Representation Learning for Sample-Efficient CATE Estimation by Leveraging Multiple Outcomes

Maitreyi Swaroop[1]*, Shikha Bhat[1], Samantha Rodriguez[2], Tamar Krishnamurti[2], Bryan Wilder[1]

[1]Machine Learning Department, Carnegie Mellon University
[2]Department of General Internal Medicine, University of Pittsburgh

Abstract

Estimating conditional average treatment effects (CATE) enables efficient targeting of interventions, but many applications have limited experimental samples, making it difficult to estimate heterogeneous effects from high-dimensional covariates. In such events, policy-makers and medical practitioners often succumb to the curse of dimensionality or apply off-the-shelf dimension reduction methods that may not preserve treatment heterogeneity. Yet these domains often come with large historical datasets measuring a wide range of outcomes – a source of supervision that is rarely exploited in practice. Following causal representation learning, we hypothesize that such domains with high-dimensional covariates have lower-dimensional underlying dynamics. We can thus leverage the diverse outcomes measured in historical data to learn a lower-dimensional representation of the covariates. Theoretically, we prove that when the auxiliary outcomes satisfy a set of surrogacy conditions and the representation retains relevant covariate information, the original CATE is identified when the high-dimensional covariates are replaced by the learned representation. Combined with existing dimension-dependent rates for CATE estimation, the result implies greater sample-efficiency on the same experimental sample. Additionally, we characterize the bias-variance tradeoff when the assumptions do not hold perfectly, and show that the representation-based estimator can still achieve lower error when the reduction in estimator variance outweighs the bias due to compression. Empirically, we evaluate the method on synthetic data and semi-synthetic medical data.

1 Introduction

Treatment effects often vary substantially across individuals, and identifying those who benefit most enables efficient allocation of limited resources [4]. Estimating the conditional average treatment effect (CATE) is therefore central to designing well-targeted interventions. However, CATE estimation faces a fundamental challenge: randomized experiments are often small-scale, while the covariates needed to capture meaningful heterogeneity are high-dimensional. This curse of dimensionality often leads researchers to abandon data-driven heterogeneity discovery in favor of manually specifying subgroups of interest [2, 17, 58]. More recently, pretraining has emerged as a solution to sample-limited tasks [35, 12, 51]. The insight to learn representations from large, related datasets that transfer effectively to specific use cases has transformed many areas of machine learning [10]. Yet, application to treatment effect estimation remains limited [44, 41, 39], and often confined to specialized settings where covariates are images or text [59, 22, 29].

Causal representation learning (CRL) [53] studies how high-dimensional observations arise from lower-dimensional causal variables, often with the goal of recovering those variables up to an

*Corresponding author: mswaroop@andrew.cmu.edu

Preprint.

equivalence class. We build upon this perspective, but use representation learning to aid statistical estimation of causal quantities, instead of identifying the ground-truth causal variables. Recent work on actionable prediction [40] supports the view that effective intervention requires measuring latent states that determine treatment heterogeneity (rather than simply optimizing single-outcome prediction). For example, effects of a mental health intervention may be driven by social determinants (access to family or community support), biological factors (genetic predisposition or chronic health conditions), and other stressors (food security or employment status). While these states are often not directly observed, EHR data provides rich observational samples where these latent factors are (noisily) reflected across many measured covariates and outcome indicators. The role of these latent factors in driving health outcomes is well-studied in the healthcare and social sciences literature [9, 49, 1].

We propose that the diverse outcomes often measured in administrative or historical data can help uncover this latent structure and enable more accurate estimation of heterogeneous treatment effects. Theoretically, we use the surrogacy framework of Athey et al. [3] to prove that when the auxiliary outcomes satisfy a set of surrogacy conditions, the learned representation will suffice to capture heterogeneous treatment effects on the target outcome. To realize this, we learn a bottleneck representation $\phi(X)$ on large-scale historical data which captures information between the covariates X and k auxiliary outcomes $(Y^{(1)},\ldots,Y^{(k)})$. These outcomes serve as indicators of underlying dynamics that predict heterogeneity in the target outcome Y^*. We then use data from a randomized experiment of the intervention of interest to learn treatment effects on a target outcome Y^* as a function of the lower-dimensional $\phi(X)$ instead of the original X. Combining our identification result with the CATE estimation error rates of Kennedy [33], our method provides gains in sample-efficiency as CATE estimation rates degrade with covariate dimension.

We emphasize that we do not require surrogacy conditions to be exactly satisfied for our method to be useful (just as pretraining yields sample efficiency gains for many tasks even if the pretrained representations are not perfectly adapted to the target domain). When the assumptions do not hold perfectly, the estimand coarsens to the average treatment effect among units with covariates sharing the same representation. The total error in CATE estimation admits the familiar bias-variance decomposition of squared bias (due to coarsening), and variance (due to the finite experimental sample). In practice, the total error is often minimized by accepting some nonzero bias in this tradeoff. Thus even when the conditions only hold approximately, we find that our lower dimensional representation achieves lower total error than the raw covariates.

Empirically, we demonstrate our method on synthetic and semi-synthetic medical data, finding substantial improvements in sample efficiency for CATE estimation.

The paper is organized as follows: Section 2 discusses related work from causal inference literature. Section 3 presents the formal problem setup and introduces notation. Section 4 provides the main identification result of the CATE estimator using the learned representation. Section 5 provides empirical evaluation of our method with both synthetic and real datasets. Section 6 concludes.

2 Related Work

Heterogeneous treatment effect estimation CATE estimation has advanced through meta-learner frameworks: the S-learner and T-learner estimate potential outcomes, the X-learner [38] is particularly effective for unbalanced treatment groups, and the R-learner [43] builds on Double Machine Learning [16] for robustness to nuisance estimation errors. Parallel to these frameworks, causal forests [60], extending causal trees [2], are standard for non-parametric CATE estimation. However, in high-dimensional settings with limited experimental samples, convergence rates degrade [19], necessitating methods that learn lower-dimensional representations that extract relevant structure from the original covariates. Existing remedies such as sufficient dimension reduction [23], or integrating out nuisance covariates over a pre-selected subset [21], require manual selection or linearity assumptions, motivating the need for methods that automatically learn lower-dimensional representations.

Causal representation learning Work at the intersection of representation learning and causal inference falls into two categories. The first learns representations for treatment effect estimation from observational data: balanced representations to reduce the covariate shift between treatment groups [30, 54], latent variable models for unobserved confounding and selection bias [42, 24, 63], shared

representations between propensity and outcome models [56], and energy-based representations [64]. However, these methods operate within a single dataset and do not leverage the multi-outcome structure available in historical data. The second category (the more common interpretation of causal representation learning) aims to recover latent causal variables from high-dimensional observations [53]. This includes identifiable nonlinear ICA [26, 34], and work on identifying latent causal structure under interventions [11]. These methods establish conditions under which latent variables can be recovered, but do not address treatment effect estimation.

Leveraging auxiliary information Our approach leverages the availability of multiple auxiliary outcomes to uncover latent structure. While this setting naturally encompasses the *surrogacy framework* where intermediate outcomes mediate the effect of treatment on a long-term target [50, 3], our motivation differs. Prior surrogacy literature addresses two main problems, (1) *identification* - where long-term outcomes are unobserved and surrogates enable estimation of otherwise unidentifiable effects [3, 28, 13], including recent extensions to heterogeneous treatment effects [14], and (2) *variance reduction*, where surrogates improve efficiency when the primary outcome is partially observed [32]. However, none of this work uses surrogacy structure to learn representations that transfer to new experiments for sample-efficient CATE estimation. Our method applies even when the primary outcome is fully observed, and where identification is not a concern but sample efficiency is.
Related directions include *data fusion*, which combines experimental data with larger observational datasets to address selection bias or increase statistical power [62], and *transfer learning* for causal inference, which focuses on generalizing treatment effects across populations with different covariate distributions [46, 37, 8]. Our setting differs from both as we do not assume that the treatment of interest appears in the historical data at all, nor that the experimental and historical samples come from different populations. Instead, we assume rich historical data from the same population, collected before the intervention was available.

Our contribution The methods discussed above address concerns about the validity of the observational or experimental sample, such as unmeasured confounding, selection bias, or covariate shift between populations. Our goal differs in that we target the setting where the experimental design is internally valid but statistically underpowered due to limited samples and high-dimensional covariates, with experimental and historical covariates drawn from the same population. We use multi-outcome historical data to learn a representation that helps transfer this structural knowledge to the experimental sample, thereby reducing the effective dimensionality of the inference problem. We thus enable accurate CATE estimation on small experimental samples where standard methods would suffer from the curse of dimensionality.

3 Problem Formulation

We consider treatment effect estimation in a setting where experiments are expensive or limited in sample size. To overcome this limitation, we leverage a larger historical dataset that captures the underlying outcome dynamics, to learn a lower dimensional representation of the covariates.

Datasets We have access to two datasets (or *populations* $P \in \{H, E\}$):

- *Historical dataset ($P = H$):* A large sample (size N_H) collected prior to the experiment. It comprises covariates $X \in \mathbb{R}^d$, k auxiliary outcomes $S = (Y^{(1)}, \ldots, Y^{(k)})$ (*surrogates*, following the surrogacy literature), and a target outcome Y^*.
- *Experimental dataset ($P = E$):* A relatively smaller sample (size N_E) collected from a randomized controlled trial (RCT). It comprises covariates X, assigned treatments T, and auxiliary outcomes S.

We do not assume that units were assigned a treatment in the historical dataset. For example, administrative data may predate an intervention, as is common in policy and health settings. We also do not require the primary outcome Y^* to be immediately observable in the experimental sample - for instance, when Y^* is a long-term outcome and treatment effects must be estimated before Y^* can be measured. Under Assumptions 3.4-3.5, observing Y^* in the experimental sample is not required for identification. However, when Y^* is observed, our reliance on surrogacy assumptions weakens. The conditions are sufficient to ensure that $\phi(X)$ captures the full CATE, but are not necessary to estimate a valid causal effect on Y^*.

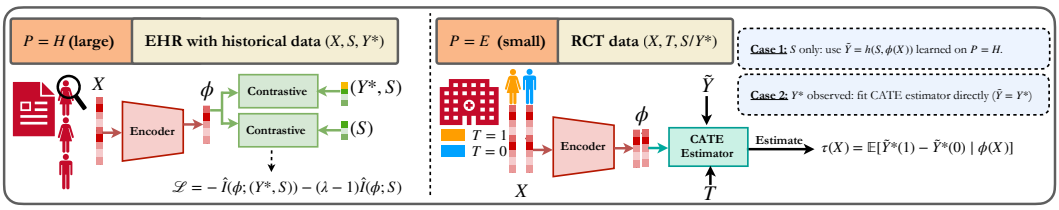


Figure 1: **Method overview.** On historical data we train ϕ via contrastive lower bounds on $I(\phi; (Y^*, S))$ and $I(\phi; S)$ (Equation 6). On the RCT we feed $(\phi(X), T, \tilde{Y})$ to a CATE meta-learner, with $\tilde{Y} = h(S, \phi(X))$ when Y^* is unobserved (Case 1) or $\tilde{Y} = Y^*$ when it is observed (Case 2).

Estimand Our goal is to learn a lower-dimensional representation $\phi : \mathbb{R}^d \to \mathbb{R}^m$ of the covariates X where $m \ll d$, using the observations $(X_i, S_i, Y_i^*, P_i = H)_{i=1}^{N_H}$. Let $Y(t)$ denote the potential outcome in treatment $t \in \{0, 1\}$ (for both the target and the auxiliary outcomes). We will then use ϕ to estimate the CATE on the target outcome in the experimental dataset $(X_i, T_i, S_i(T_i), P_i = E)_{i=1}^{N_E}$. Our target estimand is

$$\tau(x) = \mathbb{E}[Y^*(1) - Y^*(0) | X = x, P = E] \tag{1}$$

which our method will learn by estimating

$$\mathbb{E}[Y^*(1) - Y^*(0) | \phi(X), P = E]. \tag{2}$$

Assumptions We organize our assumptions from the causal inference literature into two categories: (1) standard identification assumptions, and (2) assumptions from the surrogacy framework.

Standard causal inference assumptions. (Standard in treatment effect estimation literature).

Assumption 3.1 (Stable Unit Treatment Value Assumption (SUTVA))**.** We assume that the SUTVA [52] holds, ensuring well-defined potential outcomes.

Assumption 3.2 (Unconfoundedness/Ignorability)**.**

$$(X, Y^{(j)}(1), Y^{(j)}(0)) \perp T | P = E \quad \text{for all outcomes } j \text{ (auxiliary and target).}$$

This assumption is satisfied by design since $P = E$ is a randomized controlled trial. We adopt unconditional ignorability as our primary assumption. The extension to stratified randomization, where ignorability holds conditional on a known subset $W \subset X$, is discussed in Appendix B.2.

Assumption 3.3 (Overlap)**.** (i) $0 < \Pr[T = 1 \mid X, P = E] < 1$ for all X.

(ii) The support of the experimental covariates is contained within the historical support, i.e., $\text{supp}(P_E(X)) \subseteq \text{supp}(P_H(X))$.

Surrogacy assumptions. (From the surrogacy framework of Athey et al. [3]).

Assumption 3.4 (Surrogacy)**.** $T \perp Y^* \mid S, X, P = E$.

Assumption 3.5 (Comparability)**.** $Y^* \perp P \mid S, X$

Surrogacy requires that S fully mediates the treatment effect on Y^*, while *comparability* ensures the conditional distribution of Y^* given (S, X) is consistent across populations.

4 Methodology

In this section, we establish the conditions under which a learned representation $\phi(X)$ suffices to identify heterogeneous treatment effects, and derive a training objective that encourages the representation to retain outcome-relevant information while discarding irrelevant variation in X.

Before specifying what ϕ must satisfy, we observe that ignorability is preserved under any compression of X (see Lemma B.1 in Appendix B.1 for the formal statement and proof). This distinguishes our setting from observational representation learning [30, 54], where ϕ must preserve confounders to maintain ignorability, whereas randomization removes this requirement in our setting. We now turn to the question of what additional structure ϕ must capture for sufficiency.

4.1 Representation requirements

The following assumptions specify the conditions a representation must satisfy to preserve all the information relevant for CATE estimation. We use this to motivate our training objective (Section 4.3).

Assumption 4.1 (Sufficiency)**.** (i) $Y^* \perp X | \phi(X), S, P = H$

(ii) $S \perp X | \phi(X), T, P = E$

Condition (i) requires that, given representation $\phi(X)$ and surrogates S, the full covariates X provide no additional information about the target outcome Y^*. Condition (ii) requires that, given $\phi(X)$ and treatment T, the covariates provide no additional information about the surrogates. Together, these ensure $\phi(X)$ captures all covariate information relevant for predicting treatment effect heterogeneity.

Remark 4.2 (Testability and proxy objective). Condition (i) is testable on historical data; Condition (ii) must hold for all treatment levels. Since historical data contains only controls ($T = 0$), we cannot enforce this condition for treated units during representation learning; we instead optimize a proxy objective (Section 4.3). The full condition can be verified post-hoc on experimental data, though statistical power is limited by the smaller experimental sample size.

4.2 Identification result

Our main theoretical contribution shows that under Assumptions 3.1- 4.1, the representation $\phi(X)$ learned from historical data suffices to identify the X-conditional heterogeneous treatment effects.

Theorem 4.3 (Identification of CATE under the learned representation)**.** *Let* $h(s, z) = \mathbb{E}[Y^* | S = s, \phi(x) = z, P = H]$. *Under assumptions 3.1- 4.1,*

$$\tau(x) = \mathbb{E}\left[h(S, \phi(x)) \mid \phi(x), T = 1, P = E\right] - \mathbb{E}\left[h(S, \phi(x)) \mid \phi(x), T = 0, P = E\right] \quad (3)$$

Remark 4.4. The function h is learned from historical data where we observe (X, S, Y^*) with all units untreated. Under surrogacy (3.4), the structural equation for Y^* depends on (X, S) but not on T, allowing the same h to apply to both treatment arms. Comparability (3.5) ensures this relationship is the same across populations.

Remark 4.5. Crucially, the outer expectation in Theorem 4.3 marginalizes over the distribution of S conditional on $\phi(x)$. This implies that to estimate $\tau(x)$ for a specific unit, we only require their covariates X, not their post-treatment surrogate outcomes.

