CAFE: COUNTERFACTUAL PREDICTION VIA FAST POSTERIOR ESTIMATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Counterfactual prediction estimates an individual's outcome under an alternative intervention given their factual observations. Such outcomes are generally not identifiable from observational data without additional assumptions. Even within the class of fully observed additive noise models (ANMs), different causal graphs can generate the same observational distribution yet imply different individual counterfactual outcomes. Predictions based on a single estimated graph ignore this structural uncertainty. We therefore target a Bayesian counterfactual posterior predictive distribution that combines predictions from plausible SCMs. We introduce CAFE (**C**ounterf**A**ctual Prediction via **F**ast Posterior **E**stimation), an amortized inference framework that directly approximates the Bayesian counterfactual posterior predictive distribution. We pretrain a transformer-based model on synthetic counterfactual tasks generated from a diverse prior over ANMs. Given an observational dataset, an individual's factual observations, and an intervention, CAFE approximates the corresponding posterior predictive distribution in a single forward pass. Experiments show that CAFE accurately predicts individual counterfactual outcomes in identifiable settings and approximates the posterior predictive distribution when structural uncertainty induced by observationally indistinguishable causal graphs exists. Strong performance in realistic manufacturing and viticulture settings further demonstrates its empirical robustness beyond the assumptions of the training prior.

1 INTRODUCTION

Counterfactual prediction asks what an individual's outcome would have been under an alternative intervention, given their factual observations (Pearl, 2013). The ability to reason about such alternative outcomes supports causal explanation (Halpern & Pearl, 2005), the evaluation of past actions (Oberst & Sontag, 2019), and individualized decision-making (Karimi et al., 2020). A fundamental challenge is that counterfactual outcomes are generally not identifiable from observational data without additional assumptions on the data-generating process (Zhang et al., 2015; Spirtes & Zhang, 2016; Bodik, 2026). Even when the observational distribution is known exactly, individual counterfactual non-identifiability can arise from uncertainty about the underlying causal graph, the structural equations, or the individual's exogenous noise values.

A common approach to make individual counterfactual outcomes identifiable is to assume that the correct causal graph or ordering is known and impose restrictions on the structural equations. Many SCM-based methods learn causal mechanisms conditional on a supplied or estimated graph or ordering (Pawlowski et al., 2020; Sánchez-Martin et al., 2022; Khemakhem et al., 2021; Javaloy et al., 2023; Chao et al., 2023; Scetbon et al., 2024; Mahajan et al., 2025). In practice, the true structure is often unavailable (Maathuis et al., 2009), and causal discovery may not identify it uniquely (Hiremath et al., 2024). SCMs with different graphs can generate the same observational distribution yet imply different counterfactual outcomes. Predictions based on a single estimated graph ignore this structural uncertainty and can lead to poor predictions.

To account for this structural uncertainty, we target a Bayesian counterfactual posterior predictive distribution. We remove the requirement of explicitly estimating causal graph and retain restrictions on the structural equations by considering fully observed, acyclic additive noise models (ANMs) with independent exogenous noises. Given the exact observational distribution and an in-

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.

dividual's complete factual observations, each compatible graph determines a unique counterfactual outcome under the conditions of Proposition 1. Under these conditions, individual counterfactual non-identifiability arises only when compatible graphs imply different outcomes. The posterior predictive distribution combines these outcomes according to the graphs' posterior probabilities under a prior over ANMs. It is a point mass when the outcome is identifiable and remains non-degenerate when graphs with positive posterior probability imply different outcomes.

However, exact computation of this posterior predictive distribution is generally intractable. The number of possible causal graphs grows super-exponentially with the number of variables. Computing their posterior probabilities also requires integrating over the structural mechanisms, and these integrals generally lack closed-form solutions for flexible nonlinear models. To address these challenges, we propose CAFE (**C**ounterf**A**ctual Prediction via **F**ast Posterior **E**stimation), an amortized inference framework that directly approximates the Bayesian counterfactual posterior predictive distribution. Following the prior-data fitted network approach (Hollmann et al., 2025; Robertson et al., 2026; Balazadeh Meresht et al., 2026), we pretrain a transformer on synthetic counterfactual tasks generated from the ANM prior. Each task includes an observational dataset, an individual's factual observations, and a specified intervention as inputs, with the corresponding counterfactual outcome as the prediction target. Training amortizes Bayesian integration over causal structures and mechanisms, enabling CAFE to predict a counterfactual distribution from these inputs in a single forward pass without graph search or parameter updates. Our contributions are threefold:

1. We propose CAFE, an amortized inference framework for Bayesian counterfactual prediction that requires neither a known causal graph nor explicit graph estimation. CAFE is pretrained on synthetic tasks generated from a diverse prior over causal graphs, mechanisms, and treatment–outcome relationships, supporting counterfactual prediction across varied causal structures (Section 3.1).
2. Under our modeling assumptions, CAFE accurately predicts individual counterfactual outcomes in identifiable settings (Section 4.1) and approximates the posterior predictive distribution when observationally indistinguishable causal structures imply different outcomes, preserving uncertainty that observational data alone cannot resolve (Section 4.2).
3. We evaluate CAFE on realistic manufacturing and viticulture settings. Strong predictive performance on these tasks demonstrates empirical robustness beyond the assumptions underlying our training prior (Section 4.3).

2 Problem Setup

Notations. A structural causal model (SCM) $\mathcal{M} = (G, \theta)$ describes the data-generating process of endogenous variables $\mathbf{V} = (V_1, \ldots, V_d)$ through the structural assignments

$$V_i = F_i\big(\mathbf{V}_{\mathrm{pa}_G(i)}, U_i; \theta\big), \qquad i = 1, \ldots, d. \tag{1}$$

where $\mathbf{U} = (U_1, \ldots, U_d)$ denotes the exogenous variables and $\mathrm{pa}_G(i)$ the parent indices of node i in the causal graph G, and $\mathbf{V}_{\mathrm{pa}_G(i)}$ is the corresponding subvector. The parameters θ specify both the structural functions and the joint distribution of the exogenous variables $\mathbf{U} = (U_1, \ldots, U_d)$. Let $P_{\mathbf{V}}$ denote the induced observational distribution and $\mathcal{D}_{\mathrm{obs}} = \{\mathbf{v}^{(j)}\}_{j=1}^{n}$ a dataset of n i.i.d. observations drawn from $P_{\mathbf{V}}$.

Counterfactual prediction. Counterfactual prediction asks what an individual's outcome would have been under a different treatment. For an individual with factual observations $\mathbf{V} = \mathbf{v}$, let $T = V_j$ be the intervention variable and $Y = V_k$ the outcome, with $j \neq k$. We denote by Y^t the outcome that this same individual would have experienced under $\mathrm{do}(T = t)$. The counterfactual outcome Y^t is defined as the value of Y in the modified SCM obtained by removing all incoming edges to T and replacing the structural equation for T with $T = t$, while keeping all other mechanisms and the individual's exogenous realization unchanged (Pearl, 2013). Our goal is to predict Y^t from the individual's factual observations $\mathbf{V} = \mathbf{v}$ and the observational dataset $\mathcal{D}_{\mathrm{obs}}$, without access to the underlying SCM or its causal graph. Individual counterfactual outcomes are generally not identifiable from observational data without additional constraints on the data-generating process (Bodik, 2026). Different SCMs can induce the same observational distribution yet imply

different values of Y^t for an individual with factual observations $\mathbf{V} = \mathbf{v}$. To constrain the possible data-generating processes, we consider fully observed additive noise models (ANMs), as formalized below. Notice that this assumption restricts the class of admissible SCMs but does not by itself guarantee counterfactual identifiability.

Assumption 1 (Fully observed additive noise models (ANMs))**.** *The causal graph G is acyclic, and all endogenous variables $\mathbf{V}$ are observed. For each node i, the structural function satisfies*

$$F_i\big(\mathbf{V}_{\mathrm{pa}_G(i)}, U_i; \theta\big) = f_i\big(\mathbf{V}_{\mathrm{pa}_G(i)}; \theta\big) + U_i, \qquad i = 1, \dots, d, \tag{2}$$

where f_i is a deterministic function and $U_1, \dots, U_d$ are mutually independent.

The following discussion assumes the regularity and support conditions specified in Appendix A.1, which also contains the formal statement and proof.

Proposition 1 (Individual counterfactual identifiability, informal)**.** *Within the class of SCMs satisfying Assumption 1, the observational distribution $P_{\mathbf{V}}$ and a compatible causal graph G uniquely determine the counterfactual outcome y^t_G for given factual observations $\mathbf{v}$ and intervention $\mathrm{do}(T = t)$. Consequently, if G is identifiable from $P_{\mathbf{V}}$ within this model class, the individual counterfactual outcome is identifiable.*

When $P_{\mathbf{V}}$ is known exactly, Proposition 1 implies that each compatible graph determines a unique counterfactual outcome for a given individual and intervention. Any remaining counterfactual uncertainty therefore arises from disagreement between compatible graphs.

Let Π be a prior over the SCMs considered above, and let $\{G_1, \dots, G_K\}$ denote the graphs compatible with $P_{\mathbf{V}}$ that have positive posterior probability. Write $\mathcal{Y}(\mathbf{v}, t; G, \theta)$ for the counterfactual outcome of Y under $\mathrm{do}(T = t)$ for an individual with factual observations $\mathbf{v}$ in the SCM (G, θ). For each compatible graph G, let $\theta_G(P_{\mathbf{V}})$ denote the parameters of any SCM with graph G that induces $P_{\mathbf{V}}$. The Bayesian counterfactual posterior is

$$p^\star_\Pi(\cdot \mid \mathbf{v}, t, P_{\mathbf{V}}) = \sum_{i=1}^{K} p_\Pi(G_i \mid P_{\mathbf{V}})\, \delta_{\mathcal{Y}(\mathbf{v}, t; G_i, \theta_{G_i}(P_{\mathbf{V}}))}, \tag{3}$$

where δ_a denotes a point mass at a. The mixture weights are determined by both $P_{\mathbf{V}}$ and the specified SCM prior Π. When the graph is identifiable from $P_{\mathbf{V}}$, $K = 1$ and the posterior reduces to a single point mass. When multiple graphs have positive posterior probability, i.e., $K \geq 2$, the posterior is a weighted mixture of their corresponding point masses. If these graphs imply different counterfactual outcomes, the posterior preserves uncertainty that the observational distribution alone cannot resolve. Counterfactual prediction methods should therefore account for this uncertainty rather than assume that a single estimated graph is correct.

3 METHOD

Section 2 shows that disagreement between compatible graphs can leave counterfactual outcomes uncertain even when $P_{\mathbf{V}}$ is known exactly. In practice, we have access only to a finite observational dataset $\mathcal{D}_{\mathrm{obs}}$, which introduces additional uncertainty about both the graph G and the SCM parameters θ. The corresponding Bayesian counterfactual posterior predictive distribution is

$$p^\star_\Pi(\cdot \mid \mathbf{v}, t, \mathcal{D}_{\mathrm{obs}}) = \sum_{G \in \mathcal{G}_\Pi} p_\Pi(G \mid \mathbf{v}, \mathcal{D}_{\mathrm{obs}}) \int_{\Theta_G} \delta_{\mathcal{Y}(\mathbf{v}, t; G, \theta)}\, p_\Pi(d\theta \mid \mathbf{v}, G, \mathcal{D}_{\mathrm{obs}}). \tag{4}$$

Here, $\mathcal{G}_\Pi$ denotes the graphs supported by the prior Π, and Θ_G is the SCM parameter space for graph G. Under Assumption 1, the factual observations $\mathbf{v}$ uniquely determine the individual's noise realization for each fixed SCM (G, θ). Keeping this realization unchanged under intervention yields the deterministic outcome $\mathcal{Y}(\mathbf{v}, t; G, \theta)$. Eq. 4 averages the corresponding point masses over the posterior uncertainty in G and θ. For each fixed intervention value t, under the prior Π, this posterior converges weakly almost surely to the population posterior in Eq. 3 as the observational sample size tends to infinity. The conditions and proof are provided in Appendix A.2. Exact evaluation of Eq. 4 is generally intractable. In particular, the posterior distributions over the causal graph G and mechanism parameters θ are typically not available in closed form, and computing the posterior predictive distribution requires marginalization over a combinatorial space of graph structures and a

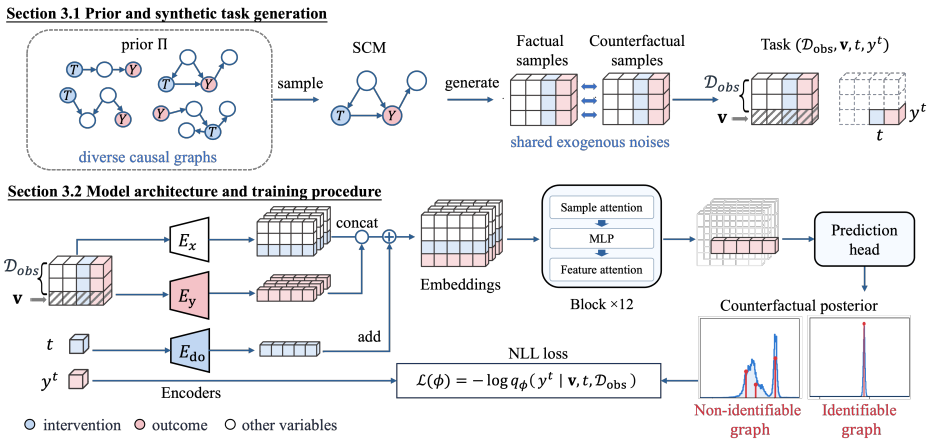

Figure 1: Overview of CAFE. Top: We sample SCMs with diverse DAG structures from the prior Π and generate paired factual and counterfactual data with shared exogenous noise. Bottom: We train a Transformer on these tasks to approximate the counterfactual posterior predictive distribution from observational data $\mathcal{D}_{\text{obs}}$, an individual's factual observations $\mathbf{v}$, and an intervention t.

potentially high-dimensional space of mechanism parameters. As illustrated in Figure 1, we amortize this computation by training a counterfactual posterior predictor $q_\phi(y^t \mid \mathbf{v}, t, \mathcal{D}_{\text{obs}})$ on synthetic counterfactual tasks. Each task $(\mathcal{D}_{\text{obs}}, \mathbf{v}, t, y^t)$ is generated from an SCM sampled from Π. The predictor receives the observational dataset $\mathcal{D}_{\text{obs}}$, factual state $\mathbf{v}$, intervention t and learns to predict the counterfactual outcome y^t. Let P_Π denote the joint distribution induced by the prior. We train q_ϕ to learn the counterfactual posterior predictive distribution by minimizing the expected negative log-likelihood:

$$\mathcal{L}(\phi) = \mathbb{E}_{(\mathcal{D}_{\text{obs}}, \mathbf{v}, t, y^t) \sim P_\Pi} \left[-\log q_\phi(y^t \mid \mathbf{v}, t, \mathcal{D}_{\text{obs}}) \right]. \tag{5}$$

Thus, with sufficient model capacity and exact optimization of the population objective, the optimal predictor matches the Bayesian posterior predictive distribution for almost every input under P_Π. At inference time, the trained model takes an observational dataset, an individual's factual state, and a specified intervention, and approximates this distribution in a single forward pass.

