contextual signal the statistic concentrates near zero.

Corollary 1 (Null behavior). *If* $X_j \perp X_i \mid X_{S\setminus\{i\}}$ *and the regrets are small, then* $\delta_{i\to j}(q;S)$ *concentrates near* 0 *at the rate in Theorem 2.*

Proof sketch. Combine Theorem 1 (oracle target $= 0$ under conditional independence), Proposition 1 (bias bound), and Theorem 2. □

Second, the decision rule provides finite-sample control when expected evidence separates true and false edges by a margin.

Corollary 2 (Finite-sample decision under a margin). *Fix node* j *and contexts* $\{S_{ij}\}$ *for candidates* $i \in \mathrm{Pred}_\pi(j)$*. Suppose there exists* $\gamma > 0$ *such that* $\mathbb{E}[\delta_{i\to j}(q;S_{ij})] \geq \tau_j^{\mathrm{MDL}}(S_{ij}, i) + \gamma$ *for all true parents and* $\mathbb{E}[\delta_{i\to j}(q;S_{ij})] \leq \tau_j^{\mathrm{MDL}}(S_{ij}, i) - \gamma$ *for all non-parents. If the sub-exponential condition holds uniformly with parameters* (ν, b)*, then the probability of any inclusion/exclusion error at node* j *is at most* $2P_j \exp(-c\,n\,\min\{\gamma^2/\nu^2, \gamma/b\})$*.*

Proof sketch. Apply Theorem 2 to each candidate and take a union bound; see Appendix F.4. □

3.3 From evidence to decision: the MDL gate

An edge is retained only when the data-compression gain, measured in nats, exceeds the code-length cost of describing it. Using the coding theorem, code length in nats approximates negative log-likelihood, so the acceptance condition is

$$\underbrace{\sum_{s=1}^{n}\Big[\log q_{j,S}(\cdot) - \log q_{j,S\setminus\{i\}}(\cdot)\Big]}_{\text{data-compression gain} = n\,\delta_{i\to j}} \;>\; \underbrace{L(M_{j,S}) - L(M_{j,S\setminus\{i\}})}_{\text{model-complexity cost}}. \tag{2}$$

For $k = |S \setminus \{i\}|$ and $P_j = |\mathrm{Pred}_\pi(j)|$, a transparent two-part code yields the per-sample gate.

$$\tau_j^{\mathrm{MDL}}(S,i) = \frac{1}{n}\Big\{\log(P_j - k) + \lambda\log(k{+}1) + \kappa\Big\}, \tag{3}$$

Algorithm 1 Prequential Evidence Pruning (order π given)

Input: data D; order π; predictive component q; folds $\{I_k\}_{k=1}^{K}$.
Initialize: For each node j, set $S_j \leftarrow \mathrm{Pred}_\pi(j)$.
for nodes j in topological order π **do**
  for each $i \in S_j$ (in any fixed order) **do**
    Compute the prequential log-likelihoods and the resulting gain $\delta_{i\to j}$, Eq. (1).
    Compute $\tau \leftarrow \tau_j^{\mathrm{MDL}}(S_j, i)$, Eq. (3).
    if $\delta_{i\to j} \le \tau$ **then**
      Remove edge (i, j) and set $S_j \leftarrow S_j \setminus \{i\}$.
    end if
  end for
end for
Output: pruned DAG $\widehat{G}$.

where the first term prices the identity of the added parent among $P_j - k$ admissible candidates and the second encodes the new set size with a universal integer code ($\lambda \in [0, 1]$), in addition to a fixed overhead κ. This penalty quantifies the *order-aware search combinatorics* and is agnostic to the parametric dimension of the predictive component, whose statistical complexity is handled prequentially. For a classical calibration under regular parametric conditions, see Appendix F.5. For a tabular contrast with CI tests and BIC, see Appendix E.

The PEP Decision Rule. We replace ad-hoc threshold tuning with a computable penalty that adapts to sample size (n), search space size (P_j), and current model complexity (k). Concretely, we keep an edge when the prequential gain exceeds the MDL gate:

$$\text{Keep edge } i \to j \qquad \Longleftrightarrow \qquad \delta_{i\to j}(q; S) \;>\; \tau_j^{\mathrm{MDL}}(S, i). \tag{4}$$

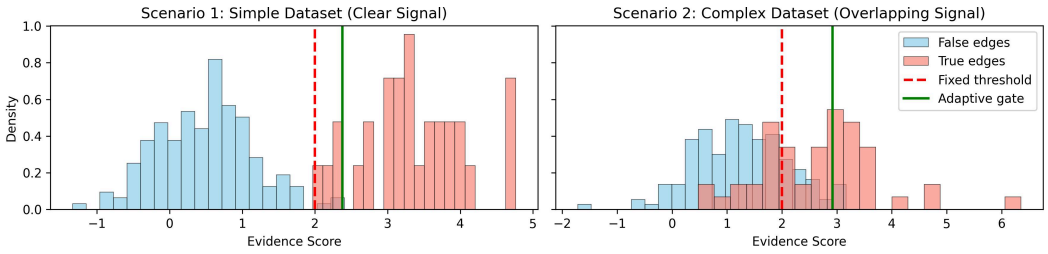

Figure 2: Fixed versus adaptive gates (schematic). Left: when the distributions of $\delta_{i\to j}$ for true and false edges are well separated, both a fixed threshold and the MDL gate succeed. Right: when the distributions overlap, a fixed threshold erroneously includes many false edges, whereas the MDL gate $\tau_j^{\mathrm{MDL}}(S, i)$ adapts to (n, P_j, k) and maintains separation without validation tuning.

Adaptive versus Fixed Gates. Figure 2 visualizes the decision rule in two stylized scenarios. When the evidence distributions for true and false edges are well separated (large population margin), both a fixed threshold and the MDL gate succeed. When the distributions overlap (small margin), a fixed threshold yields many false inclusions, whereas the MDL gate adapts to (n, P_j, k) and restores separation. This aligns with our theory: prequential scoring yields concentration (Theorem 2), and under a positive margin, the finite-sample decision error decays exponentially in n (Corollary 2). The panels are schematic; empirical evaluations are reported in Section 4.

The PEP algorithm. Algorithm 1 implements greedy backward pruning. Starting from the fully-connected forward graph consistent with π, it evaluates each candidate via Eq. (1) and removes edges that fail the condition in Eq. (4).

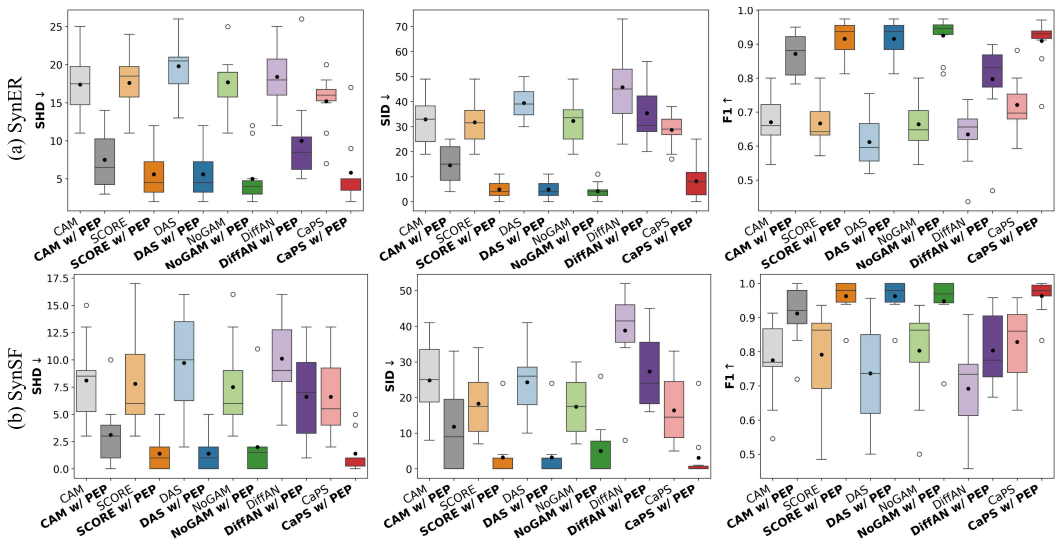

Figure 3: PEP as a plug-and-play enhancement for diverse ordering backbones. Performance comparison of six ordering algorithms using their original pruning modules (e.g., CAM pruning) versus the same backbones augmented with PEP.

4 Experiments

Experimental Setup. We evaluate PEP as a plug-in pruning module for six diverse ordering backbones across a comprehensive suite of benchmarks, encompassing synthetic graphs, challenging misspecification scenarios, and real-world datasets. To ensure statistical robustness, all results are averaged over 10 independent runs and assessed using standard metrics (SHD, SID, F1-score). A detailed description of our experimental protocol, including the specific backbones, dataset configurations, and implementation details, is provided in Appendix G.

4.1 Empirical Validation of the PEP Framework

The empirical validation of the PEP framework is structured as an investigation into four central questions: (i) Does PEP function as a general-purpose, 'plug-and-play' enhancement for diverse ordering backbones? (ii) Are its performance gains attributable to the principled framework itself, or merely to the power of its predictive component? (iii) Do these advantages persist under challenging misspecification scenarios where classical assumptions are violated? And (iv) how robust is PEP to a weak or non-informative ordering? The following analyses address each of these questions directly.

Plug-and-play Improvements Across Ordering Backbones. To validate PEP's utility as a 'plug-and-play' module, we integrated it into six diverse, state-of-the-art ordering backbones, replacing their default pruning mechanisms. The results, shown in Figure 3, demonstrate a clear and consistent pattern of improvement. Across both ER and SF graphs, the PEP-augmented pipelines systematically outperform their original counterparts, leading to substantial reductions in SHD and SID, and marked increases in F1. This finding is significant: it reveals that the performance ceiling of many modern ordering-based methods is not limited by their ordering stage alone but is bottlenecked by their reliance on marginal, additivity-constrained pruning. By evaluating edges in the context of their co-parents, PEP provides a more powerful and general mechanism that unlocks the full potential of these strong ordering algorithms.

Robustness Under Misspecification. To probe the framework's robustness, we subjected it to a stress test across six scenarios where classical causal discovery assumptions are violated (Figure 4). The results reveal a decisive and consistent advantage for PEP. This superiority is most pronounced in the *Post-Nonlinear (PNL)* setting, providing direct empirical validation for our central hypothesis: PEP's context-aware evaluation, which makes no additivity assumption, excels where marginal

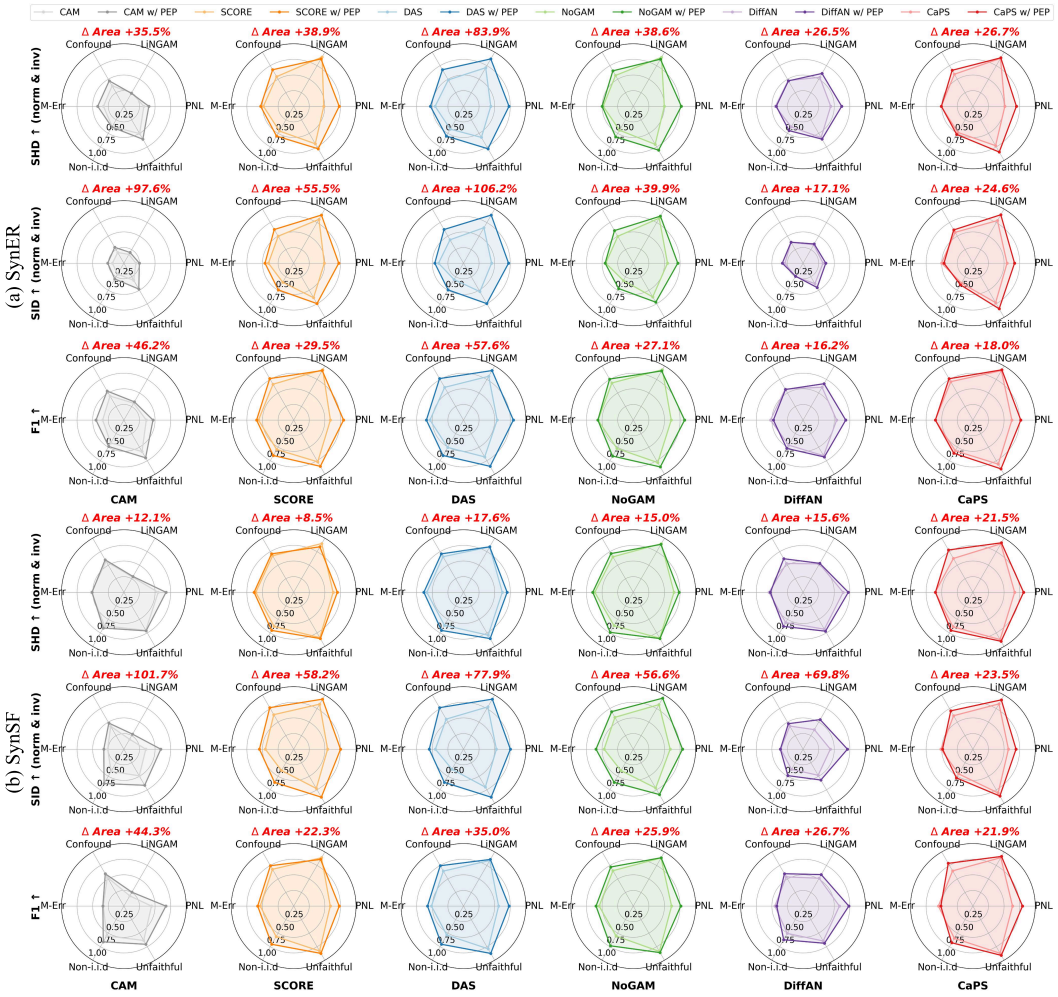

Figure 4: Misspecification stress test. Each polygon summarizes a scenario using three axes: SHD, SID, and F1. SHD and SID are normalized to $[0, 1]$ using bounds determined by the graph structure, then inverted so that higher values indicate better structure and orientation recovery. The polygon area serves as a composite score and the legend reports the *area growth rate* of PEP relative to CAM.

methods falter. More broadly, PEP's consistent performance edge across all scenarios demonstrates the practical benefit of its core design. By replacing specific statistical assumptions with a general, evidence-based principle, the PEP framework is inherently more robust to the kinds of model misspecification frequently encountered in real-world data.

Isolating the Advantage: Framework vs. Predictor. To disentangle the contribution of our framework from that of its specific implementation, we compare four distinct pruning modules: the baseline CAM-pruning, and the PEP framework instantiated with a dataset-trained Random Forest (RF), an XGBoost (XGB), and a pre-trained TabPFN. The results in Figure 5 reveal a highly informative pattern. Simply applying the PEP framework with a standard learner like Random Forest does not guarantee an improvement over CAM, and in some cases underperforms. Using a more powerful learner like XGBoost makes the PEP framework competitive with, and sometimes slightly better than, CAM. However, the most significant and consistent performance gain occurs when the framework is paired with TabPFN. This pattern illuminates the core role of the PEP framework. While a powerful predictor like TabPFN can provide high-fidelity calibrated densities, it is PEP's context-aware decision rule that fully unlocks the potential of this evidence. The results therefore demonstrate that merely replacing the statistical test with a powerful predictor is insufficient; the key

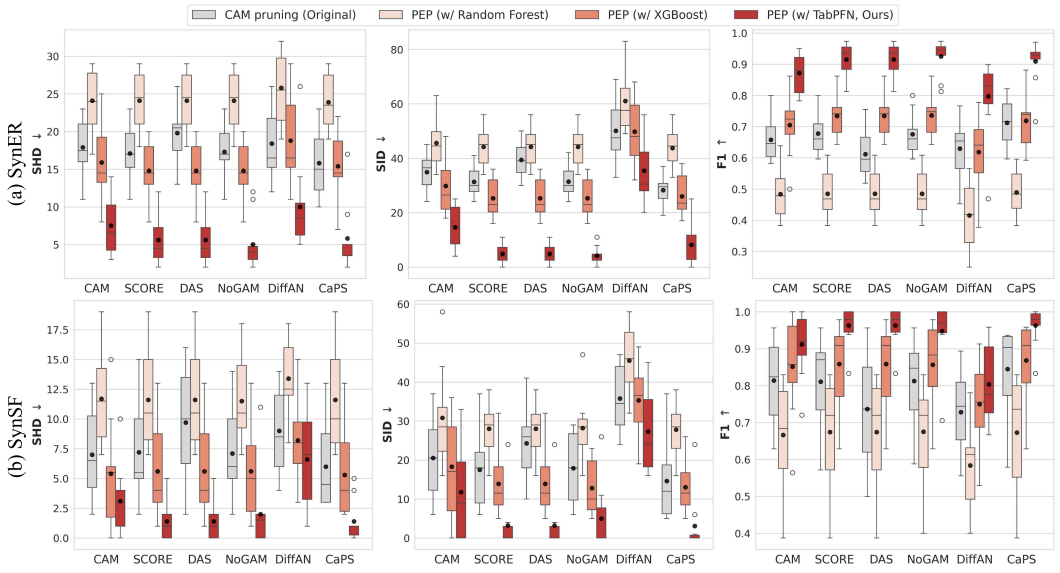

Figure 5: Isolating the source of PEP's advantage. Performance comparison between the baseline CAM pruning and the PEP framework when instantiated with three different predictive components: a Random Forest (RF), XGBoost (XGB), and the pre-trained TabPFN model.

Table 1: Performance on real-world benchmarks. CaPS w/ PEP achieves state-of-the-art performance on the Sachs and SynTReN datasets.

Dataset	Sachs			SynTReN		
Metrics	SHD ↓	SID ↓	F1 ↑	SHD ↓	SID ↓	F1 ↑
CAM	12.0 ± 0.0	55.0 ± 0.0	0.44 ± 0.00	41.3 ± 9.9	170.2 ± 45.2	0.22 ± 0.09
SCORE	12.0 ± 0.0	45.0 ± 0.0	0.44 ± 0.00	38.6 ± 7.0	187.5 ± 58.6	0.205 ± 0.091
DAS	13.0 ± 0.0	48.0 ± 0.0	0.33 ± 0.00	39.4 ± 8.0	168.3 ± 55.4	0.23 ± 0.07
NoGAM	12.0 ± 0.0	45.0 ± 0.0	0.44 ± 0.00	39.2 ± 7.0	184.9 ± 59.9	0.20 ± 0.08
DiffAN	13.0 ± 1.6	50.3 ± 7.6	0.36 ± 0.15	41.4 ± 6.9	196.7 ± 74.7	0.19 ± 0.11
CaPS	11.0 ± 0.0	42.0 ± 0.0	0.50 ± 0.00	37.2 ± 5.3	178.9 ± 58.6	0.23 ± 0.07
CaPS w/ PEP	$\mathbf{11.0 \pm 0.0}$	$\mathbf{42.0 \pm 0.0}$	$\mathbf{0.50 \pm 0.00}$	$\mathbf{33.0 \pm 7.7}$	$\mathbf{164.3 \pm 26.6}$	$\mathbf{0.24 \pm 0.03}$

Table 2: Pruning performance with a non-informative random order. Comparison of PEP and CAM pruning when both are provided with a random topological order.

Dataset	Method	SHD ↓	SID ↓	F1 ↑
SynER	CAM pruning	26.00 ± 4.58	72.40 ± 5.37	0.393 ± 0.110
	PEP	$\mathbf{24.60 \pm 7.37}$	$\mathbf{68.00 \pm 9.25}$	$\mathbf{0.443 \pm 0.176}$
SynSF	CAM pruning	19.00 ± 6.44	59.40 ± 15.29	0.495 ± 0.145
	PEP	$\mathbf{17.60 \pm 7.40}$	$\mathbf{58.80 \pm 15.65}$	$\mathbf{0.503 \pm 0.178}$

advantage lies in the synergy between a component that generates high-quality evidence (TabPFN) and a principled framework (PEP) that can effectively interpret and leverage it.

Performance on Real-World Benchmarks. To validate PEP's practical utility, we integrated it with CaPS, a state-of-the-art ordering backbone, and evaluated the pipeline on the widely-used Sachs and SynTReN benchmarks. The results in Table 1 demonstrate the framework's effectiveness. On the well-established Sachs dataset, where existing methods are highly optimized, the CaPS-PEP pipeline matches the state-of-the-art performance of the original CaPS. This result demonstrates that our principled approach incurs no performance loss on standard benchmarks. Furthermore, on the more complex SynTReN dataset, CaPS-PEP provides a clear improvement in structural accuracy (SHD), showcasing its ability to provide a significant advantage where the pruning task is more challenging. Together, these results confirm that PEP serves as a robust module that performs reliably on established problems and yields demonstrable improvements in more complex, real-world settings.

Effective Pruning Without an Informative Order. To isolate the pruning stage, we repeat the comparison under a *random* topological order so that every forward edge candidate must be vetted without informative ordering cues. PEP remains superior to CAM pruning on both ER and SF (Table 2), confirming that its improvements are not merely inherited from a strong orderer but stem from the local evidence-vs.-complexity decision rule.

