CausalPFN: Amortized Causal Effect Estimation via In-Context Learning

Vahid Balazadeh [*1] **Hamidreza Kamkari** [*2] **Valentin Thomas** [2] **Benson Li** [1] **Junwei Ma** [2] **Jesse C. Cresswell** [2] **Rahul G. Krishnan** [1]

Abstract

Causal effect estimation from observational data is fundamental in various applications. However, selecting an appropriate estimator from dozens of specialized methods demands substantial manual effort and domain expertise. We present CausalPFN, a single transformer that *amortizes* this workflow: trained once on a large library of simulated data-generating processes that satisfy ignorability, it infers causal effects for new observational datasets out-of-the-box. CausalPFN combines ideas from Bayesian causal inference with the large-scale training protocol of prior-fitted networks (PFNs), learning to map raw observations directly to causal-effects without any task-specific adjustment. Our approach achieves superior average performance on heterogeneous and average treatment effect estimation benchmarks (IHDP, Lalonde, ACIC). This ready-to-use model does not require any further training or tuning and takes a step toward automated causal inference.

1 Introduction

Causal inference—estimating the effects of interventions from data—is fundamental across numerous domains, including public policy, economics, and healthcare (Manski, 1993; Angrist & Pischke, 2014; Imbens & Rubin, 2015). The central challenge lies in estimating causal quantities from observational data: records collected without explicit interventions, where confounding factors can obscure true causal effects. Various causal identification settings have emerged to address this challenge (Angrist & Imbens, 1995; Angrist et al., 1996; Balke & Pearl, 1997; MacKinnon et al., 2007). Perhaps the most common one is to assume no unobserved confounding (ignorability) (Rubin, 1974).

[*]Equal contribution [1]Department of Computer Science, University of Toronto [2]Layer 6 AI, Toronto, Canada. Correspondence to: Vahid Balazadeh <vahid@cs.toronto.edu>, Hamidreza Kamkari <hamid@layer6.ai>.

Proceedings of the 42^{nd} International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025.

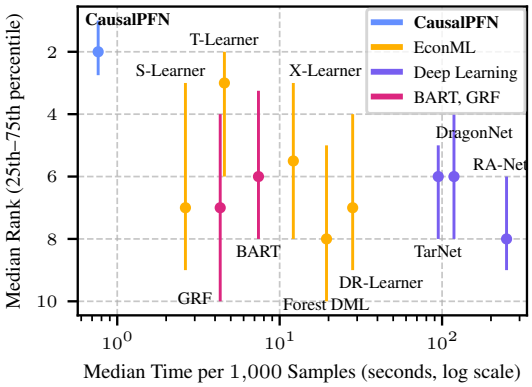

Figure 1. **Time vs. Performance.** Comparison across 130 causal inference tasks from IHDP, ACIC and Lalonde. CausalPFN achieves the best average rank (by precision in estimation of heterogeneous effect) while being orders of magnitude faster.

Even within the conceptually straightforward ignorability framework, researchers have developed dozens of specialized causal estimators over the past four decades. Prominent examples include Meta-Learners (Künzel et al., 2019), doubly robust methods (Funk et al., 2011; Kennedy, 2023), double machine learning (DML) (Chernozhukov et al., 2016; Foster & Syrgkanis, 2023), and neural network approaches (Shalit et al., 2017; Shi et al., 2019; Curth, 2021). This large number of estimators creates practical challenges as domain expertise is required to select, tune, or design the most appropriate estimator for each application (Shimoni et al., 2018; Schuler et al., 2018; Mahajan et al., 2024b).

The Bayesian paradigm offers an elegant framework to address these challenges (Rubin, 1978; Imbens & Rubin, 1997; 2015; Hill, 2011); rather than manually designing or selecting the best estimator, one can: (1) parameterize an appropriate prior distribution over plausible underlying causal mechanisms, i.e., the data-generating processes (DGPs), (2) define the causal estimand as a functional of the DGP parameters, (3) compute a posterior distribution over DGPs conditioned on data, and (4) derive the posterior-predictive distribution (PPD) of the causal estimand. However, the practical adoption of Bayesian methods remains limited. Computing posterior distributions typically requires expensive sampling methods (Oganisian & Roy, 2021), which often leads researchers to make specific assumptions about the DGPs or priors that are not necessarily reflective of the

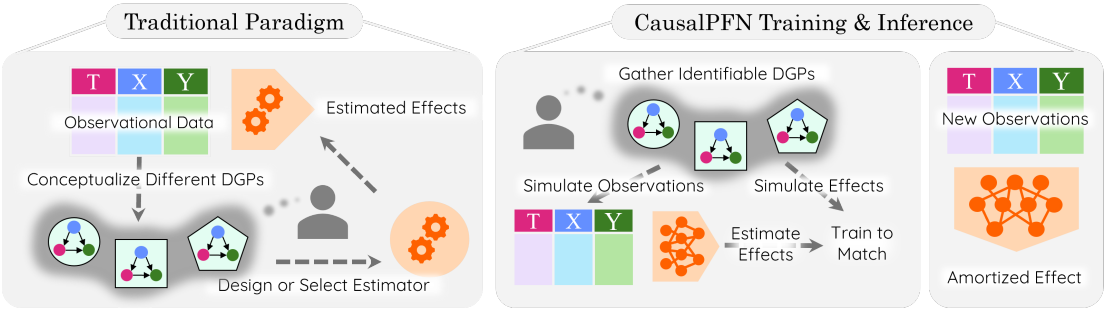

Figure 2. **Traditional Causal Inference vs. CausalPFN**. *(Left)*: A domain expert manually builds or selects an estimator for the given data. *(Right)*: The domain expert simulates diverse DGPs for pre-training, and a transformer learns to amortize causal inference automatically.

complexity of the downstream tasks (Li et al., 2023a).

Meanwhile, an emerging area in deep learning suggests using single large models that can arbitrarily approximate PPDs for different datasets by amortizing the expensive posterior inference (Garnelo et al., 2018b;a; Kim et al., 2019). A successful example is the prior-fitted network (PFN) (Müller et al., 2022) that achieved remarkable performance in tabular prediction tasks (Hollmann et al., 2023; Ma et al., 2024; Hollmann et al., 2025). PFNs employ transformer architectures trained on large-scale simulated DGPs to perform posterior-predictive inference via in-context learning; given a dataset of input-output examples as context, they can predict outputs for new inputs. PFNs shift the computational burden from inference time to training time, producing a single model that can generalize across diverse tasks. However, they are only designed for predictive tasks like regression and classification, not causal inference.

We propose to bridge the large-scale training of amortized models with Bayesian causal inference and introduce CausalPFN, a transformer-based model for causal effect estimation via in-context learning. Our framework leverages a general-purpose prior, based on the *ignorability* assumption, to generate a vast collection of simulated DGPs. By training on these diverse DGPs, our method learns to infer the causal estimands directly from observational data. While our approach requires expensive training, that can reportedly take up to seven days, once done, it is ready to use for new datasets. CausalPFN is an easy-to-use estimator with remarkably strong performance. It requires no further hyperparameter tuning or training for new tasks, in contrast to the bespoke models. Figure 1 illustrates the relative performance and efficiency of our method compared to standard baselines. Specifically, CausalPFN incurs no additional cost beyond inference time, whereas baseline methods require full pipelines, including training, hyperparameter tuning, and inference for every new dataset. We also show CausalPFN's workflow compared to traditional causal inference in Figure 2. Our main contribution is to show that, for the first time, a single transformer-based model trained on a diverse library of simulated DGPs can match or surpass specialized estimators across multiple datasets without task-specific tuning. Specifically, CausalPFN achieves superior average performance on IHDP, ACIC, and Lalonde benchmarks. Moreover, we release our model's weights with a user-friendly API, streamlining the adoption of CausalPFN as a capable estimator. CausalPFN is ready-to-use and does not require any further training or hyperparameter tuning.

2 Background

Causal Effect Estimation. We adopt the potential outcomes framework for causal inference (Rubin, 2005). Let T be the treatment assignment and $\mathbf{X}$ the observed covariates. For each $t \in \mathcal{T}$, Y_t denotes the potential outcome under treatment t; the realized (factual) outcome is $Y := Y_T$. Given observational samples $(\mathbf{X}, T, Y)$, a central goal is to recover the *conditional expected potential outcomes* (CEPOs)

$$\mu_t(\mathbf{x}) := \mathbb{E}[Y_t \mid \mathbf{X} = \mathbf{x}], \ \forall t \in \mathcal{T}. \tag{1}$$

For binary treatments, $\mathcal{T} = \{0, 1\}$, two widely-used causal estimands are the average treatment effect (ATE) and the conditional average treatment effect (CATE). Both can be derived through CEPOs:

$$\text{ATE}: \quad \lambda := \mathbb{E}[\mu_1(\mathbf{X}) - \mu_0(\mathbf{X})], \tag{2}$$

$$\text{CATE}: \quad \tau(\mathbf{x}) := \mu_1(\mathbf{x}) - \mu_0(\mathbf{x}). \tag{3}$$

We refer to CEPOs, CATE, and ATE collectively as *causal effects*. Recovering causal effects from observational data is impossible without further assumptions: different DGPs can induce the same distribution over $(\mathbf{X}, T, Y)$ but have different causal effects (Pearl, 2009; Hernán & Robins, 2010; Imbens & Rubin, 2015). Throughout this work, we assume *strong ignorability*, a standard assumption that makes estimation possible. Strong ignorability posits that, conditional on observed covariates, treatment assignment has positive probability for all $t \in \mathcal{T}$ and is independent of all potential outcomes (Rubin, 1974; Rosenbaum & Rubin, 1983); a precise statement is given in Appendix A.

Bayesian Causal Inference. A Bayesian formulation of causal inference considers an explicit likelihood model for the observed data and the unobserved poten-

tial outcomes (Rubin, 1978; Hahn et al., 2020). Let ψ be the parameter that indexes the joint distribution $\mathrm{P}^{\psi}(\mathbf{X}, T, \{Y_t\}_{t\in\mathcal{T}}, Y)$. A prior $\pi(\psi)$ encodes domain knowledge on parameters ψ. Given i.i.d. observations $\mathcal{D}_{\text{obs}} = \left\{(\mathbf{x}^{(n)}, t^{(n)}, y^{(n)})\right\}_{n=1}^{N}$ coming from the observational distribution $\mathrm{P}^{\psi}_{\text{obs}}$, Bayes' rule yields the posterior $\pi(\psi \mid \mathcal{D}_{\text{obs}})$. For any functional $g(\psi)$—for example $g(\psi) = \mathbb{E}^{\psi}[Y_1 - Y_0]$ for ATE—the posterior predictive distribution (PPD) $\pi^{g}(\cdot \mid \mathcal{D}_{\text{obs}}) := \int \mathbb{I}(g(\psi) \in \cdot)\, \mathrm{d}\pi(\psi \mid \mathcal{D}_{\text{obs}})$ is induced by the posterior distribution $\pi(\psi \mid \mathcal{D}_{\text{obs}})$. Point estimates and credible intervals therefore arise automatically from these induced posteriors. Because the posterior is rarely available in closed form, one resorts to approximate inference such as Markov-chain Monte-Carlo (MCMC) (Hill, 2011). The Bayesian paradigm offers a unified framework for inference on causal estimands with automatic uncertainty quantification.

3 CausalPFN

Objective. We adopt the Bayesian paradigm for causal inference, as discussed above. Our primary estimands of interest are the CEPOs $\mu_t(\mathbf{x}; \psi)$ from (1). As shown in (2) and (3), CEPOs directly enable estimation of both ATE and CATEs. We focus on developing an estimator that can accurately infer these quantities from observational data.

Given a suitably rich prior distribution π over the DGPs, which we will explicitly design in this section, we define our target as the posterior-predictive distribution of CEPOs:

Definition 1 (CEPO-PPD)**.** *For each $t \in \mathcal{T}$ and covariate vector $\mathbf{x}$, the* CEPO-PPD *is* $\pi^{\mu_t}(\cdot \mid \mathbf{x}, \mathcal{D}_{obs}) := \int \mathbb{I}(\mu_t(\mathbf{x}; \psi) \in \cdot)\, \mathrm{d}\pi(\psi \mid \mathcal{D}_{obs})$.

Running a new posterior inference for every dataset is computationally demanding. Recent work shows that in-context transformers can *amortize* Bayesian prediction: instead of sampling from the posterior at test time, a single network is trained to map a context set directly to the PPDs (Garnelo et al., 2018a;b; Müller et al., 2022). Inspired by those, we amortize the entire posterior-predictive inference using a single transformer model q_θ that approximates the CEPO-PPDs, that is $\pi^{\mu_t}(\cdot \mid \mathbf{x}, \mathcal{D}_{\text{obs}}) \approx q_\theta(\cdot \mid \mathbf{x}, t, \mathcal{D}_{\text{obs}})$, without requiring to compute $\pi(\psi \mid \mathcal{D}_{\text{obs}})$. To train, we introduce the following loss function:

Definition 2 (Causal Data-Prior Loss)**.** *For any $t \in \mathcal{T}$, we define the causal data-prior loss as* $\mathcal{L}_t(\theta) := \mathbb{E}_{\psi\sim\pi,\ \mathcal{D}_{obs}\cup\{\mathbf{x}\}\ \sim\ \mathrm{P}^{\psi}_{obs}}[-\log q_\theta(\mu_t(\mathbf{x}; \psi) \mid \mathbf{x}, t, \mathcal{D}_{obs})]$.

