EVICD: STRUCTURED EVIDENTIAL UNCERTAINTY FOR CAUSAL DISCOVERY

Anonymous authors
Paper under double-blind review

ABSTRACT

When causal graphs support downstream decisions, knowing *which predicted relations warrant commitment* can be as important as the graph itself. Existing methods typically return a point estimate or an edge confidence, which does not by itself reveal whether uncertainty reflects weak commitment or incompatible relation-specific assessments. We propose **EviCD**, a graph neural network that predicts per-pair Dempster–Shafer mass functions over four modeled relations: $X_i \to X_j$, $X_j \to X_i$, latent confounding without a direct edge, and no modeled relation. A restricted-focal architecture yields two complementary model-level diagnostics: vacuity $m(\Omega)$, measuring support left uncommitted, and conflict K, measuring incompatibility among channel-level assessments. These quantities support selective prediction, allowing EviCD to abstain on high-vacuity pairs under a retained-accuracy guarantee conditional on calibration. EviCD does not resolve observational non-identifiability; its outputs are simulator-relative assessments. On the main synthetic benchmark, EviCD attains the highest aggregate selective-prediction score among methods with an available selective confidence, while maintaining competitive structural accuracy and providing an explicit continuous confounding score. Controlled perturbations further show that increasing latent-confounder strength primarily changes K, whereas changes in noise, sample size, and signal strength primarily affect $m(\Omega)$.

1 INTRODUCTION

Inferring causal structure from observational data is fundamentally difficult: ground-truth graphs are rarely available, controlled interventions are often infeasible, and multiple structures may induce the same observational distribution (Spirtes et al., 2000; Peters et al., 2017). In this setting, uncertainty quantification is not merely an auxiliary diagnostic. A practitioner needs to know not only which relation a model prefers, but also when the available evidence supports committing to that relation and when the model should abstain.

Existing causal-discovery methods provide only partial support for this decision. Classical constraint-, score-, and functional-based methods typically return a graph or equivalence-class representation, while probabilistic, amortized, or resampling-based methods additionally provide edge probabilities or stability scores. Such scalar confidences are useful, but a single categorical score does not by itself preserve the internal structure of uncertainty. For example, an intermediate confidence may arise because relation-specific assessments are mutually incompatible, or because the observations provide weak support for any particular relation. These situations may motivate different follow-up actions: additional observations may help when support is weak, whereas persistent incompatibility among candidate relations may instead motivate targeted interventions or model checking. Related approaches and their output representations are further detailed in Appendix A.

We propose **EviCD** (Evidential Causal Discovery), an amortized graph neural network that represents each ordered variable pair with a Dempster–Shafer mass function (Dempster, 1967; Shafer, 1976). Under the pairwise relation space considered here, the frame contains four hypotheses: $X_i \to X_j$, $X_j \to X_i$, latent-common-cause confounding without a direct edge, and no modeled relation. Unlike a categorical probability vector, a mass function can assign support to sets of hypotheses in addition to single hypotheses. EviCD uses this structure to retain *vacuity* $m(\Omega)$, the mass left un-

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

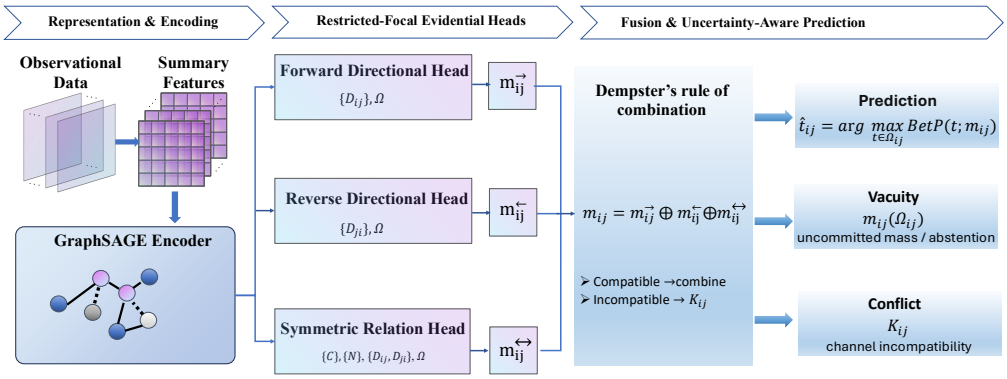

Figure 1: **Overview of EviCD.** A shared GraphSAGE encoder maps observational summary statistics to relation-specific representations for each ordered pair (i, j). Three restricted-focal heads produce directional and symmetric BPAs, which are combined by Dempster fusion into m_{ij}. The fused BPA yields categorical predictions through the pignistic transform, vacuity $m_{ij}(\Omega_{ij})$ for selective abstention, and conflict K_{ij} for channel incompatibility.

committed among the modeled relations, and a *conflict degree* K, an architecture-defined measure of incompatibility among relation-specific channel outputs. Figure 1 summarizes the overall pipeline.

The restricted-focal design is central to EviCD. Two directional heads support orientation-specific relations, while a symmetric head represents confounding, absence, orientation-unresolved direction, or vacuity. Dempster fusion preserves compatible set-valued support while exposing incompatibility between relation-specific commitments. The resulting BPA supports categorical prediction through the pignistic transform and selective abstention through vacuity.

Importantly, EviCD does *not* circumvent observational non-identifiability. The model is trained on simulator-generated graph–data pairs using a simulation-based amortized-learning paradigm, and its outputs are therefore model-level assessments relative to the training distribution (Lorch et al., 2022). The pignistic likelihood supervises the induced categorical predictor but does not uniquely identify the underlying mass decomposition. Likewise, neither $m(\Omega)$ nor K is claimed to be a certificate of causal identifiability. We make these scope restrictions explicit and evaluate the intended behavior of the two diagnostics with controlled perturbations rather than assuming their semantics from the representation alone.

The model is trained once on simulated ER/SF structures with and without latent confounders and is applied to unseen benchmarks without retraining. Our contributions are:

- We formulate pairwise causal-structure prediction using restricted-focal Dempster–Shafer mass functions, retaining set-valued support and vacuity in addition to the categorical relation prediction. We give an information-theoretic interpretation of the induced pignistic predictor while explicitly separating that predictive objective from identification of the BPA decomposition.

- We derive the algebraic decomposition of channel conflict under the restricted focal families and establish a conditional selective-risk guarantee for vacuity-based abstention. The theory characterizes model-level quantities; it does not assume that observational causal equivalence has been resolved.

- Across ten baselines, EviCD remains competitive on structural recovery while providing strong selective prediction and an explicit continuous confounding score. Controlled experiments further show a complementary empirical response: K changes strongly with latent-confounder strength, whereas $m(\Omega)$ responds primarily to noise, sample size, and signal strength.

2 METHOD

2.1 EVIDENTIAL PAIRWISE RELATIONS

Let $\mathbf{X} \in \mathbb{R}^{n \times d}$ denote n i.i.d. observational samples of d observed variables. For each ordered pair (i, j), $i \neq j$, EviCD predicts a basic probability assignment (BPA) over

$$\Omega_{ij} = \{\mathrm{D}_{ij}, \mathrm{D}_{ji}, \mathrm{C}, \mathrm{N}\}, \qquad m_{ij} : 2^{\Omega_{ij}} \to [0, 1], \quad m_{ij}(\emptyset) = 0, \quad \sum_{A \subseteq \Omega_{ij}} m_{ij}(A) = 1. \tag{1}$$

Here D_{ij} and D_{ji} denote the two direct-edge directions, C denotes latent-common-cause confounding in the absence of a direct edge, and N denotes absence of both a direct edge and the modeled latent-confounding relation. Predictions for (i, j) and (j, i) are produced separately; reversal consistency between them is encouraged during training rather than imposed by construction.

The simulator uses a direct-edge-priority labeling rule: when direct causation and latent confounding coexist, the pair is assigned to the corresponding directional class. Bow configurations are therefore included in D_{ij} or D_{ji} rather than represented by a separate joint state.

A focal set A with $|A| > 1$ represents support committed to the set but not to one of its individual hypotheses. Of particular interest is $m_{ij}(\Omega_{ij})$, the mass left *uncommitted* among all modeled relations; $m_{ij}(\Omega_{ij}) = 1$ corresponds to complete vacuity. A categorical distribution is recovered when mass is restricted to singletons, whereas a general BPA can additionally retain composite support. The associated belief and plausibility functions are given in Appendix B.

The frame is a modeling choice rather than an identifiability claim. EviCD is trained relative to a simulator-defined distribution, and observationally equivalent structures may remain indistinguishable. Moreover, the current four-state frame does not distinguish direct-only relations from bows as separate output classes.

2.2 RESTRICTED-FOCAL EVIDENTIAL ARCHITECTURE

Featurization and graph encoding. EviCD computes deterministic marginal and pairwise summaries from $\mathbf{X}$ and constructs a complete message-passing graph over the observed variables. Each node receives 13 marginal statistics, while each ordered pair receives 118 pairwise features including mutual information, HSIC, partial-correlation and separating-set summaries, direction-sensitive residual scores, and V-structure indicators (Appendix D). Thus, the encoder has access to both association statistics and explicit conditional-independence-related cues.

We use a three-layer GraphSAGE-style encoder (Hamilton et al., 2017):

$$\begin{aligned} \mathbf{z}_{uv}^{(k)} &= \sigma\Big(\mathbf{W}_z^{(k)}[\mathbf{h}_u^{(k-1)} \| \mathbf{h}_v^{(k-1)} \| \mathbf{e}_{uv}]\Big), \\ \mathbf{h}_v^{(k)} &= \sigma\Big(\mathbf{W}_h^{(k)}[\mathbf{h}_v^{(k-1)} \| \operatorname{MEAN}_{u \in N(v)} \mathbf{z}_{uv}^{(k)}]\Big). \end{aligned} \tag{2}$$

The weights are shared across variables, and the mean aggregation is permutation invariant. For each ordered pair, we then form two directional representations and one symmetric representation:

$$\begin{aligned} \mathbf{r}_{ij}^{\rightarrow} &= g_{\mathrm{dir}}([\mathbf{h}_i^{(L)} \| \mathbf{h}_j^{(L)} \| \mathbf{e}_{ij}]), \\ \mathbf{r}_{ij}^{\leftarrow} &= g_{\mathrm{dir}}([\mathbf{h}_j^{(L)} \| \mathbf{h}_i^{(L)} \| \mathbf{e}_{ji}]), \\ \mathbf{r}_{ij}^{\leftrightarrow} &= g_{\mathrm{sym}}([\mathbf{h}_i^{(L)} + \mathbf{h}_j^{(L)} \| \, |\mathbf{h}_i^{(L)} - \mathbf{h}_j^{(L)}|]). \end{aligned} \tag{3}$$

The symmetric representation is invariant to swapping i and j, whereas the two directional representations retain orientation-specific information.

Restricted focal families. The three channels are restricted to

$$\begin{aligned} &\mathcal{F}_{\rightarrow} = \{\{\mathrm{D}_{ij}\}, \Omega_{ij}\}, && \mathcal{F}_{\leftarrow} = \{\{\mathrm{D}_{ji}\}, \Omega_{ij}\}, \\ &\mathcal{F}_{\leftrightarrow} = \{\{\mathrm{C}\}, \{\mathrm{N}\}, \{\mathrm{D}_{ij}, \mathrm{D}_{ji}\}, \Omega_{ij}\}. \end{aligned} \tag{4}$$

The directional channels can support one orientation or remain vacuous. The symmetric channel can support the confounding-only state C, absence N, or the composite statement $\{\mathrm{D}_{ij}, \mathrm{D}_{ji}\}$ that

a directed relation exists without resolving its orientation. We use "channel" in an architectural sense: the channels share the same observations and encoder and are not assumed to be statistically independent causal evidence sources.

Each channel maps its representation to evidence through prototype-based simple-support functions. For a prototype p,

$$s_p(\mathbf{r}) = \sigma(\alpha_p)\exp[-\gamma_p^2\|\mathbf{r}-\mathbf{w}_p\|^2]. \tag{5}$$

Multiple same-focus directional supports are combined by a noisy-OR. In the symmetric module, each prototype distributes its non-vacuous support across $\{\mathrm{C}\}$, $\{\mathrm{N}\}$, and $\{\mathrm{D}_{ij},\mathrm{D}_{ji}\}$ using learned nonnegative normalized weights. Full prototype equations are given in Appendix B.

