CDFM: TOWARDS A GENERAL-PURPOSE CAUSAL DISCOVERY FOUNDATION MODEL

Anonymous authors
Paper under double-blind review

ABSTRACT

Causal discovery, the process of recovering underlying causal structures from observational data, is a fundamental pursuit across scientific disciplines. Over the past decades, numerous algorithms have been developed to tackle this challenge through workflows tailored to the specific causal mechanisms underlying each type of dataset, demonstrating effectiveness across a wide range of applications. However, as the volume and heterogeneity of real-world data continue to grow, this dataset-specific approach inevitably leads to a fragmented, test-driven paradigm that struggles to scale to the demands of modern scientific discovery. To address this, we formulate the Causal Discovery Foundation Model (CDFM) as a unified, general-purpose framework for zero-shot structural inference. To ensure reliable generalization across unknown domains, we first investigate the theoretical boundaries of causal identifiability, revealing the indispensable role of causal prior mechanisms in this process. Building on these insights, we formulate a principled variational framework that treats unknown causal mechanisms as latent variables and mathematically decomposes the intractable marginal likelihood into distinct, tractable learning modules. The variational decomposition provides a conceptual design principle for the CDFM architecture, while comprehensive causal knowledge guides the large-scale synthesis of our pretraining data. By pretraining on a massive, highly diverse space of synthetic structural causal models, CDFM successfully internalizes complex statistical asymmetries. Extensive experiments demonstrate that CDFM achieves strong and consistently competitive performance across heterogeneous synthetic and real-world benchmarks.

1 INTRODUCTION

Causal discovery, the process of recovering underlying causal structures from observational data, is a fundamental pursuit across scientific disciplines, with broad applications in fields such as economics (Ghysels et al., 2016), biology (Sachs et al., 2005), and social science (Cai et al., 2017). Over the past decades, numerous causal discovery algorithms have been developed, demonstrating effectiveness across a wide variety of datasets (Gamella et al., 2025; Mooij et al., 2016; Runge et al., 2020).

Despite these advancements, practical causal discovery remains fragmented because different algorithms rely on different assumptions about the data-generating process. Existing methods are typically designed for particular mechanism classes, such as linear non-Gaussian models (Shimizu et al., 2006) or additive-noise models (Hoyer et al., 2009). Their performance can therefore depend strongly on how well these assumptions match the dataset at hand.

This motivates a complementary question: can a single pretrained model amortize structural inference across a broad collection of causal mechanism classes, while still respecting the identifiability limits of observational data?

Recent foundation models for tabular data provide evidence that this strategy is plausible. Models such as TabPFN (Hollmann et al.), LimiX (Zhang et al., 2025), and TabICL (Qu et al.) learn reusable inductive biases from large synthetic or heterogeneous pretraining distributions. Related work has extended this idea to causal effect estimation (Balazadeh Meresht et al., 2025; Robertson et al., 2025; Ma et al., 2026) and amortized causal discovery (Lorch et al., 2022; Swelam et al., 2025). However,

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.

a general formulation connecting causal identifiability, mechanism uncertainty, architecture design, and pretraining coverage remains underdeveloped.

We introduce the **C**ausal **D**iscovery **F**oundation **M**odel (CDFM), a framework for direct structural inference across heterogeneous and initially unknown causal mechanisms. CDFM is pretrained on diverse synthetic data sampled from a broad space of identifiable causal assumptions, incorporating rich causal knowledge. The resulting model emerges as a general-purpose causal discovery foundation model that is capable of zero-shot structural inference across diverse, unknown mechanisms.

The main contributions of this work are summarized as follows:

- We formulate a causal discovery foundation model, a paradigm that is specifically tailored for general-purpose causal discovery, shifting the paradigm from specialized algorithm selection to a unified inference approach. To mathematically ground this, we unify foundation models and structural identifiability through a principled variational framework, treating unknown mechanisms as latent variables to translate the marginalization problem into a tractable learning objective.
- We design CDFM that aligns mechanism encoding, structure-aware reconstruction, and graph decoding with the theoretical decomposition, allowing the model to robustly infer causal structures directly from observational data.
- We pretrain CDFM on a large synthetic collection spanning diverse graph families, mechanisms, noises, data types, and perturbations, and evaluate one frozen model across heterogeneous synthetic and real-world benchmarks. We further use task-specific diagnostic experiments to probe behavior around known identifiability boundaries.

2 Related Work

The methodology of causal discovery from observational data has evolved significantly in recent years. In this section, we review the literature, tracing the progression from theoretically constrained algorithms to amortized inference and the generalized foundation models.

Constrained causal discovery. The traditional approach to causal discovery relies on algorithms derived under specific constraints about the data-generating process. Typical independence constraint-based methods, such as PC and FCI (Spirtes et al., 2000), exploit conditional independencies under the causal Markov and faithfulness assumptions, typically identifying structures only up to a Markov equivalence class (MEC). To resolve MEC ambiguities, additional Functional Causal Models (FCMs) are included for imposing strict structural constraints. Linear non-Gaussian models (LiNGAM) (Shimizu et al., 2006; 2011) break symmetry by assuming linear equations with non-Gaussian noise. Additive noise models (ANM) (Hoyer et al., 2009) and post-nonlinear models (Zhang & Hyvärinen, 2009) extend this identifiability to nonlinear regimes. This family is further enriched by causal additive models (Bühlmann et al., 2014) and causal location-scale noise models (Immer et al., 2023; Yin et al., 2024), which leverage dependent noise variance as an additional source of causal asymmetry. For discrete variables, several works have addressed ordinal, categorical, and count data Peters et al. (2010); Ni & Mallick (2022); Yao et al. (2025); Cai et al. (2018); Qiao et al. (2026; 2024); Xiang et al. (2024); Qiao et al. (2023). Later advancements transitioned from combinatorial search to continuous optimization (e.g., NOTEARS (Zheng et al., 2018)), utilizing smooth algebraic characterizations of acyclicity. These methods provide strong identification within their respective assumption classes, but they do not remove the need to specify which class is appropriate for a given dataset.

Amortized causal discovery. To overcome the bottleneck of solving dataset-specific optimization problems, amortized inference reframes causal discovery as a supervised learning task. Methods like AVICI (Lorch et al., 2022) and CSIvA (Ke et al., 2023) train neural networks to infer causal graphs directly from observational and interventional samples. This learning-based framework has been subsequently expanded across various dimensions, including applications to time-series data (Löwe et al., 2022), active causal induction using deep reinforcement learning (Annadani et al., 2024), and scalable conditional independence testing (Leiner et al., 2025). Further methodological extensions (Thompson et al., 2026; Wu et al., 2025; Yin et al., 2025; Peng et al., 2026) have demonstrated

the empirical feasibility of large-scale causal discovery. While these amortized models successfully internalize useful inductive biases from simulated data via empirical heuristics, a systematic theoretical framework for addressing structural identifiability across highly heterogeneous and unknown mechanisms remains unresolved.

Foundation models for tabular data. Large-scale pretraining has recently demonstrated remarkable in-context generalization on tabular and structured data tasks. Models such as LimiX (Zhang et al., 2025), TabPFN (Hollmann et al.), and TabICL (Qu et al.) have shown that foundation architectures can effectively model complex, heterogeneous datasets. Within the causal inference domain, efforts like CausalPFN (Balazadeh Meresht et al., 2025), Do-PFN (Robertson et al., 2025), and related foundation-model approaches amortize causal effect estimation over simulated data-generating processes. Causal discovery poses an additional challenge because the prediction target is itself a structural object whose identifiability depends on the mechanism class; scaling the pretraining distribution therefore has to be considered together with, rather than separately from, the assumptions under which different graphs are distinguishable.

3 Background and Preliminaries

We represent the data-generating process using a Structural Causal Model (SCM). Let $\mathbf{X} = \{X_1, X_2, \ldots, X_D\}$ be observed variables, where D denotes the number of variables, and $\mathbf{L}$ be unobserved latent confounders. The system is governed by structural assignments:

$$V_i := f_i(\mathbf{PA}_i, N_i), \quad \text{for } V_i \in \mathbf{V} = \mathbf{X} \cup \mathbf{L}, \tag{1}$$

where $\mathbf{PA}_i$ are direct causal parents, f_i is a deterministic mechanism, and N_i is exogenous noise. By marginalizing out $\mathbf{L}$, we represent the observed structure as an adjacency matrix $G \in \{0,1\}^{D\times D}$ encoding the causal structure over the observed variables after marginalizing out $\mathbf{L}$. In addition, to capture the effects of latent confounding, we augment the graph with bidirected edges, signifying the existence of at least one unobserved common cause $L \in \mathbf{L}$ that directly influences both variables. Note that if both $X_i \to X_j$ and latent confounding exist, we encode only the directed edge, and we restrict the target graph class to bow-free acyclic directed mixed graphs.

Given this formulation, CDFM predicts a graph over the observed variables, augmented with bidirected edges in the designated latent-confounding setting with unknown and heterogeneous mechanisms $\mathcal{M}$.

4 Identifiability

As discussed earlier by Pearl Pearl (2010), behind any causal conclusion, there must be some causal assumption, untested in observational studies. The role of a foundation model is therefore not to remove assumptions, but to learn and adopt over a broad spectrum of assumption classes. To develop principled guidelines for the design of CDFM, enabling it to generalize across diverse causal mechanisms, we therefore first investigate the theoretical boundaries of causal identifiability and guide the architecture design of CDFM.

In causal discovery, underlying assumptions are typically encapsulated implicitly within structural causal models. However, as established by Peters et al. (2017), an SCM with independent noise alone is insufficient to identify the causal direction.

Theorem 1 (Impossibility of Learning Causal Direction without Constraint)**.** *Let X and Y be two random variables. For any joint distribution $P_{X,Y}$, valid SCMs exist for both causal orderings simultaneously. That is, there exist measurable functions f_X, f_Y and real-valued noise variables N_X, N_Y such that:*

$$\begin{aligned} Y &= f_Y(X, N_Y), \quad X \perp\!\!\!\perp N_Y, \\ X &= f_X(Y, N_X), \quad Y \perp\!\!\!\perp N_X. \end{aligned} \tag{2}$$

Crucially, because this holds for any joint distribution, the variables are perfectly symmetric. Thus, identifying the true causal graph is mathematically impossible without imposing prior assumptions on the underlying structural equations. Traditional algorithms resolve this by manually asserting a specific, known causal prior mechanism (e.g., linear non-Gaussianity (Shimizu et al., 2011), or

additive noise model (Hoyer et al., 2009)). Given that the true mechanism M^* is known, identifiability transforms into a structural model selection problem. The optimal graph minimizes the Kullback-Leibler (KL) divergence between the true distribution and the model distribution.

Theorem 2 (Identifiability under Known Mechanism)**.** *Let the data $\mathbf{X}$ be generated by a true graph G^* and a known causal mechanism M^*. For any alternative graph G', the difference in expected log-likelihood under the true distribution $\mathcal{L}(G^*) := \mathbb{E}_{\mathbf{X}\sim P(\cdot|G^*,M^*)}[\log P(\mathbf{X} \mid G^*, M^*)]$ is bounded by the KL divergence:*

$$\mathcal{L}(G^*) - \mathcal{L}(G') = D_{\mathrm{KL}}\big[P(\mathbf{X}|G^*, M^*) \parallel P(\mathbf{X}|G', M^*)\big] \geq 0. \tag{3}$$

If M^* belongs to a strictly identifiable class, the inequality is strict for every $G' \neq G^*$, and equality holds if and only if $G' = G^*$.

Unlike traditional algorithms, a foundation model drives a paradigm shift in which it does not assume a single, known mechanism M^*. Instead, it aims to map an observational dataset $\mathbf{X} \in \mathbb{R}^{N\times D}$ directly to the causal graph G across a vast hypothesis space of diverse, unknown mechanisms $\mathcal{M}$. This shift means the mapping inevitably amounts to a Bayesian model selection problem, where the optimal likelihood requires marginalizing over the entire mechanism space $\mathcal{M}$:

$$\mathbb{E}_{\mathbf{X}}[\log P(\mathbf{X}, G)] = \mathbb{E}_{\mathbf{X}}\left[\log \int_{\mathcal{M}} P(\mathbf{X}|M, G)P(G|M)P(M)dM\right]. \tag{4}$$

One may still identify the causal structure using the exact marginal likelihood as long as the space of causal prior mechanisms $\mathcal{M}$ satisfies a strict identifiability property. We say the space $\mathcal{M}$ is strictly identifiable if no alternative structure and mechanism can perfectly mimic the true generating distribution. Formally, $P(\mathbf{X}|M^*, G^*) \neq P(\mathbf{X}|M_{G'}, G')$ holds for all $G' \neq G^*$, where $M_{G'} = \arg\max_{M\in\mathcal{M}} \mathbb{E}_{\mathbf{X}}[\log P(\mathbf{X}|M, G')]$ represents the optimal mechanism choice under G'. If the space possesses this intrinsic mutual exclusivity, the exact marginal likelihood uniquely identifies the true causal graph.