Sensitivity to the Pruning Gate. We analyze PEP's sensitivity to the decision gate, demonstrating its robustness to the precise threshold value. Figure 6 reveals two key properties. First, our evidence score δ is a strong ranker of true versus false edges (Right, high ROC/PR AUCs). Second, graph-level

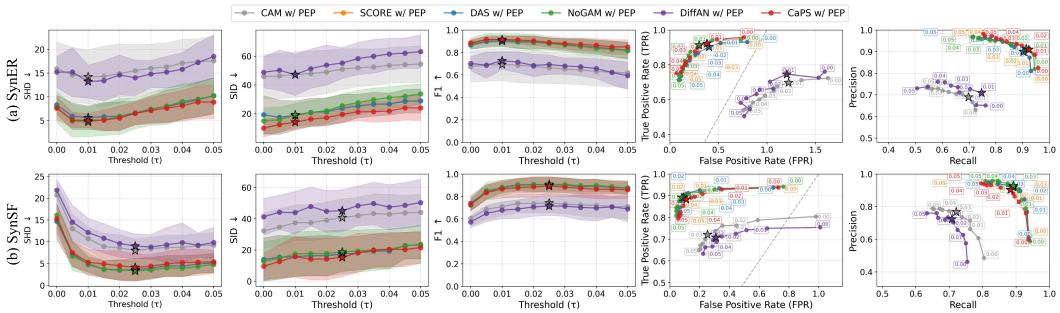

Figure 6: Sensitivity to the pruning gate. (Left) Performance metrics (SHD, SID, F1) plotted against a sweep of the decision threshold value. (Right) Receiver Operating Characteristic (ROC) and Precision-Recall (PR) curves for the evidence score δ.

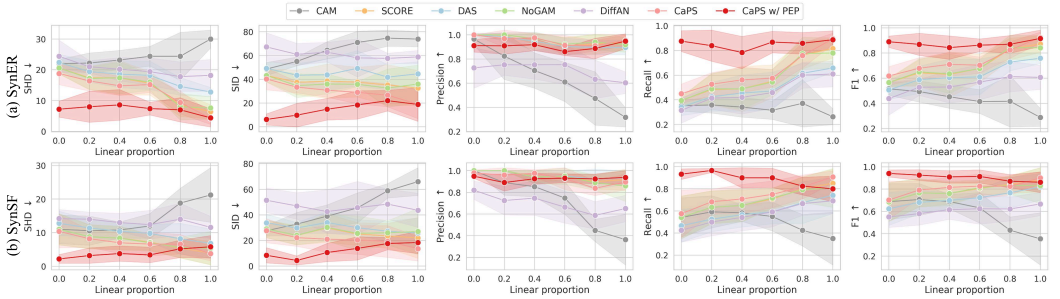

Figure 7: Robustness to functional form. Performance is evaluated as a function of the proportion of linear relationships in the data-generating process. This proportion is varied from 0.0 (fully non-linear) to 1.0 (fully linear).

metrics exhibit a wide, flat plateau, indicating that near-optimal performance persists across a broad range of thresholds, not just a single tuned point (Left). Crucially, the MDL-computed gates for both synthetic datasets (marked by $\star$) lie well within this high-performance plateau. This validates our core design: the combination of a well-calibrated evidence statistic and a principled, adaptive gate yields stable, near-optimal performance without manual threshold tuning. The differing positions of the gates for each dataset further underscore the framework's desirable adaptivity.

Robustness to Functional Form. We validate our framework's robustness to functional form by varying the data's linearity from fully non-linear to fully linear (Figure 7). The results confirm our central hypothesis: PEP's context-aware approach delivers its greatest advantage in challenging non-linear and mixed-linearity regimes where additivity-based methods falter. Critically, it remains highly competitive in the predominantly linear settings where those same methods are theoretically strongest. This demonstrates that PEP is a general-purpose framework that excels in complex scenarios without sacrificing performance in simpler ones.

5 Conclusion

This paper introduced *Prequential Evidence Pruning (PEP)*, a principled framework that reframes the pruning stage of ordering-based causal discovery. We replace marginal tests with a local cost-benefit analysis, where an edge is kept only if its context-aware, prequential log-evidence gain exceeds a computable MDL code-length penalty. Our theoretical analysis grounds this approach, showing the evidence metric targets CMI and is stable in finite samples, while our experiments demonstrate that this principled mechanism consistently improves the performance of diverse state-of-the-art ordering pipelines. By reframing pruning as a transparent trade-off between prequential evidence and model complexity, PEP offers a principled and modular building block for the field. The framework's true potential lies in its general design, opening promising avenues for future work, such as instantiating it with a broader class of calibrated density estimators.

Reproducibility Statement

We summarize steps taken to ensure reproducibility. Datasets and generation procedures are described in Appendix G.1, the compared backbones and their implementations in Appendix G.2, and evaluation metrics in Appendix G.3. Training and evaluation details, including fold splits and global hyperparameters, are provided in Appendix G, and additional experiments are reported in Appendix H.3. We will release the full codebase and scripts for all experiments upon acceptance to ensure end-to-end reproducibility. For the review phase, we submit an anonymized `.zip` archive containing the code as supplementary material.

Ethics Statement

This work focuses on methodological advances in causal discovery and is evaluated on synthetic benchmarks (SynER and SynSF) and widely used public datasets (Sachs and SynTReN). No personally identifiable information or sensitive attributes are used.

References

Vahid Balazadeh, Hamidreza Kamkari, Valentin Thomas, Benson Li, Junwei Ma, Jesse C Cresswell, and Rahul G Krishnan. Causalpfn: Amortized causal effect estimation via in-context learning. *arXiv preprint arXiv:2506.07918*, 2025.

Béla Bollobás, Christian Borgs, Jennifer T Chayes, and Oliver Riordan. Directed scale-free graphs. In *SODA*, volume 3, pp. 132–139, 2003.

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

Peter Bühlmann, Jonas Peters, and Jan Ernest. Cam: Causal additive models, high-dimensional order search and penalized regression. *The Annals of Statistics*, 42:2526—-2556, 2014.

David Maxwell Chickering. Optimal structure identification with greedy search. *Journal of machine learning research*, 3(Nov):507–554, 2002.

Anish Dhir, Ruby Sedgwick, Avinash Kori, Ben Glocker, and Mark van der Wilk. Continuous bayesian model selection for multivariate causal discovery. In *International Conference on Machine Learning*, 2025.

Bao Duong, Sunil Gupta, and Thin Nguyen. Causal discovery via bayesian optimization. In *International Conference on Learning Representations*, 2025.

Paul Erdős and Alfréd Rényi. On the evolution of random graphs. *Publ. Math. Inst. Hungar. Acad. Sci*, 5:17–61, 1960.

Tomas Geffner, Javier Antoran, Adam Foster, Wenbo Gong, Chao Ma, Emre Kiciman, 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. ISSN 2835-8856.

Peter D Grünwald. *The minimum description length principle*. MIT press, 2007.

John Hicks et al. *Causality in economics*. Australian National University Press, 1980.

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*, 637:319–326, 2025a.

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, 2025b.

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

Lars Lorch, Jonas Rothfuss, Bernhard Schölkopf, and Andreas Krause. Dibs: Differentiable bayesian structure learning. *Advances in Neural Information Processing Systems*, 34:24111–24123, 2021.

Yuchen Ma, Dennis Frauen, Emil Javurek, and Stefan Feuerriegel. Foundation models for causal inference via prior-data fitted networks. *arXiv preprint arXiv:2506.10914*, 2025.

Francesco Montagna, Atalanti Mastakouri, Elias Eulig, Nicoletta Noceti, Lorenzo Rosasco, Dominik Janzing, Bryon Aragam, and Francesco Locatello. Assumption violations in causal discovery and the robustness of score matching. *Advances in Neural Information Processing Systems*, 36: 47339–47378, 2023a.

Francesco Montagna, Nicoletta Noceti, Lorenzo Rosasco, Kun Zhang, and Francesco Locatello. Scalable causal discovery with score matching. In *Conference on Causal Learning and Reasoning*, pp. 752–771, 2023b.

Francesco Montagna, Nicoletta Noceti, Lorenzo Rosasco, Kun Zhang, and Francesco Locatello. Causal discovery with score matching on additive models with arbitrary noise. In *Conference on Causal Learning and Reasoning*, pp. 726–751, 2023c.

Samuel Müller, Noah Hollmann, Sebastian Pineda Arango, Josif Grabocka, and Frank Hutter. Transformers can do bayesian inference. In *International Conference on Learning Representations*, 2022.

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 33*, 2020.

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

Jake Robertson, Noah Hollmann, Samuel Müller, Noor Awad, and Frank Hutter. FairPFN: A tabular foundation model for causal fairness. In *International Conference on Machine Learning*. PMLR, 2025a.

Jake Robertson, Arik Reuter, Siyuan Guo, Noah Hollmann, Frank Hutter, and Bernhard Schölkopf. Do-pfn: In-context learning for causal effect estimation. *arXiv preprint arXiv:2506.06039*, 2025b.

Paul Rolland, Volkan Cevher, Matthäus Kleindessner, Chris Russell, Dominik Janzing, Bernhard Schölkopf, and Francesco Locatello. Score matching enables causal discovery of nonlinear additive noise models. In *International Conference on Machine Learning*, pp. 18741–18753. PMLR, 2022.

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

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

Peter Spirtes and Clark Glymour. An algorithm for fast recovery of sparse causal graphs. *Social science computer review*, 9(1):62–72, 1991.

Marc Teyssier and Daphne Koller. Ordering-based search: A simple and effective algorithm for learning bayesian networks. *arXiv preprint arXiv:1207.1429*, 2012.

Tim Van den Bulcke, Koenraad Van Leemput, Bart Naudts, Piet van Remortel, Hongwu Ma, Alain Verschoren, Bart De Moor, and Kathleen Marchal. Syntren: a generator of synthetic gene expression data for design and analysis of structure learning algorithms. *BMC bioinformatics*, 7:43, 2006.

Chikako Van Koten and AR Gray. An application of bayesian network for predicting object-oriented software maintainability. *Information and Software Technology*, 48:59–67, 2006.

Zhuopeng Xu, Yujie Li, Cheng Liu, and Ning Gui. Ordering-based causal discovery for linear and nonlinear relations. *Advances in Neural Information Processing Systems*, 37:4315–4340, 2024.

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 31*, 2018.

Appendices

A LLM USAGE

We used large language models only for fixing grammar and typos. All technical content, including theorems, proofs, algorithms, experiments, and analyses, was authored and verified by the paper's authors.

B NOTATIONS

We summarize symbols used throughout the paper for quick reference. Full definitions are provided in the main text.

Table B.1: Summary of key notations used in the paper.

Symbol	Definition
$\mathcal{E}_\pi$, M	Forward edge set under order π, and its size $M = \lvert\mathcal{E}_\pi\rvert$.
$X = (X_1, \dots, X_d)$, d, n, D	Random vector, dimension (#nodes), sample size, and the dataset.
$G^\star$, G, $\widehat{G}$, π	True DAG, a (candidate) graph, pruned DAG (Alg. 1 output), and a topological order.
$\mathrm{Pa}_G(j)$, $\mathrm{Ch}_G(j)$	Parent set and child set of node j in graph G.
$\mathrm{Pred}_\pi(j)$, P_j	Predecessors of j under order π; $P_j = \lvert\mathrm{Pred}_\pi(j)\rvert$.
S, $S \setminus \{i\} = S'$, k	Working parent set for j, the set after removing i, and $k = \lvert S'\rvert$.
S_j, m_j	Working parent set for node j during pruning; #candidates for j after screening.
$p(\cdot)$, $q_{j,S}(\cdot \mid \cdot)$	True conditional density and predictive conditional density for $X_j \mid X_S$.
$q_{j,S}^{(-k)}$	Out-of-fold predictor for fold k used in prequential scoring.
K, I_k, I_k^c	#folds, index set of fold k, and its complement.
$\delta_{i\to j}(q; S)$	Prequential log-evidence gain for edge $i \to j$ in context S (Eq. (1)).
r_S	Conditional log-loss regret of $q_{j,S}$ relative to $p(\cdot \mid \cdot)$.
Z_s, (ν, b), c	Per-sample log-diff, sub-exponential parameters, and an absolute constant (Thm. 2).
$\tau_j^{\mathrm{MDL}}(S, i)$, λ, κ	MDL gate and its set-size / overhead constants (Eq. (3)).
$L(\cdot)$, $M_{j,S}$	Code length in nats; local model for node j with parent set S (Eq. (2)).
$I(X; Y \mid Z)$	Conditional mutual information.
γ	Margin constant used in finite-sample decision corollaries.
d_S, Δd	Parametric dimension for context S and its difference (used in App. E.5).
$\Delta\mathrm{BIC}$	Difference in the Bayesian Information Criterion.
$\alpha(r)$, $\bar{\alpha}$	Per-sample evaluation cost (as a function of parent-set size) and its average (App. G.1).

C PRELIMINARIES

C.1 CAUSAL ADDITIVE MODELS (CAM)

CAM (Bühlmann et al., 2014) is a two–stage, *ordering-based* approach for learning DAGs under an additive structural equation model (SEM). In this model, each variable is written as $X_j = \sum_{k\in \mathrm{pa}(j)} f_{j,k}(X_k) + \varepsilon_j$ with independent noise, and the learning problem is decomposed into (i) estimating a topological order and (ii) pruning edges given that order. The key design choice in CAM is to separate these tasks: the order is obtained by maximizing (restricted) likelihood under the additive SEM, while sparsity is imposed only in the subsequent pruning step. This decoupling turns structure learning into a tractable combination of permutation search and local variable selection.

Stage 1: Order search. CAM searches over permutations (optionally restricted by a coarse skeleton) and picks the order that maximizes the additive-SEM likelihood; consistency of this maximum-likelihood order estimator is established for both low- and high-dimensional regimes. Intuitively, once the order is known, causal discovery reduces to a set of (potentially nonlinear) regressions of each node on its predecessors.

Stage 2: Pruning / feature selection ("CAM pruning"). Given an estimated order, CAM fits for each node X_j a generalized additive model (GAM) of X_j on its predecessors and then tests, for each candidate parent X_k, the null hypothesis $H_0 : f_{j,k}(\cdot) \equiv 0$. Edges failing to show a statistically significant contribution (at a user-chosen level α) are removed. Conceptually, this pruning acts as a *marginal, additivity-constrained proxy for a conditional-independence (CI) test*: under the additive SEM, $f_{j,k} \equiv 0$ corresponds to "no effect of X_k on X_j given the other covariates" in the GAM regression sense. Because hypotheses are assessed *one parent at a time* within an additive model,

CAM pruning cannot capture purely non-additive synergies (e.g., XOR-type interactions) and may therefore miss edges whose influence manifests only through interactions.

In summary, CAM provides a strong and widely used baseline: an efficient order-search via (restricted) likelihood, followed by GAM-based significance testing that serves as an additivity-constrained CI surrogate for pruning. This precisely delineates the comparison point for our work, where we retain the ordering paradigm but replace marginal, hypothesis-test pruning with a joint, context-aware evidence rule.

Relationship to Conditional Independence (CI) Testing. The pruning mechanism described in § 3.3, often referred to as "CAM pruning," can be understood as a specific and constrained form of a Conditional Independence (CI) test. The null hypothesis tested for each parent, $H_0 : f_{j,k}(\cdot) \equiv 0$, is conceptually equivalent to testing for the conditional independence $X_j \perp\!\!\!\perp X_k \mid \mathrm{Pred}_{\hat{\pi}}(j) \setminus \{k\}$. However, this equivalence holds only under the strong assumption that the relationships are correctly specified by the additive model. Because the test is performed on each parent individually within this additive structure, it is considered a *marginal, additivity-constrained proxy for a CI test*. This is a crucial distinction from general, non-parametric CI tests (like those based on CMI), as CAM pruning will fail to detect non-additive interactions (e.g., synergies) where the marginal contribution of a parent is zero.

C.2 Score-based leaf identification via the score function

Let $s(x) = \nabla_x \log p(x)$ denote the *score function*. Under additive noise models with $X_j = f_j(X_{\mathrm{Pa}(j)}) + \varepsilon_j$ and $\varepsilon_j \perp\!\!\!\perp X_{\mathrm{Pa}(j)}$, the j-th component admits the decomposition

$$s_j(x) \;=\; -\frac{x_j - f_j(x_{\mathrm{Pa}(j)})}{\sigma_j^2} \;+\; \sum_{i \in \mathrm{Ch}(j)} \frac{\partial f_i}{\partial x_j}(x_{\mathrm{Pa}(i)})\, \frac{x_i - f_i(x_{\mathrm{Pa}(i)})}{\sigma_i^2},$$

so that the contribution from children vanishes at leaves. Two practical criteria follow from properties of the score Jacobian:

- **Variance-based (SCORE).** In nonlinear settings, a node X_j is a leaf iff the variance of the j-th diagonal of the score Jacobian is zero: $\mathrm{Var}\big[\partial_{x_j} s_j(X)\big] = 0$.
- **Expectation-based (CaPS).** A leaf can be identified without relying on variance by maximizing a diagonal of the *expected* Jacobian: $j^\star = \arg\max_j \mathrm{diag}(\mathbb{E}\,[\nabla s(X)])$, under identifiability conditions that hold for linear and nonlinear cases.

These criteria yield effective order-estimation subroutines which we later combine with our pruning module.

C.3 Prequential (cross-fitted) scoring

Given data $\{x^{(s)}\}_{s=1}^n$ and a candidate parent set $S \subseteq \mathrm{Pred}_\pi(j)$ for node j, let $q_{j,S}$ be any predictive conditional density for X_j given X_S. A K-fold *prequential* (cross-fitted) evaluation proceeds as follows: partition $\{1, \ldots, n\}$ into folds $\{I_k\}_{k=1}^K$, fit a predictor on the complement I_k^c, and score *only* the held-out fold,

$$\widehat{\ell}_{\text{preq}}(j, S) \;=\; \frac{1}{n} \sum_{k=1}^{K} \sum_{s \in I_k} \log q_{j,S}^{(-k)}\big(x_j^{(s)} \mid x_S^{(s)}\big).$$

Prequential scoring avoids in-sample optimism and ensures that, conditional on fitted predictors, per-sample contributions are independent across s, enabling concentration bounds for edge-wise evidence differences.

C.4 Conditional mutual information (CMI)

For random variables (X, Y, Z) with joint density p, the conditional mutual information (CMI) is

$$I(X; Y \mid Z) \;=\; \mathbb{E}\left[\log \frac{p(X \mid Y, Z)}{p(X \mid Z)}\right] \;=\; H(X \mid Z) - H(X \mid Y, Z).$$

In our setting, the *population* target of the prequential log-evidence gain for edge $i \to j$ in context S equals $I(X_j; X_i \mid X_{S\setminus\{i\}})$ when $q = p$, which justifies interpreting the statistic as a context-aware conditional-dependence measure.

C.5 Minimum Description Length (MDL)

The *Minimum Description Length* (MDL) principle formalizes Occam's razor: the best model is the one that yields the shortest lossless description of the data. Using a two-part code,

$$L(D; M) \;=\; L(M) \;+\; L(D \mid M),$$

where $L(\cdot)$ denotes code length in *nats*. The coding theorem connects code length and probability: for any prefix-free code matched to p, $L(x) \approx -\log p(x)$, so MDL trades off model complexity against fit via (negative) log-likelihood.

A two-part code for local edge additions. In ordering-based pruning we consider local families for X_j with parent sets $S' = S \setminus \{i\}$ and S. Adding one parent incurs a transparent complexity cost with two terms: (i) *Identity cost* $\log(P_j - k)$ to name which of the remaining $P_j - k$ admissible candidates (from $\mathrm{Pred}_\pi(j)$) is added when $|S'| = k$; (ii) *Set-size cost* $\lambda \log(k+1)$ from a universal code for integers ($\lambda \in [0, 1]$); plus a fixed overhead κ. Averaging per sample yields the computable gate used in the main text:

$$\tau_j^{\mathrm{MDL}}(S, i) \;=\; \frac{1}{n}\Big[\log(P_j - k) + \lambda \log(k+1) + \kappa\Big].$$

In regular parametric regimes, adding a $\frac{1}{2}\Delta d \frac{\log n}{n}$ term to the gate recovers a local BIC-style comparison; we provide a separate calibration and proof in the appendix devoted to theoretical results.

C.6 Structural Causal Models (SCMs)

A *Structural Causal Model (SCM)* over X consists of a DAG $G^\star$ and structural assignments

$$X_j \;=\; f_j\big(X_{\mathrm{Pa}_{G^\star}(j)},\, \varepsilon_j\big), \qquad j = 1, \ldots, d,$$

with mutually independent exogenous noises $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_d)$.[1] The induced observational density factorizes as

$$p(x) \;=\; \prod_{j=1}^{d} p\big(x_j \mid x_{\mathrm{Pa}_{G^\star}(j)}\big),$$

which is the global Markov property of the DAG. Interventions $do(X_S = x_S)$ replace the assignments $\{f_j : j \in S\}$ by constants and sever incoming edges into S, enabling interventional semantics via the truncated factorization. Ordering-based discovery exploits the existence of a (possibly estimated) topological order π to constrain candidate parents for X_j to the set $\mathrm{Pred}_\pi(j) = \{i : \pi(i) < \pi(j)\}$ and reduces structure learning to *pruning* spurious edges among these forward links.

