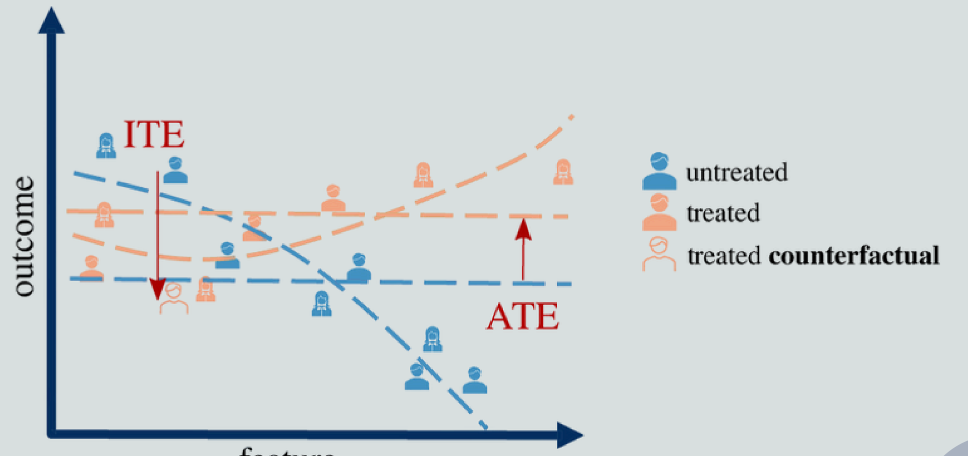
ROBUST CAUSAL INFERENCE WITH MULTI-TASK GAUSSIAN PROCESSES

ENHANCING GENERALIZATION AND CALIBRATION THROUGH DATA-AWARE KERNEL AND PRIOR DESIGN

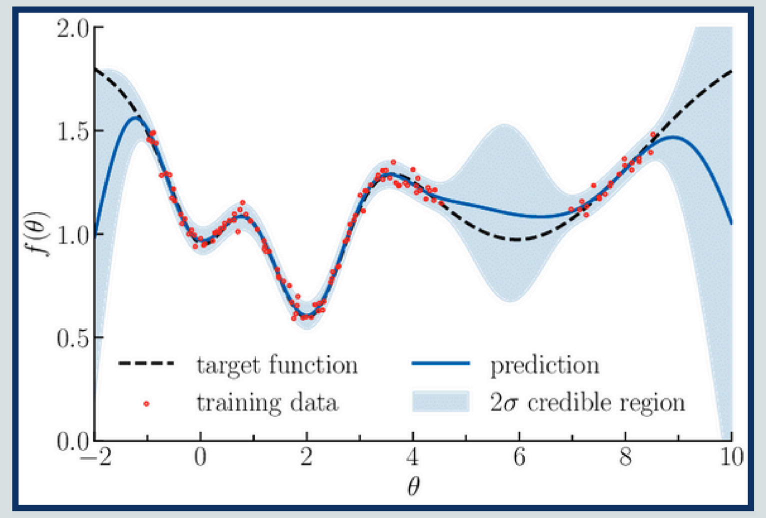
1. INTRODUCTION

Causal inference asks: *"What would have happened under a different treatment?"*—a critical question in healthcare, policy, and economics. It aims to estimate effects at the **individual treatment effect** (**ITE**) or **conditional on covariates effect** (**CATE**) to support informed decision-making [1].



Causal machine learning for healthcare and precision medicine - Scientific Figure on ResearchGate. Available from: https://www.researchgate.net/figure/We-illustrate-the-difference-between-individualized-and-average-treatment-effect-ITE_fig3_362454203 [accessed 23 Jun 2025]

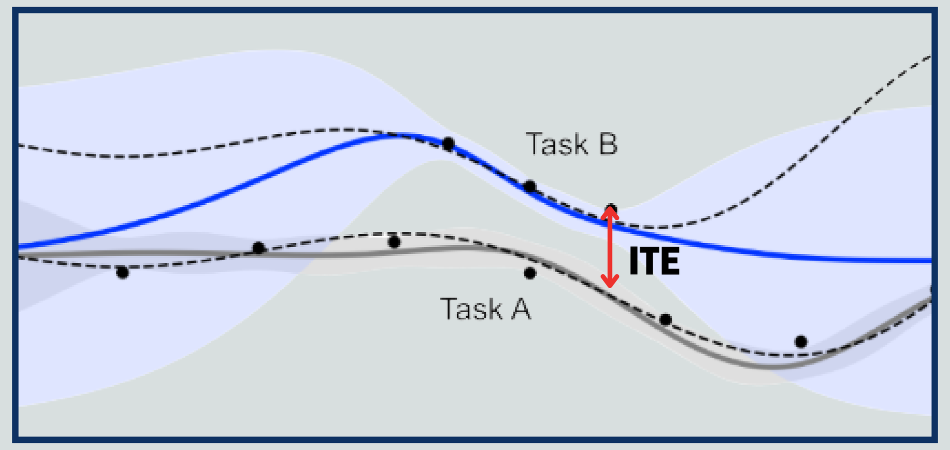
Parametric models like **CFRNet** or **TARNet** require manual architecture tuning and retraining for each dataset, and often lack reliable uncertainty estimates [2].