Theorem 3 (Identifiability by Marginal Likelihood)**.** *Let a dataset $\mathbf{X}^{(N)}$ of N i.i.d. samples be generated by the true graph G^* and true mechanism $M^* \in \mathcal{M}$. If the space of causal prior mechanisms $\mathcal{M}$ satisfies the strict space identifiability property, then the exact marginal likelihood uniquely identifies the true causal graph almost surely:*

$$\lim_{N\to\infty} \frac{1}{N} \log P(\mathbf{X}^{(N)}|G^*) > \lim_{N\to\infty} \frac{1}{N} \log P(\mathbf{X}^{(N)}|G'), \quad \textit{a.s.}, \quad \forall G' \neq G^*. \tag{5}$$

While Theorem 3 guarantees that the true graph can be identified in principle, directly computing the exact integral in Eq. (4) is intractable.

To bridge this gap, we formulate the learning process of CDFM through variational inference. Under uninformative uniform priors of $P(G)$ and $P(M)$, maximizing the marginal likelihood is mathematically equivalent to maximizing the joint likelihood. We introduce a parameterized variational distribution $Q(M \mid \mathbf{X})$, which is designed to internalize the mechanism and structural information within the architecture. This yields a tractable Evidence Lower Bound (ELBO) over the joint likelihood (see Appendix A.1 for derivation):

$$\log P(G, \mathbf{X}) \geq \mathbb{E}_{Q(M|\mathbf{X})}\big[\log P(\mathbf{X}|G, M) + \log P(G|M)\big] - D_{\mathrm{KL}}\big(Q(M|\mathbf{X}) \parallel P(M)\big). \tag{6}$$

Equation (6) is an exact Jensen lower bound for any admissible Q, but two qualifications matter for the implemented model. First, even under a uniform prior, the KL term depends on the entropy and support of $Q(M \mid \mathbf{X})$ and is therefore not a constant. Second, the exact posterior is generally $P(M \mid \mathbf{X}, G)$, whereas CDFM uses an encoder that does not condition on a candidate graph. We therefore use the ELBO as a principled design decomposition rather than claiming that the network optimizes the exact Bayesian objective. In this interpretation, $Q(M \mid \mathbf{X})$ motivates mechanism-sensitive representations, $\log P(\mathbf{X} \mid G, M)$ motivates structure-aware reconstruction, and $\log P(G \mid M)$ motivates directed graph decoding.

By this, with sufficient capacity and massive pretraining, the variational gap (D_{KL} between the true and approximate posterior) vanishes at the global optimum. Consequently, optimizing this ELBO implicitly maximizes the exact marginal likelihood. This guarantees that CDFM inherently respects the mathematical boundaries of causal identifiability:

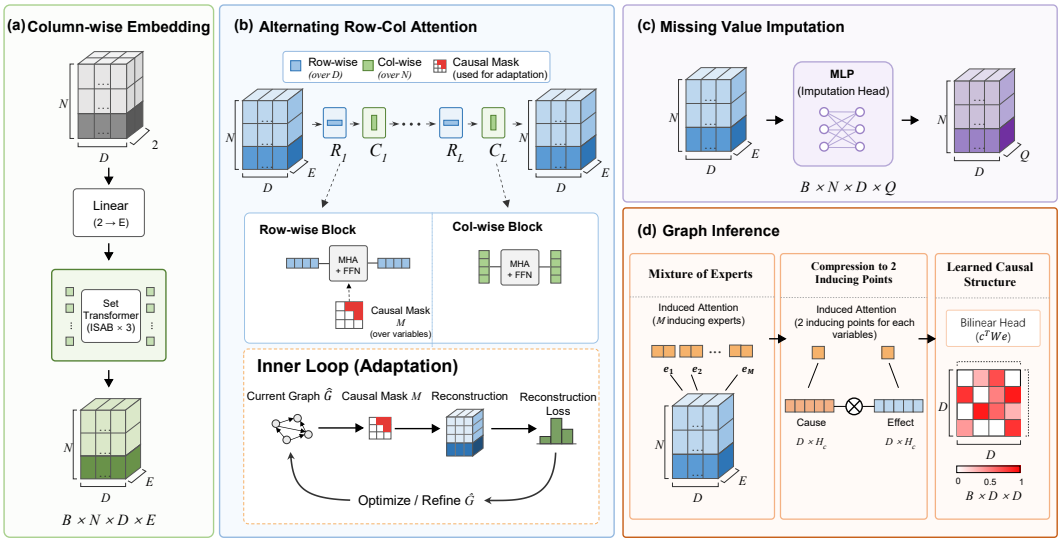


Figure 1: The overall architecture of CDFM.

Corollary 1 (Identifiability of CDFM)**.** *Suppose the foundation model has sufficient capacity to reach the optimum variational ELBO. Let* $\mathbf{X}^{(N)}$ *be any test dataset with unknown causal graph* G^**. In the inference phase, the model predicts the graph by maximizing the expected log-likelihood of the structural decoder:*

$$\hat{G} = \arg\max_{G \in \mathcal{G}} \mathbb{E}_{Q(M|\mathbf{X}^{(N)})} \big[\log P(G \mid M) \big]. \tag{7}$$

Under the strict identifiability of the mechanism space $\mathcal{M}$ *and a uniform prior* $P(G)$*, this structural inference uniquely recovers the true causal graph almost surely, i.e.,* $\hat{G} = G^*$ *as* $N \to \infty$*.*

5 Causal Discovery Foundation Model

CDFM is designed to instantiate the intractable marginalization over the mechanism space $\mathcal{M}$ through a deep neural architecture. It takes an observation matrix $\mathbf{X} \in \mathbb{R}^{N \times D}$ (consisting of N observations over D variables, with potential missing entries) and infers the underlying causal graph $G \in \{0,1\}^{D \times D}$. As formalized in Section 3, the graph would be an adjacency matrix augmented with bidirected edges when latent confounding is present.

The architectural design of CDFM explicitly mirrors the three components of the variational ELBO: mechanism inference (Section 5.1), data reconstruction (Section 5.2), and graph inference (Section 5.3). An overview of the CDFM architecture is shown in Figure 1.

5.1 Mechanism Inference

To construct the variational distribution $Q(M|\mathbf{X})$, CDFM first extracts distributional statistics from each feature through column-wise embedding. To explicitly handle missing values, each scalar x_{ij} is paired with a binary missingness indicator $r_{ij} \in \{0,1\}$. After column-wise standardization, missing values are replaced with zero and marked by $r_{ij} = 1$. The resulting pair is projected into an E-dimensional space as $u_{ij} = \text{Linear}_{\text{in}}([x_{ij}, r_{ij}]) \in \mathbb{R}^E$, allowing the model to distinguish observed and missing entries.

Inspired by (Qu et al.), we then capture high-order statistics of each marginal distribution using a shared Set Transformer (Lee et al., 2019), which treats each column as an *unordered set*. For column $c_j = \{x_{1j}, \ldots, x_{Nj}\}$, its embeddings are processed by $L_{\text{col}} = 3$ induced self-attention blocks (ISABs). Each block performs

$$H = \text{MAB}_1(V_I, U, U), \qquad V = \text{MAB}_2(U, H, H), \tag{8}$$

where $V_I \in \mathbb{R}^{K\times E}$ denotes K learnable inducing vectors and U contains the embedded samples of the column. Two linear heads then produce $W_j = \mathrm{Linear}_W(V)$ and $B_j = \mathrm{Linear}_B(V)$, yielding the final cell embedding

$$e_{ij} = W_{ij} \odot x_{ij} + B_{ij} \in \mathbb{R}^E. \tag{9}$$

Because W_{ij} and B_{ij} are inferred from the entire column, this affine transformation encodes each value relative to its empirical marginal distribution, providing distributional signatures of the underlying mechanisms.

CDFM further captures dependencies across variables through alternating row- and column-wise attention, progressively integrating information along both axes of the data matrix, as illustrated in Figure 1. Each attention layer is followed by a GELU-based feedforward network, producing a joint latent representation used to parameterize $Q(M|\mathbf{X})$.

5.2 Structure-aware data reconstruction

To instantiate the reconstruction term $\log P(\mathbf{X} \mid G, M)$, CDFM employs a self-supervised imputation objective whose information flow is constrained by the current graph hypothesis. Because CDFM alternates sample-wise and variable-wise attention, the structural constraint is applied only to the variable-wise attention blocks (i.e., the even-indexed blocks under our alternating ordering). Specifically, each such block incorporates a soft structural mask $\mathbf{S} \in \mathbb{R}^{D\times D}$ derived from the continuous edge weights W. When variable i attends to variable j ($i \neq j$), the corresponding attention logit is biased according to the strength of the hypothesized edge $j \to i$:

$$S_{ij} = -\beta \cdot \sigma\left(-\frac{W_{ji} - \tau}{T}\right) \tag{10}$$

where τ denotes an edge-strength threshold, T controls the sharpness of the relaxation, and β determines the maximum masking penalty. Under this formulation, weak candidate edges ($W_{ji} \ll \tau$) induce a penalty close to $-\beta$, suppressing information flow from variable j to i, whereas strong edges receive little penalty. The resulting attention pattern therefore encourages each variable to aggregate information primarily from its direct causes under the current graph hypothesis, while remaining fully differentiable with respect to W.

Following the final alternating block, the model gathers the contextualized embeddings $z^{(2L)}$ at positions originally marked as missing ($r_{ij} = 1$) and processes them through a residual block to predict quantile levels $\hat{y}_{ij} \in \mathbb{R}^Q$. By evaluating the discrepancy between these predictions and the masked true values, the imputation task provides a continuous gradient signal for structure learning.

The differentiability of $\mathbf{S}$ further enables test-time adaptation without access to ground-truth graph labels. During this inner-loop optimization, CDFM updates the continuous graph hypothesis W by minimizing the reconstruction objective on the target dataset. In this way, the inferred structure is refined toward graph configurations whose induced information flow better explains the observed dependencies in previously unseen data.

5.3 Graph Inference

The final component realizes the graph inference term $\log P(G \mid M)$, decoding the causal structure from the neurally encoded mechanism representations.

To make graph prediction invariant to the number of samples N, CDFM first aggregates sample-level representations into variable-level structural summaries. Specifically, an induced self-attention module uses $2K$ learnable inducing vectors, organized as K pairs of cause–effect experts, to extract complementary structural evidence from the encoded mechanisms.

The resulting expert representations are then compressed by an attention-based aggregator into two representations for each variable i: a cause embedding $\mathbf{e}_i^{\mathrm{c}}$ and an effect embedding $\mathbf{e}_i^{\mathrm{e}}$. This aggregation acts as a soft-routing mechanism that adaptively combines expert evidence according to the inferred data-generating mechanism.

These representations are evaluated to obtain the respective cause and effect vectors, yielding $\mathbf{u}_i = \mathbf{W}_c\mathbf{e}_i^{\mathrm{c}}$ and $\mathbf{v}_j = \mathbf{W}_e\mathbf{e}_j^{\mathrm{e}}$, where $\mathbf{u}_i, \mathbf{v}_j \in \mathbb{R}^p$. The continuous edge logit is then computed using a

temperature-scaled dot product augmented by a global sparsity bias $A_{ij} = \exp(\tau) \cdot \frac{\mathbf{u}_i^\top \mathbf{v}_j}{\sqrt{p}} + b_0$, where p is the projection dimension, τ is a learned log-temperature parameter governing the sharpness of the edge probabilities, and b_0 is a learned global bias representing the baseline sparsity prior of the causal graphs. The diagonal is strictly masked to $-\infty$ to explicitly preclude causal self-loops. The inherent asymmetry of this bilinear evaluation ($\mathbf{u}_i^\top \mathbf{v}_j \neq \mathbf{u}_j^\top \mathbf{v}_i$) naturally accommodates both directed causal relationships and bidirected edges.

To obtain a binary graph, CDFM further employs a lightweight graph-adaptive threshold calibrator trained once on an independent synthetic calibration set. Rather than using a fixed cutoff, the calibrator predicts a graph-specific threshold from ten summary statistics of the predicted edge-probability distribution, including moments, quantiles, graph size, gap statistics of the sorted probabilities, and a Kneedle-inspired knee feature (Satopaa et al., 2011). A logit-linear model is fitted to oracle thresholds on held-out synthetic graphs, and its parameters remain fixed at evaluation time.

