WHAT DO CAUSAL DISCOVERY FOUNDATION MODELS LEARN?

Anonymous authors
Paper under double-blind review

ABSTRACT

Recovering a unique causal structure from observational data is difficult due to finite data and non-identifiability. Recent causal discovery foundation models pretrain on synthetic causal data and directly infer structures for new datasets, avoiding repeated structure search. Yet when the causal structure is not identifiable or the true data-generating distribution lies outside the distribution family used for pretraining, what these models learn and which structures they can recover remain unclear. To answer these questions, we develop a statistical framework that characterizes both the graph distribution learned through pretraining and its large-sample recovery target. We prove that, as the number of pretraining tasks and model capacity increase, the learned graph distribution approaches the posterior over graphs induced by the pretraining prior. As the sample size increases, this posterior concentrates on the KL-best-fitting graph set under model misspecification. Under correct specification and the Markov and faithfulness assumptions, it concentrates on the Markov equivalence class of the true DAG; under additional identifiability assumptions, it concentrates on the true DAG. These results provide a unified perspective for understanding existing models and guide the design of pretraining distributions and model architectures. As a concrete instantiation, we factorize DAGs into skeletons and topological orders, use conditional diffusion to model multimodal distributions over topological orders, and orient skeletons by sampled orders to guarantee acyclicity. Across diverse synthetic benchmarks, our method achieves stronger overall structure recovery than traditional methods and other causal discovery foundation models.

1 INTRODUCTION

Discovering causal relationships from observational data is essential for understanding underlying mechanisms, designing interventions, and supporting scientific decision-making (Spirtes et al., 2000; Pearl, 2009). Traditional causal discovery methods use conditional-independence testing, graph search, or continuous optimization (Spirtes et al., 2000; Chickering, 2002; Zheng et al., 2018). Their recovery guarantees rely on assumptions about the data-generating process, and their potentially costly structure search or optimization must generally be repeated for each new dataset. Causal discovery foundation models address these limitations by pretraining on synthetic tasks generated from diverse graphs, causal mechanisms, and noise distributions (Lorch et al., 2022; Ke et al., 2023; Thompson et al., 2026). Given pairs of datasets and their generating graphs, these models learn to infer causal structures directly from new datasets without repeating the full task-specific inference procedure. Broad pretraining allows a single model to incorporate information from multiple causal model families and supports zero-shot inference across datasets. This paradigm shifts much of the computational cost of causal discovery to pretraining, making its statistical foundations increasingly important.

Existing research on causal discovery foundation models can be broadly divided into two lines. Much of the empirical work focuses on developing model architectures and expanding or diversifying the pretraining distribution, with effectiveness demonstrated primarily through benchmark performance (Peng et al., 2026; Li et al., 2026; Blöbaum et al., 2026). Benchmark performance alone, however, does not explain why the pretraining pipeline works or how its predictions on new datasets should be interpreted. A second line of work makes identifiability central to its theoretical analysis. These methods include identifiable causal model families in pretraining and establish recovery

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.

results when the test distribution and the prior-supported mechanism space satisfy the corresponding identifiability conditions (Qiao et al., 2026; Chen et al., 2026). Their theoretical guarantees do not extend to non-identifiable settings or to cases in which the test distribution falls outside the pretraining model family, limiting their applicability in practice.

These observations raise several questions: Why does this pretraining-based causal discovery pipeline work? When the test distribution falls outside the pretraining model family, or when the causal structure is non-identifiable, what causal structures can such models recover? Like classical causal discovery methods, do causal discovery foundation models enjoy convergence guarantees as the number of observations grows?

To answer these questions, we develop a unified statistical framework that studies two asymptotic regimes. First, we fix the number of observations in each dataset and consider increasingly many pretraining tasks and increasingly expressive model classes. In the resulting limit, the learned graph distribution approaches the graph posterior induced by the pretraining prior. Second, we study what this posterior recovers as the number of observations grows. We define the *KL-best-fitting graph set* as the set of graphs whose prior-supported mechanism families most closely approximate the true data distribution in KL divergence. Under suitable support and regularity conditions, the posterior concentrates on this set, even when the model is misspecified. Under correct specification and the Markov and faithfulness assumptions, this set coincides with the Markov equivalence class of the true DAG; with additional identifiability assumptions, it reduces to the true DAG.

Guided by this framework, we develop a causal discovery foundation model that decomposes a DAG into an undirected skeleton and a topological order, using diffusion to model order uncertainty while guaranteeing acyclic graph samples.

Our contributions are threefold. (1) We establish a unified statistical framework that characterizes the learning and recovery targets of causal discovery foundation models, explains their empirical effectiveness, and guides the design of pretraining distributions and posterior parameterizations. **(2)** To our knowledge, we introduce the first causal discovery foundation model that uses conditional diffusion over permutations to model topological-order posteriors, capturing multimodal order uncertainty while guaranteeing acyclic graph samples. **(3)** We demonstrate the effectiveness of our model across diverse synthetic benchmarks.

2 Related Work

Classical and Bayesian causal discovery. Classical methods recover causal structure through conditional independence testing, score-based search, or continuous optimization, as exemplified by PC, GES, and NOTEARS (Spirtes et al., 2000; Chickering, 2002; Zheng et al., 2018). Their recovery targets depend on the underlying assumptions: observational conditional independences generally identify a Markov equivalence class, while unique-DAG recovery requires additional restrictions (Verma & Pearl, 1990; Pearl, 2009). Bayesian methods represent the remaining uncertainty through a posterior over graphs (Heckerman et al., 1995; Friedman & Koller, 2003). To make posterior inference tractable, BCD Nets uses variational inference, while DiBS and BayesDAG develop differentiable representations for gradient-based inference (Cundy et al., 2021; Lorch et al., 2021; Annadani et al., 2023).

Causal discovery foundation models. Learning from synthetic graph–dataset pairs provides a way to replace repeated structure search with neural inference. AVICI learns directed-edge probabilities, while CSIvA uses an autoregressive decoder to capture dependencies when predicting adjacency matrices (Lorch et al., 2022; Ke et al., 2023). SEA combines classical estimates obtained from variable subsets, whereas Kode incorporates coarse prior knowledge alongside observations (Wu et al., 2025; Xu et al., 2026). ADAG jointly predicts structures and parameters under linear structural equation models, whereas CauScale provides a scalable architecture for causal discovery on graphs with up to 1,000 nodes (Yin et al., 2025; Peng et al., 2026). For heterogeneous mechanisms, CDFM treats unknown mechanisms as latent variables and derives a variational decomposition into mechanism-inference, reconstruction, and graph-inference modules (Qiao et al., 2026). DAG-FM instead constructs graphs sequentially through leaf and parent prediction (Chen et al., 2026).

Statistical foundations. Nagler (2023) provides a statistical analysis of prior-data fitted networks, relating their learning objective to approximation of a prior-induced posterior predictive distribu-

tion. Within causal discovery, Montagna et al. (2025) show, primarily in the bivariate setting, how assumptions encoded in synthetic training data determine which causal directions a transformer can identify. Their analysis clarifies that pretraining incorporates causal assumptions rather than removing the need for them. CDFM and DAG-FM further establish unique-graph recovery under suitable identifiability conditions on the mechanism families (Qiao et al., 2026; Chen et al., 2026). Our framework connects these questions through the prior-induced graph posterior: it separates learning this posterior from its large-sample concentration and characterizes recovery targets under both non-identifiability and model misspecification.

Order distributions and diffusion. BCNP combines a Gumbel–Sinkhorn permutation with a lower-triangular adjacency matrix, while Arrow orients a predicted skeleton using a Plackett–Luce order distribution (Dhir et al., 2025; Thompson et al., 2026). DiffAN instead uses denoising diffusion to estimate the data score and its Jacobian for iterative leaf removal under additive noise models, rather than modeling a distribution over orders (Sanchez et al., 2023). Classical Plackett–Luce and Mallows models impose structured permutation distributions, while Gumbel–Sinkhorn provides a differentiable continuous relaxation (Luce, 1959; Plackett, 1975; Mallows, 1957; Mena et al., 2018). SymmetricDiffusers provides a more flexible construction through successive discrete denoising steps on the permutation space $\mathfrak{S}_d$ (Zhang et al., 2025).

3 A Statistical Framework for Causal Discovery Foundation Models

3.1 Framework and Target

We follow the standard supervised-pretraining setup of causal discovery foundation models, in which a model is trained on synthetic graph–dataset pairs to predict causal structures for new datasets (Thompson et al., 2026; Li et al., 2026; Blöbaum et al., 2026; Chen et al., 2026). Let $G \in \mathcal{G}_d$ be a DAG over d observed variables. We denote a structural causal model (SCM) by (G, η), where η specifies the causal functions, their parameters, and the exogenous noise distributions. Together, G and η determine an observational distribution $P_{G,\eta}$. A pretraining prior Π specifies a distribution over these causal models. Each synthetic task is generated by

$$(G, \eta) \sim \Pi, \qquad X^{(n)} = (X_1, \dots, X_n) \sim P^n_{G,\eta},$$

where $X_1, \dots, X_n \in \mathbb{R}^d$ are i.i.d. observations from the sampled model. For fixed n, this procedure induces a joint distribution $P_{\Pi,n}(G, X^{(n)})$, with dataset marginal $P_{\Pi,n}(X^{(n)})$ and graph posterior $P_{\Pi,n}(G \mid X^{(n)})$. A causal discovery foundation model with learnable parameters θ takes an observational dataset $X^{(n)}$ as input and outputs a conditional distribution $q_\theta(\cdot \mid X^{(n)})$ over the candidate DAGs in $\mathcal{G}_d$.

The model approximates the target graph posterior by minimizing

$$\mathcal{L}_{\mathrm{KL}}(\theta) = \mathbb{E}_{X^{(n)} \sim P_{\Pi,n}} \Big[\mathrm{KL}\Big(P_{\Pi,n}(\cdot \mid X^{(n)}) \,\|\, q_\theta(\cdot \mid X^{(n)}) \Big) \Big].$$

By the KL–cross-entropy identity, minimizing $\mathcal{L}_{\mathrm{KL}}(\theta)$ is equivalent to minimizing the population graph log-loss

$$R_n(\theta) = \mathbb{E}_{(G, X^{(n)}) \sim P_{\Pi,n}} [-\log q_\theta(G \mid X^{(n)})].$$

Given M independently generated pretraining tasks $\{(G_m, X^{(n)}_m)\}_{m=1}^M$, the corresponding empirical graph log-loss is

$$\widehat{R}_{M,n}(\theta) = -\frac{1}{M} \sum_{m=1}^{M} \log q_\theta(G_m \mid X^{(n)}_m).$$

As illustrated in Figure 1, our framework separates posterior learning from large-sample causal recovery. The first component concerns posterior learning. For fixed n, it asks how finite pretraining data and model capacity affect the approximation of $P_{\Pi,n}(G \mid X^{(n)})$. The second studies large-sample behavior. It asks, for $X^{(n)}_0 \sim P^n_0$ generated from a distribution associated with a true DAG G_0, where $P_{\Pi,n}(G \mid X^{(n)}_0)$ concentrates as n grows.

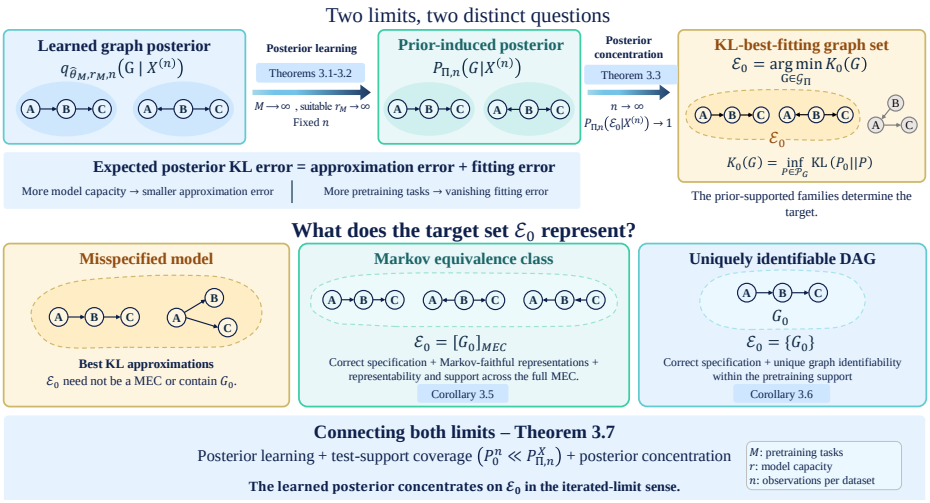


Figure 1: Overview of our statistical framework.

3.2 Learning the Target Graph Posterior

We first study whether pretraining learns the target graph posterior $P_{\Pi,n}(G \mid X^{(n)})$ for a fixed dataset size n. Let

$$\mathcal{Q}_r := \left\{ q_\theta(G \mid X^{(n)}) : \theta \in \Theta_r \right\}$$

denote the set of conditional graph distributions that can be represented by the network architecture and its allowed parameter settings at capacity level r. Here, Θ_r is the corresponding parameter space, and larger r represents greater expressive capacity.

For $\theta \in \Theta_r$, we write $R_{r,n}(\theta) := R_n(\theta)$ and $\widehat{R}_{M,r,n}(\theta) := \widehat{R}_{M,n}(\theta)$, where the subscript r makes explicit the model class over which optimization is performed.

Given M independently generated pretraining tasks, let $\widehat{\theta}_{M,r,n}$ be an M^{-1}-approximate minimizer of the empirical graph log-loss:

$$\widehat{R}_{M,r,n}\left(\widehat{\theta}_{M,r,n}\right) \le \inf_{\theta \in \Theta_r} \widehat{R}_{M,r,n}(\theta) + \frac{1}{M}. \tag{1}$$

We define the approximation and fitting errors by

$$\varepsilon_{\mathrm{app}}(r,n) := \inf_{\theta \in \Theta_r} \mathbb{E}_{X^{(n)} \sim P_{\Pi,n}} \left[\mathrm{KL}\Big(P_{\Pi,n}(\cdot \mid X^{(n)}) \,\|\, q_\theta(\cdot \mid X^{(n)}) \Big) \right],$$

$$\varepsilon_{\mathrm{fit}}(M,r,n) := R_{r,n}\left(\widehat{\theta}_{M,r,n}\right) - \inf_{\theta \in \Theta_r} R_{r,n}(\theta).$$

The approximation error is the smallest posterior error achievable by the model class, even with unlimited pretraining data. The fitting error is the excess population risk of the model learned from M tasks relative to the best model in the class. Appendix A.1 states and discusses the assumptions used below.

Theorem 3.1 (Posterior-error decomposition and best-in-class convergence)**.** *Fix r and n. Under the finite-log-loss condition in Assumption A.2,*

$$\mathbb{E}_{X^{(n)} \sim P_{\Pi,n}} \left[\mathrm{KL}\Big(P_{\Pi,n}(\cdot \mid X^{(n)}) \,\|\, q_{\widehat{\theta}_{M,r,n}}(\cdot \mid X^{(n)}) \Big) \right] = \varepsilon_{\mathrm{app}}(r,n) + \varepsilon_{\mathrm{fit}}(M,r,n).$$

If the fixed-class uniform law of large numbers in Assumption A.3 also holds, then

$$\varepsilon_{\mathrm{fit}}(M,r,n) \xrightarrow[M \to \infty]{P} 0.$$

For graph distributions p and q, write $\|p - q\|_{\mathrm{TV}} = \frac{1}{2} \sum_{G \in \mathcal{G}_d} |p(G) - q(G)|$.

Theorem 3.2 (Consistent approximation of the target graph posterior)**.** *Fix n. Suppose that the model classes are nested, $\mathcal{Q}_1 \subseteq \mathcal{Q}_2 \subseteq \cdots$, and suppose that Assumptions A.1, A.2, A.3, and A.4 hold. Then there exists a deterministic nondecreasing sequence $r_M \to \infty$ such that*

$$\mathbb{E}_{X^{(n)}\sim P_{\Pi,n}}\left[\mathrm{KL}\Big(P_{\Pi,n}(\cdot \mid X^{(n)}) \,\|\, q_{\widehat{\theta}_{M,r_M,n}}(\cdot \mid X^{(n)})\Big)\right] \xrightarrow[M\to\infty]{P} 0.$$

Moreover,

$$\mathbb{E}_{X^{(n)}\sim P_{\Pi,n}}\left[\Big\|P_{\Pi,n}(\cdot \mid X^{(n)}) - q_{\widehat{\theta}_{M,r_M,n}}(\cdot \mid X^{(n)})\Big\|_{\mathrm{TV}}\right] \xrightarrow[M\to\infty]{P} 0.$$

Together, Theorems 3.1 and 3.2 separate the roles of pretraining scale and model capacity. Increasing the number of pretraining tasks reduces the fitting error within a fixed model class, whereas eliminating the approximation error requires an increasingly expressive model class.

3.3 Large-Sample Causal Target

We now study the large-sample target of the prior-induced graph posterior. Let $X_0^{(n)} = (X_i)_{i=1}^n$, where $X_i \overset{\text{i.i.d.}}{\sim} P_0$; P_0 has DAG G_0 but may lie outside the pretraining model family. Define

$$\mathcal{G}_\Pi = \{G \in \mathcal{G}_d : \Pi(G) > 0\}, \qquad \mathcal{P}_G = \{P_{G,\eta} : \eta \in \mathrm{supp}(\Pi(\cdot \mid G))\}.$$

Thus, $\mathcal{P}_G$ contains the observational distributions that graph G can express under the pretraining prior. For each $G \in \mathcal{G}_\Pi$, define

$$K_0(G) = \inf_{P\in\mathcal{P}_G} \mathrm{KL}(P_0\|P), \qquad \mathcal{E}_0 = \arg\min_{G\in\mathcal{G}_\Pi} K_0(G).$$

The *KL-best-fitting graph set* $\mathcal{E}_0$ therefore contains the prior-supported graphs whose best available mechanisms most closely approximate P_0. The convergence below requires the graph-wise Hellinger compactness condition (Assumption A.7), which also guides the construction of the pretraining distribution. Appendix A.2 states and discusses the assumptions used below.

Theorem 3.3 (Concentration on the KL-best-fitting graph set)**.** *Suppose that Assumptions A.5, A.7, A.8, and A.9 hold. Then*

$$P_{\Pi,n}\left(\mathcal{E}_0 \mid X_0^{(n)}\right) \longrightarrow 1 \qquad P_0^\infty\text{-almost surely}.$$

Theorem 3.3 identifies $\mathcal{E}_0$, rather than necessarily a unique DAG, as the general large-sample recovery target. Under model misspecification, it contains the prior-supported graphs whose associated distribution families best approximate P_0.

Under correct specification, the KL-best-fitting graph set has a more direct causal interpretation.

Proposition 3.4 (Characterization under correct specification)**.** *Suppose that the conditions of Theorem 3.3 hold and that $P_0 \in \mathcal{P}_{G_0}$ for some $G_0 \in \mathcal{G}_\Pi$. Then $\min_{G\in\mathcal{G}_\Pi} K_0(G) = 0, \mathcal{E}_0 = \{G \in \mathcal{G}_\Pi : P_0 \in \mathcal{P}_G\}$.*

Under correct specification, $\mathcal{E}_0$ consists exactly of the prior-supported graphs that can represent the true observational distribution. Standard causal assumptions determine whether this set is a Markov equivalence class or a single DAG. We write $[G_0]_{\mathrm{MEC}}$ for the Markov equivalence class of G_0.

Corollary 3.5 (Recovery of a Markov equivalence class)**.** *Suppose that the conditions of Proposition 3.4 hold and that P_0 is Markov and faithful to G_0. Assume that every prior-supported representation of P_0 is Markov and faithful to its graph, and that*

$$P_0 \in \mathcal{P}_G \qquad \text{for every } G \in \mathcal{G}_\Pi \cap [G_0]_{\mathrm{MEC}}.$$

Then

$$\mathcal{E}_0 = \mathcal{G}_\Pi \cap [G_0]_{\mathrm{MEC}}, \qquad P_{\Pi,n}\left(\mathcal{G}_\Pi \cap [G_0]_{\mathrm{MEC}} \mid X_0^{(n)}\right) \longrightarrow 1 \quad P_0^\infty\text{-almost surely}.$$

If $[G_0]_{\mathrm{MEC}} \subseteq \mathcal{G}_\Pi$, then $\mathcal{E}_0 = [G_0]_{\mathrm{MEC}}$.

Corollary 3.6 (Recovery of a uniquely identifiable DAG)**.** *Suppose that the conditions of Proposition 3.4 hold and that P_0 is uniquely graph-identifiable within the support of the pretraining prior:*

$$P_0 \notin \mathcal{P}_G \qquad \text{for every } G \in \mathcal{G}_\Pi \setminus \{G_0\}.$$

Then

$$\mathcal{E}_0 = \{G_0\}, \qquad P_{\Pi,n}\left(G_0 \mid X_0^{(n)}\right) \longrightarrow 1 \qquad P_0^\infty\text{-almost surely}.$$

Implications for existing models. Our framework explains why existing causal discovery foundation models can perform well even when their pretraining priors are designed empirically, and clarifies their limitations: pretraining-family compactness (A.7) and model-class posterior expressiveness (A.4) are not guaranteed by empirical design. See Appendix C.3 for detailed analysis.

Identifiable families and LiNGAM. Our framework also covers pretraining priors constructed from identifiable model families (Qiao et al., 2026; Chen et al., 2026). As a concrete example, suppose that the pretraining prior is supported only on linear non-Gaussian acyclic models (LiNGAMs) (Shimizu et al., 2006). If $P_0 \in \mathcal{P}_{G_0}$ for some $G_0 \in \mathcal{G}_\Pi$, LiNGAM identifiability implies

$$P_0 \in \mathcal{P}_G \quad \Longleftrightarrow \quad G = G_0 \qquad \text{for every } G \in \mathcal{G}_\Pi.$$

Consequently,

$$\mathcal{E}_0 = \{G_0\}, \qquad P_{\Pi,n}\left(G_0 \mid X_0^{(n)}\right) \longrightarrow 1 \quad P_0^\infty\text{-almost surely.}$$

This provides a concrete instance in which an identifiable pretraining family makes the true DAG the large-sample target of the graph posterior. The corresponding regularity and prior-support conditions are discussed in Appendix C.4.

3.4 From Pretraining to Causal Recovery

We finally connect posterior learning with large-sample causal recovery.

Theorem 3.7 (Iterated-limit consistency of the learned graph posterior)**.** *Suppose that the conditions of Theorem 3.3 hold and that, for every fixed n, the conditions of Theorem 3.2 hold. Let $r_{M,n}$ be the corresponding capacity sequence, and write $P^X_{\Pi,n}(B) := P_{\Pi,n}\left(X^{(n)} \in B\right)$. Assume additionally that a fresh test dataset is independent of pretraining and that $P_0^n \ll P^X_{\Pi,n}$ for every fixed n. Then, for every $\epsilon > 0$,*

$$\lim_{n\to\infty} \limsup_{M\to\infty} \Pr\left(\mathbb{E}_{X_0^{(n)}\sim P_0^n}\left[1 - q_{\widehat{\theta}_{M,r_{M,n},n}}\left(\mathcal{E}_0 \mid X_0^{(n)}\right)\right] > \epsilon\right) = 0,$$

where the probability is over the pretraining tasks and training randomness.

Thus, in the iterated limit, the learned graph posterior assigns vanishing expected mass to graphs outside the KL-best-fitting graph set. The complete proof is provided in Appendix D.

