The Importance of Ablation Studies for Complex Nonparametric Causal Models

Hugo Gobato Souto[1], **Francisco Louzada**[1]

[1]Institute of Mathematics and Computer Sciences – University of São Paulo (USP)
Av. Trab. São Carlense 400, 13566-590 – São Carlos – SP – Brazil

hugogobatosouto@gmail.com, louzada@icmc.usp.br

Abstract. *Ablation studies are essential for understanding the contribution of individual components within complex models, yet their application in nonparametric treatment effect estimation remains limited. This paper emphasizes the importance of ablation studies by examining the Bayesian Causal Forest (BCF) model, particularly the inclusion of the estimated propensity score $\hat{\pi}(x_i)$ intended to mitigate regularization-induced confounding (RIC). Through an ablation study and three different conditional ablation studies, we demonstrate that excluding $\hat{\pi}(x_i)$ does not diminish the model's performance in estimating average and conditional average treatment effects or in uncertainty quantification. Moreover, omitting $\hat{\pi}(x_i)$ reduces computational time by approximately 21%. These findings could suggest that the BCF model's inherent flexibility suffices in adjusting for confounding without explicitly incorporating the propensity score. The study advocates for the routine use of ablation studies in treatment effect estimation to ensure model components are essential and to prevent unnecessary complexity.*

Key Words: Conditional Average Treatment Effect, Average Treatment Effect, Bayesian Additive Regression Trees, Bayesian Causal Forest Model, Continuous Treatment Effect.

1. Introduction

In recent years, the field of machine learning has witnessed a substantial proliferation of complex models, particularly within the realm of neural networks [Souto and Moradi 2024; Olivares et al. 2023; Adhikari et al. 2019; Hasanpour et al. 2016]. A critical aspect of model development and validation that has gained prominence is the use of *ablation studies*. Ablation studies involve systematically removing or altering components of a model to assess their individual contributions to overall performance [Cao et al. 2020; Gomez et al. 2017; Wu et al. 2022]. This methodology enables researchers to dissect models, understand the significance of each component, and optimize architectures for enhanced efficiency and effectiveness.

The importance of ablation studies is well-recognized in the machine learning literature [Sheikholeslami et al. 2021; Du 2020; Meyes et al. 2019], especially in the development of neural networks for tasks such as image recognition and natural language processing. For example, He et al. 2016 utilized ablation studies to demonstrate the impact of residual connections on deep convolutional neural network performance, leading to the development of the ResNet architecture. Similarly, Szegedy et al. 2015 employed ablation experiments to refine the Inception network by analyzing the contributions of various convolutional filter sizes to the network's ability to capture different levels of image features.

Despite their widespread application in general machine learning research, the practice of conducting ablation studies is noticeably lacking in the literature on treatment effect estimation, particularly with complex nonparametric models used for estimating the Average Treatment Effect (ATE) and the Conditional Average Treatment Effect (CATE). ATE represents the average difference in the potential outcomes if the entire population were treated versus if they were untreated (i.e., $\mathbb{E}[Y(1) - Y(0)]$). CATE generalizes this to subgroups, representing the expected treatment effect conditioned on observed covariates $X_i = x_i$ (i.e., $\tau(x_i) = \mathbb{E}[Y_i(1) - Y_i(0) \mid X_i = x_i]$), which captures treatment effect heterogeneity across individuals. Many studies in this domain introduce sophisticated models without thoroughly investigating the necessity and influence of their individual components [Shalit et al. 2017; Wager and Athey 2018; Hill 2011; Hahn et al. 2020]. This omission can be problematic, as it may lead to the adoption of unnecessarily complex models, potential overfitting, and inefficiencies that could be mitigated through careful analysis.

