Is Merging Worth It? Securely Evaluating the Information Gain for Causal Dataset Acquisition

Jake Fawkes[*]
University of Oxford

Lucile Ter-Minassian[*]
University of Oxford

Desi R. Ivanova
University of Oxford

Uri Shalit
Technion - Israel Institute of Technology

Chris Holmes
University of Oxford

Abstract

Merging datasets across institutions is a lengthy and costly procedure, especially when it involves private information. Data hosts may therefore want to prospectively gauge which datasets are most beneficial to merge with, without revealing sensitive information. For causal estimation this is particularly challenging as the value of a merge depends not only on reduction in epistemic uncertainty but also on improvement in overlap. To address this challenge, we introduce the first *cryptographically secure* information-theoretic approach for quantifying the value of a merge in the context of heterogeneous treatment effect estimation. We do this by evaluating the *Expected Information Gain* (EIG) using multi-party computation to ensure that no raw data is revealed. We further demonstrate that our approach can be combined with differential privacy (DP) to meet arbitrary privacy requirements whilst preserving more accurate computation compared to DP alone. To the best of our knowledge, this work presents the first privacy-preserving method for dataset acquisition tailored to causal estimation. Code is publicly available: `https://github.com/LucileTerminassian/causal_prospective_merge`.

[*]Correspondence to: `jake.fawkes@st-hughs.ox.ac.uk` and `lucile.ter-minassian@spc.ox.ac.uk`.

Proceedings of the 28[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2025, Mai Khao, Thailand. PMLR: Volume 258.

1 INTRODUCTION

As the demand for data-driven decision making grows, the question of how to optimally collect data for a given task becomes increasingly important. Data fusion (Castanedo et al., 2013), which integrates pre-existing data from various sources, is a popular method to increase sample size, reduce sampling variability, and enhance statistical power and robustness (Lenzerini, 2002; Doan et al., 2012). However, merging datasets is often a time-consuming and resource-intensive task. This is especially true in sensitive domains such as healthcare, where concerns surrounding privacy, downstream applications, and data security mean long ethical approval procedures are required before undergoing a merge (Platt and Kardia, 2015; Mello et al., 2013; European Commission, 2018). Consequently, practitioners crucially need methods to determine the value of a potential merge in advance, whilst also complying with privacy requirements (Abouelmehdi et al., 2018).

In this work, we focus on data fusion in the context of heterogeneous treatment effect estimation. As a concrete example of this problem, consider a hospital that wishes to assess the impact of a medical intervention on its patients. Upon finding that its own data is insufficient to get accurate estimates, the hospital plans to select one of K possible candidate hospitals for a potential data merge.

Given the costs involved, the hospital would like to identify *in advance* which potential dataset would provide the most information upon merging, *whilst* complying with privacy regulations. We propose a solution for such types of problems, allowing for scenarios where patient outcomes at the candidate hospitals to be unobserved or simply masked.

We quantify the value of a merge in a principled, information theoretic way by applying techniques from

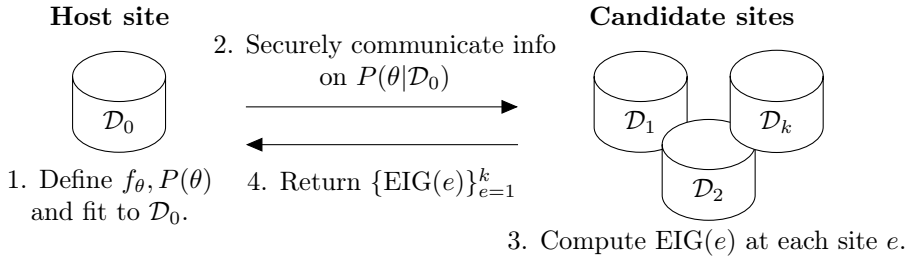


Figure 1: Flow chart depicting our method. In step 1 the *host site* chooses a parameterised model class for the conditional outcome, given by f_θ and sets the prior $P(\theta)$. They then perform a local Bayesian update, and communicate information on the posterior $P(\theta|\mathcal{D}_0)$ to the *candidate site* using secure multi party computation (MPC). The candidate applies MPC once more to privately calculate the Expected Information Gain (EIG) and communicate it back to host. This allows the host to select the best merge out of the potential candidates.

Bayesian experimental design (Rainforth et al., 2024; Lindley, 1956; Chaloner and Verdinelli, 1995). Our solution shares similarities with standard Bayesian dataset acquisition (MacKay, 1992; Kirsch et al., 2019), however in causal contexts data serve an additional, distinct purpose. Specifically acquired data should not only enhance our understanding of the outcome function, but also assist in combating the selection bias that is inherent to causal effect estimation (Holland, 1986) by balancing treatment. Put differently, whilst generic data fusion aims at reducing the epistemic uncertainty from incomplete knowledge of the outcome, causal estimation also seeks to improve treatment overlap (Rubin, 1997; Crump et al., 2009), i.e. reducing the epistemic uncertainty for counterfactual outcomes. To resolve this, we make use of favourable parameterisations available in popular Bayesian causal inference methods (Hahn et al., 2020; Alaa and Van Der Schaar, 2017), which allows us to prioritise information gain in the parameters relevant for the causal problem, rather than gaining information in irrelevant parts of the conditional outcome.

To ensure privacy, we employ Secure Multi-Party Computation (Yao, 1982; Evans et al., 2018; Knott et al., 2021). This cryptographic protocol enables multiple parties to jointly compute the output of a function without revealing any of their own private inputs. A classic example involves determining the wealthiest person in a group without anyone revealing their personal net worth. In our context it allows the different candidate sites to compute their expected information gains relative to the initial sites' data, without exposing the contents of their datasets. This ensures that the noise required for privacy guarantees can be added to the final statistic as opposed to the raw data at each site.

Our contributions can be summarised as follows:

- We propose information-theoretic methods to measure the value of a data merge in the context of heterogeneous treatment effects estimation. Our primary contribution is a novel approach that specifically targets the reduction of entropy in parameters that directly influence the conditional average treatment effect (CATE). We also present a standard approach based on expected entropy reduction in all parameters of a conditional outcome model.
- We demonstrate how both of the approaches can be used with three popular CATE estimators; Bayesian Polynomial Regression (Gelman et al., 2021), Causal Multitask Gaussian Processes (Alaa and Van Der Schaar, 2017), and Bayesian Causal Forests (Hahn et al., 2020). We derive closed form expressions for the expected entropy reduction in the first two models and give a Monte Carlo estimator for the other.
- We provide a privacy protocol for our methods based on multi-party computation (MPC; Yao, 1982). This ensures that statistic can be computed without any party revealing their raw data. Therefore, differential privacy (DP; Dwork, 2006) guarantees can be achieved by noising the *final* computed statistic, rather than to the original raw data, ensuring less loss of accuracy.
- We experimentally validate our methodology across a range of synthetic and semi-synthetic tasks, demonstrating strong agreement between our prospective rankings and the true rankings obtained after performing the merge. Moreover, we find that our proposed methodology to target CATE parameters improves over traditional Bayesian data selection and a number of other baselines. Finally, we show that, for the same level of privacy guarantees, our MPC protocol chained with DP performs better than applying DP to the raw inputs in the linear case.

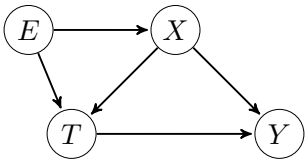


Figure 2: Assumed DAG

2 PROBLEM STATEMENT, ASSUMPTIONS & NOTATION

Notation The random variables X, T, and Y represent the covariates, treatment, and outcomes, with domains $\mathcal{X}$, $\{0,1\}$, and $\mathcal{Y}$, respectively; $\mathbf{x}$, t, and y denote realisations of these variables. We let $\mathcal{D}_e$ be the dataset comprised of elements $\{\mathbf{x}_i, t_i, y_i\}_{i=1}^{n_e}$, drawn i.i.d. from a distribution $P_e(\mathbf{x}, t, y)$, where $e \in \{0, \cdots, K\}$ indexes the datasets. Vectors of observations in $\mathcal{D}_e$ are denoted in bold, i.e. $\mathbf{y}_e = (y_i)_{i=1}^{n_e}$, $\mathbf{t}_e = (t_i)_{i=1}^{n_e}$, and $\mathbf{X}_e = (\mathbf{x}_i)_{i=1}^{n_e}$ refers to the data matrix. We use potential outcomes framework (Rubin, 1974), so that $Y(t)$ represents the outcome resulting from an intervention setting $T = t$.

Assumptions and objective Throughout we focus on estimating the *Conditional Average Treatment Effect* (CATE), given by:

$$\tau(\mathbf{x}) = \mathbb{E}\left[Y(1) - Y(0) | X = \mathbf{x}\right]$$

To estimate CATE, we begin with an initial dataset, $\mathcal{D}_0$, referred to as the *host*. Our goal is to accurately estimate CATE with respect to the distribution of this dataset, $P_0(\mathbf{x})$. Specifically, we focus on minimising the Precision in Estimation of Heterogeneous Effects (PEHE; Louizos et al., 2017) given by $\epsilon_{\text{PEHE}}(f) = \int_{P_0(\mathbf{x})} \left(\hat{\tau}_f(\mathbf{x}) - \tau(\mathbf{x})\right)^2 d\mathbf{x}$, where $\hat{\tau}_f$ is the CATE estimate arising the outcome model f.

We consider a set of potential datasets for merging, $\mathcal{D}_e$, referred to as the *candidate* sites. In these candidate datasets, we assume that the outcomes are unmeasured or masked, and so denote them by $\mathcal{D}_e = \{\mathbf{x}_i^e, t_i^e, Y_i^e\}_{i=1}^{n_e}$ to show the randomness in Y_i^e. The goal is to prospectively identify which of the candidate datasets $\mathcal{D}_e$, would reduce the uncertainty over CATE if we were to measure or unmask the Y_i^e's and merge with the host dataset $\mathcal{D}_0$.

We assume that datasets are generated according to the Directed Acyclic Graph (DAG) in Figure 2.* This implies that $P(y|\mathbf{x}, t, e)$ is fixed across environments, but both covariate distributions and treatment allocation schemes are free to vary i.e. $P(\mathbf{x}, t|e)$ can depend on e. We also assume positivity over the whole population, so that $\forall\ \mathbf{x}, 0 < P(T = 1|X = \mathbf{x}) < 1$, but allow for violations at the site level. Adding the consistency assumption (i.e. $Y(t) = Y$ when $T = t$) to positivity and the causal structure in Figure 2 (which implies no hidden confounders) we have that CATE is identifiable (Pearl, 2009; Richardson and Robins, 2013), constant across environments e, and given by:

$$\begin{aligned}\tau(\mathbf{x}) &= \mathbb{E}\left[Y|X = \mathbf{x}, T = 1\right] - \mathbb{E}\left[Y|X = \mathbf{x}, T = 0\right] \\ &= \mathbb{E}\left[Y|X = \mathbf{x}, T = 1, E = e\right] \\ &\quad - \mathbb{E}\left[Y|X = \mathbf{x}, T = 0, E = e\right]\end{aligned}$$

*We make us of the SWIG framework to combine causal graphical models with potential outcomes. More details can be found in Richardson and Robins (2013).

3 METHOD

We take a Bayesian approach to modelling the CATE. Specifically, we define $f_\theta : \mathcal{X} \times \{0,1\} \to \mathcal{Y}$ to be a function representing the expected outcome conditional on covariates and treatment, i.e., $f_\theta(\mathbf{x}, t)$ seeks to approximate $\mathbb{E}\left[Y|X = \mathbf{x}, T = t\right]$. We assign a prior distribution $P(\theta)$, to the parameters θ and denote the conditional likelihood of the outcome given parameters, covariates, and treatment as $P(Y|\theta, \mathbf{x}, t)$, which is chosen appropriately based on the type of outcome being modelled. For example, we can select a normal likelihood for continuous outcomes, or a Bernoulli likelihood for binary ones. The data-generating process can be written as

$$\theta \sim P(\theta), \quad Y|f_\theta(\mathbf{x}, t) \sim P(Y|\theta, (\mathbf{x}, t)).$$

Throughout this paper we focus on continuous y and use a normal likelihood with fixed variance σ^2, i.e. $P(Y|\mathbf{x}, t, \theta) = \mathcal{N}(Y; f_\theta(\mathbf{x}, t), \sigma^2)$. CATE is then estimated via the current posterior mean, so that if we have conditioned on data $\mathcal{D}$ our estimate is $\hat{\tau}(\mathbf{x}) = \mathbb{E}_{\theta \sim p(\theta|\mathcal{D})}[f_\theta(\mathbf{x}, 1) - f_\theta(\mathbf{x}, 0)]$.

