CONTEXT AUGMENTATION FOR CAUSAL FOUNDATION MODELS WITH INTERVENTIONAL DATA AND DOMAIN KNOWLEDGE

Anonymous authors
Paper under double-blind review

ABSTRACT

Causal effect estimation typically requires expert knowledge to select an appropriate estimator. Recently proposed causal foundation models (CFMs) amortize the prediction of interventional distributions from observational contexts, showing promise for automatic causal effect estimation. However, observational data alone may be insufficient to identify interventional distributions: the model's implicit posterior may spread across multiple causal worlds that are observationally equivalent but imply different interventional distributions. In practice, randomized trial data or structural domain knowledge are sometimes available, yet existing CFMs either provide no interface for such information or require architectural changes and retraining. We propose test-time context augmentation, which supplies this information to a pretrained CFM through additional context samples without retraining. We first consider real interventional samples added without labels distinguishing them from observational samples. Under an idealized Bayesian analysis, we provide a posterior-odds condition under which these samples increase posterior mass on the class of structural causal models (SCMs) that induce the true interventional distribution. When only observational data are available, we fit an auxiliary SCM constrained by domain knowledge and use it to generate pseudo-interventional samples. The framework does not require the ground-truth DAG, and partial structural knowledge suffices as long as it identifies the interventional distribution. We instantiate it with a fully specified DAG and a more accessible causal order. Experiments show that context augmentation breaks observational symmetry in a non-identifiable bivariate setting and strengthens CFM representations of the local causal structure around treatment. Context augmentation also enables a graph-free CFM to match its graph-conditioned counterpart, and improves Do-PFN and CausalPFN on synthetic and semi-synthetic benchmarks.

1 INTRODUCTION

Causal effect estimation typically relies on expert knowledge to determine the applicable causal setting and select an appropriate estimator.(Pearl, 2009; Hernán & Robins, 2010; Imbens & Rubin, 2015). Recently proposed causal foundation models (CFMs), such as Do-PFN, CausalPFN, and CausalFM, are pretrained on large collections of synthetic datasets sampled from priors over structural causal models (SCMs) (Robertson et al., 2026; Balazadeh et al., 2025; Ma et al., 2025). At inference time, they take an observational dataset as context and answer the intervention query in a single forward pass. Conceptually, these models approximate an implicit Bayesian model average: they weight the interventional predictions of all SCMs compatible with the context by their posterior probabilities (Dhir et al., 2025). This offers a route toward automated causal effect estimation.

However, observational data alone may be insufficient to identify the interventional distribution. Multiple SCMs can be compatible with the same observational distribution while implying different interventional distributions (Peters et al., 2017). A simple example is the bivariate linear-Gaussian setting, in which observational data cannot distinguish $T \to Y$ from $Y \to T$. For general-purpose CFMs such as Do-PFN and MACE-TNP Dhir et al. (2025), whose priors are not restricted to identifiable causal settings, this ambiguity can leave posterior mass spread across different causal worlds, resulting in uncertain or biased predictions (Robertson et al., 2026). In practice, users

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.

may have additional information that helps narrow this ambiguity, such as a randomized trial, a causal order, or partial graph structure. However, most existing CFMs have no direct way to use it. General-purpose CFMs take $(\boldsymbol{x}, t, y)$ triplets as context but have no input channel for structural knowledge. Graph-conditioned models such as Graphs4CFM can accept graph information, but this requires modifying the architecture and retraining the model (Reuter et al., 2026). Already released pretrained CFMs therefore cannot use graph information directly. Models trained for specific identifiable settings, such as CausalPFN and CausalFM, still require users to choose the appropriate setting in advance(Balazadeh et al., 2025; Ma et al., 2025). Their predictions may deteriorate when that setting is misspecified .

Retrieval augmentation and prompting methods for large language models illustrate how information supplied at test time can influence a pretrained model's predictions (Lewis et al., 2020; Wei et al., 2022). Motivated by this, we investigate whether interventional data or structural knowledge can be converted into context samples and supplied to a pretrained CFM without retraining.

We begin with the case of interventional data. Since a pretrained CFM has no interface for labels distinguishing observational from interventional samples, the added samples are treated as observational context. We conduct an idealized Bayesian analysis in which the added samples reweight the posterior over candidate SCMs according to their observational likelihood. We analyze three types of candidates: the randomized version of the true SCM, a regression-style SCM with treatment as a root node whose distribution matches the interventional samples, and the true SCM itself. The first two explain the added interventional samples exactly and induce the correct interventional distribution. Moreover, because the two kinds of samples in the augmented context differ only in their treatment-assignment mechanism, SCMs whose assignment mechanism lies between randomized and covariate-dependent may also attain high likelihood while still inducing the correct interventional distribution. We then partition the candidates according to whether their interventional predictions match those of the true SCM and derive the posterior odds of the two classes before and after augmentation. When the likelihood of the class with correct interventional predictions dominates, the model's posterior interventional prediction moves toward the true interventional distribution.

Our analysis also yields a construction principle for the case without real interventional data: we seek pseudo-interventional samples whose distribution approximates the true interventional distribution. A fully specified DAG is not required if the available domain knowledge satisfies an identification condition for the true interventional distribution. We instantiate the framework with two forms of knowledge: a causal graph, and, in the back-door setting, a causal order that suffices to identify the interventional distribution (Vashishtha et al., 2025). Given the available knowledge, we construct a compatible auxiliary graph, fit an additive-noise SCM to the observational data, and sample from the fitted SCM under $\mathrm{do}(t)$. We then append the resulting pseudo-interventional samples to the CFM's context.

Our contributions are as follows:

1. We propose a test-time context augmentation framework that incorporates domain knowledge, including DAGs and causal orders, into pretrained CFMs through pseudo-interventional samples, without retraining or modifying their architectures.

2. We extend the Bayesian analysis of CFMs from observational-only contexts to contexts containing observational and interventional samples without source labels. We derive a posterior-odds condition that characterizes when the added samples increase posterior mass on SCMs inducing the true interventional distribution. This analysis also motivates the construction of pseudo-interventional context.

3. Mechanistically, we show that context augmentation breaks observational symmetry in a non-identifiable bivariate setting, and probing analysis (Swelam et al., 2025) reveals strengthened representations of the local causal structure around treatment, consistent with the theory that interventional data orient local edges around the treatment (Hauser & Bühlmann, 2012).

4. For causal effect estimation, context augmentation enables graph-free Graphs4CFM (Reuter et al., 2026) to match its graph-conditioned counterpart on its own pretraining prior. It can also improve the CATE and ATE estimates of Do-PFN and CausalPFN on synthetic and semi-synthetic benchmarks.

2 BACKGROUND

This section introduces structural causal models, interventional distributions, and causal foundation models. Table 4 in Appendix A.1 summarizes our notation.

2.1 STRUCTURAL CAUSAL MODELS AND INTERVENTIONAL DISTRIBUTIONS

We describe the data-generating process using an SCM (Pearl, 2009; Peters et al., 2017), $\psi := (\mathcal{G}, \{f_k\}_{k=1}^K, p(\boldsymbol{\varepsilon}))$. Here, $\mathcal{G} = (\mathcal{V}, \boldsymbol{A})$ is a DAG over $\mathcal{V} = \{Z_1, \ldots, Z_K\}$, with $A_{i,j} = 1$ indicating that Z_i is a direct cause of Z_j; each variable is generated by $Z_k = f_k(Pa_{\mathcal{G}}(Z_k), \varepsilon_k)$; and $p(\boldsymbol{\varepsilon})$ is the joint distribution of the exogenous noises. The CFM itself need not assume unconfoundedness, while our augmentation framework operates on observed variables. Our experiments follow Do-PFN (Robertson et al., 2026) in using mutually independent exogenous noises, although both the Do-PFN architecture and our framework can accommodate dependent noises.

We denote treatment by T, outcome by Y, and the remaining covariates by $\boldsymbol{X} = (X_1, \ldots, X_{K-2})$, with $\boldsymbol{x}$ denoting their values. An SCM generates the observational dataset $\mathcal{D}_{\text{obs}} = \{(\boldsymbol{x}^{(i)}, t^{(i)}, y^{(i)})\}_{i=1}^N$ by ancestral sampling in DAG order. The intervention $\text{do}(T = t)$, abbreviated $\text{do}(t)$, sets T to the constant t and removes the influence of its parents. It yields the intervened SCM $\psi_{\text{do}(t)}$ and the conditional interventional distribution (CID) $p(y \mid \text{do}(t), \boldsymbol{x})$ (Shpitser & Pearl, 2006). For binary treatment, the CID defines the conditional average treatment effect (CATE):

$$\tau(\boldsymbol{x}) := \mathbb{E}[Y \mid \text{do}(1), \boldsymbol{x}] - \mathbb{E}[Y \mid \text{do}(0), \boldsymbol{x}]. \tag{1}$$

2.2 CAUSAL FOUNDATION MODELS AND OBSERVATIONAL NON-IDENTIFIABILITY

General-purpose CFMs such as Do-PFN (Robertson et al., 2026) and MACE-TNP (Dhir et al., 2025) predict interventional distributions from observational data without explicitly discovering the causal structure. They follow the PFN paradigm. During pretraining, SCMs are drawn from a prior $\psi \sim p(\psi)$, observational datasets $\mathcal{D}_{\text{obs}}$ and interventional query-target pairs $(\boldsymbol{x}_q, t_q, y_q)$ are generated, with $(\boldsymbol{x}_q, y_q) \sim p_\psi(\boldsymbol{x}, y \mid \text{do}(t_q))$, and the model q_θ minimizes the expected negative log-likelihood:

$$\min_\theta \ \mathbb{E}[-\log q_\theta(y_q \mid \text{do}(t_q), \boldsymbol{x}_q, \mathcal{D}_{\text{obs}})]\,. \tag{2}$$

Up to conditional entropy independent of θ, this objective minimizes the expected KL divergence from the posterior predictive distribution to q_θ. At inference, the CFM uses $\mathcal{D}_{\text{obs}}$ as context to approximate an implicit Bayesian model average over SCMs (Madigan & Raftery, 1994; Zhang et al., 2025b):

$$p(y_q \mid \text{do}(t_q), \boldsymbol{x}_q, \mathcal{D}_{\text{obs}}) = \int_\Psi p_\psi(y_q \mid \text{do}(t_q), \boldsymbol{x}_q)\, p(\psi \mid \mathcal{D}_{\text{obs}})\, d\psi, \tag{3}$$

where Ψ is the support of $p(\psi)$.

Observational data may be compatible with multiple SCMs that imply different interventional distributions (Spirtes et al., 1993; Peters et al., 2017).In such cases, point identification requires sufficient structural knowledge, suitable restrictions on functional forms or noise distributions, or relevant interventional data. In Eq. 3, this ambiguity appears as posterior mass across different causal worlds and can yield uncertain or biased CID estimates. Interventional data can refine an observational Markov equivalence class into a smaller interventional Markov equivalence class (Hauser & Bühlmann, 2012; 2015; Eberhardt et al., 2005; Yang et al., 2018), ruling out some candidate structures. This motivates our use of interventional samples or, when they are unavailable, pseudo-interventional samples constructed from structural knowledge to guide a pretrained CFM at test time.

3 RELATED WORK

Prior-fitted models for causal effect estimation. Existing prior-fitted causal foundation models for causal effect estimation can be roughly grouped into two lines. The first line trains models for specific identifiable causal settings. For example, CausalPFN (Balazadeh et al., 2025) focuses on amortized causal effect estimation under the back-door setting, while Ma et al. (2025) extend prior-fitted models to several classical settings such as back-door adjustment, front-door adjustment, and instrumental

variables. By restricting the training prior to causal worlds where the target effect is identifiable, these methods can achieve strong performance when the assumed setting is correct. However, they still require users to select an appropriate model or causal setting based on domain knowledge, and their performance can degrade when this knowledge is misspecified.