5.4 Training Strategy

Causal prior construction and pretraining. CDFM is pretrained on synthetically generated $(\mathbf{X}, G)$ pairs sampled from randomized structural causal models (SCMs) to acquire a broad causal prior before downstream adaptation. The generator covers diverse graph topologies, including Erdős–R'enyi (ER), scale-free (SF), Watts–Strogatz, stochastic-block, geometric, chain, tree, layered, and collider-rich structures, with varying graph sizes and edge densities. Causal mechanisms are assigned at the node level and span heterogeneous functional families, including linear structural equations with diverse noise settings (Yin et al., 2024), random Fourier feature (RFF) mechanisms, post-nonlinear models (Zhang & Hyvärinen, 2009), causal additive models (CAM) (Bühlmann et al., 2014), neural and tree-based mechanisms, as well as discrete additive-noise (Peters et al., 2010) and ordinal/mixed-type relationships (Ni & Mallick, 2022). Exogenous noise is sampled from multiple distributions, such as Gaussian, Laplace, Cauchy, Student's-t, log-normal, and Gaussian mixtures, with varying signal-to-noise ratios. The generator further incorporates measurement perturbations, discretization, latent confounding, and random missingness to expose the model to realistic observational imperfections. Together, these variations encourage CDFM to learn transferable causal representations rather than overfit to a particular graph family, functional form, or noise assumption. Full details of the graph generators, mechanism taxonomy, noise configurations, and simulation procedures are provided in Appendix B.1.

Training Objective. CDFM is trained end-to-end with binary cross-entropy (BCE) as the primary objective for causal discovery, complemented by auxiliary regularization and reconstruction terms:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{BCE}} + \alpha_{\text{margin}}\mathcal{L}_{\text{margin}} + \gamma_{\text{F}_1}\mathcal{L}_{\text{soft}-\text{F}_1} + \beta_{\text{acyc}}\mathcal{L}_{\text{acyc}} + \alpha_{\text{impute}}\mathcal{L}_{\text{quantile}}. \tag{11}$$

Here, $\mathcal{L}_{\text{BCE}}$ provides the main structural supervision, while the margin and soft-F_1 terms improve prediction confidence and alleviate class imbalance in sparse graphs. The acyclicity penalty promotes valid directed structures and is omitted for samples with latent confounding, whereas the quantile loss supervises missing-value reconstruction. Details are provided in Appendix C.

6 Experiments

Experimental setting. The experiments cover synthetic mechanisms, real-world data, identifiability diagnostics, and imputation. We compare CDFM with causal discovery methods spanning amortized inference (AVICI Lorch et al. (2022), TabCausal Li et al. (2026), Arrow Thompson et al. (2026), and FoundCause), constraint- and score-based discovery (PC Spirtes et al. (2000), GES Chickering (2002)), functional causal models (DirectLiNGAM Shimizu et al. (2011)), and continuous optimization (NOTEARS Zheng et al. (2018)), using their standard implementations and recommended hyperparameters.

For multivariate causal discovery, we evaluate edge ranking using AUROC and thresholded graph recovery using precision, recall, F_1 score. For methods that do not produce continuous edge scores suitable for ranking, we report balanced accuracy in place of AUROC. Results are averaged over independently generated datasets. For bivariate causal direction inference, we report weighted accuracy following the official evaluation protocol.

Table 1: Synthetic benchmark AUROC grouped by sample size. Results averaged over 15 mechanism families and nine graph sizes ($D \in \{10, 15, 20, 25, 30, 40, 50, 70, 100\}$).

Method	500 samples	1000 samples	2000 samples	3000 samples	4000 samples
CDFM	**0.864 ± 0.109**	**0.878 ± 0.104**	**0.885 ± 0.101**	**0.887 ± 0.100**	**0.888 ± 0.101**
TabCausal	0.767 ± 0.166	0.774 ± 0.166	0.782 ± 0.167	0.786 ± 0.164	0.786 ± 0.166
Arrow	0.676 ± 0.153	0.669 ± 0.149	0.680 ± 0.154	0.682 ± 0.154	0.685 ± 0.154
AVICI	0.756 ± 0.163	0.760 ± 0.165	0.764 ± 0.164	0.761 ± 0.161	0.760 ± 0.163
DirectLiNGAM	0.676 ± 0.150	0.689 ± 0.154	0.713 ± 0.158	0.724 ± 0.162	0.731 ± 0.164
GES	0.627 ± 0.104	0.631 ± 0.105	0.611 ± 0.101	0.645 ± 0.104	0.640 ± 0.105
PC	0.700 ± 0.122	0.715 ± 0.127	0.728 ± 0.129	0.737 ± 0.132	0.740 ± 0.133
NOTEARS	0.563 ± 0.127	0.581 ± 0.124	0.574 ± 0.151	0.583 ± 0.157	0.585 ± 0.167
FoundCause	0.827 ± 0.143	0.840 ± 0.139	0.842 ± 0.141	0.843 ± 0.138	0.840 ± 0.139

Table 2: Real-world benchmarks. Results of GES and PC on Tübingen are omitted as they are not applicable to pairwise settings.

	Chamber A1		**Chamber A2**		**Tübingen**	**RCB**	
Method	$F_1 \uparrow$	AUROC↑	$F_1 \uparrow$	AUROC↑	Acc.↑	$F_1 \uparrow$	AUROC↑
CDFM	**0.735**	0.935	**0.727**	**0.965**	**0.695**	**0.673**	**0.855**
TabCausal	0.623	**0.939**	0.635	0.939	0.632	0.257	0.778
FoundCause	0.629	0.842	0.708	0.896	0.663	0.295	0.815
Arrow	0.354	0.644	0.413	0.797	0.516	0.214	0.814
AVICI	0.189	0.357	0.198	0.503	0.516	0.121	0.755
DirectLiNGAM	0.603	0.758	0.657	0.811	0.474	0.140	0.791
GES	0.406	0.649	0.490	0.809	–	0.290	0.670
PC	0.278	0.594	0.292	0.598	–	0.382	0.712
NOTEARS	0.254	0.586	0.217	0.451	0.295	0.100	0.477

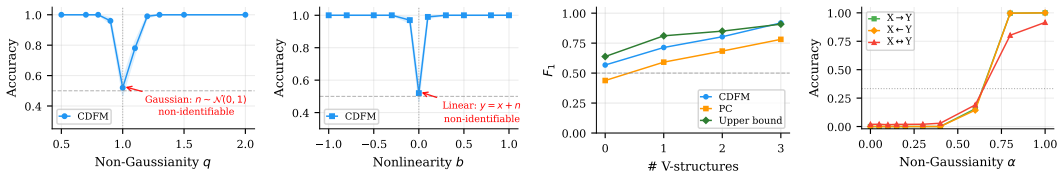
Synthetic benchmark. We begin with the synthetic experiments, where the data-generating process is known and can be varied systematically. The benchmark spans 15 mechanism families (Appendix B), nine graph sizes from $D = 10$ to 100, five sample sizes from $N = 500$ to 4000, and multiple held-out seeds. This provides the primary test of whether one frozen model can generalize across mechanisms. As shown in Table 1, CDFM achieves the highest AUROC at all sample sizes, demonstrating the effectiveness of the proposed method compared with the baseline methods.

Real-world benchmarks. Having established performance on synthetic benchmarks, we next test our methods on three complementary settings: the multivariate physical systems of Causal Chamber Gamella et al. (2025), the heterogeneous bivariate cause–effect pairs from Tübingen Mooij et al. (2016), and RCB, our curated benchmark of real tabular datasets with independently screened trusted causal relations. The results are summarized in Table 2.

On Causal Chamber, CDFM is evaluated on the observational A1 task and the 20-environment A2 task, where predictions are obtained separately for each environment and then averaged. CDFM achieves the best F_1 scores of 0.735 and 0.727 on A1 and A2, respectively.

On the 95 strictly bivariate Tübingen pairs, CDFM reaches a weighted direction accuracy of 0.695, compared with 0.663 for FoundCause and 0.632 for TabCausal.

Finally, we introduce a Real Causal Bench (RCB), which extends the evaluation to general real-world tabular data where complete causal ground truth is unavailable. We retain only relations supported by both domain knowledge and statistical evidence, resulting in 78 directed edges, 7 confounded pairs, and 187 non-edge pairs across seven datasets. CDFM achieves an AUROC of 0.855 and an F_1 score of 0.673. Details of the benchmark construction are provided in Appendix F.



(a) Non-Gaussianity ($b = 0$). (b) Nonlinearity ($q = 1$). (c) V-structures ($d = 5$). (d) Latent confounding.

Figure 2: Empirical validation of identifiability. (a) Direction accuracy vs. the non-Gaussianity parameter q for linear SCMs ($b = 0$). (b) Direction accuracy vs. the nonlinearity parameter b under Gaussian noise ($q = 1$). (c) F_1 score vs. the number of V-structures, against PC and the theoretical upper bound. (d) Three-class accuracy under latent confounding vs. non-Gaussianity.

Table 3: Imputation MAE at masked positions.

	Continuous				Discrete			
Method	**0.1**	**0.2**	**0.3**	**0.5**	**0.1**	**0.2**	**0.3**	**0.5**
CDFM	**0.416±0.158**	**0.433±0.158**	**0.451±0.157**	**0.500±0.152**	**0.713±0.098**	**0.718±0.097**	**0.723±0.097**	0.735±0.098
MissForest	0.420±0.171	0.445±0.171	0.472±0.172	0.546±0.171	0.732±0.092	0.750±0.091	0.766±0.091	0.804±0.094
MICE	0.433±0.167	0.475±0.168	0.514±0.170	0.580±0.156	0.751±0.089	0.754±0.088	0.779±0.093	0.826±0.097
SoftImpute	0.476±0.169	0.492±0.165	0.508±0.160	0.549±0.147	0.806±0.160	0.806±0.153	0.801±0.140	0.787±0.104
KNN (k=5)	0.529±0.182	0.554±0.178	0.582±0.172	0.648±0.152	0.776±0.097	0.784±0.093	0.791±0.091	0.802±0.091
MatrixFactorization	0.648±0.108	0.663±0.097	0.676±0.088	0.701±0.074	0.777±0.089	0.778±0.088	0.778±0.088	0.780±0.088
Median	0.729±0.072	0.729±0.069	0.729±0.068	0.730±0.067	0.720±0.098	0.720±0.096	0.720±0.096	**0.721±0.096**
Mean	0.736±0.070	0.737±0.068	0.737±0.066	0.737±0.065	0.777±0.088	0.777±0.087	0.777±0.087	0.777±0.087
IterativeSVD	0.766±0.211	0.765±0.191	0.764±0.177	0.762±0.147	0.940±0.326	0.936±0.322	0.923±0.289	0.876±0.222

Identifiability diagnostics. In this experiment, we aim to test whether a foundation model still respects the theoretical limits of identifiability. We therefore run three controlled diagnostic experiments with CDFM fine-tuned on bivariate additive-noise models, V-structure identification, and latent-confounding tasks. Specifically, in the linear-Gaussian regime, direction accuracy remains near chance, but increases as non-Gaussianity or nonlinearity introduces identifiable asymmetry (Fig. 2a and Fig. 2b). Orientation performance likewise improves as more V-structures become available and approaches the perfect-d-separation upper bound (Fig. 2c). Finally, distinguishing direct effects from pure confounding becomes substantially harder near Gaussianity (Fig. 2d). Taken together, these results confirm that CDFM can exploit identifiable causal signals when they are available while remaining constrained by the theoretical limits of observational identifiability. (See Appendix E for the construction details of this diagnostic.)

Missing-value imputation. Finally, the quantile-reconstruction head is trained to support representation learning rather than as a dedicated imputer, so we ask whether it nevertheless learns useful conditional structure. We evaluate one-pass median-quantile imputation on continuous RFF and discrete CPT data under MCAR missingness, spanning $D \in [5, 100]$, $N = 500$, and four missing rates. We compare with standard statistical, KNN Troyanskaya et al. (2001); Pedregosa et al. (2011), MICE van Buuren & Groothuis-Oudshoorn (2011), MissForest Stekhoven & Bühlmann (2012), SoftImpute, IterativeSVD, and MatrixFactorization Rubinsteyn & Feldman. Results are presented in Table 3, where CDFM achieves the best MAE on most data, demonstrating that the auxiliary quantile head captures useful conditional structure and can serve as an effective imputer.