This loss function enables a fundamental shift in computational paradigm; rather than computing the posterior distribution at inference, we transfer the computation burden to training. By minimizing the causal data-prior loss, the model q_θ learns to map the observational data to the corresponding predictive distribution directly, without ever explicitly computing the posterior. Given observational data $\mathcal{D}_{\text{obs}}$ from an underlying $\psi^\star$, a natural point estimate for CEPOs is the expectation of the predicted CEPO-PPD, that is, $\mathbb{E}_{\mu\sim q_\theta}[\mu \mid \mathbf{x}, t, \mathcal{D}_{\text{obs}}] \approx \mu_t(\mathbf{x}; \psi^\star)$. These CEPO estimates can also form point estimates for CATEs using (3), and for ATEs using (2) by empirical averaging across units in $\mathcal{D}_{\text{obs}}$. We provide a theoretical justification for the causal data-prior loss in Appendix B.

A Scalable Causal Prior. Here, we focus on designing an appropriate prior π for the causal data-prior loss in Definition 2. This prior must balance two factors: First, it should contain a rich set of DGPs with sufficient coverage to approximate real-world scenarios. *Second, all DGPs in our prior must satisfy strong ignorability*. Otherwise, the resulting model cannot distinguish between DPGs with different causal effects but similar observational data.

To address these requirements, we develop a procedure that can transform *any* base table from standard tabular priors into a valid causal dataset: ***(i)*** retrieve a base table with N rows from either a large library of tabular data or synthesize it; ***(ii)*** randomly select columns with a varying number of covariates as $\mathbf{X}$; ***(iii)*** pick two other columns, relabel them as $\mu_0(\mathbf{X}), \mu_1(\mathbf{X})$; ***(iv)*** add zero-mean noise to $\mu_0(\mathbf{X}), \mu_1(\mathbf{X})$ and obtain Y_0, Y_1—these four steps simulate samples from a joint $(\mathbf{X}, Y_0, Y_1)$; ***(v)*** generate a random function f to map covariates to their treatment logits; ***(vi)*** sample binary treatments $T \sim \text{Bernoulli}(\text{Sigmoid}(f(\mathbf{X})))$; ***(vii)*** finally, form the observed outcomes $Y := Y_T$. This approach guarantees strong ignorability *by design*: since treatment T is determined solely from $\mathbf{X}$, it is conditionally independent from the potential outcomes Y_0, Y_1. Also, by applying the sigmoid function, we ensure $0 < P(T \mid \mathbf{X})$, satisfying positivity. For the diversity aspect of π, we rely on sampling covariates directly from a mix of real and synthetic tables, which yields data that is more likely to reflect the scenarios the model will face at inference. Appendix C details additional mechanisms for controlling treatment effect heterogeneity and positivity, as well as the detailed configurations of the prior-generation process.

Model Architecture & Parallel Training. We model q_θ using a PFN-style transformer encoder that receives a sequence of row tokens as *context* (i.e., $\mathcal{D}_{\text{obs}}$), where each token embeds a triplet $(t^{(n)}, \mathbf{x}^{(n)}, y^{(n)})$. At every iteration, we embed B_Q batched *query* tokens $(t, \mathbf{x})$. We then apply 20 layers of self-attention and MLP layers, followed by a final projection layer to get $q_\theta(\cdot \mid \mathbf{x}, t, \mathcal{D}_{\text{obs}})$ for all the $(t, \mathbf{x})$ pairs in the batched query. The transformer uses the asymmetric masking used in PFNs: both context and query tokens attend only to the context tokens, ensuring that the predicted CEPO-PPDs are mutually independent.

To model CEPO-PPDs, we approximate each with a quan-

Table 1. **CATE & ATE results.** Columns correspond to benchmark suites: IHDP, ACIC 2016, Lalonde CPS/PSID. *(left half)* mean PEHE and the average rank when pooling all tasks. *(right half)* mean ATE relative error and its average across all tasks. Lalonde PEHE is in thousands. Top-three per column are **green**, **blue**, and **orange**. Cells with "—" indicate that the method is not applicable or has diverged.

Method	Mean PEHE ± Standard Error (↓ better)					Mean ATE Relative Error ± Standard Error (↓ better)				
	IHDP	ACIC 2016	Lalonde CPS ($\times 10^3$)	Lalonde PSID ($\times 10^3$)	Avg. Rank	IHDP	ACIC 2016	Lalonde CPS	Lalonde PSID	Avg. Error
CausalPFN	**0.58±0.07**	**0.92±0.11**	**8.83±0.04**	**13.98±0.43**	**2.22±0.16**	0.21±0.04	**0.04±0.01**	**0.08±0.02**	**0.20±0.03**	**0.18±0.03**
T-Learner	**1.90±0.34**	1.03±0.08	**9.21±0.09**	**13.43±0.42**	**4.16±0.25**	0.21±0.04	**0.04±0.01**	**0.27±0.03**	**0.03±0.01**	**0.19±0.03**
X-Learner	2.83±0.46	**0.85±0.14**	12.19±0.42	20.37±0.78	**5.78±0.27**	0.19±0.03	**0.03±0.01**	0.86±0.07	0.70±0.08	0.27±0.03
BART	2.50±0.38	**0.68±0.10**	12.78±0.11	20.86±0.43	5.99±0.25	0.48±0.11	**0.04±0.01**	1.01±0.02	0.83±0.03	0.52±0.09
DragonNet	2.16±0.25	2.11±0.18	**10.31±0.29**	**15.64±0.53**	6.08±0.22	0.22±0.03	0.07±0.02	0.52±0.08	0.40±0.07	0.24±0.03
S-Learner	3.45±0.60	1.19±0.15	12.77±0.09	22.00±0.48	6.37±0.29	0.20±0.04	**0.05±0.01**	1.01±0.02	0.95±0.02	0.31±0.04
TarNet	**1.80±0.14**	2.20±0.20	—	18.93±0.42	6.64±0.17	0.23±0.04	0.07±0.02	—	0.76±0.04	0.26 ±0.03
GRF	3.67±0.60	1.32±0.28	12.40±0.19	22.39±0.45	6.68±0.28	**0.18±0.03**	0.07±0.02	0.81±0.06	0.80±0.05	0.52±0.09
DR-Learner	3.45±0.54	1.09±0.15	20.05±1.88	33.99±8.71	6.74±0.27	**0.16±0.03**	0.06±0.02	1.68±0.67	0.75±0.07	0.31±0.07
Forest DML	4.31±0.70	1.42±0.29	171.0±50.6	22.66±0.51	7.35±0.28	**0.09±0.01**	**0.04±0.01**	2.31±0.66	0.96±0.03	0.32±0.08
RA-Net	2.35±0.19	2.35±0.24	11.50±0.31	16.95±0.78	7.57±0.19	0.23±0.03	0.07±0.02	0.75±0.06	0.44±0.05	0.27±0.03
IPW	—	—	—	—	—	0.23±0.04	0.24±0.05	**0.25±0.03**	**0.05±0.01**	**0.22 ±0.03**

tized histogram. We discretize the outcome axis into $L = 1024$ bins and let the network project the query tokens into L logits. We then apply SoftMax to turn those into a quantized distribution $q_\theta(\cdot \mid \mathbf{x}, t, \mathcal{D}_{\text{obs}})[\ell]$, for $\ell \in [L]$. At each round of gradient update, we place a Gaussian kernel with a small σ at the true CEPO $\mu_t(\mathbf{x})$ and integrate it over bins to obtain Gaussian quantized probabilities $\mathrm{P}^{\text{Gauss}}_{\mu_t(\mathbf{x})}[\ell]$ and minimize the *histogram loss*: $\texttt{HL}\big[\mu_t(\mathbf{x}) \,\|\, q_\theta\big] = -\sum_{\ell=1}^{L} \mathrm{P}^{\text{Gauss}}_{\mu_t(\mathbf{x})}[\ell] \log q_\theta[\ell]$. This loss is an approximate form of the causal data-prior loss, which matches in the limit $\sigma \to 0$ and $L \to \infty$. The histogram loss affords a tractable proxy for the continuous CEPO-PPD. See Appendix D for the complete parallel training pipeline.

4 Experiments

Baseline Estimators. We compare to a broad suite of baselines. This includes double machine learning (DML) with causal forests, doubly robust learner (DR-Learner), as well as the T-, S-, and X-Learners, all part of the `EconML` package (Battocchi et al., 2019). Moreover, we include deep-learning–based methods such as TarNet, DragonNe, and RA-Net, implemented via the `CATENets` library (Curth, 2021). Finally, we compare to inverse propensity weighting (IPW) (Rosenbaum & Rubin, 1983), Bayesian regression trees (BART) (Hill, 2011), and generalized random forests (GRF) (Athey et al., 2019). All the baselines, except for IPW, provide both CATE and ATE estimates. *We tune most of the baselines with cross-validation via grid search. The set of hyperparameter, along with the results with default hyperparameters are all detailed in Appendix E.*

Results. We report the relative error for ATE, and the precision in estimation of heterogeneous effects (PEHE) for CATE, defined as the root mean squared deviation between predicted and true CATEs (Hill, 2011). Table 1 compares CausalPFN to all baselines on four standard set of datasets: 100 realizations of IHDP (Ramey et al., 1992; Hill, 2011), 10 realizations of ACIC 2016 (Dorie et al., 2019), and the Lalonde CPS and Lalonde PSID cohorts (LaLonde, 1986) with their causal effects provided by `RealCause` (Neal et al., 2020) (we use the first 10 realizations). Our model demonstrates superior performance on both CATE and ATE tasks, remaining within the top three models across all the benchmarks, except for the ATE on IHDP datasets. To assess the overall performance of each method on CATE, we calculate the average rank of each method (based on PEHE) on all 130 realizations, as PEHEs are not standardized across different datasets. For ATEs, we average the relative errors directly. CausalPFN outperforms all the baselines on both average metrics. Notably, unlike other methods that are trained directly on the target datasets, our model is trained entirely on simulated data and *never* sees the evaluation data. Additional results on marketing datasets and uncertainty quantification are provided in Appendices F and G.

5 Conclusions & Limitations

In this paper, we introduced a practical paradigm for amortized causal effect estimation that combines Bayesian causal inference with large-scale tabular training. Despite learning solely from simulated data, CausalPFN matches, and often outperforms, specialized causal estimators across diverse real-world domains. Through amortization, we significantly reduce the burden of estimator selection at inference time, and to foster adoption, we will open-source the model.

That said, several important limitations remain: ***(i)*** We fundamentally assume strong ignorability, which is an untestable assumption. Without this condition, CausalPFN has no guarantees of validity. Domain expertise remains essential to determine whether this method is appropriate or whether alternative approaches should be employed. ***(ii)*** While CausalPFN already supports multi-arm discrete treatments, extending amortized inference to continuous treatments remains unexplored. ***(iii)*** Finally, extending our framework to richer domain-informed priors like instrumental variables can broaden the framework's reach.

References

Akyürek, E., Schuurmans, D., Andreas, J., Ma, T., and Zhou, D. What learning algorithm is in-context learning? investigations with linear models. In *The Eleventh International Conference on Learning Representations*, 2023.

Angrist, J. and Imbens, G. Identification and estimation of local average treatment effects, 1995.

Angrist, J. D. and Pischke, J.-S. *Mastering 'Metrics: The path from cause to effect*. Princeton University Press, 2014.

Angrist, J. D., Imbens, G. W., and Rubin, D. B. Identification of causal effects using instrumental variables. *Journal of the American statistical Association*, 91(434): 444–455, 1996.

Athey, S., Tibshirani, J., and Wager, S. Generalized random forests. *The Annals of Statistics*, 47(2):1148–1178, 2019. doi: 10.1214/18-AOS1709.

Bai, Y., Chen, F., Wang, H., Xiong, C., and Mei, S. Transformers as statisticians: Provable in-context learning with in-context algorithm selection. In *Advances in Neural Information Processing Systems*, volume 36, pp. 57125–57211, 2023.

Balke, A. and Pearl, J. Bounds on treatment effects from studies with imperfect compliance. *Journal of the American statistical Association*, 92(439):1171–1176, 1997.

Battocchi, K., Dillon, E., Hei, M., Lewis, G., Oka, P., Oprescu, M., and Syrgkanis, V. EconML: A Python Package for ML-Based Heterogeneous Treatment Effects Estimation, 2019. URL https://github.com/py-why/EconML. Version 0.15.0.

Bischl, B., Casalicchio, G., Feurer, M., Gijsbers, P., Hutter, F., Lang, M., Gomes Mantovani, R., van Rijn, J., and Vanschoren, J. OpenML benchmarking suites. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, 2021.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pp. 1877–1901, 2020.

Bynum, L. E., Puli, A. M., Herrero-Quevedo, D., Nguyen, N., Fernandez-Granda, C., Cho, K., and Ranganath, R. Black box causal inference: Effect estimation via meta prediction. *arXiv:2503.05985*, 2025.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. Double/debiased machine learning for treatment and causal parameters. *arXiv:1608.00060*, 2016.

Coda-Forno, J., Binz, M., Akata, Z., Botvinick, M., Wang, J., and Schulz, E. Meta-in-context learning in large language models. *Advances in Neural Information Processing Systems*, 36:65189–65201, 2023.

Curth, A. CATENets: Sklearn-style Implementations of Neural Network-based Conditional Average Treatment Effect (CATE) Estimators. https://github.com/AliciaCurth/CATENets, 2021. GitHub repository, commit 821bfb0. Accessed: 2025-05-11.