C.7 Tabular foundation model (TabPFN) and prior-data fitted networks

Prior-Data Fitted Networks (PFNs) instantiate in-context learning for supervised tasks by training a transformer to approximate the *Bayesian posterior predictive* over a prior of tasks. A PFN receives, at inference, a full dataset context and emits predictive distributions for held-out points in a single forward pass. *TabPFN* specializes this idea to tabular data: it is pre-trained on a very large corpus of synthetic datasets sampled from SCM-driven generators spanning mixed data types and diverse mechanisms. Practically, for any X_j and parent set S it returns a calibrated conditional distribution $q_{j,S}(\cdot \mid x_S)$ from which we compute prequential log-likelihoods. For regression with discretized outputs, we integrate the predictive mass over the bin containing the observed value; for categorical data we use the emitted probabilities directly. This zero-shot, calibrated density estimation is what makes TabPFN a convenient predictive component for our framework, eliminating per-dataset training while supporting mixed types.

[1] Independence of the exogenous noises (causal sufficiency) may be relaxed to allow latent confounding, but we keep the canonical acyclic, causally sufficient case for clarity.

D ILLUSTRATIVE EXAMPLES: WHY CONTEXT-AWARE PRUNING MATTERS

This appendix provides, on concrete mathematical examples, the two claims made in the Introduction and in §3: (i) pruning must be *context-aware* to capture non-additive structure and to avoid confounding, and (ii) PEP's *computed* MDL gate replaces tuned thresholds with an auditable code-length cost. Each example walks through the marginal calculation (what classical pruning would see) and the PEP calculation (the prequential log-evidence gain δ), then states the decision under the MDL rule $\delta > \tau^{\mathrm{MDL}}$ (Eq. (1)–Eq. (4)). These examples mirror the advantages emphasized in the paper's opening sections and experiments.

Notations. All logarithms are natural (nats). For $p \in (0,1)$, $h(p) = -p\log p - (1-p)\log(1-p)$ denotes the binary entropy. We write $S \subseteq \mathrm{Pred}_\pi(j)$ for the co-parents of X_j (including i when testing $i \to j$). At the oracle ($q = p$), $\mathbb{E}[\delta_{i\to j}(p; S)] = I(X_j; X_i \mid X_{S\setminus\{i\}})$ by Theorem 1; bounded log-loss regret perturbs this by at most $r_S + r_{S\setminus\{i\}}$ (Proposition 1); prequential scoring yields concentration (Theorem 2).

D.1 NOISY XOR: A CANONICAL CASE OF DISCRETE SYNERGY

We begin with the classic XOR problem, a canonical example where two parents are only informative when considered together. The data is generated by $X_3 = X_1 \oplus X_2 \oplus N$, where the parents $X_1, X_2 \overset{\text{i.i.d.}}{\sim} \mathrm{Bernoulli}(\frac{1}{2})$ and $N \sim \mathrm{Bernoulli}(\varepsilon)$ is a noise term.

A marginal analysis, which evaluates the link $X_1 \to X_3$ in isolation, would find the variables to be independent, as the influence of the random co-parent X_2 averages out any effect. This leads to a marginal mutual information of exactly zero:

$$I(X_3; X_1) = 0.$$

A single-parent test would therefore fail. In contrast, PEP's context-aware approach conditions on X_2, revealing a clear signal where the oracle evidence gain is strictly positive:

$$\mathbb{E}\big[\delta_{1\to 3}(p; \{1,2\})\big] = I(X_3; X_1 \mid X_2) = \ln 2 - h(\varepsilon) > 0.$$

This demonstrates that while the marginal signal is null, the conditional signal is strong, allowing our proposed method to correctly identify the synergistic relationship.

D.2 MULTIPLICATIVE INTERACTION: A CASE OF CONTINUOUS SYNERGY

To show this principle extends beyond discrete cases, we consider a continuous synergy defined by $X_3 = X_1X_2 + \varepsilon$, where parents $X_1, X_2 \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$ and noise $\varepsilon \sim \mathcal{N}(0, \sigma^2)$. A marginal analysis based on first-order statistics, such as linear regression or covariance, will fail. Because the variables are zero-mean, the marginal covariance is zero:

$$\mathrm{Cov}(X_3, X_1) = 0.$$

A test based on correlation would find no effect. The context-aware approach of PEP, however, targets the CMI by evaluating the full conditional distributions. This is strictly positive and correctly quantifies the information gain from the interaction:

$$\mathbb{E}\big[\delta_{1\to 3}(p; \{1,2\})\big] = I(X_3; X_1 \mid X_2) = \frac{1}{2}\mathbb{E}_{X_2}\left[\log\Big(1 + \frac{X_2^2}{\sigma^2}\Big)\right] > 0.$$

This confirms that our method can identify purely interactive signals that are invisible to common marginal tests, with an evidence gain that appropriately grows as the noise σ^2 decreases.

D.3 CONFOUNDING: A CASE OF AVOIDING SPURIOUS EDGES

Here we verify that context is crucial for avoiding false positives. Consider a common confounder $C \sim \mathcal{N}(0,1)$ generating $X_i = aC + \varepsilon_i$ and $X_j = bC + \varepsilon_j$, with no direct edge between them. A marginal analysis will be fooled by the confounder, as the common cause C induces a non-zero spurious correlation:

$$\mathrm{Cov}(X_i, X_j) = ab\,\mathrm{Var}(C) \neq 0.$$

This would lead a marginal method to incorrectly add a non-existent edge. The context-aware approach of PEP avoids this by including the confounder C in the context set. By d-separation, the variables are conditionally independent, and the oracle evidence is exactly zero:

$$\mathbb{E}\big[\delta_{i\to j}(p;\{i,C\})\big] = I(X_j; X_i \mid C) = 0.$$

This verifies that when the confounder is observed, our mechanism correctly finds zero evidence and prunes the spurious edge.

D.4 Post-Nonlinear Effects: A Case of Robustness to Warping

We next consider a case where a simple relationship is obscured by a non-linear transformation: $X_3 = g(X_1 + X_2 + \varepsilon)$, where g is an invertible, non-linear function. A marginal analysis can be easily fooled. A simple test focused on mean effects might fail because the function g distorts the underlying additive structure. The context-aware approach of PEP is robust to this distortion due to a key property of mutual information: its invariance to invertible transformations. The oracle target for PEP therefore remains strongly positive:

$$I(X_3; X_1 \mid X_2) = I(g(X_1 + X_2 + \varepsilon); X_1 \mid X_2) = I(X_1 + X_2 + \varepsilon; X_1 \mid X_2) > 0.$$

This shows our metric correctly identifies dependencies even when they are obscured by complex transformations.

D.5 Suppressor Effect: A Case of Handling Collinearity

Finally, we examine the classic suppressor effect, which occurs with highly correlated parents ($\rho \approx 1$) in the model $X_3 = \beta_1 X_1 + \beta_2 X_2 + \varepsilon$, where $\beta_1 \approx -\beta_2$. In a marginal analysis, the effects of the two parents nearly cancel, leading to a marginal covariance close to zero:

$$\mathrm{Cov}(X_3, X_1) = \beta_1 + \beta_2 \rho \approx 0.$$

A marginal test would see a weak signal and might incorrectly prune a true parent. The context-aware approach of PEP resolves this by assessing the contribution of X_1 given X_2. The conditional signal remains strong, as captured by the CMI:

$$I(X_3; X_1 \mid X_2) = \tfrac{1}{2}\log\Big(1 + \frac{\beta_1^2(1-\rho^2)}{\sigma^2}\Big) > 0.$$

This demonstrates that our method can identify the true importance of a parent even when its signal is masked by other, highly correlated parents.

D.6 The Finite-Sample Decision Gate

The preceding examples analyzed the oracle CMI, which represents the ideal signal. This final example connects this theory to the practical, finite-sample decision rule that PEP actually implements. A traditional approach might have a strong evidence metric but still rely on a heuristic or tuned threshold. In contrast, PEP provides an auditable acceptance condition. Our concentration guarantees (Thm. 2) establish a probabilistic lower bound on the empirical evidence $\delta_{i\to j}(q; S)$ that we measure from data. PEP's final step is to keep an edge only if this conservatively estimated signal exceeds the computable MDL penalty, τ^{MDL}. This transforms the pruning decision into a transparent and principled trade-off, which can be intuitively summarized as:

$$\underbrace{I(X_j; X_i \mid X_{S\setminus\{i\}})}_{\text{Signal}} - \underbrace{(2\varepsilon_{\text{reg}} + \psi_n(\alpha))}_{\text{Uncertainty}} > \underbrace{\tau_j^{\mathrm{MDL}}(S,i)}_{\text{Complexity Cost}}.$$

This provides a complete, theoretically grounded recipe for making a decision, moving beyond the simple identification of a signal.

E Additional Related Work

Continuous Optimization & Bayesian Approaches. One major paradigm in causal discovery is to cast the problem as a single, continuous optimization problem. This line of work was famously initiated by NOTEARS (Zheng et al., 2018), which introduced a fully differentiable characterization of

Table E.1: Comparison of local edge evaluation mechanisms across constraint-based tests, decomposable BIC scoring, and PEP. All code lengths are in nats.

	Conditional Independence (CI) Tests	**Decomposable BIC Scoring**	**Prequential Evidence Pruning (PEP)**
Core approach	Decide edges by testing $X_i \perp X_j \mid X_S$ with a user-chosen significance level.	Select a graph by maximizing a global decomposable score that trades off in-sample fit and parametric complexity.	Prune edges under a given order by comparing a local prequential evidence gain with a computed code-length penalty.
Evidence score	Test statistic $T(X_i, X_j \mid X_S)$ that estimates or surrogates $I(X_j; X_i \mid X_S)$.	Local in-sample log-likelihood difference under decomposability, $\ell(S) - \ell(S \setminus \{i\})$.	Prequential log-evidence gain $\delta_{i\to j}(q; S) = \frac{1}{n}\sum_{s=1}^{n} \log \frac{q_j(x_j^{(s)} \mid x_S^{(s)})}{q_j(x_j^{(s)} \mid x_{S\setminus\{i\}}^{(s)})}$, with q_j evaluated out-of-fold.
Decision rule	Reject H_0 if p-value $< \alpha$ (per-test or FDR-controlled).	Accept if $\ell(S) - \ell(S \setminus \{i\}) > \frac{1}{2}\Delta d \frac{\log n}{n}$ (parametric penalty).	Accept if $\delta_{i\to j}(q; S) > \tau_j^{\mathrm{MDL}}(S, i)$, where $\tau_j^{\mathrm{MDL}}(S, i) = \frac{1}{n}\{\log(P_j - k) + \lambda \log(k+1) + \kappa\}$.
Representative properties	Nonparametric options available; requires α; test-by-test decisions and multiple-testing control.	Consistent under correct parametric family; global, in-sample objective; decomposable local updates.	Prequential and context-aware; sample-size aware penalty; no threshold tuning; model-class agnostic.

acyclicity, enabling standard gradient-based methods. This foundational idea was extended by subsequent works to handle non-linear relationships using neural networks, such as GraNDAG (Lachapelle et al., 2020) and GOLEM (Ng et al., 2020). Further advancements include DrBO (Duong et al., 2025), which employs sophisticated search strategies like Bayesian optimization, and CGP-CDE (Dhir et al., 2025), which integrates flexible Gaussian Process models. From a more strictly Bayesian perspective, where the goal is to infer a posterior distribution over graphs rather than a single point estimate, methods like DiBS (Lorch et al., 2021) and DECI (Geffner et al., 2024) have been proposed. While powerful, these approaches typically involve complex, model-specific training procedures to learn both the graph and functional parameters.

Prior-Data Fitted Networks for Causality. Prior-Data Fitted Networks (PFNs) (Müller et al., 2022) use large-scale, synthetic pre-training to approximate Bayesian predictive inference via in-context learning. TabPFN (Hollmann et al., 2025b) realizes this idea for tabular data and provides calibrated, zero-shot predictive densities that are valuable when samples are scarce or mechanisms are heterogeneous. Building on this paradigm, several works adapt PFNs to *causal inference* tasks. These include models such as FairPFN (Robertson et al., 2025a) for fairness-aware prediction, Do-PFN (Robertson et al., 2025b) for estimating interventional outcomes without a known graph, CausalPFN (Balazadeh et al., 2025) for treatment-effect estimation with calibrated uncertainty, and the comprehensive CausalFM (Ma et al., 2025) framework, illustrating the promise of PFNs as general-purpose causal tools. *In contrast*, we shift the focus to causal discovery. Rather than building an end-to-end PFN for inference, our contribution is a new framework (PEP). It leverages the PFN as a powerful predictive engine to compute a prequential evidence score, which is then assessed by a principled MDL gate.

F Proofs for Theoretical Guarantees

F.1 Population identity: proof of Theorem 1

Proof. Let $S' = S \setminus \{i\}$. Under $q = p$,

$$\mathbb{E}[\delta_{i\to j}(p; S)] = \mathbb{E}\big[\log p(X_j \mid X_S) - \log p(X_j \mid X_{S'})\big] \tag{5}$$
$$= -H(X_j \mid X_S) + H(X_j \mid X_{S'}) \tag{6}$$
$$= I(X_j; X_i \mid X_{S'}), \tag{7}$$

where the second equality uses the definition of conditional entropy, and the last equality is the standard identity for conditional mutual information. All expectations are finite by Assumption 1(ii). □

F.2 STABILITY: PROOF OF PROPOSITION 1

Proof. Write $S' = S \setminus \{i\}$, $p_S(\cdot) = p(X_j \mid X_S)$, $q_S(\cdot) = q_{j,S}(X_j \mid X_S)$, and similarly for S'. Then

$$\begin{aligned}\mathbb{E}[\delta(q;S)] - \mathbb{E}[\delta(p;S)] &= \mathbb{E}[\log q_S - \log q_{S'}] - \mathbb{E}[\log p_S - \log p_{S'}] \\ &= \underbrace{\mathbb{E}[\log q_S - \log p_S]}_{-r_S} - \underbrace{\mathbb{E}[\log q_{S'} - \log p_{S'}]}_{-r_{S'}} \\ &= -r_S + r_{S'}.\end{aligned}$$

Hence $\big|\mathbb{E}[\delta(q;S)] - \mathbb{E}[\delta(p;S)]\big| \le r_S + r_{S'}$. If $r_S, r_{S'} \le \varepsilon$, the bias is $\le 2\varepsilon$. □

F.3 CONCENTRATION: PROOF OF THEOREM 2

Proof. Let $Z_s = \log q_{j,S}(X_j^{(s)} \mid X_S^{(s)}) - \log q_{j,S'}(X_j^{(s)} \mid X_{S'}^{(s)})$, $S' = S \setminus \{i\}$. Consider K-fold prequential scoring and denote by $\widehat{q}_{j,S}^{(k)}, \widehat{q}_{j,S'}^{(k)}$ the fitted predictors for fold k. Condition on the σ-algebra $\mathcal{F}$ generated by all fitted predictors $\{(\widehat{q}_{j,S}^{(k)}, \widehat{q}_{j,S'}^{(k)})\}_{k=1}^K$. For $s \in I_k$, Z_s is a measurable function of $X^{(s)}$ and $(\widehat{q}_{j,S}^{(k)}, \widehat{q}_{j,S'}^{(k)})$, and by construction $X^{(s)}$ is independent of $\mathcal{F}$; hence $\{Z_s : s \in [n]\}$ are *independent* given $\mathcal{F}$.

Assume the sub-exponential Orlicz ψ_1 norms are uniformly bounded almost surely: $\|Z_s - \mathbb{E}[Z_s \mid \mathcal{F}]\|_{\psi_1} \le c_1\nu$ and $|Z_s - \mathbb{E}[Z_s \mid \mathcal{F}]| \le c_2 b$ a.s. for constants (ν, b).[2] Then conditional Bernstein's inequality gives, for any $t > 0$,

$$\Pr\Big(\Big|\tfrac{1}{n}\sum_{s=1}^n Z_s - \mathbb{E}[Z_s \mid \mathcal{F}]\Big| \ge t \,\Big|\, \mathcal{F}\Big) \le 2\exp\left(-cn\min\left\{\frac{t^2}{\nu^2}, \frac{t}{b}\right\}\right).$$

Taking expectations over $\mathcal{F}$ and using $\mathbb{E}[\,\mathbb{E}[Z_s \mid \mathcal{F}]\,] = \mathbb{E}[Z_s]$ yields the unconditional tail bound with the same exponent. Since $\delta_{i\to j}(q;S) = \frac{1}{n}\sum_s Z_s$, the claim follows. □

Uniform-over-edges extension. Let $\mathcal{E}_\pi = \{(i,j) : i \in \mathrm{Pred}_\pi(j)\}$ be the forward edge set with $|\mathcal{E}_\pi| = M$. If the sub-exponential parameters (ν, b) hold uniformly for all $(i,j) \in \mathcal{E}_\pi$, then by the union bound

$$\Pr\Big(\max_{(i,j)\in\mathcal{E}_\pi} \big|\delta_{i\to j}(q;S_{ij}) - \mathbb{E}\delta_{i\to j}(q;S_{ij})\big| \ge t\Big) \le 2M\exp\left(-cn\min\left\{\frac{t^2}{\nu^2}, \frac{t}{b}\right\}\right),$$

where S_{ij} denotes the current co-parent context used for $(i \to j)$.

F.4 MDL PENALTY DERIVATION AND A FINITE-SAMPLE CONSISTENCY COROLLARY

Two-part code for one-parent augmentation. Let $P_j = |\mathrm{Pred}_\pi(j)|$ and $k = |S \setminus \{i\}|$. The augmentation $S' = S \setminus \{i\} \mapsto S$ requires (i) *identity* of the added parent among the $P_j - k$ remaining candidates, which can be encoded in $\log(P_j - k)$ nats by an optimal prefix code (Kraft inequality), and (ii) *set size* $k+1$, encodable by a universal integer code with length $\lambda\log(k+1)$ (e.g., Rissanen's code up to a constant factor). A constant header κ absorbs fixed per-edge overhead. Dividing by n gives Eq. (3):

$$\tau_j^{\mathrm{MDL}}(S,i) = \frac{1}{n}[\log(P_j - k) + \lambda\log(k+1) + \kappa].$$

This code is *model-class free*: it penalizes the *combinatorics* of adding a parent, rather than the parametric complexity of q (which is handled prequentially).

Corollary 3 (Finite-sample consistency under a margin)**.** *Fix node j and contexts $\{S_{ij}\}$ used to test candidates $i \in \mathrm{Pred}_\pi(j)$. Suppose there exists $\gamma > 0$ such that*

$$\mathbb{E}[\delta_{i\to j}(q;S_{ij})] \ge \tau_j^{\mathrm{MDL}}(S_{ij}, i) + \gamma \quad \textit{for all true parents } i \in \mathrm{pa}(j),$$

[2] A sufficient condition is that the conditional log-densities are uniformly bounded above, and $q_{j,S}, q_{j,S'}$ are bounded away from 0 on the support of p; more generally, it suffices that the conditional MGF exists in a neighborhood of 0.

and

$$\mathbb{E}[\delta_{i\to j}(q; S_{ij})] \le \tau_j^{\mathrm{MDL}}(S_{ij}, i) - \gamma \quad \textit{for all non-parents } i \notin \mathrm{pa}(j).$$

If the sub-exponential condition of Theorem 2 holds uniformly with (ν, b), then

$$\Pr\big(\textit{any decision error at node } j\big) \;\le\; 2P_j \exp\!\left(-cn \min\left\{\frac{\gamma^2}{\nu^2}, \frac{\gamma}{b}\right\}\right),$$

i.e., false inclusions and false exclusions vanish exponentially in n.

Proof. For any candidate i, Theorem 2 implies $\Pr\big(|\delta_{i\to j} - \mathbb{E}\delta_{i\to j}| \ge \gamma\big) \le 2\exp(-cn\min\{\gamma^2/\nu^2, \gamma/b\})$. If $i \in \mathrm{pa}(j)$, then $\delta_{i\to j} > \tau^{\mathrm{MDL}}$ fails only if $\delta_{i\to j} - \mathbb{E}\delta_{i\to j} \le -\gamma$; similarly for $i \notin \mathrm{pa}(j)$. Union bound over at most P_j candidates yields the claim. □

Remark (parametric add-on). If $q_{j,S}$ belongs to a parametric family with d_S free parameters trained by MLE on n samples (not our default prequential use), one could add a BIC-style term $\frac{1}{2}(d_S - d_{S\setminus\{i\}})\frac{\log n}{n}$ to Eq. (3). Our non-parametric default only penalizes combinatorics; training complexity is handled by prequential scoring and does not enter the code length.

F.5 BIC CALIBRATION UNDER REGULAR PARAMETRIC CONDITIONS

This subsection provides a classical calibration of PEP's decision rule under regular parametric assumptions. The result is intended for orientation only. It shows that the prequential evidence gain reduces to the usual in-sample likelihood gain up to $o_p((\log n)/n)$ and that, after adding the familiar $\frac{1}{2}\Delta d\,\frac{\log n}{n}$ term to the gate, the PEP rule recovers a local BIC comparison. The main guarantees of PEP in the paper do not rely on these assumptions and follow instead from the CMI target, regret stability, and prequential concentration.

