Leveraging a Simulator for Learning Causal Representations for CATE from Post-Treatment Covariates

Lokesh Nagalapatti[*] **Pranava Singhal**[*] **Avishek Ghosh** **Sunita Sarawagi**
IIT Bombay

Abstract

Treatment effect estimation involves assessing the impact of different treatments on individual outcomes. Current methods rely on observational datasets where covariates are gathered before treatments and outcomes are observed afterward. However, real-world scenarios often deviate from this protocol, leading to both covariate and outcome observed post-treatment. We first establish that this deviation renders treatment effects unidentifiable, necessitating additional assumptions for estimation. We propose *SimPONet*, which unlike prior methods that assume counterfactual supervision in the training datasets, leverages a simulator that generates related synthetic counterfactual data. This allows extraction of causal representations from post-treatment covariates that aid in identifying treatment effects. The accuracy of such estimates hinges on the quality of the simulator, and we conduct theoretical analyses to establish generalization bounds that assess the CATE error based on the distributional discrepancies between real and synthetic data. In a linear setting, we analytically derive the CATE error, demonstrating the limitations of several baseline methods. Our empirical validation on synthetic and semi-synthetic real world datasets further reinforces *SimPONet*'s effectiveness in precise treatment effect estimation from post-treatment data.

1 Introduction

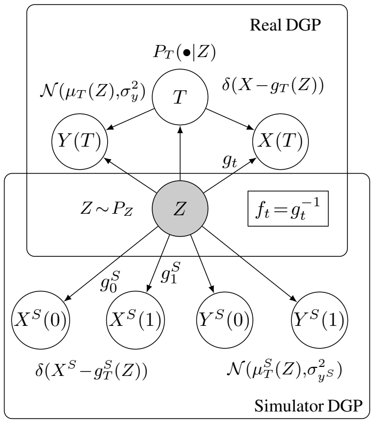

Figure 1: The Data Generating process for Real and Simulator.

Many applications require estimating the difference in outcome as a result of a change in treatment. The gold standard to estimating such effects is Randomized Control Trials, which are often expensive, and with the easy availability of observational data, there is extensive interest in harnessing them for deriving these estimates. The first step in estimating treatment effects from observational datasets is to determine the set of covariates that, when conditioned upon, make these effects identifiable. Prior works [31, 40, 12, 53, 39, 52, 7, 11, 63, 54, 55], assume that such covariates are observed, and gathered prior to treatment, with outcomes being observed afterward. However, collecting such datasets is challenging as it requires tracking the same individuals over two distinct time points. In contrast, readily available observational datasets often consist of both covariates and outcomes recorded post-treatment. For instance, in economics [2, 1], the effectiveness of policies is frequently evaluated using post-policy data for both outcomes and covariates. Similarly, in voluntary healthcare surveys, only post-treatment data about patients might be accessible. In medical imaging, an image taken under a specific instrument setting (treatment) may be evaluated to determine whether switching to a different setting would improve a subsequent diagnosis (outcome). We model our scenario using a Data Generating Process (DGP) as illustrated in the top panel of Figure 1, marked Real DGP. The figure shows **latent** variables Z that influence the observed variables—treatment

[*] Lokesh and Pranava contributed equally. Corresponding Author: Lokesh N <nlokeshiisc@gmail.com>

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

T, outcome Y, and covariates X. We begin this paper by showing that the latent nature of Z impedes the identifiability of treatment effects.

Lemma 1. *The causal effect of T on Y cannot be identified given i.i.d. samples from the real DGP.*

proof. Since X is a collider, conditioning on it opens the backdoor path $T \rightarrow X \leftarrow Z \rightarrow Y$. Since Z is not observed, the treatment effects cannot be identified from X,T, and Y alone. Thus, our only hope lies in extracting *causal* representations from X that affect Y. Such representations enable identification of treatment effects, and Z is one such representation.

The key takeaway from Lemma 1 is that certain additional assumptions are unavoidable for learning causal representations and thereby achieving identifiability. While previous works [36, 3] have assumed access to counterfactuals in observational data – an assumption that rarely holds in practice – we take a different approach by using simulators that generate synthetic counterfactuals from a related distribution. Our objective is to harness the full potential of these simulators and rigorously assess the estimation error caused by the mismatch between real-world and simulated distributions. This analysis drives the development of our algorithm, *SimPONet*, that imposes explicit regularization using the simulator supervision to enhance the accuracy of treatment effect estimates, surpassing what can be achieved using only the observational dataset. Through carefully designed experiments, we systematically vary the distributional gap between real and synthetic data across various DGPs, demonstrating that *SimPONet* consistently outperforms multiple baselines in estimating CATE.

Contributions: *(1)* We address Treatment Effect Estimation with post-treatment covariates, a problem known to be *non-identifiable* by leveraging a simulator that provides synthetic counterfactuals to learn causal representations suitable for this task. *(2)* We propose *SimPONet*, a novel training algorithm that maximizes simulator utility for both learning causal representations and effect estimation. *(3)* To our knowledge, this is the *first* work to systematically analyze the role of simulators in estimating CATE from post-treatment covariates. *(4)* We establish generalization bounds for CATE error, guiding *SimPONet*'s learning objective and highlighting the impact of simulator-real distribution mismatch. *(5)* Our experiments with diverse DGPs demonstrate the effectiveness of *SimPONet*.

We provide the related work in Appendix B and examples of real-world simulators in Appendix B.3.

2 Problem Formulation

We use random variables X,T,Y to denote covariates, binary treatments, and outcomes respectively. The observational dataset has n samples: $D_{\text{trn}} = \{(\mathbf{x}_i, t_i, y_i)\}_{i=1}^{n}$ where $t_i \in \mathcal{T} = \{0,1\}$ denotes treatment, $\mathbf{x}_i \in \mathcal{X} \subset \mathbb{R}^{n_x}$ denotes covariates observed after t_i is applied, and y_i the resulting outcome. We use the Neyman-Rubin potential outcomes framework to denote $Y_i(t), X_i(t)$ as the potential outcome and covariate for unit i under any treatment t. The main challenge is the absence of counterfactuals in D_{trn}, i.e., for each unit i, we observe covariates and outcomes under only one treatment t_i.

We use the random variable $Z \in \mathcal{Z} \subset \mathbb{R}^{n_z}$ to denote the causal representations of covariates X. Z generates X via covariate generating functions $g_t : \mathcal{Z} \mapsto \mathcal{X}$ for $t \in \{0,1\}$. We assume that g_t is diffeomorphic [33, 34, 57]; i.e., it is smooth, invertible, and has a smooth inverse. Diffeomorphism ensures that all factors involved in generating X are preserved within it so that there exists inverse functions $f_t : \mathcal{X} \mapsto \mathcal{Z}, \forall t \in \{0,1\}$ that could recover the causal representations Z back. Each training sample is obtained from the real DGP as follows: (1) $z_i \sim P_Z$, (2) $t_i \sim P(T|z_i)$, (3) $\mathbf{x}_i \sim P(X|z_i, t_i) = \delta(X - g_{t_i}(z_i))$, where δ denotes the dirac-delta distribution, (4) $y_i \sim P(Y|z_i, t_i) = \mathcal{N}(\mu_{t_i}(z_i), \sigma_y^2)$ is sampled from a Gaussian with mean $\mu_{t_i}(z_i)$ and constant variance σ_y^2. Here, $\mu_t : \mathcal{Z} \mapsto \mathcal{Y}\ \forall t$ generates responses for individual i under treatment t_i. We express the *factual* observed outcome for i as $Y_i(t_i) = \mu_{t_i}(f_{t_i}(\mathbf{x}_i))$, and the missing *counterfactual* (CF) outcome under $1-t_i$ as $Y_i(1-t_i) = \mu_{1-t_i}(f_{t_i}(\mathbf{x}_i))$.

Our Goal is Conditional Average Treatment Effect (CATE) estimation which quantifies the difference in outcomes due to a change in treatment. Given a test unit $(\mathbf{x}_j, t_j)$, its CATE is given by $\tau_j = \mathbb{E}[Y_j(T = 1) - Y_j(T = 0)|\mathbf{x}_j, t_j]$. As argued earlier, this introduces the sub goal of learning causal representations of observed covariates X using $f_t : X \mapsto Z$. We use $\tau : \mathcal{Z} \mapsto \mathcal{Y}$ to express the treatment effect using the latent z_j as $\tau(z_j) = \mu_1(z_j) - \mu_0(z_j)$. Since f_t inverts X to give Z, the same effect can also be expressed for $(\mathbf{x}_j, t_j)$ using τ_X as $\tau_X(\mathbf{x}_j, t_j) = \mu_1(f_{t_j}(\mathbf{x}_j)) - \mu_0(f_{t_j}(\mathbf{x}_j))$ where $\tau_X : \mathcal{X} \times \mathcal{T} \mapsto \mathcal{Y}$. Notice that $\tau_X(\bullet, t) = \tau \circ f_t(\bullet)$. When estimating τ_X, the factual outcome is easy, all we need to do is fit a regression model on the observation data. The main challenge lies in estimating the counterfactual outcome under treatment $1 - t_j$.

We begin by discussing Theorem 1 in [33] which presents an impossibility result stating that f_t which maps covariates to their latent factors is *not* identifiable solely using D_{trn}. The main hurdle is that multiple DGPs can yield the same marginal distribution $P(X,T)$, making it impossible to isolate the true DGP. However, prior work has shown how to learn f_t with *counterfactuals*, requiring that D_{trn} includes both covariates $X_i(t_i)$ and $X_i(1-t_i)$. Theorem 4.4 of [57] shows that such counterfactual supervision allows for recovery of Z up to a diffeomorphic transformation using contrastive learning. Proposition 2 in [65] further shows that the diffeomorphic transformation h is, in particular, a rotation in an n_z dimensional unit-normalized hypersphere.

Simulator DGP Since counterfactuals are not available in real data, we seek to leverage a simulator that generates paired instances giving rise to a counterfactual dataset $D_{\text{syn}} = \{\mathbf{x}_i^S(0), \mathbf{x}_i^S(1), y_i^S(0), y_i^S(1)\}$ generated using the DGP as shown in the lower panel in Figure 1. The simulated instances are obtained as follows: (1) $z_i \sim P_Z$; i.e., Z is sampled from the same distribution as real, (2) post-treatment covariates $\mathbf{x}_i^S(t) \sim P(X^S|Z = z_i, T = t) = \delta(X^S - g_t^S(z_i))$ under *both* treatments $t = \{0,1\}$. $g_t^S : \mathcal{Z} \mapsto \mathcal{X}\ \forall t$ are diffeomorphic functions, and (3) corresponding outcomes $y_i^S(0)$, $y_i^S(1)$ are sampled from $P(Y^S|Z = z_i, T = t) = \mathcal{N}(\mu_t^S(z_i), \sigma_{y^S}^2)$, where $\mu_t^S : \mathcal{Z} \mapsto \mathcal{Y}, \forall t$. Note that z_i remains hidden even in D_{syn}. We use "S" in the superscript to indicate a simulator component. Intuitively, if the structural equations of the simulator are close to those of real data, we can leverage them to improve the counterfactual estimates in the real distribution beyond what is possible solely using the observational data D_{trn}. Now we describe some metrics that assess the distance between real and simulator DGP.