Proof intuition. The large historical sample identifies how (X, S) predict the target outcome Y^*, while the randomized experiment sample identifies how the treatment changes the distribution of S (and, if available, Y^*) conditional on X. Comparability allows the historical outcome relationship between (X, S) and Y^* to be used in the experimental sample, and surrogacy allows for treatment-induced changes in S to be translated into changes in Y^*. The two sufficiency conditions justify replacing X with $(\phi(X), S)$ in both steps: condition (i) ensures that $\phi(X)$ captures all covariate information needed to predict Y^* from surrogates, so the outcome model h loses nothing by conditioning on $\phi(X)$ instead of X; condition (ii) ensures that $\phi(X)$ captures all covariate information needed to predict surrogates given treatment, so estimating surrogate shifts conditional on $\phi(X)$ recovers the same quantity as conditioning on X. Notably, no treatment variation is needed to learn the representation and it is trained entirely on historical controls. Full proof in Appendix A.

Implications for treatment effect estimation The identification result suggests a natural two-stage estimation procedure (see Figure 1). In the first stage, we use historical data to (i) learn the representation ϕ by optimizing Equation 6, and (ii) fit the outcome model h mapping $\phi(X), S$ to Y^*. In the second stage, we estimate treatment effects on the experimental sample using standard CATE meta-learners, but with a key modification that pseudo outcomes are regressed on $\phi(X)$ rather than the raw covariates X. This is where the sample efficiency gain is realized – Theorem 4.3 shows that, under sufficiency, the original CATE can be estimated using $\phi(X) \in \mathbb{R}^m$ rather than $X \in \mathbb{R}^d$, with $m \ll d$. Under the conditions of Corollary 1 of Kennedy [33], the DR-Learner for a γ-smooth CATE then has pointwise error $O_p\left(N_E^{-\gamma/(2\gamma+m)}\right)$, compared with $O_p\left(N_E^{-\gamma/(2\gamma+d)}\right)$ when applied directly to X. Thus the representation improves the rate governing CATE estimation from the experimental sample.

Identification under sufficiency violation. As noted in Remark 4.2, we cannot explicitly enforce Assumption 4.1(ii) during training: since historical data contains only controls ($T = 0$), this assumption may fail at $T = 1$ when treatment effect heterogeneity on S depends on features of X not predictive of S at baseline. We show that under our remaining assumptions, the estimator continues to recover a valid causal effect, and define the *ϕ-averaged CATE* as follows:

$$\tau_\phi(x) := \mathbb{E}[Y^*(1) - Y^*(0) \mid \phi(X) = \phi(x), P = E].$$

This is the average treatment effect among units sharing representation $\phi(x)$. It coincides with $\tau(x)$ when τ is constant on level sets of ϕ. When the representation compresses away features that drive heterogeneity, $\tau_\phi(x)$ averages τ over those features. Thus heterogeneity along directions ϕ retains is preserved and only heterogeneity along compressed directions is averaged away.

Theorem 4.6 (Identification under violation of Assumption 4.1(ii))**.** *Under Assumptions 3.1–3.5 and Assumption 4.1(i),*

$$\tau_\phi(x) = \mathbb{E}[h(S, \phi(x)) \mid \phi(x), T = 1, P = E] - \mathbb{E}[h(S, \phi(x)) \mid \phi(x), T = 0, P = E]. \quad (4)$$

The proof (Appendix C.1) closely follows the proof of Theorem 4.3, with the sole distinction between Theorem 4.3 and Theorem 4.6 being the invocation of Assumption 4.1(ii).

Bias-variance tradeoff. Theorem 4.6 explains why the representation remains useful even under imperfect sufficiency. As is frequently seen in representation learning, compressing X may discard some task-relevant information. Here, the compression replaces $\tau(x)$ with the coarser, but still causally interpretable $\tau_\phi(x)$. Relative to $\tau(x)$, this loss of heterogeneity is realized in bias due to compression, while estimation treatment effects over $\phi(X) \in \mathbb{R}^m$ rather than $X \in \mathbb{R}^d$ can reduce experimental-stage variance. The representation therefore reduces the mean squared error whenever the squared compression bias is smaller than the reduction in variance. This bias-variance tradeoff is especially relevant when $m \ll d$. We formalize this decomposition in Appendix C.2.

4.3 Objective for obtaining ϕ

Having established that $\phi(X)$ identifies the treatment effect, we now derive an objective for learning ϕ from historical data based on Information Bottleneck principles. The representation must satisfy both conditions of Assumption 4.1; we translate each into an optimization target.

Condition (i): Outcome-relevant information. The requirement $Y^* \perp X | \phi(X), S, P = H$, equivalent to $I(Y^*; X | \phi(X), S) = 0$, where $I(\cdot;\cdot)$ denotes mutual information (MI) [18]. By the chain rule of mutual information,

$$I(Y^*; X \mid S) = I(Y^*; \phi(X) \mid S) + I(Y^*; X \mid \phi(X), S).$$

Since $I(Y^*; X|S)$ is constant with respect to ϕ, minimizing $I(Y^*; X \mid \phi(X), S)$ is equivalent to maximizing $I(Y^*; \phi(X) \mid S)$:

$$\phi^* = \arg\min_\phi I(Y^*; X \mid \phi(X), S) = \arg\max_\phi I(Y^*; \phi(X) \mid S).$$

Condition (ii): Surrogate-relevant information. The requirement $S \perp X \mid \phi(X), T, P = E$ poses a challenge, since we learn ϕ on historical data ($P = H$) where $T = 0$ for all units. We instead impose the proxy condition $S \perp X \mid \phi(X), P = H$, equivalent to $I(S; X \mid \phi(X)) = 0$, which we encourage by maximizing $I(S; \phi(X))$:

$$\phi^* = \arg\max_\phi I(S; \phi(X)). \quad (5)$$

Appendix C characterizes the consequences when this proxy is imperfectly satisfied.

Final objective: We combine both terms as a finite-sample scalarization:

$$\phi^* = \arg\max_\phi \quad I(Y^*; \phi(X)|S) + \lambda \cdot I(S; \phi(X)) \quad (6)$$

where $\lambda > 0$ reflects differences in noise levels between Y^* and S. By the chain rule, $I((Y^*, S); \phi(X)) = I(S; \phi(X)) + I(Y^*; \phi(X) \mid S)$, so $\lambda = 1$ corresponds to maximizing the joint MI $I(\phi(X); (Y^*, S))$.

Practical implementation. We implement ϕ as a feedforward encoder mapping $X \in \mathbb{R}^d$ to $\mathbb{R}^m$. The conditional term $I(Y^*; \phi(X) \mid S)$ has no direct sample-based estimator, so we apply the chain rule above to rewrite Equation 6 as

$$I(\phi(X); (Y^*, S)) + (\lambda - 1) \cdot I(\phi(X); S)$$

which decomposes the objective into two unconditional MI terms that admit standard variational lower bounds. We train the encoder jointly with two MI estimator heads: one for joint pair (Y^*, S), and one for S alone, instantiated as either InfoNCE [57] or MINE [7]. As a simpler alternative, we can replace the MI estimators with prediction functions from $\phi(X) \to \hat{S}$ and $(\phi(X), S) \to \hat{Y}^*$ trained with BCE (binary) or MSE (continuous) losses. The prediction losses lower-bound the corresponding MI terms (see Appendix E), making this substitution theoretically valid. We report results across all three estimators in Section 5.

5 Empirical Results

In this section, we provide empirical validation for our claims that using representations that explicitly capture the predictive relationship between X and S produces more sample-efficient estimates of the CATE than other standard methods. We provide our code (linked).

Baselines We compare our method against baselines representative of standard techniques employed by practitioners, who typically either use raw covariates or apply off-the-shelf dimensionality reduction on the covariates before CATE estimation. We consider three categories of baselines:

1. **Raw covariates** X: using the high dimensional covariates X to fit the CATE learner. The most straightforward approach in which no dimensionality reduction is applied.
2. **Unsupervised reduction (PCA, ICA, Autoencoder)**: Principal Component Analysis [47], Independent Component Analysis [31, 27], and autoencoders [5, 36] are widely adopted methods for handling high-dimensional covariates without requiring outcome supervision.
3. **Supervised reduction (PLS)**: Partial Least Squares [61], a supervised dimensionality reduction of X using S as targets. PLS has access to the same surrogate information as our method, serving as a near-linear analog of our method's surrogate-relevant objective; under joint Gaussianity the two recover related subspaces.

We use the same latent representation dimension across all dimensionality-reduction-based methods (ours and the baselines). Further implementation details are provided in Appendix F.

Our method We evaluate three variants of our representation-learning objective. MI-MINE and MI-InfoNCE optimize conditional mutual-information objectives using MINE [7] and InfoNCE [57] respectively, with the joint target (Y, S). The prediction encoder learns $\phi(X)$ using prediction heads for Y and S. All three methods map X to an m-dimensional representation and otherwise use the same outcome-prediction and CATE-estimation pipeline. Architecture and other implementation details are provided in Appendix F.

Procedure All methods follow a two-stage procedure. In the first stage, we use the historical sample to (i) learn a representation (where applicable), and (ii) train an outcome model. For all methods, the outcome model $h : \phi(X) \times S \to Y$ maps representations and surrogates to predicted outcomes $\hat{Y}$. For the raw-covariate baseline, we additionally fit $h(X)$ with no surrogate information. In the second stage, we apply h to the experimental sample and use $\hat{Y}$ as the CATE target. All representation-based methods fit the CATE learner conditioned on the representation, while the raw-X baselines use X directly in the CATE learner. We use cross-fitted DML learners for CATE estimation. Further implementation details are in Appendix F.

Metrics We report normalized PEHE [25, 54] and normalized top-20% policy value (for which 0 is random targeting and 1 is oracle targeting). We repeat each experiment on 10 random data draws, using the same draws for every method. We report the mean PEHE and policy value over 10 runs along with standard error.

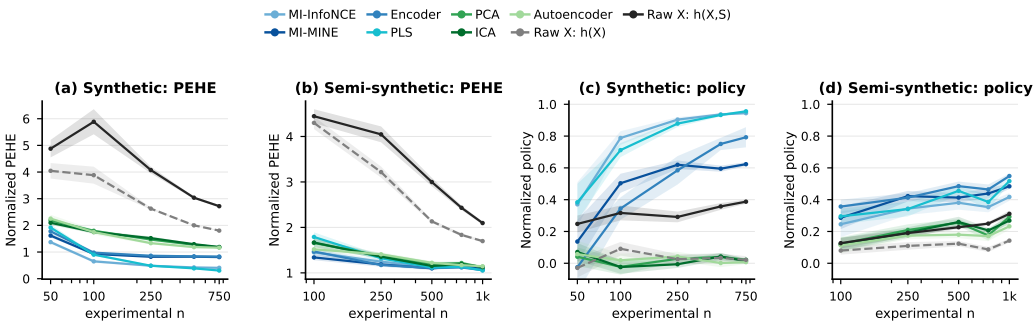


Figure 2: Normalized PEHE and normalized top-20% policy value on the synthetic and perinatal semi-synthetic datasets at $\dim(\phi) = 10$. Curves show means over 10 runs and the surrounding shaded regions show ± 1 standard error. The two raw-X curves differ only in whether the historical outcome model uses S.

5.1 Synthetic dataset

Description. We draw covariates $X \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$ with $d = 1000$. We define the ground-truth latent as a 10-dimensional vector $Z = XW^\top$, where each coordinate of Z depends on a small subset of covariates through the sparse matrix W. We split the latent into two halves, $Z = (Z_s, Z_y)$ with $Z_s, Z_y \in \mathbb{R}^5$. In the main experiment, Z_s determines the surrogates $S \in \mathbb{R}^{10}$ and the treatment effect on S, and we generate Y as a noisy linear function of S. Since Y and the treatment effect depend on X only through Z_s, our sufficiency assumption holds exactly with a five-dimensional representation.

In Appendices C.3 and D.1, we study what happens when our assumptions do not hold by letting Z_y influence the outcome and treatment effect. Further implementation details are in Appendix F.1.

Experiment design. We generate $N_H = 10{,}000$ historical samples with $T \equiv 0$ and an experimental pool of $N_E = 5{,}000$ with $T \sim \text{Bernoulli}(0.5)$. Encoders are trained on H and CATE estimation is run on stratified subsamples at $n \in \{50, 100, 250, 500, 750\}$. We set $\dim(\phi) = 10$ and use the predicted outcome $\widehat{Y}$ as the CATE target. Appendix F.2 examines sensitivity to the choice of the bottleneck dimension by varying $\dim(\phi) \in 2, 3, 5, 10, 20, 50$ for a fixed DGP. We report results averaged over 10 historical and experimental resamples from the same fixed DGP.

Results. Figure 2(a) reports normalized PEHE against n for the DML learner. The supervised methods (ours and PLS) clearly outperform the unsupervised methods across all sample sizes, and both raw-X variants have the worst performance: at $n = 100$, MI-InfoNCE achieves normalized PEHE 0.650 ± 0.059, compared with 0.902 ± 0.047 for PLS, at least 1.750 for the unsupervised baselines, and 3.883 ± 0.319 for raw X. Adding S to the raw-X outcome model does not help (5.886 ± 0.471). The policy metric shows a similar small-sample trend: at $n = 100$, MI-InfoNCE attains a top-20% policy value of 0.788 ± 0.043, versus 0.091 ± 0.043 for raw X with $h(X)$.

The linear DGP is well suited to PLS, which is accordingly the strongest baseline: MI-InfoNCE and PLS are effectively tied by $n = 250$, and PLS is modestly ahead at larger n. That our method leads at the smallest samples and substantially outperforms every unsupervised baseline throughout confirms the importance of the outcome and surrogate relevant objective, while all lower dimensional methods outperforming both variants of the raw-X baseline highlight the contribution of dimensionality reduction. Tabulated results across all sample sizes, the X-learner, and the true-experimental-Y variant are in Appendix F.1 (Tables 1–4).

5.2 Semi-synthetic medical dataset

Description. We construct a semi-synthetic dataset using real covariates from a perinatal depression study at a large academic medical center (anonymized), using real covariates and diagnosis trajectories with synthetic treatment assignment and potential outcomes. More details in Appendix F.3.

Cohorts and covariates. The historical cohort comprises $N_H = 60{,}248$ pregnancies, and the experimental pool contains $N_E = 11{,}747$ pregnancies from a perinatal mobile-app study. After preprocessing, both cohorts share $X \in \mathbb{R}^{340}$. The covariates include maternal demographics, obstetric history and pregnancy characteristics, health behaviors, clinical measurements, diagnoses during pregnancy, and health-care utilization. Further details are provided in Appendix F.3.

Auxiliary outcomes. For the auxiliary outcomes, we use diagnoses recorded in the EHR during pregnancy for conditions related to perinatal depression risk such as anxiety, depression, bipolar disorder, obsessive-compulsive disorder, trauma reactions, and substance-use disorders, along with conditions such as hypertension, diabetes, and autoimmune, cardiac, kidney, and liver conditions. Each condition is recorded at multiple points during pregnancy, giving 56 binary indicators; we drop those that are nearly always zero or one in the historical cohort, leaving 37 features. We summarize them with their first five principal components, $S \in \mathbb{R}^5$, standardizing and fitting the PCA on the historical cohort only. All of these diagnoses are observed in both cohorts.

Synthetic treatment and outcome. Treatment assignment is randomized with probability $1/2$ in the experimental sample. The synthetic treatment shifts S along an outcome-relevant direction whose magnitude depends on X. Potential outcomes are then generated from the same outcome model $P(Y \mid X, S)$ in both cohorts, so there is no direct $T \to Y$ path. This makes randomization, surrogacy, and outcome-model comparability hold by construction. Appendix F.3.2 assesses how well our assumptions hold on the semi-synthetic dataset.

Experiment design. We set $\dim(\phi) = 10$ and train the representations and outcome models on $10{,}000$ historical observations. For each run, we reserve 5,000 experimental instances for testing and draw training samples of size $n \in \{100, 250, 500, 750, 1000\}$ from the remaining pool. We estimate the CATE using three-fold cross-fitted DML, replacing the experimental outcome with its prediction $\widehat{Y}$. We repeat the experiment over 10 random data draws and use the same draws for every method. Figure 2(b,d) compares all methods, including raw-X baselines using $h(X)$ and $h(X, S)$. Tabulated results are in Appendix F.3, along with additional results for $\dim(\phi) = 5$, and runs using the true experimental outcome, along with the remaining implementation details.

Results. As in the synthetic dataset, Figure 2(b,d) shows a clear advantage of dimension-reduction methods. At $n = 100$, conditional MINE achieves a normalized PEHE of 1.340, compared with more than 4.2 for either raw-X baseline. The gap narrows as the experimental sample grows but remains substantial: at $n = 1000$, the prediction encoder and PLS both achieve approximately 1.06, compared with 1.70 for $h(X)$ and 2.10 for $h(X, S)$. The policy results show a similar trend in relative performance. Appendix F.3 reports more detailed results.