4 Method

4.1 Skeleton–Order Parameterization

Given an observational dataset $D = X^{(n)} \in \mathbb{R}^{n\times d}$, our model predicts a distribution over causal DAGs by decomposing each graph into an undirected skeleton and a topological order. Let $S \in \{0,1\}^{d\times d}$ be a symmetric skeleton with zero diagonal, and let π be a permutation of the variables. Writing $r_\pi(j)$ for the position of node j in π, we orient the skeleton by

$$g(S,\pi)_{jk} = S_{jk}\mathbf{1}\left\{r_\pi(j) < r_\pi(k)\right\}. \tag{2}$$

Every directed edge follows the same order, so $g(S,\pi)$ is acyclic by construction. A shared encoder first maps the input dataset to variable representations $H = h_\psi(D) \in \mathbb{R}^{d\times m}$, where ψ denotes the encoder parameters and m is the hidden dimension. Conditioned on H, the induced graph distribution is

$$q_{\theta,\phi}(G \mid H) = \sum_S \sum_\pi q_\phi(S \mid H)\, p_\theta(\pi \mid H)\, \mathbf{1}\left\{G = g(S,\pi)\right\}. \tag{3}$$

Skeleton distribution. The first factor in Eq. (3) models uncertainty over undirected skeletons. A symmetric pairwise decoder produces logits $\ell^S_{jk} = \ell^S_{kj}$ for unordered node pairs. We model the skeleton distribution as

$$q_\phi(S \mid H) = \prod_{1\le j<k\le d} \text{Bernoulli}\left(S_{jk}; \sigma(\ell^S_{jk})\right), \qquad S_{jk} = S_{kj}. \tag{4}$$

The complete encoder and skeleton-decoder architectures are provided in Appendix E.

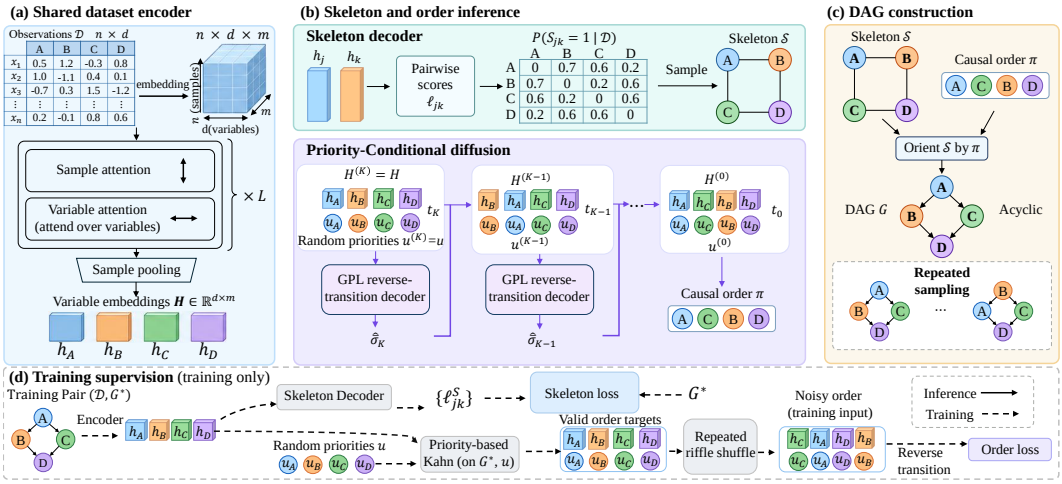


Figure 2: Overview of our proposed model. A shared encoder extracts variable and relation representations from the input dataset. A symmetric decoder predicts an undirected skeleton, while conditional diffusion models a distribution over topological orders. Orienting the sampled skeleton according to a sampled order produces an acyclic graph.

4.2 Conditional Diffusion over Topological Orders

The second factor $p_\theta(\pi \mid H)$ in Eq. (3) represents a distribution over topological orders. Because a DAG may admit multiple valid orders, choosing a fixed canonical order would impose arbitrary supervision. For each training DAG $G^\star$, we sample independent priorities $u_j \sim \mathrm{Uniform}(0,1)$ and construct a unique target order $\pi_0 = \tau(G^\star, u)$ using a priority-based Kahn algorithm (see Appendix E.3). Resampling u generates different valid target orders for the same DAG.

For an order π, define $H_\pi = (H_{\pi(1)}, \ldots, H_{\pi(d)})$ and $u_\pi = (u_{\pi(1)}, \ldots, u_{\pi(d)})$, which arrange the variable representations and priorities according to π. Starting from the target sequence (H_{π_0}, u_{π_0}), the forward process progressively randomizes the variable order through riffle shuffles. At step t,

$$\sigma_t \sim q_{\mathrm{shuf}}, \qquad (H_{\pi_t(i)}, u_{\pi_t(i)}) = (H_{\pi_{t-1}(\sigma_t(i))}, u_{\pi_{t-1}(\sigma_t(i))}), \quad i = 1, \ldots, d.$$

We use a denoising schedule $0 = t_0 < t_1 < \cdots < t_K = T$. For each k, the reverse permutation $\bar{\sigma}_k$ satisfies

$$(H_{\pi_{t_{k-1}}}, u_{\pi_{t_{k-1}}}) = \bar{\sigma}_k \cdot (H_{\pi_{t_k}}, u_{\pi_{t_k}}).$$

Thus, it maps the state at t_k back to that at t_{k-1} and may undo one or more forward shuffles. We model its distribution as

$$p_\theta\left(\bar{\sigma}_k \mid H_{\pi_{t_k}}, u_{\pi_{t_k}}, t_k\right)$$

using a generalized Plackett–Luce (GPL) distribution (Zhang et al., 2025).

Training pairs each corrupted state $(H_{\pi_{t_k}}, u_{\pi_{t_k}})$ with its reverse target $\bar{\sigma}_k$, yielding

$$\mathcal{L}_{\mathrm{ord}} = -\mathbb{E}\left[\frac{1}{K}\sum_{k=1}^{K} \log p_\theta\left(\bar{\sigma}_k \mid H_{\pi_{t_k}}, u_{\pi_{t_k}}, t_k\right)\right]. \tag{5}$$

Composing these conditional reverse transitions allows the model to capture complex, multimodal distributions over topological orders.

For skeleton prediction, $\mathcal{L}_{\mathrm{skel}}$ is the average of the ordinary and class-balanced binary cross-entropy losses. The complete training objective is

$$\mathcal{L} = \lambda_{\mathrm{skel}}\mathcal{L}_{\mathrm{skel}} + \lambda_{\mathrm{ord}}\mathcal{L}_{\mathrm{ord}}.$$

At inference time, we sample u and initialize the terminal state as $(H^{(K)}, u^{(K)}) = (H, u)$. For $k = K, \ldots, 1$, we sample and apply a reverse permutation:

$$\widehat{\bar{\sigma}}_k \sim p_\theta(\cdot \mid H^{(k)}, u^{(k)}, t_k), \qquad (H^{(k-1)}, u^{(k-1)}) = \widehat{\bar{\sigma}}_k \cdot (H^{(k)}, u^{(k)}).$$

Applying the same sequence of permutations to the node labels yields the final topological order π. Detailed constructions of the forward process, GPL reverse transitions, losses, and sampling procedure are provided in Appendices E.3 and E.4.

For pretraining, we construct a diverse synthetic distribution over Erdős–Rényi-type and scale-free DAGs that combines linear-additive, additive-MLP, non-additive-MLP, random Fourier feature (RFF), LiNGAM, and equal-variance linear Gaussian mechanisms with multiple exogenous-noise families. To align the simulator with the compactness condition used in our theory, its continuous parameters are restricted to compact sets; full details are provided in Appendix F.

5 Experiments

5.1 Controlled Evaluation of Permutation Expressivity

A broad pretraining prior designed to cover diverse data-generating processes can induce complex, multimodal posteriors over topological orders, making an expressive order model essential. We evaluate this capacity on five target distributions over four-node topological orders. Three-step GPL diffusion fits the multimodal and structured targets more accurately than the one-step GPL and Gumbel–Sinkhorn baselines. Full details are provided in Appendix J.1.

5.2 Synthetic Causal Discovery Benchmark

Setup. We evaluate causal structure recovery across graph sizes $d \in \{3, 5, 10, 20\}$, sample sizes $n \in \{20, 50, 100, 300, 1000, 2000\}$, and eight test distributions. For every distribution–node-count–sample-size cell, we generate 50 independent datasets, resampling the DAG, structural functions, mechanism parameters, and exogenous noise for each dataset. This gives $8 \times 4 \times 6 = 192$ benchmark settings and $9{,}600$ test datasets per method.

The first six distributions cover mechanism families commonly used in synthetic causal discovery generators. ID-Mix combines linear additive, additive MLP, non-additive MLP, and RFF mechanisms, while Linear, MLP-Additive, MLP-Nonadditive, and RFF isolate the four component families. Linear, neural-network, RFF, and GP-related mechanisms have all been considered in existing causal discovery foundation models (Lorch et al., 2022; Thompson et al., 2026; Li et al., 2026; Blöbaum et al., 2026; Qiao et al., 2026; Chen et al., 2026). GP uses Gaussian-process mechanisms that are absent from our pretraining distribution.

The final two distributions evaluate causal model families for which the DAG is identifiable under their respective standard conditions. LiNGAM uses linear mechanisms with independent non-Gaussian noise (Shimizu et al., 2006), whereas ANM-Gaussian uses nonlinear additive mechanisms with independent Gaussian noise (Hoyer et al., 2008; Peters et al., 2014). Both families have also been considered in prior causal discovery foundation models (Qiao et al., 2026; Chen et al., 2026). The particular ANM generator evaluated here is not included in our pretraining distribution. Definitions of all eight test SEM families and their connections to existing synthetic generators are provided in Appendix G.1.

We compare our method with six causal discovery foundation models and eight classical methods. The DAG-level comparison includes six foundation-model baselines and six directed-output classical methods, for 13 methods including ours. Because PC and GES natively output CPDAGs, they are reported only in the CPDAG-level comparison in Appendix H.4. We report directed-edge F1 and normalized structural Hamming distance (nSHD), as defined in Appendix G.2, macro-averaged over all 192 benchmark settings. Because every cell contains 50 datasets, this is also the equal-weight average over all $9{,}600$ test datasets. Baseline implementation details are provided in Appendix K.

Results. Our method achieves the highest overall DAG F1 (0.640) and the lowest DAG nSHD (0.217). Compared with TabCausal, the strongest external foundation-model baseline in DAG F1, it improves F1 from 0.559 to 0.640 and reduces nSHD from 0.257 to 0.217. It also produces acyclic graphs on all test datasets by construction.

The stratified results show broad performance gains across test conditions. Across the eight test distributions, our method achieves the highest DAG F1 on six and the lowest DAG nSHD on five.

Table 1: Macro-averaged DAG recovery over 192 benchmark settings and 9,600 test datasets.

Foundation models			Classical methods		
Method	F1 ↑	nSHD ↓	Method	F1 ↑	nSHD ↓
Ours	**0.640**	**0.217**	SCORE	0.507	0.284
TabCausal	0.559	0.257	DAS	0.488	0.271
Arrow	0.547	0.268	ICA-LiNGAM	0.485	0.301
FoundCause	0.518	0.321	Direct-LiNGAM	0.480	0.303
CDFM	0.504	0.472	DAGMA-linear	0.438	0.306
AVICI	0.391	0.370	NOTEARS-linear	0.412	0.309
CauScale	0.294	0.391			

It is best under both metrics on ID-MIX, both MLP families, RFF, and ANM-GAUSSIAN, and additionally achieves the highest F1 on the GP distribution. Across observational sample sizes, it achieves the highest F1 and lowest nSHD at every $n \in \{20, 50, 100, 300, 1000, 2000\}$. Across graph sizes, it is best under both metrics for $d \in \{3, 5, 10\}$, while remaining competitive at $d = 20$, with a DAG F1 of 0.455 and an nSHD of 0.197. Detailed results and analyses stratified by test distribution, graph size, and observational sample size are provided in Appendices H.1– H.3.

To assess stability at a larger scale, we evaluate the seven foundation models on $1{,}000$ datasets in each of the 192 benchmark cells (192,000 per method). Our method again attains the highest mean DAG F1 and lowest mean nSHD; detailed means and dataset-level standard deviations are reported in Appendix H.5.

5.3 IDENTIFIABLE RECOVERY AND REAL-WORLD EVALUATION

Recovery under an identifiable prior. When the learned graph distribution is a sufficiently accurate approximation to the prior-induced graph posterior, correct specification and identifiability lead to concentration on the true DAG as the number of observations increases. We use a controlled LiNGAM experiment to examine whether this theoretically predicted behavior can be realized by our finite pretrained model in practice. We pretrain a $d = 5$ model on $6{,}000{,}000$ independently generated LiNGAM datasets. As the test sample size increases from 20 to $2{,}000$, DAG F1 rises monotonically from 0.597 to 0.989, while DAG nSHD decreases from 0.2980 to 0.0145. At large sample sizes, its performance approaches that of Direct-LiNGAM and ICA-LiNGAM. This progression toward near-perfect recovery is consistent with the large-sample behavior predicted by our theory. Complete settings and results are provided in Appendix J.3.

Real-world evaluation. We evaluate our model on the Sachs protein-signaling dataset, which contains 853 observational samples of 11 phosphorylated proteins and phospholipids, together with a 17-edge reference DAG (Sachs et al., 2005). Our method achieves the best nSHD of 0.2727 among all evaluated methods. Our method and TabCausal also outperform every directed-output classical baseline on this dataset. The complete comparison is provided in Appendix I.

6 LIMITATIONS AND CONCLUSION

We establish a unified statistical framework for analyzing causal discovery foundation models. We characterize both the graph posterior learned through pretraining and its large-sample causal target, and instantiate this framework with a skeleton–order model using conditional diffusion. Experiments demonstrate strong structure recovery and the predicted large-sample behavior under an identifiable LiNGAM prior. Several limitations remain. Even under the regularity conditions of our concentration theorem, the KL-best-fitting graph set under misspecification need not coincide with the true DAG or its Markov equivalence class. Designing broader pretraining priors that satisfy the theoretical conditions is therefore an important direction for future work. In addition, the current model represents the skeleton and topological order separately, and jointly modeling them may increase the expressive capacity of the graph posterior. Finally, owing to computational constraints, the model is pretrained only on graphs with 3 to 25 nodes; extending it to substantially larger graphs is left for future work.

AI USE STATEMENT

In this work, we used generative AI tools to help develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, assist in the writing of proofs, propose or refine hypotheses, design or provide feedback on research methodology or experiments, implement methods, assist with translation, and interpret results. We have not used generative AI tools to generate synthetic datasets or to clean and reformat datasets, and qualitative and thematic data analysis is not applicable to this work. Additionally, we used generative AI tools to draft parts of the paper, discover research topics and search for information, and edit the paper to improve readability. We have reviewed all AI-assisted work. We checked the identified literature and citations against their original sources and independently verified all AI-assisted mathematical claims, proof arguments, methodological suggestions, and interpretations of results. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We support the reproducibility of both our theoretical and empirical results through detailed documentation in the appendix. The assumptions underlying the graph-posterior learning and large-sample concentration results are stated in Appendices A.1 and A.2, with complete proofs provided in Appendices B, C, and D. The model architecture, topological-order target construction, diffusion transitions, training objective, and posterior-sampling procedure are described in Appendix E, while the implementation configuration, optimization settings, and computational resources are reported in Appendix E.5. The graph, mechanism, noise, and parameter distributions used for pretraining are specified in Appendix F. The construction of the synthetic benchmark and the definitions of the evaluation metrics are provided in Appendix G, with detailed results reported in Appendix H. The experimental setups and results for the permutation-expressivity study, pretraining-scaling study, identifiable LiNGAM experiment, and Sachs evaluation are provided in Appendices J.1, J.2, J.3, and I, respectively.

ETHICS STATEMENT

This work develops statistical theory and learning methodology for causal discovery. The experiments use synthetic data and publicly available benchmark datasets. We did not collect new data from human participants or access personally identifiable information or private data. Therefore, we do not identify any significant ethical concerns associated with this work.

REFERENCES

Yashas Annadani, Nick Pawlowski, Joel Jennings, Stefan Bauer, Cheng Zhang, and Wenbo Gong. Bayesdag: Gradient-based posterior inference for causal discovery. *Advances in Neural Information Processing Systems*, 36:1738–1763, 2023.

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

Patrick Blöbaum, Krishnakumar Balasubramanian, and Shiva Prasad Kasiviswanathan. FoundCause: Causal discovery with latent confounders from observational data. *arXiv preprint arXiv:2606.17516*, 2026.

Yikang Chen, Zhengkang Guan, Haoyuan Qian, Xingxuan Zhang, Peng Cui, Yi Yang, Fei Wu, and Kun Kuang. DAG-FM: A foundation model for causal discovery under heterogeneous causal mechanisms. *arXiv preprint arXiv:2607.11510*, 2026. URL https://arxiv.org/abs/2607.11510.

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

Chris Cundy, Aditya Grover, and Stefano Ermon. Bcd nets: Scalable variational approaches for bayesian causal discovery. *Advances in Neural Information Processing Systems*, 34:7095–7110, 2021.

Pierpaolo De Blasi and Stephen G. Walker. Bayesian asymptotics with misspecified models. *Statistica Sinica*, 23:169–187, 2013. doi: 10.5705/ss.2010.239.

Anish Dhir, Matthew Ashman, James Requeima, and Mark van der Wilk. A meta-learning approach to bayesian causal discovery. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=eeJz7eDWKO.

Nir Friedman and Daphne Koller. Being bayesian about network structure. a bayesian approach to structure discovery in bayesian networks. *Machine learning*, 50(1):95–125, 2003.

David Heckerman, Dan Geiger, and David M Chickering. Learning bayesian networks: The combination of knowledge and statistical data. *Machine learning*, 20(3):197–243, 1995.

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 21*, pp. 689–696. Curran Associates, Inc., 2008. URL https://proceedings.neurips.cc/paper/2008/hash/f7664060cc52bc6f3d620bcedc94a4b6-Abstract.html.

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

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

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

R. Duncan Luce. *Individual Choice Behavior: A Theoretical Analysis*. Wiley, 1959.

Colin L Mallows. Non-null ranking models. i. *Biometrika*, 44(1/2):114–130, 1957.

Gonzalo E. Mena, David Belanger, Scott Linderman, and Jasper Snoek. Learning latent permutations with gumbel-sinkhorn networks. In *International Conference on Learning Representations*, 2018.

Francesco Montagna, Nicoletta Noceti, Lorenzo Rosasco, Kun Zhang, and Francesco Locatello. Scalable causal discovery with score matching. In *Proceedings of the Second Conference on Causal Learning and Reasoning*, volume 213 of *Proceedings of Machine Learning Research*, pp. 752–771. PMLR, 2023.

Francesco Montagna, Max Cairney-Leeming, Dhanya Sridhar, and Francesco Locatello. Demystifying amortized causal discovery with transformers. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL https://openreview.net/forum?id=9Lgy7IGSfp.

Thomas Nagler. Statistical foundations of prior-data fitted networks. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 25660–25676. PMLR, 2023.

Whitney K. Newey and Daniel McFadden. Large sample estimation and hypothesis testing. In Robert F. Engle and Daniel McFadden (eds.), *Handbook of Econometrics*, volume 4, chapter 36, pp. 2111–2245. Elsevier, 1994. doi: 10.1016/S1573-4412(05)80005-4.

Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2nd edition, 2009.

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, Joris M. Mooij, Dominik Janzing, and Bernhard Schölkopf. Causal discovery with continuous additive noise models. *Journal of Machine Learning Research*, 15(58):2009–2053, 2014. URL https://jmlr.org/papers/v15/peters14a.html.

Robin L Plackett. The analysis of permutations. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 24(2):193–202, 1975.

Jie Qiao, Ruichu Cai, Zijian Li, Weilin Chen, Pengfei Hua, Boyan Xu, Zhengming Chen, Zhifeng Hao, and Peng Cui. CDFM: Towards a general-purpose causal discovery foundation model. *arXiv preprint arXiv:2607.11508*, 2026.

Paul Rolland, Volkan Cevher, Matthäus Kleindessner, Chris Russell, Dominik Janzing, Bernhard Schölkopf, and Francesco Locatello. Score matching enables causal discovery of nonlinear additive noise models. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 18741–18753. PMLR, 2022.

Karen Sachs, Omar Perez, Dana Pe'er, Douglas A. Lauffenburger, and Garry P. Nolan. Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, 308(5721): 523–529, 2005. doi: 10.1126/science.1105809.

Pedro Sanchez, Xiao Liu, Alison Q. O'Neil, and Sotirios A. Tsaftaris. Diffusion models for causal discovery via topological ordering. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=Idusfje4-Wq.

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(72):2003–2030, 2006. URL https://jmlr.org/papers/v7/shimizu06a.html.

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(33):1225–1248, 2011.

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

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

Tim van Erven and Peter Harremoës. Rényi divergence and kullback–leibler divergence. *IEEE Transactions on Information Theory*, 60(7):3797–3820, 2014.

Thomas Verma and Judea Pearl. Equivalence and synthesis of causal models. In *Proceedings of the Sixth Annual Conference on Uncertainty in Artificial Intelligence*, pp. 255–270, 1990.

Menghua Wu, Yujia Bao, Regina Barzilay, and Tommi Jaakkola. Sample, estimate, aggregate: A recipe for causal discovery foundation models. *Transactions on Machine Learning Research*, 2025.

Wenbo Xu, Yue He, Yunhai Wang, Xingxuan Zhang, Kun Kuang, Yueguo Chen, and Peng Cui. A knowledge-informed pretrained model for causal discovery. *arXiv preprint arXiv:2603.20842*, 2026.

Naiyu Yin, Tian Gao, and Yue Yu. Learning causal graphs at scale: A foundation model approach. *arXiv preprint arXiv:2506.18285*, 2025.

Yongxing Zhang, Donglin Yang, and Renjie Liao. Symmetricdiffusers: Learning discrete diffusion on finite symmetric groups. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=EO8xpnW7aX.

Xun Zheng, Bryon Aragam, Pradeep K Ravikumar, and Eric P Xing. Dags with no tears: Continuous optimization for structure learning. *Advances in neural information processing systems*, 31, 2018.

A Technical Assumptions and Auxiliary Results

A.1 Assumptions for Learning the Target Graph Posterior

Assumption A.1 (Independent pretraining tasks)**.** For each M and fixed n, the pretraining tasks satisfy

$$\left\{\left(G_m, X_m^{(n)}\right)\right\}_{m=1}^{M} \overset{\text{i.i.d.}}{\sim} P_{\Pi,n}.$$

Assumption A.2 (Finite graph log-loss)**.** For every fixed dataset size n and model capacity r, every $\theta \in \Theta_r$, and every $G \in \mathcal{G}_d$,

$$P_{\Pi,n}\left(G \mid X^{(n)}\right) > 0 \quad \Longrightarrow \quad q_\theta\left(G \mid X^{(n)}\right) > 0$$

for almost every $X^{(n)}$ under the dataset marginal of $P_{\Pi,n}$. Moreover,

$$R_{r,n}(\theta) < \infty.$$

The first condition is automatically satisfied by standard finite-logit discrete decoders, such as softmax- or sigmoid-based outputs, which assign strictly positive probability to every graph. The second is a mild integrability requirement and holds, for example, when the model logits are bounded or, more generally, have finite expectation under the pretraining distribution.

Assumption A.3 (Uniform law of large numbers for the pretraining risk)**.** For every fixed model capacity r and dataset size n, define

$$\Delta_{M,r,n} := \sup_{\theta \in \Theta_r} \left| \widehat{R}_{M,r,n}(\theta) - R_{r,n}(\theta) \right|.$$

Then

$$\Delta_{M,r,n} \xrightarrow[M\to\infty]{P} 0.$$

Assumption A.3 is the standard fixed-class uniform law of large numbers used in consistency analyses of empirical risk minimizers and, more generally, M-estimators. For example, under i.i.d. pretraining tasks, it follows from familiar regularity conditions such as compactness of Θ_r, continuity of the loss in θ almost surely, and the existence of an integrable envelope for the loss class; see Newey & McFadden (1994, Lemma 2.4). We state the uniform convergence property directly because verifying these sufficient conditions for a particular neural-network architecture and an unbounded log-loss requires additional model-specific analysis.

