DATA-INFORMED ADAPTATION OF FOUNDATION MODELS FOR CAUSAL EFFECT ESTIMATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Due to the fundamental problem of causal inference – the impossibility of observing counterfactual outcomes – there is always a gap between the training goal (outcome, propensity) and the inference target (CATE, ATE). Traditional causal effect estimation revolves around setting assumptions, selecting identification strategies, and choosing causal estimators, while more recently, Causal Foundation Models (CFMs), built on Prior-Data-Fitted Networks (PFNs), ease this process by pretraining on massive synthetic prior datasets. Despite their success, such amortization makes causal inference less transparent, obscuring which assumptions the model implicitly relies on for a given dataset. The high diversity of real-world data-generating processes (DGPs) and Structural Causal Models (SCMs) makes it hard to know, a priori, which causal inference strategy is compatible with a given dataset. In this work, we focus on this compatibility between an observational dataset and method-specific assumptions, and find that geometric characteristics of data embeddings – specifically, the variogram – provide a useful signal of whether a given causal estimation strategy is appropriate. We develop MoE-Learner, a mixture-of-causal-experts framework that embeds a dataset under multiple assumption-specific views and uses the variogram as a counterfactual-free router to assess each expert's validity. Rather than eliminating causal assumptions, this shifts their validation from a fixed, a priori choice to a criterion that can be checked directly from data geometry. MoE-Learner achieves the best average rank on benchmark datasets (IHDP, ACIC, Lalonde). Being orthogonal to (causal) foundation models, our method builds a bridge from data characteristics to causal estimation strategy rather than proposing another CFM.

1 INTRODUCTION

The scientific exploration of the world can be summarized in three steps: (1) finding connections between objects (correlation); (2) identifying causation within these correlations (causal inference) (Pearl, 2009; Pearl & Mackenzie, 2018); and (3) mechanistic explanation (Machamer et al., 2000; Craver, 2007; Woodward, 2005). Artificial Intelligence (AI), especially Deep Learning (DL) and transformer-based Large Language Models (LLMs), has achieved remarkable success in correlation prediction – e.g., linking images to their labels or predicting the next word – but a gap between correlation and causation remains (Kıcıman et al., 2023; Schölkopf et al., 2021).

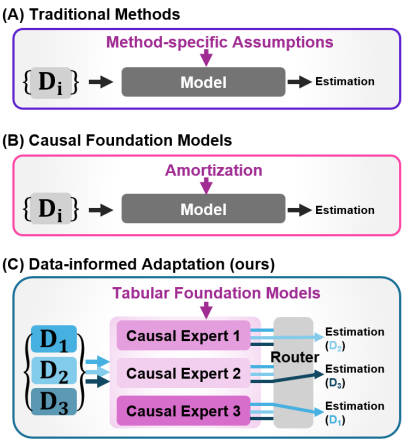


Figure 1: Comparison of different paradigms.

Causal effect estimation aims at identifying whether an action (e.g., an educational training or a drug treatment) has a causal effect on a population (e.g., income after training, or survival rate after treatment) (Rubin, 1974; Rosenbaum & Rubin, 1983). Traditionally, assumptions play a central role in causal inference, where bespoke estimators are tailored to specific settings in which certain assumptions hold. Beyond the common assumptions of Unconfoundedness/Exchangeability, Positivity/Overlap, and

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

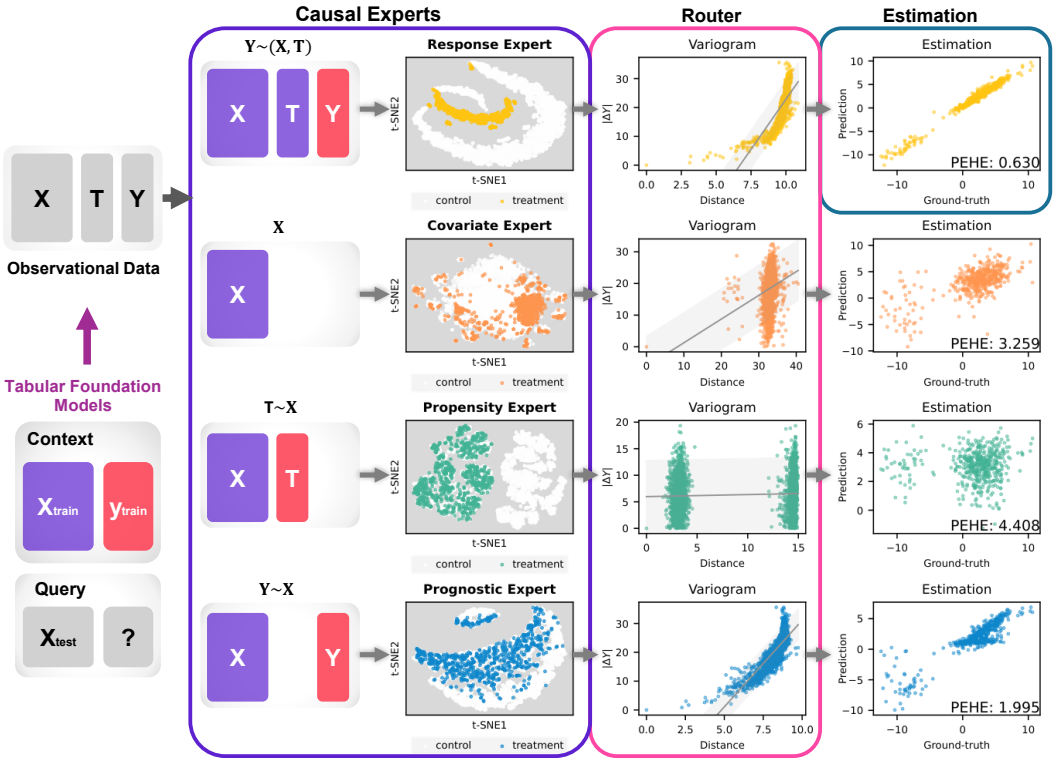


Figure 2: Illustration of the MoE-Learner workflow using ACIC data as the example. The observational data $(\boldsymbol{X}, T, Y)$ is converted to four kinds of embeddings by modeling different relationships $(Y \sim (\boldsymbol{X}, T), Y \sim \boldsymbol{X}, T \sim \boldsymbol{X}$, and a covariate modeling of $\boldsymbol{X})$. Each expert makes its own causal estimation. The Router selects the expert by evaluating the variograms of the embeddings.

SUTVA (Rubin, 1980; Rosenbaum & Rubin, 1983), each estimator typically carries additional, method-specific assumptions: for example, S-Learner assumes the response surfaces for treatment (Y_1) and control (Y_0) share a common functional form, whereas T-Learner models them as distinct functions (Künzel et al., 2019). Such an assumption need not hold for a given dataset, and a mismatch can lead to a biased estimate (Athey & Imbens, 2016).

Traditional causal inference therefore consists of two steps (Pearl, 2009; Rubin, 2008): (I) Causal modeling: identifying cause-and-effect mechanisms, isolating confounding, and making assumptions (Pearl, 2010; Rosenbaum & Rubin, 1983). (II) Model specification: choosing estimators, determining functional forms, and tuning hyperparameters (Künzel et al., 2019; Chernozhukov et al., 2018). Recently, Prior-Data-Fitted Networks (PFNs) (Hollmann et al., 2023; Helli et al., 2024), e.g., TabPFN (Hollmann et al., 2025) and TabICL (Qu et al., 2025), have introduced a groundbreaking new paradigm for tabular learning, in which a transformer is pretrained on massive synthetic tables, achieving consistently strong performance across diverse tabular data and tasks (Nagler, 2023; Adriaensen et al., 2023). Inspired by PFNs, Causal Foundation Models (CFMs) (Stith et al., 2026) have since been proposed – e.g., CausalPFN (Balazadeh Meresht et al., 2026), CausalFM (Ma et al., 2025), and Do-PFN (Robertson et al., 2025) – which are pretrained on diverse data-generating processes (DGPs) and Structural Causal Models (SCMs), and then perform inference for a new dataset by integrating over the resulting Bayesian posterior.

However, evidence suggests that CFMs, which estimate the treatment effect directly from an input dataset in a single forward pass, bypassing step-by-step intermediate mechanisms such as propensity weighting (IPW) or g-computation, can perform suboptimally – particularly on out-of-distribution real-world medical data that differs substantially from the prior distribution of SCMs used in training (Mourao et al., 2026).

In this work, we ask the following research questions regarding current challenges for causal effect estimation: (RQ1) Each estimator requires a set of model-specific assumptions. However, do they

apply to the data? To what extent? (RQ2) When settling on causal assumptions, which estimation strategy should be selected? Can a strategy adapt flexibly to the data? (RQ3) In practice, how can a general-purpose tabular foundation model be retrofitted into a causal estimator with minimal causal assumptions?

In the following sections, we first reveal the connections from causal inference to Prior-Data-Fitted Networks, Mixture of Experts, and Test-Time Adaptation, and then introduce a principled framework.

2 Background and Related Work

Causal Effect Estimation. Many causal effect estimation methods have been developed over the past several decades. Beyond common assumptions such as unconfoundedness and positivity, each method typically relies on additional assumptions, which we refer to as method-specific assumptions. For example, IPW (Rosenbaum & Rubin, 1983) relies entirely on a correctly specified propensity model and is highly sensitive to positivity violations. In contrast, methods such as RA-Net (Curth & van der Schaar, 2021) and BART (Hill, 2011) perform direct regression adjustment without any propensity correction, leaving them exposed to covariate shift between treatment arms. Other methods, such as DragonNet (Shi et al., 2019) and TarNet (Shalit et al., 2017), jointly model both arms through a shared representation and incorporate propensity estimation, implicitly assuming that a representation sufficient for predicting treatment is also sufficient for predicting the outcome. While these varied and often sophisticated assumptions pose significant challenges in selecting a proper method for a real-world dataset, formulating a coherent and correct set of assumptions itself requires profound domain expertise (Pearl, 2009; Imbens & Rubin, 2015). More importantly, once a set of assumptions is fixed, whether it aligns with the observed data – and to what extent – is usually left unexamined (Kang & Schafer, 2007; Alaa & Van Der Schaar, 2019), so the strategy remains fixed without adaptation to the real-world data.

Key Takeaway: For traditional causal inference methods, method-specific assumptions are typically tailored to the method rather than adaptive to the data.

Algorithm Selection for Causal Inference. A separate line of work addresses this fixation by treating estimator choice itself as a selection problem (Schuler et al., 2018; Nie & Wager, 2021; Alaa & Van Der Schaar, 2019). Since CATE estimators cannot be validated against ground-truth counterfactuals, prior work instead relies on *surrogate metrics*, such as R-scorer or doubly-robust proxy losses, computed on a held-out validation set of observed outcomes (Foster & Syrgkanis, 2023; Lan & Syrgkanis, 2024): Mahajan et al. (2024) benchmark 34 such metrics, using them to select among CATE estimators based on their agreement with held-out observed outcomes. This approach, however, still depends on a labeled validation split, and in small or imbalanced causal datasets, setting one aside can materially reduce the effective sample size available for estimating treatment effects, particularly for the smaller treatment arm.

Foundation Models for Causal Effect Estimation. Different causal foundation models amortize estimation over a prior of SCMs with varying assumptions to ease causal inference at test time (Stith et al., 2026; Bellot & Dhir, 2026). CausalPFN (Balazadeh Meresht et al., 2026) assumes the user has specified an identification regime restricted to the back-door criterion; CausalFM (Ma et al., 2025) generalizes this to back-door, IV, and front-door regimes. In contrast, Do-PFN (Robertson et al., 2025), MACE-TNP (Dhir et al., 2026), and ACTIVA (Sauter et al., 2025) instead specify a broad prior spanning many plausible SCMs. However, this amortization makes causal inference less transparent: among the SCMs in the pretraining prior, which align well with a given test dataset, which diverge from it, and how should such disagreements be reconciled? A prior that is too broad risks poor specificity, while one that is too narrow (e.g., models trained specifically for front-door or IV settings) risks poor coverage of real-world regimes. Reuter et al. (2026) address this by conditioning on prior causal information such as partial ancestral graphs, though such information is not always available.

Mixture of Experts for Causal Inference. Given heterogeneous causal structures, mixed assumptions, and diverse estimator choices, the mixture-of-experts (MoE) architecture is a natural fit (Shazeer et al., 2017; Künzel et al., 2019). In LLMs, MoE learns a dispatching mechanism (the router) that assigns each token to an "expert" – a dense subnetwork – enabling both test-time acceleration and smart allocation to the expert best suited to the token (e.g., grammar, math, coding) (Fedus et al., 2022). Adapting MoE to causal effect estimation, however, is not straightforward: unlike next-token

prediction, which provides a self-supervised training signal for the router, the fundamental problem of causal inference – the impossibility of observing both treated and untreated outcomes for the same unit – makes it impossible to train the router and the experts jointly, end-to-end, for CATE/ATE prediction. CausalMoE (Liu et al., 2026), a foundation model for multimodal Granger causal discovery, designs experts for heterogeneous modalities in time-series causal discovery (e.g., semantic, visual, temporal, and frequency experts), but it remains unclear how to adapt MoE to the differing, and possibly unknown, causal structures and regimes relevant to effect estimation.

Test-Time Adaptation of Pretrained Models. Deep neural networks often suffer significant performance degradation under domain shift. Test-Time Adaptation (TTA) methods address this by adapting a trained model to the test domain without labeled test data or access to information from the training procedure (e.g., the training dataset or its feature statistics) (Sun et al., 2020; Wang et al., 2020; Schneider et al., 2020; Gozeten et al., 2025). Since PFNs are pretrained solely on synthetic simulation data, a distribution shift between the prior's DAGs or SCMs and real-world ones is expected. We identify causal effect estimation as fundamentally akin to TTA: an ideal model, or meta-model, must select a proper identification strategy or estimator without knowing the ground-truth CATE/ATE, the true underlying DGP or SCM, or having access to the pretraining prior. In TTA, since no ground-truth label is available at test time, methods instead exploit data-dependent information such as nearest-neighbor structure (Jang et al., 2022) and local geometry (Lei et al., 2025). Recent work has moved toward prior-free tabular TTA (He & Shi, 2026; Lee, 2026); however, the fundamental problem of causal inference renders such methods inapplicable here, since they address a shift between training and test distributions *for the same predictive task*, not the more basic discrepancy between the training goal (outcome or propensity prediction) and the testing goal (CATE/ATE estimation).