Definition 1 [$d_{\mathbf{x}|t}(f_t, f_t^S)$] We assess the distance between the functions f_t and f_t^S using the following expected distance: $d_{\mathbf{x}|t}(f_t, f_t^S) = \mathbb{E}_{\mathbf{x} \sim P(X|t)}\left[||f_t(\mathbf{x}) - f_t^S(\mathbf{x})||_2^2\right]$.

Definition 2 [$d_z(\tau, \tau^S)$] We assess the distance between the real and simulator CATE functions on the P_Z distribution as: $d_z(\tau, \tau^S) = \mathbb{E}_{z \sim P_Z}\left[(\tau(z) - \tau^S(z))^2\right]$. Under composition with a diffeomorphic function h, we write $d_{h(z)}(\tau, \tau^S) = \mathbb{E}_{z \sim P_Z}\left[(\tau(h(z)) - \tau^S(h(z)))^2\right]$.

Assumptions for Identifying CATE τ_X. We summarise the assumptions that are needed on the real dataset D_{trn} and simulated counterfactual dataset D_{syn} to identify the CATE function τ_X: (A1) *Positivity:* $P(T = t|Z = z) > 0,\ \forall t \in \mathcal{T}, z \in \mathcal{Z}$. (A2) *Diffeomorphic Covariate Generation:* Covariates in both real and synthetic distributions are obtained through diffeomorphic transformations of Z under any treatment T. (A3) *Identifiability of τ given Z:* The causal factors Z that generate X form a sufficient adjustment set, blocking backdoor paths between T and Y, thus making τ identifiable from Z. Note that A2 and A3 together ensure that X contains information about all the relevant latent factors that affect the outcome Y and is a weaker notion of the commonly used *unconfoundedness* assumption.

CATE Error ($\mathcal{E}_{\text{CATE}}$): Given a test dataset $D_{\text{tst}} = \{(\mathbf{x}_j, t_j, y_j(0), y_j(1))\}_{j=1}^m$, with each $\mathbf{x}_j$ rendered under t_j, we assess the error in estimating CATE as: $\mathcal{E}_{\text{CATE}} = \frac{1}{m}\sum_j [\tau_j - \widehat{\tau}_j]^2$ where $\tau_j = y_j(1) - y_j(0)$ is the ground truth effect and $\widehat{\tau}_j$ is the predicted effect for the instance $(\mathbf{x}_j(t_j), t_j)$. The CATE error can be decomposed across treatment T as $\mathcal{E}_{\text{CATE}} = \sum_{t \in \mathcal{T}} P(T = t)\mathcal{E}_{\text{CATE}}^t$ where $\mathcal{E}_{\text{CATE}}^t = \int_{\mathbf{x} \in \mathcal{X}} [\tau_X(\mathbf{x}, t) - \widehat{\tau_X}(\mathbf{x}, t)]^2 P(\mathbf{x}|t) d\mathbf{x}$

Definition 3. Let us define *factual* error $\mathcal{E}_F^t$ and *counterfactual* error $\mathcal{E}_{CF}^t$ on samples with observed treatment t and missing treatment $1-t$ as follows:

$$\mathcal{E}_F^t = \int_{\mathbf{x} \in \mathcal{X}} [\mu_t(f_t(\mathbf{x})) - \widehat{\mu}_t(\widehat{f}_t(\mathbf{x}))]^2 P(\mathbf{x}|t) d\mathbf{x} \text{ and } \mathcal{E}_{CF}^t = \int_{\mathbf{x} \in \mathcal{X}} [\mu_{1-t}(f_t(\mathbf{x})) - \widehat{\mu}_{1-t}(\widehat{f}_t(\mathbf{x}))]^2 P(\mathbf{x}|t) d\mathbf{x}$$

Lemma 2. *The CATE error is related to the factual and counterfactual error as: $\mathcal{E}_{\text{CATE}}^t \leq 2\mathcal{E}_F^t + 2\mathcal{E}_{CF}^t$ [Proof in Appendix B.5.1]*

3 Learning Causal Representations for CATE

Our task involves learning two functions: $\widehat{f}_t$ that extracts the causal representation Z from X and $\widehat{\mu}_t$ that estimates the outcomes $Y(t)$. With access to *counterfactual simulated* data D_{syn} and *observational real* data D_{trn}, one can come up with the following approaches for estimating CATE: *(1)* SimOnly, which only uses D_{syn}, and *(2)* $\text{Real}_\mu\text{Sim}_f$, which uses D_{syn} to estimate f_t and subsequently, D_{trn} to estimate μ_t. We now discuss the training approach for each of these methods.

3.1 SimOnly Estimator

SimOnly solely uses counterfactual supervision in D_{syn} and identifies the simulator's DGP as follows: (Step 1) Estimating the inverse map f^S from covariate pairs $\{\mathbf{x}_i^S(0), \mathbf{x}_i^S(1)\}$ using contrastive learning [57]:

$$\{\widetilde{f}_0^S, \widetilde{f}_1^S\} = \underset{\{\widehat{f}_0^S, \widehat{f}_1^S\}}{\operatorname{argmin}} \mathbb{E}\left[-\log \frac{\exp(\operatorname{sim}(\hat{z}_i(1), \hat{z}_i(0))}{\sum_{j \neq i} \sum_{t,t'} \exp(\operatorname{sim}(\hat{z}_i(t), \hat{z}_j(t')))}\right] \text{ where } \hat{z}_i(t) = \widehat{f}_t^S(\mathbf{x}_i^S(t)) \quad (1)$$

where $\operatorname{sim}(\bullet, \bullet)$ is cosine similarity, $(\mathbf{x}_i^S(t), \mathbf{x}_i^S(1-t))$ denotes a positive pair with the same underlying latent z_i. A negative pair $(\mathbf{x}_i^S(t), \mathbf{x}_j^S(t'))$ has different (z_i, z_j). Contrastive learning increases similarity of representations of positive pairs $(\hat{z}_i(0), \hat{z}_i(1))$ while pushing apart the negative pairs $(\hat{z}_i(t), \hat{z}_{j \neq i}(t'))$. (Step 2) Estimating $\widetilde{\tau}^S(z) = \widetilde{\mu}_1^S(z) - \widetilde{\mu}_0^S(z)$ with supervision on difference of outcomes $\tau^S(f_t^S(\mathbf{x}_i^S(t))) = y_i^S(1) - y_i^S(0)$ as $\widetilde{\tau}^S = \operatorname{argmin}_{\tau^S} \mathbb{E}_{\mathbf{x}^S}\left[\tau^S(\widetilde{f}_t^S(\mathbf{x}^S(t))) - \tau^S(f_t^S(\mathbf{x}^S(t)))\right]^2$. SimOnly uses these estimates as-is on real data, i.e. $\widehat{\tau} = \widetilde{\tau}^S$ and $\widehat{f}_t = \widetilde{f}_t^S$, $\forall t \in \mathcal{T}$.

3.2 Real$_\mu$Sim$_f$ Estimator

Unlike SimOnly, which uses D_{syn} to learn both $\widehat{f}$ and $\widehat{\mu}$, this approach leverages D_{syn} solely to learn the representation extractor $\widehat{f}$. Specifically, it assumes that $\widehat{f}_t = \widetilde{f}_t^S$, as obtained from Eq. 1. Thereafter, it learns the $\widehat{\mu}$ parameters by applying a factual loss on D_{trn} to estimate $\widehat{\mu}_0, \widehat{\mu}_1 = \operatorname{argmin}_{\{\widehat{\mu}_0, \widehat{\mu}_1\}} \sum_{D_{\text{trn}}} (y_i - \widehat{\mu}_{t_i}(\widetilde{f}_{t_i}^S(\mathbf{x}_i)))^2$. We call this method Real$_\mu$Sim$_f$ since it borrows the causal representation extractor from the simulator DGP.

We will now perform a theoretical analysis to assess the above methods within a *linear* DGP framework, with the goal of identifying the conditions under which each model can accurately recover the CATE τ_X. We start by outlining the setup for the linear DGP.

Linear DGP. In this setup, all functions in both the real and simulator DGPs, as shown in Figure 1, are linear, allowing us to derive closed-form solutions for the three methods. Training datasets D_{trn} and D_{syn} are sampled as follows: (1) Latent variables $z \in \mathbb{R}^{n_z}$ are drawn from distribution P_Z. (2) Real and simulator covariates for a treatment t are computed as $g_t(z) = z\boldsymbol{R}_t$ and $g_t^S(z) = z\boldsymbol{S}_t$, with $\boldsymbol{R}_t$ and $\boldsymbol{S}_t$ being invertible matrices. (3) Outcomes are generated as $\mu_t(z) = z^\top w_t$ and $\mu_t^S(z) = z^\top w_t^S$, where w_t and w_t^S are vectors in $\mathbb{R}^{n_z}$. The closed-form CATE error expressions for each model are summarized in Table 1, with detailed derivations in Appendix B.6. The last column of Table 1 specifies conditions for zero CATE error, which we argue are challenging to meet in real-world scenarios.

Method	Estimate for CATE $\widehat{\tau_X}(\mathbf{x}^\star, 1)$	CATE Error $[\widehat{\tau_X}(\mathbf{x}^\star, 1) - \tau(\mathbf{x}^\star, 1)]^2$	Favorable Condition
SimOnly	$\mathbf{x}^\star(\boldsymbol{S}_1^{-1} w_1^S - \boldsymbol{S}_1^{-1} w_0^S)$	$\left[\mathbf{x}^\star(\boldsymbol{R}_1^{-1} w_\tau - \boldsymbol{S}_1^{-1} w_\tau^S)\right]^2$	$\tau = \tau^S, f_t = f_t^S$
Real$_\mu$Sim$_f$	$\mathbf{x}^\star \boldsymbol{S}_1^{-1} \boldsymbol{S}_1 \boldsymbol{R}_1^{-1} w_1 - \mathbf{x}^\star \boldsymbol{S}_1^{-1} \boldsymbol{S}_0 \boldsymbol{R}_0^{-1} w_0$	$\left[\mathbf{x}^\star(\boldsymbol{R}_1^{-1} - \boldsymbol{S}_1^{-1} \boldsymbol{S}_0 \boldsymbol{R}_0^{-1}) w_0\right]^2$	$f_t^S = f_t$

Table 1: RQ1: This table presents the predicted CATE and the corresponding CATE errors obtained from the baseline models computed analytically for a test instance $\mathbf{x}^\star$ observed under treatment 1. The final column indicates the conditions under which each model yields accurate CATE estimates.

1. *SimOnly* relies on the simulator perfectly matching the real world, but building such accurate simulators is highly difficult, making this method unsuitable for CATE estimation.
2. *Real$_\mu$Sim$_f$*, as expected, requires alignment between simulator and real-world covariates, i.e., $\mathbf{x}_t = g_t(z) = g_t^S(z) = \mathbf{x}_t^S$. This limitation arises because it simply learns the representation extractor from D_{syn} without adjusting for real covariates. Moreover, performing such an adjustment is not clear because D_{trn} lacks counterfactuals, which prevents the use of contrastive loss needed for training $\widehat{f}$.