F. Leclercq, "Bayesian optimization for likelihood-free cosmological inference," Physical Review D, vol. 98, 2018. doi: 10.1103/PhysRevD.98.063511.

In contrast, non-parametric models like **Gaussian Processes** (**GPs**) are flexible models predicting outcomes by treating functions as distributions, allowing them to estimate both the result as well as its uncertainty through **Confidence Intervals** (**CI**)

Causal Multi-task Gaussian Processes (**CMGPs**)extend **GPs** to model treatment and control outcomes jointly. They identify key features (via **ARD**) and self-tune parameters (via **empirical Bayes**) to estimate individual treatment effects with uncertainty [3].



https://honegumi.readthedocs.io/en/latest/curriculum/concepts/multitask/multitask.html

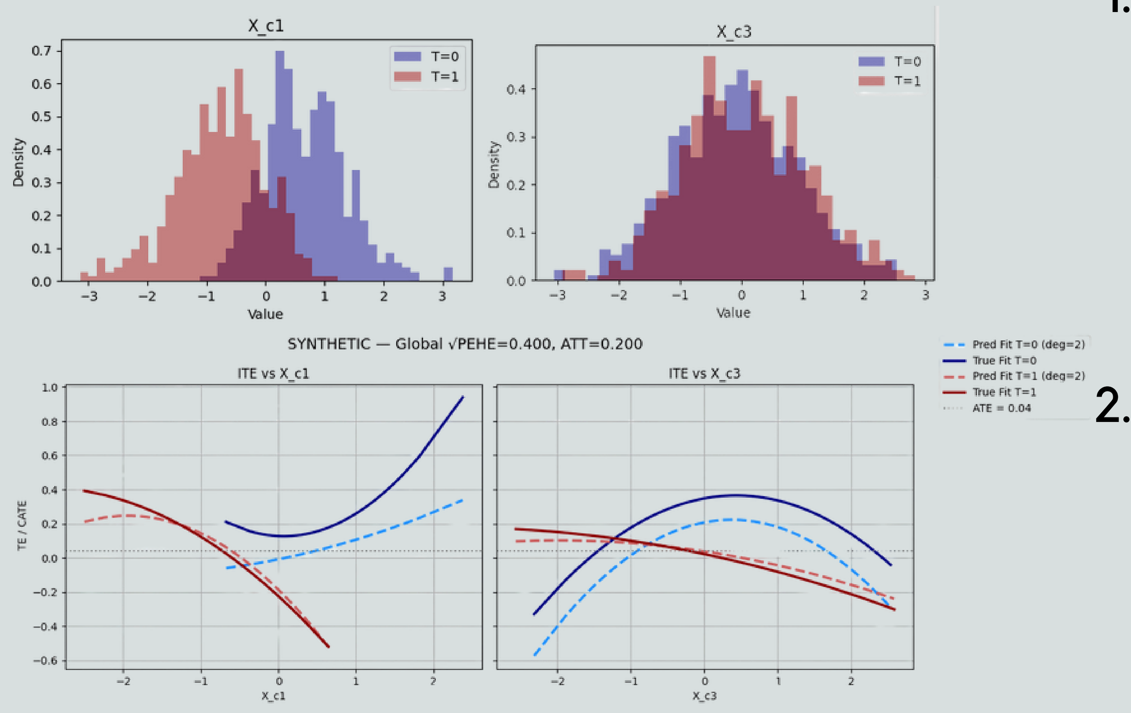
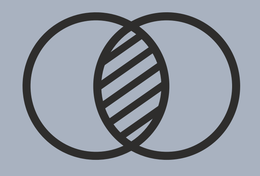


Figure: Effect of Treatment Overlap on CATE Accuracy

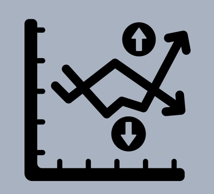
Left: Low overlap leads to inaccurate and overconfident CATE estimates. **Right:** High overlap enables accurate and well-calibrated predictions.

