DCD-PFN: A DECOUPLING-AWARE FOUNDATION MODEL FOR CAUSAL DISCOVERY

Anonymous authors
Paper under double-blind review

ABSTRACT

Causal discovery from observational data is critical for understanding complex data-generating mechanisms, yet traditional methods often struggle under strong nonlinearity or noise, and may suffer from computational bottlenecks. Recent tabular foundation models based on Prior-Data Fitted Networks (PFNs) have demonstrated remarkable zero-shot inference capabilities, but their potential for explicit structural causal discovery remains underexplored. To bridge this gap, we propose DCD-PFN, a decoupling-aware foundation model for causal discovery. Rather than directly learning a mapping from data to global graphs, DCD-PFN amortizes decoupling-based local causal discovery. Through pre-training on diverse synthetic Structural Causal Models (SCMs), the model learns to infer sample-wise decoupling weights from observational data, which guide zero-shot Markov boundary (MB) identification in a single forward pass. DCD-PFN recovers global causal structures via a principled reconstruction procedure with parallelized local discovery. Experiments on both synthetic and real-world datasets demonstrate robust zero-shot generalization.

1 INTRODUCTION

Causal discovery serves as the cornerstone for understanding the generative mechanisms underlying complex data and for enabling reliable decision-making (Spirtes & Zhang, 2016; Glymour et al., 2019). In recent decades, traditional causal discovery algorithms have evolved into diverse paradigms (Spirtes et al., 2000), including constraint-based (Spirtes & Glymour, 1991; Spirtes et al., 1995), score-based (Chickering, 2002; Zheng et al., 2018), and functional-based methods (Shimizu et al., 2006; Chen et al., 2026). While these approaches can be effective in certain settings, their performance may degrade under complex nonlinear relationships and noisy data-generating mechanisms. Although constraint-based methods equipped with modern non-parametric conditional independence tests (Zhang et al., 2011) provide a flexible solution with fewer parametric restrictions, they often suffer from substantial computational bottlenecks. Consequently, achieving robust and efficient causal discovery remains a pressing challenge.

Recently, pre-trained tabular foundation models, represented by Prior-Data Fitted Networks (PFNs) (Hollmann et al., 2025), have provided a novel perspective to address these challenges. By pre-training on massive amounts of synthetic data, these models (Zhang et al., 2025a; Qu et al., 2026) have demonstrated remarkable in-context learning capabilities and zero-shot generalization performance on tabular prediction tasks, requiring only amortized forward-pass inference. However, existing PFN models are primarily designed as powerful posterior predictive distribution fitters. While this paradigm offers an appealing combination of efficient inference and zero-shot generalization, its potential for explicit structural causal discovery remains underexplored.

To bridge this gap, we introduce DCD-PFN, a decoupling-aware foundation model for causal discovery. Rather than predicting a global graph in one step, DCD-PFN amortizes decoupling-based Markov boundary (MB) identification. Through pre-training on diverse synthetic Structural Causal Models (SCMs), the model learns to infer sample-wise decoupling weights from observational data, which guide zero-shot MB identification in a single forward pass. The learned MB identifier is further integrated into a principled local-to-global reconstruction procedure (Guan et al., 2026), enabling global causal structure recovery with parallelized local discovery.

The main contributions of this paper are summarized as follows:

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.

- To bridge the gap between zero-shot inference and principled structure learning, we propose DCD-PFN, a novel decoupling-aware foundation model for causal discovery.
- DCD-PFN amortizes decoupling-based local MB discovery and is integrated into a parallelized local-to-global reconstruction procedure to recover global causal structures.
- Experiments on both synthetic and real-world datasets demonstrate that DCD-PFN exhibits strong robustness and zero-shot generalization.

2 Related Work

2.1 Prior-Data Fitted Networks

Prior-Data Fitted Networks (PFNs) redefine Bayesian posterior predictive distribution estimation as in-context learning, utilizing Transformers (Kossen et al., 2021) trained on synthetic priors to approximate predictive distributions (Müller et al., 2022; Nagler, 2023). This paradigm has been successfully validated on tabular data, outperforming tree-based baselines by leveraging priors derived from SCMs (Hollmann et al., 2023; 2025). Recent efforts have further scaled this approach: TabPFN v2.5 (Grinsztajn et al., 2025) expands dataset capacity and provides a distillation engine for low-latency deployment, TabICL v2 (Qu et al., 2026) introduces a scalable architecture for classification and regression, LimiX (Zhang et al., 2025a) enables joint distribution modeling, and Mitra (Zhang et al., 2025b) optimizes prior diversity.

Beyond architecture scaling, concurrent research scrutinizes the reliability of PFNs under non-stationary settings (Ye et al., 2025; Cheng et al., 2025; Djilani et al., 2025), with methods like Drift-Resilient TabPFN (Helli et al., 2024) incorporating time-variant SCM priors to address distribution shifts. Critically, this paradigm has also demonstrated significant potential in causality. However, while works such as CausalFM (Ma et al., 2026) and Do-PFN (Robertson et al., 2025) show that PFNs can effectively model intervention outcomes, they remain primarily focused on approximating interventional distributions through black-box fitting. The potential of PFNs to serve as zero-shot foundation models for explicit structural causal discovery remains largely underexplored.

2.2 Classical Causal and Markov Boundary Discovery

Causal discovery seeks to reconstruct directed acyclic graphs (DAGs) from observational data. Traditional global structure learning methods often rely on conditional independence constraints (e.g., PC (Spirtes & Glymour, 1991)) or functional asymmetries (e.g., LiNGAM (Shimizu et al., 2006)). Alternatively, local discovery focuses on the MB, the minimal set rendering a target independent of all other variables (Aliferis et al., 2010). Classical algorithms, such as Grow-Shrink (GS) (Margaritis & Thrun, 1999) and Incremental Association (IA) (Tsamardinos et al., 2003), alongside modern variants (BAMB (Ling et al., 2019), EEMB (Wang et al., 2020), CCMB (Wu et al., 2020)), typically iteratively conduct conditional independence tests to discover the MB.

2.3 Amortized Inference for Causal Discovery

To bypass the sample inefficiency and high latency of iterative algorithms, recent works have shifted towards amortized causal discovery. Models such as AVICI (Lorch et al., 2022) and BCNP (Dhir et al., 2025) treat causal discovery as a supervised meta-learning problem. By pre-training on synthetic datasets generated from diverse SCMs, these models map observational data directly to causal graphs via feed-forward inference. This amortized synergy has also been extended to causal induction (Ke et al., 2023), active learning (Annadani et al., 2024), and conditional independence testing (Duong et al., 2025).

While existing amortized methods often predict the global causal graph directly, DCD-PFN focuses its amortized inference on decoupling-based local MB identification. Guided by the theoretical framework (Guan et al., 2026), DCD-PFN is trained to serve as a reusable MB identifier and is integrated into a parallel graph reconstruction procedure.

3 METHOD

3.1 PROBLEM SETUP

Consider d random variables $\mathbf{X} = (X_1, \ldots, X_d) \in \mathcal{X} \subset \mathbb{R}^d$ and an observational dataset $\mathcal{D}_n = \{\mathbf{x}^{(i)}\}_{i=1}^n$ consisting of n independent and identically distributed (i.i.d.) samples drawn from an underlying joint distribution $\mathbb{P}$. We assume the data-generating distribution $\mathbb{P}$ is Markov with respect to a DAG $\mathcal{G} = (\mathcal{V}, E)$, where $\mathcal{V} = \{X_1, \ldots, X_d\}$ serves as the set of nodes and E denotes the set of directed edges. This implies that the joint distribution factorizes as:

$$\mathbb{P}(X_1, \ldots, X_d) = \prod_{j=1}^{d} \mathbb{P}(X_j \mid \mathrm{Pa}^{\mathcal{G}}(X_j)), \tag{1}$$

where $\mathrm{Pa}^{\mathcal{G}}(X_j)$ denotes the set of parents of node X_j in $\mathcal{G}$. Within the Bayesian framework, we treat both the DAG $\mathcal{G}$ and the model parameters θ as random variables. The joint prior distribution is specified as $\Pi(\mathcal{G}, \theta) = \Pi(\theta \mid \mathcal{G})\Pi(\mathcal{G})$, and the data-generating process is modeled as $\mathbf{x}^{(i)} \mid \mathcal{G}, \theta \overset{\text{i.i.d.}}{\sim} \mathbb{P}(\cdot \mid \mathcal{G}, \theta)$.