Dai, D., Sun, Y., Dong, L., Hao, Y., Ma, S., Sui, Z., and Wei, F. Why can GPT learn in-context? language models secretly perform gradient descent as meta-optimizers. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 4005–4019. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.findings-acl.247.

Defazio, A., Yang, X., Khaled, A., Mishchenko, K., Mehta, H., and Cutkosky, A. The road less scheduled. *Advances in Neural Information Processing Systems*, 37: 9974–10007, 2024.

Dong, Q., Li, L., Dai, D., Zheng, C., Ma, J., Li, R., Xia, H., Xu, J., Wu, Z., Chang, B., et al. A survey on in-context learning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 1107–1128, 2024.

Doob, J. L. Application of the theory of martingales. *Le calcul des probabilites et ses applications*, pp. 23–27, 1949.

Dorie, V., Hill, J., Shalit, U., Scott, M., and Cervone, D. Automated versus do-it-yourself methods for causal inference: Lessons learned from a data analysis competition. *Statistical Science*, 34(1):43–68, 2019.

Fischer, S. F., Feurer, M., and Bischl, B. OpenML-CTR23 – A curated tabular regression benchmarking suite. In *AutoML Conference (Workshop)*, 2023.

Foster, D. J. and Syrgkanis, V. Orthogonal statistical learning. *The Annals of Statistics*, 51(3):879–908, 2023.

Funk, M. J., Westreich, D., Wiesen, C., Stürmer, T., Brookhart, M. A., and Davidian, M. Doubly robust estimation of causal effects. *American Journal of Epidemiology*, 173(7):761–767, 2011.

Garnelo, M., Rosenbaum, D., Maddison, C., Ramalho, T., Saxton, D., Shanahan, M., Teh, Y. W., Rezende, D., and Eslami, S. M. A. Conditional neural processes. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pp. 1704–1713, 2018a.

Garnelo, M., Schwarz, J., Rosenbaum, D., Viola, F., Rezende, D. J., Eslami, S., and Teh, Y. W. Neural processes. *arXiv:1807.01622*, 2018b.

Gijsbers, P., Bueno, M. L., Coors, S., LeDell, E., Poirier, S., Thomas, J., Bischl, B., and Vanschoren, J. AMLB: An AutoML benchmark. *Journal of Machine Learning Research*, 25(101):1–65, 2024.

Grinsztajn, L., Oyallon, E., and Varoquaux, G. Why do tree-based models still outperform deep learning on typical tabular data? In *Advances in Neural Information Processing Systems*, 2022.

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. On calibration of modern neural networks. In *International Conference on Machine Learning*, pp. 1321–1330. PMLR, 2017.

Hahn, P. R., Murray, J. S., and Carvalho, C. M. Bayesian regression tree models for causal inference: Regularization, confounding, and heterogeneous effects (with discussion). *Bayesian Analysis*, 15(3):965–1056, 2020.

Henry, A., Dachapally, P. R., Pawar, S., and Chen, Y. Query-key normalization for transformers. *arXiv preprint arXiv:2010.04245*, 2020.

Hernán, M. A. and Robins, J. M. Causal inference, 2010.

Hill, J. L. Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20(1):217–240, 2011.

Hillstrom, K. Minethatdata e-mail analytics and data mining challenge dataset. https://blog.minethatdata.com/2008/03/minethatdata-e-mail-analytics-and-data.html, 2008. Accessed: 2025-05-11.

Hollmann, N., Müller, S., Eggensperger, K., and Hutter, F. TabPFN: A transformer that solves small tabular classification problems in a second. In *The Eleventh International Conference on Learning Representations*, 2023.

Hollmann, N., Müller, S., Purucker, L., Krishnakumar, A., Körfer, M., Hoo, S. B., Schirrmeister, R. T., and Hutter, F. Accurate predictions on small data with a tabular foundation model. *Nature*, 637(8045):319–326, 2025.

Imbens, G. W. and Rubin, D. B. Bayesian inference for causal effects in randomized experiments with noncompliance. *The annals of statistics*, pp. 305–327, 1997.

Imbens, G. W. and Rubin, D. B. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, 2015.

Kamkari, H., Balazadeh, V., Zehtab, V., and Krishnan, R. G. Order-based structure learning with normalizing flows. *arXiv:2308.07480*, 2023.

Ke, N. R., Chiappa, S., Wang, J. X., Bornschein, J., Goyal, A., Rey, M., Weber, T., Botvinick, M., Mozer, M. C., and Rezende, D. J. Learning to induce causal structure. In *International Conference on Learning Representations*, 2023.

Kennedy, E. H. Towards optimal doubly robust estimation of heterogeneous causal effects. *Electronic Journal of Statistics*, 17(2):3008–3049, 2023.

Khemakhem, I., Monti, R., Leech, R., and Hyvarinen, A. Causal autoregressive flows. In *International Conference on Artificial Antelligence and Statistics*, pp. 3520–3528. PMLR, 2021.

Kim, H., Mnih, A., Schwarz, J., Garnelo, M., Eslami, A., Rosenbaum, D., Vinyals, O., and Teh, Y. W. Attentive neural processes. In *International Conference on Learning Representations*, 2019.

Kingma, D. P. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

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

LaLonde, R. J. Evaluating the econometric evaluations of training programs with experimental data. *The American Economic Review*, pp. 604–620, 1986.

Lenta LLC. Lenta uplift dataset. https://github.com/maks-sh/scikit-uplift, 2020. Accessed: 2025-05-11.

Li, F., Ding, P., and Mealli, F. Bayesian causal inference: a critical review. *Philosophical Transactions of the Royal Society A*, 381(2247):20220153, 2023a.

Li, Y., Ildiz, M. E., Papailiopoulos, D., and Oymak, S. Transformers as algorithms: Generalization and stability in in-context learning. In *International Conference on Machine Learning*, pp. 19565–19594. PMLR, 2023b.

Lorch, L., Sussex, S., Rothfuss, J., Krause, A., and Schölkopf, B. Amortized inference for causal structure learning. *Advances in Neural Information Processing Systems*, 35:13104–13118, 2022.

Louizos, C., Shalit, U., Mooij, J. M., Sontag, D., Zemel, R., and Welling, M. Causal effect inference with deep latent-variable models. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Ma, J., Thomas, V., Hosseinzadeh, R., Kamkari, H., Labach, A., Cresswell, J. C., Golestan, K., Yu, G., Volkovs, M., and Caterini, A. L. TabDPT: Scaling Tabular Foundation Models. *arXiv:2410.18164*, 2024.

MacKinnon, D. P., Fairchild, A. J., and Fritz, M. S. Mediation analysis. *Annu. Rev. Psychol.*, 58(1):593–614, 2007.

Mahajan, D., Gladrow, J., Hilmkil, A., Zhang, C., and Scetbon, M. Zero-shot learning of causal models. *arXiv:2410.06128*, 2024a.

Mahajan, D., Mitliagkas, I., Neal, B., and Syrgkanis, V. Empirical analysis of model selection for heterogeneous causal effect estimation. In *The Twelfth International Conference on Learning Representations*, 2024b.

Maksim Shevchenko, I. E. User guide for uplift modeling and casual inference. https://www.uplift-modeling.com/en/latest/user_guide/index.html, 2020.

Manski, C. F. Identification problems in the social sciences. *Sociological methodology*, pp. 1–56, 1993.

McElfresh, D., Khandagale, S., Valverde, J., Prasad C, V., Ramakrishnan, G., Goldblum, M., and White, C. When do neural nets outperform boosted trees on tabular data? In *Advances in Neural Information Processing Systems*, 2023.

Megafon PJSC. Megafon uplift dataset. https://github.com/maks-sh/scikit-uplift, 2020. Accessed: 2025-05-11.

Miller, J. W. A detailed treatment of Doob's theorem. *arXiv:1801.03122*, 2018.

Müller, S., Hollmann, N., Arango, S. P., Grabocka, J., and Hutter, F. Transformers Can Do Bayesian Inference. In *International Conference on Learning Representations*, 2022.

Neal, B., Huang, C.-W., and Raghupathi, S. Realcause: Realistic causal inference benchmarking. *arXiv:2011.15007*, 2020.

Nilforoshan, H., Moor, M., Roohani, Y., Chen, Y., Šurina, A., Yasunaga, M., Oblak, S., and Leskovec, J. Zero-shot causal learning. *Advances in Neural Information Processing Systems*, 36:6862–6901, 2023.

Oganisian, A. and Roy, J. A. A practical introduction to bayesian estimation of causal effects: Parametric and nonparametric approaches. *Statistics in Medicine*, 40(2): 518–551, 2021.

Olsson, C., Elhage, N., Nanda, N., Joseph, N., DasSarma, N., Henighan, T., Mann, B., Askell, A., Bai, Y., Chen, A., Conerly, T., Drain, D., Ganguli, D., Hatfield-Dodds, Z., Hernandez, D., Johnston, S., Jones, A., Kernion, J., Lovitt, L., Ndousse, K., Amodei, D., Brown, T., Clark, J., Kaplan, J., McCandlish, S., and Olah, C. In-context learning and induction heads. *arXiv:2209.11895*, 2022.

Ovadia, Y., Fertig, E., Ren, J., Nado, Z., Sculley, D., Nowozin, S., Dillon, J., Lakshminarayanan, B., and Snoek, J. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Pearl, J. *Causality*. Cambridge University Press, 2009.

Peters, J., Mooij, J. M., Janzing, D., and Schölkopf, B. Causal discovery with continuous additive noise models. *The Journal of Machine Learning Research*, 15(1):2009–2053, 2014.

Peters, J., Janzing, D., and Schölkopf, B. *Elements of causal inference: Foundations and learning algorithms*. The MIT Press, 2017.

Peyrard, M. and Cho, K. Meta-statistical learning: Supervised learning of statistical inference. *arXiv preprint arXiv:2502.12088*, 2025.

Qu, J., Holzmüller, D., Varoquaux, G., and Morvan, M. L. TabICL: A Tabular Foundation Model for In-Context Learning on Large Data. *arXiv:2502.05564*, 2025.

Radcliffe, N. Using control groups to target on predicted lift: Building and assessing uplift models. Technical report, Stochastic Solutions, 2007.

Ramey, C. T., Bryant, D. M., Wasik, B. H., Sparling, J. J., Fendt, K. H., and La Vange, L. M. Infant health and development program for low birth weight, premature infants: Program elements, family participation, and child intelligence. *Pediatrics*, 89(3):454–465, 1992.

Retail Hero. Retail hero (x5) uplift dataset. https://github.com/maks-sh/scikit-uplift, 2020. Accessed: 2025-05-11.

Rosenbaum, P. R. and Rubin, D. B. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

Rubin, D. B. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688, 1974.

Rubin, D. B. Bayesian inference for causal effects: The role of randomization. *The Annals of Statistics*, pp. 34–58, 1978.

Rubin, D. B. Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, 100(469):322–331, 2005.

Scetbon, M., Jennings, J., Hilmkil, A., Zhang, C., and Ma, C. A fixed-point approach for causal generative modeling. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 43504–43541, 2024.

Schuler, A., Baiocchi, M., Tibshirani, R., and Shah, N. A comparison of methods for model selection when estimating individual treatment effects. *arXiv preprint arXiv:1804.05146*, 2018.

Shalit, U., Johansson, F. D., and Sontag, D. Estimating individual treatment effect: generalization bounds and algorithms. In *International conference on machine learning*, pp. 3076–3085. PMLR, 2017.

Shazeer, N. Glu variants improve transformer. *arXiv preprint arXiv:2002.05202*, 2020.

Shi, C., Blei, D., and Veitch, V. Adapting neural networks for the estimation of treatment effects. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Shimoni, Y., Yanover, C., Karavani, E., and Goldschmnidt, Y. Benchmarking framework for performance-evaluation of causal inference analysis. *arXiv preprint arXiv:1802.05046*, 2018.

Thomas, V., Ma, J., Hosseinzadeh, R., Golestan, K., Yu, G., Volkovs, M., and Caterini, A. L. Retrieval & fine-tuning for in-context tabular models. *Advances in Neural Information Processing Systems*, 37:108439–108467, 2024.

Vetter, J., Gloeckler, M., Gedon, D., and Macke, J. H. Effortless, simulation-efficient bayesian inference using tabular foundation models. *arXiv preprint arXiv:2504.17660*, 2025.

Von Oswald, J., Niklasson, E., Randazzo, E., Sacramento, J., Mordvintsev, A., Zhmoginov, A., and Vladymyrov, M. Transformers learn in-context by gradient descent. In *International Conference on Machine Learning*, pp. 35151–35174. PMLR, 2023.

von Oswald, J., Schlegel, M., Meulemans, A., Kobayashi, S., Niklasson, E., Zucchet, N., Scherrer, N., Miller, N., Sandler, M., y Arcas, B. A., Vladymyrov, M., Pascanu, R., and Sacramento, J. Uncovering mesa-optimization algorithms in transformers. *arXiv:2309.05858*, 2023.

Xia, K., Lee, K.-Z., Bengio, Y., and Bareinboim, E. The causal-neural connection: Expressiveness, learnability, and inference. *Advances in Neural Information Processing Systems*, 34:10823–10836, 2021.

Xia, K., Pan, Y., and Bareinboim, E. Neural causal models for counterfactual identification and estimation. *arXiv preprint arXiv:2210.00035*, 2022.

Xie, S. M., Raghunathan, A., Liang, P., and Ma, T. An explanation of in-context learning as implicit bayesian inference. In *International Conference on Learning Representations*, 2022.

Yadlowsky, S., Doshi, L., and Tripuraneni, N. Pretraining data mixtures enable narrow model selection capabilities in transformer models. *arXiv:2311.00871*, 2023.