Lemma 1 (Reduction to BIC under regular parametric conditions)**.** *Fix a node j and a context $S \subseteq \mathrm{Pred}_\pi(j)$ with $i \in S$, and let $S' = S \setminus \{i\}$. Suppose $q_{j,S}$ and $q_{j,S'}$ are correctly specified, regular parametric conditionals with respective dimensions d_S and $d_{S'}$. Assume i.i.d. data, K-fold prequential (cross-fitted) scoring with fixed K, and standard regularity (MLE consistency and asymptotic normality, positive-definite Fisher information, and uniform integrability of log-likelihoods). Then*

$$\delta_{i\to j}(q; S) \;=\; \frac{1}{n}\Big(\log L_j(S) - \log L_j(S')\Big) \;+\; o_p\Big(\tfrac{\log n}{n}\Big), \tag{8}$$

where $\log L_j(\cdot)$ denotes the in-sample maximized log-likelihood for X_j given the indicated parent set. Define the augmented penalty

$$\tau_j^{\mathrm{MDL+BIC}}(S, i) \;:=\; \tau_j^{\mathrm{MDL}}(S, i) \;+\; \frac{1}{2}(d_S - d_{S'})\frac{\log n}{n}, \tag{9}$$

with $\tau_j^{\mathrm{MDL}}(S, i)$ as in Eq. (3)-Eq. (4). Then the PEP decision

$$\delta_{i\to j}(q; S) \;>\; \tau_j^{\mathrm{MDL+BIC}}(S, i) \tag{10}$$

is asymptotically equivalent to the local BIC inequality

$$\underbrace{\Big(\log L_j(S) - \log L_j(S')\Big) - \tfrac{1}{2}(d_S - d_{S'})\log n}_{\Delta\mathrm{BIC}(i\to j; S)} \;>\; \log(P_j - k) + \lambda\log(k+1) + \kappa \;+\; o_p(1), \tag{11}$$

where $k = |S'|$ and $P_j = |\mathrm{Pred}_\pi(j)|$. In particular, if $P_j - k = 1$ and $\lambda = \kappa = 0$, the rule reduces asymptotically to $\Delta\mathrm{BIC}(i\to j; S) > 0$.

Proof. Let M_S and $M_{S'}$ denote the local families for S and S', with parameters $\theta_S \in \mathbb{R}^{d_S}$ and $\theta_{S'} \in \mathbb{R}^{d_{S'}}$. For a single observation $(x_j^{(s)}, x_S^{(s)})$, write $\ell_S(\theta_S; s) = \log p_{\theta_S}(x_j^{(s)} \mid x_S^{(s)})$ and $\ell_S(\theta_S) = \sum_{s=1}^n \ell_S(\theta_S; s)$, with MLE $\hat\theta_S = \arg\max_{\theta_S} \ell_S(\theta_S)$, and analogously for S'.

Step 1 (prequential–in-sample alignment). Let $\{I_k\}_{k=1}^K$ be a fixed K-fold partition with $|I_k| = n_k \asymp n/K$. Denote foldwise MLEs by $\hat\theta_S^{(-k)}$ (trained on the complement of I_k). Standard M-estimation

stability yields $\hat{\theta}_S^{(-k)} - \hat{\theta}_S = O_p(n^{-1})$. A second-order Taylor expansion around $\hat{\theta}_S$, summed over $s \in I_k$ and $k = 1, \ldots, K$, gives

$$\mathrm{Preq}_S \;=\; \sum_{k=1}^{K} \sum_{s \in I_k} \ell_S(\hat{\theta}_S^{(-k)}; s) \;=\; \ell_S(\hat{\theta}_S) + O_p(n^{-1/2}), \quad \frac{1}{n}\mathrm{Preq}_S \;=\; \frac{1}{n}\ell_S(\hat{\theta}_S) + O_p(n^{-3/2}).$$

An identical relation holds for S'.

Step 2 (gain identity). By definition of δ in Eq. (1),

$$\delta_{i \to j}(q; S) \;=\; \frac{1}{n}\big(\mathrm{Preq}_S - \mathrm{Preq}_{S'}\big) \;=\; \frac{1}{n}\big(\ell_S(\hat{\theta}_S) - \ell_{S'}(\hat{\theta}_{S'})\big) + O_p(n^{-3/2}),$$

which equals Eq. (8) since $n^{-3/2} = o\big((\log n)/n\big)$.

Step 3 (equivalence with local BIC). Multiply 10 by n and substitute 8. After rearrangement one obtains 11, which establishes the claim. □

Scope. The calibration above relies on fixed-K cross-fitting stability of MLEs and a second-order expansion; it does not invoke Laplace approximations for marginal likelihoods. It shows that prequential (out-of-fold) gains recover the in-sample BIC regime under regular parametric families. The default operation of PEP, however, remains model-class agnostic and applies beyond this regime, with guarantees derived from its CMI target, regret stability, and concentration.

G Implementation Details

All experiments were conducted on a single NVIDIA 3090 GPU, and all reported results are the average over 10 independent runs using different random seeds for data generation. Within each run, the dataset was partitioned into a training set (used as the context for TabPFN's in-context learning) and a test set, on which the evidence scores were evaluated, adhering to the prequential principle of out-of-sample evaluation. Our approach is designed to be principled and avoid per-dataset tuning. For the PEP framework, we fix the MDL constant $\lambda = 1$ following universal coding principles and use a single global offset $\kappa = 25$ that is calibrated once for the entire study. The TabPFNv2 (Hollmann et al., 2025a) model used is the publicly available, pre-trained checkpoint without any fine-tuning. In contrast, for the XGBoost and Random Forest baselines, hyperparameters were selected for each dataset via 5-fold cross-validation to ensure a strong comparison.

G.1 Benchmark Datasets

To clarify our experimental setup, we designed two distinct settings based on the evaluation goal. For the main performance comparison on the SynER and SynSF datasets, we set the functional forms to be fully non-linear (linear proportion = 0.0). This was to ensure a fair comparison, as many score estimator-based ordering methods (e.g., SCORE, DAS, NoGAM) rely on non-linear assumptions for their identification strategies.

In contrast, for the misspecification scenario tests, the suite of tests already includes a dedicated scenario for purely linear relationships (LiNGAM). Therefore, the baseline (vanilla) case for these specific tests was set to a more general, mixed environment with a linear proportion of 0.5. This dual setup allowed us to effectively validate our method's performance under the most relevant conditions for each experimental goal.

Synthetic Dataset Generation Details. All synthetic datasets used in our experiments were generated following a two-step process. First, a ground truth Directed Acyclic Graph (DAG) was generated using one of two standard models. Second, data was sampled from a Structural Equation Model (SEM) defined by that DAG. The main results reported in our paper are based on a default setting using synthetic data with $d = 10$ nodes, $n = 2000$ samples, and dense graphs with an expected number of edges equal to $4d$. To further validate our framework's performance under different conditions, this appendix presents additional experiments focusing on a more challenging scenario with fewer samples (e.g., $n = 1000$). This allows us to assess the model's effectiveness when less data is available (See, Appendix H.3.

- **Erdös-Rényi (ER) Graphs**: The ER model (Erdős & Rényi, 1960) is a fundamental random graph model that generates a homogeneous graph structure. For a given number of nodes d, each of the $\binom{d}{2}$ possible undirected edges is created with a fixed, uniform probability p. To create a DAG, we first establish a random permutation of the nodes to define a topological order, and then orient the selected edges to follow this order. The resulting graphs are characterized by a degree distribution that approximates a Poisson distribution, meaning most nodes have a similar number of connections. The expected number of edges in the graph is controlled by the probability p.
- **Scale-Free (SF) Graphs**: The SF model (Bollobás et al., 2003) generates graphs with a heterogeneous structure, which are often considered more representative of real-world networks. We use the Barabási-Albert model, which employs a preferential attachment mechanism. The graph is grown iteratively, starting with a small number of nodes. At each step, a new node is added and connected to a fixed number of existing nodes, where the probability of connecting to an existing node is proportional to its current degree (number of connections). This "rich-get-richer" process results in a graph characterized by a power-law degree distribution, with a few highly-connected 'hub' nodes and many nodes with very few connections. Similar to the ER model, the graph is then converted to a DAG by orienting edges according to a random permutation.

Real-World Benchmark Details. To evaluate the performance of our framework in practical scenarios, we utilized two well-established real-world benchmark datasets.

- **Sachs**: The Sachs dataset (Sachs et al., 2005) is a widely-used benchmark in causal discovery, derived from a study of a protein-signaling network in human primary T cells. The dataset consists of $n = 853$ samples, with measurements for $d = 11$ phosphorylated proteins and phospholipids obtained via flow cytometry. The ground truth causal graph, which is a consensus network established from expert biological knowledge and interventional experiments (i.e., perturbing specific proteins and observing the effects on others), contains 20 edges. This dataset is a standard testbed for evaluating an algorithm's ability to recover known biological signaling pathways from observational data.
- **SynTReN**: The Syntren (Synthetic Transcriptional Regulatory Network) dataset (Van den Bulcke et al., 2006) is a pseudo-real-world benchmark designed to mimic the complexities of gene expression data. The underlying network structure is not random but is a sub-network extracted from the well-documented transcriptional regulatory network of yeast (Saccharomyces cerevisiae). The observational data, however, is generated synthetically from this real biological structure using a detailed kinetic model that simulates the dynamics of transcription and translation, including both stochastic noise and measurement error. For our experiments, we use a version with $d = 20$ nodes (genes) and $n = 500$ samples. Syntren is considered a particularly challenging benchmark as it combines a realistic, non-random graph structure with a noisy, complex data generation process.

Misspecified Scenario Details. To rigorously evaluate the robustness of our framework, we generated synthetic datasets under six challenging misspecified scenarios. These scenarios are designed to systematically violate the core assumptions upon which many causal discovery algorithms are built, following the methodology of recent benchmarks (Montagna et al., 2023a). We followed the setup of recent benchmarks (Montagna et al., 2023a), with specific parameters set as follows: a confounder probability of $\rho = 0.2$, a signal-to-noise ratio of $\gamma = 0.8$ for measurement error, a 30% probability of unfaithful distributions ($p_{\text{unfaithful}} = 0.3$), and an exponent of 3.0 for the post-nonlinear transformation.

- **Latent Confounders**: This scenario violates the causal sufficiency assumption, which states that there are no unobserved common causes. For a randomly selected subset of variable pairs (X_i, X_j) that do not have an edge between them, we introduce a latent confounder C. The data generating process for these variables is modified to $X_i = f_i(pa(i) \cup \{C\}) + \epsilon_i$ and $X_j = f_j(pa(j) \cup \{C\}) + \epsilon_j$. This induces a spurious correlation between X_i and X_j, testing an algorithm's ability to avoid inferring a non-existent direct causal link.
- **Measurement Error**: This scenario violates the assumption that variables are measured without error. We first generate the true data X according to the SEM. The observed data,

$\tilde{X}$, is then generated by adding independent Gaussian noise to the true values:

$$\tilde{X}_i := X_i + \eta_i, \quad \text{where } \eta_i \sim \mathcal{N}(0, \sigma_\eta^2).$$

The severity of the error is controlled by the signal-to-noise ratio. This scenario tests an algorithm's resilience to corrupted input data.

- **Unfaithful Distributions**: This scenario violates the faithfulness assumption, which states that all conditional independencies in the data correspond to d-separations in the causal graph. We create an unfaithful distribution by selecting a path of length two, $X_i \to X_j \to X_k$, and adding a direct edge $X_i \to X_k$. The causal mechanism for X_k is then carefully parameterized such that the causal effects along the two paths cancel each other out, resulting in a marginal independence between X_i and X_k ($X_i \perp\!\!\!\perp X_k$). This tests an algorithm's ability to recover true edges even when their statistical signal is masked.
- **Autoregressive Model (Non-i.i.d. Data)**: This scenario violates the assumption that data samples are independent and identically distributed (i.i.d.). We introduce a temporal dependency by generating data from an autoregressive model of order 1 (AR(1)). Each sample $x^{(s)}$ is generated based on the previous sample $x^{(s-1)}$:

  $$x^{(s)} = Ax^{(s-1)} + \epsilon^{(s)},$$

  where A is the adjacency matrix of the causal graph. This tests an algorithm's robustness to temporal correlations in the data.
- **Post-Nonlinear (PNL) Models**: This scenario presents a severe violation of the additivity assumption used by many methods. A PNL model introduces a final, non-linear distortion g_j applied to the entire causal mechanism for each variable X_j:

  $$X_j = g_j\left(\sum_{k \in pa(j)} f_{j,k}(X_k) + \epsilon_j\right).$$

  This creates complex, non-additive interactions between all parent variables, providing a strong test of a model's flexibility.
- **Linear Non-Gaussian Acyclic Model (LiNGAM)**: This scenario violates the assumption of Gaussian noise, which is a key requirement for the identifiability of many score-based methods. The data is generated from a purely linear SEM, but the independent noise terms ϵ_j are drawn from a non-Gaussian distribution (e.g., a uniform distribution). This tests an algorithm's dependence on the Gaussian noise assumption for identifying the correct causal direction.

G.2 Baseline Selection

In our experiments, we compared PEP against a comprehensive suite of state-of-the-art ordering-based causal discovery algorithms. While **DAS** was evaluated alongside other methods, we excluded it from the final reported results in the main paper for the sake of clarity and conciseness. The reason for this is that the core ordering mechanism of DAS is identical to that of **SCORE**. Consequently, when a deterministic pruning module like our proposed PEP is applied, the final performance metrics are identical for both backbones. To avoid presenting redundant results in the main paper, we therefore include only SCORE as the representative method for this particular family of score-based algorithms. However, for the sake of completeness, we report the full results including DAS in this appendix. The implementations for CAM, SCORE, DAS, and NoGAM were utilized from the `dodiscover` package[3]. For DiffAN and CaPS, we used the authors' original implementations[4].

- **CAM**: The Causal Additive Models (CAM) (Bühlmann et al., 2014) algorithm decouples the discovery problem into two stages: estimating a topological order and performing feature selection. To find the order, it employs a greedy search that aims to maximize a likelihood score. For the second stage, it uses the well-known CAM pruning procedure. This method

[3] `https://github.com/py-why/dodiscover`

[4] `https://github.com/vios-s/DiffAN`, `https://github.com/E2real/CaPS`

fits a Generalized Additive Model (GAM) for each variable X_j using its predecessors in the order as covariates and tests the null hypothesis $H_0 : f_{j,k}(\cdot) \equiv 0$ for each potential parent X_k. Edges are pruned based on the statistical significance (p-value) of each parent's contribution.

- **SCORE**: The SCORE algorithm (Rolland et al., 2022) recursively identifies the topological order by finding leaf nodes. Under non-linear assumptions, a node X_j is identified as a leaf if the variance of the diagonal of its score function's Jacobian is zero. The score function, $s(x) = \nabla_x \log p(x)$, is typically estimated via a Stein gradient estimator. The formal criterion for identifying a leaf node is:

$$\underset{j}{\operatorname{argmin}} \operatorname{Var}\left[\frac{\partial s_j(x)}{\partial x_j}\right].$$

  For the pruning stage, SCORE defaults to using established methods like CAM pruning on the fully-connected DAG derived from the estimated order.

- **DAS**: The Discovery At Scale (DAS) (Montagna et al., 2023b) algorithm utilizes the same variance-based criterion as SCORE to identify the topological order. Its primary innovation lies in the pruning stage, where it uses the off-diagonal elements of the score Jacobian to perform an initial, computationally efficient edge selection based on the criterion $E[|\partial_{X_k} s_j(x)|] \neq 0 \iff X_k \in pa(j)$. This step significantly reduces the number of candidate edges for the final pruning stage, which still relies on CAM pruning to refine the graph and reduce false positives.

- **NoGAM**: The NoGAM algorithm (Montagna et al., 2023c) generalizes the score-based ordering approach to accommodate arbitrary additive noise models. It identifies leaf nodes by finding the node that minimizes the mean squared error of a score prediction derived from estimated noise residuals, R_i. The formal criterion is:

$$\underset{j}{\operatorname{argmin}}\, E\left[\left(E[s_j(X)|R_j] - s_j(X)\right)^2\right].$$

  The score function is approximated via score matching based on the Stein identity. Similar to other score-based methods, it relies on a post-processing step like CAM pruning to obtain the final sparse DAG.

- **DiffAN**: The DiffAN algorithm (Sanchez et al., 2023) shares the same variance-based leaf identification criterion as SCORE for finding the topological order. Its core contribution is a more scalable method for estimating the score function. Instead of using traditional score matching, DiffAN employs probabilistic diffusion models to efficiently approximate the score and its Jacobian. After the order is established, it applies a standard post-processing pruning step, such as CAM pruning, to finalize the causal graph.

- **CaPS**: The Causal Discovery with Parent Score (CaPS) (Xu et al., 2024) algorithm introduces a new ordering criterion designed to be robust in mixed linear and non-linear settings. It identifies leaf nodes based on the expectation of the score's Jacobian, rather than its variance. The formal criterion is:

$$j^* = \underset{j}{\operatorname{argmax}}\left(\operatorname{diag}\left(\mathbb{E}\left[\frac{\partial s(x)}{\partial x}\right]\right)\right) \implies x_{j^*} \text{ is a leaf node.}$$

  A key feature of CaPS is its "parent score," which is used to perform an efficient pre-pruning of weak edges and to supplement the graph with strong edges, thereby reducing the computational burden and improving the accuracy of the final CAM pruning step.

G.3 Evaluation Metrics

We evaluate the accuracy of the recovered graph structures using a suite of standard metrics. Let TP (True Positives) be the number of correctly identified edges, FP (False Positives) be the number of incorrectly identified edges, FN (False Negatives) be the number of missed true edges, and R be the number of edges with a reversed direction.

- **Structural Hamming Distance (SHD):** The SHD measures the overall structural dissimilarity between the estimated graph and the ground truth graph. It is defined as the total number of edge operations required to make the two graphs identical:

$$\text{SHD} = \text{FP} + \text{FN} + \text{R}.$$

  A lower SHD indicates a more accurate structural recovery.

- **Structural Intervention Distance (SID):** The SID is a more causally-informed metric that quantifies the number of downstream errors in interventional reasoning that would result from using the estimated graph. It identifies pairs of variables (i, j) for which the set of causal paths from i to j is incorrectly estimated. A lower SID indicates that the estimated graph is more faithful for predicting the effects of interventions.

- **Precision, Recall, and F1 Score:** These metrics provide a balanced assessment of edge discovery accuracy, treating the problem as a binary classification task for each potential edge.
  - **Precision** measures the fraction of predicted edges that are correct:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}.$$

  - **Recall** (or True Positive Rate) measures the fraction of true edges that were correctly identified:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}.$$

  - The **F1 Score** is the harmonic mean of Precision and Recall, providing a single, balanced measure:

$$\text{F1 Score} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}.$$

  - **A Note on Reversed Edges in Metrics**: In our evaluation, we treat reversed edges (R) as a distinct type of error. For metrics such as the False Discovery Rate (FDR) and False Positive Rate (FPR) reported in our full results, we include these reversed edges in the numerator alongside false positives (e.g., $\text{FPR} = (\text{R} + \text{FP})/(\text{TN} + \text{FP})$). We adopt this stricter convention because a reversed edge, while correctly identifying an adjacency, represents a fundamentally incorrect claim about the causal direction and should be penalized as a type of false discovery.

H Experiments

H.1 Computational Time Analysis.

In this section, we analyze the computational runtime of our proposed PEP framework compared to the traditional CAM pruning method. Let n be the sample size and, for each node j, let m_j be the number of candidate parents provided to the pruner. We analyze *only* the pruning stage, conditional on a given order and candidate set.

CAM Pruning. CAM typically employs a backward elimination strategy. It begins by fitting a Generalized Additive Model (GAM) on all m_j candidate parents. A single fit, involving B backfitting sweeps over m_j smoothers with basis size s, has a cost of $C_{\text{GAM}}(n, m_j, s, B) = \Theta(B\, m_j s^2 n)$. The backward elimination process may refit the model up to m_j times as parents are removed one by one. This results in a total pruning cost for node j that is quadratic in the number of initial candidates:

$$T_{\text{CAM}}(j) \;\approx\; \sum_{r=1}^{m_j} C_{\text{GAM}}(n, r, s, B) \;=\; \Theta\big(B\, s^2 n \textstyle\sum_{r=1}^{m_j} r\big) \;=\; \Theta\big(B\, s^2 n\, m_j^2\big).$$

The total cost is the sum over all nodes, $T_{\text{CAM}} = \sum_j T_{\text{CAM}}(j)$.

Table H.1: Runtime comparison (in seconds) of CAM pruning and our proposed PEP framework across various ordering-based methods on synthetic datasets with various configurations (d: number of nodes, n: number of samples).

Dataset	Method	CAM	SCORE	DAS	NoGAM	DiffAN	CaPS
d=10, n=1000	CAM pruning	25.53	8.08	8.42	22.72	10.71	7.58
	PEP	63.63	45.53	45.81	60.21	46.11	43.39
d=10, n=2000	CAM pruning	30.53	16.96	20.08	58.56	16.79	16.76
	PEP	89.14	74.00	76.18	114.53	67.75	65.78

Prequential Evidence Pruning. PEP also performs greedy backward pruning, starting with m_j candidate parents. To decide which edge to remove at each step, it must evaluate the evidence gain $\delta_{i\to j}$ for all parents currently in the set. A full backward path requires a total of $m_j + (m_j - 1) + \cdots + 1 = \Theta(m_j^2)$ evaluations of the δ statistic.