In this paper, we highlight the importance of ablation studies in the context of treatment effect estimation by examining the Bayesian Causal Forest (BCF) model proposed by Hahn et al. 2020. The BCF model is a prominent example, already used in various applied studies [Yeager et al. 2019; Yeager et al. 2022; Bail et al. 2019], that incorporates the estimated propensity score into the baseline function to mitigate Regularization-induced Confounding (RIC). The propensity score, $\pi(x_i) = P(D_i = 1 \mid X_i = x_i)$, is the conditional probability of an individual receiving the treatment given their observed covariates, which serves as a balancing score in observational studies to adjust for confounding (systematic differences in pre-treatment covariates between treated and untreated groups). RIC is a bias that occurs when using regularized, flexible nonparametric estimators to control for confounding. When there is strong targeted selection (i.e., the treatment assignment is highly correlated with the baseline prognostic function of outcomes under control), the outcome model's regularization pressure can aggressively oversmooth the prognostic function, causing the treatment effect term $\tau(x_i)$ to absorb the residual unadjusted confounding and creating severe estimation bias [Hahn et al. 2018; Hahn et al. 2020]. The inclusion of the propensity score is presented as a critical component for accurately estimating treatment effects in observational studies [Hahn et al. 2020].

We challenge this assertion by conducting an ablation study focused on the use of the estimated propensity score within the BCF model. Utilizing synthetic data generated based on the designs proposed by Ballinari and Bearth 2024, we assess the necessity of the propensity score in varying data-generating processes (DGPs) and sample sizes. Our findings indicate that the inclusion of the estimated propensity score does not enhance the performance of the BCF model in any of the tested scenarios. Furthermore, incorporating the propensity score results in a significant decrease in computational efficiency, with the model running approximately 21% slower. These results underscore the critical role that ablation studies can play in the development and evaluation of models for treatment effect estimation. By systematically examining the contributions of individual model components, researchers can avoid unnecessary complexity, improve computational efficiency, and enhance the interpretability of their models.

The remainder of this paper is structured as follows: In Section 2, we explore the role of the estimated propensity score in the BCF model and discuss the implications of the RIC problem. Section 3 provides a brief overview of the synthetic data used for our analysis and outlines the evaluation methods employed. In Section 4 and 5, we present the results of our main ablation study and conditional ablation studies respectively, and offer a

detailed analysis of the findings. Finally, Section 6 provides a discussion of the theoretical implications, offers practical recommendations, and concludes the paper.

2. BCF Model Architecture

To mitigate the effects of RIC in observational studies with targeted selection, Hahn et al. 2020 proposed the BCF model, which incorporates the estimated propensity score $\hat{\pi}(x_i)$ directly into the outcome model as an additional covariate. The outcome model is formulated as:

$$\mathbb{E}[Y_i \mid X_i = x_i, D_i = d_i] = \mu(x_i, \hat{\pi}(x_i)) + \tau(x_i)d_i, \tag{1}$$

where $\mu(x_i, \hat{\pi}(x_i))$ is the prognostic function capturing the baseline outcome adjusted for the propensity score, and $\tau(x_i)$ is the heterogeneous treatment effect (CATE). The estimated propensity score $\hat{\pi}(x_i)$, usually obtained using a standard binary BART model, acts as a control covariate in the baseline function, effectively absorbing the strong confounding and freeing the treatment effect function $\tau(x_i)$ from regularization pressure.

For estimating $\mu(x_i, \hat{\pi}(x_i))$ and $\tau(x_i)$, BCF employs separate BART priors with different regularization parameters to reflect their distinct roles. The baseline prognostic function $\mu(x_i, \hat{\pi}(x_i))$ is modeled using 200 trees, a shrinkage parameter $\eta = 0.95$, and a depth parameter $\beta = 2$ to capture potentially complex baseline relationships [Chipman et al. 2010]. In contrast, the treatment effect function $\tau(x_i)$ is modeled with much stronger regularization to prevent overfitting and reflect the assumption that treatment effect heterogeneity is typically modest. The prior for $\tau(x_i)$ utilizes 50 trees, a shrinkage parameter $\eta = 0.25$, and a depth parameter $\beta = 3$ [Hahn et al. 2020].

By accounting for the treatment assignment mechanism within the outcome model, BCF aims to produce more reliable estimates of the ATE and CATE under strong target selection. Nonetheless, despite these theoretical advantages, the necessity of including $\hat{\pi}(x_i)$ in the BCF model has not been thoroughly examined through ablation studies. In our investigation, we aim to assess the impact of excluding $\hat{\pi}(x_i)$ from the BCF model by conducting a partial ablation study, challenging the assumption that the inclusion of the estimated propensity score is crucial for superb performance of the BCF model, even in cases of target selection and the outcome variable being largely determined by the covariates.

