JOINT SUPERVISION OF POTENTIAL OUTCOMES FOR CAUSAL FOUNDATION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Estimating treatment effects is central to science across many disciplines and requires counterfactual reasoning about alternative worlds. In the real world, however, each unit reveals its outcome only under the factual world, so the factual and counterfactual outcomes can never be jointly learned by any model trained on real data. The problem is therefore tractable only in synthetic settings, where both worlds can be generated. Recent progress on prior-data fitted networks (PFNs) has shown that models pre-trained purely on synthetic data can transfer to real-world prediction, achieving state-of-the-art tabular performance via in-context learning. This paradigm has since been carried over to causal effect estimation, predicting interventional outcomes given observational data. Despite this fully synthetic setup, existing causal PFNs carry the real-world constraint into their training objective, confining supervision to a single world's outcome per unit. We therefore propose a joint causal PFN training and inference architecture, apply it to existing causal PFNs, and show that it improves statistical efficiency and distributional information. Across semi-synthetic and synthetic benchmarks spanning diverse causal scenarios, jointly supervised models improve upon their baselines, establishing that joint learning is the preferable training setup for causal PFNs.

1 INTRODUCTION

Estimating causal treatment effects is central to medicine, economics, finance, and public policy (Imbens & Rubin, 2015; Athey & Imbens, 2017). The effect can be considered at the unit and at the population level (Holland, 1986). At the unit level, it is the difference between a unit's factual and counterfactual outcomes; its estimation is therefore inherently a joint learning problem over them. In practice, however, each unit is observed only under the factual setting, so the pair can never be jointly learned from real data (Heckman et al., 1997; Rubin, 1974; Peters et al., 2017; Holland, 1986). The only viable option is therefore learning each outcome separately (Shalit et al., 2017; Alaa et al., 2017). This recovers at best the two marginal distributions; the dependence between the outcomes remains unidentified even under randomization (Heckman et al., 1997; Fan & Park, 2010). Assuming the outcomes are independent makes learning practical and solves the joint problem but is implausible (Andrews & Didelez, 2021; Vansteelandt & Lancker, 2024), and its misspecification carries a real inferential cost (Ding & Dasgupta, 2016). Joint learning at unit level therefore remain tractable only in synthetic settings, in which both outcomes can be generated.

At the population level, in contrast to unit level, the treatment effect depends only on the two marginal distributions, which randomization identifies. Randomized controlled trials are therefore the ideal setup for it, but are often unethical, expensive, slow, or infeasible (Hernán & Robins, 2020). Hence, observational data is the only viable alternative. Recovering the marginals from observational data, however, requires strong assumptions, such as unconfoundedness, positivity, or a valid instrument (Hernán & Robins, 2020; Pearl, 2009; Chernozhukov et al., 2018; Wager & Athey, 2018). These assumptions are difficult to justify in practice: confounders are frequently unmeasured, valid instruments seldom exist, and the causal graph is rarely known fully.

To address the assumption problem, a family of causal foundation models based on PFNs has recently been proposed (Robertson et al., 2025; Reuter et al., 2026; Balazadeh et al., 2025; Ma et al., 2026; Thumm et al., 2026). A PFN is a transformer that is trained on synthetic data and returns a posterior-predictive distribution (Müller et al., 2022). PFNs achieve state-of-the-art performance on tabular prediction on observational data (Hollmann et al., 2023; 2025; Stith et al., 2026), showing that a model trained purely on synthetic set-up can transfer to real-world prediction. Causal PFNs adopt the

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.

same architecture but predict interventional rather than observational outcomes (Stith et al., 2026). They resolve the assumption problem by encoding the identification and structural assumptions in the training prior. They mostly differ in what those assumptions are. Non-PFN causal machine-learning methods (Chernozhukov et al., 2018; Wager & Athey, 2018; Nie & Wager, 2020; Kennedy, 2023; Künzel et al., 2019) target the same prediction outcome but are outperformed by PFN-based approaches on standard tabular benchmarks (Robertson et al., 2025; Balazadeh et al., 2025; Reuter et al., 2026).

While causal PFNs address the population level problem, the joint learning problem of the unit level remains: despite the fully synthetic setup, existing models carry the real-world constraint into their training objective, confining supervision to a single outcome per unit (Robertson et al., 2025; Reuter et al., 2026; Balazadeh et al., 2025). Two limitations follow. First, discarding the joint structure discards the dependence between the two. This dependence improves the prediction of both unit- and population-level treatment effects. The separate information is always recoverable from the joint, but not conversely; single-outcome training is therefore statistically less efficient, and the models report only a fraction of their effective capacity. Second, without learning the outcome dependency, the models cannot faithfully deliver the distributional quantities a decision maker requires: the response, unit-level and population-level treatment effect densities. Deriving all of them together under the current architecture is either infeasible or requires an independence assumption between the outcomes. We address both limitations by introducing a joint head training architecture, which supervises both generated outcomes of every unit, combined with MALC (Danisman et al., 2026) and an optimal-transport aggregation at inference. Our method improves the performance of existing causal PFNs, and delivers the missing distributional information required for both unit- and population-level decisions. We show empirically that, across semi-synthetic and synthetic benchmarks spanning diverse causal scenarios, our modification improves point estimation, coverage, and interval length over the unmodified counterparts, establishing that joint supervision is the preferable training setup for causal PFNs.

2 Background and Related Work

2.1 Structural Causal Model and Interventions

We work with an observational dataset $\mathcal{D}_{\text{obs}} = \{(X_i, T_i, Y_i)\}_{i=1}^{n}$, where each row records a unit's covariates $X_i \in \mathbb{R}^d$, the treatment it received $T_i \in \{0, 1\}$, and the outcome observed under that treatment, $Y_i \in \mathbb{R}$.

We assume $\mathcal{D}_{\text{obs}}$ is generated by a structural causal model (Pearl, 2009; Peters et al., 2017) (SCM) $\psi = (G, f, p_\varepsilon)$ over $\{X, T, Y\}$: G a directed graph, $f = \{f_V\}$ structural equations shared across all units, p_ε the exogenous noise distribution. Each row of $\mathcal{D}_{\text{obs}}$ is one sample $\varepsilon_i \sim p_\varepsilon$ propagated through f. An intervention $\text{do}(T = t)$ under SCMs sets the treatment to value t and severs its dependence on its causal parents, while leaving every other structural equation untouched. This defines the potential outcome:

$$Y^{\text{do}(t)}(x, \varepsilon) = f_Y\big(x, t, \text{pa}_{-T}(\psi; x, \varepsilon), \varepsilon_Y\big), \quad t \in \{0, 1\}, \tag{1}$$

where $\text{pa}_{-T}(\psi; x, \varepsilon)$ denotes the values of Y's graph-parents other than T, propagated through ψ. The conditional average treatment effect (CATE) at covariate value x follows as

$$\tau(x) := \mathbb{E}\big[Y^{\text{do}(1)} - Y^{\text{do}(0)} \mid X = x\big], \tag{2}$$

and the average treatment effect (ATE) is its average over the covariate population, $\mathbb{E}_X[\tau(X)]$.

2.2 Joint Supervision of Outcomes

The dependence between a unit's two potential outcomes is captured by the *cross-world correlation* (Bodik et al., 2026):

$$\rho(x) = \text{corr}\big(Y^{\text{do}(1)}, Y^{\text{do}(0)} \mid X = x\big), \tag{3}$$

which in additive structural models corresponds to the correlation between the exogenous noise terms entering the two outcomes. Since each unit is observed under a single treatment, $\rho(x)$ is not identifiable from real data (Rubin, 1974; Heckman et al., 1997) and hence remains unknown. In our setup, by contrast, $\rho(x)$ is known and serves as a controlled parameter: both potential outcomes in Section 2.1 are generated from the same SCM, so the noise correlation, and with it the dependence, is determined by the data-generating process. Because the training data is simulated

from a prior over such SCMs, the pair $\left(y^{\mathrm{do}(0)}, y^{\mathrm{do}(1)}\right)$ is generated for every unit, and the joint law $P^{\star}\left(Y^{\mathrm{do}(0)}, Y^{\mathrm{do}(1)} \mid X = x\right)$ becomes a direct supervision target. We refer to training on the pair as *joint supervision*.

Earlier methods also tried to model the two outcomes together. In real data, however, only one of them is observed per unit, so the dependence between them had to be assumed rather than learned. The multi-task Gaussian process of Alaa et al. (2017) places a correlated prior over the two outcome functions, yet the residual dependence between the outcomes remains unidentified and untrained. CEVAE (Louizos et al., 2017) generates both outcomes from a shared latent confounder, so the coupling follows from the latent-variable assumption rather than from data. Closest to our setting, deep twin networks (Vlontzos et al., 2023) assume both outcomes share one noise draw, infer the noise from the observed outcome, and propagate it under the other treatment. Factual data, however, does not determine the counterfactual, as many mechanisms fit it equally well while disagreeing on the counterfactuals. Uniqueness is instead obtained by a monotonicity assumption, which the real data can never verify. Similarly, GANITE (Yoon et al., 2018) outputs both outcomes together, but the missing one is generated by the model itself. It's loss never sees a true pair.

Across these approaches, due to the limitations of real data, the dependence is assumed rather than learned. In the PFN setting, the training data is generated rather than observed, and both outcomes exist for every unit. Hence, the disagreement that real data cannot settle is resolved by the generated pair. This provides, to our knowledge, the first opportunity to learn the dependence, under the assumption underlying all PFNs: that models trained on synthetic data transfer to reality, as established for tabular prediction (Hollmann et al., 2023; 2025).

2.3 Prior-data fitted networks for Bayesian inference

We define Bayesian prediction as the task of computing the posterior predictive distribution (PPD), $p(y \mid x, \mathcal{D})$, for a query x given an observed dataset $\mathcal{D}$. Prior-Data Fitted Networks (Müller et al., 2022) (PFNs) amortize the computation of posterior predictions: a neural network q_ϕ is trained once, offline, on synthetic pairs $(\theta, \mathcal{D})$ sampled from the prior $p(\theta)$, directly approximating the PPD. At inference time, a real dataset $\mathcal{D}_{\text{real}}$ is provided as context, and the network outputs an approximation to $p(y \mid x, \mathcal{D}_{\text{real}})$ in a single forward pass without gradient updates.

2.4 Causal foundation models

Causal foundation models (CFMs) apply the PFN paradigm to causal effect estimation. In this setting, the hypothesis space consists of structural causal models (SCMs) $\psi \sim p(\psi)$.

Existing CFMs differ mainly in how much causal structure their prior commit to. Do-PFN (Robertson et al., 2025) samples SCMs from a broad prior with no identifiability restrictions and the model is trained to predict interventional outcomes from $\mathcal{D}_{\text{obs}}$ alone, without knowledge of the graph. Use What You Know (UWYK) (Reuter et al., 2026) uses an unrestricted prior similar to Do-PFN, but supplies causal graph knowledge at inference time as a partially known ancestral matrix $\tilde{\mathbf{T}} \in \{-1, 0, 1\}^{(d+2)\times(d+2)}$, injected into the feature-wise attention of the transformer, so that a single model spans the range from no graph information to full ancestral knowledge. CausalPFN (Balazadeh et al., 2025) constructs its training prior from a mixture of real and synthetic base tables. Two randomly selected columns of each table are treated as the potential outcomes $Y^{\mathrm{do}(0)}$ and $Y^{\mathrm{do}(1)}$, and the treatment is drawn as $T \sim \mathrm{Bernoulli}(\sigma(f(X)))$ for a random propensity function f that depends on X alone. This construction guarantees ignorability by design. The model is then trained to predict the conditional expected potential outcome $\mu_t(x) := \mathbb{E}[Y^{\mathrm{do}(t)} \mid X = x]$ separately for each $t \in \{0, 1\}$, and these estimates are valid only when ignorability holds. CausalFM (Ma et al., 2026) uses separate priors encoding different identifiability constraints, so each model satisfies its constraint by design but covers only that setting. It predicts the conditional average treatment effect $\tau(x)$ directly, without modeling either $Y^{\mathrm{do}(0)}$ or $Y^{\mathrm{do}(1)}$; because the map $(Y^{\mathrm{do}(0)}, Y^{\mathrm{do}(1)}) \mapsto \tau(x)$ is not injective, neither potential outcome can be recovered from $\tau(x)$ alone.

Despite these different approaches, no CFM performs the joint supervision of Section 2.2: although every training prior generates both potential outcomes of each unit, all models retain single-outcome supervision. Two limitations follow.

Statistical efficiency. CFMs (Robertson et al., 2025; Reuter et al., 2026; Balazadeh et al., 2025) are trained with a single-outcome head on (T_i, Y_i) pairs, using only the outcome under the treatment

assigned to each unit. When $\rho(x) \neq 0$, observing one outcome improves the estimate of the other, and the improvement is at worst zero, in which case the joint is as informative as the single outcome. In addition, the pair simply contains more data: each unit contributes two outcomes to the loss instead of one. More generally, every marginal can be derived from the joint, but not every joint can be derived from the marginals. Training on the joint is therefore strictly more efficient than training on a single outcome. Section 3.1 quantifies this gap, and our ablations separate the effect of the dependence from the effect of the additional data.

Distributional information. The primary purpose of any model is to provide the maximum possible information for a better-informed decision. Existing CFMs, however, report only a limited view: some provide a scalar CATE together with response densities (Robertson et al., 2025; Reuter et al., 2026; Balazadeh et al., 2025), while others provide a CATE density with no response information (Ma et al., 2026), and none delivers an ATE density. A decision maker generally requires all three: the response densities support outcome decisions (e.g., whether a treated unit's outcome exceeds a clinical or economic threshold), while the CATE and ATE densities support the treatment decision itself (e.g., whether to intervene at all). One might argue that a point-wise CATE estimate combined with the marginal densities already provides sufficient information; however, the same point estimate and confidence interval can arise from different shapes of uncertainty that call for different actions. For example, a near-uniform effect density and a bimodal density with two contrasting (negative and positive) modes can yield identical point estimates and intervals: the uniform density indicates that no action can be preferred over another, whereas the bimodal density indicates that with domain knowledge one may favor one mode over the other. To recover the missing distributional information, one might attempt to reconstruct a CATE distribution from the marginals, but this amounts to setting $\rho(x) \equiv 0$, the independence assumption that Section 2.2 shows to be implausible for outcomes of the same unit (Andrews & Didelez, 2021; Vansteelandt & Lancker, 2024) and unverifiable from real data; discarding the coupling in this way yields spurious treatment-effect values.

Our proposed joint head architecture resolves these gaps, providing both statistical efficiency and the full distributional information while simultaneously refining the marginal response distribution that current models already provide; it therefore plays an additive role, both improving what existing models report and delivering the additional information decision makers require.

3 Methodology

To address the two gaps identified in Section 2.4, we replace the single-outcome head with a joint head that outputs the density $p(y^{\mathrm{do}(0)}, y^{\mathrm{do}(1)} \mid x, \mathcal{D}_{\mathrm{obs}})$ over the paired potential outcomes, trained under the joint supervision of Section 2.2. To unfold our methodology, we first establish the statistical efficiency gained by the joint objective in Section 3.1, and then show in Section 3.2 that the joint density recovers the dependence between $Y^{\mathrm{do}(0)}$ and $Y^{\mathrm{do}(1)}$ that no pair of marginals can express, enabling distributional inference over the CATE and ATE.

3.1 Statistical efficiency of the joint objective

For each query covariate X with observational context $\mathcal{D}_{\mathrm{obs}} = \{(X_i, T_i, Y_i)\}_{i=1}^{N}$, the joint head predicts a density $f(y^{\mathrm{do}(0)}, y^{\mathrm{do}(1)} \mid X, \mathcal{D}_{\mathrm{obs}})$ on $\mathbb{R}^2$. The training loss is the negative log-density against paired labels drawn from the SCM prior,

$$\mathcal{L}(y^{\mathrm{do}(0)}, y^{\mathrm{do}(1)} \mid X, \mathcal{D}_{\mathrm{obs}}) \;=\; -\log f(y^{\mathrm{do}(0)}, y^{\mathrm{do}(1)} \mid X, \mathcal{D}_{\mathrm{obs}}), \tag{4}$$

averaged over queries in a batch. Positivity of f on all of $\mathbb{R}^2$ is required for the loss to remain finite. We enforce it by combining a bounded interior $K \times K$ grid on $\mathcal{I} = [a, b]^2$, $a < b$, with Gaussian tails along the four extended-axis and four diagonal directions, producing a nine-region mixture. Most of the learning occurs on inner region. Appendix B details the construction of the training.

Learning objectives. Consider estimating $\tau(x) = \mathbb{E}[Y^{\mathrm{do}(1)} - Y^{\mathrm{do}(0)} \mid X = x]$ with calibrated uncertainty, reported as a posterior credible interval. Let $Y^{\mathrm{do}(t)} = \mu_t(x) + \varepsilon^{\mathrm{do}(t)}$ with $\mathrm{cor}(\varepsilon^{\mathrm{do}(0)}, \varepsilon^{\mathrm{do}(1)}) = \rho(x)$, $\sigma_t^2(x) = \mathrm{Var}(\varepsilon^{\mathrm{do}(t)} \mid X = x)$, and effect variance

$$v(x) \;=\; \mathrm{Var}\big(Y^{\mathrm{do}(1)} - Y^{\mathrm{do}(0)} \mid X = x\big) \;=\; \sigma_1^2(x) + \sigma_0^2(x) - 2\rho(x)\,\sigma_1(x)\,\sigma_0(x). \tag{5}$$

Causal PFNs approximate $\tau(x)$ and $v(x)$ through the posterior predictive distribution. However, their learning target can be different, and this results in differences across approximations. For example, in

contrast to joint learning in Equation 4, a single-outcome head can approximate $\tau(x)$ by learning each outcome separately under $\mathcal{L}(y^{\mathrm{do}(t)} \mid X, \mathcal{D}_{\mathrm{obs}})$, so $(\varepsilon^{\mathrm{do}(0)}, \varepsilon^{\mathrm{do}(1)})$ would enter as independent draws. Hence, even a perfect single-outcome model in this format is at minimum wrong by $\varepsilon^{\mathrm{do}(1)} - \varepsilon^{\mathrm{do}(0)}$. If the model is instead trained to learn $\mu_t(x)$ directly, a perfect model estimates $\tau(x) = \mu_1(x) - \mu_0(x)$ without error. The dependence, however, would be lost. Taking the expectation removes the noise, as $\mathbb{E}[\varepsilon^{\mathrm{do}(t)} \mid X = x] = 0$; the targets $\mu_0(x)$ and $\mu_1(x)$ therefore contain no information about $\rho(x)$. The effect variance $v(x)$ thus remains unidentified, and no calibrated credible interval can be reported. However, the joint head of Equation 4 targets the full law $f(y^{\mathrm{do}(0)}, y^{\mathrm{do}(1)} \mid x, \mathcal{D}_{\mathrm{obs}})$, which contains $\mu_0(x)$, $\mu_1(x)$, and $\rho(x)$ at once: the shared draw cancels the noise in the effect, recovering $\tau(x)$ without error, while the learned dependence makes the approximation of $v(x)$ more faithful. Therefore, of the three settings, joint learning operates on the largest information set.

Data efficiency. The single-outcome observation is a randomised function of the joint one, so any parameter that can be estimated from it can be estimated from the joint at least as well (Lemma 2, Appendix C). Hence, nothing is lost by joint training. Theorem 1, in contrast to the preservation, formalizes the advantage gained by joint data. The single-outcome observation informs $Y^{\mathrm{do}(t)}$ only for the t selected by T_i; the discarded share propagates to the error of the marginal densities and of the CATE.

Theorem 1. *Let $T_i \sim \mathrm{Bernoulli}(\pi)$ with $\pi \in (0,1)$, write $\pi_1 = \pi$ and $\pi_0 = 1 - \pi$, and assume standard MLE regularity. For every query x:*

(i) for each $t \in \{0,1\}$, the marginal density estimators satisfy

$$\lim_{N\to\infty} N \cdot \mathbb{E}\big[(\hat{p}_{\mathrm{U}}(y \mid x,t) - p^{\star}(y \mid x,t))^2\big] \;=\; \frac{1}{\pi_t} \cdot \lim_{N\to\infty} N \cdot \mathbb{E}\big[(\hat{p}_{\mathrm{R}}(y \mid x,t) - p^{\star}(y \mid x,t))^2\big]; \quad (6)$$

(ii) the CATE estimator $\hat{\tau}(x) = \hat{\mu}_1(x) - \hat{\mu}_0(x)$ satisfies

$$\lim_{N\to\infty} \frac{\mathbb{E}[(\hat{\tau}^{\,\mathrm{U}}(x) - \tau(x))^2]}{\mathbb{E}[(\hat{\tau}^{\,\mathrm{R}}(x) - \tau(x))^2]} \;=\; \frac{1}{2\pi(1-\pi)}. \quad (7)$$

Proofs are given in Appendix C. The results are asymptotic in N and hold at a fixed number of training units. Since the gain is a data efficiency gain, it is large when the training data is limited; as the data grows, both objectives converge to the truth and the gain diminishes. The gain may therefore reflect the additional data rather than the joint structure itself. To separate the two, we conduct an ablation in which the baseline is trained on the same amount of label information; any remaining difference is then attributable to the joint head.