Since none of these favorable conditions are likely to hold in real life, all three methods are prone to large CATE error. To improve upon this, we first derive generalization bounds on the CATE error that apply to arbitrary DGP settings, aiming to uncover better learning objectives from the bound.

Lemma 3. *Assume τ is K_τ-Lipschitz, and $\widetilde{f}^S$ and $\widetilde{\tau}^S$ are estimates from the simulator DGP. Then, the CATE error on the estimates $\widehat{f}_t$ and $\widehat{\tau}$ admits the following bound:*

$$\mathcal{E}_{CATE}^t(\widehat{f}_t, \widehat{\tau}) \leq [12 d_{h(z)}(\widehat{\tau}, \widetilde{\tau}^S) + 12 K_\tau^2 d_{\mathbf{x}|t}(\widehat{f}_t, \widetilde{f}_t^S) + 8\mathcal{E}_F^t] + [12 d_z(\tau, \tau^S) + 12 K_\tau^2 d_{\mathbf{x}|t}(f_t, f_t^S)]$$

where $d_{\mathbf{x}|t}, d_z, d_{h(z)}$ are distance functions defined in section 2. [Proof in Appendix B.5.2.]

The expression in blue corresponds to discrepancy between real and simulator functions, and cannot be minimised. Whereas, the remaining terms can be minimised by training on $D_{\text{trn}}, D_{\text{syn}}$.

3.3 *SimPONet* Estimator

We now introduce *SimPONet*, which adopts the first three terms from lemma 3 into its objective. This leads to a joint learning framework, leveraging supervision from both D_{trn} and D_{syn}. *SimPONet* relaxes the strict equality $\widehat{f}_t = \widetilde{f}_t^S$ used by Real$_\mu$Sim$_f$, and instead uses $\widetilde{f}_t^S$ as a regularizer, while ensuring that $\widehat{\mu}_t$ accurately predicts the factual outcomes for instances in D_{trn}. Additionally, it imposes the τ^S loss on simulated instances to leverage any potential closeness between the true treatment effect, τ, and the simulated treatment effect, τ^S. Moreover, τ^S loss is necessary to avoid degenerate solutions. Note that $\tau_X = \tau \circ f_t$ has two degrees of freedom in τ and f_t. Applying regularisation only on $\widehat{f}_t$ can drive the regulariser $||\widehat{f}_t(\mathbf{x}) - \widetilde{f}_t^S(\mathbf{x})||_2^2$ to zero by making $\widehat{f}_t = \widetilde{f}_t^S$ and still minimise factual error $\mathcal{E}_F^t$ by updating $\widehat{\mu}_t$. Thus *SimPONet* collapses to the Real$_\mu$Sim$_f$ estimator. The τ^S loss helps avoid such degeneracies. *SimPONet*'s overall loss is:

$$\min_{\{\widehat{\mu}_t, \widehat{f}_t\}} \underbrace{\sum_{D_{\text{trn}}} \Big(y_i - \widehat{\mu}_{t_i}(\widehat{f}_{t_i}(\mathbf{x}_i))\Big)^2}_{\text{Factual Loss on } D_{\text{trn}}} + \underbrace{\lambda_f \sum_{D_{\text{trn}}} \|\widetilde{f}_{t_i}^S(\mathbf{x}_i) - \widehat{f}_{t_i}(\mathbf{x}_i)\|_2^2}_{d(\widetilde{f}_t^S, \widehat{f}_t)} + \underbrace{\lambda_\tau \sum_{D_{\text{syn}}} \sum_{t \in \{0,1\}} \Big(\tau_i^S - \widehat{\tau}(\widetilde{f}_t^S(\mathbf{x}_i^S(t)))\Big)^2}_{\tau^S \text{ loss on } D_{\text{syn}}} \tag{2}$$

where $\tau_i^S = y_i^S(1) - y_i^S(0)$ and $\lambda_\tau, \lambda_f > 0$ are loss weights. We present the *SimPONet*'s pseudocode in Appendix B.4.

4 Experiments

We conduct experiments across several DGP settings, by systematically varying the gap between the real and simulator components to assess how *SimPONet* performs compared to the baselines.

4.1 Linear DGP: Linear f, Linear μ

In this experiment, we generate samples according to the Linear DGP outlined in Section 3. We test two configurations for Z: (a) when Z is sampled from a multivariate Gaussian distribution, and (b) when Z is directly taken from the real-world IHDP dataset. We control the gap between real and simulator distributions using the constants γ_R, γ_{RS}, and $\gamma_\tau > 0$ as follows: *(1)* Initialize $\boldsymbol{R}_0^{-1}, w_0 \sim \mathcal{N}(0,1)$. *(2)* To inject a distance $\gamma_R \in (0, 0.5)$ between $\boldsymbol{R}_0^{-1}$ and $\boldsymbol{R}_1^{-1}$, set $\boldsymbol{R}_1^{-1} = (1-\gamma_R)\boldsymbol{R}_0^{-1} + \gamma_R \mathcal{N}(0,1)$. *(3)* Set $w_1 \sim \gamma w_0 + (1-\gamma)\mathcal{N}(0,1)$. We use $\gamma = 0.4$ in all experiments. *(4)* Similarly, inject a γ_{RS} gap between $\boldsymbol{R}_t^{-1}$ and $\boldsymbol{S}_t^{-1}$. *(5)* For treatment effect parameters $w_\tau = w_1 - w_0$ in the real DGP, we sample its simulator counterpart with a gap γ_τ as $w_\tau^S = (1-\gamma_\tau)w_\tau + \gamma_\tau \mathcal{N}(0,1)$ and set w_t^S accordingly.

We compare *SimPONet* with SimOnly, and Real$_\mu$Sim$_f$methods. While the baselines offer closed-form solutions, *SimPONet*'s loss function is more complex; so, we optimize it to a local minimum via alternating minimization. See B.6 for details. As shown in Table 2, *SimPONet* performs either best or second-best in both synthetic and real-world settings. Its CATE error remains controlled largely due to its capability to bound errors in the counterfactual distribution. In contrast, the Real$_\mu$Sim$_f$ model perform well only in specific DGP settings but significantly underperform in other settings due to high counterfactual error, resulting in poor CATE estimates.

Table 2: Linear DGP experiment across various levels of gaps between the real and simulated data. "low" refers to a γ value of 0.1; "high" denotes 0.4.

			Synthetic-Gaussian			Real World-IHDP		
$d(f_0,f_1)$	$d(f_t,f_t^S)$	$d(\tau,\tau^S)$	SimOnly	Real$_\mu$Sim$_f$	*SimPONet*	SimOnly	Real$_\mu$Sim$_f$	*SimPONet*
0.00	high	high	2.82 (0.27)	15.75 (0.01)	2.58 (0.00)	3.57 (0.11)	48.76 (0.05)	3.20 (0.00)
low	low	low	0.63 (0.00)	1.19 (0.01)	0.54 (0.00)	1.00 (0.44)	2.73 (0.00)	0.97 (0.00)
low	low	high	1.57 (0.16)	1.19 (0.83)	1.39 (0.00)	1.62 (0.26)	2.73 (0.02)	1.49 (0.00)
low	high	low	2.14 (0.22)	15.75 (0.01)	1.85 (0.00)	3.67 (0.31)	48.76 (0.05)	3.37 (0.00)
low	high	high	2.47 (0.56)	15.75 (0.01)	2.57 (0.00)	3.57 (0.11)	48.76 (0.05)	3.19 (0.00)
high	low	low	0.63 (0.00)	1.19 (0.01)	0.54 (0.00)	1.00 (0.47)	2.73 (0.00)	0.98 (0.00)
high	low	high	1.57 (0.16)	1.19 (0.83)	1.39 (0.00)	1.62 (0.27)	2.73 (0.02)	1.50 (0.00)
high	high	low	2.14 (0.21)	15.75 (0.01)	1.85 (0.00)	3.67 (0.31)	48.76 (0.05)	3.38 (0.00)
high	high	high	2.82 (0.26)	15.75 (0.01)	2.57 (0.00)	3.57 (0.11)	48.76 (0.05)	3.19 (0.00)

4.2 Non-Linear DGP: Linear f, Non-Linear μ

In the interest of space, we defer the details of this experient to Appendix B.6.

4.3 Arbitrary DGPs: Real-world Semi-Synthetic Datasets

We perform experiments using semi-synthetic observational datasets commonly used for evaluating treatment effect estimation methods: the Infant Health Development Program (IHDP) and the Atlantic Causal Inference Conference (ACIC) datasets. These datasets provide real-world pre-treatment covariates (Z). For details, see Appendix B.7. To align these datasets with our study, we apply diffeomorphic and non-linear RealNVP Normalizing Flows [16] to transform Z into post-treatment covariates X. We use randomly initialized flows with two coupling layers: g_0 and g_1 for real data, and g_0^S and g_1^S for synthetic data. Real outcomes are taken directly from the dataset, while simulator outcomes are synthesized with a gap γ_τ as follows: (1) sample $w_\tau^S \in \mathbb{R}^{n_z} \sim \mathcal{N}(0,1)$, (2) set $\tau^S(z) = \tau(z) + (\sigma(\tau) \cdot \gamma_\tau \cdot z^\top w_\tau^S)$, where $\sigma(\tau)$ is the standard deviation of ITE labels. When $\gamma_\tau = 0$, $\tau = \tau^S$; when $\gamma_\tau = 1$, τ is challenging to recover from τ^S.

We compared *SimPONet* against baselines from the CATENets [13], a well-known ITE benchmarking library. We pass the representations extracted using $\widetilde{f}_t^S(\mathbf{x})$ as input to the baselines. We present the results in Table 3 for $\gamma_\tau = 0.1$, and make the following key observations:

Table 3: RQ3: Comparison of *SimPONet* with baselines in CATENets library. We show p-values in brackets. *SimPONet* outperforms others overall, while SimOnly performs best on ACIC-2 since $\tau = \tau^S$ there.