Amortized inference approaches frame causal structure learning as a supervised learning task. Instead of searching over the superexponential space of potential graphs, these methods parameterize a neural network $p_\phi(\mathcal{G} \mid \mathcal{D}_n)$ to directly approximate the intractable posterior $p(\mathcal{G} \mid \mathcal{D}_n)$. The optimal parameters ϕ^* are obtained by minimizing the expected negative log-likelihood (NLL) over the data-generating distribution:

$$\phi^* = \arg\min_{\phi} \mathbb{E}_{(\mathcal{G},\theta)\sim\Pi} \mathbb{E}_{\mathcal{D}_n \sim \mathbb{P}(\cdot \mid \mathcal{G}, \theta)} \left[-\log p_\phi\big(\mathcal{G} \mid \mathcal{D}_n\big) \right]. \tag{2}$$

While directly learning the full graph $\mathcal{G}$ provides a complete picture of the system, its recovery from observational data can be challenging. We instead first learn the MB (Pearl, 1988) of a target variable X_t, denoted by $M_t(\mathbb{P}) = \mathrm{MB}_{\mathbb{P}}(X_t)$, where $t \in \{1, \ldots, d\}$. $M_t(\mathbb{P})$ is an inclusion-minimal set such that

$$X_t \perp\!\!\!\perp \big(\mathcal{V} \setminus (\{X_t\} \cup M_t(\mathbb{P}))\big) \mid M_t(\mathbb{P}). \tag{3}$$

We assume strict positivity, so that $M_t(\mathbb{P})$ is unique (Pearl, 1988; Guan et al., 2026). Moreover, when $\mathbb{P}$ is Markov and faithful to $\mathcal{G}$ (Spirtes et al., 2000), the MB of X_t has the graphical characterization (Pearl, 1988):

$$M_t(\mathbb{P}) = \mathrm{Pa}^{\mathcal{G}}(X_t) \cup \mathrm{Ch}^{\mathcal{G}}(X_t) \cup \mathrm{Sp}^{\mathcal{G}}(X_t), \tag{4}$$

where $\mathrm{Ch}^{\mathcal{G}}(\cdot)$ and $\mathrm{Sp}^{\mathcal{G}}(\cdot)$ denote children and spouses in $\mathcal{G}$, respectively. This characterization supplies ground-truth MB labels M_t from the generating DAG $\mathcal{G}$. Consequently, our local amortized inference objective becomes

$$\eta^* = \arg\min_{\eta} \mathbb{E}_{(\mathcal{G},\theta)\sim\Pi} \mathbb{E}_{\mathcal{D}_n \sim \mathbb{P}(\cdot \mid \mathcal{G}, \theta)} \left[-\log p_\eta\big(M_t \mid \mathcal{D}_n, t\big) \right]. \tag{5}$$

Here $p_\eta\big(M_t \mid \mathcal{D}_n, t\big)$ denotes the model's approximation to the posterior probability of the target's MB under the synthetic prior Π.

3.2 MODEL ARCHITECTURE

The foundational component of DCD-PFN is the Dual-Axis Attention Block (DAB) (Kossen et al., 2021), an architecture proven effective in capturing complex data dependencies by alternating between Feature Attention and Sample Attention. By explicitly modeling both the interactions between different features within a single observation and the relationships across different observations in the entire dataset, the DAB allows DCD-PFN to learn highly expressive representations.

The inputs $\mathbf{D} \in \mathbb{R}^{n \times d}$ are first mapped into an e-dimensional embedding space via linear projection, supplemented with unique feature ID embeddings to maintain column permutation invariance. An encoder consisting of L_0 DAB layers maps these inputs into a high-dimensional global representation space, yielding $\mathbf{H} \in \mathbb{R}^{n \times d \times e}$. Within this representation, the features excluding the target variable X_t are denoted as $\mathbf{H}_{-t} \in \mathbb{R}^{n \times (d-1) \times e}$.

Feature-wise Decoupling. $\mathbf{H}_{-t}$ is utilized to learn sample decoupling weights $w_i, i \in \{1, \ldots, n\}$ to decouple variables other than X_t. Specifically, $\mathbf{H}_{-t}$ is first encoded through L_1 DAB layers to

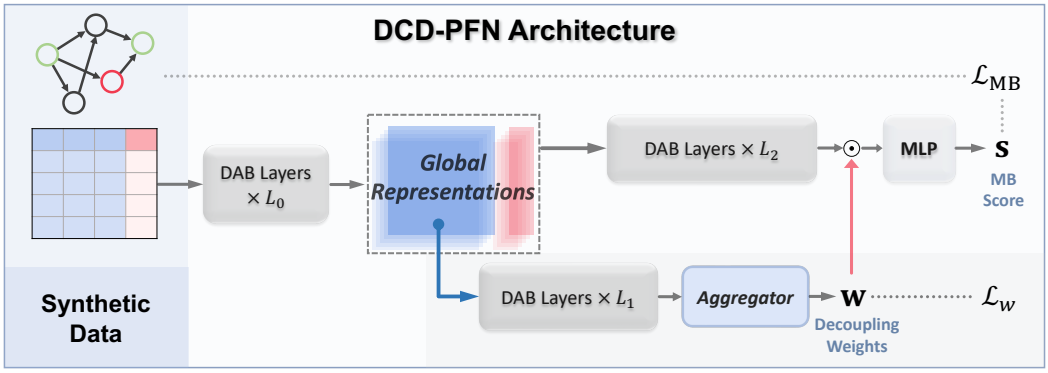

Figure 1: Architecture of DCD-PFN. Input synthetic data is encoded into global representations via L_0 Dual-Axis Attention Block (DAB) layers. The architecture then splits into two joint branches: (1) a decoupling branch that learns sample-wise weights $\mathbf{w}$ via L_1 DAB layers and an aggregator under the guidance of $\mathcal{L}_w$, and (2) an MB learning branch that refines representations via L_2 DAB layers and utilizes an MLP to output the final Markov Boundary scores $\mathbf{s}$ supervised by $\mathcal{L}_{MB}$.

obtain $\mathbf{H}_{-t}^{(1)} \in \mathbb{R}^{n\times(d-1)\times e}$, and is then processed by a cross-attention-based aggregator to compress the feature dimension. Using a learnable query token $\mathbf{q}$, the compressed representation $\mathbf{z}_i$ for each sample is computed as follows:

$$\mathbf{z}_i = \text{LayerNorm}(\text{MHA}(\mathbf{q}, \mathbf{H}_{i,-t,.}^{(1)}, \mathbf{H}_{i,-t,.}^{(1)})), \tag{6}$$

where MHA denotes Multi-Head Attention. The representation $\mathbf{z}_i$ is mapped through a linear layer, followed by a Softplus activation to ensure positivity and a normalization step to yield a valid sample weight w_i with a mean of 1.