6 Discussion

Contributions This paper aims to help ML practitioners in healthcare, policy, and industry who rely on RCTs and A/B tests to estimate heterogeneous effects of interventions but are limited by small experimental sample sizes. These settings often come with access to rich historical data: EHRs, administrative records, prior A/B tests. However, the potential of such data to aid causal inference is often overlooked. We provide theoretical justification, empirical evidence, and a systematic way to use representation learning to leverage such data, and to improve the statistical power of CATE estimation in small-scale experiments. We hope that this also motivates more deliberate collection of auxiliary outcomes in administrative and operational data, so that the structural signal needed for sample-efficient causal inference is available when new interventions are tested.

Limitations and future work Our identification result relies on surrogacy and sufficiency, and while our analysis shows that empirical gains are robust to imperfect satisfaction of these conditions – sufficiency violations cost only granularity (the estimand coarsens to the average effect among units sharing $\phi(X)$), and surrogacy violations are sidestepped whenever Y^* is observed in the experiment – characterizing performance under more general failure modes remains open. Practitioners will also benefit from extensions of our work, such as fine-tuning ϕ on experimental data, allowing treated units in the historical sample, swapping the MLP encoder for pretrained image or text encoders, supporting multiple treatments and target outcomes, and extending to observational studies.

References

[1] Nancy E. Adler and Judith Stewart. Health disparities across the lifespan: Meaning, methods, and mechanisms. *Annals of the New York Academy of Sciences*, 1186(1):5–23, 2010. doi: https://doi.org/10.1111/j.1749-6632.2009.05337.x. URL `https://nyaspubs.onlinelibrary.wiley.com/doi/abs/10.1111/j.1749-6632.2009.05337.x`.

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

[3] Susan Athey, Raj Chetty, Guido W Imbens, and Hyunseung Kang. The surrogate index: Combining short-term proxies to estimate long-term treatment effects more rapidly and precisely. Technical report, National Bureau of Economic Research, 2019.

[4] Susan Athey, Niall Keleher, and Jann Spiess. Machine learning who to nudge: causal vs predictive targeting in a field experiment on student financial aid renewal. *Journal of Econometrics*, 249:105945, 2025.

[5] Pierre Baldi and Kurt Hornik. Neural networks and principal component analysis: Learning from examples without local minima. *Neural networks*, 2(1):53–58, 1989.

[6] Keith Battocchi, Eleanor Dillon, Maggie Hei, Greg Lewis, Paul Oka, Miruna Oprescu, and Vasilis Syrgkanis. EconML: A Python Package for ML-Based Heterogeneous Treatment Effects Estimation. https://github.com/py-why/EconML, 2019. Version 0.x.

[7] Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeshwar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and Devon Hjelm. Mutual information neural estimation. In *International conference on machine learning*, pages 531–540. PMLR, 2018.

[8] Ioana Bica and Mihaela van der Schaar. Transfer learning on heterogeneous feature spaces for treatment effects estimation. *Advances in Neural Information Processing Systems*, 35: 37184–37198, 2022.

[9] Kenneth A Bollen. Latent variables in psychology and the social sciences. *Annual review of psychology*, 53(1):605–634, 2002.

[10] Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Ben Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, Julian Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Rob Reich, Hongyu Ren, Frieda Rong, Yusuf Roohani, Camilo Ruiz, Jack Ryan, Christopher Ré, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishnan Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. On the opportunities and risks of foundation models, 2022. URL `https://arxiv.org/abs/2108.07258`.

[11] Johann Brehmer, Pim De Haan, Phillip Lippe, and Taco S Cohen. Weakly supervised causal representation learning. *Advances in Neural Information Processing Systems*, 35:38319–38331, 2022.

[12] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[13] Ruichu Cai, Weilin Chen, Zeqin Yang, Shu Wan, Chen Zheng, Xiaoqing Yang, and Jiecheng Guo. Long-term causal effects estimation via latent surrogates representation learning. *Neural Netw.*, 176(C), August 2024. ISSN 0893-6080. doi: 10.1016/j.neunet.2024.106336. URL https://doi.org/10.1016/j.neunet.2024.106336.

[14] Ruichu Cai, Junjie Wan, Weilin Chen, Zeqin Yang, Zijian Li, Peng Zhen, and Jiecheng Guo. Long-term individual causal effect estimation via identifiable latent representation learning. In James Kwok, editor, *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, IJCAI-25*, pages 4788–4796. International Joint Conferences on Artificial Intelligence Organization, 8 2025. doi: 10.24963/ijcai.2025/533. URL https://doi.org/10.24963/ijcai.2025/533. Main Track.

[15] Huigang Chen, Totte Harinen, Jeong-Yoon Lee, Mike Yung, and Zhenyu Zhao. Causalml: Python package for causal machine learning, 2020.

[16] Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 01 2018.

[17] Victor Chernozhukov, Mert Demirer, Esther Duflo, and Ivan Fernandez-Val. Generic machine learning inference on heterogeneous treatment effects in randomized experiments, with an application to immunization in india. Technical report, National Bureau of Economic Research, 2018.

[18] Thomas M. Cover and Joy A. Thomas. *Entropy, Relative Entropy, and Mutual Information*, chapter 2, pages 13–55. John Wiley & Sons, Ltd, 2005. ISBN 9780471748823. doi: https://doi.org/10.1002/047174882X.ch2. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/047174882X.ch2.

[19] Alicia Curth, Richard W. Peck, Eoin McKinney, James Weatherall, and Mihaela van der Schaar. Using machine learning to individualize treatment effect estimation: Challenges and opportunities. *Clinical Pharmacology & Therapeutics*, 115(4):710–719, 2024. doi: https://doi.org/10.1002/cpt.3159. URL https://ascpt.onlinelibrary.wiley.com/doi/abs/10.1002/cpt.3159.

[20] A. P. Dawid. Conditional independence in statistical theory. *Journal of the Royal Statistical Society. Series B (Methodological)*, 41(1):1–31, 1979. ISSN 00359246. URL http://www.jstor.org/stable/2984718.

[21] Qingliang Fan, Yu-Chin Hsu, Robert P Lieli, and Yichong Zhang. Estimation of conditional average treatment effects with high-dimensional data. *Journal of Business & Economic Statistics*, 40(1):313–327, 2022.

[22] Amir Feder, Katherine A Keith, Emaad Manzoor, Reid Pryzant, Dhanya Sridhar, Zach Wood-Doughty, Jacob Eisenstein, Justin Grimmer, Roi Reichart, Margaret E Roberts, et al. Causal inference in natural language processing: Estimation, prediction, interpretation and beyond. *Transactions of the Association for Computational Linguistics*, 10:1138–1158, 2022.

[23] Trinetri Ghosh, Yanyuan Ma, and Xavier De Luna. Sufficient dimension reduction for feasible and robust estimation of average causal effect. *Statistica Sinica*, 31(2):821, 2021.

[24] Negar Hassanpour and Russell Greiner. Learning disentangled representations for counterfactual regression. In *International Conference on Learning Representations*, 2020.

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

[26] Aapo Hyvarinen and Hiroshi Morioka. Unsupervised feature extraction by time-contrastive learning and nonlinear ica. *Advances in neural information processing systems*, 29, 2016.

[27] A. Hyvärinen and E. Oja. Independent component analysis: algorithms and applications. *Neural Networks*, 13(4):411–430, 2000. ISSN 0893-6080. doi: https://doi.org/10.1016/S0893-6080(00)00026-5. URL https://www.sciencedirect.com/science/article/pii/S0893608000000265.

[28] Guido Imbens, Nathan Kallus, Xiaojie Mao, and Yuhao Wang. Long-term causal inference under persistent confounding via data combination. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 87(2):362–388, 2025.

[29] Ziyang Jiang, Zhuoran Hou, Yiling Liu, Yiman Ren, Keyu Li, and David Carlson. Estimating causal effects using a multi-task deep ensemble. In *International Conference on Machine Learning*, pages 15023–15040. PMLR, 2023.

[30] Fredrik Johansson, Uri Shalit, and David Sontag. Learning representations for counterfactual inference. In *International conference on machine learning*, pages 3020–3029. PMLR, 2016.

[31] Christian Jutten and Jeanny Herault. Blind separation of sources, part i: An adaptive algorithm based on neuromimetic architecture. *Signal processing*, 24(1):1–10, 1991.

[32] Nathan Kallus and Xiaojie Mao. On the role of surrogates in the efficient estimation of treatment effects with limited outcome data. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 87(2):480–509, 2025.

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

[34] Ilyes Khemakhem, Diederik Kingma, Ricardo Monti, and Aapo Hyvarinen. Variational autoencoders and nonlinear ica: A unifying framework. In *International conference on artificial intelligence and statistics*, pages 2207–2217. PMLR, 2020.

[35] Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Joan Puigcerver, Jessica Yung, Sylvain Gelly, and Neil Houlsby. Big transfer (bit): General visual representation learning. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, pages 491–507, Cham, 2020. Springer International Publishing.

[36] Mark A. Kramer. Nonlinear principal component analysis using autoassociative neural networks. *AIChE Journal*, 37(2):233–243, 1991. doi: https://doi.org/10.1002/aic.690370209. URL https://aiche.onlinelibrary.wiley.com/doi/abs/10.1002/aic.690370209.

[37] Sören R Künzel, Bradly C Stadie, Nikita Vemuri, Varsha Ramakrishnan, Jasjeet S Sekhon, and Pieter Abbeel. Transfer learning for estimating causal effects using neural networks. *arXiv preprint arXiv:1808.07804*, 2018.

[38] 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. doi: 10.1073/pnas.1804597116. URL https://www.pnas.org/doi/abs/10.1073/pnas.1804597116.

[39] Sheng Li. Large pre-trained models for treatment effect estimation: Are we there yet? *Patterns*, 5(6), 2024.

[40] Lydia T Liu, Solon Barocas, Jon Kleinberg, and Karen Levy. On the actionability of outcome prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 22240–22249, 2024.

[41] Ruoqi Liu, Pin-Yu Chen, and Ping Zhang. Cure: A deep learning framework pre-trained on large-scale patient data for treatment effect estimation. *Patterns*, 5(6), 2024.

[42] Christos Louizos, Uri Shalit, Joris M. Mooij, David A. Sontag, Richard S. Zemel, and Max Welling. Causal effect inference with deep latent-variable models. In *Neural Information Processing Systems*, 2017.

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

[44] Hamed Nilforoshan, Michael Moor, Yusuf Roohani, Yining Chen, Anja Šurina, Michihiro Yasunaga, Sara Oblak, and Jure Leskovec. Zero-shot causal learning. *Advances in Neural Information Processing Systems*, 36:6862–6901, 2023.

[45] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

[46] Judea Pearl and Elias Bareinboim. External validity: From do-calculus to transportability across populations. In *Probabilistic and causal inference: The works of Judea Pearl*, pages 451–482. ACM, 2022.

[47] Karl Pearson. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572, 1901. doi: 10.1080/14786440109462720. URL `https://doi.org/10.1080/14786440109462720`.

[48] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[49] Jo C. Phelan, Bruce G. Link, and Parisa Tehranifar. Social conditions as fundamental causes of health inequalities: Theory, evidence, and policy implications. *Journal of Health and Social Behavior*, 51(1_suppl):S28–S40, 2010. doi: 10.1177/0022146510383498. URL `https://doi.org/10.1177/0022146510383498`. PMID: 20943581.

[50] Ross L. Prentice. Surrogate endpoints in clinical trials: definition and operational criteria. *Statistics in medicine*, 8 4:431–40, 1989.

[51] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

[52] Donald B. Rubin. Randomization analysis of experimental data: The fisher randomization test comment. *Journal of the American Statistical Association*, 75(371):591–593, 1980. ISSN 01621459, 1537274X. URL `http://www.jstor.org/stable/2287653`.

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

[54] Uri Shalit, Fredrik D. Johansson, and David Sontag. Estimating individual treatment effect: generalization bounds and algorithms. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 3076–3085. PMLR, 06–11 Aug 2017. URL `https://proceedings.mlr.press/v70/shalit17a.html`.

[55] Claude Elwood Shannon. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423, 1948.

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

[57] Aäron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *CoRR*, abs/1807.03748, 2018. URL `http://arxiv.org/abs/1807.03748`.

[58] Ravi Varadhan, Jodi B Segal, Cynthia M Boyd, Albert W Wu, and Carlos O Weiss. A framework for the analysis of heterogeneity of treatment effect in patient-centered outcomes research. *Journal of clinical epidemiology*, 66(8):818–825, 2013.

[59] Victor Veitch, Dhanya Sridhar, and David Blei. Adapting text embeddings for causal inference. In *Conference on uncertainty in artificial intelligence*, pages 919–928. PMLR, 2020.

[60] Stefan Wager and Susan Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242, 2018. doi: 10.1080/01621459.2017.1319839. URL https://doi.org/10.1080/01621459.2017.1319839.

[61] Herman Wold. Estimation of principal components and related models by iterative least squares. *Multivariate analysis*, pages 391–420, 1966.

[62] Shu Yang, Siyi Liu, Donglin Zeng, and Xiaofei Wang. Data fusion methods for the heterogeneity of treatment effect and confounding function. *Bernoulli*, 31(4):2987–3012, 2025.

[63] Weijia Zhang, Lin Liu, and Jiuyong Li. Treatment effect estimation with disentangled latent factors. In *AAAI Conference on Artificial Intelligence*, 2020.

[64] Yao Zhang, Jeroen Berrevoets, and Mihaela Van Der Schaar. Identifiable energy-based representations: An application to estimating heterogeneous causal effects. In *International Conference on Artificial Intelligence and Statistics*, pages 4158–4177. PMLR, 2022.

A Proof of Theorem 4.3

In this section we prove that under Assumptions 3.1-4.1, the CATE $\tau(x) = \mathbb{E}[Y^*(1) - Y^*(0) \mid X = x, P = E]$ is identified by Equation (3). Since $\tau(x)$ is a difference of conditional expectations, it suffices to show that for each treatment arm $t \in \{0, 1\}$:

$$\mathbb{E}[Y^*(t) \mid X = x, P = E] = \mathbb{E}[h(S, \phi(x)) \mid \phi(x), T = t, P = E]$$

where $h(s, z) = \mathbb{E}[Y^* \mid S = s, \phi(X) = z, P = H]$. For brevity, we abbreviate Y^* as Y.

Proof. We prove Equation (3) for each treatment arm $t \in \{0, 1\}$.

$$\begin{aligned}
\mathbb{E}[Y(t) \mid X = x, P = E] &= \mathbb{E}[Y \mid X = x, T = t, P = E] && \text{(ignorability, SUTVA)} \\
&= \mathbb{E}\big[\mathbb{E}[Y \mid S, T = t, X = x, P = E] \,\big|\, X = x, T = t, P = E\big] && \text{(iterated expectations)} \\
&= \mathbb{E}\big[\mathbb{E}[Y \mid S, X = x, P = E] \,\big|\, X = x, T = t, P = E\big] && \text{(surrogacy, A3.4)} \\
&= \mathbb{E}\big[\mathbb{E}[Y \mid S, X = x, P = H] \,\big|\, X = x, T = t, P = E\big] && \text{(comparability, A3.5)} \\
&= \mathbb{E}\big[\mathbb{E}[Y \mid S, \phi(x), P = H] \,\big|\, X = x, T = t, P = E\big] && \text{(sufficiency (i), A4.1)} \\
&= \mathbb{E}\big[h(S, \phi(x)) \,\big|\, X = x, T = t, P = E\big] && \text{(definition of } h\text{)} \\
&= \mathbb{E}\big[h(S, \phi(x)) \,\big|\, \phi(x), T = t, P = E\big] && \text{(sufficiency (ii), A4.1).}
\end{aligned}$$

Taking the difference across $t \in \{0, 1\}$ yields Equation (3). □

A.1 Discussion of proof steps

Step 1 (Ignorability). By randomization of treatment assignment in the experimental sample, Assumptions 3.1– 3.2 ensures that the potential outcome $Y(t)$ equals the observed outcome Y when $T = t$. This is a standard identification step in any randomized experiment.

Step 2 (Tower property). We introduce S via the law of iterated expectations.

Step 3 (Surrogacy). Assumption 3.4 states $T \perp Y \mid S, X, P = E$. Treatment is conditionally independent of the outcome given surrogates and covariates, so we drop T from the inner expectation. The outer expectation retains T, which governs the distribution of S.

Step 4 (Comparability). Assumption 3.5 states $Y \perp P \mid S, X$. The conditional distribution of Y given (S, X) is invariant across populations, allowing us to switch from $P = E$ to $P = H$ in the inner expectation.