Evidential fusion. Prototype-level and channel-level BPAs are fused with Dempster's normalized conjunctive rule. For two BPAs with $\kappa(m',m'')<1$,

$$(m'\oplus m'')(A)=\frac{\sum_{B\cap C=A} m'(B)m''(C)}{1-\kappa(m',m'')},\qquad \kappa(m',m'')=\sum_{B\cap C=\emptyset} m'(B)m''(C). \tag{6}$$

The focal restrictions make the compatibility structure explicit. For example, $\{\mathrm{D}_{ij}\}\cap\{\mathrm{D}_{ij},\mathrm{D}_{ji}\}=\{\mathrm{D}_{ij}\}$, so orientation-unresolved directional support is compatible with a committed direction, whereas $\{\mathrm{D}_{ij}\}\cap\{\mathrm{C}\}=\emptyset$ creates direction–type incompatibility.

For the three channel-level BPAs, we define K_{ij} as the mass assigned to the empty set by their *unnormalized conjunctive combination* before normalization. This total conflict is invariant to grouping, although intermediate pairwise conflict coefficients in a sequential implementation may depend on the grouping.

Theorem 1 (Channel-conflict decomposition)**.** *Let*

$$a=m_{ij}^{\rightarrow}(\{\mathrm{D}_{ij}\}),\qquad b=m_{ij}^{\leftarrow}(\{\mathrm{D}_{ji}\}),\qquad c=m_{ij}^{\leftrightarrow}(\{\mathrm{C}\}),\qquad \nu=m_{ij}^{\leftrightarrow}(\{\mathrm{N}\}).$$

Under Eq. 4, the total channel conflict is

$$K_{ij}=ab+\big[a(1-b)+(1-a)b\big](c+\nu). \tag{7}$$

The first term is generated by opposing directional commitments, while the second arises from direction–type incompatibility. Mass on $\{\mathrm{D}_{ij},\mathrm{D}_{ji}\}$ or Ω_{ij} does not itself create additional conflict with a single directional commitment.

Theorem 1 gives K_{ij} a precise *model-level* interpretation as incompatibility among the restricted-focal channel outputs. It does not imply that K_{ij} uniquely identifies an underlying causal mechanism. Section 3.3 therefore evaluates its empirical behavior under controlled structural perturbations. The proof, fusion-order details, and additional architectural properties are provided in Appendix C.

2.3 Pignistic Learning with Evidential Regularization

For categorical prediction, a BPA induces the pignistic distribution (Smets, 2005),

$$\mathrm{BetP}(t;m)=\sum_{A\ni t}\frac{m(A)}{|A|}. \tag{8}$$

The implementation predicts a separate BPA for every ordered pair. Let

$$\mathcal{P}_d=\{(i,j):i\neq j\},\qquad |\mathcal{P}_d|=d(d-1).$$

Given simulator-provided relation labels t_{ij}^*, we minimize the mean pignistic negative log-likelihood

$$\mathcal{L}_{\mathrm{NLL}}=-\mathbb{E}_{p(G,\mathbf{X})}\left[\frac{1}{|\mathcal{P}_d|}\sum_{i\neq j}\log\mathrm{BetP}\Big(t_{ij}^*;m_{ij}^{f_\phi(\mathbf{X})}\Big)\right]. \tag{9}$$

For a fixed graph size d, let J be sampled uniformly and independently from $\mathcal{P}_d$, let $T^*=t_J^*$, and define

$$q_\phi(T\mid\mathbf{X},J)=\mathrm{BetP}(T;m_J).$$

Then the mean NLL admits the conditional mutual-information lower bound

$$I(T^*; \mathbf{X} \mid J) \geq H(T^* \mid J) - \mathcal{L}_{\mathrm{NLL}}(\phi). \tag{10}$$

Training mixes graph sizes; Eq. 10 applies conditional on each fixed d, while the implemented objective averages the corresponding ordered-pair cross-entropies across sampled instances.

Importantly, Eq. 10 concerns only the *pignistic categorical predictor*. The mapping $m \mapsto \mathrm{BetP}(m)$ is many-to-one, so predictive likelihood does not uniquely identify the underlying BPA decomposition and composite masses do not enlarge the categorical variational family. The particular decomposition learned by EviCD is instead shaped by the restricted-focal parameterization and auxiliary regularizers rather than identified by the predictive likelihood alone. Its empirical interpretation is therefore evaluated separately. Appendix C.2 gives the formal result, proof, and a simple many-to-one example.

We additionally use an error-aware vacuity regularizer

$$\mathcal{L}_{\mathrm{cal}} = \frac{1}{|\mathcal{P}_d|} \sum_{i \neq j} \mathbb{1}[\hat{t}_{ij} \neq t^*_{ij}] \left(1 - m_{ij}(\Omega_{ij})\right), \tag{11}$$

where $\hat{t}_{ij} = \arg\max_{t \in \Omega_{ij}} \mathrm{BetP}(t; m_{ij})$. The indicator acts as a non-differentiable per-batch mask; gradients from this term flow through $m_{ij}(\Omega_{ij})$ rather than through the argmax prediction. Consequently, $\mathcal{L}_{\mathrm{cal}}$ encourages misclassified pairs to receive greater vacuity, but does not by itself establish calibration or causal non-identifiability.

We further use a reversal-consistency regularizer $\mathcal{L}_{\mathrm{sym}}$ that couples the independently predicted directional evidence for (i, j) and (j, i) under index reversal. Specifically, it compares the forward evidence parameters for (i, j) with the reverse evidence parameters for (j, i), and vice versa. The total objective is

$$\mathcal{L}_{\mathrm{total}} = \mathcal{L}_{\mathrm{NLL}} + \lambda_{\mathrm{cal}} \mathcal{L}_{\mathrm{cal}} + \lambda_{\mathrm{sym}} \mathcal{L}_{\mathrm{sym}}. \tag{12}$$

All three terms are averaged over the $d(d-1)$ ordered-pair outputs. The reversal-consistency term regularizes the two independently predicted ordered representations; it should not be interpreted as an additional independent source of causal evidence.

2.4 Inference and Selective Prediction

The full *Evidential Graph* retains every predicted BPA together with its belief–plausibility bounds and vacuity. When a single categorical relation is required, the *Maximum Pignistic Graph* uses

$$\hat{t}_{ij} = \arg\max_{t \in \Omega_{ij}} \mathrm{BetP}(t; m_{ij}). \tag{13}$$

For selective prediction, EviCD retains this prediction only when

$$m_{ij}(\Omega_{ij}) < \theta, \qquad \text{and returns ABSTAIN otherwise.} \tag{14}$$

The threshold θ is an inference-time operating point rather than a training hyperparameter. Writing

$$C(\theta) = \Pr[m_{ij}(\Omega_{ij}) < \theta] \tag{15}$$

for coverage, decreasing θ retains fewer pairs and therefore lowers $C(\theta)$. Our experiments summarize the resulting retained-accuracy–coverage curve rather than selecting a single operating threshold.

If an acyclic graph is required, an optional post-processing mode projects the retained directed predictions onto a maximum-weight acyclic subgraph. This projection is not used as evidence for latent confounding; the complete inference variants are described in Appendix B.

To characterize vacuity-based selection, let

$$M = m_{ij}(\Omega_{ij}), \qquad R(\theta) = \Pr[\hat{t}_{ij} \neq t^*_{ij} \mid M < \theta], \qquad A(\theta) = 1 - R(\theta).$$

Theorem 2 (Retained-accuracy monotonicity under vacuity calibration)**.** *Let $g : [0, 1] \to [0, 1]$ be non-decreasing. Suppose that, for every θ satisfying $\Pr(M < \theta) > 0$,*

$$|R(\theta) - \mathbb{E}[g(M) \mid M < \theta]| \leq \varepsilon. \tag{16}$$

Then, for any $\theta_2 < \theta_1$ with $\Pr(M < \theta_2) > 0$,

$$A(\theta_2) \geq A(\theta_1) - 2\varepsilon. \tag{17}$$

Thus, under the stated cumulative vacuity–error calibration condition, stricter vacuity thresholding cannot reduce retained-set accuracy by more than 2ε. When $\varepsilon = 0$, retained accuracy is non-decreasing as coverage is reduced. This is a conditional guarantee: the training loss encourages useful vacuity ordering but does not itself establish Eq. 16. The formal definition and proof are given in Appendix C.5.

3 EXPERIMENTS

3.1 EXPERIMENTAL SETUP

We organize the experiments around three questions: **(Q1)** does EviCD remain competitive on structural recovery while adding selective prediction and an explicit continuous confounding score? **(Q2)** do vacuity and conflict respond differently to controlled changes in data informativeness and latent-confounder strength? **(Q3)** do these behaviors transfer beyond the training simulator?

Datasets and training. For the main synthetic comparison, we generate Erdős–Rényi (ER) and scale-free (SF) graphs with $d \in \{10, 20\}$, expected degree in $\{2, 4\}$, and $n = 1000$ observations. Each configuration is evaluated with three random seeds, giving eight structural settings and 24 test instances in total. Mechanisms are linear or random-Fourier-feature (RFF) structural equations with Gaussian, Laplace, or Student-t additive noise; test instances may also contain latent common causes according to the simulator protocol. We separately evaluate scalability at $d \in \{50, 100\}$.

External evaluation uses Sachs (Sachs et al., 2005), DREAM4 (Marbach et al., 2009), and Smith fMRI (Smith et al., 2011). EviCD is trained once on simulated ER/SF graph–data pairs with $d \in \{10, 20, 50, 100\}$ and $n \in \{500, 1000, 2000\}$, and is applied to all test benchmarks without retraining. Full simulator, relation-label, and preprocessing details are given in Appendix E; training details are in Appendix G.

Baselines and metrics. We compare against PC, FCI, GES, GRaSP, NOTEARS, DAGMA, LiNGAM, stability selection, AVICI, and DAG-GNN (Spirtes et al., 2000; 1995; Chickering, 2002; Lam et al., 2022; Zheng et al., 2018; Bello et al., 2022; Shimizu et al., 2006; Meinshausen & Bühlmann, 2010; Lorch et al., 2022; Yu et al., 2019). These span constraint-, score-, optimization-, functional-, resampling-, and amortized neural approaches.

We report relation-presence AUROC, macro-F1, and structural Hamming distance (SHD) for structural recovery. For confounding detection, we report confounding AUROC (C-AUC) where a confounding decision score is available; score granularity differs across methods, so these values are interpreted with care. Selective performance is summarized by $\mathrm{AUC}_{\mathrm{sel}}$, the area under the retained-accuracy–coverage curve, for methods with a continuous selective score. We additionally report top-label ECE for methods with an available continuous confidence mapping. Exact output mappings, score definitions, and thresholding procedures are given in Appendix F.

3.2 STRUCTURAL AND SELECTIVE PREDICTION

Main synthetic benchmark. Table 1 summarizes the eight synthetic settings with $d \in \{10, 20\}$. EviCD essentially matches AVICI in relation-presence AUROC (0.804 vs. 0.805), while other methods are stronger on some point-estimate structure metrics. In particular, DAGMA obtains the highest macro-F1 and lowest SHD, while EviCD achieves the second-lowest aggregate SHD. EviCD therefore does not uniformly dominate the evaluated baselines in structural recovery. Its additional value is that the same forward pass also produces a continuous confounding score, vacuity for selective abstention, and a channel-conflict diagnostic.

Under the method-specific confidence mappings defined in Appendix F, EviCD obtains lower top-label ECE (0.069 ± 0.010) than AVICI (0.106 ± 0.023) and stability selection (0.288 ± 0.040). This comparison concerns calibration of the induced decision confidence rather than calibration or identification of the underlying BPA decomposition.

For selective prediction, EviCD obtains $\mathrm{AUC}_{\mathrm{sel}} = 0.727 \pm 0.035$, compared with 0.714 ± 0.030 for AVICI and 0.291 ± 0.035 for stability selection. EviCD therefore attains the highest aggregate selective score, although its margin over AVICI is modest. At the individual-setting level, EviCD

Table 1: Main synthetic benchmark (mean ± SE across eight settings). "—" denotes an unavailable output. Best in **bold** and second best underlined for directly comparable columns. C-AUC is shown without ranking because score granularity differs across methods. Additional metrics and per-setting results are reported in Appendix I.