Computational cost and resolution. The learning objectice in Equation 4 requires $K^2 + 9 + 4$ trainable parameters per query, so the head's cost grows quadratically in the per-dimension resolution, whereas a single-outcome head grows linearly. In our experiments, to keep the comparison against the published baselines at similar cost, we train jointly supervised models at $K = \sqrt{J}$, where J is the grid resolution of the baseline single-outcome model. Reducing the per-dimension resolution, however, shrinks the model's representational capacity, degrading estimation quality across both point and distributional metrics. To mitigate this, we apply MALC (Danisman et al., 2026), a mixture-of-log-concave estimator, at inference, which reconstructs a continuous density from the coarse grid and reduces the resolution loss. Nevertheless, this places the jointly supervised model at a disadvantage relative to its baseline; improvements under this setup therefore reflect a conservative gain. We further conduct an ablation with baselines trained at the same resolution K for a matched comparison. The comparisons thus give insight into the trade-off between cost and performance: whether, under a fixed budget, one should prioritize resolution or joint learning, and which design is preferable when the budget is not a limitation.

3.2 Distributional inference for CATE and ATE

Sections 3.2.1 and 3.2.2 describe how the per-unit CATE density and the population-level ATE density are recovered from the joint head's output.

3.2.1 Per-unit CATE distribution via MALC

The joint head produces a discrete distribution on a $K \times K$ grid with mass p_{ij} at bin centre $(y_0^i,\, y_1^j)$. The discrete CATE distribution follows directly by the change of variables $(Y^{\mathrm{do}(0)}, Y^{\mathrm{do}(1)}) \mapsto$

$(Y^{\mathrm{do}(0)}, \tau)$ with $\tau := Y^{\mathrm{do}(1)} - Y^{\mathrm{do}(0)}$, marginalising out the first coordinate: each grid cell (y_0^i, y_1^j) contributes its mass p_{ij} to the difference value $y_1^j - y_0^i$,

$$f_\tau(t) \;=\; \sum_{ij} p_{ij}\, \mathbf{1}\{y_1^j - y_0^i = t\}. \tag{8}$$

The CATE point estimate is the mean of this discrete distribution, $\hat{\tau}(x) \;=\; \sum_{ij} p_{ij}\,\big(y_1^j - y_0^i\big)$, and uses no further approximation. MALC (Danisman et al., 2026) is applied to the discrete distribution f_τ, fitting a continuous density

$$\tilde{f}_\tau(t) \;=\; \sum_{m=1}^{M} w_m\, g_m(t), \qquad g_m \text{ log-concave}, \quad w_m \geq 0, \quad \sum_m w_m = 1, \tag{9}$$

as a mixture of M log-concave components. The smoothed $\tilde{f}_\tau$ interpolates between the discrete difference values, so the CATE density can be evaluated at any t in the range rather than only at the grid's differences; this removes the coarse-bin limitation of the reduced resolution $K = \sqrt{J}$. The full fitting procedure is given in Appendix D.

3.2.2 ATE DISTRIBUTION VIA OPTIMAL TRANSPORT

The per-query CATE densities should be used to infer the ATE density. Averaging them pointwise gives the mixture $\frac{1}{N}\sum_n \tilde{f}_\tau^{(n)}$. However, the mixture describes the effect of a randomly drawn unit, not the average effect: its spread mixes the within-query uncertainty with the heterogeneity across covariates, so narrow densities at different locations aggregate into a wide one. The average effect requires averaging the locations while preserving the shape, leading to the 2-Wasserstein barycenter as the natural choice.

Given the per-query CATE densities $\{\tilde{f}_\tau^{(n)}\}_{n=1}^N$ across a covariate sample $\{X^{(n)}\}_{n=1}^N$, we define the ATE density f_{ATE} as their 2-Wasserstein barycenter with uniform weights. For univariate densities on a common support this barycenter admits the closed-form quantile average

$$Q_{\mathrm{ATE}}(u) \;=\; \frac{1}{N}\sum_{n=1}^{N} Q_\tau^{(n)}(u), \qquad u \in (0,1), \tag{10}$$

where $Q_\tau^{(n)} = (\tilde{F}_\tau^{(n)})^{-1}$ is the quantile function of $\tilde{f}_\tau^{(n)}$. The smoothing of Section 3.2.1 makes this well posed: the continuous $\tilde{F}_\tau^{(n)}$ invert cleanly, whereas the step functions of the raw grid would reduce equation 10 to an average over the grid's difference values. The mean of f_{ATE} is the average of the per-query CATE means, so the ATE point estimate is unchanged, while providing additional density information. To our knowledge, no prior CFM has focused on ATE density before, although it remains an informative tool for population-level decisions. We hope this opens space for future consideration of how to construct it better in the PFN setting. Derivation and the grid-level algorithm are given in Appendix E, and Figure 1 illustrates the procedure.

4 EXPERIMENTAL SETUP

We evaluate our proposed joint supervision method across three causal PFN models and three benchmarks, measuring point estimation, density estimation, and uncertainty calibration. This section describes the compared baselines, ablations, benchmarks, and evaluation metrics.

Compared baselines. We evaluate the effect of joint supervision on three major causal PFN models: Do-PFN (Robertson et al., 2025), UWYK (Reuter et al., 2026), and CausalPFN (Balazadeh et al., 2025). For each model, we hold the training data generation and the transformer architecture fixed and change only the learning objective to Equation 4; the jointly supervised variant is denoted as 2D throughout the results. Another influential causal PFN model, CausalFM (Ma et al., 2026), is excluded because it trains a separate model per identification setting, which is computationally heavy to replicate across our ablation setups. We see no obvious reason why the improvements would not carry over, and treat the three selected models as representative.

We compare each 2D model against its published checkpoint except for CausalPFN. The main results for that use a re-trained baseline. The published model is trained in two stages: the first predicts observational outcomes on real datasets, and the second carries those weights over and predicts

interventional outcomes on the synthetic setup. The first stage does not transfer to the joint setting: its weights do not match the joint head's bin structure, and its real datasets contain no counterfactual outcomes to pair. We therefore apply the joint head to the second, causal stage and re-train the baseline from the same stage under the same protocol for a fair comparison. Results against the full two-stage checkpoint are nevertheless reported in the Appendix I.3, where similar improvements are observed. For UWYK, the authors provide two training options: binary and continuous treatment prediction. The published checkpoint covers only the continuous version, so for fairness we also re-train the binary version, with results in the Appendix I.4. For Do-PFN, the comparison is direct and requires no additional baseline training.

Although all three models are single-outcome based, CausalPFN's learning target differs from that of Do-PFN and UWYK. This naturally influences what joint supervision can improve for each. Do-PFN and UWYK predict the outcome densities, so we expect gains in both point estimation and coverage. CausalPFN instead predicts the conditional expectation $\mu_t(x)$, which removes the noise from its target; the point estimate is therefore already favourable, and we expect the gain to lie in the coverage, with point estimates comparable if not worse.

Ablations. We conduct two ablations, following the arguments in Section 3.1. In the first, each baseline is re-trained at the reduced resolution K of its 2D counterpart. This isolates the effect of joint supervision from the effect of resolution. In the second, each baseline is given the same amount of data information. This isolates the effect of the dependence from the effect of the additional data. Each baseline thus enters with three variants: its 2D version, a 1D version at the matched resolution K, and a 1D version with the matched label budget. Together with the CausalPFN second-stage and UWYK binary re-trainings, this amounts to twelve trained models in total. The ablation results are reported in Appendix I.1 and computation details in J.

Benchmarks. We evaluate on three benchmarks. RealCause (Neal et al., 2021) is semi-synthetic and generates the two outcomes independently, reflecting the $\rho \equiv 0$ setting. ComplexMech is adopted from UWYK and consists of randomly drawn structural causal models with nonlinear mechanisms and binarized for treatment. The case study is adopted from Do-PFN and spans six causal scenarios: observed confounding, observed mediation, their combination, unobserved confounding, frontdoor and backdoor structures. Both synthetic benchmarks carry dependence between the outcomes; construction details are given in Appendix G. The synthetic benchmarks are run across covariate dimensions d, and we additionally vary the correlation from 0 to 1; these variation results are reported in Appendix I.2.

Evaluation metrics. We report point estimation and posterior credible interval coverage with length. Improvements are summarised by geometric mean (Δ) and lengths normalised by the outcome scale (Len/σ_Y); definitions and derivations are given in Appendix F. Point estimation tests the efficiency argument of Section 3.1, and coverage and length test the dependence argument of Section 3.2. Throughout the main results, all metrics are reported for the CATE.

5 Results

We report the semi-synthetic results in Section 5.1 and the synthetic results in Section 5.2.

5.1 Semi-synthetic

RealCause, although a commonly used benchmark in causal inference, generates the factual and counterfactual outcomes independently (Appendix G). We therefore treat it as a robustness test in which the dependence assumption is violated. Point estimates are reported in Table 1 and interval calibration in Table 2; 2D (Indep) denotes the jointly supervised model with its learned dependence disabled at inference, combining the two marginals as independent.

Finding 1: 2D models improve point estimation even under independent generation. Do-PFN 2D and UWYK 2D improve upon their baselines in both unit and population level metrics, with UWYK No-Anc 2D better on all ten cells. For CausalPFN, the improvement appears only at the population level: as discussed in the evaluation metrics, changing the objective from the expectation to the outcomes themselves is not expected to improve point estimation, and trades it for coverage and length.

Finding 2: under misspecification, the joint loses coverage; its marginals remain calibrated. Under misspecification, the 2D models produce intervals 10–20% shorter than the baseline models with

Table 1: Unit- and population-level point estimation on RealCause across five datasets ($\sqrt{\text{PEHE}}$ (Precision in Estimation of Heterogeneous Effect), ϵ_{ATE}; lower is better; mean $\pm$ standard error of the mean over realizations). Bold marks the better model within each baseline. Δ summarises the relative improvement across datasets by geometric mean; positive favours 2D.

	IHDP		ACIC		CPS		PSID		PSID_bal		Δ (%)	
Model	$\sqrt{\text{PEHE}}\downarrow$	$\epsilon_{\text{ATE}}\downarrow$	$\sqrt{\text{PEHE}}\downarrow$	$\epsilon_{\text{ATE}}\downarrow$	$\sqrt{\text{PEHE}}\downarrow$	$\epsilon_{\text{ATE}}\downarrow$	$\sqrt{\text{PEHE}}\downarrow$	$\epsilon_{\text{ATE}}\downarrow$	$\sqrt{\text{PEHE}}\downarrow$	$\epsilon_{\text{ATE}}\downarrow$	$\sqrt{\text{PEHE}}$	ϵ_{ATE}
Do-PFN	6.065 $\pm$0.894	0.932 $\pm$0.398	**4.114** $\pm$0.520	**0.664** $\pm$0.039	**12014** $\pm$32	0.880 $\pm$0.006	20907 $\pm$138	0.927 $\pm$0.007	22892 $\pm$161	1.090 $\pm$0.012	**+8.6**	**+62.9**
Do-PFN 2D	**6.037** $\pm$0.802	**0.530** $\pm$0.088	4.178 $\pm$0.513	0.741 $\pm$0.055	12254 $\pm$18	**0.823** $\pm$0.002	**17238** $\pm$152	**0.107** $\pm$0.007	**17196** $\pm$153	**0.113** $\pm$0.009	–	–
UWYK No-Anc	6.288 $\pm$0.788	0.669 $\pm$0.048	3.350 $\pm$0.504	0.445 $\pm$0.083	12562 $\pm$37	0.958 $\pm$0.005	22443 $\pm$132	0.976 $\pm$0.002	21886 $\pm$136	0.935 $\pm$0.004	**+12.9**	**+37.3**
UWYK No-Anc 2D	**4.508** $\pm$0.626	**0.375** $\pm$0.088	**2.743** $\pm$0.474	**0.134** $\pm$0.042	**12299** $\pm$31	**0.876** $\pm$0.007	**21426** $\pm$145	**0.849** $\pm$0.012	**20039** $\pm$141	**0.676** $\pm$0.009	–	–
UWYK Anc	5.490 $\pm$0.773	0.490 $\pm$0.079	2.768 $\pm$0.483	0.233 $\pm$0.083	**10773** $\pm$58	**0.613** $\pm$0.017	23304 $\pm$156	1.056 $\pm$0.008	**19039** $\pm$212	**0.625** $\pm$0.017	**+1.1**	**+10.1**
UWYK Anc 2D	**4.305** $\pm$0.623	**0.325** $\pm$0.077	**2.722** $\pm$0.468	**0.117** $\pm$0.034	12475 $\pm$28	0.915 $\pm$0.005	**21850** $\pm$136	**0.903** $\pm$0.009	21467 $\pm$141	0.864 $\pm$0.006	–	–
CausalPFN-C	**0.825** $\pm$0.145	**0.052** $\pm$0.022	1.412 $\pm$0.220	0.090 $\pm$0.020	**10043** $\pm$181	**0.352** $\pm$0.042	18281 $\pm$297	0.684 $\pm$0.026	18645 $\pm$314	0.707 $\pm$0.028	-1.6	**+23.3**
CausalPFN-C 2D	1.200 $\pm$0.105	0.071 $\pm$0.014	**1.264** $\pm$0.204	**0.090** $\pm$0.018	10902 $\pm$176	0.389 $\pm$0.040	**15557** $\pm$212	**0.253** $\pm$0.016	**16769** $\pm$310	**0.334** $\pm$0.025	–	–

Table 2: 95% posterior credible interval calibration for the CATE on RealCause across five datasets (Coverage, Length). Coverage is computed per realization as the share of queries whose true CATE falls in the interval, and each column reports the mean $\pm$ standard deviation across realizations; lengths are averaged the same way. Coverage closest to the 0.95 is best; shorter length is better at comparable coverage. Average reports the mean coverage and the length normalised by the outcome scale (Len/σ_Y). 2D (Indep) disables the learned dependence at inference.

	IHDP		ACIC		CPS		PSID		PSID_bal		Average	
Model	Coverage	Length	Coverage	Length	Coverage	Length	Coverage	Length	Coverage	Length	Coverage	Len/σ_Y
Do-PFN	0.988 $\pm$0.018	25.95 $\pm$31.93	0.986 $\pm$0.021	26.39 $\pm$2.72	0.943 $\pm$0.023	48914 $\pm$1177	0.979 $\pm$0.009	86249 $\pm$2989	0.972 $\pm$0.012	86444 $\pm$5467	0.974	5.46
Do-PFN 2D	0.876 $\pm$0.168	16.60 $\pm$20.99	0.792 $\pm$0.228	10.73 $\pm$4.24	0.412 $\pm$0.022	16756 $\pm$709	0.885 $\pm$0.022	44751 $\pm$1332	0.894 $\pm$0.028	46219 $\pm$2699	0.772	2.66
Do-PFN 2D (Indep)	0.992 $\pm$0.015	26.85 $\pm$30.45	0.989 $\pm$0.021	27.17 $\pm$3.71	0.994 $\pm$0.002	50830 $\pm$186	0.972 $\pm$0.011	66848 $\pm$2134	0.968 $\pm$0.013	64829 $\pm$4220	0.983	5.01
UWYK No-Anc	0.994 $\pm$0.013	25.22 $\pm$26.18	0.993 $\pm$0.013	23.00 $\pm$4.81	0.950 $\pm$0.022	44818 $\pm$943	0.882 $\pm$0.026	59930 $\pm$2474	0.896 $\pm$0.028	59966 $\pm$3175	0.943	4.49
UWYK No-Anc 2D	0.997 $\pm$0.007	26.55 $\pm$29.74	0.979 $\pm$0.039	17.63 $\pm$7.92	0.807 $\pm$0.079	32345 $\pm$4521	0.673 $\pm$0.151	38456 $\pm$9520	0.886 $\pm$0.047	54241 $\pm$4902	0.868	3.68
UWYK No-Anc 2D (Indep)	0.999 $\pm$0.003	31.59 $\pm$38.94	0.990 $\pm$0.019	24.47 $\pm$7.97	0.904 $\pm$0.013	43378 $\pm$1165	0.920 $\pm$0.029	76185 $\pm$5184	0.966 $\pm$0.015	82357 $\pm$5774	0.956	5.29
UWYK Anc	0.991 $\pm$0.017	23.11 $\pm$24.73	0.994 $\pm$0.014	20.03 $\pm$4.32	0.977 $\pm$0.010	44641 $\pm$1196	0.924 $\pm$0.019	70949 $\pm$4628	0.960 $\pm$0.017	69246 $\pm$5575	0.969	4.58
UWYK Anc 2D	0.995 $\pm$0.010	22.76 $\pm$25.71	0.977 $\pm$0.038	16.18 $\pm$6.98	0.698 $\pm$0.125	27754 $\pm$5118	0.577 $\pm$0.154	32549 $\pm$8797	0.689 $\pm$0.090	38958 $\pm$5099	0.787	3.09
UWYK Anc 2D (Indep)	0.999 $\pm$0.005	27.07 $\pm$33.38	0.989 $\pm$0.023	21.79 $\pm$6.95	0.900 $\pm$0.011	42811 $\pm$1157	0.877 $\pm$0.036	67531 $\pm$4619	0.912 $\pm$0.031	70555 $\pm$4960	0.935	4.72
CausalPFN-C	0.997 $\pm$0.009	4.70 $\pm$1.27	0.997 $\pm$0.005	8.76 $\pm$2.59	0.976 $\pm$0.028	47788 $\pm$2631	0.949 $\pm$0.017	76381 $\pm$12426	0.934 $\pm$0.020	71295 $\pm$12172	0.971	3.63
CausalPFN-C 2D	0.999 $\pm$0.004	13.81 $\pm$15.28	0.998 $\pm$0.003	12.55 $\pm$1.52	0.952 $\pm$0.052	41861 $\pm$1944	0.970 $\pm$0.012	59139 $\pm$8881	0.960 $\pm$0.013	59309 $\pm$9152	0.976	3.58
CausalPFN-C 2D (Indep)	0.999 $\pm$0.003	13.90 $\pm$15.09	0.999 $\pm$0.003	14.62 $\pm$1.67	0.961 $\pm$0.045	43677 $\pm$1696	0.980 $\pm$0.009	71684 $\pm$10407	0.970 $\pm$0.011	66849 $\pm$10608	0.982	3.98

reduced coverage. However, when the specification is corrected, the 2D (Indep) rows recover coverage to 0.94–0.98, with interval lengths competitive with the baselines despite the lower resolution and better at matched resolution (Appendix I.1). This confirms that the marginals inside the joint head remain calibrated and that the under-coverage stems from assuming dependence.

5.2 Synthetic

Both synthetic benchmarks generate the two outcomes under dependence. Due to space constraints, we report calibration in the main text (Tables 3 and 4), and leave the point-estimation tables in Appendix H.

Finding 3: joint supervision preserves or improves coverage and shortens intervals when the dependence is present. We observe the largest improvement from joint supervision for Do-PFN. On the case studies, its coverage increases by four points over the single-outcome model while the credible length reduces to less than half, and on ComplexMech it holds near 95% coverage on average at one seventeenth of the length. UWYK shows a similar pattern. On the case studies, both the ancestral and no-ancestral variants move from severe under-coverage to higher coverage under joint supervision, gaining up to twenty points while reporting only half the credible length; on ComplexMech, the coverage improves by up to five points at around half the length. Comparable gains also appear in point estimation, between 20 and 90 percent for Do-PFN and UWYK (Appendix H).

For CausalPFN, the change is small and partly unfavourable. Coverage increases by up to three points on both benchmarks, but the length does not shorten, as the reduced resolution is most costly for this model.[1] Accordingly, it shows clear improvements in the matched-resolution ablation (Appendix I.1).

[1] During training, CausalPFN outputs the response on a fixed grid over $[-10, 10]$ with a bin width of ≈ 0.01 on standardized outcomes. For our jointly supervised variant the width grows roughly sixty-fold to 0.625, so about half of the training data falls within two bins, making the resolution reduction especially costly for this model; for Do-PFN and UWYK, the share of data within two bin widths is dramatically smaller.

Table 3: 95% posterior credible interval calibration for the CATE on the six causal scenarios of the Do-PFN case studies (Coverage, Length), computed as in Table 2. Results are aggregated across varying treatment effect magnitudes and covariate dimensions.