OVERLAP-AWARE KERNEL SCALING



SCALES KERNEL SMOOTHNESS BASED ON LOCAL TREATMENT OVERLAP, ESTIMATED USING K-NEAREST NEIGHBORS, TO REDUCE OVERCONFIDENCE IN REGIONS WITH SPARSE DATA. [5]

VARIANCE-WEIGHTED ARD REGULARIZATION



SCALES FEATURE LENGTHSCALES BASED ON MARGINAL TREATMENT EFFECT VARIANCE, ESTIMATED VIA A PLUG-IN RIDGE T-LEARNER, TO DOWN-WEIGHT UNSTABLE OR NOISY FEATURES. [6]

2. PROBLEM DESCRIPTION

CMGPs face two key challenges in real-world use:

1. **Overfitting in high dimensions:** CMGP employs ARD RBF kernels with a separate length-scale for each covariate. As the number of features increases, so does the number of hyperparameters, making optimization more difficult [4].
2. **Overconfidence under imbalance:** When treatment overlap is poor, CMGP lacks counterfactual support and relies on fixed smoothness assumptions. This causes uncertainty to shrink unrealistically, leading to overconfident predictions in sparse regions[3].

3. RESEARCH QUESTION

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?

- Overlap-aware kernel scaling improves uncertainty calibration and credible interval coverage in imbalanced regions, where traditional kernels are prone to overconfidence.
- Variance-informed ARD regularization reduces overfitting by down-weighting unstable features, aligning with recent evidence that such regularization enhances generalization in causal models.

4. METHODOLOGY

STAGE I: DIAGNOSING CMGP LIMITATIONS

Evaluated standard **CMGP** on synthetic data to explore its behavior under increasing dimensionality and treatment imbalance.

STAGE II: EVALUATING ENHANCEMENTS

Each proposed enhancement was applied independently and tested under the same synthetic dataset and experiment conditions.

STAGE III: BENCHMARKING ON IHDP

All **CMGP** variants, including individual and combined enhancements, were benchmarked on IHDP to assess performance retention or improvement.

5. EXPERIMENTAL SETUP

DATASETS

- **Synthetic** (**PolynomialDGP**): Simulated data with tunable overlap, dimensionality, and treatment effects
- **IHDP:** Widely used semi-synthetic benchmark with real covariates and simulated outcomes (100 splits)