Step 5 (Sufficiency (i)). Assumption 4.1(i) states $Y \perp X \mid \phi(X), S, P = H$. Conditional on the representation $\phi(X)$ and surrogates, the full covariates are redundant for predicting Y.

Step 6 (Outcome model). We define $h(s, z) = \mathbb{E}[Y \mid S = s, \phi(X) = z, P = H]$.

Step 7 (Sufficiency (ii)). Assumption 4.1(ii) states $S \perp X \mid \phi(X), T, P = E$. Conditional on the representation and treatment, the full covariates are redundant for predicting S, so the outer expectation depends on X only through $\phi(X)$.

Combining. We obtain:

$$\mathbb{E}[Y(t) \mid X = x, P = E] = \mathbb{E}[h(S, \phi(x)) \mid \phi(x), T = t, P = E]$$

The CATE is:

$$\tau(x) = \mathbb{E}[h(S, \phi(x)) \mid \phi(x), T = 1, P = E] - \mathbb{E}[h(S, \phi(x)) \mid \phi(x), T = 0, P = E]$$

Summary. Our main result specifies what information the representation must retain from the covariates, and what can be ignored (Steps 5 and 7). It also combines historical and experimental data in the following way: historical data is used to fit the outcome model h, which captures how the surrogates and baseline covariates jointly predict the outcome. Experimental data is used to estimate how treatment shifts the distribution of surrogates conditional on the representation $\phi(x)$. Finally, surrogacy ensures that treatment affects Y only through S and, along with comparability, thus allows us to use the historical outcome model to estimate the CATE in the experimental sample.

B On ignorability assumption

B.1 Ignorability under compression

We formalize and prove the claim that in randomized experiments, ignorability (Assumption 3.2) is preserved by any function of the covariates – including any learned representation $\phi(X)$. For brevity, we abbreviate Y^* as Y.

Lemma B.1 (Ignorability under compression)**.** *Under Assumption 3.2, for any measurable $\phi : \mathbb{R}^d \to \mathbb{R}^m$,*

$$T \perp (Y(0), Y(1)) \mid \phi(X), P = E.$$

Proof. Assumption 3.2 gives $T \perp (X, Y(0), Y(1)) \mid P = E$. Since ϕ is a measurable function of X, this implies $T \perp (\phi(X), Y(0), Y(1)) \mid P = E$. By the Weak Union axiom for conditional independence [20], $T \perp (Y(0), Y(1)) \mid \phi(X), P = E$ for any measurable ϕ. □

This contrasts with observational representation learning [30, 54], where ignorability holds conditional on X, and a learned ϕ must explicitly retain the variables responsible for that conditioning. In our setting, randomization makes treatment unconditionally independent of all covariates and potential outcomes, so any function of X (including $\phi(X)$) preserves ignorability. The representation can therefore be optimized for information relevant to the target outcome and surrogates (Section 4.3), without a separate requirement to retain information for confounding adjustment.

We discuss below the extension to stratified randomization, where ignorability holds only conditional on a known stratification subset $V \subset X$.

B.2 Extension to stratified randomizations

Assumption 3.2 corresponds to simple randomization, as seen in standard RCTs. Under stratified randomization, treatment is assigned independently within strata defined by a known subset of covariates $V \subset X$. Within strata, treatment is independent of all other covariates and potential outcomes:

$$T \perp (X, Y(1), Y(0)) \mid V, P = E \tag{7}$$

Since V contains all covariates governing treatment assignment in a stratified RCT, treatment is independent of all other covariates and potential outcomes given V. Thus we can define:

$$\tilde{\phi}(X) := (V, \phi(X))$$

where ϕ is trained on the full X using the objective in Equation 6. Concatenating V ensures it is preserved regardless of whether the encoder retains it. Because $\phi(X)$ is a function of X, (7) implies $T \perp (\phi(X), Y(1), Y(0)) \mid V, P = E$. Applying Weak Union axiom (as in Lemma B.1) yields $T \perp (Y(1), Y(0)) \mid V, \phi(X), P = E$, i.e. ignorability is preserved: $T \perp (Y(0), Y(1)) \mid \tilde{\phi}(X), P = E$. Theorem 4.3 applies with $\tilde{\phi}$ in place of ϕ, provided Assumption 4.1 is restated with respect to $\tilde{\phi}(X)$. Since $\tilde{\phi}(X)$ contains strictly more information than $\phi(X)$, the sufficiency conditions are no harder to satisfy than in the unstratified case. The proof of Theorem 4.3 is then identical with $\tilde{\phi}$ substituted throughout. The dimensional benefit is preserved: $\tilde{\phi}(X) \in \mathbb{R}^{m+|V|}$, and stratification variables are typically low-dimensional, so $m + |V| \ll d$.

C Sufficiency Violation

C.1 Identification under violation

In this section, we analyze the estimator from Theorem 4.3 when Assumption 4.1(ii) is violated. Recall that (ii) requires $S \perp X \mid \phi(X), T, P = E$ for all treatment levels $t \in \{0, 1\}$. Since the representation ϕ is learned on historical data where $T = 0$ for all units, the proxy objective in Section 4.3 can at best enforce this condition at $T = 0$; it may fail at $T = 1$ when treatment effect heterogeneity on S depends on features of X not predictive of S at baseline.

We define the ϕ-averaged CATE as:

$$\tau_\phi(x) := \mathbb{E}[Y^*(1) - Y^*(0) \mid \phi(X) = \phi(x), P = E].$$

This is the average treatment effect among units sharing representation $\phi(x)$. It coincides with $\tau(x)$ when τ is constant on level sets of ϕ (eg. trivially for bijective ϕ). When the representation compresses away features that drive treatment effect heterogeneity, $\tau_\phi(x)$ averages τ over those features. That is, heterogeneity along directions ϕ retains is preserved, only heterogeneity along compressed directions is averaged over. For brevity, we shall abbreviate Y^* as Y in the following analysis.

From Theorem 4.6, under Assumptions 3.1–3.5 and Assumption 4.1(i),

$$\tau_\phi(x) = \mathbb{E}[h(S, \phi(x)) \mid \phi(x), T = 1, P = E] - \mathbb{E}[h(S, \phi(x)) \mid \phi(x), T = 0, P = E]$$

Proof. For each treatment arm $t \in \{0, 1\}$, we show

$$\mathbb{E}[h(S, \phi(x)) \mid \phi(X) = \phi(x), T = t, P = E] = \mathbb{E}[Y(t) \mid \phi(X) = \phi(x), P = E].$$

By the law of iterated expectations, we can expand $\mathbb{E}[h(S, \phi(x)) \mid \phi(X) = \phi(x), T = t, P = E]$ as

$$\mathbb{E}\big[\mathbb{E}[h(S, \phi(x)) \mid X, T = t, P = E]\big|\phi(X) = \phi(x), T = t, P = E\big]$$

. The outer expectation integrates over X on the set $\{x : \phi(x') = \phi(x)\}$. On this set, steps 1–6 of the proof of Theorem 4.3 (which do not use Assumption 4.1(ii)) apply with $\phi(x') = \phi(x)$, giving

$$\mathbb{E}[h(S, \phi(x)) \mid X = x', T = t, P = E] = \mathbb{E}[Y(t) \mid X = x', P = E].$$

By Assumption 3.2, $T \perp X \mid P = E$, so the distribution of X given $\phi(X) = \phi(x)$ is the same with or without conditioning on $T = t$. Therefore

$$\begin{aligned}\mathbb{E}[h(S, \phi(x)) \mid \phi(X) = \phi(x), T = t, P = E] =& \mathbb{E}[\mathbb{E}[Y(t) \mid X, P = E] \mid \phi(X) = \phi(x), P = E] \\ &= \mathbb{E}[Y(t) \mid \phi(X) = \phi(x), P = E].\end{aligned}$$

Taking the difference between $t \in \{0, 1\}$ yields $\tau_\phi(x)$. □

C.2 Bias-variance tradeoff under imperfect sufficiency

We now characterize the bias-variance tradeoff discussed in Section 4. We note that we are concerned with setting in which $N_H \gg N_E$. Our target is the original CATE $\tau(x)$. When Assumption 4.1 does not hold perfectly, Theorem 4.6 instead identifies

$$\tau_\phi(x) := \mathbb{E}[Y^*(1) - Y^*(0) \mid \phi(X) = \phi(x), P = E].$$

By the law of iterated expectations, we have:

$$\tau_\phi(X) = \mathbb{E}[\tau(X) \mid \phi(X), P = E].$$

Given an experimental sample D of size N_E, let $\widehat{\tau}_{\phi,D}(\phi(x))$ denote our representation-based estimator. Since the representation and outcome model are learned from the much larger historical sample, we treat them as fixed and focus on the error arising in the experimental stage. We further assume that the experimental estimator is centered at the coarsened CATE:

$$\mathbb{E}_D\left[\widehat{\tau}_{\phi,D}(\phi(x))\right] = \tau_\phi(x).$$

Our error thus lends itself to a familiar bias-variance decomposition,

$$\mathbb{E}_{X,D}\left[\{\widehat{\tau}_{\phi,D}(\phi(X)) - \tau(X)\}^2\right] = \underbrace{\mathbb{E}_X\left[\{\tau_\phi(X) - \tau(X)\}^2\right]}_{\text{squared bias due to coarsening}} + \underbrace{\mathbb{E}_X\left[\mathrm{Var}_D\left\{\widehat{\tau}_{\phi,D}(\phi(X))\right\}\right]}_{\text{variance due to the finite experimental sample}}.$$

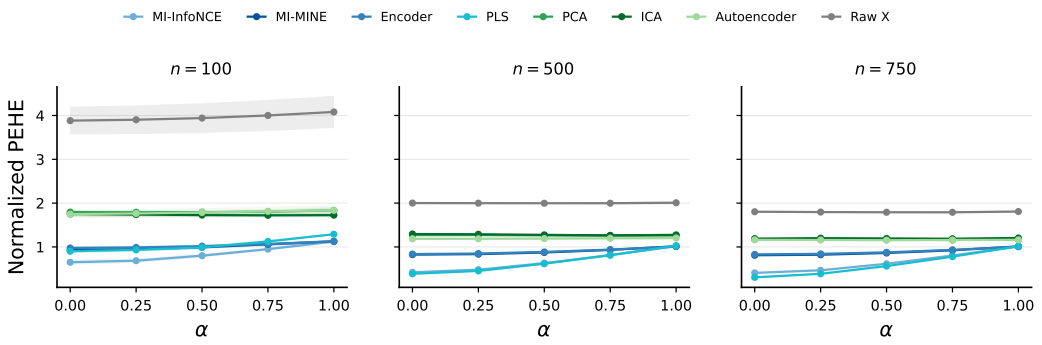
The cross term is zero because

$$\mathbb{E}\left[\tau(X) - \tau_\phi(X) \mid \phi(X), P = E\right] = 0.$$

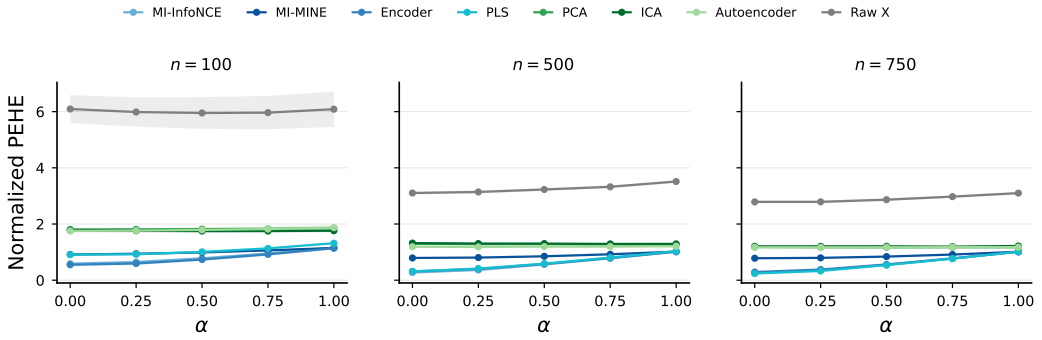
For comparison, an estimator that uses raw covariates directly would have no coarsening bias (assuming it is an unbiased estimator), and the error would consist entirely of variance due to the finite experimental sample. Let V_X and V_ϕ denote the expected variances of the raw-covariate and representation-based estimators respectively. The representation-based estimator has lower mean squared error whenever

$$\mathbb{E}_X\left[\{\tau_\phi(X) - \tau(X)\}^2\right] < V_X(N_E) - V_\phi(N_E).$$

Thus we still benefit from the representation-based estimator, even under imperfect sufficiency. The representation reduces the total error whenever the reduction in variance exceeds the squared bias due to coarsening. In the following section, we empirically examine how the tradeoff impacts performance of our method as we increase the extent of sufficiency violation.



(a) Normalized PEHE, $\hat{Y}$ target.



(b) Normalized PEHE, Y target.

Figure 3: **Sufficiency-(ii) violation, DML — normalized PEHE.** True-outcome-aware representations retain a clear advantage under partial violations, but their normalized PEHE increases with α as more treatment-effect heterogeneity depends on features that cannot be learned from the historical data.

C.3 Empirical analysis of sufficiency violation

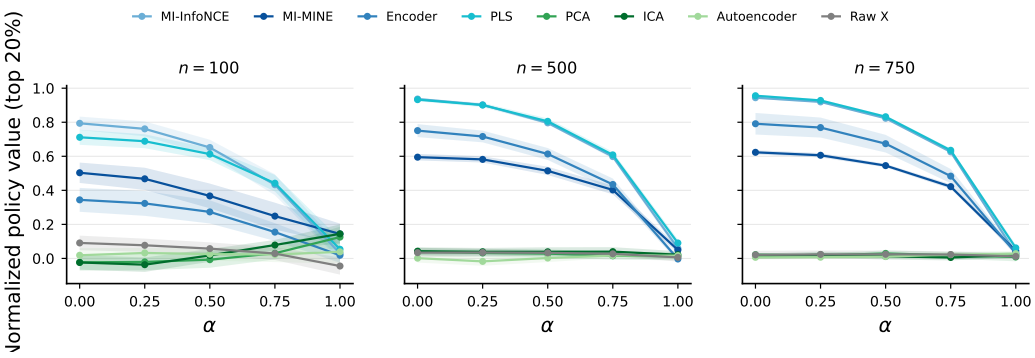
We complement Theorem 4.6 with an empirical sweep over the degree of sufficiency-(ii) violation. Building on the synthetic DGP of Section 5.1, we introduce a scalar $\alpha \in [0, 1]$ that controls how much of the treatment-induced shift in S is driven by features of X *not* predictive of S at baseline ($T = 0$):

$$S_i(t) = S_i(0) + t \cdot \left\{ \gamma_0 + \sqrt{1 - \alpha^2}\, g_s(Z_{s,i}) + \alpha\, g_y(Z_{y,i}) \right\},$$

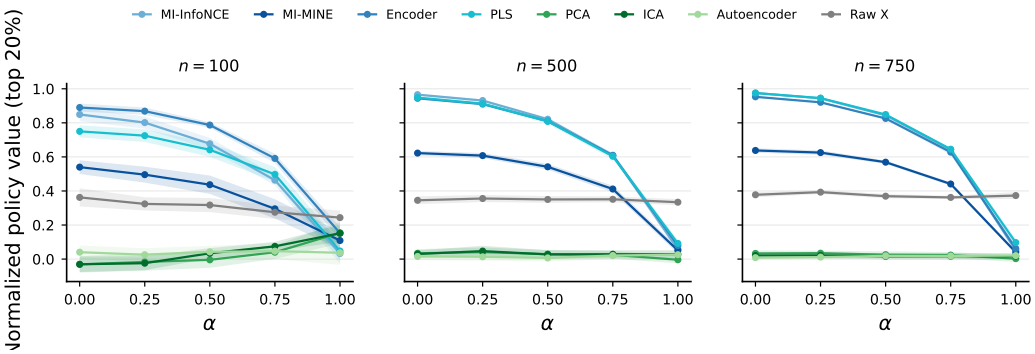
The baseline surrogate $S_i(0)$ depends on Z_s but not on Z_y. Thus, the historical control data contain information about Z_s, but provide no signal for the representation to retain Z_y. Thus, $g_y(Z_y)$ introduces treatment-effect heterogeneity that cannot be learned from the untreated historical sample. We scale $g_s(Z_s)$ and $g_y(Z_y)$ to have equal variance, so that changing α changes the source of the heterogeneity without changing its overall magnitude. At $\alpha = 0$, Assumption 4.1(ii) holds and we identify $\tau(x)$. As α increases, more heterogeneity depends on features that ϕ cannot recover from the historical data, and Theorem 4.6 instead identifies $\tau_\phi(x)$. At $\alpha = 1$, all treatment-effect heterogeneity is determined by these features.

