Robust Causal Inference with Multi-task Gaussian Processes

Enhancing Generalization and Calibration through Data-Aware Kernel and Prior Design

Logan Ritter

Supervisor(s): Dr. Jesse Krijthe, Rickard Karlsson

EEMCS, Delft University of Technology, The Netherlands

A Thesis Submitted to the EEMCS Faculty of Delft University of Technology
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering

June 22, 2025

Name of the student: Logan Ritter
Final project course: CSE3000 Research Project
Thesis committee: Dr. Jesse Krijthe, Rickard Karlsson, Ricardo Marroquim

Abstract

Causal Multi-task Gaussian Processes (CMGPs) provide a Bayesian approach for estimating individualized treatment effects by modeling potential outcomes as correlated functions. However, they struggle under high-dimensionality and treatment imbalance, leading to overfitting and unreliable uncertainty estimates. This study examines two failure modes: poor generalization in high-dimensional spaces and overconfident predictions in low-overlap regions. To address these, two data-aware enhancements are proposed: an overlap-adaptive kernel that scales similarity based on local treatment density, and a regularized prior that down-weights unstable features using marginal treatment effect variance. Evaluations on synthetic data and the IHDP benchmark show improved effect estimation, credible interval calibration, and robustness in challenging settings. These findings highlight practical strategies for enhancing CMGPs in real-world causal inference tasks.

1 Introduction

Machine learning models have achieved significant success in predictive tasks across domains such as healthcare, economics, and public policy. However, in high-stakes applications, prediction alone is often insufficient. Decision-makers require insight into the outcomes of hypothetical interventions, motivating the field of causal inference [1], which focuses on estimating the effects of treatments rather than identifying associations.

A variety of methods have been proposed for causal effect estimation. Classical strategies such as propensity score matching, inverse probability weighting, and doubly robust estimation [2], [3] rely on strong parametric assumptions and often scale poorly. More recent machine learning approaches—such as Bayesian Additive Regression Trees (BART) [4], Causal Forests [5], and representation learning methods like TARNet and CFRNet [6]—address nonlinearity and heterogeneity but typically lack calibrated uncertainty estimates, limiting their reliability in real-world settings.

The Causal Multi-task Gaussian Process (CMGP) [7] provides a Bayesian framework for individualized treatment effect estimation. CMGP models potential outcomes as correlated outputs of a multi-output Gaussian process and uses automatic relevance determination (ARD) to assign feature-specific relevance. Empirical Bayes estimation is employed to trade off predictive accuracy on observed outcomes with uncertainty on unobserved counterfactuals, enabling estimation of both point estimates and credible intervals.

Despite these advantages, CMGPs face two critical challenges in practical applications. First, in high-dimensional covariate spaces, ARD kernels become unstable due to the increasing number of hyperparameters. This makes lengthscale estimation unreliable and prone to overfitting unless ample data are available [8]. Second, in regions with low treatment overlap, CMGPs often yield overconfident and poorly calibrated predictions. Stationary kernels apply global smoothness assumptions, leading to excessive extrapolation in data-sparse areas and underestimation of uncertainty [9].

These limitations motivate the central research question of this work:

> *How can data-aware enhancements to kernel design and prior specification improve the generalization, calibration, and robustness of CMGPs in high-dimensional and imperfect observational data?*

In this context, a data-aware enhancement refers to a mechanism that adapts the model's structure or regularization based on empirical properties of the observed data, such as local density, imbalance, or residual variability. This paper proposes two such enhancements. The first modifies the CMGP kernel by introducing overlap-aware scaling, which adapts kernel smoothness based on treatment density and covariate variation to prevent overconfident extrapolation in low-overlap regions. The second introduces a variance-weighted ARD prior, which regularizes feature-specific prior variances based on smoothed residuals, thereby mitigating overfitting in high-dimensional or low-sample regimes.

Two hypotheses are investigated. First, that overlap-aware kernel scaling improves uncertainty calibration and credible interval coverage in sparse or imbalanced regions, where traditional kernels are prone to overconfidence [10]. Second, that variance-informed ARD regularization reduces overfitting by down-weighting unstable features, aligning with recent evidence that such regularization enhances generalization in causal models [11], [12].

This study contributes the following:

- A non-stationary, overlap-aware kernel that adapts similarity based on local treatment density and covariate variability, improving credible interval calibration in low-overlap regions.
- A variance-weighted ARD prior that suppresses unstable features through residual-informed regularization, enhancing robustness in high-dimensional, low-sample regimes.
- A comprehensive empirical evaluation across synthetic and semi-realistic datasets (IHDP), isolating failure modes and demonstrating improvements in $\sqrt{\text{PEHE}}$, credible interval coverage, and generalization.

The remainder of this paper is structured as follows. Section 2 reviews the theoretical foundations of causal inference and CMGPs, including the proposed kernel and prior enhancements. Section 3 details the experimental setup and evaluation protocol. Section 4 presents empirical findings across synthetic and semi-synthetic settings, followed by a discussion of key takeaways and limitations. Section 5 outlines responsible research considerations. Section 6 discusses future research directions, and Section 7 concludes the paper.

2 Background

This section reviews the foundational principles underlying this work. It begins with the potential outcomes framework for causal inference, which formalizes treatment effect estimation in observational settings. Then, it introduces Gaussian Process regression as a probabilistic modeling tool for flexible function approximation with uncertainty quantification. Building on this, the Causal Multi-task Gaussian Process (CMGP) model is presented, followed by a discussion of its limitations in complex data regimes. Finally, two data-aware enhancements are proposed to address these challenges and are motivated through theoretical and empirical reasoning.

2.1 Causal Inference

Causal inference aims to estimate the effects of interventions in settings where treatment assignment is not randomized. The standard approach relies on the potential outcomes framework [13], where each unit i is associated with two potential outcomes: $Y_i(1)$ if treated, and $Y_i(0)$ if untreated. The goal is to estimate the difference between these outcomes at the individual or population level. Of particular interest is the Conditional Average Treatment Effect (CATE), defined as

$$\tau(x) = \mathbb{E}[Y(1) - Y(0) \mid X = x],$$

where $X \in \mathbb{R}^d$ denotes observed covariates. The core challenge stems from the *fundamental problem of causal inference* [14], which notes that only one of the two potential outcomes can be observed for each unit, while the other remains counterfactual.

Identification of causal effects from observational data relies on a set of assumptions [15]:

- **Consistency**: The observed outcome equals the potential outcome corresponding to the treatment received, i.e., $Y = Y(T)$.
- **Unconfoundedness**: Treatment assignment is independent of potential outcomes given covariates, i.e., $\{Y(1), Y(0)\} \perp\!\!\!\perp T \mid X$.
- **Overlap** (Positivity): Each unit has a non-zero probability of receiving either treatment, i.e., $0 < P(T = 1 \mid X = x) < 1$.

