# CONFOUNDINGSHAP: QUANTIFYING CONFOUNDING STRENGTH IN CAUSAL INFERENCE

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

In causal inference, confounders are variables that influence both treatment decisions and outcomes. However, unlike as in randomized clinical trials, the treatment assignment mechanism in observational studies is not known, and it is thus unclear which covariates act as confounders. Here, we aim to generate insight for causal inference and answer: which of the observed covariates act as confounders? We introduce CONFOUNDINGSHAP, a Shapley-based method for attributing confounding strength to individual covariates. Our contributions are twofold. First, we propose a Shapley game targeted to infer the confounding strength of the covariates. Our resulting Shapley values differ from the standard applications of SHAP explanations on causal targets, such as understanding treatment effect heterogeneity, which are ill-suited for our task. Second, as our task requires evaluating the value function over many adjustment sets, we provide a scalable TabPFN-based estimation that avoids exhaustive refitting. We demonstrate the practical value across various datasets, where CONFOUNDINGSHAP provides informative explanations of which observed covariates drive confounding and thereby helps to provide more insight for causal inference in practice.

## 1 INTRODUCTION

In causal inference, treatment assignment is often not random, so treated and untreated groups can differ systematically in their characteristics (Imbens & Rubin, 2015; Pearl, 2009; Rubin, 1974). This gives rise to observed **confounding**, where variables influence both treatment assignment and the outcome, making it difficult to attribute observed differences to the treatment itself (VanderWeele & Shpitser, 2013).[1]

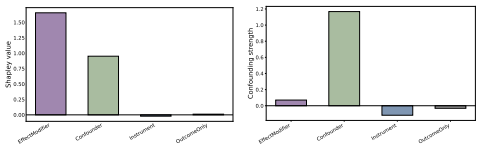



Figure 1: Shown is the difference between variable importance obtained by applying SHAP to a CATE estimator (Kennedy (2023)) vs. confounding strength (*ours*).

To generate new insight for causal inference, we ask: ***which observed covariates drive confounding strength?***

**Example:** *In oncology, targeted cancer therapies are often known to improve survival, but clinicians are more likely to prescribe these treatments to patients with more advanced disease or worse overall health, such as those with later tumor stage (Hu et al., 2023). These characteristics are also strongly related to patient outcomes and therefore act as confounders.*

Identifying which variables drive confounding strength helps determine which patient characteristics are needed to make treatment groups comparable, for example, when choosing a minimal sufficient adjustment set in observational studies or conducting prospective observational studies, such as emulated target trials (Chen et al., 2016; Feuerriegel et al., 2024; Hemkens et al., 2018; Hernán & Robins, 2016; Textor et al., 2017).

Several explanation methods have been developed for causal machine learning (Paillard et al., 2025; Rehill & Biddle, 2024), but with *different* target objects. Examples are methods for variable

[1]Throughout this paper, we mainly rely on the quantitative definition of confounding from Definition 5 in (VanderWeele & Shpitser, 2013).



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.

importance of propensity scores (to explain the treatment assignment) (Gutman et al., 2025), of outcome models (to understand the outcome prediction) (Hu et al., 2022), and of the conditional average treatment effect (CATE) itself (to understand treatment effect heterogeneity) (Svensson et al., 2025). However, none of these help explain confounding strength.

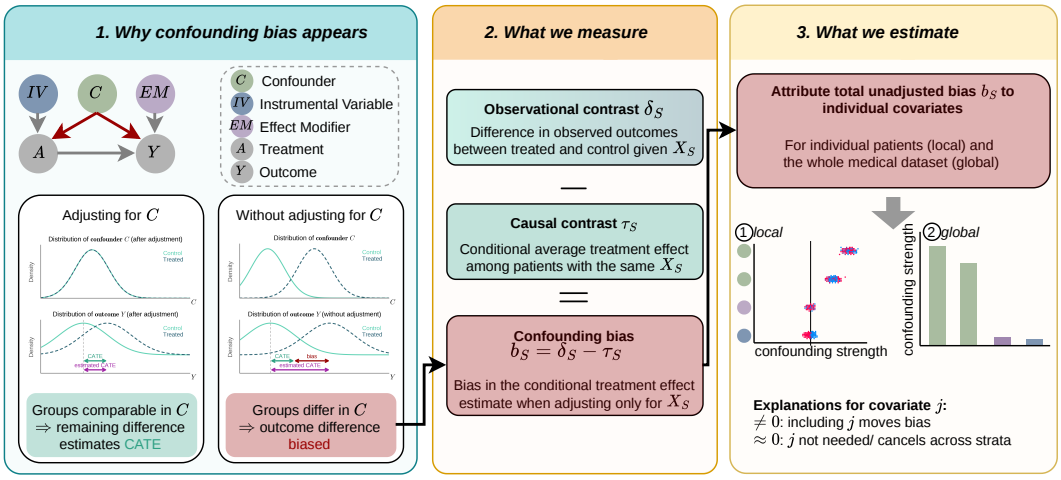



Figure 2: **Overview of our ConfoundingSHAP method.** Therein, residual confounding bias is defined as the gap between observational and causal contrasts and attributed to individual covariates.

Explaining confounding strength is fundamentally different from explaining treatment effect heterogeneity, such as applying SHAP on CATE estimators. The reason is that explanations of the CATE target variation in treatment effects, **not** the confounding strength. Thus, CATE explanations will be dominated by **effect modifiers** and include **observed confounders**, but will **not isolate** them. We demonstrate this difference in Fig. 1. In a special case, explanations of a fitted CATE estimator can even miss confounders. If a confounder shifts treated and untreated outcomes in a similar way, the contributions cancel out in the CATE, and the **observed confounder** does not show up in the variable importance for treatment effect heterogeneity (see Suppl. H.1). Hence, a naïve application of existing SHAP (Lundberg & Lee, 2017) explanations would only explain effect heterogeneity and is thus *ill-suited* for our task of quantifying confounding strength.

In this paper, we propose ConfoundingSHAP: *a Shapley-based method for attributing observed confounding strength to individual covariates*. For this, we define a cooperative game directly over the functional that captures confounding strength. In our method, the value function thus quantifies, for each coalition of covariates, the residual confounding bias that remains in CATE estimation when adjustment is restricted to that coalition. ConfoundingSHAP then decomposes the bias functional into covariate-level Shapley contributions, which yields an attribution of which observed covariates drive confounding bias and by how much. We provide a scalable computational algorithm for this task. Specifically, since our task requires evaluating the value function over many adjustment sets, we propose an efficient TabPFN-based approximation that avoids exhaustive refitting. Our method is designed for the CATE, but can also trivially yield attributions for the average treatment effect (ATE) simply by averaging the bias functional over the covariate distribution.

Our **main contributions**[2] are the following. **(1)** We introduce ConfoundingSHAP: a novel Shapley-based method to measure the confounding strength at the covariate level. **(2)** We propose a novel scalable approximation for ConfoundingSHAP based on TabPFN estimators, which avoids the exhaustive refitting required by the exact bias functional while preserving the interpretation of the underlying cooperative game. **(3)** We demonstrate the practical value across various numerical experiments, where ConfoundingSHAP provides informative explanations of which observed covariates drive confounding strength.

[2]Code: `https://anonymous.4open.science/r/ConfoundingSHAP-703E`. Our method is already integrated into an existing open-source package. To preserve anonymity during review, we provide the implementation separately in an anonymous repository.

## 2 Related Work[3]

**Explanation methods for variable importance.** Explanation methods are widely used for measuring variable importance with respect to a fitted model or predictive task. A first class are methods with local surrogates. A prominent example is LIME (Ribeiro et al., 2016), which provides instance-level explanations. A second class are *Shapley value decompositions*. A prominent example is SHAP (Lundberg & Lee, 2017), which explains local or global model outputs by attributing contributions to individual features based on cooperative game theory at the feature level. Related approaches such as SAGE (Covert et al., 2020) extend this idea to define global importance for a predictive utility. For a detailed review, see Suppl. A. ⇒ *However, these methods explain predictive models rather than confounding strength.*

**Causal Shapley values.** A different literature stream develops causal formulations of Shapley values, for example, by incorporating causal structure into feature attribution through causal graphs, asymmetric coalitions, or intervention-based value functions (Frye et al., 2020; Heskes et al., 2020; Parafita et al., 2025; Witter et al., 2026). ⇒ *However, our work is orthogonal: our focus is* ***not*** *on causal formulations of Shapley attributions, but to use Shapley attributions to understand causal questions.*

Table 1: **Differences ConfoundingSHAP vs. key related work.** Shown is the semantics of Shapley values depending on the causal target.

| Target | $\nu(S)$ | $\phi_j$ explains | Example(s) |
|---|---|---|---|
| ① Propensity score | $\mathbb{E}[\hat{\pi}(X) \mid X_S]$ | treatment assignment | (Gutman et al., 2025; Martini et al., 2022) |
| ② Outcome model | $\mathbb{E}[\hat{\mu}_a(X) \mid X_S]$ | outcome prediction | (Schrod et al., 2022) |
| ③ CATE | $\mathbb{E}[\hat{\tau}(X) \mid X_S]$ | effect heterogeneity | (Bénard & Josse, 2025; Svensson et al., 2025) |
| **Confounding bias** | $-\mathbb{E}[b_S(X_S)]$ | bias under different adjustments | *this paper* |

**Explanations for causal targets.** Several methods focus on different causal targets (see Table 1), based on which Shapley values have *different* semantics:

① **Variable importance for propensity scores (treatment assignment).** Feature importance methods have been used to explain treatment assignment/ propensity score models; for example, to identify variables predictive of treatment allocation and guide propensity score analyses (Gutman et al., 2025; Martini et al., 2022). ⇒ *However, these explanations target who receives treatment, but* ***not*** *whether these variables induce confounding (e.g., an instrument may strongly predict treatment without directly affecting the outcome).*

② **Variable importance for outcome models (outcome prediction).** In medical applications of causal machine learning, SHAP has been used to explain treatment-specific outcome models, i.e., the predicted potential outcomes under a given intervention (such as, e.g., survival or adverse event risk under alternative treatment options (Schrod et al., 2022)). In clinical terminology, such explanations primarily identify *prognostic* variables (i.e., variables associated with outcome risk), but not necessarily *predictive* variables, (i.e., variables that predict differential treatment responses) (Sechidis et al., 2018).

③ **Variable importance for effect heterogeneity (CATE).** Several works study variable importance for CATEs to identify effect modifiers that drive variation in individualized treatment effects (Bénard & Josse, 2025; Hines et al., 2025; Morzywolek et al., 2025; Paillard et al., 2025; Sechidis et al., 2025a; Svensson et al., 2025). These methods are useful for understanding the structure of the estimated CATE, but they do **not** directly explain which covariates drive confounding strength. ⇒ *Variable importance for CATE heterogeneity and for confounding strength answer distinct but complementary questions.*

**Research gap:** To the best of our knowledge, there is no Shapley-based method to explain observed confounding strength. To fill this gap, we propose a method to attribute confounding strength under different adjustment strategies to individual covariates.

[3] We provide an extended related work in Suppl. A.

## 3 SETTING

**Preliminaries: Shapley values.** For a cooperative game with $p$ players and value function $\nu : 2^{[p]} \to \mathbb{R}$, the *Shapley value* of player $j$ is the unique allocation satisfying efficiency, symmetry, linearity, and the dummy axiom (Shapley, 1953):

$$\phi_j(\nu) = \sum_{S \subseteq [p] \setminus \{j\}} \frac{1}{p\binom{p-1}{|S|}} \left[\nu(S \cup \{j\}) - \nu(S)\right]. \tag{1}$$

The efficiency property guarantees $\sum_{j=1}^{p} \phi_j(\nu) = \nu([p]) - \nu(\emptyset)$, meaning that the covariate-level Shapley values add up exactly to the change in the value function from using no covariates to using the full covariate set. In our setting, this is crucial because it allows us to interpret the Shapley values as a decomposition of the total confounding reduction across individual covariates.

**Setup.** We consider the standard causal inference setting for CATE estimation (Rubin, 1974) with population $Z = (X, A, Y) \sim \mathbb{P}$, where $X \in \mathcal{X} \subseteq \mathbb{R}^p$ are observed pre-treatment covariates, $A \in \{0, 1\}$ is a binary treatment indicator, and $Y \in \mathbb{R}$ is a scalar outcome. For a subset $S \subseteq [p]$, we write $X_S \in \mathbb{R}^{|S|}$ for the corresponding subvector and $X_{-S}$ for its complement. We observe an i.i.d. sample $\mathcal{D} = \{(x_i, a_i, y_i)\}_{i=1}^{n}$ of size $n \in \mathbb{N}$.

**CATE.** We work within the potential outcomes framework (Rubin, 1974; 2005). Let $Y(0)$ and $Y(1)$ denote the *potential outcomes* under control and treatment, respectively. Let $\pi(x) := \mathbb{P}(A = 1 \mid X = x)$ denote the *propensity score*, and let $\mu_a(x) := \mathbb{E}[Y \mid X = x, A = a]$ denote the *outcome regression* in treatment arm $a \in \{0, 1\}$. The *conditional average treatment effect (CATE)* is given by $\tau(x) := \mathbb{E}[Y(1) - Y(0) \mid X = x]$, and the *average treatment effect (ATE)* by $\tau := \mathbb{E}[\tau(X)]$.

Causal effects are identified under consistency, positivity, and no unmeasured confounding conditional on the full observed covariate vector $X$ (Rosenbaum & Rubin, 1983b; Rubin, 1974; Van Der Laan & Rubin, 2006). Identification also holds, equivalently, when the full covariate set $X$ is the valid adjustment set. Importantly, a strict subset $X_S$, $S \subsetneq [p]$, may no longer be a valid adjustment set and thus introduces confounding in the CATE, which we aim to quantify.

**Objective.** Our aim is to *quantify the contribution of individual covariates to confounding strength*. We first note the fact that restricting the covariate set to $X_S$, $S \subsetneq [p]$, changes which observed differences between treated and untreated groups are captured. The resulting change in the treatment effect estimate provides a signal of the confounding strength associated with the omitted covariates. Our goal is to attribute this observed confounding strength to individual covariates.

We formalize **confounding bias** (Imbens & Rubin, 2015; Pearl, 2009; Rubin, 1974) at the level of a covariate subset $X_S$ as the difference between two quantities defined at the same information level: the *observational contrast* and the *causal contrast*. We define both as follows.

• **Observational contrast**. Let the *observational contrast*[4] be the subgroup effect wrt. $X_S$, namely the conditional difference in observed outcomes between treated and control units *given the restricted adjustment set* $X_S$, i.e., $\delta_S(x_S) := \mathbb{E}[Y \mid A = 1, X_S = x_S] - \mathbb{E}[Y \mid A = 0, X_S = x_S]$. This quantity captures the observed treatment-control difference within groups defined by $X_S$. If adjustment by $X_S$ is insufficient, $\delta_S(x_S)$ will reflect systematic differences between treated and control units, and therefore the observational contrast is <u>not</u> equal to the causal contrast.

• **Causal contrast**. Let the *causal contrast* be the projected CATE wrt. $X_S$. We use the full covariate set $X$ as the reference adjustment set. We then define the causal contrast at information level $X_S$ as the projection of the full-adjustment CATE onto $X_S$, i.e., $\tau_S(x_S) := \mathbb{E}[\tau(X) \mid X_S = x_S]$. Here, $\tau(X)$ denotes the CATE when adjusting by the full covariate set $X$. Thus, $\tau_S(x_S)$ is the average fully-adjusted treatment effect but only among units sharing $X_S = x_S$.

Then, the *confounding bias* can be quantified as a residual bias from confounding under adjustment by $X_S$, i.e., $b_S(x_S) := \delta_S(x_S) - \tau_S(x_S)$.