MB Learning. The complete representation $\mathbf{H}$, combined with the learned decoupling weights $w_i, i \in \{1,\dots,n\}$, is employed to identify the $\text{MB}^{\mathcal{G}}(X_t)$. First, $\mathbf{H}$ passes through L_2 DAB layers to obtain a refined representation $\mathbf{H}^{(2)}$. Utilizing the learned decoupling weights, we then compute the weighted similarity between the target embeddings $\mathbf{H}_t^{(2)}$ and other variable embeddings $\mathbf{H}_{-t}^{(2)}$:

$$\mathbf{S} = \sum_{i=1}^{n} w_i \cdot (\mathbf{H}_{i,-t}^{(2)} \odot \mathbf{H}_{i,t}^{(2)}), \quad \mathbf{S}_t = \sum_{i=1}^{n} w_i \cdot (\mathbf{H}_{i,t}^{(2)} \odot \mathbf{H}_{i,t}^{(2)}), \quad \mathbf{S}_{-t} = \sum_{i=1}^{n} w_i \cdot (\mathbf{H}_{i,-t}^{(2)} \odot \mathbf{H}_{i,-t}^{(2)}). \tag{7}$$

Here, $\mathbf{S}$ captures the weighted cross-correlation in the embedding space, while $\mathbf{S}_t$ and $\mathbf{S}_{-t}$ serve as normalization metrics that account for target and other variables intensities. To ensure robustness across different magnitudes, these metrics undergo a signed logarithmic transformation:

$$\mathbf{S}_{comb} = \text{Concat}(\text{sign}(\mathbf{S})\log(1+|\mathbf{S}|), \log(1+\mathbf{S}_t), \log(1+\mathbf{S}_{-t})). \tag{8}$$

The concatenated vector is processed by an MLP with a Sigmoid activation function to yield the final probability score $\mathbf{s} \in [0,1]^{d-1}$ representing the score of each variable belonging to the MB.

3.3 Pre-training

Synthetic Data Generation. The generalization capability of a causal foundation model relies heavily on the diversity of its pre-training prior. We generate synthetic priors using a diverse range of SCMs. For each dataset, the underlying DAG is sampled using the Erdős-Rényi model. The structural equations governing node relationships are stochastically selected from a diverse set of functional components, including random neural networks with varied activation functions, decision tree structures, and discretization steps. Finally, noise from multiple distributions with randomized variances is superimposed.

The end-to-end model is optimized via a joint loss function:

Decoupling Loss ($\mathcal{L}_w$). We employ the Pairwise Weighted Hilbert-Schmidt Independence Criterion (HSIC) loss. Given the sample decoupling weights vector $\mathbf{w} = (w_1,\dots,w_n) \in \mathbb{R}^n$, we compute the

initial kernel matrix $\mathbf{K}^{(j)} \in \mathbb{R}^{n\times n}$ for any non-target feature X_j (where $j \neq t$) using a Radial Basis Function (RBF). Subsequently, a weighted centering operation is applied to the kernel matrix using normalized weights $\tilde{\mathbf{w}} = \mathbf{w}/\sum_{i=1}^{n} w_i$:

$$\tilde{\mathbf{K}}^{(j)} = (\mathbf{I} - \mathbf{1}\tilde{\mathbf{w}}^\top)\mathbf{K}^{(j)}(\mathbf{I} - \tilde{\mathbf{w}}\mathbf{1}^\top) \tag{9}$$

where $\mathbf{I}$ is the identity matrix and $\mathbf{1}$ is a vector of ones. To incorporate sample weights into the kernel inner products, we apply trace weighting to the centered kernel matrix:

$$\bar{\mathbf{K}}^{(j)} = \mathbf{W}^{1/2}\tilde{\mathbf{K}}^{(j)}\mathbf{W}^{1/2} \tag{10}$$

where $\mathbf{W} = \mathrm{diag}(\mathbf{w})$ is the diagonal matrix constructed from the sample weights.

To eliminate the inherent scale differences between the kernel matrices of different feature dimensions, we compute the normalized HSIC (i.e., Centered Kernel Alignment). The final decoupling loss is defined as the sum of the normalized HSIC scores across all feature pairs:

$$\mathcal{L}_w = \sum_{\substack{j<k \\ j,k\neq t}} \frac{\mathrm{Tr}(\bar{\mathbf{K}}^{(j)}\bar{\mathbf{K}}^{(k)})}{\sqrt{\mathrm{Tr}(\bar{\mathbf{K}}^{(j)}\bar{\mathbf{K}}^{(j)})\mathrm{Tr}(\bar{\mathbf{K}}^{(k)}\bar{\mathbf{K}}^{(k)})}} \tag{11}$$

MB Loss ($\mathcal{L}_{\text{MB}}$). A standard Binary Cross-Entropy (BCE) loss is utilized to align the predicted MB probability s_j with the ground-truth label g_j derived from the synthetic SCMs:

$$\mathcal{L}_{\text{MB}} = -\frac{1}{d-1}\sum_{\substack{j=1 \\ j\neq t}}^{d}[g_j \log(s_j) + (1-g_j)\log(1-s_j)] \tag{12}$$

Joint objective. The model is trained end-to-end with

$$\mathcal{L} = \mathcal{L}_{\text{MB}} + \lambda_w \mathcal{L}_w. \tag{13}$$

Here, $\mathcal{L}_{\text{MB}}$ supervises MB prediction, while $\mathcal{L}_w$ provides an auxiliary decoupling constraint. Since the learned weights also enter the embedding statistics used for MB prediction, both objectives contribute to training the weighting branch. This joint formulation couples the decoupling criterion with the downstream goal of identifying the target's MB.

3.4 Amortized MB Inference and Global Graph Reconstruction

We develop a parallel implementation of the reconstruction procedure of Guan et al. (2026), using DCD-PFN as an amortized MB identifier. As detailed in Algorithm 1, local discovery branches are executed independently across targets, followed by a global aggregation and orientation stage. The reconstruction guarantee relies on (Spirtes et al., 2000; Guan et al., 2026):

- *Causal sufficiency*: no unobserved confounders among the variables in $\mathcal{V}$.
- *Markov condition*: $\mathbb{P}$ is Markov with respect to the underlying DAG $\mathcal{G}$.
- *Faithfulness*: every conditional independence relationship in $\mathbb{P}$ is implied by d-separation in $\mathcal{G}$.

Under these assumptions and given an exact MB oracle for all queried variable sets, the reconstruction procedure recovers the CPDAG of $\mathcal{G}$ (Guan et al., 2026). Our parallel implementation preserves this reconstruction logic, while using DCD-PFN to approximate the required MB queries.

For each query $\mathrm{MB}(X_t \mid \mathcal{O})$, we restrict the dataset to the variables in $\mathcal{O}$ and apply the same pre-trained model with X_t as the target. Non-target variables with predicted MB scores above 0.5 are included in the estimated MB. Each query requires a forward pass without parameter updates or query-specific iterative weight optimization.

Within each target branch, the algorithm removes candidate subsets from the initial MB and examines changes in the remaining boundary to distinguish structural neighbors from spouses. After the parallel branches finish, their neighbor sets and candidate v-structures are aggregated to construct the skeleton and orient unshielded colliders. Meek's rules then propagate the implied orientations to produce an estimated CPDAG $\hat{\mathcal{G}}$. Although the full reconstruction involves multiple MB queries, amortization reduces the computation within each query, and parallel execution reduces sequential processing across targets.