Each evaluation of $\delta_{i\to j}$ requires obtaining out-of-fold log-likelihoods for two parent sets over K folds. Let $C_q(n, r)$ be the cost of training and evaluating the predictive model q on a parent set of size r. For an *amortized* predictor (e.g., a pre-trained model like TabPFN), this cost is dominated by evaluation, with a per-sample cost of $\alpha(r)$. The cost is thus $C_q(n, r) \approx \Theta(Kn\alpha(r))$. The total pruning cost for node j is the number of δ evaluations multiplied by the average cost of each evaluation:

$$T_{\text{PEP}}(j) \;=\; \Theta(m_j^2) \times \Theta(Kn\bar{\alpha}) \;=\; \Theta\big(Knm_j^2\bar{\alpha}\big),$$

where $\bar{\alpha}$ is the average per-sample evaluation cost. For non-amortized predictors that are re-trained from scratch for each evaluation, the cost would be significantly higher.

Summary. Conditional on a fixed order, the practical implementations of both CAM-pruning and PEP exhibit pruning costs that are *linear in the sample size* n and *quadratic in the number of candidate parents* m_j:

$$T_{\text{CAM}}(j) = \Theta(Bs^2nm_j^2), \qquad T_{\text{PEP}}(j) = \Theta(Knm_j^2\bar{\alpha}).$$

The constants reflect the core operations of each method: backfitting sweeps and basis size (B, s) for CAM, versus fold count and per-sample evaluation cost $(K, \bar{\alpha})$ for PEP. PEP's slightly larger constant factor is the computational price for its robust, out-of-fold, context-aware evidence evaluation, which enables its improved pruning accuracy. Table H.1 reports wall-clock times, which remain in a practical range for typical research use cases.

H.2 Statistical Significance Tests

To rigorously validate the performance improvements observed in our main experiments, we conducted statistical significance tests. We use the non-parametric Wilcoxon signed-rank test to compare the paired results of our proposed method against each baseline over 10 random seeds, reporting the p-value to assess statistical significance ($p < 0.05$).

To quantify the magnitude of these differences, we also compute Cohen's d as a standardized measure of effect size (ES). Following established conventions, we interpret the absolute value of the effect size, |ES|, as negligible ($|ES| < 0.2$), small ($0.2 \leq |ES| < 0.5$), medium ($0.5 \leq |ES| < 0.8$), or large ($|ES| \geq 0.8$). A large effect size indicates a practically meaningful and substantial difference in performance. Our tables report this absolute value to clearly convey the strength of the observed effect.

Tables H.2 through H.5 summarize these results across various synthetic data settings. The results overwhelmingly confirm that our method provides not only a statistically significant but also a practically substantial improvement over all baselines across most key metrics, as indicated by both low p-values and large effect sizes.

We also acknowledge that this statistical significance is not uniformly present across every conceivable setting. For instance, when pairing the PEP module with the DiffAN ordering backbone, the high variance inherent to its diffusion-based score estimation can lead to less stable orderings, occasionally

resulting in non-significant p-values. Similarly, in settings with a very high proportion of linear relationships, the performance gap between PEP and specialized methods like SCORE or NoGAM can narrow, which in a few runs may not be statistically significant. This aligns with our analysis that PEP's most profound advantage lies in the complex, mixed, and non-linear regimes, while it remains highly competitive in simpler, purely linear settings.

Table H.2: Statistical significance tests of CaPS-PEP against baselines on the synthetic ER dataset ($d = 10, n = 2000, 4d$ edges).

Metric	Method	CAM Pruning	CaPS w/ PEP	Effect Size	p-value	Improvement (%)
SHD (↓)	CAM	17.40	5.80	-2.68	< 0.002	66.67%
	SCORE	17.60	5.80	-2.83	< 0.002	67.05%
	DAS	19.80	5.80	-3.40	< 0.002	70.71%
	NoGAM	17.70	5.80	-2.80	< 0.002	67.23%
	DiffAN	18.40	5.80	-3.08	< 0.002	68.48%
	CaPS	15.20	5.80	-2.31	< 0.002	61.84%
SID (↓)	CAM	32.90	8.20	-2.76	< 0.002	75.08%
	SCORE	31.70	8.20	-2.85	< 0.002	74.13%
	DAS	39.40	8.20	-4.56	< 0.002	79.19%
	NoGAM	32.30	8.20	-2.89	< 0.002	74.61%
	DiffAN	45.70	8.20	-3.13	< 0.002	82.06%
	CaPS	28.70	8.20	-2.87	< 0.002	71.43%
F1 Score (↑)	CAM	0.670	0.909	3.06	< 0.002	35.65%
	SCORE	0.667	0.909	3.24	< 0.002	36.31%
	DAS	0.612	0.909	3.94	< 0.002	48.53%
	NoGAM	0.664	0.909	3.17	< 0.002	36.92%
	DiffAN	0.634	0.909	3.42	< 0.002	43.33%
	CaPS	0.721	0.909	2.44	< 0.002	26.09%

Table H.3: Statistical significance tests of CaPS-PEP against baselines on the synthetic SF dataset ($d = 10, n = 2000, 4d$ edges.

Metric	Method	CAM Pruning	CaPS w/ PEP	Effect Size	p-value	Improvement (%)
SHD (↓)	CAM	8.10	1.40	2.27	< 0.002	82.72%
	SCORE	7.80	1.40	1.84	< 0.002	82.05%
	DAS	9.70	1.40	2.37	< 0.002	85.57%
	NoGAM	7.50	1.40	1.87	< 0.002	81.33%
	DiffAN	10.10	1.40	3.11	< 0.002	86.14%
	CaPS	6.60	1.40	1.75	< 0.002	78.79%
SID (↓)	CAM	24.80	3.10	2.23	< 0.002	87.50%
	SCORE	18.30	3.10	1.77	< 0.002	83.06%
	DAS	24.30	3.10	2.60	< 0.002	87.24%
	NoGAM	17.40	3.10	1.82	< 0.002	82.18%
	DiffAN	38.80	3.10	3.49	< 0.002	92.01%
	CaPS	16.40	3.10	1.52	< 0.002	81.10%
F1 Score (↑)	CAM	0.775	0.963	2.06	< 0.002	24.22%
	SCORE	0.791	0.963	1.53	< 0.002	21.73%
	DAS	0.737	0.963	2.03	< 0.002	30.75%
	NoGAM	0.803	0.963	1.51	< 0.002	19.94%
	DiffAN	0.692	0.963	2.72	< 0.002	39.27%
	CaPS	0.829	0.963	1.47	< 0.002	16.25%

Table H.4: Statistical significance tests of the performance improvement gained by applying the PEP module to various ordering-based backbones on the synthetic ER dataset ($d = 10, n = 2000, 4d$ edges.

Metric	Method	CAM Pruning	w/ PEP	Effect Size	p-value	Improvement (%)
SHD (↓)	CAM	17.40	7.50	2.42	< 0.002	56.90%
	SCORE	17.60	5.60	3.20	< 0.002	68.18%
	DAS	17.60	5.60	3.20	< 0.002	68.18%
	NoGAM	17.70	5.00	3.32	< 0.002	71.75%
	DiffAN	18.40	10.00	1.63	< 0.010	45.65%
	CaPS	15.20	5.80	2.31	< 0.002	61.84%
SID (↓)	CAM	32.90	14.60	1.99	< 0.004	55.62%
	SCORE	31.70	4.90	3.89	< 0.002	84.54%
	DAS	31.70	4.90	3.89	< 0.002	84.54%
	NoGAM	32.30	4.20	4.11	< 0.002	87.00%
	DiffAN	45.70	35.40	0.75	0.232	22.54%
	CaPS	28.70	8.20	2.87	< 0.002	71.43%
F1 Score (↑)	CAM	0.670	0.871	2.73	< 0.002	30.02%
	SCORE	0.667	0.915	3.73	< 0.002	37.25%
	DAS	0.667	0.915	3.73	< 0.002	37.25%
	NoGAM	0.664	0.925	3.77	< 0.002	39.32%
	DiffAN	0.634	0.797	1.50	< 0.02	25.66%
	CaPS	0.721	0.909	2.44	< 0.002	26.09%

Table H.5: Statistical significance tests of the performance improvement gained by applying the PEP module to various ordering-based backbones on the synthetic SF dataset ($d = 10, n = 2000, 4d$ edges.

Metric	Method	CAM Pruning	w/ PEP	Effect Size	p-value	Improvement (%)
SHD (↓)	CAM	8.10	7.50	0.16	0.695	7.41%
	SCORE	7.80	1.40	1.85	< 0.002	82.05%
	DAS	7.80	1.40	1.85	< 0.002	82.05%
	NoGAM	7.50	2.00	1.44	< 0.020	73.33%
	DiffAN	10.10	6.60	0.93	0.160	34.65%
	CaPS	6.60	1.40	1.75	< 0.002	78.79%
SID (↓)	CAM	24.80	14.60	1.03	< 0.03	41.13%
	SCORE	18.30	3.20	1.77	< 0.002	82.51%
	DAS	18.30	3.20	1.77	< 0.002	82.51%
	NoGAM	17.40	5.00	1.49	< 0.05	71.26%
	DiffAN	38.80	27.30	1.00	0.105	29.64%
	CaPS	16.40	3.10	1.52	< 0.002	81.10%
F1 Score (↑)	CAM	0.775	0.871	1.01	< 0.002	12.38%
	SCORE	0.791	0.963	1.52	< 0.002	21.69%
	DAS	0.791	0.963	1.52	< 0.002	21.69%
	NoGAM	0.803	0.948	1.23	< 0.05	18.04%
	DiffAN	0.692	0.804	0.94	0.131	16.21%
	CaPS	0.829	0.963	1.47	< 0.002	16.25%

H.3 Additional Experimental Results

Ablation Studies. To further investigate the robustness and performance characteristics of our PEP framework, we conduct two additional ablation studies by varying the data generation environment. First, we assess the framework's performance under data scarcity. Figure H.1, Table H.6, Table H.7 replicates the main plug-in enhancement and framework-versus-predictor experiments, but with the sample size reduced from $n = 2000$ to $n = 1000$. The results demonstrate that the consistent performance gains from using PEP are robust even with more limited data. Second, we evaluate the framework in a more general, mixed-linearity setting. While our main experiments used a fully non-linear environment for fair comparison with score-based methods, Figure H.2, Table H.8, Table H.9 presents the same set of experiments on datasets generated with a linear proportion of 0.5. This shows that PEP maintains its strong performance in environments that mix both linear and non-linear relationships, highlighting its versatility.

Detailed Results. This section reports the detailed numerical results corresponding to the plots presented in the main text. While the main paper illustrates the findings through figures, here we provide the exact quantitative values (mean and standard deviation) for each experiment in tabular form. As a reference, the detailed numerical results corresponding to Fig. 3, 4, and 5 in the main text are reported in Table H.10, Table H.11, and Table H.12–Table H.13.

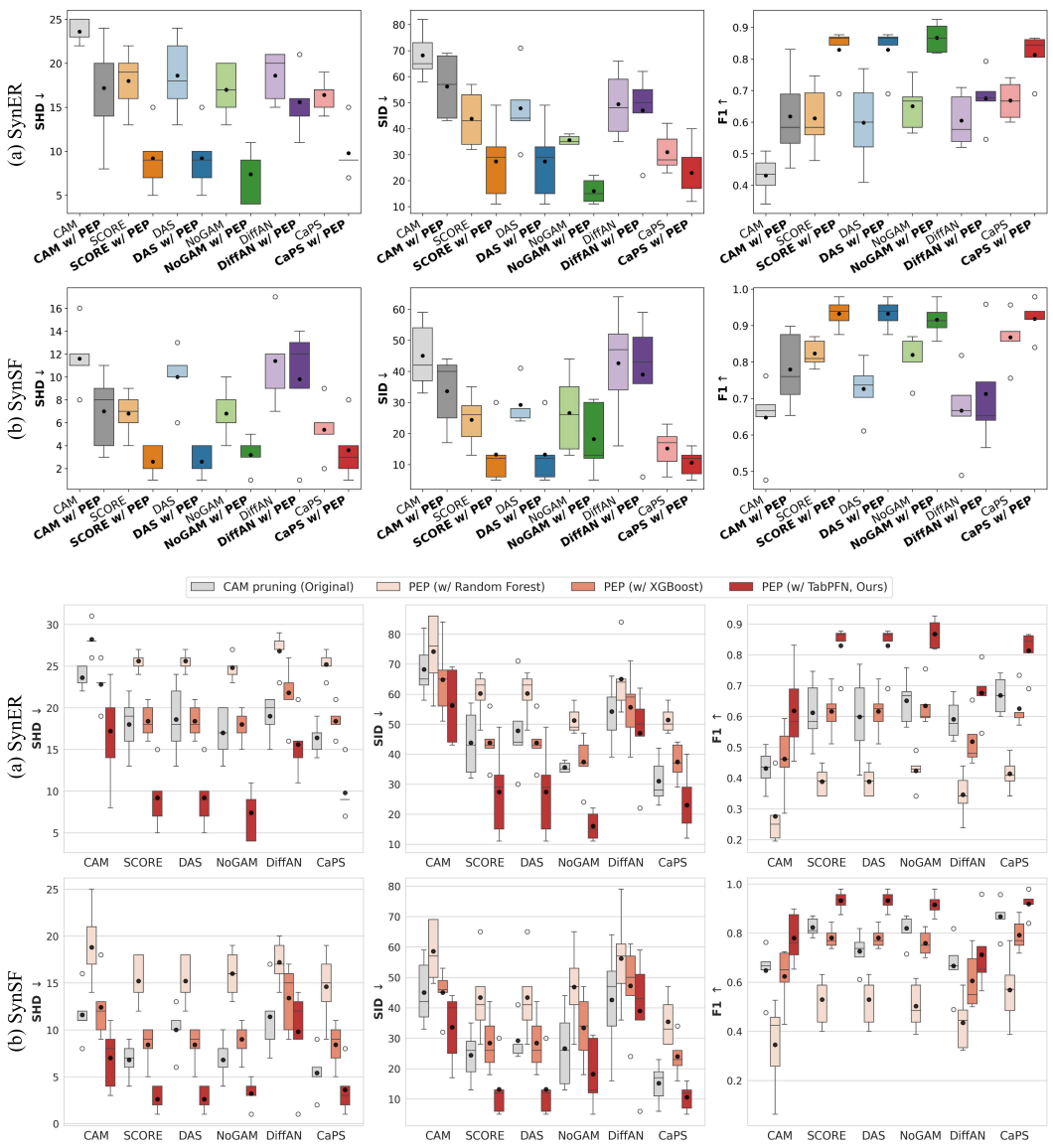

Figure H.1: Performance comparison on synthetic datasets with a reduced sample size ($n = 1000$). This figure replicates the main plug-in and framework-vs-predictor experiments, showing that PEP's advantages hold in data-scarce conditions.

Table H.6: Benchmark results at n=1000 comparing CAM pruning vs PEP across six ordering backbones. Cells report mean ± std (two decimals). Bold marks the better mean per backbone within each metric (lower is better for SHD/SID; higher is better otherwise).

Dataset	Metric	Pruning	CAM	SCORE	DAS	NoGAM	DiffAN	CaPS
SynER	SHD↓	CAM pruning	23.60 ± 1.34	18.00 ± 3.54	18.60 ± 4.45	17.00 ± 3.08	18.60 ± 2.88	16.40 ± 1.95
		PEP	**17.20 ± 6.26**	**9.20 ± 3.77**	**12.60 ± 3.66**	**7.40 ± 3.21**	**15.60 ± 3.65**	**9.80 ± 3.03**
	SID↓	CAM pruning	68.20 ± 9.42	43.80 ± 11.12	47.80 ± 15.02	44.30 ± 10.71	61.10 ± 10.55	38.40 ± 9.59
		PEP	**56.00 ± 15.90**	**26.50 ± 11.88**	**33.70 ± 12.25**	**24.60 ± 10.87**	**57.20 ± 16.69**	**22.20 ± 8.30**
	F1↑	CAM pruning	0.52 ± 0.07	0.73 ± 0.08	0.70 ± 0.07	0.74 ± 0.08	0.60 ± 0.09	0.78 ± 0.07
		PEP	**0.64 ± 0.12**	**0.86 ± 0.06**	**0.81 ± 0.07**	**0.86 ± 0.06**	**0.70 ± 0.08**	**0.89 ± 0.04**
	Precision↑	CAM pruning	0.66 ± 0.12	0.90 ± 0.08	0.88 ± 0.08	0.91 ± 0.08	0.76 ± 0.10	0.93 ± 0.05
		PEP	**0.77 ± 0.10**	**0.95 ± 0.05**	**0.92 ± 0.05**	**0.95 ± 0.05**	**0.82 ± 0.08**	**0.96 ± 0.03**
	Recall↑	CAM pruning	0.47 ± 0.10	0.64 ± 0.11	0.61 ± 0.10	0.64 ± 0.10	0.55 ± 0.12	0.72 ± 0.09
		PEP	**0.57 ± 0.14**	**0.82 ± 0.09**	**0.78 ± 0.09**	**0.83 ± 0.09**	**0.62 ± 0.10**	**0.86 ± 0.05**
SynSF	SHD↓	CAM pruning	10.30 ± 3.38	7.90 ± 2.73	13.80 ± 3.38	10.10 ± 2.08	11.70 ± 2.08	6.70 ± 3.60
		PEP	**8.80 ± 2.52**	**5.40 ± 2.72**	**10.30 ± 2.41**	**7.80 ± 2.20**	**9.10 ± 2.64**	**4.30 ± 2.94**
	SID↓	CAM pruning	51.70 ± 12.81	46.60 ± 11.88	64.60 ± 10.36	49.50 ± 10.95	49.20 ± 10.19	17.30 ± 10.49
		PEP	**43.60 ± 12.43**	**36.90 ± 11.64**	**53.30 ± 13.56**	**42.20 ± 11.76**	48.90 ± 12.47	**15.20 ± 12.24**
	F1↑	CAM pruning	0.68 ± 0.08	0.76 ± 0.08	0.65 ± 0.12	0.83 ± 0.07	0.70 ± 0.06	0.84 ± 0.09
		PEP	**0.73 ± 0.08**	**0.82 ± 0.08**	**0.69 ± 0.09**	**0.88 ± 0.06**	**0.72 ± 0.08**	**0.89 ± 0.08**
	Precision↑	CAM pruning	0.73 ± 0.07	0.81 ± 0.08	0.70 ± 0.12	0.90 ± 0.07	0.70 ± 0.07	0.86 ± 0.11
		PEP	**0.76 ± 0.08**	**0.85 ± 0.08**	**0.72 ± 0.10**	**0.92 ± 0.06**	**0.71 ± 0.07**	**0.90 ± 0.09**
	Recall↑	CAM pruning	0.65 ± 0.10	0.73 ± 0.11	0.62 ± 0.12	0.78 ± 0.08	0.70 ± 0.08	0.82 ± 0.11
		PEP	**0.71 ± 0.09**	**0.80 ± 0.10**	**0.66 ± 0.11**	**0.88 ± 0.08**	**0.72 ± 0.09**	**0.88 ± 0.09**

Table H.7: Pruning comparison at n=1000 across six ordering backbones. Per metric, four pruning variants are listed: CAM, PEP w/ RF, PEP w/ XGB, PEP w/ TabPFN. Cells report mean ± std (two decimals). Bold marks the best mean per backbone within each metric (lower is better for SHD/SID; higher is better otherwise).