**Interpretation of $b_S(x_S)$.** The quantity $b_S(x_S)$ captures the residual bias when adjusting only for $X_S$ rather than all covariates $X$. If $X_S$ contains all confounders (e.g., it is a sufficient adjustment set

[4]Importantly, $\delta_S(x_S)$ is an observational quantity: it compares observed outcomes between the actually treated and actually untreated units conditional on $X_S = x_S$, rather than comparing the potential outcomes $Y(1)$ and $Y(0)$ for the same target population.

or $S = [p]$), then

$$X_S \text{ contains all the confounders} \quad \Rightarrow \quad b_S(x_S) = 0. \tag{2}$$

Hence, $b_S(x_S) \neq 0$ implies that $X_S$ omits some confounders. The converse does not hold: $b_S(x_S) = 0$ need not imply sufficiency, since biases from omitted confounders may cancel during averaging (e.g., $b_\emptyset(x_\emptyset) = b_{[p]}(x) = 0$ while $b_S(x_S) > 0$ for $S \subset [p]$; see Suppl. E).

**Difference to SHAP for CATE estimators.** Our task is fundamentally *different* from feature attribution for fitted CATE estimators (Hines et al., 2025; Liu & Huang, 2025; Morzywolek et al., 2025) as our target is the residual bias $b_S$, not variation in $\tau_S$. This distinction is sharpest under cancellation: if a covariate shifts both potential outcomes similarly, the contributions to $\mu_1(x)$ and $\mu_0(x)$ may largely disappear in $\tau(x) = \mu_1(x) - \mu_0(x)$. Then, the causal contrast for a subset $S$ excluding all confounders may coincide with the full-covariate contrast (i.e., $\tau_S = \tau_{[p]}$) and thus yields zero CATE SHAP, even though the observational contrasts $\delta_S$ and $\delta_{[p]}$ differ (see example in Fig. 1 and Suppl. H.1).

## 4 ConfoundingSHAP

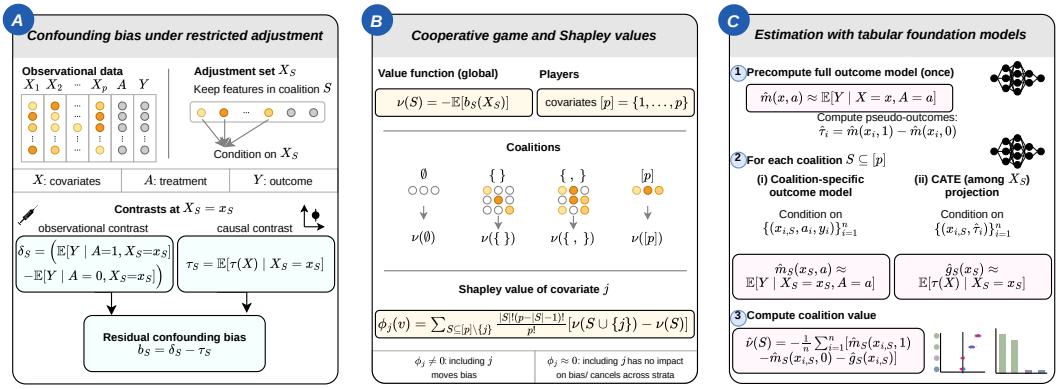



Figure 3: **Overview of ConfoundingSHAP.** Our method defines confounding bias under restricted adjustment (A), introduces a game over this bias functional (B) and then estimates coalition values with tabular foundation models (C).

**Overview.** We introduce ConfoundingSHAP, a Shapley-based method for attributing confounding strength to individual covariates. The method consists of three steps (see Fig. 3). A We first define the residual confounding bias functional $b_S$ that compares the *observational contrast* to the corresponding *causal contrast*, both with the same information level $X_S$. B We then construct a cooperative game for which the value function is defined directly by this bias functional, so that Shapley values attribute bias reduction to individual covariates. C We estimate coalition values for the value function to replace potentially expensive refitting per coalition with an in-context approach (using TabPFN as a state-of-the-art tabular foundation model). The key idea is that the explanation target is not a fitted prediction, propensity, or CATE model. Instead, ConfoundingSHAP defines feature importance with respect to a causal bias functional. The resulting attributions therefore quantify contribution to confounding strength in the data rather than predictive relevance, treatment assignment relevance, or treatment effect heterogeneity.

### A Confounding bias under restricted adjustment

First, we briefly recall the residual confounding bias from Section 3. For a covariate subset $S \subseteq [p]$, the observational contrast $\delta_S(x_S)$ compares treated and control units conditional on $X_S = x_S$, whereas the causal contrast $\tau_S(x_S)$ is the full-adjustment CATE projected onto the same subset. The difference (i.e., $b_S(x_S) = \delta_S(x_S) - \tau_S(x_S)$) measures the residual confounding bias associated with using $X_S$ as the adjustment set (see Panel A in Fig. 3). If $X_S$ is sufficient for adjustment, the two contrasts agree and $b_S(x_S) = 0$; if relevant confounders are missing, residual bias may remain. The collection $\{b_S(X_S) : S \subseteq [p]\}$ therefore describes how confounding bias changes across adjustment sets and provides the input for the cooperative game in step B.

## B A COOPERATIVE GAME OVER RESIDUAL CONFOUNDING BIAS

We now define the cooperative game in which the players are the covariates $j \in [p]$. A coalition of players $S \subseteq [p]$ represents the set of covariates that are included in the adjustment set. The goal of the cooperative game in CONFOUNDINGSHAP is to attribute the residual confounding bias to individual covariates by comparing coalitions that differ only in whether a given covariate is included. We define the game as follows:

**Coalition values.** For an individual with covariates $x$, define the datapoint-level coalition value as

$$\nu_x(S) := -b_S(x_S). \tag{3}$$

The corresponding population-level coalition value averages over the distribution of $X_S$, i.e.,

$$\nu(S) := -\mathbb{E}\left[b_S(X_S)\right]. \tag{4}$$

*Why the negative sign?* We define $\nu(S) = -\mathbb{E}[b_S(X_S)]$ with a negative sign, so that, when comparing a smaller subset $S$ to the larger subset $S \cup \{j\}$, a positive contribution means that the confounding bias is attributed to not yet including $j$. Hence, the sign of the resulting Shapley value matches the direction of the bias attributed to that covariate.

**Covariate-level decomposition.** The efficiency property of Shapley values is the main reason for using the above construction. The property states that the covariate-level attributions sum to the difference between the full and empty conditioning sets, i.e., $\sum_{j=1}^{p} \phi_j(\nu) = \nu([p]) - \nu(\emptyset)$. Since the full conditioning set is the reference adjustment set, we have $b_{[p]}(X) = 0$ and hence $\nu([p]) = 0$. Therefore, $\sum_{j=1}^{p} \phi_j(\nu) = 0 - (-\mathbb{E}[b_\emptyset(X_\emptyset)]) = b_\emptyset$. Thus, the Shapley values decompose the population-level residual bias of the unconditioned comparison into covariate-level contributions. Hence, the Shapley values explain how the difference between no conditioning and full conditioning is distributed across the covariates. Analogously, local Shapley values obtained from the datapoint-level values $\nu_x$ decompose the residual bias for a specific covariate profile $x$: $\sum_{j=1}^{p} \phi_j(\nu_x) = \nu_x([p]) - \nu_x(\emptyset) = b_\emptyset$.

**Causal interpretation of value functions, their differences, and Shapley values.** The value functions $\nu_x(S) = -b_S(x_S)$ and $\nu(S) = -\mathbb{E}[b_S(X_S)]$ measure residual confounding bias relative to the full covariate set. The local value $\nu_x(S)$ describes, for a specific covariate profile $x$, how much signed residual bias remains when the treatment-control comparison is made only within groups defined by $X_S$, or equivalently when the covariates outside $S$ are withheld from the adjustment set. At the same time, the global value $\nu(S)$ is the corresponding population average and quantifies the residual confounding bias wrt. ATE. Hence, the differences of global values (i.e., $\nu(S \cup \{j\}) - \nu(S)$) measure how the *signed residual bias changes when covariate $j$ is added to the adjustment set*. Finally, the Shapley value $\phi_j$ averages these marginal differences over all possible contexts $S$, and therefore *quantifies the average contribution of covariate $j$ to removing signed residual confounding bias*.

**Cancellations.** A zero value can arise in distinct ways: $b_S(x_S) = 0$ from absent confounding or cancellation across strata, and $\phi_j = 0$ from a null player or marginal contributions cancelling across coalitions. Neither implies the other. In practice $\phi_j \neq 0$ identifies covariates with a systematic signed contribution to residual bias, hence, the causal interpretation of our CONFOUNDINGSHAP has a practical relevance for causal estimation.

## C ESTIMATING COALITION VALUES (WITH TABPFN)

In standard SHAP (Lundberg & Lee, 2017) explanations, the cooperative game is usually defined with respect to a fitted model. To evaluate a coalition $S$, the model is queried using the features in $S$, while features outside $S$ are treated as unavailable, for example by masking them or averaging over their conditional distribution. The coalition value $\nu(S)$ is then obtained from the resulting model output. Our target is more challenging because confounding bias is not the output of a single fitted model, so coalition values cannot be obtained by simply masking features in a fixed predictor. To see this, note that changing $S$ changes the conditioning set and therefore the contrasts that define $b_S$. Thus, for each coalition $S$, computing $\nu(S)$ may require fitting new coalition-specific models for the observational contrast and the projected causal contrast. Hence, a naïve implementation would refit these models from scratch for every coalition, which can quickly become computationally prohibitive for large or expensive regressors.[5]

[5] We address the separate question of how many coalitions need to be evaluated for Shapley estimation below, and provide a computational complexity analysis in Suppl. K.

To make coalition-value evaluation scalable, we use TabPFN (Grinsztajn et al., 2026; Hollmann et al., 2023; 2025) as our regression backbone. TabPFN is well suited to this setting because each coalition evaluation can be viewed as a new tabular regression task with a different input subset $X_S$ and, depending on the contrast, a different regression target. Instead of retraining from scratch for every such task, TabPFN uses a pretrained tabular foundation model with frozen parameters and produces predictions by conditioning on the training data provided in context. In our implementation, this allows us to evaluate the coalition-specific regressions needed for $\nu(S)$ *without* updating model parameters for every coalition. Our estimation procedure has three sub-steps (see Panel C in Fig. 3).

**(1) Precompute the full outcome model.** We first estimate the full outcome regression

$$m(x, a) := \mathbb{E}[Y \mid X = x, A = a], \tag{5}$$

using the full covariate vector $X$. Here, the model $m$ is fitted once and is used to form the full-adjustment CATE estimates $\hat{\tau}_i := \hat{m}(x_i, 1) - \hat{m}(x_i, 0)$. The $\hat{\tau}_i$ (also called pseudo-outcomes) provide the reference causal signal that is projected onto covariate subsets, and the population average corresponds to $\bar{\tau} = \mathbb{E}[\tau(X)]$.

**(2) Estimate coalition-specific quantities.** For each coalition $S \subseteq [p]$, we estimate the quantities needed to evaluate $b_S$.[6] First, we estimate the coalition-specific outcome model

$$m_S(x_S, a) := \mathbb{E}[Y \mid X_S = x_S, A = a], \tag{6}$$

which gives the observational contrast $\hat{\delta}_S(x_{i,S}) = \hat{m}_S(x_{i,S}, 1) - \hat{m}_S(x_{i,S}, 0)$. Second, we estimate the projection of the full-adjustment CATE onto the same subset, given by

$$g_S(x_S) := \mathbb{E}[\tau(X) \mid X_S = x_S], \tag{7}$$

where we regress the pseudo-outcomes $\hat{\tau}_i$ on $X_{i,S}$. This yields $\hat{g}_S(x_{i,S})$ as our estimate of the projected causal contrast.

**(3) Compute coalition values.** Combining the above two estimates gives the plug-in estimate of the residual confounding bias for coalition $S$, $\hat{b}_S(x_{i,S}) = \hat{\delta}_S(x_{i,S}) - \hat{g}_S(x_{i,S})$. The population-level coalition value simplifies to $\nu(S) = -\mathbb{E}\left[\delta_S(x_S) - g_S(x_S)\right] = -\mathbb{E}[\delta_S(x_S) - \bar{\tau}]$, where $\bar{\tau} = \mathbb{E}[\tau(X)]$ making the corresponding estimator $\hat{\nu}(S)$ computationally more efficient than its local counterpart.

The plug-in construction separates the full outcome model (1) from the coalition-specific regressions (2). The full outcome model is estimated once and used to compute pseudo-outcomes $\hat{\tau}_i = \hat{m}(x_i, 1) - \hat{m}(x_i, 0)$. For each queried coalition $S$, we then estimate two quantities on $X_S$: the observational contrast $\hat{\delta}_S$ and the projection $\hat{g}_S$ of the pseudo-outcomes.

We evaluate the coalition-specific regressions with TabPFN (Grinsztajn et al., 2026; Hollmann et al., 2023; 2025). Each coalition defines an in-context regression task on $X_S$: one task for the outcome model entering $\hat{\delta}_S$, and one task for projecting the pseudo-outcomes $\hat{\tau}_i$ onto $X_S$. Thus, changing $S$ changes the regression task, *but does not require retraining*. TabPFN is used only to obtain the coalition-value estimates $\hat{\nu}(S)$ efficiently (3). The confounding-bias game itself remains unchanged and is the one defined in Section 4.

**Estimating Shapley values.** After estimating coalition values, the remaining computational challenge is the Shapley aggregation itself. Exact Shapley values require evaluating $\nu$ across all $2^p$ coalitions, which becomes infeasible as the number of covariates grows.

We therefore use Shapley estimation: instead of evaluating all coalitions, the estimator queries at most $B \in \mathbb{N}$ coalitions and reconstructs approximate Shapley values from the resulting $\hat{\nu}(S)$ evaluations. Existing approximators provide three options for this setting (Kolpaczki & Hüllermeier, 2026): Monte-Carlo estimators based on maximum sample reuse (MSR) (Kolpaczki et al., 2024a; Štrumbelj & Kononenko, 2014; Wang & Jia, 2023), regression-based estimators based on weighted least-squares formulations (Fumagalli et al., 2026a;b; Lundberg & Lee, 2017; Musco & Witter, 2025; Witter et al., 2025), and hybrid estimators that combine regression-based proxy models with Monte-Carlo correction (Butler et al., 2025; Witter et al., 2025).

We use RegressionMSR (Witter et al., 2025) for our main experiments in Section 5. RegressionMSR fits a proxy $\tilde{\nu}$ to the queried coalition values, extracts Shapley values from this proxy,

[6] For global coalition values, the projected causal contrast does not need to be estimated separately for every coalition, since $\mathbb{E}[g_S(X_S)] = \mathbb{E}[\mathbb{E}[\tau(X) \mid X_S]] = \mathbb{E}[\tau(X)]$. Thus, the projection step is only required for local coalition values, improving efficiency when only population-level values are needed.

and applies an MSR correction to the residual game $\hat{\nu} - \tilde{\nu}$. We provide an ablation of this choice in Suppl. G where we compare RegressionMSR to OddSHAP (Fumagalli et al., 2026a), a recent proxy-based estimator that regresses on the odd Fourier subspace, and to representatives of the other two classes: MSR (Wang & Jia, 2023) as an MSR estimator and KernelSHAP (Lundberg & Lee, 2017) as a weighted least-squares estimator. Importantly, all four estimators are consistent, i.e. recover the exact Shapley values in the limit $B \to 2^p$.

## 5 Experiments

We evaluate whether ConfoundingSHAP identifies covariates that contribute to residual confounding bias. Our experiments are designed for different objectives: **(1)** We use synthetic data with known covariate roles to test whether our method recovers the **true confounders** and assigns a low confounding strength to **instruments**, **effect modifiers**, and outcome predictors. **(2)** We use the real-world ACTG experiment to test whether our method can detect that confounding strength disappears under random treatment assignment in randomized controlled trials. **(3)** We apply our method to a real-world clinical study (i.e., the SUPPORT right heart catheterization study (Connors et al., 1996)) to analyze whether ConfoundingSHAP can identify variables that are clinically meaningful. We report additional experiments wrt. approximation quality, stability, and runtimes (see Suppl. H and Suppl. I). We report all implementation details in Suppl. F; details about datasets in Suppl. J; and runtimes in Suppl. I.