Method	IHDP	ACIC-2	ACIC-7	ACIC-26
RNet [40]	1.54 (0.00)	3.30 (0.00)	5.91 (0.04)	6.06 (0.18)
XNet [31]	1.0 (0.00)	0.43 (0.15)	5.49 (0.17)	5.1 (0.38)
DRNet [51]	0.96 (0.00)	0.24 (0.59)	5.53 (0.15)	5.08 (0.39)
CFRNet [53]	0.96 (0.00)	0.36 (0.26)	5.55 (0.15)	5.09 (0.38)
FlexTENet [11]	0.96 (0.00)	0.32 (0.32)	5.46 (0.19)	5.04 (0.40)
DragonNet [54]	0.96 (0.00)	0.29 (0.41)	5.57 (0.14)	5.09 (0.38)
IPW [47]	0.96 (0.00)	0.36 (0.24)	5.56 (0.15)	5.09 (0.38)
k-NN [55]	0.96 (0.00)	0.33 (0.33)	5.48 (0.18)	5.13 (0.37)
PerfectMatch [50]	0.98 (0.00)	0.56 (0.11)	5.75 (0.08)	5.13 (0.37)
StableCFR [60]	1.01 (0.00)	1.09 (0.03)	5.56 (0.15)	5.08 (0.43)
ESCFR [59]	0.96 (0.00)	0.27 (0.47)	5.55 (0.15)	5.79 (0.21)
SimOnly	0.94 (0.00)	0.00 (0.98)	6.65 (0.00)	6.60 (0.12)
Real$_\mu$Sim$_f$	0.96 (0.00)	0.17 (0.76)	5.57 (0.14)	5.09 (0.38)
SimPONet	0.79 (0.00)	0.26 (0.00)	5.04 (0.00)	4.67 (0.00)

(a) **ACIC-2** is an unusual dataset where the potential outcomes μ_t exhibit a complex non-linear pattern, yet their difference, τ, remains constant. Our sampling scheme for τ^S, as outlined earlier, gives $\tau^S = \tau$ for any γ_τ in this dataset. Consequently, SimOnly outperforms all baselines. *SimPONet*'s performance is affected due its loss on factual potential outcomes in D_{trn}. This could have been avoided by setting a large weight on τ^S regularizer in the objective. However, in general, tuning this weight without explicit τ supervision is not straightforward, and so we opt not to adjust it. (b) **IHDP, ACIC-7,26** Across these datasets, the CATENets baselines significantly underperform *SimPONet*. *SimPONet* achieves the best performance by leveraging inductive biases from the closeness between τ and τ^S in the synthetic dataset. One notable approach that stands out is FlexTENet [11], which explicitly shares parameters across the $\widehat{\mu}_0$ and $\widehat{\mu}_1$ networks, and offers the second best performance in ACIC-7, 26.

5 Conclusion

In this paper, we addressed the problem of estimating treatment effects for individuals whose covariates are influenced by the treatment, a setting not solvable using observational data alone. We proposed to solve this task using off-the-shelf simulators that synthesize counterfactuals, unlike prior work relying on real-world counterfactuals, which limits their practical applicability. Ours is the first work to systematically analyse the role of simulators in handling the limitation of lack of counterfactual supervision in real world observational data. We introduced *SimPONet*, which balances learning from real and simulator distributions to bound the rather intractable counterfactual error. Our theoretical analysis showed that *SimPONet* has better CATE generalization bounds under reasonable assumptions in contrast to other proposals that need strong assumptions on the DGP. Extensive experiments with synthetic and real-world datasets demonstrated that *SimPONet* is indeed a superior alternative.

References

[1] Joshua Angrist. Estimating the labor market impact of voluntary military service using social security data on military applicants, 1995.

[2] Orley Ashenfelter. Estimating the effect of training programs on earnings. *The Review of Economics and Statistics*, pages 47–57, 1978.

[3] Philip Bachman, R Devon Hjelm, and William Buchwalter. Learning representations by maximizing mutual information across views. *Advances in neural information processing systems*, 32, 2019.

[4] Elias Bareinboim and Judea Pearl. Controlling selection bias in causal inference. In *Artificial Intelligence and Statistics*, pages 100–108. PMLR, 2012.

[5] Elias Bareinboim and Jin Tian. Recovering causal effects from selection bias. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29, 2015.

[6] Nitay Calderon, Eyal Ben-David, Amir Feder, and Roi Reichart. DoCoGen: Domain counterfactual generation for low resource domain adaptation. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7727–7746, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.533. URL `https://aclanthology.org/2022.acl-long.533`.

[7] Vinod K Chauhan, Soheila Molaei, Marzia Hoque Tania, Anshul Thakur, Tingting Zhu, and David A Clifton. Adversarial de-confounding in individualised treatment effects estimation. In *International Conference on Artificial Intelligence and Statistics*, pages 837–849. PMLR, 2023.

[8] Zeming Chen, Qiyue Gao, Antoine Bosselut, Ashish Sabharwal, and Kyle Richardson. DISCO: Distilling counterfactuals with large language models. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5514–5528, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.302. URL `https://aclanthology.org/2023.acl-long.302`.

[9] Alexander Coppock. Avoiding post-treatment bias in audit experiments. *Journal of Experimental Political Science*, 6(1):1–4, 2019.

[10] Juan Correa, Jin Tian, and Elias Bareinboim. Generalized adjustment under confounding and selection biases. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

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

[12] Alicia Curth and Mihaela van der Schaar. In search of insights, not magic bullets: Towards demystification of the model selection dilemma in heterogeneous treatment effect estimation. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, 2023.

[13] Alicia Curth, David Svensson, Jim Weatherall, and Mihaela van der Schaar. Really doing great at estimating cate? a critical look at ml benchmarking practices in treatment effect estimation. In *Thirty-fifth conference on neural information processing systems datasets and benchmarks track (round 2)*, 2021.

[14] Chiara Dalla Man, Robert A Rizza, and Claudio Cobelli. Meal simulation model of the glucose-insulin system. *IEEE Transactions on biomedical engineering*, 54(10):1740–1749, 2007.

[15] Nikhil J Dhinagar, Sophia I Thomopoulos, Emily Laltoo, and Paul M Thompson. Counterfactual mri generation with denoising diffusion models for interpretable alzheimer's disease effect detection. *bioRxiv*, pages 2024–02, 2024.

[16] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real nvp. *arXiv preprint arXiv:1605.08803*, 2016.

[17] Tanay Dixit, Bhargavi Paranjape, Hannaneh Hajishirzi, and Luke Zettlemoyer. CORE: A retrieve-then-edit framework for counterfactual data generation. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2964–2984, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-emnlp.216. URL `https://aclanthology.org/2022.findings-emnlp.216`.

[18] Zijun Gao and Yanjun Han. Minimax optimal nonparametric estimation of heterogeneous treatment e ffects. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. L in, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 21751–21762. Curran Associates, Inc., 2020. URL `https://proceedings.neurips.cc/paper_files/paper/2020/file/f75b757d3459\c3e93e98ddab7b903938-Paper.pdf`.

[19] Muhammad Waleed Gondal, Manuel Wuthrich, Djordje Miladinovic, Francesco Locatello, Martin Breidt, Valentin Volchkov, Joel Akpo, Olivier Bachem, Bernhard Schölkopf, and Stefan Bauer. On the transfer of inductive bias from simulation to the real world: a new disentanglement dataset. *Advances in Neural Information Processing Systems*, 32, 2019.

[20] Yu Gu, Jianwei Yang, Naoto Usuyama, Chunyuan Li, Sheng Zhang, Matthew P Lungren, Jianfeng Gao, and Hoifung Poon. Medjourney: Counterfactual medical image generation by instruction-learning from multimodal patient journeys. 2023.

[21] Negar Hassanpour and Russell Greiner. Counterfactual regression with importance sampling weights. In *IJCAI*, pages 5880–5887, 2019.

[22] Negar Hassanpour and Russell Greiner. Learning disentangled representations for counterfactual regression. In *International Conference on Learning Representations*, 2019.

[23] JONATHAN HOMOLA, MIGUEL M. PEREIRA, and MARGIT TAVITS. Fixed effects and post-treatment bias in legacy studies. *American Political Science Review*, 118(1):537–544, 2024. doi: 10.1017/S0003055423001351.

[24] Qiang Huang, Defu Cao, Yi Chang, and Yan Liu. Extracting post-treatment covariates for heterogeneous treatment effect estimation. 2023.

[25] Stefano M Iacus, Gary King, and Giuseppe Porro. Causal inference without balance checking: Coarsened exact matching. *Political analysis*, 20(1):1–24, 2012.

[26] Guillaume Jeanneret, Loïc Simon, and Frédéric Jurie. Diffusion models for counterfactual explanations. In *Proceedings of the Asian Conference on Computer Vision*, pages 858–876, 2022.

[27] Yonghan Jung, Jin Tian, and Elias Bareinboim. Learning causal effects via weighted empirical risk minimization. *Advances in neural information processing systems*, 33:12697–12709, 2020.

[28] Nathan Kallus. Deepmatch: Balancing deep covariate representations for causal inference using adversarial training. In *International Conference on Machine Learning*, pages 5067–5077. PMLR, 2020.

[29] Edward H Kennedy. Towards optimal doubly robust estimation of heterogeneous causal effects. *arXiv preprint arXiv:2004.14497*, 2020.

[30] Gary King. A hard unsolved problem? post-treatment bias in big social science questions. In *Hard Problems in Social Science" Symposium, April*, volume 10, 2010.

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

[32] Oran Lang, Ilana Traynis, and Yun Liu. Explaining counterfactual images. *Nature Biomedical Engineering*, 2023. URL `https://rdcu.be/dwVKK`.

[33] Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Raetsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *international conference on machine learning*, pages 4114–4124. PMLR, 2019.

[34] Francesco Locatello, Michael Tschannen, Stefan Bauer, Gunnar R¨¨ätsch, Bernhard Schölkopf, and Olivier Bachem. Disentangling factors of variations using few labels, 2019. URL `https://openreview.net/forum?id=SkGy6hjvPE`.

[35] Nishtha Madaan, Inkit Padhi, Naveen Panwar, and Diptikalyan Saha. Generate your counterfactuals: Towards controlled counterfactual generation for text. In *AAAI Conference on Artificial Intelligence*, 2020. URL `https://api.semanticscholar.org/CorpusID:228063841`.

[36] Lokesh Nagalapatti, Guntakanti Sai Koushik, Abir De, and Sunita Sarawagi. Learning recourse on instance environment to enhance prediction accuracy. In *Advances in Neural Information Processing Systems*, 2022.

[37] Lokesh Nagalapatti, Akshay Iyer, Abir De, and Sunita Sarawagi. Continuous treatment effect estimation using gradient interpolation and kernel smoothing. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(13):14397–14404, Mar. 2024. doi: 10.1609/aaai.v38i13.29353. URL `https://ojs.aaai.org/index.php/AAAI/article/view/29353`.

[38] Lokesh Nagalapatti, Pranava Singhal, Avishek Ghosh, and Sunita Sarawagi. Pairnet: Training with observed pairs to estimate individual treatment effect. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net, 2024. URL `https://openreview.net/forum?id=o5SVr80Rgg`.

[39] Lizhen Nie, Mao Ye, Qiang Liu, and Dan Nicolae. Vcnet and functional targeted regularization for learning causal effects of continuous treatments. *arXiv preprint arXiv:2103.07861*, 2021.

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

[41] Michal Ozery-Flato, Pierre Thodoroff, and Tal El-Hay. Adversarial balancing for causal inference. *ArXiv*, abs/1810.07406, 2018.

[42] Yushu Pan and Elias Bareinboim. Counterfactual image editing. *arXiv preprint arXiv:2403.09683*, 2024.

[43] Nick Pawlowski, Daniel Coelho de Castro, and Ben Glocker. Deep structural causal models for tractable counterfactual inference. *Advances in neural information processing systems*, 33: 857–869, 2020.

[44] J. Pearl and Cambridge University Press. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2000. ISBN 9780521773621. URL `https://books.google.co.in/books?id=wnGU_TsW3BQC`.

[45] Judea Pearl. Conditioning on post-treatment variables. *Journal of Causal Inference*, 3(1): 131–137, 2015.