Assumption A.4 (Posterior expressiveness)**.** For every fixed n, the sequence of graph-posterior model classes is sufficiently expressive for the target graph posterior:

$$\varepsilon_{\text{app}}(r, n) \longrightarrow 0 \qquad \text{as } r \to \infty.$$

Equivalently, for every $\epsilon > 0$, there exists a capacity level $R_\epsilon < \infty$ such that, for every $r \geq R_\epsilon$, there exists a parameter $\theta_r \in \Theta_r$ satisfying

$$\mathbb{E}_{X^{(n)} \sim P_{\Pi,n}}\left[\mathrm{KL}\left(P_{\Pi,n}\left(\cdot \mid X^{(n)}\right) \,\|\, q_{\theta_r}\left(\cdot \mid X^{(n)}\right)\right)\right] < \epsilon.$$

Intuitively, this assumption requires that, as model capacity increases, the model can approximate the target graph posterior arbitrarily well on average over datasets generated by the pretraining distribution.

A.2 Assumptions for Large-Sample Posterior Concentration

For each $G \in \mathcal{G}_\Pi$, let

$$\mathcal{H}_G := \operatorname{supp}(\Pi(\cdot \mid G)) .$$

Let

$$\Omega := \bigsqcup_{G \in \mathcal{G}_\Pi} (\{G\} \times \mathcal{H}_G)$$

denote the prior-supported SCM space.

Assumption A.5 (Dominated and measurable causal model)**.** There exists a common σ-finite measure μ such that every distribution $P_{G,\eta}$, with $G \in \mathcal{G}_\Pi$ and $\eta \in \mathcal{H}_G$, admits a density

$$p_{G,\eta} := \frac{dP_{G,\eta}}{d\mu}.$$

The true distribution P_0 also admits a density p_0 with respect to μ.

The map

$$T : \Omega \longrightarrow \mathcal{F}, \qquad T(G, \eta) = p_{G,\eta},$$

is measurable, where

$$\mathcal{F}_G := \{p_{G,\eta} : \eta \in \mathcal{H}_G\} , \qquad \mathcal{F} := \bigcup_{G \in \mathcal{G}_\Pi} \mathcal{F}_G.$$

All posterior integrals and sets appearing below are assumed measurable.

For densities $p, q \in \mathcal{F}$, define the Hellinger distance by

$$H(p, q) := \left[\int \left(\sqrt{p(x)} - \sqrt{q(x)} \right)^2 \mu(dx) \right]^{1/2}.$$

Throughout, two densities that are equal μ-almost everywhere are identified as the same element of the density space. Accordingly, the Hellinger distance is a genuine metric on $\mathcal{F}$.

Lemma A.6 (Lower semicontinuity of the KL divergence)**.** *Let p_0 be a probability density with respect to μ, and let $p_1, p_2, \ldots, p$ be probability densities with respect to the same measure. If*

$$H(p_m, p) \longrightarrow 0,$$

then

$$\mathrm{KL}(p_0 \| p) \le \liminf_{m \to \infty} \mathrm{KL}(p_0 \| p_m),$$

where the KL divergence is allowed to take the value $+\infty$.

Proof. Let P_m, P, and P_0 denote the probability measures with densities p_m, p, and p_0, respectively, with respect to μ. For every measurable set A,

$$\begin{aligned} |P_m(A) - P(A)| &= \left| \int_A (p_m - p)\, d\mu \right| \\ &\le \int |p_m - p|\, d\mu \\ &= \int |\sqrt{p_m} - \sqrt{p}|\, (\sqrt{p_m} + \sqrt{p})\, d\mu \\ &\le H(p_m, p) \left[\int (\sqrt{p_m} + \sqrt{p})^2\, d\mu \right]^{1/2}, \end{aligned}$$

where the final inequality follows from the Cauchy–Schwarz inequality. Moreover,

$$\begin{aligned} \int (\sqrt{p_m} + \sqrt{p})^2\, d\mu &= 2 + 2 \int \sqrt{p_m p}\, d\mu \\ &\le 4, \end{aligned}$$

because

$$\int \sqrt{p_m p}\, d\mu \le \left(\int p_m\, d\mu\right)^{1/2} \left(\int p\, d\mu\right)^{1/2} = 1.$$

Consequently,

$$|P_m(A) - P(A)| \le 2H(p_m, p) \longrightarrow 0$$

for every measurable set A. Thus,

$$P_m \longrightarrow P$$

setwise.

By van Erven & Harremoës (2014, Theorem 15), Rényi divergence of every order $\alpha \in (0, \infty]$ is lower semicontinuous in the pair of probability measures under setwise convergence. Taking $\alpha = 1$, for which Rényi divergence equals the KL divergence, and observing that P_0 is fixed, we obtain

$$\mathrm{KL}(P_0 \| P) \le \liminf_{m\to\infty} \mathrm{KL}(P_0 \| P_m).$$

Expressing these divergences in terms of their densities proves the claim. □

Assumption A.7 (Graph-wise Hellinger compactness)**.** For every $G \in \mathcal{G}_\Pi$, the density family

$$(\mathcal{F}_G, H)$$

is compact.

A sufficient condition is that $\mathcal{H}_G$ is compact and that the parameter-to-density map is Hellinger continuous:

$$\eta_m \longrightarrow \eta \quad \Longrightarrow \quad H\left(p_{G,\eta_m}, p_{G,\eta}\right) \longrightarrow 0.$$

Here, Hellinger continuity means that small changes in the parameters lead to small changes in the induced observational distributions under the Hellinger metric. For example, in a linear model with independent Laplace noise, the sufficient condition above holds when the coefficients are restricted to closed, bounded intervals and the noise scales are restricted to a compact interval contained in $(0, \infty)$. This also guides the construction of pretraining distributions: restricting parameter ranges to compact sets and ensuring Hellinger continuity of the parameter-to-density map yields model families satisfying this assumption.

Assumption A.8 (Finite optimal KL divergence)**.** The optimal KL value over the prior-supported model family is finite:

$$K_0^\star := \inf_{p\in\mathcal{F}} \mathrm{KL}(p_0 \| p) = \min_{G\in\mathcal{G}_\Pi} K_0(G) < \infty.$$

This assumption only requires that the pretraining model family contain at least one candidate distribution with finite KL divergence from the true distribution. It neither requires the model family to represent the true distribution exactly nor requires every candidate model to have finite KL divergence from it.

Let

$$\overline{\Pi} := \Pi \circ T^{-1}$$

denote the density prior induced by the pretraining SCM prior.

Assumption A.9 (Prior support near the KL optimum)**.** For every $\epsilon > 0$,

$$\overline{\Pi}\left(\{p \in \mathcal{F} : \mathrm{KL}(p_0 \| p) \le K_0^\star + \epsilon\}\right) > 0.$$

This is a standard assumption on prior support. Bayesian updating reweights the prior using the observed data, so a set of models with zero prior probability also has zero posterior probability. Thus, for the posterior to concentrate on models with near-optimal KL divergence, the prior must assign positive probability, for every $\epsilon > 0$, to the set of models whose KL divergence is less than ϵ above the optimum. This condition is satisfied, for example, when a graph attaining the optimal KL value has positive prior probability, the conditional prior on its mechanism parameters is uniform over a box containing an optimal parameter, and the KL divergence is continuous at that parameter.

B Proofs for Learning the Target Graph Posterior

B.1 Proof of Theorem 3.1

Proof. By the law of iterated expectations,

$$R_{r,n}(\theta) = \mathbb{E}_{X^{(n)}\sim P_{\Pi,n}} \mathbb{E}_{G\sim P_{\Pi,n}(\cdot\mid X^{(n)})} \left[-\log q_\theta \left(G \mid X^{(n)} \right) \right].$$

Adding and subtracting $-\log P_{\Pi,n}(G \mid X^{(n)})$ yields

$$\begin{aligned} R_{r,n}(\theta) = {} & H_{\Pi,n}\left(G \mid X^{(n)} \right) \\ & + \mathbb{E}_{X^{(n)}\sim P_{\Pi,n}} \left[\mathrm{KL}\left(P_{\Pi,n}\left(\cdot \mid X^{(n)} \right) \parallel q_\theta \left(\cdot \mid X^{(n)} \right) \right) \right], \end{aligned}$$

where

$$H_{\Pi,n}\left(G \mid X^{(n)} \right) := \mathbb{E}_{X^{(n)}\sim P_{\Pi,n}} \mathbb{E}_{G\sim P_{\Pi,n}(\cdot\mid X^{(n)})} \left[-\log P_{\Pi,n}\left(G \mid X^{(n)} \right) \right]$$

does not depend on θ. Since $\mathcal{G}_d$ is finite,

$$0 \le H_{\Pi,n}\left(G \mid X^{(n)} \right) \le \log |\mathcal{G}_d| < \infty.$$

Therefore, under Assumption A.2, all terms in the preceding decomposition are finite.

It follows that

$$\inf_{\theta\in\Theta_r} R_{r,n}(\theta) = H_{\Pi,n}\left(G \mid X^{(n)} \right) + \varepsilon_{\mathrm{app}}(r,n).$$

Conditional on the pretraining sample and any training randomness, $\widehat{\theta}_{M,r,n} \in \Theta_r$ is fixed. Applying the same decomposition to this learned parameter gives

$$\begin{aligned} R_{r,n}\left(\widehat{\theta}_{M,r,n} \right) = {} & H_{\Pi,n}\left(G \mid X^{(n)} \right) \\ & + \mathbb{E}_{X^{(n)}\sim P_{\Pi,n}} \left[\mathrm{KL}\left(P_{\Pi,n}\left(\cdot \mid X^{(n)} \right) \parallel q_{\widehat{\theta}_{M,r,n}}\left(\cdot \mid X^{(n)} \right) \right) \right]. \end{aligned}$$

Subtracting the preceding two identities gives

$$\begin{aligned} & \mathbb{E}_{X^{(n)}\sim P_{\Pi,n}} \left[\mathrm{KL}\left(P_{\Pi,n}\left(\cdot \mid X^{(n)} \right) \parallel q_{\widehat{\theta}_{M,r,n}}\left(\cdot \mid X^{(n)} \right) \right) \right] \\ & \quad = \varepsilon_{\mathrm{app}}(r,n) + \varepsilon_{\mathrm{fit}}(M,r,n). \end{aligned}$$

It remains to control the fitting error. For every $\delta > 0$, choose $\theta_\delta \in \Theta_r$ such that

$$R_{r,n}(\theta_\delta) \le \inf_{\theta\in\Theta_r} R_{r,n}(\theta) + \delta.$$

By the definition of $\Delta_{M,r,n}$,

$$R_{r,n}\left(\widehat{\theta}_{M,r,n} \right) \le \widehat{R}_{M,r,n}\left(\widehat{\theta}_{M,r,n} \right) + \Delta_{M,r,n}.$$

By the M^{-1}-approximate empirical risk minimization property in Eq. (1),

$$\begin{aligned} \widehat{R}_{M,r,n}\left(\widehat{\theta}_{M,r,n} \right) & \le \inf_{\theta\in\Theta_r} \widehat{R}_{M,r,n}(\theta) + \frac{1}{M} \\ & \le \widehat{R}_{M,r,n}(\theta_\delta) + \frac{1}{M}. \end{aligned}$$

Applying the definition of $\Delta_{M,r,n}$ once more gives

$$\begin{aligned} R_{r,n}\left(\widehat{\theta}_{M,r,n} \right) & \le R_{r,n}(\theta_\delta) + 2\Delta_{M,r,n} + \frac{1}{M} \\ & \le \inf_{\theta\in\Theta_r} R_{r,n}(\theta) + \delta + 2\Delta_{M,r,n} + \frac{1}{M}. \end{aligned}$$

Since $\delta > 0$ is arbitrary,

$$0 \le \varepsilon_{\mathrm{fit}}(M, r, n) \le 2\Delta_{M,r,n} + \frac{1}{M}. \tag{6}$$

This proves the finite-sample bound Eq. (6). If Assumption A.3 additionally holds, then

$$\Delta_{M,r,n} \xrightarrow[M\to\infty]{P} 0.$$

Since $M^{-1} \to 0$, the bound implies

$$\varepsilon_{\mathrm{fit}}(M, r, n) \xrightarrow[M\to\infty]{P} 0.$$

This proves the stated convergence of the fitting error. Together with the preceding risk decomposition, it also shows that, for fixed r and n, the expected posterior KL error converges in probability to $\varepsilon_{\mathrm{app}}(r, n)$. □

B.2 Proof of Theorem 3.2

Proof. Let $\mathbb{P}_{\mathrm{tr}}$ denote probability with respect to the pretraining tasks and any training randomness. By Assumption A.4, for every integer $k \ge 1$, there exists a capacity r_k such that

$$\varepsilon_{\mathrm{app}}(r_k, n) \le \frac{1}{k}. \tag{7}$$

Because the model classes are nested, the sequence $(r_k)_{k\ge1}$ may be chosen to be strictly increasing.

For every fixed k, Theorem 3.1 applied to the fixed model class $\mathcal{Q}_{r_k}$ gives

$$\varepsilon_{\mathrm{fit}}(M, r_k, n) \xrightarrow[M\to\infty]{P} 0.$$

Hence, there exists a deterministic integer M_k such that, for every $M \ge M_k$,

$$P\left(\varepsilon_{\mathrm{fit}}(M, r_k, n) > \frac{1}{k}\right) \le \frac{1}{k}. \tag{8}$$

The sequence $(M_k)_{k\ge1}$ may be chosen strictly increasing and such that $M_k \ge k$.

Define

$$r_M := r_k \qquad \text{whenever } M_k \le M < M_{k+1}. \tag{9}$$

For the finitely many $M < M_1$, set $r_M = r_1$. By construction, r_M is nondecreasing and

$$r_M \longrightarrow \infty.$$

Let $k = k(M)$ denote the index satisfying

$$M_k \le M < M_{k+1}.$$

The risk decomposition from Theorem 3.1 gives

$$\begin{aligned}\mathbb{E}_{X^{(n)}\sim P_{\Pi,n}}\left[\mathrm{KL}\left(P_{\Pi,n}\left(\cdot \mid X^{(n)}\right) \,\|\, q_{\widehat{\theta}_{M,r_M,n}}\left(\cdot \mid X^{(n)}\right)\right)\right] \\ = \varepsilon_{\mathrm{app}}(r_k, n) + \varepsilon_{\mathrm{fit}}(M, r_k, n).\end{aligned}$$

By Eq. (7),

$$\varepsilon_{\mathrm{app}}(r_k, n) \le \frac{1}{k}.$$

Therefore,

$$\begin{aligned}P\left(\mathbb{E}_{X^{(n)}\sim P_{\Pi,n}}\left[\mathrm{KL}\left(P_{\Pi,n}\left(\cdot \mid X^{(n)}\right) \,\|\, q_{\widehat{\theta}_{M,r_M,n}}\left(\cdot \mid X^{(n)}\right)\right)\right] > \frac{2}{k}\right) \\ \le P\left(\varepsilon_{\mathrm{fit}}(M, r_k, n) > \frac{1}{k}\right) \le \frac{1}{k}.\end{aligned}$$

As $M \to \infty$, the corresponding index $k(M) \to \infty$. Consequently,

$$\mathbb{E}_{X^{(n)} \sim P_{\Pi,n}} \left[\mathrm{KL} \left(P_{\Pi,n} \left(\cdot \mid X^{(n)} \right) \parallel q_{\widehat{\theta}_{M,r_M,n}} \left(\cdot \mid X^{(n)} \right) \right) \right] \xrightarrow[M \to \infty]{P} 0.$$

For every $X^{(n)}$, Pinsker's inequality gives

$$\begin{aligned} &\left\| P_{\Pi,n} \left(\cdot \mid X^{(n)} \right) - q_{\widehat{\theta}_{M,r_M,n}} \left(\cdot \mid X^{(n)} \right) \right\|_{\mathrm{TV}} \\ &\qquad \leq \sqrt{\frac{1}{2} \mathrm{KL} \left(P_{\Pi,n} \left(\cdot \mid X^{(n)} \right) \parallel q_{\widehat{\theta}_{M,r_M,n}} \left(\cdot \mid X^{(n)} \right) \right)}. \end{aligned}$$

Taking expectations and applying Jensen's inequality yields

$$\begin{aligned} &\mathbb{E}_{X^{(n)} \sim P_{\Pi,n}} \left[\left\| P_{\Pi,n} \left(\cdot \mid X^{(n)} \right) - q_{\widehat{\theta}_{M,r_M,n}} \left(\cdot \mid X^{(n)} \right) \right\|_{\mathrm{TV}} \right] \\ &\qquad \leq \sqrt{\frac{1}{2} \mathbb{E}_{X^{(n)} \sim P_{\Pi,n}} \left[\mathrm{KL} \left(P_{\Pi,n} \left(\cdot \mid X^{(n)} \right) \parallel q_{\widehat{\theta}_{M,r_M,n}} \left(\cdot \mid X^{(n)} \right) \right) \right]}, \end{aligned}$$

which converges to zero in probability by the preceding KL convergence and the continuous mapping theorem. This proves the total-variation conclusion and completes the proof.

□

C Proofs for the Large-Sample Causal Target

Throughout this section, we write $X^{(n)} := X_0^{(n)}$ for notational simplicity.

C.1 Proof of Theorem 3.3

Proof. We first construct the posterior on observational densities induced by the Bayesian posterior on SCMs.

Let

$$\Pi_n^{\omega} \left(\cdot \mid X^{(n)} \right)$$

denote the posterior distribution on

$$\Omega = \bigsqcup_{G \in \mathcal{G}_{\Pi}} \left(\{G\} \times \mathcal{H}_G \right)$$

induced by the prior Π. Its graph marginal is precisely $P_{\Pi,n}(\cdot \mid X^{(n)})$. Thus, for every $\mathcal{A} \subseteq \mathcal{G}_d$,

$$P_{\Pi,n} \left(\mathcal{A} \mid X^{(n)} \right) = \Pi_n^{\omega} \left(\{(G, \eta) \in \Omega : G \in \mathcal{A}\} \,\middle|\, X^{(n)} \right).$$

Define the induced density posterior by the pushforward

$$\overline{\Pi}_n \left(B \mid X^{(n)} \right) := \Pi_n^{\omega} \left(T^{-1}(B) \mid X^{(n)} \right), \qquad B \subseteq \mathcal{F}.$$

Since the likelihood depends on an SCM only through its observational density, this posterior can be written as

$$\overline{\Pi}_n \left(B \mid X^{(n)} \right) = \frac{\displaystyle\int_B \prod_{i=1}^{n} p\left(X_i\right) \overline{\Pi}(dp)}{\displaystyle\int_{\mathcal{F}} \prod_{i=1}^{n} p\left(X_i\right) \overline{\Pi}(dp)}.$$

Hence $\overline{\Pi}_n$ is the ordinary Bayesian density posterior associated with the induced prior $\overline{\Pi}$.

We next verify that De Blasi & Walker (2013, Corollary 1) can be applied to $\overline{\Pi}_n$.

By Assumption A.7, each $\mathcal{F}_G$ is compact under the Hellinger metric. Since $\mathcal{G}_\Pi$ is a finite subset of $\mathcal{G}_d$,

$$\mathcal{F} = \bigcup_{G \in \mathcal{G}_\Pi} \mathcal{F}_G$$

is a finite union of compact sets and is therefore compact. In particular, $\mathcal{F}$ is Hellinger closed:

$$\overline{\mathcal{F}}^H = \mathcal{F}.$$

Define the pseudo-true density set

$$\mathcal{F}^\star := \{p \in \mathcal{F} : \mathrm{KL}(p_0 \| p) = K_0^\star\}.$$

By Lemma A.6, the map

$$p \longmapsto \mathrm{KL}(p_0 \| p)$$

is lower semicontinuous on the Hellinger-compact set $\mathcal{F}$. Assumption A.8 ensures that its infimum is finite. A lower semicontinuous extended-real-valued function on a compact space attains its infimum whenever that infimum is finite. Therefore, there exists $p^\star \in \mathcal{F}$ such that

$$\mathrm{KL}(p_0 \| p^\star) = K_0^\star,$$

and consequently

$$\mathcal{F}^\star \neq \varnothing.$$

Moreover, since

$$\mathcal{F}^\star = \{p \in \mathcal{F} : \mathrm{KL}(p_0 \| p) \leq K_0^\star\},$$

lower semicontinuity implies that $\mathcal{F}^\star$ is closed. It is therefore a compact subset of $\mathcal{F}$.

For a density p, write

$$H(\mathcal{F}^\star, p) := \inf_{q \in \mathcal{F}^\star} H(q, p).$$

It remains to verify the Hellinger-cover summability condition used by De Blasi & Walker (2013, Corollary 1). Fix any

$$\alpha \in (0, 1/2)$$

and define

$$\varepsilon_\alpha := 2\left(\frac{\alpha^2}{2}\right)^{1/(2\alpha)}.$$

Since $\mathcal{F}$ is compact, it is totally bounded. Hence there exist finitely many Hellinger balls

$$B_{1,\alpha}, \ldots, B_{J(\alpha),\alpha}$$

of radius ε_α whose union covers $\mathcal{F}$:

$$\mathcal{F} \subseteq \bigcup_{j=1}^{J(\alpha)} B_{j,\alpha}.$$

Because this cover is finite,

$$\sum_{j=1}^{J(\alpha)} \overline{\Pi}\left(B_{j,\alpha}\right)^\alpha < \infty.$$

Assumption A.9 supplies the near-optimal KL-support condition. Therefore, De Blasi & Walker (2013, Corollary 1) gives, for every $\epsilon > 0$,

$$\overline{\Pi}_n\left(\{p \in \mathcal{F} : H(\mathcal{F}^\star, p) > \epsilon\} \,\middle|\, X^{(n)}\right) \longrightarrow 0 \qquad P_0^\infty\text{-almost surely.} \tag{10}$$

We now transfer this density-level concentration result to the graph posterior.

Fix any graph

$$G \in \mathcal{G}_\Pi \setminus \mathcal{E}_0.$$

We first show that

$$\mathcal{F}_G \cap \mathcal{F}^\star = \varnothing. \tag{11}$$

Suppose, to the contrary, that there exists

$$p \in \mathcal{F}_G \cap \mathcal{F}^\star.$$

Since $p \in \mathcal{F}^\star$,

$$\mathrm{KL}(p_0 \| p) = K_0^\star.$$

Since $p \in \mathcal{F}_G$,

$$\begin{aligned} K_0(G) &= \inf_{q \in \mathcal{F}_G} \mathrm{KL}(p_0 \| q) \\ &\leq \mathrm{KL}(p_0 \| p) = K_0^\star. \end{aligned}$$

By the definition of $K_0^\star$, we also have

$$K_0(G) \geq K_0^\star.$$

Hence

$$K_0(G) = K_0^\star,$$

which implies $G \in \mathcal{E}_0$, contradicting $G \notin \mathcal{E}_0$. This proves equation 11.

Define

$$\gamma_G := \inf_{\substack{p \in \mathcal{F}_G \\ q \in \mathcal{F}^\star}} H(p, q). \tag{12}$$

Both $\mathcal{F}_G$ and $\mathcal{F}^\star$ are compact, and the map

$$(p, q) \longmapsto H(p, q)$$

is continuous. The infimum in equation 12 is therefore attained.

If $\gamma_G = 0$, there would exist $p_G \in \mathcal{F}_G$ and $q_G \in \mathcal{F}^\star$ such that

$$H(p_G, q_G) = 0.$$

Because Hellinger distance is a metric on probability densities modulo equality μ-almost everywhere, this implies

$$p_G = q_G \qquad \mu\text{-almost everywhere},$$

contradicting equation 11. Therefore,

$$\gamma_G > 0 \qquad \text{for every } G \in \mathcal{G}_\Pi \setminus \mathcal{E}_0.$$

If $\mathcal{E}_0 = \mathcal{G}_\Pi$, then

$$P_{\Pi,n}\left(\mathcal{E}_0 \mid X^{(n)}\right) = 1$$

for every n, and the conclusion is immediate. We therefore assume that $\mathcal{G}_\Pi \setminus \mathcal{E}_0 \neq \varnothing$.