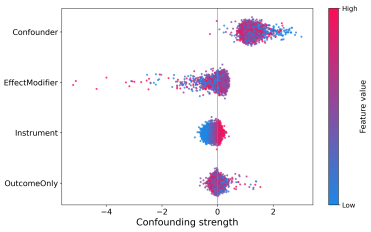
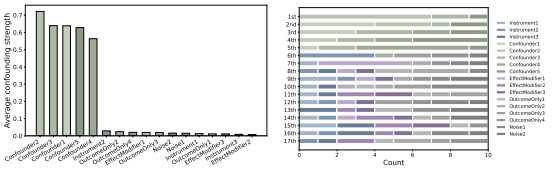



Figure 4: **Low-dimensional covariate setting.** Local values on a synthetic dataset with four covariates; global values for the same setting are in Fig. 1, right. ⇒ *The confounder has a large Shapley value; all others are close to zero.*

### 5.1 Synthetic data with known ground-truth confounders

**Aim.** Our first experiment with synthetic data is designed to benchmark whether ConfoundingSHAP can successfully recover the ground-truth confounders. For this, we use a data-generating process (DGP) from prior work on CATE estimation (Curth & van der Schaar, 2021), which contains confounders but also instruments, effect modifiers, outcome predictors, as well as noise variables.

**(a) Low-dimensional covariate setting.** Fig. 4 shows a simple setting with four covariates. Here, we can use the exact Shapley computation over all coalitions to show that our algorithm has the intended behavior and recovers the correct confounder. For the global attribution, the Shapley values concentrate correctly on the true confounder, while the Shapley value for the instrument, effect modifier, and outcome predictor is close to zero (Fig. 1, right). For local attributions, the same behavior is seen where the contributions for the confounder are large and non-zero, while the contributions for the other covariates are close to zero. ⇒ **Takeaway.** This experiment confirms the effectiveness of ConfoundingSHAP. *Our method assigns the dominant attribution to the confounding variable as intended.*




(a) Avg. global values (b) Rank stability

Figure 5: **Medium-dimensional covariate setting.** Synthetic 17-covariate dataset with five confounders. ⇒ *The confounders have the largest Shapley values and are recovered in the top five ranks across seeds.*

**(b) Medium-dimensional covariate setting.** We next evaluate a 17-covariate setting over 10 random seeds. This setting contains multiple confounders together with instruments, effect modifiers, outcome predictors, and noise variables. Here, we use the scalable algorithm from Section 4. We see that ConfoundingSHAP assigns the largest importance to the true confounders while all other covariates receive close-to-zero values (Fig. 5, left). We further evaluate the rank stability of the covariates with the highest Shapley value (i.e., which our method identifies as most important confounders; Fig. 5, right). ⇒ **Takeaway.** *Our method is robust and, across repeated runs, consistently identifies the true confounders.*

**(c) High-dimensionality covariate setting and additional experiments.** We report additional synthetic experiments with covariate dimensions ranging from 25 to 200 in Suppl. G. We also confirm the rank stability when simulating multiple DGPs with different random seeds in Suppl. H.5. Overall, our method is highly effective in identifying the correct confounders.

**Aim.** The experiment benchmarks whether ConfoundingSHAP successfully recovers ground-truth confounders from real-world covariates with artificially modified treatment and outcome, while assigning near-zero confounding strength to the same covariates when real-world treatment and outcome are used and treatment is randomly assigned.

**Randomized controlled trial setting.** We use the real-world ACTG dataset to assess whether our method can detect the disappearance of confounding strength in a randomized controlled trial. We use the ACTG (Hammer et al., 1996) data for two experiments: (i) we first artificially introduce confounding to the dataset by modifying treatment assignment and outcome and check whether our method can correctly identify the confounders; and (ii) we apply our method to the original dataset including the real-world treatment and outcome, where, given the randomized nature of the treatment assignment, we should not see any confounders. Fig. 6 shows that ConfoundingSHAP successfully assigns the largest confounding strength to the confounders on the data with introduced confounding and that under random assignment, confounding disappears for the same covariates. ⇒ **Takeaway.** *Our method works as intended. ConfoundingSHAP recovers (i) confounders when artificially introduced and (ii) the disappearance of confounders when treatment is assigned at random for the same covariates.*

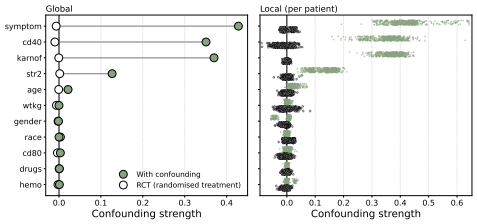



Figure 6: ⇒ *Confounding strength disappears under randomly assigned treatment, both globally (left) and locally (right).*

## 5.2 Real-world application to the SUPPORT right heart catheterization study

**Aim.** Finally, we demonstrate the clinical value of ConfoundingSHAP by using the SUPPORT right heart catheterization study (Connors et al., 1996), where the treatment is right-heart catheterization (RHC) within the first 24 hours in the intensive care unit (ICU). The outcome is 30-day mortality.

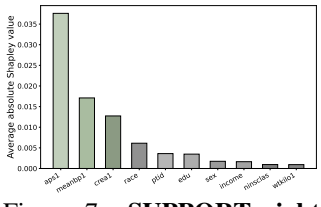



Figure 7: **SUPPORT right heart catheterization study.** Confounding strength averaged over 10 runs. ⇒ Several covariates are clinically plausible confounders.

**Results.** Fig. 7 shows that our method identifies several confounders that align with known confounders from clinical practice. For example, `aps1` (Acute Physiology Score) is a baseline severity-of-illness score summarizing acute physiological derangement, which influences both the likelihood of receiving RHC and the risk of mortality. We added the patient ID `ptid` as a refutation check: the patient ID is no confounder and is correctly attributed close-to-zero values. ⇒ **Takeaway.** *Our method identifies several variables that are plausible confounders in clinical practice, which confirms that our method can be highly informative.*

**Additional validations.** We provide additional validation experiments. We (i) compare exact and estimated Shapley values in a medium-dimensional setting in Suppl. H.3, (ii) study Shapley value estimation over increasing coalition budgets and covariate dimensions with different estimators in Suppl. G, (iii) report stability over repeated synthetic DGP draws in Suppl. H.5. (iv) We evaluate on the **ACIC 2016 benchmark** (Dorie et al., 2019): In this setting, removing the highest-ranked ConfoundingSHAP covariates increases PEHE more than removing randomly selected or low-ranked covariates. ⇒ **Takeaway.** *Our method shows a strong and robust performance.*

# 6 Conclusion

Despite the increasing use of XAI, methods for explaining causal machine learning are still scarce. We introduce a new XAI method to understand observed confounding, thereby enabling more transparent and reliable decision-making.

### AI use statement

In this work, we used generative AI tools to assist with implementing methods, developing and refining theoretical arguments including mathematical derivations, providing feedback on experimental methodology, and generating the synthetic data used in Figure 13. We have not used generative AI tools to clean or reformat datasets. For the Curth–van der Schaar synthetic datasets, we reused existing open-source simulation code based on Curth & van der Schaar (2021), without AI assistance in their generation. Qualitative and thematic data analysis is not applicable to this work. Additionally, we used generative AI tools to create or modify scientific figures, create or edit software code, search for information, and improve manuscript clarity. We have reviewed all AI-assisted work, including the implementation, mathematical arguments, synthetic-data generation procedure, and figures, and revised the material where necessary. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

### Ethics statement

This work is methodological and does not involve human subjects or the collection of personal data. We use existing datasets and do not release new sensitive data. While our method is intended to improve the analysis and interpretation of confounding effects, as with other interpretability methods, its outputs should not be treated as causal evidence without the corresponding assumptions being justified.

### Reproducibility statement

To support reproducibility, we provide anonymous downloadable source code containing the implementation of our method and the code required to reproduce the experiments. The method is also already integrated into an existing open-source software package; to preserve anonymity, the identity of the package will be revealed upon acceptance. Experimental details and implementation choices are documented in the main paper and supplementary material, while additional theoretical details are provided in the appendix.

## References

Clément Bénard and Julie Josse. Variable importance for causal forests: Breaking down the heterogeneity of treatment effects. *Journal of Causal Inference*, 13(1), 2025.

Theresa Blümlein, Joel Persson, and Stefan Feuerriegel. Learning optimal dynamic treatment regimes using causal tree methods in medicine. In *Machine Learning for Healthcare Conference*, 2022.

Landon Butler, Abhineet Agarwal, Justin Singh Kang, Yigit Efe Erginbas, Bin Yu, and Kannan Ramchandran. ProxySPEX: Inference-efficient interpretability via sparse feature interactions in LLMs. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2025.

Javier Castro, Daniel Gómez, and Juan Tejada. Polynomial calculation of the Shapley value based on sampling. *Computers & Operations Research*, 36(5):1726–1730, 2009.

Siu Lun Chau, Robert Hu, Javier Gonzalez, and Dino Sejdinovic. RKHS-SHAP: Shapley values for kernel methods. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2022.

Huigang Chen, Totte Harinen, Jeong-Yoon Lee, Mike Yung, and Zhenyu Zhao. CausalML: Python package for causal machine learning. *arXiv preprint*, arXiv:2002.11631, 2020.

Jie Chen, Arturo Vargas-Bustamante, Karoline Mortensen, and Alexander N. Ortega. Racial and ethnic disparities in health care access and utilization under the affordable care act. *Medical Care*, 54(2):140–146, 2016.

Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *International Conference on Knowledge Discovery and Data Mining (KDD)*, 2016.

Carlos Cinelli and Chad Hazlett. Making sense of sensitivity: Extending omitted variable bias. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 82(1):39–67, 2020.

Alfred F. Connors, Jr., Theodore Speroff, Neal V. Dawson, Charles Thomas, Frank E. Harrell Jr., Douglas Wagner, Norman Desbiens, Lee Goldman, Albert W. Wu, Robert M. Califf, William J. Fulkerson Jr., Humberto Vidaillet, Steven Broste, Paul Bellamy, Joanne Lynn, and William A. Knaus. The effectiveness of right heart catheterization in the initial care of critically ill patients. *JAMA*, 276(11):889–897, 1996.

Ian Covert and Su-In Lee. Improving KernelSHAP: Practical Shapley value estimation via linear regression. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2021.

Ian Covert, Scott Lundberg, and Su-In Lee. Understanding global feature contributions with additive importance measures. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2020.

Alicia Curth and Mihaela van der Schaar. Nonparametric estimation of heterogeneous treatment effects: From theory to learning algorithms. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2021.

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), 2019.

Jacob Dorn, Kevin Guo, and Nathan Kallus. Doubly-valid/Doubly-sharp sensitivity analysis for causal inference with unmeasured confounding. *Journal of the American Statistical Association*, 120(549):331–342, 2025.

Stefan Feuerriegel, Dennis Frauen, Valentyn Melnychuk, Jonas Schweisthal, Konstantin Hess, Alicia Curth, Stefan Bauer, Niki Kilbertus, Isaac S. Kohane, and Mihaela van der Schaar. Causal machine learning for predicting treatment outcomes. *Nature Medicine*, 30(4):958–968, 2024.

Joao Fonseca and Julia Stoyanovich. ExplainerPFN: Towards tabular foundation models for model-free zero-shot feature importance estimations. *arXiv preprint*, arXiv:2601.23068, 2026.

Dennis Frauen, Valentyn Melnychuk, and Stefan Feuerriegel. Sharp bounds for generalized causal sensitivity analysis. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2023.

Dennis Frauen, Fergus Imrie, Alicia Curth, Valentyn Melnychuk, Stefan Feuerriegel, and Mihaela van der Schaar. A neural framework for generalized causal sensitivity analysis. In *International Conference on Learning Representations (ICLR)*, 2024.

Christopher Frye, Colin Rowat, and Ilya Feige. Asymmetric Shapley values: Incorporating causal knowledge into model-agnostic explainability. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2020.

Fabian Fumagalli, Maximilian Muschalik, Patrick Kolpaczki, Eyke Hüllermeier, and Barbara Hammer. SHAP-IQ: Unified approximation of any-order Shapley interactions. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2023.

Fabian Fumagalli, Maximilian Muschalik, Patrick Kolpaczki, Eyke Hüllermeier, and Barbara Hammer. KernelSHAP-IQ: Weighted least-square optimization for Shapley interactions. In *International Conference on Machine Learning (ICML)*, 2024.

Fabian Fumagalli, Landon Butler, Justin Singh Kang, Kannan Ramchandran, and R. Teal Witter. An odd estimator for Shapley values. In *International Conference on Machine Learning (ICML)*, 2026a.

Fabian Fumagalli, R. Teal Witter, and Christopher Musco. PolySHAP: Extending KernelSHAP with interaction-informed polynomial regression. In *International Conference on Learning Representations (ICLR)*, 2026b.

Léo Grinsztajn, Klemens Flöge, Oscar Key, Felix Birkel, Philipp Jund, Brendan Roof, Benjamin Jäger, Dominik Safaric, Simone Alessi, Adrian Hayler, Mihir Manium, Rosen Yu, Felix Jablonski, Shi Bin Hoo, Anurag Garg, Jake Robertson, Magnus Bühler, Vladyslav Moroshan, Lennart Purucker, Clara Cornu, Lilly Charlotte Wehrhahn, Alessandro Bonetto, Bernhard Schölkopf, Sauraj Gambhir, Noah Hollmann, and Frank Hutter. TabPFN-2.5: Advancing the state of the art in tabular foundation models, February 2026.

Rom Gutman, Shimon Sheiba, Omer Noy Klein, Naama Dekel Bird, Amit Gruber, Doron Aronson, Oren Caspi, and Uri Shalit. From observational data to clinical recommendations: A causal framework for estimating patient-level treatment effects and learning policies. *arXiv preprint*, arXiv:2507.11381, 2025.

Scott M. Hammer, David A. Katzenstein, Michael D. Hughes, Holly Gundacker, Robert T. Schooley, Richard H. Haubrich, W. Keith Henry, Michael M. Lederman, John P. Phair, Manette Niu, Martin S. Hirsch, and Thomas C. Merigan. A trial comparing nucleoside monotherapy with combination therapy in HIV-infected adults with CD4 cell counts from 200 to 500 per cubic millimeter. *New England Journal of Medicine*, 335(15):1081–1090, 1996.

Lars G. Hemkens, Hannah Ewald, Florian Naudet, Aviv Ladanie, Jonathan G. Shaw, Gautam Sajeev, and John P.A. Ioannidis. Interpretation of epidemiologic studies very often lacked adequate consideration of confounding. *Journal of Clinical Epidemiology*, 93:94–102, 2018.

Miguel A. Hernán and James M. Robins. Using big data to emulate a target trial when a randomized trial is not available. *American Journal of Epidemiology*, 183(8):758–764, 2016.

Tom Heskes, Evi Sijben, Ioan Gabriel Bucur, and Tom Claassen. Causal Shapley values: Exploiting causal knowledge to explain individual predictions of complex models. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2020.

Konstantin Hess, Dennis Frauen, Valentyn Melnychuk, and Stefan Feuerriegel. Efficient and Sharp Off-Policy Learning under Unobserved Confounding. In *International Conference on Learning Representations (ICLR)*, 2026.

Oliver J. Hines, Karla Diaz-Ordaz, and Stijn Vansteelandt. Variable importance measures for heterogeneous treatment effects. *Biometrics*, 81(4):ujaf140, 2025.

Noah Hollmann, Samuel Müller, Katharina Eggensperger, and Frank Hutter. TabPFN: A transformer that solves small tabular classification problems in a second. In *International Conference on Learning Representations (ICLR)*, 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, January 2025.

Chang Hu, Lu Li, Weipeng Huang, Tong Wu, Qiancheng Xu, Juan Liu, and Bo Hu. Interpretable machine learning for early prediction of prognosis in sepsis: A discovery and validation study. *Infectious Diseases and Therapy*, 11(3):1117–1132, 2022.

Szu-Han Hu, Ming-Yii Huang, Chung-Yu Chen, and Hui-Min Hsieh. Treatment patterns of targeted and nontargeted therapies and survival effects in patients with locally advanced head and neck cancer in Taiwan. *BMC Cancer*, 23(1):567, June 2023.