3 Overview of Contributions

In real-world datasets, the DGPs and SCMs are highly diverse, making it unclear which causal inference strategy to use. We address this in two steps, corresponding to our research questions. First (RQ3), we unify and categorize causal inference strategies into four modeling types: response modeling ($Y \sim (X, T)$), propensity modeling ($T \sim X$), prognostic modeling ($Y \sim X$), and a covariate embedding (X). We design a "causal expert" for each modeling type: for the first three, an expert places the corresponding variables into the context of a tabular foundation model and extracts an embedding vector for each unit; the covariate expert instead uses a standardized embedding of X itself – a label-free, RAG-like reference view, rather than one learned through the tabular foundation model. Each expert therefore not only converts units into embeddings, but also implicitly encodes the assumptions underlying its modeling choice (e.g., the propensity expert assumes that treatment assignment is related to the covariates X). This retrofits a general-purpose tabular foundation model into a suite of causal estimators, each carrying minimal, explicit assumptions.

While the causal experts encapsulate multiple causal inference strategies, the next challenge is how to route to the proper expert for a given dataset (RQ1, RQ2). Unlike traditional MoE, we lack a direct training signal for the router: there is no observation of counterfactual outcomes, no true CATE/ATE performance for any expert, and no separate validation set. To address this, we borrow the idea of the variogram from spatial statistics – a function measuring how much sample labels differ from one another as their relative distance increases. Analogously, if a causal expert's embedding is valid, nearby units in that embedding space should also share similar outcome values; we use this local smoothness as a ground-truth-free diagnostic of how well an expert's implicit assumptions fit the data (RQ1), and route to the expert whose embedding best satisfies it (RQ2). Empirically, we find that different experts are activated for different datasets.

The framework relies on standard causal identification assumptions, but shifts the choice among method-specific estimation strategies from a fixed, a priori commitment to a criterion determined by the geometry of the joint data-assumption embeddings. We do not rely on empirical performance on a held-out validation set; instead, our routing criterion is grounded in a statistical property of the embedding space itself, making the selection more directly interpretable than validation-based alternatives.

Our method is orthogonal to causal foundation models: we do not develop yet another CFM (Figure 1). Instead, we show how to adapt a general-purpose tabular foundation model into a flexible causal

effect estimator via causal experts and a router without counterfactual supervision, achieving better performance and improved interpretability.

4 METHODOLOGY

4.1 PRELIMINARY: CAUSAL EFFECT ESTIMATION

We work within the potential outcomes framework (Rubin, 2005) and observe samples of random variables (X, T, Y), where $X \in \mathbb{R}^n$ are the observed covariates with dimension n, T is the treatment assignment drawn from a finite treatment set (binary treatments $T \in \mathbb{T} = \{0, 1\}$ here), and $Y \in \mathbb{R}$ is the outcome of interest. For every $t \in \mathbb{T}$, $Y_t \in \mathbb{R}$ is the potential outcome under treatment t, while the observed (factual) outcome is Y. The observational data $\{x, t, y\}$ are drawn from an unknown joint distribution $P(X, T, \{Y_t\}_{t\in\mathbb{T}})$.

Causal inference typically requires the propensity (treatment assignment) distribution $\pi_t(x)$ and the expected potential outcomes $\mu_t(x)$, commonly referred to together as the nuisance parameters $\eta = (\mu_0, \mu_1, \pi)$:

$$\pi_t(x) = P[T = t \mid X = x], \tag{1}$$
$$\mu_t(x) = \mathbb{E}[Y_t \mid X = x]. \tag{2}$$

Our targets of inference are two quantities: the conditional average treatment effect (CATE), which represents the average effect of intervention $(Y_1 - Y_0)$ on the subpopulation with covariates $X = x$, and the average treatment effect (ATE), which measures the average difference in outcomes between the entire population receiving treatment versus the same population not receiving it:

$$\text{CATE}: \quad \tau(x) = \mathbb{E}[Y_1 - Y_0 \mid X = x] = \mu_1(x) - \mu_0(x), \tag{3}$$
$$\text{ATE}: \quad \lambda = \mathbb{E}[Y_1 - Y_0] = \mathbb{E}[\mu_1(X) - \mu_0(X)]. \tag{4}$$

4.2 CAUSAL EXPERTS

Causal Embeddings via Tabular Foundation Models. Unlike an embedding from a standard deep neural network, where the embedding vector is simply the penultimate-layer activation, extracting an embedding $Z = g(Q; C)$ from a Prior-Data-Fitted Network – the contextualized hidden state for query Q after the transformer layers g – requires specifying a context C (Ye et al., 2025). For example, to extract the embedding $\boldsymbol{Z}_{XTY}$ from a TabPFN model predicting Y from the concatenation of the dataset matrix $\boldsymbol{X}$ and T, the query is $Q = [\boldsymbol{X}, T]$ and the context is $C = ([\boldsymbol{X}, T], Y)$, giving $\boldsymbol{Z}_{XTY} = g([\boldsymbol{X}, T]; ([\boldsymbol{X}, T], Y))$. Different choices of query and context correspond to different modeling relationships underlying traditional causal estimators. We exploit this correspondence to design four causal experts, each built from a different query-context pairing:
(1) **Response Expert** ($\boldsymbol{Z}_{XTY} = g([\boldsymbol{X}, T]; ([\boldsymbol{X}, T], Y))$): As in S-Learner, BART, DragonNet, RA-Net, and TarNet, treatment is modeled as a covariate within an outcome model, i.e., $Y \sim F(\boldsymbol{X}, T)$. The response expert embeds this relationship by taking $[\boldsymbol{X}, T]$ as the query and Y as the context, yielding a contextualized representation that jointly encodes covariates, treatment assignment, and their interaction with the outcome.
(2) **Propensity Expert** ($\boldsymbol{Z}_{XT} = g(\boldsymbol{X}; (\boldsymbol{X}, T))$): As in IPW, and as a component within DA-Learner, X-Learner, DragonNet, GRF, and Forest DML, a propensity model is estimated to reweight or adjust samples according to $T \sim F(\boldsymbol{X})$. The propensity expert embeds this relationship by taking $\boldsymbol{X}$ as the query and T as the context, producing a representation of each unit's position within the treatment-assignment mechanism, independent of the outcome.
(3) **Prognostic Expert** ($\boldsymbol{Z}_{XY} = g(\boldsymbol{X}; (\boldsymbol{X}, Y))$): As in T-Learner, X-Learner's first stage, DA-Learner, and the prognostic function in BCF, an outcome model $Y \sim F(\boldsymbol{X})$ is estimated to capture the relationship between covariates and outcome, pooling all samples into a shared representation space.
(4) **Covariate Expert** ($\boldsymbol{Z}_X = \frac{\boldsymbol{X} - \mu_X}{\sigma_X}$): The standardization of the input covariates $\boldsymbol{X}$, containing no information about T or Y. This plays a role analogous to the embedding step in retrieval-augmented generation (RAG): just as RAG embeds a query to retrieve semantically similar documents, the covariate expert embeds each unit to retrieve similar units in covariate space, providing a matching-style signal that complements the outcome- and treatment-aware experts above.

For each query unit $\boldsymbol{x}$, since we want to predict potential outcomes under both control and treatment, each unit is embedded into two representations, $\boldsymbol{z}_0$ and $\boldsymbol{z}_1$. For the Response Expert, these two embeddings differ: $\boldsymbol{z}_t = g([\boldsymbol{x}, t]; ([\boldsymbol{X}, T], Y)), t \in \{0, 1\}$. For the other three experts, $\boldsymbol{z}_0$ and $\boldsymbol{z}_1$ share the same embedding, since their query does not depend on t.

Key Takeaway: We embed a dataset under several causal experts in parallel, so that no single method-specific assumption is committed to a priori, while each expert individually remains minimally assumption-laden.

Top-K Attention from Query to Context. We refer to the generated embeddings for the TabPFN query Q and context C as the query and context embeddings, respectively. For each expert, given a context set $\boldsymbol{Z}_t = \{\boldsymbol{z}_t^{(i)}\}_{i=1}^{N_t}$ of feature-label pairs, where $t \in \{0, 1\}$ denotes the control/treatment group, we cast CATE prediction as an attention-based retrieval process. For a query embedding $\boldsymbol{z}_t$, we define a feature dissimilarity D between points $\boldsymbol{z}_t$ and $\boldsymbol{z}_t^{(k)}$ (in practice, cosine distance), and retrieve the top-K most similar context pairs $\{(\boldsymbol{z}_t^{(k)}, y_t^{(k)})\}_{k=1}^{K}$ – i.e., the K context points with smallest dissimilarity $D(\boldsymbol{z}_t, \boldsymbol{z}_t^{(k)})$ to the query. Attention scores over the retrieved neighborhood are assigned by a learnable weighting function $f_\theta : \mathbb{R} \to \mathbb{R}_{>0}$, implemented as a small MLP with Softplus activations and non-negative weight matrices, and normalized over the retrieved set:

$$a_t^{(k)} = \frac{f_\theta\big(D(\boldsymbol{z}_t, \boldsymbol{z}_t^{(k)})\big)}{\sum_{j=1}^{K} f_\theta\big(D(\boldsymbol{z}_t, \boldsymbol{z}_t^{(j)})\big)}, \qquad k = 1, \ldots, K.$$

The query label is then predicted as the attention-weighted average of the retrieved labels,

$$\hat{y}_t = \sum_{k=1}^{K} a_t^{(k)}\, y_t^{(k)}.$$

Unlike standard self-attention, we omit learned query/key projections: scores are computed directly from a fixed dissimilarity between raw feature embeddings, with f_θ as the only learned component of the attention mechanism. Finally, the CATE for unit $\boldsymbol{x}$ is predicted as

$$\hat{\tau}(\boldsymbol{x}) = \hat{y}_1 - \hat{y}_0.$$

4.3 Causal Router

Variogram: Data-Informed Evaluation of Expert Validity. When different experts implicitly encode different causal relationships, the natural next question is which expert to trust for a given dataset. Rather than committing a priori to a single, universally preferred causal estimator, MoE-Learner assesses each expert's validity directly from the geometry of its own embedding space. We borrow the idea of the *variogram* from geostatistics: a function that measures how the expected squared difference between two labels grows as the distance between their locations increases. Applied to a causal expert's embedding, the analogous question is whether units that are close in that expert's embedding space also share similar outcomes. If an expert's underlying assumption is well matched to the data, embedding proximity should track outcome similarity, so nearby units exhibit small label variation, and this variation should not shrink as embedding distance grows. An expert whose embedding shows no such relationship – where outcome similarity is unrelated to embedding proximity – encodes an assumption that fails to organize the data in a causally meaningful way, and should be trusted less.

Theoretical Analysis. We formalize this intuition for a single expert e and treatment group t; the group subscript t is suppressed below, as the analysis applies identically and independently to each group's embeddings. Let $\boldsymbol{z}^{(e)}$ denote the embedding produced by expert e, and let $f_e(\boldsymbol{z}) = \mathbb{E}[Y \mid \boldsymbol{z}^{(e)} = \boldsymbol{z}]$ be the regression function it induces. We model

$$Y = f_e(\boldsymbol{z}^{(e)}) + \varepsilon, \qquad \mathbb{E}[\varepsilon \mid \boldsymbol{z}^{(e)}] = 0, \quad \mathrm{Var}(\varepsilon) = \sigma^2,$$

and let D be the dissimilarity used for neighbor retrieval in $\boldsymbol{z}^{(e)}$-space (Section 4.2). We say f_e is β-*locally smooth* with respect to D if there exists $L_e > 0$ such that $|f_e(\boldsymbol{z}) - f_e(\boldsymbol{z}')| \le L_e D(\boldsymbol{z}, \boldsymbol{z}')^\beta$ for all $\boldsymbol{z}, \boldsymbol{z}'$ in the relevant support. For an independent pair $(\boldsymbol{z}_i^{(e)}, Y_i), (\boldsymbol{z}_j^{(e)}, Y_j)$ drawn from the same distribution, define the distance-label variogram

$$\gamma_e(d) = \tfrac{1}{2}\,\mathbb{E}\big[(Y_i - Y_j)^2 \mid D(\boldsymbol{z}_i^{(e)}, \boldsymbol{z}_j^{(e)}) = d\big].$$

Theorem 1 (Distance-Label Variogram Bound). *Under the model above, if f_e is β-locally smooth with respect to D with constant L_e, then for all d in the support of D,*

$$\sigma^2 \;\le\; \gamma_e(d) \;\le\; \sigma^2 + \tfrac{1}{2}L_e^2 d^{2\beta}.$$

The proof (Appendix B) decomposes $\gamma_e(d)$ into an irreducible noise term σ^2 and a term governed by the local variation of f_e, then bounds the latter using the smoothness condition. The theorem implies that as $d \to 0$, $\gamma_e(d)$ is squeezed toward the noise floor σ^2: for a valid expert, near-duplicate units in embedding space must have near-identical expected outcomes, up to noise. Note that the bound constrains the *ceiling* on label variation, not the exact shape of γ_e; it does not by itself force $\gamma_e(d)$ to increase in d. We therefore treat an increasing empirical trend between d and label variation as evidence *consistent with* local smoothness – and hence with a valid expert – rather than as a quantity derived directly from the theorem.