Since $\mathcal{G}_\Pi \setminus \mathcal{E}_0$ is a nonempty finite set, define

$$\gamma := \min_{G \in \mathcal{G}_\Pi \setminus \mathcal{E}_0} \gamma_G > 0.$$

Let

$$A_{\mathrm{bad}} := \{(G, \eta) \in \Omega : G \in \mathcal{G}_\Pi \setminus \mathcal{E}_0\}.$$

For every $(G, \eta) \in A_{\mathrm{bad}}$,

$$p_{G,\eta} \in \mathcal{F}_G$$

for some $G \notin \mathcal{E}_0$, and hence

$$H\left(\mathcal{F}^\star, p_{G,\eta}\right) \geq \gamma_G \geq \gamma.$$

It follows that

$$A_{\mathrm{bad}} \subseteq T^{-1}\left(\left\{p \in \mathcal{F} : H(\mathcal{F}^\star, p) > \frac{\gamma}{2}\right\}\right). \tag{13}$$

Taking posterior probabilities in equation 13 gives

$$\begin{aligned} & P_{\Pi,n}\left(\mathcal{G}_\Pi \setminus \mathcal{E}_0 \mid X^{(n)}\right) \\ &= \Pi_n^\omega\left(A_{\text{bad}} \mid X^{(n)}\right) \\ &\leq \Pi_n^\omega\left(T^{-1}\left\{p : H(\mathcal{F}^\star, p) > \frac{\gamma}{2}\right\} \,\middle|\, X^{(n)}\right) \\ &= \overline{\Pi}_n\left(\left\{p : H(\mathcal{F}^\star, p) > \frac{\gamma}{2}\right\} \,\middle|\, X^{(n)}\right). \end{aligned}$$

By equation 10, the final quantity converges to zero P_0^∞-almost surely. Therefore,

$$P_{\Pi,n}\left(\mathcal{G}_\Pi \setminus \mathcal{E}_0 \mid X^{(n)}\right) \longrightarrow 0 \qquad P_0^\infty\text{-almost surely.}$$

Equivalently,

$$P_{\Pi,n}\left(\mathcal{E}_0 \mid X^{(n)}\right) \longrightarrow 1 \qquad P_0^\infty\text{-almost surely.}$$

□

C.2 Proofs of Proposition 3.4 and Corollaries 3.5–3.6

Proof of Proposition 3.4. Since $P_0 \in \mathcal{P}_{G_0}$,

$$K_0(G_0) = \inf_{P \in \mathcal{P}_{G_0}} \mathrm{KL}(P_0 \| P) = 0.$$

Because the KL divergence is nonnegative,

$$K_0^\star = \min_{G \in \mathcal{G}_\Pi} K_0(G) = 0.$$

For any $G \in \mathcal{G}_\Pi$, if $P_0 \in \mathcal{P}_G$, then

$$K_0(G) = 0 = K_0^\star,$$

and hence $G \in \mathcal{E}_0$.

Conversely, suppose that $G \in \mathcal{E}_0$. Then

$$K_0(G) = K_0^\star = 0.$$

By Assumption A.7, the density family $\mathcal{F}_G$ is compact under the Hellinger metric. By Lemma A.6, the map

$$p \longmapsto \mathrm{KL}(p_0 \| p)$$

is lower semicontinuous on $\mathcal{F}_G$. Because

$$K_0(G) = \inf_{p \in \mathcal{F}_G} \mathrm{KL}(p_0 \| p) = 0 < \infty,$$

this infimum is attained at some $p_G^\star \in \mathcal{F}_G$.

Therefore,

$$\mathrm{KL}(p_0 \| p_G^\star) = 0.$$

The KL divergence is zero if and only if the two densities are equal μ-almost everywhere. Hence

$$p_G^\star = p_0 \qquad \mu\text{-almost everywhere.}$$

Because $p_G^\star \in \mathcal{F}_G$, the corresponding distribution belongs to $\mathcal{P}_G$. Therefore $P_0 \in \mathcal{P}_G$, proving

$$\mathcal{E}_0 = \{G \in \mathcal{G}_\Pi : P_0 \in \mathcal{P}_G\}.$$

□

Proof of Corollary 3.5. By Proposition 3.4,

$$\mathcal{E}_0 = \{G \in \mathcal{G}_\Pi : P_0 \in \mathcal{P}_G\}.$$

For a DAG H, let $\mathcal{I}(H)$ denote the set of conditional independence relations implied by d-separation in H, and let $\mathcal{I}(P)$ denote the set of conditional independence relations that hold under a distribution P.

Let $G \in \mathcal{E}_0$. Then $P_0 \in \mathcal{P}_G$, and hence P_0 is Markov and faithful to G. Since P_0 is also Markov and faithful to G_0,

$$\mathcal{I}(G) = \mathcal{I}(P_0) = \mathcal{I}(G_0).$$

Moreover, $G \in \mathcal{G}_\Pi$ because $\mathcal{E}_0 \subseteq \mathcal{G}_\Pi$. Therefore

$$G \in \mathcal{G}_\Pi \cap [G_0]_{\mathrm{MEC}},$$

which proves

$$\mathcal{E}_0 \subseteq \mathcal{G}_\Pi \cap [G_0]_{\mathrm{MEC}}.$$

Conversely, let

$$G \in \mathcal{G}_\Pi \cap [G_0]_{\mathrm{MEC}}.$$

By the assumed coverage of the prior-supported part of the Markov equivalence class,

$$P_0 \in \mathcal{P}_G.$$

Proposition 3.4 therefore implies $G \in \mathcal{E}_0$. Hence

$$\mathcal{G}_\Pi \cap [G_0]_{\mathrm{MEC}} \subseteq \mathcal{E}_0.$$

Combining the two inclusions yields

$$\mathcal{E}_0 = \mathcal{G}_\Pi \cap [G_0]_{\mathrm{MEC}}.$$

The posterior-concentration conclusion follows directly from Theorem 3.3. If, in addition, $[G_0]_{\mathrm{MEC}} \subseteq \mathcal{G}_\Pi$, then

$$\mathcal{E}_0 = [G_0]_{\mathrm{MEC}}.$$

□

Proof of Corollary 3.6. By unique graph identifiability within the support of the pretraining prior, for every $G \in \mathcal{G}_\Pi$,

$$P_0 \in \mathcal{P}_G \quad \Longleftrightarrow \quad G = G_0.$$

Applying Proposition 3.4 yields

$$\mathcal{E}_0 = \{G_0\}.$$

Theorem 3.3 therefore implies

$$P_{\Pi,n}\left(G_0 \mid X^{(n)}\right) = P_{\Pi,n}\left(\mathcal{E}_0 \mid X^{(n)}\right) \longrightarrow 1$$

P_0^∞-almost surely. □

C.3 Implications for Existing Models and Pretraining Design

Learning the simulator-induced graph posterior. Existing causal discovery foundation models typically define a simulator that first samples a graph and its mechanism parameters and then generates a dataset from the resulting SCM (Lorch et al., 2022; Ke et al., 2023; Thompson et al., 2026; Peng et al., 2026; Li et al., 2026; Blöbaum et al., 2026). This construction defines precisely the pretraining prior Π and the joint distribution $P_{\Pi,n}$ in our framework. Under graph negative log-likelihood or cross-entropy training, Theorem 3.1 decomposes the expected posterior error into approximation and fitting errors, with the latter vanishing as the number of pretraining tasks increases. For each fixed dataset size, Theorem 3.2 further shows that, with a suitably increasing model capacity, the learned graph distribution converges to the simulator-induced graph posterior. These results explain how existing models can learn a dataset-to-graph posterior mapping from synthetic tasks, and why increasing the number of pretraining tasks and model capacity improves their approximation of this posterior.

Relation to the compactness condition. For a fixed number of variables, the set of DAGs is finite. Moreover, several existing synthetic generators select causal mechanisms from a finite collection and draw many continuous hyperparameters—such as edge weights, noise scales, kernel parameters, and target signal-to-noise levels—from bounded ranges (Thompson et al., 2026; Li et al., 2026). Such constructions motivate, but do not by themselves establish, the graph-wise compactness condition in Assumption A.7.

More precisely, if the parameter support $\mathcal{H}_G$ is compact and the map

$$\eta \longmapsto p_{G,\eta}$$

is continuous in Hellinger distance, then

$$\mathcal{F}_G = \{p_{G,\eta} : \eta \in \mathcal{H}_G\}$$

is compact because it is the continuous image of a compact set. A distribution with a positive density on a bounded open interval (a, b) generally has the closed interval $[a, b]$ as its topological support. Thus, sampling a parameter from a bounded open interval does not by itself violate our compactness condition, provided that the induced density is well defined and Hellinger continuous at the endpoints. If an endpoint corresponds to a degenerate model, such as zero noise variance, one can instead restrict the parameter to a closed interval $[\epsilon, b]$ with $\epsilon > 0$.

Unbounded parameter priors can likewise be replaced by compact truncations. The following observation formalizes the relation between such a truncation and the original pretraining prior.

Lemma C.1 (Compact truncation of a pretraining prior)**.** *Suppose that $\mathcal{G}_\Pi$ is finite and that $\Pi(\cdot \mid G)$ is tight on its parameter space for every $G \in \mathcal{G}_\Pi$. Then, for every $\delta \in (0, 1)$, there exist compact sets $K_{G,\delta}$ such that*

$$\Pi\left(K_{G,\delta} \mid G\right) \geq 1 - \delta \qquad \text{for every } G \in \mathcal{G}_\Pi.$$

Define the graph-wise truncated prior by

$$\Pi^\delta(A \mid G) := \frac{\Pi(A \cap K_{G,\delta} \mid G)}{\Pi(K_{G,\delta} \mid G)}, \qquad \Pi^\delta(G) := \Pi(G).$$

Then

$$\left\|\Pi^\delta - \Pi\right\|_{\mathrm{TV}} \leq \delta$$

and, for every fixed n,

$$\left\|P_{\Pi^\delta,n} - P_{\Pi,n}\right\|_{\mathrm{TV}} \leq \delta.$$

If $\eta \mapsto p_{G,\eta}$ is Hellinger continuous on $K_{G,\delta}$, the density family induced by the truncated prior is graph-wise Hellinger compact.

Proof. Tightness gives a compact set $K_{G,\delta}$ with conditional prior probability at least $1 - \delta$ for every supported graph. The total-variation distance between a probability distribution and its conditional distribution on an event is the probability of the complement of that event. Therefore,

$$\begin{aligned}\left\|\Pi^\delta - \Pi\right\|_{\mathrm{TV}} &\leq \sum_{G \in \mathcal{G}_\Pi} \Pi(G)\Pi\left(K_{G,\delta}^c \mid G\right) \\ &\leq \delta.\end{aligned}$$

Generation of $(G, X^{(n)})$ from an SCM is a Markov kernel, under which total-variation distance cannot increase. This proves the corresponding bound for $P_{\Pi^\delta,n}$. Finally, let

$$\mathcal{H}_G^\delta := \operatorname{supp}\left(\Pi^\delta(\cdot \mid G)\right).$$

Because $K_{G,\delta}$ is closed and the truncated conditional prior assigns it probability one, $\mathcal{H}_G^\delta$ is a closed subset of $K_{G,\delta}$ and is therefore compact. The density family induced by the truncated prior is the continuous image of $\mathcal{H}_G^\delta$, and is therefore Hellinger compact. □

Consequently, our theorem applies exactly to a suitably truncated pretraining prior when the remaining regularity and prior-support conditions hold. The lemma does not imply that an untruncated implementation automatically satisfies the theorem; rather, it shows that commonly used pretraining distributions can often be approximated arbitrarily closely by ones satisfying the compactness requirement.

What the framework explains. Broad pretraining over graph structures, causal functions, and noise distributions enlarges the families $\mathcal{P}_G$. This makes it more likely that a test distribution P_0 is either represented by the pretraining family or is close in KL divergence to a prior-supported distribution. Our concentration result then identifies the graphs whose pretraining families best approximate P_0. This provides a statistical explanation for why models trained on diverse synthetic SCMs can perform well even when their pretraining distributions were selected empirically rather than derived from an explicit consistency theorem.

Limitations of existing pretraining families. A diverse simulator is not necessarily a universal simulator. For example, a pretraining family containing only particular linear, additive-noise, neural-network, or Gaussian-process mechanisms generally cannot represent every observational distribution compatible with a given DAG. Hence, $\mathcal{P}_G$ may be a strict subset of all distributions that are Markov with respect to G. When P_0 lies outside $\bigcup_{G\in\mathcal{G}_\Pi} \mathcal{P}_G$, the target is the KL-best-fitting graph set $\mathcal{E}_0$, which need not contain the true DAG or its Markov equivalence class. Increasing the number of pretraining tasks cannot remove this form of model misspecification because the target itself is determined by the pretraining prior.

Limitations of posterior parameterizations. The posterior-expressiveness condition is separate from the coverage of the pretraining distribution. Some existing models use independent-edge posteriors (Lorch et al., 2022), while others impose structured factorizations over skeletons, orders, or sequential graph decisions (Ke et al., 2023; Thompson et al., 2026). Such factorizations can be useful inductive biases, but full support over DAGs does not imply the ability to represent every probability distribution over DAGs.

For example, consider four DAGs G_{ab}, $a, b \in \{0, 1\}$, that differ only in the presence of two possible edges. Let a target posterior satisfy

$$p^\star(G_{00}) = p^\star(G_{11}) = \frac{1}{2}, \qquad p^\star(G_{01}) = p^\star(G_{10}) = 0.$$

The two edges are perfectly dependent under $p^\star$. For the family $\mathcal{Q}_{\mathrm{ind}}$ of independent Bernoulli edge distributions,

$$\inf_{q\in\mathcal{Q}_{\mathrm{ind}}} \mathrm{KL}\left(p^\star \| q\right) = \log 2 > 0.$$

Thus, increasing the size of the encoder while retaining the same independent-edge output family cannot eliminate this approximation error.

Our decomposition therefore distinguishes two limitations that can otherwise be conflated: insufficient simulator coverage changes the statistical target $\mathcal{E}_0$, whereas insufficient posterior expressiveness prevents the learned model from accurately representing the posterior associated with that target.

C.4 A Fully Verifiable LiNGAM Specialization

We now give a concrete pretraining distribution for which all conditions of Theorem 3.3 can be verified directly and the graph posterior concentrates on the true DAG.

Fix constants

$$0 < b_{\min} < b_{\max} < \infty, \qquad 0 < s_{\min} < s_{\max} < \infty.$$

For every $G \in \mathcal{G}_\Pi$, define

$$\mathcal{B}_G := \left\{ B : \begin{array}{ll} B_{ij} = 0, & \text{if } j \to i \notin E(G), \\ b_{\min} \le |B_{ij}| \le b_{\max}, & \text{if } j \to i \in E(G) \end{array} \right\}.$$

Conditionally on G, the pretraining prior is supported on

$$\Theta_G := \mathcal{B}_G \times [s_{\min}, s_{\max}]^d.$$

For $(B, s) \in \Theta_G$, observations are generated according to

$$X = BX + \varepsilon,$$

where the components of ε are mutually independent and

$$\varepsilon_i \sim \mathrm{Laplace}(0, s_i), \qquad i = 1, \dots, d.$$

We assume that the conditional prior $\Pi(\cdot \mid G)$ has full support on Θ_G, and define

$$\mathcal{P}_G^{\mathrm{LiNGAM}} := \{P_{B,s} : (B, s) \in \Theta_G\}.$$

For the LiNGAM-only prior considered here,

$$\mathcal{P}_G = \mathcal{P}_G^{\mathrm{LiNGAM}}.$$

Proposition C.2 (Posterior consistency under a compact LiNGAM prior)**.** *Suppose that the pre-training prior is constructed as above and that*

$$P_0 = P_{B_0, s_0}$$

for some

$$G_0 \in \mathcal{G}_\Pi, \qquad (B_0, s_0) \in \Theta_{G_0}.$$

Then the conditions of Theorem 3.3 hold,

$$\mathcal{E}_0 = \{G_0\},$$

and

$$P_{\Pi,n}\left(G_0 \mid X_0^{(n)}\right) \longrightarrow 1 \qquad P_0^\infty\text{-almost surely}.$$

Proof. We verify the assumptions of Theorem 3.3.

First, every $P_{B,s}$ admits a density with respect to Lebesgue measure on $\mathbb{R}^d$. Indeed, because G is acyclic, $I - B$ is invertible and

$$|\det(I - B)| = 1.$$

Therefore,

$$p_{B,s}(x) = \prod_{i=1}^{d} \frac{1}{2s_i} \exp\left(-\frac{|((I - B)x)_i|}{s_i}\right).$$

Thus, Lebesgue measure is a common dominating measure for all prior-supported models and for P_0. The map

$$(G, B, s) \longmapsto p_{B,s}$$

is also measurable.

Second, $\mathcal{B}_G$ is compact because every present coefficient is restricted to the compact set

$$[-b_{\max}, -b_{\min}] \cup [b_{\min}, b_{\max}],$$

while every absent coefficient is fixed at zero. Hence

$$\Theta_G = \mathcal{B}_G \times [s_{\min}, s_{\max}]^d$$

is compact.

If

$$(B_m, s_m) \longrightarrow (B, s),$$

then

$$p_{B_m, s_m}(x) \longrightarrow p_{B,s}(x)$$

for every $x \in \mathbb{R}^d$. Since all these functions are probability densities, Scheffé's lemma gives

$$\|p_{B_m, s_m} - p_{B,s}\|_1 \longrightarrow 0.$$

Because

$$H^2(p, q) \le \|p - q\|_1,$$

the parameter-to-density map is Hellinger continuous. Consequently,

$$\mathcal{F}_G = \{p_{B,s} : (B, s) \in \Theta_G\}$$

is Hellinger compact.

Third, because

$$P_0 = P_{B_0,s_0} \in \mathcal{P}_{G_0},$$

we have

$$K_0^\star = K_0(G_0) = 0 < \infty.$$

It remains to verify prior support near the KL optimum. For $(B, s) \in \Theta_{G_0}$,

$$-\log p_{B,s}(x) = \sum_{i=1}^{d} \left[\log(2s_i) + \frac{|((I - B)x)_i|}{s_i} \right].$$

Because B and s range over compact sets with $s_i \geq s_{\min} > 0$, there exists a finite constant C such that

$$|\log p_{B,s}(x)| \leq C\,(1 + \|x\|)$$

uniformly over $(B, s) \in \Theta_{G_0}$. Moreover,

$$\mathbb{E}_{P_0}\|X\| < \infty$$

because P_0 is generated by a finite linear transformation of independent Laplace variables. Dominated convergence therefore implies that

$$(B, s) \longmapsto \mathrm{KL}\,(P_0 \| P_{B,s})$$

is continuous at (B_0, s_0).

Hence, for every $\epsilon > 0$, there exists a relatively open neighborhood $U_\epsilon \subseteq \Theta_{G_0}$ of (B_0, s_0) such that

$$\mathrm{KL}\,(P_0 \| P_{B,s}) < \epsilon \qquad \text{for every } (B, s) \in U_\epsilon.$$

Because the conditional prior given G_0 has full support on Θ_{G_0},

$$\Pi\,(U_\epsilon \mid G_0) > 0.$$

Since $\Pi(G_0) > 0$, the induced density prior assigns positive mass to every KL neighborhood of the optimum. Thus Assumption A.9 holds.

We have therefore verified all conditions of Theorem 3.3.

Finally, every prior-supported model is a linear acyclic model with mutually independent non-Gaussian disturbances of nonzero variance and no latent confounders. LiNGAM identifiability therefore implies that if

$$P_0 \in \mathcal{P}_G,$$

then the corresponding coefficient matrix must equal B_0 (Shimizu et al., 2006). By the exact-support construction of $\mathcal{B}_G$, this further implies

$$G = G_0.$$

Consequently,

$$P_0 \notin \mathcal{P}_G \qquad \text{for every } G \in \mathcal{G}_\Pi \setminus \{G_0\}.$$

Proposition 3.4 therefore yields

$$\mathcal{E}_0 = \{G_0\}.$$

Theorem 3.3 now gives

$$P_{\Pi,n}\left(G_0 \mid X_0^{(n)}\right) \longrightarrow 1 \qquad P_0^\infty\text{-almost surely.}$$

□

The Laplace family is used here to make every regularity condition explicit and directly verifiable. The same argument extends to other compact parametric families of independent non-Gaussian disturbance densities, provided that the induced parameter-to-density map is Hellinger continuous and the KL divergence is continuous at the true parameter.

When the pretraining prior contains other mechanism families. Suppose now that the prior contains both LiNGAMs and other causal mechanism families. For each $G \in \mathcal{G}_\Pi$, let $\mathcal{P}_G^{\mathrm{LiNGAM}}$ denote the observational distributions induced by the LiNGAM component of the actual conditional prior support, and let $\mathcal{P}_G^{\mathrm{other}}$ denote those induced by all remaining mechanism components. Then

$$\mathcal{P}_G = \mathcal{P}_G^{\mathrm{LiNGAM}} \cup \mathcal{P}_G^{\mathrm{other}}.$$

LiNGAM identifiability only rules out a second LiNGAM representation of P_0 under an incorrect graph. It does not rule out the possibility that a model from another mechanism family represents the same distribution under some $G \neq G_0$. Therefore, unique-DAG recovery additionally requires cross-family graph identifiability:

$$P_0 \in \mathcal{P}_{G_0}, \qquad P_0 \notin \mathcal{P}_G \quad \text{for every } G \in \mathcal{G}_\Pi \setminus \{G_0\}.$$

If the full density family $\mathcal{F}_G$, including both the LiNGAM and other mechanism components, is Hellinger compact for every $G \in \mathcal{G}_\Pi$, then the graph-wise KL infima are attained. Consequently, the preceding cross-family identifiability condition implies

$$K_0(G_0) = 0, \qquad K_0(G) > 0 \quad \text{for every } G \in \mathcal{G}_\Pi \setminus \{G_0\},$$

and hence

$$\mathcal{E}_0 = \{G_0\}.$$

More generally, suppose that the conditions of Theorem 3.3 hold. Even under model misspecification, if $G_0 \in \mathcal{G}_\Pi$ has a strict KL advantage over every competing graph,

$$K_0(G_0) < K_0(G) \qquad \text{for every } G \in \mathcal{G}_\Pi \setminus \{G_0\},$$

then

$$\mathcal{E}_0 = \{G_0\},$$

and the posterior concentrates on G_0. This condition is stronger than LiNGAM identifiability within the LiNGAM subfamily, but it is the appropriate condition when the pretraining prior mixes several different causal mechanism families.

D Proof of the End-to-End Recovery Result

Proof of Theorem 3.7. For notational convenience, throughout this proof write

$$\widehat{q}_{M,n}(\cdot \mid x^{(n)}) := q_{\widehat{\theta}_{M,r_{M,n},n}}(\cdot \mid x^{(n)}).$$

For every fixed n, let $\mathcal{F}_{M,n}$ denote the σ-field generated by the M pretraining tasks and any randomness of the training algorithm. Conditional on $\mathcal{F}_{M,n}$, the learned posterior $\widehat{q}_{M,n}(\cdot \mid x^{(n)})$ is a fixed function of the input dataset.

Define its posterior-approximation error under the pretraining dataset marginal by

$$\mathcal{K}_{M,n} := \mathbb{E}_{X^{(n)} \sim P^X_{\Pi,n}} \left[\mathrm{KL}\left(P_{\Pi,n}\left(\cdot \mid X^{(n)}\right) \parallel \widehat{q}_{M,n}\left(\cdot \mid X^{(n)}\right)\right)\right].$$

By Theorem 3.2, the capacity sequence $r_{M,n}$ can be chosen so that, for every fixed n,

$$\mathcal{K}_{M,n} \xrightarrow[M\to\infty]{P} 0.$$

For every dataset $x^{(n)}$, define

$$Z_{M,n}\left(x^{(n)}\right) := \left\| \widehat{q}_{M,n}\left(\cdot \mid x^{(n)}\right) - P_{\Pi,n}\left(\cdot \mid x^{(n)}\right)\right\|_{\mathrm{TV}}.$$

Pinsker's inequality and Jensen's inequality give

$$\mathbb{E}_{X^{(n)} \sim P^X_{\Pi,n}}\left[Z_{M,n}\left(X^{(n)}\right)\right] \le \sqrt{\frac{1}{2}\mathcal{K}_{M,n}}.$$

Therefore, for every fixed n,

$$\mathbb{E}_{X^{(n)}\sim P^X_{\Pi,n}}\left[Z_{M,n}\left(X^{(n)}\right)\right] \xrightarrow[M\to\infty]{P} 0.$$

We next transfer this convergence from the pretraining dataset distribution to the test distribution. Fix $a > 0$ and define

$$B_{M,n}(a) := \left\{x^{(n)} : Z_{M,n}(x^{(n)}) > a\right\}.$$

By Markov's inequality,

$$P^X_{\Pi,n}\left(B_{M,n}(a)\right) \le \frac{1}{a}\mathbb{E}_{X^{(n)}\sim P^X_{\Pi,n}}\left[Z_{M,n}\left(X^{(n)}\right)\right],$$

and hence

$$P^X_{\Pi,n}\left(B_{M,n}(a)\right) \xrightarrow[M\to\infty]{P} 0.$$

Because $P_0^n \ll P^X_{\Pi,n}$ and both are probability measures, absolute continuity is uniform: for every $\delta > 0$, there exists $c_\delta > 0$ such that

$$P^X_{\Pi,n}(B) < c_\delta \quad \Longrightarrow \quad P_0^n(B) < \delta$$

for every measurable set B. It follows that

$$P_0^n\left(B_{M,n}(a)\right) \xrightarrow[M\to\infty]{P} 0.$$

Since $0 \le Z_{M,n} \le 1$,

$$\mathbb{E}_{X_0^{(n)}\sim P_0^n}\left[Z_{M,n}\left(X_0^{(n)}\right)\right] \le a + P_0^n\left(B_{M,n}(a)\right).$$

Because $a > 0$ is arbitrary,

$$\mathbb{E}_{X_0^{(n)}\sim P_0^n}\left[Z_{M,n}\left(X_0^{(n)}\right)\right] \xrightarrow[M\to\infty]{P} 0$$

for every fixed n.