Method	AUROC↑	F1↑	SHD↓	C-AUC↑	AUC_{sel}↑
PC	$0.645_{\pm.022}$	$0.327_{\pm.013}$	$194_{\pm45}$	—	—
FCI	$0.597_{\pm.031}$	$0.295_{\pm.036}$	$199_{\pm47}$	$0.531_{\pm.027}$	—
GES	$0.716_{\pm.033}$	$0.294_{\pm.010}$	$288_{\pm75}$	—	—
GRaSP	$0.763_{\pm.022}$	$0.292_{\pm.012}$	$236_{\pm58}$	—	—
DAGMA	$0.349_{\pm.035}$	$\mathbf{0.578}_{\pm.016}$	$\mathbf{136}_{\pm39}$	—	—
LiNGAM	$0.365_{\pm.023}$	$0.375_{\pm.009}$	$266_{\pm65}$	—	—
NOTEARS	$0.399_{\pm.026}$	$0.470_{\pm.013}$	$204_{\pm51}$	—	—
Stab. sel.	$0.786_{\pm.036}$	$0.307_{\pm.010}$	$272_{\pm70}$	—	$0.291_{\pm.035}$
AVICI†	$\mathbf{0.805}_{\pm.021}$	$0.476_{\pm.020}$	$255_{\pm68}$	$0.539_{\pm.029}$	$0.714_{\pm.030}$
DAG-GNN	$0.585_{\pm.025}$	$0.539_{\pm.015}$	$195_{\pm52}$	—	—
EviCD	$0.804_{\pm.035}$	$0.447_{\pm.020}$	$173_{\pm46}$	$0.629_{\pm.058}$	$\mathbf{0.727}_{\pm.035}$

†AVICI does not natively represent latent confounding. Our evaluation proxy is binary: a pair is labeled confounded only when both directional edge scores exceed 0.2. FCI also provides a discrete confounding indication. Their C-AUC values are therefore coarse decision-level references and are not directly comparable to the continuous ranking induced by EviCD's $\text{BetP}(\text{C}; m_{ij})$; see Appendix F.

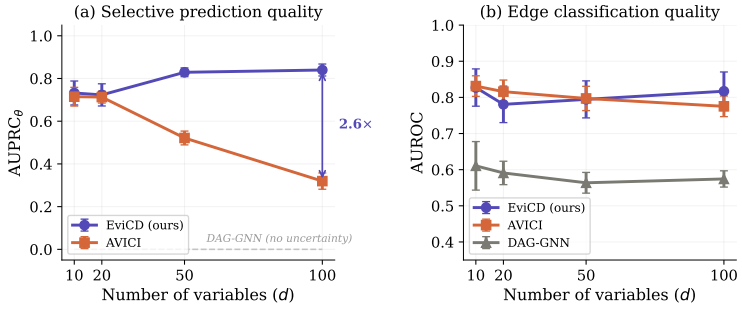

Figure 2: Scaling from $d = 10$ to 100 (mean ± SE over ER/SF × degree$\in \{2, 4\}$ with three seeds). **Left**: AUC_{sel}. **Right**: relation-presence AUROC.

and AVICI each lead in four of the eight $d \leq 20$ configurations (Appendix I), indicating that neither method uniformly dominates across settings.

Scaling with graph size. Figure 2 extends the comparison to $d = 100$. At $d = 10$, EviCD and AVICI show similar selective performance (approximately 0.73 and 0.71, respectively). With increasing graph size, their selective scores diverge in the evaluated simulation regime: EviCD remains above 0.73 and reaches approximately 0.83 at $d = 100$, whereas AVICI decreases to approximately 0.31. Both methods retain relation-presence AUROC above 0.75 across the evaluated graph sizes. We interpret this as an empirical scaling result under the present simulator and evaluation procedure rather than evidence of a fundamental limitation of probabilistic uncertainty.

3.3 Empirical Behavior of Vacuity and Conflict

EviCD exposes two model-level quantities in addition to its pignistic decision probabilities: vacuity $m(\Omega)$, measuring support left uncommitted to specific hypotheses, and channel conflict K, measuring incompatibility among the restricted-focal channel outputs. Neither quantity is assumed to be a certificate of causal identifiability. We therefore evaluate their empirical behavior directly, asking whether K is associated with prediction difficulty, whether latent-confounder strength preferentially affects K, and whether changes in data informativeness preferentially affect $m(\Omega)$.

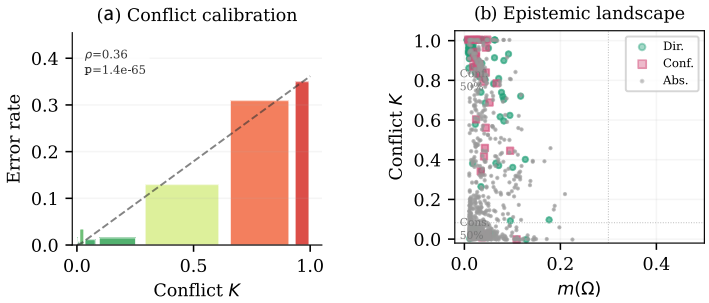

Figure 3: Behavior of EviCD's evidential diagnostics on synthetic ER graphs with $d = 20$, degree $= 2$ (5320 evaluated pairs). **Left**: empirical prediction error across bins of channel conflict K. **Right**: evaluated pairs in the $(m(\Omega), K)$ plane, colored by simulator-provided relation label. These plots characterize learned model-level diagnostics and do not imply causal identifiability from either quantity alone.

Table 2: Controlled perturbations of vacuity and channel conflict. Arrows show endpoint values; full trajectories and the $\lambda_{\text{cal}} = 0$ ablation are reported in Appendix H.

Perturbation	$m(\Omega)$	K
Confounder strength $0 \to 4$	$.021 \to .030$	$.14 \to .76$
Noise $\sigma : .2 \to 2.0$	$.0174 \to .0209$	$.203 \to .184$
Samples $n : 500 \to 2000$	$.0213 \to .0194$	$.216 \to .197$
Signal $w : .1 \to 2.0$	$.029 \to .015$	variation $\leq .08$

Conflict and prediction difficulty. On the diagnostic pool used in Figure 3, prediction error increases with K (Spearman $\rho = 0.36$, $p < 1.4 \times 10^{-65}$). This empirical association is consistent with Theorem 1, under which incompatible channel commitments contribute directly to K. The theorem itself is algebraic, however, and the observed correlation does not establish that high conflict uniquely corresponds to an underlying causal ambiguity.

Controlled dissociation. To probe the two diagnostics under controlled changes, we keep the SCM family, architecture, and training procedure fixed and vary one generative factor at a time. Table 2 summarizes the endpoint behavior.

Increasing latent-confounder strength raises K from 0.14 to approximately 0.76 across the sweep ($\Delta = 0.63$, Spearman $\rho = 0.94$, $p = 0.005$), whereas $m(\Omega)$ ranges over only 0.014. Conversely, vacuity increases overall as observation noise grows, decreases with increasing sample size, and decreases as causal signal strength increases on present edges, while the corresponding variation in K remains comparatively small. The noise trend is not strictly monotonic at every tested value, but its overall association with vacuity is positive (Spearman $\rho = 0.90$, $p = 0.037$).

Importantly, the qualitative relationships persist when the model is retrained with $\lambda_{\text{cal}} = 0$: the rank correlations are $+1.0$ between noise and vacuity, -1.0 between sample size and vacuity, and $+0.94$ between confounder strength and K. Thus, the observed empirical dissociation is not solely induced by the vacuity regularizer.

Evidential landscape and scope. Figure 3 (right) provides a complementary view of the learned diagnostics. Approximately 92% of pairs labeled C in this panel lie in the high-conflict region under the visualization thresholds used in the figure. This enrichment is compatible with the symmetric channel supporting the confounding-only state while directional channels retain incompatible directional support. It should not be interpreted as a confounding classifier based on K: the quadrant boundaries are descriptive, high conflict is not specific to latent confounding, and bow configurations receive directional labels under the simulator's direct-edge-priority rule.

Appendix H further reports a Gaussian/Laplace identifiability stress test. Mean vacuity is 0.0183 in the linear-Gaussian non-identifiable regime and 0.0152 in the Laplace identifiable control. Although

Table 3: External-benchmark results for EviCD (best/second-best in bold/underline); full comparisons are in Appendix J.

Dataset	AUROC↑	SHD↓	AUC_{sel}↑
Sachs	0.440	**36**	**0.746**
DREAM4	**0.750**	**22**	<u>0.699</u>
fMRI	0.884	<u>12.0</u>	**0.617**

the difference is in the expected direction, it is small. We therefore do not interpret $m(\Omega)$ as a certificate of observational non-identifiability; the current singleton supervision can itself suppress vacuity in such regimes.

Taken together, these experiments support a narrower operational interpretation: $m(\Omega)$ is useful for ranking the degree of model commitment, whereas K records incompatibility among relation-specific channel outputs. The two diagnostics exhibit complementary empirical responses, but their causal meaning remains relative to the learned architecture and training distribution.

3.4 External Benchmark Transfer

We finally examine whether reliability ranking learned from the synthetic ER/SF training distribution remains useful outside that simulator. Table 3 summarizes three heterogeneous benchmarks; EviCD is applied without retraining, and complete baseline results are reported in Appendix J.

Transfer is metric-dependent. On Sachs, EviCD obtains the lowest SHD and highest selective score despite weak relation-presence AUROC. On DREAM4, it achieves the highest AUROC and lowest SHD and ranks second to stability selection in AUC_{sel}. On Smith fMRI, LiNGAM is substantially stronger in point-estimate structural recovery, whereas EviCD obtains the second-lowest SHD and highest selective score. Thus, selective ranking remains useful across these benchmarks, but neither structural nor selective performance is uniformly dominant.

4 Conclusion and Limitations

We introduced **EviCD**, an amortized causal-discovery model that represents each ordered variable pair with a Dempster–Shafer mass function. Its restricted-focal architecture yields vacuity $m(\Omega)$, measuring uncommitted support, and channel conflict K, measuring incompatibility among relation-specific assessments, alongside the pignistic categorical predictor.

On the main synthetic benchmark, EviCD remains competitive in structural recovery and attains the highest aggregate selective-prediction score, although its margin over AVICI is modest and the two methods each lead in four of eight settings. Selective performance remains strong up to $d = 100$. Controlled perturbations show complementary empirical responses: latent-confounder strength primarily affects K, whereas noise, sample size, and causal-signal strength primarily affect vacuity. External results are mixed in structural accuracy but suggest that selective ranking can remain useful outside the ER/SF training simulator.

Several limitations qualify these results. Neither $m(\Omega)$ nor K is a certificate of causal identifiability, and singleton simulator labels can suppress vacuity in observationally non-identifiable regimes. The direct-edge-priority supervision also prevents the four-state output space from distinguishing direct-only relations from bows as separate atomic states. Finally, both diagnostics are simulator- and architecture-relative and may change under distribution shift. Future work should therefore explore equivalence-aware or set-valued supervision, explicit bow states, and broader training distributions including interventions, selection effects, measurement error, feedback, and temporal structure.

AI Assistance Disclosure

Generative AI tools were used to assist with language editing and polishing of the manuscript, including improvements to clarity, grammar, organization, and presentation. Generative AI was also used to assist in developing and refining parts of the synthetic-data generation workflow. The

synthetic datasets themselves were generated programmatically using the causal simulator specified in the paper, including the graph families, structural equation mechanisms, noise distributions, and latent-confounding process.

All scientific content, mathematical claims, experimental results, interpretations, simulator specifications, generation code, and final text were reviewed and verified by the authors. The authors take full responsibility for the final content of this work.

Ethics Statement

This work introduces a methodological approach and evaluates it on synthetic data and previously published benchmark datasets. We collect no new human-subject data and make no clinical or policy decisions from the learned graphs.

Reproducibility Statement

The appendix specifies the model architecture, proofs, feature construction, simulator, metrics, training protocol, controlled perturbations, and per-setting results needed to reproduce the reported analyses. The code will be available after acceptance.

References

Kevin Bello, Bryon Aragam, and Pradeep Ravikumar. DAGMA: Learning DAGs via M-matrices and a log-determinant acyclicity characterization. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

David Maxwell Chickering. Learning equivalence classes of bayesian-network structures. *Journal of machine learning research*, 2(Feb):445–498, 2002.

Chris Cundy, Aditya Grover, and Stefano Ermon. BCD Nets: Scalable variational approaches for Bayesian causal discovery. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Arthur P Dempster. Upper and lower probabilities induced by a multivalued mapping. *The Annals of Mathematical Statistics*, 38(2):325–339, 1967.

Thierry Denœux. A neural network classifier based on Dempster-Shafer theory. *IEEE Transactions on Systems, Man, and Cybernetics-Part A*, 30(2):131–150, 2000.

Thierry Denœux. Conjunctive and disjunctive combination of belief functions induced by nondistinct bodies of evidence. *Artificial Intelligence*, 172(2–3):234–264, 2008.