Correlation as an Empirical Surrogate for Expert Validity. Theorem 1 characterizes a population relationship between distance and label variation; it does not by itself yield a single scalar for comparing experts. As an empirical surrogate, for each expert e and group t we sample M anchor points from the context set $\boldsymbol{Z}_t^{(e)}$ via farthest-point sampling. For each anchor, we compute its embedding distance D and absolute label difference $|\Delta y|$ to every other point i in $\boldsymbol{Z}_t^{(e)}$, summarize their relationship with the Spearman rank correlation, and average these pairs across all M anchors

$$\rho_e^{(t)} = \tfrac{1}{M}\textstyle\sum_{k=1}^{M} \operatorname{corr}\big(\{d_{ki}\},\, \{|\Delta y_{ki}|\}\big).$$

A large, positive $\rho_e^{(t)}$ indicates that label variation tracks embedding distance – consistent with the local smoothness Theorem 1 predicts for a valid expert – while $\rho_e^{(t)} \approx 0$ indicates that expert e's embedding carries little information about outcome similarity for group t. The router selects, independently per group,

$$e_t^\star = \arg\max_{e \in \mathcal{E}} \; \rho_e^{(t)}, \qquad t \in \{0, 1\},$$

where $\mathcal{E}$ denotes the set of four causal experts. This refines the CATE prediction from Section 4.2 to use each group's selected expert:

$$\hat{\tau}(\boldsymbol{x}) = \hat{y}_1^{(e_1^\star)} - \hat{y}_0^{(e_0^\star)}.$$

Key Takeaway: If a causal expert's embedding is valid, nearby units in that embedding should also have similar outcomes – the idea underlying the variogram. The causal router scrutinizes this validity, independently per treatment group, and selects the expert for which it holds most strongly.

5 Experiments

Experiment Setup. To evaluate MoE-Learner, we conduct experiments on four benchmark datasets – IHDP (Ramey et al., 1992), ACIC 2016 (Dorie et al., 2019), Lalonde CPS/PSID (LaLonde, 1986) – each generated with a different degree of semi-synthetic construction to produce known counterfactuals for evaluation (see Appendix A.1 for details).

We compare against three categories of methods. (1) *Conventional causal estimators*: double machine learning (DML) (Chernozhukov et al., 2018; Athey et al., 2019; Foster & Syrgkanis, 2023); the doubly robust learner (DR-Learner) (Künzel et al., 2019; Kennedy, 2023); the T/S/X/DA learners (Battocchi et al., 2019); TarNet (Shalit et al., 2017), DragonNet (Shi et al., 2019), and RA-Net (Curth & van der Schaar, 2021; Curth, 2021); inverse propensity weighting (IPW) (Rosenbaum & Rubin, 1983); Bayesian additive regression trees (BART) (Hill, 2011; Chan et al., 2022); and generalized random forests (GRF) (Athey et al., 2019). (2) *Causal foundation models*: CausalPFN (Balazadeh Meresht et al., 2026), Do-PFN (Robertson et al., 2025), and CausalFM (Ma et al., 2025). (3) *General-purpose tabular foundation models*: TabPFN and TabICL. Performance is evaluated using the precision in estimation of heterogeneous effects (PEHE) for CATE and the relative ATE error for ATE (see Appendix A.2).

MoE-Learner Achieves Superior Performance on CATE and ATE. As shown in Table 1, the best and second-best methods vary considerably across datasets, and no single method wins consistently. The out-of-the-box tabular foundation models TabPFN and

Table 1: **Performance comparison for CATE and ATE.** The results are reported as the mean and standard error across all realizations of a dataset. The **best** and second best methods are highlighted. The IPW method is not applicable to CATE.

Method	PEHE (↓)					ATE Relative Error (↓)				
	IHDP	ACIC 2016	Lalonde CPS ($\times 10^3$)	Lalonde PSID ($\times 10^3$)	Avg. Rank	IHDP	ACIC 2016	Lalonde CPS	Lalonde PSID	Avg. Rank
MoE-Learner (ours)	0.80±0.15	0.64±0.08	9.05±0.02	**14.08±0.25**	**2.50**	0.22±0.04	0.02±0.00	0.19±0.01	0.12±0.02	**5.00**
Tabular Foundation Models										
TabPFN	**0.51±0.04**	**0.54±0.09**	9.38±0.05	20.48±0.09	4.00	0.20±0.04	**0.01±0.01**	0.24±0.17	0.71±0.18	6.50
TabICL	0.65±0.09	0.67±0.03	11.40±0.04	19.86±0.09	5.50	0.20±0.04	0.03±0.02	0.74±0.10	0.69±0.16	7.50
Causal Foundation Models										
CausalPFN	0.58±0.07	0.92±0.11	**8.96±0.02**	14.40±0.20	3.50	0.20±0.04	0.05±0.01	**0.13±0.01**	0.22±0.02	6.25
Do-PFN	6.06±0.94	4.25±0.68	12.01±0.03	20.91±0.14	13.00	0.57±0.95	0.67±0.04	0.87±0.01	0.92±0.01	15.25
CausalFM	8.71±0.90	5.01±0.53	12.87±0.16	22.70±0.29	15.50	1.24±0.01	1.02±0.02	1.00±0.00	1.00±0.00	16.75
Traditional Causal Inference Models										
T-Learner	1.73±0.30	0.76±0.07	9.22±0.04	15.16±0.46	4.75	0.21±0.04	0.03±0.01	0.24±0.02	0.16±0.03	5.75
DA-Learner	2.07±0.36	0.72±0.08	9.39±0.06	14.55±0.24	5.25	0.23±0.04	0.03±0.01	0.27±0.02	0.20±0.03	7.00
DragonNet	2.16±0.25	2.11±0.19	10.93±0.15	16.45±0.29	8.00	0.20±0.04	0.06±0.02	0.55±0.03	0.47±0.03	7.75
IPW	—	—	—	—	—	0.24±0.04	0.21±0.05	0.17±0.01	**0.10±0.01**	8.50
RA-Net	2.35±0.19	2.35±0.25	11.74±0.09	18.33±0.43	9.50	0.20±0.04	0.07±0.03	0.74±0.02	0.50±0.04	8.50
X-Learner	3.31±0.51	0.60±0.08	12.15±0.15	20.28±0.49	8.25	0.16±0.04	0.03±0.01	0.84±0.03	0.72±0.03	6.75
TarNet	1.82±0.14	2.20±0.21	12.88±0.02	19.19±0.18	10.50	0.20±0.04	0.05±0.02	1.00±0.00	0.78±0.01	10.50
S-Learner	2.57±0.41	0.85±0.13	12.66±0.05	21.80±0.18	11.00	0.20±0.04	0.03±0.01	0.97±0.01	0.90±0.02	9.00
BART	2.50±0.39	0.68±0.11	12.81±0.05	21.36±0.16	10.00	0.44±0.09	0.04±0.01	0.99±0.01	0.86±0.01	12.75
GRF	3.67±0.61	1.32±0.30	12.33±0.06	22.91±0.17	12.50	0.18±0.03	0.07±0.02	0.82±0.02	0.85±0.02	10.00
Forest DML	4.53±0.73	1.48±0.31	12.95±0.04	22.99±0.15	15.00	**0.08±0.01**	0.05±0.01	1.03±0.01	1.05±0.01	11.50
Forest DR Learner	4.02±0.67	1.34±0.29	15.98±0.68	22.78±0.54	14.25	0.17±0.03	0.04±0.02	1.20±0.23	3.64±2.78	11.75

TabICL are themselves very competitive, achieving average ranks of 4.00 and 5.50 on CATE, and 6.50 and 7.50 on ATE, respectively. CausalPFN attains the second-best average rank (3.50) on CATE, but underperforms on ATE except on the Lalonde CPS dataset. Traditional estimators can also achieve competitive, even best-in-class, performance in specific cases (e.g., IPW on the Lalonde datasets), but their performance is uneven across datasets, which lowers their average rank overall. MoE-Learner attains the best average ranks among all methods, 2.50 on CATE and 5.00 on ATE.

Average rank, however, only partially captures performance: when MoE-Learner is not the best or second-best (e.g., CATE on IHDP and ACIC), its performance remains on par with the top methods (Figures 5 and 6). This absence of a "winner-take-all" method suggests that real-world SCMs are diverse enough that no single, fixed estimation strategy excels across all scenarios (Figure 3 shows this directly in the varying experts activated across datasets), underscoring the need for methods that adapt to the data at hand. Our approach opens a new avenue toward this end, without requiring a choice of evaluation metric on a held-out validation set. MoE-Learner also exhibits a better uncertainty quantification result (see Appendix C for details).

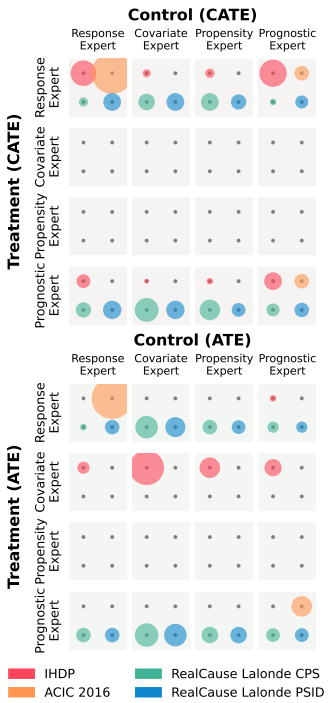


Figure 3: Expert activation percentages for CATE (upper panel) and ATE (lower panel).

Analysis of Causal Experts and Router. For the MoE architecture to be meaningful, different experts should be selected across different datasets or tasks, and the selected expert should outperform both a randomly chosen expert and the simple average of all experts. We summarize several key findings toward this end.

(1) *The variogram correlation provides an empirical signal for estimation performance.* As shown in Figure 7 (CATE on IHDP), for the Response Expert the absolute outcome difference increases almost monotonically with embedding distance, indicating that this expert captures the $Y \sim F(\boldsymbol{X}, T)$ relationship along a lower-dimensional manifold; correspondingly, its PEHE is also the lowest among experts. In contrast, for experts with weaker or noisier distance-outcome relationships

(e.g., the Prognostic Expert, whose variogram trend is positive but with higher variance), PEHE is correspondingly worse.

(2) *For the same dataset, different tasks activate different experts.* Still on IHDP, but now for the ATE task, the Response Expert exhibits poor variogram correlation, and the Covariate Expert is selected instead (Figure 11). This selection is corroborated by standalone evaluation: the Covariate Expert (0.22 ± 0.42) outperforms the Response Expert (0.28 ± 0.48) on ATE error, the reverse of their ranking on CATE (Table 4). We provide variograms for all the experts on all our datasets in Figures 7 to 14 for the inspection of different model behavior on different datasets.

(3) *For the same dataset, different treatment arms activate different experts.* We plot expert activation percentages for control and treatment outcome estimation, across both CATE and ATE, in Figure 3. Activation patterns vary substantially across datasets, arms, and tasks: for treatment-outcome estimation, only the Response and Prognostic Experts are ever activated, whereas for control-outcome estimation all four experts are activated, at differing rates. This indicates that MoE-Learner does not collapse onto a single expert across our benchmarks, but instead adapts its choice to the dataset, task, and arm.

(4) *Expert activation may reflect (unknown) aspects of the data-generating process.* The RealCause Lalonde datasets are constructed semi-synthetically, with treatment and control groups artificially concatenated rather than drawn from a real experiment. Consistent with this, the Prognostic Expert (which ignores T) is activated at a higher rate on these datasets, suggesting a treatment-assignment process that differs from IHDP or ACIC (see Section A.1 for details). This suggests our causal embeddings, variogram, and routing mechanism may be useful for probing aspects of the underlying causal mechanism itself, particularly when prior knowledge about the DGP is limited.

Ablation Experiments. (1) Our router is 1-sparse – exactly one expert is selected per treatment arm for each dataset realization – so for the router to be meaningful, it should outperform an unweighted average ensemble of all experts. Table 3 confirms this: the router substantially outperforms the ensemble, except on CATE for the Lalonde datasets, where the two are comparable. This corroborates that the routing mechanism removes the need to manually select a good, but a priori unknown, causal estimation strategy. (2) In the causal router, we use Spearman's ρ to measure the rank correlation between embedding distance and outcome difference. As a comparison, we evaluate alternative selection metrics – Pearson correlation (linear) and the intrinsic dimension of the embedding itself – both of which underperform Spearman's ρ (Table 5). This indicates that the choice of routing metric matters for selecting the proper expert, and that Spearman's rank correlation is a more robust indicator than linear correlation or geometric dimensionality alone. (3) *Ablation on the expert architecture.* Each causal expert in MoE-Learner uses simplified attention with a learnable weighting MLP (Section 4.2). Replacing this with a fixed K-Nearest-Neighbor (KNN) regressor generally performs worse (Table 4), confirming the value of the learned weighting function over uniform neighbor averaging.

6 Discussion and Conclusion

Causal effect estimation has traditionally required committing to a set of assumptions and an estimator largely independent of the dataset at hand, leaving the fit between assumption and data unexamined (Section 2). In this work, we integrate four causal estimation strategies – Response, Covariate, Propensity, and Prognostic – as parallel causal experts, so that no single method-specific assumption need be committed to a priori. A data-informed router, based on the distance-label variogram, then selects among these experts per dataset, task, and treatment arm, offering a more data-adaptive alternative to the fixed inference process of amortized causal foundation models.

In contrast to algorithm selection, we treat causal estimator selection as a dataset-dependent adaptation problem and show that characteristics observable directly from the data can inform which CATE estimation strategy is likely to perform well, without requiring a separate outcome-based validation set. Because our criterion is derived from the geometry of the data itself rather than from held-out predictive performance, it is also more directly interpretable – one can inspect *why* an expert was judged valid, rather than only observing that it scored well on a validation split. We discuss limitations and future direction in Appendix D. The code is available on GitHub.

AI USE STATEMENT

In this work, we used generative AI tools for grammar checking/correction and for retrieval and discovery (e.g., finding related work). We have not used generative AI tools for research ideation, execution, experimentation, or writing, and the rest of the required disclosure tasks are not applicable to this work. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

All the code, scripts, Jupyter Notebooks, and data files that are necessary for the reproduction of the work will be made publicly available on GitHub.