3.1 Quantifying Data Merge Utility through Expected Information Gain

In order to quantify the value of a data merge, we draw inspiration from Bayesian experimental design (BED; Chaloner and Verdinelli, 1995; Rainforth et al., 2024). BED applies information theory to provide a measure of what performing a particular experiment would tell us about a parameter of interest, relative to our current beliefs about the parameter value. In our context, the 'design' of the experiment corresponds to choosing a dataset $\mathcal{D}_e$, and the outcome is observing $\{Y_i^e\}_{i=1}^{n_e}$. The value of an experiment is quantified via the expected information gain (EIG; Lindley, 1956), which measures the expected reduction in uncertainty in the parameters, as measured by Shannon entropy, when moving from the post host posterior, $P(\theta|\mathcal{D}_0)$ to the post merge posterior, $P(\theta|\mathcal{D}_0, \mathcal{D}_e)$:

$$\text{EIG}_{\theta|\mathcal{D}_0}(e) = \mathbb{E}[H[P(\theta|\mathcal{D}_0)] - H[P(\theta|\mathcal{D}_0, \mathcal{D}_e)]],$$

where the expectation is over $P(\mathbf{y}_e|\mathbf{X}_e, \mathbf{t}_e, \mathcal{D}_0) = \mathbb{E}_{P(\theta|\mathcal{D}_0)}[P(\mathbf{y}_e|\theta, \mathbf{X}_e, \mathbf{t}_e)]$—the Bayesian marginal distribution of $\mathbf{y}_e$. The EIG is equivalent to the mutual information between parameters and outcomes under the design and can be equivalently written as:

$$\mathrm{EIG}_{\theta|\mathcal{D}_0}(e) = \mathbb{E}\left[\log \frac{P(\mathbf{y}_e|\theta, \mathbf{X}_e, \mathbf{t}_e)}{P(\mathbf{y}_e|\mathbf{X}_e, \mathbf{t}_e, \mathcal{D}_0)}\right], \quad (1)$$

where the expectation is over $P(\mathbf{y}_e, \theta|\mathbf{X}_e, \mathbf{t}_e, \mathcal{D}_0)$. This form known as Bayesian Active Learning by Disagreement (BALD) in the active learning literature (Houlsby et al., 2011). For certain classes of models, such as polynomial regression or Gaussian processes, the EIG is available in closed form (Sebastiani and Wynn, 2000). In other cases if the likelihood function isn't analytically available, we can approximate it using nested Monte Carlo (NMC; Rainforth et al., 2018; Foster, 2021) as follows:

$$\widehat{\mathrm{EIG}}^{\mathrm{NMC}}_{\theta|\mathcal{D}_0}(e) = \frac{1}{N}\sum_{i=1}^{N} \log \frac{P(\mathbf{y}_e^{(i)}|\theta^{(i)}, \mathbf{X}_e, \mathbf{t}_e)}{\widehat{P}(\mathbf{y}_e^{(i)}|\mathbf{X}_e, \mathbf{t}_e, \mathcal{D}_0)}, \quad (2)$$

where $\theta^{(i)}, \mathbf{y}_e^{(i)} \sim P(\theta|\mathcal{D}_0)P(\mathbf{y}_e|\theta^{(i)}, \mathbf{X}_e, \mathbf{t}_e)$, and further M_1 samples $\theta'_j \sim P(\theta|\mathcal{D}_0)$ for the denominator

$$\widehat{P}(\mathbf{y}_e^{(i)}|\mathbf{X}_e, \mathbf{t}_e, \mathcal{D}_0) = \frac{1}{M_1}\sum_{j=1}^{M_1} P\left(\mathbf{y}_e^{(i)}|\theta'_j, \mathbf{X}_e, \mathbf{t}_e\right). \quad (3)$$

Note that since we assume a normal likelihood, we can construct a Rao-Blackwellised estimator by analytically computing the entropy of the likelihood in Eq. 1 to give:

$$\widehat{\mathrm{EIG}}^{\mathrm{RB}}_{\theta|\mathcal{D}_0}(e) = -\frac{1}{N}\sum_{i=1}^{N} \log\left(\frac{1}{M_1}\sum_{j=1}^{M_1} P(\mathbf{y}_e^{(i)}|\mathbf{X}_e, \mathbf{t}_e, \theta'_j)\right) - \frac{n_e}{2}(1 + \log(2\pi\sigma^2)).$$

Nested estimators are biased but consistent, converging to the true EIG at a rate $\mathcal{O}((N^{-1} + cM_1^{-2})^{\frac{1}{2}})$ (Rainforth et al., 2018). A detailed algorithm for both estimators is given in Appendix A.1.

These estimators allow us gauge the value of a merge before observing outcomes $\{Y_i^e\}_{i=1}^{n_e}$ in the dataset $\mathcal{D}_e$. However, whilst this approach provides us with information on the parameters for the conditional outcome, it may not necessarily lead to improved CATE predictions. This is because $\mathrm{EIG}_{\theta|\mathcal{D}_0}$ encourages *uniform entropy reduction* across all dimensions of θ, not just the ones relevant for CATE estimation. For instance, a dataset with untreated individuals only would provide information about the conditional outcome, but less for CATE, which requires viewing treated individuals as well. This motivates our improved methodology, which we present in the next section.

3.2 EIG Targeting CATE Parameters

Many causal inference models have additional parameter structure that allows us to target CATE estimation more directly. Specifically, the parameter set, θ can often be split as $\theta = \theta_{\mathrm{c}} \cup \theta_{\mathrm{nc}}$ where θ_{c} parameterises the CATE model and θ_{nc} is a set of nuisance parameters. For example, in Bayesian Causal Forests (Hahn et al., 2020) the model is parameterised as $f_\theta(\mathbf{x}, t) = \mu_{\theta_{\mathrm{nc}}}(\mathbf{x}) + t\tau_{\theta_{\mathrm{c}}}(\mathbf{x})$, where $\mu_{\theta_{\mathrm{nc}}}$ and $\tau_{\theta_{\mathrm{c}}}(\mathbf{x})$ jointly model the conditional outcome, and $\tau_{\theta_{\mathrm{c}}}(\mathbf{x})$ directly models CATE. Leveraging such a parameterisation, we can prioritise uncertainty reduction in θ_c, and therefore CATE, by maximising

$$\mathrm{EIG}_{\theta_{\mathrm{c}}|\mathcal{D}_0}(e) = \mathbb{E}\left[\log \frac{P(\mathbf{y}_e|\theta_{\mathrm{c}}, \mathbf{X}_e, \mathbf{t}_e, \mathcal{D}_0)}{P(\mathbf{y}_e|\mathbf{X}_e, \mathbf{t}_e, \mathcal{D}_0)}\right], \quad (4)$$

where expectation is over $P(\mathbf{y}_e, \theta_{\mathrm{c}}|\mathbf{X}_e, \mathbf{t}_e, \mathcal{D}_0)$. Here $P(\mathbf{y}_e|\theta_c, \mathbf{X}_e, \mathbf{t}_e, \mathcal{D}_0) = \mathbb{E}_{P(\theta_{nc}|\theta_c, \mathcal{D}_0)}[P(\mathbf{y}_e|\theta, \mathbf{X}_e, \mathbf{t}_e)]$ is the Bayesian distribution of the outcomes conditional on the CATE-related parameters only, and is generally not available in closed form. To deal with this intractability, we suggest approximating both the numerator and denominator empirically, yielding the following estimator:

$$\widehat{\mathrm{EIG}}^{\mathrm{NMC}}_{\theta_{\mathrm{c}}|\mathcal{D}_0}(e) = \frac{1}{N}\sum_{i=1}^{N} \log \frac{\widehat{P}(\mathbf{y}_e^{(i)}|\theta_{\mathrm{c}}^{(i)}, \mathbf{X}_e, \mathbf{t}_e, \mathcal{D}_0)}{\widehat{P}(\mathbf{y}_e^{(i)}|\mathbf{X}_e, \mathbf{t}_e, \mathcal{D}_0)}, \quad (5)$$

where $\theta_c^{(i)}, \mathbf{y}_e^{(i)} \sim P(\theta_{\mathrm{c}}|\mathcal{D}_0)P(\mathbf{y}_e|\theta_{\mathrm{c}}, \mathbf{X}_e, \mathbf{t}_e)$, the denominator $\widehat{P}(\mathbf{y}_e^{(i)}|\mathbf{X}_e, \mathbf{t}_e, \mathcal{D}_0)$ is as in Eq. 3, and use further M_2 samples $\theta_{nc}^{(ik)} \sim P(\theta_{nc}^{(ik)}|\theta_c^{(i)}, \mathcal{D}_0)$ for the numerator:

$$\widehat{P}(\mathbf{y}_e^{(i)}|\theta_{\mathrm{c}}^{(i)}, \mathbf{X}_e, \mathbf{t}_e, \mathcal{D}_0) = \frac{1}{M_2}\sum_{k=1}^{M_2} P\left(\mathbf{y}_e^{(i)}|\theta_{nc}^{(ik)} \cup \theta_c^{(i)}, \mathbf{X}_e, \mathbf{t}_e\right)$$

This ensures that we prioritise a gain in information in the part of the model directly responsible for CATE. We again give an algorithm in Appendix A.1.

3.3 Procedure and Model Classes

We now apply both procedures to three popular Bayesian causal inference methods: Bayesian Polynomial Regression, Bayesian Causal Forests (Hahn et al., 2020), and Causal Multi-task Gaussian Processes (Alaa and Van Der Schaar, 2017). We describe the standard predictive method as well as the parameter split used to target CATE.

Bayesian Polynomial Regression Due to its ubiquity across numerous fields, we first apply our method to Bayesian Polynomial Regression model. This involves specifying an initial polynomial transformation[†]

[†]This could be an arbitrary non-linear function but we focus on polynomial transformations.

$\phi : \mathcal{X} \times \mathcal{T} \to \mathbb{R}^p$, a mean μ_0, and precision matrix Λ_0 to parameterise the prior as:

$$f_\theta(\mathbf{x}, t) = \phi(\mathbf{x}, t)^\top \theta, \quad \theta \sim \mathcal{N}(\mu_0, \sigma^2 \Lambda_0^{-1}). \tag{6}$$

ϕ allows for higher order, or interaction terms. We further assume $\phi(\mathbf{x}, t)$ and θ can be split as:

$$\phi(\mathbf{x}, t) = \begin{bmatrix} \phi_{\text{nc}}(\mathbf{x}) & t\phi_{\text{c}}(\mathbf{x}) \end{bmatrix}^\top, \quad \theta = \begin{bmatrix} \theta_{\text{nc}} & \theta_{\text{c}} \end{bmatrix}^\top$$

This covers a broad range of Bayesian polynomial regressions, including the selection used in Gelman et al. (2021). In Appendix B.2.1 we show that both EIGs can be computed in closed form:

Proposition 3.1. *For the Bayesian Polynomial Regression model defined in Eq. 6 we have:*

$$\text{EIG}_{\theta|\mathcal{D}_0}(e) = \log\det\left(\Phi_e^\top \Phi_e + \Phi_0^\top \Phi_0 + \Lambda_0\right) + C$$
$$\text{EIG}_{\theta_c|\mathcal{D}_0}(e) = \log\det(\Phi_{c,e}^\top \Phi_{c,e} + \Phi_{c,0}^\top \Phi_{c,0} + \Lambda_0^{[c,c]}) + C',$$

where $\Phi_e = \phi(\mathbf{X}_e, \mathbf{t}_e)$, $\Phi_0 = \phi(\mathbf{X}_0, \mathbf{t}_0)$, $\Phi_{c,e} = \mathbf{t}_e \odot \phi_{\text{c}}(\mathbf{X}_e)$, $\Phi_{c,0} = \mathbf{t}_0 \odot \phi_{\text{c}}(\mathbf{X}_0)$, *with* ϕ, ϕ_c *applied row-wise and* $\odot$ *denoting element-wise multiplication;* C, C' *are constant in* e.

Bayesian Causal Forest Bayesian Causal Forests (BCF; Hahn et al., 2020) are one of the most popular causal inference methods, building upon Bayesian Additive Regression Trees (BART; Chipman et al., 2012), which are themselves a mainstay in observational causal inference (Hill, 2011). The BCF model can be expressed as $f_\theta(\mathbf{x}, t) = \mu_{\theta_{\text{nc}}}(\mathbf{x}) + t\tau_{\theta_{\text{c}}}(\mathbf{x})$, where $\mu_{\theta_{\text{nc}}}$ and $\tau_{\theta_{\text{c}}}(\mathbf{x})$ are independent BART models; further details and alternative parameterisations can be found in Appendix B.1. As the posterior is only available via sampling, we estimate both EIGs using NMC as given in Eq. 2 and Eq. 5.

Causal Multi-task Gaussian Processes Causal multi-task Gaussian Processes (Alaa and Van Der Schaar, 2017) use a vector-valued GP (Alvarez et al., 2012) to jointly model the conditional outcomes, allowing information sharing between them:

$$\mathbf{f} = \begin{bmatrix} f(\mathbf{x}, 0) & f(\mathbf{x}, 1) \end{bmatrix}^\top, \text{ where } \mathbf{f} \sim \mathcal{GP}(0, \mathbf{K}) \tag{7}$$

for a vector-valued kernel $\mathbf{K} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}^{2\times 2}$. The outcomes are then modelled by evaluating the relevant portion of the GP. Under this setting CATE is given by $\tilde{\tau} = \mathbf{ef}$ for $\mathbf{e} = \begin{bmatrix} -1 & 1 \end{bmatrix}$, meaning that $\tilde{\tau}$ is also a GP given by $\tilde{\tau} \sim \mathcal{GP}(0, \mathbf{e}^\top \mathbf{K}\mathbf{e})$. The advantage of causal multi-task GPs is that they allow us to get a closed form posterior for τ without having to observe any samples from $Y(1) - Y(0)$.