[46] Marcel Robeer, Floris Bex, and Ad Feelders. Generating realistic natural language counterfactuals. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3611–3625, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-emnlp.306. URL `https://aclanthology.org/2021.findings-emnlp.306`.

[47] James M Robins, Andrea Rotnitzky, and Lue Ping Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American statistical Association*, 89(427):846–866, 1994.

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

[49] Axel Sauer and Andreas Geiger. Counterfactual generative networks. In *International Conference on Learning Representations*, 2021. URL `https://openreview.net/forum?id=BXewfAYMmJw`.

[50] Patrick Schwab, Lorenz Linhardt, and Walter Karlen. Perfect match: A simple method for learning representations for counterfactual inference with neural networks. *arXiv preprint arXiv:1810.00656*, 2018.

[51] Patrick Schwab, Lorenz Linhardt, Stefan Bauer, Joachim M Buhmann, and Walter Karlen. Learning counterfactual representations for estimating individual dose-response curves. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 5612–5619, 2020.

[52] Uri Shalit, Fredrik D. Johansson, and David Sontag. Estimating individual treatment effect: generalization bounds and algorithms, 2016. URL `https://arxiv.org/abs/1606.03976`.

[53] Uri Shalit, Fredrik D Johansson, and David Sontag. Estimating individual treatment effect: generalization bounds and algorithms. In *International Conference on Machine Learning*, pages 3076–3085. PMLR, 2017.

[54] Claudia Shi, David Blei, and Victor Veitch. Adapting neural networks for the estimation of treatment effects. *Advances in neural information processing systems*, 32, 2019.

[55] Elizabeth A Stuart. Matching methods for causal inference: A review and a look forward. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 25(1):1, 2010.

[56] Jayaraman Thiagarajan, Vivek Sivaraman Narayanaswamy, Deepta Rajan, Jia Liang, Akshay Chaudhari, and Andreas Spanias. Designing counterfactual generators using deep model inversion. *Advances in Neural Information Processing Systems*, 34:16873–16884, 2021.

[57] Julius Von Kügelgen, Yash Sharma, Luigi Gresele, Wieland Brendel, Bernhard Schölkopf, Michel Besserve, and Francesco Locatello. Self-supervised learning with data augmentations provably isolates content from style. *Advances in neural information processing systems*, 34:16451–16467, 2021.

[58] Clinton J. Wang, Natalia S. Rost, and Polina Golland. Spatial-intensity transform gans for high fidelity medical image-to-image translation. In Anne L. Martel, Purang Abolmaesumi, Danail Stoyanov, Diana Mateus, Maria A. Zuluaga, S. Kevin Zhou, Daniel Racoceanu, and Leo Joskowicz, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*, pages 749–759, Cham, 2020. Springer International Publishing. ISBN 978-3-030-59713-9.

[59] Hao Wang, Jiajun Fan, Zhichao Chen, Haoxuan Li, Weiming Liu, Tianqiao Liu, Quanyu Dai, Yichao Wang, Zhenhua Dong, and Ruiming Tang. Optimal transport for treatment effect estimation. *Advances in Neural Information Processing Systems*, 36, 2024.

[60] Anpeng Wu, Kun Kuang, Ruoxuan Xiong, Bo Li, and Fei Wu. Stable estimation of heterogeneous treatment effects. In *International Conference on Machine Learning*, pages 37496–37510. PMLR, 2023.

[61] Liuyi Yao, Sheng Li, Yaliang Li, Mengdi Huai, Jing Gao, and Aidong Zhang. Representation learning for treatment effect estimation from observational data. *Advances in Neural Information Processing Systems*, 31, 2018.

[62] Jinsung Yoon, James Jordon, and Mihaela Van Der Schaar. Ganite: Estimation of individualized treatment effects using generative adversarial nets. In *International Conference on Learning Representations*, 2018.

[63] Yao Zhang, Alexis Bellot, and Mihaela Schaar. Learning overlapping representations for the estimation of individualized treatment effects. In *International Conference on Artificial Intelligence and Statistics*, pages 1005–1014. PMLR, 2020.

[64] Yi-Fan Zhang, Hanlin Zhang, Zachary C. Lipton, Li Erran Li, and Eric P. Xing. Exploring transformer backbones for heterogeneous treatment effect estimation, 2022. URL `https://arxiv.org/abs/2202.01336`.

[65] Roland S. Zimmermann, Yash Sharma, Steffen Schneider, Matthias Bethge, and Wieland Brendel. Contrastive learning inverts the data generating process. 139:12979–12990, 2021. URL http://proceedings.mlr.press/v139/zimmermann21a.html.

A Appendix / supplemental material

B Related Work

CATE with Pre-Treatment Covariates has been widely researched where the primary challenge is to handle confounding that arises out of biased treatment assignment in observational datasets. The main ideas explored include: estimating pseudo-outcomes for missing treatments in the training dataset and then using these to train effect predictors [18, 11, 40, 29, 62, 63, 37]; adding targeted regularizers to ensure consistent ITE [54, 39, 64]; learning balanced representation of covariates across treatment groups [52, 53, 61, 7, 59, 60]; matching to near-by covariates [55, 48, 25, 50, 28, 38]; and weighing losses to mitigate confounding [21, 22, 27, 41]. In our problem, if Z could be recovered perfectly, all of these methods could be applied, and we will present a comparison with imperfectly extracted Z.

CATE with Post-Treatment Covariates is more challenging and falls into the third rung (counterfactual) of Pearl's causal ladder [44]; a formal proof is in [45]. In economics, post-treatment variables in trials are known to exacerbate estimated causal effects [9, 23, 30]. Post-treatment variables have been used to estimate selection bias $P(T|Z)$ in observational data [4, 5, 10]. A very closely related work is [24] that leverages post-treatment variables for estimating treatment effects but differs from us since they assume: (1) covariates X causally affect Y, and (2) an entangled version of X,Z is observed; they simply focus on disentangling Z through representation learning. Our setting is more challenging as Z is latent.

Prior Work on Counterfactual Simulators In Appendix B.2, we discuss previous works that leverage recent foundation models to generate pseudo-counterfactuals in the real distribution. Furthermore, we describe two real-world applications where proprietary toolboxes were used to construct simulators for estimating treatment effects. One such example uses the SimBiology toolbox to build a pharmacological simulator, which assesses the effect of SGLT2 inhibitors (T) in managing type-2 diabetes (Y), considering post-treatment covariates such as glucose levels across various body parts. Another example involves the Ansys Battery and Electrode Simulator, which analyzes how different materials (T) impact battery longevity (Y), based on post-treatment variables (X) such as charge/discharge rates, temperature, and other factors influenced by material properties in an electrochemical setting. Further details are provided in Appendix B.3.

B.1 Code

We have released the code in the anonynous URL `https://anonymous.4open.science/r/catenets-simponet/README.md`. We have also uploaded the code along with our submission.

B.2 Counterfactual Simulators

Here we discuss prior works that train generative models for generating counterfactuals. In general, to obtain counterfactuals in the real distribution, we need to follow three steps [43]: (a) *abduction*, inverting X to obtain Z, (b) *action*, changing the observed treatment, and (c) *prediction* generating a new X under the new treatment. These steps require prior knowledge of the DGP specifications, which are often difficult to define and cannot be learned from observational data alone [43]. Consequently, many methods bypass the principled approach and use pre-trained models like Diffusion models and Large Language models to generate pseudo counterfactuals from a related synthetic domain. Such simulators are proposed across various modalities, including images [43, 20, 56, 42, 49, 26], text [35, 6, 8, 17, 46], and healthcare [32, 58, 15]. However, the simulated data should be used with caution. Prior research [19] shows that such counterfactual data is not directly usable for downstream tasks but provides strong inductive biases that transfer well to the real distribution. Our method can incorporate any such counterfactual generators as simulators, as long as they guide the learning of the Z extractor and ensure that the treatment effects estimated from simulated data closely match the effects in real data.

B.3 Real-World Applications of Simulators for Estimating CATE

We provide two examples for showcasing how simulators are used in medicine and electrochemistry below:

Medicine. Simulators play a crucial role in pharmacology, particularly for assessing drug efficacies. For instance, the SimBiology toolbox [2] in MATLAB is commonly used to predict the effects of `SGLT2`

[2] `https://in.mathworks.com/videos/series/simbiology-tutorials-for-qsp-pbpk-and-pk-pd-modeling-and-analysis.html`

inhibitors (T) on type-2 diabetes (Y) while considering post-treatment covariates (X) such as plasma glucose levels, gut glucose levels, urinary glucose excretion, and liver insulin levels. SimBiology enables modeling these effects using differential and algebraic equations that are often calibrated on target populations to minimize the real-simulator mismatch. Despite not perfectly replicating reality, such simulators are invaluable for early-stage clinical trial decisions and have demonstrated utility in modeling short-term treatment effects [14].

Electrochemistry. Another application involves recommending optimal electrode materials to maximize battery capacity (Y). By observing Y under various electrode materials (T) and post-treatment variables like charge/discharge rate, internal resistance, and temperature distribution (X), the Ansys Battery Cell and Electrode Simulator [3] provides realistic electrochemical simulations. This tool has been used by Volkswagen Motorsport for comprehensive multiphysics simulations to design and validate battery models. Such simulators are highly relevant for practical decision-making in industries.

These examples illustrate the practical relevance of simulators across different fields. While simulators cannot fully replace real data or randomized controlled trials (RCTs), they offer valuable insights that can reduce the number of RCTs needed for optimal treatment identification. Our paper aims to characterize the CATE error when using imperfect simulators in conjunction with real observational data. Additionally, *SimPONet* maximizes the utility of simulators by leveraging the highly correlated simulator's treatment effects with real-world effects, without relying on the exact correlation of individual potential outcomes.

B.4 *SimPONet* Pseudocode

Here, we present the *SimPONet* pseudocode. The steps involved in our algorithm are:

line 1 First we use the simulator dataset D_{syn} to apply contrastive losses on the counterfactual covariates using Eq. 1. This optimization gives us a Z extractor in the simulator distribution, which we denote as $\widetilde{f}_t^S$.

line 2 We partition the training dataset into train, validation dataset using stratified split based on T. We then initialize the loss weigts λ_f, λ_τ to their defaults.

line 3 We can now apply gradient descent algorithm on the *SimPONet*'s objective in Eq. 2 to train the $\widehat{\mu}_t, \widehat{f}_t$ parameters of the model.

Algorithm 1 *SimPONet* Algorithm

```
Require: Observational Data $D_{\text{trn}}$: $\{(\mathbf{x}_i, t_i, y_i)\}$, Simulator Data $D_{\text{syn}}$: $\{(\mathbf{x}_i^S(0), \mathbf{x}_i^S(1), y_i^S(0), y_i^S(1))\}$
1: Let $\widetilde{f}_t^S \leftarrow$ Eq. 1 (Minimize Contrastive loss on $D_{\text{syn}}$)
2: Set $D_{\text{trn}}, D_{\text{val}} \leftarrow \text{SPLIT}(D_{\text{trn}}, pc = 0.3, \text{stratify}=T)$, and init default hyperparameters $\lambda_f, \lambda_\tau \leftarrow 1, 1$
3: $\{\widehat{f}_t, \widehat{\mu}_t\} \leftarrow$ Eq. 2 (perform gradient descent on SimPONet's objective using $D_{\text{trn}}, D_{\text{syn}}$ while early stopping using Factual Error on $D_{\text{val}}$)
4: Return $\{\widehat{f}_t, \widehat{\mu}_t\}$ for $t = 0, 1$
```

We present the pseudocode for *SimPONet* in Alg. 1.