Zhang, J., Jennings, J., Hilmkil, A., Pawlowski, N., Zhang, C., and Ma, C. Towards causal foundation model: on duality between causal inference and attention. *arXiv:2310.00809*, 2023.

Zheng, X., Aragam, B., Ravikumar, P. K., and Xing, E. P. DAGs with NO TEARS: Continuous Optimization for Structure Learning. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Émilie Diemert, Teytaud, O., Oblé, G., and Meynet, F. A large scale benchmark for uplift modeling. In *Proceedings of the AdKDD Workshop*, 2018. URL https://www.cs.cornell.edu/~diemert/criteo-uplift-dataset/.

Appendix Contents

A Assumptions for Causal Effect Estimation

Let P be the joint distribution on the covariates $\mathbf{X}$, treatment T, potential outcomes $\{Y_t\}_{t\in\mathcal{T}}$, and the observed outcome Y. Denote the observational distribution of the variables $(\mathbf{X}, T, Y)$ by P_{obs}. Recall that our main quantities of interests are conditional expected potential outcomes (CEPOs), defined as $\mu_t(\mathbf{x}) = \mathbb{E}\left[Y_t \mid \mathbf{X} = \mathbf{x}\right]$.

To make estimating CEPOs from observational data feasible, we make the following assumption:

Assumption 0. [Strong Ignorability & SUTVA]

1. (Unconfoundedness or Ignorability) $Y_t \perp\!\!\!\perp T \mid \mathbf{X}$ for all $t \in \mathcal{T}$,
2. (Positivity or Overlap) $\mathrm{P}(T = t \mid \mathbf{X}) > 0$ a.e. for all $t \in \mathcal{T}$,
3. (Consistency) $Y = Y_t$ if $T = t$ for all $t \in \mathcal{T}$,
4. (No-Interference) The treatment assigned to one unit does not affect the outcomes of other units.

Under the above assumption, it can be proven that $\mathrm{P}^1 \neq \mathrm{P}^2$ implies $\mathrm{P}^1_{\text{obs}} \neq \mathrm{P}^2_{\text{obs}}$ (Peters et al., 2017). In particular, CEPOs can be re-written as conditional expectations:

$$\mu_t(\mathbf{x}) = \mathbb{E}\left[Y_t \mid \mathbf{X} = \mathbf{x}\right] = \mathbb{E}\left[Y \mid \mathbf{X} = \mathbf{x}, T = t\right]. \tag{4}$$

B Asymptotic Justification of CausalPFN

In this section, we provide a theoretical motivation for our approach and the causal data-prior loss function by providing an asymptotic consistency result. We first provide an informal statement and give a proof sketch to develop some intuition.

B.1 Informal Statement and Proof

Proposition 1 (Informal)**.** *Assume **(i)** the model is sufficiently expressive, i.e., there exists parameters $\theta^\star$ that attain the **global** minimum of $\mathcal{L}_t(\theta)$ for every $t \in \mathcal{T}$, and **(ii)** the support of the prior π only consists of DGPs that **satisfy strong ignorability**. Then, under mild regularity assumptions, for almost all $\psi^\star \sim \pi$, and any i.i.d. set of observational samples $\mathcal{D}_{obs} \sim \mathrm{P}^{\psi^\star}_{obs}$ that $|\mathcal{D}_{obs}| \to \infty$, we have*

$$\mathbb{E}_{\mu\sim q_{\theta^\star}}\left[\mu \,\middle|\, \mathbf{x}, t, \mathcal{D}_{obs}\right] \xrightarrow{a.s.} \mu_t(\mathbf{x}\,;\,\psi^\star) \quad \textit{for almost every} \quad \mathbf{x} \sim \mathrm{P}^{\psi^\star}(\mathbf{X}). \tag{5}$$

Although the theorem assumes the idealized conditions $\psi^\star \sim \pi$ and a globally optimal $\theta^\star$, it still yields practical guidance: enlarging model capacity and broadening the prior π increases the chance of consistently recovering the CEPO. The proof sketch is as follows: first, we prove that optimizing the causal data-prior loss is equivalent to minimizing the KL divergence between q_θ and the true CEPO-PPD. Second, we draw from the Bayesian consistency theory and link its identification to the causal identification under strong ignorability. Once this link is established, the theorem follows naturally from Doob's theorem (Doob, 1949), with the mean of $\pi^{\mu_t}(\cdot \mid \mathcal{D}_{\text{obs}})$ almost surely converging to the true CEPO as observations grow.

B.2 Formal Proof

We now outline the concrete notation and set of assumptions we need for the formal statement and proof of Proposition 1.

Let $\mathcal{B}$ be the Borel sets of the real line. Let the random variable $Z = (\mathbf{X}, T, Y)$ denote all the observed data, defined on the sample space $\mathcal{Z}$ with the σ-algebra $\mathcal{B}_{\mathcal{Z}}$. Similarly, let $\tilde{Z} = (\mathbf{X}, T, \{Y_t\}_{t\in\mathcal{T}})$ be the random variable denoting all the observed and unobserved variables, defined on the sample space $\tilde{\mathcal{Z}}$ with the σ-algebra $\mathcal{B}_{\tilde{\mathcal{Z}}}$. We use $\mathcal{D}^n_{\text{obs}}$ to represent a set of n observed samples $(Z_1, Z_2, \ldots, Z_n)$.

In addition to the random variables above, we define parameters ψ as random variables taking values in Ψ, with σ-algebra $\mathcal{B}_\Psi$. We abuse the notation and use ψ to refer to both a random variable and a fixed value in Ψ. Denote the probability measure on $(\Psi, \mathcal{B}_\Psi)$ by π. For each $\psi \in \Psi$, let $\mathrm{P}^\psi_{\text{obs}}$ and P^ψ be probability measures on $(\mathcal{Z}, \mathcal{B}_{\mathcal{Z}})$ and $(\tilde{\mathcal{Z}}, \mathcal{B}_{\tilde{\mathcal{Z}}})$, respectively. We also define (parametric) CEPOs, and the posterior-predictive distribution (PPD) of CEPOs as

$$\mu_t\left(\mathbf{x}\,;\,\psi\right) \coloneqq \mathbb{E}_{Y_t\sim\mathrm{P}^\psi}\left[Y_t \mid \mathbf{X} = \mathbf{x}\right], \; \pi^{\mu_t}\left(B \mid \mathbf{x}, \mathcal{D}^n_{\text{obs}}\right) \coloneqq \int \mathbb{I}\left(\mu_t\left(\mathbf{x}\,;\,\psi\right) \in B\right) \mathrm{d}\pi\left(\psi \mid \mathcal{D}^n_{\text{obs}}\right), \tag{6}$$

respectively, for all $\psi \in \Psi, t \in \mathcal{T}$, and $B \in \mathcal{B}$. Finally, let $q_\theta(\cdot \mid \mathbf{x}, t, \mathcal{D}^n_{\text{obs}})$ denote the predictive distribution produced by a model with parameter θ, given the query $(t, \mathbf{x})$ and the context $\mathcal{D}^n_{\text{obs}}$. We use q_θ and π^{μ_t} for both the measures and their Lebesgue's densities, which we assume have full support over $\mathbb{R}$. We now state our assumptions:

- **Assumption 1. [Measurability]** $\mathcal{Z}$ is a Borel measurable subset of a Polish space. An analogous statement holds for Ψ. Moreover, $\psi \mapsto \mathrm{P}^{\psi}_{\text{obs}}(A)$ is measurable for every $A \in \mathcal{B}_{\mathcal{Z}}$.

Given **Assumption 1**, we can define $\mathrm{P}^{\pi}_{\text{obs}}$ as the joint distribution of $\left((Z_1, Z_2, \dots,), \psi\right)$. We abuse the notation of $\mathrm{P}^{\psi}_{\text{obs}}$ and write $\mathbf{x} \sim \mathrm{P}^{\pi}_{\text{obs}}$ as the marginal distribution of $\mathbf{x}$, where we first sample $\psi \sim \pi$, and then $z = (\mathbf{x}, t, y) \sim \mathrm{P}^{\psi}_{\text{obs}}$.

- **Assumption 2. [No Redundancy]** $\mathrm{P}^{\psi_1} \neq \mathrm{P}^{\psi_2}$ for all $\psi_1, \psi_2 \in \Psi$ that $\psi_1 \neq \psi_2$.
- **Assumption 3. [Existence of Conditional Expectations]** There exists a $\mathcal{X}_0$ with $\mathrm{P}^{\pi}_{\text{obs}}(\mathcal{X}_0) = 1$, such that for any $t_0 \in \mathcal{T}$, $\mathbf{x}_0 \in \mathcal{X}_0$, there exists a measurable function $g(\psi) \coloneqq \mathbb{E}_{Y \sim \mathrm{P}^{\psi}_{\text{obs}}}[Y \mid \mathbf{X} = \mathbf{x}_0, T = t_0]$ that $\mathbb{E}_{\psi \sim \pi}[|g(\psi)|] < \infty$.

We refer to Assumptions 1-3 collectively as the *regularity* assumptions. Besides these assumptions, we make the following two idealistic assumptions:

- **Assumption 4. [Well-Specified Prior]** The support of prior π only consists of parameters ψ such that P^{ψ} satisfies **Assumption 0** (strong ignorability & SUTVA).

To state the final assumption, we re-state a more formal definition of causal data-prior loss:

Definition 2 (Formal). *For any $t \in \mathcal{T}$, we define the causal data-prior loss for datasets of size n as*

$$\mathcal{L}_{t,n}(\theta) \;\coloneqq\; \mathbb{E}_{\psi \sim \pi,\ \mathcal{D}^n_{obs} \cup \{\mathbf{x}\} \sim \mathrm{P}^{\psi}_{obs}} \left[-\log q_\theta(\mu_t(\mathbf{x}\,;\,\psi) \mid \mathbf{x}, t, \mathcal{D}^n_{obs}) \right]. \tag{7}$$

- **Assumption 5. [Expressive Model]** There exists a parameter $\theta^\star$ that attains the global minimum of $\mathcal{L}_{t,n}(\theta)$ defined in (7), for all $t \in \mathcal{T}$ and all $n \in [\mathbb{N}]$.

Proposition 1 (Formal). *Assume the regularity Assumptions 1-3 hold. Moreover, assume the prior distribution π satisfies Assumption 4, and there exists $\theta^\star$ that satisfies Assumption 5. Then, there exists a set $\mathcal{X}_0$ with $\mathrm{P}^{\pi}_{obs}(\mathcal{X}_0) = 1$, such that for almost all $\psi^\star \sim \pi$, if $Z_1, Z_2, \dots \overset{i.i.d.}{\sim} \mathrm{P}^{\psi^\star}_{obs}$, we have*

$$\lim_{n \to \infty} \mathbb{E}_{\mu \sim q_{\theta^\star}} \left[\mu \,\middle|\, \mathbf{x}_0, t_0, \mathcal{D}^n_{obs} \right] \overset{a.s.}{=} \mu_{t_0}(\mathbf{x}_0\,;\,\psi^\star) \;\; \textit{for all} \;\; t_0 \in \mathcal{T},\ \mathbf{x}_0 \in \mathcal{X}_0, \tag{8}$$

where $\mathcal{D}^n_{obs} = (Z_1, Z_2, \dots, Z_n)$.

Proof. We outline the proof as follows: first, we prove that optimizing the causal data-prior loss is equivalent to minimizing the KL divergence between q_θ and the true CEPO-PPD. Second, we draw from the Bayesian consistency theorem of Doob (Doob, 1949; Miller, 2018) to show that under strong ignorability, as the observational data grows, the mean of the predictive distribution $\pi^{\mu_t}(\cdot \mid \mathbf{x}, \mathcal{D}^n_{\text{obs}})$ almost surely converges to the true CEPO.

Step 0. *Defining the expected forward-KL divergence between $\pi^{\mu_t}(\cdot \mid \mathbf{x}, \mathcal{D}^n_{obs})$ and $q_\theta(\cdot \mid \mathbf{x}, t, \mathcal{D}^n_{obs})$:*

Let $\mathrm{P}^{\pi}_{\text{obs}}(Z_1, Z_2, \dots) \coloneqq \int_\Psi \mathrm{P}^{\psi}_{\text{obs}}(Z_1, Z_2, \dots)\, \mathrm{d}\pi(\psi)$, be the marginal observational distribution under the prior π. Then, the expected forward-KL divergence between the PPD π^{μ_t} and q_θ, for datasets of size n, is defined as

$$\mathcal{L}^{\mathrm{KL}}_{t,n}(\theta) \coloneqq \mathbb{E}_{\mathcal{D}^n_{\text{obs}} \cup \{\mathbf{x}\} \sim \mathrm{P}^{\pi}_{\text{obs}}} \left[\mathrm{D}_{\mathrm{KL}}\left(\pi^{\mu_t}(\cdot \mid \mathbf{x}, \mathcal{D}^n_{\text{obs}}) \,\|\, q_\theta(\cdot \mid \mathbf{x}, t, \mathcal{D}^n_{\text{obs}}) \right) \right], \tag{9}$$

where $\mathcal{D}^n_{\text{obs}} \cup \{\mathbf{x}\} \sim \mathrm{P}^{\pi}_{\text{obs}}$ means to sample $\psi \sim \pi$ first and then sample $\mathcal{D}^n_{\text{obs}} \cup \{\mathbf{x}\} \sim \mathrm{P}^{\psi}_{\text{obs}}$.