Guido W. Imbens and Donald B. Rubin. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press, 1 edition, April 2015.

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

Patrick Kolpaczki and Eyke Hüllermeier. Approximation algorithms for the Shapley value: Taxonomy and properties. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 2026.

Patrick Kolpaczki, Viktor Bengs, Maximilian Muschalik, and Eyke Hüllermeier. Approximating the Shapley value without marginal contributions. In *Conference on Artificial Intelligence (AAAI)*, 2024a.

Patrick Kolpaczki, Maximilian Muschalik, Fabian Fumagalli, Barbara Hammer, and Eyke Hüllermeier. SVARM-IQ: Efficient approximation of any-order Shapley interactions through stratification. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2024b.

Lauren D. Liao and Samuel D. Pimentel. jointVIP: Prioritizing variables in observational study design with joint variable importance plot in R. *Journal of Open Source Software*, 9(103):6093, 2024. doi: 10.21105/joss.06093. URL https://joss.theoj.org/papers/10.21105/joss.06093.

Zihuan Liu and Xin Huang. Explaining individualized treatment rules: Integrating LIME and SHAP with XGBOOST in precision medicine. *Statistics in Medicine*, 44(28-30):e70322, 2025.

Alexander R. Luedtke, Ivan Diaz, and Mark J. van der Laan. The statistics of sensitivity analyses. *U.C. Berkeley Division of Biostatistics Working Paper Series*, pp. Working Paper 341, 2015.

Scott Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2017.

Scott M. Lundberg, Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan M. Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, and Su-In Lee. From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence*, 2(1):56–67, 2020.

David Martens, Galit Shmueli, Theodoros Evgeniou, Kevin Bauer, Christian Janiesch, Stefan Feuerriegel, Sebastian Gabel, Sofie Goethals, Travis Greene, Nadja Klein, Mathias Kraus, Niklas Kühl, Claudia Perlich, Wouter Verbeke, Alona Zharova, Patrick Zschech, and Foster Provost. Beware of "Explanations" of AI. *Business & Information Systems Engineering (BISE)*, 2026.

Michael L. Martini, Sean N. Neifert, William H. Shuman, Emily K. Chapman, Alexander J. Schüpper, Eric K. Oermann, J Mocco, Michael Todd, James C. Torner, Andrew Molyneux, Stephan Mayer, Peter Le Roux, Mervyn D. I. Vergouwen, Gabriel J. E. Rinkel, George K. C. Wong, Peter Kirkpatrick, Audrey Quinn, Daniel Hänggi, Nima Etminan, Walter M. Van Den Bergh, Blessing N. R. Jaja, Michael Cusimano, Tom A. Schweizer, Jose I. Suarez, Hitoshi Fukuda, Sen Yamagata, Benjamin Lo, Airton Leonardo De Oliveira Manoel, Hieronymus D. Boogaarts, and R. Loch Macdonald. Rescue therapy for vasospasm following aneurysmal subarachnoid hemorrhage: A propensity score–matched analysis with machine learning. *Journal of Neurosurgery*, 136(1): 134–147, 2022.

Alec McClean, Zach Branson, and Edward H. Kennedy. Calibrated sensitivity models. *Biometrika*, pp. 113(2):asag001, 2026.

Valentyn Melnychuk, Dennis Frauen, and Stefan Feuerriegel. Partial Counterfactual Identification of Continuous Outcomes with a Curvature Sensitivity Model. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2023.

Pawel Morzywolek, Peter B. Gilbert, and Alex Luedtke. Inference on local variable importance measures for heterogeneous treatment effects. *arXiv preprint*, arXiv:2510.18843, 2025.

Maximilian Muschalik, Hubert Baniecki, Fabian Fumagalli, Patrick Kolpaczki, Barbara Hammer, and Eyke Hüllermeier. Shapiq: Shapley interactions for machine learning. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2024a.

Maximilian Muschalik, Fabian Fumagalli, Barbara Hammer, and Eyke Hüllermeier. Beyond Tree-SHAP: Efficient computation of any-order Shapley interactions for tree ensembles. In *Conference on Artificial Intelligence (AAAI)*, volume 38, pp. 14388–14396, 2024b.

Maximilian Muschalik, Fabian Fumagalli, Paolo Frazzetto, Janine Strotherm, Luca Hermes, Alessandro Sperduti, Eyke Hüllermeier, and Barbara Hammer. Exact computation of any-order Shapley interactions for graph neural networks. In *International Conference on Learning Representations (ICLR)*, 2025.

Christopher Musco and R. Teal Witter. Provably accurate Shapley value estimation via leverage score sampling. In *International Conference on Learning Representations (ICLR)*, 2025.

Alexander Nadel and Ron Wettenstein. From decision trees to boolean logic: A fast and unified SHAP algorithm. In *Conference on Artificial Intelligence (AAAI)*, volume 40, pp. 24476–24485, 2026.

Joseph Paillard, Angel Reyero Lobo, Vitaliy Kolodyazhniy, Bertrand Thirion, and Denis A. Engemann. Measuring variable importance in heterogeneous treatment effects with confidence. In *International Conference on Machine Learning (ICML)*, 2025.

Álvaro Parafita, Tomas Garriga, Axel Brando, and Francisco J. Cazorla. Practical Do-Shapley explanations with estimand-agnostic causal inference. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2025.

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

Patrick Rehill and Nicholas Biddle. Transparency challenges in policy evaluation with causal machine learning: Improving usability and accountability. *Data & Policy*, 6:e43, 2024.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why should I trust you?": Explaining the predictions of any classifier. In *International Conference on Knowledge Discovery and Data Mining (KDD)*, 2016.

James M. Robins, Andrea Rotnitzky, and Daniel O. Scharfstein. Sensitivity analysis for selection bias and unmeasured confounding in missing data and causal inference models. In Willard Miller, M. Elizabeth Halloran, and Donald Berry (eds.), *Statistical Models in Epidemiology, the Environment, and Clinical Trials*, volume 116, pp. 1–94. Springer New York, New York, NY, 2000.

P. R. Rosenbaum and D. B. Rubin. Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 45(2):212–218, January 1983a.

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, 1983b.

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

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

S Schrod, A Schäfer, S Solbrig, R Lohmayer, W Gronwald, P J Oefner, T Beißbarth, R Spang, H U Zacharias, and M Altenbuchinger. BITES: Balanced individual treatment effect for survival data. *Bioinformatics*, 38(Supplement_1):i60–i67, June 2022.

Konstantinos Sechidis, Konstantinos Papangelou, Paul D Metcalfe, David Svensson, James Weatherall, and Gavin Brown. Distinguishing prognostic and predictive biomarkers: An information theoretic approach. *Bioinformatics*, 34(19):3365–3376, 2018.

Konstantinos Sechidis, Sophie Sun, Yao Chen, Jiarui Lu, Cong Zhang, Mark Baillie, David Ohlssen, Marc Vandemeulebroecke, Rob Hemmings, Stephen Ruberg, and Björn Bornkamp. WATCH: A workflow to assess treatment effect heterogeneity in drug development for clinical trial sponsors. *Pharmaceutical Statistics*, 24(2):e2463, 2025a.

Konstantinos Sechidis, Cong Zhang, Sophie Sun, Yao Chen, Asher Spector, and Björn Bornkamp. Using individualized treatment effects to assess treatment effect heterogeneity. *Statistics in Medicine*, 44(28-30):e70324, 2025b.

L. S. Shapley. A value for n-person games. In Harold William Kuhn and Albert William Tucker (eds.), *Contributions to the Theory of Games (AM-28), Volume II*, pp. 307–318. Princeton University Press, 1953.

Yan Shi, Ying Zou, Jihua Liu, Yuanyuan Wang, Yingbin Chen, Fang Sun, Zhi Yang, Guanghe Cui, Xijun Zhu, Xu Cui, and Feifei Liu. Ultrasound-based radiomics XGBoost model to assess the risk of central cervical lymph node metastasis in patients with papillary thyroid carcinoma: Individual application of SHAP. *Frontiers in Oncology*, 12:897596, 2022.

Yishai Shimoni, Ehud Karavani, Sivan Ravid, Peter Bak, Tan Hung Ng, Sharon Hensley Alford, Denise Meade, and Yaara Goldschmidt. An evaluation toolkit to guide model selection and cohort definition in causal inference, June 2019.

Susan M. Shortreed and Ashkan Ertefaie. Outcome-adaptive lasso: Variable selection for causal inference. *Biometrics*, 73(4):1111–1122, 2017. doi: 10.1111/biom.12679. URL `https://pubmed.ncbi.nlm.nih.gov/28273693/`.

Erik Štrumbelj and Igor Kononenko. Explaining prediction models and individual predictions with feature contributions. *Knowledge and Information Systems*, 41(3):647–665, 2014.

Qingfeng Sun, Kai Zhang, Yuanlong Xu, Mengmei Luo, Zhouzhou Yang, Qianyu Liu, Sang Liu, and Aimei Liu. Explainable machine learning for predicting clinical outcomes in HIV/TB co-infection: A comparative retrospective study. *BMC Infectious Diseases*, 25(1):1589, 2025.

David Svensson, Erik Hermansson, Nikolaos Nikolaou, Konstantinos Sechidis, and Ilya Lipkovich. Overview and practical recommendations on using Shapley values for identifying predictive biomarkers via CATE modeling. *arXiv preprint*, arXiv:2505.01145, 2025.

Zhiqiang Tan. A distributional approach for causal inference using propensity scores. *Journal of the American Statistical Association*, 101(476):1619–1637, 2006.

Constantine Tarabanis, Evangelos Kalampokis, Mahmoud Khalil, Carlos L. Alviar, Larry A. Chinitz, and Lior Jankelson. Explainable SHAP-XGBoost models for in-hospital mortality after myocardial infarction. *Cardiovascular Digital Health Journal*, 4(4):126–132, 2023.

Johannes Textor, Benito Van Der Zander, Mark S. Gilthorpe, Maciej Liśkiewicz, and George T.H. Ellison. Robust causal inference using directed acyclic graphs: The R package 'dagitty'. *International Journal of Epidemiology*, pp. dyw341, 2017.

Mark J. Van Der Laan and Daniel Rubin. Targeted maximum likelihood learning. *The International Journal of Biostatistics*, 2(1), 2006.

Tyler J. VanderWeele and Peng Ding. Sensitivity analysis in observational research: Introducing the E-value. *Annals of Internal Medicine*, 167(4):268–274, 2017.

Tyler J. VanderWeele and Ilya Shpitser. On the definition of a confounder. *The Annals of Statistics*, 41(1), 2013.

Jiachen T. Wang and Ruoxi Jia. Data Banzhaf: A robust data valuation framework for machine learning. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2023.

Jiachen T. Wang, Prateek Mittal, and Ruoxi Jia. Efficient data Shapley for weighted nearest neighbor algorithms. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2024.

R. Teal Witter, Yurong Liu, and Christopher Musco. Regression-adjusted monte carlo estimators for Shapley values and probabilistic values. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2025.

R. Teal Witter, Álvaro Parafita, Tomas Garriga, Maximilian Muschalik, Fabian Fumagalli, Axel Brando, and Lucas Rosenblatt. Exactly computing Do-Shapley values. *arXiv preprint*, arXiv:2602.07203, 2026.

Peng Yu, Chao Xu, Albert Bifet, and Jesse Read. Linear TreeShap. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2022.

Artjom Zern, Klaus Broelemann, and Gjergji Kasneci. Interventional SHAP values and interaction values for piecewise linear regression trees. In *Conference on Artificial Intelligence (AAAI)*, volume 37, pp. 11164–11173, June 2023.

# A Extended related work

**Explainability.** Explainability methods for machine learning provide the methodological backbone for our attribution construction. SHAP formalizes feature attributions as an additive explanation problem and assigns feature-level contributions to individual model predictions using Shapley values (Lundberg & Lee, 2017). LIME explains individual predictions through an interpretable local surrogate model fitted around the instance of interest (Ribeiro et al., 2016). SAGE is especially close in spirit because it also defines a Shapley-value game through a deliberately chosen value function (Covert et al., 2020); however, its value function is based on predictive performance, whereas ours is based on a causal bias functional. Thus, our work builds on the game-theoretic attribution perspective of explainable machine learning, but changes the object being explained: rather than attributing to a prediction or predictive performance, we attribute to confounding.

**Explaining treatment assignment.** A related line of applied work uses explainability for treatment assignment or propensity score modeling. In clinical applications, SHAP values have been used to interpret machine learning models for treatment allocation, for example, by identifying variables predictive of rescue therapy after aneurysmal subarachnoid hemorrhage and using these variables in subsequent propensity score matched analysis (Martini et al., 2022). SHAP has also been proposed as a clinical face-validity check for propensity score models in frameworks for estimating patient-level treatment effects and learning treatment policies from observational data, where it can summarize the main predictors of an estimated treatment assignment mechanism (Gutman et al., 2025). These workflows show that explainability can help inspect whether treatment assignment models reflect plausible clinical decision factors. Nevertheless, treatment assignment importance is not equivalent to confounding strength: an instrumental variable may strongly predict treatment while having no direct effect on the outcome, whereas confounding requires relevance to both treatment assignment and outcome.

**Explaining outcome prediction.** In applied machine learning, SHAP is widely used to interpret outcome-prediction models. For example, SHAP has been used to explain early prediction of in-hospital mortality in sepsis, in-hospital mortality after myocardial infarction, clinical outcomes in HIV/TB co-infection, and ultrasound-radiomics XGBoost predictions of central cervical lymph-node metastasis in papillary thyroid carcinoma (Hu et al., 2022; Shi et al., 2022; Sun et al., 2025; Tarabanis et al., 2023). These studies illustrate the practical importance of interpretable prediction in medicine. However, outcome-prediction explanations answer a prognostic question: which covariates drive the predicted risk? This differs from the causal question of which covariates induce confounding bias in a treatment effect estimate. A strong prognostic variable need not be a strong confounder if it does not also affect treatment assignment.

**Treatment effect heterogeneity and CATE explanations.** Explainability has also been studied in the context of treatment effect estimation and individualized decision making. A growing literature develops variable-importance measures for heterogeneous treatment effects. PermuCATE uses conditional permutation importance to assess variable importance for CATE estimation, while variable-importance measures for causal forests have been proposed to quantify how covariates contribute to treatment effect heterogeneity (Bénard & Josse, 2025; Paillard et al., 2025). Related work develops treatment effect variable importance measures and inference procedures for local variable importance in heterogeneous treatment effects (Hines et al., 2025; Morzywolek et al., 2025). Several recent papers also combine SHAP or related explainability tools with CATE or individualized treatment rule modeling. For example, SHAP has been used to identify predictive biomarkers through CATE models, to support covariate ranking for effect modification in clinical drug development, and to interpret XGBoost-based individualized treatment rule models in precision medicine (Blümlein et al., 2022; Sechidis et al., 2025b; Svensson et al., 2025). Here, the target is effect modification, individualized treatment benefit, or learned treatment rules. In contrast, our target is the bias induced by restricted adjustment, not heterogeneity of the causal effect itself.

**Causal Shapley explanations.** Our work is also orthogonal to causal variants of Shapley explanations. Causal Shapley values incorporate causal knowledge into feature attributions for individual predictions, while asymmetric Shapley values restrict coalitions according to causal ordering information (Frye et al., 2020; Heskes et al., 2020). More recently, do-Shapley explanations use intervention-based reasoning to define causal attributions (Parafita et al., 2025; Witter et al., 2026). However, the target remains a model output, prediction, or intervention-based explanation of a causal

estimand. Our contribution is different: we define the cooperative game directly on a confounding bias functional, so that the resulting Shapley values quantify contributions to confounding strength.