Dataset	Metric	Pruning	CAM	SCORE	DAS	NoGAM	DiffAN	CaPS
SynER	SHD↓	CAM	23.60 ± 1.34	18.00 ± 3.54	18.60 ± 4.45	17.00 ± 3.08	18.60 ± 2.88	16.40 ± 1.95
		PEP w/ RF	34.60 ± 1.90	34.60 ± 1.90	34.90 ± 1.85	34.90 ± 1.85	35.70 ± 2.06	34.90 ± 1.85
		PEP w/ XGB	21.70 ± 6.17	18.80 ± 5.07	18.80 ± 5.07	17.50 ± 4.53	23.10 ± 5.66	19.20 ± 4.60
		PEP w/ TabPFN	**15.50 ± 6.36**	**8.50 ± 3.78**	**8.50 ± 3.78**	**6.80 ± 3.28**	**14.10 ± 3.66**	**9.30 ± 3.25**
	SID↓	CAM	68.20 ± 9.42	43.80 ± 11.12	47.80 ± 15.02	44.30 ± 10.71	61.10 ± 10.55	38.40 ± 9.59
		PEP w/ RF	111.50 ± 11.60	107.90 ± 11.06	108.20 ± 10.93	108.70 ± 11.16	116.00 ± 12.04	105.20 ± 9.73
		PEP w/ XGB	54.10 ± 17.55	33.40 ± 11.21	33.40 ± 11.21	27.70 ± 11.23	60.00 ± 20.38	30.40 ± 10.47
		PEP w/ TabPFN	**45.60 ± 17.58**	**19.40 ± 9.50**	**19.40 ± 9.50**	**14.70 ± 9.24**	**49.70 ± 16.85**	**23.30 ± 8.41**
	F1↑	CAM	0.52 ± 0.07	0.73 ± 0.08	0.70 ± 0.07	0.74 ± 0.08	0.60 ± 0.09	0.78 ± 0.07
		PEP w/ RF	0.27 ± 0.12	0.33 ± 0.08	0.33 ± 0.08	0.35 ± 0.07	0.23 ± 0.10	0.39 ± 0.08
		PEP w/ XGB	0.61 ± 0.11	0.74 ± 0.07	0.74 ± 0.07	0.77 ± 0.07	0.58 ± 0.10	0.81 ± 0.05
		PEP w/ TabPFN	**0.74 ± 0.10**	**0.88 ± 0.06**	**0.88 ± 0.06**	**0.90 ± 0.06**	**0.71 ± 0.08**	**0.92 ± 0.03**
	Precision↑	CAM	0.66 ± 0.12	0.90 ± 0.08	0.88 ± 0.08	0.91 ± 0.08	0.76 ± 0.10	0.93 ± 0.05
		PEP w/ RF	0.25 ± 0.10	0.26 ± 0.07	0.26 ± 0.07	0.27 ± 0.07	0.22 ± 0.08	0.30 ± 0.06
		PEP w/ XGB	0.65 ± 0.11	0.80 ± 0.07	0.80 ± 0.07	0.83 ± 0.07	0.70 ± 0.09	0.87 ± 0.06
		PEP w/ TabPFN	**0.78 ± 0.10**	**0.92 ± 0.05**	**0.92 ± 0.05**	**0.93 ± 0.06**	**0.75 ± 0.16**	**0.96 ± 0.03**
	Recall↑	CAM	0.47 ± 0.10	0.64 ± 0.11	0.61 ± 0.10	0.64 ± 0.10	0.55 ± 0.12	0.72 ± 0.09
		PEP w/ RF	0.25 ± 0.15	0.38 ± 0.10	0.38 ± 0.10	0.36 ± 0.10	0.31 ± 0.10	0.42 ± 0.14
		PEP w/ XGB	0.54 ± 0.12	0.66 ± 0.08	0.66 ± 0.08	0.65 ± 0.09	0.52 ± 0.12	0.68 ± 0.09
		PEP w/ TabPFN	**0.78 ± 0.12**	**0.92 ± 0.04**	**0.92 ± 0.04**	**0.90 ± 0.04**	**0.72 ± 0.16**	**0.93 ± 0.03**
SynSF	SHD↓	CAM	10.30 ± 3.38	7.90 ± 2.73	13.80 ± 3.38	10.10 ± 2.08	11.70 ± 2.08	6.70 ± 3.60
		PEP w/ RF	9.60 ± 2.31	9.60 ± 2.31	9.60 ± 2.31	9.50 ± 2.32	10.80 ± 2.48	8.70 ± 2.22
		PEP w/ XGB	8.40 ± 3.36	7.60 ± 2.68	7.60 ± 2.68	7.20 ± 2.45	9.50 ± 2.76	6.60 ± 2.74
		PEP w/ TabPFN	**6.60 ± 3.21**	**4.40 ± 2.63**	**4.40 ± 2.63**	**4.10 ± 2.29**	**8.30 ± 2.34**	**4.60 ± 2.77**
	SID↓	CAM	51.70 ± 12.81	46.60 ± 11.88	64.60 ± 10.36	49.50 ± 10.95	49.20 ± 10.19	17.30 ± 10.49
		PEP w/ RF	58.20 ± 13.69	55.10 ± 12.98	55.10 ± 12.98	52.40 ± 12.70	66.00 ± 14.12	47.10 ± 11.33
		PEP w/ XGB	45.20 ± 11.56	40.40 ± 10.32	40.40 ± 10.32	38.30 ± 9.65	55.30 ± 12.68	36.50 ± 10.06
		PEP w/ TabPFN	**38.30 ± 12.01**	**32.20 ± 11.81**	**32.20 ± 11.81**	**29.50 ± 11.55**	**50.60 ± 13.10**	**31.60 ± 11.84**
	F1↑	CAM	0.68 ± 0.08	0.76 ± 0.08	0.65 ± 0.12	0.83 ± 0.07	0.70 ± 0.06	0.84 ± 0.09
		PEP w/ RF	0.71 ± 0.07	0.79 ± 0.07	0.79 ± 0.07	0.82 ± 0.07	0.68 ± 0.08	0.86 ± 0.08
		PEP w/ XGB	0.78 ± 0.08	0.86 ± 0.06	0.86 ± 0.06	0.89 ± 0.06	0.74 ± 0.09	0.91 ± 0.05
		PEP w/ TabPFN	**0.80 ± 0.08**	**0.90 ± 0.05**	**0.90 ± 0.05**	**0.92 ± 0.05**	**0.77 ± 0.09**	**0.94 ± 0.05**
	Precision↑	CAM	0.73 ± 0.07	0.81 ± 0.08	0.70 ± 0.12	0.90 ± 0.07	0.70 ± 0.07	0.86 ± 0.11
		PEP w/ RF	0.74 ± 0.08	0.82 ± 0.08	0.82 ± 0.08	0.86 ± 0.07	0.72 ± 0.07	0.90 ± 0.09
		PEP w/ XGB	0.79 ± 0.08	0.86 ± 0.06	0.86 ± 0.06	0.89 ± 0.06	0.76 ± 0.09	0.93 ± 0.07
		PEP w/ TabPFN	**0.81 ± 0.08**	**0.89 ± 0.05**	**0.89 ± 0.05**	**0.92 ± 0.05**	**0.75 ± 0.10**	**0.96 ± 0.03**
	Recall↑	CAM	0.65 ± 0.10	0.73 ± 0.11	0.62 ± 0.12	0.78 ± 0.08	0.70 ± 0.08	0.82 ± 0.11
		PEP w/ RF	0.66 ± 0.10	0.76 ± 0.10	0.76 ± 0.10	0.79 ± 0.10	0.66 ± 0.10	0.83 ± 0.11
		PEP w/ XGB	0.73 ± 0.11	0.82 ± 0.08	0.82 ± 0.08	0.86 ± 0.08	0.71 ± 0.09	0.89 ± 0.08
		PEP w/ TabPFN	**0.77 ± 0.10**	**0.87 ± 0.05**	**0.87 ± 0.05**	**0.91 ± 0.05**	**0.74 ± 0.10**	**0.97 ± 0.05**

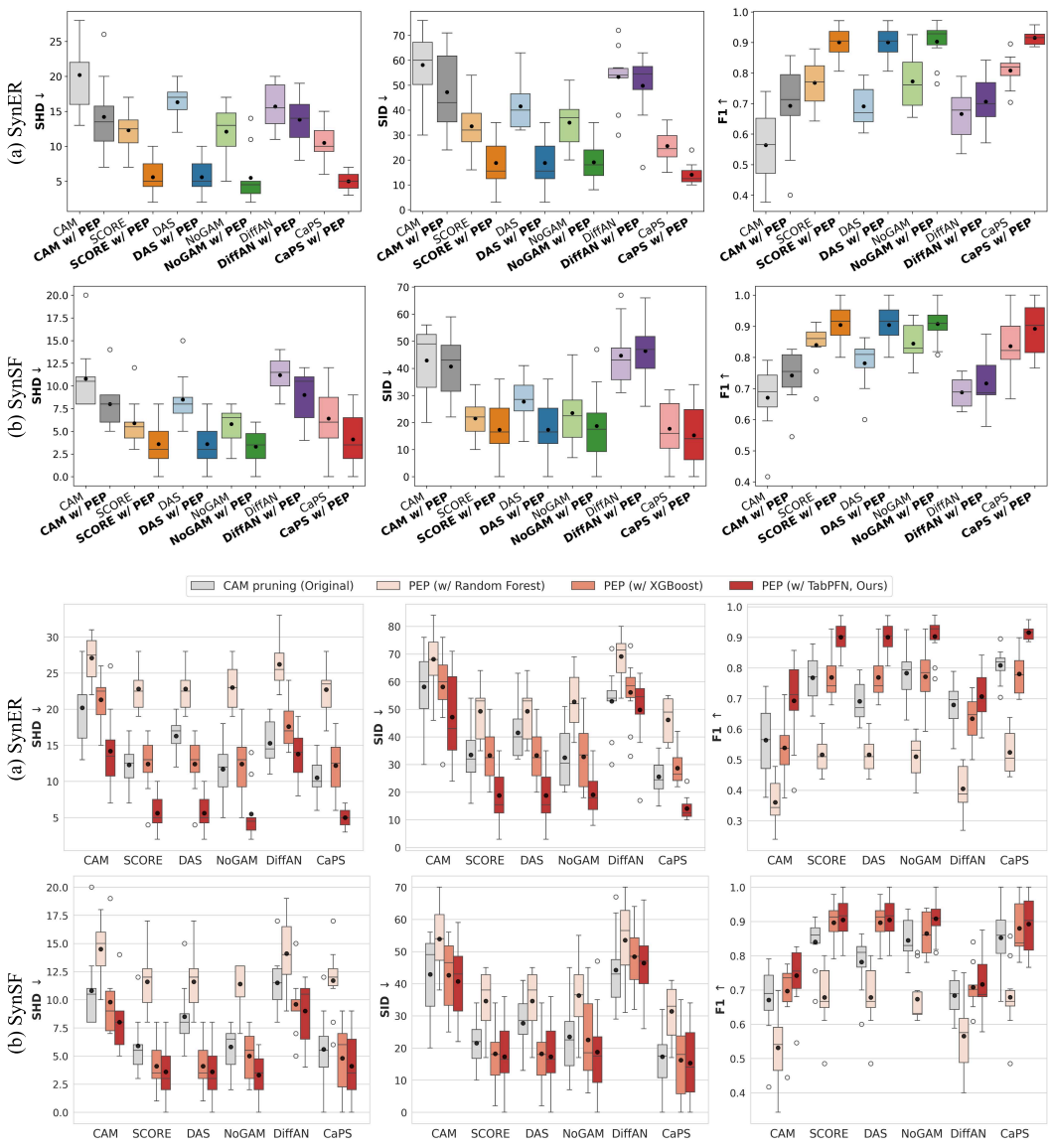

Figure H.2: Performance comparison on synthetic datasets with a mixed environment (linear proportion = 0.5). This figure demonstrates PEP's strong performance in a more general setting beyond the fully non-linear environment used for the main comparisons.

Table H.8: Benchmark results at *linear proposition* = 0.5 comparing CAM pruning vs PEP across six ordering backbones. Cells report mean ± std (two decimals). Bold marks the better mean per backbone within each metric (lower is better for SHD/SID; higher is better otherwise).

Dataset	Metric	Pruning	CAM	SCORE	DAS	NoGAM	DiffAN	CaPS
SynER	SHD↓	CAM pruning	20.20 ± 4.66	12.30 ± 3.34	16.30 ± 2.41	12.10 ± 4.07	15.70 ± 3.47	10.50 ± 2.55
		PEP	**14.20 ± 5.65**	**5.60 ± 2.55**	**8.90 ± 2.57**	**7.60 ± 2.63**	**13.80 ± 3.94**	**5.00 ± 1.49**
	SID↓	CAM pruning	58.10 ± 13.76	33.50 ± 10.54	41.50 ± 9.72	35.00 ± 11.22	53.30 ± 12.07	25.60 ± 6.83
		PEP	**47.20 ± 15.81**	**18.80 ± 6.59**	**26.80 ± 9.09**	**26.70 ± 9.40**	**49.80 ± 13.53**	**14.10 ± 4.36**
	F1↑	CAM pruning	0.56 ± 0.19	0.80 ± 0.11	0.78 ± 0.10	0.82 ± 0.07	0.64 ± 0.12	0.81 ± 0.05
		PEP	**0.69 ± 0.14**	**0.90 ± 0.07**	**0.86 ± 0.07**	**0.86 ± 0.07**	**0.71 ± 0.09**	**0.92 ± 0.02**
	Precision↑	CAM pruning	0.67 ± 0.15	0.91 ± 0.08	0.90 ± 0.08	0.93 ± 0.07	0.77 ± 0.11	0.94 ± 0.04
		PEP	**0.77 ± 0.12**	**0.95 ± 0.05**	**0.93 ± 0.05**	**0.95 ± 0.05**	**0.82 ± 0.09**	**0.97 ± 0.02**
	Recall↑	CAM pruning	0.51 ± 0.21	0.74 ± 0.15	0.72 ± 0.14	0.73 ± 0.10	0.61 ± 0.13	0.71 ± 0.08
		PEP	**0.63 ± 0.18**	**0.86 ± 0.11**	**0.81 ± 0.10**	**0.80 ± 0.10**	**0.67 ± 0.11**	**0.88 ± 0.04**
SynSF	SHD↓	CAM pruning	9.80 ± 2.82	5.80 ± 1.99	14.60 ± 4.16	11.20 ± 1.93	12.10 ± 2.46	6.40 ± 3.69
		PEP	**8.00 ± 2.62**	**3.60 ± 2.11**	**10.40 ± 3.25**	**7.90 ± 2.09**	**9.00 ± 2.79**	**4.10 ± 3.14**
	SID↓	CAM pruning	49.00 ± 13.73	42.90 ± 14.08	62.30 ± 10.50	45.50 ± 12.46	44.70 ± 11.81	17.70 ± 10.53
		PEP	**40.70 ± 12.01**	**33.20 ± 13.16**	**51.60 ± 14.12**	**40.00 ± 12.56**	46.40 ± 12.76	**15.30 ± 12.37**
	F1↑	CAM pruning	0.69 ± 0.07	0.76 ± 0.08	0.63 ± 0.11	0.84 ± 0.06	0.69 ± 0.05	0.84 ± 0.10
		PEP	**0.74 ± 0.09**	**0.83 ± 0.09**	**0.68 ± 0.10**	**0.89 ± 0.06**	**0.72 ± 0.08**	**0.89 ± 0.09**
	Precision↑	CAM pruning	0.74 ± 0.07	0.82 ± 0.08	0.67 ± 0.10	0.91 ± 0.08	0.69 ± 0.07	0.86 ± 0.12
		PEP	**0.77 ± 0.09**	**0.86 ± 0.09**	**0.69 ± 0.11**	**0.92 ± 0.06**	**0.72 ± 0.09**	**0.90 ± 0.09**
	Recall↑	CAM pruning	0.65 ± 0.08	0.72 ± 0.10	0.61 ± 0.13	0.79 ± 0.07	0.69 ± 0.06	0.82 ± 0.10
		PEP	**0.72 ± 0.09**	**0.80 ± 0.10**	**0.66 ± 0.12**	**0.89 ± 0.08**	**0.72 ± 0.09**	**0.88 ± 0.09**

Table H.9: Pruning comparison at *linear proposition* = 0.5 across six ordering backbones. Per metric, four pruning variants are listed: CAM, PEP w/ RF, PEP w/ XGB, PEP w/ TabPFN. Cells report mean ± std (two decimals). Bold marks the best mean per backbone within each metric (lower is better for SHD/SID; higher is better otherwise).

Dataset	Metric	Pruning	CAM	SCORE	DAS	NoGAM	DiffAN	CaPS
SynER	SHD↓	CAM	17.90 ± 3.84	17.10 ± 3.81	19.80 ± 3.79	17.30 ± 3.74	18.40 ± 4.79	15.80 ± 4.26
		PEP w/ RF	24.10 ± 4.09	24.10 ± 3.96	24.10 ± 4.17	24.10 ± 3.96	25.80 ± 4.83	23.90 ± 3.93
		PEP w/ XGB	15.90 ± 4.95	14.80 ± 3.94	14.80 ± 4.74	14.80 ± 3.94	18.80 ± 5.85	15.40 ± 4.77
		PEP w/ TabPFN	**7.50 ± 3.92**	**5.60 ± 3.09**	**7.30 ± 4.37**	**5.00 ± 3.56**	**10.00 ± 6.27**	**5.80 ± 4.42**
	SID↓	CAM	34.90 ± 7.28	31.30 ± 5.46	39.40 ± 6.08	31.40 ± 5.66	50.10 ± 11.40	28.30 ± 6.53
		PEP w/ RF	74.40 ± 10.28	71.90 ± 9.85	74.07 ± 11.03	72.06 ± 10.37	77.90 ± 13.69	69.70 ± 8.85
		PEP w/ XGB	29.80 ± 11.40	27.20 ± 8.93	29.20 ± 12.57	26.30 ± 9.19	47.60 ± 13.72	24.00 ± 7.29
		PEP w/ TabPFN	**14.60 ± 8.15**	**4.90 ± 3.20**	**15.60 ± 9.71**	**7.80 ± 5.59**	**35.40 ± 12.33**	**8.20 ± 7.54**
	F1↑	CAM	0.67 ± 0.08	0.68 ± 0.08	0.61 ± 0.08	0.66 ± 0.08	0.60 ± 0.09	0.71 ± 0.08
		PEP w/ RF	0.39 ± 0.07	0.39 ± 0.07	0.39 ± 0.07	0.39 ± 0.07	0.35 ± 0.10	0.41 ± 0.07
		PEP w/ XGB	0.74 ± 0.09	0.76 ± 0.08	0.73 ± 0.10	0.76 ± 0.08	0.66 ± 0.10	0.78 ± 0.07
		PEP w/ TabPFN	**0.90 ± 0.05**	**0.94 ± 0.04**	**0.90 ± 0.08**	**0.93 ± 0.04**	**0.83 ± 0.06**	**0.95 ± 0.03**
	Precision↑	CAM	0.96 ± 0.04	1.00 ± 0.02	0.98 ± 0.02	0.99 ± 0.02	0.87 ± 0.07	0.97 ± 0.04
		PEP w/ RF	0.37 ± 0.07	0.32 ± 0.07	0.31 ± 0.07	0.31 ± 0.07	0.39 ± 0.10	0.35 ± 0.08
		PEP w/ XGB	0.79 ± 0.09	0.82 ± 0.07	0.81 ± 0.09	0.84 ± 0.07	0.72 ± 0.10	0.85 ± 0.06
		PEP w/ TabPFN	**0.95 ± 0.03**	**0.97 ± 0.02**	**0.94 ± 0.05**	**0.97 ± 0.02**	**0.91 ± 0.05**	**0.98 ± 0.02**
	Recall↑	CAM	0.50 ± 0.08	0.52 ± 0.08	0.45 ± 0.09	0.51 ± 0.08	0.50 ± 0.11	0.59 ± 0.10
		PEP w/ RF	0.51 ± 0.09	0.58 ± 0.10	0.56 ± 0.09	0.56 ± 0.09	0.43 ± 0.13	0.53 ± 0.11
		PEP w/ XGB	0.70 ± 0.12	0.72 ± 0.13	0.64 ± 0.14	0.70 ± 0.12	0.65 ± 0.11	0.77 ± 0.12
		PEP w/ TabPFN	**0.92 ± 0.09**	**0.97 ± 0.05**	**0.88 ± 0.10**	**0.95 ± 0.05**	**0.88 ± 0.08**	**0.97 ± 0.05**
SynSF	SHD↓	CAM	8.10 ± 3.81	7.80 ± 4.61	9.70 ± 4.64	7.50 ± 4.28	10.10 ± 3.57	6.60 ± 3.84
		PEP w/ RF	7.70 ± 3.31	7.70 ± 3.18	7.70 ± 3.39	7.70 ± 3.18	9.40 ± 4.06	7.60 ± 3.06
		PEP w/ XGB	6.30 ± 3.36	6.30 ± 2.99	6.30 ± 3.38	6.30 ± 2.99	9.20 ± 4.33	6.00 ± 3.06
		PEP w/ TabPFN	**3.10 ± 3.03**	**1.40 ± 1.63**	**2.80 ± 3.29**	**1.20 ± 1.30**	**6.60 ± 3.98**	**1.40 ± 1.71**
	SID↓	CAM	24.80 ± 11.47	18.30 ± 9.49	24.30 ± 8.72	17.40 ± 8.17	38.80 ± 12.35	16.40 ± 9.78
		PEP w/ RF	44.80 ± 13.18	42.20 ± 12.62	44.50 ± 13.33	42.20 ± 12.62	56.80 ± 15.29	41.20 ± 11.47
		PEP w/ XGB	21.60 ± 8.25	19.30 ± 7.42	20.80 ± 8.47	19.40 ± 7.85	36.90 ± 11.76	15.90 ± 9.11
		PEP w/ TabPFN	**11.80 ± 12.88**	**3.20 ± 3.41**	**10.30 ± 10.12**	**3.10 ± 2.85**	**27.30 ± 10.61**	**3.10 ± 7.58**
	F1↑	CAM	0.78 ± 0.12	0.79 ± 0.15	0.74 ± 0.15	0.80 ± 0.14	0.69 ± 0.13	0.83 ± 0.12
		PEP w/ RF	0.83 ± 0.08	0.88 ± 0.08	0.84 ± 0.08	0.88 ± 0.08	0.77 ± 0.09	0.89 ± 0.08
		PEP w/ XGB	0.89 ± 0.09	0.93 ± 0.06	0.90 ± 0.08	0.93 ± 0.06	0.80 ± 0.11	0.94 ± 0.05
		PEP w/ TabPFN	**0.91 ± 0.09**	**0.96 ± 0.05**	**0.91 ± 0.08**	**0.96 ± 0.05**	**0.80 ± 0.11**	**0.96 ± 0.05**
	Precision↑	CAM	**0.96 ± 0.07**	**0.96 ± 0.05**	**0.96 ± 0.09**	**0.99 ± 0.03**	**0.82 ± 0.11**	**0.96 ± 0.06**
		PEP w/ RF	0.95 ± 0.06	0.94 ± 0.06	0.95 ± 0.07	0.96 ± 0.05	0.81 ± 0.11	0.95 ± 0.06
		PEP w/ XGB	0.95 ± 0.06	0.95 ± 0.05	0.95 ± 0.06	0.96 ± 0.05	0.82 ± 0.10	0.96 ± 0.05
		PEP w/ TabPFN	0.95 ± 0.06	0.96 ± 0.05	0.95 ± 0.06	0.97 ± 0.05	0.82 ± 0.09	0.97 ± 0.05
	Recall↑	CAM	0.67 ± 0.16	0.69 ± 0.19	0.62 ± 0.18	0.70 ± 0.19	0.61 ± 0.16	0.75 ± 0.18
		PEP w/ RF	0.72 ± 0.12	0.80 ± 0.10	0.74 ± 0.12	0.80 ± 0.10	0.71 ± 0.11	0.83 ± 0.11
		PEP w/ XGB	0.76 ± 0.12	0.84 ± 0.08	0.79 ± 0.11	0.84 ± 0.08	0.73 ± 0.09	0.89 ± 0.08
		PEP w/ TabPFN	**0.92 ± 0.09**	**0.97 ± 0.05**	**0.90 ± 0.09**	**0.97 ± 0.05**	**0.82 ± 0.09**	**0.97 ± 0.05**

Table H.10: Benchmark results comparing CAM pruning vs PEP across six ordering backbones. Cells report mean ± std (two decimals). Bold marks the better mean per backbone within each metric (lower is better for SHD/SID; higher is better otherwise).