Step 1. *The causal data-prior loss $\mathcal{L}_{t,n}$ in (7) is equivalent to the expected forward-KL $\mathcal{L}^{\mathrm{KL}}_{t,n}$:*

First, note that

$$\mathcal{L}^{\mathrm{KL}}_{t,n}(\theta) = \mathbb{E}_{\mathcal{D}^n_{\text{obs}} \cup \{\mathbf{x}\} \sim \mathrm{P}^{\pi}_{\text{obs}}} \left[\int_{\mathbb{R}} \log \frac{\pi^{\mu_t}(\mu \mid \mathbf{x}, \mathcal{D}^n_{\text{obs}})}{q_\theta(\mu \mid \mathbf{x}, t, \mathcal{D}^n_{\text{obs}})}\, \mathrm{d}\pi^{\mu_t}(\mu \mid \mathbf{x}, \mathcal{D}^n_{\text{obs}}) \right]. \tag{10}$$

From (6), we can see that the measure $\pi^{\mu_t}(\cdot \mid \mathbf{x}, \mathcal{D}^n_{\text{obs}})$ is the pushforward of the measure $\pi(\cdot \mid \mathcal{D}^n_{\text{obs}})$ by the function $f(\psi) = \mu_t(\mathbf{x}\,;\,\psi)$. Hence, for any measurable function $h : \mathbb{R} \to \mathbb{R}$, we get

$$\mathbb{E}_{\mu \sim \pi^{\mu_t}(\cdot|\mathbf{x},\mathcal{D}^n_{\text{obs}})}\left[h(\mu)\right] = \mathbb{E}_{\psi \sim \pi\left(\cdot|\mathcal{D}^n_{\text{obs}}\right)}\left[h(f(\psi))\right] = \mathbb{E}_{\psi \sim \pi\left(\cdot|\mathcal{D}^n_{\text{obs}}\right)}\left[h(\mu_t(\mathbf{x}\,;\,\psi))\right]. \tag{11}$$

Setting $h(\mu) = \log \frac{\pi^{\mu_t}(\mu|\mathbf{x},\mathcal{D}^n_{\text{obs}})}{q_\theta(\mu|\mathbf{x},t,\mathcal{D}^n_{\text{obs}})}$, and combining with (10) and (11) yields

$$\mathcal{L}^{\text{KL}}_{t,n}(\theta) = \mathbb{E}_{\mathcal{D}^n_{\text{obs}} \cup \{\mathbf{x}\} \sim \mathrm{P}^{\pi}_{\text{obs}}}\left[\mathbb{E}_{\psi \sim \pi(\cdot|\mathcal{D}^n_{\text{obs}})}\left[\log \frac{\pi^{\mu_t}(\mu_t(\mathbf{x}\,;\,\psi) \mid \mathbf{x}, \mathcal{D}^n_{\text{obs}})}{q_\theta(\mu_t(\mathbf{x}\,;\,\psi) \mid \mathbf{x}, t, \mathcal{D}^n_{\text{obs}})}\right]\right] \tag{12}$$

$$= \mathbb{E}_{\mathcal{D}^n_{\text{obs}} \cup \{\mathbf{x}\} \sim \mathrm{P}^{\pi}_{\text{obs}},\, \psi \sim \pi(\cdot|\mathcal{D}^n_{\text{obs}})}\left[\log \frac{\pi^{\mu_t}(\mu_t(\mathbf{x}\,;\,\psi) \mid \mathbf{x}, \mathcal{D}^n_{\text{obs}})}{q_\theta(\mu_t(\mathbf{x}\,;\,\psi) \mid \mathbf{x}, t, \mathcal{D}^n_{\text{obs}})}\right]. \tag{13}$$

Next, we use the Bayes' rule to derive

$$\underbrace{\mathrm{P}^{\pi}_{\text{obs}}\left(\mathcal{D}^n_{\text{obs}}\right)}_{\text{evidence}} \underbrace{\pi(\psi \mid \mathcal{D}^n_{\text{obs}})}_{\text{posterior}} = \underbrace{\pi(\psi)}_{\text{prior}} \underbrace{\mathrm{P}^{\psi}\left(\mathcal{D}^n_{\text{obs}}\right)}_{\text{likelihood}}. \tag{14}$$

Combining (13) and (14), we get

$$\mathcal{L}^{\text{KL}}_{t,n}(\theta) = \mathbb{E}_{\psi \sim \pi,\ \mathcal{D}^n_{\text{obs}} \cup \{\mathbf{x}\} \sim \mathrm{P}^{\psi}_{\text{obs}}}\left[\log \frac{\pi^{\mu_t}(\mu_t(\mathbf{x}\,;\,\psi) \mid \mathbf{x}, \mathcal{D}^n_{\text{obs}})}{q_\theta(\mu_t(\mathbf{x}\,;\,\psi) \mid \mathbf{x}, t, \mathcal{D}^n_{\text{obs}})}\right] \tag{15}$$

$$= \mathbb{E}_{\psi \sim \pi,\ \mathcal{D}^n_{\text{obs}} \cup \{\mathbf{x}\} \sim \mathrm{P}^{\psi}_{\text{obs}}}\left[-\log q_\theta(\mu_t(\mathbf{x}\,;\,\psi) \mid \mathbf{x}, t, \mathcal{D}^n_{\text{obs}})\right] + \text{constant term in } \theta \tag{16}$$

$$= \mathcal{L}_{t,n}(\theta) + \text{constant term in } \theta. \tag{17}$$

Step 2. *Attaining the global optima for the causal data-prior loss $\mathcal{L}_{t,n}$ at $\theta^\star$ for all $t \in \mathcal{T}$ and $n \in [\mathbb{N}]$ means that $q_{\theta^\star}\left(\cdot \mid \mathbf{x}, t, \mathcal{D}^n_{obs}\right) \overset{a.e.}{=} \pi^{\mu_t}\left(\cdot \mid \mathbf{x}, \mathcal{D}^n_{obs}\right)$ for all $t \in \mathcal{T}$, $n \in \mathbb{N}$, and almost all $\mathcal{D}^n_{\text{obs}} \cup \{\mathbf{x}\} \sim \mathrm{P}^{\pi}_{\text{obs}}$.*

This result directly follows from the Step 1, and the fact that the KL divergence between two distributions is globally minimized and equal to zero iff the two are equal a.e.

Now that we have established the equivalence of the model q_θ with the PPD π^{μ_t}, we show the asymptotic consistency of π^{μ_t}.

Step 3. *Re-stating Doob's consistency with our notation:*

Corollary 2 (Corollary 2.3 in Miller (2018))**.** *Suppose $\mathcal{Z}$ and Ψ are Borel measurable subsets of two Polish spaces. Suppose $g : \Psi \to \mathbb{R}$ is a measurable function with $\mathbb{E}\left[|g(\psi)|\right] < \infty$. For every $\psi \in \Psi$, let P^{ψ}_{obs} be a probability measure on $(\mathcal{Z}, \mathcal{B}_{\mathcal{Z}})$. Moreover, assume $\psi \mapsto \mathrm{P}^{\psi}_{obs}(A)$ is measurable for every $A \in \mathcal{B}_{\mathcal{Z}}$, and $\psi \neq \psi' \implies \mathrm{P}^{\psi}_{obs} \neq \mathrm{P}^{\psi'}_{obs}$. There exists $\Psi_0 \subseteq \Psi$ with $\pi(\Psi_0) = 1$ such that for all $\psi^\star \in \Psi_0$, if $Z_1, Z_2, \ldots \sim \mathrm{P}^{\psi^\star}_{obs}$ i.i.d., then*

$$\lim_{n\to\infty} \mathbb{E}_{\psi \sim \pi(\cdot|Z_1,\ldots,Z_n)}\left[g(\psi) \mid Z_1, \ldots, Z_n\right] \overset{a.s.}{=} g(\psi^\star). \tag{18}$$

Final Step. *All the assumptions for Doob's Theorem are satisfied and we can reach the main result*:

- **Assumption 1** states that $\mathcal{Z}$ and Ψ are Borel measurable subsets of two Polish spaces. Moreover, $\psi \mapsto \mathrm{P}^{\psi}_{\text{obs}}(A)$ is measurable for every $A \in \mathcal{B}_{\mathcal{Z}}$.
- Fix a value of $\mathbf{x}_0 \in \mathcal{X}_0$, and any $t_0 \in \mathcal{T}$. Define $g(\psi) \coloneqq \mathbb{E}_{Y \sim \mathrm{P}^{\psi}_{\text{obs}}}\left[Y \mid \mathbf{X} = \mathbf{x}_0, T = t_0\right]$. According to **Assumption 3**, $g(\psi)$ is measurable and satisfies $\mathbb{E}\left[|g(\psi)|\right] < \infty$.
- **Assumptions 2** and **4** combined yield $\psi \neq \psi' \implies \mathrm{P}^{\psi}_{\text{obs}} \neq \mathrm{P}^{\psi'}_{\text{obs}}$ for all ψ, ψ' in the support of π.

We can now directly use Corollary 2 to deduce that there exists $\Psi_0 \subseteq \Psi$ with $\pi(\Psi_0) = 1$, such that for all $\psi^\star \in \Psi_0$ and i.i.d. $\mathcal{D}^n_{\text{obs}} \sim \mathrm{P}^{\psi^\star}_{\text{obs}}$, we have

$$\lim_{n\to\infty} \mathbb{E}_{\psi \sim \pi\left(\cdot|\mathcal{D}^n_{\text{obs}}\right)}\left[g(\psi) \mid \mathcal{D}^n_{\text{obs}}\right] \overset{a.s.}{=} g(\psi^\star). \tag{19}$$

Finally, note that since the support of π satisfies strong ignorability (under **Assumption 4**), for any i.i.d. $\mathcal{D}^n_{\text{obs}} =$

$(Z_1, \ldots, Z_n) \sim \mathrm{P}_{\text{obs}}^{\psi^\star}$, the support of the posterior $\pi(\cdot \mid \mathcal{D}_{\text{obs}}^n)$ also satisfies strong ignorability. Hence,

$$\mathbb{E}_{\psi\sim\pi\left(\cdot|\mathcal{D}_{\text{obs}}^n\right)}\left[g(\psi) \mid \mathcal{D}_{\text{obs}}^n\right] = \mathbb{E}_{\psi\sim\pi\left(\cdot|\mathcal{D}_{\text{obs}}^n\right)}\left[\mathbb{E}_{Y\sim\mathrm{P}_{\text{obs}}^{\psi}}\left[Y \mid \mathbf{X}=\mathbf{x}_0, T=t_0\right] \;\middle|\; \mathcal{D}_{\text{obs}}^n\right] \tag{20}$$

$$\text{(By Strong Ignorability)} \quad = \mathbb{E}_{\psi\sim\pi\left(\cdot|\mathcal{D}_{\text{obs}}^n\right)}\left[\mathbb{E}_{Y_{t_0}\sim\mathrm{P}^{\psi}}\left[Y_{t_0} \mid \mathbf{X}=\mathbf{x}_0\right] \;\middle|\; \mathcal{D}_{\text{obs}}^n\right] \tag{21}$$

$$= \mathbb{E}_{\psi\sim\pi\left(\cdot|\mathcal{D}_{\text{obs}}^n\right)}\left[\mu_{t_0}\left(\mathbf{x}_0\,;\,\psi\right)\right] \tag{22}$$

$$= \mathbb{E}_{\mu\sim\pi^{\mu_{t_0}}\left(\cdot|\mathbf{x}_0,\mathcal{D}_{\text{obs}}^n\right)}\left[\mu\right]. \tag{23}$$

Combining (23) with (19), we get

$$\lim_{n\to\infty} \mathbb{E}_{\mu\sim\pi^{\mu_{t_0}}\left(\cdot|\mathbf{x}_0,\mathcal{D}_{\text{obs}}^n\right)}\left[\mu\right] = \mu_{t_0}\left(\mathbf{x}_0\,;\,\psi^\star\right). \tag{24}$$

for almost all $\psi^\star$ in the support of π.

We have now shown consistency guarantees for the true CEPO-PPD π^{μ_t}. To obtain the final result, we can simply plug in $q_\theta \overset{a.e.}{=} \pi^{\mu_t}$ (from **Step 2**) to prove Proposition 1. □

C Prior Generation & Simulating DGPs

As illustrated in Figure 3, our prior generation consists of retrieving or synthesizing a base table, subsampling covariates $\mathbf{X}$ and CEPOs μ_0 and μ_1, synthesizing treatments T, potential outcomes Y_t, and finally, observed outcomes Y. We break down each of the components:

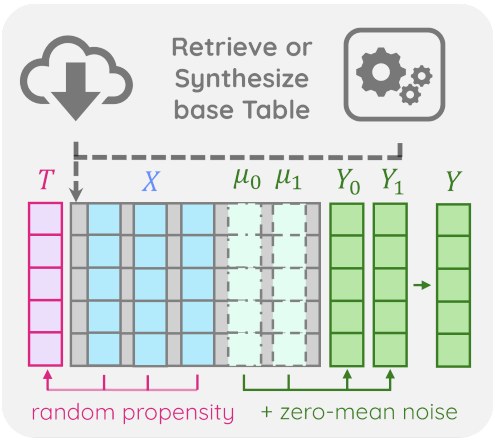

Figure 3. **Prior construction**. Sample diverse base tables (OpenML or synthetic TabPFN), select covariates X, draw treatment T with a random propensity model, select columns μ_0, μ_1 and add zero-mean noise to form Y_0, Y_1, and Y.