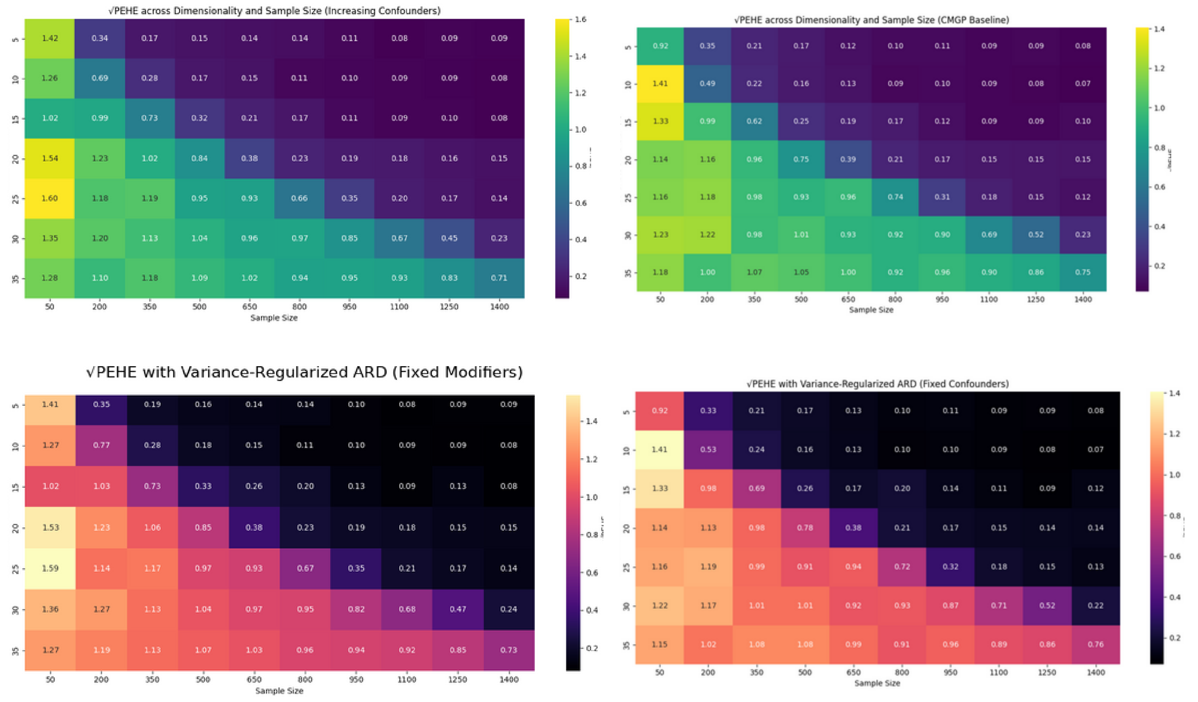
METRICS

- **Root PEHE:** Measures average error in individual treatment effect predictions.
- **Credible Intervals:** Proportion of true treatment effects that fall within the model's 95% credible intervals.
- **Run-to-Run Variability:** Standard deviation across seeds or splits; reflects robustness.

6. RESULTS

1. Failure mode 1: Sample Complexity and Effect of Variance-Regularized ARD

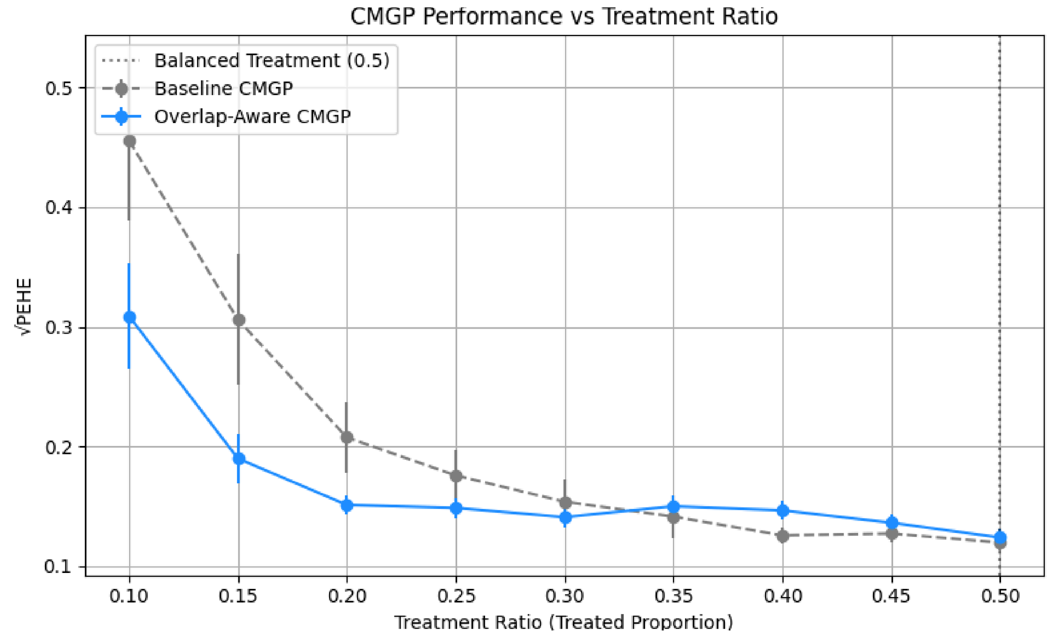
Baseline CMGP is compared with the variance-weighted ARD variant under increasing covariate dimensionality (5 to 30). Sample sizes range from 50 to 1400. Experiments are run separately under fixed effect modifiers and fixed confounders. Each configuration is averaged over 5 random seeds.



- CMGP's accuracy worsens with more features unless the sample size increases.
- Standard ARD fails to ignore less relevant features in small-data settings.
- Variance-weighted ARD shows minimal gains in this noiseless, deterministic setting, where flat variance signals limit its effect. Still, it remains safe and shows more promise in noisy, uncertain settings.