3.1 Prior and Synthetic Task Generation

We instantiate prior Π through a sampling procedure over causal graphs, structural functions, and exogenous distributions. Below, we describe the main components of the prior and the task-generation procedure. Sampling hyperparameters are provided in Appendix B.1.

Causal graphs G. We first sample the number of nodes d and a target average in-degree $\bar{k}$. We then sample a random node ordering and select $\min\{\lfloor d\bar{k} \rfloor, d(d-1)/2\}$ edges uniformly without replacement from those pointing forward in this ordering, ensuring the graph is acyclic. To cover diverse causal relationships, we randomly select two distinct nodes as the intervention variable $T = V_j$ and the outcome variable $Y = V_k$. Unlike the back-door priors used by Balazadeh Meresht et al. (2026), our prior accommodates a broader range of practical settings where the remaining observed variables include treatment descendants, rather than only pre-treatment covariates. For example, when predicting coronary events (Y) under simvastatin treatment (T), a patient's factual observations may include LDL cholesterol levels measured during treatment (X_1), which are themselves affected by the medication (Group, 1994). A different configuration arises when predicting progression to severe acute kidney injury (Y) under an alternative furosemide dose (T) (Chawla et al., 2013). Urine output following treatment (X_1) depends on both the dose and the extent of renal tubular injury (X_2), which also influences subsequent disease progression. These relationships form the path $T \to X_1 \leftarrow X_2 \to Y$. By including such causal structures, our prior accounts for the diverse causal roles of observed variables. An ablation study in Appendix D.6 shows that restricting pretraining to back-door structures can reduce prediction accuracy on other causal structures.

Structural functions and exogenous distributions θ. For each SCM, we assign each root node $V_i = U_i$ a distribution chosen from the Gaussian, uniform, and Zipf families, then sample root-node

values independently from their assigned distributions. For each non-root node i, we instantiate f_i as a randomly initialized MLP with $|\,\mathrm{pa}_G(i)|$ inputs and one scalar output. We then sample a signal-to-noise ratio (SNR) to vary the strength of the additive noise relative to the structural signal. The noise family is selected from Gaussian, Laplace, Gumbel, uniform, and exponential distributions; we also include a noiseless option with $U_i = 0$. These choices allow the prior to represent both symmetric and asymmetric disturbances, as well as deterministic mechanisms.

Factual and counterfactual generation. Given the sampled SCM $\mathcal{M} = (G, \theta)$, we jointly generate factual and counterfactual samples in topological order. At each non-intervened node, we apply the same structural function to the factual and counterfactual parent values, sharing one noise realization within each sample pair:

$$v_i = f_i\left(\mathbf{v}_{\mathrm{pa}_G(i)}\right) + u_i, \qquad v_i^t = f_i\left(\mathbf{v}^t_{\mathrm{pa}_G(i)}\right) + u_i, \qquad i \neq j, \tag{6}$$

where $f_i \equiv 0$ for root nodes. Before selecting the intervention value, we randomly assign individuals to an observational group and a target group. Upon reaching the treatment node $T = V_j$, we generate its factual values for all individuals and sample an intervention value t uniformly from the factual treatment values in the observational group. We then set intervention value $v_j^t = t$ for all individuals and continue the traversal, propagating the intervention to generate the counterfactual values. The factual observations of the observational group form $\mathcal{D}_{\mathrm{obs}}$. Each target individual's factual state $\mathbf{v}$ and paired counterfactual outcome $y^t = v_k^t$ yield a training task $(\mathcal{D}_{\mathrm{obs}}, \mathbf{v}, t, y^t)$.

3.2 Model architecture and training procedure

We implement q_ϕ using a tabular Transformer as illustrated in Figure 1 adapted for counterfactual prediction. The model takes an observational dataset $\mathcal{D}_{\mathrm{obs}}$, an individual's complete factual observations $\mathbf{v}$ (including the observed outcome y), and an intervention value t. It predicts the distribution of that same individual's outcome Y^t under $\mathrm{do}(T = t)$.

Input encoding. Let d denote the number of observed variables, including the treatment $T = V_j$ and outcome $Y = V_k$. We form an $(N + M) \times d$ input table, where the first N rows contain observations from $\mathcal{D}_{\mathrm{obs}}$ and the remaining M rows contain the factual observations of the individuals whose counterfactual outcomes we predict. Three separate two-layer MLP encoders map scalar inputs to 192-dimensional embeddings. Both E_x and E_y are applied to all $N + M$ rows: E_x is shared across all factual variables other than Y, while E_y encodes factual outcomes. We apply E_{do} only to the intervention values specified for the last M rows and add the resulting embeddings to the corresponding factual treatment embeddings. This associates the intervention with its target variable while retaining the individual's observed treatment and outcome, yielding an embedding tensor of shape $(N + M) \times d \times 192$.

Transformer backbone. We use the 12-layer Transformer architecture of LimiX (Zhang et al., 2025). Each block applies sample attention, a position-wise MLP, and feature attention in sequence (the SMF configuration), with residual connections and layer normalization. Sample attention gathers information from the N observational rows separately for each variable. Feature attention combines information across variables within the same row. Together, these operations combine each individual's factual observations and intervention with information from $\mathcal{D}_{\mathrm{obs}}$.

Training procedure. We jointly train the encoders, Transformer backbone, and an MLP prediction head on synthetic tasks generated from Π using the procedure described in Section 3.1. We discretize the continuous counterfactual outcome space into $B = 5000$ bins using the boundaries provided by TabPFN2.5 (Grinsztajn et al., 2025). For each individual, the prediction head outputs a probability for each bin, representing the probability that their counterfactual outcome falls within that interval. We minimize the negative log-probability assigned to the bin containing the counterfactual outcome, averaged over the M individuals. This is the discretized version of the objective in Eq. 5. Training hyperparameters are provided in Appendix B.2.

4 Experiments

We design our experiments to answer four questions: (1) Can CAFE accurately predict individual counterfactual outcomes when the causal graph is identifiable and the conditions of Proposition 1

Table 1: Counterfactual prediction RMSE on 7 C-Suite datasets with additive noise (mean ± std). Bold indicates the lowest mean RMSE in each column, excluding the factual-outcome baseline.

Method	collider lingauss	fork lin nongauss	fork nonlin gauss	large backdoor	nonlin simpson	symprod simpson	weak arrows
y factual	0.57 ±0.00	0.00 ±0.00	0.00 ±0.00	0.74 ±0.00	1.36 ±0.00	1.37 ±0.00	1.07 ±0.00
CAREFL	0.11 ±0.15	0.34 ±0.00	0.02 ±0.00	0.61 ±0.10	1.36 ±0.00	1.37 ±0.00	0.21 ±0.00
CausalNF	0.08 ±0.01	0.55 ±0.43	0.09 ±0.11	1.57 ±0.82	2.48 ±0.16	1.37 ±0.00	1.86 ±0.89
VCI	0.51 ±0.00	0.34 ±0.00	0.28 ±0.01	0.70 ±0.01	1.58 ±0.12	1.37 ±0.01	1.02 ±0.01
SCIGAN	0.81 ±0.01	0.62 ±0.04	0.47 ±0.00	1.20 ±0.06	1.07 ±0.09	1.51 ±0.05	0.98 ±0.03
DECI	0.39 ±0.30	0.60 ±0.46	0.03 ±0.07	0.42 ±0.23	3.04 ±1.07	1.45 ±0.10	0.42 ±0.17
DiBS	0.36 ±0.08	0.28 ±0.07	0.03 ±0.03	0.62 ±0.14	1.89 ±0.24	1.34 ±0.10	0.59 ±0.14
Fip	0.40 ±0.08	0.89 ±0.13	**0.00 ±0.00**	0.74 ±0.00	2.56 ±0.57	1.49 ±0.34	1.07 ±0.00
CAFE	**0.05 ±0.01**	**0.01 ±0.00**	0.01 ±0.00	**0.15 ±0.02**	**0.93 ±0.01**	**1.20 ±0.01**	**0.14 ±0.00**

hold (Section 4.1)? (2) Within the same class of Assumption 1, can it approximate the counterfactual posterior predictive distribution when observationally equivalent graphs imply different counterfactual outcomes, preserving the uncertainty that observational data alone cannot resolve (Section 4.2)? (3) How robust is its predictive performance in realistic settings where unobserved variables or non-additive noise may violate our modeling assumptions (Section 4.3)? (4) Can methods for estimating treatment effects be used to predict individual counterfactual outcomes (Appendix D.5)?

Baselines We compare CAFE against 7 methods with different approaches to causal modeling and inference. All methods are evaluated without access to the ground-truth graph or causal ordering. CAREFL (Khemakhem et al., 2021) and CausalNF (Javaloy et al., 2023) start from a partially directed graph estimated using PC: CAREFL orients the remaining edges using its likelihood-ratio criterion, while CausalNF models unresolved components as blocks with fixed internal orderings. FiP predicts a causal ordering using a pretrained ordering model. DECI (Geffner et al., 2022) and DiBS (Lorch et al., 2021) take a Bayesian approach to causal structure learning: DECI learns a variational distribution over graphs together with point estimates of additive-noise mechanisms, while DiBS jointly approximates a posterior over graphs and mechanism parameters. We also evaluate VCI (Wu et al., 2025) and SCIGAN (Bica et al., 2020), which assume that observed covariates suffice to control treatment–outcome confounding, avoiding explicit graph estimation. Implementation details for all baselines are provided in Appendix C.1.

Evaluation protocol. Across all experiments, we randomly partition the individuals in each dataset into two equal folds. We use the factual observations from one fold as an observational dataset of size N and evaluate counterfactual predictions for the $M = N$ individuals in the other, then swap their roles so that every individual is evaluated once. We repeat this procedure with five partition seeds and report the mean and standard deviation of each evaluation metric across seeds.

4.1 Counterfactual Prediction under Identifiability

C-Suite We first examine whether CAFE can accurately predict identifiable individual counterfactual outcomes under our causal assumptions. We evaluate individual counterfactual prediction on C-Suite (Geffner et al., 2022), a synthetic benchmark covering classical causal structures and inference scenarios. We consider additive-noise settings with at least three variables. For each dataset, the causal graph is uniquely determined by the exact observational distribution within the class of faithful independent ANMs with C^3 mechanisms. The remaining conditions of Proposition 1 also hold, making each individual's counterfactual outcome identifiable from the observational distribution, factual observations, and intervention (see Appendix A.4 for proofs). These are population-level guarantees; our experiments assess how accurately the outcomes can be predicted from finite observational data. Additional evaluations on C-Suite settings outside the additive-noise assumption are reported in Appendix D.2.

Metric For each C-Suite dataset, we use $N = 2000$. In these identifiable settings, individual counterfactual outcomes are uniquely determined, and the predictive distributions shown in Appendix D.1 concentrate sharply near these outcomes, supporting our focus on point-prediction accuracy. We use

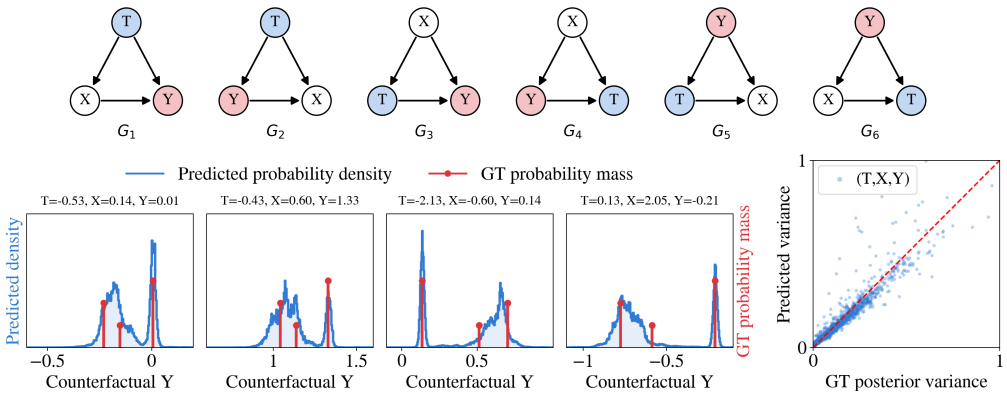

Figure 2: Counterfactual uncertainty under non-identifiable causal structure. Top: six observationally indistinguishable SCMs. Bottom: predicted versus ground-truth counterfactual distributions for randomly selected individuals (left) and posterior variances across individuals (right) at $N = 1000$.

the posterior predictive mean as the point prediction and report counterfactual RMSE. We also include a factual-outcome baseline that predicts $\hat{y}^t = y$, providing a reference for whether a method improves upon retaining the observed outcome.

Results Table 1 shows that CAFE achieves the lowest reported RMSE among seven baselines on six of the seven datasets, demonstrating strong individual counterfactual prediction across diverse causal structures. In the fork settings, e.g., $T \leftarrow X \rightarrow Y$, intervening on T leaves the individual's outcome unchanged. CAFE yields an RMSE of approximately 0.01 in both settings, closely preserving the factual outcome. CAFE also exhibits low variation across seeds on all seven datasets, whereas CAREFL, CausalNF, DECI, and FiP vary substantially in several settings. Although the causal graphs are identifiable from the exact observational distributions under the stated assumptions, this does not guarantee accurate recovery of their structures or mechanisms by a learning algorithm given finite data (Rahmadi et al., 2017; Faller et al., 2024). Errors in estimated edges, causal orderings, or mechanisms can affect counterfactual predictions.