We use the same three-fold DML learner as in Section 5. We run each method using both the historical outcome-model prediction $\widehat{Y}$ and the true experimental outcome Y as the CATE target. We vary $\alpha \in \{0, 0.25, 0.5, 0.75, 1\}$ and $n \in \{50, 100, 250, 500, 750\}$ and report mean $\pm$ one standard error over the same ten paired historical and experimental samples used in Section 5.1. Remaining implementation details in Appendix F.1.

Results. Figures 3 and 4 report normalized PEHE and policy value across α, respectively. At $n = 500$, increasing α from zero to 0.5 raises MI-InfoNCE's normalized PEHE from 0.423 to 0.549,



(a) Normalized policy value (top 20%), $\hat{Y}$ target.



(b) Normalized policy value (top 20%), Y target.

Figure 4: **Sufficiency-(ii) violation, DML — normalized top-20% policy value.** The policy advantage declines gradually under partial violation and disappears at $\alpha = 1$, when the representation no longer contains the information that determines treatment-effect rankings.

while its policy value remains high (0.936 to 0.862). At $\alpha = 1$, however, its PEHE reaches 1.031 and its policy value falls to 0.041; Other supervised methods follow the same pattern, and the trend is similar across sample sizes. We observe that compression can still reduce squared error relative to raw X, whose normalized PEHE is approximately 2.0 in this setting, but it cannot recover heterogeneity absent from the representation, as seen at $\alpha = 1.0$. The results support our method's robustness to significant violations of sufficiency, although not to complete violations.

D Surrogacy Violation

Appendix C.3 characterizes what happens to the estimator when sufficiency-(ii) is violated. Here we treat the second identification assumption that may not hold exactly in practice: *surrogacy* (Assumption 3.4), which requires $T \perp Y^* \mid S, X, P = E$ – i.e., that the surrogates S fully mediate the effect of treatment on the target outcome. When this fails, treatment exerts an additional direct effect on Y^* that bypasses S.

Figure 5 illustrates both violations within the broader graphical model assumed by our framework. The top row shows the unviolated case: the historical DAG has no treatment node, the experimental DAG introduces T acting on S (with S in turn driving Y^*), and the latent Z affects both S and Y^* throughout. The bottom row depicts the two violations we are concerned with. Surrogacy violation (left) adds a direct $T \to Y^*$ edge, breaking the mediation through S. Sufficiency-(ii) violation (right) is treatment-arm-specific: at $T = 0$ the structure matches what ϕ was trained on, while at $T = 1$ a violating $X \to S$ edge appears, encoding heterogeneity in the treatment effect on S that ϕ cannot recover from historical (untreated) data. The two violations affect CATE estimation in different ways: sufficiency-(ii) coarsens the estimand to $\tau_\phi(x)$ (Theorem 4.6), while surrogacy violation introduces

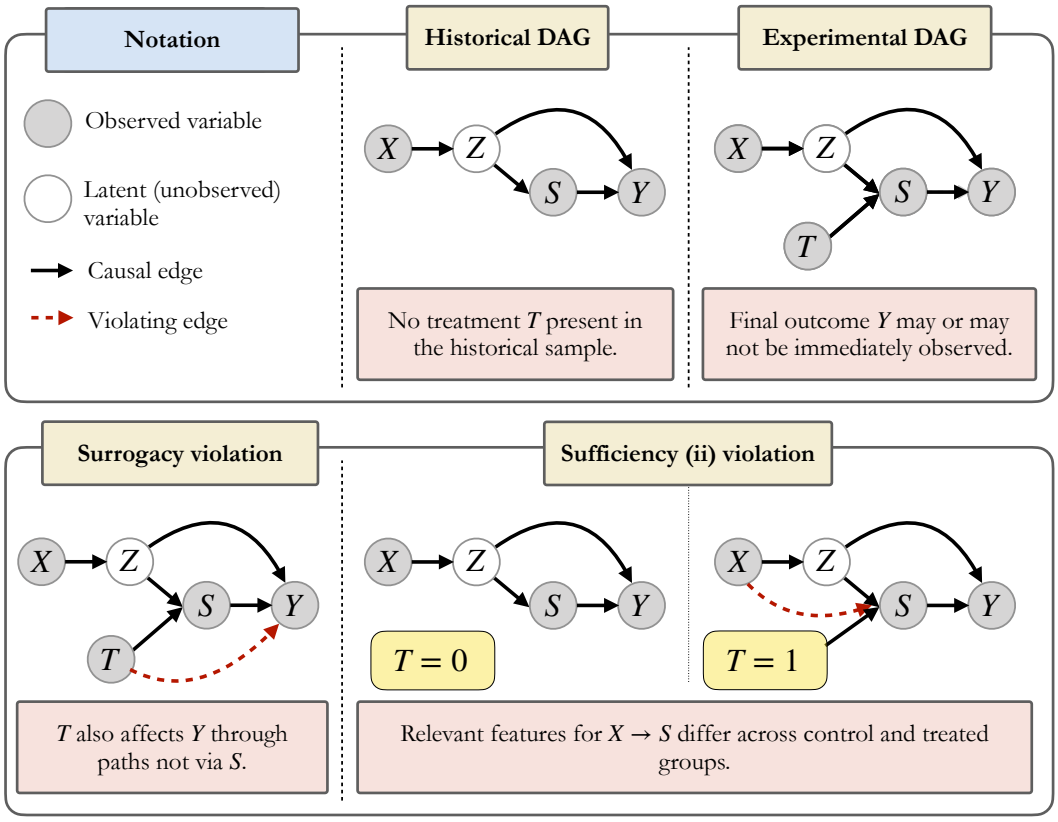


Figure 5: **Causal structure: default and two identification violations.** *Top row, left to right:* notation; the historical DAG, where no treatment is present; the experimental DAG, where T acts on S and S mediates the effect on Y^*. Z is the ground-truth latent representation of X on which S and Y^* depend; $\phi(X)$ is trained to recover Z. *Bottom row:* the two violations. *Left, surrogacy violation:* a direct $T \to Y^*$ edge (dashed red) breaks the mediation through S; analyzed in Appendix D.1. *Right, sufficiency-(ii) violation:* the violation is specific to the treated group. At $T = 0$ (left sub-panel) the structure matches what ϕ was trained on, so sufficiency-(ii) holds; at $T = 1$ (right sub-panel) a violating $X \to S$ edge (dashed red) appears, encoding heterogeneity in the treatment effect on S that depends on features of X unrecoverable from untreated data. Analyzed in Appendix C.3.

bias unless Y^* is observed in the experimental sample, in which case randomization identifies the representation-conditional CATE directly.

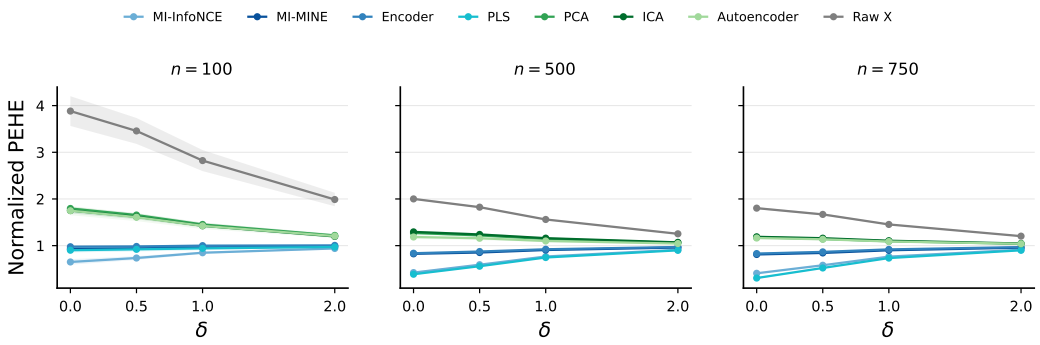
In the following section, we run an empirical sweep over a scalar δ controlling the strength of the direct $T \to Y^*$ path, and study the robustness of our method to such a violation.

D.1 Empirical analysis of surrogacy violation (δ)

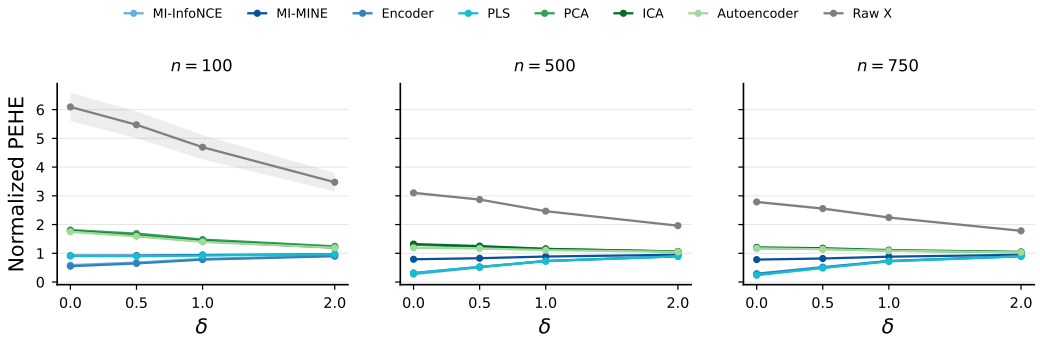
We complement the above discussion with an empirical sweep over the degree of *surrogacy violation*, i.e. the extent to which Y^* depends on covariates X through paths that do not pass through the surrogate S. Building on the synthetic DGP of Section 5.1, we introduce a scalar $\delta \geq 0$ that controls the strength of the direct effect of treatment on Y^* that does not pass through S:

$$Y_i^*(t) = h(S_i(t)) + \delta \cdot t \cdot g_\perp(Z_i^Y) + \varepsilon_i,$$

where Z^Y denotes latent directions that do not affect S, and $g_\perp$ is fixed across δ. Since the direct term is multiplied by treatment, it is absent from the untreated historical data and cannot be learned by the historical outcome model. At $\delta = 0$, surrogacy holds and the treatment effect on Y^* is fully mediated by S. As δ increases, a larger share of the treatment effect bypasses S. We scale the direct and mediated components to have equal variance at $\delta = 1$; at $\delta = 2$, the direct component accounts for 80% of the CATE variance.



(a) Normalized PEHE, $\hat{Y}$ target.



(b) Normalized PEHE, Y target.

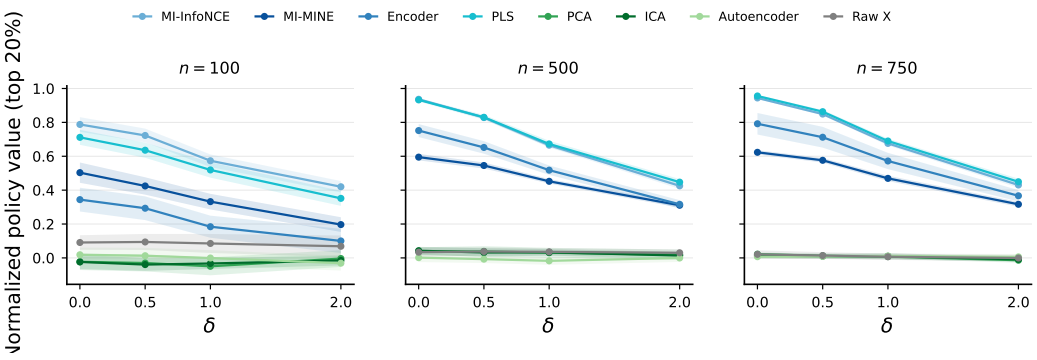
Figure 6: **Surrogacy violation, DML – normalized PEHE.** The outcome-aware representations lose accuracy as more of the treatment effect bypasses S, but remain more accurate than unsupervised compression and raw X throughout the sweep.

We use the same three-fold DML learner as in Section 5. We run each method using both the historical outcome-model prediction $\widehat{Y}$ and the true experimental outcome Y as the CATE target. We vary $\delta \in \{0, 0.5, 1, 2\}$ and $n \in \{50, 100, 250, 500, 750\}$ and report mean $\pm$ one standard error over the same ten historical and experimental samples used in Section 5.1. Remaining implementation details are in Appendix F.1.

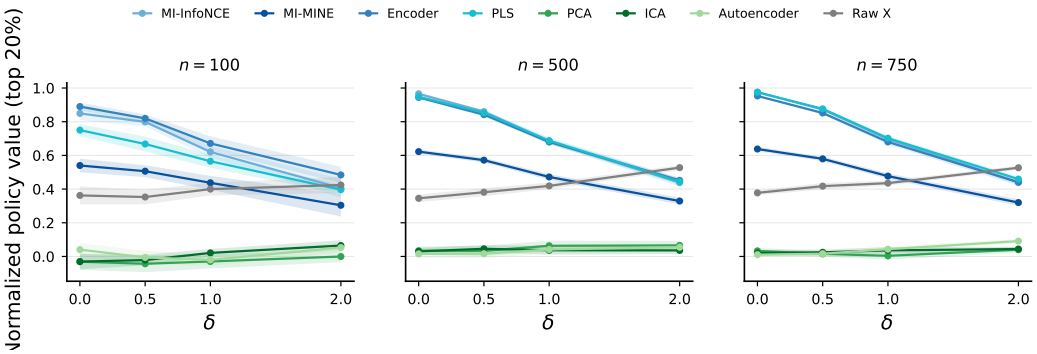
Results. Figures 6 and 7 report normalized PEHE and policy value across δ, respectively. At $n = 500$, increasing δ from zero to one raises MI-InfoNCE's normalized PEHE from 0.423 to 0.767, while its policy value remains 0.665. At $\delta = 2$, its PEHE reaches 0.911 and its policy value falls to 0.425; PLS follows the same pattern, and the trend is similar across sample sizes. When we use the true experimental outcome, raw X increasingly recovers the direct-effect ranking, although its normalized PEHE at $\delta = 2$ remains substantially higher than that of MI-InfoNCE (1.961 versus 0.898). For raw X and unsupervised methods, normalized PEHE declines with δ even though absolute PEHE increases because the standard deviation of the CATE also increases. Thus historical supervision remains useful under substantial surrogacy violations, but its advantage narrows as more of the treatment effect bypasses S.

E Prediction losses as mutual information lower bounds

In addition to neural MI estimators (InfoNCE, MINE), we also report results using simple prediction losses (BCE for binary outcomes, MSE for continuous). Here we show that both BCE and MSE losses lower-bound the mutual information $I(X;Y)$, providing principled justification for their use.



(a) Normalized policy value (top 20%), $\hat{Y}$ target.



(b) Normalized policy value (top 20%), Y target.

Figure 7: **Surrogacy violation, DML – normalized top-20% policy value.** The policy advantage declines as the direct effect grows, but remains positive at $\delta = 2$ because the surrogate-mediated component remains informative.

Setup. Since $I(X;Y) = H(Y) - H(Y \mid X)$ and $H(Y)$ is fixed, maximizing MI is equivalent to minimizing $H(Y \mid X)$. We show each loss upper-bounds $H(Y \mid X)$, so minimizing the loss tightens a lower bound on MI. Conditioning on additional covariates Z follows identically.

BCE ($Y \mid X$ Bernoulli). The expected BCE loss decomposes via $H(P,Q) = H(P) + D_{\mathrm{KL}}(P \,\|\, Q)$:

$$\mathcal{L}_{\mathrm{BCE}}(\theta) = \mathbb{E}_X[H(P_{\mathrm{true}}(Y \mid X), P_\theta(Y \mid X))] = H(Y \mid X) + \mathbb{E}_X[D_{\mathrm{KL}}(P_{\mathrm{true}} \,\|\, P_\theta)].$$

Since $D_{\mathrm{KL}} \geq 0$, $\mathcal{L}_{\mathrm{BCE}} \geq H(Y \mid X)$, hence $H(Y) - \mathcal{L}_{\mathrm{BCE}} \leq I(X;Y)$. Minimizing $\mathcal{L}_{\mathrm{BCE}}$ drives $D_{\mathrm{KL}} \to 0$ (given sufficient capacity), making the bound tight.