Data Sources for the Base Tables. We draw the base tables from two sources: ***(i)*** real-world tables from OpenML, and ***(ii)*** fully synthetic data.

(i) We use the OpenML collections used in Grinsztajn et al. (2022), AMLB (Gijsbers et al., 2024), and TabZilla (McElfresh et al., 2023), all listed in Ma et al. (2024). To widen coverage, we also add tables from CTR23 (Fischer et al., 2023) and CC18 (Bischl et al., 2021). All OpenML IDs are in this link.[1] Data leakage is ruled out as none of the tables that share covariates or outcomes with our test sets (Lalonde, IHDP, ACIC, Criteo, Megafon, Hillstrom, Lenta, X5) are included in training. Moreover, the propensities are sampled purely synthetically, following the approach described below.

(ii) For additional diversity, we generate synthetic tables using the random neural networks used to train TabPFN v1, with the same hyperparameters described in Hollmann et al. (2023). Inputs, from a standard Gaussian distribution, are fed into the network, and a subset of the outputs and hidden neurons are selected to construct the tabular data. Some columns are discretized at random to produce categorical and ordinal variables to reflect the structure of real-world tabular domains. While TabPFN v2 (Hollmann et al., 2025) is a newer and stronger model, its training data is not publicly available, so we restrict ourselves to the v1 generator to ensure transparent evaluation and leakage control.

CEPOs with Heterogeneity Control. Once the base table is given, we randomly select two columns and name them $\mu_{\text{raw},0}$ and $\mu_{\text{raw},1}$. However, in practice, we observe that directly using such columns for CEPOs can result in large variances (*heterogeneity*) for CATEs. We therefore apply a light-weight post-processing inspired by `RealCause` (Neal et al., 2020).

The post-processing requires a heterogeneity hyperparameter γ, which we sample uniformly from $[0, 1]$ during prior generation. Then, for N units (rows) extracted from the base table, let $\tau_{\text{raw}}^{(n)} = \mu_{\text{raw},1}^{(n)} - \mu_{\text{raw},0}^{(n)}$ be the CATE for unit $n \in [N]$, and $\lambda_{\text{raw}} = \frac{1}{N}\sum_{n=1}^{N} \tau_{\text{raw}}^{(n)}$ the sample ATE. We draw i.i.d. $\{\alpha^{(n)}\}_{n=1}^{N} \sim \mathrm{Unif}[0, 1]$ and construct the final γ-augmented

[1] https://drive.google.com/file/d/1NXib83Lc7jGOPJx554p-I3sxFrcWeF52

CEPOs as

$$\mu_1^{(n)} \coloneqq \left[\alpha^{(n)} + (1-\alpha^{(n)})\gamma\right]\mu_{\text{raw},1}^{(n)} + (1-\gamma)(1-\alpha^{(n)})(\mu_{\text{raw},0}^{(n)} + \lambda_{\text{raw}}), \tag{25}$$

$$\mu_0^{(n)} \coloneqq \left[(1-\alpha^{(n)}) + \alpha^{(n)}\gamma\right]\mu_{\text{raw},0}^{(n)} + (1-\gamma)\alpha^{(n)}(\mu_{\text{raw},1}^{(n)} - \lambda_{\text{raw}}). \tag{26}$$

A simple algebraic check shows

$$\tau^{(n)} \;\coloneqq\; \mu_1^{(n)} - \mu_0^{(n)} \;=\; \gamma\,\tau_{\text{raw}}^{(n)} + (1-\gamma)\lambda_{\text{raw}}, \quad \operatorname{Var}[\tau \mid \mathbf{x}] = \gamma^2 \operatorname{Var}[\tau_{\text{raw}} \mid \mathbf{x}]. \tag{27}$$

Hence, while preserving the average treatment effect, $\gamma = 0$ yields a dataset with a zero variance CATE (fully homogeneous), whereas $\gamma = 1$ recovers the original heterogeneity.

Outcomes. After constructing the CEPO columns $\mu_0(\mathbf{x})$ and $\mu_1(\mathbf{x})$, we need to turn them into potential outcomes by adding zero-mean noises. To avoid tying the data to a specific parametric noise model, we introduce two additional *nuisance* columns, $\eta_0(\mathbf{x})$ and $\eta_1(\mathbf{x})$, sampled from the base table. Let ϵ_t be random scalars, independent from $\mathbf{x}$, with $\mathbb{E}[\epsilon_t] = 0$. We define the potential outcomes as

$$Y_t \;=\; \mu_t(\mathbf{x}) \;+\; \eta_t(\mathbf{x})\,\epsilon_t, \quad t \in \mathcal{T}. \tag{28}$$

This construction preserves the conditional means, that is $\mathbb{E}[Y_t \mid \mathbf{x}] = \mu_t(\mathbf{x})$. The input-dependent scale factors $\eta_t(\mathbf{x})$ allow for heteroscedastic noises and capture a richer family of outcome distributions than additive parametric noise models. For our training, we sample ϵ_t from a Gaussian with a variance uniformly drawn from $(0, \operatorname{Var}(\mu_t)]$. This choice of noise values ensures a similar noise scale to the scale of CEPOs, resulting in training data with a more informative signal-to-noise ratio.

Propensities with Positivity Control. Given a covariate vector $\mathbf{x}$, the strong ignorability assumption requires the propensity values $0 < \mathrm{P}(T = 1 \mid \mathbf{X} = \mathbf{x}) < 1$. Hence, due to the invertibility of the sigmoid function, it is sufficient to generate treatment logits, through any function $f : \mathbf{X} \to \mathbb{R}$, and then apply a sigmoid function to get values within $(0, 1)$. To simulate different degrees of confounding, we choose f by randomly selecting one of the following mechanisms:

- ***(i)*** **Randomized treatments (RCT).** Treatments are independent of covariates, i.e., f is constant. We sample $c \sim \texttt{Logistic}(0, 1)$ and set $f(\mathbf{x}) = c$ to get uniform propensities.
- ***(ii)*** **Linear logits.** Draw the random vector $\mathbf{w}$ from a standard Gaussian and set $f(\mathbf{x}) = \mathbf{w}^\top \mathbf{x}$.
- ***(iii)*** **Non-linear logits.** Feed $\mathbf{x}$ into a randomly initialized MLP, similar architecture of Hollmann et al. (2023), to get $f(\mathbf{x})$.

Empirically, we observe that the above procedure yields an artificially high level of positivity, which is not reflective of real-world scenarios. We therefore apply a light-weight post-processing transform, inspired by `RealCause` (Neal et al., 2020), to better control the positivity level. Concretely, we sample a parameter $\xi \in [0, 1]$ and *exacerbate* extreme propensity scores to mimic poor positivity:

$$\mathrm{P}(T = 1 \mid \mathbf{X} = \mathbf{x}) \;\coloneqq\; \xi\,\text{Sigmoid}\left(f(\mathbf{x})\right) \;+\; (1-\xi)\,\mathbb{I}\big[f(\mathbf{x}) > 0\big]. \tag{29}$$

Here, $\xi = 1$ leaves the original positivity intact. However, for smaller ξ values, the support of the treated and control groups become increasingly disjoint, leading to low-positivity scenarios.

Treatment Assignment. Finally, each unit's treatment is drawn as $T \sim \text{Bernoulli}\left(\text{Sigmoid}\left(f(\mathbf{X})\right)\right)$, and the observed outcome Y is also derived by selecting the assigned potential outcome $Y \coloneqq Y_T$.

Collectively, all of the steps above simulate different DGPs, with various levels of positivity and heterogeneity, extracted from real and synthetic sources of tabular data. This procedure creates a broad prior π for CausalPFN, which is necessary for the model to work well in practice.

D Architecture & Training Details

D.1 A High-Level Overview of the Training Pipeline

Figure 4 illustrates the abstract training pipeline: at each iteration, we sample a DGP $\psi_i \sim \pi$, generate an observational dataset $\mathcal{D}_{\text{obs}}$ from this DGP, and select a query point $(\mathbf{x}, t)$. We compute (simulate) the ground-truth CEPO $\mu_t\left(\mathbf{x}\,;\,\psi_i\right)$ and feed both the observational data and query to the model. The model outputs a CEPO-PPD, and we update θ to increase the probability assigned to the true CEPO value. Through stochastic gradient descent, θ minimizes the data-prior loss and implicitly learns to perform posterior-predictive inference, without ever explicitly computing the posterior.

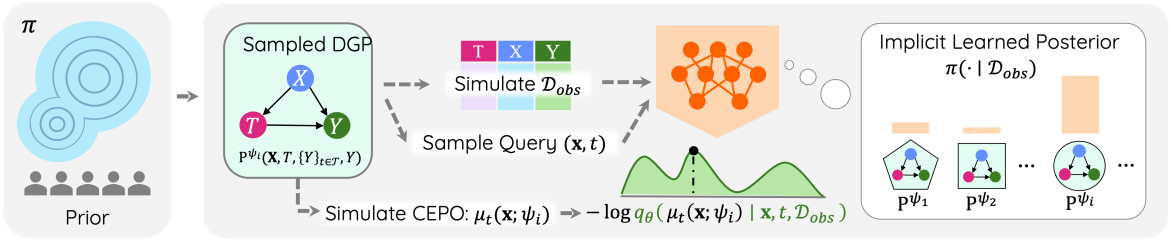

Figure 4. **Causal Data-prior Training**. At each iteration a DGP $\psi_i \sim \pi$ is sampled *(left)*, yielding the joint law $\mathrm{P}^{\psi_i}(X, T, \{Y_t\}_{t\in\mathcal{T}}, Y)$. From this DGP we simulate an observational context $\mathcal{D}_{\text{obs}}$ and a query $(\mathbf{x}, t)$ with its true $\mu_t(\mathbf{x}\,;\,\psi_i)$ *(centre)*. Passing $(\mathbf{x}, t, \mathcal{D}_{\text{obs}})$ through the transformer predicts the CEPO-PPD $q_\theta(\cdot \mid \mathbf{x}, t, \mathcal{D}_{\text{obs}})$ *(in green)*, which is derived from an implicit posterior $\pi(\cdot \mid \mathcal{D}_{\text{obs}})$ that is *never* explicitly computed *(right)*. We train θ to minimize the causal data-prior loss $-\log q_\theta\big(\mu_t(\mathbf{x}\,;\,\psi_i) \mid \mathbf{x}, t, \mathcal{D}_{\text{obs}}\big)$ *(bottom)*.

Algorithm 1 Parallel training of CausalPFN.

Input: Prior π, DGPs and CEPO values $\mathrm{P}^{\psi}_{\text{obs}}$, $\mu_t(\cdot;\,\psi)$, model q_θ, DGP batch size B_t, query batch size B_q, fixed feature length F, and histogram loss $\texttt{HL}$.
while not converged **do**
  Sample $\psi[1], \ldots, \psi[B_t] \sim \pi$
  Sample $\mathcal{D}_{\text{obs}}[i] \sim \mathrm{P}^{\psi[i]}_{\text{obs}}, \forall 1 \le i \le B_t$
  Randomly sample query treatments $t^{(i,j)}$ for $1 \le i \le B_t, 1 \le j \le B_q$
  Sample query covariates $\mathbf{x}^{(i,j)} \sim \mathrm{P}^{\psi}_{\text{obs}}[i]$ for $1 \le i \le B_t, 1 \le j \le B_q$
  Set $\mu^{(i,j)} \leftarrow \mu_{t^{(i,j)}}\left(\mathbf{x}^{(i,j)}\,;\,\psi[i]\right)$
  Pad $\mathbf{x}^{(i,j)}$ with zeros such that $\mathbf{x}^{(i,j)} \in \mathbb{R}^F$
  $\widehat{\mathcal{L}} \leftarrow \frac{1}{B_t \cdot B_q} \sum_{i,j} \texttt{HL}\left[\mu^{(i,j)} \| q_\theta(\cdot \mid \mathbf{x}^{(i,j)}, t^{(i,j)}, \mathcal{D}_{\text{obs}}[i])\right]$
  Update θ using the gradients $\nabla_\theta \widehat{\mathcal{L}}$
end while

D.2 Architecture & Training

We represent each context row $(t, \mathbf{x}, y)$ and query row $(t, \mathbf{x})$ as single tokens by summing up (1) a treatment embedding for t, (2) a covariate embedding for $\mathbf{x}$ (padded to length $F = 100$), and (3) an outcome embedding for y (only for context rows). We use linear layers for embeddings and omit the positional encodings to preserve the permutation invariance of the context set, similar to other PFN-style transformers. All tokens—context and query—are passed into a 20-layer transformer, with a hidden size of 384, QK-normalization (RMS)[2], and a parallel SwiGLU-activated (Shazeer, 2020) feed-forward block.

The transformer's query outputs are then projected to a 1024-dimensional logit vector, then softmaxed at a fixed temperature of $\theta_T = 1.0$ to form a discrete CEPO posterior over the interval $[-10, 10]$. We then scale the interval to match the scale of the outcomes and clip the out-of-range values. At inference time, we return the posterior mean as the point estimate and sample 10,000 times to estimate credible intervals at any desired significance level α.

The full model has approximately 20M parameters and is trained in two stages: ***(i)*** a predictive phase that mimics standard predictive PFN training from Ma et al. (2024), and ***(ii)*** a causal phase that optimizes the CEPO loss. We use AdamW (Kingma, 2014) with warmup and cosine annealing for the predictive phase, and switch to the schedule-free optimizer (Defazio et al., 2024) in the causal phase. The model is trained with a maximum context length of 16K in the first phase and 2,048 in the second. We use four A100 GPUs trained for one week for the initial phase, and two days on an H100 for the second phase.