REFERENCES

Steven Adriaensen, Herilalaina Rakotoarison, Samuel Müller, and Frank Hutter. Efficient bayesian learning curve extrapolation using prior-data fitted networks. *Advances in Neural Information Processing Systems*, 36:19858–19886, 2023.

Ahmed Alaa and Mihaela Van Der Schaar. Validating causal inference models via influence functions. In *International Conference on Machine Learning*, pp. 191–201. PMLR, 2019.

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

Luc Anselin. Local indicators of spatial association—lisa. *Geographical analysis*, 27(2):93–115, 1995.

Susan Athey and Guido Imbens. Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113(27):7353–7360, 2016.

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

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

Keith Battocchi, Eleanor Dillon, Maggie Hei, Greg Lewis, Paul Oka, Miruna Oprescu, and Vasilis Syrgkanis. EconML: A Python Package for ML-Based Heterogeneous Treatment Effects Estimation. https://github.com/py-why/EconML, 2019. Version 0.15.0.

Alexis Bellot and Anish Dhir. Foundation models for partial causal identification, 2026. URL https://arxiv.org/abs/2608.20841.

Marianne Bertrand, Esther Duflo, and Sendhil Mullainathan. How much should we trust differences-in-differences estimates? *The Quarterly journal of economics*, 119(1):249–275, 2004.

Stephanie Chan, Adam Santoro, Andrew Lampinen, Jane Wang, Aaditya Singh, Pierre Richemond, James McClelland, and Felix Hill. Data distributional properties drive emergent in-context learning in transformers. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1), 2018. doi: 10.1111/ectj.12097.

Carl F Craver. *Explaining the brain: Mechanisms and the mosaic unity of neuroscience*. Clarendon press, 2007.

Alicia Curth. 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.

Alicia Curth and Mihaela van der Schaar. On inductive biases for heterogeneous treatment effect estimation. In *Advances in Neural Information Processing Systems*, volume 34, pp. 15883–15894, 2021.

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

Vincent Dorie, Jennifer Hill, Uri Shalit, Marc Scott, and Dan Cervone. Automated versus do-it-yourself methods for causal inference: Lessons learned from a data analysis competition. *Statistical Science*, 34(1):43–68, 2019.

William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39, 2022.

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

Halil Alperen Gozeten, M Emrullah Ildiz, Xuechen Zhang, Mahdi Soltanolkotabi, Marco Mondelli, and Samet Oymak. Test-time training provably improves transformers as in-context learners. *Proceedings of machine learning research*, 267:20266, 2025.

Rundong He and Jieming Shi. Prior-free tabular test-time adaptation. In *International Conference on Learning Representations*, volume 2026, pp. 69478–69494, 2026.

Kai Helli, David Schnurr, Noah Hollmann, Samuel Müller, and Frank Hutter. Drift-resilient TabPFN: In-context learning temporal distribution shifts on tabular data. *Advances in Neural Information Processing Systems*, 37:98742–98781, 2024.

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

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

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

Guido W Imbens and Thomas Lemieux. Regression discontinuity designs: A guide to practice. *Journal of econometrics*, 142(2):615–635, 2008.

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

Minguk Jang, Sae-Young Chung, and Hye Won Chung. Test-time adaptation via self-training with nearest neighbor information. *arXiv preprint arXiv:2207.10792*, 2022.

Joseph D. Y. Kang and Joseph L. Schafer. Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science*, 22(4):574–580, 2007. ISSN 08834237.

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

Emre Kıcıman, Robert Ness, Amit Sharma, and Chenhao Tan. Causal reasoning and large language models: Opening a new frontier for causality. *arXiv preprint arXiv:2305.00050*, 2023.

Samory Kpotufe. k-nn regression adapts to local intrinsic dimension. *Advances in neural information processing systems*, 24, 2011.

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

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

Hui Lan and Vasilis Syrgkanis. Causal q-aggregation for cate model selection. In *International Conference on Artificial Intelligence and Statistics*, pp. 4366–4374. PMLR, 2024.

Seunghan Lee. Mitigating label shift in tabular in-context learning via test-time posterior adjustment. *arXiv preprint arXiv:2605.04363*, 2026.

Mingxi Lei, Chunwei Ma, Meng Ding, Yufan Zhou, Ziyun Huang, and Jinhui Xu. Ttvd: Towards a geometric framework for test-time adaptation based on voronoi diagram. In *International Conference on Learning Representations*, volume 2025, pp. 30339–30361, 2025.

Elizaveta Levina and Peter Bickel. Maximum likelihood estimation of intrinsic dimension. *Advances in neural information processing systems*, 17, 2004.

Bo Liu, Di Dai, Jingwei Liu, Jiarui Jin, Xiaocheng Fang, Guangkun Nie, Hongyan Li, and Shenda Hong. Causalmoe: A billion-scale multimodal foundation model for granger causal discovery with pattern-routed heterogeneous experts. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pp. 11422–11431, 2026.

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

Peter Machamer, Lindley Darden, and Carl F Craver. Thinking about mechanisms. *Philosophy of science*, 67(1):1–25, 2000.

Divyat Mahajan, Ioannis Mitliagkas, Brady Neal, and Vasilis Syrgkanis. Empirical analysis of model selection for heterogeneous causal effect estimation. In *International Conference on Learning Representations*, volume 2024, pp. 26673–26702, 2024.

Francisco Mourao, David Hajage, Daria Bystrova, Bertrand Bouvarel, Nathanaël Lapidus, Fabrice Carrat, and Benjamin Glemain. Prior-data fitted networks for causal inference: a simulation study with real-world scenarios, 2026. URL https://arxiv.org/abs/2603.15928.

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

Xinkun Nie and Stefan Wager. Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika*, 108(2):299–319, 2021.

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

Judea Pearl. An introduction to causal inference. *The international journal of biostatistics*, 6(2):7, 2010.

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

Jingang Qu, David Holzmüller, Gaël Varoquaux, and Marine Le Morvan. TabICL: A Tabular Foundation Model for In-Context Learning on Large Data. In *Forty-second International Conference on Machine Learning*, 2025.

Craig T Ramey, Donna M Bryant, Barbara H Wasik, Joseph J Sparling, Kaye H Fendt, and Lisa M La Vange. Infant health and development program for low birth weight, premature infants: Program elements, family participation, and child intelligence. *Pediatrics*, 89(3):454–465, 1992.

Arik Reuter, Anish Dhir, Cristiana Diaconu, Jake Robertson, Ole Ossen, Frank Hutter, Adrian Weller, Mark van der Wilk, and Bernhard Schölkopf. Use what you know: Causal foundation models with partial graphs. *arXiv preprint arXiv:2602.14972*, 2026.

Jake Robertson, Arik Reuter, Siyuan Guo, Noah Hollmann, Frank Hutter, and Bernhard Schölkopf. Do-pfn: In-context learning for causal effect estimation. *arXiv preprint arXiv:2506.06039*, 2025.

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

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

Donald B Rubin. Randomization analysis of experimental data: The fisher randomization test comment. *Journal of the American statistical association*, 75(371):591–593, 1980.

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

Donald B Rubin. For objective causal inference, design trumps analysis. *The Annals of Applied Statistics*, 2(3):808–840, 2008.

Andreas Sauter, Saber Salehkaleybar, Frank van Harmelen, Aske Plaat, and Erman Acar. Activa: Amortized causal effect estimation via transformer-based variational autoencoder. *arXiv preprint arXiv:2503.01290*, 2025.

Steffen Schneider, Evgenia Rusak, Luisa Eck, Oliver Bringmann, Wieland Brendel, and Matthias Bethge. Improving robustness against common corruptions by covariate shift adaptation. *Advances in neural information processing systems*, 33:11539–11551, 2020.

Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021.

Alejandro Schuler, Michael Baiocchi, Robert Tibshirani, and Nigam Shah. A comparison of methods for model selection when estimating individual treatment effects. *arXiv:1804.05146*, 2018.

Uri Shalit, Fredrik D. Johansson, and David Sontag. Estimating individual treatment effect: generalization bounds and algorithms. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 3076–3085, 2017.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*, 2017.

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

Christopher Stith, Hossein Rahmani, and Jesse C Cresswell. Causal foundation models. *arXiv preprint arXiv:2609.03003*, 2026.

Yu Sun, Xiaolong Wang, Zhuang Liu, John Miller, Alexei Efros, and Moritz Hardt. Test-time training with self-supervision for generalization under distribution shifts. In *International conference on machine learning*, pp. 9229–9248. PMLR, 2020.

Gábor J Székely, Maria L Rizzo, and Nail K Bakirov. Measuring and testing dependence by correlation of distances. 2007.

Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. *arXiv preprint arXiv:2006.10726*, 2020.

James Woodward. *Making things happen: A theory of causal explanation*. Oxford university press, 2005.

Han-Jia Ye, Si-Yang Liu, and Wei-Lun Chao. A closer look at TabPFN v2: Strength, limitation, and extension. *arXiv:2502.17361*, 2025.

A Implementation Details

A.1 Benchmarking Datasets

All methods are tested on four benchmarking datasets, as shown in Table 2, with diverse numbers of covariates and sample sizes. All datasets are semi-synthetic but synthesis occurs at different stages of the construction. Their key characteristics and challenges are summarized as follows.

IHDP (Ramey et al., 1992) evaluates treatment effects of specialist home visits on the cognitive development of low birth-weight infants. It uses real pre-treatment covariates from a randomized experiment, while selection bias and potential outcomes are synthetically introduced using non-linear response surface equations. Its primary challenge stems from induced confounding coupled with a small sample size, making estimators prone to overfitting.

ACIC 2016 (Dorie et al., 2019), derived from the Collaborative Perinatal Project, models health and demographic factors influencing maternal and child outcomes. The covariates are real, but the treatment assignments and potential outcomes are synthetically generated across 77 distinct data-generating processes. The main difficulty lies in handling extreme cross-setting diversity and high-dimensional noise, which severely tests whether causal overlap and unconfoundedness assumptions hold uniformly.

RealCause Lalonde CPS (LaLonde, 1986) evaluates job training interventions by pairing experimental treated workers from the National Supported Work demonstration with a non-experimental comparison group from the Current Population Survey. A deep generative model is fitted on empirical data to synthesize joint distributions and ground-truth potential outcomes with known counterfactuals. Its main challenge is severe class imbalance and extreme propensity score polarization, where positivity assumptions break down for high-earning control subjects.

RealCause Lalonde PSID (LaLonde, 1986) assesses the same job training program by combining the experimental treated cohort with observational controls from the Panel Study of Income Dynamics. Similar to the CPS benchmark, ground-truth outcomes are constructed via a fitted deep generative model to enable counterfactual evaluation. The primary hurdle is substantial distribution shift and severe positivity violations between treated and control groups, forcing models to rely on untestable structural extrapolation.

Dataset Group	IHDP	ACIC 2016	Lalonde CPS	Lalonde PSID
Realizations	100	10	100	100
Data Sizes	747	4802	16177	2675
Treated Group Sizes	129	1326	156	144
Control Group Sizes	618	3476	16021	2531
Treated Fractions	17.27%	27.61%	0.96%	5.38%
Covariate Dimensions	25	58	8	8
Training Data Sizes	672	4321	14559	2407
Evaluation Data Sizes	75	481	1618	268

Table 2: Statistics for the benchmarking datasets used in the experiment. The numbers of the treated examples are highlighted, showing a diverse range of percentages of the treated group.

A.2 Evaluation Metrics

Given a dataset of N units with covariates and ground-truth CATE values $\{(\boldsymbol{x}^{(n)}, \tau(\boldsymbol{x}^{(n)}))\}_{n=1}^{N}$, and a ground-truth ATE λ, we evaluate models using the relative ATE error and the precision in estimation of heterogeneous effects (PEHE) (Hill, 2011):

$$\text{RelativeError} = \frac{|\hat{\lambda} - \lambda|}{|\lambda|}, \qquad \text{PEHE} = \sqrt{\frac{1}{N}\sum_{n=1}^{N}\left(\tau(\boldsymbol{x}^{(n)}) - \hat{\tau}(\boldsymbol{x}^{(n)})\right)^2}. \tag{5}$$

Here, $\hat{\tau}$ and $\hat{\lambda}$ denote the estimated CATE and ATE, respectively.

Table 3: Comparison of MoE-Learner with averaging ensemble. Better results are highlighted.

Method	Mean PEHE (↓ better)					Mean ATE Relative 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. Rank
MoE-Learner (ours)	0.80±0.15	0.64±0.08	9.05±0.02	14.08±0.25	2.50	0.22±0.04	0.02±0.00	0.19±0.01	0.12±0.02	5.00
Ensemble	2.35±0.86	1.73±0.14	9.04±0.21	14.00±0.30	6.00	0.26±0.46	0.09±0.09	0.22±0.12	0.13±0.15	8.75

Table 4: Comparison of causal experts with (★) and without the MLP-based weighting function

Expert	Weighting Function	Mean PEHE (↓ better)					Mean ATE Relative 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. Rank
Response Expert	★	0.64±1.05	0.65±0.26	9.20±0.22	13.80±1.76	2.25	0.28±0.48	0.02±0.01	0.30±0.14	0.09±0.09	6.00
		1.54±3.36	0.97±0.37	9.19±0.21	13.75±1.76	4.00	0.27±0.48	0.02±0.01	0.30±0.14	0.09±0.08	6.00
Covariate Expert	★	3.87±6.30	3.16±1.14	9.10±0.24	14.37±1.87	7.75	0.22±0.42	0.25±0.15	0.15±0.11	0.06±0.06	8.00
		4.24±6.88	2.82±1.43	9.43±0.23	14.66±1.72	9.25	0.23±0.43	0.19±0.09	0.16±0.11	0.06±0.06	7.75
Prognostic Expert	★	0.85±1.54	1.31±0.67	9.23±0.21	13.92±1.91	4.25	0.27±0.50	0.08±0.06	0.32±0.13	0.11±0.12	9.50
		1.62±3.26	1.61±0.91	9.23±0.21	13.79±1.77	5.00	0.27±0.48	0.08±0.05	0.31±0.14	0.10±0.11	9.25
Propensity Expert	★	5.25±8.05	3.09±1.65	9.04±0.25	14.96±4.83	8.75	0.26±0.46	0.19±0.23	0.16±0.14	0.38±0.39	9.25
		5.28±8.12	3.06±1.69	9.01±0.21	14.08±1.85	8.25	0.27±0.50	0.11±0.05	0.17±0.14	0.33±0.15	9.25

A.3 Causal Experts

Causal Embedding. Pretrained on a vast collection of synthetic datasets and performing classification or regression at test time via in-context learning, TabPFN unifies heterogeneous datasets into a fixed-dimensional representation. For the causal embeddings, we use TabPFN version 8.2.0. Following the causal-expert construction described in Section 4.2, TabPFN is used to embed each unit under the corresponding query-context relationship. To avoid label leakage when constructing these representations, we use 5-fold cross-validation: for each fold, the embedding of a unit is generated using a TabPFN model whose context does not contain that unit. Thus, the representation used for routing does not directly incorporate the unit's own observed outcome through the context. Empirically, we find that cross-validated embeddings perform better than embeddings generated without cross-validation. TabPFN is only one possible choice for constructing the causal embeddings. Other tabular foundation models, such as TabICL, could be used within the same framework. We leave a systematic comparison and selection of the optimal embedding model for future work.