William L Hamilton, Rex Ying, and Jure Leskovec. Inductive representation learning on large graphs. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Wai-Yin Lam, Bryan Andrews, and Joseph Ramsey. Greedy relaxations of the sparsest permutation algorithm. In *Proceedings of the 38th Conference on Uncertainty in Artificial Intelligence*, pp. 1052–1062. PMLR, 2022.

Lars Lorch, Jonas Rothfuss, Bernhard Schölkopf, and Andreas Krause. DiBS: Differentiable Bayesian structure learning. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Lars Lorch, Scott Sussex, Jonas Rothfuss, Andreas Krause, and Bernhard Schölkopf. Amortized inference for causal structure learning. *Advances in Neural Information Processing Systems*, 35: 13104–13118, 2022.

Daniel Marbach, Thomas Schaffter, Claudio Mattiussi, and Dario Floreano. Generating realistic in silico gene networks for performance assessment of reverse engineering methods. *Journal of Computational Biology*, 16(2):229–239, 2009.

Marie-Hélène Masson and Thierry Denœux. ECM: An evidential version of the fuzzy c-means algorithm. *Pattern Recognition*, 41(4):1384–1397, 2008.

Nicolai Meinshausen and Peter Bühlmann. Stability selection. *Journal of the Royal Statistical Society: Series B*, 72(4):417–473, 2010.

Jonas Peters and Peter Bühlmann. Structural intervention distance for evaluating causal graphs. *Neural Computation*, 27(3):771–799, 2015.

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

Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems*, volume 20, 2007.

Karen Sachs, Omar Perez, Dana Pe'er, Douglas A Lauffenburger, and Garry P Nolan. Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, 308(5721): 523–529, 2005.

Natalie Schluter. On maximum spanning DAG algorithms for semantic DAG parsing. In *Proceedings of the ACL 2014 Workshop on Semantic Parsing*, pp. 61–65, 2014.

Murat Sensoy, Lance Kaplan, and Melih Kandemir. Evidential deep learning to quantify classification uncertainty. *Advances in Neural Information Processing Systems*, 31, 2018.

Glenn Shafer. *A Mathematical Theory of Evidence*. Princeton University Press, Princeton, NJ, 1976.

Shohei Shimizu, Patrik O Hoyer, Aapo Hyvärinen, Antti Kerminen, and Michael Jordan. A linear non-Gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7(10): 2003–2030, 2006.

Philippe Smets. Decision making in the TBM: the necessity of the pignistic transformation. *International Journal of Approximate Reasoning*, 38(2):133–147, 2005.

Stephen M Smith, Karla L Miller, Gholamreza Salimi-Khorshidi, Matthew Webster, Christian F Beckmann, Thomas E Nichols, Joseph D Ramsey, and Mark W Woolrich. Network modelling methods for fMRI. *NeuroImage*, 54(2):875–891, 2011.

Peter Spirtes, Christopher Meek, and Thomas Richardson. Causal inference in the presence of latent variables and selection bias. In *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, pp. 499–506, 1995.

Peter Spirtes, Clark N Glymour, and Richard Scheines. *Causation, Prediction, and Search*. MIT Press, Cambridge, MA, 2nd edition, 2000.

Yue Yu, Jie Chen, Tian Gao, and Mo Yu. DAG-GNN: DAG structure learning with graph neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, pp. 7154–7163. PMLR, 2019.

Xun Zheng, Bryon Aragam, Pradeep K Ravikumar, and Eric P Xing. DAGs with NO TEARS: Continuous optimization for structure learning. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

A Related Work

A.1 Causal Discovery Methods

We organize prior work by the form of structural output and the type of uncertainty information it exposes. Methods used as experimental baselines are described in §3.1; additional uncertainty-aware methods are discussed here for context.

Constraint-, score-, and functional-based methods. Constraint-based methods use conditional-independence information to recover equivalence-class structure. PC (Spirtes et al., 2000) assumes causal sufficiency, whereas FCI (Spirtes et al., 1995) returns a partial ancestral graph that accommodates latent confounding and selection effects through endpoint marks. These outputs encode structural ambiguity but do not provide a continuous confidence score for each relation.

Score-based methods optimize a structural objective over candidate graphs. GES (Chickering, 2002) searches over equivalence classes, while GRaSP (Lam et al., 2022) performs greedy search over variable orderings. Functional approaches such as LiNGAM (Shimizu et al., 2006) exploit distributional asymmetries, including non-Gaussianity, to orient causal relations. These methods primarily return a structural estimate rather than a calibrated per-pair uncertainty representation.

Continuous-optimization methods. NOTEARS (Zheng et al., 2018) formulates DAG learning as continuous optimization under an acyclicity constraint. DAGMA (Bello et al., 2022) uses a log-determinant characterization of acyclicity, and DAG-GNN (Yu et al., 2019) combines structural-equation learning with graph neural networks. These approaches can achieve strong structural recovery but do not natively expose relation-level abstention scores or latent-confounding hypotheses in the form used by EviCD.

Amortized neural causal discovery. AVICI (Lorch et al., 2022) learns an amortized causal-discovery model from simulated graph–data pairs and produces edge probabilities in a single forward pass. EviCD adopts the same broad simulation-based amortization paradigm but predicts Dempster–Shafer mass functions over a four-state pairwise relation frame. Training is still driven by the categorical predictor induced through the pignistic transformation; as shown in Proposition 1, the predictive objective acts at the BetP level and does not uniquely identify the underlying BPA. The additional contribution of EviCD is therefore not a larger categorical variational family, but a structured parameterization that retains uncommitted mass and architecture-defined channel incompatibility alongside the categorical relation prediction.

A.2 Uncertainty in Causal Discovery

Bayesian and resampling approaches. Bayesian causal-discovery methods such as DiBS (Lorch et al., 2021) and BCD Nets (Cundy et al., 2021) represent uncertainty through distributions over graphs or marginal edge probabilities. Such posterior summaries are informative about structural uncertainty, but a single marginal confidence does not by itself retain the distinction between support left uncommitted and incompatibility among relation-specific assessments that is explicitly represented by EviCD. More expressive probabilistic models could in principle encode richer uncertainty structure; our comparison is therefore specifically to the standard marginal confidence outputs exposed by these methods.

Stability selection (Meinshausen & Bühlmann, 2010) quantifies how often a relation is recovered across resampled datasets. In our experiments, GES is used as the base learner. Selection frequency provides an empirical stability score that can support abstention, but it does not separately encode relation-type incompatibility within a single fitted model.

Evidential learning. Dempster–Shafer theory has been used in classification (Denœux, 2000; Sensoy et al., 2018), clustering (Masson & Denœux, 2008), and evidence fusion (Denœux, 2008). These works establish mass-function representations and evidential combination mechanisms, but are not designed for the pairwise relation space and symmetry structure considered here.

EviCD adapts evidential modeling to this setting through restricted focal families for forward, reverse, and symmetric relation channels. The symmetric channel is exactly invariant to exchang-

Table 4: Comparison of output capabilities for the causal-discovery methods used in our experiments. "Confounding" denotes whether the method exposes a native or directly mappable latent-confounding representation; AVICI is shown separately because its reported confounding result uses an evaluation proxy rather than a native output.

Method	Relation output	Score / uncertainty	Confounding	Selective	Channel conflict
PC / GES / GRaSP	dir., abs.	Binary / score output	✗	✗	✗
FCI	PAG marks	Discrete marks	✓[§]	✗	✗
NOTEARS / DAGMA	dir., abs.	Edge weights	✗	✗	✗
LiNGAM	dir., abs.	Structural estimate	✗	✗	✗
DAG-GNN	dir., abs.	Edge weights	✗	✗	✗
AVICI	dir., abs.	$P(e_{ij} \mid \mathbf{X})$	Proxy[‡]	✓[†]	✗
Stability sel.	dir., abs.	Selection frequency	✗	✓[†]	✗
EviCD	dir., conf., abs.	$m : 2^{\Omega} \to [0, 1]$	✓	✓	✓

[†] Selective prediction is obtained by thresholding the method's continuous confidence or stability score.
[‡] AVICI does not natively contain a confounding hypothesis; our evaluation uses the binary rule $\mathbb{1}[p_{ij} > 0.2 \wedge p_{ji} > 0.2]$; see Appendix F.
[§] FCI represents latent-variable ambiguity through PAG endpoint marks; our evaluation maps the relevant marks to a discrete confounding indicator.

ing the two variables, whereas the directional channels retain orientation-specific information. The resulting channel-level BPAs are fused with Dempster's rule, and acyclicity can optionally be imposed at inference time through the selective-DAG projection described in Appendix B. A separate reversal-consistency regularizer encourages the independently predicted directional evidence for (i, j) and (j, i) to agree under index reversal.

Table 4 summarizes the output capabilities of the methods used in our empirical comparison.

B Architecture and Additional Inference Details

B.1 Belief-Function Quantities

For a BPA m on frame Ω, the belief and plausibility functions are

$$\mathrm{Bel}(A) = \sum_{B \subseteq A} m(B), \qquad \mathrm{Pl}(A) = \sum_{B \cap A \neq \emptyset} m(B). \tag{18}$$

For any probability measure compatible with the belief function, $\mathrm{Bel}(A) \leq P(A) \leq \mathrm{Pl}(A)$ (Shafer, 1976). The special case in which all mass is assigned to singletons recovers an ordinary categorical distribution. A general BPA can additionally retain support on composite hypotheses or the full frame; this representational freedom does not imply that the decomposition is identified by observational data or by the pignistic likelihood.

B.2 Encoder and Prototype Mapping

Each node receives 13 marginal summary features and each ordered pair receives 118 pairwise features (Appendix D). The GraphSAGE-style encoder uses

$$\mathbf{z}_{uv}^{(k)} = \sigma\Big(\mathbf{W}_z^{(k)}[\mathbf{h}_u^{(k-1)} \| \mathbf{h}_v^{(k-1)} \| \mathbf{e}_{uv}]\Big), \tag{19}$$

$$\mathbf{h}_v^{(k)} = \sigma\Big(\mathbf{W}_h^{(k)}\left[\mathbf{h}_v^{(k-1)} \middle\| \mathrm{MEAN}_{u \in N(v)}\, \mathbf{z}_{uv}^{(k)}\right]\Big). \tag{20}$$

For each ordered pair,

$$\mathbf{r}_{ij}^{\rightarrow} = g_{\mathrm{dir}}([\mathbf{h}_i^{(L)} \| \mathbf{h}_j^{(L)} \| \mathbf{e}_{ij}]), \tag{21}$$

$$\mathbf{r}_{ij}^{\leftarrow} = g_{\mathrm{dir}}([\mathbf{h}_j^{(L)} \| \mathbf{h}_i^{(L)} \| \mathbf{e}_{ji}]), \tag{22}$$

$$\mathbf{r}_{ij}^{\leftrightarrow} = g_{\mathrm{sym}}([\mathbf{h}_i^{(L)} + \mathbf{h}_j^{(L)} \| |\mathbf{h}_i^{(L)} - \mathbf{h}_j^{(L)}|]). \tag{23}$$

Table 5: Qualitative model-level interpretations of vacuity and channel conflict. These are operational heuristics, not causal-identifiability guarantees.

$m(\Omega)$	K	**State**	**Possible response**
Low	Low	Committed agreement	Retain prediction
Low	High	Channel incompatibility	Inspect competing hypotheses
High	Low	High vacuity	Abstain; seek more evidence
High	High	Vacuous and incompatible	Abstain; inspect mismatch

Each prototype p has center $\mathbf{w}_p$, amplitude α_p, and bandwidth γ_p,

$$s_p(\mathbf{r}) = \sigma(\alpha_p)\exp[-\gamma_p^2\|\mathbf{r}-\mathbf{w}_p\|^2]. \tag{24}$$

For a directional channel, activations combine as $s(\mathbf{r}) = 1-\prod_p(1-s_p(\mathbf{r}))$, yielding singleton mass s and vacuous mass $1-s$.

For the symmetric channel, prototype p distributes its evidence strength over $\{\mathrm{C}\}$, $\{\mathrm{N}\}$, and $\{\mathrm{D}_{ij}, \mathrm{D}_{ji}\}$ using

$$u_{p,k} = \frac{\beta_{p,k}^2}{\sum_l \beta_{p,l}^2}, \qquad m_p(A_k) = u_{p,k}s_p, \qquad m_p(\Omega_{ij}) = 1-s_p. \tag{25}$$

Prototype BPAs are aggregated using Dempster's normalized conjunctive rule. For two BPAs satisfying $\kappa(m', m'') < 1$,

$$(m' \oplus m'')(A) = \frac{\sum_{B\cap C=A} m'(B)m''(C)}{1-\kappa(m',m'')}, \qquad \kappa(m',m'') = \sum_{B\cap C=\emptyset} m'(B)m''(C). \tag{26}$$

Conflict arising while aggregating prototypes within a single head is absorbed by that head's Dempster normalization and is not included in K_{ij}. The diagnostic K_{ij} is defined only from the three channel-level BPAs after within-head aggregation.