	Observed Confounder		**Observed Mediator**		**Observed M+C**		**Unobserved Confounder**		**Frontdoor**		**Backdoor**		**Average**	
Model	Coverage	Length	Coverage	Length	Coverage	Length	Coverage	Length	Coverage	Length	Coverage	Length	Coverage	Len/σ_Y
Do-PFN	0.996 ±0.026	5.70 ±4.25	0.662 ±0.381	0.72 ±0.55	0.921 ±0.176	4.05 ±7.75	0.995 ±0.036	6.18 ±4.28	1.000 ±0.000	4.35 ±3.97	1.000 ±0.003	4.54 ±5.39	0.929	3.01
Do-PFN 2D	0.983 ±0.109	2.62 ±2.23	0.953 ±0.160	0.90 ±1.08	0.929 ±0.198	2.34 ±1.63	0.983 ±0.103	2.84 ±2.36	0.995 ±0.058	1.49 ±1.53	0.990 ±0.087	1.54 ±1.56	0.972	1.32
UWYK No-Anc	0.960 ±0.102	5.37 ±4.71	0.225 ±0.380	0.69 ±0.51	0.483 ±0.457	2.94 ±2.53	0.987 ±0.046	6.11 ±5.17	1.000 ±0.007	4.23 ±5.34	0.999 ±0.011	4.24 ±5.27	0.776	2.79
UWYK No-Anc 2D	0.999 ±0.012	3.19 ±2.21	0.916 ±0.274	0.98 ±0.59	0.997 ±0.028	3.54 ±1.98	0.998 ±0.021	3.59 ±2.41	0.998 ±0.015	2.06 ±1.47	0.997 ±0.022	2.20 ±1.60	0.984	1.77
UWYK Anc	0.998 ±0.019	4.25 ±4.35	0.492 ±0.351	1.06 ±0.72	0.640 ±0.391	3.19 ±2.34	1.000 ±0.006	5.28 ±5.25	1.000 ±0.011	3.80 ±4.94	1.000 ±0.003	4.13 ±5.18	0.855	2.60
UWYK Anc 2D	0.995 ±0.042	2.43 ±2.05	0.921 ±0.266	0.84 ±0.47	0.997 ±0.029	3.28 ±1.70	0.998 ±0.020	3.57 ±2.43	0.998 ±0.013	2.53 ±1.69	0.997 ±0.020	1.52 ±1.44	0.984	1.63
CausalPFN-C	0.988 ±0.032	1.49 ±2.90	0.808 ±0.245	1.86 ±1.16	0.920 ±0.195	3.34 ±3.70	0.987 ±0.053	2.75 ±3.38	0.998 ±0.009	1.24 ±2.15	0.999 ±0.007	1.24 ±2.12	0.950	1.33
CausalPFN-C 2D	0.999 ±0.007	2.85 ±1.65	0.920 ±0.200	1.86 ±1.11	0.973 ±0.084	2.54 ±1.38	0.997 ±0.013	3.18 ±2.06	0.999 ±0.007	2.27 ±1.88	0.999 ±0.004	2.35 ±1.83	0.981	1.77

Table 4: 95% posterior credible interval calibration for the CATE on ComplexMech, reported per graph size (Coverage, Length), computed as in Table 2. The treatment is binarized from the original continuous benchmark (Appendix G).

	n=5		n=10		n=20		n=30		n=40		n=50		**Average**	
Model	Coverage	Length	Coverage	Length	Coverage	Length	Coverage	Length	Coverage	Length	Coverage	Length	Coverage	Len/σ_Y
Do-PFN	0.955 ±0.194	6.71 ±19.50	0.961 ±0.132	7.15 ±21.79	0.979 ±0.108	2.05 ±3.79	0.990 ±0.088	2.51 ±5.96	0.999 ±0.010	3.42 ±13.49	0.985 ±0.074	3.19 ±10.67	0.978	22.06
Do-PFN 2D	0.899 ±0.247	0.26 ±0.27	0.949 ±0.189	0.23 ±0.19	0.962 ±0.146	0.24 ±0.22	0.971 ±0.097	0.25 ±0.25	0.974 ±0.130	0.24 ±0.23	0.983 ±0.089	0.26 ±0.26	0.956	1.32
UWYK No-Anc	0.879 ±0.271	0.74 ±0.57	0.924 ±0.182	0.73 ±0.66	0.920 ±0.230	0.61 ±0.46	0.964 ±0.146	0.78 ±0.79	0.962 ±0.155	0.62 ±0.61	0.966 ±0.143	0.66 ±0.57	0.936	3.74
UWYK No-Anc 2D	0.972 ±0.153	0.47 ±0.48	0.977 ±0.133	0.38 ±0.37	0.969 ±0.144	0.31 ±0.35	0.992 ±0.036	0.30 ±0.36	0.981 ±0.127	0.30 ±0.39	0.992 ±0.030	0.30 ±0.37	0.981	1.84
UWYK Anc	0.950 ±0.172	0.80 ±0.52	0.993 ±0.028	0.81 ±0.54	0.959 ±0.162	0.66 ±0.44	0.975 ±0.123	0.81 ±0.81	0.973 ±0.141	0.62 ±0.62	0.981 ±0.095	0.62 ±0.56	0.972	3.89
UWYK Anc 2D	0.967 ±0.155	0.40 ±0.41	0.972 ±0.137	0.25 ±0.24	0.967 ±0.144	0.22 ±0.24	0.978 ±0.101	0.20 ±0.23	0.974 ±0.131	0.21 ±0.30	0.987 ±0.042	0.23 ±0.28	0.974	1.34
CausalPFN-C	0.861 ±0.286	0.54 ±0.57	0.907 ±0.204	0.47 ±0.50	0.931 ±0.169	0.43 ±0.44	0.952 ±0.119	0.54 ±0.65	0.962 ±0.120	0.43 ±0.54	0.953 ±0.128	0.43 ±0.49	0.928	2.57
CausalPFN-C 2D	0.963 ±0.154	0.56 ±0.44	0.985 ±0.071	0.55 ±0.37	0.982 ±0.108	0.53 ±0.39	0.998 ±0.012	0.58 ±0.50	0.990 ±0.090	0.50 ±0.43	0.999 ±0.004	0.50 ±0.40	0.986	2.89

6 Discussion and Conclusion

Although causal foundation models are a step forward, at the current stage they do not resolve causal inference. We discuss what this means for their evaluation and training.

Causal inference carries strict identifiability limits. The Markov equivalence class and the function class of the underlying SCM cannot be recovered from observational data (Spirtes et al., 2001; Peters et al., 2017). Two PFN models of equal capacity can arrive at different members of the same equivalence class, and point estimation alone cannot separate them. Evaluation should therefore prioritise uncertainty estimation, with coverage and density estimation given a higher standing as standard metrics in the field.

Beyond how we evaluate these models, what we evaluate them on should also be scrutinized. Real datasets offer no objectively accepted ground truth for causal inference, so evaluation mostly falls to semi-synthetic and synthetic benchmarks. There, what one believes about the connection of the factual and counterfactual worlds shapes the design of the data generation process. This belief can be better informed by reality. Humans rarely treat the two worlds as independent when making decisions, and in settings such as drug assignment, latent traits affect both outcomes similarly, making their dependence naturally positive (Bodik et al., 2026). Benchmarks may therefore benefit from reflecting this by spanning correlation levels, and bringing the setup closer to reality.

This connection between the two worlds should also be carried to the training side. Jointly learning both factual and counterfactual worlds for every unit on the outcome level captures the dependency and makes the credible intervals more calibrated. We see this reflected in the point estimation and coverage results on both independent and dependent cases, even at a 50 to 60 times lower resolution. Hence, under a limited budget, joint supervision should be prioritised over resolution. As our results suggest, providing the same amount of data is not equivalent: seeing the same number of worlds does not substitute for learning them together.

Taken together, these considerations provide three main recommendations for where the field should move. Evaluation should measure the uncertainty that identifiability leaves behind, benchmarks should carry the connection between the worlds that reality carries, and training should learn the two worlds jointly. As evaluation and benchmarks move in this direction, the gains of joint learning would only become clearer. Although the field has much progress ahead, we conclude that joint supervision is the preferable training setup for causal foundation models.

AI USE STATEMENT

In this work, we used generative AI tools to assist in writing and formatting, editing software code, summarizing existing literature, and targeted information retrieval in papers verified by the authors.

We did not use generative AI tools to generate datasets, develop theoretical models or conceptual frameworks, formulate mathematical claims or critical ingredients for proofs, assist in writing proofs, propose or refine research hypotheses, or design or provide feedback on the research methodology or experiments. The rest of the required disclosure tasks are not applicable to this work. Synthetic data used in training and evaluation were generated using the benchmark mechanisms described in Sections 4, 5 and Appendix G rather than by generative AI tools.

We have reviewed all AI-assisted work. AI-generated code was verified and tested for correctness before any interpretation of results. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We provide the methodological and experimental details needed to reproduce the work throughout the paper and appendices. Section 3 and Appendix B specify the jointly supervised architecture and training objective, while Appendix C provides the proofs of the theoretical results. Appendices D and E describe the CATE- and ATE-density inference procedures, including MALC smoothing and optimal-transport aggregation. Section 4 and Appendix G describe the experimental setup and benchmark construction, including modifications to existing benchmarks, and Appendix F defines the evaluation protocol and metrics. Appendices H and I provide the additional results and ablations, including the matched-resolution, matched-label, correlation, and checkpoint comparisons. Finally, Appendix J reports the computational environment, software versions, hardware, model sizes, training steps, throughput, and compute cost.

REFERENCES

Ahmed M. Alaa, Michael Weisz, and Mihaela van der Schaar. Deep counterfactual networks with propensity-dropout, 2017.

Ryan M. Andrews and Vanessa Didelez. Insights into the cross-world independence assumption of causal mediation analysis. *Epidemiology*, 32(2):209–219, 2021.

Susan Athey and Guido W. Imbens. The state of applied econometrics: Causality and policy evaluation. *Journal of Economic Perspectives*, 31(2):3–32, May 2017. doi: 10.1257/jep.31.2.3.

Vahid Balazadeh, Hamidreza Kamkari, Valentin Thomas, Junwei Ma, Bingru Li, Jesse C. Cresswell, and Rahul Krishnan. CausalPFN: Amortized causal effect estimation via in-context learning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Juraj Bodik, Yaxuan Huang, and Bin Yu. Cross-world assumption and refining prediction intervals for individual treatment effects, 2026.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 02 2018. ISSN 1368-4221.

Furkan Danisman, Hanna Jankowski, and Camila P. E. de Souza. Bandwidth-free nonparametric density estimation for grouped data, 2026.

Peng Ding and Tirthankar Dasgupta. A potential tale of two-by-two tables from completely randomized experiments. *Journal of the American Statistical Association*, 111(513):157–168, 2016.

Yanqin Fan and Sang Soo Park. Sharp bounds on the distribution of treatment effects and their statistical inference. *Econometric Theory*, 26(3):931–951, 2010. ISSN 02664666, 14694360.

P. Richard Hahn, Vincent Dorie, and Jared S. Murray. Atlantic causal inference conference (acic) data analysis challenge 2017, 2019.

James J. Heckman, Jeffrey Smith, and Nancy Clements. Making the most out of programme evaluations and social experiments: Accounting for heterogeneity in programme impacts. *The Review of Economic Studies*, 64(4):487–535, 1997. ISSN 00346527, 1467937X.

Miguel A. Hernán and James M. Robins. *Causal Inference: What If*. Chapman & Hall/CRC, Boca Raton, 2020.

Jennifer L. Hill. Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20(1):217–240, 2011. doi: 10.1198/jcgs.2010.08162.

Paul W. Holland. Statistics and causal inference. *Journal of the American Statistical Association*, 81 (396):945–960, 1986. ISSN 01621459, 1537274X.

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

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

Guido W. Imbens and Donald B. Rubin. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press, 2015.

Edward H. Kennedy. Towards optimal doubly robust estimation of heterogeneous causal effects. *Electronic Journal of Statistics*, 17(2):3008 – 3049, 2023. doi: 10.1214/23-EJS2157.

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

Christos Louizos, Uri Shalit, Joris Mooij, David Sontag, Richard Zemel, and Max Welling. Causal effect inference with deep latent-variable models, 2017.

Yuchen Ma, Dennis Frauen, Emil Javurek, and Stefan Feuerriegel. Foundation models for causal inference via prior-data fitted networks, 2026.

Samuel Müller, Noah Hollmann, Sebastian Pineda-Arango, Josif Grabocka, and Frank Hutter. Transformers can do bayesian inference. In *ICLR*, 2022.

Brady Neal, Chin-Wei Huang, and Sunand Raghupathi. Realcause: Realistic causal inference benchmarking, 2021.

Xinkun Nie and Stefan Wager. Quasi-oracle estimation of heterogeneous treatment effects, 2020.

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

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of Causal Inference: Foundations and Learning Algorithms*. Adaptive Computation and Machine Learning. MIT Press, Cambridge, MA, 2017. ISBN 978-0-262-03731-0.

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

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

Donald B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688–701, 1974. doi: 10.1037/h0037350.

Uri Shalit, Fredrik D. Johansson, and David Sontag. Estimating individual treatment effect: generalization bounds and algorithms, 2017.

Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, Prediction, and Search*. The MIT Press, 01 2001. ISBN 9780262284158.

Christopher Stith, Hossein Rahmani, and Jesse C. Cresswell. Causal foundation models, 2026.

Dennis Thumm, Arik Reuter, Jake Robertson, Shi Bin Hoo, Adrian Weller, Frank Hutter, Ying Chen, and Bernhard Schölkopf. Causal foundation models for time series based on prior-data fitted networks. In *2nd ICML Workshop on Foundation Models for Structured Data*, 2026.

Stijn Vansteelandt and Kelly Van Lancker. Chasing shadows: How implausible assumptions skew our understanding of causal estimands, 2024.

Athanasios Vlontzos, Bernhard Kainz, and Ciaran M. Gilligan-Lee. Estimating categorical counterfactuals via deep twin networks, 2023.

Stefan Wager and Susan Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242, 2018.

Jinsung Yoon, James Jordon, and Mihaela van der Schaar. GANITE: Estimation of individualized treatment effects using generative adversarial nets. In *International Conference on Learning Representations*, 2018.

A Appendix

The appendix is organised as follows. Appendix B details the joint head architecture and its training loss. Appendix C contains the proofs of Lemma 2 and Theorem 1. Appendix D describes how the CATE density is obtained for each model type and the MALC smoothing applied to it, and Appendix E the optimal-transport aggregation behind the ATE density. Appendix F defines the evaluation protocol and metrics, and Appendix G the construction of the three benchmarks. Appendix H reports the point-estimation results on the synthetic benchmarks. Appendix I collects the additional results, including the matched-resolution and matched-label ablations, the correlation ablation, the comparisons against the published CausalPFN and re-trained UWYK checkpoints, and the decomposition of the case studies by dimension. Appendix J lists the computational details.

B Joint Supervision Architecture

The joint learning defined in Section 3.1 must return a finite density at every point of $\mathbb{R}^2$ so that the log-loss in equation 4 is well-posed on unbounded outcomes, while remaining expressive enough on the bounded region $\mathcal{I} = [a, b]^2$ [2] . We meet both requirements with a single mixture that places a $K \times K$ histogram on $\mathcal{I}$ and Gaussian tails outside it. The construction below provide the nine regions of the mixture, the $K^2 + 9 + 4$ raw outputs the transformer emits per query, the link functions that map raw outputs to a valid density, and the approximation the design introduces.

Grid and indexing. Partition $\mathcal{I}$ into a uniform $K \times K$ grid of square bins of width $\delta := 2/K$. Bins are indexed by $(j_0, j_1) \in \{0, \dots, K-1\}^2$; bin B_{j_0,j_1} is the $\delta \times \delta$ square with lower-left corner $\big(a + j_0\delta,\ a + j_1\delta\big)$. For any $y \in [a, b]$ and coordinate $k \in \{0, 1\}$, the containing bin index along that coordinate is

$$j_k(y) \ := \ \lfloor (y + 1)/\delta \rfloor. \tag{11}$$

Per-query outputs. For each query the transformer emits $K^2 + 9 + 4$ real numbers, grouped into three tensors mapped to normalised parameters by fixed link functions:

- *Interior logits* $\ell \in \mathbb{R}^{K^2}$ softmaxed into a $K \times K$ probability table over the bins,

$$p_{j_0,j_1}(X, \mathcal{D}_{\text{obs}}) \ = \ \frac{\exp(\ell_{j_0,j_1})}{\sum_{i_0,i_1} \exp(\ell_{i_0,i_1})}; \tag{12}$$

- *Region logits* $u \in \mathbb{R}^9$ softmaxed into non-negative mixture weights across the nine regions defined below,

$$w_r(X, \mathcal{D}_{\text{obs}}) \ = \ \frac{\exp(u_r)}{\sum_{s=1}^{9} \exp(u_s)}, \qquad r \in \{1, \dots, 9\}; \tag{13}$$

[2] Outcomes are rescaled per task so that the observed range of Y_T maps to $[a, b]$. Rescaling is different for each model by design choice.

- *Raw tail scales* $\big(\sigma_L^{(0),\mathrm{raw}},\ \sigma_R^{(0),\mathrm{raw}},\ \sigma_L^{(1),\mathrm{raw}},\ \sigma_R^{(1),\mathrm{raw}}\big)$, one for each coordinate–direction pair, passed through softplus with a small floor $\varepsilon > 0$ and scaled by δ,

$$\sigma_d^{(k)}(X, \mathcal{D}_{\mathrm{obs}}) \;=\; \delta \cdot \big(\mathrm{softplus}\big(\sigma_d^{(k),\mathrm{raw}}\big) + \varepsilon\big), \qquad k \in \{0,1\},\ d \in \{L,R\}. \tag{14}$$

The δ prefactor in equation 14 makes the tail scales compatible with the interior resolution: doubling K halves both the bin width and the initial tail width, so the density is invariant under resolution changes at initialisation.

Nine regions. The plane $\mathbb{R}^2$ is partitioned into nine regions in three groups: (1) the interior square $\mathcal{I}$, where both $y_0, y_1 \in [a,b]$; (2) four edge strips, where exactly one of y_0, y_1 lies outside $[a,b]$; and (3) four corner quadrants, where both lie outside. Every query point by default falls in one region, so at inference the mixture weight w_r selects a single component; only its density contributes to f and to $\log f$.

Interior region ($y_0, y_1 \in [a,b]$). The histogram is converted to a piecewise-constant density by dividing by the bin area,

$$f_{\mathrm{inner}}(y_0, y_1 \mid X, \mathcal{D}_{\mathrm{obs}}) \;=\; \frac{p_{j_0(y_0),\, j_1(y_1)}}{\delta^2}. \tag{15}$$

The δ^2 factor is the Jacobian from bin probability to density on $[a,b]^2$.

Edge regions ($y_i \in [a,b]$, $y_{1-i} \notin [a,b]$). Consider the representative subcase $y_0 \in [a,b]$ and $y_1 < a$; the other three edges are symmetric. We factor the edge density as $p(y_1)\, p(y_0 \mid y_1)$, model the marginal $p(y_1)$ by a half-Gaussian anchored at the boundary, and approximate the conditional $p(y_0 \mid y_1)$ by the inner region at the boundary:

$$f_{\mathrm{edge}}(y_0, y_1 \mid X, \mathcal{D}_{\mathrm{obs}}) \;=\; p_L^{(1)}(y_1)\, g(y_0 \mid y_1 = a), \tag{16}$$

with

$$p_L^{(1)}(y_1) \;=\; \frac{2}{\sigma_L^{(1)}}\, \phi\!\left(\frac{y_1 + b}{\sigma_L^{(1)}}\right), \qquad g(y_0 \mid y_1 = a) \;=\; \frac{p_{j_0(y_0),\,0}}{\big(\sum_{i_0} p_{i_0,\,0}\big)\,\delta}, \tag{17}$$

where ϕ is the standard normal density.

Holding $p(y_0 \mid y_1)$ fixed at $y_1 = a$ is approximation when $Y_0 \perp Y_1$ and otherwise introduces a bias that grows with the strength of dependence and the distance past the boundary; for a bivariate normal with correlation ρ and tail scale σ the excess NLL per edge observation is approximately $\rho^2\sigma^2/\pi(1-\rho^2)$. Two effects keep this small in practice: (i) the half-Gaussian factor $p_L^{(1)}$ assigns exponentially decreasing weight to y_1 far from the boundary, so the approximation is applied most heavily where it is most accurate; and (ii) the tail scales $\sigma_d^{(k)}$ are trained end-to-end and shrink whenever the interior distribution concentrates away from the boundary, further suppressing the region where the approximation would matter.

Corner regions ($y_0, y_1 \notin [a,b]$). Let $(s_0, s_1) \in \{L,R\}^2$ select the quadrant with corner $(c_0, c_1) \in \{-1,+1\}^2$. The corner density is a bivariate normal centred at the corner, with correlation ρ (defined below) and standardised coordinates $\tilde{y}_k = (y_k - c_k)/\sigma_{s_k}^{(k)}$:

$$f_{\mathrm{corner}}(y_0, y_1 \mid X, \mathcal{D}_{\mathrm{obs}}) \;=\; \frac{1}{Z(s_0, s_1, \rho)}\, \phi_2\big(\tilde{y}_0, \tilde{y}_1;\, \rho\big). \tag{18}$$

The truncation constant Z is the probability mass the bivariate normal places in the quadrant selected by (s_0, s_1); by Sheppard's identity,

$$Z(s_0, s_1, \rho) \;=\; \begin{cases} \frac{1}{4} + \frac{\arcsin\rho}{2\pi} & \text{if } s_0 = s_1, \\ \frac{1}{4} - \frac{\arcsin\rho}{2\pi} & \text{if } s_0 \neq s_1. \end{cases} \tag{19}$$

Correlation inherited from the interior. The correlation ρ in equation 18–equation 19 is not a separate trained parameter by design choice. It is defined as the Pearson correlation of the discrete

distribution $\{p_{j_0,j_1}\}$ evaluated at bin midpoints $m_k(j_k) \coloneqq a + \delta\big(j_k + \frac{1}{2}\big)$:

$$\begin{aligned} \bar{y}_k &= \sum_{j_0,j_1} p_{j_0,j_1}\, m_k(j_k), \\ v_k &= \sum_{j_0,j_1} p_{j_0,j_1}\big(m_k(j_k) - \bar{y}_k\big)^2, \\ c &= \sum_{j_0,j_1} p_{j_0,j_1}\big(m_0(j_0) - \bar{y}_0\big)\big(m_1(j_1) - \bar{y}_1\big), \\ \rho &= \frac{c}{\sqrt{v_0\, v_1} + \varepsilon}. \end{aligned} \tag{20}$$

Tying ρ to the interior grid removes one degree of freedom, keeps the corner coupling consistent with what the interior region already represents, and prevents the tail and interior components from disagreeing about the dependence between $Y_{\mathrm{do}(0)}$ and $Y_{\mathrm{do}(1)}$.

Full mixture and training loss. The joint density is

$$f(y_0, y_1 \mid X, \mathcal{D}_{\mathrm{obs}}) \;=\; w_1\, f_{\mathrm{inner}} \;+\; \sum_{r=2}^{5} w_r\, f^{(r)}_{\mathrm{edge}} \;+\; \sum_{r=6}^{9} w_r\, f^{(r)}_{\mathrm{corner}}, \tag{21}$$

and the loss in equation 4 is evaluated at the region of the observed label $(y_0^\star, y_1^\star)$; the region indicator selects a single component, so only its weight and parameters receive gradient at that point. Because the region assignment is deterministic in the label, the mixture weights w are trained purely from which region observations fall into, while the interior grid, tail scales, and derived correlation are trained within each region.