Algorithm 1 Parallel DCD-Construction

Require: Set of variables $\mathcal{V} = \{X_1, \ldots, X_d\}$, MB identifier $\text{MB}(\cdot \mid \cdot)$
Ensure: Completed Partially Directed Acyclic Graph (CPDAG) $\hat{\mathcal{G}} = (\mathcal{V}, \hat{E})$

Initialize graph $\hat{\mathcal{G}} = (\mathcal{V}, \hat{E})$ with empty edge set $\hat{E} \leftarrow \emptyset$
parallel for $X_t \in \mathcal{V}$
    Initialize local v-structure candidates $\mathcal{K}_t \leftarrow \varnothing$
    Obtain initial MB: $\mathcal{M}_t \leftarrow \text{MB}(X_t \mid \mathcal{V})$
    Initialize neighbor set $\mathcal{N}_t \leftarrow \mathcal{M}_t$
    for $k = 1$ to $|\mathcal{M}_t| - 1$ **do**
        for each subset $\mathcal{Z} \subset \mathcal{M}_t$ with $|\mathcal{Z}| = k$ **do**
            $\mathcal{M}_{\setminus \mathcal{Z}} \leftarrow \text{MB}(X_t \mid (\mathcal{M}_t \cup \{X_t\}) \setminus \mathcal{Z})$
            $\mathcal{D}_{\text{drop}} \leftarrow (\mathcal{M}_t \setminus \mathcal{Z}) \setminus \mathcal{M}_{\setminus \mathcal{Z}}$
            if $\mathcal{D}_{\text{drop}} \neq \emptyset$ **then**
                for each $Y \in \mathcal{D}_{\text{drop}}$ **do**
                    for each $Z \in \mathcal{Z}$ **do**
                        $\mathcal{K}_t \leftarrow \mathcal{K}_t \cup \{(X_t, Z, Y)\}$
                    end for
                    $\mathcal{N}_t \leftarrow \mathcal{N}_t \setminus \{Y\}$
                end for
            end if
        end for
    end for
end parallel for
$\mathcal{K} \leftarrow \bigcup_{X_t \in V} \mathcal{K}_t$
for each $X_t \in V$ **do**
    for each $W \in \mathcal{N}_t$ **do**
        Add undirected edge $X_t - W$ to $\hat{\mathcal{E}}$
    end for
end for
for each $(X, Z, Y) \in \mathcal{K}$ **do**
    if X and Y are both adjacent to Z in $\hat{\mathcal{G}}$, but are not adjacent to each other **then**
        Orient edges in $\hat{E}$ as: $X \to Z$ and $Y \to Z$
    end if
end for
Apply Meek rules (Meek, 1995) to $\hat{\mathcal{G}}$ to exhaustively orient remaining undirected edges
return $\hat{\mathcal{G}}$

4 EXPERIMENTS

We evaluate DCD-PFN on synthetic datasets (Section 4.2), semi-synthetic datasets constructed from benchmark topologies (Section 4.3), and the real-world Sachs flow-cytometry dataset (Section 4.4). Additionally, we conduct ablation studies on the decoupling mechanism in Appendix A.5.

4.1 EXPERIMENTAL SETUP

Baselines. We evaluate DCD-PFN against methods from several representative paradigms: (i) *Constraint-based:* PC (Spirtes & Glymour, 1991), implemented with the Kernel-based conditional independence test (KCIT) (Zhang et al., 2011) and its ensemble variant E-KCIT (Guan & Kuang, 2026). (ii) *Score-based:* Greedy Equivalence Search (GES) (Chickering, 2002). (iii) *Functional-based:* ICALiNGAM (Shimizu et al., 2006) and DirectLiNGAM (Shimizu et al., 2011). (iv) *Continuous optimization:* Notears (Zheng et al., 2018), GraNDAG (Lachapelle et al., 2020), and GOLEM (Ng et al., 2020). (v) *Amortized methods:* AVICI (Lorch et al., 2022), a pre-trained amortized transformer-based causal discovery method.

Implementation details. For GES, ICALiNGAM, DirectLiNGAM, Notears, GraNDAG, and GOLEM, we use the implementations provided by the `gCastle` (Zhang et al., 2021) library with their default hyperparameter settings. PC with KCIT is implemented using `causal-learn` (Zheng et al., 2024), while E-KCIT is evaluated using the official implementation with its default configuration (Guan & Kuang, 2026). For AVICI, we use the official pretrained `scm-v0` model recommended

Table 1: Performance evaluation on synthetic datasets ($d \in \{10, 15, 20\}$). Results are reported as mean $\pm$ standard deviation over 25 repetitions for each graph size. Bold and underline denote the best and second-best mean values, respectively (ties receive the same formatting).

Method	$\text{SHD}_c \downarrow$	Precision $\uparrow$	Recall $\uparrow$	F1 $\uparrow$
		$d = 10$		
DCD-PFN	**6.560 ± 1.121**	**0.949 ± 0.129**	**0.289 ± 0.116**	**0.431 ± 0.140**
AVICI	7.440 ± 1.325	0.695 ± 0.459	0.173 ± 0.147	0.264 ± 0.208
PC (KCIT)	8.680 ± 1.492	0.152 ± 0.159	0.120 ± 0.128	0.133 ± 0.140
PC (E-KCIT)	7.960 ± 1.338	0.212 ± 0.199	0.142 ± 0.130	0.167 ± 0.151
GES	8.640 ± 2.196	0.230 ± 0.271	0.173 ± 0.198	0.196 ± 0.227
ICALiNGAM	7.560 ± 1.158	0.800 ± 0.408	0.160 ± 0.129	0.257 ± 0.180
DirectLiNGAM	7.560 ± 1.193	0.760 ± 0.436	0.160 ± 0.133	0.255 ± 0.188
Notears	8.240 ± 0.779	0.600 ± 0.500	0.084 ± 0.087	0.145 ± 0.140
GraNDAG	13.320 ± 4.905	0.323 ± 0.372	0.151 ± 0.106	0.174 ± 0.133
GOLEM	7.720 ± 0.980	0.800 ± 0.408	0.142 ± 0.109	0.234 ± 0.161
		$d = 15$		
DCD-PFN	**11.000 ± 1.581**	0.883 ± 0.166	**0.260 ± 0.092**	**0.391 ± 0.117**
AVICI	12.560 ± 4.234	0.691 ± 0.397	0.186 ± 0.111	0.281 ± 0.167
PC (KCIT)	14.400 ± 1.633	0.106 ± 0.100	0.094 ± 0.089	0.100 ± 0.094
PC (E-KCIT)	12.880 ± 1.424	0.169 ± 0.139	0.114 ± 0.092	0.135 ± 0.110
GES*	13.708 ± 2.476	0.205 ± 0.170	0.161 ± 0.130	0.178 ± 0.145
ICALiNGAM	11.920 ± 1.038	0.870 ± 0.302	0.151 ± 0.075	0.253 ± 0.114
DirectLiNGAM	11.360 ± 1.150	0.950 ± 0.204	0.191 ± 0.082	0.313 ± 0.118
Notears	12.640 ± 1.114	0.787 ± 0.407	0.100 ± 0.080	0.172 ± 0.127
GraNDAG	19.680 ± 6.866	0.386 ± 0.349	0.166 ± 0.087	0.187 ± 0.096
GOLEM	11.680 ± 1.030	**0.990 ± 0.050**	0.169 ± 0.074	0.281 ± 0.104
		$d = 20$		
DCD-PFN	**14.520 ± 2.740**	0.826 ± 0.184	**0.301 ± 0.101**	**0.435 ± 0.162**
AVICI	21.760 ± 11.061	0.483 ± 0.395	0.173 ± 0.152	0.231 ± 0.189
PC (KCIT)	20.280 ± 2.923	0.125 ± 0.091	0.101 ± 0.076	0.111 ± 0.082
PC (E-KCIT)	18.040 ± 1.837	0.177 ± 0.121	0.109 ± 0.066	0.133 ± 0.082
GES	19.840 ± 3.118	0.170 ± 0.148	0.149 ± 0.131	0.159 ± 0.138
ICALiNGAM	16.200 ± 1.323	0.930 ± 0.221	0.149 ± 0.071	0.251 ± 0.108
DirectLiNGAM	15.640 ± 1.524	0.982 ± 0.070	0.183 ± 0.083	0.300 ± 0.113
Notears	17.120 ± 1.013	0.920 ± 0.277	0.099 ± 0.053	0.176 ± 0.089
GraNDAG	31.000 ± 19.636	0.290 ± 0.304	0.099 ± 0.046	0.123 ± 0.072
GOLEM	15.800 ± 1.414	**0.983 ± 0.063**	0.175 ± 0.080	0.288 ± 0.109

for out-of-the-box inference (Lorch et al., 2022). We do not perform dataset-specific tuning for any baseline or our DCD-PFN. The architecture and optimization settings of DCD-PFN are provided in Appendix A.1.