**Approximation of Shapley values and interactions.** Exact Shapley value computation is generally exponential in the number of features, so a large literature studies scalable approximations. Existing methods can broadly be divided into model-specific and model-agnostic approaches. Model-specific methods exploit structural assumptions on the value function $\nu$, with tree-based games being the most prominent case and enabling polynomial- or linear-time extraction algorithms (Lundberg et al., 2020; Nadel & Wettenstein, 2026; Yu et al., 2022; Zern et al., 2023). Efficient procedures have also been proposed for kernel models and $k$-nearest-neighbor predictors (Chau et al., 2022; Wang et al., 2024). Model-agnostic methods instead treat $\nu$ as a black box, typically using Monte Carlo estimation, regression-based estimation, or hybrid surrogate approaches (Castro et al., 2009; Covert et al., 2020; Fumagalli et al., 2026a;b; Kolpaczki et al., 2024a; Lundberg & Lee, 2017; Musco & Witter, 2025; Štrumbelj & Kononenko, 2014; Wang & Jia, 2023) as follows: Monte Carlo methods estimate expected marginal contributions, while regression-based methods recover Shapley values as coefficients in a weighted regression problem. Recent work improves these estimators through stratified sampling, paired sampling, leverage-score designs, and more expressive polynomial regressions (Covert & Lee, 2021; Fumagalli et al., 2026b; Kolpaczki et al., 2024a; Musco & Witter, 2025). Hybrid methods such as ProxySPEX and RegressionMSR learn an approximation $\hat{\nu}$ before extracting Shapley values, with RegressionMSR additionally accounting for the residual $\nu - \hat{\nu}$ (Butler et al., 2025; Witter et al., 2025). Related approximation ideas have also been extended from feature attributions to Shapley interactions, including Monte Carlo, KernelSHAP-style, tree-based, and graph-neural-network-specific methods (Fumagalli et al., 2023; 2024; Kolpaczki et al., 2024b; Muschalik et al., 2024b; 2025).

**Unmeasured confounding and sensitivity analysis.** Sensitivity analysis provides a complementary perspective on causal inference under unmeasured confounding. Rather than point-identifying a causal estimand under no unobserved confounding, it relaxes this assumption through sensitivity models and derives bounds for causal quantities (Cinelli & Hazlett, 2020; Robins et al., 2000; Rosenbaum & Rubin, 1983a; VanderWeele & Ding, 2017). A prominent line of work uses marginal sensitivity models, where hidden confounding is parameterized by bounding deviations between the observed and complete treatment assignment mechanisms, with recent extensions deriving sharp bounds and generalizing this perspective to continuous treatments and broader causal queries (Dorn et al., 2025; Frauen et al., 2023; 2024; Hess et al., 2026; Melnychuk et al., 2023; Tan, 2006). Our work differs in the object being quantified. Sensitivity analysis asks how causal conclusions could change under hypothetical unmeasured confounding. We instead focus on measured covariates and attribute the bias that arises when adjustment is restricted to subsets of those covariates.

**TabPFN-based approximation.** From a computational perspective, our work is related to prior-data fitted networks and tabular foundation models. TabPFN demonstrates that prior-data fitted networks can provide strong performance for small tabular prediction tasks by training on synthetic tabular data (Grinsztajn et al., 2026; Hollmann et al., 2023; 2025). ExplainerPFN investigates whether meaningful Shapley-value estimates can be obtained in a zero-shot setting using a TabPFN-style model trained on synthetic structural causal models (Fonseca & Stoyanovich, 2026). Our use of TabPFN is different: we use TabPFN-based estimation to make repeated evaluation of coalition-specific causal quantities scalable, while the Shapley game itself remains defined on confounding bias rather than predictive-model explanations.

**Observed-covariate prioritization and selection.** Existing approaches also identify observed covariates relevant for confounding adjustment. The joint variable importance plot (jointVIP) combines treatment imbalance with outcome correlation to provide variable-wise unadjusted bias measures under linear working models (Liao & Pimentel, 2024). Outcome-Adaptive Lasso uses outcome associations to guide variable selection for propensity score estimation, retaining confounders and prognostic variables while excluding unnecessary treatment predictors (Shortreed & Ertefaie, 2017). Unlike these approaches, ConfoundingSHAP defines a cooperative game directly over residual confounding bias, measured as the difference between observational and projected causal contrasts under restricted adjustment. Its Shapley values attribute changes in this bias across adjustment coalitions to individual covariates, providing local and global decompositions rather than variable-wise prioritization or adjustment-set selection.

## B Relation to sensitivity models

Interestingly, our quantifier of observed confounding (i.e., residual bias $b_S$) has a connection to existing sensitivity models (McClean et al., 2026) (where the main aim is to upper-bound the strength of unobserved confounding). Specifically, our residual bias $b_S$ is the closest to the sensitivity model of Luedtke et al. (2015): it is easy to see that their sensitivity parameter upper-bounds the *absolute* average of our residual bias $|\mathbb{E}[b_S(X_S)]|$. However, our *signed* values of residual bias allow to quantify not only the magnitude but also the direction of the confounding bias.

## C JointVIP

We additionally compare ConfoundingSHAP against the published jointVIP bias measure (Liao & Pimentel, 2024), which we reimplement in Python following the reference R implementation. For each covariate, jointVIP multiplies the standardized mean difference between treatment groups by its Pearson outcome correlation. Following the published procedure, we estimate outcome correlations and standard deviations using a separate control-only pilot sample comprising 20% of controls, which is excluded from the analysis sample. Both methods use the same original synthetic datasets with 4, 11, and 17 covariates, although ConfoundingSHAP uses all 5,000 observations. We apply jointVIP to the original covariates without nonlinear transformations and normalize absolute importance scores separately for each method and setting.

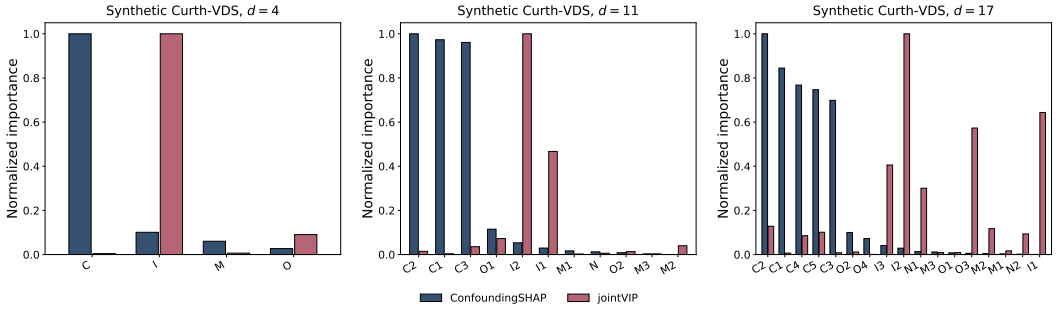



Figure 8: Comparison of ConfoundingSHAP and the published jointVIP bias measure, reimplemented in Python, on synthetic datasets with 4, 11, and 17 covariates. Absolute scores are normalized independently within each method and setting to compare rankings; their magnitudes are not directly comparable.

As shown in Fig. 8, ConfoundingSHAP concentrates attribution on the ground-truth confounders, whereas jointVIP assigns relatively high normalized scores to non-confounding variables. This reflects the quadratic structure of our synthetic DGP: treatment assignment and outcomes depend on squared confounders, whose raw-variable mean differences and Pearson outcome correlations vanish at the population level. The comparison thus illustrates a limitation of the published jointVIP measure when applied without nonlinear feature transformations. Unlike jointVIP's variable-wise bias measure, ConfoundingSHAP attributes changes in residual confounding bias across adjustment coalitions.

# D Relations between SHAP for nuisance functions and our ConfoundingSHAP

In the following, we briefly clarify the relation between ConfoundingSHAP and standard SHAP explanations of the nuisance functions $\pi(x) = \mathbb{P}(A = 1 \mid X = x)$ and $\mu_a(x) = \mathbb{E}[Y \mid X = x, A = a]$, $a \in \{0, 1\}$. Although these nuisance functions enter the definition of confounding bias, their individual SHAP values are **not** sufficient to recover ConfoundingSHAP values.

**Alternative expression for $b_S$.** Recall that, for an adjustment set $S \subseteq [p]$, the residual confounding bias is

$$b_S(x_S) = \delta_S(x_S) - \tau_S(x_S), \tag{8}$$

where $\delta_S(x_S)$ is the observational treated-control contrast conditional on $X_S = x_S$, and where $\tau_S(x_S) = \mathbb{E}[\tau(X) \mid X_S = x_S]$ is the corresponding projected causal contrast. ConfoundingSHAP then uses residual confounding bias $b_S$ to define the coalition values. These coalition values differ from standard coalition values for $\pi$, $\mu_0$, or $\mu_1$, which explain treatment assignment or outcome prediction rather than residual bias. To see that, we make use of the following identity. Let

$$e_S(x_S) = \mathbb{E}[\pi(X) \mid X_S = x_S]. \tag{9}$$

Then, under the full-adjustment identification assumptions, the residual bias can be written as

$$b_S(x_S) = \frac{\mathrm{Cov}(\pi(X), \mu_1(X) \mid X_S = x_S)}{e_S(x_S)} + \frac{\mathrm{Cov}(\pi(X), \mu_0(X) \mid X_S = x_S)}{1 - e_S(x_S)}. \tag{10}$$

Thus, residual confounding is governed by the conditional alignment between treatment assignment and the potential-outcome regressions. This shows why **SHAP values for the nuisance functions alone cannot determine ConfoundingSHAP**: ordinary nuisance SHAP values describe separate predictive relevance, whereas confounding depends on coalition-specific conditional covariance between treatment assignment and outcome levels. Hence, ordinary SHAP values are neither sufficient nor necessary for ConfoundingSHAP.

**Why not sufficient.** Consequently, large nuisance SHAP values are not sufficient for large ConfoundingSHAP values. A variable may strongly predict treatment assignment but not outcomes, as in the case of an instrument; or it may strongly predict outcomes but not treatment assignment, as in the case of an outcome-only prognostic variable. Such variables can have large SHAP values for $\pi$ or $\mu_a$ while contributing little to residual confounding bias.

**Why not necessary.** Large nuisance SHAP values are also not necessary in general. For example, suppose $X_1 = U + \varepsilon_1$, $X_2 = U + \varepsilon_2$, and $X_3 = U$, with mutually independent noise terms. Let treatment depend only on $X_1$, so that $\pi(X)$ is a function of $X_1$, and let the outcome regressions depend only on $X_2$. Then, under standard model-output SHAP, $X_3$ may receive zero attribution for all nuisance functions $\pi, \mu_0, \mu_1$. Nevertheless, adjusting for $X_3$ can remove the association between $X_1$ and $X_2$, and therefore substantially reduce residual confounding. Hence, $X_3$ can have a large ConfoundingSHAP value despite small or zero nuisance SHAP values. This illustrates that ConfoundingSHAP measures adjustment value, and not direct predictive relevance.

**Absence of monotonicity.** There is therefore **no general monotonicity relation** of the form

$$|\phi_j^{\mathrm{conf}}| \text{ large} \Rightarrow |\phi_j^{\pi}|,\ |\phi_j^{\mu_0}|,\ |\phi_j^{\mu_1}| \text{ large}, \tag{11}$$

nor the reverse implication. Shapley monotonicity applies when comparing marginal contributions within the same cooperative game. Here, however, nuisance SHAP and ConfoundingSHAP are Shapley values of different games with different value functions.

**When can monotonicity hold?** Only under restrictive orthogonal-additive settings does a product-like relation emerge. For instance, if the covariates are independent, the nuisance functions are additive, and the denominators $e_S(x_S)$ are treated as approximately constant, then the contribution of covariate $j$ to residual confounding is proportional to a treatment-relevance term times an outcome-relevance term. Outside such special cases, correlated covariates, proxies, nonlinearities, interactions, and cancellations break any monotone relationship between nuisance SHAP values and ConfoundingSHAP.

**Summary.** In summary, SHAP values for $\pi, \mu_0, \mu_1$ can be useful diagnostics for understanding which variables are predictive of treatment assignment or outcomes. They are, however, neither

sufficient nor necessary for ConfoundingSHAP. ConfoundingSHAP targets the coalition-specific residual-bias functional and therefore captures how covariates reduce treatment–outcome imbalance under adjustment.

## E Cancellation example

We give a simple population example showing that $b_\emptyset = 0$ need not imply that all partially adjusted biases are zero. Let $X_1, X_2 \in \{0, 1\}$ be independent Bernoulli random variables with $\Pr(X_1 = x_1, X_2 = x_2) = 1/4$ for all $(x_1, x_2)$. Let treatment be generated as

$$A \mid X_1 = x_1, X_2 = x_2 \sim \text{Bernoulli}(\pi_{x_1 x_2}), \tag{12}$$

with

| $X_1$ | $X_2$ | $\pi_{x_1 x_2}$ | $m(x_1, x_2)$ |
|---|---|---|---|
| 0 | 0 | $1/3$ | $0$ |
| 0 | 1 | $3/10$ | $-3/2$ |
| 1 | 0 | $1/10$ | $-4/3$ |
| 1 | 1 | $9/10$ | $-7/6$. |

(13)

Let the potential outcomes be

$$Y(0) = m(X_1, X_2) + \varepsilon, \qquad Y(1) = m(X_1, X_2) + \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, \sigma^2), \tag{14}$$

with $\varepsilon$ independent of $(X_1, X_2, A)$ conditional on $(X_1, X_2)$. Then, $Y$ is continuous, and the true treatment effect is zero:

$$\tau(X) = Y(1) - Y(0) = 0. \tag{15}$$

Hence, for every subset $S \subseteq \{1, 2\}$, we have

$$\tau_S(x_S) = \mathbb{E}[\tau(X) \mid X_S = x_S] = 0, \tag{16}$$

and therefore

$$b_S(x_S) = \delta_S(x_S). \tag{17}$$

First, consider the empty adjustment set. The treated mean is

$$\mathbb{E}[Y \mid A = 1] = \frac{\sum_{x_1, x_2} \Pr(X_1 = x_1, X_2 = x_2)\pi_{x_1 x_2} m(x_1, x_2)}{\sum_{x_1, x_2} \Pr(X_1 = x_1, X_2 = x_2)\pi_{x_1 x_2}} = \frac{-49/120}{49/120} = -1. \tag{18}$$

The untreated mean is

$$\mathbb{E}[Y \mid A = 0] = \frac{\sum_{x_1, x_2} \Pr(X_1 = x_1, X_2 = x_2)(1 - \pi_{x_1 x_2}) m(x_1, x_2)}{\sum_{x_1, x_2} \Pr(X_1 = x_1, X_2 = x_2)(1 - \pi_{x_1 x_2})} = \frac{-71/120}{71/120} = -1. \tag{19}$$

Thus,

$$b_\emptyset = \delta_\emptyset = \mathbb{E}[Y \mid A = 1] - \mathbb{E}[Y \mid A = 0] = 0. \tag{20}$$

Now consider the adjustment for $X_1$ only. For $X_1 = 0$, we arrive at

$$\begin{aligned} b_{\{1\}}(0) &= \frac{\pi_{00} m_{00} + \pi_{01} m_{01}}{\pi_{00} + \pi_{01}} - \frac{(1 - \pi_{00}) m_{00} + (1 - \pi_{01}) m_{01}}{(1 - \pi_{00}) + (1 - \pi_{01})} \\ &= -\frac{27}{38} + \frac{63}{82} = \frac{45}{779} > 0. \end{aligned} \tag{21}$$

For $X_1 = 1$, we yield

$$\begin{aligned} b_{\{1\}}(1) &= \frac{\pi_{10} m_{10} + \pi_{11} m_{11}}{\pi_{10} + \pi_{11}} - \frac{(1 - \pi_{10}) m_{10} + (1 - \pi_{11}) m_{11}}{(1 - \pi_{10}) + (1 - \pi_{11})} \\ &= -\frac{71}{60} + \frac{79}{60} = \frac{2}{15} > 0. \end{aligned} \tag{22}$$

Hence, the population-level partially adjusted bias is also positive:

$$\mathbb{E}[b_{\{1\}}(X_1)] = \frac{1}{2}\left(\frac{45}{779} + \frac{2}{15}\right) = \frac{2233}{23370} > 0. \tag{23}$$

Similarly, for adjustment for $X_2$ only, we obtain

$$\begin{aligned} b_{\{2\}}(0) &= \frac{\pi_{00} m_{00} + \pi_{10} m_{10}}{\pi_{00} + \pi_{10}} - \frac{(1 - \pi_{00}) m_{00} + (1 - \pi_{10}) m_{10}}{(1 - \pi_{00}) + (1 - \pi_{10})} \\ &= -\frac{4}{13} + \frac{36}{47} = \frac{280}{611} > 0, \end{aligned} \tag{24}$$

and

$$\begin{aligned} b_{\{2\}}(1) &= \frac{\pi_{01}m_{01} + \pi_{11}m_{11}}{\pi_{01} + \pi_{11}} - \frac{(1-\pi_{01})m_{01} + (1-\pi_{11})m_{11}}{(1-\pi_{01}) + (1-\pi_{11})} \\ &= -\frac{5}{4} + \frac{35}{24} = \frac{5}{24} > 0. \end{aligned} \tag{25}$$

Thus, we have

$$\mathbb{E}[b_{\{2\}}(X_2)] = \frac{1}{2}\left(\frac{280}{611} + \frac{5}{24}\right) = \frac{9775}{29328} > 0. \tag{26}$$

Finally, under full adjustment, it follows that

$$\begin{aligned} \delta_{\{1,2\}}(x_1, x_2) &= \mathbb{E}[Y \mid A = 1, X_1 = x_1, X_2 = x_2] - \mathbb{E}[Y \mid A = 0, X_1 = x_1, X_2 = x_2] \\ &= m(x_1, x_2) - m(x_1, x_2) = 0. \end{aligned} \tag{27}$$

Since $\tau_{\{1,2\}}(x_1, x_2) = 0$, we have

$$b_{\{1,2\}}(x_1, x_2) = 0 \qquad \text{for all } (x_1, x_2). \tag{28}$$

Therefore, this data-generating process satisfies

$$b_\emptyset = 0, \qquad b_{\{1,2\}}(x_1, x_2) = 0, \tag{29}$$

but

$$b_{\{1\}}(0) > 0, \quad b_{\{1\}}(1) > 0, \qquad b_{\{2\}}(0) > 0, \quad b_{\{2\}}(1) > 0. \tag{30}$$

Thus, the unadjusted residual bias is zero because of cancellation in the crude comparison, even though conditioning on either single covariate reveals positive residual confounding bias.