Dissimilarity. We instantiate D_{ij} as cosine distance, $D(\boldsymbol{z}_i, \boldsymbol{z}_j) = 1 - \frac{\boldsymbol{z}_i^\top \boldsymbol{z}_j}{\|\boldsymbol{z}_i\| \|\boldsymbol{z}_j\|}$, which bounds $D \in [0, 2]$ and is invariant to feature norm, so retrieval depends only on directional alignment between query and context points.

Monotonic weighting function. f_θ is a multilayer perceptron mapping the one-dimensional input D (the cosine distance; equivalently, $1 - D$ is the cosine similarity) to a scalar weight. Because larger similarity corresponds to smaller distance, the network first transforms the input to $1 - D$. Every weight matrix $W^{(l)}$ is reparameterized as $\mathrm{Softplus}(\widetilde{W}^{(l)})$, ensuring all coefficients are non-negative, and Softplus activations are applied after every layer, including the output. Since a non-negative linear map composed with a non-decreasing activation is itself non-decreasing, the resulting network is non-increasing in D. The final Softplus guarantees $f_\theta(D) > 0$ everywhere, avoiding degenerate zero-weight neighborhoods. In our experiments f_θ uses one hidden layer of width 32.

Top-K retrieval. For a context set of size N, exact top-K selection is performed with a partial selection algorithm rather than a full sort, giving $O(N)$ average-case retrieval per query instead of $O(N \log N)$; this becomes relevant as N grows beyond the scale used here.

Training. f_θ is trained on the training split itself: at each step we sample a random subset of P points from $\boldsymbol{Z}$ to serve as queries, and for each query retrieve its top-K neighbors from the remaining $N - 1$ points under leave-one-out (the query point is masked out of its own context before retrieval, preventing the trivial shortcut $D(\boldsymbol{z}, \boldsymbol{z}) = 0$). The model is optimized with Adam to minimize mean squared error between $\hat{y}$ and the true label of each sampled query, $\mathcal{L} = \frac{1}{P} \sum_{p=1}^{P} (\hat{y}_p - y_p)^2$. Query subsets are resampled each training step. We use $K = 100$, $P = 400$, learning rate 0.1, for 150 steps.

Table 5: Comparison of three metrics for the router. We estimate the intrinsic dimension using the maximum-likelihood estimator of Levina and Bickel (Levina & Bickel, 2004), which models the points in a local neighborhood as a homogeneous Poisson process and estimates the dimension from the distances to the k nearest neighbors.

Method	Mean PEHE (↓ better)					Mean ATE Relative 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. Rank
Spearman's r	0.80±0.15	0.64±0.08	9.05±0.02	14.08±0.25	2.50	0.22±0.04	0.02±0.00	0.19±0.01	0.12±0.02	5.00
Pearson's r	0.71±1.05	0.63±0.27	9.11±0.02	14.56±0.87	3.00	0.22±0.42	0.02±0.01	0.25±0.16	0.16±0.03	5.50
Intrinsic Dimension	0.85±1.09	1.27±0.82	9.11±0.02	14.37±0.87	4.00	0.27±0.48	0.06±0.06	0.15±0.01	0.07±0.07	7.50

Inference. At test time, the full training set $\boldsymbol{Z}$ (all N points) serves as the context; no leave-one-out masking is applied, since test queries are disjoint from $\boldsymbol{Z}$.

A.4 Causal Router

Representative-Point Sampling. To evaluate the geometry of each embedding efficiently, we compute the variogram statistics using a subset of maximally $M = 150$ representative samples. We select these points using farthest-point sampling with cosine distance. Starting from an initial point, we iteratively select the point in $\boldsymbol{Z}$ whose nearest selected point has the largest cosine distance, and continue until M representative points have been obtained. This procedure encourages the representative set to cover diverse regions of the embedding space while avoiding an excessive concentration of samples in dense regions.

Correlation-Based Routing. After selecting the representative points, we characterize the relationship between embedding distance and label variation using correlation measures. Specifically, we consider both Pearson's correlation and Spearman's rank correlation as empirical measures of the distance–label relationship. Pearson's correlation captures linear association, whereas Spearman's correlation captures monotonic association and is less sensitive to the precise functional form of the relationship.

In addition to these correlation-based measures, we compute the intrinsic dimension (Levina & Bickel, 2004) of each embedding as a complementary geometric diagnostic. Intrinsic dimension characterizes the effective dimensionality of the manifold on which the embedding lies and has been shown to influence the convergence behavior of nearest-neighbor regression (Kpotufe, 2011; Székely et al., 2007; Anselin, 1995). A lower intrinsic dimension may therefore indicate that the embedding occupies a lower-dimensional structure, potentially making local retrieval more statistically efficient. However, intrinsic dimension alone does not determine local smoothness or predictive quality: an embedding can have low intrinsic dimension while still failing to organize samples according to their outcomes. We therefore treat intrinsic dimension as a complementary geometric property rather than a replacement for the distance–label correlation used by the router. As shown in Table 5, the Spearman's rank correlation is the most robust indicator for the expert performance.

B Proof of Theorem 1

Proof of Theorem 1 Fix expert e and, as in the main text, suppress the group subscript t and expert superscript (e); write $Z_i = \boldsymbol{z}_i^{(e)}$ and $f = f_e$. Let (Z_i, Y_i) and (Z_j, Y_j) be independent draws from the same distribution, with $Y_i = f(Z_i) + \varepsilon_i$, $\mathbb{E}[\varepsilon_i \mid Z_i] = 0$, and $\mathrm{Var}(\varepsilon_i) = \sigma^2$.

Since $Y_i - Y_j = (f(Z_i) - f(Z_j)) + (\varepsilon_i - \varepsilon_j)$, squaring and conditioning on $D(Z_i, Z_j) = d$ gives

$$\begin{aligned} V(d) &:= \mathbb{E}[(Y_i - Y_j)^2 \mid D_{ij} = d] \\ &= \mathbb{E}[(f(Z_i) - f(Z_j))^2 \mid D_{ij} = d] \\ &+ \mathbb{E}[(\varepsilon_i - \varepsilon_j)^2 \mid D_{ij} = d] \\ &+ 2\,\mathbb{E}[(f(Z_i) - f(Z_j))(\varepsilon_i - \varepsilon_j) \mid D_{ij} = d]. \end{aligned} \tag{6}$$

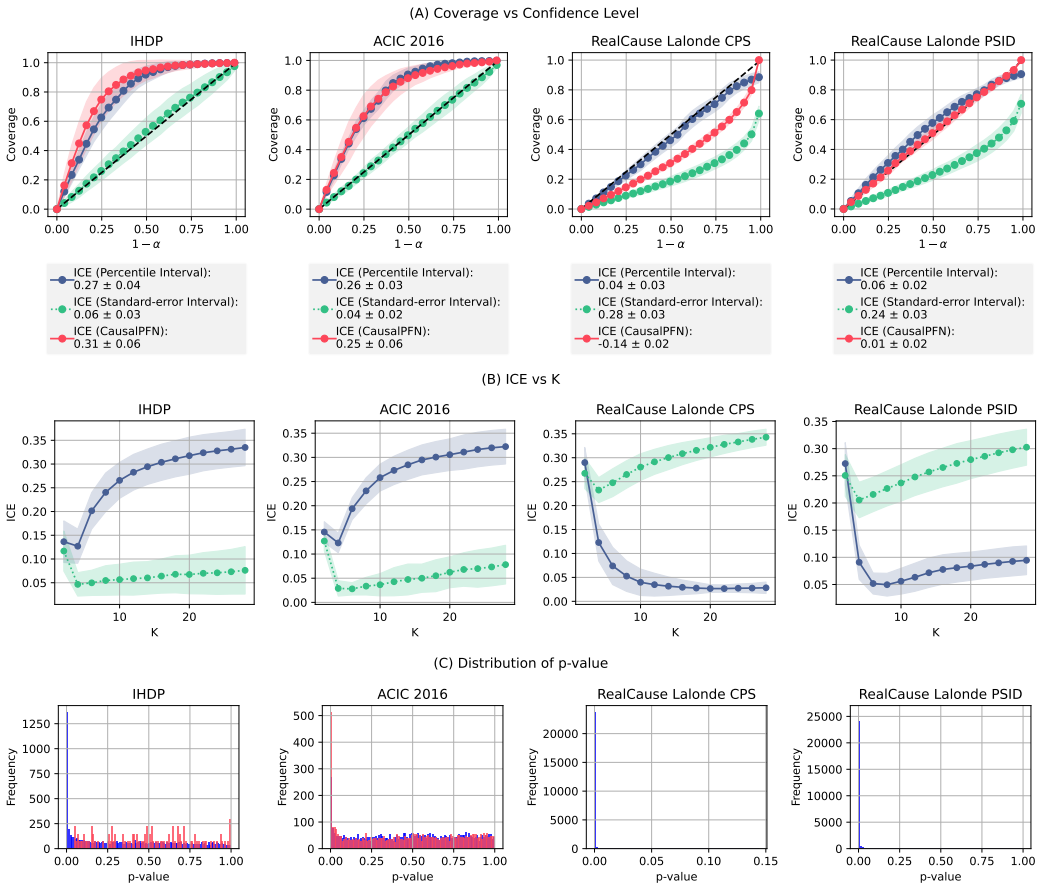


Figure 4: (A) ICE from the percentile interval, standard-error interval, and CausalPFN. (B) Sensitivity of ICE to K. (C) Different distribution of the p-values from Shapiro–Wilk test.

The cross term vanishes since $\varepsilon_i, \varepsilon_j$ are conditionally mean-zero given (Z_i, Z_j), and $\mathbb{E}[(\varepsilon_i - \varepsilon_j)^2] = \mathrm{Var}(\varepsilon_i) + \mathrm{Var}(\varepsilon_j) = 2\sigma^2$ by independence. Hence

$$\begin{aligned} V(d) &= 2\sigma^2 + \mathbb{E}[(f(Z_i) - f(Z_j))^2 \mid D_{ij} = d], \\ \gamma_e(d) &= \sigma^2 + \tfrac{1}{2}\mathbb{E}[(f(Z_i) - f(Z_j))^2 \mid D_{ij} = d]. \end{aligned} \tag{7}$$

Since the second term is a conditional expectation of a nonnegative quantity, $\gamma_e(d) \geq \sigma^2$.

By β-local smoothness, almost surely

$$(f(Z_i) - f(Z_j))^2 \leq L_e^2 D(Z_i, Z_j)^{2\beta},$$

so conditioning on $D_{ij} = d$ gives

$$\mathbb{E}[(f(Z_i) - f(Z_j))^2 \mid D_{ij} = d] \leq L_e^2 d^{2\beta}.$$

Substituting into the decomposition above yields

$$\gamma_e(d) \leq \sigma^2 + \tfrac{1}{2} L_e^2 d^{2\beta},$$

which together with the lower bound gives the claim. ■

C Uncertainty Quantification

Beyond a point estimate, the top-K retrieval mechanism from Section 4.2 also supports uncertainty quantification, using the same neighborhood – from each group's selected expert $e_t^\star$ – already retrieved for prediction.

For each query point, recall the retrieved top-K pairs $\{(\boldsymbol{z}_t^{(k)}, y_t^{(k)})\}_{k=1}^K$ and their normalized attention weights $\{a_t^{(k)}\}_{k=1}^K$. We estimate the local variance from the retrieved labels as

$$\hat{\sigma}_t^2 = \sum_{k=1}^{K} a_t^{(k)} \left(y_t^{(k)} - \hat{y}_t\right)^2,$$

noting that the denominator $\sum_k a_t^{(k)}$ is omitted since the weights are already normalized to sum to one. We define the effective number of neighbors, following Kish's effective sample size, as

$$\tilde{K}_t = \frac{1}{\sum_{k=1}^{K} \left(a_t^{(k)}\right)^2},$$

which discounts K toward the number of neighbors actually carrying weight when attention is concentrated on a few points. The standard error of the weighted prediction is then $\mathrm{se}(\hat{y}_t) = \hat{\sigma}_t / \sqrt{\tilde{K}_t}$.

Assuming the local label distribution is approximately Gaussian, a two-sided $(1-\alpha)$ uncertainty interval for y_t is given by

$$[L_t, U_t] = \left[\hat{y}_t - z_{1-\alpha/2}\,\mathrm{se}(\hat{y}_t),\ \hat{y}_t + z_{1-\alpha/2}\,\mathrm{se}(\hat{y}_t)\right],$$

where $z_{1-\alpha/2}$ is the $(1-\alpha/2)$-quantile of the standard normal distribution.