4.2 Counterfactual uncertainty under non-identifiable causal structure.

Observationally equivalent SCMs. We next evaluate whether CAFE can capture counterfactual uncertainty due to causal graph non-identifiability from observational data. We construct six fully observed linear Gaussian SCMs satisfying Assumption 1 that induce the same observational distribution due to Markov equivalent causal structures. This construction provides an analytically tractable ground-truth posterior for evaluating uncertainty arising from observationally indistinguishable causal structures. Consider three observed variables $\mathbf{V} = (T, X, Y)$ with joint distribution

$$\mathbf{V} \sim \mathcal{N}(\mathbf{0}, \Sigma), \qquad \Sigma = \begin{pmatrix} 1 & 1/2 & 1/2 \\ 1/2 & 1 & 1/2 \\ 1/2 & 1/2 & 1 \end{pmatrix}. \tag{7}$$

For each permutation π of the three variables, we define an SCM with mutually independent exogenous noises:

$$\begin{aligned} V_{\pi_1} &= U_{\pi_1}, & U_{\pi_1} &\sim \mathcal{N}(0, 1), \\ V_{\pi_2} &= \tfrac{1}{2} V_{\pi_1} + U_{\pi_2}, & U_{\pi_2} &\sim \mathcal{N}(0, \tfrac{3}{4}), \\ V_{\pi_3} &= \tfrac{1}{3} V_{\pi_1} + \tfrac{1}{3} V_{\pi_2} + U_{\pi_3}, & U_{\pi_3} &\sim \mathcal{N}(0, \tfrac{2}{3}). \end{aligned} \tag{8}$$

These six complete DAGs are the only DAGs compatible with $P_{\mathbf{V}}$ under Assumption 1. The six SCMs differ only in the ordering of T, X, and Y; their structural coefficients and noise distributions follow the same specification in each ordering. Our prior samples these orderings uniformly and applies the same rules for generating mechanisms and noise distributions in each ordering. Permuting the variables leaves $P_{\mathbf{V}}$ unchanged, so conditioning on $P_{\mathbf{V}}$ preserves this symmetry. The six compatible DAGs therefore receive equal posterior weights: $p_{\Pi}(G_i \mid P_{\mathbf{V}}) = 1/6$ for $i = 1, \ldots, 6$.

For an individual with factual observations $\mathbf{v} = (t_\mathrm{f}, x, y)$, where t_f is the observed treatment value, we compute the counterfactual outcome under $\mathrm{do}(T = t)$ in each SCM while keeping the individual's exogenous noise unchanged. Substituting these outcomes and their weights into Eq. 3 gives

$$p^\star_\Pi(\cdot \mid \mathbf{v}, t, P_\mathbf{V}) = \underbrace{\tfrac{1}{2}\delta_y}_{G_4, G_5, G_6} + \underbrace{\tfrac{1}{3}\delta_{y+\frac{1}{2}(t-t_\mathrm{f})}}_{G_1, G_2} + \underbrace{\tfrac{1}{6}\delta_{y+\frac{1}{3}(t-t_\mathrm{f})}}_{G_3}. \tag{9}$$

We evaluate how closely CAFE's predictions from finite observational dataset approximate this analytical population posterior as the observational dataset size increases.

Metrics For each observational dataset size $N \in \{50, 100, 250, 500, 1000\}$, we set $t = -1$. We assess first- and second-order accuracy using posterior mean RMSE and posterior variance RMSE, respectively. Beyond these moment summaries, we use the integrated quadratic distance (IQD) (Thorarinsdottir et al., 2013) to measure the discrepancy between the model's continuous predictive distribution and the discrete ground-truth posterior:

$$\mathrm{IQD} = \frac{1}{2N} \sum_{i=1}^{2N} \int_{\mathbb{R}} \left[\widehat{F}_i(z) - F_i^{\mathrm{GT}}(z) \right]^2 dz, \tag{10}$$

where $\widehat{F}_i$ and F_i^{GT} are the predicted and ground-truth counterfactual CDFs for individual i.

Results Table 2 shows that CAFE's errors relative to the analytical population posterior generally decrease as the observational dataset grows. Additional observations help estimate the observational distribution more accurately, but cannot distinguish the six SCMs, which predict different counterfactual outcomes for the same individual. At $N = 1000$, Figure 2 (bottom left) shows that CAFE's predicted distributions closely match the analytical posteriors for the selected individuals. The variance comparison (bottom right) further supports the accuracy of CAFE's posterior variance estimates across individuals. Additional comparisons with DECI and DiBS, two Bayesian methods that can model graph uncertainty, show that CAFE achieves the lowest errors on all three posterior metrics (Appendix D.3). Together, these results support CAFE's ability to approximate the ground-truth counterfactual posterior and capture uncertainty due to graph non-identifiability.

Table 2: Counterfactual posterior prediction under non-identifiable causal structure (mean ± std).

N	Mean RMSE	Variance RMSE	IQD
50	0.464 ± 0.120	0.558 ± 0.132	0.161 ± 0.050
100	0.342 ± 0.191	0.565 ± 0.267	0.104 ± 0.092
250	0.153 ± 0.048	0.085 ± 0.017	0.031 ± 0.017
500	0.110 ± 0.020	0.073 ± 0.015	0.020 ± 0.008
1000	$\mathbf{0.094 \pm 0.012}$	$\mathbf{0.055 \pm 0.010}$	$\mathbf{0.018 \pm 0.004}$

4.3 Robustness to violations of modeling assumptions.

CausalMan. We evaluate predictive robustness on CausalMan (Tagliapietra et al., 2025), a physics-based simulator modeled after a real production line that assembles magnetic valves into hydraulic units. Given a component's factual measurements, the task is to predict whether it would have passed quality inspection (`Sec_C2_Machine1_ProcessResult`) under a different press-fitting force (`PF_M1_T1_Force`). We use CausalMan-Micro, which simulates a single product and retains 24 observed variables after removing constants, and CausalMan-Small, which includes multiple products and 53 observed variables. Both have 104 unobserved variables. Both datasets contain mixed variable types, and their unobserved process variables and non-additive noise place them outside the fully observed ANMs assumption. Moreover, the observed variables include descendants of the intervention variable, so these tasks do not satisfy the common assumption that all covariates are unaffected by treatment. For each dataset, we use $N = 1000$ observational samples and evaluate interventions at $t \in \{15000, 30000\}$. We threshold continuous counterfactual predictions at 0.5 to obtain binary labels and report counterfactual prediction accuracy. Table 3 shows that CAFE achieves the highest accuracy in all four settings. Several baselines achieve accuracies close to those of the factual-outcome baseline. CAREFL infers no directed path from the intervention variable to the outcome, while FiP places the outcome before the intervention variable in its inferred ordering. Both structural errors prevent the intervention from affecting the predicted outcome. On CausalMan-Small, DECI struggles to learn acyclic graphs; stronger acyclicity regularization yields overly sparse structures, with counterfactual accuracy matching the factual-outcome baseline. DiBS

Table 3: Counterfactual prediction performance on CausalMan and Sangiovese (mean ± std). We report accuracy (%, higher is better) on CausalMan under two press-fitting force interventions, and MAPE (%, lower is better) for Brix and pH on Sangiovese.

Method	CausalMan Micro Accuracy (%) ↑		CausalMan Small Accuracy (%) ↑		Sangiovese MAPE (%) ↓	
	$t = 15000$	$t = 30000$	$t = 15000$	$t = 30000$	Brix	pH
y factual	92.85 ±0.00	12.15 ±0.00	94.05 ±0.00	6.70 ±0.00	6.83 ±0.00	2.07 ±0.00
CAREFL	92.85 ±0.00	12.15 ±0.00	94.05 ±0.00	6.70 ±0.00	2.52 ±0.14	1.88 ±0.23
CausalNF	92.98 ±0.26	20.86 ±17.42	94.40 ±0.54	24.30 ±21.61	4.20 ±0.25	1.93 ±0.17
VCI	92.79 ±0.06	12.09 ±0.06	94.04 ±0.02	6.69 ±0.02	5.70 ±0.25	1.83 ±0.07
SCIGAN	91.48 ±2.67	12.15 ±0.00	92.80 ±2.50	6.68 ±0.02	6.58 ±0.08	2.76 ±0.12
DECI	95.85 ±0.79	53.87 ±14.58	94.05 ±0.00	6.70 ±0.00	4.44 ±2.14	1.85 ±0.28
DiBS	92.85 ±0.00	**100.00 ±0.00**	94.05 ±0.00	**100.00 ±0.00**	5.12 ±0.38	1.77 ±0.08
FiP	92.85 ±0.00	12.15 ±0.00	94.05 ±0.00	6.70 ±0.00	6.83 ±0.00	2.07 ±0.00
CAFE	**98.05 ±0.07**	**100.00 ±0.00**	**97.98 ±0.06**	**100.00 ±0.00**	**1.99 ±0.01**	**0.72 ±0.02**

can recover correct causal relationships in both datasets, but inaccuracies in its learned structural equations limit its sensitivity to small deviations from the factual treatment value, as observed at $t = 15000$. It is also computationally expensive, requiring approximately $48{,}000s$ on CausalMan-Small, compared with about $0.12s$ for CAFE (Appendix Table 6). VCI and SCIGAN also fail to improve counterfactual prediction over retaining the factual outcome. Both methods assume that covariates are unaffected by treatment, an assumption violated by the treatment descendants present in these datasets. Using the factual values of these descendants under intervention may explain why their predictions remain close to the factual outcome. These results demonstrate CAFE's robustness in realistic scenarios that violate the fully observed ANM assumptions underlying its training prior.

Sangiovese. We further evaluate CAFE on counterfactual samples generated from a Bayesian network fitted to vineyard experiments in Tuscany (Magrini et al., 2017). The task is to predict how an alternative harvest time would have changed the sugar concentration (`Brix`) and acidity (`pH`) of an individual grape batch. These batches differ in vegetative vigour, crop load, and leaf condition. For each batch, we toggle the binary `HarvestTime` indicator, setting $t = 1 - t_{\mathrm{f}}$ to switch between standard and late harvesting. We use $N = 1000$ samples to form the observational dataset and evaluate counterfactual predictions using mean absolute percentage error (MAPE). Table 3 shows that our model achieves the lowest MAPE on both Sangiovese outcomes, reaching 1.99% for Brix and 0.72% for pH. These correspond to relative error reductions of approximately 21% and 59% over the strongest competing baseline for each setting, respectively. CAFE also exhibits low variation across partition seeds, which reflects its ability to predict counterfactuals directly without relying on a separately estimated causal graph or ordering.

Taken together, CAFE retains strong individual counterfactual prediction performance across manufacturing and viticulture settings, even when unobserved variables or non-additive noise place the data-generating process outside its training assumptions. This provides empirical support for the practical applicability of counterfactual prediction learned from a synthetic SCM prior.

5 Conclusion

We developed CAFE for individual counterfactual prediction when the causal graph is unknown and observational data may support conflicting counterfactual outcomes. Our analysis characterizes counterfactual identifiability and structural uncertainty under fully observed ANMs. Building on this foundation, CAFE amortizes Bayesian counterfactual inference through in-context learning on synthetic tasks drawn from a diverse SCM prior. Experiments demonstrate accurate predictions in identifiable settings, preservation of uncertainty from observationally equivalent models, and empirical robustness beyond the training assumptions. Our theoretical analysis is based on fully observed ANMs with independent exogenous noises. Although Section 4.3 provides empirical evidence of robustness to violations of these assumptions, extending the theoretical characterization of counterfactual identifiability and posterior uncertainty to more general classes of SCMs remains an important direction for future work.

AI USE STATEMENT

We used generative AI tools to assist in the writing of proofs, assist with translation. We have not used generative AI tools to generate synthetic data sets, help develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, propose or refine hypotheses, design or provide feedback on research methodology or experiments, implement methods, clean and reformat dataset, support qualitative and thematic data analysis, interpret results. Additionally, we used generative AI tools for summarizing or analyse existing literature, sourcing/searching for information and identify relevant literature. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Vahid Balazadeh Meresht, Hamidreza Kamkari, Valentin Thomas, Junwei Ma, Bingru Li, Jesse Cresswell, and Rahul Krishnan. Causalpfn: Amortized causal effect estimation via in-context learning. *Advances in Neural Information Processing Systems*, 38:154945–154984, 2026.

Elias Bareinboim, Juan D Correa, Duligur Ibeling, and Thomas Icard. On pearl's hierarchy and the foundations of causal inference. In *Probabilistic and causal inference: the works of judea pearl*, pp. 507–556. 2022.

Ioana Bica, James Jordon, and Mihaela van der Schaar. Estimating the effects of continuous-valued interventions using generative adversarial networks. *Advances in Neural Information Processing Systems*, 33:16434–16445, 2020.

Juraj Bodik. Retrospective counterfactual prediction by conditioning on the factual outcome: A cross-world approach. *arXiv preprint arXiv:2603.27320*, 2026.

Patrick Chao, Patrick Blöbaum, Sapan Patel, and Shiva Prasad Kasiviswanathan. Modeling causal mechanisms with diffusion models for interventional and counterfactual queries. *arXiv preprint arXiv:2302.00860*, 2023.

Lakhmir S Chawla, Danielle L Davison, Ermira Brasha-Mitchell, Jay L Koyner, John M Arthur, Andrew D Shaw, James A Tumlin, Sharon A Trevino, Paul L Kimmel, and Michael G Seneff. Development and standardization of a furosemide stress test to predict the severity of acute kidney injury. *Critical care*, 17(5):R207, 2013.

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.

Riccardo Crupi, Alessandro Castelnovo, Daniele Regoli, and Beatriz San Miguel Gonzalez. Counterfactual explanations as interventions in latent space. *Data Mining and Knowledge Discovery*, 38(5):2733–2769, 2024. doi: 10.1007/s10618-022-00889-2. URL https://doi.org/10.1007/s10618-022-00889-2.

Anish Dhir, Cristiana Diaconu, Valentinian Lungu, James Requeima, Richard Turner, and Mark van der Wilk. Estimating interventional distributions with uncertain causal graphs through meta-learning. *Advances in Neural Information Processing Systems*, 38:140060–140096, 2026.

Philipp M Faller, Leena C Vankadara, Atalanti A Mastakouri, Francesco Locatello, and Dominik Janzing. Self-compatibility: Evaluating causal discovery without ground truth. In *International Conference on Artificial Intelligence and Statistics*, pp. 4132–4140. PMLR, 2024.

Tomas Geffner, Javier Antoran, Adam Foster, Wenbo Gong, Chao Ma, Emre Kiciman, Amit Sharma, Angus Lamb, Martin Kukla, Nick Pawlowski, et al. Deep end-to-end causal inference. *arXiv preprint arXiv:2202.02195*, 2022.