The second line trains general-purpose CFMs over broad priors of causal worlds. Do-PFN (Robertson et al., 2026), MACE-TNP (Dhir et al., 2025), and ACTIVA (Sauter et al., 2025) are representative examples. These models are trained on diverse SCMs with different DAGs and mechanisms, and directly predict interventional distributions from observational data. This allows them to perform implicit causal adjustment and to represent uncertainty caused by observational non-identifiability (Robertson et al., 2026; Dhir et al., 2025). However, when domain knowledge is available, these models do not provide a direct mechanism to use it. As a result, they may still assign posterior mass to causal worlds that are observationally plausible but inconsistent with the available structural knowledge.

Conditioning CFMs on structural knowledge. Recent work incorporates domain knowledge into CFMs by conditioning the model on graph information (Reuter et al., 2026; Zhu et al., 2026). Reuter et al. (2026) represent causal knowledge using full or partial ancestral matrices and develop graph-conditioned architectures that can use structural information. We refer to this model as Graphs4CFM throughout the paper. During training, parts of the graph information are masked so that the model can handle incomplete causal knowledge at test time. While effective, these approaches require training a new model with the graph-conditioning mechanism built into the architecture.

Our work is complementary but targets a different setting. Rather than training a new graph-conditioned CFM, we study how to incorporate domain knowledge into already pretrained CFMs such as Do-PFN and CausalPFN. Motivated by the success of test-time tools for large foundation models, such as retrieval-augmented generation (Lewis et al., 2020), chain-of-thought prompting (Wei et al., 2022), and tool use (Schick et al., 2023), we study whether a similar test-time mechanism can benefit causal foundation models.

4 Context Augmentation for Causal Foundation Models

4.1 Context Augmentation with Interventional Data

Existing causal foundation models, such as Do-PFN and CausalPFN, typically predict interventional outcomes from observational contexts. We study the effect of adding real randomized controlled trial (RCT) samples when the model does not explicitly distinguish observational from interventional samples. Let $\boldsymbol{z} = (\boldsymbol{x}, t, y)$ denote a sample and ψ_0 the true structural causal model. The observational dataset $\mathcal{D}_{\text{obs}}$ contains N samples, while the RCT dataset $\mathcal{D}_{\text{int}}$ contains $M = rN$ samples, with treatment randomized according to ρ.

For any candidate SCM $\psi \in \Psi$, define the observational and interventional mappings by

$$\Phi_{\text{obs}}(\psi) := P_\psi^{\text{obs}}, \qquad \Phi_{\text{int}}(\psi) := P_\psi^{\rho}, \qquad P_\psi^{\rho}(d\boldsymbol{x}, dt, dy) := \rho(dt) P_\psi(d\boldsymbol{x}, dy \mid \text{do}(t)). \tag{4}$$

Here, $P_\psi(\cdot \mid \text{do}(t))$ denotes the distribution of the SCM $\psi_{\text{do}(t)}$ obtained by intervening on T at the fixed value t. Thus, observational samples follow $P_{\text{obs}} = \Phi_{\text{obs}}(\psi_0)$, whereas RCT samples follow $P_{\text{int}} = \Phi_{\text{int}}(\psi_0)$.

We concatenate the two datasets into $\mathcal{D}_{\text{aug}} = \mathcal{D}_{\text{obs}} \cup \mathcal{D}_{\text{int}}$. To analyze the added samples, we consider an idealized Bayesian update that treats all context samples as observational. For brevity, denote $p(\psi \mid \mathcal{D}_{\text{obs}})$ and $p(\psi \mid \mathcal{D}_{\text{aug}})$ by $\pi_{\text{obs}}(\psi)$ and $\pi_{\text{aug}}(\psi)$. The augmented interventional prediction is

$$p(y_q \mid \text{do}(t_q), \boldsymbol{x}_q, \mathcal{D}_{\text{aug}}) = \int_\Psi p_\psi(y_q \mid \text{do}(t_q), \boldsymbol{x}_q) \pi_{\text{aug}}(\psi)\, d\psi. \tag{5}$$

Since the model does not explicitly accept labels distinguishing observational from interventional samples, the likelihood assigned to the added samples in this analysis is

$$L_{\text{int}}(\psi) := \prod_{\boldsymbol{z}^{(j)} \in \mathcal{D}_{\text{int}}} p_\psi^{\text{obs}}(\boldsymbol{z}^{(j)}). \tag{6}$$

Consequently,

$$\pi_{\mathrm{aug}}(\psi) = \frac{\pi_{\mathrm{obs}}(\psi)L_{\mathrm{int}}(\psi)}{\int_{\Psi} \pi_{\mathrm{obs}}(\psi')L_{\mathrm{int}}(\psi')\,d\psi'}, \qquad \frac{\pi_{\mathrm{aug}}(\psi_a)}{\pi_{\mathrm{aug}}(\psi_b)} = \frac{\pi_{\mathrm{obs}}(\psi_a)}{\pi_{\mathrm{obs}}(\psi_b)}\,\frac{L_{\mathrm{int}}(\psi_a)}{L_{\mathrm{int}}(\psi_b)}. \tag{7}$$

Under suitable density and integrability conditions, the law of large numbers gives

$$\frac{1}{M}\log L_{\mathrm{int}}(\psi) \longrightarrow C_{\mathrm{int}} - D_{\mathrm{KL}}(P_{\mathrm{int}} \| \Phi_{\mathrm{obs}}(\psi))\,, \tag{8}$$

where C_{int} does not depend on ψ. Hence, the added rows tend to increase the relative weight of candidates whose observational distributions are closer to P_{int}. We discuss several candidates that may receive high likelihood.

The randomized SCM. Let ψ_0^ρ be obtained from ψ_0 by replacing only its treatment-assignment mechanism with an independent draw $T \sim \rho$, leaving the remaining structural mechanisms unchanged. If this candidate is supported by the prior, then

$$\Phi_{\mathrm{obs}}(\psi_0^\rho) = P_{\mathrm{int}}, \qquad \Phi_{\mathrm{int}}(\psi_0^\rho) = P_{\mathrm{int}}. \tag{9}$$

It therefore explains the RCT samples exactly through its observational distribution and retains the correct interventional prediction.

A regression-style SCM. The prior may also contain a candidate ψ_{reg} with $T \sim \rho$ as an independent root node, whose remaining mechanisms satisfy

$$P^{\mathrm{obs}}_{\psi_{\mathrm{reg}}}(d\boldsymbol{x}, dt, dy) = \rho(dt)P_{\psi_0}(d\boldsymbol{x}, dy \mid \mathrm{do}(t)) = P_{\mathrm{int}}(d\boldsymbol{x}, dt, dy). \tag{10}$$

It follows that $\Phi_{\mathrm{int}}(\psi_{\mathrm{reg}}) = P_{\mathrm{int}}$ for treatment values supported by ρ.

The original SCM and intermediate assignment mechanisms. The true SCM satisfies

$$\Phi_{\mathrm{obs}}(\psi_0) = P_{\mathrm{obs}}, \qquad \Phi_{\mathrm{int}}(\psi_0) = P_{\mathrm{int}}. \tag{11}$$

Although its observational likelihood for the added RCT samples is evaluated under P_{obs}, its other structural mechanisms are the same as those generating the RCT samples. It may therefore assign substantial likelihood to these samples.

The prior may further contain candidates that differ from ψ_0 only in their treatment-assignment mechanism. For example, the influence of pretreatment covariates on treatment assignment may lie between that in ψ_0 and that in the randomized ψ_0^ρ. We denote this class of intermediate-assignment candidates by $\Psi_{\mathrm{mid}} \subseteq \Psi$. If such a candidate explains the original observational samples and assigns high observational likelihood to the added RCT samples, it may also receive high augmented posterior weight. Because the remaining structural mechanisms are unchanged, $\mathrm{do}(t)$ removes the modified assignment mechanism, giving $\Phi_{\mathrm{int}}(\psi) = P_{\mathrm{int}}$.

Under the Bayesian interpretation, CFMs such as Do-PFN and Graphs4CFM can be viewed as implicitly implementing $\mathrm{do}(t)$ by approximating the posterior average of predictions from intervened SCMs. For treatments covered by ρ, each candidate's prediction depends on ψ only through $\Phi_{\mathrm{int}}(\psi)$. We therefore study the pushforward of the SCM posterior through Φ_{int}, focusing on the mass assigned to the true interventional distribution P_{int}. The preimage of P_{int} and its complement in SCM space are

$$\mathcal{A} := \Phi_{\mathrm{int}}^{-1}(\{P_{\mathrm{int}}\}) = \{\psi \in \Psi : \Phi_{\mathrm{int}}(\psi) = P_{\mathrm{int}}\}, \qquad \mathcal{A}^c := \Psi \setminus \mathcal{A}. \tag{12}$$

As discussed above, $\mathcal{A}$ includes ψ_0 and may also contain ψ_0^ρ, ψ_{reg}, and members of Ψ_{mid}. We use $\pi_{\mathrm{obs}}(\mathcal{A})$ and $\pi_{\mathrm{aug}}(\mathcal{A})$ to denote its posterior mass before and after augmentation, respectively, equivalently the mass assigned to the true interventional distribution P_{int}.

For $\mathcal{B} \in \{\mathcal{A}, \mathcal{A}^c\}$, define the class-averaged likelihood of the added samples as

$$\bar{L}_{\mathcal{B}} := \frac{\int_{\mathcal{B}} \pi_{\mathrm{obs}}(\psi)L_{\mathrm{int}}(\psi)\,d\psi}{\pi_{\mathrm{obs}}(\mathcal{B})}. \tag{13}$$

When both classes have positive observational posterior mass,

$$\frac{\pi_{\mathrm{aug}}(\mathcal{A})}{1-\pi_{\mathrm{aug}}(\mathcal{A})} = \frac{\pi_{\mathrm{obs}}(\mathcal{A})}{1-\pi_{\mathrm{obs}}(\mathcal{A})}\,\frac{\bar{L}_{\mathcal{A}}}{\bar{L}_{\mathcal{A}^c}}. \tag{14}$$

Thus, if $\bar{L}_{\mathcal{A}} > \bar{L}_{\mathcal{A}^c}$, RCT samples increase posterior mass on the true interventional distribution. Although these class-averaged likelihoods are generally intractable under $p(\psi)$, RCT augmentation substantially improves causal effect prediction in our experiments.

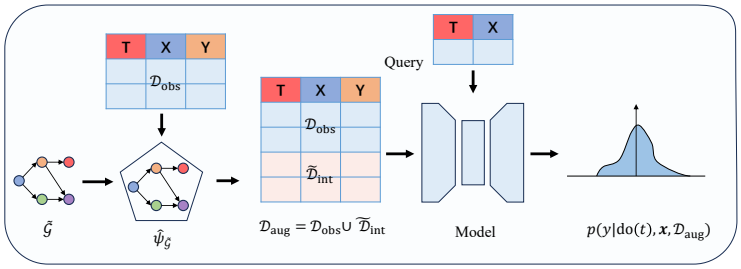


Figure 1: Overview of context augmentation with domain knowledge.