MSE ($Y \mid X$ real-valued, $\mu(X) = \mathbb{E}[Y \mid X]$). The bias-variance decomposition gives:

$$\mathcal{L}_{\mathrm{MSE}}(\theta) = \mathbb{E}_X[\mathrm{Var}(Y \mid X)] + \mathbb{E}_X[(\mu(X) - f_\theta(X))^2] \geq \mathbb{E}_X[\mathrm{Var}(Y \mid X)].$$

To connect to MI, we bound the conditional differential entropy $h(Y \mid X)$. The Gaussian distribution has maximum differential entropy among distributions with a given variance [55], so pointwise: $h(Y \mid X = x) \leq \frac{1}{2}\ln(2\pi e \cdot \mathrm{Var}(Y \mid X = x))$. Taking $\mathbb{E}_X$ and applying Jensen's inequality:

$$h(Y \mid X) \leq \tfrac{1}{2}\ln(2\pi e \cdot \mathbb{E}_X[\mathrm{Var}(Y \mid X)]) \leq \tfrac{1}{2}\ln(2\pi e \cdot \mathcal{L}_{\mathrm{MSE}}(\theta)).$$

Hence $h(Y) - \frac{1}{2}\ln(2\pi e \cdot \mathcal{L}_{\mathrm{MSE}}) \leq I(X;Y)$, and minimizing $\mathcal{L}_{\mathrm{MSE}}$ tightens this bound.

Tightness. The BCE bound is tight at convergence (with sufficient capacity). The MSE bound involves two additional gaps: the max-entropy gap (tight when $Y \mid X$ is Gaussian) and Jensen's gap (tight when $\mathrm{Var}(Y \mid X)$ is constant in X). Both close under additive Gaussian noise $Y = \mu(X) + \varepsilon$, $\varepsilon \sim \mathcal{N}(0, \sigma^2)$. In general the MSE bound is valid but looser, which is consistent with our empirical finding (Section 5) that MINE and InfoNCE achieve modestly better performance than the prediction-loss variant.

Application to the combined objective. The bounds above apply to each term in Equation 6: setting $(X, Y \mid Z) = (\phi(X), Y^* \mid S)$ yields the outcome-relevant MI bound, and $(X, Y) = (\phi(X), S)$ yields the surrogate-relevant bound.

F Empirical results (cont.)

Libraries used. All experiment code is written in Python. We implement the neural-network encoder architecture using PyTorch [45]. For baseline machine-learning models, preprocessing, and evaluation metrics we use scikit-learn [48]; the X-learner is implemented via CausalML [15], and the DML estimator via EconML [6]. Synthetic outcome and propensity nuisance models use scikit-learn's `HistGradientBoostingRegressor`; dataset-specific settings are given below.

Our method: encoder backbone. For the synthetic experiments, all neural methods use the same feedforward encoder with ReLU activations, $X \in \mathbb{R}^d \to 128 \to 64 \to 32 \to m$, where $m = 10$ is the bottleneck dimension. We train with Adam (lr 10^{-3}, batch 256, up to 200 epochs) and early stopping on training loss with patience 15 and stopping criterion of 10^{-4}. The semi-synthetic experiment uses the smaller backbone and validation-based early stopping described in Appendix F.3, with $m = 10$ in the main analysis and $m = 5$ as a sensitivity analysis.

Our method: training variants. We instantiate the objective in Equation 6 in three ways, sharing the encoder backbone but using different heads.

- `encoder_pred`: prediction-based (MSE bound on MI). A surrogate head $\phi(X) \to \hat{S}$ and an outcome head $(\phi(X), S) \to \hat{Y}$, both with one 64-wide hidden layer and ReLU. Loss $\mathcal{L} = \mathrm{MSE}(\hat{Y}, Y) + \lambda_S \cdot \mathrm{MSE}(\hat{S}, S)$.
- `mi_infonce_cond`: InfoNCE on the joint (Y, S) anchor. Projection heads $g_\phi : \phi(X) \to \mathbb{R}^p$ and $g_{YS} : (Y, S) \to \mathbb{R}^p$ feed a symmetric in-batch contrastive loss.
- `mi_mine_cond`: Donsker–Varadhan estimator on the joint (Y, S) anchor: a critic network $T_{(Y,S)}(\phi(X), (Y, S)) \to \mathbb{R}$ scores joint vs. shuffled pairs.

Both MI variants additionally include a marginal S-anchor term weighted by $(\lambda_S - 1)$, recovering the chain-rule decomposition of Section 4.3. We use $\lambda_S = 2$ throughout the reported experiments.

F.1 Synthetic dataset: details

F.1.1 Description

Data-generating process (full). We construct the ten-dimensional latent variable $Z = X\, W_{XZ}^\top$ from linear combinations of a small subset of the covariates. We split $Z = (Z_s, Z_y)$, where $Z_s, Z_y \in \mathbb{R}^5$ depend on two disjoint sets of ten covariates. We draw the coefficients defining these relationships once and hold them fixed across all data samples. In the main synthetic experiment in Section 5.1, we set $\alpha = \delta = 0$, so only Z_s affects the surrogates, outcome, and treatment effect. We use Z_y only in the sufficiency- and surrogacy-violation experiments.

The surrogate $S \in \mathbb{R}^{10}$ is a noisy linear function of Z_s, shifted by a Z_s-driven treatment effect:

$$S_i(t) = \beta_0 + Z_{s,i}\beta_1^\top + t \cdot \tau_S(Z_{s,i}) + \epsilon_{S,i}, \qquad \epsilon_{S,i} \sim \mathcal{N}(\mathbf{0}, \sigma_S^2 \mathbf{I}),$$

with $\tau_S(z) = \gamma_0 + z\gamma_z^\top$, $\beta_0 \sim \mathcal{N}(0, 0.5^2)$, $\beta_1 \sim \mathcal{N}(0, 1)$, $\gamma_0 \sim \mathcal{N}(0, 0.5^2)$, $\gamma_z \sim \mathcal{N}(0, 0.3^2)$, and $\sigma_S = 0.1$. The primary outcome depends on X only through S:

$$Y_i = S_i w_{SY} + \epsilon_{Y,i}, \qquad \epsilon_{Y,i} \sim \mathcal{N}(0, \sigma_Y^2),$$

with $w_{SY} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and $\sigma_Y = 0.1$. Both the baseline value of S and the treatment effect on S depend only on Z_s in the main experiment. For the sufficiency-violation experiment, we also allow Z_y to affect the treatment-induced change in S. We scale the Z_s and Z_y components so that they contribute equal variance to the scalar CATE; the weights $\sqrt{1-\alpha^2}$ and α therefore keep its variance fixed as α varies. For the surrogacy-violation experiment, we add a direct treatment effect on Y that depends on Z_y and scale it to contribute the same CATE variance as the mediated component at $\delta = 1$. The w_{SY} weights mapping $S \to Y$ are $\mathcal{N}(\mathbf{0}, \mathbf{I})$; observation noise is $\sigma_S = \sigma_Y = 0.1$ on both surrogate and outcome.

Method	$\hat{Y}$ target					Y target				
	$n = 50$	$n = 100$	$n = 250$	$n = 500$	$n = 750$	$n = 50$	$n = 100$	$n = 250$	$n = 500$	$n = 750$
MI-InfoNCE	1.141±0.048	**0.818±0.052**	0.587±0.028	0.500±0.012	0.474±0.008	**1.167±0.058**	**0.789±0.048**	0.508±0.028	0.395±0.012	0.357±0.008
MI-MINE	1.138±0.044	0.999±0.039	0.841±0.010	0.829±0.009	0.816±0.007	**1.167±0.058**	0.978±0.038	0.836±0.013	0.818±0.013	0.806±0.010
Encoder	1.141±0.057	1.000±0.040	0.853±0.020	0.833±0.018	0.824±0.018	**1.167±0.058**	0.789±0.054	**0.437±0.012**	**0.364±0.006**	**0.329±0.007**
PLS	1.145±0.046	0.832±0.041	**0.534±0.022**	**0.485±0.021**	**0.423±0.012**	**1.167±0.058**	0.809±0.042	0.498±0.017	0.441±0.018	0.393±0.008
PCA	1.159±0.052	1.313±0.060	1.542±0.047	1.529±0.039	1.454±0.019	**1.167±0.058**	1.314±0.059	1.564±0.049	1.551±0.031	1.469±0.022
ICA	1.156±0.050	1.286±0.045	1.498±0.027	1.494±0.029	1.424±0.018	**1.167±0.058**	1.299±0.043	1.517±0.024	1.516±0.025	1.447±0.018
Autoencoder	1.158±0.051	1.260±0.047	1.414±0.055	1.411±0.030	1.400±0.023	**1.167±0.058**	1.257±0.044	1.434±0.063	1.429±0.032	1.413±0.023
Raw X	**1.053±0.028**	1.177±0.040	1.091±0.024	1.068±0.015	1.038±0.011	**1.167±0.058**	1.134±0.044	0.997±0.015	0.892±0.011	0.813±0.010

Table 1: Normalized PEHE on the synthetic data, X-learner ($\dim(\phi) = 10$). Entries are means ± standard errors over 10 paired data resamples; lower is better. At $n = 50$ with the true Y target, the X-learner returns the same constant-effect estimate for every representation.

Method	$\hat{Y}$ target					Y target				
	$n = 50$	$n = 100$	$n = 250$	$n = 500$	$n = 750$	$n = 50$	$n = 100$	$n = 250$	$n = 500$	$n = 750$
MI-InfoNCE	**-0.006±0.085**	0.611±0.062	0.851±0.015	0.890±0.010	0.911±0.005	0.012±0.084	0.659±0.042	0.881±0.020	**0.931±0.010**	0.944±0.005
MI-MINE	-0.117±0.109	0.451±0.053	0.594±0.022	0.590±0.023	0.613±0.013	**0.084±0.111**	0.496±0.048	0.611±0.024	0.610±0.020	0.634±0.010
Encoder	-0.075±0.140	0.321±0.075	0.632±0.079	0.709±0.063	0.773±0.060	0.060±0.102	**0.691±0.060**	**0.908±0.008**	0.928±0.005	**0.950±0.003**
PLS	-0.144±0.105	**0.618±0.070**	**0.858±0.009**	**0.894±0.009**	**0.915±0.008**	0.072±0.084	0.637±0.066	0.872±0.011	0.902±0.012	0.923±0.006
PCA	-0.032±0.081	0.055±0.059	0.076±0.029	0.086±0.023	0.101±0.021	-0.031±0.050	0.049±0.057	0.080±0.031	0.088±0.029	0.120±0.011
ICA	-0.008±0.084	0.079±0.045	0.118±0.025	0.117±0.025	0.135±0.022	0.021±0.070	0.085±0.040	0.135±0.030	0.114±0.026	0.135±0.018
Autoencoder	-0.015±0.080	0.020±0.041	0.141±0.035	0.148±0.023	0.142±0.015	-0.180±0.063	0.015±0.046	0.142±0.034	0.161±0.020	0.134±0.023
Raw X	-0.120±0.073	-0.049±0.075	0.043±0.065	0.000±0.038	0.024±0.039	-0.020±0.082	0.348±0.041	0.533±0.028	0.611±0.016	0.684±0.012

Table 2: Normalized top-20% policy value on the synthetic data, X-learner ($\dim(\phi) = 10$). Entries are means ± standard errors over 10 paired data resamples; higher is better.

Methods, training, and evaluation. We compare our three representations (`encoder_pred` – MSE prediction heads on S and Y, `mi_mine_cond` – MINE on the joint (Y, S) anchor, and `mi_infonce_cond` – InfoNCE on the joint (Y, S) anchor) against `pls` (supervised linear), `autoencoder`, `pca`, and `ica` (unsupervised), and two raw-feature comparisons with $\phi(X) \equiv X$, whose outcome models fit either of $h(X)$ or $h(X, S)$. Neural encoders use the MLP backbone $1000 \to 128 \to 64 \to 32 \to \phi$ with ReLU activations, trained with Adam (lr 10^{-3}, batch 256, up to 200 epochs, early stopping on training loss with patience 15 and min-delta 10^{-4}); the surrogate-prediction weight is $\lambda_S = 2$, and the InfoNCE temperature is 0.07. Every reported synthetic experiment uses a 50-tree random forest for the outcome prediction model trained on historical data and `HistGradientBoostingRegressor` with 40 iterations inside the CATE learners. We report both X-learner and DML using either the predicted outcome $\hat{Y}$ or the observed outcome Y as the CATE target.

Metrics, seeds, and compute. We report normalized PEHE $\sqrt{\mathbb{E}[(\hat{\tau} - \tau)^2]}/\mathrm{std}(\tau)$ (scale-invariant across DGPs) and normalized top-k policy value $(V_\pi - V_{\mathrm{random}})/(V_{\mathrm{oracle}} - V_{\mathrm{random}})$, where V_π is the mean of τ_{true} over the top-k units ranked by $\hat{\tau}$, V_{random} is the population mean, and V_{oracle} is V_π with $\hat{\tau}$ replaced by τ_{true}. By construction 0 corresponds to a random ranker and 1 to the oracle policy. All synthetic experiments fix one DGP (`dgp_seed` $= 42$) and average over the same 10 historical and experimental resamples. The main synthetic and violation experiments use $n \in \{50, 100, 250, 500, 750\}$; error bars show ± 1 standard error.

F.1.2 Additional results: figures

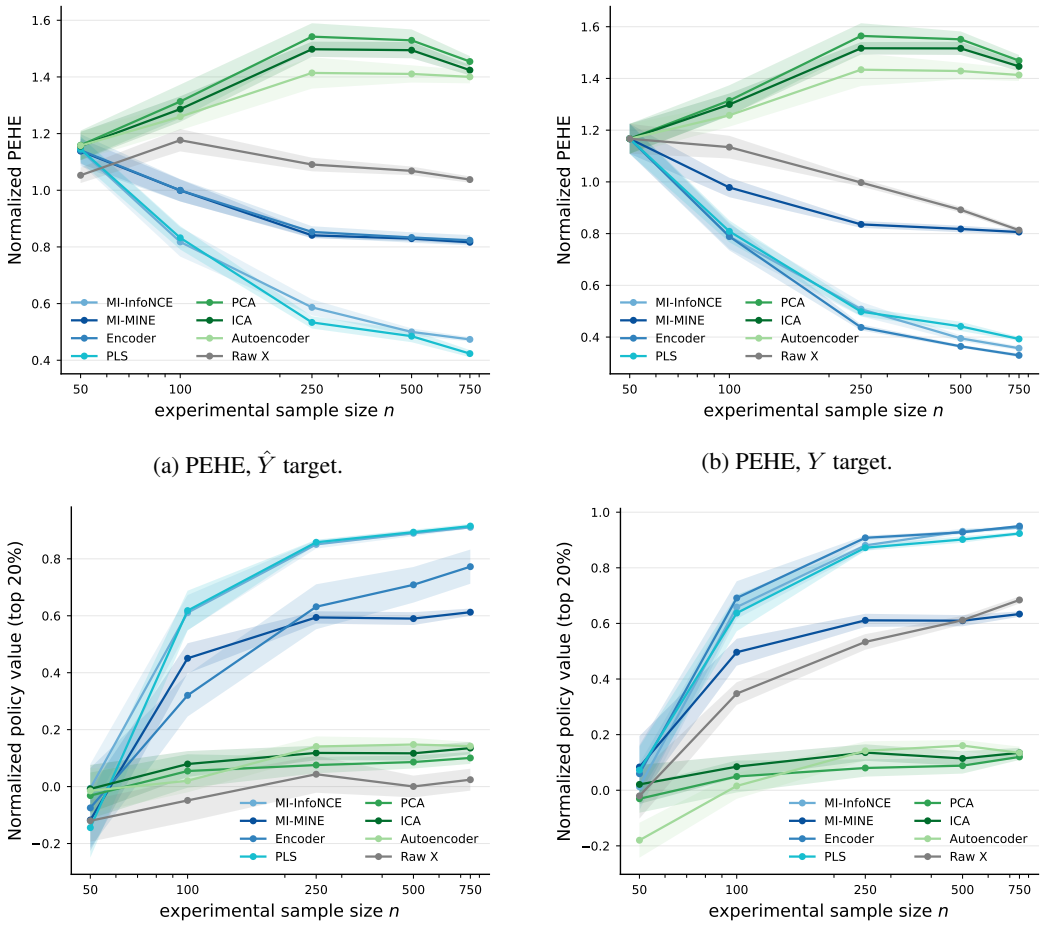
Figure 8 reports the X-learner results across both CATE targets ($\hat{Y}$ and Y); Figure 9 reports the corresponding DML results. Every panel uses the same DGP, sample-size grid, methods, and 10 historical and experimental resamples as the main analysis.

F.1.3 Additional results: tabulated PEHE and top-20% policy value

Tables 1–4 report X-learner and DML results for the eight methods evaluated under both CATE targets. All entries use the same 10 historical and experimental resamples.

F.2 Empirical analysis of representation dimension (m)

We probe the sensitivity of the estimators to the representation dimension $m = \dim(\phi(X))$ at fixed latent dimensionality $z_{\mathrm{dim}} = 10$. By construction, only the five coordinates in Z_s affect the surrogates and outcome, so $m = 5$ is the smallest dimension that can support exact sufficiency. We sweep



(a) PEHE, $\hat{Y}$ target.