Define

$$\mathcal{A}_{M,n} := \mathbb{E}_{X_0^{(n)}\sim P_0^n}\left[1 - \widehat{q}_{M,n}\left(\mathcal{E}_0 \mid X_0^{(n)}\right)\right],$$
$$\rho_n := \mathbb{E}_{X_0^{(n)}\sim P_0^n}\left[1 - P_{\Pi,n}\left(\mathcal{E}_0 \mid X_0^{(n)}\right)\right].$$

The defining property of total variation gives

$$|\mathcal{A}_{M,n} - \rho_n| \le \mathbb{E}_{X_0^{(n)}\sim P_0^n}\left[Z_{M,n}\left(X_0^{(n)}\right)\right].$$

Combining this inequality with the preceding convergence yields

$$\mathcal{A}_{M,n} \xrightarrow[M\to\infty]{P} \rho_n$$

for every fixed n.

By Theorem 3.3,

$$P_{\Pi,n}\left(\mathcal{E}_0 \mid X_0^{(n)}\right) \longrightarrow 1 \qquad P_0^\infty\text{-almost surely.}$$

Since the posterior mass is bounded between zero and one, the dominated convergence theorem implies

$$\rho_n \longrightarrow 0 \qquad \text{as } n \to \infty.$$

Finally, fix $\epsilon > 0$. For all sufficiently large n, $\rho_n < \epsilon/2$, and therefore

$$\Pr\left(\mathcal{A}_{M,n} > \epsilon\right) \le \Pr\left(|\mathcal{A}_{M,n} - \rho_n| > \frac{\epsilon}{2}\right).$$

Taking $M \to \infty$ and then $n \to \infty$ gives

$$\lim_{n\to\infty} \limsup_{M\to\infty} \Pr\left(\mathcal{A}_{M,n} > \epsilon\right) = 0,$$

which is the claimed iterated-limit result. □

E MODEL ARCHITECTURE

This appendix provides the detailed model construction and training procedure for the skeleton and topological-order components summarized in Section 4.

E.1 SHARED DATA ENCODER

The input dataset $D \in \mathbb{R}^{n\times d}$ can be viewed as a table with n samples and d variables. The first stage of the model extracts representations from this table by alternating attention along the sample dimension and the variable dimension. This design matches two requirements of causal discovery: the model should aggregate distributional information across exchangeable samples, and it should also capture statistical relationships among variables.

We first embed each scalar observation D_{ij} into a latent vector,

$$Z_{ij}^{(0)} = e_\xi(D_{ij}) \in \mathbb{R}^m, \qquad i = 1, \dots, n, \quad j = 1, \dots, d. \tag{14}$$

Thus $Z^{(0)} \in \mathbb{R}^{n\times d\times m}$ is a representation of the full data table. We then apply a stack of alternating attention blocks. In each block, the model first performs sample-wise attention: for each variable j, attention is applied across the n samples

$$\widetilde{Z}_{:j}^{(\ell)} = \mathrm{Attn}_{\mathrm{sample}}\left(Z_{:j}^{(\ell-1)}\right), \qquad j = 1, \dots, d. \tag{15}$$

This aggregates information about each variable's empirical distribution across samples.

The model then performs variable-wise attention: for each sample i, attention is applied across the d variables

$$Z_{i:}^{(\ell)} = \mathrm{Attn}_{\mathrm{var}}\left(\widetilde{Z}_{i:}^{(\ell)}\right), \qquad i = 1, \dots, n. \tag{16}$$

This captures dependence patterns and possible asymmetries among variables within each sample. Repeating these two operations for several layers lets the model combine distributional information across samples with relational information across variables.

After the final alternating-attention layer, we aggregate the sample dimension to obtain one representation for each variable:

$$H_j = \mathrm{Pool}_{i=1}^{n}\left(Z_{ij}^{(L)}\right), \qquad H \in \mathbb{R}^{d\times m}. \tag{17}$$

Here H_j summarizes variable j's marginal behavior and its statistical relationships with the other variables. The pooling operation is permutation-invariant over samples, so reordering the observations does not change the resulting variable representations. We denote the complete shared encoder by

$$H = h_\psi(D) \in \mathbb{R}^{d\times m}. \tag{18}$$

E.2 SKELETON DECODER

The skeleton module estimates whether each unordered pair of variables is connected in the underlying causal graph. For each pair (j, k), it computes an adjacency logit from the learned representations:

$$a_{jk} = s_\phi(H_j, H_k). \tag{19}$$

Because the skeleton is undirected, the pairwise logits are symmetrized:

$$\ell_{jk}^S = \frac{1}{2}(a_{jk} + a_{kj}), \qquad \ell_{jj}^S = -\infty. \tag{20}$$

The skeleton distribution is modeled over unordered pairs:

$$q_\phi(S \mid D) = \prod_{1\le j<k\le d} \mathrm{Bernoulli}\left(S_{jk}; \sigma(\ell_{jk}^S)\right), \qquad S_{jk} = S_{kj}. \tag{21}$$

During training, the skeleton target is obtained from the ground-truth DAG $G^\star$ by removing edge directions:

$$S_{jk}^\star = \mathbf{1}\{G_{jk}^\star = 1 \text{ or } G_{kj}^\star = 1\}. \tag{22}$$

For each unordered node pair, define

$$b_{jk} = -\left[S^\star_{jk} \log \sigma(\ell^S_{jk}) + (1 - S^\star_{jk}) \log\left(1 - \sigma(\ell^S_{jk})\right)\right]. \tag{23}$$

Let

$$N_+ = \sum_{1 \le j < k \le d} S^\star_{jk}, \qquad N_- = \sum_{1 \le j < k \le d} (1 - S^\star_{jk}) \tag{24}$$

denote the numbers of edges and non-edges, respectively.

The ordinary binary cross-entropy averages equally over all unordered node pairs:

$$\mathcal{L}_{\text{ordinary}} = \frac{1}{N_+ + N_-} \sum_{1 \le j < k \le d} b_{jk}. \tag{25}$$

The class-balanced loss assigns equal total weight to edges and non-edges:

$$\mathcal{L}_{\text{balanced}} = \frac{1}{2}\left[\frac{1}{N_+} \sum_{1 \le j < k \le d} S^\star_{jk} b_{jk} + \frac{1}{N_-} \sum_{1 \le j < k \le d} (1 - S^\star_{jk}) b_{jk}\right]. \tag{26}$$

The final skeleton loss is

$$\mathcal{L}_{\text{skel}} = 0.5\, \mathcal{L}_{\text{ordinary}} + 0.5\, \mathcal{L}_{\text{balanced}}. \tag{27}$$

E.3 Priority-Conditioned Diffusion over Topological Orders

The topological-order module learns a distribution over valid causal orders. A DAG generally admits multiple valid topological orders. Using a fixed canonical order would therefore introduce an arbitrary training target that is not intrinsic to the graph. For each training DAG $G^\star$, we sample independent node priorities

$$u = (u_1, \ldots, u_d), \qquad u_j \sim \text{Uniform}(0, 1). \tag{28}$$

We then run a priority-based Kahn algorithm. At each step, the algorithm selects the node with the smallest priority among the currently available source nodes. Because the sampled priorities are distinct almost surely, this procedure produces a unique topological order

$$\pi_0 = \tau(G^\star, u). \tag{29}$$

For fixed $G^\star$ and u, the target order is deterministic, while resampling u produces different valid orders of the same DAG.

Let

$$H = (H_1, \ldots, H_d)$$

be the variable representations produced by the shared encoder. For an order $\pi = (\pi(1), \ldots, \pi(d))$, define

$$H_\pi = \big(H_{\pi(1)}, \ldots, H_{\pi(d)}\big), \qquad u_\pi = \big(u_{\pi(1)}, \ldots, u_{\pi(d)}\big). \tag{30}$$

Thus, each priority u_j remains attached to the corresponding variable representation H_j throughout the diffusion process.

Forward shuffling. The clean diffusion state is (H_{π_0}, u_{π_0}). At each forward step, we sample

$$\sigma_t \sim q_{\text{shuf}} \tag{31}$$

and apply the same shuffle jointly to the variable representations and their priorities:

$$(H_{\pi_t}, u_{\pi_t}) = \sigma_t \cdot (H_{\pi_{t-1}}, u_{\pi_{t-1}}), \qquad t = 1, \ldots, T. \tag{32}$$

Specifically, the action of a permutation σ on the current ordered state is

$$\sigma \cdot (H_\pi, u_\pi) = \big((H_{\pi(\sigma(1))}, \ldots, H_{\pi(\sigma(d))}),\ (u_{\pi(\sigma(1))}, \ldots, u_{\pi(\sigma(d))})\big). \tag{33}$$

Under this convention, the resulting order satisfies

$$\pi_t = \pi_{t-1} \circ \sigma_t. \tag{34}$$

We use the riffle-shuffle distribution for q_{shuf}. Repeated riffle shuffles progressively remove information about the clean order, with the resulting distribution approaching the uniform distribution over the permutation space $\mathfrak{S}_d$.

Scheduled reverse transitions. The reverse process need not undo exactly one forward shuffle at each prediction. Let

$$0 = t_0 < t_1 < \cdots < t_K = T \tag{35}$$

be a denoising schedule. For each $k = 1, \ldots, K$, let $\bar{\sigma}_k$ denote the reverse transition from t_k to t_{k-1}, defined by

$$(H_{\pi_{t_{k-1}}}, u_{\pi_{t_{k-1}}}) = \bar{\sigma}_k \cdot (H_{\pi_{t_k}}, u_{\pi_{t_k}}). \tag{36}$$

Equivalently,

$$\bar{\sigma}_k = \pi_{t_k}^{-1} \circ \pi_{t_{k-1}}. \tag{37}$$

Therefore, $\bar{\sigma}_k$ may undo either one forward shuffle or the cumulative effect of several forward shuffles.

The reverse model predicts

$$p_\theta \left(\bar{\sigma}_k \mid H_{\pi_{t_k}}, u_{\pi_{t_k}}, t_k \right). \tag{38}$$

The current order π_{t_k} is represented implicitly through $H_{\pi_{t_k}}$ and $u_{\pi_{t_k}}$; it is not separately supplied to the reverse model. Because the reverse-time grid is fixed, t_{k-1} is determined by t_k.

We parameterize each reverse transition using a generalized Plackett–Luce (GPL) distribution. Given $(H_{\pi_{t_k}}, u_{\pi_{t_k}}, t_k)$, the reverse network outputs d^2 scores

$$s_{\theta,i,j} \left(H_{\pi_{t_k}}, u_{\pi_{t_k}}, t_k \right), \qquad i, j \in \{1, \ldots, d\}. \tag{39}$$

Here, i denotes the output position currently being selected, while j indexes an object in the current ordered state. Thus, $s_{\theta,i,j}$ measures the preference for placing the object currently at position j into output position i.

For a reverse permutation $\sigma \in \mathfrak{S}_d$, the corresponding GPL probability is

$$\begin{aligned} &p_\theta \left(\sigma \mid H_{\pi_{t_k}}, u_{\pi_{t_k}}, t_k \right) \\ &\quad = \prod_{i=1}^{d} \frac{\exp s_{\theta,i,\sigma(i)} \left(H_{\pi_{t_k}}, u_{\pi_{t_k}}, t_k \right)}{\sum_{j=i}^{d} \exp s_{\theta,i,\sigma(j)} \left(H_{\pi_{t_k}}, u_{\pi_{t_k}}, t_k \right)}. \end{aligned} \tag{40}$$

The learned reverse process induces a priority-conditioned distribution $p_\theta(\pi \mid D, u)$. The order distribution used in the graph posterior is obtained by marginalizing over the auxiliary priorities:

$$p_\theta(\pi \mid D) = \int p_\theta(\pi \mid D, u) p(u)\, du. \tag{41}$$

Theorem E.1 (Universal expressiveness of GPL order diffusion)**.** *Fix the number of variables $d \geq 2$ and a conditioning context*

$$c = (H, u).$$

Let ν_c be an arbitrary probability distribution on the permutation space $\mathfrak{S}_d$. Consider a reverse diffusion process initialized from the uniform distribution,

$$\pi_K \sim \mathrm{Unif}(\mathfrak{S}_d),$$

with reverse transitions

$$\rho_t \sim p_{\mathrm{GPL}} \left(\cdot; s_t(H_{\pi_t}, u_{\pi_t}, c) \right), \qquad \pi_{t-1} = \pi_t \circ \rho_t, \qquad t = K, \ldots, 1,$$

where $s_t(H_{\pi_t}, u_{\pi_t}, c)$ denotes a GPL score matrix that may depend on the currently ordered variable representations, the correspondingly ordered priority vector, the conditioning context, and the diffusion timestep.

If the GPL scores are allowed to take values in $\mathbb{R} \cup \{-\infty\}$, then there exist a finite K and GPL score functions $s_1, \ldots, s_K$ such that

$$p(\pi_0 \mid c) = \nu_c.$$

If the GPL scores are restricted to finite real values, then for every $\epsilon > 0$, there exist a finite K and finite GPL score functions such that

$$\| p(\pi_0 \mid c) - \nu_c \|_{\mathrm{TV}} < \epsilon.$$

Proof. The exact-representation statement follows directly from Zhang et al. (2025, Theorem 2; Appendix E, Theorem 6). Their result shows that, starting from an arbitrary distribution on $\mathfrak{S}_d$, a finite sequence of GPL-parameterized reverse transitions can produce any prescribed target distribution on $\mathfrak{S}_d$. In particular, the initial distribution may be chosen as $\mathrm{Unif}(\mathfrak{S}_d)$.

The exact construction permits GPL scores equal to $-\infty$, thereby allowing selected transitions to have exactly zero probability. Replacing every $-\infty$ score by a sufficiently large negative finite value produces transition kernels converging to those in the exact construction. Since both $\mathfrak{S}_d$ and the number of reverse transitions are finite, the corresponding terminal distributions converge in total variation to ν_c. □

Theorem E.1 applies pointwise to every fixed dataset representation and priority vector. In particular, any conditional distribution supported on valid topological orders is a probability distribution on $\mathfrak{S}_d$, with zero mass assigned to invalid orders. Therefore, with sufficiently expressive GPL transitions and sufficiently many reverse steps, the proposed order diffusion model can represent arbitrary multimodal distributions over valid topological orders. This universality concerns the order-distribution family itself; it does not require the target distribution to be induced by a single vector of node scores.

E.4 Training Objective and Posterior Sampling

Order-diffusion loss. For each training pair $(D, G^\star)$, we first compute $H = h_\psi(D)$, sample a priority vector u, and construct $\pi_0 = \tau(G^\star, u)$. We then sample a forward trajectory using Eq. (32). For each adjacent pair $t_{k-1} < t_k$ in the denoising schedule, the sampled trajectory determines the reverse target $\bar{\sigma}_k$ in Eq. (36).

The order-diffusion loss is

$$\mathcal{L}_{\mathrm{ord}} = -\mathbb{E}\left[\frac{1}{K}\sum_{k=1}^{K}\log p_\theta\left(\bar{\sigma}_k \mid H_{\pi_{t_k}}, u_{\pi_{t_k}}, t_k\right)\right], \tag{42}$$

where the expectation is over the pretraining task, the sampled priority vector, and the forward shuffles.

In practice, we sample $k \sim \mathrm{Uniform}\{1, \ldots, K\}$ and use

$$\widetilde{\mathcal{L}}_{\mathrm{ord}} = -\log p_\theta\left(\bar{\sigma}_k \mid H_{\pi_{t_k}}, u_{\pi_{t_k}}, t_k\right) \tag{43}$$

as an unbiased stochastic estimate of Eq. (42).

Joint objective. The complete training objective is

$$\mathcal{L} = \lambda_{\mathrm{skel}}\mathcal{L}_{\mathrm{skel}} + \lambda_{\mathrm{ord}}\mathcal{L}_{\mathrm{ord}}, \tag{44}$$

where λ_{skel} and λ_{ord} control the relative contributions of skeleton recovery and topological-order prediction.

Posterior sampling. Given a new dataset D, the shared encoder first computes

$$H = h_\psi(D).$$

We sample a skeleton

$$S \sim q_\phi(S \mid D)$$

and an independent priority vector $u \sim p(u)$. The supplied variable arrangement is used as the initial current state:

$$(H^{(K)}, u^{(K)}) = (H, u). \tag{45}$$

Although one may formally associate this arrangement with a permutation π_T, that permutation is neither known nor supplied to the reverse model.

For $k = K, \ldots, 1$, we sample

$$\widehat{\bar{\sigma}}_k \sim p_\theta\left(\cdot \mid H^{(k)}, u^{(k)}, t_k\right) \tag{46}$$

and jointly reorder the current representations and priorities:

$$(H^{(k-1)}, u^{(k-1)}) = \widehat{\sigma}_k \cdot (H^{(k)}, u^{(k)}). \tag{47}$$

The reordered pair is then used as the input to the next reverse transition.

To record the resulting variable order, we initialize a label list

$$\ell^{(K)} = (1, \dots, d)$$

and apply the same sampled transitions:

$$\ell^{(k-1)} = \widehat{\sigma}_k \cdot \ell^{(k)}. \tag{48}$$

This label list is used only for bookkeeping and is not an input to the reverse model. After the final transition, the sampled topological order is

$$\pi = \ell^{(0)}.$$

Finally, we orient the sampled skeleton according to this order:

$$G = g(S, \pi). \tag{49}$$

Because all edges are oriented consistently with a single topological order, every sampled graph is acyclic.

E.5 Implementation and Computational Resources

The shared data encoder consists of four attention layers with hidden dimension 256, feed-forward dimension 512, eight attention heads, and dropout 0.1. The topological-order diffusion denoiser contains seven layers, the skeleton decoder contains two layers, and the auxiliary precedence head has hidden dimension 64. The permutation diffusion uses 11 forward timesteps and the reverse-time grid $11 \to 7 \to 4 \to 0$, with riffle-shuffle forward transitions and generalized Plackett–Luce reverse distributions. A single forward-diffusion sample is used for each training example.

We train the model for five epochs with batch size 16 using AdamW with an initial learning rate of 2×10^{-4} and weight decay 10^{-5}. The learning rate is linearly warmed up over the first 2% of optimization steps and then held constant. Gradients are clipped to norm 1.0, and all training and inference computations use FP32.

Pretraining the main model used in the synthetic benchmark took approximately 72 hours using one NVIDIA RTX PRO 6000 Blackwell GPU with 96 GB of memory and 22 virtual CPU cores from an Intel Xeon Platinum 8470Q processor.

F Pretraining Distribution Details

This appendix provides the full specification of the synthetic pretraining distribution introduced in Section 4. In addition to listing the graph, mechanism, and noise distributions, we state the bounded-support modifications used to obtain a compact simulator parameter space.

F.1 Task and Graph-Size Distribution

The pretraining corpus contains 4,377,500 independently generated graph–dataset pairs with 3–25 variables, each containing 2,000 observational samples. Of these, 4,250,000 are drawn from the compact main distribution, 106,250 are identifiable LiNGAM tasks, and 21,250 are equal-variance linear Gaussian tasks. The latter two components correspond respectively to 2.5% and 0.5% of the size of the main component.

Within the main distribution, linear-additive, additive-MLP, non-additive-MLP, and random Fourier feature mechanisms are sampled with probabilities 20%, 40%, 30%, and 10%, respectively. All continuous simulator parameters across the three components are restricted to the bounded supports below.

F.2 Graph Generation

We use two graph families, Erdős–Rényi-type graphs and scale-free graphs, selected with equal probability. For a graph with d nodes, we first sample its number of edges as

$$m \sim \mathrm{Unif}\left\{0, \ldots, \min\left(4d, \binom{d}{2}\right)\right\}.$$

Erdős–Rényi-type graphs. Among the $\binom{d}{2}$ possible unordered node pairs, we sample exactly m pairs uniformly without replacement. These pairs form the undirected skeleton. Thus, conditional on d and m, every skeleton with m edges has the same probability.

Scale-free graphs. To generate heterogeneous node degrees, we associate each node j with a positive propensity W_j. The propensities are sampled from a truncated power-law distribution with exponent α, where both α and W_j lie in fixed closed and bounded sets. We then sample m unordered node pairs without replacement, with pair-selection probabilities proportional to

$$W_i W_j.$$

Nodes with larger propensities are therefore more likely to participate in multiple edges, producing a heavy-tailed degree pattern.

Acyclic orientation. After constructing the undirected skeleton, we sample a permutation

$$\pi \sim \mathrm{Unif}(\mathfrak{S}_d)$$

and orient every skeleton edge from the earlier node to the later node in π. Since every edge follows the same order, the resulting directed graph is acyclic by construction.

F.3 Structural Equation Models

For a sampled DAG G, variables are generated in topological order. We use six structural equation model families. In every family, all continuously valued mechanism parameters are sampled from prespecified closed and bounded sets, and all architectural choices are drawn from finite sets.

Linear additive model. For each non-root node j, the structural equation is

$$X_j = \sum_{k \in \mathrm{pa}(j)} w_{kj} X_k + s_j Z_j,$$

where w_{kj} are nonzero structural coefficients and Z_j is a standardized exogenous-noise variable. The coefficients and noise scales are restricted to fixed compact sets.

Additive MLP model. The linear structural function is replaced by a randomly initialized multilayer perceptron:

$$X_j = f_j^{\mathrm{MLP}}\left(X_{\mathrm{pa}(j)}\right) + s_j Z_j.$$

The number of layers, hidden width, and activation function are selected from finite sets, while all network weights are sampled from a fixed compact set. Different causal nodes use independently sampled network weights.

Non-additive MLP model. To allow interactions between causes and exogenous noise, the noise variable is supplied directly as an input to the mechanism:

$$X_j = f_j^{\mathrm{MLP}}\left(\left[X_{\mathrm{pa}(j)}, a_j Z_j\right]\right).$$

The noise scale a_j, network weights, and other continuous mechanism parameters are restricted to closed and bounded sets. No additional noise is added after the MLP transformation.

Random Fourier feature model. For each non-root node, the structural function is

$$X_j = f_j^{\mathrm{RFF}}\left(X_{\mathrm{pa}(j)}\right) + s_j Z_j,$$

where

$$f_j^{\mathrm{RFF}}(x) = \sqrt{\frac{2}{K}} \sum_{r=1}^{K} c_{jr} \cos\left(\omega_{jr}^\top x + b_{jr}\right).$$

The feature count K is selected from a finite set, while the length scale, frequencies, phases, and output coefficients are restricted to compact sets.

LiNGAM. The LiNGAM family uses the linear structural equations

$$X_j = \sum_{k \in \mathrm{pa}(j)} w_{kj} X_k + s_j Z_j,$$

where the disturbances Z_j are mutually independent and non-Gaussian. The structural coefficients and node-specific noise scales are sampled from fixed closed and bounded sets.