4.2 Context Augmentation with Domain Knowledge

We now turn to the setting where only observational data are available and study how to incorporate domain knowledge into pretrained CFMs. Based on the analysis in Section 4.1, our goal is to construct pseudo-interventional samples whose distribution approximates P_{int}. Figure 1 illustrates the overall workflow. When the true DAG is known and there is no unobserved confounding, the interventional distribution is identifiable from P_{obs} (Pearl, 2009). We can therefore fit an auxiliary SCM to $\mathcal{D}_{\text{obs}}$ under this DAG and use it to generate such samples. In practice, however, the true DAG is difficult to obtain, and a misspecified DAG can introduce substantial bias (Vashishtha et al., 2025). In fact, a true DAG is not required for generating the target distribution: the auxiliary SCM need only belong to the interventional equivalence class $\mathcal{A}$ defined in Eq. (12). More generally, the available knowledge need only satisfy an identification condition for the interventional distribution.

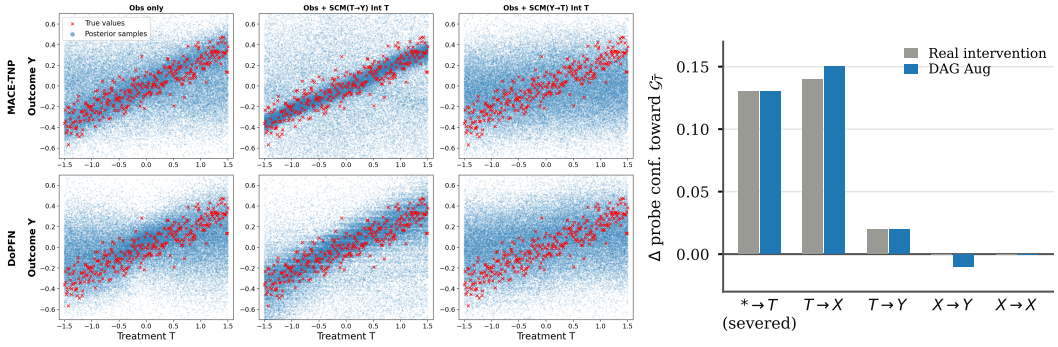
Our context augmentation framework can use any partial domain knowledge satisfying such a condition. Here, we use causal order in the back-door setting as a concrete example. Causal order contains less structural information than a complete DAG, yet is easier to obtain from domain experts or LLMs (Vashishtha et al., 2025; Kiciman et al., 2024). We represent causal order by levels: root variables with no parents are assigned to level 0, and every other variable is assigned the length of the longest directed path from any level-0 variable to it (Vashishtha et al., 2025). Let $\ell(Z_k)$ denote the level of Z_k. Under causal sufficiency, a correct causal order is sufficient to identify a valid back-door adjustment set: all variables preceding T in the order form such a set (Vashishtha et al., 2025).

Constructing an auxiliary DAG. Given the level order, we construct a knowledge-consistent DAG $\widetilde{\mathcal{G}}$ in one of two ways. The first uses the order as prior knowledge for a causal discovery algorithm, such as PC, PC-simple, or FCI (Spirtes et al., 1993; Bühlmann et al., 2010; Zhang, 2008; Vashishtha et al., 2025). Only edges directed from lower to higher levels are permitted, while the data determine each variable's parent set. The second connects every variable in a lower level to every variable in a higher level. This dense graph makes no structural assumptions beyond the order, but its high-dimensional local regressions can be difficult to fit when there are many variables. As an approximation that retains indirect pathways while being easier to fit, we also consider connecting only adjacent levels.

Fitting a knowledge-guided SCM. Given $\widetilde{\mathcal{G}}$, we fit an auxiliary SCM $\widehat{\psi}_{\widetilde{\mathcal{G}}} = (\widetilde{\mathcal{G}}, \{\widehat{f}_k\}_{k=1}^{K}, \widehat{p}(\boldsymbol{\varepsilon}))$ to $\mathcal{D}_{\text{obs}}$ under the additive-noise assumption $Z_k = f_k(Pa_{\widetilde{\mathcal{G}}}(Z_k)) + \varepsilon_k$. Each local mechanism $\widehat{f}_k$ is fitted using only the parents of Z_k specified by $\widetilde{\mathcal{G}}$, and the noise distribution is estimated from the residuals. Root variables are modeled by their empirical marginal distributions. In our implementation, the SCM is instantiated with DoWhy, and its local mechanisms are fitted using either DoWhy's built-in autofit Sharma & Kiciman (2020); Blöbaum et al. (2024) or TabPFN Hollmann et al. (2025) as the supervised predictor.

Generating pseudo-interventional context. We generate $M = rN$ pseudo-interventional samples, where r controls the augmentation strength. For each sample, we draw $\widetilde{t}^{(j)} \sim \rho$, apply $\text{do}(T = \widetilde{t}^{(j)})$ to $\widehat{\psi}_{\widetilde{\mathcal{G}}}$, and sample the remaining variables from the intervened SCM:

$$(\widetilde{\boldsymbol{x}}^{(j)}, \widetilde{y}^{(j)}) \sim p_{\widehat{\psi}_{\widetilde{\mathcal{G}}}}\big(\boldsymbol{x}, y \mid \text{do}(T = \widetilde{t}^{(j)})\big). \tag{15}$$



(a) Posterior predictive under three contexts.

(b) Probing causal structure (Do-PFN).

Figure 2: Mechanistic analysis. (a) Posterior predictive samples (blue) vs. ground-truth interventional samples (red) in a bivariate non-identifiable setting ($T \sim \mathcal{N}(0,1)$, $Y = 0.25T + \varepsilon$, $\varepsilon \sim \mathcal{N}(0, 0.1^2)$). Left: observational context only; middle/right: context augmented with pseudo-interventional rows from a fitted SCM oriented $T \to Y$ / $Y \to T$. (b) Change in probe confidence toward the intervened graph $\mathcal{G}_{\bar{T}}$ relative to the observational-only context, after augmentation with either real interventional data or DAG-guided pseudo-interventional data.

This yields $\widetilde{\mathcal{D}}_{\text{int}} = \{(\widetilde{\boldsymbol{x}}^{(j)}, \widetilde{t}^{(j)}, \widetilde{y}^{(j)})\}_{j=1}^{M}$. We concatenate the datasets into $\mathcal{D}_{\text{aug}} = \mathcal{D}_{\text{obs}} \cup \widetilde{\mathcal{D}}_{\text{int}}$ and pass the augmented context to the pretrained CFM q_θ without source labels. Its prediction is

$$\widehat{p}(y_q \mid \text{do}(t_q), \boldsymbol{x}_q) = q_\theta\big(y_q \mid \text{do}(t_q), \boldsymbol{x}_q, \mathcal{D}_{\text{aug}}\big). \tag{16}$$

The procedure requires no retraining and applies to CFMs that take a set of $(\boldsymbol{x}, t, y)$ triplets as context.

Sources of error. The error of the interventional prediction can arise from two sources.

Pseudo-interventional versus real interventional samples. The pseudo-interventional samples may deviate from the true interventional distribution: mechanisms fitted with small N may have high variance, the additive-noise assumption may be violated, and the domain knowledge, whether a full DAG or a causal order, may be misspecified. Even when $\widehat{\psi}_{\widetilde{\mathcal{G}}}$ is imperfect, augmentation may still help: a pretrained CFM may use statistical structure in the augmented context, such as conditional-independence relations, to infer the causal structure implicitly. In the experiments, we probe how the causal information encoded in CFM representations changes before and after augmentation.

Residual posterior mass. As shown by the posterior-odds update in Eq. (7), the added samples reweight candidates according to their likelihoods. With a finite number of added samples, observationally plausible candidates outside $\mathcal{A}$ may retain posterior mass, so the prediction can combine contributions from both classes. The choice of r therefore balances the influence of pseudo-interventional samples generated from a potentially biased $\widehat{\psi}_{\widetilde{\mathcal{G}}}$ against the posterior mass remaining on candidates that fit P_{obs} but lie outside $\mathcal{A}$.

5 Experiment

We evaluate the effectiveness of context augmentation through mechanistic analysis and predictive performance. Our experiments include synthetic datasets augmented using the true DAG and semi-synthetic datasets augmented using a level order.

Can structural information help break observational equivalence? As shown in Figure 2a, with observational context alone, the posterior predictive samples of Do-PFN and MACE-TNP reflect non-identifiability, with probability mass associated with both the true causal direction $T \to Y$ and the alternative direction $Y \to T$. After augmenting the context with pseudo-interventional samples generated by an SCM fitted under $T \to Y$, the predictions concentrate closer to the true interventional distribution, although a small amount of the predictive pattern associated with $Y \to T$ remains.

Does augmentation improve the model's internal representation of causal structure? Following Swelam et al. (2025), we train probes to read structural information from Do-PFN representations.

Table 1: Results on synthetic datasets drawn from the Graphs4CFM prior. We report the mean $\pm$ standard error, and Rank is the average rank over the three metrics. The best result in each column is in **bold** and the second best is underlined. Graphs4CFM† conditions on the true DAG. The augmentation variants fit an SCM with the true DAG and add pseudo-observational, -interventional, or -counterfactual rows at a ratio of r=1.0.

Method	MSE ↓	R^2 ↑	|ATE bias| ↓	Rank ↓
Graphs4CFM (no graph)	0.058$\pm$0.012	0.597$\pm$0.017	0.052$\pm$0.004	5.67
DoWhy (SCM)	0.088$\pm$0.014	0.411$\pm$0.034	0.050$\pm$0.005	7.33
Ensemble	0.061$\pm$0.012	0.588$\pm$0.017	0.044$\pm$0.003	6.17
Graphs4CFM† (true DAG)	**0.055**$\pm$0.012	0.610$\pm$0.017	0.041$\pm$0.003	3.33
Graphs4CFM + Obs Aug	0.059$\pm$0.012	0.588$\pm$0.017	0.056$\pm$0.004	6.83
Graphs4CFM + Int Aug	0.056$\pm$0.012	**0.617**$\pm$0.016	**0.039**$\pm$0.003	**2.67**
Graphs4CFM + CF Aug	0.056$\pm$0.012	0.614$\pm$0.016	**0.039**$\pm$0.003	3.00
Graphs4CFM + RCT Aug	0.052$\pm$0.012	0.636$\pm$0.016	0.034$\pm$0.002	1.00

Table 2: Results on synthetic datasets in the backdoor setting. We report the mean $\pm$ standard error. The augmentation variants fit an SCM with the true DAG and add pseudo-observational, -interventional, or -counterfactual rows at a ratio of r=1.0. The best result in each column is in **bold** and the second best is underlined.

Method	PEHE ↓	|ATE bias| ↓
CausalPFN	0.149$\pm$0.0075	0.051$\pm$0.0035
DoWhy (SCM)	0.095$\pm$0.0062	0.029$\pm$0.0018
Ensemble	0.101$\pm$0.0050	0.034$\pm$0.0021
CausalPFN + Obs Aug	0.123$\pm$0.0060	0.037$\pm$0.0023
CausalPFN + Int Aug	0.088$\pm$0.0042	0.029$\pm$0.0018
CausalPFN + CF Aug	**0.082**$\pm$0.0042	**0.028**$\pm$0.0017
CausalPFN + RCT Aug	0.045$\pm$0.0019	0.007$\pm$0.00033

As shown in Figure 2b, after adding pseudo-interventional rows, probe confidence in the removal of incoming edges to T increases substantially relative to the observational-only context. Confidence in outgoing edges from T, including $T \to X$ and $T \to Y$, also increases. These changes indicate a stronger representation of the local structure around T. They are consistent with the result that interventional data can help determine edge directions around the intervention target (Hauser & Bühlmann, 2012), although that theory assumes that observational and interventional data are distinguished. The comparison with genuine interventional data further suggests that structural information can be conveyed through context even when the auxiliary SCM's mechanisms are fitted imperfectly. Corresponding probing results for Graphs4CFM are reported in Appendix A.6.2.