C Proofs

In this Section we report the proofs of Lemma 2, and Theorem 1. To avoid collision with the interior square $\mathcal{I} = [a, b]^2$ notation of Appendix B, we denote the Fisher information matrix evaluated at the true joint law $\mathbb{P}^\star$ as J in this Section.

C.1 Proof of Lemma 2

Lemma 2. *Let $Z_i^{\mathrm{R}} \coloneqq \big(X_i,\, Y^{\mathrm{do}(0)}(x_i, \varepsilon_i),\, Y^{\mathrm{do}(1)}(x_i, \varepsilon_i)\big)$ denote the per-unit observation under the joint objective and $Z_i^{\mathrm{U}} \coloneqq \big(X_i,\, T_i,\, Y^{\mathrm{do}(T_i)}(x_i, \varepsilon_i)\big)$ the per-unit observation under the single-outcome objective, with T_i drawn from any distribution on $\{0, 1\}$ independent of ε_i. Then Z^{U} is a randomised function of Z^{R}: sampling T from the same distribution and setting $Z^{\mathrm{U}} = \big(X, T, Y^{\mathrm{do}(T)}\big)$ recovers the law of the single-outcome observation from that of the paired one. Consequently, any parameter of the joint law $P^\star\big(Y^{\mathrm{do}(0)}, Y^{\mathrm{do}(1)} \mid X\big)$ estimable from Z^{U} is estimable from Z^{R} at least as well.*

Proof. Let λ denote the law of T on $\{0, 1\}$, chosen independently of ε_i. Define the measurable map $\Phi(x, y_0, y_1;\, t) \coloneqq (x, t, y_t)$; then for any Z^{R} with joint law $\mathbb{P}^\star$,

$$Z^{\mathrm{U}} \;\overset{d}{=}\; \Phi\big(Z^{\mathrm{R}};\, T\big), \qquad T \sim \lambda \perp Z^{\mathrm{R}}, \tag{22}$$

recovers the law of the marginal observation from the joint one. The map is measurable and marginalises over T without reference to $\mathbb{P}^\star$, so Z^{R} is a sufficient statistic for every parameter of $\mathbb{P}^\star$. By the Rao–Blackwell theorem, any $\mathbb{P}^\star$-estimator based on Z^{U} is dominated in mean-squared error by its conditional expectation given Z^{R}; equivalently, any parameter estimable from Z^{U} is estimable from Z^{R} at least as well. □

C.2 Proof of Theorem 1

The two parts share a common Fisher-information calculation.

Proof of part (i). Fix a query point x and a treatment level $t \in \{0, 1\}$, and let ϕ_t parameterise the marginal density $p^\star(y \mid x, t)$. Under the joint objective, each unit contributes an observation

$(X_i, Y_{\mathrm{do}(0)}(x_i, \varepsilon_i), Y_{\mathrm{do}(1)}(x_i, \varepsilon_i))$; the arm-$t$ coordinate is observed at every unit, so the per-unit Fisher information for ϕ_t equals $\mathrm{J}_{Y_{\mathrm{do}(t)}|X}(\phi_t)$.

Under the marginal objective, the arm-t coordinate is observed only when $T_i = t$, which occurs with probability π_t. Since T_i is independent of ε_i, the arm-$(1-t)$ observations contribute zero information about ϕ_t, and the per-unit Fisher information decomposes as

$$\mathrm{J}_{\mathrm{U}}(\phi_t) \;=\; \pi_t \, \mathrm{J}_{Y_{\mathrm{do}(t)}|X}(\phi_t) \;+\; (1-\pi_t)\cdot 0 \;=\; \pi_t \, \mathrm{J}_{\mathrm{R}}(\phi_t). \tag{23}$$

Inverting the Cramér–Rao inequality gives $\mathrm{CRB}_{\mathrm{U}}(\phi_t) = \pi_t^{-1}\mathrm{CRB}_{\mathrm{R}}(\phi_t)$ per unit, and this ratio is preserved after aggregating over N units. Under MLE regularity, both $\hat{p}_{\mathrm{U}}$ and $\hat{p}_{\mathrm{R}}$ are asymptotically efficient, so their $N \cdot \mathrm{MSE}$ limits equal their respective Cramér–Rao bounds; the ratio of these limits is equation 6. □

Proof of part (ii). The CATE estimator $\hat{\tau}(x) = \hat{\mu}_1(x) - \hat{\mu}_0(x)$ is a linear functional of the two mean estimators; under MLE asymptotic normality the CR bound on $\mathrm{Var}[\hat{\tau}]$ equals the sum of the per-arm CR bounds, since the two mean estimators are constructed from disjoint units and are asymptotically independent.

Write $v_t \coloneqq \mathrm{CRB}_{\mathrm{R}}(\hat{\mu}_t)$ for the per-arm CRB on the arm-t mean under the joint objective. Part (i) with the mean as a smooth functional of the marginal density gives $\mathrm{CRB}_{\mathrm{U}}(\hat{\mu}_t) = \pi_t^{-1} v_t$, so

$$\frac{\mathrm{Var}[\hat{\tau}^{\mathrm{U}}]}{\mathrm{Var}[\hat{\tau}^{\mathrm{R}}]} \;\longrightarrow\; \frac{\pi_1^{-1} v_1 + \pi_0^{-1} v_0}{v_1 + v_0}. \tag{24}$$

Under the standing prior over SCMs the arms are exchangeable, so $v_0 = v_1 =: v$; substituting,

$$\frac{\mathrm{Var}[\hat{\tau}^{\mathrm{U}}]}{\mathrm{Var}[\hat{\tau}^{\mathrm{R}}]} \;\longrightarrow\; \frac{\pi_1^{-1} + \pi_0^{-1}}{2} \;=\; \frac{1}{2\pi(1-\pi)}, \tag{25}$$

where the last equality uses $\pi_1^{-1} + \pi_0^{-1} = \pi^{-1} + (1-\pi)^{-1} = 1/\big(\pi(1-\pi)\big)$. This is equation 7, and the special case $\pi = 1/2$ recovers the factor of 2 often quoted informally. □

Remark 3. *Equation 6 is an asymptotic-variance statement, not a finite-sample inequality. Under MLE regularity both $\mathbb{E}[(\hat{p}_{\mathrm{U}} - p^\star)^2]$ and $\mathbb{E}[(\hat{p}_{\mathrm{R}} - p^\star)^2]$ shrink as $\mathcal{O}(1/N)$; multiplying by N produces finite limiting constants $C_{\mathrm{U}}, C_{\mathrm{R}}$ whose ratio equals π_t^{-1}.*

D CATE Density Estimation

This section details how the CATE density f_τ is obtained for the single-outcome and the jointly supervised models, and how MALC is applied to the result.

Single-outcome models. A single-outcome model outputs the two marginal densities $\hat{p}(y \mid x, do(0))$ and $\hat{p}(y \mid x, do(1))$ separately. The CATE density is then obtained by convolving the marginals under the independence assumption,

$$f_\tau(t) = \int \hat{p}(y + t \mid x, do(1)) \, \hat{p}(y \mid x, do(0)) \, dy, \tag{26}$$

computed on the grid as the discrete cross-correlation of the two bin vectors. Independence is the only option available to these models.

Jointly supervised models. The joint head outputs $\hat{p}(y^{do(0)}, y^{do(1)} \mid x, D_{obs})$ directly. The CATE density follows as the pushforward of the joint under the difference map,

$$f_\tau(t) = \sum_{i,j} p_{ij} \, \mathbb{1}\{y_1^j - y_0^i = t\}, \tag{27}$$

which requires no assumption on the dependence: the learned coupling enters through the joint probabilities p_{ij}.

D.1 MALC

Both constructions yield a discrete density on the grid of achievable differences. We write q_j for the probability mass at difference value t_j, so that f_τ is represented by the pairs $\{(t_j, q_j)\}$. MALC is applied to this discrete density, producing the smoothed mixture $\tilde{f}_\tau(t)$, following the base construction of Danisman et al. (2026).

Base class. Each mixture component is a single mean-adjusted log-concave (MALC) density $g \in \mathcal{F}_{\mathrm{LC}}$, where $\mathcal{F}_{\mathrm{LC}}$ is the class of densities on $\mathbb{R}$ with concave log. We take the constructor, sampling proposal, and convergence guarantees of a single MALC fit as given by Danisman et al. (2026).

Mixture family. For a fixed order $M \geq 1$, the mixture family is

$$\mathcal{F}_{\mathrm{MALC}}^{(M)} \;=\; \Big\{ f \;=\; \sum_{m=1}^{M} w_m \, g_m \;:\; w \in \Delta^{M-1},\;\; g_m \in \mathcal{F}_{\mathrm{LC}} \;\; \forall m \Big\}. \tag{28}$$

$\mathcal{F}_{\mathrm{MALC}}^{(M)}$ permits up to M modes while retaining the shape constraint componentwise.

Mixture EM on binned data. For $M > 1$, the components are combined by expectation–maximisation directly on the difference grid. Let $\gamma_{j,m}^{(s)}$ denote the responsibility of component m for bin j at iteration s, and $G_m^{(s)}(j) \coloneqq \int_{B_j} g_m^{(s)}$ the bin-integrated component density over the bin B_j around t_j. The updates are

$$\gamma_{j,m}^{(s+1)} \;=\; \frac{w_m^{(s)} \, G_m^{(s)}(j)}{\sum_l w_l^{(s)} \, G_l^{(s)}(j)}, \tag{29}$$

$$w_m^{(s+1)} \;=\; \sum_j q_j \, \gamma_{j,m}^{(s+1)}, \tag{30}$$

$$g_m^{(s+1)} \;=\; \mathrm{MALC}\Big(q_j \, \gamma_{j,m}^{(s+1)} \Big). \tag{31}$$

Initialisation is a Voronoi assignment around the modal peaks of $\{q_j\}$, and the iterations stop after five consecutive non-improving log-likelihoods.

Order selection. The number of components is chosen by BIC on an effective sample size n_{eff} derived from the grid probabilities themselves:

$$n_{\mathrm{eff}} \;\coloneqq\; \Big(\sum_j q_j^2 \Big)^{-1}, \qquad \mathrm{BIC}(M) \;=\; -2 \, n_{\mathrm{eff}} \sum_j q_j \, \log \hat{f}^{(M)}(t_j) \;+\; (2M-1) \, \log n_{\mathrm{eff}}, \tag{32}$$

where $\hat{f}^{(M)} \in \mathcal{F}_{\mathrm{MALC}}^{(M)}$ is the mixture obtained by iterating equation 29–equation 31 at order M. We select $M^\star = \arg\min_{M \in \{1,\dots,M_{\max}\}} \mathrm{BIC}(M)$ with $M_{\max} = 5$. The parameter count $2M-1$ tallies the $M-1$ mixture weights plus a scalar mean per component; the log-concave shape itself does not contribute a finite parameter count. Each component refit draws $B = 1000$ synthetic samples.

Each per-order fit is $\mathcal{O}(S \cdot G + S \cdot M \cdot B \log B)$, where S is the EM iteration count (typically ≤ 20 in our experiments) and G the number of difference bins; the outer sweep over orders adds a factor $M_{\max}$. Because the fit uses the binned probabilities directly, no resampling of the original context is required and the whole procedure runs on a single CPU per query.

Output and downstream use. The MALC estimator is the smooth density

$$\tilde{f}_\tau(t) \;=\; \sum_{m=1}^{M^\star} \hat{w}_m^\star \, \hat{g}_m^\star(t), \tag{33}$$

evaluable at any $t \in \mathbb{R}$. It interpolates between the grid points of the discrete f_τ and removes the coarse-bin limitation.

Effect on calibration. Tables 5, 6, and 7 report both the raw and the MALC-smoothed versions for the main models. The single-outcome models' resolution is already high, and smoothing degraded their performance. Therefore, in the main results, the 2D models are displayed with the MALC-smoothed density, and the single-outcome models with their raw discrete densities. For consistency, MALC is also applied to the matched-resolution models of Appendix I.1.

E ATE Density via Optimal Transport

Section 3.2.2 defines the ATE density f_{ATE} as the uniformly-weighted 2-Wasserstein barycenter of the per-query CATE densities $\{f_\tau^{(n)}\}_{n=1}^N$ and states its closed-form quantile representation in equation 10. In this segment we report the derivation, the grid-level algorithm, and finally a illustration of the process (Figure 1).

Table 5: 95% posterior credible interval calibration for the CATE on RealCause across five datasets (Coverage, Length), computed as in Table 2. Each model is reported with its raw discrete density and with MALC smoothing applied at inference (+MALC). Coverage closest to the 0.95 is best; shorter length is better at comparable coverage. Average reports the mean coverage and the length normalised by the outcome scale (Len/σ_Y).

	IHDP		ACIC		CPS		PSID		PSID_bal		Average	
Model	Cov	Len	Cov	Len	Cov	Len	Cov	Len	Cov	Len	Cov	Len/σ_Y
Do-PFN	0.988 ±0.018	25.95 ±31.93	0.986 ±0.021	26.39 ±2.72	0.943 ±0.023	48914 ±1177	0.979 ±0.009	86249 ±2989	0.972 ±0.012	86444 ±5467	0.974	5.46
+MALC	0.988 ±0.017	26.30 ±32.12	0.986 ±0.021	26.94 ±2.95	0.952 ±0.028	48126 ±3164	0.982 ±0.008	85579 ±2879	0.975 ±0.011	85241 ±5172	0.977	5.45
Do-PFN 2D	0.861 ±0.163	16.60 ±21.50	0.780 ±0.246	10.52 ±5.00	0.501 ±0.053	18236 ±1275	0.852 ±0.029	42025 ±1969	0.869 ±0.037	44058 ±3426	0.773	2.62
+MALC	0.876 ±0.168	16.60 ±20.99	0.792 ±0.228	10.73 ±4.24	0.412 ±0.022	16756 ±709	0.885 ±0.022	44751 ±1332	0.894 ±0.028	46219 ±2699	0.772	2.66
Do-PFN 2D (indep)	0.991 ±0.015	26.99 ±31.13	0.986 ±0.028	27.08 ±3.69	0.934 ±0.007	48875 ±193	0.968 ±0.012	66872 ±2346	0.966 ±0.012	65647 ±4484	0.969	4.98
+MALC	0.992 ±0.015	26.85 ±30.45	0.989 ±0.021	27.17 ±3.71	0.994 ±0.002	50830 ±186	0.972 ±0.011	66848 ±2134	0.968 ±0.013	64829 ±4220	0.983	5.01
UWYK No-Anc	0.994 ±0.013	25.22 ±26.18	0.993 ±0.013	23.00 ±4.81	0.950 ±0.022	44818 ±943	0.882 ±0.026	59930 ±2474	0.896 ±0.028	59966 ±3175	0.943	4.49
+MALC	0.993 ±0.015	25.49 ±26.66	0.993 ±0.013	22.66 ±5.02	0.882 ±0.019	39736 ±1285	0.876 ±0.027	61222 ±2478	0.894 ±0.031	61220 ±3199	0.928	4.41
UWYK No-Anc 2D	0.997 ±0.010	26.61 ±28.60	0.981 ±0.039	18.30 ±8.17	0.903 ±0.030	41222 ±3956	0.746 ±0.153	41339 ±10756	0.936 ±0.025	58312 ±4532	0.913	4.01
+MALC	0.997 ±0.007	26.55 ±29.74	0.979 ±0.039	17.63 ±7.92	0.807 ±0.079	32345 ±4521	0.673 ±0.151	38456 ±9520	0.886 ±0.047	54241 ±4902	0.868	3.68
UWYK No-Anc 2D (indep)	0.999 ±0.003	32.31 ±40.76	0.993 ±0.014	25.50 ±8.96	0.934 ±0.012	48255 ±613	0.936 ±0.024	78426 ±5849	0.971 ±0.013	85276 ±6773	0.967	5.53
+MALC	0.999 ±0.003	31.59 ±38.94	0.990 ±0.019	24.47 ±7.97	0.904 ±0.013	43378 ±1165	0.920 ±0.029	76185 ±5184	0.966 ±0.015	82357 ±5774	0.956	5.29
UWYK Anc	0.991 ±0.017	23.11 ±24.73	0.994 ±0.014	20.03 ±4.32	0.977 ±0.010	44641 ±1196	0.924 ±0.019	70949 ±4628	0.960 ±0.017	69246 ±5575	0.969	4.58
+MALC	0.988 ±0.021	23.08 ±24.61	0.992 ±0.017	19.75 ±4.40	0.906 ±0.017	40125 ±1244	0.916 ±0.022	70389 ±3982	0.971 ±0.013	69725 ±4988	0.955	4.46
UWYK Anc 2D	0.995 ±0.015	22.99 ±24.62	0.979 ±0.038	16.69 ±7.11	0.865 ±0.068	36547 ±5737	0.634 ±0.171	34669 ±10160	0.786 ±0.083	42734 ±5318	0.852	3.40
+MALC	0.995 ±0.010	22.76 ±25.71	0.977 ±0.038	16.18 ±6.98	0.698 ±0.125	27754 ±5118	0.577 ±0.154	32549 ±8797	0.689 ±0.090	38958 ±5099	0.787	3.09
UWYK Anc 2D (indep)	0.998 ±0.006	26.98 ±33.61	0.991 ±0.019	22.20 ±7.18	0.932 ±0.011	47890 ±745	0.887 ±0.035	67765 ±4961	0.927 ±0.028	72244 ±5351	0.947	4.88
+MALC	0.999 ±0.005	27.07 ±33.38	0.989 ±0.023	21.79 ±6.95	0.900 ±0.011	42811 ±1157	0.877 ±0.036	67531 ±4619	0.912 ±0.031	70555 ±4960	0.935	4.72
CausalPFN-C	0.997 ±0.009	4.70 ±1.27	0.997 ±0.005	8.76 ±2.59	0.976 ±0.028	47788 ±2631	0.949 ±0.017	76381 ±12426	0.934 ±0.020	71295 ±12172	0.971	3.63
+MALC	0.997 ±0.009	4.72 ±1.29	0.996 ±0.007	8.64 ±2.40	0.953 ±0.058	46439 ±5303	0.976 ±0.011	74450 ±12383	0.967 ±0.013	68583 ±11366	0.978	3.53
CausalPFN-C 2D	0.998 ±0.006	12.91 ±14.15	0.997 ±0.005	11.77 ±1.52	0.944 ±0.045	42618 ±1882	0.956 ±0.014	57606 ±10700	0.943 ±0.019	58404 ±10153	0.968	3.50
+MALC	0.999 ±0.004	13.81 ±15.28	0.998 ±0.003	12.55 ±1.52	0.952 ±0.052	41861 ±1944	0.970 ±0.012	59139 ±8881	0.960 ±0.013	59309 ±9152	0.976	3.58
CausalPFN-C 2D (indep)	0.998 ±0.005	13.04 ±13.98	0.998 ±0.005	13.88 ±1.76	0.959 ±0.034	44608 ±2033	0.975 ±0.010	68575 ±13197	0.963 ±0.014	65860 ±11479	0.979	3.89
+MALC	0.999 ±0.003	13.90 ±15.09	0.999 ±0.003	14.62 ±1.67	0.961 ±0.045	43677 ±1696	0.980 ±0.009	71684 ±10407	0.970 ±0.011	66849 ±10608	0.982	3.98

Table 6: 95% posterior credible interval calibration for the CATE on the six causal scenarios of the case studies (Coverage, Length), computed as in Table 2. Each model is reported with its raw discrete density and with MALC smoothing applied at inference (+MALC). Results are aggregated across varying treatment effect magnitudes and covariate dimensions. Coverage closest to the nominal 0.95 is best; shorter length is better at comparable coverage. Average reports the mean coverage and the length normalised by the outcome scale (Len/σ_Y).