Léo Grinsztajn, Klemens Flöge, Oscar Key, Felix Birkel, Philipp Jund, Brendan Roof, Benjamin Jäger, Dominik Safaric, Simone Alessi, Adrian Hayler, et al. Tabpfn-2.5: Advancing the state of the art in tabular foundation models. *arXiv preprint arXiv:2511.08667*, 2025.

Léo Grinsztajn, Klemens Flöge, Oscar Key, Felix Birkel, Philipp Jund, Brendan Roof, Mihir Manium, Shi Bin Hoo, Magnus Bühler, Anurag Garg, et al. Tabpfn-3: Technical report. *arXiv preprint arXiv:2605.13986*, 2026.

Scandinavian Simvastatin Survival Study Group. Randomised trial of cholesterol lowering in 4444 patients with coronary heart disease: the scandinavian simvastatin survival study (4s). *The Lancet*, 344(8934):1383–1389, 2026/09/09 1994. doi: 10.1016/S0140-6736(94)90566-5. URL https://doi.org/10.1016/S0140-6736(94)90566-5.

Joseph Y Halpern and Judea Pearl. Causes and explanations: A structural-model approach. part ii: Explanations. *The British journal for the philosophy of science*, 2005.

Sujai Hiremath, Jacqueline Maasch, Mengxiao Gao, Promit Ghosal, and Kyra Gan. Hybrid top-down global causal discovery with local search for linear and nonlinear additive noise models. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 133147–133179. Curran Associates, Inc., 2024. doi: 10.52202/079017-4231.

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.

Adrián Javaloy, Pablo Sánchez-Martín, and Isabel Valera. Causal normalizing flows: from theory to practice. *Advances in Neural Information Processing Systems*, 36:58833–58864, 2023.

Amir-Hossein Karimi, Julius Von Kügelgen, Bernhard Schölkopf, and Isabel Valera. Algorithmic recourse under imperfect causal knowledge: a probabilistic approach. *Advances in neural information processing systems*, 33:265–277, 2020.

Ilyes Khemakhem, Ricardo Monti, Robert Leech, and Aapo Hyvarinen. Causal autoregressive flows. In *International conference on artificial intelligence and statistics*, pp. 3520–3528. PMLR, 2021.

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

Yuchen Ma, Dennis Frauen, Emil Javurek, and Stefan Feuerriegel. Foundation models for causal inference via prior-data fitted networks. In *International Conference on Learning Representations*, volume 2026, pp. 79065–79098, 2026.

Marloes H. Maathuis, Markus Kalisch, and Peter Bühlmann. Estimating high-dimensional intervention effects from observational data. *The Annals of Statistics*, 37(6A), December 2009. ISSN 0090-5364. doi: 10.1214/09-aos685. URL http://dx.doi.org/10.1214/09-AOS685.

Alessandro Magrini, Stefano Di Blasi, and Federico Mattia Stefanini. A conditional linear gaussian network to assess the impact of several agronomic settings on the quality of tuscan sangiovese grapes. *Biometrical Letters*, 54(1):25–42, 2017.

Divyat Mahajan, Jannes Gladrow, Agrin Hilmkil, Cheng Zhang, and Meyer Scetbon. Amortized inference of causal models via conditional fixed-point iterations. *Trans. Mach. Learn. Res.*, 2025, 2025. URL https://openreview.net/forum?id=D9pq25PGc5.

Hossein Mohebbi, Oliver Schulte, Ke Li, and Pascal Poupart. Counterfactual residual data augmentation for regression. In *Proceedings of the 43rd International Conference on Machine Learning (ICML)*, 2026. URL https://openreview.net/forum?id=mqd3zyHQj9.

Arash Nasr-Esfahany and Emre Kiciman. Counterfactual (non-) identifiability of learned structural causal models. *arXiv preprint arXiv:2301.09031*, 2023.

Michael Oberst and David Sontag. Counterfactual off-policy evaluation with gumbel-max structural causal models. In *International Conference on Machine Learning*, pp. 4881–4890. PMLR, 2019.

Nick Pawlowski, Daniel C. Castro, and Ben Glocker. Deep structural causal models for tractable counterfactual inference. In *Advances in Neural Information Processing Systems*, 2020.

Judea Pearl. Structural counterfactuals: A brief introduction. *Cognitive science*, 37(6):977–985, 2013.

Judea Pearl and Dana Mackenzie. The book of why. 2018.

Jonas Peters, Joris M Mooij, Dominik Janzing, and Bernhard Schölkopf. Causal discovery with continuous additive noise models. 2014.

Ridho Rahmadi, Perry Groot, Marianne Heins, Hans Knoop, Tom Heskes, et al. Causality on cross-sectional data: Stable specification search in constrained structural equation modeling. *Applied Soft Computing*, 52:687–698, 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. *arXiv preprint arXiv:2602.14972*, 2026.

Jake Robertson, Arik Reuter, Siyuan Guo, Noah Hollmann, Frank Hutter, and Bernhard Schölkopf. Do-pfn: In-context learning for causal effect estimation. *Advances in Neural Information Processing Systems*, 38:174811–174848, 2026.

Pablo Sánchez-Martin, Miriam Rateike, and Isabel Valera. Vaca: Designing variational graph autoencoders for causal queries. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 8159–8168, 2022.

Meyer Scetbon, Joel Jennings, Agrin Hilmkil, Cheng Zhang, and Chao Ma. A fixed-point approach for causal generative modeling. *arXiv preprint arXiv:2404.06969*, 2024.

Patrick Schwab, Lorenz Linhardt, Stefan Bauer, Joachim M Buhmann, and Walter Karlen. Learning counterfactual representations for estimating individual dose-response curves. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 5612–5619, 2020.

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

Christopher Stith, Medha Barath, Vahid Balazadeh, Jesse C Cresswell, and Rahul G Krishnan. Causal foundation models with continuous treatments. *arXiv preprint arXiv:2605.15133*, 2026.

Nicholas Tagliapietra, Juergen Luettin, Lavdim Halilaj, Moritz Willig, Tim Pychynski, and Kristian Kersting. Causalman: A physics-based simulator for large-scale causality, 2025. URL https://arxiv.org/abs/2502.12707.

Thordis L. Thorarinsdottir, Tilmann Gneiting, and Nadine Gissibl. Using proper divergence functions to evaluate climate models. *SIAM/ASA Journal on Uncertainty Quantification*, 1(1):522–534, 2013. doi: 10.1137/130907550. URL https://doi.org/10.1137/130907550.

Yulun Wu, Louis McConnell, and Claudia Iriondo. Counterfactual generative modeling with variational causal inference. In *International Conference on Learning Representations*, volume 2025, pp. 86921–86955, 2025.

Kun Zhang, Zhikun Wang, Jiji Zhang, and Bernhard Schölkopf. On estimation of functional causal models: general results and application to the post-nonlinear causal model. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 7(2):1–22, 2015.

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

Max Zhu, Martino Mansoldo, Ching-Hao Wang, and Stefan Groha. A causal foundation model for structure and outcome prediction, 2026. URL https://arxiv.org/abs/2606.26467.

APPENDIX CONTENTS

A Proof **14**

A.1 Proof of Proposition 1 14

A.2 Population limit of the counterfactual posterior 15

A.3 Proof of Proposition 4 16

A.4 Identifiability of 7 C-Suite settings 17

B Pretraining Details **20**

B.1 Prior data generation 20

B.2 Training hyperparameters 21

C Related Works **21**

C.1 Counterfactual Prediction Baselines 21

C.2 Relation to other causal foundation models 22

D Additional Results **23**

D.1 Predictive Distributions under Identifiable Causal Structure 23

D.2 Non additive noise settings in C-Suite 23

D.3 Comparison with Bayesian Structure-Learning Methods 24

D.4 Running Time 24

D.5 Can CEPO Methods Predict Individual Counterfactuals? 25

D.6 Ablation study on the training prior 28

A PROOF

A.1 PROOF OF PROPOSITION 1

Proposition 2 (Individual counterfactual identifiability)**.** *Let $\mathfrak{M}$ be the class of SCMs satisfying Assumption 1, and let $\mathcal{M} \in \mathfrak{M}$ have graph G and induce $P_{\mathbf{V}}$. For each node i, let*

$$S_i := \operatorname{supp}\big(P_{\mathbf{V}_{\mathrm{pa}_G(i)}}\big)$$

denote the joint observational support of its parent variables. Fix an intervention variable $T = V_j$, an outcome $Y = V_k$ with $j \neq k$, and a factual state $\mathbf{v} \in \operatorname{supp}(P_{\mathbf{V}})$. Write $\mathbf{v}^t$ for the counterfactual state under $\mathrm{do}(T = t)$ in $\mathcal{M}$. Suppose that:

(i) G is identifiable from $P_{\mathbf{V}}$ within $\mathfrak{M}$;

(ii) every model in $\mathfrak{M}$ inducing $P_{\mathbf{V}}$ satisfies $\mathbb{E}|V_i| < \infty$, $\mathbb{E}|U_i| < \infty$, and continuity of f_i on S_i for every node i;

(iii) $\mathbf{v}^t_{\mathrm{pa}_G(i)} \in S_i$ for every $i \neq j$.

Then $\mathbf{v}^t$ is uniquely determined by $(P_{\mathbf{V}}, \mathbf{v}, j, t)$ within $\mathfrak{M}$. In particular, the individual counterfactual outcome $y^t = v^t_k$ is identifiable.

Proof. We show that observationally equivalent additive mechanisms differ only by constants on their observational supports, and that these constants cancel when computing counterfactuals. This cancellation is a form of invariance under invertible reparameterizations of the exogenous variables (Nasr-Esfahany & Kiciman, 2023).

Let $\widetilde{\mathcal{M}} \in \mathfrak{M}$ be any model inducing $P_{\mathbf{V}}$, with structural functions $\widetilde{f}_i$ and exogenous variables $\widetilde{U}_i$. By condition (i), both models have the same graph G. We show that their counterfactual states agree for the given factual state and intervention.

In an acyclic SCM, the parents of node i are functions of exogenous variables other than U_i. Mutual independence of the exogenous variables therefore gives

$$U_i \perp \mathbf{V}_{\mathrm{pa}_G(i)}$$

in $\mathcal{M}$, and the analogous independence holds in $\widetilde{\mathcal{M}}$. Define

$$\mu_i = \mathbb{E}_{\mathcal{M}}[U_i], \qquad \widetilde{\mu}_i = \mathbb{E}_{\widetilde{\mathcal{M}}}[\widetilde{U}_i].$$

By integrability and equality of the observational distributions,

$$f_i(\mathbf{x}) + \mu_i = \widetilde{f}_i(\mathbf{x}) + \widetilde{\mu}_i \quad P_{\mathbf{V}_{\mathrm{pa}_G(i)}}\text{-almost surely},$$

because both sides are versions of the same conditional mean.

Both sides are continuous on S_i by condition (ii). If they differed at a point of S_i, continuity would imply a difference on a relative neighborhood with positive parent probability, contradicting their almost-sure equality. Hence

$$\widetilde{f}_i(\mathbf{x}) = f_i(\mathbf{x}) + c_i, \qquad \mathbf{x} \in S_i, \qquad c_i := \mu_i - \widetilde{\mu}_i. \tag{11}$$

Since $\mathbf{v} \in \operatorname{supp}(P_{\mathbf{V}})$, its parent subvectors belong to the corresponding sets S_i. The exogenous realizations recovered from $\mathbf{v}$ in the two models therefore satisfy

$$u_i = v_i - f_i(\mathbf{v}_{\mathrm{pa}_G(i)}),$$
$$\widetilde{u}_i = v_i - \widetilde{f}_i(\mathbf{v}_{\mathrm{pa}_G(i)}) = u_i - c_i.$$

Let $\mathbf{v}^t$ and $\widetilde{\mathbf{v}}^t$ denote the counterfactual states obtained by keeping these respective realizations fixed under $\mathrm{do}(T = t)$.

Proceed in a topological order of G. At the intervened node, both models assign $v^t_j = \widetilde{v}^t_j = t$. At any other node i, suppose the two counterfactual states agree on its parents; this condition is vacuous for

a root. Their common parent configuration equals $\mathbf{v}^t_{\mathrm{pa}_G(i)}$, which belongs to S_i by condition (iii). Eq. 11 therefore gives

$$\begin{aligned}\widetilde{v}_i^t &= \widetilde{f}_i\left(\mathbf{v}^t_{\mathrm{pa}_G(i)}\right) + \widetilde{u}_i \\ &= f_i\left(\mathbf{v}^t_{\mathrm{pa}_G(i)}\right) + c_i + u_i - c_i \\ &= v_i^t.\end{aligned}$$

Induction yields $\widetilde{\mathbf{v}}^{\,t} = \mathbf{v}^t$. Since $\widetilde{\mathcal{M}}$ was arbitrary, all models in $\mathfrak{M}$ inducing $P_{\mathbf{V}}$ give the same counterfactual state and, in particular, the same outcome $y^t = v_k^t$. □

To obtain Eq. 3, apply the proposition separately to each fixed-graph subclass of SCMs supported by Π. Within each subclass, condition (i) is automatic. Assuming conditions (ii)–(iii) hold for the given observational distribution and counterfactual query in every subclass with positive graph posterior probability, each graph G determines a unique outcome y^t_G. The conditional counterfactual posterior given G is therefore a point mass at y^t_G. Averaging these point masses with the graph posterior probabilities gives the stated mixture.

A.2 Population limit of the counterfactual posterior

Proposition 3 (Population limit of the counterfactual posterior)**.** *Let Π be a prior over fully observed, acyclic ANMs with mutually independent exogenous noises. Suppose the SCM parameter spaces are standard Borel, the observational and counterfactual maps are measurable, and $\mathbf{v}$ and the observational samples are conditionally i.i.d. given the SCM. Fix an intervention value t. Assume that, almost surely under the joint model induced by Π, conditions (ii) and (iii) of the formal identifiability proposition in Appendix A.1 hold for $P_{\mathbf{V}}$ and $(\mathbf{v}, t)$ within every fixed-graph subclass with $p_\Pi(G \mid P_{\mathbf{V}}) > 0$. Then*

$$p^\star_\Pi(\cdot \mid \mathbf{v}, t, \mathcal{D}_{\mathrm{obs}}) \;\Rightarrow\; p^\star_\Pi(\cdot \mid \mathbf{v}, t, P_{\mathbf{V}})$$

as $n \to \infty$, almost surely under this joint model. Here, $\Rightarrow$ denotes weak convergence. Graph identifiability is not required.