As GPs are inherently non-parametric they do not fit directly into the framework laid out in Section 3.1 and Section 3.2. This is not a problem for the predictive case as we can replace θ with $\mathbf{f}$ in Eq. 1 and get closed form expressions (Houlsby et al., 2011). However, for the causal case this creates challenges as we cannot directly evaluate the expressions in Eq. 4 with $\tilde{\tau}$ in the place of θ_c. To resolve this we instead focus on entropy reduction in CATE predictions on the host dataset. We denote this by $\tilde{\tau}(\mathbf{X}_0)$, where $\mathbf{X}_0$ is the host data matrix. The information gains e denote these by $\text{EIG}_{\mathbf{f}|\mathcal{D}_0}(e)$ and $\text{EIG}_{\tilde{\tau}(\mathbf{X}_0)|\mathcal{D}_0}(e)$ respectively. As the following proposition shows, both of these are now available in closed form:

Proposition 3.2. *Let* $n_e^{(t)}$ *be the number of subjects receiving treatment* t *in dataset* e. *For the causal multi-task GP model, defined in Eq. 7 we have*

$$\begin{aligned} \text{EIG}_{\mathbf{f}|\mathcal{D}_0} &= \frac{1}{2}\log\det\left(\Sigma_1\right) \\ &\quad - n_e^{(0)}\log(\sigma_0) - n_e^{(1)}\log(\sigma_1) \\ \text{EIG}_{\tilde{\tau}(\mathbf{X}_0)|\mathcal{D}_0}(e) &= \frac{1}{2}\log\left(\det(\Sigma_1)\det(\Sigma_2)\right) \\ &\quad - \frac{1}{2}\log\left(\det(\Sigma)\right), \end{aligned}$$

where $\Sigma_1, \Sigma_2, \Sigma$ *and the proof are given Appendix B.3.*

4 PRIVACY

For privacy we use *Multi-Party Computation* (MPC; Evans et al., 2018). First introduced by Yao (1982), MPC focuses on a setting where m separate parties wish to compute the value of a function $f(x_1, \cdots, x_m)$ where the i^{th} party inputs x_i and wishes to keep this private. To resolve this, MPC involves the specification of a protocol of message passing between parties which if followed would lead to the computation of f. In this work, we focus on the *semi-honest* setting, in which all parties follow the specified protocol, but some *corrupt* parties will try to learn as much about their peers inputs in the process. The goal is to devise a protocol which will preserve the privacy of the non-corrupt party's inputs, up to a given computational budget by the adversary. In our setting this means that any collection of corrupt sites are unable to learn anything about the other sites data during the EIG calculation, so any noise needed for privacy guarantees can be added to the final statistic.

For implementing multi-party computation, we employ the open source library CrypTen (Knott et al., 2021). CrypTen builds upon PyTorch (Paszke et al., 2019) allowing for standard tensor operations to be performed in an MPC protocol. For arithmetic operations on floating-point values this is achieved as follows: A float, x_F, is multiplied by some large scaling factor $B = 2^L$ and rounded to the nearest integer $\lfloor x_F \rceil$, where L is the number of precision bits. The integer $\lfloor x_F \rceil$ is then associated with its equivalence class $x \in \mathbb{Z}/Q\mathbb{Z}$ where

$\mathbb{Z}/Q\mathbb{Z}$ is a ring of Q elements. The value x can then be shared across all m parties using Shamir secret sharing (Shamir, 1979), in which each party gets access to a share of x is given by $[x]_i \in \mathbb{Z}/Q\mathbb{Z}$ which is generated such that the sum of all shares recovers the original value, so $x = \sum_{i=1}^{m} [x]_i \mod Q$. At any point all parties can combine their shares to decode the output as $x_F \approx x/B$. We let $[x] = \{[x]_i\}_{i=1}^{m}$ denote the set of all shares corresponding to the secret value x.

Arithmetic operations building on addition are performed locally, so that for two secret values $[x], [y]$ each party performs $[z]_i = [x]_i + [y]_i$ and the result, z is obtained by all parties summing their share. M ultiplication is implemented using Beaver triples (Beaver, 1992), logarithms are approximated using householder iterations (Householder, 1970), and reciprocals use Newton-Raphson. We implement log-determinants using Cholesky LDL decompositions, which are preferred to standard Cholesky factorisations as they avoid the use of square roots, which would require additional approximation in Crypten. This is possible as we only compute the log determinant of positive semi-definite matrices. This provides all the operations necessary to implement the above EIG calculations in a private manner using MPC.

When returning EIG statistics to the host, we add a small amount of noise to prevent information leakage. If we only need to output the best site, we use the exponential protocol (Dwork, 2006) ensuring minimal information leakage. Finally, we note that the discretisation required to represent a float in the ring $\mathbb{Z}/Q\mathbb{Z}$ involves some degree of precision loss. Nevertheless, as we empirically demonstrate in the next section, this leads to minimal depreciation in performance compared to differential privacy.

5 EXPERIMENTS & RESULTS

We experimentally validate our approach in a setting where the host has to rank a number of candidate datasets based on the estimated gain from merging. We use selection of synthetic and semi-synthetic benchmark datasets as this allows us to use the known CATE to get a ground truth ranking. For each model, we do this by ranking datasets on the true PEHE on $P_0(\mathbf{x})$ of the relevant model trained on the merged dataset $\mathcal{D}_0 \cup \mathcal{D}_e$. This is then compared against implied EIG rankings. Throughout, we tune any model hyper-parameters on the host site when measuring the information gain as well as re-tuning paramters on each merged dataset when getting the ground truth ranking.

To generate the datasets, we begin with an initial, large dataset $\mathcal{D}$ from which we subsample the host and candidate sites. We do this by choosing a selection function, $S_e(\mathbf{x}, t)$, and using it to subsample a dataset of size n_e, where $S_e(\mathbf{x}_i, t_i)$ is the probability of subsampling point i for dataset $\mathcal{D}_e$. Varying the selection functions across sites ensures heterogeneity amongst datasets whilst complying with the independences given by the causal structure in Figure 2. We begin by providing an illustrative example on synthetic data before evaluating on the causal benchmarks, specifically: Lalonde (LaLonde, 1986) and Infant Health and Development Program (IHDP; Louizos et al., 2017). Further details and results can be found in Appendix C and D, respectively.

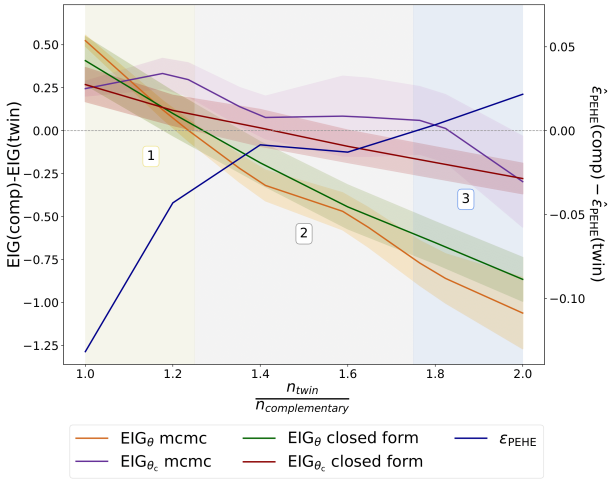


Figure 3: Difference in post host EIG and $\hat{\epsilon}_{\text{PEHE}}$ for a linear CATE model trained on $\mathcal{D}_0 \cup \mathcal{D}_{comp}$ and $\mathcal{D}_0 \cup \mathcal{D}_{twin}$ for increasing $\frac{n_{twin}}{n_{comp}}$ and fixed n_{host}=n_{comp}=100. PEHE is evaluated on hold out data from the host distribution. Lines show mean ±1 s.d. across 50 seeds. The three regions show different datasets preferences. In region 2, $\text{EIG}_{\theta|\mathcal{D}_0}$ incorrectly favours $\mathcal{D}_{twin}$, whilst $\text{EIG}_{\theta_c|\mathcal{D}_0}$ correctly selects $\mathcal{D}_{comp}$ for the causal task.

5.1 Illustrative experiment

Concept To illustrate the difference between our two methods—the standard approach based on the full parameter set, EIG_θ (Eq. 1), and the CATE-targeting one, EIG_{θ_c} (Eq. 4)—we start with a simple example where the host must choose between two candidate sites. We design these sites as follows: a *complementary* site, representing the "ideal" merge for causal purposes, and a *twin* site, containing information similar to the host. To create these, we first simulate an initial large dataset $\mathcal{D}$ as if it were a randomised controlled trial with an equal probability of treatment and subsample the host dataset, $\mathcal{D}_0$, using a selection function $S_0(\mathbf{x}, t)$. The data of the complementary site, $\mathcal{D}_{comp}$, is subsampled using $1 - S_0(\mathbf{x}, t)$ as a selection function, whilst the twin site uses $S_0(\mathbf{x}, t)$. This ensures the twin dataset mirrors the host's distribution and the complementary causally "complements" it. Assuming equal sizes, merging $\mathcal{D}_0$

Model	Objective	$\rho(\uparrow)$	p@1 ($\uparrow$)	p@3 ($\uparrow$)	p@5 ($\uparrow$)
Polynomial	$\text{EIG}_{\theta_{c\|\mathcal{D}_0}}$	$\mathbf{0.70 \pm 0.08}$	$\mathbf{0.50 \pm 0.15}$	$\mathbf{0.70 \pm 0.04}$	$\mathbf{0.78 \pm 0.04}$
	$\text{EIG}_{\theta\|\mathcal{D}_0}$	$\mathbf{0.68 \pm 0.06}$	$\mathbf{0.50 \pm 0.15}$	$\mathbf{0.70 \pm 0.06}$	$\mathbf{0.76 \pm 0.04}$
	Best baseline	0.40 ± 0.11	0.40 ± 0.15	0.60 ± 0.15	0.66 ± 0.04
Causal GP	$\text{EIG}_{\tilde{\tau}(X_0)\|\mathcal{D}_0}$	$\mathbf{0.49 \pm 0.06}$	$\mathbf{0.50 \pm 0.15}$	$\mathbf{0.50 \pm 0.08}$	$\mathbf{0.62 \pm 0.06}$
	$\text{EIG}_{\mathbf{f}\|\mathcal{D}_0}$	0.33 ± 0.06	0.30 ± 0.15	0.43 ± 0.05	$\mathbf{0.60 \pm 0.04}$
	Best baseline	0.31 ± 0.12	0.10 ± 0.20	0.20 ± 0.07	0.46 ± 0.05
Bayesian CF	$\text{EIG}_{\theta_{c\|\mathcal{D}_0}}$	$\mathbf{0.54 \pm 0.10}$	$\mathbf{0.60 \pm 0.15}$	$\mathbf{0.63 \pm 0.08}$	$\mathbf{0.70 \pm 0.04}$
	$\text{EIG}_{\theta\|\mathcal{D}_0}$	0.36 ± 0.10	0.30 ± 0.14	0.50 ± 0.07	$\mathbf{0.66 \pm 0.05}$
	Best baseline	0.45 ± 0.11	$\mathbf{0.60 \pm 0.14}$	$\mathbf{0.63 \pm 0.08}$	$\mathbf{0.70 \pm 0.04}$

Table 1: Ranking experiment for the IHDP dataset, measured by Spearman ρ and precision at k (p@k). We include the best performing baseline method, which is different for different models.

with $\mathcal{D}_{comp}$ would recreate the initial randomised trial $\mathcal{D}$, as the variable e acts as a collider for $\mathbf{x}$ and t, removing their conditional dependency (see causal DAG in Appendix 4). Therefore, $\mathcal{D}_{comp}$ represents an ideal merge as it balances treatment allocation, whereas $\mathcal{D}_{twin}$ covers similar regions of the data space to those in the host, $\mathcal{D}_0$, potentially amplifying pre-existing biases and imbalances.

For the illustrative experiment, we vary the ratio of sample sizes, $\frac{n_{twin}}{n_{comp}}$, in order to compare which dataset is chosen by $\text{EIG}_{\theta|\mathcal{D}_0}$ and the causally targeted $\text{EIG}_{\theta_c|\mathcal{D}_0}$ for linear regression. The aim is to demonstrate that $\text{EIG}_{\theta|\mathcal{D}_0}$ will prefer $\mathcal{D}_{twin}$ at points where $\mathcal{D}_{comp}$ dataset is still the preferable dataset for CATE estimation. Indeed, for large values of the ratio $\frac{n_{twin}}{n_{comp}}$, $\mathcal{D}_{twin}$ provides significant information about the conditional outcome, but not in the regions that are most relevant for causal estimation. On the other hand, EIG_{θ_c} should continue to select $\mathcal{D}_{twin}$ whilst it remains preferable for CATE estimation. We simulate $\mathbf{x} \in \mathbb{R}^3$ where x_1 is Bernoulli and other covariates are normal. The true outcome is sampled from a normal linear model, and the selection functions are logistic regressions. We also include sample based estimates for each EIG to demonstrate how they differ from their closed form counterparts. Experimental details provided in Appendix C.