Finally, to enhance parallel training, we batch both the queries and the tables. That is, rather than sampling only one DGP and one query token, each gradient update samples B_t DGPs, draws B_q queries per DGP, and concatenates everything into a single tensor. The tensor is then passed through the transformer to get $B_t B_q$ CEPO-PPDs. The final loss is averaged over all the batches. See Figure 5 and Algorithm 1 for a detailed demonstration of CausalPFN's training pipeline.

D.3 Handling Large Tables at Inference Time

CausalPFN's default maximum context length is set to 4,096 at inference, but real-world tables may contain millions of rows. Training PFN-style transformers on such long contexts can be challenging due to hardware or architectural constraints.

[2]Different from Henry et al. (2020), we perform normalization *after* the query and key projection.

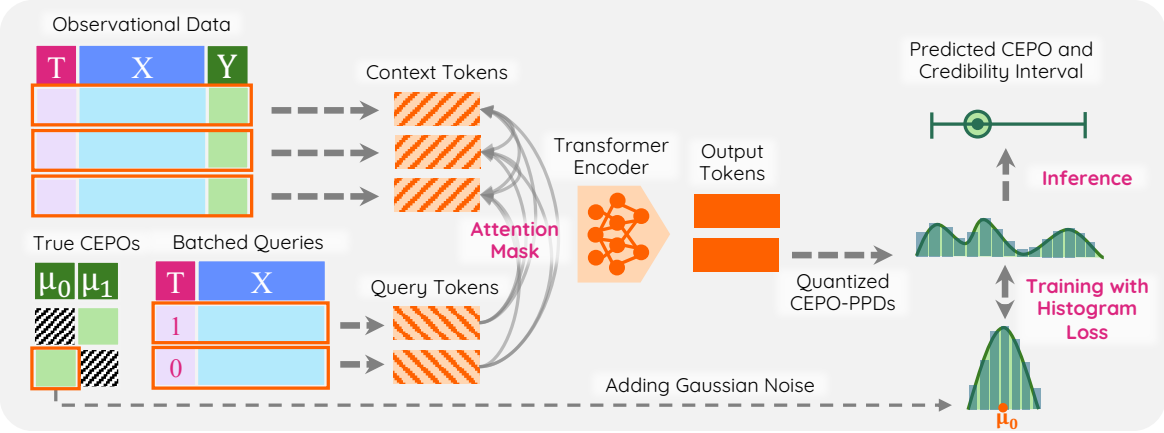

Figure 5. **Training pipeline.** *(Left)*: An observational data, and a batch of queries along with their true CEPO values are sampled from the prior. Each observational row, containing the treatment, covariates, and the outcome form a context token, while query tokens consist of only the treatment and covariates. *(Middle)*: The context and query tokens are fed into a transformer encoder with an asymmetric attention masking, where both context and query tokens attend only to the context tokens. *(Bottom-Right)*: The output tokens are projected into a 1024-dimensional logit vector and softmaxed to form a discrete posterior-predictive distribution (CEPO-PPD). Then, the true CEPO value corresponding to each output token is smoothed by adding narrow-width Gaussian noise, and training is done by minimizing the cross-entropy (histogram) loss; this is a trick commonly used for stabilizing training in histogram losses. *(Top-Right)*: At inference time, we return the CEPO-PPD mean as the point estimate and sample from CEPO-PPD to estimate credible intervals.

While some tabular foundation models such as TabICL (Qu et al., 2025) modify the architecture itself, Thomas et al. (2024) show that, retrieving a small relevant subset of rows for each query at inference time allows a model with a short context length to better generalize to longer contexts.

We adopt this retrieval philosophy in CausalPFN to enable causal effect estimation on large tables. First, we fit a lightweight gradient boosting regressor on the context data to produce weak CATE estimates for each covariate. This regressor estimates CATE by regressing outcomes on the treatment and covariates and then taking the difference in predicted outcomes between $T = 1$ and $T = 0$. This step is applied *only* when the table is too large to fit within the model's maximum context window. We then ***(i)*** sort both the context rows and the queries based on their weak CATE estimates, which effectively stratifies the data; ***(ii)*** partition the ordered queries into consecutive mini-batches; and ***(iii)*** for each query batch, use a fast bisection search to select a contiguous window of context rows whose weak CATE estimate range most closely matches that of the batch. As a result, each batch is exposed only to a neighborhood of rows with similar causal effects, allowing all CEPO predictions to be computed with short forward passes.

E Baseline Hyperparameters and Results without Hyperparameter Tuning

No Hyperparameter Tuning. Table 2 summarizes the performance of all methods without hyperparameter tuning. Indeed, in this setting, CausalPFN consistently attains the first or second best results on heterogeneous treatment effect, and overall minimum average relative error on ATE.

EconML Hyperparameters. For the results without hyper parameter tuning in Table 2, we ran the models with the recommended hyper parameters in the Jupyter notebooks from EconML (Battocchi et al., 2019). For the tuned results in Table 1, we performed a grid search on both the propensity and outcome models with the following search space:

- Model family $\in$ { Random forest (with 100 trees), Gradient boosting }
- Maximum depth $\in \{3, 5\}$
- Minimum samples per leaf $\in \{10, 50\}$

Each candidate configuration was evaluated using three-fold cross-validation. For DR-Learner and Forest DML, we additionally expanded the covariates with quadratic terms (polynomial degree 2).

CATE Nets. For the results without hyper parameter tuning in Table 2, we ran the models with the default hyperparameters

and a batch size of 512. For the tuned results in Table 1, we perform a grid search on the hyperparameters for the neural architecture:

- Number of layers $\in \{2, 3\}$
- Representation dimension $\in \{128, 256\}$
- Number of hidden output layers $\in \{1, 2\}$
- Width of the hidden output layers $\in \{128, 256\}$

the rest of the hyperaparameters are left unchanged.

BART & GRF. The GRF implementation includes an internal `tune` option. We enable this option for the tuned experiment reported in Table 1 and disable it for the untuned experiment in Table 2. BART, on the other hand, offers no comparable hyperparameter-tuning routine. Its only alternative, a full cross-fit, is prohibitively slow uses a rudimentary Bayesian routine. Thus, the BART scores appear unchanged in both Table 1 and Table 2.

Table 2. **CATE & ATE results.** PEHE *(left half)* alongside ATE relative error and its overall average *(right half)*. PEHE for Lalonde CPS/PSID is shown in thousands. Best numbers are in **green**; second best are in **blue**.

Method	PEHE ± Standard Error (↓ better)				ATE Relative Error ± Standard Error (↓ better)				
	IHDP	ACIC 2016	Lalonde CPS ($\times 10^3$)	Lalonde PSID ($\times 10^3$)	IHDP	ACIC 2016	Lalonde CPS	Lalonde PSID	Avg.
CausalPFN	**0.58±0.07**	**0.92±0.11**	**8.83±0.04**	**13.98±0.43**	0.20±0.04	**0.04±0.01**	**0.07±0.02**	0.20±0.04	**0.18±0.03**
T Learner	2.28±0.34	1.41±0.11	**9.24±0.05**	**14.01±0.41**	0.21±0.04	0.05±0.01	0.37±0.02	**0.09±0.03**	**0.20±0.03**
DragonNet	2.13±0.24	2.23±0.20	10.3±0.39	16.2±0.78	0.19±0.04	0.09±0.02	0.55±0.10	0.40±0.07	0.23±0.03
GRF	4.26±0.69	1.36±0.30	12.1±0.22	21.2±0.48	**0.18±0.03**	0.07±0.02	0.81±0.06	0.80±0.05	0.27±0.03
DR Learner	3.82±0.49	1.09±0.09	14.34±0.72	34.04±3.64	0.19±0.03	**0.04±0.01**	0.91±0.10	46.8±42.9	0.28±0.03
RA Net	2.08±0.19	2.08±0.19	12.5±0.37	19.9±1.71	0.20±0.03	0.07±0.03	0.93±0.05	0.67±0.06	0.28±0.03
TarNet	**1.88±0.15**	2.26±0.20	11.9±0.13	18.4±0.45	0.20±0.04	0.06±0.02	0.95±0.02	0.68±0.03	0.29±0.03
S Learner	3.06±0.52	1.36±0.12	12.86±0.04	21.81±0.44	0.23±0.05	0.05±0.01	1.01±0.01	0.94±0.01	0.33±0.04
Forest DML	3.73±0.61	1.20±0.24	133.9±13.8	27.51±1.83	**0.11±0.02**	0.05±0.01	3.91±0.63	0.96±0.09	0.50±0.10
BART	2.50±0.39	**0.68±0.11**	12.7±0.11	20.8±0.45	0.50±0.11	**0.04±0.01**	1.01±0.02	0.83±0.03	0.53±0.08
IPW	5.70±0.89	3.21±0.62	10.94±0.06	18.20±0.45	0.23±0.04	0.24±0.05	**0.25±0.03**	**0.05±0.01**	0.81±0.03
X Learner	3.00±0.47	1.02±0.16	13.01±0.10	20.27±0.69	0.19±0.03	**0.03±0.01**	1.06±0.02	0.74±0.06	0.92±0.03

F Marketing Experiments

Policy Evaluation on Marketing Randomized Trials. Ground-truth CATEs are only available for synthetic or semi-synthetic datasets. However, if a randomized controlled trial (RCT) is available, we can still evaluate the quality of a CATE estimator by assessing the performance of policies derived from it. A common tool for evaluating such policies is the *Qini curve* (Radcliffe, 2007), which plots the cumulative treatment effect when units are ranked in descending order of their predicted CATE.

Formally, let $(y^{(n)}, t^{(n)})_{n=1}^{N}$ denote outcomes and binary treatments from an RCT, and let $\widehat{\tau}_n$ be the corresponding CATE estimates, ordered so that $\widehat{\tau}_1 \geq \cdots \geq \widehat{\tau}_N$. Define

$$\lambda(q) := \sum_{n=1}^{\lfloor qN \rfloor} \left(\frac{t^{(n)} y^{(n)}}{r(q)} - \frac{(1-t^{(n)}) y^{(n)}}{1-r(q)} \right), \qquad Q(q) := q \cdot \lambda(q)/\lambda, \qquad 0 \leq q \leq 1, \tag{30}$$

where $r(q) = \frac{1}{\lfloor qN \rfloor} \sum_{n}^{\lfloor qN \rfloor} t^{(n)}$ is the empirical treatment rate for the first q-quantile of units. Because the data comes from an RCT, $\lambda(q)$ unbiasedly estimates the ATE for the top q-quantile of units ranked by predicted CATEs. Plotting $Q(q)$ against the treated fraction q yields the (normalized) Qini curve, and the area under this curve is called the *Qini score*. A random ranking produces a baseline curve as a straight line from $(0, 0)$ to $(1, 1)$. The higher the Qini curve lies above this line, the better the model prioritizes high-impact units with larger CATE values, leading to greater lift and policy benefit.

We benchmark CausalPFN on five large marketing RCTs from the `scikit-uplift` library (Maksim Shevchenko, 2020). The first dataset, Hillstrom (Hillstrom, 2008), includes 64,000 customers randomly assigned to one of three treatments: no

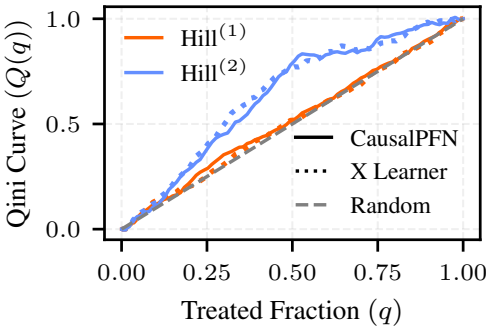

Figure 6. Hill$^{(1)}$ & Hill$^{(2)}$ Qini curves.

Table 3. **Normalized Qini scores** (↑ better). All datasets use 50k stratified subsamples, except Hill$^{(1)}$ and Hill$^{(2)}$, which use the full 64k rows. Columns are normalized to 1.0 for the best model and winners are **bolded**.

Method	Hill$^{(1)}$	Hill$^{(2)}$	Criteo	X5	Lenta	Mega	Avg.
CausalPFN	0.992	0.968	0.859	0.922	**1.000**	0.970	**0.952**
X Learner	0.975	0.980	**1.000**	0.937	0.771	**1.000**	0.944
S Learner	**1.000**	**1.000**	0.881	**1.000**	0.651	0.941	0.912
DA Learner	0.985	0.964	0.626	0.929	0.781	0.998	0.881
T Learner	0.991	0.972	0.701	0.964	0.644	0.986	0.876

Table 4. **Normalized Qini scores** (↑ better). Scores are normalized per dataset such that the top-performing model achieves 1.0 (highlighted in **bold**). All datasets use full stratified subsamples: Hill$^{(1)}$ and Hill$^{(2)}$ (64K rows), Criteo (2.5M rows), X5 (200K rows), Lenta (687K rows), and Mega (600K rows).