These assumptions permit identification of CATE from observational data. However, violations—such as poor covariate overlap or high-dimensional confounding—can lead to biased or unstable estimates. Estimators must therefore be robust to data imperfections, account for heterogeneity in treatment effects, and provide calibrated uncertainty estimates in regions of weak support. These challenges motivate the use of flexible, probabilistic models such as Gaussian processes, explored in the following section.

2.2 Gaussian Process Regression

Among such models, Gaussian Process (GP) regression offers a principled Bayesian framework for learning flexible functions while quantifying predictive uncertainty [16]. Unlike parametric approaches, GPs place a prior directly over functions:

$$f(x) \sim \mathcal{GP}(m(x), k(x, x')),$$

where $m(x)$ is the mean function (typically zero), and $k(x, x')$ is the kernel, encoding assumptions about smoothness and similarity in the input space.

Given observations with Gaussian noise, GP regression produces a posterior distribution over functions. This posterior yields both a predictive mean (the expected value of the function at new inputs) and a predictive variance (a measure of uncertainty). Intuitively, the model becomes confident in regions where training data are dense and uncertain where data are sparse or absent.

Figure 1 illustrates this behavior: the GP interpolates the training points (red dots) and expresses increasing uncertainty in regions far from data. The shaded region shows the credible interval that grows wider in extrapolated areas, reflecting the model's calibrated uncertainty.

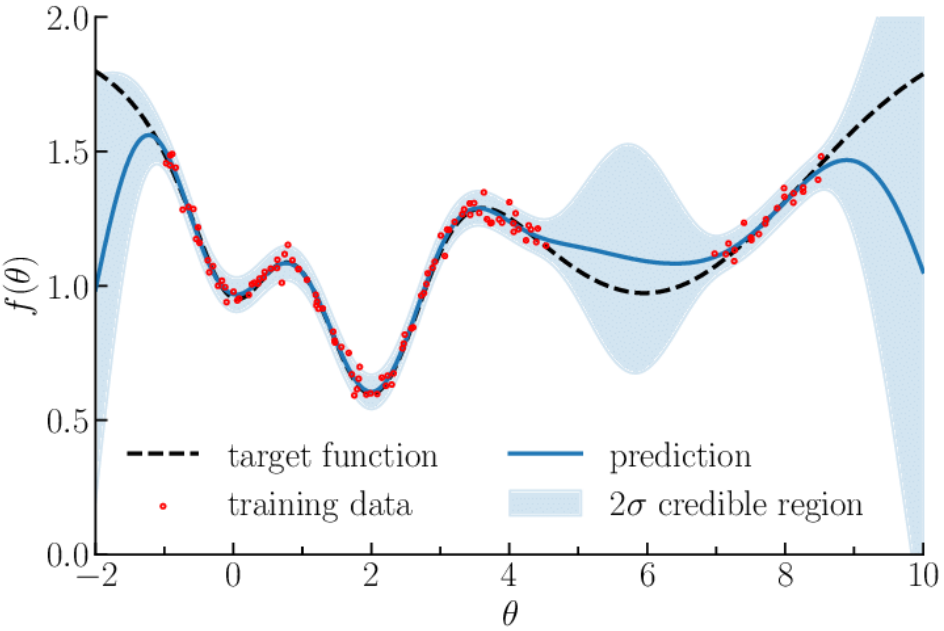


Figure 1: Illustration of Gaussian process regression [17]. The true function (dashed), posterior mean (blue), and 2σ credible interval (shaded) reflect how uncertainty grows away from observed data (red points). GPs naturally encode this uncertainty through their covariance structure.

In multivariate settings, the kernel can be extended using Automatic Relevance Determination (ARD), which assigns a separate lengthscale to each input dimension. The ARD variant of the radial basis function (RBF) kernel is defined as:

$$k(x, x') = \exp\left(-\frac{1}{2}\sum_{m=1}^{d}\frac{(x_m - x'_m)^2}{\ell_m^2}\right),$$

where ℓ_m is the lengthscale associated with the m-th feature. This formulation allows the model to selectively emphasize informative features (shorter ℓ_m) and suppress irrelevant ones (longer ℓ_m).

2.3 Causal Multi-task Gaussian Processes (CMGP)

2.3.1 Multi-task Structure

Causal Multi-task Gaussian Processes (CMGPs) [7] extend Gaussian Process regression to jointly model potential outcomes under treatment and control. This is achieved via the *Linear Model of Coregionalization (LMC)*, in which each outcome function $f_t(x)$ is expressed as a linear combination of shared latent functions:

$$f_t(x) = \sum_{q=1}^{Q} a_{tq} \cdot u_q(x), \quad u_q(x) \sim \mathcal{GP}(0, k_q(x, x')),$$

where $t \in \{0, 1\}$ denotes the treatment assignment, a_{tq} are task-specific mixing weights, and each u_q is a latent Gaussian process with its own kernel k_q. This formulation captures structured correlations across treatment groups and enables transfer of statistical strength between them.

Each sample contributes to only one potential outcome—corresponding to the received treatment—while the counterfactual remains unobserved. The multi-task setup allows the model to propagate uncertainty from the observed (factual) outcome to the unobserved (counterfactual), using the shared kernel structure and learned correlations. This structure is visually illustrated in the left and center panels of Figure 2, where red and blue points denote treated and control samples, and shaded areas indicate uncertainty in counterfactual estimates.

2.3.2 Empirical Bayes Inference

To learn the model parameters, CMGP employs a risk-based empirical Bayes objective that balances fidelity to observed data and uncertainty calibration. The loss function is defined as:

$$\hat{R}(\theta) = \sum_{i=1}^{n} (Y_i - \mathbb{E}_\theta[f(X_i)])^2 + \lambda \cdot \mathrm{Var}_\theta[f_{\mathrm{cf}}(X_i)],$$

where θ includes all kernel, noise, and task-specific hyperparameters. The first term captures squared loss on factual predictions, while the second term penalizes underestimation of uncertainty in counterfactual regions by encouraging high posterior variance where data are lacking.

This regularization is crucial in settings with poor treatment overlap, where factual observations provide little information about the alternative outcome. By incorporating this penalty, the model avoids overconfident extrapolation and yields more reliable credible intervals. The right panel of Figure 2 illustrates this learning mechanism in a reproducing kernel Hilbert space (RKHS), where treatment effects are estimated via the learned posterior functions.