How does augmentation compare with explicit graph conditioning? Table 1 reports results on datasets drawn from the Graphs4CFM pretraining prior, where the model is expected to approximate its corresponding Bayes-optimal predictor. Its true-DAG-conditioned variant outperforms the version without graph information. After interventional or counterfactual context augmentation, Graphs4CFM without graph information improves and achieves performance competitive with the explicitly conditioned variant, Graphs4CFM†. In contrast, observational augmentation slightly degrades performance, potentially because it introduces errors from the fitted SCM.

Does augmentation further improve CausalPFN, a model with an identifiable prior? In the backdoor setting, context augmentation improves CausalPFN's predictions (Table 2). Unlike the Graphs4CFM results, observational augmentation also helps CausalPFN, possibly because the additional DoWhy-generated samples improve effective treatment overlap. The augmented CausalPFN matches or slightly outperforms direct DoWhy prediction. Genuine interventional samples produce

Table 3: Results on three semi-synthetic datasets and the Causal Chamber dataset. We report the mean ± standard error. Methods in the lower block have access to causal structure. The SCM is fitted using the causal order. Int Aug adds pseudo-interventional augmented samples at a ratio of r=0.5. For each augmentation variant, the second line reports the relative change over its base model.

	PEHE ↓				**ATE Relative Error ↓**			
Method	Sales ($\times 10^4$)	Chamber	PSID ($\times 10^3$)	CPS ($\times 10^3$)	Sales	Chamber	PSID	CPS
S-Learner	108±1.1	110±4.2	19.6±0.26	11.8±0.05	0.857±0.01	0.206±0.014	0.599±0.02	0.828±0.01
T-Learner	39.9±1.4	80.2±4.9	**13.8**±0.20	9.21±0.04	0.150±0.01	**0.066**±0.012	0.098±0.01	0.281±0.02
X-Learner	62.5±1.1	115±15	19.7±0.50	12.6±0.21	0.470±0.02	0.128±0.019	0.664±0.04	0.937±0.04
DML	148±11	218±3.2	22.4±0.16	12.2±0.09	1.04±0.04	0.589±0.011	0.756±0.01	**0.229**±0.02
CausalFM	105±0.31	205±2.1	18.1±0.19	10.9±0.10	0.837±0.0041	0.686±0.0068	0.549±0.02	0.569±0.03
Do-PFN	34.7±1.6	179±2.2	20.9±0.14	12.3±0.04	0.182±0.03	0.567±0.009	0.927±0.01	0.961±0.01
CausalPFN	58.0±1.6	91.1±3.1	14.4±0.20	**9.19**±0.03	0.433±0.03	0.211±0.013	0.222±0.02	0.314±0.02
Graphs4CFM†	75.8±2.4	209±3.4	19.7±0.20	11.2±0.10	0.591±0.01	0.517±0.01	0.650±0.02	0.700±0.02
DoWhy (SCM)	35.2±13	65.0±3.8	18.6±0.14	11.2±0.04	**0.041**±0.017	0.083±0.012	0.142±0.01	0.419±0.02
Do-PFN + Int Aug	**27.7**±3.3	124±1.9	17.4±0.16	11.4±0.02	0.075±0.019	0.302±0.009	0.202±0.01	0.258±0.01
vs. Do-PFN	↓20.3%	↓30.8%	↓16.9%	↓7.8%	↓58.8%	↓46.7%	↓78.2%	↓73.2%
CausalPFN + Int Aug	34.4±1.8	**61.3**±3.7	13.9±0.17	9.44±0.04	0.163±0.012	0.087±0.013	**0.080**±0.01	0.409±0.02
vs. CausalPFN	↓40.7%	↓32.7%	↓3.2%	↑2.7%	↓62.4%	↓58.8%	↓64.0%	↑30.3%

much larger gains in both synthetic benchmarks. Median results with error bars are reported in the appendix.

How does augmentation perform on semi-synthetic benchmarks? Table 3 shows that augmentation improves Do-PFN's PEHE and ATE relative error on all four benchmarks. On PSID and CPS, its augmented ATE relative error is even lower than that of the setting-specific CausalPFN. CausalPFN also improves on three benchmarks, but CPS is a failure case. Upon inspection, only 1.03% of its samples are treated, indicating severe treatment imbalance and raising concerns about overlap. Pseudo-interventional rows from the poorly fitted auxiliary SCM then dominate the treated samples in the augmented context and degrade CausalPFN's performance.

6 Conclusion

We studied how pretrained CFMs can use interventional data and structural knowledge through test-time context augmentation. Under an idealized Bayesian analysis, we derived a condition under which interventional samples supplied without source labels increase posterior mass on SCMs that induce the true interventional distribution. When real interventional data are unavailable, our framework fits an auxiliary SCM constrained by domain knowledge such as a DAG or a causal order and adds pseudo-interventional samples to the context. Mechanistic analysis showed that augmentation breaks observational symmetry in a non-identifiable bivariate setting and increases the information about the local structure around the treatment encoded in CFM representations. On synthetic data drawn from its pretraining prior, augmentation enables graph-free Graphs4CFM to match its graph-conditioned counterpart. On synthetic and semi-synthetic benchmarks, context augmentation improves Do-PFN and CausalPFN in most settings. These results suggest that context augmentation offers a practical way to use structural knowledge with pretrained CFMs without retraining.

Limitations. Our analysis is based on an idealized Bayesian update. Although the CFM training objective targets the posterior predictive distribution under the pretraining prior, a pretrained model only approximates this predictor and need not implement it exactly. The analysis also represents an interventional query as applying $\mathrm{do}(t)$ to each candidate SCM before posterior averaging. Empirical studies support the ability of CFMs to answer interventional queries (Robertson et al., 2026; Reuter et al., 2026; Zhu et al., 2026), but do not establish that they implement this mechanism exactly. Moreover, augmentation reweights rather than eliminates observationally plausible candidates that imply incorrect interventional distributions.

AI USE STATEMENT

In this work, we used generative AI tools to assist with writing and refining code for synthetic dataset generation and experimental implementation. We used AI to propose causal orders for the LaLonde–PSID and LaLonde–CPS benchmarks based on variable descriptions. AI tools also assisted with formalizing mathematical arguments, checking author-written proofs, and translating, editing, and proofreading the manuscript. The central mathematical claims and proof ideas were developed by the authors. We reviewed and tested AI-assisted code, checked the generated datasets and proposed causal orders against the intended experimental setup, and verified the mathematical arguments and manuscript text. We take responsibility for all methods, results, claims, and content in this paper.

REPRODUCIBILITY STATEMENT

To facilitate reproducibility, Appendix A.3 documents dataset generation and processing, including the use of the publicly available Graphs4CFM data generator. Appendix A.4 describes the comparison methods, Appendix A.5 details auxiliary SCM fitting and context augmentation, Appendix A.6 specifies the probing protocols, and Appendix A.7 defines the evaluation metrics. We report means and standard errors across all benchmarks, as well as 95% bootstrap confidence intervals for median performance on the two synthetic benchmarks (Figures 4 and 5). Code for selected experiments is provided as supplementary material; we plan to release the complete code upon acceptance.

REFERENCES

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

Patrick Blöbaum, Peter Götz, Kailash Budhathoki, Atalanti A Mastakouri, and Dominik Janzing. Dowhy-gcm: An extension of dowhy for causal inference in graphical causal models. *Journal of Machine Learning Research*, 25(147):1–7, 2024.

P. Bühlmann, M. Kalisch, and M. H. Maathuis. Variable selection in high-dimensional linear models: partially faithful distributions and the pc-simple algorithm. *Biometrika*, 97(2):261–278, 06 2010. ISSN 0006-3444. doi: 10.1093/biomet/asq008. URL https://doi.org/10.1093/biomet/asq008.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters, 2018.

Anish Dhir, Cristiana Diaconu, Valentinian Mihai Lungu, James Requeima, Richard E Turner, and Mark van der Wilk. Estimating interventional distributions with uncertain causal graphs through meta-learning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Frederick Eberhardt, Clark Glymour, and Richard Scheines. On the number of experiments sufficient and in the worst case necessary to identify all causal relations among n variables. In *Proceedings of the Twenty-First Conference on Uncertainty in Artificial Intelligence*, pp. 178–184, 2005.

Juan L. Gamella, Jonas Peters, and Peter Bühlmann. Causal chambers as a real-world physical testbed for AI methodology. *Nature Machine Intelligence*, 2025. doi: 10.1038/s42256-024-00964-x.

Alain Hauser and Peter Bühlmann. Characterization and greedy learning of interventional markov equivalence classes of directed acyclic graphs. *The Journal of Machine Learning Research*, 13(1): 2409–2464, 2012.

Alain Hauser and Peter Bühlmann. Jointly interventional and observational data: estimation of interventional markov equivalence classes of directed acyclic graphs. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 77(1):291–318, 2015.

Miguel A Hernán and James M Robins. Causal inference, 2010.

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

Noah Hollmann, Samuel Müller, Lennart Purucker, Arjun Krishnakumar, Max Körfer, Shi Bin Hoo, Robin Tibor Schirrmeister, and Frank Hutter. Accurate predictions on small data with a tabular foundation model. *Nature*, 637(8045):319–326, 2025.

Guido W Imbens and Donald B Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge university press, 2015.

Emre Kiciman, Robert Ness, Amit Sharma, and Chenhao Tan. Causal reasoning and large language models: Opening a new frontier for causality. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=mqoxLkX210. Featured Certification.

Sören R Künzel, Jasjeet S Sekhon, Peter J Bickel, and Bin Yu. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the national academy of sciences*, 116(10):4156–4165, 2019.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33: 9459–9474, 2020.

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

David Madigan and Adrian E Raftery. Model selection and accounting for model uncertainty in graphical models using occam's window. *Journal of the American Statistical Association*, 89(428): 1535–1546, 1994.

Brady Neal, Chin-Wei Huang, and Sunand Raghupathi. Realcause: Realistic causal inference benchmarking. *arXiv preprint arXiv:2011.15007*, 2020.

Judea Pearl. *Causality*. Cambridge university press, 2009.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: foundations and learning algorithms*. The MIT Press, 2017.

Arik Reuter, Anish Dhir, Cristiana Diaconu, Jake Robertson, Ole Ossen, Frank Hutter, Adrian Weller, Mark van der Wilk, and Bernhard Schölkopf. Use what you know: Causal foundation models with partial graphs. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=l1PVCATRCx.

Jake Robertson, Arik Reuter, Siyuan Guo, Noah Hollmann, Frank Hutter, and Bernhard Schölkopf. Do-PFN: In-context learning for causal effect estimation. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026. URL https://openreview.net/forum?id=OaNbl9b56B.

Andreas Sauter, Saber Salehkaleybar, Aske Plaat, and Erman Acar. Activa: Amortized causal effect estimation without graphs via transformer-based variational autoencoder. *arXiv preprint arXiv:2503.01290*, 2025.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools. *Advances in neural information processing systems*, 36:68539–68551, 2023.

Amit Sharma and Emre Kiciman. Dowhy: An end-to-end library for causal inference. *arXiv preprint arXiv:2011.04216*, 2020.

Ilya Shpitser and Judea Pearl. Identification of conditional interventional distributions. In *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence*, pp. 437–444, 2006.