By systematically analyzing the model's performance, we provide insights into the practical significance of incorporating $\hat{\pi}(x_i)$ and contribute to the discourse on model simplification and efficiency in treatment effect estimation, making the BCF model without the propensity score 21% faster than the standard BCF model.

3. Methodology

3.1. Considered Synthetic Data

To evaluate the impact of including the estimated propensity score $\hat{\pi}(x_i)$ in the BCF model, we conducted a partial ablation study comparing three model variants: the original BCF model (BCF ($\hat{\pi}(\mathbf{X})$)), BCF with the true propensity score (BCF ($\pi(\mathbf{X})$)), and BCF without the propensity score (BCF (no $\hat{\pi}(\mathbf{X})$)), which sets $\hat{\pi}(x_i) = 0.5$ globally. Synthetic data allows rigorous evaluation against known ground-truth causal effects.

The synthetic DGPs are inspired by Ballinari and Bearth 2024 and designed to vary target selection and outcome determination by covariates, which dictate the severity of

regularization-induced confounding (RIC) [Hahn et al. 2020]. The systems are defined as:

$$X_i \sim \text{Uniform}(0,1)^5, \quad D_i \mid X_i \sim \text{Bernoulli}(\pi(X_i)), \quad \varepsilon_i \sim \mathcal{N}(0,1),$$
$$Y_i = b(X_i) + (D_i - 0.5) \cdot \left(\frac{X_{i,1} + X_{i,2}}{2\alpha} \right) + \varepsilon_i,$$

where X_i are covariates, D_i is the treatment indicator, Y_i is the outcome, and $\alpha \in \{1, 2, 4\}$ controls outcome determination by covariates. Larger α values correspond to more dominant baseline effects (roughly 3, 7, and 13 times the treatment effect size), compounding the negative impact of RIC. The baseline prognostic function $b(X_i)$ and propensity functions $\pi(X_i)$ are:

$$\begin{aligned}
b(X_i) &= \sin(\pi X_{i,1} X_{i,2}) + 2(X_{i,3} - 0.5)^2 + X_{i,4} + 0.5 X_{i,5}, \\
\pi_{\text{extreme}}(X_i) &= 0.05 + 0.9 \cdot \text{BetaCDF}_{2,4}(\text{sigmoid}(b(X_i))) \\
\pi_{\text{moderate}}(X_i) &= 0.05 + 0.75 \cdot \text{BetaCDF}_{2,4}(\text{sigmoid}(b(X_i))) \\
&\quad + 0.15 \cdot \text{BetaCDF}_{2,4}(\min(X_{i,1}, X_{i,2})), \\
\pi_{\text{slight}}(X_i) &= 0.05 + 0.9 \cdot \text{BetaCDF}_{2,4}(\min(X_{i,1}, X_{i,2})).
\end{aligned}$$

These formulations correspond to DGP1 (extreme target selection), DGP2 (moderate), and DGP3 (slight). Figure 1 visualizes these configurations. Following Hahn et al. 2020, we use $n = 250$ for the main studies. To ensure conciseness, the main text focuses on the most challenging confounding setting ($\alpha = 4$), while results for $\alpha = 1$ and $\alpha = 2$ are hosted in the online repository (and yield identical conclusions).

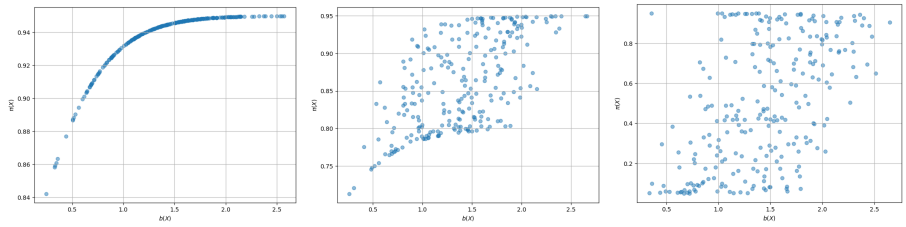


Figure 1. Scatter plot of $\pi(X)$ by $b(X)$ for different configurations of $\pi(X)$