	Obs. Conf.		Obs. Med.		Obs. M+C		Unobs. Conf.		Frontdoor		Backdoor		Average	
Model	Cov	Len	Cov	Len	Cov	Len	Cov	Len	Cov	Len	Cov	Len	Cov	Len/σ_Y
Do-PFN	0.996 ±0.026	5.70 ±4.25	0.662 ±0.381	0.72 ±0.55	0.921 ±0.176	4.05 ±7.75	0.995 ±0.036	6.18 ±4.28	1.000 ±0.000	4.35 ±3.97	1.000 ±0.003	4.54 ±5.39	0.929	3.01
+MALC	0.991 ±0.044	6.09 ±4.76	0.577 ±0.400	0.65 ±0.50	0.912 ±0.181	4.21 ±5.88	0.994 ±0.036	6.60 ±4.85	0.986 ±0.082	4.34 ±3.93	0.993 ±0.050	4.49 ±4.64	0.909	3.08
Do-PFN 2D	0.910 ±0.218	2.46 ±2.31	0.344 ±0.447	0.81 ±1.35	0.897 ±0.222	2.38 ±1.86	0.905 ±0.250	2.68 ±2.41	0.855 ±0.290	1.36 ±1.52	0.872 ±0.281	1.41 ±1.55	0.797	1.24
+MALC	0.983 ±0.109	2.62 ±2.23	0.953 ±0.160	0.90 ±1.08	0.929 ±0.198	2.34 ±1.63	0.983 ±0.103	2.84 ±2.36	0.995 ±0.058	1.49 ±1.53	0.990 ±0.087	1.54 ±1.56	0.972	1.32
UWYK No-Anc	0.960 ±0.102	5.37 ±4.71	0.225 ±0.380	0.69 ±0.51	0.483 ±0.457	2.94 ±2.53	0.987 ±0.046	6.11 ±5.17	1.000 ±0.007	4.23 ±5.34	0.999 ±0.011	4.24 ±5.27	0.776	2.79
+MALC	0.953 ±0.113	5.20 ±4.40	0.197 ±0.389	0.59 ±0.42	0.476 ±0.457	2.83 ±2.36	0.982 ±0.055	5.91 ±4.86	1.000 ±0.006	4.00 ±4.95	0.999 ±0.013	4.03 ±4.90	0.768	2.66
UWYK No-Anc 2D	0.999 ±0.008	3.64 ±2.48	0.878 ±0.326	1.55 ±0.85	0.998 ±0.025	3.85 ±2.05	0.998 ±0.023	4.11 ±2.66	0.997 ±0.031	2.50 ±1.74	0.997 ±0.020	2.66 ±1.85	0.978	2.10
+MALC	0.999 ±0.012	3.19 ±2.21	0.916 ±0.274	0.98 ±0.59	0.997 ±0.028	3.54 ±1.98	0.998 ±0.021	3.59 ±2.41	0.998 ±0.015	2.06 ±1.47	0.997 ±0.022	2.20 ±1.60	0.984	1.77
UWYK Anc	0.998 ±0.019	4.25 ±4.35	0.492 ±0.351	1.06 ±0.72	0.640 ±0.391	3.19 ±2.34	1.000 ±0.006	5.28 ±5.25	1.000 ±0.011	3.80 ±4.94	1.000 ±0.003	4.13 ±5.18	0.855	2.60
+MALC	0.998 ±0.019	4.15 ±4.16	0.288 ±0.400	0.89 ±0.64	0.613 ±0.401	3.07 ±2.23	0.999 ±0.007	5.09 ±4.92	1.000 ±0.009	3.64 ±4.65	1.000 ±0.004	3.98 ±4.91	0.816	2.49
UWYK Anc 2D	0.994 ±0.047	2.68 ±2.28	0.878 ±0.326	1.53 ±0.86	0.997 ±0.028	3.53 ±1.75	0.998 ±0.020	4.08 ±2.68	0.998 ±0.023	2.97 ±1.97	0.995 ±0.037	1.88 ±1.79	0.977	1.93
+MALC	0.995 ±0.042	2.43 ±2.05	0.921 ±0.266	0.84 ±0.47	0.997 ±0.029	3.28 ±1.70	0.998 ±0.020	3.57 ±2.43	0.998 ±0.013	2.53 ±1.69	0.997 ±0.020	1.52 ±1.44	0.984	1.63
CausalPFN-C	0.988 ±0.032	1.49 ±2.90	0.808 ±0.245	1.86 ±1.16	0.920 ±0.195	3.34 ±3.70	0.987 ±0.053	2.75 ±3.38	0.998 ±0.009	1.24 ±2.15	0.999 ±0.007	1.24 ±2.12	0.950	1.33
+MALC	0.990 ±0.029	1.43 ±2.67	0.913 ±0.160	1.67 ±1.01	0.920 ±0.195	3.08 ±3.26	0.988 ±0.050	2.63 ±3.11	0.998 ±0.009	1.22 ±2.06	0.999 ±0.007	1.22 ±2.03	0.968	1.26
CausalPFN-C 2D	0.999 ±0.010	2.66 ±1.57	0.862 ±0.282	1.78 ±1.06	0.965 ±0.101	2.40 ±1.34	0.996 ±0.020	3.00 ±1.99	0.999 ±0.009	2.14 ±1.80	0.999 ±0.006	2.22 ±1.76	0.970	1.67
+MALC	0.999 ±0.007	2.85 ±1.65	0.920 ±0.200	1.86 ±1.11	0.973 ±0.084	2.54 ±1.38	0.997 ±0.013	3.18 ±2.06	0.999 ±0.007	2.27 ±1.88	0.999 ±0.004	2.35 ±1.83	0.981	1.77

Notation. Let F_n and $Q_n := F_n^{-1}$ be the CDF and quantile function of the per-query CATE density $f_\tau^{(n)}$, and let $\alpha \in \Delta^{N-1}$ be nonnegative weights summing to 1. On $\mathbb{R}$ the weighted 2-Wasserstein barycenter of $\{f_\tau^{(n)}\}_{n=1}^N$ admits the closed-form quantile representation

$$Q^\star(t) \;=\; \sum_{n=1}^{N} \alpha_n \, Q_n(t), \qquad t \in (0,1), \tag{34}$$

of which equation 10 is the uniform-weight case $\alpha_n = 1/N$.

Grid-level implementation. Because $Q^\star$ is defined explicitly by equation 34, no iterative optimisation is needed. Given the per-query densities sampled on a common grid $x_1 < \cdots < x_M$ with spacing Δx, the barycenter density is produced in three quantile passes and one density-recovery pass:

1. Discretise the probability axis by $\tau_j := (j - \frac{1}{2})/M_\tau$ for $j = 1, \ldots, M_\tau$ (we use $M_\tau = 4001$).

Table 7: 95% posterior credible interval calibration for the CATE on ComplexMech, reported per graph size (Coverage, Length), computed as in Table 2. Each model is reported with its raw discrete density and with MALC smoothing applied at inference (+MALC). Coverage closest to the 0.95 is best; shorter length is better at comparable coverage. Average reports the mean coverage and the length normalised by the outcome scale (Len/σ_Y).

	n=5		n=10		n=20		n=30		n=40		n=50		**Average**	
Model	Cov	Len	Cov	Len	Cov	Len	Cov	Len	Cov	Len	Cov	Len	Cov	Len/σ_Y
Do-PFN	0.955 ±0.194	6.71 ±19.50	0.961 ±0.132	7.15 ±21.79	0.979 ±0.108	2.05 ±3.79	0.990 ±0.088	2.51 ±5.96	0.999 ±0.010	3.42 ±13.49	0.985 ±0.074	3.19 ±10.67	0.978	22.06
+MALC	0.763 ±0.377	5.82 ±17.10	0.824 ±0.312	6.12 ±18.86	0.837 ±0.316	2.05 ±3.16	0.820 ±0.325	2.69 ±5.77	0.832 ±0.319	3.01 ±9.15	0.819 ±0.312	3.05 ±8.90	0.816	20.12
Do-PFN 2D	0.690 ±0.392	0.20 ±0.25	0.750 ±0.349	0.18 ±0.17	0.768 ±0.318	0.21 ±0.23	0.789 ±0.338	0.24 ±0.28	0.803 ±0.283	0.24 ±0.26	0.842 ±0.286	0.26 ±0.30	0.774	1.21
+MALC	0.899 ±0.247	0.26 ±0.27	0.949 ±0.189	0.23 ±0.19	0.962 ±0.146	0.24 ±0.22	0.971 ±0.097	0.25 ±0.25	0.974 ±0.130	0.24 ±0.23	0.983 ±0.089	0.26 ±0.26	0.956	1.32
UWYK No-Anc	0.879 ±0.271	0.74 ±0.57	0.924 ±0.182	0.73 ±0.66	0.920 ±0.230	0.61 ±0.46	0.964 ±0.146	0.78 ±0.79	0.962 ±0.155	0.62 ±0.61	0.966 ±0.143	0.66 ±0.57	0.936	3.74
+MALC	0.868 ±0.285	0.64 ±0.48	0.929 ±0.178	0.65 ±0.55	0.933 ±0.205	0.56 ±0.42	0.974 ±0.122	0.70 ±0.69	0.973 ±0.124	0.58 ±0.57	0.969 ±0.143	0.62 ±0.52	0.941	3.38
UWYK No-Anc 2D	0.820 ±0.362	0.54 ±0.62	0.847 ±0.328	0.45 ±0.53	0.811 ±0.335	0.33 ±0.45	0.822 ±0.339	0.34 ±0.45	0.784 ±0.371	0.34 ±0.49	0.764 ±0.400	0.32 ±0.46	0.808	2.07
+MALC	0.972 ±0.153	0.47 ±0.48	0.977 ±0.133	0.38 ±0.37	0.969 ±0.144	0.31 ±0.35	0.992 ±0.036	0.30 ±0.36	0.981 ±0.127	0.30 ±0.39	0.992 ±0.030	0.30 ±0.37	0.981	1.84
UWYK Anc	0.950 ±0.172	0.80 ±0.52	0.993 ±0.028	0.81 ±0.54	0.959 ±0.162	0.66 ±0.44	0.975 ±0.123	0.81 ±0.81	0.973 ±0.141	0.62 ±0.62	0.981 ±0.095	0.62 ±0.56	0.972	3.89
+MALC	0.944 ±0.186	0.68 ±0.42	0.984 ±0.087	0.72 ±0.44	0.960 ±0.158	0.60 ±0.40	0.983 ±0.103	0.73 ±0.71	0.979 ±0.119	0.58 ±0.59	0.987 ±0.070	0.59 ±0.53	0.973	3.52
UWYK Anc 2D	0.800 ±0.363	0.44 ±0.53	0.775 ±0.375	0.28 ±0.37	0.738 ±0.382	0.23 ±0.33	0.774 ±0.370	0.21 ±0.31	0.698 ±0.415	0.21 ±0.36	0.734 ±0.410	0.24 ±0.36	0.753	1.43
+MALC	0.967 ±0.155	0.40 ±0.41	0.972 ±0.137	0.25 ±0.24	0.967 ±0.144	0.22 ±0.24	0.978 ±0.101	0.20 ±0.23	0.974 ±0.131	0.21 ±0.30	0.987 ±0.042	0.23 ±0.28	0.974	1.34
CausalPFN-C 2D	0.960 ±0.161	0.55 ±0.43	0.980 ±0.101	0.54 ±0.37	0.980 ±0.110	0.52 ±0.39	0.997 ±0.014	0.58 ±0.51	0.989 ±0.091	0.49 ±0.42	0.999 ±0.004	0.49 ±0.40	0.984	2.86
+MALC	0.963 ±0.154	0.56 ±0.44	0.985 ±0.071	0.55 ±0.37	0.982 ±0.108	0.53 ±0.39	0.998 ±0.012	0.58 ±0.50	0.990 ±0.090	0.50 ±0.43	0.999 ±0.004	0.50 ±0.40	0.986	2.89

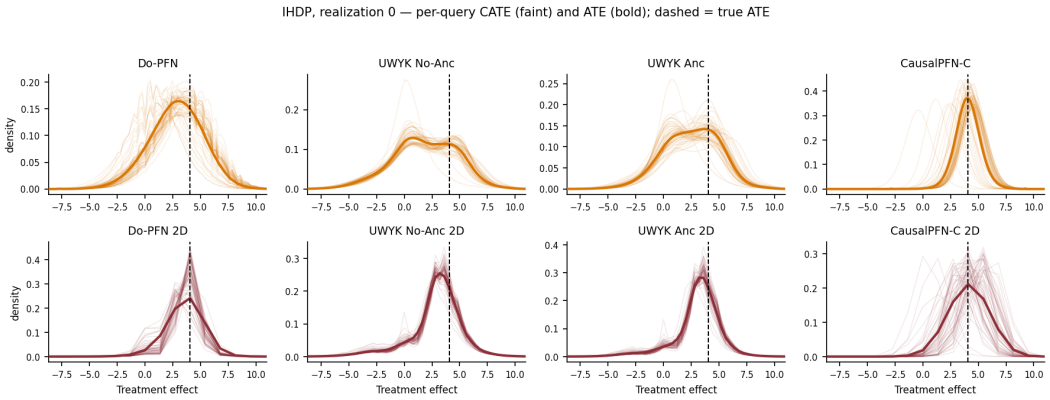

Figure 1: ATE density via optimal transport on one IHDP realization. Faint curves are per-query CATE densities, the bold curve their 2-Wasserstein barycenter (the ATE density), and the dashed line the true ATE. Top row shows the single-outcome models, bottom row their jointly supervised (2D) variants.

2. For each query $n \in \{1, \dots, N\}$, compute the CDF by trapezoidal integration, renormalise, and invert to obtain the quantile function at every τ_j:

$$F_n(x_k) \leftarrow \sum_{\ell \le k} \tfrac{1}{2}\big(\hat{p}_n(x_{\ell-1}) + \hat{p}_n(x_\ell)\big)\,\Delta x, \quad F_n \leftarrow F_n / F_n(x_M), \quad Q_n(\tau_j) \leftarrow F_n^{-1}(\tau_j).$$

3. Average the quantile functions pointwise per equation 34: $Q^\star(\tau_j) = \sum_n \alpha_n\, Q_n(\tau_j)$.

4. Recover the barycenter density on the original grid: invert $Q^\star$ back to a CDF $F^\star$, take a central-difference derivative, and renormalise to unit integral,

$$\hat{f}_{\mathrm{ATE}}(x_k) \leftarrow \frac{F^\star(x_{k+1}) - F^\star(x_{k-1})}{2\,\Delta x}, \qquad \hat{f}_{\mathrm{ATE}} \leftarrow \hat{f}_{\mathrm{ATE}} \Big/ \Big(\Delta x \sum_k \hat{f}_{\mathrm{ATE}}(x_k)\Big).$$

Every step is $\mathcal{O}(N \cdot M)$ vector arithmetic; producing $\hat{f}_{\mathrm{ATE}}$ is therefore much cheaper than the per-query MALC fits it aggregates.

Illustration. Figure 1 shows the aggregation on one IHDP realization. The faint curves are the per-query CATE densities, and the bold curve is their 2-Wasserstein barycenter, the ATE density. The barycenter averages the locations of the per-query densities while preserving their shape. The jointly supervised models concentrate the per-query densities, and their ATE density tightens around the true value.

Table 8: Composition of the six case studies as we vary the observed covariate count d. Hidden nodes scale as $h = \max(1, \lfloor d/10 \rfloor)$ and are not considered covariates, so $|X| = d$ in every cell. A node counts as a confounder (C) if it is a parent of both T and Y, and as a mediator (M) if it lies on a $T \to \cdot \to Y$ path; subscripts denote observed, unobserved, and hidden nodes.

	Observed Confounder	Backdoor Criterion		Observed Mediator	Observed Mediator + Confounder		Unobserved Confounder		Frontdoor Criterion	
d	C_{obs}	C_{obs}	M_{uonbs}	M_{obs}	C_{obs}	M_{obs}	C_{obs}	C_{unobs}	C_{obs}	M_{uonbs}
5	5	5	1	5	3	2	5	1	5	1
10	10	10	1	10	5	5	10	1	10	1
20	20	20	2	20	10	10	20	2	20	2
30	30	30	3	30	15	15	30	3	30	3
40	40	40	4	40	20	20	40	4	40	4
50	50	50	5	50	25	25	50	5	50	5

F Evaluation Metrics

Evaluation protocol. For each benchmark setting we draw B SCMs $\psi_1, \ldots, \psi_B$ independently. From each ψ_b we generate a single observational set $\mathcal{D}_{\text{obs}}^{(b)}$, independently across SCMs, and Q query covariates $x_{b1}, \ldots, x_{bQ}$ with known ground-truth effects $\tau_b(x_{bq})$ and ATE_b. Each model is conditioned on $\mathcal{D}_{\text{obs}}^{(b)}$ and returns, for every query, a CATE estimate $\hat{\tau}_b(x_{bq})$ and a CATE density, and, per SCM, an ATE estimate $\widehat{\text{ATE}}_b$.

Point estimation. Per SCM we compute

$$\sqrt{\text{PEHE}}_b = \left(\frac{1}{Q} \sum_{q=1}^{Q} \big(\hat{\tau}_b(x_{bq}) - \tau_b(x_{bq}) \big)^2 \right)^{1/2}, \qquad \epsilon_{\text{ATE},b} = \frac{\big| \widehat{\text{ATE}}_b - \text{ATE}_b \big|}{\big| \text{ATE}_b \big|}, \tag{35}$$

and report their means over the B SCMs.

Empirical credible interval coverage. For each query, the 95% credible interval $\text{CrI}_b(x_{bq})$ is the central interval between the 2.5% and 97.5% quantiles of the predicted CATE density. Within each SCM we record the proportion of queries whose true CATE lies in its interval,

$$A_b = \frac{1}{Q} \sum_{q=1}^{Q} \mathbf{1}\big\{ \tau_b(x_{bq}) \in \text{CrI}_b(x_{bq}) \big\}, \tag{36}$$

yielding one coverage value per observational set. The reported coverage is the mean across SCMs, $\bar{A} = \frac{1}{B} \sum_{b=1}^{B} A_b$. Since the SCMs and their observational sets are drawn independently, $A_1, \ldots, A_B$ are i.i.d., so $\bar{A}$ is an unbiased estimate of the expected average coverage over the benchmark's SCM distribution, and its standard error is $s_A/\sqrt{B}$, with s_A the sample standard deviation of the A_b. Queries within an SCM share one posterior and are therefore dependent; aggregating them into A_b before computing uncertainty accounts for this.

Interval length. Lengths are averaged in the same way, per SCM and then across SCMs. For averages across datasets with different outcome scales we report the length normalised by the outcome standard deviation, Len/σ_Y.

Relative improvement. For a metric m where lower is better, with values m_k^{base} and m_k^{2D} on the K datasets or settings of a table,

$$\Delta = 100 \Big(1 - \Big(\prod_{k=1}^{K} m_k^{\text{2D}} / m_k^{\text{base}} \Big)^{1/K} \Big) \%, \tag{37}$$

so $\Delta > 0$ favours the 2D model.

G Benchmark Construction

RealCause. We use the RealCause datasets (Neal et al., 2021) directly without any modification. Its generative models fit the outcome distribution under each treatment separately and sample the factual and counterfactual outcomes as independent draws. Table 9 verifies this empirically.

Table 9: Estimated cross-world correlation per benchmark. $\hat{\rho}$ is the inverse-variance combination of the per-realization estimates, with its 95% CI and the p-value of the test $\rho = 0$; $\bar{\rho} \pm \mathrm{sd}$ is the unweighted mean and spread of those same per-realization estimates.

Dataset	$\hat{\rho}$	95% CI	$p\ (\rho = 0)$	$\bar{\rho} \pm \mathrm{sd}$
IHDP	$+0.0023$	$[-0.0052, +0.0099]$	0.541	$+0.0023 \pm 0.037$
ACIC	$+0.003$	$[-0.0066, +0.0113]$	0.603	$+0.003 \pm 0.013$
CPS	-0.001	$[-0.0023, +0.0007]$	0.313	-0.001 ± 0.009
PSID	$+0.003$	$[-0.0012, +0.0064]$	0.183	$+0.003 \pm 0.020$
ComplexMech	$+0.9991$	$[+0.9990, +0.9991]$	$< 10^{-300}$	$+0.9967 \pm 0.0028$
ComplexMech Ablation	$+0.8138$	$[+0.8049, +0.8223]$	$< 10^{-300}$	$+0.4529 \pm 0.5247$
Case study	$+1.0000$	$[+1.0000, +1.0000]$	$< 10^{-300}$	$+1.0000 \pm 0.0000$

Table 10: Unit- and population-level point estimation on the six causal scenarios of the Do-PFN case studies ($\sqrt{\mathrm{PEHE}}$, ϵ_{ATE}; lower is better; mean $\pm$ standard error of the mean over realizations). Results are aggregated across varying treatment effect magnitudes and covariate dimensions. Bold marks the better model within each baseline. Δ summarises the relative improvement across scenarios by geometric mean; positive favours 2D.

	Observed Confounder		**Observed Mediator**		**Observed M+C**		**Unobserved Confounder**		**Frontdoor**		**Backdoor**		**Δ (%)**	
Model	$\sqrt{\mathrm{PEHE}}\downarrow$	$\epsilon_{\mathrm{ATE}}\downarrow$	$\sqrt{\mathrm{PEHE}}\downarrow$	$\epsilon_{\mathrm{ATE}}\downarrow$	$\sqrt{\mathrm{PEHE}}\downarrow$	$\epsilon_{\mathrm{ATE}}\downarrow$	$\sqrt{\mathrm{PEHE}}\downarrow$	$\epsilon_{\mathrm{ATE}}\downarrow$	$\sqrt{\mathrm{PEHE}}\downarrow$	$\epsilon_{\mathrm{ATE}}\downarrow$	$\sqrt{\mathrm{PEHE}}\downarrow$	$\epsilon_{\mathrm{ATE}}\downarrow$	$\sqrt{\mathrm{PEHE}}$	ϵ_{ATE}
Do-PFN	0.811 ±0.017	0.565 ±0.015	0.796 ±0.019	0.465 ±0.009	0.874 ±0.018	0.587 ±0.012	0.853 ±0.016	0.603 ±0.015	**0.101** ±0.003	**0.497** ±0.018	**0.120** ±0.005	**0.660** ±0.028	**+34.7**	**+18.8**
Do-PFN 2D	**0.416** ±0.010	**0.401** ±0.017	**0.324** ±0.014	**0.212** ±0.007	**0.585** ±0.014	**0.486** ±0.014	**0.452** ±0.010	**0.459** ±0.017	0.106 ±0.003	0.613 ±0.023	**0.120** ±0.004	0.754 ±0.033	–	–
UWYK No-Anc	0.937 ±0.016	0.619 ±0.006	1.363 ±0.021	0.873 ±0.006	1.381 ±0.020	0.927 ±0.006	0.912 ±0.016	0.597 ±0.007	**0.081** ±0.002	**0.357** ±0.007	**0.081** ±0.002	**0.410** ±0.008	**+51.3**	**+22.8**
UWYK No-Anc 2D	**0.258** ±0.004	**0.395** ±0.020	**0.165** ±0.004	**0.110** ±0.002	**0.330** ±0.005	**0.418** ±0.020	**0.275** ±0.005	**0.413** ±0.021	0.181 ±0.004	1.024 ±0.032	0.200 ±0.005	1.207 ±0.040	–	–
UWYK Anc	0.304 ±0.005	0.360 ±0.017	1.262 ±0.019	0.800 ±0.006	1.237 ±0.018	0.830 ±0.007	0.324 ±0.007	**0.308** ±0.015	**0.095** ±0.003	**0.386** ±0.009	0.178 ±0.005	1.049 ±0.034	**+42.7**	**+29.3**
UWYK Anc 2D	**0.264** ±0.005	**0.300** ±0.011	**0.133** ±0.002	**0.086** ±0.001	**0.345** ±0.005	**0.422** ±0.019	**0.269** ±0.004	0.425 ±0.022	0.197 ±0.004	1.062 ±0.033	**0.143** ±0.004	**0.760** ±0.023	–	–
CausalPFN-C	**0.181** ±0.006	**0.119** ±0.009	0.568 ±0.010	0.381 ±0.005	0.610 ±0.012	0.589 ±0.026	**0.346** ±0.010	**0.434** ±0.024	**0.149** ±0.006	**0.338** ±0.016	**0.150** ±0.006	**0.328** ±0.016	−8.4	−56.0
CausalPFN-C 2D	0.208 ±0.007	0.294 ±0.018	**0.465** ±0.013	**0.324** ±0.007	**0.475** ±0.013	**0.513** ±0.018	0.394 ±0.012	0.536 ±0.028	0.202 ±0.007	0.788 ±0.026	0.214 ±0.007	0.895 ±0.029	–	–

Case study. We take the six causal scenarios of Do-PFN (Robertson et al., 2025), keeping both the causal graphs and the structural equations. Two types of variations are added. First, a constant is added to the treated outcome equation, shifting the mean CATE to -2, 0, and $+2$; a zero-mean benchmark rewards a model that predicts no effect. Hence, this variation remove that favorism. Reported numbers pool the three mean base as one dataset. Second, the covariate dimension is varied over $d \in \{5, 10, 20, 30, 40, 50\}$ by adding covariates; Table 8 reports the resulting composition per scenario. Both potential outcomes of a unit are generated by propagating the same exogenous draw under $\mathrm{do}(T{=}0)$ and $\mathrm{do}(T{=}1)$, so $\tau(x)$ is noiseless given x and $\rho(x) = 1$ by construction, as Table 9 confirms.