Dataset	Metric	Pruning	CAM	SCORE	DAS	NoGAM	DiffAN	CaPS
SynER	SHD↓	CAM pruning	17.40 ± 4.25	17.60 ± 3.92	19.80 ± 3.79	17.70 ± 4.08	18.40 ± 3.75	15.20 ± 3.68
		PEP	**7.50 ± 3.92**	**5.60 ± 3.57**	**5.60 ± 3.57**	**5.00 ± 3.56**	**10.00 ± 6.27**	**5.80 ± 4.42**
	SID↓	CAM pruning	32.90 ± 10.14	31.70 ± 8.88	39.40 ± 6.08	32.30 ± 9.06	45.70 ± 15.14	28.70 ± 6.73
		PEP	**14.60 ± 8.15**	**4.90 ± 3.98**	**4.90 ± 3.98**	**4.20 ± 3.36**	**35.40 ± 12.33**	**8.20 ± 7.54**
	F1↑	CAM pruning	0.67 ± 0.08	0.67 ± 0.07	0.61 ± 0.08	0.66 ± 0.08	0.63 ± 0.09	0.72 ± 0.08
		PEP	**0.87 ± 0.07**	**0.92 ± 0.06**	**0.92 ± 0.06**	**0.93 ± 0.06**	**0.80 ± 0.13**	**0.91 ± 0.07**
	Precision↑	CAM pruning	**0.97 ± 0.04**	**1.00 ± 0.01**	**0.98 ± 0.02**	**1.00 ± 0.01**	**0.87 ± 0.13**	**0.97 ± 0.04**
		PEP	0.88 ± 0.06	0.92 ± 0.04	0.92 ± 0.04	0.93 ± 0.04	0.79 ± 0.11	0.93 ± 0.06
	Recall↑	CAM pruning	0.52 ± 0.09	0.51 ± 0.09	0.45 ± 0.09	0.50 ± 0.10	0.50 ± 0.08	0.58 ± 0.10
		PEP	**0.86 ± 0.08**	**0.92 ± 0.09**	**0.92 ± 0.09**	**0.92 ± 0.09**	**0.80 ± 0.15**	**0.90 ± 0.10**
SynSF	SHD↓	CAM pruning	8.10 ± 3.81	7.80 ± 4.61	9.70 ± 4.64	7.50 ± 4.28	10.10 ± 3.57	6.60 ± 3.84
		PEP	**3.10 ± 3.03**	**1.40 ± 1.65**	**1.40 ± 1.65**	**2.00 ± 3.30**	**6.60 ± 3.98**	**1.40 ± 1.71**
	SID↓	CAM pruning	24.80 ± 11.47	18.30 ± 9.49	24.30 ± 8.72	17.40 ± 8.17	38.80 ± 12.35	16.40 ± 9.78
		PEP	**11.80 ± 12.88**	**3.20 ± 7.50**	**3.20 ± 7.50**	**5.00 ± 8.46**	**27.30 ± 10.61**	**3.10 ± 7.58**
	F1↑	CAM pruning	0.78 ± 0.12	0.79 ± 0.15	0.74 ± 0.15	0.80 ± 0.14	0.69 ± 0.13	0.83 ± 0.12
		PEP	**0.91 ± 0.09**	**0.96 ± 0.05**	**0.96 ± 0.05**	**0.95 ± 0.09**	**0.80 ± 0.11**	**0.96 ± 0.05**
	Precision↑	CAM pruning	**0.96 ± 0.07**	**0.96 ± 0.05**	**0.96 ± 0.09**	**0.99 ± 0.03**	**0.82 ± 0.11**	**0.96 ± 0.06**
		PEP	0.90 ± 0.10	0.96 ± 0.06	0.96 ± 0.06	0.94 ± 0.10	0.79 ± 0.14	0.95 ± 0.06
	Recall↑	CAM pruning	0.67 ± 0.16	0.69 ± 0.19	0.62 ± 0.18	0.70 ± 0.19	0.61 ± 0.16	0.75 ± 0.18
		PEP	**0.92 ± 0.09**	**0.97 ± 0.06**	**0.97 ± 0.06**	**0.96 ± 0.08**	**0.82 ± 0.09**	**0.97 ± 0.05**

Table H.11: Pruning comparison across ordering backbones. Per metric, four pruning variants are listed: CAM, PEP w/ RF, PEP w/ XGB, PEP w/ TabPFN. Cells report mean ± std (two decimals). Bold marks the better mean per backbone within each metric (lower is better for SHD/SID; higher is better otherwise).

Dataset	Metric	Pruning	CAM	SCORE	DAS	NoGAM	DiffAN	CaPS
SynER	SHD↓	CAM	17.90 ± 3.84	17.10 ± 3.81	19.80 ± 3.79	17.30 ± 3.74	18.40 ± 4.79	15.80 ± 4.26
		PEP w/ RF	24.10 ± 4.09	24.10 ± 3.96	24.10 ± 4.17	24.10 ± 3.96	25.80 ± 4.83	23.90 ± 3.93
		PEP w/ XGB	15.90 ± 4.95	14.80 ± 3.94	14.80 ± 4.74	14.80 ± 3.94	18.80 ± 5.85	15.40 ± 4.77
		PEP w/ TabPFN	**7.50 ± 3.92**	**5.60 ± 3.09**	**7.30 ± 4.37**	**5.00 ± 3.56**	**10.00 ± 6.27**	**5.80 ± 4.42**
	SID↓	CAM	34.90 ± 7.28	31.30 ± 5.46	39.40 ± 6.08	31.40 ± 5.66	50.10 ± 11.40	28.30 ± 6.53
		PEP w/ RF	74.40 ± 10.28	71.90 ± 9.85	74.07 ± 11.03	72.06 ± 10.37	77.90 ± 13.69	69.70 ± 8.85
		PEP w/ XGB	29.80 ± 11.40	27.20 ± 8.93	29.20 ± 12.57	26.30 ± 9.19	47.60 ± 13.72	24.00 ± 7.29
		PEP w/ TabPFN	**14.60 ± 8.15**	**4.90 ± 3.20**	**15.60 ± 9.71**	**7.80 ± 5.59**	**35.40 ± 12.33**	**8.20 ± 7.54**
	F1↑	CAM	0.67 ± 0.08	0.68 ± 0.08	0.61 ± 0.08	0.66 ± 0.08	0.60 ± 0.09	0.71 ± 0.08
		PEP w/ RF	0.39 ± 0.07	0.39 ± 0.07	0.39 ± 0.07	0.39 ± 0.07	0.35 ± 0.10	0.41 ± 0.07
		PEP w/ XGB	0.74 ± 0.09	0.76 ± 0.08	0.73 ± 0.10	0.76 ± 0.08	0.66 ± 0.10	0.78 ± 0.07
		PEP w/ TabPFN	**0.90 ± 0.05**	**0.94 ± 0.04**	**0.90 ± 0.08**	**0.93 ± 0.04**	**0.83 ± 0.06**	**0.95 ± 0.03**
	Precision↑	CAM	0.96 ± 0.04	1.00 ± 0.02	0.98 ± 0.02	0.99 ± 0.02	0.87 ± 0.07	0.97 ± 0.04
		PEP w/ RF	0.37 ± 0.07	0.32 ± 0.07	0.31 ± 0.07	0.31 ± 0.07	0.39 ± 0.10	0.35 ± 0.08
		PEP w/ XGB	0.79 ± 0.09	0.82 ± 0.07	0.81 ± 0.09	0.84 ± 0.07	0.72 ± 0.10	0.85 ± 0.06
		PEP w/ TabPFN	**0.95 ± 0.03**	**0.97 ± 0.02**	**0.94 ± 0.05**	**0.97 ± 0.02**	**0.91 ± 0.05**	**0.98 ± 0.02**
	Recall↑	CAM	0.50 ± 0.08	0.52 ± 0.08	0.45 ± 0.09	0.51 ± 0.08	0.50 ± 0.11	0.59 ± 0.10
		PEP w/ RF	0.51 ± 0.09	0.58 ± 0.10	0.56 ± 0.09	0.56 ± 0.09	0.43 ± 0.13	0.53 ± 0.11
		PEP w/ XGB	0.70 ± 0.12	0.72 ± 0.13	0.64 ± 0.14	0.70 ± 0.12	0.65 ± 0.11	0.77 ± 0.12
		PEP w/ TabPFN	**0.92 ± 0.09**	**0.97 ± 0.05**	**0.88 ± 0.10**	**0.95 ± 0.05**	**0.88 ± 0.08**	**0.97 ± 0.05**
SynSF	SHD↓	CAM	8.10 ± 3.81	7.80 ± 4.61	9.70 ± 4.64	7.50 ± 4.28	10.10 ± 3.57	6.60 ± 3.84
		PEP w/ RF	7.70 ± 3.31	7.70 ± 3.18	7.70 ± 3.39	7.70 ± 3.18	9.40 ± 4.06	7.60 ± 3.06
		PEP w/ XGB	6.30 ± 3.36	6.30 ± 2.99	6.30 ± 3.38	6.30 ± 2.99	9.20 ± 4.33	6.00 ± 3.06
		PEP w/ TabPFN	**3.10 ± 3.03**	**1.40 ± 1.63**	**2.80 ± 3.29**	**1.20 ± 1.30**	**6.60 ± 3.98**	**1.40 ± 1.71**
	SID↓	CAM	24.80 ± 11.47	18.30 ± 9.49	24.30 ± 8.72	17.40 ± 8.17	38.80 ± 12.35	16.40 ± 9.78
		PEP w/ RF	44.80 ± 13.18	42.20 ± 12.62	44.50 ± 13.33	42.20 ± 12.62	56.80 ± 15.29	41.20 ± 11.47
		PEP w/ XGB	21.60 ± 8.25	19.30 ± 7.42	20.80 ± 8.47	19.40 ± 7.85	36.90 ± 11.76	15.90 ± 9.11
		PEP w/ TabPFN	**11.80 ± 12.88**	**3.20 ± 3.41**	**10.30 ± 10.12**	**3.10 ± 2.85**	**27.30 ± 10.61**	**3.10 ± 7.58**
	F1↑	CAM	0.78 ± 0.12	0.79 ± 0.15	0.74 ± 0.15	0.80 ± 0.14	0.69 ± 0.13	0.83 ± 0.12
		PEP w/ RF	0.83 ± 0.08	0.88 ± 0.08	0.84 ± 0.08	0.88 ± 0.08	0.77 ± 0.09	0.89 ± 0.08
		PEP w/ XGB	0.89 ± 0.09	0.93 ± 0.06	0.90 ± 0.08	0.93 ± 0.06	0.80 ± 0.11	0.94 ± 0.05
		PEP w/ TabPFN	**0.91 ± 0.09**	**0.96 ± 0.05**	**0.91 ± 0.08**	**0.96 ± 0.05**	**0.80 ± 0.11**	**0.96 ± 0.05**
	Precision↑	CAM	**0.96 ± 0.07**	**0.96 ± 0.05**	**0.96 ± 0.09**	**0.99 ± 0.03**	**0.82 ± 0.11**	**0.96 ± 0.06**
		PEP w/ RF	0.95 ± 0.06	0.94 ± 0.06	0.95 ± 0.07	0.96 ± 0.05	0.81 ± 0.11	0.95 ± 0.06
		PEP w/ XGB	0.95 ± 0.06	0.95 ± 0.05	0.95 ± 0.06	0.96 ± 0.05	0.82 ± 0.10	0.96 ± 0.05
		PEP w/ TabPFN	0.95 ± 0.06	0.96 ± 0.05	0.95 ± 0.06	0.97 ± 0.05	0.82 ± 0.09	0.97 ± 0.05
	Recall↑	CAM	0.67 ± 0.16	0.69 ± 0.19	0.62 ± 0.18	0.70 ± 0.19	0.61 ± 0.16	0.75 ± 0.18
		PEP w/ RF	0.72 ± 0.12	0.80 ± 0.10	0.74 ± 0.12	0.80 ± 0.10	0.71 ± 0.11	0.83 ± 0.11
		PEP w/ XGB	0.76 ± 0.12	0.84 ± 0.08	0.79 ± 0.11	0.84 ± 0.08	0.73 ± 0.09	0.89 ± 0.08
		PEP w/ TabPFN	**0.92 ± 0.09**	**0.97 ± 0.05**	**0.90 ± 0.09**	**0.97 ± 0.05**	**0.82 ± 0.09**	**0.97 ± 0.05**

Table H.12: Scenario comparison (SynER). Cells report mean ± std (two decimals). Bold marks the better mean per backbone within each metric (lower is better for SHD/SID; higher is better otherwise).