P. Spirtes, C. Glymour, and R. Scheines. *Causation, prediction, and search*. Springer-Verlag. (2nd edition MIT Press 2000), 1993.

Omar Swelam, Lennart Purucker, Jake Robertson, Hanne Raum, Joschka Boedecker, and Frank Hutter. Does tabPFN understand causal structures? In *EurIPS 2025 Workshop: AI for Tabular Data*, 2025. URL https://openreview.net/forum?id=A8n3nvAYWl.

Aniket Vashishtha, Abbavaram Gowtham Reddy, Abhinav Kumar, Saketh Bachu, Vineeth N. Balasubramanian, and Amit Sharma. Causal order: The key to leveraging imperfect experts in causal inference. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=9juyeCqL0u.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

Karren Yang, Abigail Katcoff, and Caroline Uhler. Characterizing and learning equivalence classes of causal dags under interventions. In *International Conference on Machine Learning*, pp. 5541–5550. PMLR, 2018.

Jiji Zhang. On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias. *Artificial Intelligence*, 172(16-17):1873–1896, 2008.

Qiong Zhang, Yan Shuo Tan, Qinglong Tian, and Pengfei Li. Tabpfn: One model to rule them all? *arXiv preprint arXiv:2505.20003*, 2025a.

Yufeng Zhang, Fengzhuo Zhang, Zhuoran Yang, and Zhaoran Wang. What and how does in-context learning learn? bayesian model averaging, parameterization, and generalization. In *International Conference on Artificial Intelligence and Statistics*, pp. 1684–1692. PMLR, 2025b.

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

Max Zhu, Martino Mansoldo, Ching-Hao Wang, and Stefan Groha. A causal foundation model for structure and outcome prediction. In *2nd ICML Workshop on Foundation Models for Structured Data*, 2026. URL https://openreview.net/forum?id=GOf9c4lOCf.

A Appendix

A.1 Notation

Table 4 summarizes the notation used throughout the paper. Uppercase italic letters denote random variables (e.g., T, Y, Z_k), and the corresponding lowercase letters denote their values (e.g., t, y). Bold lowercase letters denote vectors (e.g., $\boldsymbol{x}$), bold italic uppercase letters denote matrices (e.g., $\boldsymbol{A}$), and calligraphic letters denote graphs, datasets, and other sets; $\mathcal{N}$ and $\mathcal{U}$ denote the Gaussian and uniform distributions, with $\mathcal{N}(\mu, \sigma^2)$ parameterized by its variance. A superscript (i) or (j) indexes samples, and a subscript q marks a query. A tilde marks objects constructed from domain knowledge or pseudo-interventions (e.g., $\widetilde{\mathcal{G}}$, $\widetilde{\mathcal{D}}_{\text{int}}$), and a hat marks quantities fitted to data or predicted by a model (e.g., $\widehat{\psi}_{\widetilde{\mathcal{G}}}$, $\hat{\tau}$).

A.2 Derivation of the Posterior Updates

We derive the two identities in Eq. (7) and the posterior odds of $\mathcal{A}$ in Section 4.1. In the idealized analysis, given ψ, every row of $\mathcal{D}_{\text{aug}} = \mathcal{D}_{\text{obs}} \cup \mathcal{D}_{\text{int}}$ is treated as an independent draw from P_ψ^{obs}. The likelihoods of the two parts of the context are

$$L_{\text{obs}}(\psi) := \prod_{\boldsymbol{z}^{(i)} \in \mathcal{D}_{\text{obs}}} p_\psi^{\text{obs}}(\boldsymbol{z}^{(i)}), \qquad L_{\text{int}}(\psi) := \prod_{\boldsymbol{z}^{(j)} \in \mathcal{D}_{\text{int}}} p_\psi^{\text{obs}}(\boldsymbol{z}^{(j)}), \tag{17}$$

and the likelihood of $\mathcal{D}_{\text{aug}}$ is $L_{\text{obs}}(\psi) L_{\text{int}}(\psi)$. Densities on Ψ are taken with respect to the base measure $d\psi$, and we assume

$$Z_{\text{obs}} := \int_\Psi p(\psi) L_{\text{obs}}(\psi)\, d\psi \in (0, \infty), \qquad Z_{\text{aug}} := \int_\Psi \pi_{\text{obs}}(\psi) L_{\text{int}}(\psi)\, d\psi \in (0, \infty). \tag{18}$$

Step 1: observational posterior. By Bayes' rule,

$$\pi_{\text{obs}}(\psi) = \frac{p(\psi) L_{\text{obs}}(\psi)}{Z_{\text{obs}}}, \qquad \text{i.e.,} \qquad p(\psi) L_{\text{obs}}(\psi) = Z_{\text{obs}}\, \pi_{\text{obs}}(\psi) \quad \text{for every } \psi \in \Psi. \tag{19}$$

Step 2: augmented posterior. Applying Bayes' rule to $\mathcal{D}_{\text{aug}}$ gives

$$\pi_{\text{aug}}(\psi) = \frac{p(\psi) L_{\text{obs}}(\psi) L_{\text{int}}(\psi)}{\int_\Psi p(\psi') L_{\text{obs}}(\psi') L_{\text{int}}(\psi')\, d\psi'}. \tag{20}$$

Substituting Eq. (19) into the numerator, and pointwise in ψ' into the integrand of the denominator, yields

$$\begin{aligned} \pi_{\text{aug}}(\psi) &= \frac{Z_{\text{obs}}\, \pi_{\text{obs}}(\psi) L_{\text{int}}(\psi)}{\int_\Psi Z_{\text{obs}}\, \pi_{\text{obs}}(\psi') L_{\text{int}}(\psi')\, d\psi'} = \frac{Z_{\text{obs}}\, \pi_{\text{obs}}(\psi) L_{\text{int}}(\psi)}{Z_{\text{obs}} \int_\Psi \pi_{\text{obs}}(\psi') L_{\text{int}}(\psi')\, d\psi'} \\ &= \frac{\pi_{\text{obs}}(\psi) L_{\text{int}}(\psi)}{Z_{\text{aug}}}, \end{aligned} \tag{21}$$

where Z_{obs} can be taken outside the integral because it does not depend on ψ'. This is the first identity in Eq. (7).

Step 3: pairwise posterior ratio. Let $\psi_a, \psi_b \in \Psi$ with $\pi_{\text{obs}}(\psi_b) L_{\text{int}}(\psi_b) > 0$. Taking the ratio of Eq. (21) at ψ_a and ψ_b cancels Z_{aug}:

$$\frac{\pi_{\text{aug}}(\psi_a)}{\pi_{\text{aug}}(\psi_b)} = \frac{\pi_{\text{obs}}(\psi_a)}{\pi_{\text{obs}}(\psi_b)} \cdot \frac{L_{\text{int}}(\psi_a)}{L_{\text{int}}(\psi_b)}, \tag{22}$$

which is the second identity in Eq. (7).

Table 4: Summary of notation.

Symbol	Description
Variables and data	
$Z_1, \dots, Z_K, \mathcal{V}$	Random variables of an SCM and their set $\mathcal{V} = \{Z_1, \dots, Z_K\}$
$T, Y, X_1, \dots, X_{K-2}$	Treatment, outcome, and covariates
t, y, $\boldsymbol{x}$	Values of the treatment, the outcome, and the covariate vector
$\boldsymbol{z} = (\boldsymbol{x}, t, y)$	A sample, i.e., one row of a context
$\boldsymbol{x}_q$, t_q, y_q	Query covariates, query treatment, and target outcome
$\mathcal{D}_{\text{obs}}$	Observational dataset $\{(\boldsymbol{x}^{(i)}, t^{(i)}, y^{(i)})\}_{i=1}^{N}$
$\mathcal{D}_{\text{int}}$	Real interventional (RCT) dataset with M samples
$\widetilde{\mathcal{D}}_{\text{int}}$	Pseudo-interventional dataset $\{(\widetilde{\boldsymbol{x}}^{(j)}, \widetilde{t}^{(j)}, \widetilde{y}^{(j)})\}_{j=1}^{M}$
$\mathcal{D}_{\text{aug}}$	Augmented context: $\mathcal{D}_{\text{obs}} \cup \mathcal{D}_{\text{int}}$ (Sec. 4.1) or $\mathcal{D}_{\text{obs}} \cup \widetilde{\mathcal{D}}_{\text{int}}$ (Sec. 4.2)
N; $M = rN$; r	Number of observational samples; number of added samples; augmentation ratio
ρ	Distribution of the treatment values assigned in interventional and pseudo-interventional samples
Structural causal models	
$\psi = (\mathcal{G}, \{f_k\}_{k=1}^{K}, p(\boldsymbol{\varepsilon}))$	SCM with DAG $\mathcal{G}$, mechanisms f_k, and noise distribution $p(\boldsymbol{\varepsilon})$
$\mathcal{G} = (\mathcal{V}, \boldsymbol{A})$	DAG with adjacency matrix $\boldsymbol{A} \in \{0,1\}^{K \times K}$, where $A_{i,j} = 1$ iff $Z_i \to Z_j$
$Pa_{\mathcal{G}}(Z_k)$	Parents of Z_k in $\mathcal{G}$
ε_k, $\boldsymbol{\varepsilon}$	Exogenous noise of Z_k; $\boldsymbol{\varepsilon} = (\varepsilon_1, \dots, \varepsilon_K)$
$\mathrm{do}(T = t)$, $\mathrm{do}(t)$	Intervention that sets T to t
$\psi_{\mathrm{do}(t)}$; $\mathcal{G}_{\bar{T}}$	Intervened SCM; its graph, obtained by removing all incoming edges of T
$\tau(\boldsymbol{x})$	Conditional average treatment effect (CATE)
$\ell(Z_k)$; $\prec$	Level of Z_k in the causal order; precedence between levels
Distributions and posteriors	
P, p	A distribution and its density
P_ψ^{obs}, p_ψ^{obs}	Observational distribution and density induced by ψ
$P_\psi(\cdot \mid \mathrm{do}(t))$, $p_\psi(\cdot \mid \mathrm{do}(t))$	Interventional distribution and density induced by ψ
P_ψ^ρ	Distribution of RCT samples from ψ with $T \sim \rho$
Φ_{obs}, Φ_{int}	Maps $\psi \mapsto P_\psi^{\text{obs}}$ and $\psi \mapsto P_\psi^\rho$
ψ_0; P_{obs}, P_{int}	True SCM; $P_{\text{obs}} = \Phi_{\text{obs}}(\psi_0)$ and $P_{\text{int}} = \Phi_{\text{int}}(\psi_0)$
ψ_0^ρ, ψ_{reg}, Ψ_{mid}	Randomized, regression-style, and intermediate-assignment candidates (Sec. 4.1)
$p(\psi)$, Ψ	Prior over SCMs and its support
$\pi_{\text{obs}}(\psi)$, $\pi_{\text{aug}}(\psi)$	Posteriors $p(\psi \mid \mathcal{D}_{\text{obs}})$ and $p(\psi \mid \mathcal{D}_{\text{aug}})$
$\mathcal{A}$, $\mathcal{A}^c$	Candidates with $\Phi_{\text{int}}(\psi) = P_{\text{int}}$ (Eq. (12)) and their complement in Ψ
$L_{\text{int}}(\psi)$; $\bar{L}_{\mathcal{B}}$	Likelihood of the added samples under P_ψ^{obs}; its π_{obs}-weighted average over $\mathcal{B} \in \{\mathcal{A}, \mathcal{A}^c\}$
$D_{\text{KL}}(P \Vert Q)$	Kullback–Leibler divergence
Models and context augmentation	
q_θ	Pretrained CFM with parameters θ
$\widetilde{\mathcal{G}}$	Auxiliary DAG consistent with the domain knowledge
$\widehat{\psi}_{\widetilde{\mathcal{G}}}$, $\widehat{f}_k$, $\widehat{p}(\boldsymbol{\varepsilon})$	Auxiliary SCM fitted to $\mathcal{D}_{\text{obs}}$ under $\widetilde{\mathcal{G}}$, with its mechanisms and noise distribution
$\widehat{p}(y_q \mid \mathrm{do}(t_q), \boldsymbol{x}_q)$	Prediction of q_θ given the augmented context $\mathcal{D}_{\text{aug}}$
$\tau^{(i)}$, $\hat{\tau}^{(i)}$; N_{test}	True and predicted effect of test unit i; number of test queries (Appendix A.7)