B.5 Theoretical Analysis

In this section, we present the proofs for our theoretical results.

B.5.1 Proof of Lemma 2

The CATE error is related to the factual and counterfactual error as: $\mathcal{E}_{\text{CATE}}^t \leq 2\mathcal{E}_F^t + 2\mathcal{E}_{CF}^t$

[3] https://www.ansys.com/applications/battery/battery-cell-and-electrode

Proof. We decompose the CATE error into factual and counterfactual estimation error as follows:

$$\begin{aligned}\mathcal{E}^t_{\text{CATE}} &= \int_{\mathbf{x}\in\mathcal{X}} [\tau_X(\mathbf{x},t) - \widehat{\tau_X}(\mathbf{x},t)]^2 P(\mathbf{x}|t) d\mathbf{x} = \int_{\mathbf{x}\in\mathcal{X}} [\tau(f_t(\mathbf{x})) - \widehat{\tau}(\widehat{f}_t(\mathbf{x}))]^2 P(\mathbf{x}|t) d\mathbf{x} \\ &= \int_{\mathbf{x}\in\mathcal{X}} [(\mu_1(f_t(\mathbf{x})) - \mu_0(f_t(\mathbf{x}))) - (\widehat{\mu}_1(\widehat{f}_t(\mathbf{x})) - \widehat{\mu}_0(\widehat{f}_t(\mathbf{x})))]^2 P(\mathbf{x}|t) d\mathbf{x} \\ &= \int_{\mathbf{x}\in\mathcal{X}} [(\mu_1(f_t(\mathbf{x})) - \widehat{\mu}_1(\widehat{f}_t(\mathbf{x}))) - (\mu_0(f_t(\mathbf{x})) - \widehat{\mu}_0(\widehat{f}_t(\mathbf{x})))]^2 P(\mathbf{x}|t) d\mathbf{x}\end{aligned}$$

Let $t' = 1 - t$ denote the *counterfactual* treatment. We can then rewrite the above expression as:

$$\mathcal{E}^t_{\text{CATE}} = \int_{\mathbf{x}\in\mathcal{X}} [(\mu_t(f_t(\mathbf{x})) - \widehat{\mu}_t(\widehat{f}_t(\mathbf{x}))) - (\mu_{t'}(f_t(\mathbf{x})) - \widehat{\mu}_{t'}(\widehat{f}_t(\mathbf{x})))]^2 P(\mathbf{x}|t) d\mathbf{x}$$

Now using the inequality $(a-b)^2 \le 2a^2 + 2b^2$, we can separate the *factual* and *counterfactual* terms:

$$\begin{aligned}\mathcal{E}^t_{\text{CATE}} &\le 2\int_{\mathbf{x}\in\mathcal{X}} [\mu_t(f_t(\mathbf{x})) - \widehat{\mu}_t(\widehat{f}_t(\mathbf{x}))]^2 P(\mathbf{x}|t) d\mathbf{x} + 2\int_{\mathbf{x}\in\mathcal{X}} [\mu_{t'}(f_t(\mathbf{x})) - \widehat{\mu}_{t'}(\widehat{f}_t(\mathbf{x}))]^2 P(\mathbf{x}|t) d\mathbf{x} \\ &= 2\mathcal{E}^t_F + 2\mathcal{E}^t_{CF}\end{aligned}$$

B.5.2 Recovery of f^S upto a diffeomorphic transformation

Lemma 4. *As $|D_{syn}| \to \infty$, contrastive training with paired covariates recovers $\widetilde{f}^S_t = h \circ f^S_t$ while paired outcome supervision recovers $\widetilde{\tau}^S = \tau^S \circ h^{-1}$ where h is a diffeomorphic transformation. Moreover, when the latent space $\mathcal{Z} \subset \mathbb{S}^{(n_z - 1)}$ (unit-norm hypersphere in $\mathbb{R}^{n_z}$), h is a rotation transform by Extended Mazur-Ulam Theorem as shown in [65] (Proposition 2).*

Proof. Theorem 4.4 of [57] shows that contrastive training with covariate pairs $\{\mathbf{x}^S_i(0), \mathbf{x}^S_i(1)\}$ recovers Z upto a diffeomorphic transformation h, i.e. for the simulator DGP our estimate $\hat{z}_i = \widetilde{f}^S(\mathbf{x}^S_i(t), t) = h(z_i) = h(f^S(\mathbf{x}^S_i(t), t), \forall t \in \mathcal{T}$. Moreover for unit-norm latent representations, $\mathcal{Z} \subset \mathbb{S}^{d_z - 1}$, [65] show that h is an isometric (norm-preserving) function and therefore, a rotation transform by an extension of Mazur-Ulam Theorem. Mazur-Ulam Theorem states that any smooth, invertible and isometric function is necessarily affine. Moreover, in our setting, the norm of z as well as $\hat{z}$ is always one and thus, h is necessarily a rotation. Therefore, we recover $\widetilde{f}^S = h \circ f^S$ upto a rotation of the true inverse map f^S with sufficient paired samples from the simulator.

Next, we recover $\widetilde{\tau}^S$ from the following minimisation:

$$\widetilde{\tau}^S = \underset{\widehat{\tau}^S}{\text{argmin}} \mathbb{E}_{\mathbf{x}^S} \left[\widehat{\tau}^S(\widetilde{f}^S(\mathbf{x}^S(t), t)) - \tau^S(f^S(\mathbf{x}^S(t), t))\right]^2 = \underset{\widehat{\tau}^S}{\text{argmin}} \mathbb{E}_z \left[\widehat{\tau}^S(h(z)) - \tau^S(z)\right]^2$$

The above optimization gives $\widetilde{\tau}^S = \tau^S \circ h^{-1}$ and hence we recover the CATE function τ^S for the simulator DGP composed with h^{-1}.

Proof of Lemma 3.

Assume τ is K_τ-Lipschitz, and $\widetilde{f}^S$ and $\widetilde{\tau}^S$ are estimates from the simulator DGP. Then, the CATE error on the estimates $\widehat{f}_t$ and $\widehat{\tau}$ admits the following bound:

$$\mathcal{E}^t_{\text{CATE}}(\widehat{f}_t, \widehat{\tau}) \le [12 d_{h(z)}(\widehat{\tau}, \widetilde{\tau}^S) + 12 K^2_\tau d_{\mathbf{x}|t}(\widehat{f}_t, \widetilde{f}^S_t) + 8\mathcal{E}^t_F] + [12 d_z(\tau, \tau^S) + 12 K^2_\tau d_{\mathbf{x}|t}(f_t, f^S_t)]$$

where $d_{\mathbf{x}|t}, d_z, d_{h(z)}$ are distance functions defined in section 2.

Proof. We now construct at upper bound on *counterfactual* error $\mathcal{E}^t_{CF}$ that relies on both observational data and simulator estimates to motivate the *SimPONet* objective:

$$\begin{aligned}
\mathcal{E}^t_{CF} &= \int_{\mathbf{x}\in\mathcal{X}} [\mu_{t'}(f_t(\mathbf{x})) - \widehat{\mu}_{t'}(\widehat{f}_t(\mathbf{x}))]^2 P(\mathbf{x}|t)d\mathbf{x} \\
&= \int_{\mathbf{x}\in\mathcal{X}} [(\mu_{t'}(f_t(\mathbf{x})) - \mu_t(f_t(\mathbf{x}))) - (\widehat{\mu}_{t'}(\widehat{f}_t(\mathbf{x})) - \widehat{\mu}_t(\widehat{f}_t(\mathbf{x}))) + \mu_t(f_t(\mathbf{x})) - \widehat{\mu}_t(\widehat{f}_t(\mathbf{x}))]^2 P(\mathbf{x}|t)d\mathbf{x} \\
&= \int_{\mathbf{x}\in\mathcal{X}} [(2\mathbf{1}_{t=0} - 1)\cdot(\tau(f_t(\mathbf{x})) - \widehat{\tau}(\widehat{f}_t(\mathbf{x}))) + \mu_t(f_t(\mathbf{x})) - \widehat{\mu}_t(\widehat{f}_t(\mathbf{x}))]^2 P(\mathbf{x}|t)d\mathbf{x}
\end{aligned}$$

Where $\mathbf{1}_{t=0} = 1$ when $t = 0$ and zero otherwise, and thus, $(2\mathbf{1}_{t=0} - 1) = \pm 1$ adjusting the sign of CATE terms. Now we utilise the inequality $(a+b+c)^2 \leq 3(a^2+b^2+c^2)$ to obtain:

$$\begin{aligned}
\mathcal{E}^t_{CF} &= \int_{\mathbf{x}\in\mathcal{X}} [(2\mathbf{1}_{t=0} - 1)\cdot(\tau(f_t(\mathbf{x})) - \widehat{\tau}(h \circ f_t(\mathbf{x})) + \widehat{\tau}(h \circ f_t(\mathbf{x})) - \widehat{\tau}(\widehat{f}_t(\mathbf{x}))) + \mu_t(f_t(\mathbf{x})) - \widehat{\mu}_t(\widehat{f}_t(\mathbf{x}))]^2 P(\mathbf{x}|t)d\mathbf{x} \\
&\leq 3\int_{\mathbf{x}\in\mathcal{X}} [\tau(f_t(\mathbf{x})) - \widehat{\tau}(h \circ f_t(\mathbf{x}))]^2 P(\mathbf{x}|t)d\mathbf{x} + 3\int_{\mathbf{x}\in\mathcal{X}} [\widehat{\tau}(h \circ f_t(\mathbf{x})) - \widehat{\tau}(\widehat{f}_t(\mathbf{x}))]^2 P(\mathbf{x}|t)d\mathbf{x} \\
&\quad + 3\int_{\mathbf{x}\in\mathcal{X}} [\mu_t(f_t(\mathbf{x})) - \widehat{\mu}_t(\widehat{f}_t(\mathbf{x}))]^2 P(\mathbf{x}|t)d\mathbf{x} \\
&= 3\int_{z\in\mathcal{Z}} [\tau(z) - \widehat{\tau}(h(z))]^2 P(z|t)dz + 3\int_{\mathbf{x}\in\mathcal{X}} [\widehat{\tau}(h(f_t(\mathbf{x}))) - \widehat{\tau}(\widehat{f}_t(\mathbf{x}))]^2 P(\mathbf{x}|t)d\mathbf{x} + 3\mathcal{E}^t_F
\end{aligned}$$