7 Conclusions

In this work, we introduce CDFM, a causal foundation model that reframes causal discovery as learning across families of causal mechanisms rather than relying on a single assumed model class. Experiments show that CDFM generalizes across heterogeneous synthetic settings and transfers effectively to real-world benchmarks, while controlled identifiability studies exhibit behavior consistent with classical causal limits. Together, these results suggest a shift from increasingly specialized causal discovery procedures toward models that learn reusable causal representations across mechanisms, with CDFM taking a step toward more general-purpose causal discovery.

REFERENCES

Yashas Annadani, Stefan Bauer, Adam Foster, and Panagiotis Tigas. Amortized active causal induction with deep reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS 2024)*, 2024.

Vahid Balazadeh Meresht, Hamidreza Kamkari, Valentin Thomas, Junwei Ma, Bingru Li, Jesse Cresswell, and Rahul Krishnan. Causalpfn: Amortized causal effect estimation via in-context learning. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, pp. 154945–154984. Curran Associates, Inc., 2025.

Peter Bühlmann, Jonas Peters, and Jan Ernest. CAM: Causal additive models, high-dimensional order search and penalized regression. *The Annals of Statistics*, 42(6):2526–2556, 2014.

Ruichu Cai, Zhenjie Zhang, Zhifeng Hao, and Marianne Winslett. Understanding social causalities behind human action sequences. *IEEE Transactions on Neural Networks and Learning Systems*, 28(8):1801–1813, 2017. doi: 10.1109/TNNLS.2016.2556724.

Ruichu Cai, Jie Qiao, Kun Zhang, Zhenjie Zhang, and Zhifeng Hao. Causal discovery from discrete data using hidden compact representation. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

David Maxwell Chickering. Optimal structure identification with greedy search. *Journal of Machine Learning Research*, 3:507–554, 2002.

Paolo Cons. Small dataset about used Fiat 500 sold in italy. Kaggle, 2019. https://www.kaggle.com/datasets/paolocons/small-dataset-about-used-fiat-500-sold-in-italy.

FICO. FICO HELOC dataset — explainable machine learning challenge. https://community.fico.com/s/explainable-machine-learning-challenge, 2018. Accessed: 2026-09-26.

Abbas El Gamal and Young-Han Kim. *Network Information Theory*. Cambridge University Press, Cambridge, U.K., 2011. ISBN 978-1-107-00873-1. doi: 10.1017/CBO9781139030687.

Juan L Gamella, Jonas Peters, and Peter Bühlmann. Causal chambers as a real-world physical testbed for ai methodology. *Nature Machine Intelligence*, 7(1):107–118, 2025.

Eric Ghysels, Jonathan B Hill, and Kaiji Motegi. Testing for granger causality with mixed frequency data. *Journal of Econometrics*, 192(1):207–230, 2016.

Hass Avocado Board. Avocado Prices. Kaggle, 2018. Compiled from weekly retail scan data. https://www.kaggle.com/datasets/neuromusic/avocado-prices.

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. 637(8045):319–326. ISSN 0028-0836, 1476-4687. doi: 10.1038/s41586-024-08328-6.

Patrik O. Hoyer, Dominik Janzing, Joris M. Mooij, Jonas Peters, and Bernhard Schölkopf. Nonlinear causal discovery with additive noise models. In *Advances in Neural Information Processing Systems*, volume 21, pp. 689–696, 2009.

Alexander Immer, Christoph Schultheiss, Julia E Vogt, Bernhard Schölkopf, Peter Bühlmann, and Alexander Marx. On the identifiability and estimation of causal location-scale noise models. In *International Conference on Machine Learning*, pp. 14316–14332. PMLR, 2023.

Andras Janosi, William Steinbrunn, Matthias Pfisterer, and Robert Detrano. Heart Disease. UCI Machine Learning Repository, 1989.

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

Juho Lee, Yoonho Lee, Jungtaek Kim, Adam R. Kosiorek, Seungjin Choi, and Yee Whye Teh. Set transformer: A framework for attention-based permutation-invariant neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3744–3753. PMLR, 2019.

James Leiner, Brian Manzo, Aaditya Ramdas, and Wesley Tansey. Scalable causal structure learning via amortized conditional independence testing. In *Proceedings of the Fourth Conference on Causal Learning and Reasoning*, volume 275 of *Proceedings of Machine Learning Research*, pp. 174–200. PMLR, 07–09 May 2025.

Zi-Rong Li, Si-Yang Liu, Tian-Zuo Wang, and Han-Jia Ye. Tabcausal: Pretraining across causal environments for tabular causal discovery. *arXiv preprint arXiv:2605.31156*, 2026.

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

Sindy Löwe, David Madras, Richard Zemel, and Max Welling. Amortized causal discovery: Learning to infer causal graphs from time-series data. In Bernhard Schölkopf, Caroline Uhler, and Kun Zhang (eds.), *Proceedings of the First Conference on Causal Learning and Reasoning*, volume 177 of *Proceedings of Machine Learning Research*, pp. 509–525. PMLR, 11–13 Apr 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.

Joris M Mooij, Jonas Peters, Dominik Janzing, Jakob Zscheischler, and Bernhard Schölkopf. Distinguishing cause from effect using observational data: methods and benchmarks. *Journal of Machine Learning Research*, 17(32):1–102, 2016.

Yang Ni and Bani Mallick. Ordinal causal discovery. In *Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence*, volume 180 of *Proceedings of Machine Learning Research*, pp. 1530–1540. PMLR, 2022.

R. Kelley Pace and Ronald Barry. Sparse spatial autoregressions. *Statistics & Probability Letters*, 33(3):291–297, 1997. doi: 10.1016/S0167-7152(96)00140-X.

Judea Pearl. An introduction to causal inference. *The international journal of biostatistics*, 6(2):7, 2010.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Bo Peng, Sirui Chen, Jiaguo Tian, Yu Qiao, and Chaochao Lu. Causcale: Neural causal discovery at scale. *arXiv preprint arXiv:2602.08629*, 2026.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. Identifying cause and effect on discrete data using additive noise models. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, pp. 597–604. PMLR, 2010.

Jonas Peters, Dominik Janzing, and Bernhard Scholkopf. *Elements of causal inference: foundations and learning algorithms*. MIT press, 2017.

Jie Qiao, Ruichu Cai, Siyu Wu, Yu Xiang, Keli Zhang, and Zhifeng Hao. Structural hawkes processes for learning causal structure from discrete-time event sequences. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, 2023. doi: 10.24963/ijcai.2023/633.

Jie Qiao, Yu Xiang, Zhengming Chen, Ruichu Cai, and Zhifeng Hao. Causal discovery from poisson branching structural causal model using high-order cumulant with path analysis. In *Proceedings of the Thirty-Sixth AAAI Conference on Artificial Intelligence (AAAI 2024)*, 2024.

Jie Qiao, Zihuai Zeng, Ruichu Cai, Zhengming Chen, and Zhifeng Hao. On the Identifiability of Poisson Branching Structural Causal Model Under Latent Confounding. In *Proceedings of the Forty-Third International Conference on Machine Learning International Conference on Machine Learning (ICML-2026)*, 2026.

Jingang Qu, David Holzmüller, Gaël Varoquaux, and Marine Le Morvan. TabICL: A tabular foundation model for in-context learning on large data. In *International Conference on Machine Learning*, pp. 50817–50847. PMLR.

Jake Robertson, Arik Reuter, Siyuan Guo, Noah Hollmann, Frank Hutter, and Bernhard Schölkopf. Do-pfn: In-context learning for causal effect estimation. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, pp. 174811–174848. Curran Associates, Inc., 2025.

Alex Rubinsteyn and Sergey Feldman. fancyimpute: A variety of matrix completion and imputation algorithms implemented in python. URL `https://github.com/iskandr/fancyimpute`. Version 0.7.0.

Jakob Runge, Xavier-Andoni Tibau, Matthias Bruhns, Jordi Muñoz-Marí, and Gustau Camps-Valls. The causality for climate competition. In *NeurIPS 2019 Competition and Demonstration Track*, pp. 110–120. Pmlr, 2020.

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.

Ville Satopaa, Jeannie Albrecht, David Irwin, and Barath Raghavan. Finding a Kneedle in a haystack: Detecting knee points in system behavior. In *2011 31st International Conference on Distributed Computing Systems Workshops*, pp. 166–171. IEEE, 2011. doi: 10.1109/ICDCSW.2011.20.

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

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

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

Daniel J. Stekhoven and Peter Bühlmann. Missforest—non-parametric missing value imputation for mixed-type data. *Bioinformatics*, 28(1):112–118, 2012. doi: 10.1093/bioinformatics/btr597.

Omar Swelam, Lennart Purucker, Jake Robertson, Hanne Raum, Joschka Boedecker, and Frank Hutter. Does tabPFN understand causal structures? In *EurIPS 2025 Workshop: AI for Tabular Data*, 2025.

Ryan Thompson, He Zhao, Daniel M. Steinberg, and Edwin V. Bonilla. Arrow: A foundation model for causal discovery. *arXiv preprint arXiv:2605.07204*, 2026.

Olga Troyanskaya, Michael Cantor, Gavin Sherlock, Pat Brown, Trevor Hastie, Robert Tibshirani, David Botstein, and Russ B. Altman. Missing value estimation methods for DNA microarrays. *Bioinformatics*, 17(6):520–525, 2001. doi: 10.1093/bioinformatics/17.6.520.

Luca Urbinati et al. Enhanced machine-learning flow for microwave-sensing systems to detect contaminants in food. In *2023 IEEE Conference on Microwave Sensing*, 2023. doi: 10.1109/10291198. Dataset available via OpenML.

U.S. Bureau of Labor Statistics. National Longitudinal Survey of Youth 1979. `https://www.bls.gov/nls/nlsy79.htm`, 1979. Accessed: 2026-09-26.

Stef van Buuren and Karin Groothuis-Oudshoorn. mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, 45(3):1–67, 2011. doi: 10.18637/jss.v045.i03.

Menghua Wu, Bo Qiang, Wenxian Shi, Yuxuan Song, et al. Sample, estimate, aggregate: A recipe for causal discovery foundation models. *Transactions on Machine Learning Research*, 2025. Published March 2025.

Yu Xiang, Jie Qiao, Zefeng Liang, Zihuai Zeng, Ruichu Cai, and Zhifeng Hao. On the identifiability of poisson branching structural causal model using probability generating function. In *Advances in Neural Information Processing Systems*, volume 37, pp. 11664–11699, 2024.

Ruicong Yao, Tim Verdonck, and Jakob Raymaekers. Causal discovery in mixed additive noise models. In *Proceedings of The 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pp. 3088–3096. PMLR, 2025.

Naiyu Yin, Tian Gao, Yue Yu, and Qiang Ji. Effective causal discovery under identifiable heteroscedastic noise model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 16486–16494, 2024. doi: 10.1609/aaai.v38i15.29586.

Naiyu Yin, Tian Gao, and Yue Yu. Learning causal graphs at scale: A foundation model approach. 2025.

Kun Zhang and Aapo Hyvärinen. On the identifiability of the post-nonlinear causal model. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, pp. 647–655. AUAI Press, 2009.

Kun Zhang, Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. Kernel-based conditional independence test and application in causal discovery. In *Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence*, UAI'11, pp. 804–813, Arlington, Virginia, USA, 2011. AUAI Press.

Xingxuan Zhang, Gang Ren, Han Yu, Hao Yuan, Hui Wang, Jiansheng Li, Jiayun Wu, Lang Mo, Li Mao, Mingchao Hao, et al. Limix: Unleashing structured-data modeling capability for generalist intelligence. *arXiv preprint arXiv:2509.03505*, 2025.

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 Proofs and Derivation

A.1 Derivation of the Variational ELBO

To approximate the intractable marginal likelihood where the mechanism M is integrated out, we introduce a variational distribution $Q(M|\mathbf{X})$ over the mechanism space $\mathcal{M}$. The joint log-probability $\log P(G, \mathbf{X})$ can be expanded and bounded as follows:

$$\log P(G, \mathbf{X}) = \log \int_{\mathcal{M}} P(G, M, \mathbf{X}) dM \tag{12}$$

$$= \log \int_{\mathcal{M}} P(G, \mathbf{X}|M) P(M) dM \tag{13}$$

$$= \log \int_{\mathcal{M}} P(G, \mathbf{X}|M) \frac{P(M)}{Q(M|\mathbf{X})} Q(M|\mathbf{X}) dM \tag{14}$$

$$\geq \mathbb{E}_{Q(M|\mathbf{X})} \left[\log \left(P(G, \mathbf{X}|M) \frac{P(M)}{Q(M|\mathbf{X})} \right) \right] \quad \text{(by Jensen's Inequality)} \tag{15}$$

$$= \mathbb{E}_{Q(M|\mathbf{X})} \big[\log P(G, \mathbf{X}|M) \big] - \mathbb{E}_{Q(M|\mathbf{X})} \left[\log \frac{Q(M|\mathbf{X})}{P(M)} \right] \tag{16}$$

$$= \mathbb{E}_{Q(M|\mathbf{X})} \big[\log P(\mathbf{X}|G, M) + \log P(G|M) \big] - D_{\mathrm{KL}} \big(Q(M|\mathbf{X}) \parallel P(M) \big). \tag{17}$$

This completes the derivation of Eq. (6) in the main text.