Step 4: posterior odds of $\mathcal{A}$. Assume Φ_{int} is measurable, so that $\mathcal{A}$ in Eq. (12) and $\mathcal{A}^c$ are measurable subsets of Ψ. For $\pi \in \{\pi_{\text{obs}}, \pi_{\text{aug}}\}$, write $\pi(\mathcal{B}) := \int_{\mathcal{B}} \pi(\psi)\, d\psi$. For $\mathcal{B} \in \{\mathcal{A}, \mathcal{A}^c\}$ with $\pi_{\text{obs}}(\mathcal{B}) > 0$, the class-averaged likelihood is

$$\bar{L}_{\mathcal{B}} = \int_{\mathcal{B}} \frac{\pi_{\text{obs}}(\psi)}{\pi_{\text{obs}}(\mathcal{B})}\, L_{\text{int}}(\psi)\, d\psi, \tag{23}$$

that is, the expectation of L_{int} under the observational posterior restricted to $\mathcal{B}$. Integrating Eq. (21) over $\mathcal{B}$ gives

$$\pi_{\text{aug}}(\mathcal{B}) = \frac{1}{Z_{\text{aug}}} \int_{\mathcal{B}} \pi_{\text{obs}}(\psi) L_{\text{int}}(\psi)\, d\psi = \frac{\pi_{\text{obs}}(\mathcal{B})\, \bar{L}_{\mathcal{B}}}{Z_{\text{aug}}}. \tag{24}$$

Since $\mathcal{A}$ and $\mathcal{A}^c$ partition Ψ, $Z_{\text{aug}} = \pi_{\text{obs}}(\mathcal{A})\bar{L}_{\mathcal{A}} + \pi_{\text{obs}}(\mathcal{A}^c)\bar{L}_{\mathcal{A}^c}$. Assume $\pi_{\text{obs}}(\mathcal{A}) \in (0, 1)$ and $\bar{L}_{\mathcal{A}^c} > 0$. Taking the ratio of Eq. (24) for $\mathcal{A}$ and $\mathcal{A}^c$ cancels Z_{aug}. Using $\pi(\mathcal{A}^c) = 1 - \pi(\mathcal{A})$ for both posteriors, we obtain

$$\frac{\pi_{\text{aug}}(\mathcal{A})}{1 - \pi_{\text{aug}}(\mathcal{A})} = \frac{\pi_{\text{obs}}(\mathcal{A})}{1 - \pi_{\text{obs}}(\mathcal{A})} \cdot \frac{\bar{L}_{\mathcal{A}}}{\bar{L}_{\mathcal{A}^c}}. \tag{25}$$

The map $u \mapsto u/(1-u)$ is strictly increasing on $[0, 1)$. Hence $\pi_{\text{aug}}(\mathcal{A}) > \pi_{\text{obs}}(\mathcal{A})$ if and only if $\bar{L}_{\mathcal{A}} > \bar{L}_{\mathcal{A}^c}$, with equality if and only if $\bar{L}_{\mathcal{A}} = \bar{L}_{\mathcal{A}^c}$.

A.3 Datasets and Data Processing

General-purpose synthetic datasets. Our general-purpose synthetic benchmark is generated from the prior used by Graphs4CFM Reuter et al. (2026). Graphs4CFM publicly provides its pretraining data-generation procedure, allowing us to generate evaluation datasets from the same prior. Thus, both its graph-conditioned and unconditioned variants are evaluated in their in-prior regime, where they are expected to closely approximate their corresponding Bayes-optimal predictors.

We consider graphs with 5, 10, 15, and 20 nodes and generate 120 datasets for each graph size. For continuous treatment, the treatment effect is defined as the difference in mean outcomes between the upper and lower quartiles of the interventional treatment distribution. Letting q_{25} and q_{75} denote the 25th and 75th percentiles of T, we compare samples with $T \leq q_{25}$ and $T \geq q_{75}$. A dataset is discarded if either group contains fewer than 20 samples. After removing 75 datasets with degenerate treatment distributions, 405 datasets remain. Each dataset contains 512 context samples and 1,000 test samples.

Backdoor synthetic datasets. To verify that CFMs tailored to specific causal settings can also benefit from context augmentation, we construct a second synthetic benchmark in which the causal effect is identifiable via backdoor adjustment. Each graph contains eight variables: treatment T, outcome Y, and six covariates $X_1, \ldots, X_6$. The edge $T \to Y$ is always present. Each covariate is assigned one of four roles: a confounder c ($X \to T$ and $X \to Y$), an instrument-like variable i ($X \to T$ only), a prognostic variable p ($X \to Y$ only), or a noise variable n with no incident edges. We consider five graph configurations:

Table 5: Graph configurations used in the backdoor synthetic benchmark, where c denotes a confounder ($X \to T$, $X \to Y$), i an instrument-like variable ($X \to T$), p a prognostic variable ($X \to Y$), and n a noise variable with no incident edges.

Setting	c	i	p	n
c2n4	X_1, X_2	–	–	$X_3, \ldots, X_6$
c2i2n2	X_1, X_2	X_3, X_4	–	X_5, X_6
c2p2n2	X_1, X_2	–	X_5, X_6	X_3, X_4
c2i2p2	X_1, X_2	X_3, X_4	X_5, X_6	–
c4n2	$X_1, \ldots, X_4$	–	–	X_5, X_6

The structural mechanisms follow the mechanism family used in the Do-PFN pretraining prior. Each root covariate is independently sampled as $X_j \sim \mathcal{N}(0, 2^2)$. For a non-root variable Z_k with $m_k =$

$|Pa_{\mathcal{G}}(Z_k)|$ parents, we use $Z_k = \mathrm{act}(\boldsymbol{w}_k^\top Pa_{\mathcal{G}}(Z_k)) + \varepsilon_k$, where $\boldsymbol{w}_k \sim \mathcal{U}(-1/\sqrt{m_k}, 1/\sqrt{m_k})$ and $\varepsilon_k \sim \mathcal{N}(0, \sigma_\varepsilon^2)$ with $\sigma_\varepsilon = 0.05$.

We consider three mechanism variants: (1) $\mathrm{act} = \tanh$, with noise added after the activation; (2) the activation is independently sampled for each node from $\{x^2, \mathrm{ReLU}, \tanh, \mathrm{id}\}$, again with noise added after the activation; and (3) the activation is sampled from the same set but noise is added before the activation, $Z_k = \mathrm{act}(\boldsymbol{w}_k^\top Pa_{\mathcal{G}}(Z_k) + \varepsilon_k)$, deliberately violating the additive-noise assumption.

Treatment is binarized following the Do-PFN prior construction. We first generate a continuous latent treatment $T^* = f_T(Pa_{\mathcal{G}}(T)) + \varepsilon_T$, then randomly sample a threshold κ from the realized T^* values after excluding the minimum and maximum, and define $T = \mathbf{1}_{T^* \geq \kappa}$. Since ε_T is Gaussian, the propensity score $e(\boldsymbol{x}) = 1 - F_{\mathcal{N}}((\kappa - f_T(\boldsymbol{x}))/\sigma_\varepsilon)$, where $F_{\mathcal{N}}$ is the standard normal CDF, lies strictly in $(0, 1)$. We additionally require the treatment proportion to lie in $[0.10, 0.90]$ and the standard deviation of the ground-truth CATEs to satisfy $\mathrm{sd}(\tau^{(i)}) \geq 0.005$; otherwise, the dataset is regenerated.

Combining the five graph configurations with the three mechanism variants gives 15 settings. For each setting, we generate 60 datasets, resulting in 900 datasets in total. Each dataset contains 640 context samples and 160 test samples. For both synthetic benchmarks, the auxiliary SCM is fitted using the ground-truth DAG.

SALES. For the SALES benchmark, we directly use the dataset introduced in the Do-PFN paper. The causal order used to construct the auxiliary SCM is

$$\{\text{Shopping Event}\} \prec \{\text{Ad Spend}, \text{Unit Price}\} \prec \{\text{Page Views}\} \prec \{\text{Sold Units}\} \\ \prec \{\text{Revenue}, \text{Operational Cost}\} \prec \{\text{Profit}\}.$$

LaLonde–PSID and LaLonde–CPS. We use the semi-synthetic datasets released by RealCause (Neal et al., 2020), constructed from the LaLonde NSW treatment group together with either the PSID or CPS control group. For each benchmark, we use the 100 released realizations, `lalonde_{psid,cps}_sample{0..99}.csv`, which contain ground-truth individual treatment effects. The data split follows the official CausalPFN benchmark implementation.

For PSID and CPS, we obtain the causal order from variable semantics using an LLM, following the procedure of Vashishtha et al. (2025) (implemented with GPT-5.6). Both datasets share the same order across all realizations:

$$\{\text{age}, \text{black}, \text{hispanic}\} \prec \{\text{education}\} \prec \{\text{nodegree}, \text{married}\} \prec \{\text{re74}\} \prec \{\text{re75}\} \prec \{T\} \prec \{Y\}.$$

The auxiliary SCM uses an order graph obtained by fully connecting variables between adjacent levels.

For Graphs4CFM, following its graph-conditioning setup, we treat all observed covariates as confounders connecting treatment and outcome: $X_j \to T$ and $X_j \to Y$ for every X_j, together with $T \to Y$, and include no $X \to X$ edges.

Causal Chamber. We use real measurements from the Causal Chamber lighttunnel dataset `lt_spice_v1 / exp_1_low_noise` (Gamella et al., 2025). We use `green` as the observed covariate, median-binarized `ir_1` as treatment T, and `ir_2` as outcome Y. We additionally retain `current_ls_raw`, which is a child of `green`.

The ground-truth DAG is shown in Fig. 3.

The corresponding causal order is

$$\{\texttt{green}\} \prec \{T, \texttt{current_ls_raw}\} \prec \{Y\}.$$

We apply a level-constrained variant of PC-simple (Bühlmann et al., 2010). For each variable Z_k, candidate parents are restricted to variables in strictly earlier levels. Starting from conditioning sets of size zero and increasing the size up to three, a candidate Z_j is removed if some $\mathcal{S} \subseteq \mathcal{C}_k \setminus \{Z_j\}$ of that size renders Z_k and Z_j conditionally independent, where $\mathcal{C}_k$ denotes the current candidate set

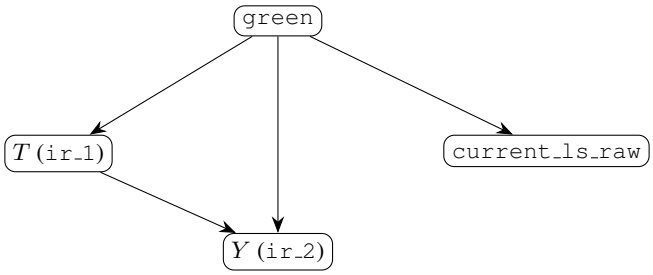


Figure 3: Ground-truth DAG of the Causal Chamber light-tunnel task.

of Z_k. Independence is assessed by Fisher-z tests implemented in causal-learn (Zheng et al., 2024) (p-value $> \alpha = 0.05$). The remaining candidates are taken as parents of Z_k, with edge directions given by the level order, and the edge $T \to Y$ is always retained.