Results Figure 3 shows the results of the experiment divided into three regions. In region one, both methods choose the complementary dataset over the twin dataset which is consistent with ground truth ranking given by the PEHE upon merging. In region two, $\text{EIG}_{\theta|\mathcal{D}_0}$ chooses $\mathcal{D}_{twin}$ whilst $\text{EIG}_{\theta_{c|\mathcal{D}_0}}$ opts for $\mathcal{D}_{comp}$. Here, the complementary dataset is still the optimal in terms of CATE, but $\text{EIG}_{\theta|\mathcal{D}_0}$ preferences $\mathcal{D}_{twin}$ as it leads to a greater entropy reduction in the full set of parameters. This result shows that by focusing on the causal parameters alone, $\text{EIG}_{\theta_{c|\mathcal{D}_0}}$ is able to make the correct decision in selecting $\mathcal{D}_{comp}$. In the final region we see all lines have crossed the x axis, showing that all methods agree with the ground truth in choosing $\mathcal{D}_{twin}$. Finally, we note the MCMC estimates agree with the closed form counterparts, with increased variability due to sampling.

5.2 Ranking experiment

Concept For our main experiment we validate our framework in a setting where the host must choose between many potential candidates, each with different distributions. To do so we begin with a standard causal inference benchmark dataset, $\mathcal{D}$, and form the host and candidates datasets using subsampling functions, $S_e(\mathbf{x}, t)$ as detailed above where subsampling function is a logistic regression with random parameters. This ensures that each site has a different covariate distribution. We apply the three methods described in Section Section 3.3 to estimate both the two expected information gains for each candidate site, $\mathcal{D}_e$. Ultimately, like before, we compare the implied rankings with the ground truth ranking given by the PEHE. Full details are provided in Appendix C.

Baselines We provide a number of simple comparison methods as baselines for our task. Specifically, (a) ranking by sample size, (b) ranking by similarity of covariate distribution measured by a multivariate Gaussian fit to the host, and (c) ranking by dissimilarity of treatment allocation measured by the error of a propensity model fit on the host. We compare using Spearman rho(ρ) and precision at k (p@k).

Results Table 1 shows the results of our experiment on the IHDP dataset (Louizos et al., 2017) where the host has to rank the best datasets out of 10 candidates. We report the average performance of the rankings across 20 repeated experiments, and according to Spearman ρ (Spearman, 1961) and precision at k. We include the best baseline by Spearman ρ performance. Additional metrics and baseline performance can be

Method	MSE (↓)	ρ(↑)
MPC Linear	$\mathbf{(3.80 \pm 0.04) \times 10^{-6}}$	$\mathbf{0.797 \pm 0.06}$
DP Linear	9.80 ± 0.30	0.06 ± 0.11

Table 2: Multi-Party Computation Results for EIG_{θ_c}.

found in Appendix D. These results demonstrate that across all models, our EIG_{θ_c} based approach, focusing on causal parameters, outperforms the standard approach which seeks expected entropy reduction in all parameters. Further, our method works significantly better than all baselines.

5.3 Multi-Party Computation Experiments

Finally, we experimentally demonstrate the performance of our cryptographic protocol. We repeat a similar experiment as in Section 5.2 on the IHDP dataset, this time choosing between twenty datasets. Results are given using a linear model as this allows us to compare our MPC based approach against differential privacy (DP; Dwork, 2006, see Appendix A.2 for definition). We use $\epsilon = 100$ and Laplace noising (Dwork et al., 2014), following existing work on DP linear regression (Bernstein and Sheldon, 2019; Awan and Slavković, 2021). In order to ensure a fair comparison, we noise the final statistics from the MPC computation with an appropriate amount of Laplace noise. This comes from the sensitivity of the EIG statistic which we derive in Appendix A.3. Table 2 shows both the MSE induced by the privacy method and the Spearman ρ against the noise-free ranking. Results show that MPC vastly outperforms differential privacy, both in terms of MSE and ranking, with DP producing a near random ranking, despite the relaxed noising.

6 RELATED WORK

Federated Learning Via Multi-Party Computation Our setting bares large with federated learning (Li et al., 2020a), a distributed machine learning approach that enables multiple parties to collaboratively train a shared model while keeping their raw data decentralised and private. Our goal differs from this as we focus not on learning a model across sites, but deciding which sites to pool. The application of multiparty computation within federated learning is a growing area (Li et al., 2020b; Mugunthan et al., 2019; Kanagavelu et al., 2020), with much of the work focusing on how to learn predictive models in a secure cross site fashion. Most similar to our work is Muazu et al. (2024), who develop a similar federated approach to data fusion focusing on healthcare, however they do not take a causal approach. To the best of our knowledge, there are no existing applications of multi-party computation within causal inference.

Federated/Private Causal Estimation There are, however, federated approaches to estimating causal effects. In this area, a majority of works partition the loss function into multiple components, with each component corresponding to a specific data source (Vo et al., 2022; Liu et al., 2024; Vo et al., 2023). However, modelling complex, non-linear relationships remains challenging (Almodóvar et al., 2023). Many of these algorithms come without privacy guarantees, with the exception of Niu et al. (2022), who add DP guarantees to various popular CATE estimation techniques. The latter approach however contrasts with ours as the use of sample splitting reduces data efficiency.

Causal Bayesian Active Learning Bayesian Active Learning by Disagreement (BALD; Houlsby et al., 2011) is a framework for strategic training data acquisition, focusing on regions of high uncertainty. Our method based on maximising $\text{EIG}_{\theta|\mathcal{D}_0}$ (Eq. 1) can be viewed as applying BALD after an initial update to an entire datasets rather than individual datapoints. Most similar to our work is CausalBALD (Jesson et al., 2021), which applies an active learning approach to CATE estimation. However, the acquisition function cannot easily be extended to datasets without ignoring the correlation in information provided by different points. Most applications of active learning in causality focus on causal discovery and intervention selection (Toth et al., 2022; Hauser and Bühlmann, 2014; Annadani et al., 2023).

7 DISCUSSION

Limitations Due to the cost of computing high dimensional determinants and performing multiple rounds of conditional sampling, our method can be computationally costly in high dimensions. Moreover, the cost of multi-party computation will also increase as higher order approximations are required to retain accuracy. However, both of these remain negligible compared to the data engineering and expenses of data fusion (Brodie, 2010; Kadadi et al., 2014). Finally, whilst our method offers a secure and principled way to prospectively quantify the value of dataset merges, the amount of information needed to justify such expenses may vary depending on application. It might, therefore, be beneficial to consider introducing a problem-specific threshold to determine when a proposed merger is worthwhile.

Conclusion We introduce an information-theoretic, cryptograpically secure framework for evaluating the utility of potential data merges for causal estimation. To the best of our knowledge, this is the first work addressing this relevant challenge. Through empirical evaluation, we demonstrate that our framework can reliably rank datasets according to their ability

to improve CATE estimation. We show that entropy reduction in the CATE parameters alone gives an improvement when compared to a more standard approach to Bayesian dataset acquisition. Finally, we demonstrate that our cryptographic procedure can be applied in conjunction with DP, resulting in lower loss of accuracy, compared to applying DP alone.

Acknowledgements

We would like to thank Andrew Yiu, Amartya Sanyal, Shahine Bouhabid, Xi Lin, Patrick Hough, and Sahra Ghalebikesabi for their comments. JF gratefully acknowledges funding from the EPSRC. LTM is supported by EPSRC through the Modern Statistics and Statistical Machine Learning (StatML) CDT programme, grant no. EP/S023151/1.

References

K. Abouelmehdi, A. Beni-Hessane, and H. Khaloufi. Big healthcare data: preserving security and privacy. *Journal of big data*, 5(1):1–18, 2018.

A. M. Alaa and M. Van Der Schaar. Bayesian inference of individualized treatment effects using multi-task gaussian processes. *Advances in neural information processing systems*, 30, 2017.

A. Almodóvar, J. Parras, and S. Zazo. Federated learning for causal inference using deep generative disentangled models. In *Deep Generative Models for Health Workshop NeurIPS 2023*, 2023.

M. A. Alvarez, L. Rosasco, N. D. Lawrence, et al. Kernels for vector-valued functions: A review. *Foundations and Trends® in Machine Learning*, 4(3): 195–266, 2012.

Y. Annadani, P. Tigas, D. R. Ivanova, A. Jesson, Y. Gal, A. Foster, and S. Bauer. Differentiable multi-target causal bayesian experimental design. *arXiv preprint arXiv:2302.10607*, 2023.

J. Awan and A. Slavković. Structure and sensitivity in differential privacy: Comparing k-norm mechanisms. *Journal of the American Statistical Association*, 116 (534):935–954, 2021.

D. Beaver. Efficient multiparty protocols using circuit randomization. In *Advances in Cryptology—CRYPTO'91: Proceedings 11*, pages 420–432. Springer, 1992.

G. Bernstein and D. R. Sheldon. Differentially private bayesian linear regression. *Advances in Neural Information Processing Systems*, 32, 2019.

E. V. Bonilla, K. Chai, and C. Williams. Multi-task gaussian process prediction. *Advances in neural information processing systems*, 20, 2007.

M. L. Brodie. Data integration at scale: From relational data integration to information ecosystems. In *2010 24th IEEE International Conference on Advanced Information Networking and Applications*, pages 2–3. IEEE, 2010.

F. Castanedo et al. A review of data fusion techniques. *The scientific world journal*, 2013, 2013.

K. Chaloner and I. Verdinelli. Bayesian experimental design: A review. *Statistical science*, pages 273–304, 1995.

H. A. Chipman, E. I. George, and R. E. McCulloch. Bart: Bayesian additive regression trees. *Annals of Applied Statistics*, 6(1):266–298, 2012.

R. K. Crump, V. J. Hotz, G. W. Imbens, and O. A. Mitnik. Dealing with limited overlap in estimation of average treatment effects. *Biometrika*, 96(1):187–199, 2009.

R. H. Dehejia and S. Wahba. Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *Journal of the American statistical Association*, 94(448):1053–1062, 1999.

A. Doan, A. Halevy, and Z. Ives. *Principles of data integration*. Elsevier, 2012.

C. Dwork. Differential privacy. In *International colloquium on automata, languages, and programming*, pages 1–12. Springer, 2006.

C. Dwork, A. Roth, et al. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407, 2014.

European Commission. Data protection rules for the public sector. Retrieved from https://ec.europa.eu/info/law/law-topic/data-protection/reform/rules-public-sector_en, 2018.

D. Evans, V. Kolesnikov, M. Rosulek, et al. A pragmatic introduction to secure multi-party computation. *Foundations and Trends® in Privacy and Security*, 2(2-3):70–246, 2018.

A. E. Foster. *Variational, Monte Carlo and policy-based approaches to Bayesian experimental design*. PhD thesis, University of Oxford, 2021.

A. Gelman, J. Hill, and A. Vehtari. *Regression and other stories*. Cambridge University Press, 2021.

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

A. Hauser and P. Bühlmann. Two optimal strategies for active learning of causal models from interventional data. *International Journal of Approximate Reasoning*, 55(4):926–939, 2014.

J. He, S. Yalov, and P. R. Hahn. Xbart: Accelerated bayesian additive regression trees. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1130–1138. PMLR, 2019.

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

P. W. Holland. Statistics and causal inference. *Journal of the American statistical Association*, 81(396):945–960, 1986.

N. Houlsby, F. Huszár, Z. Ghahramani, and M. Lengyel. Bayesian active learning for classification and preference learning. *arXiv preprint arXiv:1112.5745*, 2011.

A. S. Householder. The numerical treatment of a single nonlinear equation. *(No Title)*, 1970.

A. Jesson, P. Tigas, J. van Amersfoort, A. Kirsch, U. Shalit, and Y. Gal. Causal-bald: Deep bayesian active learning of outcomes to infer treatment-effects from observational data. *Advances in Neural Information Processing Systems*, 34:30465–30478, 2021.

A. Kadadi, R. Agrawal, C. Nyamful, and R. Atiq. Challenges of data integration and interoperability in big data. In *2014 IEEE international conference on big data (big data)*, pages 38–40. IEEE, 2014.

R. Kanagavelu, Z. Li, J. Samsudin, Y. Yang, F. Yang, R. S. M. Goh, M. Cheah, P. Wiwatphonthana, K. Akkarajitsakul, and S. Wang. Two-phase multi-party computation enabled privacy-preserving federated learning. In *2020 20th IEEE/ACM International Symposium on Cluster, Cloud and Internet Computing (CCGRID)*, pages 410–419. IEEE, 2020.