Here h denotes the unknown rotation transformation that relates the estimated simulator functions $(\widetilde{f}^S, \widetilde{\tau}^S)$ with the ground-truth simulator functions (f^S, τ^S) as shown in Lemma 4. Let K_τ be the Lipschitz constant for $\widehat{\tau}$. We can bound the second term in the above expression as follows:

$$\begin{aligned}
\mathcal{E}^t_{CF} &\leq 3\int_{z\in\mathcal{Z}} [\tau(z) - \widehat{\tau}(h(z))]^2 P(z|t)dz + 3K_\tau^2 \int_{\mathbf{x}\in\mathcal{X}} ||h(f_t(\mathbf{x})) - \widehat{f}_t(\mathbf{x})||^2 P(\mathbf{x}|t)d\mathbf{x} + 3\mathcal{E}^t_F \\
&= 3d_z(\tau, \widehat{\tau} \circ h) + 3K_\tau^2 d_{\mathbf{x}|t}(h \circ f_t, \widehat{f}_t) + 3\mathcal{E}^t_F
\end{aligned}$$

Now we can add and subtract simulator function estimates to bound the two distance terms as follows:

$$\begin{aligned}
\mathcal{E}^t_{CF} &\leq 3\int_{z\in\mathcal{Z}} [\tau(z) - \tau^S(z) + \tau^S(z) - \widehat{\tau}(h(z))]^2 P(z|t)dz \\
&\quad + 3K_\tau^2 \int_{\mathbf{x}\in\mathcal{X}} ||h(f_t(\mathbf{x})) - h(f_t^S(\mathbf{x})) + h(f_t^S(\mathbf{x})) - \widehat{f}_t(\mathbf{x})||^2 P(\mathbf{x}|t)d\mathbf{x} + 3\mathcal{E}^t_F \\
&\leq 6\int_{z\in\mathcal{Z}} [\tau(z) - \tau^S(z)]^2 P(z|t)dz + 6\int_{z\in\mathcal{Z}} [\tau^S(z) - \widehat{\tau}(h(z))]^2 P(z|t)dz \\
&\quad + 6K_\tau^2 \int_{\mathbf{x}\in\mathcal{X}} ||h(f_t(\mathbf{x})) - h(f_t^S(\mathbf{x})||^2 P(\mathbf{x}|t)d\mathbf{x} + 6K_\tau^2 \int_{\mathbf{x}\in\mathcal{X}} ||h(f_t^S(\mathbf{x})) - \widehat{f}_t(\mathbf{x})||^2 P(\mathbf{x}|t)d\mathbf{x} + 3\mathcal{E}^t_F \\
&= 6d_z(\tau, \tau^S) + 6d_z(\widehat{\tau} \circ h, \tau^S) + 6K_\tau^2 d_{\mathbf{x}|t}(h \circ f_t, h \circ f_t^S) + 6K_\tau^2 d_{\mathbf{x}|t}(\widehat{f}_t, h \circ f_t^S) + 3\mathcal{E}^t_F
\end{aligned}$$

Now, using Lemma 4, we can rewrite $\tau^S = \widetilde{\tau}^S \circ h$ in the second term. Thus, $d_z(\widehat{\tau} \circ h, \tau^S) = d_z(\widehat{\tau} \circ h, \widetilde{\tau}^S \circ h)$. Now making use of Definition 2, we can rewrite this as $d_{h(z)}(\widehat{\tau}, \widetilde{\tau}^S)$ which is a distance function defined on the space of rotated latents $h(z)$. We also rewrite $h \circ f^S$ as $\widetilde{f}^S$ in the fourth term.

Moreover, $d_{\mathbf{x}|t}(h \circ f_t, h \circ f_t^S) = d_{\mathbf{x}|t}(f_t, f_t^S)$ since h is a rotation transform and preserves the distance between any two vectors. Thus, $||f_t(\mathbf{x}) - f_t^S(\mathbf{x})||_2 = ||h \circ f_t(\mathbf{x}) - h \circ f_t^S(\mathbf{x})||_2$. Combining these results, we can evaluate the above bound to the following:

$$\mathcal{E}_{CF}^{t} \leq [6d_{h(z)}(\widehat{\tau},\widetilde{\tau}^{S})+6K_{\tau}^{2}d_{\mathbf{x}|t}(\widehat{f}_t,\widetilde{f}_t^{S})+3\mathcal{E}_{F}^{t}]+[6d_{z}(\tau,\tau^{S})+6K_{\tau}^{2}d_{\mathbf{x}|t}(f_t,f_t^{S})]$$

B.6 Linear DGP Derivation

We derive expressions for CATE estimates $\widehat{\tau_X}(\mathbf{x},t)$ as well as $\mathcal{E}_{\text{CATE}}^{t}$ for each of our proposed estimators in the linear setting below. Note that ground truth CATE $\tau_X(\mathbf{x},t)=\mathbf{x}\boldsymbol{R}_t^{-1}(w_1-w_0)$. We consider factual treatment $t=1$ to illustrate the errors.

B.6.1 SimOnly

For SimOnly, we use $\hat{\boldsymbol{R}}_t^{-1}=\boldsymbol{S}_t^{-1}$ and $\hat{w}_t=w_t^S$ which are obtained by training on simulator data. Thus, the CATE estimate $\widehat{\tau_X}(\mathbf{x}^*,t)=\mathbf{x}^*\boldsymbol{S}_t^{-1}(w_1^S-w_0^S)$. The CATE error on a sample $\mathbf{x}^*$, with treatment $t=1$ is given by $[\widehat{\tau_X}(\mathbf{x}^*,1)-\tau_X(\mathbf{x}^*,1)]^2=[(\mathbf{x}^*(\boldsymbol{S}_1^{-1}(w_1^S-w_0^S)-\boldsymbol{R}_1^{-1}(w_1-w_0))]^2$

B.6.2 RealOnly

For RealOnly, the factual objective $\mathcal{E}_F^t=||\mathbf{x}\hat{\boldsymbol{R}}_t^{-1}\hat{w}_t-y||_2^2=||\mathbf{x}\hat{\boldsymbol{R}}_t^{-1}\hat{w}_t-\mathbf{x}\boldsymbol{R}_t^{-1}w_t||_2^2$. Thus, the closed form solution of the estimator $\hat{\boldsymbol{R}}_t^{-1}\hat{w}_t=\boldsymbol{R}_t^{-1}w_t,\forall t\in\mathcal{T}$. Since we can't decouple the terms $\hat{\boldsymbol{R}}_t^{-1}$ and $\hat{w}_t$, the CATE estimate is given by $\widehat{\tau_X}(\mathbf{x}^*,t)=\mathbf{x}^*\hat{\boldsymbol{R}}_1^{-1}\hat{w}_1-\mathbf{x}^*\hat{\boldsymbol{R}}_0^{-1}\hat{w}_0=\mathbf{x}^*\boldsymbol{R}_1^{-1}w_1-\mathbf{x}^*\boldsymbol{R}_0^{-1}w_0$. CATE error on sample $\mathbf{x}^*$ with treatment $t=1$ is given by $[\widehat{\tau_X}(\mathbf{x}^*,1)-\tau_X(\mathbf{x}^*,1)]^2=[(\mathbf{x}^*\boldsymbol{R}_1^{-1}w_1-\mathbf{x}^*\boldsymbol{R}_0^{-1}w_0)-\mathbf{x}\boldsymbol{R}_1^{-1}(w_1-w_0)]^2=[\mathbf{x}(\boldsymbol{R}_1^{-1}-\boldsymbol{R}_0^{-1})w_0]^2$

B.6.3 Real$_\mu$Sim$_f$

For Real$_\mu$Sim$_f$, we first set $\hat{\boldsymbol{R}}_t^{-1}=\boldsymbol{S}_t^{-1}$ which is obtained by training on simulator data. Next, we train $\hat{w}_t$ on the factual objective: $||\mathbf{x}\hat{\boldsymbol{R}}_t^{-1}\hat{w}_t-\mathbf{x}\boldsymbol{R}_t^{-1}w_t||_2^2=||\mathbf{x}\boldsymbol{S}_t^{-1}\hat{w}_t-\mathbf{x}\boldsymbol{R}_t^{-1}w_t||_2^2$. This, gives us a closed form solution for the minimising $\hat{w}_t=\boldsymbol{S}_t\boldsymbol{R}_t^{-1}w_t$. The CATE estimate $\widehat{\tau_X}(\mathbf{x}^*,t)=\mathbf{x}^*\boldsymbol{S}_t^{-1}(\hat{w}_1-\hat{w}_0)=\mathbf{x}^*\boldsymbol{S}_t^{-1}(\boldsymbol{S}_1\boldsymbol{R}_1^{-1}w_1-\boldsymbol{S}_0\boldsymbol{R}_0^{-1}w_0)$. Fixing treatment $t=1$, this simplifies further: $\widehat{\tau_X}(\mathbf{x}^*,1)=\mathbf{x}^*\boldsymbol{S}_1^{-1}(\boldsymbol{S}_1\boldsymbol{R}_1^{-1}w_1-\boldsymbol{S}_0\boldsymbol{R}_0^{-1}w_0)=\mathbf{x}^*(\boldsymbol{R}_1^{-1}w_1-\boldsymbol{S}_1^{-1}\boldsymbol{S}_0\boldsymbol{R}_0^{-1}w_0)$. CATE Error is given by $[\widehat{\tau_X}(\mathbf{x}^*,1)-\tau_X(\mathbf{x}^*,1)]^2=[\mathbf{x}^*(\boldsymbol{R}_1^{-1}w_1-\boldsymbol{S}_1^{-1}\boldsymbol{S}_0\boldsymbol{R}_0^{-1}w_0)-\mathbf{x}^*\boldsymbol{R}_1^{-1}(w_1-w_0)]^2=[\mathbf{x}^*\boldsymbol{R}_1^{-1}w_0-\mathbf{x}^*\boldsymbol{S}_1^{-1}\boldsymbol{S}_0\boldsymbol{R}_0^{-1}w_0]^2=[\mathbf{x}^*(\boldsymbol{R}_1^{-1}-\boldsymbol{S}_1^{-1}\boldsymbol{S}_0\boldsymbol{R}_0^{-1})w_0]^2$

B.6.4 *SimPONet*

We train both $\hat{\boldsymbol{R}}_t^{-1},\hat{w}_t$ on the following objective jointly:

$$\mathcal{L}(\{\hat{\boldsymbol{R}}_t^{-1},\hat{w}_t\}_{t=0,1})=\left[\sum_{t=0,1}||\mathbf{x}\hat{\boldsymbol{R}}_t^{-1}\hat{w}_t-\mathbf{x}\boldsymbol{R}_t^{-1}w_t||_2^2+\lambda_f\sum_{t=0,1}||\mathbf{x}\hat{\boldsymbol{R}}_t^{-1}-\mathbf{x}\boldsymbol{S}_t^{-1}||_F^2+\lambda_\tau||\mathbf{z}(\hat{w}_1-\hat{w}_0)-\mathbf{z}(w_1-w_0)||_2^2\right]$$

Here, $\mathbf{z}=\mathbf{x}_{t'}^S\boldsymbol{S}_{t'}^{-1}$ are the latents for simulated covariates $\mathbf{x}_{t'}^S$ (which are identifiable from D_{syn}). Due to the joint nature of this optimisation, it is not possible to derive closed form solutions for the optimum. However, one can compuet gradients of the objective with respect to $\hat{\boldsymbol{R}}_t^{-1}$ and $\hat{w}_t$ separately. This, gives us an alternating minimisation algorithm with closed form updates.