When the local normality assumption is not supported, we instead construct an empirical percentile interval directly from the retrieved labels. For a nominal $(1-\alpha)$ coverage level, the interval is defined as

$$[L_t, U_t] = \left[Q_{\alpha/2}\left(\{y_t^{(k)}\}_{k=1}^K\right),\ Q_{1-\alpha/2}\left(\{y_t^{(k)}\}_{k=1}^K\right)\right],$$

where $Q_p(\cdot)$ denotes the empirical p-th percentile. This approach makes no Gaussian assumption, instead characterizing uncertainty directly from the observed local label distribution.

We construct a conservative interval for the CATE by subtracting the two group-level intervals:

$$[L, U] = [L_1 - U_0,\ U_1 - L_0]. \tag{8}$$

We evaluate interval calibration using the Integrated Calibration Error (ICE), defined over a set of significance levels $\mathcal{A}$ as

$$\mathrm{ICE} = \frac{1}{|\mathcal{A}|} \sum_{\alpha \in \mathcal{A}} \left|\widehat{\mathrm{coverage}}(\alpha) - (1-\alpha)\right|,$$

where $\widehat{\mathrm{coverage}}(\alpha)$ is the empirical fraction of test examples for which the true τ_t falls within the corresponding interval $[L_t, U_t]$. Lower ICE indicates closer agreement between empirical and nominal coverage.

Unlike approaches such as TabPFN and CausalPFN, which obtain uncertainty estimates through repeated sampling or additional calibration procedures such as temperature scaling, our approach derives uncertainty directly from the local neighborhood already used for prediction, avoiding additional sampling or calibration overhead.

The uncertainty-quantification results are shown in Figure 4. We evaluate two types of confidence intervals: a standard-error-based interval and an empirical percentile interval. The choice between these intervals is motivated by the distribution of the retrieved outcomes within each local neighborhood.

To assess whether a normal approximation is appropriate, we perform a Shapiro–Wilk test on the retrieved outcome values for each local neighborhood. The null hypothesis is that the values within a neighborhood are normally distributed. A high p-value therefore provides no evidence against the normality assumption, whereas a p-value below 0.05 leads us to reject it.

On IHDP and ACIC 2016, we observe consistently high p-values across local neighborhoods, indicating that the normality hypothesis is generally not rejected. In these settings, the standard-error-based intervals provide better ICE, achieving 0.06 ± 0.03 on IHDP and 0.04 ± 0.02 on ACIC 2016. In contrast, the Lalonde benchmarks exhibit markedly different behavior: for almost all realizations, the retrieved outcomes yield p-values below 0.05, providing evidence against the normality assumption.

Correspondingly, empirical percentile intervals achieve better ICE, with 0.04 ± 0.03 on Lalonde CPS and 0.06 ± 0.02 on Lalonde PSID.

These results support the use of a data-dependent choice of confidence-interval construction. In particular, when the retrieved outcomes are approximately compatible with a normal distribution, a standard-error-based interval is effective; when this approximation is not supported, empirical percentile intervals provide a more appropriate alternative.

D LIMITATIONS

In this work we restrict our attention to method-specific assumptions under an unconfoundedness/back-door regime, and leave broader identification-strategy selection to future work. For example, settings where unconfoundedness itself fails call for different identification strategies altogether, such as Instrumental Variables (IV) (Angrist et al., 1996), Regression Discontinuity (RDD) (Imbens & Lemieux, 2008), or Difference-in-Differences (DiD) (Bertrand et al., 2004).

As discussed in Section 4.3, the variogram-based router is a surrogate diagnostic for local smoothness, not a direct certificate of causal assumption validity – an embedding can appear smooth while still violating the assumption it is meant to encode, so routing decisions should be interpreted as data-informed evidence rather than formal guarantees. We also note that, unlike a standard MoE architecture for LLMs, our Mixture-of-Causal-Experts operates at the treatment-arm level, not the instance level, because the router must consider the local population of units. This points to future development of unit-level routing for more flexible causal inference.

We do not propose yet another causal foundation model; instead, we show how a general-purpose tabular foundation model can be retrofitted, via causal experts and a counterfactual-free causal router, into a causal effect estimator that outperforms both the underlying tabular foundation model and specialized causal foundation models. That said, Causal Foundation Models (e.g., CausalPFN, CausalFM, and Do-PFN) themselves are natively tabular transformer models, and can be used to extract various embeddings, which may possibly lead to better performance. We leave this question for future investigation.

E ADDITIONAL FIGURES

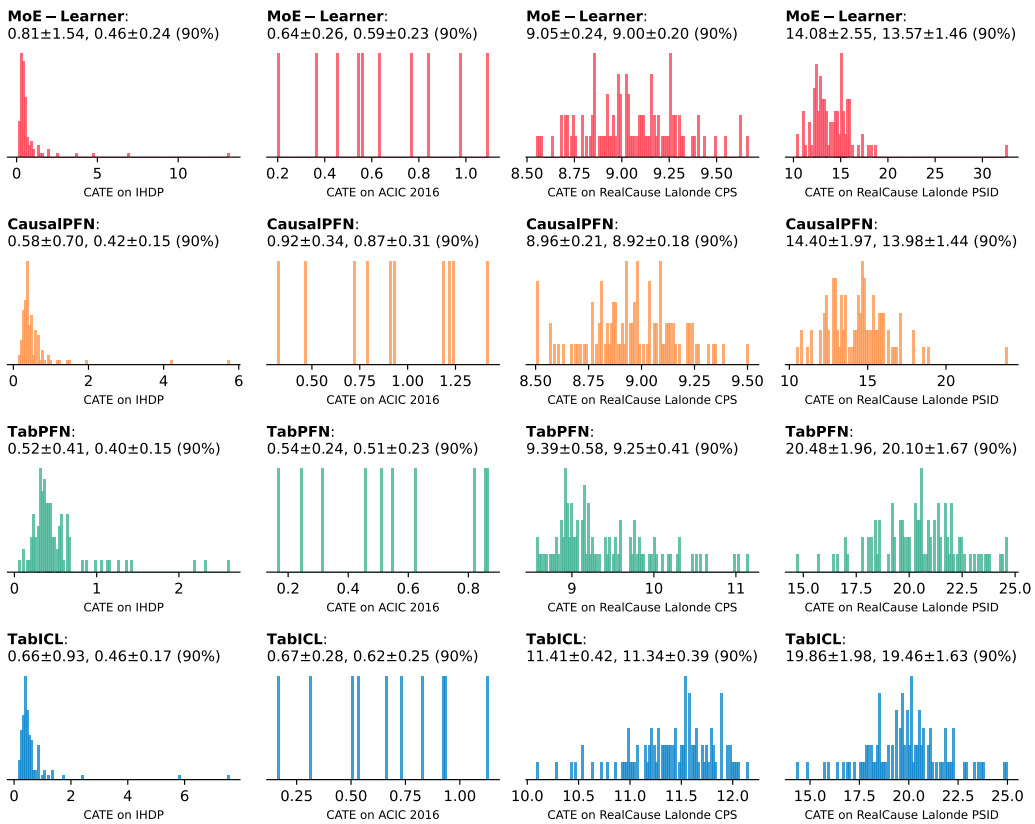


Figure 5: The distributions of PEHE scores from MoE-Learner, CausalPFN, TabPFN, and TabICL on four datasets, respectively. Lalonde PEHE is in thousands.

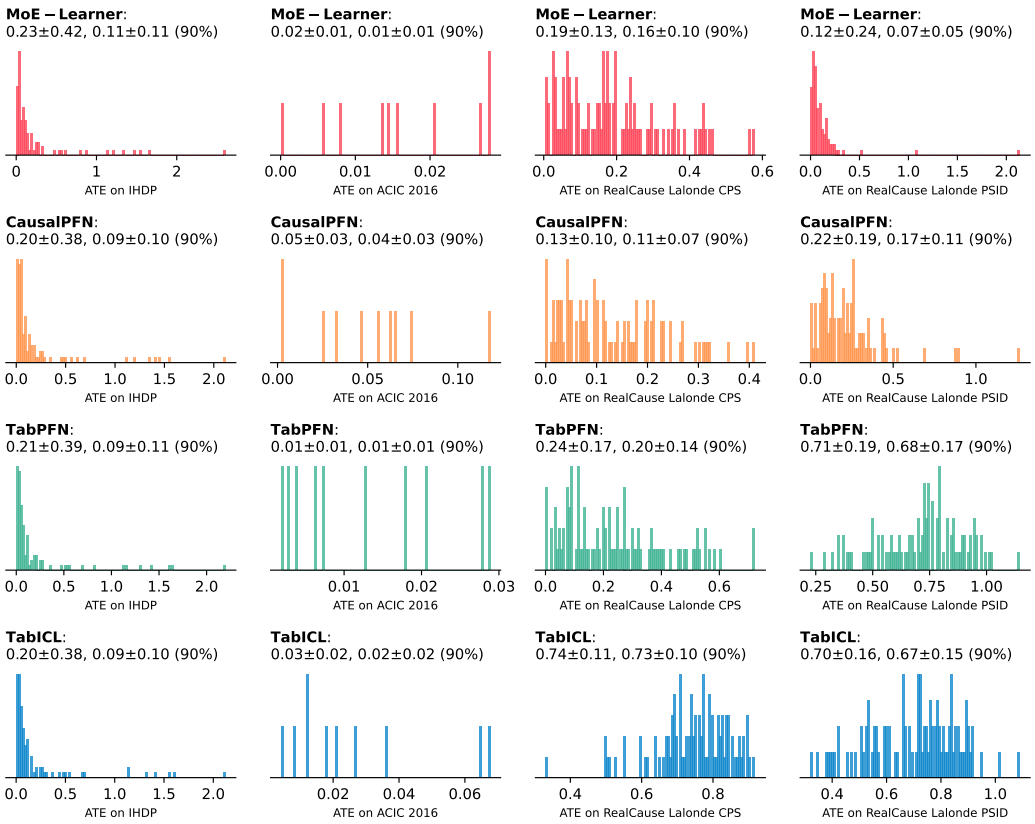


Figure 6: The distributions of the relative ATE errors from MoE-Learner, CausalPFN, TabPFN, and TabICL on four datasets, respectively.

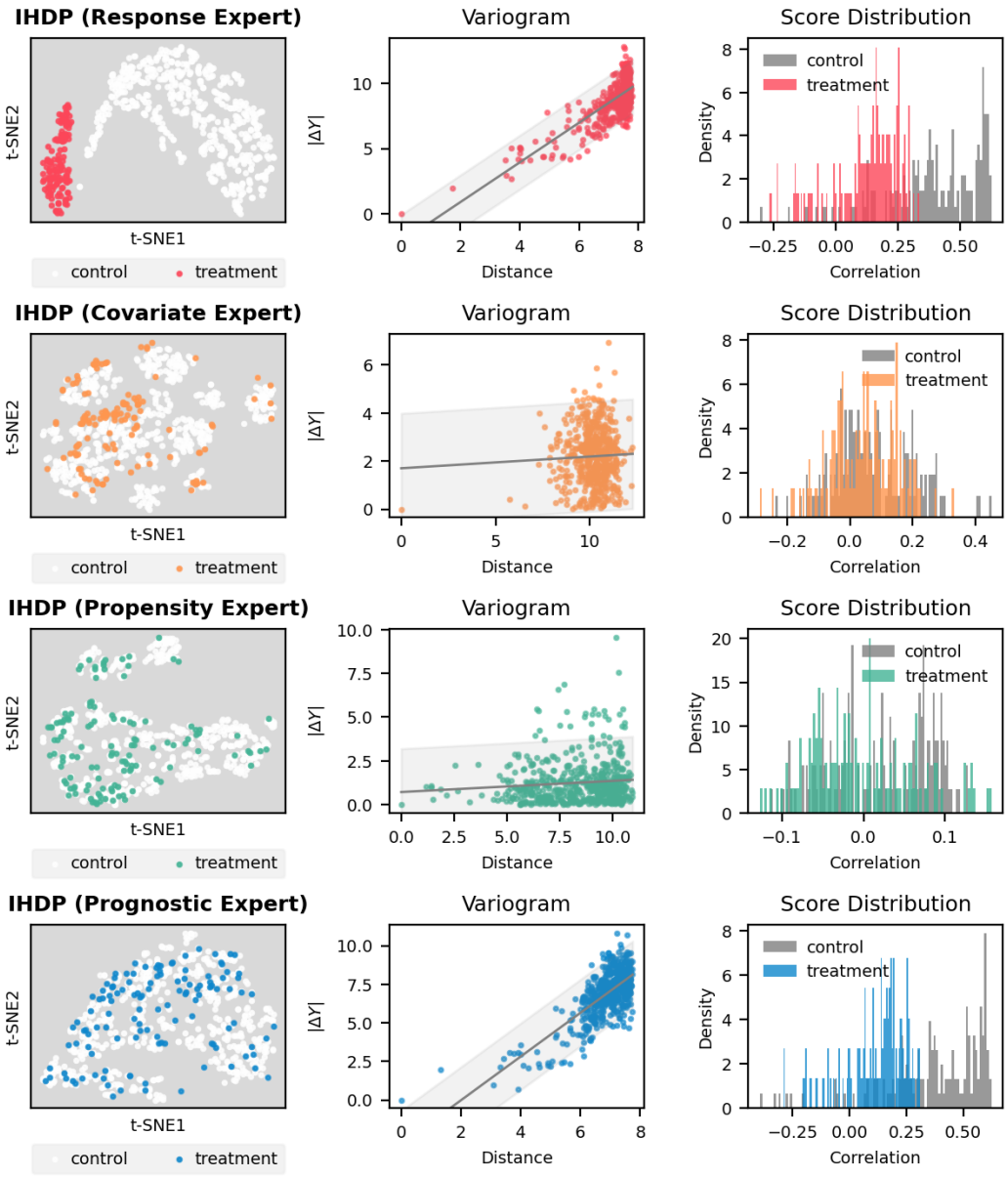


Figure 7: The t-SNE visualizations of embeddings (left column), the variograms (central column), and the distribution of the correlations in the variograms (right column) from the four experts on IHDP dataset for CATE.