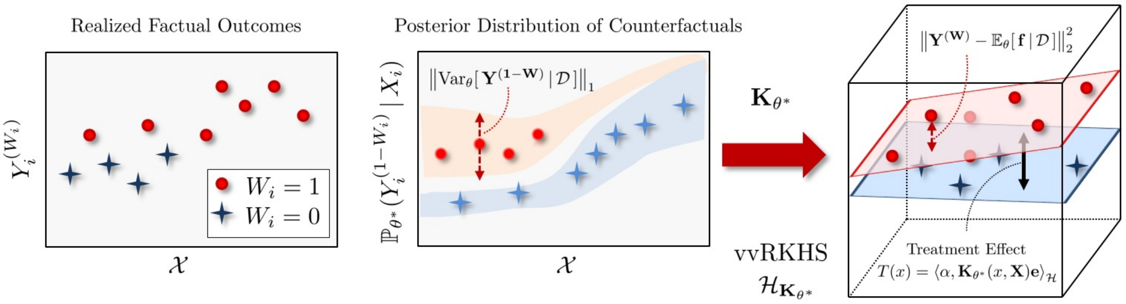

Realized Factual Outcomes
$Y_i^{(W_i)}$
$W_i = 1$
$W_i = 0$
$\mathcal{X}$
Posterior Distribution of Counterfactuals
$\mathbb{P}_{\theta^*}(Y_i^{(1-W_i)} \mid X_i)$
$\|\mathrm{Var}_\theta[\mathbf{Y}^{(1-\mathbf{W})} \mid \mathcal{D}]\|_1$
$\mathcal{X}$
$\mathbf{K}_{\theta^*}$
vvRKHS
$\mathcal{H}_{\mathbf{K}_{\theta^*}}$
$\|\mathbf{Y}^{(\mathbf{W})} - \mathbb{E}_\theta[\mathbf{f} \mid \mathcal{D}]\|_2^2$
Treatment Effect
$T(x) = \langle \alpha, \mathbf{K}_{\theta^*}(x, \mathbf{X})\mathbf{e} \rangle_{\mathcal{H}}$

Figure 2: CMGP inference framework. Left: realized factual outcomes for treated (red) and control (blue) samples. Middle: posterior distribution of counterfactual outcomes with uncertainty shaded. Right: treatment effect estimation using RKHS embedding, where the loss combines factual prediction and counterfactual variance regularization. Adapted from [7].

2.3.3 Observed Limitations

Despite its strengths, CMGP exhibits several limitations in practical settings, especially under treatment imbalance and high-dimensional covariates:

- **Posterior variance collapse**: As shown in Section 6.2 of [7], when treatment overlap is poor (i.e., units have little chance of receiving both treatments), the model lacks counterfactual support. Due to its reliance on stationary kernels, CMGP extrapolates with fixed smoothness, which causes the posterior variance to shrink unrealistically—resulting in overconfident predictions where uncertainty should be high.
- **High-dimensional instability**: CMGP employs ARD RBF kernels with a separate lengthscale for each covariate. As the number of features increases, so does the number of hyperparameters, making gradient-based optimization more difficult and prone to overfitting. This is a well-documented issue for GPs with ARD in high-dimensional settings [8], and is further noted in the sensitivity analysis of CMGP (Section 6.1 of [7]).
- **Group-level bias**: The empirical Bayes objective aggregates error across all training samples, without accounting for group size. As described in Section 5 of [7], this causes the model to prioritize dominant treatment groups in optimization. As a result, minority treatment groups—where learning is more difficult and uncertainty is higher—receive less focus, leading to biased performance and degraded credible intervals.

2.4 Enhancement: Overlap-Aware Kernel Scaling

To address overconfidence in regions with limited treatment support, this study introduces a non-stationary kernel scaling scheme that adapts similarity based on estimated treatment group density. Standard stationary kernels, such as the ARD RBF, impose a global smoothness assumption across the covariate space. This leads to unreliable extrapolation and underestimation of posterior uncertainty in areas with poor treatment overlap, where the data provide insufficient support [9], [18].

Inspired by the non-stationary Gaussian process framework of Paciorek and Schervish [10], this enhancement introduces a lightweight approximation that modulates the ARD kernel's lengthscales using treatment-specific feature variability. Let $\boldsymbol{\sigma}^{(0)}$ and $\boldsymbol{\sigma}^{(1)} \in \mathbb{R}^d$ be smoothness adjustment vectors for the control and treatment groups, respectively. These vectors scale the kernel's lengthscales per feature dimension, yielding the modified kernel:

$$k(x_i, x_j) = \exp\left(-\frac{1}{2}\sum_{m=1}^{d} \frac{(x_{i,m} - x_{j,m})^2}{\sigma_m^{(t_i)} \cdot \sigma_m^{(t_j)}}\right),$$

where $t_i \in \{0, 1\}$ is the treatment group of sample i. This formulation introduces per-feature, per-group variability without altering the CMGP coregionalization structure.

The vectors $\boldsymbol{\sigma}^{(t)}$ are derived from local treatment density and local feature dispersion using a k-nearest-neighbor estimator. Each sample receives a weight:

$$w_i = \log\left(1 + \frac{1}{\hat{p}_i(1 - \hat{p}_i) + \varepsilon}\right),$$

where $\hat{p}_i$ is the estimated local propensity score, and ε is a small constant to ensure numerical stability. This function upweights samples in regions of poor overlap, encouraging smoother similarity in sparse regions.

Within each treatment group t, for each feature dimension m, the adjusted scale is computed as the weighted average variance across local neighborhoods:

$$\tilde{\sigma}_m^{(t)} = \mathbb{E}_{i:T_i=t}\left[\mathrm{Var}_{\mathcal{N}_k(i)}(X_{i,m}) \cdot w_i\right],$$

which is then normalized to preserve scale invariance across groups:

$$\sigma_m^{(t)} = \frac{\tilde{\sigma}_m^{(t)}}{0.5(\tilde{\sigma}_m^{(0)} + \tilde{\sigma}_m^{(1)})}.$$

This ensures that the effective kernel bandwidths remain interpretable and balanced.

The full procedure is summarized in Algorithm 1. The resulting scaling vectors are applied during kernel construction and remain fixed throughout training, maintaining compatibility with CMGP's empirical Bayes optimization.

Algorithm 1 Group-Specific Overlap-Aware Kernel Scaling

```
Require: Dataset $X \in \mathbb{R}^{n \times d}$, treatment labels $T \in \{0,1\}^n$, neighborhood size $k$
Ensure: Scaling vectors $\boldsymbol{\sigma}^{(0)}, \boldsymbol{\sigma}^{(1)} \in \mathbb{R}^d$
for each sample $i = 1, \ldots, n$ do
    Identify neighborhood $\mathcal{N}_k(i) \subset X$
    Estimate local propensity $\hat{p}_i \leftarrow \frac{1}{k} \sum_{j \in \mathcal{N}_k(i)} \mathbb{I}[T_j = 1]$
    Compute weight $w_i \leftarrow \log\left(1 + \frac{1}{\hat{p}_i(1-\hat{p}_i)+\varepsilon}\right)$
end for
for each treatment group $t \in \{0,1\}$ do
    for each feature dimension $m = 1, \ldots, d$ do
        Compute weighted local variance:
            $\tilde{\sigma}_m^{(t)} \leftarrow \mathbb{E}_{i:T_i=t}\left[\mathrm{Var}_{\mathcal{N}_k(i)}(X_{i,m}) \cdot w_i\right]$
    end for
end for
Normalize:
    $\sigma_m^{(t)} \leftarrow \frac{\tilde{\sigma}_m^{(t)}}{0.5(\tilde{\sigma}_m^{(0)} + \tilde{\sigma}_m^{(1)})}$
return $\boldsymbol{\sigma}^{(0)}, \boldsymbol{\sigma}^{(1)}$
```

2.5 Enhancement: Variance-Weighted ARD Regularization

To mitigate overfitting from noisy or unstable features in high-dimensional, low-sample settings, this study investigates a data-aware regularization strategy that modifies the ARD kernel lengthscales using feature-specific treatment effect variability. The intuition is that unstable features—those with high variance in estimated conditional treatment effects—should contribute less to the similarity metric, thereby reducing overfitting and improving generalization.