The three channel-level BPAs are then fused by the same operator. Equivalently, K_{ij} can be defined as the empty-set mass of their unnormalized conjunctive combination prior to the final normalization. This total channel conflict is invariant to grouping, although the intermediate pairwise conflict coefficients in a sequential implementation can depend on the chosen grouping.

B.3 Additional Inference Modes

The main paper focuses on pignistic prediction and vacuity-based selective prediction. The full BPA also supports the following outputs.

Evidential graph. Retain each m_{ij} together with Bel/Pl intervals and $m_{ij}(\Omega_{ij})$ without forcing a categorical decision.

Maximum-pignistic graph. Return $\hat{t}_{ij} = \arg\max_t \mathrm{BetP}(t; m_{ij})$ for every pair. This collapses the evidential representation to a point prediction.

Selective DAG. After vacuity-based abstention, retained directed predictions can be projected onto a maximum-weight acyclic subgraph (Schluter, 2014). A directed prediction removed by the projection is assigned C only if its pre-projection confounding belief independently exceeds a chosen threshold; otherwise it is marked ABSTAIN. Acyclicity violation is therefore not itself treated as evidence of confounding.

C Proofs and Formal Definitions

C.1 Cumulative Vacuity–Error Calibration

The selective-prediction result in Theorem 2 is conditional on the vacuity score ordering prediction difficulty. We formalize this requirement below. This condition concerns the relationship between

the scalar vacuity score $m(\Omega)$ and empirical prediction error; it should not be confused with calibration of the full mass function or with standard confidence calibration of $\mathrm{BetP}(m)$.

Definition 1 (Cumulative ε-vacuity calibration)**.** *Let*

$$M_{ij} := m_{ij}(\Omega_{ij}), \tag{27}$$

and let $g : [0,1] \to [0,1]$ be a non-decreasing reference function relating vacuity to prediction error. We say that f_ϕ satisfies cumulative ε-vacuity calibration with respect to g *if, for every $\theta \in (0,1]$ such that $\Pr[M_{ij} < \theta] > 0$,*

$$\left|\Pr\left[\hat{t}_{ij} \neq t^*_{ij} \,\middle|\, M_{ij} < \theta\right] - \mathbb{E}\left[g(M_{ij}) \,|\, M_{ij} < \theta\right]\right| \leq \varepsilon. \tag{28}$$

The probability and expectation are taken over graph–data pairs $(G, \mathbf{X}) \sim p(G, \mathbf{X})$ and the evaluated variable pairs.

Definition 1 requires the retained-set error at each vacuity threshold to track, up to tolerance ε, the average value of a non-decreasing error-reference function g. Consequently, lower-vacuity subsets correspond to no larger reference error in expectation, which is the ordering property used in Theorem 2.

This is a sufficient condition for the selective-risk guarantee rather than a property established by the training objective. In particular, $\mathcal{L}_{\mathrm{cal}}$ encourages prediction errors to receive larger $m(\Omega)$ but does not imply Eq. 28. The empirical selective-prediction experiments in §3.3 evaluate whether the learned vacuity score exhibits the required ordering behavior.

C.2 Proof of Proposition 1 (Information-Theoretic Interpretation of the Pignistic Predictor)

Proposition 1. *Fix a graph size d and let $\mathcal{P}_d = \{(i,j) : i \neq j\}$. Let J be sampled uniformly and independently from $\mathcal{P}_d$, and define the categorical predictor*

$$q_\phi(T \mid \mathbf{X}, J) = \mathrm{BetP}\Big(T; m_J^{f_\phi(\mathbf{X})}\Big). \tag{29}$$

Then the mean ordered-pair negative log-likelihood

$$\mathcal{L}_{\mathrm{NLL}}(\phi) = \mathbb{E}_{p(G,\mathbf{X})}\left[-\frac{1}{|\mathcal{P}_d|}\sum_{i\neq j}\log \mathrm{BetP}\Big(t^*_{ij}; m_{ij}^{f_\phi(\mathbf{X})}\Big)\right] \tag{30}$$

satisfies

$$I(T^*; \mathbf{X} \mid J) \geq H(T^* \mid J) - \mathcal{L}_{\mathrm{NLL}}(\phi). \tag{31}$$

Since $H(T^* \mid J)$ is independent of ϕ, minimizing the mean NLL maximizes this variational lower bound. When training mixes graph sizes, Eq. 31 applies conditional on each fixed d, while the implemented objective averages the corresponding ordered-pair cross-entropies across sampled training instances.

Proof. For a fixed ordered pair index J, the conditional mutual information can be written as

$$\begin{aligned} I(T^*; \mathbf{X} \mid J) &= H(T^* \mid J) - H(T^* \mid \mathbf{X}, J) \\ &= H(T^* \mid J) + \mathbb{E}\left[\log p(T^* \mid \mathbf{X}, J)\right]. \end{aligned} \tag{32}$$

Adding and subtracting the log predictive density gives

$$\begin{aligned} I(T^*; \mathbf{X} \mid J) &= H(T^* \mid J) + \mathbb{E}\left[\log q_\phi(T^* \mid \mathbf{X}, J)\right] \\ &\quad + \mathbb{E}_{\mathbf{X},J}\left[D_{\mathrm{KL}}(p(T^* \mid \mathbf{X}, J) \| q_\phi(T^* \mid \mathbf{X}, J))\right] \\ &\geq H(T^* \mid J) + \mathbb{E}\left[\log q_\phi(T^* \mid \mathbf{X}, J)\right] \\ &= H(T^* \mid J) - \mathcal{L}_{\mathrm{NLL}}(\phi), \end{aligned} \tag{33}$$

where the inequality follows from non-negativity of the KL divergence and the last equality follows because J is uniform over the $d(d-1)$ ordered pairs. The bound is tight when $q_\phi(T \mid \mathbf{X}, J) = p(T \mid \mathbf{X}, J)$ almost surely. □

Remark 1 (What the pignistic objective identifies)**.** *Proposition 1 concerns the categorical predictor induced by* BetP*, rather than the underlying BPA itself. The mapping*

$$m \longmapsto \mathrm{BetP}(m) \tag{34}$$

is many-to-one: distinct mass functions can induce identical pignistic probabilities and therefore have identical predictive likelihood.

For example, on a four-element frame, both the vacuous BPA $m_1(\Omega) = 1$ *and the singleton-only BPA* $m_2(\{t\}) = 1/4$ *for every* $t \in \Omega$ *induce the same uniform pignistic distribution* $\mathrm{BetP}(t) = 1/4$*. Thus, composite or vacuous masses do not enlarge the categorical variational family appearing in Proposition 1, nor are they uniquely identified by predictive likelihood. Their decomposition in EviCD is instead induced by the restricted-focal parameterization and auxiliary regularization, and its empirical interpretation is evaluated separately.*

C.3 Proof of Proposition 2 (Architectural Properties)

Proposition 2. *The EviCD architecture satisfies the following structural properties:*

(i) sample-permutation invariance and variable-permutation equivariance;

(ii) symmetric-channel invariance: $m_{ij}^{(\leftrightarrow)} = m_{ji}^{(\leftrightarrow)}$*;*

(iii) focal-set scope restriction: each channel assigns zero mass to focal elements outside its predefined scope;

(iv) compatibility-based composite resolution under Dempster intersection.

Proof. Property (i). The node and edge features are permutation-invariant functions of the multiset of observations: reordering the n samples does not change statistics such as entropy, skewness, mutual information, partial correlation, or HSIC. The message-passing layers use shared parameters and permutation-invariant mean aggregation, so the resulting node representations are also invariant to sample ordering.

For variable permutations, let π be a permutation of the d observed variables. Applying π to the columns of $\mathbf{X}$ consistently permutes node features, pairwise features, and graph indices. Because the same feature maps and message-passing functions are applied to all nodes and pairs, the output assessments are permuted accordingly. Hence f_ϕ is permutation equivariant with respect to the variable axis.

Property (ii). The symmetric representation is

$$\mathbf{r}_{ij}^{\leftrightarrow} = g_{\mathrm{sym}}\left([\mathbf{h}_i^{(L)} + \mathbf{h}_j^{(L)} \| |\mathbf{h}_i^{(L)} - \mathbf{h}_j^{(L)}|]\right). \tag{35}$$

Both the sum and element-wise absolute difference are invariant to exchanging i and j, so

$$\mathbf{r}_{ij}^{\leftrightarrow} = \mathbf{r}_{ji}^{\leftrightarrow}. \tag{36}$$

Since the symmetric DST module is deterministic with shared parameters,

$$m_{ij}^{(\leftrightarrow)} = m_{ji}^{(\leftrightarrow)}. \tag{37}$$

In particular, its mass assigned to $\{\mathrm{C}\}$ is identical for the two orderings.

This statement concerns the symmetric channel itself. This exact invariance concerns the symmetric channel itself. The directional channels are predicted separately for (i, j) and (j, i); their agreement under reversal is encouraged by $\mathcal{L}_{\mathrm{sym}}$ in Section 2.3. No exact symmetry of the complete fused BPA is assumed by this proposition. *Property (iii).* By construction, the forward directional module has focal family

$$\mathcal{F}_{\rightarrow} = \{\{\mathrm{D}_{ij}\}, \Omega_{ij}\}, \tag{38}$$

and therefore assigns zero mass to all other focal elements. The reverse module analogously assigns mass only to $\{\mathrm{D}_{ji}\}$ and Ω_{ij}.

The symmetric module has focal family

$$\mathcal{F}_{\leftrightarrow} = \{\{\mathrm{C}\}, \{\mathrm{N}\}, \{\mathrm{D}_{ij}, \mathrm{D}_{ji}\}, \Omega_{ij}\}, \tag{39}$$

so no parameter or computation path assigns singleton mass directly to $\{\mathrm{D}_{ij}\}$ or $\{\mathrm{D}_{ji}\}$. Thus the focal-set scope restrictions hold identically by construction.

Property (iv). Suppose the fused directional BPA assigns positive mass to $\{\mathrm{D}_{ij}\}$ and the symmetric channel assigns positive mass to the composite $\{\mathrm{D}_{ij}, \mathrm{D}_{ji}\}$. Their intersection is

$$\{\mathrm{D}_{ij}\} \cap \{\mathrm{D}_{ij}, \mathrm{D}_{ji}\} = \{\mathrm{D}_{ij}\}, \tag{40}$$

which is non-empty. The corresponding product therefore contributes to the singleton $\{\mathrm{D}_{ij}\}$ rather than to conflict. The same argument holds for $\{\mathrm{D}_{ji}\}$. Hence a composite expressing "directed relation with unresolved orientation" can be refined to a singleton direction by compatible directional support. □

C.4 Proof of Theorem 1 (Algebraic Decomposition of Channel Conflict)

Theorem 1. Let

$$a = m_{ij}^{(\rightarrow)}(\{\mathrm{D}_{ij}\}), \qquad b = m_{ij}^{(\leftarrow)}(\{\mathrm{D}_{ji}\}), \tag{41}$$

$$c = m_{ij}^{(\leftrightarrow)}(\{\mathrm{C}\}), \qquad \nu = m_{ij}^{(\leftrightarrow)}(\{\mathrm{N}\}). \tag{42}$$

Under the restricted focal families $\mathcal{F}_{\rightarrow}$, $\mathcal{F}_{\leftarrow}$, and $\mathcal{F}_{\leftrightarrow}$, let K_{ij} denote the mass assigned to $\emptyset$ by the unnormalized conjunctive combination of the three channel-level BPAs. Then

$$K_{ij} = ab + \big[a(1-b) + (1-a)b\big](c + \nu). \tag{43}$$

Proof. The forward and reverse directional BPAs assign masses

$$m_{ij}^{(\rightarrow)}(\{\mathrm{D}_{ij}\}) = a, \qquad m_{ij}^{(\rightarrow)}(\Omega_{ij}) = 1 - a, \tag{44}$$

$$m_{ij}^{(\leftarrow)}(\{\mathrm{D}_{ji}\}) = b, \qquad m_{ij}^{(\leftarrow)}(\Omega_{ij}) = 1 - b. \tag{45}$$

Write the remaining symmetric-channel masses as

$$u = m_{ij}^{(\leftrightarrow)}(\{\mathrm{D}_{ij}, \mathrm{D}_{ji}\}), \qquad v = m_{ij}^{(\leftrightarrow)}(\Omega_{ij}), \tag{46}$$

so that

$$c + \nu + u + v = 1. \tag{47}$$

The total conjunctive conflict is the sum of all products of channel-level focal masses whose three-way intersection is empty.