A.2 Proof of Theorem 1

Proof. To provide an intuitive construction, we first assume that X and Y are continuous random variables. We will demonstrate that a valid Structural Causal Model can always be constructed for the causal direction $X \to Y$.

Let $F_{Y|X}(y|x)$ denote the conditional cumulative distribution function of Y given X. According to the inverse transform sampling theorem, there exists an auxiliary noise variable $N_Y \sim \mathcal{U}[0, 1]$ that is independent of X, such that $Y = F^{-1}_{Y|X}(N_Y \mid X)$. Therefore, by defining the structural function as $f_Y(X, N_Y) := F^{-1}_{Y|X}(N_Y \mid X)$, we obtain $Y = f_Y(X, N_Y)$ with $N_Y \perp\!\!\!\perp X$. Hence, a valid SCM corresponding to the causal direction $X \to Y$ exists.

Applying the same argument to the reverse direction, there exists an independent noise variable $N_X \sim \mathcal{U}[0, 1]$ such that $X = F^{-1}_{X|Y}(N_X \mid Y)$. Equivalently, defining $f_X(Y, N_X) := F^{-1}_{X|Y}(N_X \mid Y)$, we obtain $X = f_X(Y, N_X)$ where $N_X \perp\!\!\!\perp Y$. Therefore, for any continuous joint distribution $P_{X,Y}$, valid SCMs can be constructed for both causal directions $X \to Y$ and $Y \to X$.

For more general cases, including discrete or mixed variables, the Functional Representation Lemma Gamal & Kim (2011) guarantees the existence of such independent uniform noise variables and measurable structural functions, yielding the exact same result. Consequently, without additional assumptions on the causal mechanisms, the causal direction is inherently unidentifiable from observational data alone.

□

A.3 Proof of Theorem 2

Proof. By definition, the expected log-likelihood of a graph G under the true data-generating distribution is

$$\mathcal{L}(G) = \mathbb{E}_{\mathbf{X} \sim P(\mathbf{X}|G^*, M^*)} \left[\log P(\mathbf{X}|G, M^*) \right].$$

Therefore, the difference between the expected log-likelihood of the true graph and that of an arbitrary alternative graph G' is

$$\begin{aligned}\mathcal{L}(G^*) - \mathcal{L}(G') &= \mathbb{E}_{\mathbf{X}\sim P(\mathbf{X}|G^*,M^*)}\left[\log P(\mathbf{X}|G^*,M^*) - \log P(\mathbf{X}|G',M^*)\right] \\ &= \mathbb{E}_{\mathbf{X}\sim P(\mathbf{X}|G^*,M^*)}\left[\log \frac{P(\mathbf{X}|G^*,M^*)}{P(\mathbf{X}|G',M^*)}\right] \\ &= D_{\mathrm{KL}}\left[P(\mathbf{X}|G^*,M^*) \parallel P(\mathbf{X}|G',M^*)\right].\end{aligned}$$

Since the Kullback–Leibler divergence is always non-negative, we have

$$D_{\mathrm{KL}}\left[P(\mathbf{X}|G^*,M^*) \parallel P(\mathbf{X}|G',M^*)\right] \geq 0,$$

which proves that

$$\mathcal{L}(G^*) \geq \mathcal{L}(G').$$

To prove the strictness of the inequality, suppose that there exists a graph $G' \neq G^*$ such that $\mathcal{L}(G^*) = \mathcal{L}(G')$. This implies that, $P(\mathbf{X}|G^*,M^*) = P(\mathbf{X}|G',M^*)$. Therefore, the two distinct graphs G^* and G' induce the same observational distribution under the known mechanism M^*. However, this contradicts the assumption that M^* belongs to a strictly identifiable mechanism class, under which each graph corresponds to a unique induced distribution. Hence, no graph $G' \neq G^*$ can achieve the same expected log-likelihood as G^*.

Therefore, the equality holds if and only if $G' = G^*$. □

A.4 Proof of Theorem 3

To show the identifiability under the broad hypothesis space of causal mechanisms $\mathcal{M}$, we will need to first define the strict identifiability property from the set of causal prior mechanisms. Formally, we say the space of causal prior mechanisms is strictly identifiable if $P(\mathbf{X}|M^*,G^*) \neq P(X|M_{G'},G')$ holds for all alternative graphs $G' \neq G^*$, where M^* and G^* denote the true causal mechanism and true causal structure, respectively, and $M_{G'} = \arg\max_{M\in\mathcal{M}} \mathbb{E}_{\mathbf{X}}[\log P(\mathbf{X}|M,G')]$ represents the optimal mechanism choice under G'. This formulation serves as a direct extension from the traditional known causal mechanism setting, further suggesting that the entire space of causal prior mechanisms must possess the intrinsic property of mutual exclusivity. Under this structural foundation, the exact marginal likelihood uniquely isolates the true causal graph.

Proof. Let $\mathbf{X}^{(N)} = \{x_1, x_2, \ldots, x_N\}$ be a dataset of N independent and identically distributed (i.i.d.) samples drawn from the true causal distribution $P^*(x) = P(x|M^*,G^*)$. For any candidate graph G, the exact marginal likelihood of the dataset is given by the integral over the mechanism space $\mathcal{M}$:

$$P(\mathbf{X}^{(N)}|G) = \int_{\mathcal{M}} P(\mathbf{X}^{(N)}|M,G)P(M|G)dM = \int_{\mathcal{M}} \left(\prod_{i=1}^{N} P(x_i|M,G)\right) P(M|G)dM. \quad (18)$$

We evaluate the scaled log marginal likelihood as $N \to \infty$. Following standard asymptotic Bayesian theory, the integral is asymptotically dominated by the maximum likelihood estimate over the parameter space, and we have:

$$\lim_{N\to\infty} \frac{1}{N} \log P(\mathbf{X}^{(N)}|G) = \sup_{M\in\mathcal{M}} \mathbb{E}_{\mathbf{X}}[\log P(\mathbf{X}|M,G)], \text{ a.s.} \quad (19)$$

We now evaluate Eq. (19) under two distinct structural scenarios. In case 1, the candidate graph is the true graph $G = G^*$. Since the true data-generating distribution is contained in the model class under G^*, the supremum is attained by any mechanism inducing $P(\mathbf{X} \mid M^*, G^*)$. By definition, we have

$$\sup_{M\in\mathcal{M}} \mathbb{E}_{\mathbf{X}}\left[\log P(\mathbf{X}|M,G^*)\right] = \mathbb{E}_{\mathbf{X}}\left[\log P(\mathbf{X}|M^*,G^*)\right].$$

In case 2, an alternative candidate graph $G = G' \neq G^*$. Let $M_{G'} \in \mathcal{M}$ be the optimal mechanism that the alternative graph topology can possibly choose to fit the data, and we have

$$\lim_{N\to\infty} \frac{1}{N} \log P(\mathbf{X}^{(N)}|G') = \mathbb{E}_{\mathbf{X}}[\log P(\mathbf{X}|M_{G'},G')]. \quad (20)$$

We then examine the difference between the two asymptotic limits:

$$\begin{aligned}\Delta\mathcal{L} &= \mathbb{E}_{\mathbf{X}}\left[\log P(\mathbf{X}|M^*, G^*)\right] - \mathbb{E}_{\mathbf{X}}[\log P(\mathbf{X}|M_{G'}, G')] \\ &= \mathbb{E}_{\mathbf{X}}\left[\log \frac{P(\mathbf{X}|M^*, G^*)}{P(\mathbf{X}|M_{G'}, G')}\right] \\ &= D_{\mathrm{KL}}\big(P(\mathbf{X}|M^*, G^*) \parallel P(\mathbf{X}|M_{G'}, G')\big).\end{aligned} \tag{21}$$

Since the mechanism class $\mathcal{M}$ is strictly identifiable, we have:

$$P(\mathbf{X}|M_{G'}, G') \neq P(\mathbf{X}|M^*, G^*), \tag{22}$$

guaranteeing that the KL divergence is strictly positive:

$$D_{\mathrm{KL}}\big(P(\mathbf{X}|M^*, G^*) \parallel P(\mathbf{X}|M_{G'}, G')\big) > 0. \tag{23}$$

Consequently, we conclude that:

$$\lim_{N\to\infty} \frac{1}{N} \log P(\mathbf{X}^{(N)}|G^*) > \lim_{N\to\infty} \frac{1}{N} \log P(\mathbf{X}^{(N)}|G'), \quad \text{a.s.}, \quad \forall G' \neq G^*. \tag{24}$$

This completes the proof that the exact marginal likelihood uniquely identifies the true causal graph G^* almost surely as the sample size approaches infinity. □

A.5 Proof of Corollary 1

Proof. Let $\mathcal{J}(G; \mathbf{X}^{(N)}) := \mathbb{E}_{Q^*(M|\mathbf{X}^{(N)})}\big[\log P(G \mid M)\big]$ denote the structural inference in CDFM. We prove that its maximizer converges almost surely to the true graph G^*, and Q^* denotes the optimum variational posterior obtained by pretraining.

We begin with the connection between the structural inference and the joint likelihood $\log P(\mathbf{X}^{(N)}, G)$, which can be expressed by,

$$\log P(\mathbf{X}^{(N)}, G) = \mathcal{L}(G, \mathbf{X}^{(N)}; Q) + D_{\mathrm{KL}}\left(Q(M \mid \mathbf{X}^{(N)}) \parallel P(M \mid G, \mathbf{X}^{(N)})\right),$$

where

$$\mathcal{L}(G, \mathbf{X}; Q) = \mathbb{E}_{Q(M|\mathbf{X})}\big[\log P(\mathbf{X} \mid G, M) + \log P(G \mid M)\big] - D_{\mathrm{KL}}\big(Q(M \mid \mathbf{X}) \parallel P(M)\big)$$

is the fitted ELBO of CDFM given in Eq. (6):

Then, given that the foundation model has sufficient capacity to reach the optimum variational ELBO, we obtain

$$\log P(\mathbf{X}^{(N)}, G) = \mathcal{L}(G, \mathbf{X}^{(N)}; Q^*), \tag{25}$$

where the optimal variational posterior $Q^*(M \mid \mathbf{X}^{(N)})$ approaches $P(M \mid G, \mathbf{X}^{(N)})$. Thus, since

$$\log P(\mathbf{X}^{(N)}, G) = \log P(\mathbf{X}^{(N)} \mid G) + \log P(G), \tag{26}$$

and under a uniform graph prior $P(G)$, we have

$$\arg\max_G \log P(\mathbf{X}^{(N)}, G) = \arg\max_G \log P(G \mid \mathbf{X}^{(N)}) = \arg\max_G \mathcal{L}(G, \mathbf{X}^{(N)}; Q^*). \tag{27}$$

Subsequently, we establish the connection between $\arg\max_G \mathcal{J}(G; \mathbf{X}^{(N)})$ and $\arg\max_G \mathcal{L}(G, \mathbf{X}^{(N)}; Q^*)$. First, by Theorem 3, strict identifiability of the mechanism space implies that the exact marginal likelihood uniquely identifies the true graph:

$$\lim_{N\to\infty} \frac{1}{N} \log P(\mathbf{X}^{(N)} \mid G^*) > \lim_{N\to\infty} \frac{1}{N} \log P(\mathbf{X}^{(N)} \mid G'), \ \ \forall G' \neq G^*, \ \ \text{a.s.} \tag{28}$$

Since the graph prior $P(G)$ is uniform, Bayes' rule gives

$$P(G \mid \mathbf{X}^{(N)}) \propto P(\mathbf{X}^{(N)} \mid G).$$

Therefore, for any $G' \neq G^*$,

$$\frac{P(G' \mid \mathbf{X}^{(N)})}{P(G^* \mid \mathbf{X}^{(N)})} = \frac{P(\mathbf{X}^{(N)} \mid G')}{P(\mathbf{X}^{(N)} \mid G^*)}. \tag{29}$$