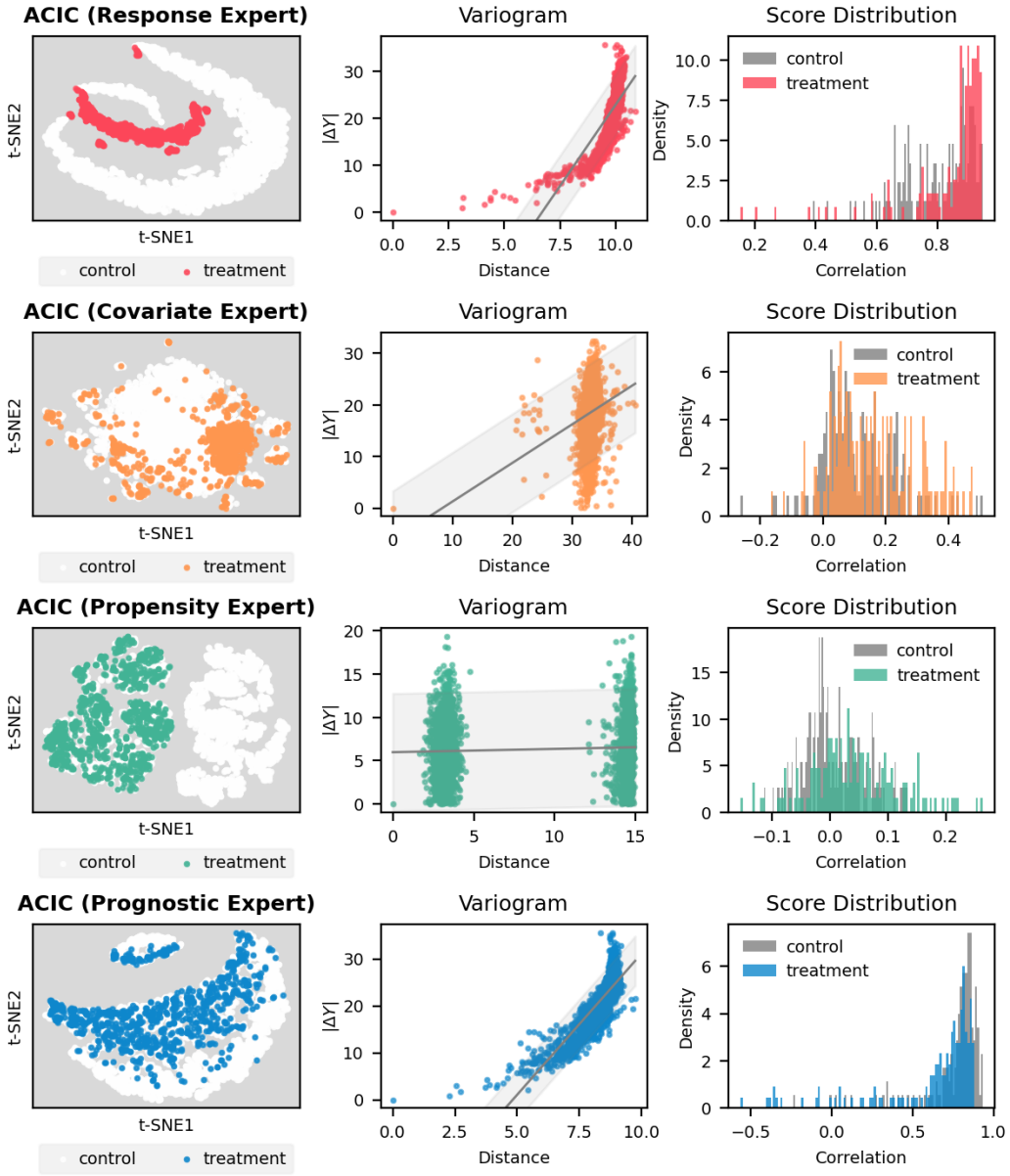


Figure 8: The t-SNE visualizations of embeddings (left column), the variograms (central column), and the distribution of the correlations in the variograms (right column) from the four experts on ACIC 2016 dataset for CATE.

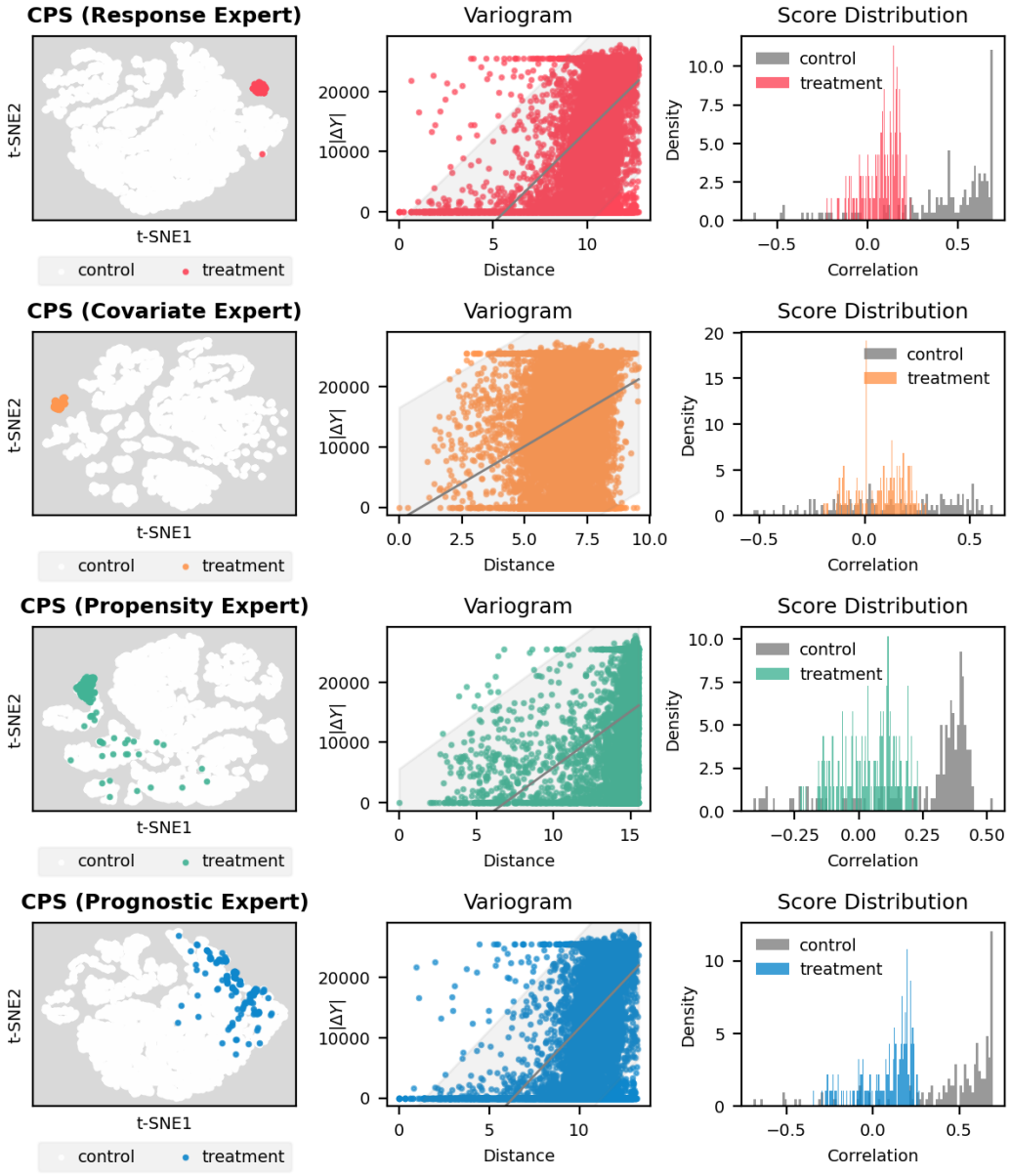


Figure 9: The t-SNE visualizations of embeddings (left column), the variograms (central column), and the distribution of the correlations in the variograms (right column) from the four experts on Lalonde CPS dataset for CATE.

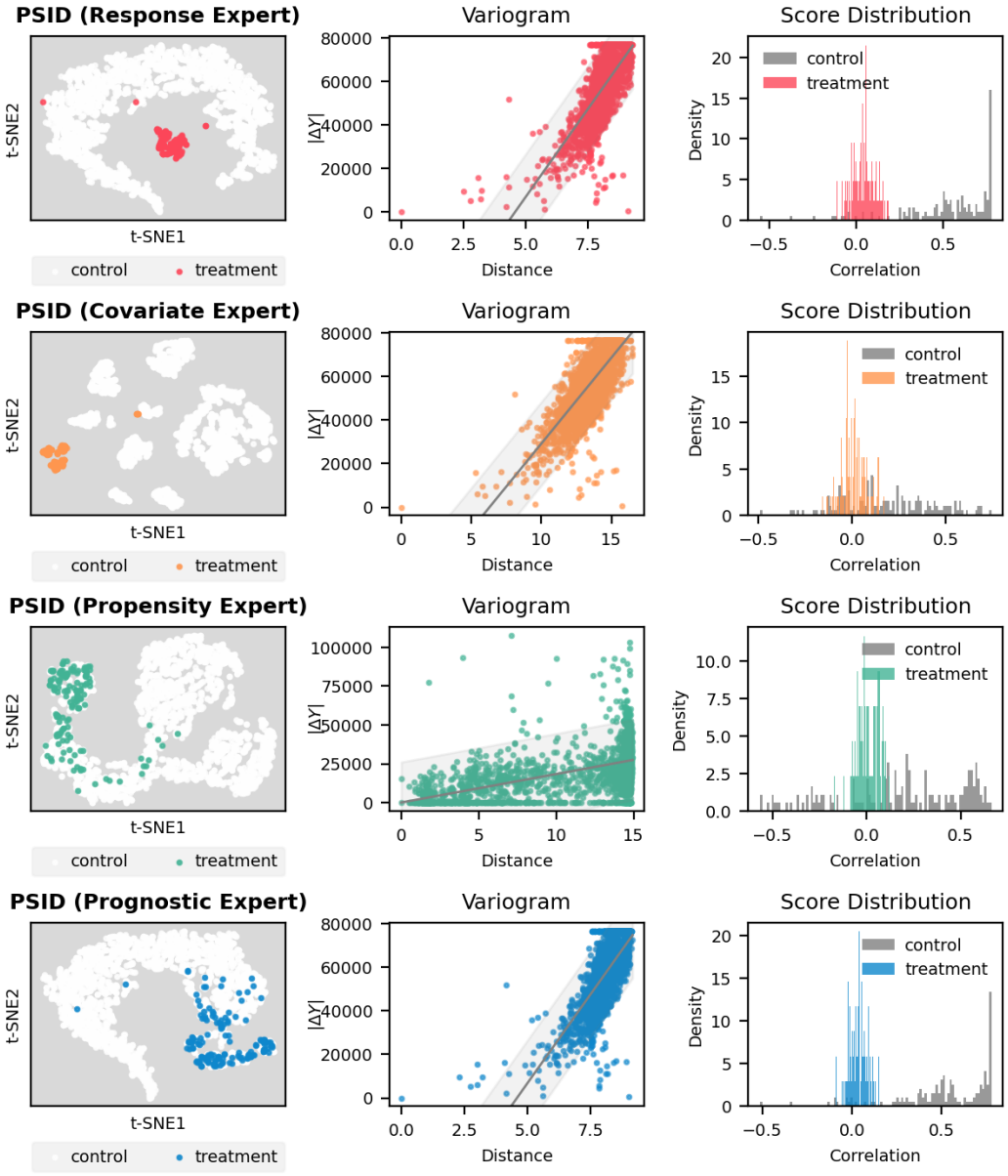


Figure 10: The t-SNE visualizations of embeddings (left column), the variograms (central column), and the distribution of the correlations in the variograms (right column) from the four experts on Lalonde PSID dataset for CATE.

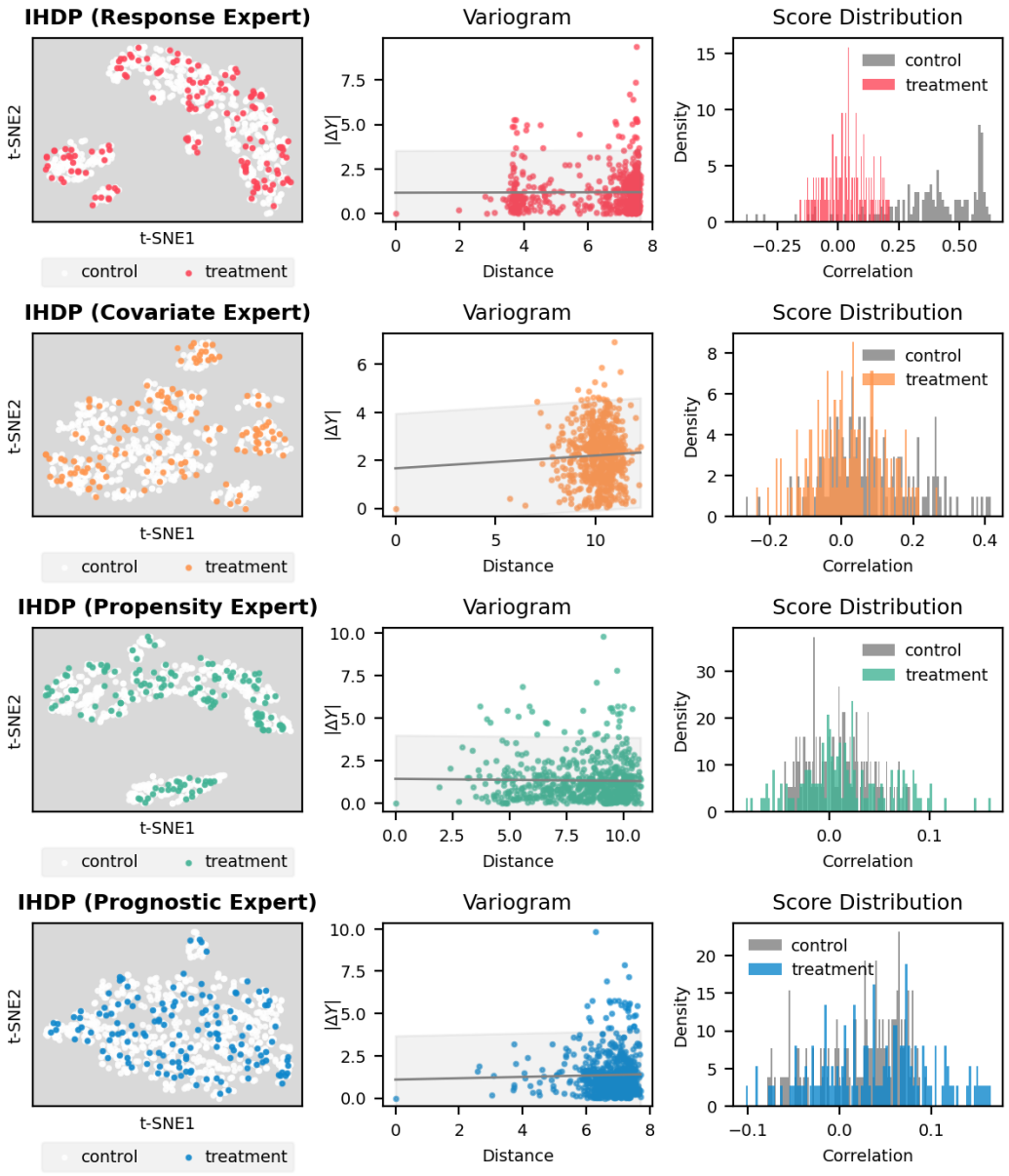


Figure 11: The t-SNE visualizations of embeddings (left column), the variograms (central column), and the distribution of the correlations in the variograms (right column) from the four experts on IHDP dataset for ATE.