(b) PEHE, Y target.

(c) Policy@20%, $\hat{Y}$ target.

(d) Policy@20%, Y target.

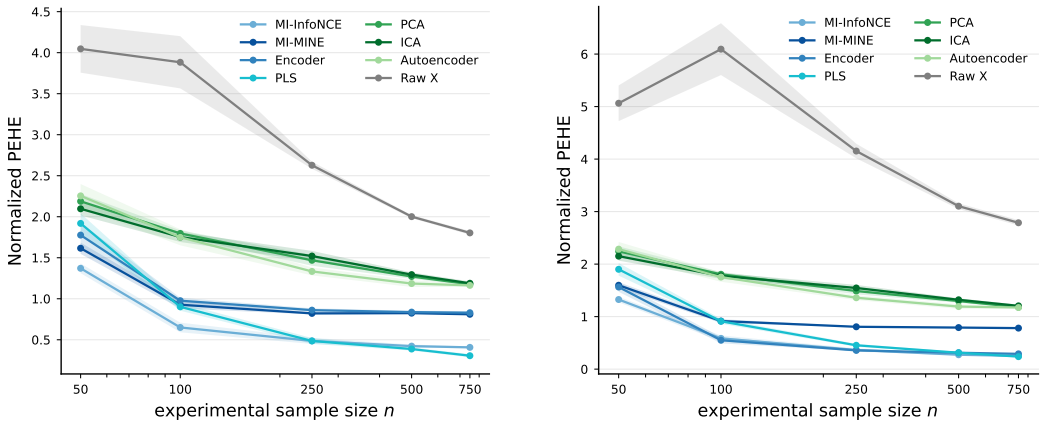
Figure 8: **Synthetic data** results using the X-learner. Rows: error metric (PEHE; normalized top-20% policy value). Columns: CATE target ($\hat{Y}$ vs. Y). Curves show means over 10 historical and experimental resamples from the same DGP; shaded regions show ± 1 standard error.

Method	$\hat{Y}$ target					Y target				
	$n = 50$	$n = 100$	$n = 250$	$n = 500$	$n = 750$	$n = 50$	$n = 100$	$n = 250$	$n = 500$	$n = 750$
MI-InfoNCE	**1.371±0.059**	**0.650±0.059**	0.486±0.037	0.423±0.019	0.408±0.011	**1.326±0.053**	0.586±0.056	0.364±0.032	**0.274±0.012**	0.252±0.010
MI-MINE	1.616±0.069	0.928±0.046	0.822±0.013	0.825±0.008	0.811±0.008	1.598±0.059	0.916±0.036	0.806±0.008	0.791±0.008	0.781±0.007
Encoder	1.776±0.151	0.977±0.036	0.861±0.023	0.837±0.017	0.830±0.019	1.563±0.091	**0.550±0.060**	**0.357±0.013**	0.312±0.008	0.291±0.005
PLS	1.919±0.123	0.902±0.047	**0.486±0.036**	**0.387±0.019**	**0.306±0.020**	1.901±0.120	0.911±0.048	0.454±0.028	0.313±0.014	**0.238±0.011**
PCA	2.188±0.089	1.796±0.037	1.469±0.067	1.271±0.028	1.177±0.016	2.243±0.094	1.806±0.042	1.487±0.063	1.296±0.034	1.180±0.015
ICA	2.097±0.081	1.752±0.060	1.521±0.063	1.295±0.027	1.189±0.023	2.151±0.086	1.776±0.057	1.546±0.059	1.321±0.028	1.205±0.023
Autoencoder	2.254±0.144	1.750±0.098	1.333±0.048	1.185±0.042	1.163±0.020	2.285±0.137	1.753±0.089	1.360±0.045	1.191±0.044	1.173±0.018
Raw X	4.047±0.290	3.883±0.319	2.628±0.043	2.001±0.019	1.803±0.022	5.063±0.340	6.094±0.494	4.153±0.139	3.103±0.055	2.787±0.051

Table 3: Normalized PEHE on the synthetic data, DML ($\dim(\phi) = 10$). Entries are means $\pm$ standard errors over 10 paired data resamples; lower is better.

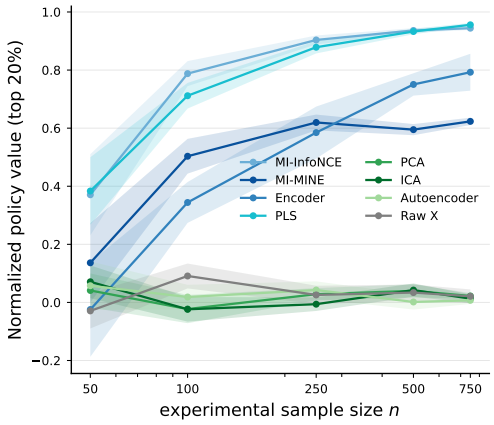
$m \in \{2, 3, 5, 10, 20, 50\}$ at the canonical setting $\alpha = \delta = 0$ with $n \in \{50, 100, 250, 500, 750\}$, using three-fold DML and both CATE targets; results are means $\pm$ one standard error over the same 10 historical and experimental resamples as the main synthetic experiment. Raw X does not depend on m and provides a constant reference. All $4{,}800$ DML fits completed successfully.

Results. Figures 10 and 11 report normalized PEHE and policy value across representation dimensions, respectively. The results show a familiar underfitting vs. estimation tradeoff. At $n = 500$, increasing m from 2 to 5 lowers MI-InfoNCE's normalized PEHE from 0.961 to 0.538 and raises its policy value from 0.341 to 0.940. Increasing m beyond 5 provides no uniform gain: MI-InfoNCE

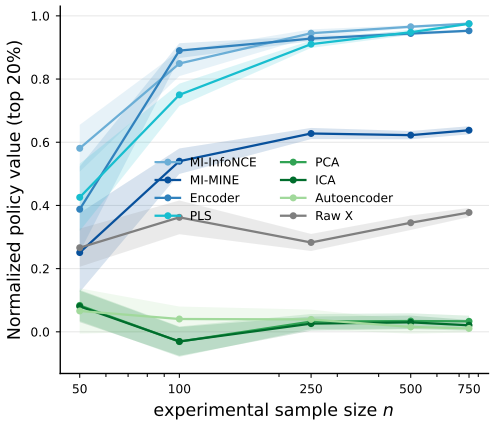


(a) PEHE, $\hat{Y}$ target.

(b) PEHE, Y target.



(c) Policy@20%, $\hat{Y}$ target.



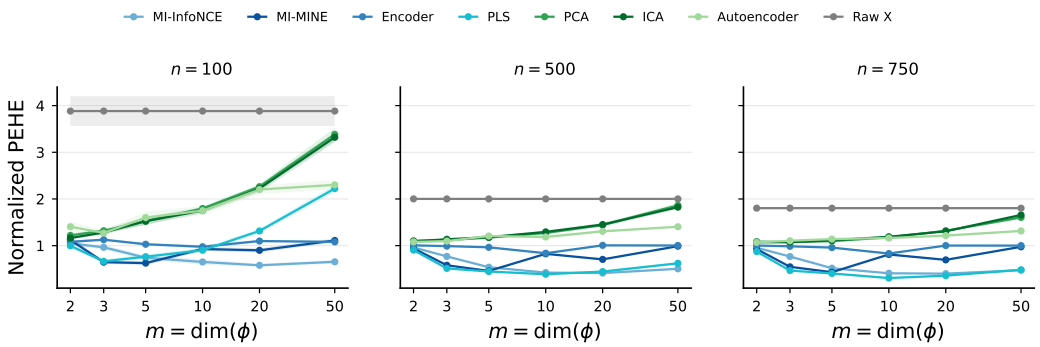
(d) Policy@20%, Y target.

Figure 9: **Synthetic data** results using the DML learner using predicted and true outcomes as CATE targets. Curves show means over the same 10 historical and experimental resamples as the main analysis; shaded regions show ± 1 standard error.

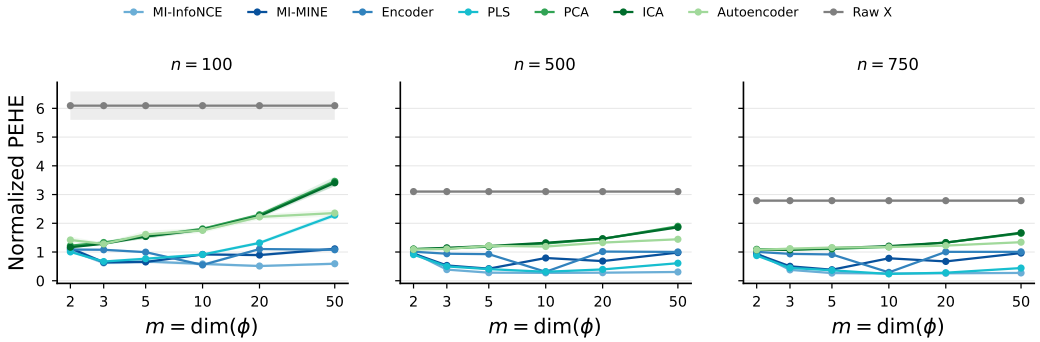
Method	$\hat{Y}$ target					Y target				
	$n = 50$	$n = 100$	$n = 250$	$n = 500$	$n = 750$	$n = 50$	$n = 100$	$n = 250$	$n = 500$	$n = 750$
MI-InfoNCE	0.371±0.140	**0.788±0.043**	**0.904±0.015**	**0.936±0.008**	0.944±0.006	**0.581±0.074**	0.849±0.040	**0.945±0.012**	**0.966±0.005**	**0.976±0.002**
MI-MINE	0.136±0.136	0.503±0.060	0.620±0.027	0.595±0.019	0.623±0.013	0.251±0.125	0.540±0.040	0.628±0.018	0.622±0.013	0.638±0.013
Encoder	-0.025±0.163	0.344±0.070	0.585±0.090	0.750±0.039	0.793±0.064	0.388±0.135	**0.890±0.023**	0.928±0.008	0.944±0.006	0.953±0.004
PLS	**0.384±0.116**	0.712±0.044	0.879±0.021	0.933±0.009	**0.956±0.007**	0.426±0.107	0.750±0.035	0.910±0.012	0.948±0.008	0.975±0.002
PCA	0.041±0.059	-0.023±0.049	0.028±0.021	0.040±0.023	0.022±0.010	0.084±0.049	-0.031±0.048	0.032±0.024	0.034±0.024	0.033±0.018
ICA	0.071±0.056	-0.024±0.042	-0.006±0.023	0.042±0.023	0.013±0.015	0.080±0.050	-0.030±0.046	0.026±0.023	0.030±0.021	0.021±0.017
Autoencoder	0.058±0.083	0.019±0.042	0.043±0.030	0.002±0.026	0.007±0.012	0.066±0.072	0.041±0.039	0.039±0.031	0.015±0.021	0.011±0.013
Raw X	-0.029±0.061	0.091±0.043	0.026±0.032	0.034±0.028	0.021±0.024	0.266±0.060	0.362±0.053	0.283±0.027	0.345±0.023	0.378±0.015

Table 4: Normalized top-20% policy value on the synthetic data, DML ($\dim(\phi) = 10$). Entries are means $\pm$ standard errors over 10 paired data resamples; higher is better.

remains consistently good at $m = 10$ and 20, whereas PLS begins to degrade at $m = 50$ and the prediction encoder and MI-MINE are more sensitive to the choice of dimension. The main choice $m = 10$ therefore lies within the effective range for the best-performing methods, but increasing representation dimensions has diminishing marginal returns, and for some methods, at small sample sizes, is even detrimental.



(a) Normalized PEHE, $\hat{Y}$ target.



(b) Normalized PEHE, Y target.

Figure 10: **Representation dimension sweep, DML — normalized PEHE.** Increasing m through the outcome-relevant latent dimension substantially improves the best supervised methods. Beyond that point, additional dimensions provide no uniform benefit and can increase error for some methods.

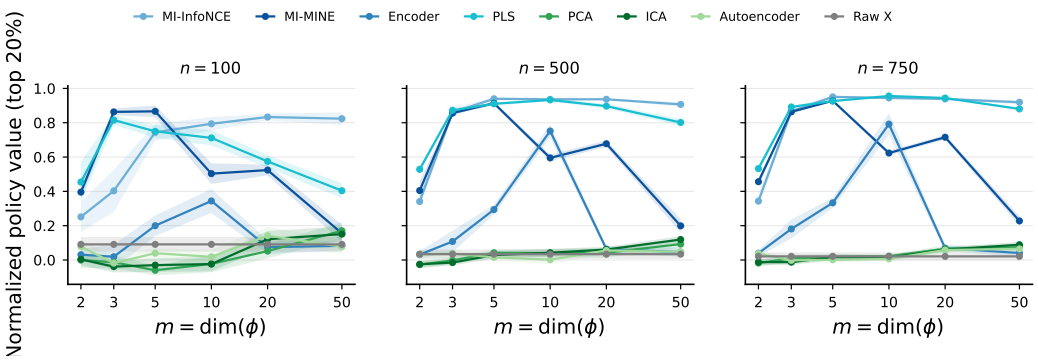
F.3 Semi-synthetic (Medical) dataset

F.3.1 Construction

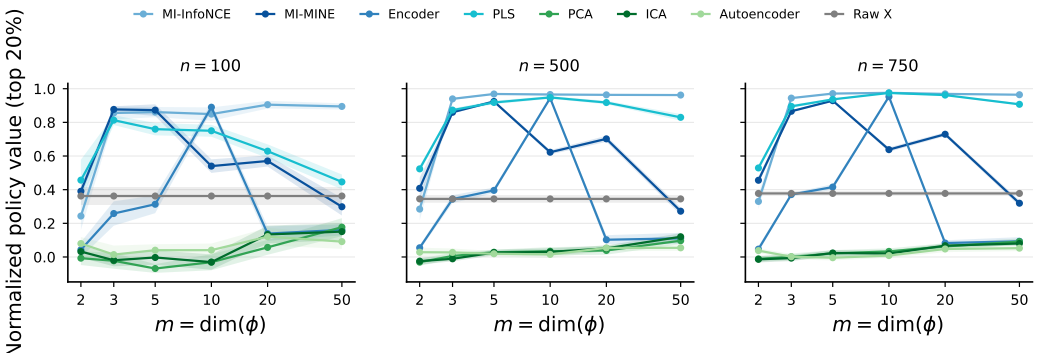
Data disclosure and integrity checks. The empirical records underlying this semi-synthetic dataset were anonymized before access and were obtained under IRB approval. We use them to obtain covariate, diagnosis, and outcome distributions. We synthetically generate treatment assignment and potential outcomes.

Cohorts and covariates. The historical (EHR) cohort H has $N_H = 60{,}248$ samples. The experimental pool (mobile-app cohorts) E contains $N_E = 11{,}747$ samples. We retain covariates X after excluding identifiers, variables that may lead to leakage of targets, features with excessive missingness, and all 37 diagnosis trajectories used to construct S. The resulting $X \in \mathbb{R}^{340}$ spans maternal demographics and insurance; obstetric history and pregnancy characteristics; prenatal measurements and health behaviors; diagnoses recorded across pregnancy; and emergency, inpatient, prenatal, and behavioral-health utilization. We use an ℓ_1-logistic model fit on H to identify 40 relevant coordinates with high signal used to parameterize the semi-synthetic outcome and treatment mechanisms.

Auxiliary outcomes. We define a pre-specified pool of 56 diagnosis and problem-list trajectories from the first, second, and third trimesters and the post-baseline/pre-postpartum window. These include anxiety, depression, hypertension, gestational hypertension, substance-use and behavioral diagnoses, bipolar disorder, obsessive-compulsive disorder, trauma reactions, diabetes, autoimmune conditions, cardiomyopathy, kidney and liver conditions, and gestational diabetes. To deal with missingness, we retain only 37 outcomes whose prevalence in H lies between 0.5% and 99.5%. They are standardized using H only, after which we fit PCA to obtain the surrogates. The first five



(a) Normalized policy value (top 20%), $\hat{Y}$ target.



(b) Normalized policy value (top 20%), Y target.

Figure 11: **Representation dimension sweep, DML — normalized top-20% policy value.** MI-InfoNCE and PLS recover most of the available ranking signal once m is large enough to represent the outcome-relevant latent structure. Unsupervised methods remain near random across dimensions.

components define $S \in \mathbb{R}^5$ and explain 41.3% of the historical outcome variance. These auxiliary outcomes are observed in both cohorts.