# F Implementation details

ConfoundingSHAP is implemented in Python using `numpy` and `pandas` for data handling, `TabPFN` (Grinsztajn et al., 2026; Hollmann et al., 2023; 2025) for regression nuisance estimation, and `shapiq` (Muschalik et al., 2024a) for exact and approximate Shapley-value computation. Experiment settings are stored in configuration files, including dataset parameters, random seeds, coalition budgets, approximation methods, and plotting options.

For each dataset, we first estimate the full-adjustment outcome model with a TabPFN-based S-learner on the augmented design matrix $(X, A)$. This yields plug-in pseudo-outcomes

$$\hat{\tau}_i = \hat{\mu}_1(X_i) - \hat{\mu}_0(X_i). \tag{31}$$

For each coalition $S$, we then estimate two coalition-specific quantities: the observational treated-control contrast $\hat{\delta}_S$ and the projected CATE $\hat{g}_S$. The signed coalition value passed to the Shapley routine is

$$\hat{\nu}(S) = -\frac{1}{n}\sum_{i=1}^{n}\left[\hat{\delta}_S(X_{i,S}) - \hat{g}_S(X_{i,S})\right]. \tag{32}$$

All main experiments use this signed value function. Absolute and squared value modes are implemented as alternatives but are not used for the reported main results.

Low-dimensional synthetic experiments use exact Shapley computation by enumerating all $2^p$ coalitions. Higher-dimensional experiments use budgeted approximations through `shapiq`, primarily RegressionMSR, with coalition budgets specified per experiment. We use first-order Shapley values for all main attributions. Local Shapley values are computed only for exact runs, where all coalition values are available.

All nuisance regressions use `TabPFNRegressor`. To reduce the cost of repeated coalition evaluation, most synthetic and ACTG experiments use a single TabPFN estimator per regression call. We disable the regression target power transform via `REGRESSION_Y_PREPROCESS_TRANSFORMS=(None,)` to avoid numerical overflow during inverse target transformation. Other TabPFN settings are left at their package defaults unless specified in the experiment configuration.

In the TabPFN configuration used for our experiments, the default pretraining-limit checks allow up to 10,000 samples and 500 input columns. We do not bypass these limits or truncate observations or covariates. The full-adjustment S-learner uses the complete augmented matrix $(X, A)$, while coalition-specific regressions use $(X_S, A)$ or $X_S$. All reported experiments remain within these limits, including SUPPORT/RHC with 5,735 observations and synthetic experiments with up to 200 covariates. These limits are specific to the nuisance estimator used in our implementation rather than the ConfoundingSHAP framework. Larger datasets can be accommodated using a compatible TabPFN version.

We use the `TabPFNRegressor.fit` routine for in-context learning, which does not retrain TabPFN's pretrained weights, but sets the labeled in-context reference set used for prediction. We use TabPFN as a plug-in nuisance estimator as follows: for each coalition, the corresponding design matrix defines the context, and the resulting nuisance predictions are averaged to form the empirical coalition value. A cross-fitted variant could instead build the TabPFN context on training folds and evaluate coalition values on held-out folds; this changes only the finite-sample nuisance-estimation scheme, not the ConfoundingSHAP game itself.

All reported runs were executed on a GPU cluster. The hardware consisted of two NVIDIA H200 NVL GPUs with approximately 144GB of memory each, NVIDIA driver version 580.95.05, and CUDA version 13.0. Runtime measurements in Section I report end-to-end wall-clock times under this cluster environment, including coalition-value evaluation and Shapley reconstruction.

## G Ablation studies

This section evaluates the budgeted Shapley approximators used to scale ConfoundingSHAP beyond exact enumeration. Since the synthetic benchmarks provide ground-truth covariate roles, we can directly assess whether approximation methods concentrate attribution on the true confounders. We report two metrics. First, the *confounder Shapley mass* measures the fraction of absolute attribution assigned to true confounders, i.e.,

$$M(\hat{\phi}; \mathcal{C}) := \frac{\sum_{j \in \mathcal{C}} |\hat{\phi}_j|}{\sum_{j=1}^{p} |\hat{\phi}_j|} \in [0, 1], \tag{33}$$

where $\mathcal{C} \subseteq [p]$ denotes the set of true confounders. Second, we report *confounder recovery*, defined as the fraction of true confounders recovered among the top-$|\mathcal{C}|$ variables ranked by $|\hat{\phi}_j|$. Let

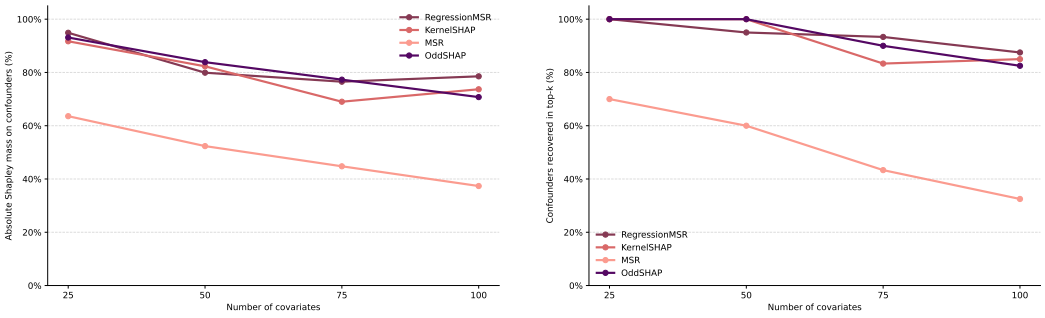
$$\mathrm{Top}_{|\mathcal{C}|}(\hat{\phi}) \tag{34}$$

denote the set of $|\mathcal{C}|$ covariates with largest absolute attribution $|\hat{\phi}_j|$. Confounder recovery is then given by

$$R(\hat{\phi}; \mathcal{C}) := \frac{\left|\mathcal{C} \cap \mathrm{Top}_{|\mathcal{C}|}(\hat{\phi})\right|}{|\mathcal{C}|} \in [0, 1]. \tag{35}$$

For example, if there are ten true confounders and nine of them appear among the top ten ranked covariates, then $R = 0.9$. Higher values indicate that the approximation preserves the confounder ranking induced by the exact Shapley game.

All four approximators recover the exact Shapley values in the limit $B \to 2^p$, but they differ substantially at finite coalition budgets. At a fixed budget of $B = 1024$, Fig. 9 shows that RegressionMSR and KernelSHAP assign substantially more attribution mass to true confounders than MSR, and the difference becomes more pronounced as the number of covariates increases. The same pattern is observed for confounder recovery: proxy- and kernel-based approximations preserve the confounder ranking more reliably at matched budget.



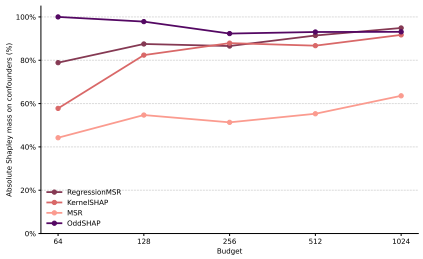

(a) Confounder Shapley mass.

(b) Confounder recovery.

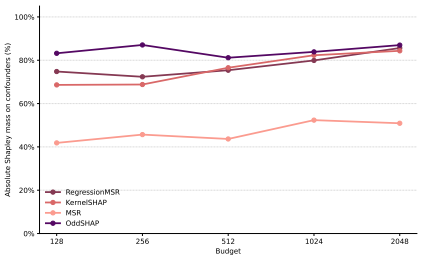
Figure 9: Approximation quality at fixed budget $B = 1024$ across increasing numbers of covariates.

We next vary the coalition budget for synthetic problems across 25, 50, 75, and 100 covariates, each containing 40% true confounders. Figures 10 and 11 show that increasing the budget improves both attribution mass and confounder recovery. This confirms the expected accuracy-compute tradeoff: larger budgets provide more information about the set function and yield more reliable rankings, especially in higher-dimensional settings.

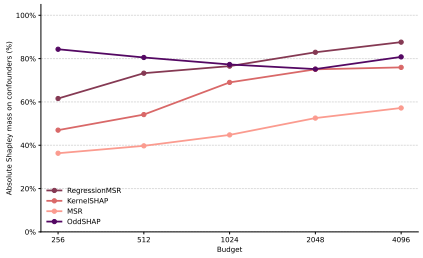
Finally, we consider a 200-covariate setting. Fig. 12 shows that approximation quality continues to improve with budget, but the absolute values remain lower than in smaller problems. This highlights the main computational limitation of the current implementation: very high-dimensional settings require large coalition budgets to obtain stable rankings, and these budgets can lead to long runtimes; see Section I.



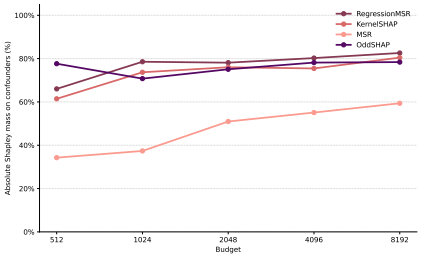

(a) 25 covariates




(b) 50 covariates

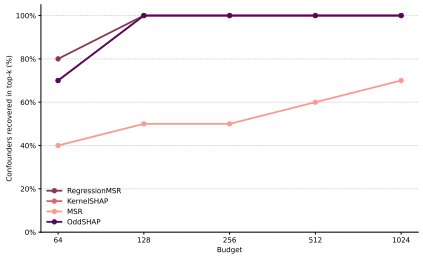



(c) 75 covariates

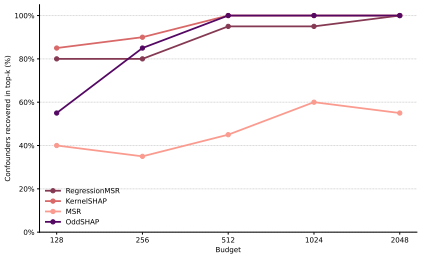



(d) 100 covariates

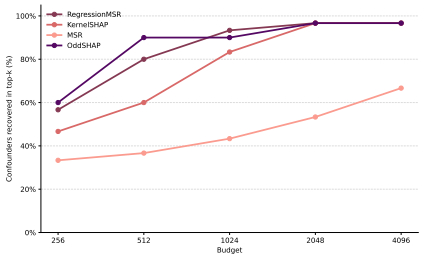
Figure 10: Absolute Shapley mass assigned to true confounders as a function of coalition budget. Each setting contains 40% confounders.



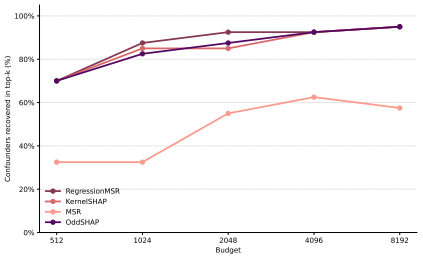

(a) 25 covariates




(b) 50 covariates

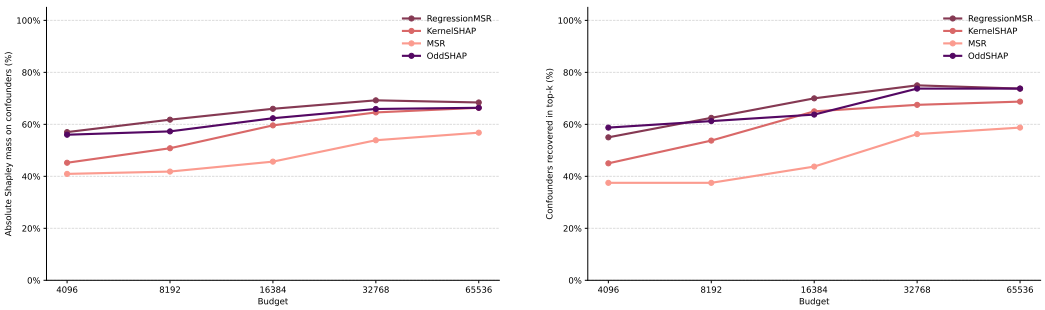



(c) 75 covariates




(d) 100 covariates

Figure 11: Fraction of true confounders recovered among the top 40% of covariates as a function of coalition budget.




(a) Confounder Shapley mass. (b) Confounder recovery.

Figure 12: Approximation quality for 200 covariates across increasing budgets.

# H Additional experiments

## H.1 Confounders can cancel when explaining a CATE estimator

We construct a synthetic observational dataset with $n = 5000$ independent observations and four covariates with distinct causal roles: an instrument $Z$, a confounder $C$, an effect modifier $M$, and an outcome-only prognostic variable $O$. All covariates are sampled independently as

$$Z_i, C_i, M_i, O_i \sim \mathcal{N}(0, 1). \tag{36}$$

Treatment assignment is observational and follows

$$\Pr(A_i = 1 \mid X_i) = \text{logit}^{-1}(3C_i + Z_i), \tag{37}$$

so both the confounder $C_i$ and the instrument $Z_i$ affect treatment uptake.