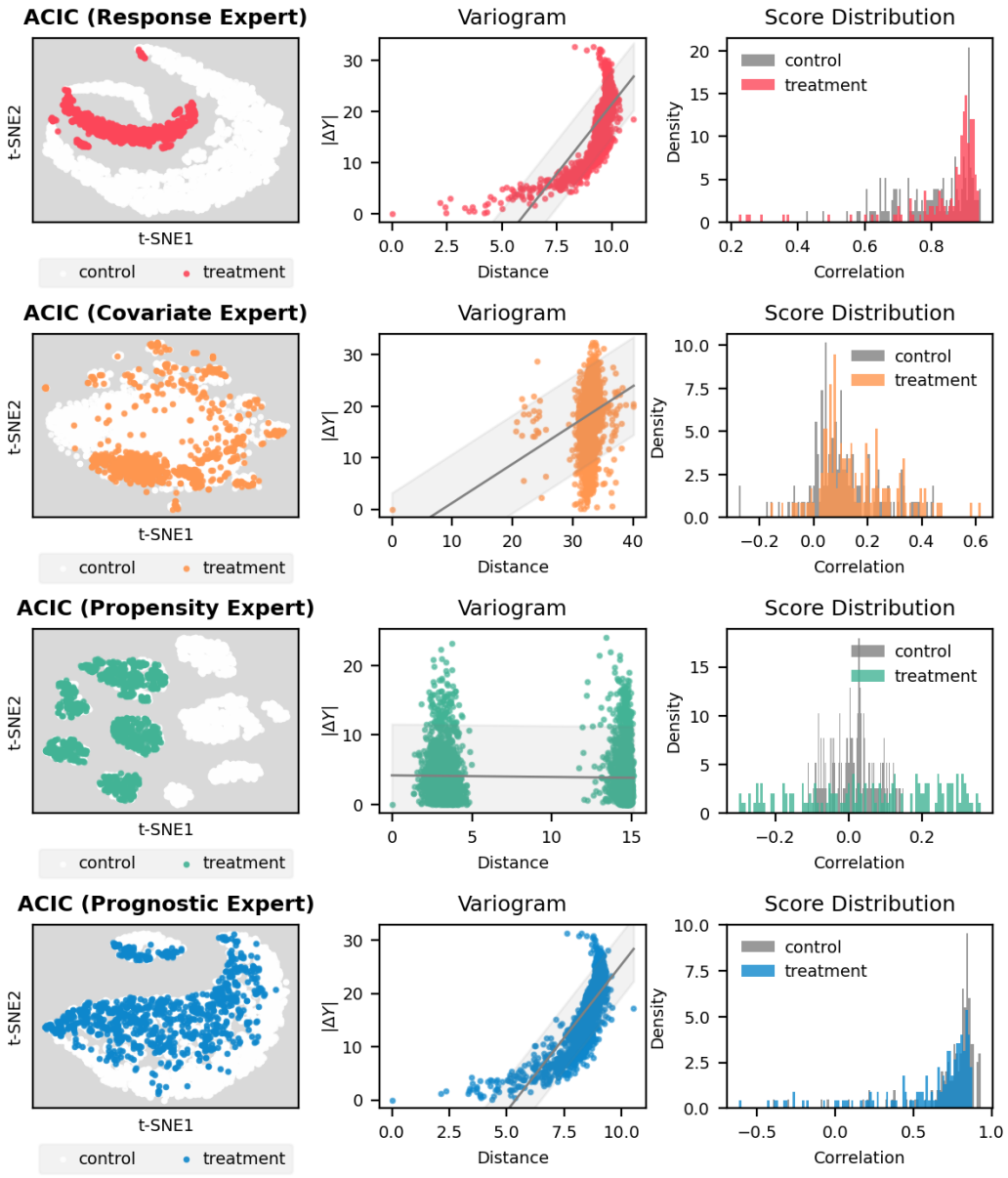


Figure 12: The t-SNE visualizations of embeddings (left column), the variograms (central column), and the distribution of the correlations in the variograms (right column) from the four experts on ACIC 2016 dataset for ATE.

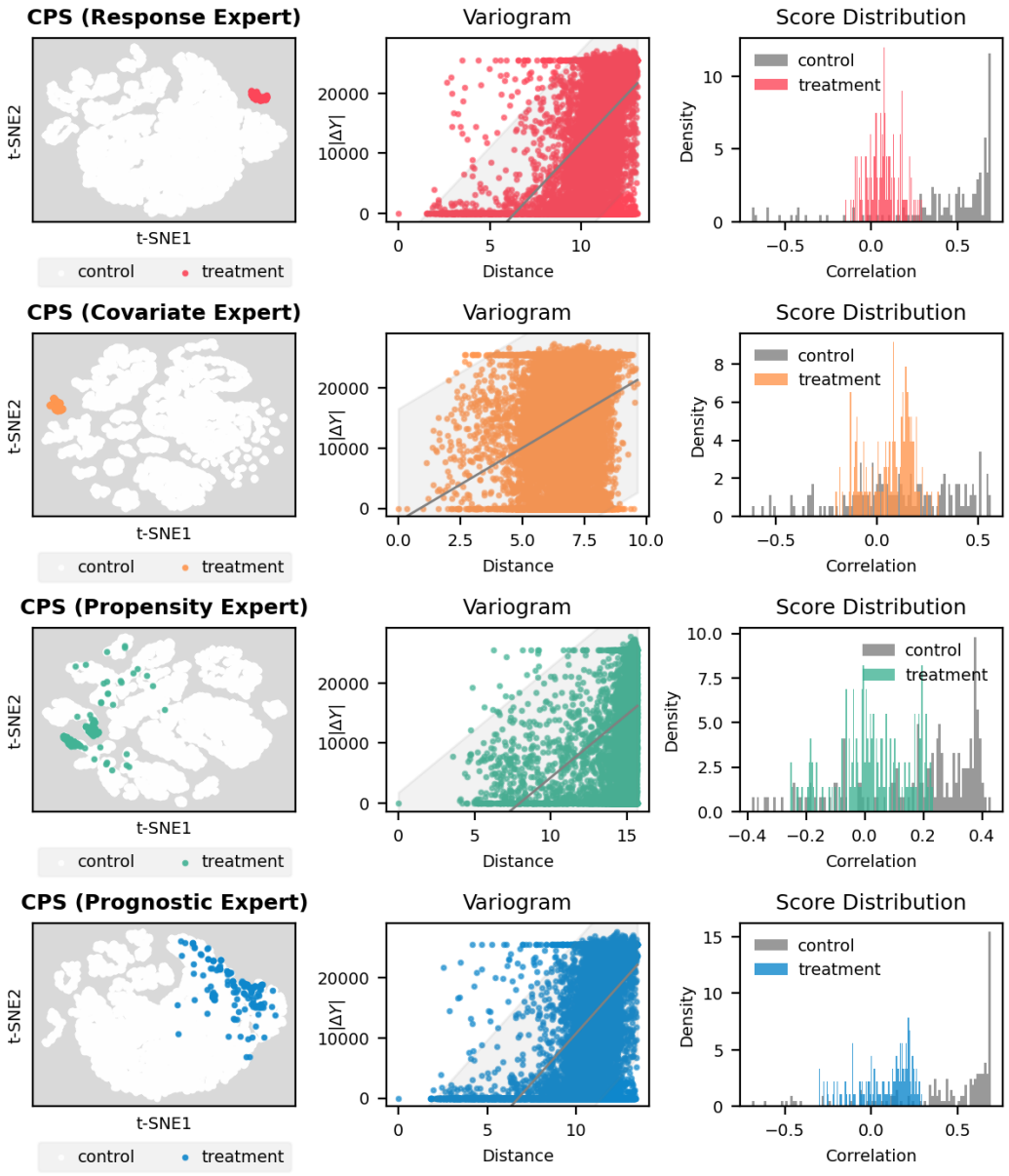


Figure 13: The t-SNE visualizations of embeddings (left column), the variograms (central column), and the distribution of the correlations in the variograms (right column) from the four experts on Lalonde CPS dataset for ATE.

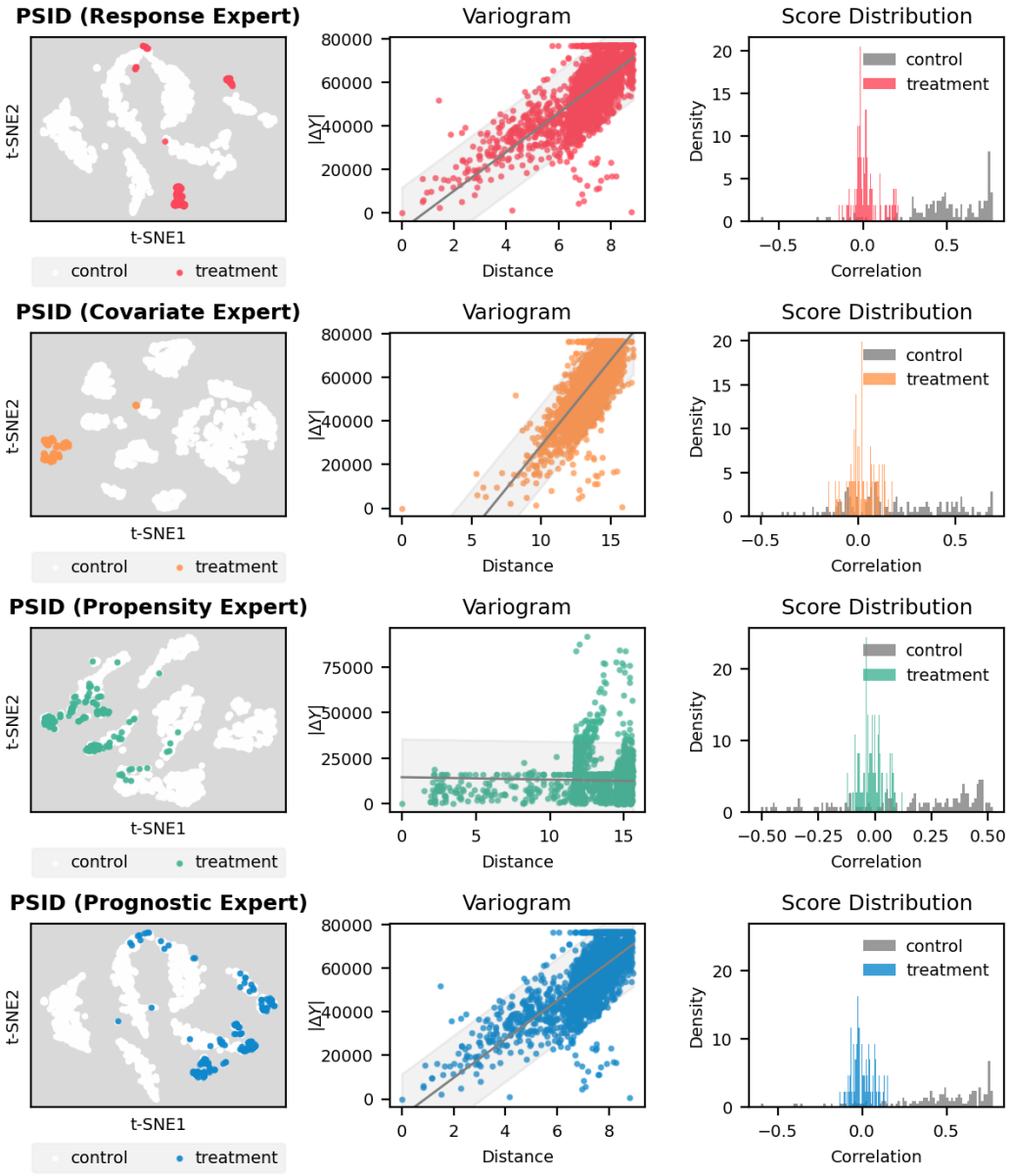


Figure 14: The t-SNE visualizations of embeddings (left column), the variograms (central column), and the distribution of the correlations in the variograms (right column) from the four experts on Lalonde PSID dataset for ATE.