ComplexMech. ComplexMech is adopted from UWYK (Reuter et al., 2026) and samples a fresh SCM with nonlinear mechanisms for every realization. We modify it by binarizing the treatment effect. A uniformly selected treatment node is binarised by thresholding at the median of its own sample, which keeps the two treatment groups balanced; the threshold is frozen and reused for the interventional passes, and the observational context is re-drawn after binarisation. Realizations without a directed path $T \to Y$ are discarded and similar to case study benchmark the two interventional passes share the exogenous draw.

The generator mixes two noise types. In the additive case, $Y = \sigma(f(\mathrm{pa})) + \varepsilon$, the shared noise cancels when the two outcomes are subtracted, so $\rho(x) = 1$. In the non-additive case, the noise enters before the activation, and does not cancel, consequently, $\rho(x) < 1$. Each realization therefore has its own correlation level. This lets us group realizations by their correlation, which is how the correlation ablation of Appendix I.2 is built. The main results report the realizations with $\rho > 0.99$; the ablation study lowers this threshold and spans the remaining levels.

H Point Estimation on the Synthetic Benchmarks

This section reports the point-estimation results summarised in Section 5.2. Table 10 covers the six causal scenarios of the case studies and Table 11 ComplexMech under near-perfect correlation.

For Do-PFN and UWYK, joint supervision improves point estimation throughout, as anticipated based on the explanatin in Section 3.1. For CausalPFN, learning the expectation directly yields better point estimates overall; the gain from joint supervision lies in capturing the dependence, and consequently in the coverage results.

Table 11: Unit- and population-level point estimation on ComplexMech ($\sqrt{\text{PEHE}}$, ϵ_{ATE}; lower is better; mean $\pm$ standard error of the mean over realizations), reported per graph size. The treatment is binarized from the original continuous benchmark (Appendix G). Bold marks the better model within each baseline. Δ summarises the relative improvement across graph sizes by geometric mean; positive favours 2D.

	n=5		n=10		n=20		n=30		n=40		n=50		Δ (%)	
Model	$\sqrt{\text{PEHE}}\downarrow$	$\epsilon_{\text{ATE}}\downarrow$	$\sqrt{\text{PEHE}}\downarrow$	$\epsilon_{\text{ATE}}\downarrow$	$\sqrt{\text{PEHE}}\downarrow$	$\epsilon_{\text{ATE}}\downarrow$	$\sqrt{\text{PEHE}}\downarrow$	$\epsilon_{\text{ATE}}\downarrow$	$\sqrt{\text{PEHE}}\downarrow$	$\epsilon_{\text{ATE}}\downarrow$	$\sqrt{\text{PEHE}}\downarrow$	$\epsilon_{\text{ATE}}\downarrow$	$\sqrt{\text{PEHE}}$	ϵ_{ATE}
Do-PFN	1.484 ±0.380	4.363 ±1.141	1.250 ±0.284	3.867 ±1.069	0.608 ±0.135	1.527 ±0.260	0.467 ±0.085	1.259 ±0.218	0.378 ±0.086	1.319 ±0.381	0.389 ±0.086	1.196 ±0.232	**+93.2**	**+90.7**
Do-PFN 2D	**0.091** ±0.022	**0.239** ±0.038	**0.059** ±0.012	**0.217** ±0.038	**0.045** ±0.008	**0.189** ±0.029	**0.035** ±0.005	**0.168** ±0.027	**0.028** ±0.003	**0.132** ±0.021	**0.032** ±0.004	**0.154** ±0.023	–	–
UWYK No-Anc	0.204 ±0.040	0.426 ±0.036	0.107 ±0.025	0.307 ±0.032	0.071 ±0.020	0.208 ±0.029	0.035 ±0.005	0.157 ±0.024	0.028 ±0.004	0.129 ±0.024	0.040 ±0.010	0.143 ±0.024	**+34.6**	**+29.3**
UWYK No-Anc 2D	**0.083** ±0.022	**0.241** ±0.061	**0.056** ±0.009	**0.215** ±0.045	**0.051** ±0.010	**0.172** ±0.029	**0.029** ±0.004	**0.100** ±0.014	**0.025** ±0.003	**0.095** ±0.018	**0.028** ±0.004	**0.117** ±0.023	–	–
UWYK Anc	0.128 ±0.027	0.334 ±0.059	0.076 ±0.012	0.234 ±0.027	0.059 ±0.014	0.167 ±0.022	0.036 ±0.005	0.149 ±0.019	0.028 ±0.003	0.118 ±0.020	0.031 ±0.004	0.131 ±0.020	**+31.2**	**+39.1**
UWYK Anc 2D	**0.078** ±0.021	**0.215** ±0.046	**0.045** ±0.008	**0.142** ±0.024	**0.041** ±0.010	**0.104** ±0.013	**0.026** ±0.004	**0.083** ±0.012	**0.021** ±0.003	**0.064** ±0.011	**0.024** ±0.003	**0.092** ±0.017	–	–
CausalPFN-C	0.120 ±0.025	0.343 ±0.064	0.076 ±0.010	0.267 ±0.037	**0.058** ±0.010	**0.150** ±0.017	**0.050** ±0.006	**0.154** ±0.020	**0.040** ±0.004	**0.146** ±0.023	**0.042** ±0.004	**0.130** ±0.020	−15.8	−13.4
CausalPFN-C 2D	**0.102** ±0.022	**0.307** ±0.056	**0.075** ±0.011	**0.247** ±0.030	0.068 ±0.010	0.180 ±0.025	0.067 ±0.007	0.188 ±0.023	0.054 ±0.006	0.184 ±0.026	0.058 ±0.006	0.179 ±0.027	–	–

Table 12: 95% posterior credible interval calibration for the CATE on RealCause across five datasets (Coverage, Length), computed as in Table 2. 2D (Indep) disables the learned dependence at inference. The matched-label and matched-resolution variants are denoted in brackets.

	IHDP		**ACIC**		**CPS**		**PSID**		**PSID_bal**		**Average**	
Model	Coverage	Length	Coverage	Length	Coverage	Length	Coverage	Length	Coverage	Length	Coverage	Len/σ_Y
Do-PFN	0.988 ±0.018	25.95 ±31.93	0.986 ±0.021	26.39 ±2.72	0.943 ±0.023	48914 ±1177	0.979 ±0.009	86249 ±2989	0.972 ±0.012	86444 ±5467	0.974	5.46
Do-PFN (matched label)	0.989 ±0.002	22.41 ±2.64	0.990 ±0.006	27.31 ±0.94	0.991 ±0.000	49371 ±60	0.967 ±0.001	68057 ±169	0.965 ±0.001	67694 ±414	0.980	4.87
Do-PFN (matched resolution)	0.985 ±0.028	29.75 ±20.48	0.994 ±0.011	33.61 ±2.08	0.884 ±0.008	37491 ±239	0.902 ±0.020	48298 ±1651	0.901 ±0.025	50876 ±3416	0.933	4.61
Do-PFN 2D	0.876 ±0.168	16.60 ±20.99	0.792 ±0.228	10.73 ±4.24	0.412 ±0.022	16756 ±709	0.885 ±0.022	44751 ±1332	0.894 ±0.028	46219 ±2699	0.772	2.66
Do-PFN 2D (Indep)	0.992 ±0.015	26.85 ±30.45	0.989 ±0.021	27.17 ±3.71	0.994 ±0.002	50830 ±186	0.972 ±0.011	66848 ±2134	0.968 ±0.013	64829 ±4220	0.983	5.01
UWYK No-Anc	0.994 ±0.013	25.22 ±26.18	0.993 ±0.013	23.00 ±4.81	0.950 ±0.022	44818 ±943	0.882 ±0.026	59930 ±2474	0.896 ±0.028	59966 ±3175	0.943	4.49
UWYK No-Anc (matched label)	0.879 ±0.012	19.34 ±1.78	0.970 ±0.013	19.92 ±1.28	0.910 ±0.002	45477 ±206	0.815 ±0.003	55346 ±243	0.899 ±0.003	59030 ±347	0.895	4.09
UWYK No-Anc (matched resolution)	0.934 ±0.007	17.53 ±1.82	0.987 ±0.007	20.19 ±1.96	0.922 ±0.001	45281 ±107	0.823 ±0.003	54038 ±282	0.849 ±0.003	55418 ±318	0.903	3.96
UWYK No-Anc 2D	0.997 ±0.007	26.55 ±29.74	0.979 ±0.039	17.63 ±7.92	0.807 ±0.079	32345 ±4521	0.673 ±0.151	38456 ±9520	0.886 ±0.047	54241 ±4902	0.868	3.68
UWYK No-Anc 2D (Indep)	0.999 ±0.003	31.59 ±38.94	0.990 ±0.019	24.47 ±7.97	0.904 ±0.013	43378 ±1165	0.920 ±0.029	76185 ±5184	0.966 ±0.015	82357 ±5774	0.956	5.29
UWYK Anc	0.991 ±0.017	23.11 ±24.73	0.994 ±0.014	20.03 ±4.32	0.977 ±0.010	44641 ±1196	0.924 ±0.019	70949 ±4628	0.960 ±0.017	69246 ±5575	0.969	4.58
UWYK Anc (matched label)	0.894 ±0.011	20.17 ±1.97	0.972 ±0.013	19.95 ±1.33	0.904 ±0.002	44207 ±199	0.824 ±0.003	56049 ±262	0.926 ±0.003	59335 ±368	0.904	4.11
UWYK Anc (matched resolution)	0.937 ±0.010	16.17 ±1.69	0.992 ±0.004	19.37 ±1.98	0.928 ±0.001	44946 ±113	0.888 ±0.003	64791 ±585	0.916 ±0.002	60445 ±402	0.932	4.09
UWYK Anc 2D	0.995 ±0.010	22.76 ±25.71	0.977 ±0.038	16.18 ±6.98	0.698 ±0.125	27754 ±5118	0.577 ±0.154	32549 ±8797	0.689 ±0.090	38958 ±5099	0.787	3.09
UWYK Anc 2D (Indep)	0.999 ±0.005	27.07 ±33.38	0.989 ±0.023	21.79 ±6.95	0.900 ±0.011	42811 ±1157	0.877 ±0.036	67531 ±4619	0.912 ±0.031	70555 ±4960	0.935	4.72
CausalPFN-C	0.997 ±0.009	4.70 ±1.27	0.997 ±0.005	8.76 ±2.59	0.976 ±0.028	47788 ±2631	0.949 ±0.017	76381 ±12426	0.934 ±0.020	71295 ±12172	0.971	3.63
CausalPFN-C (matched label)	0.996 ±0.014	4.72 ±1.41	0.998 ±0.004	8.84 ±2.47	0.962 ±0.056	51740 ±8101	0.970 ±0.012	62529 ±8429	0.964 ±0.017	64467 ±11140	0.978	3.44
CausalPFN-C (matched resolution)	1.000 ±0.002	14.36 ±14.72	1.000 ±0.001	13.90 ±1.63	0.973 ±0.040	55337 ±7142	0.984 ±0.008	87607 ±11461	0.981 ±0.010	79227 ±14299	0.988	4.63
CausalPFN-C 2D	0.999 ±0.004	13.81 ±15.28	0.998 ±0.003	12.55 ±1.52	0.952 ±0.052	41861 ±1944	0.970 ±0.012	59139 ±8881	0.960 ±0.013	59309 ±9152	0.976	3.58
CausalPFN-C 2D (Indep)	0.999 ±0.003	13.90 ±15.09	0.999 ±0.003	14.62 ±1.67	0.961 ±0.045	43677 ±1696	0.980 ±0.009	71684 ±10407	0.970 ±0.011	66849 ±10608	0.982	3.98

I Additional Results

I.1 Matched-resolution and matched-label ablations

In Section 3.1 we discuss that joint supervision is more efficient due to two reasons: learning the dependence of outcomes and simply learning more by additional labels. The two ablations separate them. In the matched-resolution version, each baseline is re-trained at its 2D counterpart's grid resolution (J=10 for Do-PFN, J=32 for UWYK and CausalPFN), removing the resolution advantage of the baseline. In the matched-label version, each baseline is trained on both generated outcomes of every unit, receiving the same data information as the 2D model without learning the pair together. Tables 12, 13, and 14 report the calibration with posterior credible intervals.

Providing the additional data information improves the baseline for Do-PFN and CausalPFN, and matching the resolution decreases the coverage for Do-PFN and UWYK. For CausalPFN models, for all datasets for all variants including the joint supervision coverage rates are usually preserved. On RealCause, when we compare the changes with the 2D, we observe that the marginal calibration is similar to the matched-resolution variants for Do-PFN and better for UWYK and CausalPFN.

On the case studies, the 2D models of Do-PFN and UWYK reach higher coverage at shorter lengths in comparison to both ablations. Do-PFN 2D improves coverage by four points over both variants while reporting one fifth of the matched-resolution length and half of the matched-label length. UWYK 2D increases the coverage around twenty points over both variants at shorter length. CausalPFN shows similar coverage percentages for both ablations, and its 2D length improves only against the matched-resolution variant.

On ComplexMech, we observe similar patterns. For Do-PFN, when the data efficiency is removed, the gain reduces from around seventeen times to around four times, showing that for ComplexMech benchmark, portion of the improvement was due to additional data. However, reporting a length four times shorter still remains substantial. When the resolutions are matched, the increased amount this time goes from seventeen times to twentyfour times. For UWYK, when the data efficiency removed we still observe similar levels of improvement, and when the resolution is matched, joint

Table 13: 95% posterior credible interval calibration for the CATE on the six causal scenarios of the Do-PFN case studies (Coverage, Length), computed as in Table 2. Results are aggregated across varying treatment effect magnitudes and covariate dimensions. The matched-label and matched-resolution variants are denoted in brackets.

	Obs. Conf.		Obs. Med.		Obs. M+C		Unobs. Conf.		Frontdoor		Backdoor		Average	
Model	Coverage	Length	Coverage	Length	Coverage	Length	Coverage	Length	Coverage	Length	Coverage	Length	Coverage	Len/σ_Y
Do-PFN	0.996 ±0.026	5.70 ±4.25	0.662 ±0.381	0.72 ±0.55	0.921 ±0.176	4.05 ±7.75	0.995 ±0.036	6.18 ±4.28	1.000 ±0.000	4.35 ±3.97	1.000 ±0.003	4.54 ±5.39	0.929	3.01
Do-PFN (matched label)	0.997 ±0.001	5.35 ±0.10	0.642 ±0.010	0.65 ±0.01	0.915 ±0.005	3.36 ±0.06	0.998 ±0.000	5.85 ±0.10	0.998 ±0.000	4.67 ±0.11	0.999 ±0.000	4.71 ±0.10	0.925	2.96
Do-PFN (matched resolution)	0.949 ±0.142	12.28 ±5.86	0.714 ±0.406	3.95 ±4.19	0.931 ±0.140	9.26 ±5.16	0.978 ±0.081	13.16 ±5.06	0.994 ±0.031	11.29 ±5.49	0.995 ±0.025	11.41 ±5.43	0.927	7.41
Do-PFN 2D	0.983 ±0.109	2.62 ±2.23	0.953 ±0.160	0.90 ±1.08	0.929 ±0.198	2.34 ±1.63	0.983 ±0.103	2.84 ±2.36	0.995 ±0.058	1.49 ±1.53	0.990 ±0.087	1.54 ±1.56	0.972	1.32
UWYK No-Anc	0.960 ±0.102	5.37 ±4.71	0.225 ±0.380	0.69 ±0.51	0.483 ±0.457	2.94 ±2.53	0.987 ±0.046	6.11 ±5.17	1.000 ±0.007	4.23 ±5.34	0.999 ±0.011	4.24 ±5.27	0.776	2.79
UWYK No-Anc (matched label)	0.885 ±0.005	4.82 ±0.10	0.278 ±0.009	0.80 ±0.02	0.439 ±0.011	2.45 ±0.05	0.941 ±0.003	5.46 ±0.11	0.997 ±0.001	3.35 ±0.11	0.997 ±0.001	3.34 ±0.10	0.756	2.35
UWYK No-Anc (matched resolution)	0.983 ±0.001	4.71 ±0.11	0.219 ±0.009	0.81 ±0.01	0.553 ±0.009	2.83 ±0.06	0.995 ±0.001	5.60 ±0.12	0.995 ±0.001	3.12 ±0.11	0.997 ±0.000	3.11 ±0.11	0.790	2.31
UWYK No-Anc 2D	0.999 ±0.012	3.19 ±2.21	0.916 ±0.274	0.98 ±0.59	0.997 ±0.028	3.54 ±1.98	0.998 ±0.021	3.59 ±2.41	0.998 ±0.015	2.06 ±1.47	0.997 ±0.022	2.20 ±1.60	0.984	1.77
UWYK Anc	0.998 ±0.019	4.25 ±4.35	0.492 ±0.351	1.06 ±0.72	0.640 ±0.391	3.19 ±2.34	1.000 ±0.006	5.28 ±5.25	1.000 ±0.011	3.80 ±4.94	1.000 ±0.003	4.13 ±5.18	0.855	2.60
UWYK Anc (matched label)	0.884 ±0.004	4.92 ±0.11	0.294 ±0.009	0.82 ±0.02	0.483 ±0.010	2.73 ±0.06	0.953 ±0.003	5.59 ±0.11	0.999 ±0.000	3.66 ±0.11	0.997 ±0.001	3.61 ±0.11	0.768	2.50
UWYK Anc (matched resolution)	0.998 ±0.000	3.41 ±0.10	0.297 ±0.009	0.80 ±0.02	0.599 ±0.010	2.69 ±0.06	0.999 ±0.000	4.32 ±0.12	0.994 ±0.001	2.93 ±0.10	0.998 ±0.000	3.14 ±0.11	0.814	2.05
UWYK Anc 2D	0.995 ±0.042	2.43 ±2.05	0.921 ±0.266	0.84 ±0.47	0.997 ±0.029	3.28 ±1.70	0.998 ±0.020	3.57 ±2.43	0.998 ±0.013	2.53 ±1.69	0.997 ±0.020	1.52 ±1.44	0.984	1.63
CausalPFN-C	0.988 ±0.032	1.49 ±2.90	0.808 ±0.245	1.86 ±1.16	0.920 ±0.195	3.34 ±3.70	0.987 ±0.053	2.75 ±3.38	0.998 ±0.009	1.24 ±2.15	0.999 ±0.007	1.24 ±2.12	0.950	1.33
CausalPFN-C (matched label)	0.994 ±0.017	1.37 ±2.37	0.987 ±0.058	2.24 ±1.60	0.920 ±0.196	3.62 ±3.84	0.991 ±0.038	2.53 ±2.86	0.999 ±0.005	1.20 ±1.94	0.999 ±0.004	1.19 ±1.86	0.982	1.36
CausalPFN-C (matched resolution)	1.000 ±0.000	3.62 ±3.43	0.911 ±0.205	5.80 ±4.29	0.988 ±0.064	5.52 ±4.26	1.000 ±0.001	4.44 ±3.95	1.000 ±0.000	2.87 ±3.29	1.000 ±0.000	2.91 ±3.02	0.983	2.89
CausalPFN-C 2D	0.999 ±0.007	2.85 ±1.65	0.920 ±0.200	1.86 ±1.11	0.973 ±0.084	2.54 ±1.38	0.997 ±0.013	3.18 ±2.06	0.999 ±0.007	2.27 ±1.88	0.999 ±0.004	2.35 ±1.83	0.981	1.77

Table 14: 95% posterior credible interval calibration for the CATE on ComplexMech, reported per graph size (Coverage, Length), computed as in Table 2. The matched-label and matched-resolution variants are denoted in brackets.