A. Kirsch, J. Van Amersfoort, and Y. Gal. Batchbald: Efficient and diverse batch acquisition for deep bayesian active learning. *Advances in neural information processing systems*, 32, 2019.

B. Knott, S. Venkataraman, A. Hannun, S. Sengupta, M. Ibrahim, and L. van der Maaten. Crypten: Secure multi-party computation meets machine learning. *Advances in Neural Information Processing Systems*, 34:4961–4973, 2021.

N. Krantsevich, J. He, and P. R. Hahn. Stochastic tree ensembles for estimating heterogeneous effects. In *International Conference on Artificial Intelligence and Statistics*, pages 6120–6131. PMLR, 2023.

R. J. LaLonde. Evaluating the econometric evaluations of training programs with experimental data. *The American economic review*, pages 604–620, 1986.

M. Lenzerini. Data integration: A theoretical perspective. In *Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 233–246, 2002.

L. Li, Y. Fan, M. Tse, and K.-Y. Lin. A review of applications in federated learning. *Computers & Industrial Engineering*, 149:106854, 2020a.

Y. Li, Y. Zhou, A. Jolfaei, D. Yu, G. Xu, and X. Zheng. Privacy-preserving federated learning framework based on chained secure multiparty computing. *IEEE Internet of Things Journal*, 8(8):6178–6186, 2020b.

D. V. Lindley. On a measure of the information provided by an experiment. *The Annals of Mathematical Statistics*, 27(4):986–1005, 1956.

Y. Liu, H. Wang, S. Wang, Z. He, W. Xu, J. Zhu, and F. Yang. Disentangle estimation of causal effects from cross-silo data. *arXiv preprint arXiv:2401.02154*, 2024.

C. Louizos, U. Shalit, J. M. Mooij, D. Sontag, R. Zemel, and M. Welling. Causal effect inference with deep latent-variable models. *Advances in neural information processing systems*, 30, 2017.

D. J. MacKay. Information-based objective functions for active data selection. *Neural computation*, 4(4): 590–604, 1992.

B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pages 1273–1282. PMLR, 2017.

M. M. Mello, J. K. Francer, M. Wilenzick, P. Teden, B. E. Bierer, and M. Barnes. Preparing for responsible sharing of clinical trial data, 2013.

T. Muazu, Y. Mao, A. U. Muhammad, M. Ibrahim, U. M. M. Kumshe, and O. Samuel. A federated learning system with data fusion for healthcare using multi-party computation and additive secret sharing. *Computer Communications*, 216:168–182, 2024.

V. Mugunthan, A. Polychroniadou, D. Byrd, and T. H. Balch. Smpai: Secure multi-party computation for federated learning. In *Proceedings of the NeurIPS 2019 Workshop on Robust AI in Financial Services*, volume 21. MIT Press Cambridge, MA, USA, 2019.

F. Niu, H. Nori, B. Quistorff, R. Caruana, D. Ngwe, and A. Kannan. Differentially private estimation of heterogeneous causal effects. In *Conference on Causal*

Learning and Reasoning, pages 618–633. PMLR, 2022.

A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

J. Pearl. Causal inference in statistics: An overview. 2009.

J. Platt and S. Kardia. Public trust in health information sharing: implications for biobanking and electronic health record systems. *Journal of personalized medicine*, 5(1):3–21, 2015.

T. Rainforth, R. Cornish, H. Yang, A. Warrington, and F. Wood. On nesting monte carlo estimators. In *International Conference on Machine Learning*, pages 4267–4276. PMLR, 2018.

T. Rainforth, A. Foster, D. R. Ivanova, and F. Bickford Smith. Modern bayesian experimental design. *Statistical Science*, 39(1):100–114, 2024.

T. S. Richardson and J. M. Robins. Single world intervention graphs (swigs): A unification of the counterfactual and graphical approaches to causality. *Center for the Statistics and the Social Sciences, University of Washington Series. Working Paper*, 128(30):2013, 2013.

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

D. B. Rubin. Estimating causal effects from large data sets using propensity scores. *Annals of internal medicine*, 127(8_Part_2):757–763, 1997.

F. Sattler, S. Wiedemann, K.-R. Müller, and W. Samek. Robust and communication-efficient federated learning from non-iid data. *IEEE transactions on neural networks and learning systems*, 31(9):3400–3413, 2019.

P. Sebastiani and H. P. Wynn. Maximum entropy sampling and optimal bayesian experimental design. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62(1):145–157, 2000.

A. Shamir. How to share a secret. *Communications of the ACM*, 22(11):612–613, 1979.

R. Shokri and V. Shmatikov. Privacy-preserving deep learning. In *Proceedings of the 22nd ACM SIGSAC conference on computer and communications security*, pages 1310–1321, 2015.

C. Spearman. The proof and measurement of association between two things. 1961.

S. Tarumi, M. Suzuki, H. Yoshida, S. Miyauchi, and R. Kurazume. Personalized federated learning for institutional prediction model using electronic health records: A covariate adjustment approach. In *2023 45th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 1–4. IEEE, 2023.

C. Toth, L. Lorch, C. Knoll, A. Krause, F. Pernkopf, R. Peharz, and J. Von Kügelgen. Active bayesian causal inference. *Advances in Neural Information Processing Systems*, 35:16261–16275, 2022.

T. V. Vo, A. Bhattacharyya, Y. Lee, and T.-Y. Leong. An adaptive kernel approach to federated learning of heterogeneous causal effects. *Advances in Neural Information Processing Systems*, 35:24459–24473, 2022.

T. V. Vo, T.-Y. Leong, et al. Federated learning of causal effects from incomplete observational data. *arXiv preprint arXiv:2308.13047*, 2023.

H. Wang, Z. Kaplan, D. Niu, and B. Li. Optimizing federated learning on non-iid data with reinforcement learning. In *IEEE INFOCOM 2020-IEEE conference on computer communications*, pages 1698–1707. IEEE, 2020.

A. C. Yao. Protocols for secure computations. In *23rd annual symposium on foundations of computer science (sfcs 1982)*, pages 160–164. IEEE, 1982.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes]
2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]
3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Yes]
   (b) The license information of the assets, if applicable. [Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]
5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

A Mathematical Details

A.1 Algorithms for Computing EIG

Algorithm 1 Algorithm for $\widehat{\text{EIG}}^{\text{NMC}}_{\theta|\mathcal{D}_0}(e)$

Input: Data Matrix $\mathbf{X}_e$, Treatment Vector $\mathbf{t}_e$, Variance σ^2
Output: $\widehat{\text{EIG}}^{\text{NMC}}_{\theta|\mathcal{D}_0}(e)$
Require: $l = NM_1$ for some $n, M_1 > 0$
$S \leftarrow 0$
Sample $\{\theta_i\}_{i=1}^{l} \sim P(\theta \mid \mathcal{D}_0)$ for $l = NM_1$
Split $\{\theta_i\}_{i=1}^{l}$ as $\{(\theta_i), (\theta_{i,j})_{j=1}^{M_1}\}_{i=1}^{N}$
for $i \in \{1, \cdots, N\}$ **do**
  Sample $\mathbf{y}_e^{(i)} \sim P(\mathbf{y}_e|\mathbf{X}_e, \mathbf{t}_e, \theta_i)$
  $S \leftarrow S + \log \frac{P(\mathbf{y}_e^{(i)}|\theta_i, \mathbf{X}_e, \mathbf{t}_e)}{\frac{1}{M_1}\sum_{j=1}^{M_1} P(\mathbf{y}_e^{(i)}|\theta_{i,j}, \mathbf{X}_e, \mathbf{t}_e)}$
end for
$\widehat{\text{EIG}}^{\text{NMC}}_{\theta|\mathcal{D}_0}(e) \leftarrow \frac{1}{N} S$
return $\widehat{\text{EIG}}^{\text{NMC}}_{\theta|\mathcal{D}_0}(e)$

Algorithm 2 Algorithm for $\widehat{\text{EIG}}^{\text{RB}}_{\theta|\mathcal{D}_0}(e)$

Input:$\{\theta_i\}_{i=1}^{l} \sim P(\theta \mid \mathcal{D}_0)$, Data Matrix $\mathbf{X}_e$, Treatment Vector $\mathbf{t}_e$, Variance σ^2
Output: $\widehat{\text{EIG}}^{\text{RB}}_{\theta|\mathcal{D}_0}(e)$
$S \leftarrow 0$
Sample $\{\theta_i\}_{i=1}^{l} \sim P(\theta \mid \mathcal{D}_0)$ for $l = NM_1$
Split $\{\theta_i\}_{i=1}^{l}$ as $\{(\theta_i), (\theta_{i,j})_{j=1}^{M_1}\}_{i=1}^{N}$
for $i \in \{1, \cdots, N\}$ **do**
  Sample $\mathbf{y}_e^{(i)} \sim P(\mathbf{y}_e|\mathbf{X}_e, \mathbf{t}_e, \theta_i)$
  $S \leftarrow S - \log \frac{1}{M_1}\sum_{j=1}^{M_1} P(\mathbf{y}_e^{(i)}|\theta_{i,j}, \mathbf{X}_e, \mathbf{t}_e)$
end for
$\widehat{\text{EIG}}^{\text{RB}}_{\theta|\mathcal{D}_0}(e) \leftarrow \frac{1}{N} S$
return $\widehat{\text{EIG}}^{\text{NMC}}_{\theta|\mathcal{D}_0}(e)$

Algorithm 3 Algorithm for $\widehat{\mathrm{EIG}}_{\theta_c|\mathcal{D}_0}(e)$

Input Data Matrix $\mathbf{X}_e$, Treatment Vector $\mathbf{t}_e$, Variance σ^2
Output: $\widehat{\mathrm{EIG}}_{\theta_c|\mathcal{D}_0}(e)$

$S \leftarrow 0$
Sample $\{\theta_i\}_{i=1}^N \sim P(\theta \mid \mathcal{D}_0)$
for $i \in \{1, \cdots, N\}$ **do**
  Sample $\mathbf{y}_e \sim P(\mathbf{y}_e|\theta, \mathbf{X}_e, \mathbf{t}_e)$
  Sample $\{\theta'_j\}_{j=1}^{M_1} \sim P(\theta \mid \mathcal{D}_0)$
  Sample $\{(\theta_{\mathrm{nc}})^{(ik)}\}_{k=1}^{M_2} \sim P(\theta_{\mathrm{nc}} \mid (\theta_c)_i, \mathcal{D}_0)$
  $\widehat{P}(\mathbf{y}_e^{(i)}|\mathbf{X}_e, \mathbf{t}_e, \mathcal{D}_0) \leftarrow \frac{1}{M_1}\sum_{j=1}^{M_1} P\left(\mathbf{y}_e^{(i)}|\theta'_j, \mathbf{X}_e, \mathbf{t}_e\right)$
  $\widehat{P}(\mathbf{y}_e^{(i)}|\theta_{\mathrm{c}}^{(i)}, \mathbf{X}_e, \mathbf{t}_e, \mathcal{D}_0) \leftarrow \frac{1}{M_2}\sum_{k=1}^{M_2} P\left(\mathbf{y}_e^{(i)}|\theta_{nc}^{(ik)} \cup \theta_c^{(i)}, \mathbf{X}_e, \mathbf{t}_e\right)$
  $S \leftarrow S + \log\left(\frac{\widehat{P}(\mathbf{y}_e^{(i)}|\theta_{\mathrm{c}}^{(i)}, \mathbf{X}_e, \mathbf{t}_e, \mathcal{D}_0)}{\widehat{P}(\mathbf{y}_e^{(i)}|\mathbf{X}_e, \mathbf{t}_e, \mathcal{D}_0)}\right)$
end for
return $\frac{1}{N}S$

A.2 Differential Privacy Definition

Definition A.1. *We say a randomised algorithm, $\mathcal{A}$ satisfies ϵ differential privacy if for any input dataset $\mathcal{D}$ and dataset $\mathcal{D}'$ differing by a single entry, we have*

$$P(\mathcal{A}(\mathcal{D}) \in \mathcal{O}) \leq \exp(\epsilon) P(\mathcal{A}(\mathcal{D}') \in \mathcal{O}).$$

A.3 Sensitivity of Linear Statistic

We derive the sensitivity of the linear EIG statistic in order to give a fair comparison with naive differential privacy in Section 5.3.

Proposition A.1. *Let:*

$$f(\mathbf{X}_e) = \log\det(\mathbf{X}_e^\top \mathbf{X}_e + \mathbf{X}_0^\top \mathbf{X}_0 + \Lambda_0) \tag{8}$$

If $\Lambda_0 = cI$ and $\|\mathbf{x}\|_\infty \leq M$ for all $\mathbf{x} \sim P_e(\mathbf{x})$ and e. Then we have that:

$$\Delta_f = \max_{\mathbf{X}'_e, \mathbf{X}_e} |f(\mathbf{X}'_e) - f(\mathbf{X}_e)| \leq \frac{Md}{\sqrt{c}} \tag{9}$$

Where $\mathbf{X}'_e, \mathbf{X}_e$ differ in at most one row. This implies $f(\mathbf{X}_e) + Z$ for $Z \sim \mathrm{Laplace}(\frac{Md}{\epsilon\sqrt{c}})$ is a ϵ differentially private release of $f(\mathbf{X}_e)$.