Metrics. Since the evaluated methods may return either DAGs or CPDAGs, we convert estimated DAGs to their corresponding CPDAGs and evaluate all methods against the reference CPDAGs. We report standard evaluation metrics: Structural Hamming Distance adapted for CPDAGs (SHD_c), Precision, Recall, and F1-score, as detailed in Appendix A.2. For repeated experiments, we report the mean and standard deviation of each metric across datasets.

4.2 Synthetic Datasets

We construct synthetic datasets from continuous Structural Equation Models (SEMs) on random DAGs with $d \in \{10, 15, 20\}$ nodes. Graph structures are sampled using the Barabási-Albert model, with each new node attaching to one existing node. Edge directions are assigned according to a topological order to ensure acyclicity, followed by a random permutation of node labels. The observational data is generated sequentially via the following additive noise model:

$$x_i = f(W_i^\top x_{\text{pa}(i)} + b_i) + \epsilon_i \tag{14}$$

Table 2: Performance evaluation on the SynTReN semi-synthetic benchmark. Each of the 10 evaluation graphs contains 20 nodes and 19–34 edges. Results are reported as mean ± standard deviation across all 10 datasets. Bold and underline denote the best and second-best mean values, respectively (ties receive the same formatting).

Method	$\text{SHD}_c \downarrow$	Precision ↑	Recall ↑	F1 ↑
DCD-PFN	**21.100 ± 5.507**	0.504 ± 0.211	**0.166 ± 0.115**	**0.244 ± 0.151**
AVICI	26.300 ± 10.307	0.219 ± 0.271	0.064 ± 0.084	0.098 ± 0.127
PC (KCIT)	22.700 ± 4.877	0.190 ± 0.114	0.132 ± 0.077	0.155 ± 0.091
PC (E-KCIT)	22.300 ± 4.547	0.167 ± 0.113	0.092 ± 0.057	0.117 ± 0.074
GES	23.600 ± 5.700	0.214 ± 0.130	<u>0.155 ± 0.104</u>	<u>0.177 ± 0.112</u>
ICALiNGAM	21.600 ± 4.695	0.590 ± 0.270	0.086 ± 0.060	0.146 ± 0.093
DirectLiNGAM	<u>21.500 ± 4.649</u>	<u>0.617 ± 0.201</u>	0.090 ± 0.058	0.152 ± 0.088
Notears	22.500 ± 4.552	0.483 ± 0.412	0.046 ± 0.050	0.083 ± 0.086
GraNDAG	35.100 ± 12.679	0.177 ± 0.168	0.078 ± 0.062	0.082 ± 0.051
GOLEM	<u>21.500 ± 4.649</u>	**0.688 ± 0.213**	0.090 ± 0.054	0.155 ± 0.083

where $x_{\text{pa}(i)}$ denotes the feature vector of node i's parents. The weight matrix W_i is initialized using the Glorot uniform scheme, and the bias term b_i is sampled from a standard normal distribution. To reflect nonlinear causal mechanisms, the transformation function $f(\cdot)$ is uniformly sampled from the set $\{z, z^2, z^3, \tanh(z), \exp(z), \log(|z|)\}$. The intermediate output of each node is standardized to zero mean and unit variance to prevent numerical overflow and maintain a stable signal scale.

We model the additive noise ϵ_i using four standard distributions: Gaussian, Laplace, Uniform and Student's t. For each node independently, the noise distribution is chosen uniformly at random, and its standard deviation is sampled uniformly from the interval $[1.0, 5.0]$. Finally, for each simulated causal graph, we draw $n = 800$ observational samples and conduct 25 repetitions with different random seeds, resampling both the graph structure and data for each run. The synthetic evaluation changes the graph-generation distribution and the data-generating mechanism used during DCD-PFN pre-training, thereby evaluating its generalization capabilities.

As shown in Table 1, DCD-PFN achieves the lowest mean SHD_c and highest mean F1-scores at all three graph sizes. Its mean F1-scores are 0.431, 0.391, and 0.435 for $d = 10$, 15, and 20, respectively, compared with 0.264, 0.281, and 0.231 for AVICI. These correspond to relative improvements of approximately 63%, 39%, and 88%. The comparison with GOLEM illustrates the precision-recall trade-off: at $d = 15$ and $d = 20$, GOLEM achieves higher precision, whereas DCD-PFN achieves higher recall and F1-score. Nevertheless, DCD-PFN's mean recall remains between 0.260 and 0.301, indicating that substantial portions of the reference structures remain unrecovered. (*For GES at $d = 15$, one of the 25 runs failed to produce a valid graph. The reported metrics summarize the remaining 24 valid runs.)

4.3 Semi-Synthetic Datasets

We evaluate DCD-PFN on two semi-synthetic benchmarks based on SynTReN and Child graph structures. Specifically, we use the ten SynTReN graphs used in the GraNDAG study (Lachapelle et al., 2020), each containing 20 nodes and 19–34 directed edges, and the Child Bayesian network (Spiegelhalter & Cowell, 1992), containing 20 nodes and 25 directed edges. For both benchmarks, we retain the graph structures and generate continuous observations using the same SEM procedure, mechanism family, and noise settings described in Section 4.2, with $n = 800$ samples per dataset. This evaluation therefore assesses recovery on benchmark topologies under controlled synthetic mechanisms.

For SynTReN, we generate one dataset per graph and report the mean and standard deviation across the ten datasets in Table 2. For Child, we keep the graph fixed and perform 25 independent data-generation repetitions, with the full results provided in Table 4 of Appendix A.3.

On SynTReN, DCD-PFN achieves the lowest mean SHD_c (21.100) and the highest mean recall (0.166) and F1-score (0.244). GOLEM achieves the highest mean precision (0.688), while DCD-PFN attains a mean precision of 0.504 with higher recall. These results support the effectiveness of DCD-PFN across the evaluated benchmark topologies.

Table 3: Performance evaluation on the Sachs Flow-Cytometry Dataset ($n = 1755$, `cd3cd28` and `cd3cd28icam2` conditions). Bold and underline denote the best and second-best values, respectively (ties receive the same formatting).

Method	$\text{SHD}_c \downarrow$	Precision ↑	Recall ↑	F1 ↑
DCD-PFN	**14**	**1.000**	**0.300**	**0.462**
AVICI	19	0.286	0.100	0.148
PC (KCIT)	19	0.333	0.200	0.250
PC (E-KCIT)	16	0.545	**0.300**	0.387
GES	19	0.308	0.200	0.242
ICALiNGAM	<u>15</u>	0.714	<u>0.250</u>	0.370
DirectLiNGAM	<u>15</u>	<u>0.833</u>	<u>0.250</u>	0.385
Notears	18	0.500	0.150	0.231
GraNDAG	20	0.000	0.000	0.000
GOLEM	<u>15</u>	0.667	**0.300**	<u>0.414</u>

4.4 Real-world Dataset

We evaluate DCD-PFN on the Sachs Flow-Cytometry dataset (Sachs et al., 2005), a widely used benchmark in causal discovery. The dataset contains measurements of $d = 11$ phosphorylated proteins and phospholipids derived from primary human CD4+ T cells. Following prior work, we use the consensus signaling network containing 11 nodes and 20 directed edges, derived from the original experimental study (Sachs et al., 2005) as the reference causal graph.

We adopt two commonly used observational configurations of this dataset:

- $n = 1755$: samples from the `cd3cd28` and `cd3cd28icam2` conditions are combined (Lorch et al., 2022; Li et al., 2023; Guan & Kuang, 2026).
- $n = 853$: only samples from the `cd3cd28` condition are used (Lachapelle et al., 2020; Ng et al., 2020; Mukherjee et al., 2020).