Scenario	Metric	Pruning	CAM	SCORE	DAS	NoGAM	DiffAN	CaPS
Vanilla	SHD↓	CAM pruning	17.90 ± 3.84	17.10 ± 3.81	19.80 ± 3.79	17.30 ± 3.74	18.40 ± 4.79	15.80 ± 4.26
		PEP	**14.40 ± 3.65**	**6.40 ± 2.95**	**9.60 ± 3.44**	**7.40 ± 4.83**	**14.80 ± 4.60**	**5.50 ± 2.51**
	SID↓	CAM pruning	34.90 ± 7.28	31.30 ± 5.46	39.40 ± 6.08	31.40 ± 5.66	50.10 ± 11.40	28.30 ± 6.53
		PEP	**31.70 ± 10.56**	**13.70 ± 7.88**	**22.00 ± 9.33**	**13.40 ± 8.53**	**46.60 ± 14.09**	**9.90 ± 7.18**
	F1↑	CAM pruning	0.67 ± 0.08	0.68 ± 0.08	0.61 ± 0.08	0.66 ± 0.08	0.60 ± 0.09	0.71 ± 0.08
		PEP	**0.76 ± 0.06**	**0.90 ± 0.04**	**0.84 ± 0.07**	**0.89 ± 0.05**	**0.71 ± 0.07**	**0.92 ± 0.03**
	Precision↑	CAM pruning	0.96 ± 0.04	1.00 ± 0.02	0.98 ± 0.02	0.99 ± 0.02	0.87 ± 0.07	0.97 ± 0.04
		PEP	**0.77 ± 0.05**	**0.95 ± 0.03**	**0.92 ± 0.05**	**0.95 ± 0.03**	**0.81 ± 0.06**	**0.97 ± 0.02**
	Recall↑	CAM pruning	0.50 ± 0.08	0.52 ± 0.08	0.45 ± 0.09	0.51 ± 0.08	0.50 ± 0.11	0.59 ± 0.10
		PEP	**0.70 ± 0.08**	**0.87 ± 0.06**	**0.80 ± 0.10**	**0.86 ± 0.07**	**0.66 ± 0.07**	**0.88 ± 0.05**
PNL	SHD↓	CAM pruning	21.40 ± 4.40	21.20 ± 4.18	23.30 ± 4.11	20.90 ± 4.10	22.40 ± 5.28	19.50 ± 4.37
		PEP	**18.00 ± 3.77**	**8.60 ± 2.54**	**12.10 ± 2.85**	**9.70 ± 3.46**	**18.10 ± 4.87**	**7.00 ± 2.22**
	SID↓	CAM pruning	40.10 ± 10.28	38.20 ± 8.68	47.00 ± 8.07	38.60 ± 9.55	56.00 ± 12.85	35.40 ± 8.50
		PEP	**39.40 ± 11.57**	**18.50 ± 6.66**	**29.90 ± 9.35**	**20.50 ± 7.40**	**49.50 ± 8.90**	**15.20 ± 6.32**
	F1↑	CAM pruning	0.61 ± 0.09	0.62 ± 0.09	0.55 ± 0.09	0.62 ± 0.09	0.57 ± 0.12	0.64 ± 0.09
		PEP	**0.68 ± 0.08**	**0.86 ± 0.06**	**0.80 ± 0.08**	**0.84 ± 0.07**	**0.67 ± 0.08**	**0.89 ± 0.05**
	Precision↑	CAM pruning	0.94 ± 0.04	0.99 ± 0.02	0.97 ± 0.02	0.99 ± 0.02	0.84 ± 0.08	0.96 ± 0.04
		PEP	**0.76 ± 0.06**	**0.94 ± 0.03**	**0.91 ± 0.05**	**0.94 ± 0.03**	**0.81 ± 0.06**	**0.97 ± 0.02**
	Recall↑	CAM pruning	0.43 ± 0.08	0.45 ± 0.08	0.38 ± 0.09	0.45 ± 0.08	0.43 ± 0.11	0.51 ± 0.09
		PEP	**0.66 ± 0.08**	**0.84 ± 0.06**	**0.77 ± 0.10**	**0.81 ± 0.07**	**0.64 ± 0.07**	**0.87 ± 0.05**
LinGAM	SHD↓	CAM pruning	10.20 ± 2.56	9.90 ± 2.58	12.00 ± 2.79	9.90 ± 2.37	12.00 ± 3.25	7.90 ± 3.26
		PEP	**8.60 ± 2.32**	**3.80 ± 2.18**	**6.00 ± 2.38**	**3.90 ± 2.14**	**8.70 ± 2.68**	**2.80 ± 1.54**
	SID↓	CAM pruning	24.20 ± 7.19	21.60 ± 6.52	26.10 ± 6.59	21.40 ± 6.33	33.10 ± 10.41	18.40 ± 6.35
		PEP	**22.90 ± 7.69**	**9.10 ± 4.21**	**16.40 ± 5.71**	**9.60 ± 4.42**	**30.30 ± 10.57**	**7.80 ± 4.50**
	F1↑	CAM pruning	0.78 ± 0.08	0.80 ± 0.08	0.76 ± 0.08	0.81 ± 0.07	0.72 ± 0.09	0.83 ± 0.06
		PEP	**0.86 ± 0.06**	**0.93 ± 0.04**	**0.89 ± 0.06**	**0.91 ± 0.05**	**0.80 ± 0.06**	**0.94 ± 0.03**
	Precision↑	CAM pruning	0.93 ± 0.05	0.97 ± 0.03	0.96 ± 0.04	0.98 ± 0.03	0.79 ± 0.08	0.95 ± 0.04
		PEP	**0.84 ± 0.06**	**0.95 ± 0.04**	**0.92 ± 0.05**	**0.94 ± 0.04**	**0.79 ± 0.06**	**0.96 ± 0.03**
	Recall↑	CAM pruning	0.66 ± 0.10	0.69 ± 0.11	0.62 ± 0.12	0.69 ± 0.11	0.64 ± 0.12	0.74 ± 0.10
		PEP	**0.88 ± 0.06**	**0.96 ± 0.04**	**0.90 ± 0.09**	**0.94 ± 0.04**	**0.83 ± 0.06**	**0.96 ± 0.04**
Confounded	SHD↓	CAM pruning	25.20 ± 4.60	24.60 ± 4.59	26.60 ± 4.52	24.50 ± 4.31	25.80 ± 5.64	22.80 ± 4.82
		PEP	**22.40 ± 4.10**	**10.00 ± 2.63**	**14.10 ± 2.72**	**11.70 ± 3.01**	**22.10 ± 5.13**	**9.10 ± 2.35**
	SID↓	CAM pruning	49.50 ± 12.41	45.20 ± 11.03	55.80 ± 9.78	45.30 ± 10.31	65.10 ± 14.66	41.20 ± 10.08
		PEP	**49.30 ± 14.97**	**23.80 ± 8.61**	**37.30 ± 10.31**	**27.30 ± 9.11**	**58.40 ± 11.49**	**22.50 ± 8.53**
	F1↑	CAM pruning	0.56 ± 0.09	0.57 ± 0.10	0.51 ± 0.10	0.58 ± 0.10	0.52 ± 0.12	0.62 ± 0.10
		PEP	**0.62 ± 0.08**	**0.83 ± 0.06**	**0.77 ± 0.08**	**0.81 ± 0.07**	**0.63 ± 0.08**	**0.86 ± 0.05**
	Precision↑	CAM pruning	0.90 ± 0.05	0.96 ± 0.03	0.94 ± 0.04	0.97 ± 0.03	0.82 ± 0.08	0.94 ± 0.04
		PEP	**0.72 ± 0.06**	**0.93 ± 0.04**	**0.90 ± 0.05**	**0.93 ± 0.04**	**0.79 ± 0.06**	**0.96 ± 0.03**
	Recall↑	CAM pruning	0.44 ± 0.09	0.45 ± 0.10	0.40 ± 0.10	0.47 ± 0.10	0.46 ± 0.12	0.53 ± 0.11
		PEP	**0.74 ± 0.07**	**0.89 ± 0.06**	**0.83 ± 0.09**	**0.86 ± 0.07**	**0.68 ± 0.07**	**0.88 ± 0.05**
Measure-Err	SHD↓	CAM pruning	20.00 ± 4.43	19.80 ± 4.24	22.70 ± 4.33	19.50 ± 4.30	21.70 ± 4.83	18.50 ± 4.29
		PEP	**18.30 ± 4.10**	**8.40 ± 2.31**	**12.50 ± 3.04**	**9.40 ± 3.40**	**18.70 ± 4.81**	**7.40 ± 2.06**
	SID↓	CAM pruning	52.50 ± 12.19	49.50 ± 10.66	60.20 ± 8.45	49.90 ± 12.28	63.89 ± 8.07	44.80 ± 10.08
		PEP	**49.70 ± 12.33**	**24.80 ± 8.70**	**38.70 ± 9.58**	**27.60 ± 8.65**	**60.00 ± 19.43**	**48.30 ± 9.78**
	F1↑	CAM pruning	0.51 ± 0.11	0.52 ± 0.10	0.45 ± 0.10	0.52 ± 0.10	0.47 ± 0.12	0.60 ± 0.12
		PEP	**0.57 ± 0.10**	**0.79 ± 0.07**	**0.73 ± 0.09**	**0.76 ± 0.07**	**0.59 ± 0.12**	**0.81 ± 0.08**
	Precision↑	CAM pruning	0.88 ± 0.05	0.94 ± 0.03	0.92 ± 0.04	0.95 ± 0.03	0.78 ± 0.07	0.92 ± 0.04
		PEP	**0.66 ± 0.08**	**0.91 ± 0.05**	**0.88 ± 0.06**	**0.90 ± 0.05**	**0.67 ± 0.12**	**0.89 ± 0.08**
	Recall↑	CAM pruning	0.39 ± 0.10	0.40 ± 0.10	0.34 ± 0.11	0.41 ± 0.10	0.41 ± 0.11	0.48 ± 0.12
		PEP	**0.51 ± 0.13**	**0.75 ± 0.08**	**0.69 ± 0.14**	**0.73 ± 0.08**	**0.55 ± 0.20**	**0.73 ± 0.14**
Non-i.i.d	SHD↓	CAM pruning	9.70 ± 2.33	9.40 ± 2.39	11.40 ± 2.62	9.30 ± 2.33	11.20 ± 3.02	7.40 ± 2.67
		PEP	**8.40 ± 2.50**	**3.60 ± 2.13**	**5.80 ± 2.13**	**3.80 ± 2.14**	**8.20 ± 2.65**	**2.90 ± 1.46**
	SID↓	CAM pruning	24.70 ± 7.56	22.00 ± 6.80	26.40 ± 6.82	21.80 ± 6.42	32.40 ± 10.56	18.10 ± 6.17
		PEP	**23.60 ± 7.71**	**9.20 ± 4.18**	**16.30 ± 5.38**	**9.60 ± 4.53**	**29.80 ± 10.84**	**8.10 ± 4.64**
	F1↑	CAM pruning	0.78 ± 0.08	0.80 ± 0.08	0.76 ± 0.07	0.81 ± 0.07	0.72 ± 0.08	0.83 ± 0.06
		PEP	**0.86 ± 0.07**	**0.93 ± 0.04**	**0.89 ± 0.07**	**0.91 ± 0.05**	**0.80 ± 0.06**	**0.94 ± 0.03**
	Precision↑	CAM pruning	0.93 ± 0.05	0.97 ± 0.03	0.96 ± 0.04	0.98 ± 0.03	0.79 ± 0.07	0.95 ± 0.04
		PEP	**0.84 ± 0.06**	**0.95 ± 0.04**	**0.92 ± 0.05**	**0.94 ± 0.04**	**0.79 ± 0.06**	**0.96 ± 0.03**
	Recall↑	CAM pruning	0.66 ± 0.10	0.68 ± 0.10	0.62 ± 0.10	0.69 ± 0.10	0.64 ± 0.11	0.73 ± 0.10
		PEP	**0.88 ± 0.06**	**0.96 ± 0.04**	**0.90 ± 0.09**	**0.94 ± 0.04**	**0.83 ± 0.06**	**0.96 ± 0.04**
Unfaithful	SHD↓	CAM pruning	22.50 ± 3.57	22.10 ± 3.44	24.50 ± 3.71	22.20 ± 3.64	24.10 ± 4.37	20.60 ± 3.61
		PEP	**19.70 ± 3.08**	**9.70 ± 2.45**	**13.10 ± 2.67**	**11.10 ± 3.12**	**19.30 ± 3.90**	**8.50 ± 2.14**
	SID↓	CAM pruning	66.00 ± 11.04	60.90 ± 10.38	75.20 ± 9.63	59.80 ± 11.29	83.10 ± 13.72	55.50 ± 9.33
		PEP	**60.80 ± 12.44**	**29.40 ± 9.20**	**44.60 ± 9.67**	**33.60 ± 9.69**	**72.20 ± 12.47**	**27.60 ± 8.65**
	F1↑	CAM pruning	0.48 ± 0.07	0.50 ± 0.07	0.43 ± 0.07	0.50 ± 0.07	0.46 ± 0.10	0.57 ± 0.09
		PEP	**0.55 ± 0.07**	**0.79 ± 0.06**	**0.72 ± 0.09**	**0.77 ± 0.07**	**0.57 ± 0.08**	**0.82 ± 0.05**
	Precision↑	CAM pruning	0.86 ± 0.06	0.93 ± 0.04	0.91 ± 0.05	0.94 ± 0.04	0.77 ± 0.07	0.92 ± 0.05
		PEP	**0.68 ± 0.07**	**0.92 ± 0.05**	**0.89 ± 0.06**	**0.91 ± 0.05**	**0.76 ± 0.06**	**0.94 ± 0.04**
	Recall↑	CAM pruning	0.35 ± 0.08	0.36 ± 0.08	0.30 ± 0.09	0.37 ± 0.08	0.36 ± 0.11	0.45 ± 0.10
		PEP	**0.64 ± 0.07**	**0.83 ± 0.06**	**0.77 ± 0.10**	**0.82 ± 0.07**	**0.62 ± 0.07**	**0.85 ± 0.06**

Table H.13: Scenario comparison (SynSF). Cells report mean ± std (two decimals). Bold marks the better mean per backbone within each metric (lower is better for SHD/SID; higher is better otherwise).

Scenario	Metric	Pruning	CAM	SCORE	DAS	NoGAM	DiffAN	CaPS
Vanilla	SHD↓	CAM pruning	10.10 ± 3.90	5.60 ± 2.59	**7.50 ± 3.27**	5.50 ± 2.07	11.10 ± 4.48	5.10 ± 3.41
		PEP	**7.40 ± 1.52**	**3.60 ± 3.05**	9.00 ± 3.54	**2.60 ± 2.41**	**9.40 ± 2.30**	**3.80 ± 2.94**
	SID↓	CAM pruning	41.40 ± 11.38	21.20 ± 7.19	**24.40 ± 8.68**	23.10 ± 12.14	**39.50 ± 12.70**	15.80 ± 8.02
		PEP	**32.40 ± 10.38**	**13.00 ± 7.81**	26.40 ± 9.76	**15.40 ± 18.72**	45.20 ± 13.22	**15.60 ± 12.82**
	F1↑	CAM pruning	0.69 ± 0.11	0.85 ± 0.08	**0.81 ± 0.09**	0.85 ± 0.06	0.70 ± 0.11	0.87 ± 0.09
		PEP	**0.77 ± 0.04**	**0.91 ± 0.07**	0.77 ± 0.10	**0.93 ± 0.08**	**0.71 ± 0.06**	**0.90 ± 0.08**
	Precision↑	CAM pruning	0.74 ± 0.13	**0.91 ± 0.10**	**0.92 ± 0.08**	0.91 ± 0.09	**0.70 ± 0.16**	0.90 ± 0.11
		PEP	**0.77 ± 0.05**	0.91 ± 0.09	0.89 ± 0.10	**0.93 ± 0.08**	0.70 ± 0.07	**0.92 ± 0.08**
	Recall↑	CAM pruning	0.65 ± 0.10	0.80 ± 0.07	**0.72 ± 0.10**	0.80 ± 0.06	0.71 ± 0.08	0.85 ± 0.10
		PEP	**0.77 ± 0.04**	**0.91 ± 0.06**	0.67 ± 0.10	**0.93 ± 0.08**	**0.73 ± 0.05**	**0.88 ± 0.10**
PNL	SHD↓	CAM pruning	18.00 ± 5.40	14.70 ± 3.16	15.10 ± 2.69	13.90 ± 3.51	14.33 ± 4.27	13.20 ± 3.26
		PEP	**13.00 ± 8.04**	**12.00 ± 1.58**	**12.00 ± 1.58**	**10.60 ± 2.30**	**11.00 ± 6.08**	**7.50 ± 2.07**
	SID↓	CAM pruning	58.40 ± 18.40	41.00 ± 8.26	42.50 ± 6.35	36.80 ± 11.17	50.67 ± 16.03	38.70 ± 7.82
		PEP	**36.75 ± 20.25**	**22.40 ± 7.77**	**22.40 ± 7.77**	**19.00 ± 12.41**	**26.00 ± 7.00**	**27.75 ± 8.01**
	F1↑	CAM pruning	0.44 ± 0.18	0.59 ± 0.09	0.56 ± 0.10	0.61 ± 0.11	0.58 ± 0.12	0.64 ± 0.08
		PEP	**0.67 ± 0.21**	**0.73 ± 0.05**	**0.73 ± 0.05**	**0.76 ± 0.06**	**0.73 ± 0.12**	**0.80 ± 0.05**
	Precision↑	CAM pruning	**0.56 ± 0.23**	**0.77 ± 0.12**	**0.83 ± 0.16**	**0.80 ± 0.14**	**0.69 ± 0.15**	0.75 ± 0.11
		PEP	0.62 ± 0.25	0.64 ± 0.04	0.64 ± 0.04	0.67 ± 0.06	0.66 ± 0.14	**0.79 ± 0.09**
	Recall↑	CAM pruning	0.37 ± 0.15	0.48 ± 0.11	0.43 ± 0.10	0.50 ± 0.12	0.51 ± 0.11	0.56 ± 0.09
		PEP	**0.75 ± 0.16**	**0.86 ± 0.06**	**0.86 ± 0.06**	**0.88 ± 0.07**	**0.83 ± 0.07**	**0.80 ± 0.05**
LiNGAM	SHD↓	CAM pruning	28.83 ± 1.83	**4.00 ± 3.23**	6.50 ± 2.88	**4.10 ± 2.51**	19.00 ± 4.57	4.20 ± 2.94
		PEP	**28.33 ± 1.15**	6.40 ± 4.56	**6.40 ± 4.56**	4.60 ± 2.30	**18.60 ± 5.13**	**3.44 ± 3.64**
	SID↓	CAM pruning	73.50 ± 8.69	15.20 ± 14.05	19.60 ± 10.99	15.90 ± 12.25	57.50 ± 6.41	14.70 ± 12.68
		PEP	**65.00 ± 5.29**	**6.80 ± 10.43**	**6.80 ± 10.43**	**5.20 ± 5.40**	**40.80 ± 11.63**	**8.00 ± 9.11**
	F1↑	CAM pruning	0.19 ± 0.04	**0.89 ± 0.09**	0.84 ± 0.08	0.89 ± 0.07	0.51 ± 0.09	0.89 ± 0.08
		PEP	**0.26 ± 0.05**	0.86 ± 0.10	**0.86 ± 0.10**	**0.89 ± 0.06**	**0.58 ± 0.10**	**0.92 ± 0.08**
	Precision↑	CAM pruning	0.17 ± 0.03	**0.89 ± 0.10**	**0.90 ± 0.06**	**0.90 ± 0.08**	0.46 ± 0.10	**0.90 ± 0.10**
		PEP	**0.21 ± 0.04**	0.79 ± 0.13	0.79 ± 0.13	0.84 ± 0.08	**0.49 ± 0.11**	0.89 ± 0.11
	Recall↑	CAM pruning	0.22 ± 0.06	0.90 ± 0.08	0.79 ± 0.10	0.88 ± 0.07	0.59 ± 0.08	0.89 ± 0.07
		PEP	**0.32 ± 0.06**	**0.95 ± 0.05**	**0.95 ± 0.05**	**0.96 ± 0.03**	**0.72 ± 0.08**	**0.94 ± 0.06**
Confounded	SHD↓	CAM pruning	17.20 ± 5.05	13.00 ± 3.92	13.70 ± 3.47	13.80 ± 3.26	18.70 ± 5.23	15.00 ± 3.03
		PEP	**16.00 ± 4.74**	**11.60 ± 3.36**	**11.60 ± 3.36**	**11.40 ± 3.65**	**15.20 ± 4.15**	**8.75 ± 3.95**
	SID↓	CAM pruning	52.00 ± 11.55	32.60 ± 12.94	40.00 ± 10.28	37.00 ± 12.44	51.60 ± 10.71	34.33 ± 6.86
		PEP	**46.60 ± 11.82**	**21.00 ± 14.51**	**21.00 ± 14.51**	**27.60 ± 19.55**	**47.60 ± 13.22**	**26.25 ± 14.52**
	F1↑	CAM pruning	0.53 ± 0.15	0.68 ± 0.12	0.65 ± 0.10	0.65 ± 0.09	0.54 ± 0.13	0.65 ± 0.07
		PEP	**0.60 ± 0.12**	**0.75 ± 0.09**	**0.75 ± 0.09**	**0.72 ± 0.11**	**0.60 ± 0.13**	**0.79 ± 0.10**
	Precision↑	CAM pruning	**0.57 ± 0.18**	**0.71 ± 0.11**	**0.76 ± 0.13**	0.69 ± 0.09	0.52 ± 0.14	0.63 ± 0.07
		PEP	0.57 ± 0.10	0.70 ± 0.09	0.70 ± 0.09	**0.69 ± 0.09**	**0.58 ± 0.10**	**0.87 ± 0.11**
	Recall↑	CAM pruning	0.50 ± 0.14	0.66 ± 0.15	0.58 ± 0.12	0.63 ± 0.11	0.57 ± 0.13	0.68 ± 0.08
		PEP	**0.63 ± 0.17**	**0.81 ± 0.16**	**0.81 ± 0.16**	**0.78 ± 0.17**	**0.63 ± 0.19**	**0.73 ± 0.12**
Measure-Err	SHD↓	CAM pruning	**19.20 ± 1.75**	15.70 ± 2.71	16.50 ± 1.84	15.30 ± 2.58	**18.14 ± 2.79**	**15.83 ± 1.47**
		PEP	19.60 ± 1.14	**14.60 ± 2.51**	**14.60 ± 2.51**	**14.00 ± 3.54**	19.00 ± 2.16	16.00 ± 0.82
	SID↓	CAM pruning	66.60 ± 10.28	49.10 ± 10.33	49.40 ± 7.53	47.90 ± 13.36	**57.71 ± 9.74**	**43.83 ± 9.83**
		PEP	**61.00 ± 9.90**	**40.60 ± 11.44**	**40.60 ± 11.44**	**36.20 ± 15.66**	57.25 ± 13.89	46.25 ± 7.54
	F1↑	CAM pruning	0.33 ± 0.09	0.54 ± 0.10	0.51 ± 0.07	0.55 ± 0.09	**0.45 ± 0.08**	**0.55 ± 0.01**
		PEP	**0.34 ± 0.07**	**0.58 ± 0.09**	**0.58 ± 0.09**	**0.60 ± 0.14**	0.42 ± 0.08	0.51 ± 0.04
	Precision↑	CAM pruning	**0.49 ± 0.14**	**0.78 ± 0.15**	**0.79 ± 0.15**	**0.81 ± 0.15**	**0.57 ± 0.14**	0.68 ± 0.09
		PEP	0.46 ± 0.09	0.76 ± 0.09	0.76 ± 0.09	0.79 ± 0.13	0.52 ± 0.13	**0.82 ± 0.07**
	Recall↑	CAM pruning	0.25 ± 0.07	0.42 ± 0.11	0.39 ± 0.09	0.42 ± 0.10	**0.38 ± 0.07**	**0.47 ± 0.04**
		PEP	**0.27 ± 0.06**	**0.48 ± 0.12**	**0.48 ± 0.12**	**0.49 ± 0.15**	0.35 ± 0.05	0.38 ± 0.03
Non-i.i.d	SHD↓	CAM pruning	15.20 ± 4.66	14.80 ± 4.29	15.60 ± 3.57	15.30 ± 4.30	17.60 ± 5.76	14.50 ± 4.50
		PEP	**13.60 ± 6.07**	**11.80 ± 4.97**	**11.80 ± 4.97**	**10.40 ± 4.67**	**14.80 ± 4.32**	**11.83 ± 2.32**
	SID↓	CAM pruning	53.20 ± 18.34	54.60 ± 21.80	56.70 ± 19.00	54.40 ± 21.16	56.50 ± 17.75	48.90 ± 20.79
		PEP	**32.80 ± 14.11**	**34.80 ± 21.12**	**34.80 ± 21.12**	**34.80 ± 24.16**	**46.00 ± 11.77**	**42.33 ± 12.64**
	F1↑	CAM pruning	0.55 ± 0.15	0.56 ± 0.14	0.54 ± 0.13	0.55 ± 0.14	0.51 ± 0.15	0.60 ± 0.15
		PEP	**0.66 ± 0.17**	**0.71 ± 0.12**	**0.71 ± 0.12**	**0.73 ± 0.12**	**0.62 ± 0.12**	**0.68 ± 0.04**
	Precision↑	CAM pruning	**0.64 ± 0.20**	0.66 ± 0.15	0.66 ± 0.15	0.64 ± 0.16	0.54 ± 0.18	0.64 ± 0.16
		PEP	0.64 ± 0.18	**0.68 ± 0.14**	**0.68 ± 0.14**	**0.69 ± 0.11**	**0.60 ± 0.11**	**0.77 ± 0.09**
	Recall↑	CAM pruning	0.49 ± 0.12	0.50 ± 0.16	0.46 ± 0.14	0.49 ± 0.14	0.48 ± 0.13	0.56 ± 0.16
		PEP	**0.69 ± 0.17**	**0.74 ± 0.13**	**0.74 ± 0.13**	**0.78 ± 0.15**	**0.66 ± 0.15**	**0.61 ± 0.05**
Unfaithful	SHD↓	CAM pruning	11.90 ± 3.84	6.50 ± 0.93	8.30 ± 1.77	6.10 ± 2.08	12.75 ± 1.67	5.10 ± 1.52
		PEP	**11.50 ± 1.91**	**5.80 ± 1.64**	**5.80 ± 1.64**	**5.80 ± 3.03**	**11.25 ± 1.71**	**3.78 ± 1.86**
	SID↓	CAM pruning	45.70 ± 12.22	24.75 ± 8.31	26.90 ± 9.36	25.40 ± 12.59	46.62 ± 12.16	17.20 ± 5.20
		PEP	**30.00 ± 7.62**	**10.20 ± 9.20**	**10.20 ± 9.20**	**14.60 ± 17.97**	**38.75 ± 19.35**	**12.11 ± 7.13**
	F1↑	CAM pruning	0.63 ± 0.11	0.83 ± 0.03	0.79 ± 0.05	0.84 ± 0.07	0.64 ± 0.06	0.87 ± 0.04
		PEP	**0.71 ± 0.04**	**0.87 ± 0.04**	**0.87 ± 0.04**	**0.86 ± 0.09**	**0.69 ± 0.08**	**0.91 ± 0.04**
	Precision↑	CAM pruning	**0.67 ± 0.13**	**0.87 ± 0.05**	**0.89 ± 0.07**	**0.89 ± 0.10**	**0.64 ± 0.05**	0.90 ± 0.04
		PEP	0.65 ± 0.05	0.81 ± 0.05	0.81 ± 0.05	0.80 ± 0.10	0.62 ± 0.06	**0.91 ± 0.05**
	Recall↑	CAM pruning	0.61 ± 0.10	0.80 ± 0.04	0.71 ± 0.05	0.79 ± 0.07	0.66 ± 0.08	0.85 ± 0.05
		PEP	**0.77 ± 0.02**	**0.95 ± 0.03**	**0.95 ± 0.03**	**0.93 ± 0.09**	**0.77 ± 0.12**	**0.92 ± 0.05**