Proof. To prove this we will use the fact that if $\max_{\|\mathbf{X}\| \leq M} \|Df(\mathbf{X})\|_F = L$ we have that $|f(\mathbf{X}'_e) - f(\mathbf{X}_e)| \leq L\|\mathbf{X}'_e - \mathbf{X}_e\|_F$ for all $\mathbf{X}'_e, \mathbf{X}_e$ with norm bounded by M where $\|\cdot\|_F$ is the Frobenious norm. Write $\mathbf{E} = \mathbf{X}_0^\top \mathbf{X}_0 + \Lambda_0$, via the chain rule we have that $Df(\mathbf{X}) = \left(\mathbf{X}_e^\top \mathbf{X}_e + \mathbf{E}\right)^{-1} \mathbf{X}_e^\top$ Now:

$$\left\|\left(\mathbf{X}_e^\top \mathbf{X}_e + \mathbf{E}\right)^{-1} \mathbf{X}_e^\top\right\|_F = \sqrt{\mathrm{tr}\left(\left(\mathbf{X}_e^\top \mathbf{X}_e + \mathbf{E}\right)^{-1} \mathbf{X}_e^\top \mathbf{X}_e \left(\mathbf{X}_e^\top \mathbf{X}_e + \mathbf{E}\right)^{-1}\right)} \tag{10}$$

$$= \sqrt{\mathrm{tr}\left(\left(\mathbf{X}_e^\top \mathbf{X}_e + \mathbf{E}\right)^{-1} - \left(\mathbf{X}_e^\top \mathbf{X}_e + \mathbf{E}\right)^{-1} \mathbf{E} \left(\mathbf{X}_e^\top \mathbf{X}_e + \mathbf{E}\right)^{-1}\right)} \tag{11}$$

$$\leq \sqrt{\mathrm{tr}\left(\left(\mathbf{X}_e^\top \mathbf{X}_e + \mathbf{E}\right)^{-1}\right)} \leq \sqrt{\frac{d}{c}} \tag{12}$$

We have used the fact that $\left(\mathbf{X}_e^\top \mathbf{X}_e + \mathbf{E}\right)^{-1} \mathbf{E} \left(\mathbf{X}_e^\top \mathbf{X}_e + \mathbf{E}\right)^{-1}$ is positive semi definite and so has positive trace, and that the eigenvalues of $\mathbf{X}_e^\top \mathbf{X}_e + \mathbf{E}$ are bounded below by c so the eigenvalues of $\left(\mathbf{X}_e^\top \mathbf{X}_e + \mathbf{E}\right)^{-1}$ are bounded above by $\frac{1}{c}$. Finally for neighbouring datasets we can change at most d entries by M so $\|\mathbf{X}'_e - \mathbf{X}_e\|_F \leq \sqrt{d}M$ □

B Model Details

Throughout all methods will aim to model the outcome as some function plus an error, so:

$$Y_i = f(\mathbf{x}_i, t_i) + \epsilon_i \tag{13}$$

B.1 Models via Sampling

Bayesian Additive Regression Trees (BART) BART models (Chipman et al., 2012) f as the sum of L piecewise constant binary regression trees, so we have:

$$f(\mathbf{x}, t) = \sum_{l=1}^{L} g_i(\mathbf{x}, t, T_l, \mathbf{m}_l) \tag{14}$$

where T_l is a regression tree given by a partition $(\mathcal{A}_1, \cdots, \mathcal{A}_{\mathcal{B}(l)})$ of $\mathcal{X} \times \mathcal{T}$ and the set of leaf parameter values $\mathbf{m}_l = (m_{l1}, \cdots, m_{l\mathcal{B}(l)})$ so that:

$$g_l(\mathbf{x}, t) = m_j \text{ if } \mathbf{x}, t \in \mathcal{A}_j \tag{15}$$

The mean parameters are given with independent normal parameters $m_{lj} \sim \mathcal{N}(0, \sigma_m)$. Over trees, the prior is such that the probability of a node having children at depth d is given by:

$$\alpha(1+d)^{-\theta} \text{ for } \alpha \in (0,1), \theta \in [0, \infty) \tag{16}$$

The original BART model explores this space using Metropolis-Hastings Markov chain Monte Carlo, but we make use of XBART (He et al., 2019) for accelerated posterior sampling.

Bayesian Causal Forest Bayesian Causal Forests (BCF; Hahn et al., 2020) build upon BART models utilising specific parameterisations for causal inference tasks. The two parameterisations suggested in Hahn et al. (2020) are firstly:

$$f(\mathbf{x}, t) = \mu(\mathbf{x}) + t\tau(\mathbf{x}) \tag{17}$$

Where μ, τ are independent BART models. To draw specific attention to their parameters will write them as $\mu_{\theta_{\text{nc}}}, \tau_{\theta_{\text{c}}}$ noting that $\tau_{\theta_{\text{c}}}$ is a model for CATE. Hahn et al. (2020) note that this parameterisation is not invariant to which treatment is assigned as positive or negative, leading them to propose the following invariant parameterisation:

$$f_\theta(\mathbf{x}, t) = \tilde{\mu}_{\tilde{\theta}_{\text{nc}}}(\mathbf{x}) + b_t \tilde{\tau}_{\tilde{\theta}_{\text{c}}}(\mathbf{x}) \tag{18}$$

Where $b_t \sim \mathcal{N}(0, \frac{1}{2})$. Under this parameterisation a CATE estimate is given by $(b_1 - b_0)\tilde{\tau}_{\tilde{\theta}_{\text{c}}}(\mathbf{x})$. When sampling we make use of the accelerated BCF approach (Krantsevich et al., 2023) which builds upon XBART and uses the following slightly modified model:

$$f_\theta(\mathbf{x}, t) = a\tilde{\mu}_{\tilde{\theta}_{\text{nc}}}(\mathbf{x}) + b_t \tilde{\tau}_{\tilde{\theta}_{\text{c}}}(\mathbf{x}) \tag{19}$$

Where $a \sim \mathcal{N}(0, 1)$.

We define the set θ_{c} to be any parameters affiliated with the τ model, including b_t for the invariant parameterisation. In order to sample $P(\theta_{\text{nc}} | \theta_{\text{c}}, \mathcal{D}_0)$ we refit θ_{nc} parameters on the dataset $\mathcal{D}_0$ to the residuals resulting from subtracting the τ portion of the model. So this refitting μ is as follows for the standard parametisation:

$$Y - t\tau_{\theta_{\text{c}}}(\mathbf{x}) = \mu_{\theta_{\text{nc}}}(\mathbf{x}) \tag{20}$$

Or for the accelerated BCF approach (Krantsevich et al., 2023):

$$Y - b_t \tilde{\tau}_{\tilde{\theta}_{\text{c}}}(\mathbf{x}) = a\tilde{\mu}_{\tilde{\theta}_{\text{nc}}}(\mathbf{x}) \tag{21}$$

Where any parameters on the left hand side are fixed.

B.2 Closed Form Models

In this section we give details of models for which the EIG is available in closed form. We provide details of the models as well as proofs for the expressions.

B.2.1 Bayesian Polynomial Regression Derivations

In this we derive the results for Bayesian polynomial regression. We have modelled our data as:

$$y \sim \mathcal{N}(\phi(\mathbf{x}, t)^\top \theta, \sigma^2), \quad \theta \sim \mathcal{N}(\mu_0, \sigma^2 \Lambda_0^{-1}) \tag{22}$$

In this context, the posterior is available in closed form as:

$$\theta|\mathcal{D}_0 \sim \mathcal{N}(\mu_0, \sigma^2 \tilde{\Lambda}_0^{-1}) \tag{23}$$

$$\tilde{\Lambda}_0 = \left(\phi\left(\mathbf{X}_0, \mathbf{t}_0\right)^\top \phi\left(\mathbf{X}_0, \mathbf{t}_0\right) + \Lambda_0^{-1}\right) \tag{24}$$

$$\mu_0 = \left(\Lambda_0\right)^{-1} \left(\phi\left(\mathbf{X}_0, \mathbf{t}_0\right)^\top \phi\left(\mathbf{X}_0, \mathbf{t}_0\right) \hat{\theta}_0 + \Lambda_0 \mu_0\right) \tag{25}$$

$$\hat{\theta}_0 = \left(\phi\left(\mathbf{X}_0, \mathbf{t}_0\right)^\top \phi\left(\mathbf{X}_0, \mathbf{t}_0\right)\right)^{-1} \phi\left(\mathbf{X}_0, \mathbf{t}_0\right)^\top \mathbf{y}_0 \tag{26}$$

Expected Information Gain Over all parameters We use the fact that $\text{EIG}_{\theta|\mathcal{D}_0}$ can be written as:

$$\text{EIG}_{\theta|\mathcal{D}_0}(e) = \mathbb{E}_{P(\mathbf{y}_e|\mathbf{X}_e, \mathbf{t}_e, \mathcal{D}_0)}\left[H[P(\theta|\mathcal{D}_0)] - H[P(\theta|\mathcal{D}_0, \mathcal{D}_e)]\right],$$

As the posterior over θ is Gaussian we can directly evaluate these expressions using the closed form entropy for a Gaussian distribution as:

$$H[P(\theta|\mathcal{D}_0)] = \frac{n_e}{2}\left(1 + \log(2\pi)\right) + \frac{1}{2}\left(\log\det\left(\tilde{\Lambda}_0^{-1}\right)\right)$$

The distribution $\theta|\mathcal{D}_0, \mathcal{D}_e)$ can be obtained as above, where we now use $\tilde{\Lambda}_0$ as the prior precision matrix before updating on $\mathcal{D}_0$. This gives:

$$H[P(\theta|\mathcal{D}_0)] = \frac{n_e}{2}\left(1 + \log(2\pi)\right) + \frac{1}{2}\left(\log\det\left(\left(\phi\left(\mathbf{X}_e, \mathbf{t}_e\right)^\top \phi\left(\mathbf{X}_e, \mathbf{t}_e\right) + \tilde{\Lambda}_0\right)^{-1}\right)\right)$$

Using the above form for $\tilde{\Lambda}_0$ and collecting all constants gives the expression presented in the text.

$\text{EIG}_{\theta_c|\mathcal{D}_0}$: Expected Information Gain Over all parameters This follows as above but using the fact that the block form of the matrices to allow us write the covariance precision matrix for the post host posterior over θ_c as follows:

$$\left(\mathbf{t}_0 \odot \phi_c(\mathbf{X}_0)\right)^\top \left(\mathbf{t}_0 \odot \phi_c(\mathbf{X}_0)\right) + \left(\Lambda_0\right)_{[c,c]}$$

Where $\left(\Lambda_0\right)_{[c,c]}$ corresponds to block of Λ_0 with entries after $[c, c]$ in the row and column.

B.3 Causal Multitask Gaussian Processes (Alaa and Van Der Schaar, 2017)

In this work, CATE is modelled using a multitask Gaussian process (Bonilla et al., 2007). Multitask Gaussian Processes use a GP in vector-valued Reproducing Kernel Hilbert Space (vv-RKHS) to share information between tasks (Alvarez et al., 2012). In Alaa and Van Der Schaar (2017), learning the conditional outcome function for each treatment is seen as a separate task, so we jointly model:

$$Y|\mathbf{x}, t \sim \mathcal{N}(0, f_t(\mathbf{x}), \sigma_t^2) \tag{27}$$

Where each f_t is a Gaussian Process. The kernel $\mathbf{K}_\eta : \mathcal{X} \times \mathcal{X} \to \mathbb{R}^{2\times 2}$ is now a symmetric positive semi-definite matrix-valued function, with hyper-parameters η. In the case of Alaa and Van Der Schaar (2017) they use a *linear model of coregionalization*[‡], giving the kernel as:

$$\mathbf{K}_\eta(\mathbf{x}, \mathbf{x}') = \mathbf{A}_0 k_0(\mathbf{x}, \mathbf{x}') + \mathbf{A}_1 k_1(\mathbf{x}, \mathbf{x}') \tag{28}$$

[‡]See Alvarez et al. (2012) for more details.