$$\begin{aligned}\frac{\partial\mathcal{L}}{\partial\hat{\boldsymbol{R}}_t^{-1}}&=\frac{\partial}{\partial\hat{\boldsymbol{R}}_t^{-1}}\left[||\mathbf{x}\hat{\boldsymbol{R}}_t^{-1}\hat{w}_t-y||_2^2+\lambda_f||\mathbf{x}\hat{\boldsymbol{R}}_t^{-1}-\mathbf{x}\boldsymbol{S}_t^{-1}||_F^2\right]\\&=2\mathbf{x}^T\mathbf{x}\hat{\boldsymbol{R}}_t^{-1}(\hat{w}_t\hat{w}_t{}^T+\lambda_f\boldsymbol{I})-2\mathbf{x}^Ty\hat{w}_t+-2\lambda_f\mathbf{x}^T\mathbf{x}\boldsymbol{S}_t^{-1}\end{aligned}$$

Setting the derivative to zero, we obtain the following update rule:

$$\hat{\boldsymbol{R}}_t^{-1}\leftarrow(\mathbf{x}^\dagger y\hat{w}_t+\lambda_f\boldsymbol{S}_t^{-1})\cdot(\hat{w}_t\hat{w}_t{}^T+\lambda_f\boldsymbol{I})^{-1}$$

where $\mathbf{x}^\dagger=(\mathbf{x}^T\mathbf{x})^{-1}\mathbf{x}^T$ is the pseudoinverse of $\mathbf{x}$.

$$\begin{aligned}\frac{\partial\mathcal{L}}{\partial\hat{w}_t}&=\frac{\partial}{\partial\hat{w}_t}\left[||\mathbf{x}\hat{\boldsymbol{R}}_t^{-1}\hat{w}_t-y||_2^2+\lambda_\tau||\mathbf{z}(\hat{w}_t-\hat{w}_{t'})-(y_1^S-y_0^S)||_2^2\right]\\&=2(\hat{z}^T\hat{z})\hat{w}_t-2\hat{z}^Ty+2\lambda_\tau(z^Tz\hat{w}_t-z^T(z\hat{w}_{t'}+(y_1^S-y_0^S)))\\&=2[(\hat{z}^T\hat{z})+\lambda_\tau(z^Tz)]\hat{w}_t-2(\hat{z}^Ty+\lambda_\tau z^T(z\hat{w}_{t'}+(y_1^S-y_0^S)))\end{aligned}$$

Where $\hat{z} = \mathbf{x}\hat{\boldsymbol{R}}_t^{-1}$. Setting the derivative to zero, we obtain the following update rule:

$$\hat{w_t} \leftarrow ((\hat{z}^T\hat{z}) + \lambda_\tau(z^T z))^{-1} \cdot (\hat{z}^T y + \lambda_\tau z^T(z\hat{w_{t'}} + (y_1^S - y_0^S)))$$

For *SimPONet*, we perform alternating updates of $\hat{w}_t$ and $\hat{\boldsymbol{R}}_t^{-1}$ fixing the other estimate.

B.7 Summary of Datasets

IHDP. The Infant Health and Development Program (IHDP) is a randomized controlled trial designed to assess the impact of physician home visits on the cognitive test performance of premature infants. The dataset exhibits selection bias due to the deliberate removal of non-random subsets of treated individuals from the training data. Since outcomes are observed for only one treatment, we generate both observed and counterfactual outcomes using a synthetic outcome generation function based on the original covariates for both treatments, making the dataset suitable for causal inference.

The IHDP dataset includes 747 subjects and 25 variables. While the original dataset discussed in [53] had 1000 versions, our work uses a smaller version with 100 iterations, aligning with the CATENets benchmark. Each version varies in the complexity of the assumed outcome generation function, treatment effect heterogeneity, etc. As outlined in [13], reporting the standard deviation of performance across the 100 different seeds is inappropriate. Therefore, we calculate p-values through paired t-tests between our method (*SimPONet*) and other baseline methods, using *SimPONet* as the baseline for all experiments. We follow setting D of the IHDP dataset as mentioned in [11] where response surfaces are modified to suppress the extremely high variance of potential outcomes in certain versions of the IHDP dataset.

ACIC. The Atlantic Causal Inference Conference (ACIC) competition dataset (2016)[4] consists of 77 datasets, all containing the same 58 covariates derived from the Collaborative Perinatal Project. Each dataset simulates binary treatment assignments and continuous outcome variables, with variations in the complexity of the treatment assignment mechanism, treatment effect heterogeneity, the ratio of treated to control observations, overlap between treatment and control groups, dimensionality of the confounder space, and the magnitude of the treatment effect.

All datasets share common characteristics, such as independent and identically distributed observations conditional on covariates, adherence to the ignorability assumption (selection on observables with all confounders measured and no hidden bias), and the presence of non-true confounding covariates. Of the 77 datasets, we selected a subset of three: versions 2, 7, and 26, aligning with the CATENets benchmark. These versions present non-linear covariate-to-outcome relationships and maximum variability in treatment effect heterogeneity. Version 2, notably, exhibits no heterogeneity, meaning the treatment effect is constant across all individuals. However, accurately estimating outcome differences even for this version is challenging due to the inherent noise in potential outcome realizations in the dataset.

[4] `https://jenniferhill7.wixsite.com/acic-2016/competition`

B.8 Table of Symbols

Symbol	Definition
X	Real post-treatment covariates: Random Variable
Y	Real outcomes: Random Variable
X^S	Simulator post-treatment covariates: Random Variable
Y^S	Simulator outcomes: Random Variable
T	Treatment: Random Variable
Z	Latent (unobserved) pre-treatment representations: Random Variable
D_{trn}	Observational training dataset from Real DGP
D_{syn}	Counterfactual dataset from Simulator DGP
D_{tst}	Test dataset from Real DGP
$\mathbf{x},\mathbf{x}^S,z,t,y,y^S$	Realisations of random variables X,X^S,Z,T,Y,Y^S respectively
$\mathcal{X}$	Space of post-treatment covariate values: Set
$\mathcal{T}$	Space of treatment values: Set $=\{0,1\}$
$\mathcal{Z}$	Space of latents: Set
$\mathcal{Y}$	Space of outcomes: Set
n_z,n_x	Dimensions of vector spaces in which $\mathcal{Z},\mathcal{X}$ lie
$Y_i(t)$	Potential outcome for i^{th} unit under treatment t
$X_i(t)$	Potential post-treatment covariate for i^{th} unit under treatment t
g_t	Mapping from $\mathcal{Z} \mapsto \mathcal{X}$, transforms latents to real post-treatment covariates under t
g_t^S	Mapping from $\mathcal{Z} \mapsto \mathcal{X}$, transforms latents to simulated post-treatment covariates under t
f_t	Mapping from $\mathcal{X} \mapsto \mathcal{Z}$, transforms real post-treatment covariates under t to latents
f_t^S	Mapping $\mathcal{X} \mapsto \mathcal{Z}$, transforms simulated post-treatment covariates under t to latents
P_Z	Probability distribution of latents Z
μ_t	Outcome function for real data under t
μ_t^S	Outcome function for simulated data under t
τ	Conditional Average Treatment Effect for real data, $\mu_1 - \mu_0$, Mapping $\mathcal{Z} \mapsto \mathcal{Y}$
τ^S	Conditional Average Treatment Effect for simulated data, $\mu_1^S - \mu_0^S$, Mapping $\mathcal{Z} \mapsto \mathcal{Y}$
$\circ$	Composition of functions
$\tau_X(\mathbf{x},t)$	Conditional Average Treatment Effect for real data, $\tau \circ f_t(\mathbf{x})$, Mapping $\mathcal{X} \times \mathcal{T} \mapsto \mathcal{Y}$
$\tau_X^S(\mathbf{x}^S,t)$	Conditional Average Treatment Effect for simulated data, $\tau^S \circ f_t^S(\mathbf{x}^S)$, Mapping $\mathcal{X} \times \mathcal{T} \mapsto \mathcal{Y}$
h	Diffeomorphic transformation, arises due to contrastive learning
$\mathbb{S}^d$	Unit-norm hypersphere of dimension d, Subset of $\mathbb{R}^{(d+1)}$
$d_{\mathbf{x} \mid t}$	Expected squared-distance between two functions on $P(X \mid T)$, see Section 2 for definition
d_z	Expected squared-distance between two functions on P_Z, see Section 2 for definition
$d_{h(z)}$	d_z under transformation h on z, see Section 2 for definition
sim(●,●)	Cosine similarity
$\widehat{f}_t$	Estimate for f_t
$\widehat{f}_t^S$	Estimate for f_t^S
$\widehat{\mu}_t$	Estimate for μ_t
$\widehat{\mu}_t^S$	Estimate for μ_t^S
$\widetilde{f}_t^S$	Estimate for f_t^S recovered from contrastive learning
$\widetilde{\mu}_t^S$	Estimate for μ_t^S on recovering Simulator DGP
$\mathcal{E}_{\text{CATE}}$	CATE estimation error
$\mathcal{E}_{\text{CATE}}^t$	CATE estimation error on covariates $\mathbf{x}$ under treatment t
$\mathcal{E}_F^t$	Factual error on treatment t samples
$\mathcal{E}_{CF}^t$	Counterfactual error on treatment t samples
K_μ	Lipschitz constant for $\mu_t,\widehat{\mu}_t$
K_τ	Lipschitz constant for $\tau,\widehat{\tau}$
K_{μ^S}	Lipschitz constant for $\mu_t^S,\widehat{\mu}_t^S,\widetilde{\mu}_t^S$
K_{τ^S}	Lipschitz constant for $\tau^S,\widehat{\tau}^S,\widetilde{\tau}^S$
$\boldsymbol{R}_t$	g_t for linear DGP: Matrix
$\boldsymbol{S}_t$	g_t^S for linear DGP: Matrix
w_t	μ_t for linear DGP: Vector
w_t^S	μ_t^S for linear DGP: Vector
w_τ	τ for linear DGP: Vector
w_τ^S	τ^S for linear DGP: Vector

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Yes, we have listed all the claims and contributions clearly.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We explained limitations in Section 5.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: We have provided detailed proofs in the Appendix. Each Lemma statements clearly mention the assumptions under which the theoretical results hold.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We have uploaded the code along with our submission. All our datasets are publicly available. Our code encompasses scripts that can be fired to produce an XL sheet with the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: All used public datasets and uploaded the code.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We did not perform hyperparameter tuning. We fixed all the hyperparameters to CATENets defaults [13]. We adjusted just one hyperparameter for ACIC-2 dataset. We provide a clear justification for why that change was required.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We report std. deviations and p-values wherever appropriate.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

Justification: All our experiments are run on standard benchmarks and are not compute intensive.

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We conform to the Neurips Code of Ethics.

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: We use existing benchmarks and so our work does not have any negative impact.

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: We do not use such high risk data.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We will release the code and make it public after acceptance.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer:[NA]

    Justification: No new assets are introduced.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer:[NA]

    Justification: No crowdsourcing.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: No crowdsourcing.