To further scrutinize $\hat{\pi}(x_i)$'s role and model robustness under stress, we performed conditional ablation studies systematically varying key hyperparameters and data properties for DGP1 ($\alpha = 4$):

1. **Sample Size (N)**: Varying $n \in \{100, 500, 1000, 10000\}$ to assess propensity score utility in both high-variance small-sample regimes and high-precision large-sample asymptotic regimes.
2. **Prognostic Complexity (μ trees)**: Varying the number of BART trees for the prognostic component $\mu(X_i)$ across $\{5, 25, 100, 400\}$. A simpler prognostic model (e.g., 5 trees) limits model flexibility, potentially forcing greater reliance on $\hat{\pi}(x_i)$ to mitigate RIC.
3. **Covariate Space Dimension (p)**: Varying influential covariates $p \in \{10, 50, 100, 200\}$ by dividing covariates into blocks of 5 and summing their block-wise baseline contributions. Higher dimensionality increases regularization pressure, theoretically amplifying RIC and magnifying the potential benefit of $\hat{\pi}(x_i)$.

All simulations are repeated 100 times to obtain robust distribution statistics.

3.2. Evaluation Methods

We evaluate the BCF variants on ATE and CATE estimation accuracy and uncertainty quantification using two key metrics: Root Mean Squared Error (RMSE) to assess point estimation precision, and Coverage (Cover) of the nominal 95% credible intervals to evaluate uncertainty quantification. Results from the 100 simulations are reported as means and standard deviations.

To establish statistical significance of performance differences between BCF with $\hat{\pi}(\mathbf{X})$ and BCF without $\hat{\pi}(\mathbf{X})$, we employ Welch's t-test [WELCH 1947], which is robust to unequal variances [Souto and Louzada 2024b]. The normality assumption is justified by the Central Limit Theorem given the 100 independent simulation replicates.

4. Results and Analysis: Main Ablation Studies

Before going to the results of the specifics DGPs, it is worth mentioning that all analyses were performed using the same computational environment to maintain consistency. The computational time for fitting each model was recorded to assess the impact of including the estimated propensity score on computational efficiency. The inclusion of $\hat{\pi}(x_i)$ increased the computational time by approximately 21%, demonstrating a trade-off between model complexity and computational resources. As we will see below, the additional computational costs do not lead to a better performance at CATE and ATE estimation nor uncertainty quantification.

Table 1 presents the results of the evaluation metrics for DGP1 ($b(X); \pi_{\text{extreme}}(X)$), DGP2 ($b(X); \pi_{\text{moderate}}(X)$), and DGP3 ($b(X); \pi_{\text{slight}}(X)$). Across all three DGPs, the three model variants perform remarkably similarly. For DGP3, the point estimation error (RMSE) of BCF (no $\hat{\pi}(\mathbf{X})$) is slightly higher, although statistical tests are required to confirm if these differences are significant.

Moving to the statistical tests, Figure 2 presents the p-values comparing BCF ($\hat{\pi}(\mathbf{X})$) and BCF (no $\hat{\pi}(\mathbf{X})$). For all metrics, the p-values are overwhelmingly high, showing that omitting the estimated propensity score does not lead to any statistically significant difference in treatment effect estimation accuracy (RMSE) or uncertainty quantification (Coverage) across any of the DGPs.

Hence, it can be concluded that the estimation and use of the propensity score is unnecessary for the BCF model's performance (even in extreme target selection and prognostic-dominant outcome settings). Omitting it avoids unnecessary complexity and reduces training time by approximately 21%.

5. Results and Analysis: Conditional Ablation Studies

5.1. Varying Number of Samples

The results are presented in Figures 3A and 3B. Across all tested sample sizes, BCF without $\hat{\pi}(\mathbf{X})$ performed comparably to BCF with $\hat{\pi}(\mathbf{X})$ and BCF with the true $\pi(\mathbf{X})$ in terms of RMSE and Coverage. The corresponding Welch's t-test p-values (available in the online repository) are overwhelmingly high, confirming no statistically significant difference in accuracy or uncertainty quantification between the variants with and without estimated propensity score, even at extremely small ($N = 100$) or large ($N = 10000$) sample sizes.