Proof. All conditional laws below are understood as regular conditional distributions under the joint model induced by Π, and all expectations are taken under this model. Consider an infinite observational sequence, with $\mathcal{D}_{\mathrm{obs}}$ containing its first n samples, and keep the same individual as n increases. Define

$$\mathcal{F}_n = \sigma\big(\mathbf{v}, \mathbf{v}^{(1)}, \ldots, \mathbf{v}^{(n)}\big), \qquad \mathcal{F}_\infty = \sigma\left(\bigcup_{n\geq 1} \mathcal{F}_n\right).$$

Under the ANM assumption, $Y^t = \mathcal{Y}(\mathbf{v}, t; G, \theta)$. Conditioning on the SCM therefore gives

$$\begin{aligned}\mathcal{L}_\Pi(Y^t \mid \mathcal{F}_n) &= \sum_{G\in\mathcal{G}_\Pi} p_\Pi(G \mid \mathbf{v}, \mathcal{D}_{\mathrm{obs}}) \int_{\Theta_G} \delta_{\mathcal{Y}(\mathbf{v},t;G,\theta)}\, p_\Pi(d\theta \mid G, \mathbf{v}, \mathcal{D}_{\mathrm{obs}}) \\ &= p^\star_\Pi(\cdot \mid \mathbf{v}, t, \mathcal{D}_{\mathrm{obs}}),\end{aligned}$$

which is Eq. 4. The infinite observational sequence determines $P_{\mathbf{V}}$ almost surely. Indeed, the strong law of large numbers, applied to a countable convergence-determining family of bounded continuous functions, yields

$$\frac{1}{n}\sum_{r=1}^{n} \delta_{\mathbf{v}^{(r)}} \;\Rightarrow\; P_{\mathbf{V}} \qquad \text{almost surely.}$$

Thus, $P_{\mathbf{V}}$ has an $\mathcal{F}_\infty$-measurable version.

Let $\mathcal{D}_\infty$ denote the infinite observational sequence. By conditional i.i.d. sampling,

$$\mathcal{L}_\Pi(\mathcal{D}_\infty \mid G, \theta, \mathbf{v}) = P_{\mathbf{V}}^{\otimes\mathbb{N}}.$$

This conditional law depends on the SCM only through $P_{\mathbf{V}}$. Since Y^t is determined by $(G, \theta, \mathbf{v})$, $\mathcal{D}_\infty$ and Y^t are conditionally independent given $(\mathbf{v}, P_{\mathbf{V}})$. Together with the measurability established above, this gives

$$\mathcal{L}_\Pi(Y^t \mid \mathcal{F}_\infty) = \mathcal{L}_\Pi(Y^t \mid \mathbf{v}, P_{\mathbf{V}}).$$

Moreover, $\mathcal{L}_\Pi(\mathbf{v} \mid G, P_{\mathbf{V}}) = P_{\mathbf{V}}$, so $p_\Pi(G \mid \mathbf{v}, P_{\mathbf{V}}) = p_\Pi(G \mid P_{\mathbf{V}})$. We now apply the identifiability argument in Appendix A.1 separately within each fixed-graph subclass. Within such a subclass, condition (i) is automatic, while conditions (ii) and (iii) hold by assumption. Thus, each compatible graph with positive posterior probability, together with $P_{\mathbf{V}}$, uniquely determines the counterfactual outcome for $(\mathbf{v}, t)$. Hence,

$$\begin{aligned}\mathcal{L}_\Pi(Y^t \mid \mathcal{F}_\infty) &= \sum_{i=1}^{K} p_\Pi(G_i \mid P_{\mathbf{V}})\, \delta_{\mathcal{Y}(\mathbf{v}, t; G_i, \theta_{G_i}(P_{\mathbf{V}}))} \\ &= p^\star_\Pi(\cdot \mid \mathbf{v}, t, P_{\mathbf{V}}),\end{aligned}$$

which is Eq. 3.

Finally, for any bounded continuous function $h : \mathbb{R} \to \mathbb{R}$, L'evy's upward theorem gives

$$\begin{aligned}&\int h(y)\, p^\star_\Pi(dy \mid \mathbf{v}, t, \mathcal{D}_{\text{obs}}) = \mathbb{E}_\Pi[h(Y^t) \mid \mathcal{F}_n] \\ &\qquad \longrightarrow \mathbb{E}_\Pi[h(Y^t) \mid \mathcal{F}_\infty] = \int h(y)\, p^\star_\Pi(dy \mid \mathbf{v}, t, P_{\mathbf{V}}) \qquad \text{almost surely.}\end{aligned}$$

Applying this result simultaneously to a countable convergence-determining family of bounded continuous functions on $\mathbb{R}$ establishes the claimed almost-sure weak convergence. □

A.3 Proof of Proposition 4

Proposition 4 (restated) (Exact counterfactual recovery by CATE adjustment)**.** *Consider a fully observed, acyclic SCM with mutually independent additive noises. Let* $\mathbf{X}$ *contain all endogenous variables other than* T *and* Y*, and assume that the CEPOs are finite and continuous in the treatment value.*

If $\mathbf{X}$ *contains no descendants of* T *and* Y *is not a parent of* T*, then, for almost every factual observation* $(\mathbf{x}, t_{\text{f}}, y)$ *and any intervention value* t *in the conditional support of* $T \mid \mathbf{X} = \mathbf{x}$*, the CATE* $\mu(\mathbf{x}, t) - \mu(\mathbf{x}, t_{\text{f}})$ *is identifiable from the observational distribution, and the counterfactual outcome under* $\text{do}(T = t)$ *satisfies*

$$y^t = y + \mu(\mathbf{x}, t) - \mu(\mathbf{x}, t_{\text{f}}).$$

Consequently, CATE adjustment recovers the individual counterfactual exactly whenever the CEPO estimates are exact at both the factual and intervention treatment values.

Proof. We first establish identification of the CATE. Since Y is not a parent of T, all parents of T are contained in $\mathbf{X}$, so its structural equation can be written as

$$T = h(\mathbf{X}) + U_T.$$

Because $\mathbf{X}$ contains no descendants of T, it depends only on exogenous noises other than U_T. The same holds for Y^a under $\text{do}(T = a)$, which replaces the structural equation for T. Independence of the exogenous noises therefore implies $U_T \perp (\mathbf{X}, Y^a)$ and hence $Y^a \perp T \mid \mathbf{X}$. Together with consistency, this gives

$$\mu(\mathbf{x}, a) = \mathbb{E}[Y \mid \mathbf{X} = \mathbf{x}, T = a]$$

for $P_{\mathbf{X},T}$-almost every $(\mathbf{x}, a)$. For $P_{\mathbf{X}}$-almost every $\mathbf{x}$, continuity of $\mu(\mathbf{x}, \cdot)$ uniquely determines its values throughout the conditional support of $T \mid \mathbf{X} = \mathbf{x}$. Thus, the CEPOs at t and t_{f}, and their difference, are identifiable from the observational distribution.

We next establish individual counterfactual recovery. Since $\mathbf{X}$ contains all endogenous variables other than T and Y, we can write

$$Y = g(\mathbf{X}, T) + U_Y,$$

where g depends only on the parents of Y. Intervening on T preserves both $\mathbf{X}$ and the individual's exogenous noise, so

$$y = g(\mathbf{x}, t_{\mathrm{f}}) + u_Y, \qquad y^t = g(\mathbf{x}, t) + u_Y.$$

Taking conditional expectations in $Y^a = g(\mathbf{X}, a) + U_Y$ yields

$$\mu(\mathbf{x}, a) = g(\mathbf{x}, a) + \mathbb{E}[U_Y \mid \mathbf{X} = \mathbf{x}].$$

The conditional noise mean cancels when comparing the two treatment values:

$$\mu(\mathbf{x}, t) - \mu(\mathbf{x}, t_{\mathrm{f}}) = g(\mathbf{x}, t) - g(\mathbf{x}, t_{\mathrm{f}}) = y^t - y.$$

Rearranging establishes Eq. 14. The final claim follows by substituting exact CEPO estimates into Eq. 13. □

A.4 Identifiability of 7 C-Suite settings

We use the zero-based indices and executable generating equations of C-Suite (Geffner et al., 2022). Let $s(z) = \log(1 + e^z)$, $a = \sqrt{2/3}$, $b = 1/\sqrt{3}$, and $\alpha = \sqrt{1 - 6(1/\sqrt{5} - 1/3)}$. Within each setting, the primitive noises are mutually independent: $Z_i \sim \mathcal{N}(0, 1)$, $L_i \sim \mathrm{Laplace}(0, 1)$, and $W_i \sim t_3$. Transformed primitive noises are absorbed into U_i. We consider scalar variables and the continuous-treatment versions. The sets S_i are population parent supports.

For the graph-identification arguments below, we additionally assume faithfulness of the generating and competing models to their DAGs, and restrict $\mathfrak{M}$ to independent ANMs with C^3 mechanisms. Under faithfulness, competing graphs have the same skeleton and unshielded colliders. We also use the graphical fact that two distinct Markov-equivalent DAGs have an oppositely oriented edge $A \to B$, $B \to A$, with $\mathrm{pa}_G(B) \setminus \{A\} = \mathrm{pa}_{G'}(A) \setminus \{B\} = S$; conditioning on S would therefore give additive-noise representations in both directions (Peters et al., 2014, Proposition 29(ii)).

For condition (ii), it suffices to establish finite first moments of all observed variables: in any observationally equivalent independent ANM, $X_i = \widetilde{f}_i(\widetilde{\mathbf{PA}}_i) + \widetilde{U}_i$ is an integrable sum of independent random variables, so $\mathbb{E}|\widetilde{U}_i| < \infty$. Indeed, conditioning at a parent value p with finite conditional absolute moment gives $\mathbb{E}|\widetilde{U}_i| \le \mathbb{E}|\widetilde{f}_i(p) + \widetilde{U}_i| + |\widetilde{f}_i(p)| < \infty$. Continuity of every competing mechanism follows from the model-class restriction above.

Collider lingauss. The equations and treatment are

$$X_0 = Z_0, \qquad X_2 = Z_2, \qquad X_1 = (X_0 + X_2 + Z_1)/\sqrt{3}, \qquad T = X_0.$$

Proof. (i) Under the additional faithfulness assumption, the only nontrivial independence, $X_0 \perp X_2$, identifies $X_0 \to X_1 \leftarrow X_2$ uniquely. Specifically, $\rho_{01} = \rho_{12} = 1/\sqrt{3}$, $\rho_{02|1} = -1/2$, and $\rho_{01|2} = \rho_{12|0} = 1/\sqrt{2}$. Hence its Markov equivalence class contains only this DAG.

(ii) All variables are Gaussian with finite absolute moments. Consequently every observationally equivalent ANM has integrable errors, and its mechanisms are continuous by the definition of $\mathfrak{M}$.

(iii) The joint density is positive on $\mathbb{R}^3$, so $S_1 = \mathbb{R}^2$. Replacing x_0 by any finite t preserves parent-support membership; the other non-intervened node is a root. Proposition 1 therefore applies. □

Fork lin nongauss. The equations and treatment are

$$X_1 = Z_1, \qquad X_0 = aX_1 + b\{s(1.8Z_0) - 1\}, \qquad X_2 = aX_1 + b\{s(1.8Z_2) - 1\}, \qquad T = X_0.$$

Proof. (i) Under faithfulness, the skeleton is $X_0 - X_1 - X_2$ with no collider. For either leaf $B \in \{X_0, X_2\}$, write $B = aX_1 + N$, where N has positive variance and support $[-b, \infty)$. The conditional support of $X_1 \mid B = y$ is $(-\infty, (y + b)/a]$. A reverse ANM $X_1 = g(B) + E$, $E \perp B$, would make these supports translates of one fixed support, forcing $g(y) = y/a + c$. Its residual would then equal $-N/a - c$, whose covariance with B is $-\mathrm{Var}(N)/a \neq 0$. Thus neither fork edge can reverse, identifying $X_0 \leftarrow X_1 \to X_2$.

(ii) Since $s(z) \le \log 2 + |z|$, all variables have finite second moments. The preliminary integrability argument gives finite first moments of every competing error; all competing mechanisms are continuous.

(iii) The only nonempty parent supports are $S_0 = S_2 = \mathbb{R}$. Intervening on the leaf X_0 changes no parent of another node. Hence all non-intervened parent configurations remain in support, and the proposition applies. □

Fork nonlin gauss. The equations and treatment are

$$X_1 = Z_1, \qquad X_0 = \sqrt{6}e^{-X_1^2} + \alpha Z_0, \qquad X_2 = s(1 - X_1) - 1.5 + \sqrt{0.15}Z_2, \qquad T = X_0.$$

Proof. (i) Under faithfulness, it remains to orient the two edges of the fork skeleton. Each forward bivariate model has a Gaussian cause, independent nondegenerate Gaussian noise, and a smooth nonlinear mechanism. The bivariate ANM identifiability result excludes both reverse directions (Peters et al., 2014, Theorem 20 and Corollary 22). Thus the graph is $X_0 \leftarrow X_1 \rightarrow X_2$.

(ii) The exponential term is bounded, softplus grows at most linearly, and Gaussian noises have finite second moments. Therefore all observed first moments and all competing error first moments are finite. Continuity holds throughout $\mathfrak{M}$.

(iii) The joint density is positive on $\mathbb{R}^3$. In particular, $S_0 = S_2 = \mathbb{R}$, and intervening on the leaf X_0 changes no parent of another node. The proposition applies. □

Large backdoor. The equations are

$$X_0 = s(1.8Z_0) - 1, \qquad X_1 = 1.5\{s(X_0 + 1) + s(0.5) - 3\} + 0.25Z_1,$$

$$X_2 = s(X_0 + 1) + s(0.5 + Z_2) - 3, \qquad X_i = s(X_{i-2} + 1) + s(0.5 + Z_i) - 3 \quad (i = 3, 4, 5, 6),$$

$$X_7 = 1.5s(X_5 + 1) - 1 + 0.3Z_7, \qquad X_8 = 2 - s\left(1 + \frac{-1.3X_6 + X_7}{3}\right) + 0.6L_8,$$

with $T = X_7$.