Taking logarithms and dividing by N, Eq. (29) implies that this ratio converges to zero almost surely. Hence

$$P(G^* \mid \mathbf{X}^{(N)}) \to 1, \quad P(G' \mid \mathbf{X}^{(N)}) \to 0, \ \ \forall G' \neq G^*, \ \ \text{a.s.} \tag{30}$$

Thus, the posterior distribution over graphs converges almost surely to a Dirac measure concentrated at G^*. Then, we have

$$\begin{aligned} P(M \mid \mathbf{X}^{(N)}) &= \sum_{G \in \mathcal{G}} P(M \mid G, \mathbf{X}^{(N)}) P(G \mid \mathbf{X}^{(N)}) \\ &= P(M \mid G^*, \mathbf{X}^{(N)}) P(G^* \mid \mathbf{X}^{(N)}) + \sum_{G' \neq G^*} P(M \mid G', \mathbf{X}^{(N)}) P(G' \mid \mathbf{X}^{(N)}). \end{aligned} \tag{31}$$

By Eq. (30), the first posterior weight converges to one and all remaining posterior weights vanish almost surely, and we obtain

$$P(M \mid G^*, \mathbf{X}^{(N)}) \overset{\text{a.s.}}{\sim} P(M \mid \mathbf{X}^{(N)}), \quad N \to \infty. \tag{32}$$

Therefore, since the optimal variational posterior $Q^*(M \mid \mathbf{X}^{(N)})$ approachs to $P(M \mid G, \mathbf{X}^{(N)})$, we have $Q^*(M \mid \mathbf{X}^{(N)}) \overset{\text{a.s.}}{\sim} P(M \mid \mathbf{X}^{(N)})$. This justifies using the amortized encoder $Q(M \mid \mathbf{X}^{(N)})$ in place of the true graph-conditioned mechanism posterior asymptotically.

Asymptotically, the structural inference simplifies to $\mathcal{J}(G; \mathbf{X}^{(N)}) = \mathbb{E}_{P(M|\mathbf{X}^{(N)})}[\log P(G|M)]$. By applying Jensen's inequality, we have:

$$\mathcal{J}(G; \mathbf{X}^{(N)}) = \mathbb{E}_{P(M|\mathbf{X}^{(N)})}\big[\log P(G \mid M)\big] \leq \log \mathbb{E}_{P(M|\mathbf{X}^{(N)})}\big[P(G \mid M)\big] = \log P(G \mid \mathbf{X}^{(N)}).$$

For the true graph G^*, the posterior $P(G^* \mid \mathbf{X}^{(N)}) \to 1$, which implies $\log P(G^* \mid \mathbf{X}^{(N)}) \to 0$, and thus, $\mathcal{J}(G^*; \mathbf{X}^{(N)}) \to 0$. Thus, for G^*, both sides of Jensen's inequality approach zero, making the bound asymptotically tight:

$$\mathcal{J}(G; \mathbf{X}^{(N)}) = \log P(G \mid \mathbf{X}^{(N)}). \tag{33}$$

Since the uniform prior ensures $\arg\max_G \log P(G \mid \mathbf{X}^{(N)}) = \arg\max_G \log P(\mathbf{X}^{(N)} \mid G)$, and Theorem 3 guarantees the marginal likelihood is strictly optimum at G^*, it follows that:

$$\hat{G} = \arg\max_{G \in \mathcal{G}} \mathcal{J}(G; \mathbf{X}^{(N)}) = G^*, \ \ \text{a.s. as } N \to \infty.$$

This completes the proof. □

B Synthetic Mechanism Abbreviations

Table A1: Abbreviations of synthetic mechanism families used in Table A2.

Abbreviation	Description
CAM	Causal additive model
CPT	Conditional probability table for discrete variables
Discrete ANM	Discrete additive noise model
Linear	Linear structural equation model
Linear-Hetero	Linear model with heteroscedastic noise
Linear-Ordinal	Linear model with ordinal or discretized observations
Meas. Error	Mechanism with additional measurement error
Physical	Physics-inspired nonlinear mechanism
Physical-Hetero	Physical mechanism with heteroscedastic noise
PNL	Post-nonlinear causal model
RFF	Random Fourier feature nonlinear mechanism
RFF-Hetero	RFF mechanism with heteroscedastic noise
RFF-Ordinal	RFF mechanism with ordinal or discretized observations
Rounded	Mechanism with rounded observations
Time-Lag	Time-lagged causal mechanism

B.1 Detailed Causal Prior Construction and Pretraining Processes

CDFM is pretrained on synthetically generated $(\mathbf{X}, G)$ pairs, with each instance sampled from a randomized structural causal model (SCM). The prior is designed to expose the model to diverse graph structures, functional mechanisms, marginal distributions, and observational imperfections.

Graph and mechanism diversity. The primary generator samples DAGs from ER, scale-free (SF/SFT), Watts–Strogatz, stochastic-block, and geometric graph families, with $D \in [2, 100]$, $N \in [50, 4096]$, and varying edge density. We further incorporate a heterogeneous SCM generator containing additional chain, tree, layered, bipartite-layered, block-sparse, dense-local-module, and collider-rich structures. Importantly, mechanisms are assigned at the node level, allowing a single SCM to mix linear, polynomial, multiplicative, rational, periodic, piecewise, saturation, regime-switching, neural, RFF, tree-based, aggregation, and discretized relationships. The generator additionally includes additive activation mechanisms with nonlinearities such as ReLU, Leaky-ReLU, Softplus, Sigmoid, ELU, SELU, SiLU, absolute-value, step, modulo, rounding, and rank transforms.

Mechanism type diversity. The core of our data generator is a broad taxonomy of structural causal mechanisms, each encoding distinct functional-form and distribution. We categorize these mechanisms into several families:

Continuous additive mechanisms include linear structural equation models (e.g., $X_j = \sum_{i \in \mathrm{PA}_j} w_{ij} X_i + N_j$) with both homoscedastic and heteroscedastic noise (Yin et al., 2024), random Fourier feature (RFF) models that introduce highly nonlinear dependencies; post-nonlinear causal models (Zhang & Hyvärinen, 2009) where a nonlinear distortion is applied after additive mixing ($X_j = g_2(g_1(\mathbf{PA}_j) + N_j)$), causal additive models (Bühlmann et al., 2014) where each parent contributes through an independent univariate nonlinearity, and physics-inspired mechanisms.

Discrete and mixed-type mechanisms target the prevalent non-continuous data in real-world applications. These include conditional probability tables for purely categorical variables; discrete additive noise models (Peters et al., 2010) for ordinal discrete data; and ordinal causal discovery formulations (Ni & Mallick, 2022) that bridge continuous and discrete regimes within a single mechanism.

Post-processing transformations simulate measurement and recording imperfections. Independent of the base SCM, each batch may undergo additive measurement error (perturbing a subset of variables with independent Gaussian noise) or discretization (rounding a subset of child variables to coarse bins). These transformations can co-occur within the same batch, producing data that reflects the messy, multi-step nature of real-world measurement.

Noise and edge-weight diversity. To ensure that CDFM can disentangle causal signal from exogenous variation, we vary both the noise distribution and the signal-to-noise ratio per mechanism instance. The exogenous noise N_j for each variable is drawn from one of several different distribution families, such as Gaussian, Laplace, Cauchy, Beta, Exponential, Student's-t, log-normal, and Gaussian mixtures.

Latent confounding. Causal sufficiency rarely holds in practice. To equip CDFM with the ability to reason under unobserved confounding, 15% of training batches include latent variables. The generator selects d_{latent} non-leaf nodes as latent, and then generates the full SCM, and finally drops the latent columns. In the resulting ground-truth graph, observed variable pairs that share a latent parent receive bidirected edges $X_i \leftrightarrow X_j$, while direct causal links retain directed edges $X_i \to X_j$. When a pair exhibits both a direct causal effect and a shared latent confounder, the directed edge takes priority. Acyclic constraint is disabled for these batches since bidirected edges inherently form cycles in the adjacency matrix.

Missing values. We further introduce random missingness to provide the auxiliary imputation task with a training signal and to prepare the model for incomplete real-world datasets. With 50% probability, a batch undergoes masking where some portion of the entries are missing completely at random (MCAR). The original values are retained exclusively for computing the quantile imputation loss, while the model receives only the masked matrix with an explicit binary mask channel.

C Detailed Training Objective

CDFM is trained end-to-end using a joint objective function that encompasses standard classification supervision, margin-based regularization, sparsity constraints, and self-supervised reconstruction terms:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{BCE}} + \alpha_{\text{margin}} \cdot \mathcal{L}_{\text{margin}} + \gamma_{\text{F}_1} \cdot \mathcal{L}_{\text{soft}-\text{F}_1} + \beta_{\text{acyc}} \cdot \mathcal{L}_{\text{acyc}} + \alpha_{\text{impute}} \cdot \mathcal{L}_{\text{quantile}}, \quad (34)$$

where α_{margin}, γ_{F_1}, β_{acyc}, and α_{impute} balance the respective loss components.

Structural supervision. To ensure reliable numerical convergence and sharp directional confidence, structural supervision combines standard binary cross-entropy ($\mathcal{L}_{\text{BCE}}$) and a margin-based variant ($\mathcal{L}_{\text{margin}}$). The BCE loss establishes baseline probabilistic alignment:

$$\mathcal{L}_{\text{BCE}} = -\frac{1}{|\mathcal{E}_{\text{valid}}|} \sum_{(i,j)\in\mathcal{E}_{\text{valid}}} \Big[w_{\text{pos}} \cdot G_{ij} \log \sigma(W_{ij}) + (1 - G_{ij}) \log(1 - \sigma(W_{ij})) \Big], \quad (35)$$

where w_{pos} is a positive-class weight adjusted for graph sparsity. The margin-based component acts as a structural regularizer that explicitly penalizes ambiguous predictions near the classification threshold, utilizing a strict margin parameter m:

$$\mathcal{L}_{\text{margin}} = -\frac{1}{|\mathcal{E}_{\text{valid}}|} \sum_{(i,j)\in\mathcal{E}_{\text{valid}}} \Big[w_{\text{pos}} \cdot G_{ij} \log \sigma(W_{ij} - m) + (1 - G_{ij}) \log(1 - \sigma(W_{ij} + m)) \Big], \quad (36)$$

Soft F_1 loss. Because causal adjacency matrices are inherently sparse, standard cross-entropy objectives frequently converge to trivial sparse solutions. We incorporate a differentiable relaxation of the F_1-score:

$$\mathcal{L}_{\text{soft}-\text{F}_1} = 1 - \frac{2 \cdot \text{TP}_{\text{soft}}}{2 \cdot \text{TP}_{\text{soft}} + \text{FP}_{\text{soft}} + \text{FN}_{\text{soft}}}, \quad (37)$$

where the soft true positives (TP_{soft}), false positives (FP_{soft}), and false negatives (FN_{soft}) are evaluated continuously over the predicted sigmoid probabilities $P = \sigma(W)$ and the ground-truth matrix G:

$$\text{TP}_{\text{soft}} = \sum_{i\neq j} P_{ij} G_{ij}, \quad \text{FP}_{\text{soft}} = \sum_{i\neq j} P_{ij}(1 - G_{ij}), \quad \text{FN}_{\text{soft}} = \sum_{i\neq j} (1 - P_{ij}) G_{ij}. \quad (38)$$

Acyclicity constraint. To enforce the structural validity of directed graphs, CDFM integrates a spectral radius acyclicity penalty $\mathcal{L}_{\text{acyc}} = \max(0, \rho(\sigma(W)) - \delta)$, estimated dynamically via power iteration. This constraint is systematically omitted for batches containing latent confounding, where bidirected edges naturally permit cyclic representations in the observed mixed graph.