Across all 20 seeds, the resulting graph is consistently

$$\texttt{green} \to T, \qquad \texttt{green} \to \texttt{current_ls_raw}, \qquad \texttt{current_ls_raw} \to Y, \qquad T \to Y.$$

This graph is used to fit the auxiliary SCM, and Graphs4CFM uses the same graph for explicit graph conditioning.

To construct semi-synthetic outcomes with unit-level ground-truth effects in the RealCause style, we fit a HistGradientBoosting predictor $\hat{\mu}(T, \texttt{green})$ on all real observations and define $\tau^{(i)} = \hat{\mu}(1, \texttt{green}^{(i)}) - \hat{\mu}(0, \texttt{green}^{(i)})$. Outcomes are regenerated as $y^{(i)} = \hat{\mu}(t^{(i)}, \texttt{green}^{(i)}) + \varepsilon^{(i)}$ with $\varepsilon^{(i)} \sim \mathcal{N}(0, \hat{\sigma}^2_{\text{resid}})$, where $\hat{\sigma}^2_{\text{resid}}$ is the residual variance of the predictor on the full real dataset. This homoscedastic noise preserves the magnitude of outcome variation not explained by $(T, \texttt{green})$, while $\tau^{(i)}$ does not depend on it. For each of 20 seeds (10–29), we randomly sample 2000 context rows and 200 test rows.

A.4 Comparison Methods

For all non-CFM treatment-effect estimators, namely the S-learner, T-learner, and X-learner (Künzel et al., 2019) and DML (Chernozhukov et al., 2018), we use TabPFN (Hollmann et al., 2023) as the regression or classification base learner, following prior work that pairs TabPFN with CATE meta-learners (Zhang et al., 2025a; Robertson et al., 2026).

For CausalFM (Ma et al., 2025), we follow its latest public GitHub implementation available as of September 10, 2026, together with its user documentation. We use the backdoor model variant in our experiments.

For Graphs4CFM (Reuter et al., 2026), we use the fully graph-conditioned checkpoint `full_conditioned_model`. Its structural conditioning encodes pairwise relations using $+1$ for a known ancestral relation, -1 for a known absence of such a relation, and 0 for an unknown relation. For the graph-unconditioned variant, all graph-conditioning entries are set to 0.

For Do-PFN (Robertson et al., 2026) and CausalPFN (Balazadeh et al., 2025), we use their publicly released pretrained models and open-source implementations.

For the Ensemble baseline, we average, with equal weights, the point predictions of the DoWhy (SCM) baseline and of the corresponding base CFM given only the observational context $\mathcal{D}_{\text{obs}}$.

For Dowhy (Sharma & Kiciman, 2020), we use the fitted auxiliary SCM directly as an SCM-only baseline. At prediction time, predictions are made from the pre-treatment variables using the counterfactual version of DoWhy. The SCM used by this baseline is exactly the same as that used for context augmentation.

A.5 Context Augmentation Implementation

Auxiliary SCM fitting. Given a causal graph, we construct a DoWhy-GCM `InvertibleStructuralCausalModel` under an `AdditiveNoiseModel` pa-

rameterization. We consider two mechanism-fitting procedures. The first uses `gcm.auto.assign_causal_mechanisms`, which selects causal mechanisms by cross-validation over the scikit-learn model pool. The second uses a TabPFN-based SCM: root-node distributions are represented by empirical resampling, each non-root structural mechanism is fitted using TabPFN, and the exogenous noise distribution is represented by empirical residuals. For a given experiment, the same fitted SCM is used both to generate augmented context and for the DoWhy baseline.

Context construction. Given N real context rows and a fitted auxiliary SCM, we generate $M = rN$ synthetic rows and concatenate them with the original context before passing the resulting context to the base model. For the synthetic experiments, we consider three augmentation variants:

- **obs** (observational rows): Pseudo-observational rows are sampled from the joint distribution of the fitted SCM using `gcm.draw_samples`. For binary treatment, T is first binarized according to the fitted treatment mechanism, after which all descendants of T are regenerated in topological order using fresh noise.
- **int** (interventional rows): Pseudo-interventional rows are sampled from the fitted interventional distribution using `gcm.interventional_samples`. For binary treatment, half of the augmented rows are generated under $\mathrm{do}(T = 0)$ and the other half under $\mathrm{do}(T = 1)$, i.e., ρ places equal mass on $\{0, 1\}$. For continuous treatment, the intervention value is sampled uniformly over the observed treatment range, $\widetilde{t}^{(j)} \sim \rho = \mathcal{U}[\min_i t^{(i)}, \max_i t^{(i)}]$.
- **cf** (counterfactual rows): For each real context row, we perform abduction under the invertible ANM to recover its exogenous noise variables. For binary treatment, T is switched to the opposite treatment arm and all descendants of T are recomputed while retaining the same abducted noise. For continuous treatment, the counterfactual treatment value is taken from another randomly selected context row. Pre-treatment covariates remain unchanged, so each real row is paired with a counterfactual twin sharing the same covariates.

A.6 Causal-Structure Probing

A.6.1 Do-PFN

Backbone and hooks. The Do-PFN backbone follows the PerFeature Transformer architecture of TabPFN v2, with 12 layers, hidden dimension 192, and four attention heads. Following the probing protocol of Swelam et al. (2025), we freeze the pretrained model and attach a forward hook to the output of encoder layer l. For a context with s samples and $K - 1$ input features, we extract the complete sample-by-feature token grid $\mathbf{H}_l \in \mathbb{R}^{s \times K \times 192}$, where the final column corresponds to the Y token.

Probe training. The probes are trained on an independent synthetic datasets that is disjoint from all evaluation datasets. The probe-training data follows the prior used in Appendix C of Swelam et al. (2025) , including six random graph families, linear or RFF-based nonlinear mechanisms, and Gaussian, Laplace, or Cauchy additive noise. We make the following modifications: the number of nodes is sampled from $[2, 10]$; there are no hidden variables; and node 0 is a binary treatment T generated using a logistic mechanism, with the treatment rate constrained to $[0.15, 0.85]$.

75% of the training datasets use mixed contexts consisting of one half observational rows and one half interventional rows from $\mathrm{do}(T = 0)$ and $\mathrm{do}(T = 1)$. The remaining 25% use purely observational contexts. The context length cycles over $\{200, 400, 600, 800, 1000, 1280\}$. Training data are generated online, and each SCM is used only once. We train for 30k steps, corresponding to approximately 0.5M synthetic datasets, and use a fixed set of 256 datasets for validation.

Each dataset is passed to the frozen Do-PFN with all rows used as context. At layer l, we extract sample-wise and column-wise tokens ordered as $[T, X_1, \ldots, Y]$. The probe is a lightweight cross-attention decoder. It first aggregates the sample tokens of each column into a variable-level representation $\boldsymbol{h}_i$, and then applies an asymmetric bilinear head to produce the directed-edge logits $\eta_{i,j} = U(\boldsymbol{h}_i)^\top V(\boldsymbol{h}_j)$, where $\eta_{i,j}$ corresponds to $i \to j$ and diagonal entries are masked. The target is the direct-edge adjacency matrix of the ground-truth DAG. The loss consists of positive-class-weighted binary cross entropy together with an acyclicity penalty.

We train separate probes for layers $l \in \{4, 7, 11\}$ using identical hyperparameters and select checkpoints according to validation average precision.

Evaluation protocol. The evaluation set consists of 420 datasets generated using the Do-PFN prior generator, with 5, 8, or 10 nodes. The target causal structure is the intervened graph $\mathcal{G}_{\bar{T}}$, obtained by removing all incoming edges to T.

We first apply each probe to an observational-only context containing 1280 rows and record the mean predicted probability for each edge category, including $X \to T$, $T \to X$, $T \to Y$, and other edges. This provides the observational baseline.

We then evaluate an equally sized augmented context containing 640 observational rows and 640 synthetic pseudo-interventional rows and report the paired change $\Delta = p_{\text{aug}} - p_{\text{obs}}$. As a reference, we additionally evaluate 640 observational rows combined with 640 real interventional rows. Confidence intervals are obtained by cluster bootstrap, treating each dataset as one cluster. Statistical significance is assessed using paired Wilcoxon tests with Holm correction. The main figure reports the mean result across the probes at layers $l = 4, 7, 11$.

A.6.2 GRAPHS4CFM

Graphs4CFM uses a different backbone from TabPFN, consisting of 6 two-way attention blocks with hidden dimension 256, together with a sample-attention sink row and a graph-conditioning channel. The probe implementation follows Sec. A.6.1, with minor adjustments to accommodate this architecture.

Representations are extracted from the output of the third block, selected according to validation average precision. We use only tokens corresponding to observed rows, excluding the sink row, and reorder columns as $[T, Y, X_1, \ldots]$. During representation extraction, the graph-conditioning input is set entirely to zero, so that all structural relations are treated as unknown. This matches the augmentation experiments, in which the base model receives no explicit graph.

The probe architecture is unchanged except that its representation dimension is adjusted to 256. The probe-training data is generated from the Graphs4CFM pretraining prior, using graphs with 3–10 nodes and no hidden edges. The data are generated online, each SCM is used only once, and approximately 216k synthetic datasets are used for training.

A.7 METRICS

General-purpose synthetic benchmark (continuous T).

- $\text{MSE} = \frac{1}{N_{\text{test}}} \sum_i (\hat{y}^{(i)} - y^{(i)})^2$, where N_{test} is the number of test queries.
- $R^2 = 1 - \dfrac{\sum_i (\hat{y}^{(i)} - y^{(i)})^2}{\sum_i (y^{(i)} - \bar{y})^2}$, where $\bar{y}$ is the mean of the true y on the interventional test set.
- |ATE bias|: let q_{25}, q_{75} be the quantiles of the queried T, $\mathcal{H} = \{i : t^{(i)} \geq q_{75}\}$, $\mathcal{L} = \{i : t^{(i)} \leq q_{25}\}$, $\text{ATE}_{\text{true}} = \bar{y}_{\mathcal{H}} - \bar{y}_{\mathcal{L}}$, $\text{ATE}_{\text{pred}} = \bar{\hat{y}}_{\mathcal{H}} - \bar{\hat{y}}_{\mathcal{L}}$. We report $|\text{ATE}_{\text{pred}} - \text{ATE}_{\text{true}}|$.

Binary-T datasets.

- $\text{PEHE} = \sqrt{\frac{1}{N_{\text{test}}} \sum_i (\hat{\tau}^{(i)} - \tau^{(i)})^2}$;
- $|\text{ATE bias}| = \left| \frac{1}{N_{\text{test}}} \sum_i \hat{\tau}^{(i)} - \frac{1}{N_{\text{test}}} \sum_i \tau^{(i)} \right|$ (backdoor synthetic benchmark);
- ATE relative error $|\widehat{\text{ATE}} - \text{ATE}|/|\text{ATE}|$ (semi-synthetic benchmarks, following the CausalPFN / RealCause convention).

A.8 ADDITIONAL EXPERIMENTAL RESULTS

Since the standard errors on the synthetic benchmarks in the main text are relatively large, we additionally report the median performance with 95% bootstrap confidence intervals in Figure 4 (Graphs4CFM prior) and Figure 5 (backdoor setting).