Equal-variance linear model. The equal-variance family is defined by

$$X_j = \sum_{k \in \mathrm{pa}(j)} w_{kj} X_k + \sigma Z_j,$$

where the same task-level noise scale σ is shared by all variables. Consequently,

$$\mathrm{Var}(\sigma Z_j) = \sigma^2$$

for every node j. The common variance σ^2 and all structural coefficients are restricted to fixed compact sets.

For a root node, the structural equation contains only its exogenous-noise term.

F.4 Exogenous Noise

Before being used in a structural equation, every base noise variable is centered and rescaled to have population mean zero and variance one. The simulator uses the following bounded noise families. All shape and scale parameters are sampled from prespecified closed and bounded sets.

Truncated Gaussian noise. Let

$$\widetilde{Z} \sim \mathcal{N}(0,1) \quad \text{conditioned on} \quad |\widetilde{Z}| \le 8.$$

Writing

$$v_8 = \mathrm{Var}(\widetilde{Z}),$$

we define

$$Z = \frac{\widetilde{Z}}{\sqrt{v_8}}.$$

The resulting noise has population mean zero, population variance one, and bounded support.

Uniform noise. We use the standardized uniform distribution

$$Z \sim \mathrm{Unif}[-\sqrt{3}, \sqrt{3}],$$

which has population mean zero and variance one.

Beta noise. Let

$$U \sim \mathrm{Beta}(a_Z, b_Z),$$

where a_Z and b_Z range over a fixed compact set. We standardize it as

$$Z = \frac{U - \mathbb{E}[U]}{\sqrt{\mathrm{Var}(U)}}.$$

Because the Beta distribution and its parameter set are bounded, the union of the resulting standardized supports is also bounded.

Truncated Laplace noise. Let

$$\widetilde{Z} \sim \text{Laplace}\left(0, \frac{1}{\sqrt{2}}\right) \quad \text{conditioned on} \quad |\widetilde{Z}| \leq 24.$$

Writing $v_{\text{Lap}} = \text{Var}(\widetilde{Z})$, we set

$$Z = \frac{\widetilde{Z}}{\sqrt{v_{\text{Lap}}}}.$$

This produces bounded, standardized, and non-Gaussian noise.

Truncated centered-exponential noise. Let

$$\widetilde{Z} \sim \text{Exp}(1) \quad \text{conditioned on} \quad 0 \leq \widetilde{Z} \leq 36.$$

Writing

$$\mu_{\text{Exp}} = \mathbb{E}[\widetilde{Z}], \qquad v_{\text{Exp}} = \text{Var}(\widetilde{Z}),$$

we define

$$Z = \frac{\widetilde{Z} - \mu_{\text{Exp}}}{\sqrt{v_{\text{Exp}}}}.$$

The resulting noise is bounded, standardized, and non-Gaussian.

Noise realizations are independent across nodes and observations. Depending on the sampled task, noise-family parameters may be shared across nodes or sampled independently for individual nodes.

F.5 Noise Scaling

For the linear-additive, additive-MLP, and random-Fourier-feature models, we sample a target local signal-to-total-variance ratio

$$R_j^2 \in [0.1, 0.9]$$

for each non-root node. The noise scale s_j is chosen so that the generated finite dataset satisfies

$$\frac{\widehat{\text{Var}}\left(f_j(X_{\text{pa}(j)})\right)}{\widehat{\text{Var}}(X_j)} = R_j^2.$$

The empirical covariance between the signal and noise is included when solving for s_j.

For the non-additive MLP model, the input noise scale a_j is sampled directly from a fixed compact interval. The LiNGAM and equal-variance linear models use their explicitly sampled structural-noise scales and do not apply local-R^2 calibration.

F.6 Compact Parameter Support and Hellinger Compactness

For every fixed graph G with $d \in \{3, \ldots, 25\}$, the simulator contains only finitely many choices of mechanism family, network architecture, activation function, noise family, and random-feature dimension. Conditional on each such discrete configuration, all primitive continuous parameters are restricted to the closed and bounded sets described above. Distributions that would otherwise have unbounded support, including the Gaussian, scale-free-propensity, Laplace, exponential, and random-feature distributions, are truncated to fixed bounded intervals. Since the simulator contains only finitely many discrete configurations, its complete finite-dimensional parameter space can be taken to be compact.

This construction verifies the compact-parameter-space component of the sufficient condition for graph-wise Hellinger compactness stated in Assumption A.7. Specifically, for each graph G, the construction above yields a compact parameter domain $\mathcal{H}_G$. If, in addition, the parameter-to-density map

$$\eta \longmapsto p_{G,\eta}$$

is Hellinger continuous, then the induced density family

$$\mathcal{F}_G = \{p_{G,\eta} : \eta \in \mathcal{H}_G\}$$

is Hellinger compact as the continuous image of a compact set. Thus, the compact-support construction establishes one part of the regularity condition required by our theory, while Hellinger continuity is retained as a separate assumption.

G Synthetic Benchmark Construction Details

G.1 Structural Equation Models Used for Test Data

Conditional on a sampled DAG, the structural equations and their parameters are resampled independently for every dataset. The linear, additive-MLP, non-additive-MLP, RFF, and LiNGAM families use the same functional forms as their corresponding pretraining families in Appendix F.3; we specify their benchmark instantiations below for completeness.

Signed-weight construction. For the linear-additive, additive-MLP, non-additive-MLP, and LiNGAM mechanisms, we sample one shared weight-amplitude parameter per dataset:

$$a_{\mathrm{w}} \sim \mathrm{Unif}(0, 0.9).$$

Conditional on a_{w}, every nonzero linear coefficient and every hidden-layer or output weight in these MLP mechanisms is sampled independently as

$$\begin{aligned} \Pr(S_\ell = -1) = \Pr(S_\ell = +1) &= \frac{1}{2}, \\ M_\ell \mid a_{\mathrm{w}} &\sim \mathrm{Unif}\left(1 - a_{\mathrm{w}}, 1 + a_{\mathrm{w}}\right), \\ W_\ell &= S_\ell M_\ell. \end{aligned} \tag{50}$$

Conditional on a_{w}, the signs and magnitudes are mutually independent across weights. Consequently, every nonzero weight lies in

$$[-1.9, -0.1] \cup [0.1, 1.9].$$

This construction permits both positive and negative effects, excludes coefficients arbitrarily close to zero, and keeps all weights within a fixed bounded set.

Mixture and component distributions. For ID-Mix, one mechanism family is sampled at the dataset level:

$$M \sim \mathrm{Categorical}\left(0.20,\ 0.40,\ 0.30,\ 0.10\right),$$

where the four probabilities correspond respectively to the linear-additive, additive-MLP, non-additive-MLP, and RFF families.

For these five distributions, the exogenous-noise specification is either homogeneous or heterogeneous, each with probability $1/2$. In the homogeneous case, all nodes share the same noise family and its parameters, although their noise realizations remain independent. In the heterogeneous case, the noise family and its parameters are sampled independently for each node. The noise family is selected uniformly from Gaussian, uniform, and Beta noise:

$$Z_j \sim \mathcal{N}(0, 1), \qquad Z_j \sim \mathrm{Unif}[-\sqrt{3}, \sqrt{3}],$$

or

$$U_j \sim \mathrm{Beta}(\alpha_j, \beta_j), \qquad Z_j = \frac{U_j - \mathbb{E}[U_j]}{\sqrt{\mathrm{Var}(U_j)}}, \qquad \alpha_j, \beta_j \sim \mathrm{Unif}(1, 10).$$

Consequently, every Z_j has population mean zero and population variance one.

Linear additive mechanisms. For a non-root node j,

$$X_j = \sum_{k \in \mathrm{pa}(j)} w_{kj} X_k + s_j Z_j.$$

For every dataset, two shape parameters are sampled independently:

$$\alpha_R, \beta_R \overset{\mathrm{ind}}{\sim} \mathrm{Unif}(1, 10).$$

Conditional on these shared shape parameters, each non-root node independently samples a target empirical signal-to-total-variance ratio:

$$R_j^2 = 0.1 + 0.8 B_j, \qquad B_j \sim \mathrm{Beta}(\alpha_R, \beta_R).$$

The scale $s_j > 0$ is chosen so that the generated finite dataset satisfies

$$\frac{\widehat{\mathrm{Var}}\left(\sum_{k \in \mathrm{pa}(j)} w_{kj} X_k\right)}{\widehat{\mathrm{Var}}(X_j)} = R_j^2.$$

The empirical signal–noise covariance is retained when solving for s_j.

Additive MLP mechanisms. For a non-root node j,

$$X_j = f_j^{\mathrm{MLP}}\left(X_{\mathrm{pa}(j)}\right) + s_j Z_j.$$

The number of hidden layers, their common width, and the activation function are sampled independently once per dataset:

$$L \in \{1, 2\}, \qquad h \in \{8, 16, 32, 64\},$$
$$\rho \in \{\tanh,\ \mathrm{hardtanh},\ \mathrm{sigmoid},\ \mathrm{hardsigmoid},\ \mathrm{leakyReLU}\}.$$

These discrete choices are sampled uniformly. Conditional on the shared architecture and weight-amplitude parameter, network weights are sampled independently for each non-root node using Eq. (50). All hidden layers and the scalar linear output layer have zero biases, and no input-dimension-dependent weight scaling is applied. The hardtanh and hardsigmoid activations are defined by

$$\mathrm{hardtanh}(u) = \max\{-1, \min\{1, u\}\}, \qquad \mathrm{hardsigmoid}(u) = \max\{0, \min\{1, u/6 + 1/2\}\}.$$

The leaky-ReLU negative slope is 0.2. The additive-noise scale s_j is selected using the same local-R^2 calibration as in the linear family.

Non-additive MLP mechanisms. For a non-root node j, exogenous noise is provided as an input to the structural function:

$$X_j = f_j^{\mathrm{MLP}}\left(\left[X_{\mathrm{pa}(j)},\, a_j Z_j\right]\right), \qquad a_j \sim \mathrm{Unif}(0.1, 2.0).$$

The MLP architecture and weight distributions are the same as in the additive MLP family. No additional noise is added after the MLP transformation, and local-R^2 calibration is not applied.

Random Fourier feature mechanisms. For a non-root node j,

$$X_j = f_j^{\mathrm{RFF}}\left(X_{\mathrm{pa}(j)}\right) + s_j Z_j,$$

where

$$f_j^{\mathrm{RFF}}(x) = \sqrt{\frac{2}{K}} \sum_{r=1}^{K} c_{jr} \cos\left(\omega_{jr}^\top x + b_{jr}\right).$$

The feature count and length scale are sampled independently once per dataset:

$$K \sim \mathrm{Unif}\{16, 32, 64\}, \qquad \ell \sim \mathrm{LogUnif}(0.5, 2.0).$$

For each non-root node j, the feature parameters are sampled as

$$\omega_{jr} \sim \mathcal{N}\left(0, \ell^{-2} I_{|\,\mathrm{pa}(j)|}\right), \qquad b_{jr} \sim \mathrm{Unif}(0, 2\pi), \qquad c_{jr} \sim \mathcal{N}(0, 1).$$

Conditional on K and ℓ, the frequencies, phases, and output coefficients are sampled independently across nodes and features. The additive-noise scale s_j is determined by the same local-R^2 calibration used above.

Gaussian-process mechanisms. The GP distribution uses a Gaussian-process structural function for every non-root node:

$$f_j \sim \mathcal{GP}(0, k_j), \qquad X_j = f_j\left(X_{\mathrm{pa}(j)}\right) + \varepsilon_j, \qquad \varepsilon_j \sim \mathcal{N}(0, \sigma_j^2).$$

Let $\mathbf{X}_j = (X_j^{(1)}, \ldots, X_j^{(n)})^\top$ denote the vector of observations at node j. Conditional on the parent observations and the sampled kernel and noise parameters,

$$\mathbf{X}_j \mid \mathbf{X}_{\mathrm{pa}(j)}, k_j, \sigma_j^2 \sim \mathcal{N}(0, K_j + \sigma_j^2 I_n),$$

where K_j is the Gram matrix induced by k_j. The kernel is

$$k_j = k_j^{\mathrm{base}} + k_j^{\mathrm{linear}},$$

and the base kernel is selected uniformly from

$$\{\text{Matérn-}1/2,\ \text{Matérn-}3/2,\ \text{Matérn-}5/2,\ \text{rational quadratic},\ \text{squared exponential}\}.$$

The kernel and observation-noise parameters are sampled as

$$v_j \sim \mathrm{Unif}(0.1, 10), \qquad \ell_{jk} \sim \mathrm{Unif}(0.1, 10),$$
$$v_j^{\mathrm{linear}} \sim \mathrm{Unif}(0.1, 10), \qquad \sigma_j^2 = \gamma_j^{-1}, \qquad \gamma_j \sim \mathrm{Unif}(1, 100).$$

Here v_j is the base-kernel variance, ℓ_{jk} is the length scale for parent coordinate k, and

$$k_j^{\mathrm{linear}}(x, x') = v_j^{\mathrm{linear}} x^\top x'.$$

Root nodes are generated using a white-noise kernel. This GP family is absent from our pretraining distribution and is therefore used as a mechanism-level out-of-distribution test.

LiNGAM mechanisms. The LiNGAM distribution uses

$$X_j = \sum_{k \in \mathrm{pa}(j)} w_{kj} X_k + \sigma_j Z_j.$$

The structural coefficients follow the signed-weight construction above. For every node,

$$\sigma_j \sim \mathrm{LogUnif}(0.5, 2.0).$$

For each node, independently sample a noise-family indicator

$$F_j \sim \mathrm{Unif}\{1, 2, 3\}.$$

Conditional on this indicator,

$$Z_j \mid F_j \sim \begin{cases} \mathrm{Laplace}(0, 1/\sqrt{2}), & F_j = 1, \\ \mathrm{Unif}[-\sqrt{3}, \sqrt{3}], & F_j = 2, \\ \mathcal{L}(E-1), \quad E \sim \mathrm{Exp}(1), & F_j = 3. \end{cases}$$

Conditional on the noise specifications, disturbances are independent across nodes and observations. The node-specific scales σ_j produce heterogeneous structural-noise variances.

Nonlinear additive Gaussian-noise mechanisms. The ANM-GAUSSIAN distribution uses

$$X_j = f_j\left(X_{\mathrm{pa}(j)}\right) + \varepsilon_j, \qquad \varepsilon_j \sim \mathcal{N}(0, \sigma_j^2).$$

For each non-root node, let $p_j = |\,\mathrm{pa}(j)|$. The structural function is a one-hidden-layer tanh network of width 16:

$$f_j(x) = s_j v_j^\top \tanh(W_j^\top x + b_j), \qquad W_j \in \mathbb{R}^{p_j \times 16}, \quad v_j, b_j \in \mathbb{R}^{16}.$$

Independently for each non-root node, we sample

$$a_j^{\mathrm{in}}, a_j^{\mathrm{out}} \overset{\mathrm{ind}}{\sim} \mathrm{Unif}(0, 0.5).$$

Conditional on these amplitudes, the entries of $\widetilde{W}_j$ and $\widetilde{v}_j$ are sampled independently using the signed-weight rule in equation 50, with amplitudes a_j^{in} and a_j^{out}, respectively. We then set

$$W_j = \frac{\widetilde{W}_j}{\sqrt{p_j}}, \qquad v_j = \frac{\widetilde{v}_j}{\sqrt{16}}.$$

The output scale and hidden biases are sampled independently as

$$s_j \sim \mathrm{Unif}(0.5, 2.0), \qquad b_{jr} \overset{\mathrm{iid}}{\sim} \mathrm{Unif}(-1, 1), \quad r = 1, \ldots, 16.$$

For every node, including roots, the noise standard deviation is sampled independently as

$$\sigma_j \sim \mathrm{LogUnif}(0.5, 1.5).$$

Conditional on these scales, the Gaussian disturbances are independent across nodes and observations and independent of the mechanism parameters. Root nodes satisfy $X_j = \varepsilon_j$. This particular ANM generator is not included in our pretraining distribution.

Relation to existing synthetic generators. These mechanism families are also represented in existing causal discovery foundation models, although their precise parameterizations and mixture weights differ across studies. Linear mechanisms are used by AVICI and Arrow (Lorch et al., 2022; Thompson et al., 2026). Additive MLP mechanisms appear in Arrow, TabCausal, and FoundCause (Thompson et al., 2026; Li et al., 2026; Blöbaum et al., 2026), while FoundCause additionally uses a non-additive MLP in which the parent variables and exogenous noise are jointly provided as network inputs (Blöbaum et al., 2026). RFF-based mechanisms, including GP approximations implemented with random Fourier features, appear in AVICI, TabCausal, CDFM, and DAG-FM (Lorch et al., 2022; Li et al., 2026; Qiao et al., 2026; Chen et al., 2026). DAG-FM further considers LiNGAM and ANM as explicit identifiable mechanism families (Chen et al., 2026), while CDFM includes linear mechanisms with non-Gaussian noise and nonlinear additive mechanisms (Qiao et al., 2026).

G.2 Evaluation Metrics

Let $A \in \{0,1\}^{d\times d}$ denote the ground-truth DAG adjacency matrix. For a method that returns a directed graph, let $\widehat{A} \in \{0,1\}^{d\times d}$ denote its predicted adjacency matrix, where

$$A_{ij} = 1 \quad \Longleftrightarrow \quad X_i \to X_j.$$

Both A and $\widehat{A}$ have zero diagonal. Methods that natively return a CPDAG are treated separately below.

Directed-edge metrics. Directed-edge classification is performed over all $d(d-1)$ off-diagonal ordered node pairs. Let

$$E = \{(i,j) : i \neq j,\ A_{ij} = 1\}, \qquad \widehat{E} = \{(i,j) : i \neq j,\ \widehat{A}_{ij} = 1\}.$$

We define

$$\mathrm{TP} = |E \cap \widehat{E}|, \qquad \mathrm{FP} = |\widehat{E} \setminus E|, \qquad \mathrm{FN} = |E \setminus \widehat{E}|.$$

Precision, recall, F1, and Jaccard are computed as

$$\begin{aligned}
\mathrm{Precision} &= \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}},\\
\mathrm{Recall} &= \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}},\\
\mathrm{F1} &= \frac{2\mathrm{TP}}{2\mathrm{TP}+\mathrm{FP}+\mathrm{FN}},\\
\mathrm{Jaccard} &= \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}+\mathrm{FN}}.
\end{aligned}$$

Whenever a denominator is zero, the corresponding metric is defined to be zero. Consequently, reversing a true edge produces one false positive and one false negative.

Directed exact match and DAG validity are defined by

$$\mathrm{Exact}_{\mathrm{DAG}} = \mathbf{1}\{\widehat{A} = A\}, \qquad \mathrm{Validity}_{\mathrm{DAG}} = \mathbf{1}\{\widehat{A} \text{ is acyclic}\}.$$

Structural Hamming distance. For each unordered pair $\{i,j\}$, define its directed state by

$$s_A(i,j) = (A_{ij}, A_{ji}).$$

The pairwise structural Hamming distance is

$$\mathrm{SHD}_{\mathrm{DAG}} = \sum_{1\le i<j\le d} \mathbf{1}\big\{s_A(i,j) \neq s_{\widehat{A}}(i,j)\big\}.$$

Thus, an edge addition, deletion, or reversal contributes one unit to SHD. Its normalized version is

$$n\,\mathrm{SHD}_{\mathrm{DAG}} = \frac{\mathrm{SHD}_{\mathrm{DAG}}}{\binom{d}{2}}.$$

Skeleton metrics. For each unordered pair $i < j$, the true skeleton is

$$S_{ij} = \max\{A_{ij}, A_{ji}\}.$$

Let $\widehat{S}_{ij} = 1$ if i and j are adjacent in the predicted graph, regardless of edge direction, and let $\widehat{S}_{ij} = 0$ otherwise. This definition applies to both directed-graph and native-CPDAG predictions. Skeleton precision, recall, F1, and Jaccard are obtained by applying the preceding classification definitions to the $\binom{d}{2}$ unordered node pairs. Skeleton SHD, normalized SHD, and exact match are

$$\begin{aligned}
\mathrm{SHD}_{\mathrm{skel}} &= \sum_{i<j} \mathbf{1}\{S_{ij} \neq \widehat{S}_{ij}\},\\
n\,\mathrm{SHD}_{\mathrm{skel}} &= \frac{\mathrm{SHD}_{\mathrm{skel}}}{\binom{d}{2}},\\
\mathrm{Exact}_{\mathrm{skel}} &= \mathbf{1}\{\widehat{S} = S\}.
\end{aligned}$$

CPDAG metrics. The ground-truth DAG is converted to its CPDAG. Methods that natively return a CPDAG are evaluated directly, whereas a predicted directed graph is converted to a CPDAG only when it is acyclic.

Let C and $\widehat{C}$ denote the true and predicted CPDAGs. For each $i < j$, let

$$s_C(i,j) \in \{\varnothing,\ i - j,\ i \to j,\ j \to i\}$$

denote the state of the unordered pair $\{i, j\}$ in C, where $\varnothing$ means no edge and $i - j$ means an undirected edge. Define $s_{\widehat{C}}(i,j)$ analogously. The CPDAG metrics are

$$\mathrm{SHD}_{\mathrm{CPDAG}} = \sum_{i<j} \mathbf{1}\{s_C(i,j) \neq s_{\widehat{C}}(i,j)\},$$

$$n\,\mathrm{SHD}_{\mathrm{CPDAG}} = \frac{\mathrm{SHD}_{\mathrm{CPDAG}}}{\binom{d}{2}},$$

$$\mathrm{Exact}_{\mathrm{CPDAG}} = \mathbf{1}\{\widehat{C} = C\}.$$

Precision, recall, F1, and Jaccard are additionally computed separately for directed and undirected CPDAG edges. Directed edges are evaluated over ordered node pairs, whereas undirected edges are evaluated over unordered node pairs. If a predicted directed graph is cyclic, its CPDAG metrics are left undefined.

Score-based metrics. For methods that return a continuous directed-edge score matrix, AUROC and AUPRC are computed over all $d(d-1)$ off-diagonal ordered node pairs. In accordance with the implementation, AUROC is undefined when either class is absent, while AUPRC is undefined when the ground-truth graph contains no positive directed edges. Undefined quantities are recorded as missing values.

H Detailed Synthetic Benchmark Results

This section reports detailed synthetic-benchmark results stratified by test distribution, graph size, and sample size. For each stratification, we macro-average over the remaining benchmark-cell dimensions, assigning equal weight to every distribution–node-count–sample-size cell. Since every cell contains 50 independently generated datasets, these macro-averages are also equal-weight averages over the corresponding test datasets.

The directed-graph comparisons include our final model, six external causal discovery foundation models, and six classical methods, for a total of 13 methods. PC and GES are excluded from the directed-graph tables because they natively output CPDAGs. They are included, together with all 13 directed-output methods, in the skeleton and CPDAG comparisons in Appendix H.4.

Definitions of all DAG, skeleton, and CPDAG evaluation metrics used in the following comparisons are provided in Appendix G.2.

H.1 Results across Test Distributions

Table 2 reports directed-edge F1 and normalized DAG SHD separately for the eight test distributions. Each distribution contains $4 \times 6 = 24$ node-count–sample-size cells, corresponding to $1{,}200$ test datasets per method. For each method, DAG F1 and normalized DAG SHD are reported on separate rows.

Our method achieves the highest DAG F1 on six of the eight test distributions and the lowest nSHD on five. It performs best on ID-Mix, both MLP families, RFF, and ANM-Gaussian under both metrics, and attains the highest F1 on the mechanism-level GP distribution. Arrow is strongest on the linear distribution. On LiNGAM, Direct-LiNGAM obtains the highest F1 and ICA-LiNGAM the lowest nSHD, consistent with their specialization to linear non-Gaussian models.

H.2 Results across Graph Sizes

Our method achieves the highest F1 and lowest nSHD for $d \in \{3, 5, 10\}$. At $d = 20$, TabCausal obtains the highest F1 (0.494) and the lowest nSHD (0.147). FoundCause obtains a higher F1 than our

Table 2: Directed-graph recovery across the eight test distributions. For each method, the first row reports DAG F1 and the second reports normalized DAG SHD. Higher F1 and lower nSHD are better. Bold values denote the best result for each metric within each distribution.