Quantile imputation loss. Corresponding with the data reconstruction term $\log P(\mathbf{X} \mid G, M)$ defined within the variational lower bound, a pinball loss is evaluated at a set of Q designated quantile levels (denoted by $q_k \in (0, 1)$ for $k = 1, \ldots, Q$) across the set of missing entries $\mathcal{R}$:

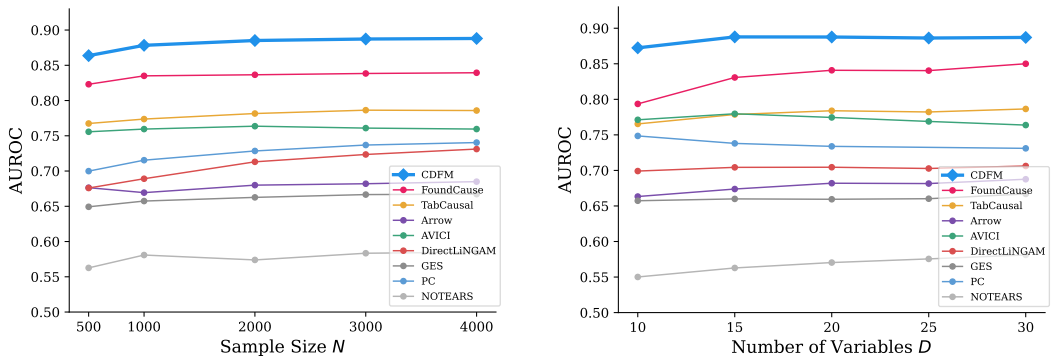
$$\mathcal{L}_{\text{quantile}} = \frac{1}{|\mathcal{R}| \cdot Q} \sum_{(i,j)\in\mathcal{R}} \sum_{k=1}^{Q} 2 \cdot (x_{ij}^{\text{true}} - \hat{y}_{ij}^{q_k}) \cdot \big(q_k - \mathbf{1}[x_{ij}^{\text{true}} \leq \hat{y}_{ij}^{q_k}]\big). \quad (39)$$

D Additional Results: Synthetic Benchmark

This section provides the full synthetic results, including scaling with sample size and dimensionality and mechanism-wise results.

D.1 Sensitivity with sample size and dimensionality

Figure 3 summarizes the synthetic benchmark along the sample-size and graph-size axes, demonstrating sensitivity to the growth of sample size and dimensionality.



(a) AUROC vs. sample size N.

(b) AUROC vs. number of variables D.

Figure 3: Performance over sample size and the number of variables. (a) AUROC as a function of sample size N (full grid); (b) AUROC as a function of the number of variables D, over the range $D \in \{10, 15, 20, 25, 30\}$ covered by the baseline runs. Results are averaged over all 15 mechanism families.

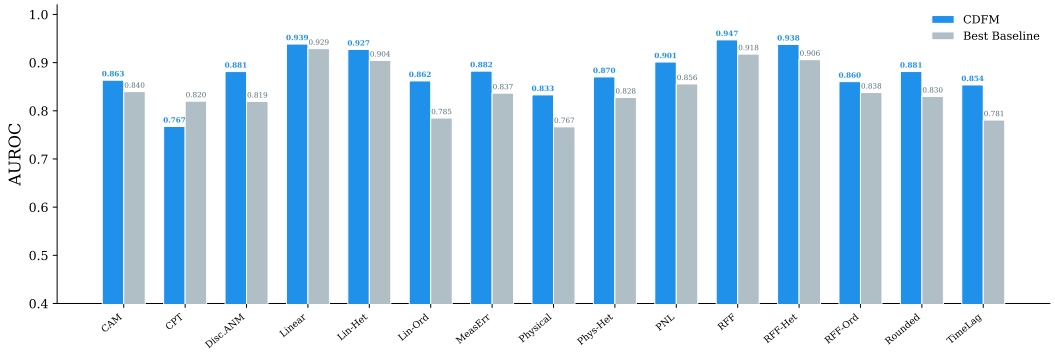


Figure 4: **Mechanism robustness.** AUROC comparison between CDFM and the best-performing baseline for each of the 15 mechanism families. CDFM has the highest AUROC on all 15 mechanism families.

D.2 Robustness across mechanism families

Table A2 breaks performance down by causal mechanism. CDFM has the highest AUROC on all 15 families. Among the non-CDFM methods, FoundCause has the highest AUROC, and TabCausal, PC, and AVICI are strongest on the remaining families. The F_1 comparison is more mechanism-dependent: DirectLiNGAM is strongest on the two exactly linear families, while PC is strongest on several discrete or corrupted-observation settings. Figures 4 and 5 visualize these differences.

E Details of Identifiability Diagnostic Experiments

These diagnostics test whether CDFM responds to known sources of observational identifiability. The pretrained model is fine-tuned on the corresponding synthetic task in each diagnostic, so the experiments are intended to probe behavior under controlled coverage rather than zero-shot transfer.

E.1 Functional asymmetry

We use the bivariate additive-noise model

$$y = x + bx^3 + n,$$

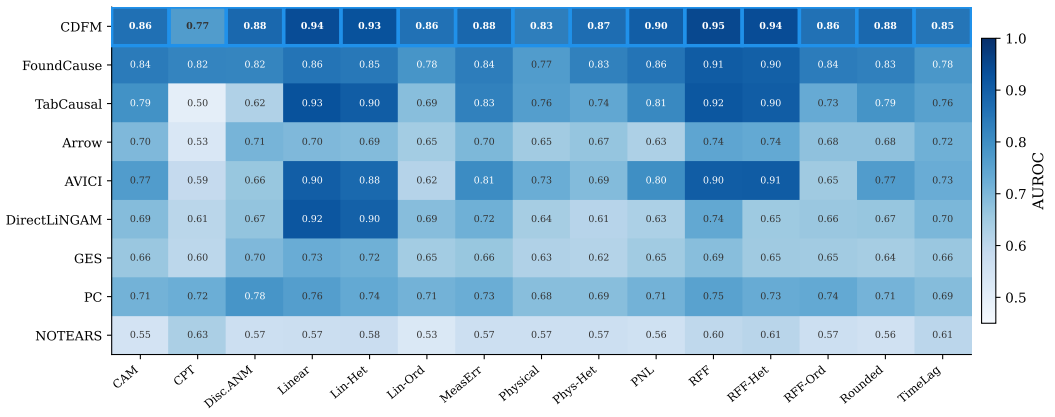


Figure 5: **Method × mechanism AUROC heatmap.** Rows correspond to methods, columns to mechanism families. CDFM (top row) has the highest AUROC across all 15 mechanism families.

Table A2: Synthetic robustness across 15 mechanism families. Results are averaged over nine graph sizes ($D \in \{10, 15, 20, 25, 30, 40, 50, 70, 100\}$) and five sample sizes ($N \in \{500, 1000, 2000, 3000, 4000\}$). Best baseline denotes the strongest non-CDFM method under each metric (FoundCause is considered; it covers 14 of the 15 families). Bold indicates best overall. Mechanism abbreviations are explained in Appendix B.

Mechanism	CDFM AUROC↑	Best Baseline AUROC↑	AUROC baseline	CDFM F_1 ↑	Best Baseline F_1 ↑	F_1 baseline
CAM	**0.863 ± 0.085**	0.843 ± 0.099	FoundCause	**0.591 ± 0.156**	0.508 ± 0.146	PC
CPT	**0.767 ± 0.136**	0.725 ± 0.122	PC	0.373 ± 0.212	**0.468 ± 0.177**	PC
Discrete ANM	**0.881 ± 0.082**	0.826 ± 0.131	FoundCause	**0.600 ± 0.171**	0.598 ± 0.150	PC
Linear	**0.939 ± 0.071**	0.929 ± 0.089	TabCausal	0.668 ± 0.192	**0.759 ± 0.190**	DirectLiNGAM
Linear-Hetero	**0.927 ± 0.074**	0.904 ± 0.098	TabCausal	0.641 ± 0.187	**0.729 ± 0.173**	DirectLiNGAM
Linear-Ordinal	**0.862 ± 0.113**	0.791 ± 0.185	FoundCause	**0.553 ± 0.196**	0.477 ± 0.208	PC
Meas. Error	**0.882 ± 0.098**	0.827 ± 0.128	TabCausal	**0.591 ± 0.193**	0.516 ± 0.191	PC
Physical	**0.833 ± 0.097**	0.767 ± 0.134	FoundCause	**0.522 ± 0.187**	0.446 ± 0.167	PC
Physical-Hetero	**0.870 ± 0.081**	0.831 ± 0.112	FoundCause	**0.558 ± 0.173**	0.445 ± 0.168	FoundCause
PNL	**0.901 ± 0.096**	0.859 ± 0.132	FoundCause	**0.589 ± 0.211**	0.453 ± 0.218	PC
RFF	**0.947 ± 0.054**	0.918 ± 0.090	TabCausal	**0.680 ± 0.178**	0.647 ± 0.204	TabCausal
RFF-Hetero	**0.938 ± 0.057**	0.906 ± 0.096	AVICI	**0.654 ± 0.178**	0.647 ± 0.215	AVICI
RFF-Ordinal	**0.860 ± 0.100**	0.846 ± 0.128	FoundCause	**0.547 ± 0.188**	0.512 ± 0.176	PC
Rounded	**0.881 ± 0.096**	0.834 ± 0.141	FoundCause	**0.575 ± 0.193**	0.456 ± 0.196	PC
Time-Lag	**0.854 ± 0.115**	0.785 ± 0.139	FoundCause	**0.558 ± 0.191**	0.463 ± 0.189	PC

where $b \in [-1, 1]$ controls nonlinearity and $q \in [0.5, 2.0]$ controls the non-Gaussianity of the noise. The point $(b, q) = (0, 1)$ is the linear-Gaussian case and is observationally non-identifiable. CDFM is fine-tuned with $b \sim \mathcal{U}[-1, 1]$, $q \sim \mathcal{U}[0.5, 2.0]$, and $N \in [100, 500]$.

For linear SCMs ($b = 0$), direction accuracy is near chance at $q = 1$ and approaches 1.0 as the noise becomes more non-Gaussian. Under Gaussian noise ($q = 1$), accuracy rises from approximately 0.5 in the linear case to 0.97–1.00 as nonlinearity increases, consistent with additive-noise identifiability Hoyer et al. (2009).

E.2 Orientation from V-structures

We next test whether CDFM uses orientation information available from conditional-independence structure. We generate linear-Gaussian SEMs, where directions are identifiable only when compelled by graphical structure such as unshielded colliders. We enumerate non-isomorphic DAGs with $d = 5$ and $N = 100$, group them by the number of V-structures, and compare CDFM with PC and a perfect-d-separation upper bound.

As the number of V-structures increases, CDFM's orientation F_1 increases and approaches the perfect-d-separation upper bound. PC is more affected by finite-sample conditional-independence errors. The diagnostic therefore separates the use of identifiable graphical information from orientation that would exceed the Markov-equivalence boundary.

E.3 LATENT CONFOUNDING

Finally, we compare three bivariate data-generating structures with only (X, Y) observed: pure confounding $X \leftarrow L \rightarrow Y$, a directed effect $X \rightarrow Y$ under confounding, and the reverse direction $Y \rightarrow X$ under confounding.

Pure confounding is most distinguishable in the sub-Gaussian regime, with accuracy 0.86 at $q = 0.5$, and is hardest near Gaussian noise, with accuracy 0.64 at $q = 1$. The directed cases reach their highest accuracy under strongly super-Gaussian noise (0.66 at $q = 2.0$). These results show that non-Gaussian signatures provide useful information for separating direct effects from latent confounding, while the distinction becomes substantially weaker near Gaussianity.

F REAL CAUSAL BENCHMARK

Reliable evaluation is important to causal discovery, yet most existing benchmarks rely heavily on synthetic data where the ground-truth graph is known by construction. Extending evaluation to real-world tabular data is substantially harder, because complete causal ground truth is rarely available and neither domain annotations nor statistical tests alone provide a sufficiently reliable reference.

To address this, we construct RCB by drawing on seven publicly available datasets from diverse domains: a microwave scattering dataset for food contaminant detection Urbinati et al. (2023), the FICO HELOC credit-risk dataset from the Explainable Machine Learning Challenge FICO (2018), the UCI Heart Disease dataset Janosi et al. (1989), weekly Hass avocado retail sales data Hass Avocado Board (2018), the National Longitudinal Survey of Youth 1979 (NLSY79) U.S. Bureau of Labor Statistics (1979), a used Fiat 500 listing dataset Cons (2019), and the California Housing dataset derived from the 1990 U.S. Census Pace & Barry (1997).

We then construct RCB using a two-stage validation procedure that combines LLM-assisted causal auditing with KCI-based statistical testing. The audit proposes and reviews candidate causal relations and their orientations from domain semantics, while KCI tests whether the corresponding dependencies are supported by the observed data. Only relations supported by both sources are retained for evaluation, and ambiguous pairs are left unlabeled, yielding a conservative but more credible benchmark for real-world causal discovery.

F.1 BENCHMARK CONSTRUCTION

We begin with dataset-specific domain annotations and remove deterministic or derived variables before evaluation. For the remaining candidates, KCI Zhang et al. (2011) at $\alpha = 0.05$ is used to test nonlinear conditional dependence.