Where k_t is the RBF kernel, given by:

$$k_t(\mathbf{x}, \mathbf{x}') = \exp(-\frac{1}{2}(\mathbf{x} - \mathbf{x}')^\top \mathbf{R}_t^{-1}(\mathbf{x} - \mathbf{x}')) \tag{29}$$

Where $\mathbf{R}_t^{-1} = \text{diag}\left(\ell_{1,t}^2, \cdots, \ell_{d,t}^2\right)$ and $\ell_{j,t}$ is the length-scale parameter for the treatment $T = t$ in the j^{th} coordinate of $\mathbf{x}$. $\mathbf{A}_t$ is given by:

$$\mathbf{A}_0 = \begin{bmatrix} \theta_{00}^2 & \rho_0 \\ \rho_0 & \theta_{01}^2 \end{bmatrix}, \mathbf{A}_1 = \begin{bmatrix} \theta_{10}^2 & \rho_1 \\ \rho_1 & \theta_{11}^2 \end{bmatrix}. \tag{30}$$

Where $\theta_i j$ and ρ_i determine the variances and covariances of the shared tasks f_t. So we have that the full set of hyper-parameters $\eta = (\theta_0, \theta_1, \mathbf{R}_0, \mathbf{R}_1, \mathbf{A}_0, \mathbf{A}_1)$. Once all these hyper-parameters have been learnt we have that the covariance between different function evaluations, $f_t(\mathbf{x}), f_{t'}(\mathbf{x}')$, is given the t, t' coordinate of $\mathbf{K}_\eta(\mathbf{x}, \mathbf{x}')$. So:

$$\text{cov}(f_t(\mathbf{x}), f_{t'}(\mathbf{x}')) = \mathbf{K}_\eta(\mathbf{x}, \mathbf{x}')_{[t,t']} \tag{31}$$

Now, if we let $K((\mathbf{x}, t), (\mathbf{x}', t') = \mathbf{K}_\eta(\mathbf{x}, \mathbf{x}')_{[t,t']}$ then we can obtain the posterior kernel in a similar way to the standard case. Precisely if we the training data be given by:

$$\tilde{\mathbf{X}} = \left[\{\mathbf{x}_i\}_{T_i=0}, \{\mathbf{x}_i\}_{T_i=1}\right]^T, \tag{32}$$

$$\tilde{\mathbf{Y}} = \left[\left\{y_i^{(T_i)}\right\}_{T_i=0}, \left\{y_i^{(T_i)}\right\}_{t_i=1}\right]^T, \tag{33}$$

$$\mathbf{\Sigma} = \text{diag}\left(\sigma_0^2 \mathbf{I}_{n-n_1}, \sigma_1^2 \mathbf{I}_{n_1}\right) \tag{34}$$

$$n_1 = \sum_i W_i, \tag{35}$$

$$\mathbf{K}_\eta(x) = (\mathbf{K}_\eta(x, X_i)_i .) \tag{36}$$

Then we have that the posterior mutlitask GP has mean and posterior kernel given by:

$$m^{\text{post}}(\mathbf{x}) = \mathbf{K}_\eta^T(\mathbf{x})\left(\mathbf{K}_\eta(\mathbf{X}, \mathbf{X}) + \Sigma\right)^{-1}\tilde{\mathbf{Y}} \tag{37}$$

$$\mathbf{K}_{\eta^*}^{\text{post}}(\mathbf{x}, \mathbf{x}') = \mathbf{K}_{\eta^*}(\mathbf{x}, \mathbf{x}') - \mathbf{K}_{\eta^*}(\mathbf{x})\left(\mathbf{K}_\eta(\mathbf{X}, \mathbf{X}) + \Sigma\right)^{-1}\mathbf{K}_{\eta^*}^T(\mathbf{x}') \tag{38}$$

$$\tag{39}$$

This leads to the following posterior over CATE, where $\mathbf{e} = [-1, 1]^\top$:

$$\tilde{\tau}(x) \sim \mathcal{N}(m^{\text{post}}(\mathbf{x})^\top \mathbf{e}, \mathbf{e}^\top \mathbf{K}_{\eta^*}^{\text{post}}(\mathbf{x}, \mathbf{x}')\mathbf{e}) \tag{40}$$

B.3.1 Expected Information Gain

First, let $\mathbf{X}_e^{(1)}, \mathbf{X}_e^{(0)}$ and $\mathbf{y}_e^{(1)}, \mathbf{y}_e^{(0)}$ be the covariance and outcomes for environment e that is treated and untreated respectively. To avoid confusion with treatment we will use $\mathbf{X}_{e^*}$ to refer to the host environment for this derivation. We will also use $\mathbf{K}_{|\mathcal{D}_0}$ to refer to the posterior kernel. Now to derive the Expected Information Gain in closed form we need the distribution of the following vector:

$$\begin{bmatrix} \mathbf{y}_e^{(1)} \\ \mathbf{y}_e^{(0)} \\ \tilde{\tau}(\mathbf{X}_{e^*}) \end{bmatrix} | \mathbf{X}_e, \mathcal{D}_0 \sim \mathcal{N}(\mathbf{m}, \Sigma) \tag{41}$$

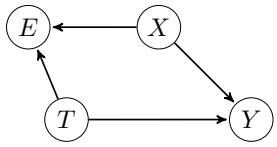
Where we have that:

$$\Sigma_1 = \begin{bmatrix} \mathbf{K}_{|\mathcal{D}_0}(\mathbf{X}_e^{(0)}, \mathbf{X}_e^{(0)}) + \sigma_0^2 I_{n_e^0} & \mathbf{K}_{|\mathcal{D}_0}(\mathbf{X}_e^{(1)}, \mathbf{X}_e^{(0)})) \\ \mathbf{K}_{|\mathcal{D}_0}(\mathbf{X}_e^{(0)}, \mathbf{X}_e^{(1)}) & \mathbf{K}_{|\mathcal{D}_0}(\mathbf{X}_e^{(1)}, \mathbf{X}_e^{(1)}) + \sigma_1^2 I_{n_e^1} \end{bmatrix} \tag{42}$$

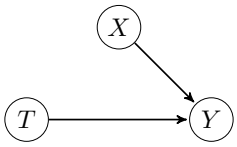
$$\Sigma_2 = \mathbf{K}_{|\mathcal{D}_0}(\mathbf{X}_{e^*}^{(1)}, \mathbf{X}_{e^*}^{(1)}) + \mathbf{K}_{|\mathcal{D}_0}(\mathbf{X}_{e^*}^{(0)}, \mathbf{X}_{e^*}^{(0)}) - 2\mathbf{K}_{|\mathcal{D}_0}(\mathbf{X}_{e^*}^{(1)}, \mathbf{X}_{e^*}^{(0)}) \tag{43}$$

$$\Sigma = \begin{bmatrix} \Sigma_1 & \Sigma_{12} \\ \Sigma_{12}^\top & \Sigma_2 \end{bmatrix} \text{ where } \Sigma_{12} = \begin{bmatrix} \mathbf{K}_{|\mathcal{D}_0}(\mathbf{X}_{e^*}^{(1)}, \mathbf{X}_e^{(0)}) - \mathbf{K}_{|\mathcal{D}_0}(\mathbf{X}_{e^*}^{(0)}, \mathbf{X}_e^{(0)}) \\ \mathbf{K}_{|\mathcal{D}_0}(\mathbf{X}_{e^*}^{(1)}, \mathbf{X}_e^{(1)}) - \mathbf{K}_{|\mathcal{D}_0}(\mathbf{X}_{e^*}^{(0}, \mathbf{X}_e^{(1)}) \end{bmatrix} \tag{44}$$

Now from the covariance matrix we can derive the standard Expected Information Gain $\text{EIG}_{\theta|\mathcal{D}_0}$ and CATE-specific $\text{EIG}_{\theta_{c|\mathcal{D}_0}}$. Throughout we will use $\mathbf{K}$ to the posterior kernel irrespective if it has been fit or not.



(a) Before merging: causal structure in D_{host}, or D_{twin} or D_{comp} *taken separately.* E acts as a collider and thus creates a dependency between X and T



(b) After merging: causal structure in $D_{host} \cup D_{comp}$

Figure 4: Causal structure for the illustrative experiment.

Expected Information Gain over the conditional outcome parameters For the $\text{EIG}_{\mathbf{f}}$ we use the BALD form:

$$H(\mathbf{y}_e|\mathbf{X}_e, \mathcal{D}_0) - H(\mathbf{y}|\mathbf{X}_e, \mathbf{f}, \mathcal{D}_0) \tag{45}$$

Using the standard form of entropy for a Gaussian distribution we can read $H(\mathbf{y}_e|\mathbf{X}_e, \mathcal{D}_0)$ off of the covariance matrix above as:

$$H(\mathbf{y}_e|\mathbf{X}_e, \mathcal{D}_0) = \frac{n_e}{2}\left(1 + \log(2\pi)\right) + \frac{1}{2}\left(\log|(\Sigma_1|)\right) \tag{46}$$

$$\text{where } \Sigma_1 = \begin{bmatrix} \mathbf{K}_\eta(\mathbf{X}_e^{(0)}, \mathbf{X}_e^{(0)}) + \sigma_0^2 I_{n_e^0} & \mathbf{K}_\eta(\mathbf{X}_e^{(1)}, \mathbf{X}_e^{(0)})) \\ \mathbf{K}_\eta(\mathbf{X}_e^{(0)}, \mathbf{X}_e^{(1)}) & \mathbf{K}_\eta(\mathbf{X}_e^{(1)}, \mathbf{X}_e^{(1)}) + \sigma_1^2 I_{n_e^1} \end{bmatrix} \tag{47}$$

And $H(\mathbf{y}_e|\mathbf{f}, \mathbf{X}_e, \mathcal{D}_0)$ being:

$$H(\mathbf{y}_e|\mathbf{f}, \mathbf{X}_e, \mathcal{D}_0) = \frac{n_e}{2}\left(1 + \log(2\pi)\right) + \frac{1}{2}\left(n_e^{(0)}\log(\sigma_0^2) + n_e^{(1)}\log(\sigma_1^2)\right) \tag{48}$$

This gives the Expected Information Gain as:

$$\mathcal{I}_{\mathbf{f}}(e) = \frac{1}{2}\log|(\Sigma_1|) - \frac{1}{2}\left(n_e^{(0)}\log(\sigma_0^2) + n_e^{(1)}\log(\sigma_1^2)\right) \tag{49}$$

Expected Information Gain on the CATE parameters We now target an Expected Information Gain on the CATE parameters on our host dataset, so $\tilde{\tau}(\mathbf{X}_0)$. By using the fact that the Expected Information Gain can be written as the mutual information between $\tilde{\tau}(X_0)$ and the observed outcomes in dataset e, we use the closed form mutual information for Gaussian's to directly write this as:

$$\mathcal{I}_{\tilde{\tau}(\mathbf{X}_0)}(e) = \frac{1}{2}\log\left(\frac{|\Sigma_1||\Sigma_2|}{|\Sigma|}\right) \tag{50}$$

$$\text{where } \Sigma_2 = \mathbf{K}_\eta(\mathbf{X}_{e^*}^{(1)}, \mathbf{X}_{e^*}^{(1)}) + \mathbf{K}_\eta(\mathbf{X}_{e^*}^{(0)}, \mathbf{X}_{e^*}^{(0)}) - 2\mathbf{K}_\eta(\mathbf{X}_{e^*}^{(1)}, \mathbf{X}_{e^*}^{(0)}) \tag{51}$$

C Experimental Details

General experimental settings and hyperparameters All standard deviations and precisions were taken equal to 1. Throughout experiments, priors were taken as zero-valued vector.

In the **illustrative experiment**, 400 samples were used for computing outer expectations, and 800 samples for inner expectations. Here, we consider $X = (X_0, X_1, X_2) \in \mathbb{R}^3$. We use the sampling selection function $P_{\text{host}}(x,t) = \text{sigmoid}(1+2\times x_0 - x_1 + 2\times t) + \epsilon$ and outcome model $y = 1 + x_0 - x_1 + x_2 + 5\times t + 2\times x_0 + 2\times x_0 - 4\times x_2 + \epsilon$ with $X_0 \sim \mathcal{B}(12,3)$, $X_1 \sim \mathcal{N}(4,1)$, $X_2 \sim \mathcal{B}(1,7)$ and $\epsilon \sim \mathcal{N}(0,1)$.

In both **ranking experiments**, the selection functions are randomly generated. We first generate a binary vector of the size of the dimension of X to define the subset of covariates that would be impact selection. We then generate two other random vectors, one for the multiplicative coefficients for each selected covariate, and another to define a power for each term in the sum. Ultimately, the probability of selection is taken as the sigmoid of this randomly generated polynomial.

In the **ranking experiment with IHDP**, 10 candidates were generated with a sample size ranging from 300 to 500. The host sample size was equal to 400. The experiment was across 20 seeds. We kept a minimum of 50 subjects in each treatment group. The hold out test dataset had a sample size of 2000. For the linear model, X, T and $X \times T$ were included. For the Gaussian Process model, a maximum of 1000 iterations was set. In CBF, both the predictive and conditional models were used with a maximum depth of 250, and a shrinkage $\alpha = 0.95$.

In the **ranking experiment with Lalonde**, 15 candidates were generated with a sample size ranging from 200 to 400. The host sample size was equal to 600. The experiment was across 20 seeds. We kept a minimum of 50 subjects in each treatment group. The hold out test dataset had a sample size of 2000. For the linear model, X, T and $X \times T$ were included. For the Gaussian Process model, a maximum of 1000 iterations was set. In CBF, both the predictive and conditional models were used with a maximum depth of 200, and a shrinkage $\alpha = 0.9$.

Datasets We describe the two datasets used in our experiments, with high-level summary given in Table 3.