2. Failure Mode 2: Effect of Overlap-Aware Kernel Scaling

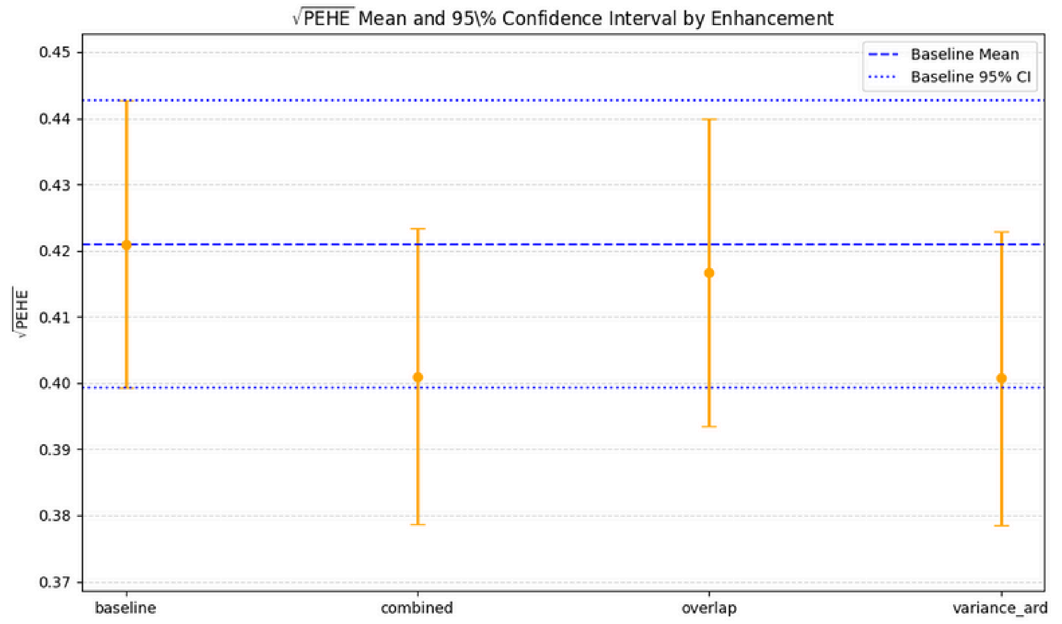
Baseline CMGP is compared with the overlap-aware kernel variant under varying treatment imbalance. The proportion of treated samples ranges from 10% to 50% (perfect balance), with synthetic data generated using fixed covariates and modifiers. Each point averages results across 100 seeds.



- Baseline CMGP becomes overconfident in regions with poor treatment-control overlap, especially when the treated group is small.
- The overlap-aware kernel reduces this overconfidence and improves uncertainty calibration.
- Gains are most noticeable in severely imbalanced settings (e.g., 10–20% treated).
- The method introduces no downside in balanced settings, confirming its robustness.

3. IHDP Benchmark

All CMGP variants are evaluated on the IHDP semi-synthetic dataset. The benchmark uses 100 random train-test splits with 25 real covariates and simulated treatment outcomes. Results are aggregated across all splits to assess generalization performance in a realistic, noisy setting.



- All enhanced variants perform comparably to or slightly better than the baseline CMGP.
- Variance-weighted ARD shows mild improvements in error and stability, confirming its utility in noisy, real-world data.
- Overlap-aware kernel maintains competitive performance without degrading accuracy or calibration.
- The combined model performs consistently well across splits, supporting the robustness of both enhancements.
- Results validate that neither modification harms generalization, reinforcing their benefit in uncertain, underdetermined settings.

7. CONCLUSIONS AND FUTURE WORK

KEY FINDINGS

CMGP struggles in complex settings.

- Performance drops when there are too many features or when treatment groups are imbalanced.

Enhancements improve reliability without hurting accuracy.

- Variance-weighted ARD helps in high-dimensional, noisy data.
- Overlap-aware kernels give better predictions and uncertainty estimates under imbalance.

Both methods are stable and generalizable.

- They perform consistently across synthetic and real-world datasets without degrading results.

LIMITATIONS & FUTURE WORK

Limited use of feature relevance

- Regularization based on treatment effect variance is only applied at initialization; future work should integrate it throughout training for stronger guidance.

Simplistic imbalance handling

- Overlap estimation relies on local treatment ratios, more flexible methods like density-based or learned measures could better reflect causal relevance.

Lack of real-world validation

- Current experiments use synthetic data, applying these methods to real observational datasets is needed to test reliability in practical settings.

RELATED LITERATURE

[1] C. E. Rasmussen and C. K. I. Williams, Gaussian Processes for Machine Learning. MIT Press, 2005.
[2] 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.
[3] 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, pp. 3424–3432, 2017.
[4] Z. Wang et al., "Bayesian optimization in high dimensions via random embeddings," Proc. IJCAI, pp. 1778–1784, 2013.
[5] C. J. Paciorek and M. J. Schervish, "Nonstationary covariance functions for gaussian process regression," in Advances in Neural Information Processing Systems, vol. 16, 2004.
[6] Y. Liu, Dynamic regularized cbdt: Variance-calibrated causal boosting for interpretable heterogeneous treatment effects, https : / / arxiv . org / abs / 2504 . 13733, arXiv:2504.13733, 2025.

AUTHORS

Logan Ritter

SUPERVISORS

Dr. Jesse Krijthe, Rickard Karlsson

AFFILIATIONS

EEMCS, Delft University of Technology, The Netherlands