Method	Metric	ID-Mix	Linear	MLP Add.	MLP Nonadd.	RFF	GP	LiNGAM	ANM Gaussian
Proposed method									
Ours	F1	**0.679**	0.667	**0.689**	**0.665**	**0.642**	**0.575**	0.687	**0.516**
	nSHD	**0.188**	0.200	**0.183**	**0.200**	**0.204**	0.291	0.200	**0.273**
Causal discovery foundation models									
TabCausal	F1	0.542	0.564	0.579	0.493	0.590	0.561	0.634	0.506
	nSHD	0.259	0.260	0.238	0.287	0.240	**0.272**	0.213	0.286
Arrow	F1	0.603	**0.705**	0.629	0.534	0.431	0.463	0.562	0.450
	nSHD	0.239	**0.193**	0.221	0.264	0.317	0.324	0.282	0.308
FoundCause	F1	0.540	0.484	0.522	0.529	0.526	0.568	0.558	0.421
	nSHD	0.311	0.337	0.315	0.321	0.318	0.299	0.308	0.360
CDFM	F1	0.500	0.492	0.501	0.492	0.522	0.516	0.553	0.455
	nSHD	0.475	0.518	0.485	0.450	0.418	0.476	0.433	0.519
AVICI	F1	0.388	0.320	0.372	0.417	0.418	0.536	0.321	0.358
	nSHD	0.361	0.408	0.374	0.357	0.346	0.307	0.420	0.385
CauScale	F1	0.283	0.301	0.284	0.281	0.198	0.256	0.476	0.274
	nSHD	0.386	0.390	0.379	0.399	0.408	0.446	0.331	0.391
Classical causal discovery methods									
SCORE	F1	0.481	0.501	0.538	0.361	0.575	0.447	0.698	0.454
	nSHD	0.288	0.273	0.248	0.354	0.241	0.375	0.175	0.315
DAS	F1	0.469	0.488	0.525	0.347	0.551	0.437	0.650	0.441
	nSHD	0.276	0.263	0.238	0.339	0.234	0.339	0.184	0.295
ICA-LiNGAM	F1	0.463	0.560	0.470	0.390	0.374	0.396	0.801	0.426
	nSHD	0.314	0.252	0.296	0.363	0.349	0.391	**0.115**	0.325
Direct-LiNGAM	F1	0.489	0.672	0.494	0.408	0.320	0.338	**0.807**	0.309
	nSHD	0.302	0.197	0.288	0.352	0.370	0.423	0.116	0.378
DAGMA-linear	F1	0.405	0.460	0.421	0.364	0.292	0.367	0.780	0.418
	nSHD	0.321	0.294	0.308	0.344	0.369	0.348	0.132	0.329
NOTEARS-linear	F1	0.372	0.419	0.392	0.335	0.255	0.366	0.768	0.385
	nSHD	0.325	0.308	0.310	0.343	0.372	0.340	0.146	0.329

method (0.464 versus 0.455) but a higher nSHD (0.242 versus 0.197), while Arrow is essentially tied with our method in F1 (0.456 versus 0.455). One plausible contributor to our weaker performance at $d = 20$ is the scale and graph-size coverage of pretraining under our current computational budget. Our pretraining corpus contains $4{,}377{,}500$ independently generated graph–dataset pairs spanning $d = 3, \ldots, 25$. FoundCause reports approximately $12{,}000$ newly generated tasks per epoch over $1{,}000$ epochs—approximately 12 million task instances in total—with graph sizes sampled from $d \in [2, 50]$, using eight NVIDIA A100 GPUs (Blöbaum et al., 2026). TabCausal does not state a directly comparable total number of pretraining tasks, but its reported training dimensions extend to $d = 100$, and its final training configuration runs for approximately 20 days on two NVIDIA RTX 4090 GPUs (Li et al., 2026).

These quantities should not be interpreted as a controlled comparison: the methods differ in their data generators, architectures, objectives, batching schemes, and hardware utilization. They nevertheless show that FoundCause and TabCausal are exposed to larger graph sizes during training and that FoundCause uses a larger reported number of synthetic task instances. This is consistent with pretraining coverage and scale being possible contributors to the $d = 20$ gap, but the present experiments do not isolate their effects from the other methodological differences.

The controlled scaling experiment in Appendix J.2 complements this comparison. With the architecture and data distribution fixed at $d = 3$, structure recovery continues to improve as the number of independently generated pretraining datasets increases. This indicates that performance has not

Table 3: Directed-graph recovery across graph sizes. For each method, the first row reports DAG F1 and the second reports normalized DAG SHD. Higher F1 and lower nSHD are better. Bold values denote the best result for each metric and graph size.

Method	Metric	$d = 3$	$d = 5$	$d = 10$	$d = 20$
Proposed method					
Ours	F1	**0.794**	**0.715**	**0.596**	0.455
	nSHD	**0.161**	**0.220**	**0.292**	0.197
Causal discovery foundation models					
TabCausal	F1	0.637	0.585	0.519	**0.494**
	nSHD	0.277	0.301	0.303	**0.147**
Arrow	F1	0.601	0.608	0.525	0.456
	nSHD	0.312	0.289	0.305	0.168
FoundCause	F1	0.597	0.525	0.487	0.464
	nSHD	0.321	0.367	0.354	0.242
CDFM	F1	0.558	0.542	0.505	0.411
	nSHD	0.550	0.522	0.484	0.331
AVICI	F1	0.499	0.442	0.347	0.278
	nSHD	0.413	0.437	0.414	0.215
CauScale	F1	0.300	0.265	0.278	0.333
	nSHD	0.484	0.461	0.405	0.215
Classical causal discovery methods					
SCORE	F1	0.629	0.557	0.465	0.376
	nSHD	0.268	0.294	0.314	0.258
DAS	F1	0.629	0.557	0.431	0.336
	nSHD	0.268	0.294	0.329	0.194
ICA-LiNGAM	F1	0.619	0.518	0.436	0.368
	nSHD	0.278	0.327	0.349	0.249
Direct-LiNGAM	F1	0.581	0.504	0.441	0.395
	nSHD	0.305	0.332	0.338	0.237
DAGMA-linear	F1	0.563	0.467	0.383	0.340
	nSHD	0.317	0.355	0.350	0.200
NOTEARS-linear	F1	0.541	0.435	0.341	0.331
	nSHD	0.329	0.372	0.358	0.178

saturated with respect to pretraining-data scale in the controlled setting and motivates larger-scale pretraining with broader coverage of large graphs.

H.3 Results across Sample Sizes

Table 4 reports directed-edge F1 and normalized DAG SHD separately for each observational sample size. For a fixed sample size, each entry averages over $8 \times 4 = 32$ distribution–node-count cells, corresponding to $1{,}600$ test datasets per method.

Our method achieves the highest F1 and lowest nSHD at every sample size. Its F1 increases from 0.384 at $n = 20$ to 0.797 at $n = 2{,}000$, while nSHD decreases from 0.343 to 0.135. The consistent improvement shows that the model increasingly exploits the additional observational information as the test sample size grows.

H.4 Skeleton and CPDAG Results

We compare all methods at the skeleton and CPDAG levels. Skeleton metrics are obtained by ignoring edge directions and are therefore defined for all methods. For methods that return a directed graph, each acyclic prediction is converted to its corresponding CPDAG before evaluation. PC and GES are evaluated directly using their native CPDAG outputs. CPDAG metrics are undefined for

Table 4: Directed-graph recovery across observational sample sizes. For each method, the first row reports DAG F1 and the second reports normalized DAG SHD. Higher F1 and lower nSHD are better. Bold values denote the best result for each metric and sample size.

Method	Metric	$n = 20$	$n = 50$	$n = 100$	$n = 300$	$n = 1000$	$n = 2000$
Proposed method							
Ours	F1	**0.384**	**0.538**	**0.626**	**0.721**	**0.774**	**0.797**
	nSHD	**0.343**	**0.266**	**0.227**	**0.180**	**0.153**	**0.135**
Causal discovery foundation models							
TabCausal	F1	0.226	0.455	0.561	0.663	0.710	0.737
	nSHD	0.377	0.305	0.261	0.220	0.200	0.179
Arrow	F1	0.354	0.485	0.550	0.603	0.637	0.654
	nSHD	0.368	0.291	0.268	0.240	0.226	0.218
FoundCause	F1	0.281	0.335	0.440	0.626	0.707	0.721
	nSHD	0.457	0.425	0.360	0.259	0.218	0.208
CDFM	F1	0.322	0.403	0.475	0.559	0.620	0.645
	nSHD	0.729	0.547	0.466	0.407	0.354	0.327
AVICI	F1	0.199	0.312	0.408	0.463	0.479	0.488
	nSHD	0.503	0.486	0.366	0.296	0.288	0.279
CauScale	F1	0.199	0.257	0.272	0.302	0.370	0.364
	nSHD	0.453	0.407	0.386	0.384	0.359	0.359
Classical causal discovery methods							
SCORE	F1	0.303	0.429	0.514	0.581	0.608	0.605
	nSHD	0.445	0.308	0.262	0.239	0.223	0.224
DAS	F1	0.275	0.407	0.491	0.565	0.596	0.597
	nSHD	0.359	0.311	0.267	0.243	0.225	0.223
ICA-LiNGAM	F1	0.341	0.422	0.483	0.535	0.567	0.562
	nSHD	0.391	0.315	0.291	0.274	0.264	0.268
Direct-LiNGAM	F1	0.326	0.397	0.477	0.537	0.573	0.569
	nSHD	0.412	0.329	0.295	0.270	0.255	0.259
DAGMA-linear	F1	0.335	0.418	0.460	0.467	0.471	0.479
	nSHD	0.409	0.308	0.280	0.285	0.279	0.272
NOTEARS-linear	F1	0.343	0.417	0.424	0.426	0.426	0.433
	nSHD	0.351	0.302	0.300	0.305	0.301	0.296

cyclic directed predictions, so Table 6 also reports the proportion of directed predictions that are acyclic.

Our method achieves the highest overall skeleton F1 and Jaccard, the lowest skeleton nSHD, and the highest skeleton exact-match rate. PC-FisherZ has the highest skeleton precision, while CDFM has the highest skeleton recall. At the CPDAG level, our method obtains the lowest nSHD, the highest exact-match rate, and the highest F1 and Jaccard for both directed and undirected components. TabCausal attains the highest directed-component precision. CPDAG conversion is undefined for cyclic predictions; in particular, CDFM's CPDAG results are based on the 23.8% of its predictions that are acyclic.

H.5 Large-Scale Comparison of Causal Discovery Foundation Models

The main benchmark evaluates both foundation models and classical methods using 50 independently generated datasets in each benchmark cell. Because extending all classical methods to a much larger test corpus would be computationally expensive, we conduct an additional comparison restricted to our method and the six external causal discovery foundation models.

We retain the same eight test distributions, four graph sizes, six observational sample sizes, and data-generation protocol as in the main benchmark. For each of the $8 \times 4 \times 6 = 192$ distribution–

Table 5: Overall skeleton recovery, macro-averaged over all 192 benchmark cells. Higher precision, recall, F1, Jaccard, and exact match are better; lower nSHD is better.

Method	Precision	Recall	F1	Jaccard	nSHD	Exact
Ours	0.801	0.777	**0.769**	**0.682**	**0.154**	**0.334**
Arrow	0.804	0.683	0.708	0.613	0.191	0.258
AVICI	0.702	0.578	0.578	0.479	0.269	0.186
TabCausal	0.813	0.644	0.684	0.596	0.196	0.256
CauScale	0.578	0.341	0.382	0.303	0.368	0.116
FoundCause	0.712	0.666	0.635	0.530	0.269	0.194
CDFM	0.606	**0.823**	0.665	0.553	0.286	0.156
Direct-LiNGAM	0.766	0.774	0.743	0.651	0.178	0.298
ICA-LiNGAM	0.756	0.771	0.736	0.641	0.182	0.289
SCORE	0.746	0.744	0.704	0.608	0.192	0.256
DAS	0.764	0.662	0.669	0.574	0.189	0.255
NOTEARS-linear	0.844	0.542	0.613	0.512	0.234	0.226
DAGMA-linear	0.819	0.639	0.674	0.574	0.211	0.252
PC-FisherZ	**0.854**	0.611	0.666	0.568	0.199	0.246
GES-BIC	0.783	0.736	0.733	0.635	0.180	0.264

Table 6: Overall CPDAG recovery, macro-averaged over all 192 benchmark cells. The CPDAG structure is evaluated using normalized structural Hamming distance (nSHD), i.e., the fraction of unordered node pairs whose states differ between the predicted and true CPDAGs, where the possible states are no edge, an undirected edge, or either directed orientation. Exact denotes the exact-CPDAG match rate. The directed component consists of compelled directed edges and is evaluated over ordered node pairs, whereas the undirected component consists of unoriented CPDAG edges and is evaluated over unordered node pairs. D and U denote the directed and undirected components, respectively; P, R, F1, and J denote precision, recall, F1 score, and Jaccard index. For directed-output methods, CPDAG metrics are computed only for acyclic predictions, and Acyclic reports the proportion of predictions that are acyclic. PC and GES natively output CPDAGs. Higher values are better except for nSHD.

	CPDAG structure		Directed component				Undirected component				Acyclic rate
Method	nSHD ↓	Exact ↑	D-P ↑	D-R ↑	D-F1 ↑	D-J ↑	U-P ↑	U-R ↑	U-F1 ↑	U-J ↑	↑
Ours	**0.230**	**0.322**	0.637	**0.618**	**0.617**	**0.559**	**0.625**	**0.688**	**0.617**	**0.550**	1.000
Arrow	0.267	0.252	0.638	0.560	0.579	0.528	0.573	0.619	0.557	0.486	1.000
AVICI	0.353	0.180	0.515	0.449	0.457	0.422	0.461	0.494	0.433	0.370	0.947
TabCausal	0.248	0.264	**0.660**	0.583	0.601	0.557	0.605	0.605	0.571	0.509	0.881
CauScale	0.404	0.115	0.536	0.452	0.465	0.434	0.363	0.337	0.315	0.266	0.997
FoundCause	0.332	0.191	0.598	0.576	0.564	0.513	0.519	0.584	0.498	0.421	0.945
CDFM	0.262	0.268	0.583	0.614	0.588	0.552	0.555	0.625	0.560	0.510	0.238
Direct-LiNGAM	0.271	0.292	0.577	0.584	0.570	0.524	0.563	0.570	0.535	0.480	1.000
ICA-LiNGAM	0.278	0.282	0.558	0.559	0.548	0.501	0.541	0.571	0.521	0.461	1.000
SCORE	0.272	0.249	0.584	0.565	0.557	0.509	0.538	0.622	0.532	0.460	1.000
DAS	0.258	0.248	0.590	0.532	0.543	0.499	0.524	0.574	0.508	0.440	1.000
NOTEARS-linear	0.287	0.218	0.555	0.469	0.481	0.458	0.557	0.569	0.514	0.436	0.989
DAGMA-linear	0.282	0.243	0.547	0.486	0.494	0.461	0.555	0.591	0.526	0.451	1.000
PC-FisherZ	0.279	0.236	0.561	0.476	0.493	0.450	0.543	0.499	0.484	0.420	–
GES-BIC	0.282	0.256	0.558	0.572	0.553	0.493	0.534	0.460	0.467	0.421	–

node-count–sample-size cells, we generate 1,000 independent datasets. This yields 192,000 test datasets per method and 1,344,000 method–dataset evaluations in total.

Table 7 confirms the overall conclusion of the main benchmark. Our method achieves the highest mean DAG F1 (0.640) and the lowest mean normalized DAG SHD (0.220). The corresponding

Table 7: Large-scale comparison of causal discovery foundation models over 192 benchmark cells and 192,000 test datasets per method. Entries report means, with dataset-level standard deviations in parentheses. Higher DAG F1 and lower normalized DAG SHD are better. Bold values denote the best mean result.

Method	F1 ↑	nSHD ↓
Ours	**0.640** (0.307)	**0.220** (0.217)
Arrow	0.542 (0.325)	0.272 (0.249)
AVICI	0.391 (0.325)	0.370 (0.276)
TabCausal	0.560 (0.345)	0.257 (0.250)
CauScale	0.301 (0.328)	0.389 (0.294)
FoundCause	0.518 (0.324)	0.320 (0.250)
CDFM	0.502 (0.249)	0.474 (0.293)

values in Table 1 are 0.640 and 0.217, respectively. The standard deviations are relatively large because this overall comparison pools datasets with different graph sizes, sample sizes, structural mechanisms, and graph realizations. They therefore reflect variation in task difficulty rather than uncertainty in the reported means.

Across the eight test distributions, our method achieves the highest DAG F1 on seven and the lowest nSHD on seven. It is best under both metrics on ID-Mix, both MLP distributions, RFF, LiNGAM, and ANM-Gaussian. Arrow obtains the highest F1 on the linear distribution, while our method obtains the lowest nSHD. On the GP distribution, our method achieves the highest F1, whereas TabCausal obtains the lowest nSHD.

The graph-size results reproduce the pattern observed in the main benchmark. Our method achieves the highest F1 and lowest nSHD for $d \in \{3, 5, 10\}$.

Our method achieves the highest F1 and the lowest nSHD at every observational sample size. Its F1 increases from 0.392 at $n = 20$ to 0.788 at $n = 2{,}000$, while its nSHD decreases from 0.348 to 0.144.

I Real-World Evaluation on the Sachs Dataset

I.1 Dataset and Experimental Protocol

The Sachs dataset contains single-cell measurements of 11 phosphorylated proteins and phospholipids involved in a human immune-cell signaling network (Sachs et al., 2005). The standard reference DAG contains 17 directed edges. The original study collected measurements under multiple experimental regimes, including both observational and interventional conditions. In our evaluation, we use the complete 853-sample observational subset.

We evaluate our final model against six causal discovery foundation models—AVICI, TabCausal, CauScale, FoundCause, CDFM, and Arrow—and six classical causal discovery methods: DirectLiNGAM, ICA-LiNGAM, SCORE, DAS, NOTEARS-linear, and DAGMA-linear. Each method receives the same 853 observations and follows its native preprocessing and inference procedure. All predictions are evaluated against the same reference graph.

I.2 Results

Table 11 reports normalized DAG structural Hamming distance. Since the Sachs graph contains $d = 11$ variables, nSHD is obtained by dividing DAG SHD by $\binom{11}{2} = 55$. Each method is evaluated once on the complete observational dataset.

Unlike the synthetic benchmarks, the Sachs dataset does not provide an exact ground-truth DAG generated by a fully known SCM. Its 17-edge reference graph was established using prior biological knowledge and evidence from interventional experiments, whereas the 853 measurements used here are treated as i.i.d. observational samples (Sachs et al., 2005). The reference graph may be incomplete, and the observations may be affected by measurement noise, latent variables, or devia-

Table 8: Large-scale foundation-model comparison across the eight test distributions. Each distribution contains 24,000 test datasets per method. Entries report means, with dataset-level standard deviations in parentheses. For each method, DAG F1 and normalized DAG SHD are reported on separate rows. Higher F1 and lower nSHD are better. Bold values denote the best mean result for each metric within each distribution.

Panel A: ID-Mix, Linear, and MLP distributions

Method	Metric	ID-Mix	Linear	MLP Add.	MLP Nonadd.
Proposed method					
Ours	F1	**0.671** (0.297)	0.674 (0.295)	**0.689** (0.291)	**0.666** (0.289)
	nSHD	**0.196** (0.199)	**0.194** (0.201)	**0.186** (0.192)	**0.203** (0.198)
External causal discovery foundation models					
TabCausal	F1	0.553 (0.345)	0.564 (0.356)	0.586 (0.343)	0.494 (0.330)
	nSHD	0.260 (0.252)	0.256 (0.265)	0.244 (0.245)	0.291 (0.252)
Arrow	F1	0.593 (0.317)	**0.694** (0.298)	0.635 (0.304)	0.520 (0.305)
	nSHD	0.242 (0.231)	0.197 (0.219)	0.223 (0.220)	0.272 (0.234)
FoundCause	F1	0.523 (0.324)	0.494 (0.336)	0.527 (0.323)	0.532 (0.320)
	nSHD	0.318 (0.252)	0.332 (0.269)	0.315 (0.250)	0.315 (0.245)
CDFM	F1	0.495 (0.247)	0.489 (0.236)	0.502 (0.241)	0.489 (0.252)
	nSHD	0.475 (0.293)	0.522 (0.307)	0.480 (0.294)	0.459 (0.287)
AVICI	F1	0.379 (0.321)	0.316 (0.306)	0.370 (0.319)	0.417 (0.316)
	nSHD	0.373 (0.276)	0.409 (0.290)	0.376 (0.275)	0.356 (0.266)
CauScale	F1	0.287 (0.322)	0.318 (0.336)	0.287 (0.324)	0.297 (0.310)
	nSHD	0.387 (0.293)	0.378 (0.295)	0.384 (0.296)	0.389 (0.287)

Panel B: RFF, GP, LiNGAM, and ANM-Gaussian distributions

Method	Metric	RFF	GP	LiNGAM	ANM Gaussian
Proposed method					
Ours	F1	**0.648** (0.335)	**0.581** (0.290)	**0.679** (0.288)	**0.508** (0.323)
	nSHD	**0.200** (0.212)	0.296 (0.251)	**0.211** (0.214)	**0.270** (0.232)
External causal discovery foundation models					
TabCausal	F1	0.590 (0.360)	0.562 (0.325)	0.626 (0.336)	0.504 (0.344)
	nSHD	0.234 (0.241)	**0.277** (0.258)	0.219 (0.227)	0.278 (0.254)
Arrow	F1	0.441 (0.343)	0.471 (0.295)	0.550 (0.325)	0.435 (0.317)
	nSHD	0.315 (0.265)	0.326 (0.261)	0.289 (0.267)	0.314 (0.257)
FoundCause	F1	0.527 (0.322)	0.577 (0.298)	0.549 (0.323)	0.418 (0.321)
	nSHD	0.314 (0.245)	0.296 (0.230)	0.315 (0.245)	0.358 (0.261)
CDFM	F1	0.516 (0.269)	0.520 (0.233)	0.561 (0.258)	0.447 (0.236)
	nSHD	0.427 (0.276)	0.477 (0.287)	0.431 (0.290)	0.520 (0.298)
AVICI	F1	0.423 (0.343)	0.550 (0.313)	0.323 (0.305)	0.347 (0.313)
	nSHD	0.346 (0.268)	0.299 (0.250)	0.417 (0.291)	0.384 (0.273)
CauScale	F1	0.211 (0.317)	0.266 (0.305)	0.472 (0.335)	0.268 (0.316)
	nSHD	0.408 (0.306)	0.448 (0.287)	0.330 (0.279)	0.388 (0.293)

tions from the assumed acyclic causal model. This experiment should therefore be interpreted as a realistic stress test rather than a strict verification of model performance.

Our method and TabCausal obtain the lowest nSHD of 0.2727, corresponding to disagreements with the reference graph on 15 of the 55 unordered node pairs. Our method, TabCausal, Arrow, and CauScale all achieve lower nSHD than the best classical baseline in this comparison, although the remaining structural errors indicate that Sachs is still challenging for current causal discovery methods.

Table 9: Large-scale foundation-model comparison across graph sizes. Each graph size contains $48{,}000$ test datasets per method. Entries report means, with dataset-level standard deviations in parentheses. For each method, DAG F1 and normalized DAG SHD are reported on separate rows. Higher F1 and lower nSHD are better. Bold values denote the best mean result for each metric and graph size.