Method	Hill$^{(1)}$	Hill$^{(2)}$	Criteo	X5	Lenta	Mega	Avg.
S Learner	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	0.913	**0.985**
X Learner	0.975	0.980	0.994	0.965	0.868	0.997	0.963
DA Learner	0.985	0.964	0.955	0.969	0.903	**1.000**	0.963
T Learner	0.991	0.972	0.902	0.953	0.833	0.987	0.940
CausalPFN	0.992	0.968	0.939	0.746	0.947	0.954	0.924

e-mail, an e-mail advertising men's merchandise, or an e-mail advertising women's merchandise. The outcome is whether a website visit occurred within two weeks (binary). We consider two causal tasks: **Hill**$^{(1)}$ – Men's-merchandise e-mail (treatment) vs. no e-mail (control), and **Hill**$^{(2)}$ – Women's-merchandise e-mail vs. no e-mail. We estimate CATEs using CausalPFN (five-fold honest splitting) and X Learner. Figure 6 shows Qini curves where CausalPFN closely matches X Learner across the targeting range. Notably, Hill$^{(2)}$ shows much greater gains, *suggesting women's-merchandise ads, compared to men's, drive more visits*. We also evaluate CausalPFN on four larger campaigns—**Lenta**, Retail Hero (**X5**), Megafon (**Mega**), and **Criteo** (Lenta LLC, 2020; Retail Hero, 2020; Megafon PJSC, 2020; Émilie Diemert et al., 2018)—each with $\sim 10^6$ rows. For tractability, we compute Qini scores on stratified 50k subsamples; Table 3 shows CausalPFN achieves the best mean performance. However, when we run it on full tables in Table 4, we observe a drop in performance, which aligns with known context-length limitations of PFN-style models on large tables (Thomas et al., 2024). Still, the strong subsample results highlight the potential of scaling CausalPFN to longer contexts, remaining an important future direction.

G Uncertainty Quantification and Calibration

Beyond point estimation, the CEPO-PPDs capture the epistemic uncertainty about the true CEPO given the observed data; a sharply peaked CEPO-PPD indicates that the available data confidently pins down the causal effects, whereas a high-variance distribution signals substantial uncertainty, which can justify making conservative decisions. Thus, we can use q_θ to construct credible intervals around CEPOs, CATEs, and ATEs via sampling from $q_\theta(\cdot \mid \mathbf{x}, t = 1, \mathcal{D}_{\text{obs}})$ and $q_\theta(\cdot \mid \mathbf{x}, t = 0, \mathcal{D}_{\text{obs}})$ and using them to quantify the epistemic uncertainty of our causal-effects.

For each unit covariate $\mathbf{x}$, CausalPFN can produce both point estimates and credible intervals for the CATE and CEPO. In particular, we do so by drawing 10,000 samples from the quantized distribution $q_\theta(\cdot \mid \mathbf{x}, t, \mathcal{D}_{\text{obs}})$ and construct credible intervals at any desired level α. Here, we evaluate these intervals, focusing on the model's calibration. We also assess a key assumption from Proposition 1—whether the inference-time DGP $\psi^\star$ lies within the prior π, and critically, how the model behaves when this assumption is violated.

We define families of synthetic DGPs to simulate both in-distribution and out-of-distribution (OOD) scenarios. We use two families of synthetic DGPs, polynomials and sinusoidals. CausalPFN is trained either on the same family it is tested on, or on a different one (OOD). As a general recipe, each DGP defines a treatment logit function $f(\mathbf{x}) \in \mathbb{R}$ and assigns treatments by sampling from the $\text{Bernoulli}\left(\text{Sigmoid}(f(\mathbf{x}))\right)$. Moreover, each DGP specifies two CEPO functions $\mu_0, \mu_1 : \mathbf{x} \to \mathbb{R}$. It then samples the potential outcomes by $Y_t = \mu_t(\mathbf{x}) + \epsilon_t$ for $t \in \{0, 1\}$, where the noise terms $\epsilon_t \sim \text{Normal}(0, 1)$, $\text{Laplace}(0, 1)$, or $\text{Uniform}(-1, 1)$ with equal probability. We now describe each DGP family in more detail:

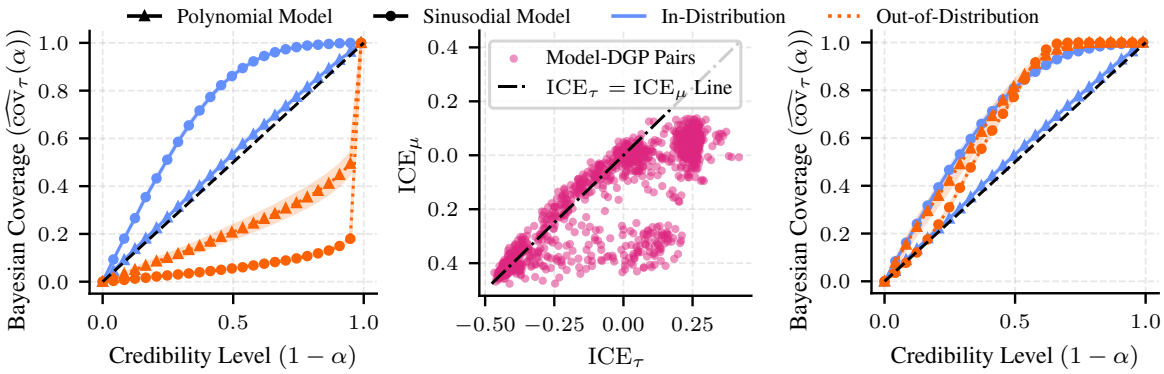

Figure 7. **Calibration.** *(Left)*: CATE coverage vs. nominal credibility. In-distribution DGPs (blue) lie on or above the diagonal (calibrated/conservative), while OOD DGPs (orange) fall below it (overconfident). *(Middle)*: Across model–DGP pairs, CATE ICE (x-axis) exceeds regression ICE (y-axis). *(Right)*: Temperature scaling based on regression ICE ensures the model is either calibrated or conservative for both in- and out-of-distribution DGPs.

(a) **Polynomial.** We first draw the number of features $d \sim \text{Unif}\{10, \dots, 20\}$ and sample covariate vectors $\mathbf{x} \sim \text{Unif}[-2, 2]^d$. We then fix a maximum degree $\deg \in \{1, 2, 3, 4\}$, augment covariates with powers $\mathbf{x}_{\text{ext}} = (x_1, \dots, x_d, x_1^2, \dots, x_d^{\deg})$, sample weights $\mathbf{w}_{\mu_0}, \mathbf{w}_{\mu_1}, \mathbf{w}_T \sim \text{Unif}[-5, 5]^{d \times \deg + 1}$, and define

$$f(\mathbf{x}) = \mathbf{w}_T^\top \mathbf{x}_{\text{ext}}, \quad \mu_t(\mathbf{x}) = \mathbf{w}_{\mu_t}^\top \mathbf{x}_{\text{ext}} \text{ for } t \in \{0, 1\}. \tag{31}$$

Degrees 1, 2, 3, and 4 give the Linear, Quadratic, Cubic, and Quartic sub-families; each degree adds new terms and is therefore a super-set of all lower degrees. We train on one degree family and test on the others to probe generalization.

(b) **Sinusoidal.** We draw the number of features $d \sim \text{Unif}\{5, \dots, 10\}$ and sample covariate vectors $\mathbf{x} \sim \text{Unif}[-3, 5]^d$. We then sample weight vectors $\mathbf{w}_{\mu_0}, \mathbf{w}_{\mu_1}, \mathbf{w}_T \sim \text{Unif}[-10, 6]^d$, and a frequency $\omega \in \mathbb{R}^+$. We define the treatment logit function and the CEPOs as

$$f(\mathbf{x}) = \sin\big(\omega \left\{\mathbf{w}_T^\top \mathbf{x}\right\}\big) + \mathbf{w}_T^\top \mathbf{x}, \quad \mu_t(\mathbf{x}) = \sin\big(\omega \left\{\mathbf{w}_{\mu_t}^\top \mathbf{x}\right\}\big) + \mathbf{w}_{\mu_t}^\top \mathbf{x} \text{ for } t \in \{0, 1\}. \tag{32}$$

For training DGPs, we create three sub-families: Linear ($\omega = 0$), L1 ($\omega \in [0, 1]$) and L2 ($\omega \in (1, 2]$). For test-time DPGs, we use the following: Linear ($\omega = 0$), L1 ($\omega \in [0.5, 1]$), L2 ($\omega \in (1.5, 2]$), and L3 ($\omega \in (2.5, 3]$). This allows us to measure extrapolation to unseen frequencies. E.g., an L2-trained model has seen DGPs from L1 and L2, but not L3.

For a unit with covariates $\mathbf{x}$ and credibility level α, we say the true CATE is *covered* if $\tau(\mathbf{x})$ lies within the predicted $100(1-\alpha)\%$ interval obtained from samples of q_θ. Plotting Bayesian coverage against nominal levels of α yields the CATE calibration curve $\widehat{\text{cov}}_\tau(\alpha)$. As shown in Figure 7 (left), CausalPFN is reliably calibrated under in-distribution settings but becomes severely overconfident when evaluated on OOD DGPs ($\psi^\star \not\sim \pi$). This aligns with prior observations that neural models often exhibit pathological overconfidence under distribution shift (Guo et al., 2017; Ovadia et al., 2019).

To correct this, we apply a temperature parameter θ_T to the SoftMax that outputs the quantized CEPO-PPD $q_\theta(\cdot \mid \mathbf{x}, t, \mathcal{D}_C^{\text{obs}})[\ell]$ from the logits of the model. Typically, one can tune θ_T to minimize the calibration error; however, in our case, direct CATE calibration is impossible because $\tau(\mathbf{x})$ is never observed at test-time. Instead, we introduce the *regression calibration* for observational data: an observed triple $(t, \mathbf{x}, y)$ is covered by the predicted credible interval when y lies inside the model's predicted interval for the CEPO-PPD. With that in mind, we let $\widehat{\text{cov}}_\mu(\alpha)$ and $\widehat{\text{cov}}_\tau(\alpha)$ denote the Bayesian coverage at level α for the regression and CATE calibration curves, respectively, and define

$$\text{ICE}_\mu := \int_0^1 \left(\widehat{\text{cov}}_\mu(\alpha) - \alpha\right) \, \mathrm{d}\alpha, \text{ and} \qquad \text{ICE}_\tau := \int_0^1 \left(\widehat{\text{cov}}_\tau(\alpha) - \alpha\right) \, \mathrm{d}\alpha, \tag{33}$$

as the integrated coverage error (ICE) for regression and CATE (negative values = overconfidence).

Ideally, we do not expect $\widehat{\text{cov}}_\mu$ to be calibrated: regression intervals combine epistemic uncertainty of the CEPO with the irreducible (aleatoric) noise in Y, so ICE_μ is biased downward. Still, it holds a useful signal. Across all model–DGP pairs in Figure 7 (middle) we consistently observe $\text{ICE}_\mu \leq \text{ICE}_\tau$: the regression curve sits at or below the CATE curve. While ICE_τ is inaccessible without having the true CATE, ICE_μ is computable from observational data. Consequently, temperature-scaling the logits to lift $\widehat{\text{cov}}_\mu$ to the diagonal also calibrates, or at worst makes conservative, the CATE intervals. We thus tune θ_T by grid search to drive ICE_μ to zero using a 5-fold calibration on the observational data. The calibrated

curves in Figure 7 (right) on the unseen test-set confirm that, after temperature scaling, overconfidence disappears and the model reliably knows when it does not know.

H Related Work

Single–dataset Estimators. Common methods for causal effect estimation are trained and applied on a single dataset. Representative examples include the X-, S-, DR-, and RA-Learners, as well as IPW and DML (Battocchi et al., 2019). Alongside these approaches, several neural variants such as TARNet (Shalit et al., 2017), DragonNet (Shi et al., 2019), CEVAE (Louizos et al., 2017), and NCMs (Xia et al., 2021; 2022) have been proposed; however, all of them still require per-dataset training and do not amortize across various datasets.

Amortized Causal Inference. Amortized causal inference methods train a *single* network to map observational data to causal quantities across *multiple* DGPs. Existing methods either first recover a causal graph representing the observational data and then compute interventions on that graph (Scetbon et al., 2024; Mahajan et al., 2024a), mirroring ideas from causal-discovery (Peters et al., 2014; Zheng et al., 2018; Khemakhem et al., 2021; Lorch et al., 2022; Ke et al., 2023; Kamkari et al., 2023), while others infer effects end-to-end (Nilforoshan et al., 2023; Zhang et al., 2023; Bynum et al., 2025). While all these methods can conceptually be used for amortized causal effect estimation, none of them provide a ready-to-use estimator that can surpass the specialized single-dataset estimators on standard benchmarks. They either rely on synthetic datasets as proof-of-concept or require multiple datasets similar to the one given at inference to train their estimators. Contrary to prior work, our method is trained once and produces causal effects without any access to, or adaptation on, the test-time DGPs. CausalPFN, through large-scale training, yields out-of-the-box performance that surpasses the specialized single-dataset estimators, setting a new milestone for amortized methods.

Scaling In-Context Transformers. In-context learning with transformers has shown impressive results across a range of domains (Brown et al., 2020; Xie et al., 2022; Coda-Forno et al., 2023; Dong et al., 2024; Vetter et al., 2025). Although the underlying mechanisms responsible for this success remain an active area of research (Akyürek et al., 2023; Dai et al., 2023; Olsson et al., 2022; Von Oswald et al., 2023; Li et al., 2023b; Yadlowsky et al., 2023; von Oswald et al., 2023; Bai et al., 2023; Peyrard & Cho, 2025), increasing model size and training data have consistently and undoubtedly led to stronger performance. This success has recently extended to tabular prediction with models such as TabPFN (Hollmann et al., 2023; 2025), TabDPT (Ma et al., 2024), and TabICL (Qu et al., 2025), which are trained on broad prior distributions and perform well on real-world data without fine-tuning. CausalPFN complements these works, demonstrating that—with sufficient scale and training—in-context learning can also be effectively adapted to causal inference.