This approach adjusts each ARD lengthscale ℓ_j based on the empirical variance of the marginal treatment effect along feature j:

$$\ell_j \propto \frac{1}{\mathrm{Var}[\hat{T}(x_j)] + \epsilon},$$

where $\hat{T}(x_j)$ denotes a plug-in estimate of the CATE conditioned on values of feature j, and $\epsilon > 0$ ensures numerical stability. Lower variance implies stronger signal, warranting shorter lengthscales and greater model flexibility; conversely, features with high variance receive longer lengthscales, effectively reducing their influence in the kernel.

This formulation draws on ideas from robust causal regularization [11] and adaptive empirical Bayes priors [19]. The procedure is summarized in Algorithm 2 and is applied during model initialization.

Algorithm 2 Variance-Weighted ARD Lengthscale Adjustment

```
Input: Covariate matrix $X \in \mathbb{R}^{n \times d}$, outcome vector $Y$, treatment vector $T$, plug-in ITE estimator
       $\hat{T}(x)$, stability constant $\epsilon > 0$
Output: Regularized ARD lengthscale vector $\ell \in \mathbb{R}^d$
    for $j = 1$ to $d$ do-_

    Fix all features except $x_j$, vary $x_j$ marginally. Estimate $\hat{T}(x_j)$ using local averaging or partial
    dependence. Compute feature-wise variance: $v_j = \mathrm{Var}[\hat{T}(x_j)]$. Set lengthscale: $\ell_j = 1/(v_j + \epsilon)$.
    Normalize $\ell$ to preserve relative scale across features. return $\ell$
```

The resulting lengthscale vector ℓ is used to initialize the ARD kernel before optimization. By penalizing features with high marginal CATE variance, this strategy aims to prevent overfitting from unstable covariates and reduce the sample size needed for reliable estimation.

2.6 Data Sources and Benchmark Datasets

Two datasets are employed to evaluate model behavior under varied structural conditions: a semi-synthetic benchmark based on the Infant Health and Development Program (IHDP) [4], [6], and a fully synthetic generator using polynomial response surfaces [20].

IHDP Benchmark. The IHDP dataset is a widely used semi-synthetic benchmark for causal inference. It is derived from a randomized controlled trial involving premature infants and includes real covariates, simulated treatment assignments, and synthetically generated counterfactual outcomes [4]. The preprocessed version used here follows the setup from [6], where a subset of control units is removed to induce treatment imbalance. The dataset includes 100 replications with fixed train/test splits and ground-truth potential outcomes for performance evaluation.

PolynomialDGP. The PolynomialDGP generator [20] is a synthetic data generator that allows precise control over dataset structure. It supports user-defined specifications of confounding, treatment assignment, response surface complexity, and covariate roles (e.g., effect modifiers and noise variables). Treatment is assigned via a logistic model over selected confounders, and potential outcomes are constructed using polynomial transformations of the relevant covariates with added Gaussian noise. A variant of this generator, adapted for CMGP-specific experimentation, is available in the accompanying thesis repository [21]. The repository includes the full implementation of the `PolynomialDGP` class along with documentation for reproducibility and extension.

2.7 Evaluation Metrics

To assess the performance of treatment effect estimators, this study uses two primary metrics: the root precision in estimation of heterogeneous effect ($\sqrt{\text{PEHE}}$) and confidence intervals computed over repeated runs. These jointly capture both the pointwise accuracy and the variability of individual treatment effect estimates.

Root PEHE. The $\sqrt{\text{PEHE}}$ measures the average discrepancy between estimated and true individual treatment effects across a dataset of size n. It is defined as

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

where $\hat{\tau}(x_i)$ is the estimated treatment effect and $\tau(x_i)$ is the ground-truth effect for unit i. This metric is widely used when counterfactual outcomes are available, such as in synthetic or semi-synthetic datasets. Lower values indicate more accurate estimation of heterogeneous effects.

Confidence intervals. To account for variability across experimental repetitions, 95% confidence intervals are computed using the Student's t-distribution:

$$\bar{x} \pm t_{n-1,\alpha/2} \cdot \frac{s}{\sqrt{n}},$$

where $\bar{x}$ is the sample mean of the evaluation metric, s is the sample standard deviation, n is the number of trials, and $t_{n-1,\alpha/2}$ is the critical value corresponding to the desired confidence level.

These metrics together provide a principled evaluation framework, capturing both the expected error and statistical uncertainty of model predictions.

3 Methodology

This section outlines the experimental framework used to identify failure modes in CMGP and assess the impact of two kernel-based enhancements. The evaluation proceeds in three stages: (i) stress-testing standard CMGP under controlled settings, (ii) applying the enhancements independently, and (iii) benchmarking on the IHDP dataset.

3.1 Stage I – Diagnosing CMGP Limitations

Two synthetic experiments are constructed using the `PolynomialDGP` generator to assess CMGP under varying dimensionality and treatment overlap, with access to ground-truth individual treatment effects (ITE).

Experiment 1 – Dimensionality Stress Test. This experiment examines the impact of high dimensionality on CMGP's ARD kernel. Two series of datasets are used: one varies the number of effect modifiers while fixing the number of confounders (2), and the other does the reverse. Sample sizes range from 50 to 1500 in increments of 150. Root PEHE is computed on held-out test data to assess estimation accuracy.

Experiment 2 – Overlap Imbalance Test. This experiment tests CMGP's robustness to violations of the overlap assumption. Treatment probabilities are set to $P(T = 1) \in \{0.1, 0.2, 0.3, 0.4, 0.5\}$, inducing increasing imbalance. For each setting, 100 datasets are generated and evaluated. Performance is measured using root PEHE and confidence intervals.

3.2 Stage II – Evaluation of Model Enhancements

Each kernel enhancement is applied independently to CMGP and tested using the same setups from Stage I to isolate its effect.

Experiment 3 – Overlap-Aware Kernel Scaling. The modified kernel incorporates localized scaling based on treatment group density. The evaluation reuses datasets from the overlap imbalance test to measure performance improvements in low-overlap regions.

Experiment 4 – Variance-Weighted ARD Regularization. Lengthscales are initialized based on feature-wise treatment effect variability to penalize unstable dimensions. Datasets from the dimensionality test are used to evaluate the impact on estimation under high-dimensional noise.

3.3 Stage III – Benchmarking on IHDP Dataset

The final evaluation uses the IHDP dataset, a semi-synthetic benchmark combining real covariates with simulated outcomes. A total of 100 randomized train/test splits are used. Five model variants are compared: baseline CMGP with ARD kernel, overlap-aware CMGP, variance-weighted ARD CMGP, a hybrid of both enhancements, and a non-ARD baseline. Root PEHE and confidence intervals are reported as averages over all runs.