Proof. (i) Under faithfulness, the skeleton and the collider $X_6 \rightarrow X_8 \leftarrow X_7$ are identified. A reversal of either edge adjacent to X_0 would make that neighbor its sole parent, since a second incoming edge would introduce an unshielded collider at X_0. For $X_0 - X_1$, the support of $X_0 \mid X_1 = y$ is always $[-1, \infty)$. A reverse ANM would therefore have a constant mechanism, contradicting dependence. For $X_0 - X_2$, the conditional support is

$$\operatorname{supp}(X_0 \mid X_2 = y) = [-1, s^{-1}(y + 3) - 1],$$

whose length varies with y; these intervals cannot be translates of a fixed noise support. Thus $X_0 \rightarrow X_1$ and $X_0 \rightarrow X_2$. Preserving the absence of further unshielded colliders then orients both branches away from X_0, identifying the entire DAG.

(ii) Every structural expression is a finite composition of affine maps and softplus functions, with Gaussian or Laplace primitive noise. The linear-growth bound for softplus gives finite second moments by topological induction. Hence all required first moments are finite in every competing ANM, and continuity holds.

(iii) Under $\mathrm{do}(X_7 = t)$, only the parents of X_8 change among non-intervened nodes. Given $\mathbf{W} = (X_0, \ldots, X_6)$, the variable X_7 is Gaussian with variance 0.3^2 and continuous conditional mean. Consequently $S_8 = \operatorname{supp}(X_6) \times \mathbb{R}$, so $(x_6, t) \in S_8$ for every finite t and every factual observation in support. All other required parent configurations are unchanged. The proposition applies. □

Nonlin Simpson. The equations are

$$X_0 = Z_0, \qquad X_1 = s(1 - X_0) - 1.5 + 0.15Z_1,$$

$$X_2 = \tanh(2X_1) + 1.5X_0 - 1 + \tanh(Z_2), \qquad X_3 = 5\tanh\left(\frac{X_2 - 4}{5}\right) + 3 + 0.1L_3,$$

with $T = X_1$ and edges $0 \rightarrow 1$, $0 \rightarrow 2$, $1 \rightarrow 2$, $2 \rightarrow 3$.

Proof. (i) Under faithfulness, apply the conditional-edge reversal criterion stated above. Each possible reversed edge is excluded: $X_0 \to X_1$ is identifiable by the nonlinear Gaussian bivariate result (Peters et al., 2014, Theorem 20 and Corollary 22). For $X_0 \to X_2$ conditional on X_1, the forward model is $B = 1.5A + c + N$ with A of full support and $N = \tanh(Z_2)$. Reverse conditional supports are bounded intervals of fixed length, forcing a reverse mechanism $g(B) = B/1.5 + c'$; the resulting residual has covariance $-\operatorname{Var}(N)/1.5$ with B, a contradiction.

For $X_1 \to X_2$ conditional on $X_0 = x_0$, put $r = x_2 - (1.5x_0 - 1)$. The reverse conditional support is a lower half-line when $-2 < r < 0$ and an upper half-line when $0 < r < 2$. These supports cannot be translates of one noise support. Finally, for $X_2 \to X_3$, write $h(x) = 5\tanh((x-4)/5) + 3$, whose range is $(-2, 8)$. The smooth positive density of X_2, combined with Laplace noise, makes $p(X_2 \mid X_3 = y)$ have a cusp at $h^{-1}(y)$ for $y \in (-2, 8)$, whereas it is smooth for $y \notin [-2, 8]$. A location family cannot change between these two regularity types. Thus no required reverse ANM exists, and the DAG is identified.

(ii) Gaussian and Laplace noises have finite moments, $\tanh$ is bounded, and softplus has at most linear growth. Therefore every observed variable has finite first moment, as does every competing additive error. All competing mechanisms are continuous.

(iii) Here $S_1 = \mathbb{R}$, $S_2 = \mathbb{R}^2$, and $S_3 = \mathbb{R}$: (X_0, X_1) has an everywhere-positive density, and $1.5X_0$ gives X_2 full marginal support. Under $\operatorname{do}(X_1 = t)$, both (x_0, t) and the propagated x_2^t therefore lie in their required parent supports. The proposition applies despite the bounded error at X_2. □

Symprod Simpson. The equations are

$$X_0 = Z_0, \qquad X_1 = 2\tanh(2X_0) + 0.1W_1,$$

$$X_2 = 0.5X_0X_1 + 0.5L_2, \qquad X_3 = \tanh(1.5X_0) + 0.3Z_3,$$

with $T = X_1$ and edges $0 \to 1$, $0 \to 2$, $1 \to 2$, $0 \to 3$.

Proof. (i) Under faithfulness, again exclude the conditional reversals of the four edges. The edge $X_0 \to X_3$ is identifiable by the nonlinear Gaussian bivariate result (Peters et al., 2014, Theorem 20 and Corollary 22). For $X_0 \to X_1$, let q be the density of $0.1W_1$. Since $f(x) = 2\tanh(2x)$ is bounded, $q(y - f(x))/q(y) \to 1$ uniformly in x as $y \to \infty$. Thus $X_0 \mid X_1 = y$ converges to $\mathcal{N}(0, 1)$, including its first two moments. If these conditional laws formed a location family, every member would consequently be Gaussian with variance one. But its density is proportional to $e^{-x^2/2}q(y - f(x))$, whose second factor is bounded, positive, and nonconstant in x; this cannot be a Gaussian density. Hence the reverse ANM is impossible.

For either edge entering X_2, fix the other parent to any nonzero value. The conditional model is $B = cA + N$, where $c \neq 0$, $N \sim \text{Laplace}(0, 0.5)$, and A has a smooth positive density and finite variance. The density $p(A \mid B = y)$ has exactly one cusp, at $A = y/c$. A reverse location family must therefore have $g(y) = y/c + c'$, so its residual is $-N/c - c'$, with covariance $-\operatorname{Var}(N)/c \neq 0$ with B. Thus neither conditional edge can reverse, and the DAG is identified.

(ii) All primitive noises have finite second moments. Moreover,

$$X_0X_1 = 2X_0\tanh(2X_0) + 0.1X_0W_1$$

has finite second moment by independence and boundedness of $\tanh$. Therefore all observed first moments, and hence all competing error first moments, are finite. Continuity holds throughout $\mathfrak{M}$.

(iii) Every additive error has an everywhere-positive density on $\mathbb{R}$, so the joint density is positive on $\mathbb{R}^4$. All parent supports are full Euclidean spaces, and every finite intervention satisfies condition (iii). The proposition applies. □

Weak arrows. The equations are

$$X_0 = s(1.8Z_0) - 1, \qquad X_1 = 0.75\{s(X_0 + 1) + s(0.5) - 3\} + 0.75Z_1,$$

$$X_2 = s(1.5 - X_0) + s(0.5 + Z_2) - 3, \qquad X_3 = s(1.5 - X_1) + s(0.5 + Z_3) - 3,$$

$$X_i = s(X_{i-2} + 1) + s(0.5 + Z_i) - 3 \quad (i = 4, 5, 6), \qquad X_7 = 1.5s(X_5 + 1) - 1 + 0.3Z_7,$$

$$X_8 = s\left(1 + 0.1\sum_{i=0}^{5} X_i + 0.5X_6 + 0.7X_7\right) - 2 + 0.5L_8,$$

with $T = X_7$.

Proof. *(i)* Under faithfulness, every edge into X_8 is compelled: for each $i < 8$, some other parent $k < 8$ is nonadjacent to i, giving the unshielded collider $X_i \to X_8 \leftarrow X_k$. As in large backdoor, the fixed conditional support $\mathrm{supp}(X_0 \mid X_1 = y) = [-1, \infty)$ excludes $X_1 \to X_0$. For $X_0 - X_2$, the reverse conditional support is

$$[\ell(y), \infty), \qquad \ell(y) = \max\{-1,\ 1.5 - s^{-1}(y+3)\}.$$

For $y > s(2.5) - 3$, this endpoint is constantly -1. A reverse ANM would therefore require a constant mechanism on that entire upper interval, and hence identical conditional laws there. This is impossible: writing $h(x) = s(1.5 - x) - 3$ and q for the density of $s(0.5 + Z_2)$, the conditional density ratio at two distinct $x, x' > -1$ is proportional to $q(y-h(x))/q(y-h(x'))$. Its log derivative tends to $h(x) - h(x') \neq 0$ as $y \to \infty$, since $(\log q)'(u) = -(u - 0.5) + o(1)$. Thus $X_0 \to X_2$ also holds. The absence of additional unshielded colliders then orients both upstream branches away from X_0, identifying the full graph.

(ii) The affine and softplus mechanisms have at most linear growth. Topological induction gives finite second moments for every variable. Hence all observed and competing-error first moments are finite; all competing mechanisms are continuous.

(iii) Let $\mathbf{W} = (X_0, \ldots, X_6)$. The conditional distribution of X_7 given $\mathbf{W}$ is Gaussian with positive variance and continuous mean, so

$$S_8 = \mathrm{supp}(\mathbf{W}) \times \mathbb{R}.$$

Intervening on X_7 changes only the parent configuration of X_8 among non-intervened nodes, replacing $(\mathbf{w}, x_7)$ by $(\mathbf{w}, t) \in S_8$. All other required parent configurations are unchanged. The proposition applies. □

B Pretraining Details

B.1 Prior data generation

We sample the graph size d by rounding a draw from $\mathrm{LogUniform}(3, 64)$ to the nearest integer. We then sample the target average in-degree as $\bar{k} \sim \mathrm{Gamma}(\alpha = 3, \theta = 0.4)$, where α and θ denote the shape and scale, respectively. The target edge count is determined by $d\bar{k}$, and edges are sampled under a random node ordering as described in Section 3.1.

For each root node, we sample its distribution family as

$$F_{\mathrm{root}} \sim \mathrm{Categorical}\big(\{\mathrm{Gaussian}, \mathrm{Uniform}, \mathrm{Zipf}\}; \{0.5, 0.3, 0.2\}\big),$$

where $\mathrm{Categorical}(S; \mathbf{p})$ denotes a distribution over the choices in S with corresponding probabilities $\mathbf{p}$. For each non-root node, the structural function is a randomly initialized MLP with one input per parent and a scalar output. Its number of linear layers L and hidden width H are sampled as

$$L \sim \mathrm{Categorical}\big(\{1, 2, 3\};\ \{0.3, 0.4, 0.3\}\big),$$
$$H \sim \mathrm{Categorical}\big(\{8, 12, 16, 20, 32, 64\};\ \{0.2, 0.2, 0.2, 0.2, 0.1, 0.1\}\big).$$

All hidden layers share width H, and $L = 1$ gives an affine mechanism. Activations are sampled independently for each hidden layer from tanh, sine, negation, identity, LeakyReLU, SELU, SiLU, ReLU, Softplus, and Hardtanh. For each MLP, we sample a weight scale s, a bias standard deviation τ, and a weight-masking probability ρ:

$$s \sim \mathrm{LogUniform}(0.7, 2.0), \qquad \tau \sim \mathrm{Uniform}(0, 0.5),$$
$$\rho \sim 0.7\,\delta_0 + 0.3\ \mathrm{Uniform}(0.05, 0.5).$$

For a layer with n_{in} inputs, weights and biases are initialized independently as

$$W_{ab} = A_{ab} Z_{ab}, \qquad A_{ab} \sim \mathrm{Bernoulli}(q),$$
$$Z_{ab} \sim \mathcal{N}\left(0, \frac{c^2}{n_{\mathrm{in}} q}\right), \qquad b_a \sim \mathcal{N}(0, \tau^2).$$

We use $c = s$ for hidden layers and $c = 1$ for the output layer. Weights in intermediate linear layers are independently masked with probability ρ ($q = 1 - \rho$), while weights in the first and output linear layers are not masked ($q = 1$).

For each non-root node, we sample the noise family as

$$F_{\text{noise}} \sim \text{Categorical}\big(\text{Gauss}, \text{Lap}, \text{Gum}, \text{Uni}, \text{Exp}, \text{None}; 0.5, 0.1, 0.1, 0.1, 0.1, 0.1\big).$$

The SNR r is sampled independently for each mechanism through

$$z \sim \text{Uniform}(\log 0.5, \log 20), \qquad r = 20.5 - \exp(z).$$

This reflected log-uniform distribution spans $[0.5, 20]$ and favors larger SNRs. Each SCM generates 2000 factual–counterfactual pairs. Before selecting the intervention value, we sample $\eta \sim \text{Uniform}(0.4, 0.6)$ and randomly assign a fraction η of the individuals to the observational group and the remainder to the target group. The factual observations of the observational group form $\mathcal{D}_{\text{obs}}$. We sample one intervention value uniformly from the factual treatment values in this group and apply it to all individuals, sharing each individual's exogenous noise between factual and counterfactual evaluations.

B.2 Training hyperparameters

We train CAFE for 300 epochs on 8 NVIDIA RTX 5090 GPUs, with 1000 training steps per epoch. We use the Adam optimizer with a learning rate of 2×10^{-4} and weight decay of 10^{-3}. The first 100 epochs are used for learning-rate warmup, and gradients are accumulated over eight steps before each optimizer update.

C Related Works

C.1 Counterfactual Prediction Baselines

CAREFL and CausalNF perform counterfactual inference given a causal ordering or DAG, and both papers describe extensions for partially known causal graphs. For a fair comparison, we use the same PC procedure to estimate a partially directed graph from observational data before applying each method's extension. For both methods, we follow the model architectures and hyperparameter settings provided in their respective official repositories.[1][2]

CAREFL. Following the multivariate discovery procedure of CAREFL (Khemakhem et al., 2021), we use its likelihood-ratio criterion to orient the edges left undirected by PC. We then fit an affine autoregressive flow according to the resulting causal ordering and use it for counterfactual prediction.

CausalNF. Following the partial-graph extension of CausalNF (Javaloy et al., 2023), we represent unresolved edges in both directions and group the resulting strongly connected components into blocks. Each block's joint distribution is modeled using an autoregressive flow with an arbitrary fixed ordering. The flow respects the causal ordering between blocks, with inter-block connections expanded to block-level dependencies.

VCI. We follow the network architecture and training hyperparameters used in the original VCI (Wu et al., 2025) single-cell experiments, as in the official implementation.[3] We use the designated treatment and outcome as T and Y, respectively, and the remaining factual measurements as covariates X.