	n=5		n=10		n=20		n=30		n=40		n=50		Average	
Model	Coverage	Length	Coverage	Length	Coverage	Length	Coverage	Length	Coverage	Length	Coverage	Length	Coverage	Len/σ_Y
Do-PFN	0.955 ±0.194	6.71 ±19.50	0.961 ±0.132	7.15 ±21.79	0.979 ±0.108	2.05 ±3.79	0.990 ±0.088	2.51 ±5.96	0.999 ±0.010	3.42 ±13.49	0.985 ±0.074	3.19 ±10.67	0.978	22.06
Do-PFN (matched label)	0.915 ±0.022	0.69 ±0.09	0.952 ±0.015	0.74 ±0.07	0.966 ±0.013	0.91 ±0.10	0.988 ±0.005	1.10 ±0.11	0.988 ±0.008	0.96 ±0.09	0.981 ±0.008	0.86 ±0.08	0.965	4.79
Do-PFN (matched resolution)	0.889 ±0.247	4.80 ±4.43	0.953 ±0.149	4.92 ±4.36	0.978 ±0.121	5.65 ±3.71	0.997 ±0.028	6.42 ±4.32	1.000 ±0.003	6.54 ±4.66	0.998 ±0.021	6.68 ±4.09	0.969	31.90
Do-PFN 2D	0.899 ±0.247	0.26 ±0.27	0.949 ±0.189	0.23 ±0.19	0.962 ±0.146	0.24 ±0.22	0.971 ±0.097	0.25 ±0.25	0.974 ±0.130	0.24 ±0.23	0.983 ±0.089	0.26 ±0.26	0.956	1.32
UWYK No-Anc	0.879 ±0.271	0.74 ±0.57	0.924 ±0.182	0.73 ±0.66	0.920 ±0.230	0.61 ±0.46	0.964 ±0.146	0.78 ±0.79	0.962 ±0.155	0.62 ±0.61	0.966 ±0.143	0.66 ±0.57	0.936	3.74
UWYK No-Anc (matched label)	0.863 ±0.026	0.77 ±0.06	0.896 ±0.022	0.68 ±0.06	0.920 ±0.021	0.56 ±0.05	0.949 ±0.015	0.70 ±0.06	0.956 ±0.016	0.58 ±0.05	0.965 ±0.012	0.59 ±0.05	0.925	3.49
UWYK No-Anc (matched resolution)	0.771 ±0.034	0.64 ±0.05	0.830 ±0.029	0.61 ±0.06	0.828 ±0.029	0.50 ±0.04	0.857 ±0.024	0.62 ±0.05	0.838 ±0.028	0.53 ±0.05	0.866 ±0.024	0.57 ±0.05	0.832	3.13
UWYK No-Anc 2D	0.972 ±0.153	0.47 ±0.48	0.977 ±0.133	0.38 ±0.37	0.969 ±0.144	0.31 ±0.35	0.992 ±0.036	0.30 ±0.36	0.981 ±0.127	0.30 ±0.39	0.992 ±0.030	0.30 ±0.37	0.981	1.84
UWYK Anc	0.950 ±0.172	0.80 ±0.52	0.993 ±0.028	0.81 ±0.54	0.959 ±0.162	0.66 ±0.44	0.975 ±0.123	0.81 ±0.81	0.973 ±0.141	0.62 ±0.62	0.981 ±0.095	0.62 ±0.56	0.972	3.89
UWYK Anc (matched label)	0.926 ±0.020	0.94 ±0.07	0.934 ±0.016	0.74 ±0.06	0.924 ±0.021	0.56 ±0.05	0.950 ±0.015	0.69 ±0.07	0.955 ±0.016	0.56 ±0.06	0.956 ±0.015	0.58 ±0.05	0.941	3.64
UWYK Anc (matched resolution)	0.850 ±0.030	0.68 ±0.05	0.884 ±0.026	0.58 ±0.04	0.856 ±0.026	0.49 ±0.04	0.871 ±0.023	0.63 ±0.06	0.843 ±0.028	0.52 ±0.05	0.879 ±0.024	0.55 ±0.05	0.864	3.11
UWYK Anc 2D	0.967 ±0.155	0.40 ±0.41	0.972 ±0.137	0.25 ±0.24	0.967 ±0.144	0.22 ±0.24	0.978 ±0.101	0.20 ±0.23	0.974 ±0.131	0.21 ±0.30	0.987 ±0.042	0.23 ±0.28	0.974	1.34
CausalPFN-C	0.861 ±0.286	0.54 ±0.57	0.907 ±0.204	0.47 ±0.50	0.931 ±0.169	0.43 ±0.44	0.952 ±0.119	0.54 ±0.65	0.962 ±0.120	0.43 ±0.54	0.953 ±0.128	0.43 ±0.49	0.928	2.57
CausalPFN-C (matched label)	0.912 ±0.226	0.71 ±0.87	0.946 ±0.162	0.53 ±0.53	0.963 ±0.122	0.48 ±0.47	0.965 ±0.100	0.54 ±0.63	0.974 ±0.100	0.42 ±0.51	0.964 ±0.120	0.44 ±0.46	0.954	2.79
CausalPFN-C (matched resolution)	0.954 ±0.170	0.86 ±1.13	0.998 ±0.009	0.71 ±0.54	0.989 ±0.096	0.68 ±0.59	0.999 ±0.004	0.72 ±0.65	0.992 ±0.090	0.61 ±0.56	0.998 ±0.016	0.60 ±0.49	0.988	3.74
CausalPFN-C 2D	0.963 ±0.154	0.56 ±0.44	0.985 ±0.071	0.55 ±0.37	0.982 ±0.108	0.53 ±0.39	0.998 ±0.012	0.58 ±0.50	0.990 ±0.090	0.50 ±0.43	0.999 ±0.004	0.50 ±0.40	0.986	2.89

has better coverage around 10 points with still around half of the length. For CausalPFN, when the data efficiency is removed, the coverage is better calibrated around 0.95 with similar length; however, when the resolutions are matched, joint supervision again provides a shorter length.

For CausalPFN, as explained before, due to its output range and standardization choices, the resolution is a serious bottleneck. A different training architecture, such as changing the standardization in the training, could make it less important. However, in this paper, we want to preserve the baseline as much as possible. Hence, for CausalPFN, the improvements are mostly observed in the matched-resolution case.

From these ablation studies we can conclude two things. Matching the labels, giving same amount of data points, does not recover the 2D performance, so the improvement of the joint supervision is not the additional data alone. Seeing the same number of worlds does not substitute for learning them together. Also, improvements grow across all baselines when the resolutions match, which is a fairer comparison where the true difference can be observed.

I.2 Varying the correlation $\rho(x)$

The main results restrict ComplexMech to $\rho > 0.99$. This ablation removes the restriction and spans all correlation levels. Point estimation is reported in Table 15 and calibration in Table 16.

Joint supervision has lower improvement rate in this setting comparison to main results. This does not indicate a weakness of the joint setup. As ρ approaches 0, the outcomes starts to decouple, and the advantage of the learned dependence reduces by construction. Nevertheless, we still observe substantial point-estimation improvements: $62\% - 69\%$ for Do-PFN and 10–32% for UWYK, while CausalPFN shows no difference. Coverage decreases for all models in this regime. The miscalibration grows where ρ varies across realizations. In the main results, ρ is fixed near one, and calibration is markedly better; when it varies, each realization requires a different dependence, which is harder to calibrate.

Table 15: Unit- and population-level point estimation on ComplexMech across all correlation levels ($\sqrt{\text{PEHE}}$, ϵ_{ATE}; lower is better; mean $\pm$ standard error of the mean over 100 realizations), reported per graph size, computed as in Table 1. Unlike Table 11, no ρ restriction is applied. Bold marks the better model within each baseline. Δ summarises the relative improvement across graph sizes by geometric mean; positive favours 2D.

	n=5		n=10		n=20		n=30		n=40		n=50		Δ (%)	
Model	$\sqrt{\text{PEHE}}\downarrow$	$\epsilon_{\text{ATE}}\downarrow$	$\sqrt{\text{PEHE}}\downarrow$	$\epsilon_{\text{ATE}}\downarrow$	$\sqrt{\text{PEHE}}\downarrow$	$\epsilon_{\text{ATE}}\downarrow$	$\sqrt{\text{PEHE}}\downarrow$	$\epsilon_{\text{ATE}}\downarrow$	$\sqrt{\text{PEHE}}\downarrow$	$\epsilon_{\text{ATE}}\downarrow$	$\sqrt{\text{PEHE}}\downarrow$	$\epsilon_{\text{ATE}}\downarrow$	$\sqrt{\text{PEHE}}$	ϵ_{ATE}
Do-PFN	0.820 ±0.384	1.940 ±0.764	0.566 ±0.124	1.272 ±0.295	0.536 ±0.074	1.144 ±0.175	0.636 ±0.110	1.690 ±0.447	0.360 ±0.073	1.223 ±0.514	0.370 ±0.053	0.765 ±0.124	**+61.8**	**+69.5**
Do-PFN 2D	0.183 ±0.020	0.330 ±0.035	0.200 ±0.021	0.381 ±0.052	0.208 ±0.020	0.389 ±0.043	0.239 ±0.024	0.425 ±0.044	0.156 ±0.016	0.285 ±0.035	0.231 ±0.028	0.607 ±0.148	–	–
UWYK No-Anc	0.295 ±0.034	0.551 ±0.038	0.280 ±0.029	0.620 ±0.040	0.241 ±0.023	0.500 ±0.038	0.256 ±0.026	0.465 ±0.039	0.171 ±0.021	0.348 ±0.036	0.215 ±0.025	0.374 ±0.038	**+22.0**	**+31.1**
UWYK No-Anc 2D	0.173 ±0.020	0.331 ±0.041	0.190 ±0.018	0.334 ±0.043	0.192 ±0.019	0.320 ±0.061	0.228 ±0.023	0.347 ±0.045	0.149 ±0.016	0.265 ±0.035	0.197 ±0.023	0.342 ±0.046	–	–
UWYK Anc	0.203 ±0.025	0.332 ±0.040	0.224 ±0.022	0.467 ±0.042	0.218 ±0.020	0.375 ±0.033	0.243 ±0.025	0.386 ±0.036	0.162 ±0.020	0.285 ±0.027	0.209 ±0.025	0.351 ±0.035	**+10.9**	**+20.6**
UWYK Anc 2D	0.170 ±0.020	0.300 ±0.036	0.183 ±0.018	0.278 ±0.039	0.196 ±0.019	0.297 ±0.040	0.228 ±0.024	0.297 ±0.032	0.150 ±0.017	0.240 ±0.032	0.196 ±0.024	0.319 ±0.043	–	–
CausalPFN-C	0.210 ±0.024	0.368 ±0.042	0.225 ±0.022	0.478 ±0.061	0.221 ±0.019	0.438 ±0.050	0.251 ±0.024	0.502 ±0.065	0.165 ±0.017	0.310 ±0.033	0.208 ±0.024	0.341 ±0.042	−0.0	−4.5
CausalPFN-C 2D	0.197 ±0.023	0.343 ±0.039	0.213 ±0.019	0.472 ±0.064	0.218 ±0.019	0.425 ±0.065	0.255 ±0.023	0.483 ±0.056	0.171 ±0.017	0.391 ±0.059	0.225 ±0.024	0.410 ±0.049	–	–

Table 16: 95% posterior credible interval calibration for the CATE on ComplexMech across all correlation levels, reported per graph size (Coverage, Length), computed as in Table 2. Unlike Table 4, no ρ restriction is applied. Average reports the mean coverage and the length normalised by the outcome scale (Len/σ_Y).

	n=5		n=10		n=20		n=30		n=40		n=50		**Average**	
Model	Coverage	Length	Coverage	Length	Coverage	Length	Coverage	Length	Coverage	Length	Coverage	Length	Coverage	Len/σ_Y
Do-PFN	0.886 ±0.234	5.44 ±30.99	0.917 ±0.178	5.72 ±30.64	0.913 ±0.178	2.16 ±4.55	0.937 ±0.137	2.94 ±6.44	0.973 ±0.092	3.15 ±16.94	0.950 ±0.129	4.42 ±26.51	0.929	14.62
Do-PFN 2D	0.770 ±0.321	0.39 ±0.32	0.791 ±0.269	0.41 ±0.29	0.812 ±0.280	0.46 ±0.32	0.792 ±0.270	0.56 ±0.49	0.828 ±0.233	0.44 ±0.40	0.835 ±0.256	0.67 ±0.71	0.805	1.78
UWYK No-Anc	0.851 ±0.262	1.04 ±0.69	0.805 ±0.279	0.89 ±0.66	0.819 ±0.281	0.87 ±0.75	0.835 ±0.257	0.98 ±0.78	0.891 ±0.209	0.77 ±0.58	0.907 ±0.195	1.02 ±0.85	0.851	3.38
UWYK No-Anc 2D	0.926 ±0.204	0.85 ±0.69	0.956 ±0.115	0.90 ±0.76	0.922 ±0.201	0.73 ±0.62	0.906 ±0.210	0.87 ±0.76	0.937 ±0.125	0.65 ±0.62	0.909 ±0.194	0.76 ±0.72	0.926	2.89
UWYK Anc	0.927 ±0.186	1.03 ±0.60	0.934 ±0.136	1.00 ±0.66	0.897 ±0.192	0.98 ±0.74	0.867 ±0.228	1.01 ±0.76	0.922 ±0.192	0.77 ±0.59	0.907 ±0.189	0.98 ±0.83	0.909	3.51
UWYK Anc 2D	0.909 ±0.208	0.70 ±0.61	0.913 ±0.179	0.70 ±0.66	0.872 ±0.242	0.53 ±0.53	0.839 ±0.260	0.59 ±0.61	0.899 ±0.159	0.47 ±0.51	0.870 ±0.239	0.61 ±0.65	0.884	2.20
CausalPFN-C	0.817 ±0.280	0.79 ±0.66	0.847 ±0.230	0.85 ±0.80	0.820 ±0.253	0.71 ±0.66	0.814 ±0.238	0.69 ±0.59	0.840 ±0.223	0.57 ±0.56	0.854 ±0.221	0.71 ±0.77	0.832	2.63
CausalPFN-C 2D	0.909 ±0.197	0.84 ±0.55	0.916 ±0.169	0.87 ±0.62	0.913 ±0.173	0.86 ±0.60	0.888 ±0.191	0.90 ±0.55	0.929 ±0.129	0.67 ±0.44	0.901 ±0.188	0.87 ±0.64	0.909	3.05

I.3 CausalPFN two-stage comparison

The main results compare against the re-trained second stage of CausalPFN (Section 4). Table 17 compares the 2D model against the published two-stage checkpoint instead.

I.4 UWYK binary re-training

The published UWYK checkpoint is trained for continuous treatment prediction (Section 4). Table 18 compares the 2D model against the re-trained binary version, UWYK-B, for both the ancestral and no-ancestral variants.

I.5 Density estimation on RealCause

Beyond the point-estimation and calibration results of Section 5.1, we evaluate each models predicted densities directly with ground-truth RealCause targets. This evaluation is done on IHDP (Hill, 2011) and ACIC 2016 (Hahn et al., 2019), which provide the true conditional means $\mu_t(x)$ and simulated potential outcomes under both treatments. Table 19 reports the evaluation across marginals $f_{Y_0} + f_{Y_1}$, CATE f_τ, and ATE f_{ATE}. We additionally evaluate the treatment-effect densities after applying MALC-T to the native discrete f_τ in Table 20. The response densities f_{Y_0} and f_{Y_1} are unchanged, while f_{ATE} is constructed from the MALC-T effect densities using the Wasserstein barycenter of Section 3.2.1.

For IHDP, we use all 100 realizations of IHDP-100. In each realization, we use the pre-split 672 units form context and 75 units form the queries. For ACIC, we use all 10 realizations in the ACIC 2016 `causallib` benchmark. Each realization has 4802 units, which we split following (Balazadeh et al., 2025): 4,321 training, sampled to 1,000 units as context and 481 test units.

For query covariates $x_1, \ldots, x_Q$, we report three negative log-likelihoods:

$$\mathcal{L}_{Y_0+Y_1} = -\frac{1}{Q}\sum_{q=1}^{Q}\left[\log \hat{f}_0(\mu_0(x_q) \mid x_q) + \log \hat{f}_1(\mu_1(x_q) \mid x_q)\right]. \tag{38}$$

$$\mathcal{L}_\tau = -\frac{1}{Q}\sum_{q=1}^{Q}\log \hat{f}_\tau(\tau_q^\star \mid x_q), \qquad \tau_q^\star = Y_q^{\text{do}(1)} - Y_q^{\text{do}(0)}. \tag{39}$$

$$\mathcal{L}_{\text{ATE}} = -\log \hat{f}_{\text{ATE}}(\theta^\star), \qquad \theta^\star = \frac{1}{Q}\sum_{q=1}^{Q}[\mu_1(x_q) - \mu_0(x_q)]. \tag{40}$$

Table 17: Comparison of CausalPFN-C 2D against the published two-stage CausalPFN checkpoint across all benchmarks. The main results instead compare against the re-trained second stage (Section 4). Left columns report point estimation ($\sqrt{\text{PEHE}}$, ϵ_{ATE}; lower is better), right columns 95% posterior credible interval calibration (Coverage, Len/σ_Y), computed as in Tables 1 and 2. Δ / Average rows summarise per benchmark; positive Δ favours 2D.

	CausalPFN		**CausalPFN-C 2D**		**CausalPFN**		**CausalPFN-C 2D**	
Dataset	$\sqrt{\text{PEHE}}\downarrow$	$\epsilon_{\text{ATE}}\downarrow$	$\sqrt{\text{PEHE}}\downarrow$	$\epsilon_{\text{ATE}}\downarrow$	Coverage	Len/σ_Y	Coverage	Len/σ_Y
RealCause								
IHDP	0.632 ±0.063	0.077 ±0.018	1.200 ±0.105	0.071 ±0.014	1.000 ±0.000	3.44 ±3.74	0.999 ±0.004	2.62 ±2.90
ACIC	1.119 ±0.124	0.052 ±0.014	1.264 ±0.204	0.090 ±0.018	1.000 ±0.001	3.16 ±0.34	0.998 ±0.003	2.35 ±0.28
CPS	10084 ±178	0.436 ±0.037	10902 ±176	0.389 ±0.040	0.991 ±0.008	5.67 ±0.20	0.952 ±0.052	4.87 ±0.23
PSID	14453 ±184	0.244 ±0.014	15557 ±212	0.253 ±0.016	0.982 ±0.008	5.33 ±0.74	0.970 ±0.012	4.03 ±0.61
PSID_bal	14791 ±213	0.305 ±0.018	16769 ±310	0.334 ±0.025	0.982 ±0.009	5.46 ±0.88	0.960 ±0.013	4.05 ±0.62
Δ / Average	–	–	−23.1	−10.0	0.991	4.61	0.976	3.58
Case studies								
Obs. Conf.	0.166 ±0.006	0.084 ±0.006	0.208 ±0.007	0.294 ±0.018	1.000 ±0.000	2.11 ±1.32	0.999 ±0.007	1.49 ±0.86
Obs. Med.	0.528 ±0.009	0.349 ±0.005	0.465 ±0.013	0.324 ±0.007	1.000 ±0.000	3.18 ±2.05	0.920 ±0.200	1.30 ±0.78
Obs. M+C	0.549 ±0.012	0.529 ±0.025	0.475 ±0.013	0.513 ±0.018	1.000 ±0.004	3.19 ±2.27	0.973 ±0.084	1.51 ±0.82
Unobs. Conf.	0.370 ±0.011	0.427 ±0.022	0.394 ±0.012	0.536 ±0.028	1.000 ±0.000	2.47 ±1.70	0.997 ±0.013	1.65 ±1.07
Frontdoor	0.119 ±0.005	0.212 ±0.009	0.202 ±0.007	0.788 ±0.026	1.000 ±0.000	3.19 ±2.76	0.999 ±0.007	2.30 ±1.90
Backdoor	0.124 ±0.005	0.219 ±0.010	0.214 ±0.007	0.895 ±0.029	1.000 ±0.000	3.29 ±2.62	0.999 ±0.004	2.37 ±1.85
Δ / Average	–	–	−20.1	−97.9	1.000	2.91	0.981	1.77
ComplexMech								
n=5	0.112 ±0.024	0.322 ±0.065	0.102 ±0.022	0.307 ±0.056	0.953 ±0.183	4.49 ±4.66	0.963 ±0.154	2.68 ±2.10
n=10	0.074 ±0.010	0.230 ±0.033	0.075 ±0.011	0.247 ±0.030	0.999 ±0.003	4.36 ±3.27	0.985 ±0.071	2.80 ±1.91
n=20	0.057 ±0.011	0.134 ±0.016	0.068 ±0.010	0.180 ±0.025	0.998 ±0.013	4.07 ±3.64	0.982 ±0.108	2.73 ±2.03
n=30	0.053 ±0.006	0.175 ±0.023	0.067 ±0.007	0.188 ±0.023	1.000 ±0.001	4.75 ±3.90	0.998 ±0.012	3.35 ±2.86
n=40	0.041 ±0.005	0.154 ±0.029	0.054 ±0.006	0.184 ±0.026	0.991 ±0.090	4.04 ±3.54	0.990 ±0.090	2.91 ±2.52
n=50	0.043 ±0.005	0.135 ±0.023	0.058 ±0.006	0.179 ±0.027	1.000 ±0.000	4.18 ±3.14	0.999 ±0.004	2.90 ±2.35
Δ / Average	–	–	−16.1	−15.3	0.990	4.31	0.986	2.89

Table 18: Comparison of UWYK 2D against the re-trained binary-treatment baseline (UWYK-B) across all benchmarks, for the ancestral and no-ancestral variants. The published checkpoint covers only continuous treatment prediction (Section 4). Left columns report point estimation ($\sqrt{\text{PEHE}}$, ϵ_{ATE}; lower is better), right columns 95% posterior credible interval calibration (Coverage, Len/σ_Y), computed as in Tables 1 and 2. Δ / Average rows summarise per benchmark; positive Δ favours 2D.