The potential outcome means share the same prognostic component

$$h(C_i, O_i) = 2C_i + O_i + 0.5O_i^2. \tag{38}$$

The control mean is

$$\mu_0(X_i) = h(C_i, O_i), \tag{39}$$

and the treated mean is

$$\mu_1(X_i) = h(C_i, O_i) + \tau(X_i), \tag{40}$$

where the treatment effect is

$$\tau(X_i) = 1 + 3M_i^2. \tag{41}$$

Observed potential outcomes are generated as

$$Y_i(a) = \mu_a(X_i) + \epsilon_i, \qquad \epsilon_i \sim \mathcal{N}(0, 0.5^2), \tag{42}$$

and the observed outcome is

$$Y_i = (1 - A_i)Y_i(0) + A_i Y_i(1). \tag{43}$$

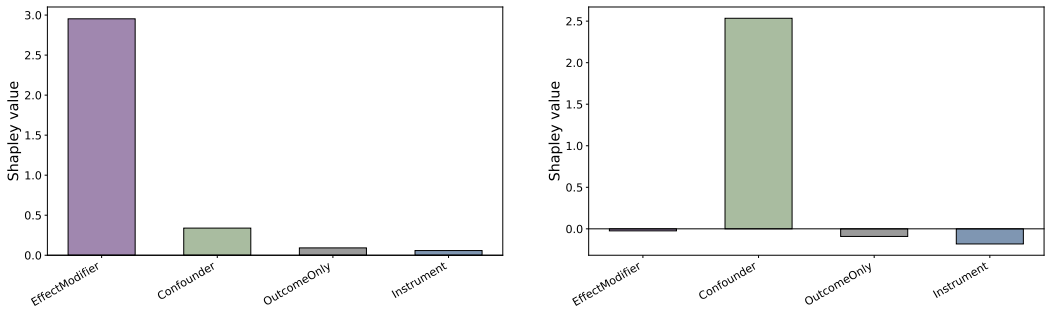
This construction makes $C_i$ a genuine confounder: it affects treatment assignment, and it also affects the outcome level through the shared prognostic term $h(C_i, O_i)$. However, $C_i$ does not modify the treatment effect, because it enters the treated and untreated potential outcome means identically. Consequently, in the conditional treatment contrast, the shared prognostic component cancels out:

$$\mu_1(X_i) - \mu_0(X_i) = \{h(C_i, O_i) + \tau(X_i)\} - h(C_i, O_i) = \tau(X_i) = 1 + 3M_i^2. \tag{44}$$

Thus, the true CATE depends only on the effect modifier $M_i$, not on the confounder $C_i$. The confounder is therefore essential for adjustment in the observational treatment-outcome relationship, but it is not a source of treatment effect heterogeneity.

For the CATE baseline, we fit a doubly robust learner (Kennedy, 2023) using `causalml.inference.meta.BaseDRRegressor` (Chen et al., 2020). The propensity model, treated and control outcome models, and final treatment effect model are implemented with gradient-boosted trees from `xgboost` (Chen & Guestrin, 2016). We then apply SHAP to the fitted CATE predictions $\hat{\tau}(x)$, so the resulting feature importance reflects variation in the estimated treatment effect function. This is the natural target for explaining CATE heterogeneity, but it is not the same estimand as attributing confounding bias.

The experiment is designed so that CATE-based SHAP and ConfoundingSHAP answer different questions. Since the true treatment effect is $\tau(X_i) = 1 + 3M_i^2$, SHAP on a CATE function should primarily attribute variation to the effect modifier $M_i$. The confounder $C_i$ can have little or no CATE-based SHAP importance because its direct outcome contribution is shared across both potential outcomes and cancels in the treatment contrast. In contrast, ConfoundingSHAP targets confounding in the observational treatment contrast. Since $C_i$ drives both treatment assignment and the outcome level, it receives the dominant ConfoundingSHAP attribution. This illustrates the central distinction: CATE-based SHAP explains heterogeneity in the fitted treatment effect, whereas ConfoundingSHAP attributes confounding bias induced by variables that jointly affect treatment assignment and outcomes.




(a) SHAP on a DR-learner

(b) ConfoundingSHAP

Figure 13: CATE explanation versus confounding bias attribution on a synthetic dataset with 4 covariates.

### H.2 ACIC 2016 data challenge

We also evaluate whether the ranking induced by ConfoundingSHAP is meaningful for downstream CATE estimation. For this purpose, we use the ACIC 2016 data challenge benchmark (Dorie et al., 2019), where covariates are real but treatment assignments and potential outcomes are simulated, making ground-truth treatment effects available. This allows us to assess whether covariates ranked highly by ConfoundingSHAP are important for estimating individual treatment effects.

We rank covariates by their absolute ConfoundingSHAP values and compare three feature-removal strategies. First, we remove the top-$k$ ranked covariates. Second, we remove $k$ randomly selected covariates. Third, we remove the lowest-ranked $k$ covariates. After removal, we refit the same CATE learner used in Section H.1 and evaluate the precision in estimation of heterogeneous effects (PEHE),

$$\text{PEHE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\hat{\tau}(x_i) - \tau(x_i)\right)^2}, \tag{45}$$

where $\hat{\tau}(x_i)$ is the estimated CATE and $\tau(x_i)$ is the known ground-truth treatment effect in the ACIC benchmark.

If ConfoundingSHAP identifies covariates important for downstream treatment effect estimation, removing the top-ranked variables should lead to a larger increase in PEHE than removing randomly selected or low-ranked variables. Fig. 14 shows this pattern: dropping the top-ranked covariates substantially increases PEHE, whereas removing low-ranked variables has little effect. Since PEHE also captures errors in treatment effect heterogeneity, we interpret this experiment as an additional downstream validation rather than a direct measure of confounding bias.

The magnitude of the PEHE increase also depends on the learner used after feature removal, since different CATE estimators may be more or less robust to omitted adjustment variables. Here, the purpose is to test whether the ConfoundingSHAP ranking identifies variables whose removal harms downstream treatment effect estimation under a fixed learner.

We use the same DR-learner setup as in Section H.1, with `xgboost` models for the propensity, outcome, and final treatment effect stages. Propensity scores are clipped before fitting, and PEHE is evaluated on a held-out test set using the ACIC ground-truth treatment effects.

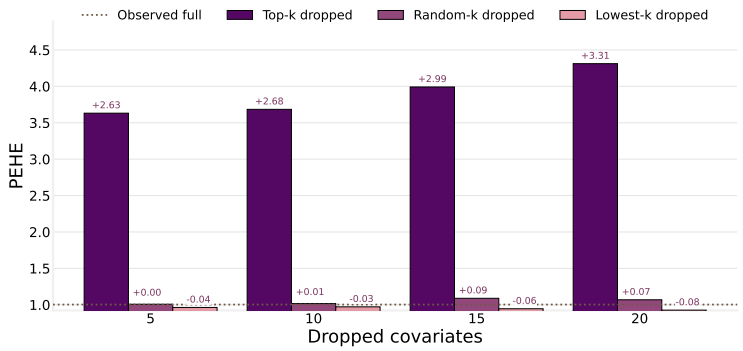
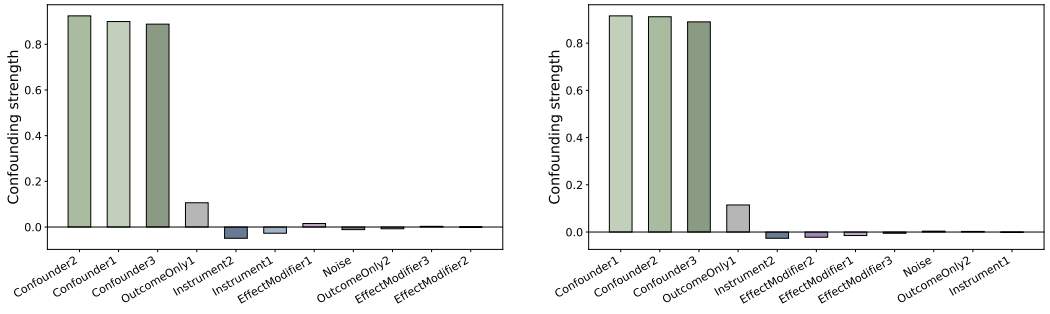



Figure 14: PEHE after removing top-ranked, random, or lowest-ranked covariates on the ACIC 2016 benchmark.

### H.3 Additional synthetic results

We further compare exact and approximate computation in the 11-covariate synthetic setting. Fig. 15 shows that the budgeted approximation preserves the main qualitative ranking: true confounders receive the largest attributions, while non-confounding variables remain small. Fig. 16 reports the corresponding local explanations, showing that confounder contributions are visible at the individual level as well. Fig. 17 provides an additional 17-covariate result, again showing that attribution is concentrated on the true confounders.




(a) Exact computation

(b) RegressionMSR approximation

Figure 15: Global Shapley values for synthetic datasets with 11 covariates: exact versus approximate computation.

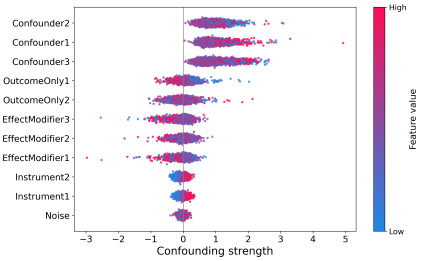



Figure 16: Local Shapley values for the 11-covariate synthetic dataset with exact computation.

### H.4 Stability over approximation seeds

We examine whether the approximate rankings are stable across repeated runs. The rank-stability plot in Fig. 18 records how often each covariate appears at each rank across 10 runs, while the bar plot reports the average absolute attribution. The true confounders are consistently among the highest

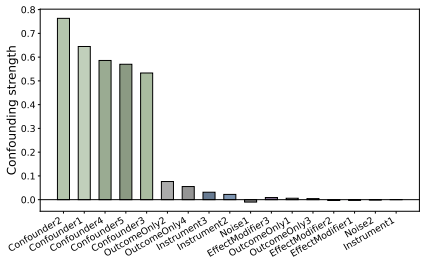
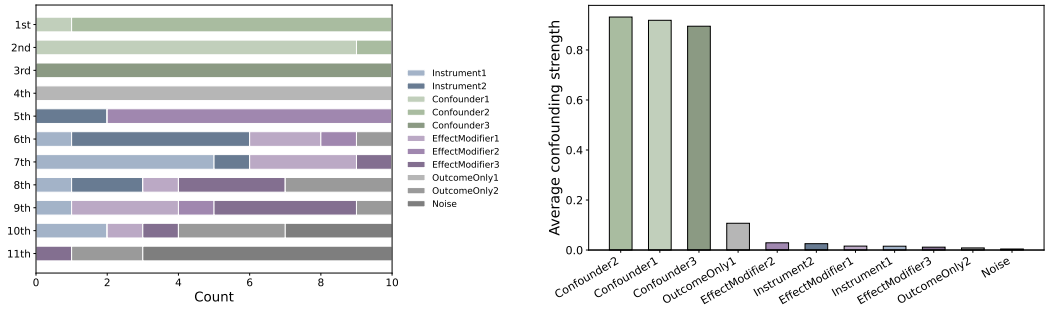



Figure 17: Global Shapley values for the 17-covariate synthetic dataset with RegressionMSR approximation.

ranks, indicating that the approximation does not merely assign importance to arbitrary correlated features.



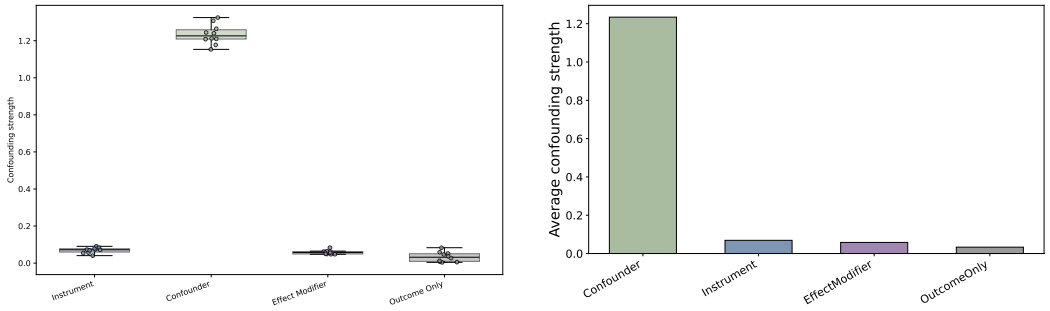

(a) Rank stability

(b) Average attribution

Figure 18: Stability of approximate ConfoundingSHAP values over 10 runs in the 11-covariate synthetic setting.

### H.5 Stability over repeated synthetic datasets

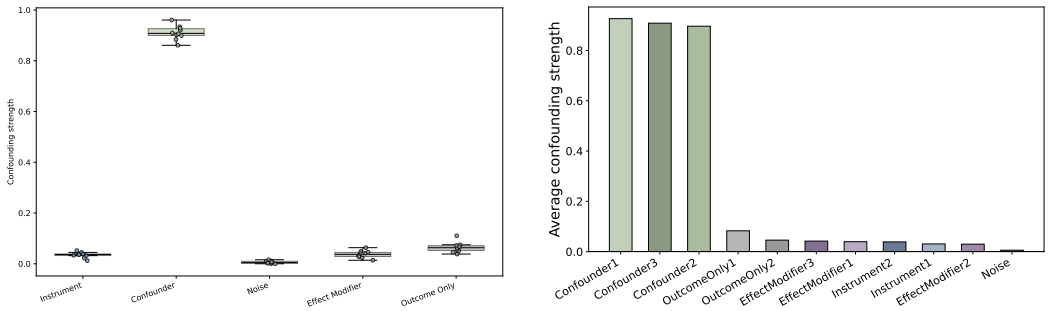
We additionally test stability across independently generated synthetic datasets. This evaluates whether the method recovers the intended covariate roles across different samples from the same data-generating process, rather than relying on a favorable random draw. Figures 19, 20, and 21 show that true confounders remain the dominant contributors across repeated datasets for both exact and approximate computation.




(a) Role scores over datasets

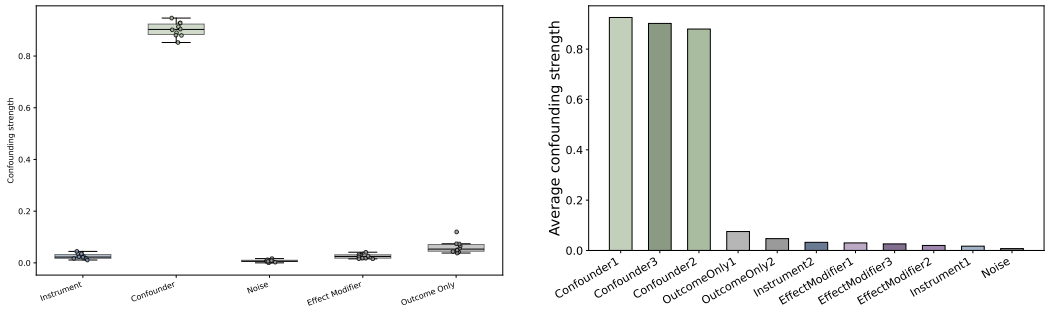
(b) Average attribution

Figure 19: Dataset-level stability for the four-covariate synthetic setting with exact computation.




(a) Role scores over datasets (b) Average attribution

Figure 20: Dataset-level stability for the 11-covariate synthetic setting with exact computation.




(a) Role scores over datasets (b) Average attribution

Figure 21: Dataset-level stability for the 11-covariate synthetic setting with RegressionMSR approximation.

# I Runtimes

Table 2 reports wall-clock runtimes for the main experiments and ablations. The timings reflect end-to-end coalition evaluation and Shapley reconstruction under the reported budgets. They should be interpreted together with the budget ablations in Section G: larger coalition budgets improve confounder recovery but can become expensive in high-dimensional settings.