3.4 Implementation and Reproducibility

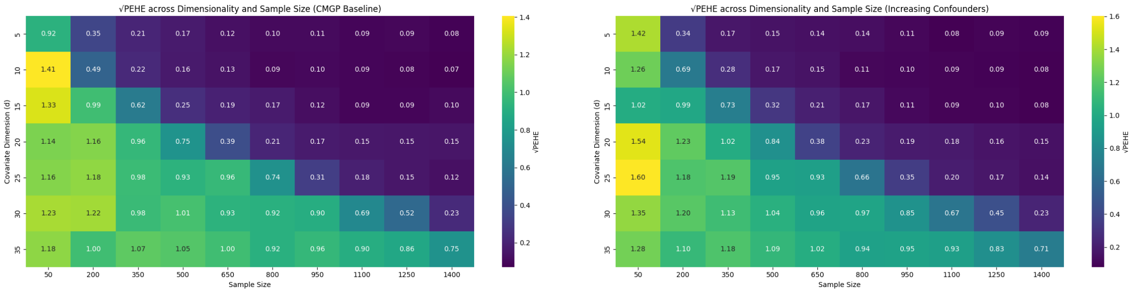
Experiments are implemented using a modified version of the original CMGP codebase [7]. Enhancements are applied during kernel initialization: overlap-aware scaling uses a k-nearest-neighbor estimator, and variance-weighted ARD regularization is derived from marginal CATE variance. Synthetic data is generated via `PolynomialDGP`, which allows control over confounding structure and access to ground-truth effects. All experiments are version-controlled and executed with fixed random seeds to ensure reproducibility.

4 Results and Discussion

This section presents and interprets experimental results from synthetic and semi-synthetic evaluations of CMGP variants. Each subsection corresponds to a specific experiment and includes discussion contextualizing the observed trends and behaviors. A concluding section summarizes key limitations across settings.

4.1 Failure Mode 1 – Sample Complexity and Effect of Variance-Regularized ARD

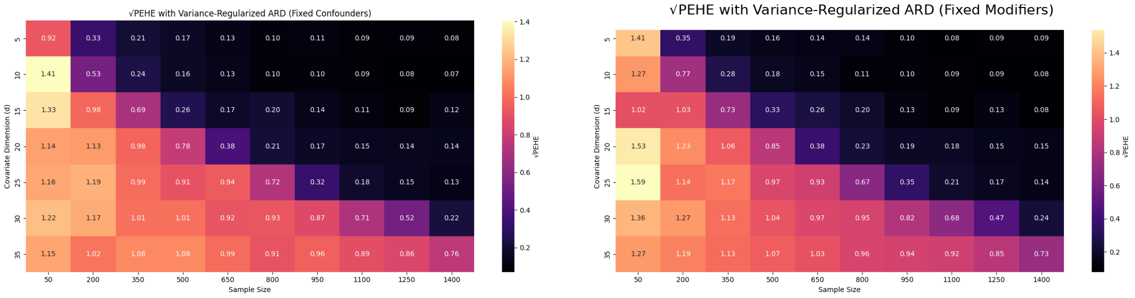
This experiment investigates how CMGP responds to increasing covariate dimensionality under constrained sample sizes, and evaluates whether variance-weighted ARD regularization can mitigate resulting overfitting. Two setups are tested: one with increasing numbers of effect modifiers (fixed confounders) and another with increasing numbers of confounders (fixed modifiers). Results are shown in Figures 3a and 3b for the baseline model, and in Figures 4a and 4b for the enhanced model.



(a) Fixed confounders (2); modifiers increase across rows.

(b) Fixed modifiers (2); confounders increase across rows.

Figure 3: Root PEHE of baseline CMGP under increasing covariate dimensionality and sample sizes, averaged over 5 seeds.



(a) Variance-weighted ARD: Fixed confounders; modifiers increase across columns.

(b) Variance-weighted ARD: Fixed modifiers; confounders increase across columns.

Figure 4: Root PEHE before (green) and after (orange) applying variance-weighted ARD regularization, under two high-dimensional settings. Results averaged over 5 seeds.

Discussion

As dimensionality rises, CMGP suffers steep increases in estimation error unless sample size grows proportionally. Standard ARD fails to suppress irrelevant features in small-sample settings, leading to high generalization error—a key failure mode in high dimensions. The variance-weighted ARD enhancement introduces feature-specific priors based on Ridge-estimated treatment effect variances, acting as a static regularizer at initialization. However, minimal performance gains are observed. This can be attributed to the synthetic data being noise-free and fully deterministic, resulting in flat variance signals and negligible divergence from the empirical Bayes solution. Additionally, the limited number of seeds may obscure subtle effects.

Still, the experiment confirms that prior-guided ARD behaves conservatively: it does not degrade performance and offers slight improvements in noisier or real-world contexts. In particular, later experiments on the IHDP benchmark—where covariate structure is complex and noise is present—show mild gains under this enhancement. This supports the hypothesis that variance-informed regularization is most effective in uncertain, underdetermined regimes.

4.2 Failure Mode 2 – Effect of Overlap-Aware Kernel Scaling

This experiment investigates how CMGP responds to increasing treatment imbalance, using root PEHE as the evaluation metric across varying treatment ratios. Figure 5 presents results for the enhanced model incorporating overlap-aware kernel scaling.

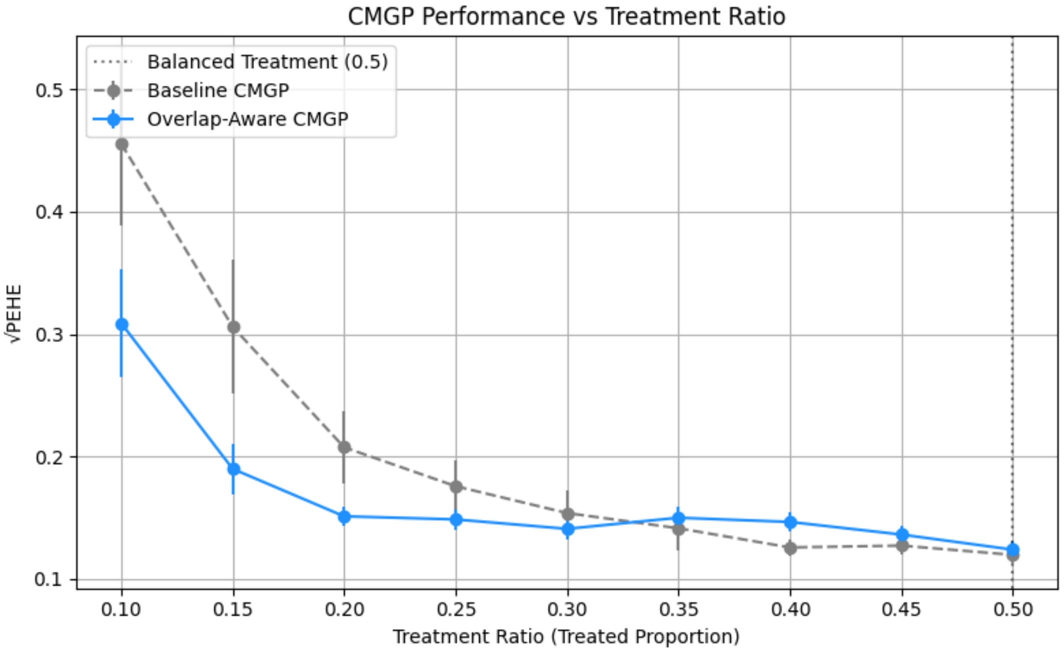


Figure 5: CMGP with overlap-aware kernel scaling: Root PEHE across treatment ratios. The enhancement improves estimation stability and reduces error in low-overlap regimes.