Table 3 reports the results for the $n = 1755$ configuration, while the corresponding results for $n = 853$ are provided in Table 5 of Appendix A.4. Under the $n = 1755$ configuration, DCD-PFN achieves the lowest structural error ($\text{SHD}_c = 14$) together with a precision of 1.000, indicating that the recovered adjacencies contain no false positives with respect to the reference CPDAG. DCD-PFN also achieves the highest F1-score among the evaluated methods, demonstrating a favorable balance between precision and recall. In contrast, the pretrained amortized baseline AVICI exhibits lower structural recovery performance in this real-world setting. Overall, the results support the zero-shot generalization capability of DCD-PFN from synthetic pre-training distributions to real-world biological data.

5 Conclusion

In this paper, we introduced DCD-PFN, a novel foundation model tailored for explicit structural causal discovery. By leveraging the powerful in-context learning capabilities of Prior-Data Fitted Networks on tabular data, DCD-PFN achieves end-to-end identification of local causal structures through a decoupling-based mechanism. To circumvent the inherent challenges of direct amortized inference for global graph reconstruction, we adopted a local-to-global approach. By pre-training on a diverse range of SCMs, DCD-PFN implicitly internalizes complex decoupling rules. Experiments on synthetic and real-world tabular datasets demonstrate that our foundation model achieves robust zero-shot generalization and out-of-the-box structural reconstruction.

Limitations and future directions. DCD-PFN's learned MB predictions can introduce errors into graph reconstruction, and parallelization does not eliminate the subset-search cost for large boundaries. Moreover, pairwise decoupling is a surrogate for mutual independence. Future work includes more efficient query selection, stronger decoupling objectives, and improved robustness beyond the pre-training distribution. Further discussion appears in Appendix B.

AI USE STATEMENT

In this work, we used generative AI tools for assisting with translation and assisting in the writing of proofs. We have not used generative AI tools for helping develop theoretical models or conceptual frameworks, or any other tasks with required disclosure. Additionally, we used generative AI tools for editing the paper to improve readability. We have reviewed all AI-assisted work. All revised text was manually checked for clarity and technical accuracy. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Constantin F. Aliferis, Alexander Statnikov, Ioannis Tsamardinos, Subramani Mani, and Xenofon D. Koutsoukos. Local causal and Markov blanket induction for causal discovery and feature selection for classification part i: Algorithms and empirical evaluation. *Journal of Machine Learning Research*, 11(7):171–234, 2010. URL http://jmlr.org/papers/v11/aliferis10a.html.

Yashas Annadani, Panagiotis Tigas, Stefan Bauer, and Adam Foster. Amortized active causal induction with deep reinforcement learning. In *ICML 2024 Workshop on Structured Probabilistic Inference & Generative Modeling*, 2024.

Yikang Chen, Xingzhe Sun, and Dehui du. Causal discovery via quantile partial effect. In *International Conference on Learning Representations*, 2026.

Zi-Jian Cheng, Zi-Yi Jia, Zhi Zhou, Yu-Feng Li, and Lan-Zhe Guo. Realistic evaluation of tabpfn v2 in open environments, 2025.

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

Anish Dhir, Matthew Ashman, James Requeima, and Mark van der Wilk. A meta-learning approach to bayesian causal discovery. In *The Thirteenth International Conference on Learning Representations*, 2025.

Mohamed Djilani, Thibault Simonetto, Karim Tit, Florian Tambon, Paul Récamier, Salah Ghamizi, Maxime Cordy, and Mike Papadakis. On the robustness of tabular foundation models: Test-time attacks and in-context defenses, 2025.

Bao Duong, Nu Hoang, and Thin Nguyen. Amortized conditional independence testing. In Xintao Wu, Myra Spiliopoulou, Can Wang, Vipin Kumar, Longbing Cao, Yanqiu Wu, Yu Yao, and Zhangkai Wu (eds.), *Advances in Knowledge Discovery and Data Mining*, pp. 410–423, Singapore, 2025. Springer Nature Singapore. ISBN 978-981-96-8170-9.

Clark Glymour, Kun Zhang, and Peter Spirtes. Review of causal discovery methods based on graphical models. *Frontiers in genetics*, 10:524, 2019.

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

Zhengkang Guan and Kun Kuang. Efficient ensemble conditional independence test framework for causal discovery. In *International Conference on Learning Representations*, 2026.

Zhengkang Guan, Fei Wu, and Kun Kuang. Decoupled causal discovery, 2026. URL https://arxiv.org/abs/2609.23535.

Kai Helli, David Schnurr, Noah Hollmann, Samuel Müller, and Frank Hutter. Drift-resilient tabpfn: In-context learning temporal distribution shifts on tabular data. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 98742–98781. Curran Associates, Inc., 2024. doi: 10.52202/079017-3134.

Noah Hollmann, Samuel Müller, Katharina Eggensperger, and Frank Hutter. TabPFN: A transformer that solves small tabular classification problems in a second. In *The Eleventh International Conference on Learning Representations*, 2023.

Noah Hollmann, Samuel Müller, Lennart Purucker, Arjun Krishnakumar, Max Körfer, Shi Bin Hoo, Robin Tibor Schirrmeister, and Frank Hutter. Accurate predictions on small data with a tabular foundation model. *Nature*, 637(8045):319–326, January 2025. ISSN 1476-4687. doi: 10.1038/s41586-024-08328-6.

Nan Rosemary Ke, Silvia Chiappa, Jane X Wang, Jorg Bornschein, Anirudh Goyal, Melanie Rey, Theophane Weber, Matthew Botvinick, Michael Curtis Mozer, and Danilo Jimenez Rezende. Learning to induce causal structure. In *International Conference on Learning Representations*, 2023.

Jannik Kossen, Neil Band, Clare Lyle, Aidan Gomez, Tom Rainforth, and Yarin Gal. Self-attention between datapoints: Going beyond individual input-output pairs in deep learning. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, 2021.

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

Shuai Li, Yingjie Zhang, Hongtu Zhu, Christina Wang, Hai Shu, Ziqi Chen, Zhuoran Sun, and Yanfeng Yang. K-nearest-neighbor local sampling based conditional independence testing. *Advances in Neural Information Processing Systems*, 36, 2023.

Zhaolong Ling, Kui Yu, Hao Wang, Lin Liu, Wei Ding, and Xindong Wu. Bamb: A balanced Markov blanket discovery approach to feature selection. *ACM Trans. Intell. Syst. Technol.*, 10(5): 52:1–52:25, 2019.

Lars Lorch, Scott Sussex, Jonas Rothfuss, Andreas Krause, and Bernhard Schölkopf. Amortized inference for causal structure learning. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 13104–13118. Curran Associates, Inc., 2022.

Yuchen Ma, Dennis Frauen, Emil Javurek, and Stefan Feuerriegel. Foundation models for causal inference via prior-data fitted networks. In *The Fourteenth International Conference on Learning Representations*, 2026.

Dimitris Margaritis and Sebastian Thrun. Bayesian network induction via local neighborhoods. In S. Solla, T. Leen, and K. Müller (eds.), *Advances in Neural Information Processing Systems*, volume 12. MIT Press, 1999.

Christopher Meek. Causal inference and causal explanation with background knowledge. In Philippe Besnard and Steve Hanks (eds.), *UAI '95: Proceedings of the Eleventh Annual Conference on Uncertainty in Artificial Intelligence, Montreal, Quebec, Canada, August 18-20, 1995*, pp. 403–410. Morgan Kaufmann, 1995.

Sudipto Mukherjee, Himanshu Asnani, and Sreeram Kannan. Ccmi: Classifier based conditional mutual information estimation. In *Uncertainty in artificial intelligence*, pp. 1083–1093. PMLR, 2020.

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

Thomas Nagler. Statistical foundations of prior-data fitted networks. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023.