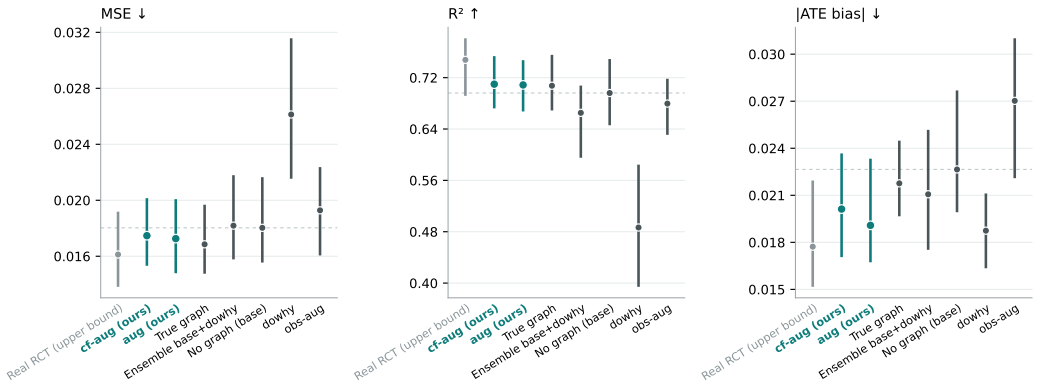


Figure 4: Median performance on synthetic datasets drawn from the Graphs4CFM prior. We report MSE (↓), R^2 (↑), and absolute ATE bias (↓). Points denote median performance across datasets, and error bars represent 95% bootstrap confidence intervals for the median. *True graph* denotes the Graphs4CFM variant explicitly conditioned on the ground-truth DAG. The dashed horizontal lines indicate the performance of Graphs4CFM without graph conditioning.

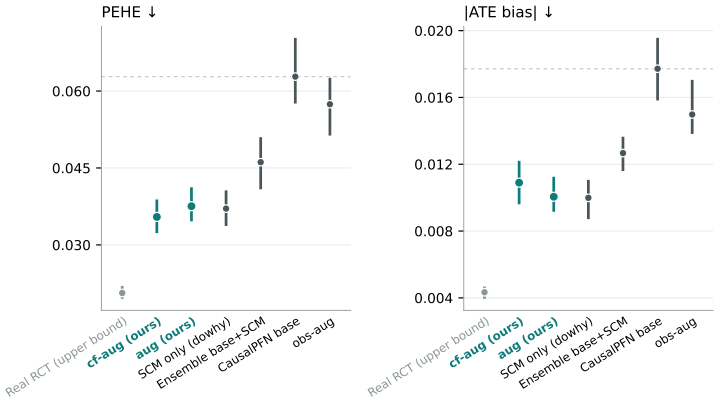


Figure 5: Median performance on synthetic datasets in the backdoor setting. Points denote median performance across datasets, with error bars showing 95% bootstrap confidence intervals for the median. The dashed horizontal lines indicate the performance of the base CausalPFN without context augmentation.

For the ablation on the augmentation ratio r, see Tables 6, 7, and 8. For the synthetic benchmarks, the auxiliary SCM is fitted using the ground-truth DAG, so we report results with $r = 1.0$ in the main text. For the semi-synthetic benchmarks, it is not possible to verify whether the structural knowledge and the fitted mechanisms are correct, so we recommend the more conservative choice $r = 0.5$.

For the probing analysis of Graphs4CFM, see Figure 7. For Graphs4CFM, context augmentation also improves the representation of the local structure around T. The corresponding Do-PFN results grouped by the number of variables are shown in Figure 6.

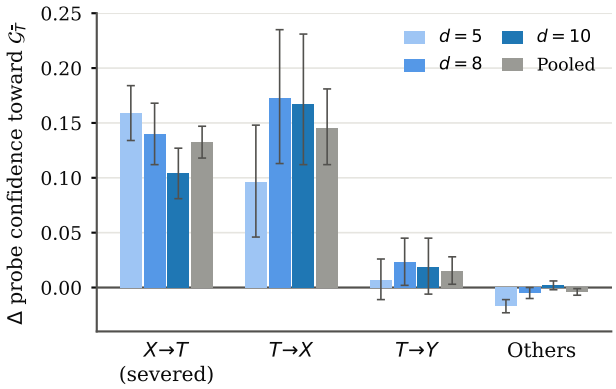


Figure 6: Probing causal structure in Do-PFN representations. Bars show the change in a "representation→DAG" probe's confidence toward the $\mathrm{do}(T)$-mutilated graph $\mathcal{G}_{\bar{T}}$ after DAG-guided context augmentation, relative to the observation-only context (averaged over layers 4/7/11 of the Do-PFN encoder), grouped by the number of variables K (denoted by d in the legend); *Pooled* aggregates all datasets. For the severed edge $X \to T$ we report the change in $1 - p(X \to T)$, so a positive value always indicates a shift toward $\mathcal{G}_{\bar{T}}$. Error bars show 95% bootstrap confidence intervals.

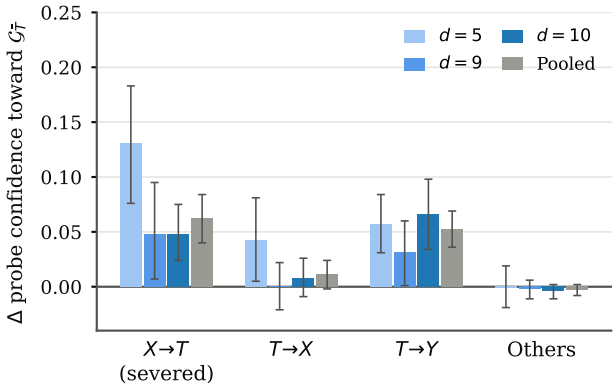


Figure 7: Probing causal structure in Graphs4CFM representations. Same setup as Figure 6, with the graph-free Graphs4CFM as the backbone.

Table 6: Effect of the augmentation ratio r on synthetic datasets drawn from the Graphs4CFM prior (setting of Table 1). We report the mean ± standard error. The best value per column among the base model and its augmented variants is in **bold**. DoWhy (SCM) is shown for reference.

Method	r	MSE ↓	R^2 ↑	|ATE bias| ↓
Graphs4CFM (no graph)	–	0.058±0.012	0.597±0.017	0.052±0.004
+ Int Aug	0.5	**0.056**±0.012	0.615±0.016	0.041±0.003
+ Int Aug	1.0	**0.056**±0.012	**0.617**±0.016	**0.039**±0.003
+ Int Aug	1.5	**0.056**±0.012	0.613±0.016	0.040±0.003
+ Int Aug	2.0	**0.056**±0.012	0.612±0.016	**0.039**±0.003
DoWhy (SCM)	–	0.088±0.014	0.411±0.034	0.050±0.005

Table 7: Effect of the augmentation ratio r on synthetic datasets in the backdoor setting (setting of Table 2). We report the mean $\pm$ standard error. Within each augmentation variant, the best value per column is in **bold**. DoWhy (SCM) is shown for reference.

Method	r	PEHE ↓	|ATE bias| ↓
CausalPFN	–	0.149$\pm$0.0075	0.051$\pm$0.0035
+ Int Aug	0.5	0.092$\pm$0.0042	0.029$\pm$0.0017
+ Int Aug	1.0	0.088$\pm$0.0042	0.029$\pm$0.0018
+ Int Aug	1.5	**0.086**$\pm$0.0042	**0.028**$\pm$0.0017
+ Int Aug	2.0	**0.086**$\pm$0.0042	**0.028**$\pm$0.0017
+ CF Aug	0.5	0.085$\pm$0.0042	0.029$\pm$0.0018
+ CF Aug	1.0	**0.082**$\pm$0.0042	**0.028**$\pm$0.0017
+ CF Aug	1.5	0.083$\pm$0.0042	**0.028**$\pm$0.0017
+ CF Aug	2.0	0.083$\pm$0.0042	**0.028**$\pm$0.0018
DoWhy (SCM)	–	0.095$\pm$0.0062	0.029$\pm$0.0018

Table 8: Effect of the augmentation ratio r on the three semi-synthetic datasets and the Causal Chamber dataset (setting of Table 3). We report the mean $\pm$ standard error. The SCM is fitted using the causal order, and Int Aug adds pseudo-interventional samples at ratio r. For each base model, the best value per column among the base model and its augmented variants is in **bold**. DoWhy (SCM) is shown for reference.

		PEHE ↓				**ATE Relative Error ↓**			
Method	r	Sales ($\times 10^4$)	Chamber	PSID ($\times 10^3$)	CPS ($\times 10^3$)	Sales	Chamber	PSID	CPS
Do-PFN	–	34.7$\pm$1.8	179$\pm$2.2	20.9$\pm$0.14	12.3$\pm$0.042	0.182$\pm$0.031	0.567$\pm$0.009	0.927$\pm$0.007	0.961$\pm$0.009
+ Int Aug	0.5	**27.7**$\pm$3.3	124$\pm$1.9	**17.3**$\pm$0.16	11.4$\pm$0.021	**0.075**$\pm$0.019	0.302$\pm$0.009	0.190$\pm$0.009	**0.267**$\pm$0.015
+ Int Aug	1.0	32.8$\pm$3.7	115$\pm$2.1	17.7$\pm$0.15	11.4$\pm$0.023	0.150$\pm$0.025	0.250$\pm$0.010	0.152$\pm$0.008	0.286$\pm$0.016
+ Int Aug	1.5	40.6$\pm$1.8	111$\pm$2.1	17.9$\pm$0.15	**11.3**$\pm$0.026	0.239$\pm$0.029	0.237$\pm$0.008	0.123$\pm$0.008	0.297$\pm$0.016
+ Int Aug	2.0	54.0$\pm$2.5	**110**$\pm$2.3	18.0$\pm$0.15	**11.3**$\pm$0.026	0.381$\pm$0.035	**0.223**$\pm$0.008	**0.109**$\pm$0.007	0.302$\pm$0.016
CausalPFN	–	58.0$\pm$1.8	91.1$\pm$3.1	14.4$\pm$0.20	**9.19**$\pm$0.032	0.433$\pm$0.029	0.211$\pm$0.013	0.222$\pm$0.017	**0.314**$\pm$0.017
+ Int Aug	0.5	**34.4**$\pm$1.8	**61.3**$\pm$3.7	**14.0**$\pm$0.17	9.43$\pm$0.041	**0.163**$\pm$0.012	0.087$\pm$0.013	0.074$\pm$0.006	0.399$\pm$0.019
+ Int Aug	1.0	37.5$\pm$1.8	62.5$\pm$3.7	14.2$\pm$0.17	9.55$\pm$0.040	0.165$\pm$0.010	0.076$\pm$0.011	0.072$\pm$0.006	0.417$\pm$0.018
+ Int Aug	1.5	39.8$\pm$1.4	62.2$\pm$3.9	14.5$\pm$0.16	9.54$\pm$0.035	0.184$\pm$0.020	**0.075**$\pm$0.010	**0.071**$\pm$0.006	0.376$\pm$0.017
+ Int Aug	2.0	40.2$\pm$1.5	64.0$\pm$4.1	14.7$\pm$0.16	9.55$\pm$0.035	0.177$\pm$0.019	0.077$\pm$0.011	0.078$\pm$0.006	0.362$\pm$0.017
DoWhy (SCM)	–	35.2$\pm$13	65.0$\pm$3.8	18.5$\pm$0.15	11.2$\pm$0.035	0.041$\pm$0.017	0.083$\pm$0.012	0.105$\pm$0.009	0.419$\pm$0.017