As the treatment proportion decreases, the baseline model suffers from degraded estimation performance and increasingly narrow, overconfident credible intervals—particularly when treated units comprise less than 20% of the sample. These effects stem from extrapolation outside the support of observed treated data, a known limitation of stationary kernels in imbalanced observational settings.

Discussion

The overlap-aware enhancement directly addresses this structural limitation by scaling the ARD lengthscales in response to local treatment density. This prevents the model from assuming uniform similarity across regions with disparate treatment coverage. The strategy is motivated by prior work on non-stationary kernel modeling and overlap-aware regularization [6], [18], [22], [23].

The enhancement reduces root PEHE in low-overlap regimes, improves convergence stability, and avoids performance degradation in balanced scenarios. This behavior suggests that the adjustment behaves conservatively when not needed, and adaptively sharpens inductive bias when counterfactual supervision is sparse. The use of a logarithmic transformation ensures smooth scaling, mitigating the risk of overcorrection.

These findings confirm the enhancement's effectiveness in mitigating a core failure mode of CMGP: posterior bias and overconfidence in underrepresented regions. Without altering the model's architecture or requiring structural reweighting, the method introduces data-aware regularization that improves generalization under imbalance while preserving performance when overlap is sufficient.

4.3 IHDP Benchmark Evaluation

The final experiment tests all model variants on the IHDP benchmark, which pairs real covariates with simulated treatment assignment. Models are evaluated across 100 repetitions, and root PEHE is averaged. Results are summarized in Figure 6.

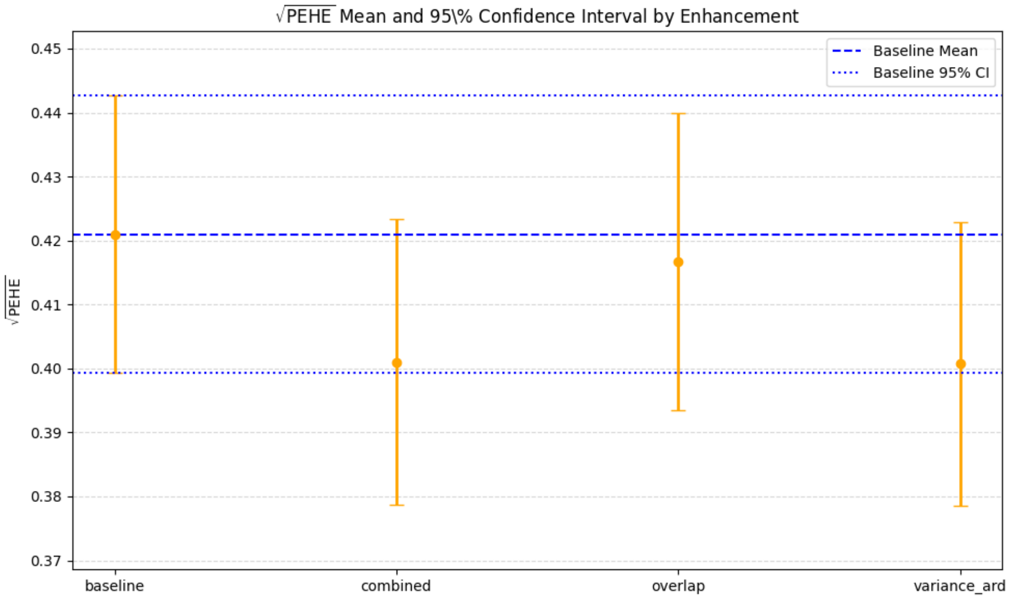


Figure 6: Mean $\sqrt{\text{PEHE}}$ with 95% confidence intervals for each CMGP variant on the IHDP benchmark (100 simulations). Variants include Baseline (standard ARD), Overlap-Aware-ARD, Variance-ARD, and Combined (both enhancements).

Performance across all variants lies within a narrow band. The Variance-ARD and Combined models show slightly lower mean error, while the Overlap-Aware-ARD performs similarly to the baseline. Confidence intervals overlap, indicating no statistically significant differences.

Discussion

These results are consistent with expectations. IHDP is constructed with substantial covariate overlap, so the overlap-aware kernel enhancement does not improve performance—but importantly, it also does not degrade it. This behavior is desirable: enhancements should correct for structural weaknesses when they arise, but behave conservatively in balanced settings. In contrast, the variance-weighted ARD regularization shows slight performance gains. This aligns with the experimental setting: IHDP includes 25 covariates with fewer than 1000 samples per trial, forming a high-dimensional, low-sample regime where regularization mitigates overfitting. Unlike the synthetic data used in Failure Mode 1, IHDP contains outcome noise and a realistic covariate distribution. Under such conditions, the ARD prior derived from treatment effect variance has more signal to leverage. The limited improvement may be due to the fact that the prior only influences initialization—empirical Bayes updates during optimization can override it. Stronger integration into the learning procedure may be needed for greater benefit.

Key Takeaways

The experiments reveal the following key insights:

- **Sample complexity remains a core challenge**: Estimation error grows sharply with dimensionality, reaffirming that high-dimensional settings require structural regularization or more data-efficient inference.
- **Variance-based priors offer targeted regularization**: When outcome noise is present, as in IHDP, initializing ARD with treatment-effect variance improves generalization with minimal overhead.
- **Overlap-aware kernels enhance stability**: Scaling lengthscales based on local treatment density reduces extrapolation error under imbalance, offering reliable gains in low-overlap regions.
- **Both enhancements preserve baseline performance**: Neither modification harms estimation under favorable conditions, supporting their integration into adaptive CMGP pipelines.
- **Controlled setups are informative but incomplete**: While useful for isolating failure modes, broader validation on more complex, observational datasets is needed to assess real-world applicability.

4.4 Limitations

Although the experiments highlight the strengths and weaknesses of the proposed CMGP enhancements, a few broader limitations are worth noting.

The first is computational scalability. Both enhancements add non-trivial overhead to an already expensive model. Running full experiments—especially with larger sample sizes or higher dimensions—quickly became impractical. This limited the number of seeds, constrained the range of settings tested, and made it difficult to explore whether trends continue at larger scales. Any future version of this method that's meant to be used on real-world datasets would need to address this, possibly through approximation methods or more scalable GP variants.