Table 2: Reported runtimes

| Dataset | Method | Covariates | Budget | Samples | Runtime in s | Runtime in h:m:s |
|---|---|---|---|---|---|---|
| Synthetic | Exact | 4 | – | 5000 | 67.07 | 0h 1m 7.07s |
| Synthetic | Exact | 11 | – | 5000 | 7750.94 | 2h 9m 10.94s |
| Synthetic | RegressionMSR[†] | 11 | 128 | 5000 | 326.57 | 0h 5m 26.57s |
| Synthetic | RegressionMSR | 17 | 128 | 5000 | 359.88 | 0h 5m 59.88s |
| Synthetic | RegressionMSR | 25 | 64 | 5000 | 176.75 | 0h 2m 56.75s |
| Synthetic | OddSHAP | 25 | 64 | 5000 | 148.81 | 0h 2m 28.81s |
| Synthetic | RegressionMSR | 25 | 128 | 5000 | 307.04 | 0h 5m 7.04s |
| Synthetic | OddSHAP | 25 | 128 | 5000 | 282.70 | 0h 4m 42.70s |
| Synthetic | RegressionMSR | 25 | 256 | 5000 | 586.84 | 0h 9m 46.84s |
| Synthetic | OddSHAP | 25 | 256 | 5000 | 540.06 | 0h 9m 0.06s |
| Synthetic | RegressionMSR | 25 | 512 | 5000 | 1118.70 | 0h 18m 38.70s |
| Synthetic | OddSHAP | 25 | 512 | 5000 | 1080.38 | 0h 18m 0.38s |
| Synthetic | RegressionMSR | 25 | 1024 | 5000 | 2168.24 | 0h 36m 8.24s |
| Synthetic | OddSHAP | 25 | 1024 | 5000 | 2127.84 | 0h 35m 27.84s |
| Synthetic | RegressionMSR | 50 | 128 | 5000 | 295.21 | 0h 4m 55.21s |
| Synthetic | OddSHAP | 50 | 128 | 5000 | 292.19 | 0h 4m 52.19s |
| Synthetic | RegressionMSR | 50 | 256 | 5000 | 606.69 | 0h 10m 6.69s |
| Synthetic | OddSHAP | 50 | 256 | 5000 | 570.75 | 0h 9m 30.75s |
| Synthetic | RegressionMSR | 50 | 512 | 5000 | 1154.41 | 0h 19m 14.41s |
| Synthetic | OddSHAP | 50 | 512 | 5000 | 1115.66 | 0h 18m 35.66s |
| Synthetic | RegressionMSR | 50 | 1024 | 5000 | 2248.64 | 0h 37m 28.64s |
| Synthetic | OddSHAP | 50 | 1024 | 5000 | 2228.51 | 0h 37m 8.51s |
| Synthetic | RegressionMSR | 50 | 2048 | 5000 | 4452.17 | 1h 14m 12.17s |
| Synthetic | OddSHAP | 50 | 2048 | 5000 | 4423.25 | 1h 13m 43.25s |
| Synthetic | RegressionMSR | 75 | 256 | 5000 | 615.13 | 0h 10m 15.13s |
| Synthetic | OddSHAP | 75 | 256 | 5000 | 596.11 | 0h 9m 56.11s |
| Synthetic | RegressionMSR | 75 | 512 | 5000 | 1206.70 | 0h 20m 6.70s |
| Synthetic | OddSHAP | 75 | 512 | 5000 | 1161.65 | 0h 19m 21.65s |
| Synthetic | RegressionMSR | 75 | 1024 | 5000 | 2329.67 | 0h 38m 49.67s |
| Synthetic | OddSHAP | 75 | 1024 | 5000 | 2312.64 | 0h 38m 32.64s |
| Synthetic | RegressionMSR | 75 | 2048 | 5000 | 4594.82 | 1h 16m 34.82s |
| Synthetic | OddSHAP | 75 | 2048 | 5000 | 4602.55 | 1h 16m 42.55s |
| Synthetic | RegressionMSR | 75 | 4096 | 5000 | 9203.27 | 2h 33m 23.27s |
| Synthetic | OddSHAP | 75 | 4096 | 5000 | 9212.50 | 2h 33m 32.50s |
| Synthetic | RegressionMSR | 100 | 512 | 5000 | 1244.95 | 0h 20m 44.95s |
| Synthetic | OddSHAP | 100 | 512 | 5000 | 1188.13 | 0h 19m 48.13s |
| Synthetic | RegressionMSR | 100 | 1024 | 5000 | 2419.14 | 0h 40m 19.14s |
| Synthetic | OddSHAP | 100 | 1024 | 5000 | 2409.28 | 0h 40m 9.28s |
| Synthetic | RegressionMSR | 100 | 2048 | 5000 | 4815.07 | 1h 20m 15.07s |
| Synthetic | OddSHAP | 100 | 2048 | 5000 | 4847.47 | 1h 20m 47.47s |
| Synthetic | RegressionMSR | 100 | 4096 | 5000 | 9559.37 | 2h 39m 19.37s |
| Synthetic | OddSHAP | 100 | 4096 | 5000 | 9561.90 | 2h 39m 21.90s |
| Synthetic | RegressionMSR | 100 | 8192 | 5000 | 19065.38 | 5h 17m 45.38s |
| Synthetic | OddSHAP | 100 | 8192 | 5000 | 19216.52 | 5h 20m 16.52s |
| Synthetic | RegressionMSR | 200 | 4096 | 5000 | 12693.24 | 3h 31m 33.24s |
| Synthetic | OddSHAP | 200 | 4096 | 5000 | 13084.93 | 3h 38m 4.93s |
| Synthetic | RegressionMSR | 200 | 8192 | 5000 | 25235.86 | 7h 0m 35.86s |
| Synthetic | OddSHAP | 200 | 8192 | 5000 | 23375.14 | 6h 29m 35.14s |
| Synthetic | RegressionMSR | 200 | 16384 | 5000 | 48402.61 | 13h 26m 42.61s |
| Synthetic | OddSHAP | 200 | 16384 | 5000 | 47985.27 | 13h 19m 45.27s |
| Synthetic | RegressionMSR | 200 | 32768 | 5000 | 94600.79 | 26h 16m 40.79s |
| Synthetic | OddSHAP | 200 | 32768 | 5000 | 91864.63 | 25h 31m 4.63s |
| Synthetic | RegressionMSR | 200 | 65536 | 5000 | 185475.45 | 51h 31m 15.45s |
| Synthetic | OddSHAP | 200 | 65536 | 5000 | 122975.39 | 34h 9m 35.39s |
| ACTG RCT | Exact | 11 | – | 1054 | 6331.66 | 1h 45m 31.66s |
| ACTG semisynth | Exact | 11 | – | 1054 | 7132.27 | 1h 58m 52.27s |
| SUPPORT | RegressionMSR | 10 | 256 | 5735 | 594.06 | 0h 9m 54.06s |

[†] MSR and KernelSHAP reused the precomputed coalition values from RegressionMSR and are therefore not reported separately.

## J DATASETS

We evaluate CONFOUNDINGSHAP using synthetic, semi-synthetic, and real-world datasets. The synthetic and semi-synthetic experiments are designed so that the confounding structure is known by construction. The SUPPORT/RHC benchmark is a less controlled observational stress test in which the exact covariate-level confounding ground truth is unavailable.

**Synthetic datasets.** Our controlled experiments use a synthetic data-generating process based on the simulation design of (Curth & van der Schaar, 2021). Covariates are drawn independently from a standard normal distribution and partitioned into interpretable roles: instruments, confounders, effect modifiers, outcome-only variables, and noise variables. We use several benchmark dimensions with known role composition. The 4-covariate setting contains one instrument, one confounder, one effect modifier, and one outcome-only variable. The 11-covariate setting contains two instruments, three confounders, three effect modifiers, two outcome-only variables, and one noise variable. The 17-covariate setting contains three instruments, five confounders, three effect modifiers, four outcome-only variables, and two noise variables. In the high-dimensional 100-covariate setting, we use 15 instruments, 40 confounders, 15 effect modifiers, 15 outcome-only variables, and 15 noise variables.

Let

$$X = (X_Z, X_C, X_\tau, X_O, X_N), \qquad X_j \sim \mathcal{N}(0,1), \tag{46}$$

where $X_Z$ denotes instruments, $X_C$ confounders, $X_\tau$ effect modifiers, $X_O$ outcome-only variables, and $X_N$ noise variables. The untreated outcome surface follows the sum-of-squares structure in (Curth & van der Schaar, 2021), i.e.,

$$\mu_0(X) = \sum_{j \in C \cup O} X_j^2, \tag{47}$$

and the treated outcome surface is

$$\mu_1(X) = \mu_0(X) + \sum_{j \in \tau} X_j^2. \tag{48}$$

Thus, the conditional average treatment effect is

$$\tau(X) = \mu_1(X) - \mu_0(X) = \sum_{j \in \tau} X_j^2, \tag{49}$$

so only the effect-modifier block changes the individual treatment effect.

Treatment assignment is generated from a logistic propensity score. We define the confounding score

$$m_C(X) = \frac{1}{|C|} \sum_{j \in C} X_j^2, \qquad \omega = Q_{0.5}\{m_C(X)\}, \tag{50}$$

where $\omega$ is the sample median of $m_C(X)$. To include variables that affect treatment but not outcome, we add an explicit instrument term

$$z(X) = \frac{1}{|Z|} \sum_{j \in Z} X_j. \tag{51}$$

The propensity score is

$$e(X) = \mathbb{P}(A = 1 \mid X) = \text{logit}^{-1}\left(\xi\{m_C(X) - \omega\} + \gamma_Z z(X)\right), \tag{52}$$

with $\xi = 3$ and $\gamma_Z = 1$ in the main experiments. The treatment is sampled via

$$A \sim \text{Bernoulli}(e(X)). \tag{53}$$

Finally, with $\varepsilon \sim \mathcal{N}(0, \sigma^2)$, potential and factual outcomes are generated as

$$Y(0) = \mu_0(X) + \varepsilon, \qquad Y(1) = \mu_1(X) + \varepsilon, \qquad Y = AY(1) + (1 - A)Y(0). \tag{54}$$

This construction separates variables that predict treatment only, variables that truly confound treatment effect estimation, variables that modify treatment effects, variables that predict outcomes only, and irrelevant noise variables.

**ACTG 175 semi-synthetic dataset.** We construct a semi-synthetic benchmark from ACTG 175, a randomized clinical trial comparing nucleoside monotherapy and combination therapy in HIV-infected adults with baseline CD4 counts between 200 and 500 cells per cubic millimeter (Hammer et al., 1996). We restrict the data to the zidovudine-only control arm and the zidovudine-plus-didanosine treatment arm. The retained baseline covariates are age, weight, hemophilia status, history of intravenous drug use, Karnofsky score, race, gender, symptomatic status, prior antiretroviral experience, baseline CD4 count, and baseline CD8 count.

To obtain a known confounding structure while preserving realistic covariate distributions, we replace the randomized treatment assignment with a semi-synthetic mechanism. We designate clinically plausible variables as confounders, namely, baseline severity and treatment-history variables: age, Karnofsky score, symptomatic status, prior antiretroviral experience, and baseline CD4 count. These variables determine both the synthetic propensity score and the synthetic potential outcome model. The remaining variables are retained in the feature matrix but excluded from the assignment and outcome mechanisms. They therefore act as non-confounding covariates in the semi-synthetic benchmark. Continuous covariates are standardized before generating treatment and outcomes.

Let $X_C$ denote the designated confounder block and $X_N$ the remaining retained baseline covariates. The treatment assignment is generated by

$$e(X_C) = \mathbb{P}(A = 1 \mid X_C) = \text{logit}^{-1}(\alpha_0 + X_C^\top \alpha), \qquad A \sim \text{Bernoulli}(e(X_C)). \tag{55}$$

In the continuous outcome experiments, potential outcomes are generated by

$$\mu_0(X_C) = \beta_0 + X_C^\top \beta, \qquad \mu_1(X_C) = \mu_0(X_C) + \tau, \tag{56}$$

and

$$Y(0) = \mu_0(X_C) + \varepsilon, \qquad Y(1) = \mu_1(X_C) + \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, \sigma^2), \tag{57}$$

with factual outcome

$$Y = AY(1) + (1 - A)Y(0). \tag{58}$$

The non-confounding ACTG covariates $X_N$ remain in the observed feature matrix but do not enter $e(X_C)$ or $\mu_a(X_C)$. This produces realistic empirical covariate correlations while keeping the true confounding roles known by construction.

In the reported ACTG 175 semi-synthetic continuous outcome experiment, we use the following parameter values:

$$\alpha_0 = -0.2, \qquad \alpha = (0.2, -0.8, 0.9, 0.5, -0.9), \tag{59}$$

$$\beta_0 = 0, \qquad \beta = (0.1, -0.5, 0.5, 0.25, -0.55), \qquad \tau = 0.45, \qquad \sigma = 0.35. \tag{60}$$

The coefficients are ordered according to the confounder block (age, Karnofsky score, symptomatic status, prior antiretroviral experience, baseline CD4 count). We use random seed 42 and retain all available observations after filtering to the two treatment arms and dropping missing values. Both the real data and the semi-synthetic data contain $n = 1054$ observations.

**ACIC 2016 benchmark.** For the ACIC experiment, we use the ACIC Challenge 2016 data distributed with the `causallib` repository Shimoni et al. (2019) (`BiomedSciAI/causallib`, path `causallib/datasets/data/acic_challenge_2016`). We use `x.csv` for the covariates and the first simulation instance, `zymu_1.csv`, for treatment assignments, factual outcomes, and potential-outcome means. The files are cached locally under `data/acic_challenge_2016`. Categorical variables are one-hot encoded and missing numeric values are median-imputed, yielding $n = 4802$ observations and 82 encoded covariates. The reported run uses random seed 42, a RegressionMSR coalition budget of 2048, PEHE test fraction 0.3, and feature-drop sizes $\{5, 10, 15, 20\}$.

**SUPPORT/RHC observational dataset.** The SUPPORT/RHC benchmark is derived from the observational study of right-heart catheterization in critically ill patients (Connors et al., 1996). Here, treatment corresponds to receiving a right-heart catheterization during the first 24 hours of ICU care, while the outcome is 30-day mortality. We use the following clinical variables as confounders: acute physiology score, mean blood pressure, and serum creatinine. Additional demographic, socioeconomic, and administrative variables are retained (with only small confounding strength). Unlike the synthetic and ACTG semi-synthetic benchmarks, SUPPORT/RHC does not provide exact feature-role ground-truth. We therefore use it as a qualitative observational benchmark to assess whether ConfoundingSHAP assigns low attribution to background descriptors and higher attribution to clinically plausible adjustment variables.

## K Computational complexity

We measure computational cost in terms of coalition-value evaluations. A coalition-value evaluation estimates the nuisance quantities needed to compute $\hat{\nu}(S)$ for one coalition $S$. With TabPFN, this is implemented through a constant number of in-context conditioning steps on a frozen pretrained model, rather than through gradient-based refitting. We therefore omit constant factors and focus on the number of queried coalitions.

**Exact computation.** Exact Shapley computation evaluates every coalition, i.e.,

$$C_{\text{exact}} = 1 + 2^p = \mathcal{O}(2^p), \tag{61}$$

where the additive constant corresponds to the full-covariate outcome model used to form pseudo-outcomes.

**Approximate computation.** With coalition budget $B$, only $B$ coalitions are queried, i.e.,

$$C_{\text{approx}} = 1 + B = \mathcal{O}(B). \tag{62}$$

This expression counts coalition evaluations; the actual wall-clock time also depends on sample size, covariate dimension, hardware, and the chosen Shapley reconstruction method. Empirical runtimes in Section I show the practical cost of this approximation for the datasets considered in our experiments.

## L Limitations

ConfoundingSHAP attributes observed confounding bias relative to the available full covariate set. The interpretation is therefore conditional on the measured variables: if important confounders are unobserved, the attributions should be understood relative to the observed reference set rather than as ground-truth causal bias. As with other plug-in causal estimation procedures, attribution quality can also be affected by limited overlap, finite-sample variability, and nuisance-estimation error. In this sense, ConfoundingSHAP is complementary to sensitivity analysis, causal diagrams, domain knowledge, and robustness checks.

A second limitation is computational. Although TabPFN-based coalition evaluation avoids gradient-based refitting, reliable approximation in high-dimensional settings still requires sufficiently large coalition budgets. The 200-covariate ablations show that attribution quality improves with increasing budget, but such runs can be computationally expensive in practice.

## M Broader impacts

ConfoundingSHAP is intended to improve transparency in causal machine learning by showing which measured covariates contribute most to observed confounding. This is particularly relevant in clinical and other observational studies, where treatment groups often differ systematically in baseline characteristics and where transparent adjustment decisions are important. By identifying measured variables that drive treatment–control incomparability, ConfoundingSHAP may support the construction of minimal or parsimonious adjustment sets, improve communication between methodological and domain experts, and inform the design of future observational studies or trials by highlighting variables that should be measured, stratified on, or checked for balance. The main risk is overinterpretation (Martens et al., 2026): ConfoundingSHAP values should not be read as automatically certifying that an adjustment set is sufficient or that a causal estimate is unbiased. Rather, the values provide diagnostic evidence about measured confounding and should be interpreted together with domain knowledge, causal reasoning, sensitivity analyses, and robustness checks.