Synthetic treatment and potential outcomes. The baseline outcome score μ_X and auxiliary-outcome direction β_S are fit using H only. Let $S_i(0)$ denote the observed untreated auxiliary-outcome representation after a shared measurement-noise draw. Treatment shifts it along the outcome-relevant direction,

$$\tau_S(X_i) = g(X_i)\frac{\beta_S}{\|\beta_S\|_2^2}, \qquad S_i(t) = S_i(0) + t\,\tau_S(X_i),$$

where $g(X)$ is a heterogeneous effect score depending on 10 of the 40 selected coordinates and normalized to have standard deviation 0.5. Let $D_i(t) = 1$ denote postpartum depression. Its potential-outcome probability satisfies

$$\operatorname{logit} P\{D_i(t) = 1 \mid X_i, S_i(t)\} = \mu_X(X_i) + \lambda_S S_i(t)^\top \beta_S, \qquad \lambda_S = 2.$$

We report $Y_i(t) = 1 - D_i(t)$, so positive treatment effects correspond to a reduced probability of postpartum depression. The same equation is used for H and E. Experimental treatment is $T \sim \text{Bernoulli}(1/2)$, independently of X, and the observed pair is $(S_i(T_i), Y_i(T_i))$. Since T affects Y only through $S(T)$, surrogacy holds by construction. Using the corresponding noisy potential surrogate in each potential-outcome equation avoids a measurement-error-induced direct association between T and Y conditional on the observed S.

Methods and evaluation. We standardize X using H, then train representations and 50-tree random-forest outcome prediction models on an i.i.d. historical sample of size 10,000. Neural methods use the encoder $340 \to 64 \to 32 \to \dim(\phi)$, Adam with learning rate 10^{-3} and batch size 256, and validation-based early stopping over at most 200 epochs. The main analysis uses

Cohort overlap			
Cohort-classification AUC	0.752		
NN coverage in (X_Y, S), $T = 0$ / $T = 1$	0.940 / 0.951		
Representation sufficiency			
Representation	Outcome Δ log loss	$S(0)$ ΔR^2	$S(1)$ ΔR^2
Prediction encoder	0.0065	0.0146	0.0379
Conditional MINE	0.0089	0.0501	0.0772
Conditional InfoNCE	0.0102	0.0568	0.0837
PLS	0.0018	0.0261	0.0443
PCA	0.0083	0.1656	0.1868
ICA	0.0102	0.1653	0.1853
Autoencoder	0.0104	0.1747	0.1902

Table 5: Diagnostics for cohort overlap and representation sufficiency. The final three columns report the cross-validated improvement from appending raw X to the representation; smaller values indicate that less predictive information remains beyond the representation.

$\dim(\phi) = 10$; $\dim(\phi) = 5$ is a sensitivity analysis. We evaluate the prediction encoder, conditional MINE, conditional InfoNCE, PLS, PCA, ICA, an autoencoder, and two raw-X variants whose outcome models fit $h(X)$ or $h(X, S)$. For each seed, we reserve 5,000 of the 11,747 experimental-pool pregnancies for testing and draw every training sample from the disjoint remaining pool. We standardize the CATE features using each experimental training sample, remove constant coordinates, and clip standardized training and test values to $[-5, 5]$. We then fit three-fold DML with 40-iteration histogram-gradient-boosted outcome and treatment nuisances and a ridge final stage. Experimental sample sizes are $n \in \{100, 250, 500, 750, 1000\}$. We repeat each experiment on 10 random data draws, using the same draws for every method, and report the mean and its standard error. We report normalized PEHE and normalized top-20% policy value.

F.3.2 Overlap and representation diagnostics

Unlike the fully synthetic experiment, the historical and experimental samples come from different empirical cohorts. Thus randomization, surrogacy, and the outcome-model comparability hold by construction, whereas overlap between the two cohorts and the sufficiency of the learned representation need not hold exactly. Hence we empirically assess overlap between H and E and whether raw X retains predictive information not retained by $\phi(X)$. We summarize the empirical assessment in Table 5.

We assess overlap between the historical and experimental samples using a cross-validated classifier trained to distinguish the two cohorts based on X. The classifier has an AUC of 0.752, indicating some distribution shift. We then compare nearest-neighbor distances in (X_Y, S), where $X_Y \in \mathbb{R}^{40}$ are the covariates used to construct the semi-synthetic outcomes. We do not use the full covariates $X \in \mathbb{R}^{340}$ since Euclidean nearest-neighbor distance becomes increasingly difficult to interpret in higher dimensions. We split our historical samples into a reference set and a held-out set. For each held-out historical sample, we compute its distance to the nearest-neighbor in the reference set. The 95-th percentile of these distances defines our coverage threshold. We consider an experimental sample to be "covered" when its distance to historical reference set does not exceed this threshold. Coverage is 94.0% among controls and 95.1% among treated observations, indicating substantial but imperfect overlap in the variables used by the DGP.

To assess sufficiency (i), we test whether raw X improves prediction of Y^* after conditioning on $(\phi(X), S)$. We compare three-fold cross-validated predictions using $(\phi(X), S)$ and $(\phi(X), S, X)$ and report the resulting reduction in log loss. To assess sufficiency (ii), we similarly compare predictions of $S(0)$ and $S(1)$ using $\phi(X)$ and $(\phi(X), X)$ and report the increase in R^2. Both potential surrogates are available by construction in the semi-synthetic experiment. Under exact sufficiency, adding X would provide no population-level improvement. Values near zero therefore indicate that little predictive information remains outside the representation.

For sufficiency (i), adding X reduces outcome log loss by at most 0.0104 across all representations. For sufficiency-(ii), the improvements for the supervised representations range from 0.0146 to

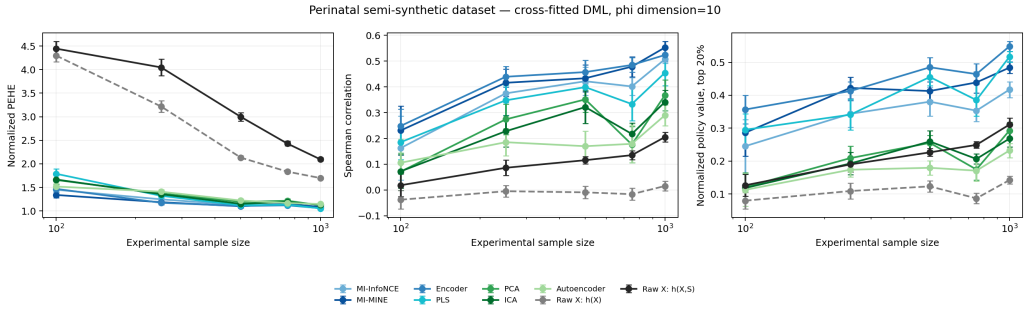


Figure 12: The nine methods used in Section 5.2 on the perinatal semi-synthetic dataset at $\dim(\phi) = 10$. Points show means over 10 runs; error bars show ± 1 standard error. Supervised representations learned from historical outcomes outperform both raw-X variants and generally outperform unsupervised compression.

Method	$n = 100$	$n = 250$	$n = 500$	$n = 750$	$n = 1000$
MI-InfoNCE	1.451±0.076	1.247±0.047	1.124±0.031	1.133±0.020	1.090±0.019
MI-MINE	**1.340±0.062**	1.186±0.052	**1.099±0.035**	1.152±0.039	1.082±0.022
Encoder	1.470±0.068	**1.173±0.034**	1.106±0.026	**1.119±0.028**	1.059±0.017
PLS	1.790±0.102	1.322±0.062	1.118±0.030	1.135±0.031	**1.055±0.015**
PCA	1.659±0.093	1.372±0.061	1.193±0.049	1.216±0.036	1.123±0.022
ICA	1.668±0.102	1.356±0.044	1.151±0.046	1.192±0.023	1.130±0.017
Autoencoder	1.518±0.065	1.408±0.044	1.221±0.034	1.170±0.023	1.146±0.019
Raw X, $h(X)$	4.299±0.137	3.217±0.122	2.133±0.035	1.835±0.033	1.699±0.026
Raw X, $h(X, S)$	4.446±0.154	4.047±0.175	2.998±0.093	2.431±0.050	2.096±0.034

Table 6: Normalized PEHE on the perinatal semi-synthetic dataset at $\dim(\phi) = 10$. Entries are means $\pm$ standard errors over 10 runs; lower is better.

0.0837, compared with 0.1653 to 0.1902 for PCA, ICA, and the autoencoder. The prediction encoder retains the most information relevant to the potential surrogates, with PLS close behind it. These improvements are close to zero, particularly for the supervised representations, suggesting that raw X contributes little predictive information beyond $\phi(X)$ and that sufficiency holds approximately in this experiment.

F.3.3 Results and representation dimension

Figure 12 and Tables 6–7 report the complete results at $\dim(\phi) = 10$. Figure 13 reports the corresponding results at $\dim(\phi) = 5$, while Table 8 directly compares the prediction encoder at dimensions 5 and 10.

For the prediction encoder, dimensions 5 and 10 give similar results. The ten-dimensional representation has lower normalized PEHE at four of the five sample sizes, although the differences are small, and neither dimension has uniformly higher policy value. We retain $\dim(\phi) = 10$ for consistency with the synthetic experiment and report dimension 5 as a sensitivity analysis.

F.3.4 True-outcome sensitivity

Figure 14 and Tables 9–10 report results using the true experimental outcome Y. To separate representation quality from outcome-prediction error, we repeat the same out-of-sample DML evaluation using the true experimental Y directly. At $n = 100$, conditional MINE has the lowest normalized PEHE (1.717 ± 0.108), compared with 6.833 ± 0.323 for raw X. At $n = 1000$, PLS and the prediction encoder are nearly tied (1.059 ± 0.039 and 1.063 ± 0.023), while raw X remains at 3.414 ± 0.062. The advantage of low-dimensional supervised representations therefore persists when the true experimental outcome, rather than $\widehat{Y}$, is used as the CATE target.

Method	$n = 100$	$n = 250$	$n = 500$	$n = 750$	$n = 1000$
MI-InfoNCE	0.245±0.081	0.344±0.040	0.381±0.043	0.353±0.033	0.417±0.024
MI-MINE	0.286±0.072	**0.422±0.033**	0.413±0.028	0.439±0.032	0.484±0.017
Encoder	**0.357±0.044**	0.413±0.028	**0.485±0.029**	**0.464±0.031**	**0.549±0.015**
PLS	0.295±0.048	0.341±0.047	0.456±0.027	0.386±0.049	0.517±0.016
PCA	0.121±0.043	0.209±0.036	0.255±0.023	0.175±0.032	0.293±0.025
ICA	0.118±0.042	0.193±0.034	0.260±0.033	0.207±0.015	0.268±0.025
Autoencoder	0.113±0.050	0.174±0.022	0.180±0.022	0.171±0.032	0.232±0.022
Raw X, $h(X)$	0.080±0.026	0.109±0.024	0.123±0.017	0.087±0.016	0.142±0.012
Raw X, $h(X, S)$	0.127±0.033	0.190±0.023	0.227±0.011	0.249±0.010	0.311±0.020

Table 7: Normalized top-20% policy value on the perinatal semi-synthetic dataset at $\dim(\phi) = 10$. Entries are means ± standard errors over 10 runs; higher is better.

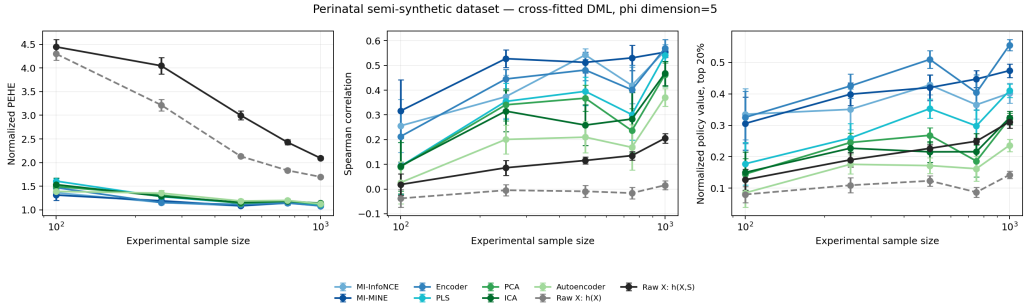


Figure 13: The nine methods used in the main comparison at $\dim(\phi) = 5$. The prediction encoder gives similar PEHE and policy value at dimensions 5 and 10.

	Normalized PEHE		Policy@20%	
n	$\dim(\phi) = 5$	$\dim(\phi) = 10$	$\dim(\phi) = 5$	$\dim(\phi) = 10$
100	1.471±0.105	1.470±0.068	0.325±0.083	0.357±0.044
250	1.153±0.025	1.173±0.034	0.425±0.039	0.413±0.028
500	1.114±0.029	1.106±0.026	0.510±0.028	0.485±0.029
750	1.148±0.029	1.119±0.028	0.404±0.056	0.464±0.031
1000	1.084±0.025	1.059±0.017	0.555±0.018	0.549±0.015

Table 8: Representation-dimension sensitivity for the prediction encoder on the perinatal semi-synthetic dataset. Entries are means ± standard errors over 10 runs.

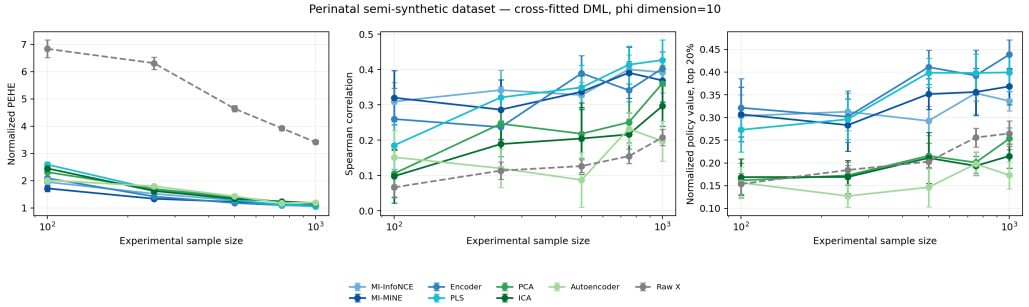


Figure 14: The eight distinct methods at $\dim(\phi) = 10$ using true experimental Y as the CATE target. The $h(X, S)$ raw-X variant is omitted because it is identical to raw X when the true outcome is used directly. Points show means over 10 runs; error bars show ±1 standard error.

Method	$n = 100$	$n = 250$	$n = 500$	$n = 750$	$n = 1000$
MI-InfoNCE	1.952±0.117	1.522±0.077	1.282±0.041	**1.096±0.026**	1.104±0.043
MI-MINE	**1.717±0.108**	**1.339±0.051**	1.224±0.059	1.121±0.037	1.088±0.023
Encoder	2.094±0.153	1.420±0.078	**1.183±0.032**	1.102±0.023	1.063±0.023
PLS	2.594±0.088	1.686±0.140	1.292±0.058	1.116±0.023	**1.059±0.039**
PCA	2.315±0.172	1.698±0.093	1.400±0.031	1.214±0.027	1.157±0.035
ICA	2.451±0.168	1.624±0.096	1.340±0.039	1.237±0.031	1.181±0.037
Autoencoder	2.018±0.076	1.797±0.060	1.427±0.035	1.179±0.057	1.202±0.029
Raw X	6.833±0.323	6.309±0.225	4.638±0.113	3.922±0.080	3.414±0.062

Table 9: Normalized PEHE on the perinatal semi-synthetic dataset at $\dim(\phi) = 10$ using true experimental Y. Entries are means ± standard errors over 10 runs; lower is better.

Method	$n = 100$	$n = 250$	$n = 500$	$n = 750$	$n = 1000$
MI-InfoNCE	0.303±0.047	**0.313±0.041**	0.293±0.032	0.354±0.046	0.336±0.022
MI-MINE	0.307±0.060	0.283±0.057	0.352±0.034	0.356±0.052	0.368±0.040
Encoder	**0.321±0.064**	0.302±0.056	**0.411±0.038**	0.392±0.056	**0.438±0.032**
PLS	0.273±0.049	0.296±0.045	0.399±0.032	**0.398±0.042**	0.399±0.043
PCA	0.162±0.038	0.173±0.022	0.216±0.028	0.201±0.023	0.253±0.024
ICA	0.169±0.040	0.169±0.036	0.211±0.056	0.193±0.021	0.215±0.026
Autoencoder	0.157±0.035	0.127±0.025	0.146±0.043	0.198±0.026	0.173±0.030
Raw X	0.154±0.023	0.184±0.018	0.203±0.013	0.256±0.021	0.265±0.028

Table 10: Normalized top-20% policy value on the perinatal semi-synthetic dataset at $\dim(\phi) = 10$ using true experimental Y. Entries are means ± standard errors over 10 runs; higher is better.