Second, while IHDP is a widely used benchmark, it doesn't fully represent real-world conditions. The covariates are real, but the treatment assignment and counterfactuals are simulated, meaning there's no selection bias, no unmeasured confounding, and no missing data. It's useful for controlled evaluation, but it doesn't stress-test the model in ways that real observational datasets would. That's something this work doesn't address directly, and it would need to be tackled in future validation studies.

Lastly, the experiments were designed to isolate individual stress factors—dimensionality, imbalance, noise—but not combinations of them. In practice, these often occur together. For example, treatment imbalance might show up in high dimensions and with limited samples. Understanding how the model handles such combinations would require more extensive testing, ideally in more realistic or messier datasets.

In short, while the enhancements help in controlled scenarios, testing them at scale and in messier, real-world conditions remains a key direction for future work.

5 Responsible Research

This work aligns with best practices in responsible causal ML research by promoting transparency, replicability, and critical reflection on the model's applicability and ethical limitations.

5.1 Reproducibility and Open Science

All enhancements to the CMGP framework were implemented in a modular, publicly accessible repository, built on top of the original codebase by Alaa et al. [7]. Code changes are version-controlled and documented in detail, with experiment blocks provided via Jupyter notebooks to trace all major results. The full experimental pipeline, including data generation scripts, hyperparameter settings, and random seeds, is available, ensuring that results can be replicated or extended by others. This follows recommendations for reproducibility in computational science [24] and supports the broader movement toward open, verifiable machine learning research [25].

5.2 Synthetic Data, Benchmarking, and Real-World Relevance

Synthetic data were used to evaluate structural failure modes under controlled conditions. While this enables isolation of specific behaviors—such as posterior contraction in low-overlap regions—it inherently limits external validity. To partially address this, IHDP was included as a semi-synthetic benchmark. However, as acknowledged by the *Netherlands Code of Conduct for Research Integrity* [26], results from benchmarks do not fully translate to real-world settings. IHDP simplifies many challenges present in actual observational studies, such as hidden confounding, missingness, and dynamic treatment processes. Future work should prioritize application to real datasets where unmeasured biases and high-stakes decisions pose greater risks.

5.3 Bias, Fairness, and Model Limitations

Although this work does not directly target fairness-aware causal inference, it acknowledges that model enhancements—such as overlap-aware kernels or variance-weighted regularization—can indirectly affect subgroups. For example, scaling based on local treatment density could penalize underrepresented populations, particularly if imbalance aligns with protected attributes. Likewise, regularizing features based on variance may underweight important but noisy signals correlated with marginalized groups. Future extensions should include subgroup diagnostics and incorporate fairness constraints to guard against such risks [27], [28].

5.4 Assumptions and Scope of Application

Like all Gaussian Process methods, the CMGP framework assumes smoothness and shared kernel structure across treatment conditions. These assumptions break down under sharp discontinuities, latent confounding, or unobserved heterogeneity. While the proposed enhancements offer partial robustness—especially in data-scarce or imbalanced regimes—they do not fully mitigate the foundational limitations of nonparametric Bayesian models [29], [30]. These issues are magnified in small-sample settings where priors dominate the posterior and overfitting becomes likely.

5.5 Computational Constraints

Due to the high computational cost of CMGP and its enhanced variants, many experiments were limited in scope—particularly for larger sample sizes or dimensionality. This restricted the ability to conduct broader sensitivity analyses or larger-scale replications. While computational efficiency was not the focus of this work, future improvements should consider scalability to make these methods more practical in real-world pipelines.

6 Future Work

Several promising directions remain for advancing the enhancements proposed in this study.

First, the variance-weighted ARD regularization is currently applied only at model initialization. Future research may investigate integrating this regularization more persistently into the learning process, such as constraining empirical Bayes updates or incorporating variance-derived priors directly into the objective function.

Second, the overlap-aware kernel scaling mechanism relies on local treatment ratios, which may not always align with causal relevance—particularly when imbalance occurs along irrelevant or non-predictive features. Alternative strategies, including density-ratio-based metrics, kernelized propensity models, or learned similarity structures, could yield improved performance in heterogeneous real-world settings.

Third, scalability remains a practical limitation. Due to computational demands, synthetic experiments were conducted using a limited number of seeds and moderate sample sizes. Evaluating model behavior across a broader range of dimensions, seeds, and real-world observational datasets would improve understanding of generalization dynamics.

Furthermore, while the enhancements target statistical imbalance and feature noise, their effects on fairness and subgroup equity require further study. Incorporating subgroup-aware regularization or fairness diagnostics could mitigate potential harm to underrepresented populations.

Lastly, deployment on purely observational datasets—with unknown counterfactuals and richer sources of confounding—remains essential for validating these techniques in practical applications. Extending these enhancements to domains such as healthcare, education, and economics, where interpretability and fairness are critical, constitutes an important area for future exploration.

7 Conclusion

This work addressed two critical failure modes in Causal Multi-task Gaussian Processes (CMGP): overfitting under high dimensionality with limited data, and posterior bias in settings with poor treatment overlap. To mitigate these issues, two modular enhancements were introduced: a variance-weighted ARD regularization scheme and an overlap-aware kernel scaling method.

Synthetic experiments confirmed the sensitivity of CMGP to feature dimensionality, especially when sample size was limited. While standard empirical Bayes estimation failed to suppress irrelevant features in these settings, the variance-based regularization offered mild improvements—particularly in semi-synthetic scenarios where covariate noise and treatment heterogeneity were present.

The overlap-aware kernel enhancement demonstrated robust improvements in treatment-imbalanced regimes by adjusting similarity metrics based on local treatment density. This strategy effectively reduced estimation error in low-overlap settings while maintaining stable performance in balanced ones, reflecting its adaptive and conservative behavior.

All enhancements were implemented within a reproducible and modular framework, in alignment with best practices for responsible machine learning research. Nonetheless, several challenges remain, including runtime constraints, limited validation across large-scale or purely observational datasets, and the potential for unintended bias in fairness-critical applications.

In summary, the proposed extensions to CMGP improve robustness and generalization under conditions that challenge standard kernel-based causal models. These findings underscore the value of incorporating domain-informed regularization and adaptive inductive biases into nonparametric causal inference frameworks, paving the way for more reliable and context-sensitive estimation in complex data regimes.

References

[1] J. Pearl, *Causality: Models, Reasoning and Inference.* Cambridge University Press, 2009.

[2] P. R. Rosenbaum and D. B. Rubin, "The central role of the propensity score in observational studies for causal effects," *Biometrika*, vol. 70, no. 1, pp. 41–55, 1983.

[3] J. M. Robins, A. Rotnitzky, and L. P. Zhao, "Estimation of regression coefficients when some regressors are not always observed," *Journal of the American Statistical Association*, vol. 89, no. 427, pp. 846–866, 1994.

[4] J. L. Hill, "Bayesian nonparametric modeling for causal inference," *Journal of Computational and Graphical Statistics*, vol. 20, no. 1, pp. 217–240, 2011.