First, simultaneous singleton commitments to opposite directions are incompatible regardless of the symmetric-channel focal element:

$$\{\mathrm{D}_{ij}\} \cap \{\mathrm{D}_{ji}\} = \emptyset. \tag{48}$$

Summing over all symmetric-channel focal elements therefore contributes

$$ab(c + \nu + u + v) = ab. \tag{49}$$

Second, if exactly one directional channel commits to a singleton while the other remains vacuous, conflict occurs precisely when the symmetric channel commits to either $\{\mathrm{C}\}$ or $\{\mathrm{N}\}$. The corresponding contribution is

$$\big[a(1-b) + (1-a)b\big](c + \nu). \tag{50}$$

If both directional channels are vacuous, no symmetric focal element creates an empty intersection. Likewise, $\{\mathrm{D}_{ij}, \mathrm{D}_{ji}\}$ and Ω_{ij} are compatible with either single directional commitment and therefore create no additional conflict.

Adding the two incompatible cases gives

$$K_{ij} = ab + \big[a(1-b) + (1-a)b\big](c + \nu), \tag{51}$$

which proves Eq. 43. □

This expression is the total empty-set mass before normalization and is therefore defined even in the limiting case of complete directional opposition, $a = b = 1$, for which $K_{ij} = 1$. Normalized Dempster fusion is itself undefined when total conflict equals one.

C.5 Proof of Theorem 2 (Retained-Accuracy Monotonicity)

Theorem 2. Let

$$M_{ij} := m_{ij}(\Omega_{ij}), \tag{52}$$

and define the retained-set risk

$$R(\theta) = \Pr\left[\hat{t}_{ij} \neq t^*_{ij} \mid M_{ij} < \theta\right], \tag{53}$$

with retained-set accuracy

$$A(\theta) = 1 - R(\theta). \tag{54}$$

Suppose f_ϕ is ε-calibrated with respect to a non-decreasing function $g : [0,1] \to [0,1]$ in the sense of Definition 1. Then, for any $0 \leq \theta_2 < \theta_1 \leq 1$ satisfying $\Pr[M_{ij} < \theta_2] > 0$,

$$A(\theta_2) \geq A(\theta_1) - 2\varepsilon. \tag{55}$$

Proof. Define

$$G(\theta) = \mathbb{E}\left[g(M_{ij}) \mid M_{ij} < \theta\right]. \tag{56}$$

By ε-calibration,

$$|R(\theta_k) - G(\theta_k)| \leq \varepsilon, \qquad k \in \{1, 2\}. \tag{57}$$

We first show that $G(\theta)$ is non-decreasing in θ. If $\Pr[\theta_2 \leq M_{ij} < \theta_1] = 0$, then $G(\theta_1) = G(\theta_2)$ and the claim is immediate. Otherwise, let

$$\alpha = \frac{\Pr[M_{ij} < \theta_2]}{\Pr[M_{ij} < \theta_1]} \in (0, 1). \tag{58}$$

By the law of total expectation,

$$G(\theta_1) = \alpha G(\theta_2) + (1 - \alpha)\mathbb{E}\left[g(M_{ij}) \mid \theta_2 \leq M_{ij} < \theta_1\right]. \tag{59}$$

Since g is non-decreasing,

$$g(M_{ij}) \leq g(\theta_2) \quad \text{for } M_{ij} < \theta_2, \tag{60}$$

and

$$g(M_{ij}) \geq g(\theta_2) \quad \text{for } \theta_2 \leq M_{ij} < \theta_1. \tag{61}$$

Hence

$$\mathbb{E}\left[g(M_{ij}) \mid \theta_2 \leq M_{ij} < \theta_1\right] \geq g(\theta_2) \geq G(\theta_2), \tag{62}$$

which implies

$$G(\theta_1) \geq G(\theta_2). \tag{63}$$

Combining Eqs. 57 and 63,

$$\begin{aligned} R(\theta_2) &\leq G(\theta_2) + \varepsilon && (64)\\ &\leq G(\theta_1) + \varepsilon && (65)\\ &\leq R(\theta_1) + 2\varepsilon. && (66) \end{aligned}$$

Therefore,

$$A(\theta_2) = 1 - R(\theta_2) \geq 1 - R(\theta_1) - 2\varepsilon = A(\theta_1) - 2\varepsilon. \tag{67}$$

□

Interpretation. When $\varepsilon = 0$, decreasing the vacuity threshold cannot decrease retained-set accuracy. For $\varepsilon > 0$, deviations from monotonicity are bounded by 2ε. This is a conditional guarantee: the training objective does not itself establish Definition 1; the corresponding empirical relationship is evaluated in §3.3.

D Feature Lists

The input features are deterministic summary statistics computed from the observational sample $\mathbf{X}$; they are not learned separately from the causal-discovery objective. In addition to marginal and pairwise dependence statistics, the edge representation explicitly includes conditional-independence-related quantities, including partial correlations, candidate separating-set summaries, and collider-style marginal-versus-conditional dependence indicators. It also includes direction-sensitive residual and regression statistics. These features provide the encoder with both association and structural cues before the restricted-focal evidential heads are applied.

Table 6: Node features ($d_n = 13$): marginal summary statistics computed separately for each observed variable from n samples.

#	Feature group	Formula / description
1–2	Range	$\min(X_i)$, $\max(X_i)$
3–4	Uniqueness	$\lvert\text{unique}(X_i)\rvert$, $\lvert\text{unique}(X_i)\rvert/n$
5	Sample size	$\log n$
6–9	Entropy	$H(X_i)$ and normalized / baseline variants
10–11	Discrete entropy	Binned entropy and normalized variant
12–13	Distribution shape	Skewness, kurtosis

Table 7: Edge features ($d_e = 118$): 61 base statistics computed for each ordered pair (X_i, X_j), followed by directional swaps and deterministic transformations. Features marked with † provide conditional-independence or direction-sensitive structural cues.

Grp	Category	Dim	Key features
1	Correlation	3	r, $\lvert r\rvert$, Gaussian MI
2	Regression	6	MSE($i\to j$), MSE($j\to i$), difference, ratio
3	Moments	6	$\mathbb{E}[X_iX_j]$, $\mathbb{E}[X_i^2X_j]$, covariance
4	Residual statistics	10	Residual skewness / kurtosis per direction
5	Information theory	8	$H(X_i, X_j)$, conditional entropies, MI
6	Partial correlation†	6	Min./max. $\lvert\rho_{\text{partial}}\rvert$, separating-set summaries
7	Direction†	6	Residual-kurtosis asymmetry (LiNGAM-style)
8	HSIC	1	RBF-kernel dependence statistic
9	Variance	6	$\text{Var}(X_i)$, $\text{Var}(X_j)$, ratio, difference
10	Nonlinearity	2	Quadratic-versus-linear MSE gap
11	Collider / V-structure†	1	Marginal-versus-conditional dependence indicator
12	Node comparison	6	Pairwise differences in entropy, skewness, and kurtosis
	Total base	**61**	
	+ augmentations	**118**	Directional swaps, absolute values, logarithmic transforms

E Data Generation

Synthetic simulator. Our synthetic simulator follows the general simulation-based training protocol of Lorch et al. (2022), augmented with latent confounders and the pairwise relation labels used by EviCD. For each instance, a directed acyclic graph is sampled from either an Erdős–Rényi (ER) or scale-free (SF) graph family with d observed variables and a specified expected degree. Conditional on the sampled graph, observations are generated from either a linear or nonlinear structural equation model.

Linear mechanisms. Each observed variable is generated according to

$$x_j = \sum_{i \in \text{pa}(j)} w_{ij} x_i + b_j + \epsilon_j, \tag{68}$$

where

$$w_{ij} \sim \text{Unif}_{\pm}(1, 3) \tag{69}$$

denotes an equal mixture of $\text{Unif}(1, 3)$ and $\text{Unif}(-3, -1)$, and $b_j \sim \text{Unif}(-3, 3)$. The weight range avoids near-zero direct effects in the simulated generating graph.

Nonlinear mechanisms. For nonlinear instances, we use random Fourier feature (RFF) approximations to functions drawn from a Gaussian process with a squared-exponential kernel (Rahimi & Recht, 2007):

$$x_j = f_j(\mathbf{x}_{\text{pa}(j)}) + b_j + \epsilon_j. \tag{70}$$

The random function is parameterized as

$$f(\mathbf{x}) = \frac{c}{\sqrt{M}} \cos\left(\frac{\mathbf{x}\mathbf{W}}{\ell} + \mathbf{b}_{\text{rff}}\right)^{\top} \boldsymbol{\alpha}, \tag{71}$$

with

$$\mathbf{W} \sim \mathcal{N}(0, \mathbf{I}), \tag{72}$$
$$\mathbf{b}_{\mathrm{rff}} \sim \mathrm{Unif}(0, 2\pi), \tag{73}$$
$$\boldsymbol{\alpha} \sim \mathcal{N}(0, \mathbf{I}), \tag{74}$$
$$\ell \sim \mathrm{Unif}(7, 10), \tag{75}$$
$$c \sim \mathrm{Unif}(10, 20), \tag{76}$$

and $M = 100$ random features. A per-variable bias $b_j \sim \mathrm{Unif}(-3, 3)$ is added.

Noise. Each simulated instance uses one additive-noise family, sampled uniformly from Gaussian, Laplace, and Student-t distributions. For Gaussian and Laplace noise, $\sigma \sim \mathrm{Unif}(0.2, 2)$; for Student-$t$ noise we use $\nu = 5$ degrees of freedom and $\sigma \sim \mathrm{Unif}(0.2, 2)$. The simulator therefore covers several tail behaviors while retaining additive, homoscedastic noise. Heteroscedastic, interventional, and nonstationary mechanisms are outside the current training distribution.

Latent confounding. With probability 0.5, latent-confounder augmentation is activated. Conditional on augmentation, the number of latent common causes is sampled as

$$N_{\mathrm{lat}} \sim \mathrm{Poisson}(d/5). \tag{77}$$

Each latent variable is connected to 2–4 observed variables and is excluded from the observed data matrix $\mathbf{X}$. Its observed children are sampled without excluding pairs that are already adjacent in the observed DAG, so bow configurations—a direct relation together with a latent common cause—can occur.

In the evaluated simulator, bows account for approximately 2.2% of all variable pairs. Among pairs sharing a latent parent, their prevalence ranges from approximately 4% to 45% depending on graph density.

Relation labels and representation scope. The relation labels used for training are obtained from the *generating graph*, rather than from an observational equivalence class. Thus, a directional label describes the simulator's ground-truth generating relation even when that direction is not identifiable from the resulting observational distribution. This distinction is particularly relevant for linear-Gaussian instances and motivates the simulator-relative interpretation discussed in Appendix G.

The four-state frame used by EviCD assigns one simulator label to each ordered pair according to

$$t^*_{ij} = \begin{cases} \mathrm{D}_{ij}, & i \to j, \\ \mathrm{D}_{ji}, & j \to i, \\ \mathrm{C}, & \text{no direct edge and } i, j \text{ share a latent parent}, \\ \mathrm{N}, & \text{otherwise}. \end{cases} \tag{78}$$

Thus a direct edge takes precedence over latent confounding when both are present. Bow pairs are therefore absorbed into the appropriate directional class, while C denotes latent-common-cause confounding without a direct edge. The current label space cannot distinguish a direct-only pair from a bow as separate joint states; the bow stress test in Appendix H is consequently a diagnostic of channel incompatibility rather than an evaluation of an explicitly represented joint label.

Main synthetic test grid. The primary synthetic comparison uses eight configurations:

$$\{\mathrm{ER}, \mathrm{SF}\} \times d \in \{10, 20\} \times \mathrm{degree} \in \{2, 4\}, \tag{79}$$

with $n = 1000$ observations and three random seeds per configuration, yielding 24 test instances. Noise families are randomized across instances, and latent confounding is included according to the test-generation protocol.

Scalability grid. For the scaling analysis in Figure 2, we additionally evaluate $d \in \{50, 100\}$ under the same ER/SF graph families and degree settings. These larger graphs are reported separately from the eight-setting main benchmark.

External benchmark preprocessing. We additionally evaluate on three benchmarks outside the EviCD training simulator.