Directed and confounded relations. A directed relation is retained only when both semantic and statistical evidence are available: domain knowledge provides the orientation, while KCI rejects $X \perp Y \mid S$ for a local adjustment set S formed from the declared parents of the two endpoints. KCI therefore supports the existence of the relation, whereas the direction comes from temporal order, a physical mechanism, or another domain-specific argument. For latent-confounding annotations, we retain only specific shared-cause groups for which marginal KCI rejects independence, and exclude broad latent groups. If a directed annotation overlaps a broader latent annotation, the directed relation is used.

Trusted non-edges. An unannotated pair is not automatically treated as a non-edge. Instead, we search conditioning sets up to order four directly with KCI. A pair is retained as a trusted non-edge when KCI does not reject conditional independence for at least one conditioning set. Pairs that conflict with an annotated relation remain unlabeled, and marginally independent pairs are excluded to avoid trivial negatives.

F.2 EVALUATION PROTOCOL AND BENCHMARK COMPOSITION

RCB is evaluated over ordered variable pairs. For a trusted directed edge $a \rightarrow b$, (a, b) is positive and (b, a) is negative; a trusted confounded pair contributes two positive cells, while a trusted non-

Relation in RCB	Ordered-cell labels
Trusted directed edge $a \to b$	(a, b) is positive; (b, a) is negative
Trusted confounded pair $a \leftrightarrow b$	Both (a, b) and (b, a) are positive
Trusted non-edge $\{a, b\}$	Both (a, b) and (b, a) are negative
Unresolved pair	No label

Table A3: RCB scoring scheme. Unresolved pairs are excluded from evaluation.

Dataset	directed	confounded	non-edges
`Contaminant`	30	0	0
`FICO-HELOC`	13	1	76
`Cardiovascular`	11	3	17
`avocado_sales`	10	0	25
`NLSY`	8	2	62
`Fiat-500`	3	1	5
`houses`	3	0	2
Total	**78**	**7**	**187**

Table A4: Composition of RCB. The combined evaluation contains 544 labeled ordered cells: 92 positives and 452 negatives.

edge contributes two negative cells. Unresolved pairs are excluded. We report the combined score over all labeled cells, together with orientation and adjacency diagnostics that isolate direction and edge existence, respectively (Table A3).

F.3 Dataset-level semantics and edge evidence

The following tables list every trusted directed edge and confounded pair. For directed edges, the semantic interpretation supplies the orientation, and the evidence column reports the KCI dependence screen; $|S|$ denotes the size of the local adjustment set. Confounded pairs are reported with the shared-cause annotation used to interpret them.

F.3.1 Contaminant

`Contaminant` contains a binary contamination label and 30 non-diagonal entries of a six-port microwave scattering matrix measured at 10.0 GHz. The semantic annotation places contamination status before the measured electromagnetic responses, and all 30 label–response relations pass the dependence screen.

Table A5: Trusted directed edges for `Contaminant`. All thirty edges share the same semantic interpretation and dependence test.

Edge(s)	Meaning	Evidence
$\mathtt{y} \to s_{ij}$ for all thirty non-diagonal ports (30 edges)	$\mathtt{y}$ is the binary contamination label, and s_{ij} is the measured electromagnetic response at 10.0 GHz. In the domain semantics, the contamination state encoded by $\mathtt{y}$ is the condition that affects the response.	marginal $p < 0.003$ on all thirty; $\lvert S\rvert = 0$

F.3.2 Fiat-500

`Fiat-500` contains used-car listings with vehicle age, mileage, dealer location, and asking price. Three directed relations survive screening—age $\to$ mileage, age $\to$ price, and mileage $\to$ price—with orientation supplied by temporal order. The latitude–longitude pair is retained as a confounded relation representing shared dealer location.

Table A6: Trusted relations for `Fiat-500`.

Edge(s)	Meaning	Evidence
`age_in_days` → `km`	Vehicle age precedes accumulated mileage, providing the temporal orientation.	$p < 10^{-3}$; $\lvert S\rvert = 1$
`age_in_days` → `y`; `km` → `y` (2 edges)	Vehicle age and accumulated mileage are pre-existing characteristics at the time a listing price is set. They can enter a seller's valuation of the vehicle, whereas the asking price cannot change the vehicle's prior age or accumulated mileage.	$p < 10^{-3}$; $\lvert S\rvert = 3$
`lat` ↔ `lon`	Latitude and longitude are two coordinate measurements of the same underlying dealer location. Neither coordinate is naturally interpreted as causing the other, so the pair is represented through a shared-location relation rather than a directed edge.	shared-location rationale

F.3.3 AVOCADO SALES

`avocado_sales` is a weekly US retail panel indexed by region, date, and product type. Ten directed relations survive screening: product type and region precede the observed sales composition, while calendar variables precede the weekly price. The dataset also contributes 25 trusted non-edge pairs.

Table A7: Trusted relations for `avocado_sales`.

Edge(s)	Meaning	Evidence
`type` → `4770`; `type` → `Small Bags`; `type` → `Large Bags`; `type` → `XLarge Bags` (4 edges)	Product type is fixed before weekly sales and affects the observed PLU and bag-format mix.	$p < 10^{-3}$; $\lvert S\rvert = 1$
`type` → `y`	Product type is fixed before the weekly price is observed and defines the market segment in which the item is sold. Price can respond to systematic differences between product types, whereas the observed price does not determine the product's pre-existing type.	$p < 10^{-3}$; $\lvert S\rvert = 3$
`region` → `4046`; `region` → `4225`; `region` → `4770` (3 edges)	Region is fixed before the weekly sales observation and affects the product-size mix.	$p < 10^{-3}$, $p = 0.0018$, $p = 0.011$; $\lvert S\rvert = 1$
`year` → `y`; `month` → `y` (2 edges)	Calendar variables precede the observed weekly price and capture temporal variation.	$p < 10^{-3}$, $p = 4 \times 10^{-4}$; $\lvert S\rvert = 3$

F.3.4 NLSY

The NLSY table contains family background, ability, schooling, 1990 labor-market variables, anthropometry, and 1996 income. Eight directed relations and two confounded pairs survive screening. The directed relations are oriented primarily by temporal order, while the two confounded pairs represent shared body-size and cognitive-endowment factors.

Table A8: Trusted relations for NLSY.

Edge(s)	Meaning	Evidence
age → famsize	Age precedes the household composition recorded in 1990.	$p < 10^{-3}$; $\lvert S\rvert = 0$
race → grade90	The recorded demographic background is fixed before the schooling outcome is completed, so the reverse direction from later educational attainment to the earlier baseline attribute is not temporally plausible.	$p < 10^{-3}$; $\lvert S\rvert = 0$
gender → height; gender → weight (2 edges)	The recorded gender variable is treated as a baseline attribute that precedes adult height and weight in the life-course ordering. Adult anthropometric measurements cannot determine that earlier baseline attribute.	$p < 10^{-3}$; $\lvert S\rvert = 0$
grade90 → typejob90	Completed schooling precedes the recorded job type.	$p < 10^{-3}$; $\lvert S\rvert = 1$
afqt89 → income96; grade90 → income96; jobsnum90 → income96 (3 edges)	Ability, schooling, and the 1990 job count are observed before the 1996 income outcome and are therefore oriented toward later income.	$p = 2 \times 10^{-4}$, $p < 10^{-3}$, $p = 0.016$; $\lvert S\rvert =$ 5–6
height ↔ weight; afqt89 ↔ grade90 (2 confounded pairs)	Height and weight can share unmeasured body-size determinants, while AFQT score and schooling can share earlier cognitive and family-background determinants; the benchmark therefore avoids assigning either pair a single directed effect.	specific shared-cause rationale

F.3.5 FICO-HELOC

FICO−HELOC contains credit-bureau attributes and a later default outcome. The retained relations include documented feature-construction links into ExternalRiskEstimate and temporally ordered links from application-time credit-history variables to default. Thirteen directed relations and one confounded pair survive screening.

Table A9: Trusted relations for FICO−HELOC.

Edge(s)	Meaning	Evidence
MSinceOldestTradeOpen, AverageMInFile, NumTrades60Ever2DerogPubRec, NumTrades90Ever2DerogPubRec, PercentInstallTrades, NetFractionRevolvingBurden, NetFractionInstallBurden, PercentTradesWBalance, MaxDelq2PublicRecLast12M, MaxDelqEver → ExternalRiskEstimate (10 edges)	These credit-history fields are documented inputs to ExternalRiskEstimate, which fixes the construction direction.	$p \in [0.0016, 0.037]$; $\lvert S\rvert = 19$
PercentTradesNeverDelq → y; MaxDelqEver → y (2 edges)	Repayment and delinquency history are recorded before the later default outcome.	$p = 0.0076$, $p = 0.025$; $\lvert S\rvert = 12$
MSinceMostRecentInqexcl7days → y	Recent credit-seeking activity is recorded before the default outcome window.	$p < 10^{-3}$; $\lvert S\rvert = 12$

Table A9 (continued)

Edge(s)	Meaning	Evidence
NumSatisfactoryTrades ↔ NumTotalTrades	Both variables summarize the same underlying credit-account history, so the benchmark treats their association as arising from shared account-history structure rather than assigning one summary measure as the cause of the other.	shared-history rationale

F.3.6 HOUSES

The houses data contain California census block-group attributes and median house value. Three directed relations survive screening, linking housing age to room count and both room count and median income to house value.

Table A10: Trusted directed edges for houses.

Edge(s)	Meaning	Evidence
housing_median_age → total_rooms	Median housing age summarizes the construction history of the existing housing stock, whereas total rooms measures the resulting stock observed at census time. This makes construction history an upstream descriptor of the observed inventory rather than an effect of the contemporaneous room count.	$p < 10^{-3}$; $\lvert S\rvert = 0$
total_rooms → y	Total rooms is a characteristic of the housing stock already present when median value is observed.	$p < 10^{-3}$; $\lvert S\rvert = 2$
median_income → y	Median income is treated as an upstream neighborhood demand and socioeconomic characteristic that can influence local housing values.	$p < 10^{-3}$; $\lvert S\rvert = 1$

F.3.7 CARDIOVASCULAR

The cardiovascular dataset contains age, anthropometry, blood pressure, lifestyle indicators, cholesterol and glucose grades, gender, and a disease label. RCB retains 11 directed relations, three confounded pairs, and 17 trusted non-edge pairs. The directed relations are oriented using temporal precedence for fixed baseline variables and risk-factor-to-outcome reasoning for clinical variables.

Table A11: Trusted relations for Cardiovascular.

Edge(s)	Meaning	Evidence
age → ap_hi; age → ap_lo (2 edges)	Age is temporally prior to the blood-pressure measurements and provides the orientation.	$p < 10^{-3}$, $p = 0.004$; $\lvert S\rvert = 1$
height → weight	Adult stature is largely established before the contemporaneous body-weight measurement and provides a structural constraint on body mass. The reverse direction from current weight to already-attained adult height is therefore less plausible for this snapshot.	$p < 10^{-3}$; $\lvert S\rvert = 0$
weight → ap_hi; weight → ap_lo (2 edges)	Body mass is treated as an upstream physiological risk factor that can affect vascular load and blood pressure.	$p < 10^{-3}$; $\lvert S\rvert = 2$

Table A11 (continued)

Edge(s)	Meaning	Evidence
`cholesterol` → `y`; `smoke` → `y` (2 edges)	Cholesterol status and smoking are treated as risk factors that precede the disease outcome.	$p < 10^{-3}$, $p = 0.030$; $\lvert S\rvert = 7$
`ap_hi` → `y`; `ap_lo` → `y`; `age` → `y`; `weight` → `y` (4 edges)	Age and weight are upstream risk factors, while elevated blood pressure is treated as a cardiovascular risk state that can contribute to disease.	$p < 10^{-3}$, $p = 0.015$, $p < 10^{-3}$, $p = 0.040$; $\lvert S\rvert = 7$–8
`height` ↔ `gender`; `cholesterol` ↔ `gluc`; `smoke` ↔ `alco` (3 confounded pairs)	The benchmark uses bidirected relations where the observed association is plausibly explained by shared unmeasured determinants: constitutional factors for height–gender, metabolic factors for cholesterol–glucose, and behavioral or social factors for smoking–alcohol use.	specific shared-cause rationale

G LLM USE STATEMENT

We used large language models (LLMs) as general-purpose assistants for drafting and polishing parts of the manuscript, providing feedback on methodological ideas, and suggesting code snippets. All substantive decisions regarding the research problem, methodology, and experimental design were made by us. We reviewed and revised the AI-assisted content and take full responsibility for the final manuscript.