	No-Anc								**Anc**							
	UWYK-B		**UWYK 2D**		**UWYK-B**		**UWYK 2D**		**UWYK-B**		**UWYK 2D**		**UWYK-B**		**UWYK 2D**	
Dataset	$\sqrt{\text{PEHE}}\downarrow$	$\epsilon_{\text{ATE}}\downarrow$	$\sqrt{\text{PEHE}}\downarrow$	$\epsilon_{\text{ATE}}\downarrow$	Coverage	Len/σ_Y	Coverage	Len/σ_Y	$\sqrt{\text{PEHE}}\downarrow$	$\epsilon_{\text{ATE}}\downarrow$	$\sqrt{\text{PEHE}}\downarrow$	$\epsilon_{\text{ATE}}\downarrow$	Coverage	Len/σ_Y	Coverage	Len/σ_Y
RealCause																
IHDP	6.283 ±0.787	0.678 ±0.050	4.508 ±0.626	0.375 ±0.088	0.994 ±0.013	4.79 ±4.97	0.997 ±0.007	5.03 ±5.64	5.477 ±0.771	0.489 ±0.079	4.305 ±0.623	0.325 ±0.077	0.991 ±0.017	4.38 ±4.67	0.995 ±0.010	4.32 ±4.88
ACIC	3.350 ±0.504	0.445 ±0.083	2.743 ±0.474	0.134 ±0.042	0.993 ±0.013	4.30 ±0.90	0.979 ±0.039	3.30 ±1.48	2.768 ±0.483	0.233 ±0.083	2.722 ±0.468	0.117 ±0.034	0.994 ±0.014	3.75 ±0.81	0.977 ±0.038	3.03 ±1.30
CPS	12921 ±17	1.008 ±0.001	12299 ±31	0.876 ±0.007	0.941 ±0.017	5.17 ±0.09	0.807 ±0.079	3.76 ±0.53	12835 ±17	1.005 ±0.001	12475 ±28	0.915 ±0.005	0.910 ±0.014	5.03 ±0.14	0.698 ±0.125	3.23 ±0.59
PSID	22486 ±129	0.985 ±0.000	21426 ±145	0.849 ±0.012	0.874 ±0.026	4.03 ±0.16	0.673 ±0.151	2.62 ±0.65	22070 ±129	0.973 ±0.001	21850 ±136	0.903 ±0.009	0.875 ±0.026	4.24 ±0.20	0.577 ±0.154	2.22 ±0.60
PSID_bal	22752 ±129	1.005 ±0.002	20039 ±141	0.676 ±0.009	0.878 ±0.026	4.08 ±0.22	0.886 ±0.047	3.70 ±0.33	19487 ±207	0.484 ±0.020	21467 ±141	0.864 ±0.006	0.957 ±0.016	4.37 ±0.35	0.689 ±0.090	2.66 ±0.35
Δ / Average	–	–	**+14.0**	**+39.1**	0.936	4.47	0.868	3.68	–	–	**+3.9**	**+12.8**	0.945	4.35	0.787	3.09
Case studies																
Obs. Conf.	0.945 ±0.016	0.627 ±0.006	0.258 ±0.004	0.395 ±0.020	0.959 ±0.103	2.79 ±2.48	0.999 ±0.012	1.66 ±1.15	0.347 ±0.006	0.345 ±0.013	0.264 ±0.005	0.300 ±0.011	0.996 ±0.026	2.23 ±2.31	0.995 ±0.042	1.27 ±1.07
Obs. Med.	1.370 ±0.021	0.879 ±0.006	0.165 ±0.004	0.110 ±0.002	0.213 ±0.381	0.43 ±0.35	0.916 ±0.274	0.69 ±0.41	1.300 ±0.020	0.830 ±0.006	0.133 ±0.002	0.086 ±0.001	0.403 ±0.382	0.62 ±0.51	0.921 ±0.266	0.59 ±0.33
Obs. M+C	1.395 ±0.020	0.936 ±0.006	0.330 ±0.005	0.418 ±0.020	0.476 ±0.459	1.73 ±1.51	0.997 ±0.028	2.10 ±1.17	1.282 ±0.019	0.857 ±0.007	0.345 ±0.005	0.422 ±0.019	0.586 ±0.420	1.82 ±1.38	0.997 ±0.029	1.94 ±1.01
Unobs. Conf.	0.921 ±0.015	0.612 ±0.006	0.275 ±0.005	0.413 ±0.021	0.986 ±0.050	3.16 ±2.70	0.998 ±0.021	1.86 ±1.25	0.356 ±0.008	0.330 ±0.014	0.269 ±0.004	0.425 ±0.022	0.999 ±0.007	2.74 ±2.76	0.998 ±0.020	1.85 ±1.26
Frontdoor	0.083 ±0.002	0.389 ±0.007	0.181 ±0.004	1.024 ±0.032	1.000 ±0.008	4.29 ±5.41	0.998 ±0.015	2.09 ±1.49	0.096 ±0.003	0.404 ±0.008	0.197 ±0.004	1.062 ±0.033	1.000 ±0.011	3.85 ±4.99	0.998 ±0.013	2.57 ±1.72
Backdoor	0.084 ±0.002	0.436 ±0.008	0.200 ±0.005	1.207 ±0.040	0.999 ±0.011	4.28 ±5.32	0.997 ±0.022	2.22 ±1.61	0.159 ±0.005	0.857 ±0.028	0.143 ±0.004	0.760 ±0.023	1.000 ±0.000	4.15 ±5.20	0.997 ±0.020	1.54 ±1.46
Δ / Average	–	–	**+52.1**	**+25.3**	0.772	2.78	0.984	1.77	–	–	**+44.5**	**+28.6**	0.831	2.57	0.984	1.63
ComplexMech																
n=5	0.201 ±0.039	0.420 ±0.034	0.083 ±0.022	0.241 ±0.061	0.847 ±0.272	2.92 ±2.43	0.972 ±0.153	2.26 ±2.32	0.178 ±0.036	0.346 ±0.034	0.078 ±0.021	0.215 ±0.046	0.906 ±0.208	3.24 ±2.52	0.967 ±0.155	1.93 ±1.97
n=10	0.104 ±0.022	0.299 ±0.031	0.056 ±0.009	0.215 ±0.045	0.884 ±0.213	2.96 ±2.63	0.977 ±0.133	1.94 ±1.90	0.096 ±0.019	0.259 ±0.024	0.045 ±0.008	0.142 ±0.024	0.937 ±0.135	3.10 ±2.55	0.972 ±0.137	1.26 ±1.22
n=20	0.070 ±0.019	0.204 ±0.029	0.051 ±0.010	0.172 ±0.029	0.925 ±0.220	2.75 ±2.43	0.969 ±0.144	1.61 ±1.80	0.066 ±0.017	0.192 ±0.025	0.041 ±0.010	0.104 ±0.013	0.933 ±0.207	2.75 ±2.31	0.967 ±0.144	1.15 ±1.22
n=30	0.035 ±0.005	0.161 ±0.025	0.029 ±0.004	0.100 ±0.014	0.958 ±0.152	4.03 ±4.46	0.992 ±0.036	1.72 ±2.05	0.034 ±0.005	0.147 ±0.021	0.026 ±0.004	0.083 ±0.012	0.964 ±0.141	4.00 ±4.54	0.978 ±0.101	1.15 ±1.35
n=40	0.028 ±0.004	0.128 ±0.025	0.025 ±0.003	0.095 ±0.018	0.956 ±0.180	3.36 ±3.63	0.981 ±0.127	1.74 ±2.28	0.028 ±0.004	0.126 ±0.022	0.021 ±0.003	0.064 ±0.011	0.957 ±0.179	3.23 ±3.49	0.974 ±0.131	1.24 ±1.75
n=50	0.036 ±0.007	0.134 ±0.022	0.028 ±0.004	0.117 ±0.023	0.961 ±0.152	3.36 ±3.11	0.992 ±0.030	1.75 ±2.14	0.036 ±0.007	0.133 ±0.022	0.024 ±0.003	0.092 ±0.017	0.965 ±0.138	3.30 ±3.15	0.987 ±0.042	1.32 ±1.61
Δ / Average	–	–	**+32.9**	**+28.0**	0.922	3.23	0.981	1.84	–	–	**+39.6**	**+42.5**	0.944	3.27	0.974	1.34

For the baseline models, we obtain both 1D marginal densities from separate predictions and the effect density is obtained by convolution (matching the RealCause data-generating process). For 2D models, the response and effect densities are obtained directly from the predicted joint distribution.

I.6 Case study: decomposition by dimension

The results in Table 10 aggregate the case studies across covariate dimensions $d \in \{5, 10, 20, 30, 40, 50\}$. Figures 2 and 3 decompose the point estimation by dimension.

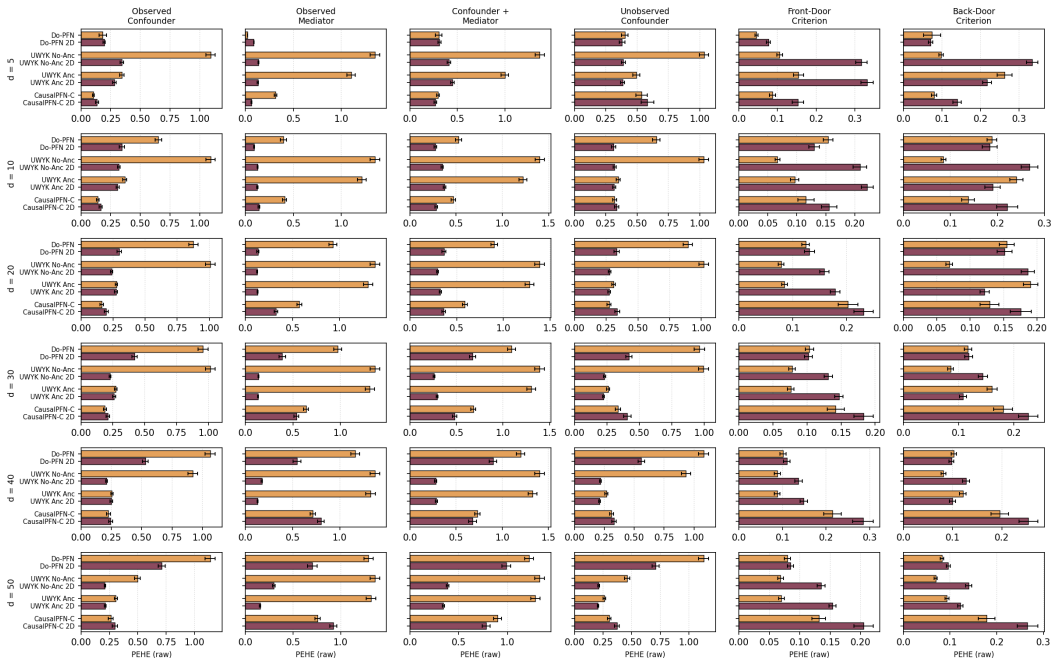

Figure 2: Unit-level point estimation ($\sqrt{\text{PEHE}}$; lower is better) on the six causal scenarios of the case studies, decomposed by covariate dimension d (rows) and scenario (columns), at context size $N = 1000$. Bars are means and whiskers standard errors of the mean over realizations. Each panel compares the single-outcome models against their jointly supervised (2D) variants.

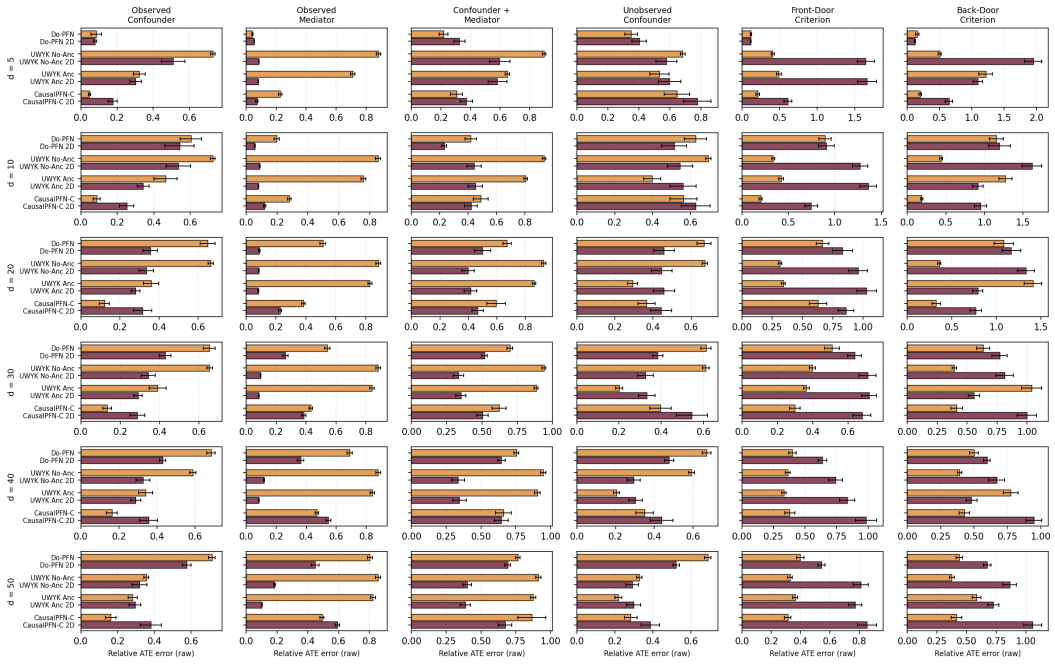

Figure 3: Population-level point estimation (ϵ_{ATE}; lower is better) on the six causal scenarios of the case studies, decomposed by covariate dimension d (rows) and scenario (columns), at context size $N = 1000$. Bars are means and whiskers standard errors of the mean over realizations. Each panel compares the single-outcome models against their jointly supervised (2D) variants.

Table 19: Negative log-likelihood (NLL) of the models' native predictive densities on IHDP and ACIC 2016, without MALC refitting. Lower is better. Results are mean ± standard error across 100 IHDP and 10 ACIC realizations

Method	IHDP: $\mathcal{L}_{Y_0+Y_1}\downarrow$	IHDP: $\mathcal{L}_{\tau}\downarrow$	IHDP: $\mathcal{L}_{\mathrm{ATE}}\downarrow$	ACIC: $\mathcal{L}_{Y_0+Y_1}\downarrow$	ACIC: $\mathcal{L}_{\tau}\downarrow$	ACIC: $\mathcal{L}_{\mathrm{ATE}}\downarrow$
Do-PFN	-0.6100 ± 0.0649	0.1848 ± 0.0354	-0.1361 ± 0.0345	-0.2837 ± 0.1703	-0.0245 ± 0.1144	-0.1918 ± 0.0912
Do-PFN 2D	-0.2419 ± 0.0532	0.4711 ± 0.0466	-0.1941 ± 0.0615	-0.1038 ± 0.1680	0.2889 ± 0.2049	-0.1759 ± 0.1609
UWYK No-Anc	-0.7887 ± 0.0640	0.3415 ± 0.0330	0.0775 ± 0.0431	-1.2416 ± 0.1829	-0.2536 ± 0.1118	-0.5013 ± 0.0944
UWYK No-Anc 2D	-0.9272 ± 0.0739	0.0103 ± 0.0358	-0.3798 ± 0.0290	-1.2508 ± 0.2121	-0.4638 ± 0.1481	-0.9971 ± 0.1901
UWYK Anc	-1.0910 ± 0.0494	0.1598 ± 0.0243	-0.1829 ± 0.0325	-1.4673 ± 0.2005	-0.4253 ± 0.1200	-0.6910 ± 0.1024
UWYK Anc 2D	-1.0857 ± 0.0736	-0.0504 ± 0.0364	-0.5157 ± 0.0289	-1.3360 ± 0.2157	-0.4862 ± 0.1496	-1.0176 ± 0.1882
CausalPFN-C	-3.8387 ± 0.1688	-0.5666 ± 0.0829	-1.5436 ± 0.0788	-2.9809 ± 0.2365	-0.9017 ± 0.1125	-1.3009 ± 0.1298
CausalPFN-C 2D	-1.6910 ± 0.0704	-0.3248 ± 0.0463	-0.6344 ± 0.0271	-2.0666 ± 0.1799	-0.7026 ± 0.0888	-0.8750 ± 0.0809

Table 20: Negative log-likelihood (NLL) on IHDP and ACIC 2016 after MALC-T refinement of the treatment-effect density. Response densities are evaluated natively; $\mathcal{L}_{\tau}$ uses the MALC-T effect density, and $\mathcal{L}_{\mathrm{ATE}}$ uses its Wasserstein-barycenter aggregation. Lower is better. Results are mean ± standard error across 100 IHDP and 10 ACIC realizations.

Method	IHDP: $\mathcal{L}_{Y_0+Y_1}\downarrow$	IHDP: $\mathcal{L}_{\tau}\downarrow$	IHDP: $\mathcal{L}_{\mathrm{ATE}}\downarrow$	ACIC: $\mathcal{L}_{Y_0+Y_1}\downarrow$	ACIC: $\mathcal{L}_{\tau}\downarrow$	ACIC: $\mathcal{L}_{\mathrm{ATE}}\downarrow$
Do-PFN	-0.6100 ± 0.0649	0.1866 ± 0.0354	-0.1356 ± 0.0345	-0.2837 ± 0.1703	-0.0238 ± 0.1140	-0.1920 ± 0.0889
Do-PFN 2D	-0.2419 ± 0.0532	0.5148 ± 0.0505	-0.2115 ± 0.0678	-0.1038 ± 0.1680	0.3352 ± 0.2158	-0.1236 ± 0.1493
UWYK No-Anc	-0.7887 ± 0.0640	0.3492 ± 0.0327	0.0741 ± 0.0436	-1.2416 ± 0.1829	-0.2537 ± 0.1094	-0.5005 ± 0.0926
UWYK No-Anc 2D	-0.9272 ± 0.0739	0.0187 ± 0.0354	-0.3641 ± 0.0283	-1.2508 ± 0.2121	-0.4553 ± 0.1486	-0.9835 ± 0.1998
UWYK Anc	-1.0910 ± 0.0494	0.1682 ± 0.0242	-0.1867 ± 0.0326	-1.4673 ± 0.2005	-0.4238 ± 0.1179	-0.6908 ± 0.1015
UWYK Anc 2D	-1.0857 ± 0.0736	-0.0490 ± 0.0362	-0.4970 ± 0.0283	-1.3360 ± 0.2157	-0.4755 ± 0.1503	-1.0315 ± 0.1968
CausalPFN-C	-3.8387 ± 0.1688	-0.5690 ± 0.0828	-1.5447 ± 0.0789	-2.9809 ± 0.2365	-0.8992 ± 0.1119	-1.3005 ± 0.1301
CausalPFN-C 2D	-1.6910 ± 0.0704	-0.3524 ± 0.0496	-0.7052 ± 0.0283	-2.0666 ± 0.1799	-0.7503 ± 0.0905	-0.9473 ± 0.0820

J Computational Details

Some models were trained on different GPUs. Do-PFN and CausalPFN variants were trained on single NVIDIA H100 (80 GB, SXM5) GPUs. All UWYK variations were trained on NVIDIA RTX PRO 6000 Blackwell Server Edition (96 GB) GPUs. Table 21 reports the additional specifications per model to compute cost.

For evaluation on both raw and MALC variants, inference was done using a mix of single NVIDIA L40S GPUs (48 GB) per task, or Intel Granite Rapids Xeon CPUs (8 cores, 64GB at a time for raw model inference and 16 cores, 96GB at a time for MALC calirbated inference).

The implementation uses Python 3.11, PyTorch 2.11.0, NumPy 1.26.4, SciPy 1.15.2, scikit-learn 1.2.2, and XGBoost 3.2.0, the last of which enters results through the synthetic-prior mechanism sampler. MALC uses the implementation of Danisman et al. (2026).

Table 21: Training cost per baseline, ablation, and jointly supervised versions. Head output dimension counts the outputs per query. The grid logits plus the models' auxiliary outputs, such as tail and region parameters (+13 for the 2D head, Appendix B; +0, +4, and +10 for the Do-PFN, UWYK, and CausalPFN 1D heads). Parameter counts are deduplicated tensor counts of the checkpoints. Published models are not trained by us, so their training columns are not applicable.

Model	Variant	Grid	Head output dim	Total params	Throughput (s/step)	Steps	GPU-hours
Do-PFN	1D (published)	100	100	7,336,721	—	—	—
	2D	10^2	$10^2+13 = 113$	7,346,718	0.103 (H100)	150,000	4.30
	1D, matched resolution	10	10	7,267,241	0.093 (H100)	150,000	3.86
	1D, matched labels	100	100	7,336,721	0.15 (Blackwell)	150,000	6.3
UWYK	1D continuous (published)	1000	$1000+4 = 1004$	11,186,463	—	—	—
	1D binary (retrained)	1000	$1000+4 = 1004$	11,186,463	1.63 (H100)	50,000	22.7
	2D	32^2	$32^2+13 = 1037$	14,010,945	1.90 (H100)	50,000	26.4
	1D, matched resolution	32	$32+4 = 36$	10,937,687	1.71 (Blackwell)	50,000	23.8
	1D, matched labels	1000	$1000+4 = 1004$	11,186,463	1.69 (Blackwell)	50,000	23.5
CausalPFN	1D two-stage (published)	1024	$1024+10 = 1034$	18,844,034	—	—	—
	1D stage-2 (reproduction)	1024	$1024+10 = 1034$	18,844,034	2.57 (H100)	50,000	35.7
	2D	32^2	$32^2+13+10 = 1047$	18,852,002	1.96 (H100)	50,000	27.2
	1D, matched resolution	32	$32+10 = 42$	18,080,194	1.90 (H100)	50,000	26.4
	1D, matched labels	1024	$1024+10 = 1034$	18,844,034	2.60 (H100)	50,000	36.1