Method	Metric	$d = 3$	$d = 5$	$d = 10$	$d = 20$
Proposed method					
Ours	F1	**0.800** (0.339)	**0.714** (0.298)	**0.593** (0.241)	0.451 (0.215)
	nSHD	**0.160** (0.259)	**0.221** (0.226)	**0.295** (0.205)	0.202 (0.134)
External causal discovery foundation models					
TabCausal	F1	0.647 (0.413)	0.590 (0.339)	0.512 (0.295)	**0.490** (0.298)
	nSHD	0.274 (0.316)	0.295 (0.262)	0.308 (0.231)	**0.152** (0.116)
Arrow	F1	0.600 (0.421)	0.602 (0.331)	0.517 (0.258)	0.449 (0.233)
	nSHD	0.313 (0.328)	0.290 (0.254)	0.312 (0.219)	0.174 (0.121)
FoundCause	F1	0.603 (0.426)	0.527 (0.325)	0.483 (0.250)	0.461 (0.243)
	nSHD	0.315 (0.335)	0.361 (0.258)	0.360 (0.207)	0.246 (0.140)
CDFM	F1	0.562 (0.328)	0.536 (0.249)	0.504 (0.186)	0.408 (0.174)
	nSHD	0.543 (0.358)	0.523 (0.283)	0.488 (0.251)	0.340 (0.219)
AVICI	F1	0.506 (0.428)	0.430 (0.326)	0.349 (0.241)	0.278 (0.213)
	nSHD	0.405 (0.361)	0.438 (0.282)	0.418 (0.222)	0.220 (0.124)
CauScale	F1	0.314 (0.446)	0.273 (0.347)	0.281 (0.242)	0.335 (0.227)
	nSHD	0.474 (0.368)	0.454 (0.303)	0.408 (0.242)	0.220 (0.134)

Table 10: Large-scale foundation-model comparison across observational sample sizes. Each sample size contains $32{,}000$ test datasets per method. Entries report means, with dataset-level standard deviations in parentheses. For each method, DAG F1 and normalized DAG SHD are reported on separate rows. Higher F1 and lower nSHD are better. Bold values denote the best mean result for each metric and sample size. External methods are ordered by their overall macro-averaged DAG F1.

Method	Metric	$n = 20$	$n = 50$	$n = 100$	$n = 300$	$n = 1000$	$n = 2000$
Proposed method							
Ours	F1	**0.392** (0.311)	**0.532** (0.313)	**0.626** (0.292)	**0.725** (0.254)	**0.775** (0.229)	**0.788** (0.221)
	nSHD	**0.348** (0.246)	**0.272** (0.231)	**0.227** (0.212)	**0.177** (0.186)	**0.150** (0.170)	**0.144** (0.167)
External causal discovery foundation models							
TabCausal	F1	0.236 (0.330)	0.449 (0.323)	0.561 (0.305)	0.664 (0.285)	0.718 (0.271)	0.730 (0.267)
	nSHD	0.380 (0.278)	0.306 (0.256)	0.263 (0.241)	0.217 (0.227)	0.192 (0.219)	0.187 (0.219)
Arrow	F1	0.347 (0.292)	0.477 (0.316)	0.541 (0.318)	0.603 (0.313)	0.637 (0.307)	0.649 (0.302)
	nSHD	0.382 (0.262)	0.300 (0.252)	0.268 (0.244)	0.239 (0.236)	0.225 (0.232)	0.220 (0.228)
FoundCause	F1	0.279 (0.263)	0.341 (0.280)	0.441 (0.297)	0.620 (0.287)	0.710 (0.257)	0.719 (0.252)
	nSHD	0.463 (0.220)	0.421 (0.227)	0.357 (0.237)	0.257 (0.237)	0.214 (0.230)	0.209 (0.227)
CDFM	F1	0.323 (0.228)	0.397 (0.230)	0.471 (0.230)	0.558 (0.223)	0.622 (0.205)	0.645 (0.198)
	nSHD	0.729 (0.199)	0.551 (0.268)	0.474 (0.274)	0.400 (0.270)	0.353 (0.271)	0.336 (0.272)
AVICI	F1	0.201 (0.237)	0.306 (0.265)	0.419 (0.310)	0.470 (0.338)	0.473 (0.341)	0.475 (0.341)
	nSHD	0.503 (0.260)	0.491 (0.255)	0.360 (0.248)	0.292 (0.264)	0.288 (0.266)	0.286 (0.266)
CauScale	F1	0.210 (0.295)	0.264 (0.309)	0.278 (0.326)	0.319 (0.338)	0.365 (0.335)	0.368 (0.336)
	nSHD	0.455 (0.265)	0.403 (0.285)	0.386 (0.299)	0.372 (0.302)	0.360 (0.299)	0.359 (0.299)

Table 11: Normalized DAG structural Hamming distance on the complete 853-sample observational subset of the Sachs dataset. Lower is better, and the best result is shown in bold.

Method	nSHD
Proposed method	
Ours	**0.2727**
Causal discovery foundation models	
Arrow	0.2909
AVICI	0.3273
TabCausal	**0.2727**
CauScale	0.2909
FoundCause	0.3091
CDFM	0.3455
Classical causal discovery methods	
Direct-LiNGAM	0.3091
ICA-LiNGAM	0.3091
SCORE	0.3818
DAS	0.3818
NOTEARS-linear	0.3091
DAGMA-linear	0.3273

J Additional Results

J.1 Controlled Evaluation of Permutation Expressivity

Setup. We compare the ability of different permutation parameterizations to fit prespecified distributions on $\mathfrak{S}_4$. Because $|\mathfrak{S}_4| = 4! = 24$, the complete permutation space can be enumerated. This allows us to evaluate each learned distribution p against its target q using the exact total variation distance

$$\mathrm{TV}(p, q) = \frac{1}{2} \sum_{\sigma \in \mathfrak{S}_4} |p(\sigma) - q(\sigma)| .$$

We compare a three-step state-conditioned GPL diffusion model, a one-step GPL model, and a Gumbel–Sinkhorn model. We additionally report an unrestricted categorical distribution over all 24 permutations as a saturated reference. The categorical parameterization can represent any distribution on $\mathfrak{S}_4$, and therefore has an analytical optimum of zero TV for every target, but its output dimension grows as $d!$.

Target distributions. We consider the following five targets:

- **Single.** All probability is concentrated on the identity permutation:
$$q(0123) = 1.$$
- **Close two.** Probability is distributed uniformly over 0123 and 0132, which differ only by swapping the final two nodes:
$$q(0123) = q(0132) = \frac{1}{2}.$$
- **Opposite two.** Probability is distributed uniformly over the identity permutation and its reversal:
$$q(0123) = q(3210) = \frac{1}{2}.$$
- **CPDAG uniform orders.** Consider the undirected chain CPDAG $0-1-2-3$. There are eight orders for which orienting each chain edge according to the order produces a DAG in the represented Markov equivalence class. The target assigns probability $1/8$ to each of these orders.

Table 12: Total variation distance to the target permutation distributions (mean ± standard deviation over three seeds; lower is better). For each trained model, the first line reports the mean and the second reports the standard deviation. The categorical column reports the analytical optimum.

Target distribution	Categorical	Three-step diffusion	One-step GPL	Gumbel–Sinkhorn
Single	0	9.19×10^{-6} $\pm 8.13\times10^{-7}$	9.45×10^{-5} $\pm 1.71\times10^{-7}$	2.00×10^{-6} $\pm 2.00\times10^{-6}$
Close two	0	1.48×10^{-5} $\pm 1.30\times10^{-5}$	7.76×10^{-5} $\pm 8.79\times10^{-6}$	1.96×10^{-3} $\pm 1.68\times10^{-3}$
Opposite two	0	1.07×10^{-4} $\pm 1.60\times10^{-4}$	0.5000 $\pm 6.7\times10^{-8}$	0.4182 ± 0.0058
CPDAG uniform orders	0	4.43×10^{-4} $\pm 2.25\times10^{-4}$	0.2743 $\pm 4.9\times10^{-8}$	0.2416 ± 0.0192
CPDAG uniform DAGs	0	4.74×10^{-4} $\pm 3.57\times10^{-4}$	0.3727 $\pm 3.3\times10^{-5}$	0.2819 ± 0.0025

- **CPDAG uniform DAGs.** We first sample uniformly from the four DAGs in the Markov equivalence class of $0-1-2-3$, and then sample uniformly from the topological orders of the selected DAG. The resulting order distribution is nonuniform: 0123 and 3210 each receive probability $1/4$, while each of the other six compatible orders receives probability $1/12$.

The first two targets test whether a model can represent concentrated probability and nearby modes. The opposite-two target instead places mass on two maximally separated permutations. The two CPDAG-induced targets capture the structured multimodality that arises when multiple orders or multiple Markov-equivalent DAGs are compatible with the same observational structure.

Results. Table 12 reports the mean TV distance and standard deviation over three random seeds. The categorical entries are analytical optima rather than trained results.

The three-step diffusion model achieves TV distances below 5×10^{-4} across all five target distributions, nearly matching the unrestricted categorical reference. While all models fit the *Single* and *Close two* targets well, the one-step GPL and Gumbel–Sinkhorn models retain TV errors of 0.24–0.50 on the *Opposite two* and CPDAG-induced targets. These results show that composing multiple state-conditioned reverse transitions substantially improves the ability to represent separated and structurally complex multimodal distributions over permutations.

J.2 Scaling with Pretraining Data

We investigate whether our model design and pretraining distribution support continued performance improvements as the amount of pretraining data increases. Because training on larger graphs is substantially more expensive, this controlled $d = 3$ experiment does not isolate the cause of the $d = 20$ performance gap. Instead, it tests whether performance has saturated with respect to pretraining-data scale.

Setup. We fix the graph size to $d = 3$ and train one model on a continuous stream of $3{,}000{,}000$ independently generated causal datasets. The training corpus follows our compact pretraining distribution. We save 27 checkpoints at training stages spanning $5{,}000$ to $3{,}000{,}000$ observed datasets. All checkpoints belong to the same training trajectory; the model architecture, optimization configuration, inference procedure, and test datasets are kept unchanged.

We evaluate ID-Mix and LiNGAM at $n \in \{20, 50, 100, 300, 1000, 2000\}$, using 50 fixed test datasets for each distribution and sample size. The horizontal axis N_{train} denotes the cumulative number of independent pretraining datasets. The shaded regions show 95% confidence intervals computed over the 50 test datasets.

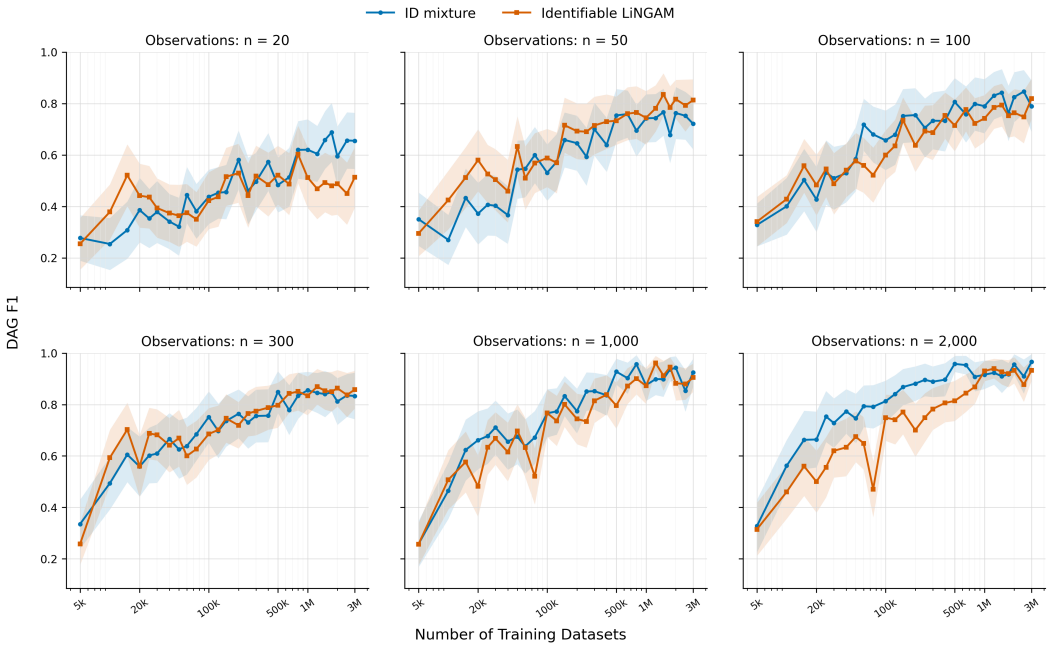


Figure 3: DAG F1 as the number of pretraining datasets increases. Each panel corresponds to a different test sample size n. Lines report the mean over 50 fixed test datasets, and shaded regions show 95% confidence intervals. Higher is better.

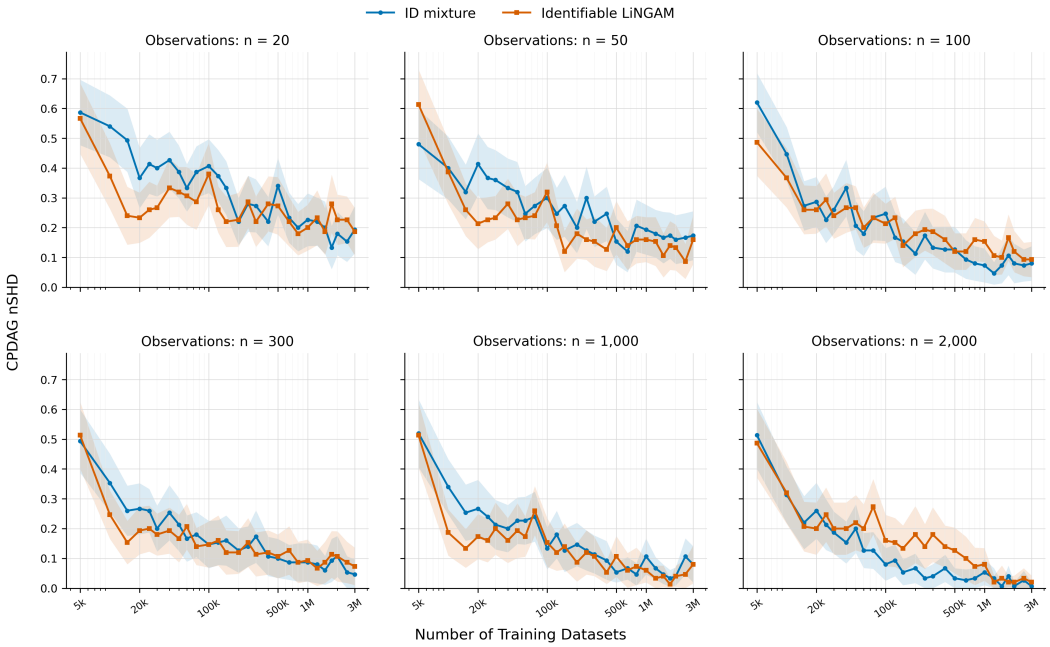


Figure 4: Normalized CPDAG structural Hamming distance (nSHD) as the number of pretraining datasets increases. Each panel corresponds to a different test sample size n. Lines report the mean over 50 fixed test datasets, and shaded regions show 95% confidence intervals. Lower is better.

Results. Figures 3 and 4 show that performance improves consistently as the model observes more pretraining datasets. Across both distributions and all test sample sizes, DAG F1 exhibits a clear increasing trend, while CPDAG nSHD decreases. The improvement is especially pronounced when

Table 13: DAG F1 under the identifiable LiNGAM distribution. Standard deviations over 200 test datasets are reported in parentheses. Higher is better.

n	Ours	Direct-LiNGAM	ICA-LiNGAM	SCORE	DAS	NOTEARS-linear	DAGMA-linear
20	0.597 (0.311)	0.538 (0.310)	0.593 (0.297)	0.595 (0.292)	0.595 (0.292)	0.615 (0.287)	0.625 (0.304)
50	0.825 (0.192)	0.779 (0.280)	0.747 (0.275)	0.717 (0.269)	0.717 (0.269)	0.758 (0.261)	0.749 (0.279)
100	0.910 (0.130)	0.876 (0.223)	0.875 (0.204)	0.781 (0.248)	0.781 (0.248)	0.816 (0.222)	0.816 (0.256)
300	0.960 (0.079)	0.963 (0.107)	0.959 (0.113)	0.822 (0.254)	0.822 (0.254)	0.834 (0.226)	0.886 (0.204)
500	0.973 (0.069)	0.966 (0.145)	0.974 (0.131)	0.809 (0.276)	0.809 (0.276)	0.830 (0.229)	0.879 (0.211)
1000	0.986 (0.041)	0.987 (0.058)	0.993 (0.040)	0.802 (0.272)	0.802 (0.272)	0.838 (0.225)	0.877 (0.216)
1500	0.987 (0.038)	0.991 (0.045)	0.994 (0.042)	0.792 (0.281)	0.792 (0.281)	0.840 (0.225)	0.874 (0.214)
2000	0.989 (0.034)	0.989 (0.059)	0.994 (0.038)	0.780 (0.281)	0.780 (0.281)	0.838 (0.225)	0.876 (0.214)

more observations are available at test time. At $n = 2{,}000$, the final checkpoint achieves DAG F1 scores of 0.966 on ID-MIX and 0.933 on LINGAM, with corresponding CPDAG nSHD values of 0.02 and 0.06.

These trends indicate that performance has not saturated over the observed range of pretraining-data scales and support further investigation of larger-scale pretraining.

J.3 Recovery under an Identifiable LiNGAM Prior

We empirically examine the large-sample recovery behavior predicted by our theory in an identifiable setting. Specifically, this experiment demonstrates that, when the true DAG is identifiable and the model is sufficiently pretrained on the corresponding data-generating distribution, our model can recover the true DAG as the number of observations increases.

Setup. We use LiNGAMs to generate all pretraining and test datasets, following the LiNGAM distribution described in Appendix F.3. We fix the number of variables to $d = 5$ and pretrain our model on $6{,}000{,}000$ independently generated causal datasets. The test datasets are sampled independently from the same distribution, with sample sizes

$$n \in \{20, 50, 100, 300, 500, 1000, 1500, 2000\}.$$

For each sample size, we generate 200 test datasets. We compare our method with Direct-LiNGAM, ICA-LiNGAM, SCORE, DAS, NOTEARS-linear, and DAGMA-linear, and report the mean and standard deviation of DAG F1 and DAG nSHD.

Direct-LiNGAM and ICA-LiNGAM are specifically designed for linear non-Gaussian acyclic models. Under the data-generating assumptions used in this experiment, both methods can identify the true DAG in the population limit and therefore serve as theoretically matched baselines. SCORE and DAS recover causal structure using derivatives of the observational score function, with their identification results developed for nonlinear additive Gaussian-noise models. Their assumptions therefore do not directly match the linear non-Gaussian test distribution considered here. NOTEARS-linear and DAGMA-linear estimate a weighted adjacency matrix by optimizing a linear least-squares objective subject to a continuous acyclicity constraint. Although they produce fully directed graphs, they do not explicitly exploit the non-Gaussianity that makes the true DAG identifiable in the present experiment and consequently do not have the same LiNGAM identification guarantee.

Results. Tables 13 and 14 show that our method improves consistently as the number of observations increases. Its DAG F1 rises monotonically from 0.597 at $n = 20$ to 0.989 at $n = 2{,}000$, while

Table 14: Normalized DAG structural Hamming distance (nSHD) under the identifiable LiNGAM distribution, obtained by dividing SHD by 10. Standard deviations over 200 test datasets are reported in parentheses. Lower is better.

n	Ours	Direct-LiNGAM	ICA-LiNGAM	SCORE	DAS	NOTEARS-linear	DAGMA-linear
20	0.2980 (0.2353)	0.3300 (0.2368)	0.2845 (0.2050)	0.2980 (0.2148)	0.2980 (0.2148)	0.2800 (0.1876)	0.2820 (0.1936)
50	0.1675 (0.1851)	0.1675 (0.1952)	0.1810 (0.1726)	0.2045 (0.1729)	0.2045 (0.1729)	0.1855 (0.1794)	0.1705 (0.1581)
100	0.0995 (0.1395)	0.0955 (0.1525)	0.0995 (0.1441)	0.1525 (0.1497)	0.1525 (0.1497)	0.1590 (0.1808)	0.1220 (0.1554)
300	0.0475 (0.0874)	0.0305 (0.0758)	0.0295 (0.0678)	0.1100 (0.1322)	0.1100 (0.1322)	0.1410 (0.1748)	0.0835 (0.1355)
500	0.0340 (0.0792)	0.0220 (0.0857)	0.0125 (0.0511)	0.1085 (0.1370)	0.1085 (0.1370)	0.1410 (0.1808)	0.0890 (0.1462)
1000	0.0195 (0.0555)	0.0125 (0.0593)	0.0055 (0.0269)	0.1110 (0.1302)	0.1110 (0.1302)	0.1370 (0.1808)	0.0895 (0.1495)
1500	0.0170 (0.0482)	0.0085 (0.0446)	0.0055 (0.0287)	0.1145 (0.1254)	0.1145 (0.1254)	0.1370 (0.1836)	0.0940 (0.1516)
2000	0.0145 (0.0430)	0.0100 (0.0585)	0.0045 (0.0252)	0.1205 (0.1261)	0.1205 (0.1261)	0.1375 (0.1833)	0.0920 (0.1515)

Table 15: Released pretrained models used in the benchmark.

Method	Model identifier or checkpoint file
AVICI	scm-v0
TabCausal	tabcausal-base.pt
CauScale	auprc=0.905_migrated.ckpt
FoundCause	checkpoint.pt (v1.0 release)
CDFM	model.safetensors
Arrow	arrowfm-base.pt

its nSHD decreases monotonically from 0.2980 to 0.0145. At larger sample sizes, our method substantially outperforms SCORE, DAS, NOTEARS-linear, and DAGMA-linear, whose assumptions do not directly match the LiNGAM data-generating process. It also approaches Direct-LiNGAM and ICA-LiNGAM, which are specifically designed to exploit the linear non-Gaussian assumptions used in this experiment.

Our theory gives an end-to-end prediction: with sufficient model capacity and pretraining data, the learned graph distribution approaches the graph posterior induced by the pretraining distribution; under correct specification and graph identifiability, this posterior further concentrates on the true DAG as the number of observations increases. The continued improvement toward near-perfect DAG recovery observed in this experiment is consistent with this prediction.

K Baseline Implementations and Versions

We compare our method with six pretrained causal discovery models and eight classical methods. The pretrained models are AVICI (Lorch et al., 2022), TabCausal (Li et al., 2026), CauScale (Peng et al., 2026), FoundCause (Blöbaum et al., 2026), CDFM (Qiao et al., 2026), and Arrow (Thompson et al., 2026). Table 15 identifies the released models used in our evaluation.

All models in Table 15 use weights released by their respective authors. For AVICI, scm-v0 is the identifier accepted by its pretrained-model loader rather than a checkpoint filename. For CauScale, we use the checkpoint released for synthetic data. We run all pretrained baselines using the inference code released by their authors, retaining each implementation's preprocessing, standardization, and

graph-decoding procedures. None of these pretrained models is retrained or fine-tuned on the test datasets.

The classical methods are fitted separately on each test dataset. Direct-LiNGAM (Shimizu et al., 2011) and ICA-LiNGAM (Shimizu et al., 2006) use lingam version 1.13.0. PC (Spirtes et al., 2000) and GES (Chickering, 2002) use causal-learn version 0.1.4.8. PC uses the Fisher-z test with significance level 0.05, and GES uses the local BIC score. SCORE (Rolland et al., 2022) and DAS (Montagna et al., 2023) use PyWhy DoDiscover version 0.0.0.dev0. NOTEARS-linear (Zheng et al., 2018) uses the original authors' notears implementation, version 3.0, and DAGMA-linear (Bello et al., 2022) uses dagma version 1.1.1. We sparsify the weighted adjacency matrices from NOTEARS-linear and DAGMA-linear using a threshold of 0.3.

All methods receive the same complete observational data matrix before method-specific preprocessing, including any normalization, row selection, or feature construction described above. We apply no additional common standardization to the classical methods and retain their native preprocessing. Where inference is stochastic, we use random seed 42.

We report skeleton metrics for all methods. PC and GES are evaluated using their native CPDAG outputs. The remaining baselines produce directed adjacency estimates, which need not be acyclic, and are evaluated using directed-edge metrics. For acyclic directed predictions, we additionally compute CPDAG metrics by converting both the predicted graph and the ground-truth DAG to their corresponding CPDAGs. CPDAG metrics are undefined for cyclic predictions, and we report the proportion of predictions that are acyclic. All metrics are defined in Appendix G.2.