Ignavier Ng, AmirEmad Ghassami, and Kun Zhang. On the role of sparsity and dag constraints for learning linear dags. *Advances in Neural Information Processing Systems*, 33:17943–17954, 2020.

Judea Pearl. *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan Kaufmann, 1988.

Jingang Qu, David Holzmüller, Gaël Varoquaux, and Marine Le Morvan. TabICLv2: A better, faster, scalable, and open tabular foundation model. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*. PMLR, 2026. URL https://openreview.net/forum?id=SxsyLjIfWB.

Jake Robertson, Arik Reuter, Siyuan Guo, Noah Hollmann, Frank Hutter, and Bernhard Schölkopf. Do-PFN: In-context learning for causal effect estimation. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

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.

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), 2006.

Shohei Shimizu, Takanori Inazumi, Yasuhiro Sogawa, Aapo Hyvarinen, Yoshinobu Kawahara, Takashi Washio, Patrik O Hoyer, and Kenneth Bollen. DirectLiNGAM: A direct method for learning a linear non-gaussian structural equation model. *Journal of Machine Learning Research-JMLR*, 12(Apr):1225–1248, 2011.

David J Spiegelhalter and Robert G Cowell. Learning in probabilistic expert systems. In J M Bernardo, J O Berger, P Dawid, A F M Smith, J M Bernardo, J O Berger, P Dawid, and A F M Smith (eds.), *Bayesian Statistics 4: Proceedings of the Fourth Valencia International Meeting, Dedicated to the memory of Morris H. DeGroot, 1931–1989*. Oxford University Press, 08 1992. ISBN 9780198522669. doi: 10.1093/oso/9780198522669.003.0025. URL https://doi.org/10.1093/oso/9780198522669.003.0025.

Peter Spirtes and Clark Glymour. An algorithm for fast recovery of sparse causal graphs. *Social Science Computer Review*, 9:62–72, 1991.

Peter Spirtes and Kun Zhang. Causal discovery and inference: concepts and recent methodological advances. In *Applied informatics*, volume 3, pp. 1–28. SpringerOpen, 2016.

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 (UAI)*, pp. 499–506. Morgan Kaufmann, 1995.

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

I. Tsamardinos, Constantin F. Aliferis, and Alexander R. Statnikov. Time and sample efficient discovery of Markov blankets and direct causal relations. In *Knowledge Discovery and Data Mining*, 2003.

Hao Wang, Zhaolong Ling, Kui Yu, and Xindong Wu. Towards efficient and effective discovery of Markov blankets for feature selection. *Inf. Sci.*, 509:227–242, 2020.

Xingyu Wu, Bingbing Jiang, Kui Yu, Chunyan Miao, and Huanhuan Chen. Accurate Markov boundary discovery for causal feature selection. *IEEE Trans. Cybern.*, 50(12):4983–4996, 2020.

Han-Jia Ye, Si-Yang Liu, and Wei-Lun Chao. A closer look at tabPFN v2: Understanding its strengths and extending its capabilities. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Keli Zhang, Shengyu Zhu, Marcus Kalander, Ignavier Ng, Junjian Ye, Zhitang Chen, and Lujia Pan. gcastle: A python toolbox for causal discovery, 2021.

Kun Zhang, Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. Kernel-based conditional independence test and application in causal discovery. In Fábio Gagliardi Cozman and Avi Pfeffer (eds.), *UAI 2011, Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence, Barcelona, Spain, July 14-17, 2011*, pp. 804–813. AUAI Press, 2011.

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

Xiyuan Zhang, Danielle C. Maddix, Junming Yin, Nick Erickson, Abdul Fatir Ansari, Boran Han, Shuai Zhang, Leman Akoglu, Christos Faloutsos, Michael W. Mahoney, Cuixiong Hu, Huzefa Rangwala, George Karypis, and Bernie Wang. Mitra: Mixed synthetic priors for enhancing tabular foundation models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025b.

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

Yujia Zheng, Biwei Huang, Wei Chen, Joseph Ramsey, Mingming Gong, Ruichu Cai, Shohei Shimizu, Peter Spirtes, and Kun Zhang. Causal-learn: Causal discovery in python. *Journal of Machine Learning Research*, 25(60):1–8, 2024.

A Additional Experimental Details and Results

A.1 Model Configuration and Optimization

DCD-PFN uses $L_0 = 2$ Dual-Axis Attention Blocks in the shared encoder, $L_1 = 3$ in the decoupling branch, and $L_2 = 3$ in the Markov boundary (MB) learning branch. The embedding dimension is 128, and multi-head attention uses 8 heads. We optimize all model parameters jointly using AdamW with an initial learning rate of 10^{-4} and weight decay of 10^{-3}. The learning rate follows a cosine annealing schedule with $T_{\max}$ set to the total number of training steps. The joint objective uses a decoupling coefficient of $\lambda_w = 0.1$.

A.2 Evaluation Metrics

We evaluate graph recovery by comparing the estimated and reference CPDAGs over unordered node pairs. For each pair of nodes $i < j$, let $s_{ij}, \hat{s}_{ij} \in \{0, 1, 2, 3\}$ denote the reference and estimated edge states, respectively, where 0 indicates no edge, 1 indicates $i \to j$, 2 indicates $j \to i$, and 3 indicates $i - j$. Thus, an undirected edge is treated as a single edge state rather than two directed edges.

Structural Hamming distance. We define the CPDAG structural Hamming distance as

$$\mathrm{SHD}_c = \sum_{i<j} \mathbb{I}[\hat{s}_{ij} \neq s_{ij}], \tag{15}$$

where $\mathbb{I}[\cdot]$ is the indicator function. Each mismatched node pair contributes one unit, whether the mismatch is an extra edge, a missing edge, a reversed direction, or a discrepancy between a directed and an undirected edge. In particular, reversing an edge incurs a cost of one rather than two.

Precision, recall, and F1-score. These metrics require an exact match of a nonzero edge state and therefore account for both adjacency and orientation. We define

$$\mathrm{TP} = \sum_{i<j} \mathbb{I}[\hat{s}_{ij} = s_{ij} \neq 0], \tag{16}$$

$$\mathrm{FP} = \sum_{i<j} \mathbb{I}[\hat{s}_{ij} \neq 0 \ \wedge \ \hat{s}_{ij} \neq s_{ij}], \tag{17}$$

$$\mathrm{FN} = \sum_{i<j} \mathbb{I}[s_{ij} \neq 0 \ \wedge \ \hat{s}_{ij} \neq s_{ij}]. \tag{18}$$

An extra edge contributes one false positive, and a missing edge contributes one false negative. If both graphs contain an edge between a node pair but disagree on its direction or directed/undirected status, the mismatch contributes both one false positive and one false negative, while contributing only one unit to SHD_c. The resulting metrics are

$$\mathrm{Precision} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}, \tag{19}$$

$$\mathrm{Recall} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}, \tag{20}$$

$$\mathrm{F1} = \frac{2\,\mathrm{Precision}\,\mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}}. \tag{21}$$

Any metric with a zero denominator is assigned a value of zero. Node pairs with no edge in either graph do not contribute to these metrics.

A.3 Additional Semi-Synthetic Results on Child

We report additional results on the Child Bayesian network, which contains 20 nodes and 25 directed edges (Spiegelhalter & Cowell, 1992). Keeping its graph structure fixed, we generate 25 datasets with different random seeds using the SEM procedure described in the main text, each containing $n = 800$ observations. Table 4 reports the mean and standard deviation of each evaluation metric across these 25 repetitions.

Table 4: Performance evaluation on the Child semi-synthetic benchmark. The fixed Child network topology consists of 20 nodes and 25 edges. Results are reported as mean ± standard deviation over 25 independent data-generation repetitions. Bold and underline denote the best and second-best mean values, respectively (ties receive the same formatting).