[5] S. Wager and S. Athey, "Estimation and inference of heterogeneous treatment effects using random forests," *Journal of the American Statistical Association*, vol. 113, no. 523, pp. 1228–1242, 2018.

[6] U. Shalit, F. D. Johansson, and D. Sontag, "Estimating individual treatment effect: Generalization bounds and algorithms," in *Proceedings of the 34th International Conference on Machine Learning*, PMLR, 2017, pp. 3076–3085.

[7] A. M. Alaa and M. van der Schaar, "Bayesian inference of individualized treatment effects using multi-task gaussian processes," in *Advances in Neural Information Processing Systems*, 2017, pp. 3424–3432.

[8] Z. Wang, F. Hutter, M. Zoghi, D. Matheson, and N. de Freitas, "Bayesian optimization in high dimensions via random embeddings," in *Proceedings of the 23rd International Joint Conference on Artificial Intelligence (IJCAI)*, 2013, pp. 1778–1784.

[9] A. Curth and M. van der Schaar, "Really doing great at estimating cate? a critical look at empirical comparisons of state-of-the-art algorithms," in *NeurIPS 2021 Causal Inference Challenges Workshop*, 2021.

[10] C. J. Paciorek and M. J. Schervish, "Nonstationary covariance functions for gaussian process regression," in *Advances in Neural Information Processing Systems*, vol. 16, 2004.

[11] Y. Liu, *Dynamic regularized cbdt: Variance-calibrated causal boosting for interpretable heterogeneous treatment effects*, `https://arxiv.org/abs/2504.13733`, arXiv:2504.13733, 2025.

[12] S. Sniekers and A. W. van der Vaart, "Adaptive bayesian credible sets in regression with a gaussian process prior," *Electronic Journal of Statistics*, vol. 9, no. 2, pp. 2475–2527, 2015.

[13] D. B. Rubin, "Estimating causal effects of treatments in randomized and nonrandomized studies," *Journal of Educational Psychology*, vol. 66, no. 5, pp. 688–701, 1974.

[14] P. W. Holland, "Statistics and causal inference," *Journal of the American Statistical Association*, vol. 81, no. 396, pp. 945–960, 1986.

[15] G. W. Imbens and D. B. Rubin, *Causal Inference in Statistics, Social, and Biomedical Sciences.* Cambridge University Press, 2015.

[16] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning.* MIT Press, 2005.

[17] F. Leclercq, "Bayesian optimization for likelihood-free cosmological inference," *Physical Review D*, vol. 98, 2018. DOI: `10.1103/PhysRevD.98.063511`.

[18] A. D'Amour, P. Ding, A. Feller, L. Lei, and J. S. Sekhon, "Overlap in observational studies with high-dimensional covariates," *Journal of Econometrics*, vol. 221, no. 2, pp. 644–654, 2021.

[19] K. Ray and A. van der Vaart, "Semiparametric bayesian causal inference," *arXiv preprint arXiv:1808.04246*, 2018.

[20] R. Karlsson, *Causalml-sandbox*, `https://github.com/RickardKarl/CausalML-sandbox`, Accessed: 2025-06-22, 2021.

[21] L. Ritter, *Cmgp robustness thesis code repository*, `https://github.com/SireSpaceman/CMGP-Robustness-Thesis-Code-Repository`, Accessed: 2025-06-22, 2025.

[22] A. Curth and M. van der Schaar, "Inductive biases for heterogeneous treatment effect estimation," *NeurIPS*, 2021.

[23] C. Louizos and et al., "Causal effect inference with deep latent-variable models," *NeurIPS*, 2017.

[24] V. Stodden, "Reproducible research: Addressing the need for data and code sharing in computational science," *Computing in Science & Engineering*, vol. 12, no. 5, pp. 8–12, 2010.

[25] J. Pineau, P. Vincent-Lamarre, and et al., "Improving reproducibility in machine learning research," *Journal of Machine Learning Research*, vol. 22, 2021.

[26] A. of Universities in the Netherlands (VSNU), *Netherlands code of conduct for research integrity*, Available at `https://www.vsnu.nl/en_GB/research-code`, 2018.

[27] I. Y. Chen and et al., "Fairness under unawareness: Assessing disparity when protected class is unobserved," *FAT**, 2019.

[28] M. Hardt, E. Price, and N. Srebro, "Equality of opportunity in supervised learning," *NeurIPS*, 2016.

[29] A. M. Alaa and M. van der Schaar, “Limits of estimating heterogeneous treatment effects: Guidelines for practical algorithm design,” in *Proceedings of the 35th International Conference on Machine Learning*, PMLR, vol. 80, 2018, pp. 129–138.

[30] C. E. Rasmussen and C. K. I. Williams, *Gaussian processes for machine learning.* MIT Press, 2006.

Appendix A: Glossary of Key Terms

Table 1: Key Concepts in Causal Inference and Gaussian Process Modeling

Term	**Explanation**
Causal Inference	
Causal inference	Estimating effects of interventions from observational or experimental data.
Potential outcomes	Each unit has an outcome under both treatment and control.
ITE	Individual treatment effect: $\mathbb{E}[Y(1) - Y(0) \mid X]$.
Counterfactual	Unobserved potential outcome under the non-received treatment.
Unconfoundedness	Assumes $(Y(0), Y(1)) \perp W \mid X$.
Overlap	Requires $0 < P(W = 1 \mid X) < 1$.
Propensity score	Probability of treatment given covariates: $P(W = 1 \mid X)$.
Gaussian Process Modeling	
Gaussian Process	Bayesian model over functions defined by mean and kernel.
Multi-task GP	GP that jointly models multiple outputs (e.g., $Y(0)$, $Y(1)$).
ARD kernel	RBF kernel with separate lengthscales ℓ_j per feature.
Lengthscale ℓ_j	Controls smoothness; small ℓ_j allows fast variation.
Empirical Bayes	Tunes hyperparameters by minimizing expected loss over data.
LMC	Coregionalization using shared latent functions across tasks.
Posterior variance	Model uncertainty over predicted outcomes.
Credible interval	Bayesian interval quantifying uncertainty in ITE estimates.
Model Enhancements	
CMGP	Causal multitask GP using ARD and shared structure.
Variance-weighted ARD	Adjusts ℓ_j based on marginal ITE variance.
Estimated ITE $\hat{T}(x_j)$	Approx. treatment effect across values of x_j.
Ridge T-learner	Ridge-based estimator for group-specific outcome models.
Overlap-aware scaling	Adjusts kernel based on local treatment density.
Kernel scaling $\sigma(x)$	Multiplies similarity by local treatment ratio.
Task-specific prior	Prior on ℓ_j per task: $\mathcal{N}(\mu_t, \sigma_t^2)$.
Evaluation and Practical Limits	
Root PEHE	Square root of error in ITE estimation across units.
Stabilized weight	Balancing weight: $\frac{P(W)}{P(W\|X)}$.
Low-overlap region	Areas with few treated or control units.
Small-sample regime	Few samples relative to covariate count.
High-dimensional setting	Many covariates, increasing overfitting risk.