Sachs. The Sachs protein-signaling dataset (Sachs et al., 2005) contains $d = 11$ variables and $n = 5{,}846$ observational samples. Each variable is standardized to zero mean and unit variance.

DREAM4. We use the DREAM4 In Silico Size-10 gene-regulation benchmark (Marbach et al., 2009) with $d = 10$ variables.

Smith fMRI. We use 28 simulated BOLD-signal scenarios from Smith et al. (2011), with $d \in \{5, 10, 15, 50\}$ and 50 subjects per scenario. Reported results are aggregated over the 28 scenarios.

All evaluated methods receive the same preprocessed observations within each benchmark.

F Metric Definitions

We evaluate three distinct aspects of performance: (i) structural prediction, (ii) confounding detection, and (iii) selective prediction. Because the evaluated methods expose different output types, we distinguish metrics that operate on hard structural predictions from metrics that require continuous confidence or uncertainty scores. Unless stated otherwise, metrics are computed over the same evaluated variable pairs for all methods.

Structural prediction. For EviCD, categorical decisions are obtained from the pignistic predictor

$$\hat{t}_{ij} = \arg\max_{t \in \Omega_{ij}} \mathrm{BetP}(t; m_{ij}). \tag{80}$$

For each baseline, hard predictions are obtained from its native structural output and mapped to the common evaluation labels according to the procedure described below.

- **relation-presence AUROC.** We treat a pair as relation-present if its reference label is D_{ij}, D_{ji}, or C, and relation-absent if its label is N. For EviCD, the continuous relation-presence score is

  $$s_{ij}^{\mathrm{rel}} = 1 - \mathrm{BetP}(\mathrm{N}; m_{ij}). \tag{81}$$

  For AVICI we use its predicted edge score $P(e_{ij} \mid \mathbf{X})$; for NOTEARS, DAGMA, and DAG-GNN we use the absolute learned edge weight; and for PC, FCI, GES, and GRaSP we use their binary adjacency indicators. The latter therefore provide only two distinct score levels, so their AUROC should be interpreted as an adjacency-ranking summary rather than a continuously scored ROC curve.
- **Macro-F1.** Macro-F1 is computed over the modeled pairwise relation labels $\{\mathrm{D}_{ij}, \mathrm{D}_{ji}, \mathrm{C}, \mathrm{N}\}$ after mapping each method's native output to this common label space. Methods that do not explicitly represent latent confounding cannot predict the C class.
- **SHD.** Structural Hamming distance counts structural disagreements between the predicted and reference graphs under the common evaluation mapping.[1]
- **Macro-precision.** Precision is macro-averaged over the same common relation-label space.
- **ECE.** Expected calibration error is reported only for methods that provide a meaningful continuous predictive confidence. For EviCD, the top-label confidence is

  $$c_{ij} = \max_{t \in \Omega_{ij}} \mathrm{BetP}(t; m_{ij}), \tag{82}$$

  and ECE is computed using 10 equal-width bins:

  $$\mathrm{ECE} = \sum_{b=1}^{10} \frac{|B_b|}{N} \left| \mathrm{acc}(B_b) - \mathrm{conf}(B_b) \right|. \tag{83}$$

  AVICI uses its corresponding predictive probability, and stability selection uses its selection frequency as confidence. Point-estimate methods without a probabilistic or frequency-based confidence are reported as "—" for ECE.

[1] We omit SID (Peters & Bühlmann, 2015) because the evaluated setting includes latent confounding, whereas standard SID assumes causally sufficient DAGs.

Confounding detection. We treat confounding detection as a binary task in which the positive class is the simulator-provided C label, i.e., latent-common-cause confounding without a direct edge under the precedence rule in Appendix E. Bow pairs carry a directional label and are therefore not positives for this metric. EviCD natively provides the continuous score

$$s_{ij}^{\text{conf}} = \text{BetP}(\text{C}; m_{ij}). \tag{84}$$

FCI provides a discrete confounding indication through its edge marks. AVICI does not natively contain a confounding hypothesis; where reported, its confounding result is based on the proxy score used in our implementation rather than an explicit AVICI output.

- **Confounding AUROC (C-AUC).** For EviCD, C-AUC is the AUROC for the simulator C label versus all other labels using $\text{BetP}(\text{C}; m_{ij})$. For FCI, the binary confounding indicator is used as the score. For AVICI, we use the binary proxy defined below. Because FCI and AVICI provide only two score levels for this evaluation, their AUROC is equivalent to balanced accuracy and should be read as a coarse decision-level reference rather than a continuous-ranking metric directly comparable to EviCD.
- **Confounding F1 (C-F1).** For EviCD, the confounding label is obtained by thresholding $\text{BetP}(\text{C}; m_{ij})$. C-F1* denotes the best F1 obtained by sweeping $\tau \in [0.01, 0.50]$ over 50 values on the evaluated predictions. We use it as an oracle threshold-sweep diagnostic rather than as a deployable operating point. FCI uses its native discrete confounding decision. AVICI's proxy is already binary, so any threshold in $(0, 1)$ yields the same prediction. Consequently, C-F1* is not a like-for-like threshold-selection comparison: the sweep can improve EviCD's continuous score but does not change the FCI or AVICI decisions.

AVICI confounding proxy. Let p_{ij} and p_{ji} denote AVICI's two directional edge scores. The evaluation code first thresholds each score at the fixed AVICI decision threshold 0.2 and labels a pair as confounded only when both directions are present:

$$s_{ij,\text{AVICI}}^{\text{conf}} = \mathbb{1}[p_{ij} > 0.2 \ \wedge \ p_{ji} > 0.2]. \tag{85}$$

Hence the AVICI confounding "score" is binary, not continuous. Its reported C-AUC is therefore the AUROC of a two-level decision score (equivalently, balanced accuracy when both classes are present), and should not be interpreted as directly comparable to the ranking quality of EviCD's continuous $\text{BetP}(\text{C}; m_{ij})$. We retain it only as a coarse reference for the heuristic bidirectional-edge proxy.

Selective prediction. Selective prediction evaluates whether a method's uncertainty score ranks incorrect predictions above correct ones for abstention. For an uncertainty score u_{ij} and threshold θ, define the retained set

$$S(\theta) = \{(i, j) : u_{ij} < \theta\}. \tag{86}$$

Its coverage and retained-set accuracy are

$$\text{Cov}(\theta) = \frac{|S(\theta)|}{N}, \qquad \text{Acc}_{\text{ret}}(\theta) = \frac{\sum_{(i,j)\in S(\theta)} \mathbb{1}[\hat{t}_{ij} = t_{ij}^*]}{|S(\theta)|}. \tag{87}$$

For EviCD,

$$u_{ij} = m_{ij}(\Omega_{ij}); \tag{88}$$

for AVICI we use

$$u_{ij} = 1 - \max_t p(t \mid \mathbf{X}); \tag{89}$$

and for stability selection we use

$$u_{ij} = 1 - \text{frequency}_{ij}. \tag{90}$$

We summarize selective performance by AUC_{sel}, the area under the retained-accuracy–coverage curve obtained by sweeping the corresponding uncertainty threshold (100 values over the evaluated score range). Higher values indicate that a method maintains higher accuracy as coverage is increased. Methods without a continuous uncertainty score are reported as "—" rather than assigned an artificial score of zero.

G Training Details and Hyperparameters

Architecture. EviCD uses a three-layer GraphSAGE encoder with hidden dimension 64. Each variable is represented by 13 marginal features, and each ordered variable pair is associated with 118 pairwise features (Appendix D). The directional heads consume the encoded node representations together with the corresponding ordered-pair features, while the symmetric head uses a permutation-symmetric combination of the two node representations, as described in §2.2. The two directional DST channels use $P = 4$ prototypes each; the symmetric channel uses the restricted focal family in Eq. 4. The three channel-level BPAs are combined using Dempster's rule.

Algorithm 1 Training the EviCD inference model f_ϕ

Require: Simulator $\mathcal{S}$, hyperparameters $\lambda_{\text{cal}}, \lambda_{\text{sym}}$, initial learning rate η

Initialize model parameters ϕ
for step $= 1$ to 200,000 **do**
  Sample topology $\in \{\text{ER}, \text{SF}\}$, $d \in \{10, 20, 50, 100\}$, degree $\in \{1, 2, 4\}$, and $n \in \{500, 1000, 2000\}$
  Sample mechanism type and noise distribution according to Appendix E
  Sample whether latent confounders are included
  Sample a causal structure G and generate $\mathbf{X} \sim p_{\mathcal{S}}(\mathbf{X} \mid G)$ under the sampled configuration
  Extract node features $\{\mathbf{x}_i\}_{i=1}^d$ and ordered-pair features $\{\mathbf{e}_{ij}\}_{i \neq j}$ from $\mathbf{X}$
  Compute $\{m_{ij}\}_{i \neq j} \leftarrow f_\phi(\{\mathbf{x}_i\}, \{\mathbf{e}_{ij}\})$
  Construct simulator-provided ordered-pair labels $\{t^*_{ij}\}_{i \neq j}$ using the direct-edge-priority rule in Appendix E
  Compute $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{NLL}} + \lambda_{\text{cal}}\mathcal{L}_{\text{cal}} + \lambda_{\text{sym}}\mathcal{L}_{\text{sym}}$
  Update ϕ with Adam and clip the gradient norm at 1
  if step $= 150{,}000$ **then**
    $\eta \leftarrow \eta/10$
  end if
end for
  return f_ϕ

Training algorithm.

Training protocol. The model is trained for 200,000 optimization steps on procedurally generated graph–data pairs. Training configurations include $d \in \{10, 20, 50, 100\}$, $n \in \{500, 1000, 2000\}$, ER and SF graph families, and expected degrees in $\{1, 2, 4\}$. A fresh simulated instance is generated at each training step. We use Adam with initial learning rate 3×10^{-4}, reduce the learning rate by a factor of ten after 150,000 steps, and clip the ℓ_2 gradient norm at 1. The reported model uses $\lambda_{\text{cal}} = 0.1$ and $\lambda_{\text{sym}} = 0.1$. All three objective terms are averaged over the $d(d-1)$ ordered-pair outputs. The reversal-consistency term is computed from the independently predicted ordered representations: it couples the forward directional evidence for (i, j) with the reverse directional evidence for (j, i), and vice versa.

Simulator-relative supervision. The training labels t^*_{ij} describe the relation encoded by the simulator's generating graph; they should not be interpreted as relations that are necessarily identifiable from the corresponding observational distribution. In particular, the current simulator assigns singleton directional labels according to the generating graph even in regimes, such as linear-Gaussian models, where observational data alone may not identify that direction. Consequently, the pignistic loss rewards prediction of the simulator label rather than equivalence-class recovery, and can discourage large vacuity on observationally non-identifiable examples. This supervision choice is one reason that $m(\Omega)$ should not be interpreted as a certificate of causal identifiability.

H Diagnostic and Robustness Analyses

H.1 Controlled Perturbations

The following analyses were run with the trained EviCD model while holding the model architecture fixed and varying one data-generating factor at a time.

Table 8: Latent-confounder-strength sweep on confounded pairs ($n = 1000$).

Strength	0	0.25	0.5	1	2	4
K	0.14	0.51	0.63	0.72	0.77	0.76
$m(\Omega)$	0.021	0.029	0.034	0.035	0.031	0.030

Conflict rises by 0.63 across the sweep (Spearman $\rho = 0.94$, $p = 0.005$), whereas vacuity varies by only 0.014. This supports the interpretation of K as a model-level response to increasingly incompatible structural signals; it does not establish that K uniquely identifies latent confounding.

Table 9: Noise and sample-size sweeps.

Noise σ	0.2	0.5	1.0	1.5	2.0
$m(\Omega)$	0.0174	0.0191	0.0205	0.0211	0.0209
K	0.203	0.195	0.190	0.189	0.184
Sample size n	500	1000	2000		
$m(\Omega)$	0.0213	0.0205	0.0194		
K	0.216	0.190	0.197		

Vacuity increases overall as observation noise grows (Spearman $\rho = 0.90$, $p = 0.037$) and decreases as sample size increases within the training support. The corresponding changes in K are small.

Table 10: Signal-strength sweep on present edges.

Signal strength w	0.1	0.3	0.5	1.0	2.0
$m(\Omega)$	0.029	0.020	0.020	0.018	0.015

Weaker direct effects produce larger vacuity on present edges. Across the noise, sample-size, and signal-strength studies, K varies by at most 0.08. On the same diagnostic data, the correlation between K and prediction error is Spearman $\rho = 0.39$ ($p \ll 10^{-100}$), replicating the trend in the original diagnostic pool.