Method	$\text{SHD}_c \downarrow$	Precision ↑	Recall ↑	F1 ↑
DCD-PFN	<u>23.440 ± 1.557</u>	0.453 ± 0.151	0.109 ± 0.048	**0.173 ± 0.071**
AVICI	29.720 ± 8.801	0.318 ± 0.345	0.075 ± 0.060	0.099 ± 0.068
PC (KCIT)	24.600 ± 3.122	0.203 ± 0.098	<u>0.142</u> ± 0.068	<u>0.166 ± 0.079</u>
PC (E-KCIT)	**23.240 ± 2.204**	0.192 ± 0.107	0.110 ± 0.064	0.139 ± 0.079
GES	25.960 ± 3.285	0.207 ± 0.109	**0.152 ± 0.088**	**0.173 ± 0.095**
ICALiNGAM	23.880 ± 1.236	0.490 ± 0.357	0.048 ± 0.045	0.086 ± 0.075
DirectLiNGAM	23.600 ± 1.384	<u>0.525 ± 0.337</u>	0.061 ± 0.052	0.107 ± 0.086
Notears	24.360 ± 0.810	0.493 ± 0.447	0.030 ± 0.029	0.056 ± 0.052
GraNDAG	38.120 ± 15.759	0.181 ± 0.235	0.067 ± 0.060	0.077 ± 0.062
GOLEM	23.720 ± 1.242	**0.533 ± 0.372**	0.056 ± 0.046	0.098 ± 0.077

Table 5: Performance evaluation on the Sachs Flow-Cytometry Dataset ($n = 853$, `cd3cd28` condition). Bold and underline denote the best and second-best values, respectively (ties receive the same formatting).

Method	$\text{SHD}_c \downarrow$	Precision ↑	Recall ↑	F1 ↑
DCD-PFN	<u>16</u>	0.667	<u>0.200</u>	<u>0.308</u>
AVICI	18	0.429	0.150	0.222
PC (KCIT)	**14**	<u>0.700</u>	**0.350**	**0.467**
PC (E-KCIT)	**14**	<u>0.700</u>	**0.350**	**0.467**
GES	<u>16</u>	0.500	<u>0.200</u>	0.286
ICALiNGAM	<u>16</u>	0.667	<u>0.200</u>	<u>0.308</u>
DirectLiNGAM	17	**0.750**	0.150	0.250
Notears	18	0.500	0.100	0.167
GraNDAG	20	0.000	0.000	0.000
GOLEM	24	0.200	<u>0.200</u>	0.200

DCD-PFN achieves a mean F1-score of 0.173, matching that of GES, while obtaining a lower mean structural error of $\text{SHD}_c = 23.440$. PC with E-KCIT achieves the lowest mean structural error (23.240), while GOLEM and GES attain the highest mean precision (0.533) and recall (0.152), respectively. These results provide additional evidence of competitive structural recovery by DCD-PFN.

A.4 Additional Results on the Sachs Dataset

We additionally evaluate all methods on the commonly used $n = 853$ configuration of the Sachs dataset, which contains samples only from the `cd3cd28` experimental condition. This setting has been adopted in several previous causal discovery studies (Lachapelle et al., 2020; Ng et al., 2020; Mukherjee et al., 2020). The results are reported in Table 5.

Under the $n = 853$ configuration, the relative performance across methods differs from that observed for the $n = 1755$ configuration reported in the main text. PC with KCIT and E-KCIT achieve the lowest structural error ($\text{SHD}_c = 14$) and the highest F1-score (0.467). Among the remaining methods, DCD-PFN achieves the best performance, with an F1-score of 0.308 and $\text{SHD}_c = 16$, demonstrating competitive structural recovery and providing further evidence of its generalization capability.

Additionally, under the default `gCastle` configuration (Zhang et al., 2021) used consistently throughout our experiments, GraNDAG returns a graph containing a single edge under both Sachs observational configurations. In each case, the predicted edge does not overlap with the reference graph, resulting in zero Precision, Recall, and F1-score.

Table 6: Ablation of the decoupling mechanism. "w/o L_w" sets $\lambda_w = 0$ during training while keeping the architecture unchanged. Results on repeated datasets are reported as mean $\pm$ standard deviation.

Dataset	Method	$\text{SHD}_c \downarrow$	Precision $\uparrow$	Recall $\uparrow$	F1 $\uparrow$
Synthetic ($d = 10$)	DCD-PFN	6.56 ± 1.12	0.95 ± 0.13	0.29 ± 0.12	0.43 ± 0.14
	w/o L_w	9.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
Synthetic ($d = 15$)	DCD-PFN	11.00 ± 1.58	0.88 ± 0.17	0.26 ± 0.09	0.39 ± 0.12
	w/o L_w	14.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
Synthetic ($d = 20$)	DCD-PFN	14.52 ± 2.74	0.83 ± 0.18	0.30 ± 0.10	0.44 ± 0.16
	w/o L_w	19.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
SynTReN	DCD-PFN	21.10 ± 5.51	0.50 ± 0.21	0.17 ± 0.12	0.24 ± 0.15
	w/o L_w	23.50 ± 4.17	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
Child	DCD-PFN	23.44 ± 1.56	0.45 ± 0.15	0.11 ± 0.05	0.17 ± 0.07
	w/o L_w	25.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
Sachs ($n = 1755$)	DCD-PFN	14	1.00	0.30	0.46
	w/o L_w	20	0.00	0.00	0.00
Sachs ($n = 853$)	DCD-PFN	16	0.67	0.20	0.31
	w/o L_w	20	0.00	0.00	0.00

A.5 Justification and Ablation of the Decoupling Mechanism

Guan et al. (2026) show that (under its theoretical assumptions) an appropriate reweighting of the observational distribution can isolate the MB of a target variable through its remaining dependencies. This result motivates our use of sample-wise weights and an auxiliary decoupling objective. Our pairwise HSIC loss serves as a practical surrogate for encouraging the desired decoupling during pre-training.

To examine its empirical role, we retrain DCD-PFN with $\lambda_w = 0$, keeping the architecture, pre-training procedure, and evaluation protocol unchanged. As shown in Table 6, the ablated model consistently produces empty graphs across the evaluated datasets, whereas the full model recovers nontrivial structure. Thus, under our training setup, MB supervision alone does not yield effective graph recovery. The auxiliary decoupling loss is critical to avoiding this degenerate outcome. This ablation evaluates the contribution of the loss, while the weighting branch remains present in both variants.

B Limitations and Future Directions

Learned local inference. The soundness and completeness guarantees of DCD assume exact MB queries under the assumptions of its theoretical framework. DCD-PFN instead supplies learned predictions, whose errors can affect both skeleton recovery and edge orientation. The fixed decision threshold of 0.5 also does not account explicitly for uncertainty across datasets or queries. Future work could investigate calibrated MB scores and uncertainty-aware aggregation of local predictions.

Decoupling and dependence measurement. Our pairwise HSIC objective encourages independence between pairs of non-target variables, but pairwise independence does not generally imply mutual independence. Likewise, the finite-dimensional embedding statistics used by the MB readout do not necessarily characterize all forms of statistical dependence. These components should therefore be understood as learned surrogates for the population decoupling principle. Exploring higher-order independence objectives and richer dependence representations is a promising direction.

Reconstruction complexity. Parallel execution distributes local discovery across targets but does not remove the combinatorial subset search within each target branch. Consequently, reconstruction can remain costly when the initial MBs are large, and practical speedups depend on available

computational resources. Adaptive query selection, pruning strategies, and batching of independent queries may further improve scalability.

Transfer beyond the pre-training distribution. Although pre-training uses diverse synthetic SCMs, this prior cannot cover all mechanisms encountered in practice. Performance may deteriorate when new datasets differ substantially in their mechanisms, noise distributions, or structural properties. Extending the prior and developing adaptation methods for such distribution shifts are important directions. Latent confounding and other departures from the assumptions of the underlying reconstruction framework also require additional methodological development.