SCIGAN. SCIGAN (Bica et al., 2020) is designed to estimate conditional mean treatment responses, its inference model does not condition on the individual's factual treatment or outcome. For individual counterfactual prediction, we omit this model and directly use the generator, conditioning on the individual's observations, including factual treatment and outcome, to predict under

[1] https://github.com/piomonti/carefl

[2] https://github.com/psanch21/causal-flows

[3] https://github.com/yulun-rayn/variational-causal-inference

the queried intervention. We use the generator and dosage discriminator architectures and training hyperparameters provided in the official SCIGAN repository.[4]

We include two Bayesian baselines, DECI and DiBS. DECI captures graph uncertainty through a variational posterior over causal graphs. We use the JointDiBS variant of DiBS, which approximates a joint posterior over graphs and mechanism parameters.

DECI. We implement DECI (Geffner et al., 2022) using the official Causica library.[5] We jointly learn a variational distribution over causal graphs and nonlinear additive-noise mechanisms from observational data, using spline flows to model the noise distributions. Following the official evaluation procedure, we sample 10 graphs from the learned variational posterior. For each graph, we infer the individual's exogenous noise from their factual observations and reuse it to predict the outcome under the intervention. We average the resulting counterfactual predictions across the sampled graphs to obtain the final point prediction.

DiBS. We implement DiBS (Lorch et al., 2021) using the official library.[6] We use JointDiBS to approximate the joint posterior over causal graphs and mechanism parameters from observational data. We model the causal mechanisms using neural networks with additive Gaussian noise. We retain the official defaults for the network architecture, noise variance, and optimizer settings. We standardize variables using the training data and run Stein variational gradient descent with 20 particles for 1000 steps.

FiP. We implement FiP (Scetbon et al., 2024) using the official repository.[7] We first train a causal ordering model using the provided pretraining code. For each evaluation dataset, this pretrained model predicts a causal ordering from the observational training data. Given the predicted ordering, we fit FiP's Transformer-based SCM to the same training data, following the architecture and training hyperparameters in the repository.

C.2 Relation to other causal foundation models

Pearl's ladder of causation distinguishes three levels: association ($\mathcal{L}_1$), intervention ($\mathcal{L}_2$), and counterfactual reasoning ($\mathcal{L}_3$) (Pearl & Mackenzie, 2018). These levels form a directional hierarchy: answering questions at a lower level does not generally suffice to answer questions at a higher one (Bareinboim et al., 2022).

Conventional tabular foundation models primarily address ($\mathcal{L}_1$), predicting outcomes from observed features. Most existing causal foundation models address ($\mathcal{L}_2$), estimating outcome distributions and treatment effects under interventions. CausalPFN (Balazadeh Meresht et al., 2026) learns the posterior distribution of conditional expected potential outcomes (CEPOs), $\mu_t(\mathbf{x}) = \mathbb{E}[Y(t) \mid \mathbf{X} = \mathbf{x}]$, implemented for binary treatments under strong ignorability. CCPFN (Stith et al., 2026) extends this target to continuous treatments, estimating treatment-response curves $t \mapsto \mu_t(\mathbf{x})$. CausalFM (Ma et al., 2026) learns predictive distributions of interventional outcomes or outcome contrasts with binary-treatment implementations for effect estimation under back-door, front-door, and instrumental-variable assumptions. Do-PFN (Robertson et al., 2026) targets conditional interventional distributions $p(y \mid \mathrm{do}(t), \mathbf{x})$ for binary treatments. MACE-TNP (Dhir et al., 2026) targets marginal interventional distributions $p(y \mid \mathrm{do}(t))$, averaging over uncertain graphs and mechanisms. Reuter et al. (2026) additionally condition interventional predictions on complete or partial graph and ancestral information. Even accurate predictions of these interventional quantities do not generally determine what would have happened to a particular individual given their factual observations.

CAFE addresses this individual counterfactual prediction task at $\mathcal{L}_3$. The work most closely related to CAFE is TabPFN-CFM (Zhu et al., 2026), which also claims to support counterfactual prediction. However, TabPFN-CFM requires the counterfactual values of all other input variables to predict an individual's counterfactual outcome (as stated in Section 3.2 of the paper). For treatment descendants, these values are generally unavailable and reveal intervention-induced changes that would

[4] https://github.com/ioanabica/SCIGAN

[5] https://github.com/microsoft/causica

[6] https://github.com/larslorch/dibs

[7] https://github.com/microsoft/causica/tree/main/research_experiments/fip

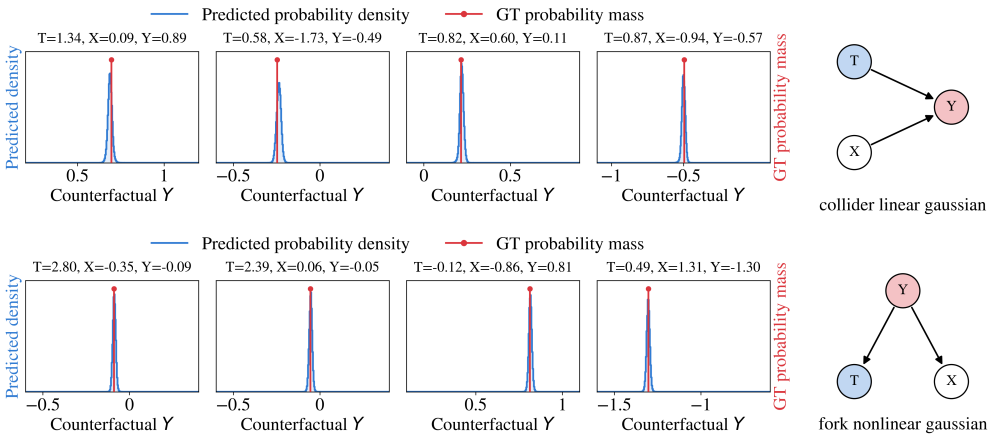

Figure 3: Counterfactual predictive distributions under identifiable causal structure. Predicted densities (blue) and ground-truth point masses (red) for four individuals from each of two C-Suite datasets: collider linear Gaussian (top) and fork nonlinear Gaussian (bottom). Panel titles report factual observations, and the corresponding causal graphs appear on the right. Each red marker represents unit probability mass at the individual's true counterfactual outcome; its height is read on a separate scale from the predicted density.

otherwise require inference, simplifying outcome prediction. CAFE instead uses only observational data, the individual's factual state, and the specified intervention. Furthermore, TabPFN-CFM does not establish conditions for individual counterfactual identifiability. Our analysis provides such conditions and characterizes the uncertainty arising when observationally equivalent graphs imply different counterfactual outcomes. We demonstrate that CAFE can capture this uncertainty by comparing its predicted counterfactual distributions with analytic reference posteriors.

D Additional Results

D.1 Predictive Distributions under Identifiable Causal Structure

Figure 3 complements the point-prediction results in Section 4.1 by examining individual predictive distributions in two identifiable C-Suite settings. Under the conditions of Proposition 1, the exact observational distribution, factual observations, and intervention uniquely determine each individual's counterfactual outcome, yielding a point-mass counterfactual distribution. Across the displayed individuals, our model produces sharply concentrated densities with peaks close to these ground-truth locations. In the collider setting, the predictions capture intervention-induced changes from the factual outcome. In the fork setting, where Y is an ancestor of T, intervening on T leaves Y unchanged, and the predicted densities concentrate near the factual outcome. The small remaining spread and offsets are compatible with finite-context uncertainty and approximation in the learned, discretized predictive distribution. Together with the non-identifiable experiment, these examples provide qualitative evidence that the model can produce concentrated predictions when the counterfactual is identifiable while retaining uncertainty when observationally indistinguishable SCMs imply different outcomes.

D.2 Non Additive Noise Settings in C-Suite

We further evaluate on three C-Suite datasets with non-additive discrete mechanisms: Large Backdoor Binary Treatment, Mixed Confounding, and Mixed Simpson. These settings contain binary or categorical nodes whose parents affect the probabilities of fixed outcome categories. Their structural equations therefore fall outside Assumption 1. Using the same evaluation protocol as in Section 4.1, we compare counterfactual RMSE across all methods. Our method achieves the lowest RMSE on all three datasets and consistently improves upon the factual-outcome baseline, demonstrating robustness to these departures from the additive-noise assumption.

Table 4: Counterfactual prediction RMSE on C-Suite datasets with non-additive noise (mean ± standard deviation). Bold indicates the lowest mean RMSE in each column.

Method	large backdoor binary treatment	mixed confounding	mixed simpson
y factual	0.17 ± 0.00	0.86 ± 0.00	0.50 ± 0.00
CAREFL	0.17 ± 0.00	0.86 ± 0.00	0.50 ± 0.00
CausalNF	0.23 ± 0.11	0.86 ± 0.00	0.64 ± 0.19
VCI	0.14 ± 0.01	1.25 ± 0.09	0.93 ± 0.34
SCIGAN	1.16 ± 0.04	0.71 ± 0.07	0.43 ± 0.06
DECI	0.14 ± 0.03	0.73 ± 0.23	1.13 ± 0.32
DiBS	0.15 ± 0.01	0.74 ± 0.08	0.56 ± 0.07
FiP	0.65 ± 0.08	0.56 ± 0.00	1.24 ± 0.00
CAFE	$\mathbf{0.05 \pm 0.00}$	$\mathbf{0.41 \pm 0.00}$	$\mathbf{0.38 \pm 0.00}$

D.3 Comparison with Bayesian Structure-Learning Methods

We compare CAFE with DECI (Geffner et al., 2022) and DiBS (Lorch et al., 2021) on the three-variable, six-DAG linear Gaussian system in Section 4.2. DECI learns a variational posterior over causal graphs, while DiBS jointly approximates a posterior over graphs and mechanism parameters. This comparison examines whether these methods accurately preserve counterfactual uncertainty when the causal structure is not identifiable.

Table 5 shows that CAFE achieves the lowest average error on all three metrics. DiBS improves on DECI in mean RMSE and IQD, but has a larger variance RMSE, indicating that more accurate posterior means do not necessarily imply more accurate uncertainty estimates. The individual counterfactual predictions illustrate how the baseline posteriors depart from the ground truth. DiBS assigns excessive probability to the dominant regions of the ground truth posterior, leaving too little mass on alternative counterfactual outcomes (Figure 4, top). Thus, despite representing uncertainty over both graphs and parameters, it does not faithfully preserve the relative probabilities of plausible outcomes in these examples. For DECI, the learned graph posterior concentrates on sparse structures (Figure 4, bottom). Its two most probable DAGs, $T \to X \to Y$ and $Y \to T \to X$, receive 48.42% and 45.02% of the probability, respectively. These two-edge graphs receive 93.44% of the probability, although neither belongs to the six compatible complete DAGs. Most of the learned graph probability therefore falls on structures that cannot reproduce the population observational distribution. This preference for sparse graphs is consistent with DECI's sparsity-promoting graph prior, which may contribute to its inaccurate counterfactual predictions in this fully connected system. These results show that, under the evaluated configurations, maintaining a distribution over causal graphs is not sufficient to recover the ground truth counterfactual posterior. CAFE provides more accurate posterior predictions and better preserves the uncertainty arising from structural non-identifiability.

Table 5: Counterfactual posterior prediction under non-identifiable causal structure (mean ± standard deviation) at $N = 1000$.

Method	Mean RMSE	Variance RMSE	IQD
DECI	0.235 ± 0.045	0.101 ± 0.030	0.082 ± 0.050
DiBS	0.142 ± 0.050	0.175 ± 0.125	0.034 ± 0.024
CAFE	$\mathbf{0.094 \pm 0.012}$	$\mathbf{0.055 \pm 0.010}$	$\mathbf{0.018 \pm 0.004}$

D.4 Running Time

Table 6 reports the time needed to apply each method to a new dataset, including graph estimation, model fitting, and counterfactual prediction where required. We exclude pretraining for CAFE and FiP because the pretrained models can be reused across downstream tasks.

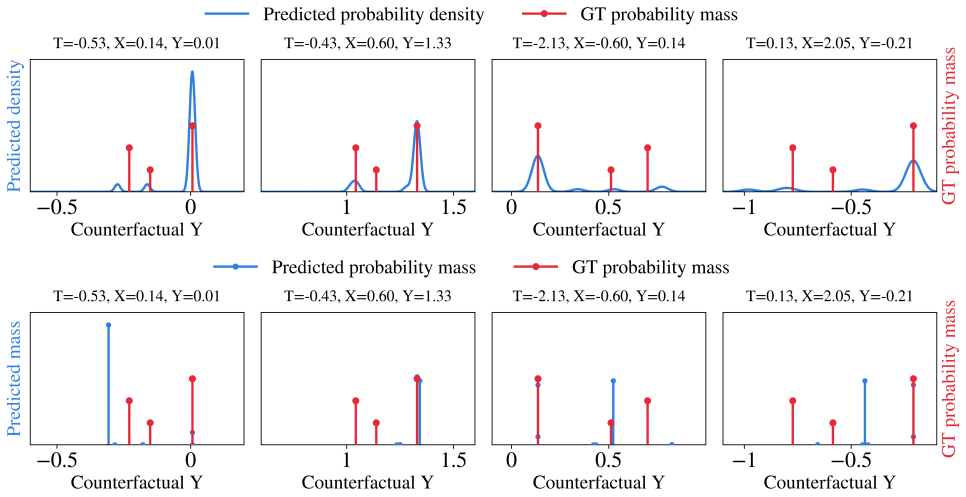

Figure 4: Estimated and ground-truth counterfactual posteriors in the three-variable, six-DAG linear Gaussian system ($N = 1000$). Top: Counterfactual posterior estimates from DiBS. Bottom: Counterfactual posterior estimates from DECI.

CAFE achieves the lowest runtime across all three datasets, as it requires neither graph search nor parameter updates on the evaluation data. We find that DiBS is particularly time-consuming. DiBS approximates the joint posterior over causal graphs and mechanism parameters by iteratively updating a set of particles. Each update requires Monte Carlo sampling of graphs and repeated likelihood and gradient computations. The resulting computation becomes more expensive as the number of variables increases, contributing to the long runtime observed on CausalMan Small.

Table 6: Runtime comparison on Sangiovese and CausalMan (seconds, mean ± standard deviation; lower is better). The number of observed variables is shown for each dataset.

Method	Sangiovese $d = 15$	CausalMan Micro $d = 24$	CausalMan Small $d = 53$
CAREFL	143.27 ±15.45	981.12 ±62.54	1064.83 ±526.84
CausalNF	21.13 ±0.42	1453.74 ±141.83	628.77 ±57.50
VCI	276.73 ±11.12	211.85 ±1.12	208.36 ±1.13
SCIGAN	233.44 ±9.16	117.91 ±2.24	122.23 ±1.94
DECI	78.33 ±1.04	1065.45 ±9.23	214.36 ±3.72
DiBS	6638.62 ±165.85	9419.14 ±983.09	48251 ±1354.46
FiP	270.44 ±5.36	84.71 ±0.63	4405.92 ±119.63
CAFE	**0.203** ±0.003	**0.068** ±0.001	**0.122** ±0.002

D.5 Can CEPO Methods Predict Individual Counterfactuals?