Removing the vacuity regularizer. When the model is retrained with $\lambda_{\text{cal}} = 0$, the qualitative relationships persist: Spearman$(\sigma, m(\Omega)) = +1.0$, Spearman$(n, m(\Omega)) = -1.0$, and Spearman(confounder strength, K) $= +0.94$. This indicates that the observed response of vacuity to data informativeness and conflict to confounder strength is not solely created by $\mathcal{L}_{\text{cal}}$.

H.2 Evidential Landscape

In this panel, approximately 92% of confounded pairs fall in the high-conflict region under the visualization threshold used in the figure. This enrichment is consistent with the architecture producing incompatible direction- and type-level assessments, but high K should not be interpreted as a confounding certificate.

H.3 Identifiability Stress Test

To probe the relationship between vacuity and observational identifiability, the same linear SCM was instantiated with Gaussian noise (direction non-identifiable from observational data in this setting) and Laplace noise (an identifiable control under LiNGAM-style assumptions).

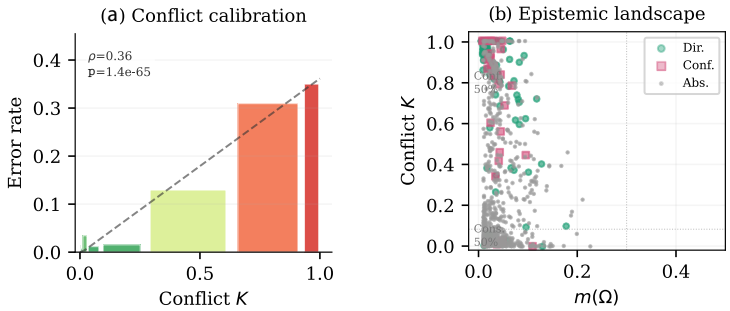

Figure 4: Diagnostic behavior on synthetic ER graphs with $d = 20$, degree $= 2$, 14 instances, and 5320 evaluated pairs. Left: empirical error across bins of K. Right: pairs in the $(m(\Omega), K)$ plane colored by simulator-provided relation type. The visualization thresholds are descriptive rather than universal decision rules.

Table 11: Gaussian/Laplace identifiability stress test.

Regime	mean $m(\Omega)$	confident-but-wrong direction
Gaussian, non-identifiable	0.0183	0.0%
Laplace, identifiable control	0.0152	1.6%

The vacuity difference is in the expected direction but small. We therefore do *not* interpret $m(\Omega)$ as a reliable certificate of causal non-identifiability. One reason is the current supervision: the simulator supplies singleton directional targets even in non-identifiable regimes, which discourages fully vacuous outputs. Equivalence-aware or set-valued supervision is a natural extension.

Bow configurations: simultaneous direct effect and latent confounding. A separate stress test on pairs containing both a direct relation and a latent common cause produces high mean conflict ($K \approx 0.98$). This shows that the current architecture can flag incompatibility, but it does not represent the conjunction as a distinct ground-truth state. The four-state frame therefore remains a genuine modeling limitation rather than a complete representation of latent-variable causal relations.

H.4 Calibration of Predictive Confidence

For methods with directly interpretable continuous confidence scores, top-label ECE on the main synthetic benchmark is 0.069 ± 0.010 for EviCD, 0.106 ± 0.023 for AVICI, and 0.288 ± 0.040 for stability selection. This comparison concerns calibration of the induced decision confidence, not identification or calibration of the BPA decomposition.

I Extended Synthetic Results

Tables 13 and 14 report the per-setting structural results for the primary synthetic benchmark with $d \in \{10, 20\}$. Table 15 reports selective-prediction performance for the same eight settings. These results complement the averages reported in the main paper and show that EviCD's advantage is concentrated in uncertainty-aware evaluation rather than uniform dominance in structural recovery.

Across the eight structural settings, no single method dominates all metrics. EviCD achieves the highest relation-presence AUROC in two settings, while AVICI, stability selection, and GRaSP lead in the remaining cases. DAG-GNN consistently obtains low SHD and strong macro-F1, whereas AVICI also performs strongly on structural recovery. This heterogeneity is consistent with the main-paper conclusion that EviCD is competitive, rather than uniformly superior, on point-estimate structure metrics.

For selective prediction, EviCD and AVICI each achieve the highest $\mathrm{AUC}_{\mathrm{sel}}$ in four of the eight settings. EviCD performs best on ER-10-deg2, ER-20-deg2, SF-10-deg4, and SF-20-deg2, whereas

Table 12: Full summary for the main synthetic benchmark (mean ± SE across eight settings). ECE is shown only for methods with a directly interpretable continuous confidence or stability score.

Method	AUROC↑	F1↑	SHD↓	Prec.↑	ECE↓	C-AUC↑	AUC_{sel}↑
PC	$0.645_{\pm.022}$	$0.327_{\pm.013}$	$194_{\pm45}$	$0.368_{\pm.017}$	—	—	—
FCI	$0.597_{\pm.031}$	$0.295_{\pm.036}$	$199_{\pm47}$	$0.465_{\pm.041}$	—	$0.531_{\pm.027}$	—
GES	$0.716_{\pm.033}$	$0.294_{\pm.010}$	$288_{\pm75}$	$0.294_{\pm.009}$	—	—	—
GRaSP	$0.763_{\pm.022}$	$0.292_{\pm.012}$	$236_{\pm58}$	$0.287_{\pm.013}$	—	—	—
DAGMA	$0.349_{\pm.035}$	$0.578_{\pm.016}$	$136_{\pm39}$	$0.546_{\pm.020}$	—	—	—
LiNGAM	$0.365_{\pm.023}$	$0.375_{\pm.009}$	$266_{\pm65}$	$0.363_{\pm.007}$	—	—	—
NOTEARS	$0.399_{\pm.026}$	$0.470_{\pm.013}$	$204_{\pm51}$	$0.443_{\pm.013}$	—	—	—
Stab. sel.	$0.786_{\pm.036}$	$0.307_{\pm.010}$	$272_{\pm70}$	$0.305_{\pm.010}$	$0.288_{\pm.040}$	—	$0.291_{\pm.035}$
AVICI	$0.805_{\pm.021}$	$0.476_{\pm.020}$	$255_{\pm68}$	$0.487_{\pm.021}$	$0.106_{\pm.023}$	$0.539_{\pm.029}$	$0.714_{\pm.030}$
DAG-GNN	$0.585_{\pm.025}$	$0.539_{\pm.015}$	$195_{\pm52}$	$0.574_{\pm.014}$	—	—	—
EviCD	$0.804_{\pm.035}$	$0.447_{\pm.020}$	$173_{\pm46}$	$0.481_{\pm.016}$	$0.069_{\pm.010}$	$0.629_{\pm.058}$	$0.727_{\pm.035}$

AVICI leads on the complementary four settings. Thus, at $d \leq 20$, the aggregate advantage of EviCD arises from its overall retained-accuracy–coverage behavior rather than dominance in every individual configuration.

Table 13: Erdős–Rényi results by setting for the primary synthetic benchmark. Best in **bold**, second best underlined.

	d=10, deg=2			d=10, deg=4			d=20, deg=2			d=20, deg=4		
Method	AUROC	F1	SHD	AUROC	F1	SHD	AUROC	F1	SHD	AUROC	F1	SHD
PC	0.692	0.288	54	0.687	0.337	104	0.683	0.344	176	0.679	0.398	242
FCI	0.799	0.495	42	0.544	0.248	112	0.631	0.327	162	0.595	0.296	262
GES	0.785	0.256	62	0.742	0.271	146	0.809	0.302	262	0.798	0.294	420
GRaSP	0.687	0.229	60	0.792	0.308	118	0.802	0.269	214	0.839	0.287	342
NOTEARS	0.441	0.451	70	0.338	0.534	80	0.297	0.470	216	0.369	0.493	294
Stab. sel.	0.912	0.258	60	0.794	0.326	130	0.866	0.290	256	**0.857**	0.304	394
AVICI	0.898	**0.550**	25	**0.859**	0.464	47	**0.953**	**0.600**	44	0.779	0.444	106
DAG-GNN	0.770	0.536	**19**	0.403	**0.551**	**29**	0.538	0.595	**39**	0.695	**0.594**	**67**
EviCD	**0.958**	0.517	36	0.826	0.470	92	0.902	0.447	147	0.800	0.380	252

Table 14: Scale-Free results per setting. Best in **bold**, second underlined.

	d=10, deg=2			d=10, deg=4			d=20, deg=2			d=20, deg=4		
Method	AUROC	F1	SHD	AUROC	F1	SHD	AUROC	F1	SHD	AUROC	F1	SHD
PC	0.699	0.326	106	0.591	0.316	154	0.583	0.332	258	0.542	0.275	462
FCI	0.592	0.367	98	0.536	0.172	180	0.557	0.244	270	0.523	0.213	466
GES	0.680	0.334	130	0.582	0.330	152	0.750	0.290	450	0.580	0.272	678
GRaSP	0.758	0.329	108	0.732	0.337	146	**0.822**	0.286	350	0.670	0.293	546
NOTEARS	0.387	0.485	90	0.524	0.465	104	0.367	0.450	312	0.470	0.411	464
Stab. sel.	0.766	0.344	123	0.643	0.325	154	0.812	0.315	415	0.637	0.291	642
AVICI	0.793	0.451	37	**0.775**	0.410	45	0.813	0.437	108	**0.719**	0.364	167
DAG-GNN	0.720	**0.577**	**22**	0.549	0.398	**44**	0.522	**0.573**	**68**	0.609	**0.479**	**106**
EviCD	**0.820**	0.508	78	0.705	**0.472**	102	0.764	0.431	248	0.657	0.354	432

J Full External-Benchmark Results

Table 15: Selective-prediction performance ($\mathrm{AUC}_{\mathrm{sel}}$) by setting. Only methods with an explicit continuous uncertainty or stability score used for selective prediction are shown. Best in **bold**, second best underlined. EviCD and AVICI each lead in four of the eight settings.

	Erdős–Rényi				Scale-Free			
Method	d=10,2	d=10,4	d=20,2	d=20,4	d=10,2	d=10,4	d=20,2	d=20,4
Stab. sel.	0.206	0.208	0.236	0.253	0.404	0.229	0.467	0.323
AVICI	0.846	**0.675**	0.662	**0.818**	**0.749**	0.590	0.682	**0.690**
EviCD	**0.896**	0.643	**0.806**	0.790	0.699	**0.690**	**0.719**	0.578

Higher $\mathrm{AUC}_{\mathrm{sel}}$ indicates better retained-set accuracy across coverage levels. Methods without a corresponding continuous selective-prediction score are reported as unavailable in the main comparison rather than assigned a score of zero.

Table 16: Full external-benchmark results. Sachs contains real experimental measurements; DREAM4 and fMRI are simulation-based benchmarks. "—" denotes an unavailable selective score. Best in **bold**, second best underlined.

	Sachs (d=11, n=5,846)				DREAM4 (d=10)				fMRI (28 sims, $d \in \{5, 10, 15, 50\}$)			
Method	AUROC↑	F1↑	SHD↓	$\mathrm{AUC}_{\mathrm{sel}}$↑	AUROC↑	F1↑	SHD↓	$\mathrm{AUC}_{\mathrm{sel}}$↑	AUROC↑	F1↑	SHD↓	$\mathrm{AUC}_{\mathrm{sel}}$↑
PC	0.502	0.361	62	—	0.500	0.282	24	—	0.678	0.361	22.1	—
FCI	0.474	0.264	38	—	0.568	0.213	26	—	0.605	0.257	18.9	—
GES	0.527	0.185	84	—	0.553	0.333	45	—	0.637	0.286	26.1	—
GRaSP	0.472	**0.392**	60	—	0.500	0.282	24	—	0.680	0.319	18.4	—
NOTEARS	0.598	0.355	42	—	0.629	0.344	35	—	0.759	0.544	12.9	—
DAGMA	0.558	0.313	42	—	0.533	0.310	49	—	0.821	0.578	12.2	—
LiNGAM	0.588	0.303	72	—	0.500	0.282	24	—	**0.985**	**0.895**	**2.40**	—
AVICI	0.560	0.337	54	0.484	0.630	0.151	64	0.460	0.886	0.186	23.57	0.434
DAG-GNN	**0.649**	0.363	40	—	0.440	0.262	30	—	0.949	0.327	16.35	—
Stab. sel.	0.581	0.151	90	0.153	0.739	**0.489**	35	**0.785**	0.664	0.291	24.1	0.144
EviCD	0.440	0.320	**36**	**0.746**	**0.750**	0.458	**22**	0.699	0.884	0.463	12.0	**0.617**