Table 3: Description of the datasets: Lalonde (LaLonde, 1986) and IHDP (Louizos et al., 2017).

	ihdp	**lalonde**
Number of samples	747	16,177
Number of features	24	8

The **Infant Health and Development Program, or IHDP** is a randomized controlled study designed to assess how home visits by specialist doctors impact the cognitive test scores of premature infants. Initially, the dataset serves as a benchmark for evaluating treatment effect estimation algorithms, as described in Hill (2011). This evaluation introduces selection bias by excluding non-random subsets of treated individuals to construct an observational dataset, with outcomes derived from the original covariates and treatments.

The **Lalonde** originates from the National Supported Work Demonstration used by Dehejia and Wahba (1999) to evaluate propensity score matching methods. The data consists of demographic variables (age, race, academic background, and previous real earnings), as well as a treatment indicator. The outcome is the real earnings in the year 1978.

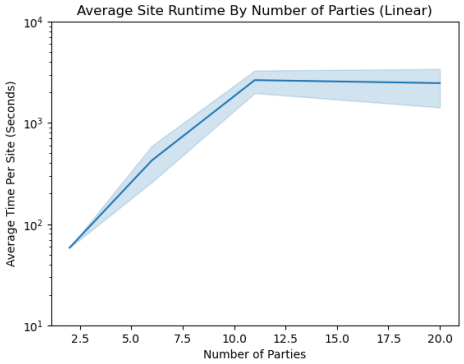
Compute times Approximate compute times for the ranking experiment on causal benchmark datasets are given in Table 4. Experiments were performed on an Apple M3 chip with a 12-core CPU and 18 GB of RAM.

Table 4: Approximate compute times.

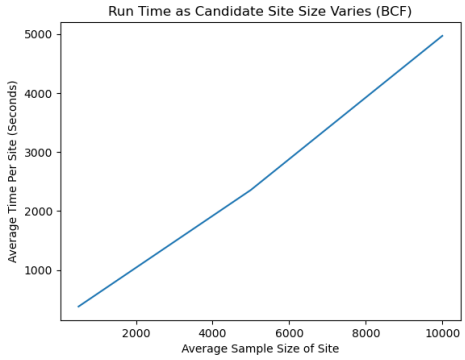
	ihdp	**lalonde**
Polynomial	$<$ 1 min	$<$ 1 min
Causal GP	3 mins	3 mins
BART	14 mins	9 mins

D Further experimental results

For completeness, we include the performance of all baselines on the IHDP dataset in Table 5 and the Lalonde in Table 6.



(a) In this we repeat the multi-party computation experiments from Section 5.3 varying the number of sites. Each site is treated as a separate party in the multi-party computation and the average runtime per site is reported. Runtime recorded on an M3 macbook. Averaged over 5 runs.



(b) Runtime of Bayesian Causal Forest as the candidate sample size increases. We run the algorithm for the task of selecting between 5 sites with average size 500, 5000, and 10,000.

Table 5: IHDP dataset ranking experiment results with 10 candidate datasets

Model	**Objective**	$\boldsymbol{\rho}(\uparrow)$	**p@1** $(\uparrow)$	**p@3** $(\uparrow)$	**p@5** $(\uparrow)$
Polynomial	$\text{EIG}_{\theta_{c}\mid\mathcal{D}_0}$	$\mathbf{0.70 \pm 0.08}$	$\mathbf{0.50 \pm 0.15}$	$\mathbf{0.70 \pm 0.04}$	$\mathbf{0.78 \pm 0.04}$
	$\text{EIG}_{\theta\mid\mathcal{D}_0}$	$\mathbf{0.68 \pm 0.06}$	$\mathbf{0.50 \pm 0.15}$	$\mathbf{0.70 \pm 0.06}$	$\mathbf{0.76 \pm 0.04}$
	PropScore Error	0.40 ± 0.11	0.40 ± 0.15	0.60 ± 0.15	0.66 ± 0.04
	Sample Size	0.34 ± 0.08	0.10 ± 0.08	0.27 ± 0.04	0.50 ± 0.06
	CovDist	0.03 ± 0.07	0.10 ± 0.08	0.23 ± 0.06	0.48 ± 0.04
Causal GP	$\text{EIG}_{\tilde{\tau}(X_0)\mid\mathcal{D}_0}$	$\mathbf{0.49 \pm 0.06}$	$\mathbf{0.50 \pm 0.15}$	$\mathbf{0.50 \pm 0.08}$	$\mathbf{0.62 \pm 0.06}$
	$\text{EIG}_{\mathbf{f}\mid\mathcal{D}_0}$	0.33 ± 0.06	0.30 ± 0.15	0.43 ± 0.05	$\mathbf{0.60 \pm 0.04}$
	Sample Size	0.31 ± 0.12	0.10 ± 0.20	0.20 ± 0.07	0.46 ± 0.05
	PropScore Error	0.21 ± 0.09	0.10 ± 0.20	0.30 ± 0.07	0.43 ± 0.05
	CovDist	0.03 ± 0.06	0.10 ± 0.20	0.160 ± 0.04	0.46 ± 0.05
Bayesian CF	$\text{EIG}_{\theta_{c}\mid\mathcal{D}_0}$	$\mathbf{0.54 \pm 0.10}$	$\mathbf{0.60 \pm 0.15}$	$\mathbf{0.63 \pm 0.08}$	$\mathbf{0.70 \pm 0.04}$
	$\text{EIG}_{\theta\mid\mathcal{D}_0}$	0.36 ± 0.10	0.30 ± 0.14	0.50 ± 0.07	$\mathbf{0.66 \pm 0.05}$
	PropScore Error	0.45 ± 0.11	$\mathbf{0.60 \pm 0.14}$	$\mathbf{0.63 \pm 0.08}$	$\mathbf{0.70 \pm 0.04}$
	Sample Size	0.16 ± 0.09	0.20 ± 0.11	0.26 ± 0.06	0.52 ± 0.05
	CovDist	0.07 ± 0.09	0.00 ± 0.00	$0.26 \pm 0.06a$	0.46 ± 0.05

E Related work: further details

On Causal Federated Learning Federated learning is a distributed machine learning approach that enables multiple parties to collaboratively train a shared model while keeping their raw data decentralised and private. Various federated learning approaches have been proposed, including federated stochastic gradient descent (Shokri and Shmatikov, 2015), federated averaging (McMahan et al., 2017), and more recently, methods for joint learning of deep neural network models (Sattler et al., 2019; Wang et al., 2020). However, these algorithms do not inherently support causal inference, as the dissimilar distributions across different data sources may lead to biased causal effect estimation. To date, limited research has been conducted on the federated estimation of causal effects, highlighting the need for further exploration in this area. Due to the scope of our work, in the following paragraphs, we will focus on presenting Federated Learning methods for CATE estimation, where covariate distribution and treatment allocation are not assumed to be identical across datasets.

Several methods propose disentangling the loss function to facilitate federated learning. Vo et al. (2022) propose CausalRFF, an adaptive kernel approach for causal inference that utilises Random Fourier Features to partition the loss function into multiple components, with each component corresponding to a specific data source. However, CausalRFF approach lacks strong privacy guarantees to prevent data recovery, and modeling complex non-linear

Table 6: Lalonde dataset ranking experiment results with 15 candidate datasets

Model	Objective	$\rho(\uparrow)$	p@1 ($\uparrow$)	p@3 ($\uparrow$)	p@5 ($\uparrow$)
Polynomial	$\text{EIG}_{\theta_{c\mid\mathcal{D}_0}}$	$\mathbf{0.47 \pm 0.05}$	$\mathbf{0.40 \pm 0.11}$	$\mathbf{0.60 \pm 0.05}$	$\mathbf{0.79 \pm 0.04}$
	$\text{EIG}_{\theta\mid\mathcal{D}_0}$	$\mathbf{0.43 \pm 0.05}$	$\mathbf{0.35 \pm 0.13}$	0.48 ± 0.04	0.53 ± 0.03
	PropScore Error	0.19 ± 0.10	0.20 ± 0.17	0.32 ± 0.06	0.49 ± 0.05
	Sample Size	0.24 ± 0.10	0.25 ± 0.08	0.38 ± 0.08	0.58 ± 0.06
	CovDist	0.20 ± 0.04	0.25 ± 0.07	0.38 ± 0.06	0.48 ± 0.07
Causal GP	$\text{EIG}_{\tilde{\tau}(X_0)\mid\mathcal{D}_0}$	$\mathbf{0.42 \pm 0.07}$	$\mathbf{0.5 \pm 0.12}$	$\mathbf{0.55 \pm 0.05}$	$\mathbf{0.72 \pm 0.03}$
	$\text{EIG}_{\mathbf{f}\mid\mathcal{D}_0}$	$\mathbf{0.41 \pm 0.04}$	$\mathbf{0.4 \pm 0.1}$	0.43 ± 0.04	0.58 ± 0.07
	PropScore Error	0.19 ± 0.05	0.21 ± 0.15	0.32 ± 0.06	0.43 ± 0.07
	Sample Size	0.13 ± 0.07	0.15 ± 0.08	0.31 ± 0.09	0.53 ± 0.06
	CovDist	0.22 ± 0.04	0.25 ± 0.07	0.36 ± 0.08	0.48 ± 0.04
Bayesian CF	$\text{EIG}_{\theta_{c\mid\mathcal{D}_0}}$	$\mathbf{0.44 \pm 0.05}$	$\mathbf{0.45 \pm 0.08}$	$\mathbf{0.55 \pm 0.05}$	$\mathbf{0.78 \pm 0.04}$
	$\text{EIG}_{\theta\mid\mathcal{D}_0}$	$\mathbf{0.39 \pm 0.05}$	$\mathbf{0.45 \pm 0.06}$	0.43 ± 0.04	0.52 ± 0.03
	PropScore Error	0.22 ± 0.06	0.2 ± 0.07	0.32 ± 0.06	0.47 ± 0.07
	Sample Size	0.18 ± 0.07	0.2 ± 0.06	0.35 ± 0.09	0.41 ± 0.06
	CovDist	0.34 ± 0.03	0.3 ± 0.07	0.42 ± 0.08	0.61 ± 0.04

relationships remains challenging (Almodóvar et al., 2023). Liu et al. (2024) introduce a Bayesian method where parameters refer to a local disentangled loss and are updated cross-silo using server aggregation. Similarly, Vo et al. (2023) divide the loss function into site-specific functions, and specify a variational posterior distribution for each local loss. Instead of tackling the loss function, Almodóvar et al. (2023)) introduce a method based on disentanglement of latent factors into instrumental, confounding, and risk factors, which are then used for treatment effect estimation. They apply federated averaging on a neural network-based generative causal inference model. Ultimately, FedCov (Tarumi et al., 2023) is a parametric method for federated adjustment of covariate distributions between sites, where sample weights are derived from a propensity-like model. In all the aforementioned methods, the accuracy of causal estimation is reduced due to the constraints of federated learning. Conversely, our approach does not alter the causal estimation step, thereby maintaining optimal estimation accuracy. The framework we propose focuses on federated learning of the Expected Information Gain that would be obtained by merging with a dataset. While some Federated Causal Learning methods (Vo et al., 2022, 2023) provide uncertainty bounds, which could potentially be used to decide which dataset to merge with by comparing the uncertainty in these bounds, the provided bounds apply to the federated estimate and not the causal estimate potentially obtained after merging. Ultimately, none of these methods provide strong privacy guarantees, such as differential privacy (DP), which would ensure that raw data cannot be recovered from the model parametrisation or summary statistics. Moreover, all these methods use the outcome values for training their federated model, and outcome values tend to be more sensitive in nature.

On Causal Differential Privacy Contrasting with previous approaches, Niu et al. (2022) introduce a meta-algorithm that adds differential privacy (DP) guarantees to various popular CATE estimation frameworks, addressing the privacy concerns mentioned earlier. However, their method relies on multiple sample splitting, where separate subsets of the data are used for estimating the propensity score and the joint response model. This approach allows for parallel composition, a property of differential privacy. In contrast, our work prioritises data efficiency, and aims to utilise the entire dataset for CATE estimation without the need for sample splitting.

On Bayesian Experimental Design Bayesian Active Learning by Disagreement (BALD) (Houlsby et al., 2011) is a method designed to strategically acquire training data by focusing on regions of high uncertainty. BALD introduces an acquisition function rooted in information theory, which guides the data acquisition process. When reducing entropy towards all parameters in Section 3.1, we introduce a new setting for BALD where dataset are considered as data points. In the CausalBALD (Jesson et al., 2021) approach, the acquisition function is altered to specifically target areas where the distributions of different treatment groups overlap, thereby maximizing sample efficiency for learning personalised treatment effects. CausalBALD is also made for the acquisition of individual data points. However, contrarily to BALD, CausalBALD's acquisition function cannot provide a scalar measure if we compute it for a dataset (i.e. a matrix $\{\mathbf{x}_i, t_i\}_{i=1}^{n_e}$) instead of data points (i.e. a vector $\mathbf{x}_i, t_i$ for a

given i). To apply CausalBALD in our setting, one would need to approximate the higher-order interaction terms between all combinations of data points within each dataset, thus making the computation intractable.