We investigate whether methods used to estimate conditional expected potential outcomes (CEPOs) at the interventional level ($\mathcal{L}_2$) can be adapted to predict individual counterfactual outcomes ($\mathcal{L}_3$). For an individual with factual observations $(\mathbf{x}, t_\mathrm{f}, y)$, let t denote the treatment value specified by the intervention $\mathrm{do}(T = t)$. For pretreatment covariates $\mathbf{X}$, the CEPO is

$$\mu(\mathbf{x}, t) = \mathbb{E}[Y^t \mid \mathbf{X} = \mathbf{x}]. \tag{12}$$

A natural adaptation adds the estimated CATE between the intervention and factual treatment values to the observed outcome (Crupi et al., 2024; Mohebbi et al., 2026):

$$\hat{y}^t_\mathrm{cal} = y + \hat{\mu}(\mathbf{x}, t) - \hat{\mu}(\mathbf{x}, t_\mathrm{f}). \tag{13}$$

We refer to this procedure as *CATE adjustment* and examine the conditions under which it recovers the true individual counterfactual.

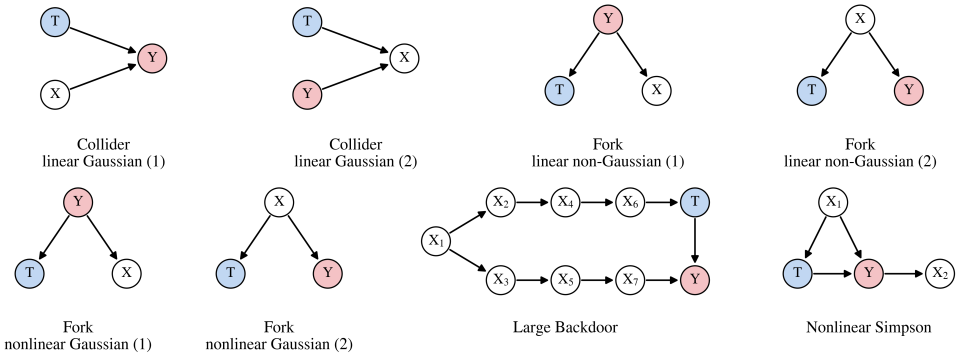

Figure 5: Eight causal graphs used to evaluate CATE adjustment.

Proposition 4 (Exact counterfactual recovery by CATE adjustment)**.** *Consider a fully observed, acyclic SCM with mutually independent additive noises. Let* $\mathbf{X}$ *contain all endogenous variables other than* T *and* Y*, and assume that the CEPOs are finite and continuous in the treatment value.*

If $\mathbf{X}$ *contains no descendants of* T *and* Y *is not a parent of* T*, then, for almost every factual observation* $(\mathbf{x}, t_{\mathrm{f}}, y)$ *and any intervention value* t *in the conditional support of* $T \mid \mathbf{X} = \mathbf{x}$*, the CATE* $\mu(\mathbf{x}, t) - \mu(\mathbf{x}, t_{\mathrm{f}})$ *is identifiable from the observational distribution, and the counterfactual outcome under* $\mathrm{do}(T = t)$ *satisfies*

$$y^t = y + \mu(\mathbf{x}, t) - \mu(\mathbf{x}, t_{\mathrm{f}}). \tag{14}$$

Consequently, CATE adjustment recovers the individual counterfactual exactly whenever the CEPO estimates are exact at both the factual and intervention treatment values.

The ability of CATE adjustment to recover individual counterfactuals depends on the causal structure. Proposition 4 establishes that an accurate CATE estimate guarantees exact recovery when (i) none of the covariates is a descendant of treatment; and (ii) the outcome is not a parent of treatment. We provide the proof of Proposition 4 in Appendix A.3.

Settings In Section 4.1, we pool prediction errors across targets within each graph when computing RMSE. Here, we report RMSE separately for each target because different choices of outcome within the same graph can satisfy or violate the conditions for exact CATE adjustment. This yields eight treatment–outcome settings across five C-Suite datasets, as illustrated in Figure 5. With all remaining observed variables used as covariates, four settings satisfy the conditions in Proposition 4: Collider linear Gaussian (1), both Fork datasets with outcome (2), and Large Backdoor. The other four settings violate these conditions. In Collider linear Gaussian (2), the covariate is a common child of treatment and outcome, so adjustment conditions on a collider that is also a treatment descendant. In both Fork datasets with outcome (1), the outcome directly causes treatment. In Nonlinear Simpson, the covariates include a descendant of the outcome and, consequently, of treatment. Thus, in these four settings, CEPO methods with CATE adjustment are not guaranteed to recover individual counterfactual outcomes exactly.

Baselines We compare CausalForestDML (Chernozhukov et al., 2018), DRNet (Schwab et al., 2020), CCPFN (Stith et al., 2026), and S-learners based on LimiX (Zhang et al., 2025) and TabPFN v3 (Grinsztajn et al., 2026). CausalForestDML directly estimates CATE. We obtain counterfactual predictions by adding the estimated CATE to the factual outcome. Other baseline methods estimate CEPO. DRNet is a neural network trained on the observational data, whereas CCPFN is a pretrained model that uses observational data as context at inference time. We also evaluate S-learners based on LimiX and TabPFN v3, using each model to predict the outcome from the covariates and treatment. For each CEPO method, we report both the estimated CEPO $\hat{\mu}(\mathbf{x}, t)$ used directly as the counterfactual prediction and the prediction obtained through CATE adjustment in Eq. 13.

Results Table 7 shows that CATE adjustment substantially improves counterfactual prediction in all four settings satisfying Proposition 4. For example, adjustment reduces CCPFN's RMSE from

Table 7: Counterfactual prediction RMSE on C-Suite datasets (mean ± standard deviation). Bold indicates the lowest mean RMSE in each column, excluding the factual-outcome baseline. Numbers in parentheses denote outcome indices. The superscript "cal" denotes an adjustment variant. Light blue marks settings satisfying the two structural conditions in Proposition 4.

	Collider linear Gaussian		Fork linear non-Gaussian	
Method	(1)	(2)	(1)	(2)
y factual	0.81 ±0.00	0.00 ±0.00	0.00 ±0.00	0.00 ±0.00
CausalForestDML	0.16 ±0.01	0.72 ±0.02	0.51 ±0.01	0.30 ±0.03
DRNet	0.64 ±0.04	0.95 ±0.03	0.76 ±0.09	0.75 ±0.07
DRNet^{cal}	0.13 ±0.01	0.56 ±0.09	0.65 ±0.10	0.46 ±0.13
CCPFN	0.58 ±0.00	0.98 ±0.00	0.56 ±0.00	0.57 ±0.00
CCPFN^{cal}	0.04 ±0.01	0.68 ±0.00	0.45 ±0.00	0.06 ±0.02
S-Learner + LimiX	0.58 ±0.00	0.99 ±0.00	0.60 ±0.00	0.56 ±0.00
S-Learner + LimiX^{cal}	0.05 ±0.01	0.68 ±0.00	0.49 ±0.00	0.06 ±0.01
S-Learner + TabPFN v3	0.58 ±0.00	1.01 ±0.00	0.61 ±0.00	0.55 ±0.00
S-Learner + TabPFN v3^{cal}	0.04 ±0.01	0.71 ±0.00	0.49 ±0.00	**0.01 ±0.00**
CAFE	**0.03 ±0.00**	**0.06 ±0.01**	**0.00 ±0.00**	**0.01 ±0.00**

	Fork nonlinear Gaussian		Large	Nonlinear
Method	(1)	(2)	backdoor	Simpson
y factual	0.00 ±0.00	0.00 ±0.00	0.74 ±0.00	1.36 ±0.00
CausalForestDML	0.38 ±0.01	0.14 ±0.01	0.58 ±0.06	1.21 ±0.01
DRNet	0.56 ±0.00	0.42 ±0.04	0.97 ±0.03	3.77 ±0.17
DRNet^{cal}	0.40 ±0.00	0.14 ±0.05	0.47 ±0.06	3.76 ±0.17
CCPFN	0.54 ±0.00	0.39 ±0.00	0.93 ±0.01	1.29 ±0.02
CCPFN^{cal}	0.34 ±0.00	0.02 ±0.00	0.42 ±0.01	1.26 ±0.03
S-Learner + LimiX	0.54 ±0.00	0.39 ±0.00	0.98 ±0.01	1.50 ±0.04
S-Learner + LimiX^{cal}	0.37 ±0.00	0.03 ±0.01	0.53 ±0.02	1.47 ±0.04
S-Learner + TabPFN v3	0.55 ±0.00	0.39 ±0.00	0.92 ±0.02	1.24 ±0.05
S-Learner + TabPFN v3^{cal}	0.37 ±0.00	0.02 ±0.00	0.41 ±0.04	1.19 ±0.05
CAFE	**0.00 ±0.00**	**0.01 ±0.00**	**0.15 ±0.02**	**0.93 ±0.01**

0.58 to 0.04 on Collider linear Gaussian (1), from 0.57 to 0.06 on Fork linear non-Gaussian (2), and from 0.39 to 0.02 on Fork nonlinear Gaussian (2). Large Backdoor also shows substantial improvements across all CEPO methods. Nevertheless, none of the adjusted methods outperforms CAFE. The remaining prediction errors may reflect inaccuracies in the estimated CEPOs.

In the four settings that violate these conditions, adjustment yields more limited improvements and leaves a substantial gap to CAFE. Collider linear Gaussian (2) illustrates both the limitation of this adaptation and the source of its partial improvement. Its graph is $T \to X \leftarrow Y$, so intervening on T leaves Y unchanged and the true counterfactual is $y^t = y$. The CEPO is therefore $\mu(x,t) = \mathbb{E}[Y \mid X = x]$, which does not depend on t. However, conditioning on the collider X induces an association between T and Y, so the observational regression $m(x,t) = \mathbb{E}[Y \mid X = x, T = t]$ varies with t. In this linear Gaussian setting, write $m(x,t) = \alpha x + \beta t + c$, with $\beta \neq 0$, and let $r = y - m(x, t_\mathrm{f})$ denote the factual prediction residual. The prediction errors before and after adjustment are then

$$m(x,t) - y^t = \beta(t - t_\mathrm{f}) - r,$$
$$y + m(x,t) - m(x,t_\mathrm{f}) - y^t = \beta(t - t_\mathrm{f}).$$

Table 8: Counterfactual prediction RMSE on C-Suite datasets (mean ± standard deviation). Full Prior denotes CAFE trained with its original prior over diverse DAGs. Backdoor Prior denotes the ablation model trained only on back-door graphs. Bold indicates the lowest mean RMSE in each column, excluding the factual-outcome baseline. Numbers in parentheses denote outcome indices. Light blue marks settings satisfying the two structural conditions in Proposition 4.

	Collider linear Gaussian		Fork linear non-Gaussian	
Method	(1)	(2)	(1)	(2)
y factual	0.81 ±0.00	0.00 ±0.00	0.00 ±0.00	0.00 ±0.00
Backdoor Prior	**0.02 ±0.00**	0.39 ±0.00	1.20 ±0.06	**0.01 ±0.00**
Full Prior	0.03 ±0.00	**0.06 ±0.01**	**0.00 ±0.00**	**0.01 ±0.00**

	Fork nonlinear Gaussian		Large	Nonlinear
Method	(1)	(2)	backdoor	Simpson
y factual	0.00 ±0.00	0.00 ±0.00	0.74 ±0.00	1.36 ±0.00
Backdoor Prior	0.26 ±0.01	**0.01 ±0.00**	**0.13 ±0.01**	1.01 ±0.02
Full Prior	**0.00 ±0.00**	**0.01 ±0.00**	0.15 ±0.02	**0.93 ±0.01**

Adjustment restores the individual's factual residual, which explains why it can reduce prediction error. It nevertheless retains the spurious treatment effect induced by conditioning on the collider. Thus, accurate observational prediction alone cannot eliminate the remaining error in this setting.

D.6 Ablation study on the training prior

We examine how restricting the causal structures used for pretraining affects counterfactual prediction. Specifically, we train an ablation model restricted to the standard back-door setting considered by CausalPFN (Balazadeh Meresht et al., 2026) and CCPFN (Stith et al., 2026). We retain the original DAG sampling procedure but change how the intervention and outcome variables are selected. For each sampled DAG, we choose T uniformly from nodes with exactly one descendant and designate that descendant as Y. If no such node exists, we resample the DAG. This yields a back-door structure: T directly affects Y, while all remaining variables $\mathbf{X}$ are pretreatment covariates that may influence treatment assignment and the outcome. The sampling procedures for structural functions and exogenous noise, the model architecture, and the training hyperparameters remain unchanged.

Results. We find that the ablation model trained with back-door prior (Table 8) shows a similar pattern of results to the CATE-adjusted baselines (Table 7). Under the back-door prior, the factual treatment t_f and outcome y provide information about the individual's outcome (centered) noise, i.e., $y - \mu(\mathbf{x}, t_\mathrm{f})$ in Eq.14, which is shared between the factual and counterfactual outcomes. The ablation model can learn to infer this noise term from its factual inputs t_f and y, implicitly performing abduction, while the CATE adjustment restores it explicitly via $y - \hat{\mu}(\mathbf{x}, t_\mathrm{f})$ in Eq.13. Both approaches can therefore recover the individual noise missing from a CEPO prediction and use the same noise under intervention to predict the individual's counterfactual outcome, explaining their strong performance in the four settings satisfying the proposition's conditions.

The other four settings involve covariates affected by treatment ($T \to X$) or outcomes that cause treatment ($Y \to T$), which the back-door prior excludes. In these cases, the adjustment above is no longer guaranteed to recover the true counterfactual. The ablation model's RMSE increases relative to full CAFE in all four settings, including from 0.06 to 0.39 on Collider linear Gaussian (2) and from 0.00 to 1.20 on Fork linear non-Gaussian (1). These results show that providing factual observations and directly supervising counterfactual predictions does not overcome the limitations of a restricted training prior. The full model's advantage, with the same architecture and prediction objective, highlights the importance of covering diverse causal relationships during pretraining.