Table 1. Key Causal Inference Metrics for Main Ablation Study ($\alpha = 4$). Best results per row are highlighted in bold (lowest RMSE, Coverage closest to 0.95; ties broken by lower standard deviation).

DGP	Metric	BCF (no $\hat{\pi}(\mathbf{X})$)	BCF ($\pi(\mathbf{X})$)	BCF ($\hat{\pi}(\mathbf{X})$)
DGP1	RMSE_{CATE}	0.157 ± 0.119	0.155 ± 0.112	$\mathbf{0.155 \pm 0.106}$
	Cover_{CATE}	$\mathbf{0.99 \pm 0.04}$	1.00 ± 0.02	1.00 ± 0.02
	RMSE_{ATE}	0.139 ± 0.127	0.139 ± 0.120	$\mathbf{0.138 \pm 0.112}$
	Cover_{ATE}	$\mathbf{0.96 \pm 0.20}$	0.98 ± 0.14	0.99 ± 0.10
DGP2	RMSE_{CATE}	0.130 ± 0.083	$\mathbf{0.129 \pm 0.081}$	0.134 ± 0.090
	Cover_{CATE}	$\mathbf{1.00 \pm 0.01}$	1.00 ± 0.02	1.00 ± 0.02
	RMSE_{ATE}	0.110 ± 0.088	$\mathbf{0.108 \pm 0.089}$	0.113 ± 0.096
	Cover_{ATE}	$\mathbf{0.95 \pm 0.22}$	0.97 ± 0.17	0.97 ± 0.17
DGP3	RMSE_{CATE}	0.147 ± 0.099	0.131 ± 0.087	$\mathbf{0.129 \pm 0.083}$
	Cover_{CATE}	$\mathbf{0.98 \pm 0.07}$	0.98 ± 0.08	0.99 ± 0.06
	RMSE_{ATE}	0.117 ± 0.105	0.100 ± 0.094	$\mathbf{0.100 \pm 0.087}$
	Cover_{ATE}	0.91 ± 0.29	$\mathbf{0.94 \pm 0.24}$	0.97 ± 0.17

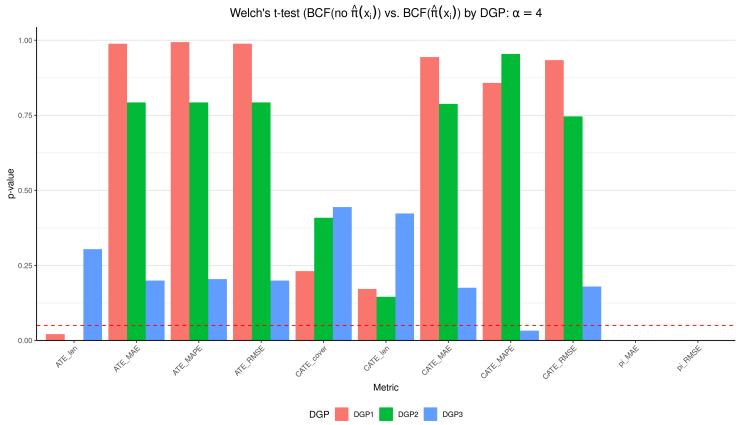


Figure 2. Welch t-test p-values for BCF (no $\hat{\pi}(\mathbf{X})$) vs BCF ($\hat{\pi}(\mathbf{X})$)

5.2. Varying Number of Prognostic Trees

Next, we varied the number of trees in the prognostic component (μ trees $\in \{5, 25, 100, 400\}$), with results presented in Figures 3C and 3D. It was hypothesized that a simpler prognostic model (e.g., 5 μ trees) might benefit more from the explicit inclusion of $\hat{\pi}(\mathbf{X})$. However, our empirical results do not support this hypothesis: BCF ($\hat{\pi}(\mathbf{X})$) and BCF (no $\hat{\pi}(\mathbf{X})$) perform indistinguishably. While BCF ($\pi(\mathbf{X})$) with the true propensity score does show slight benefits under extreme tree constraints (5 μ trees), this advantage disappears once we use the estimated propensity score $\hat{\pi}(\mathbf{X})$ or use a more realistic forest size, confirming that incorporating $\hat{\pi}(\mathbf{X})$ remains unnecessary.

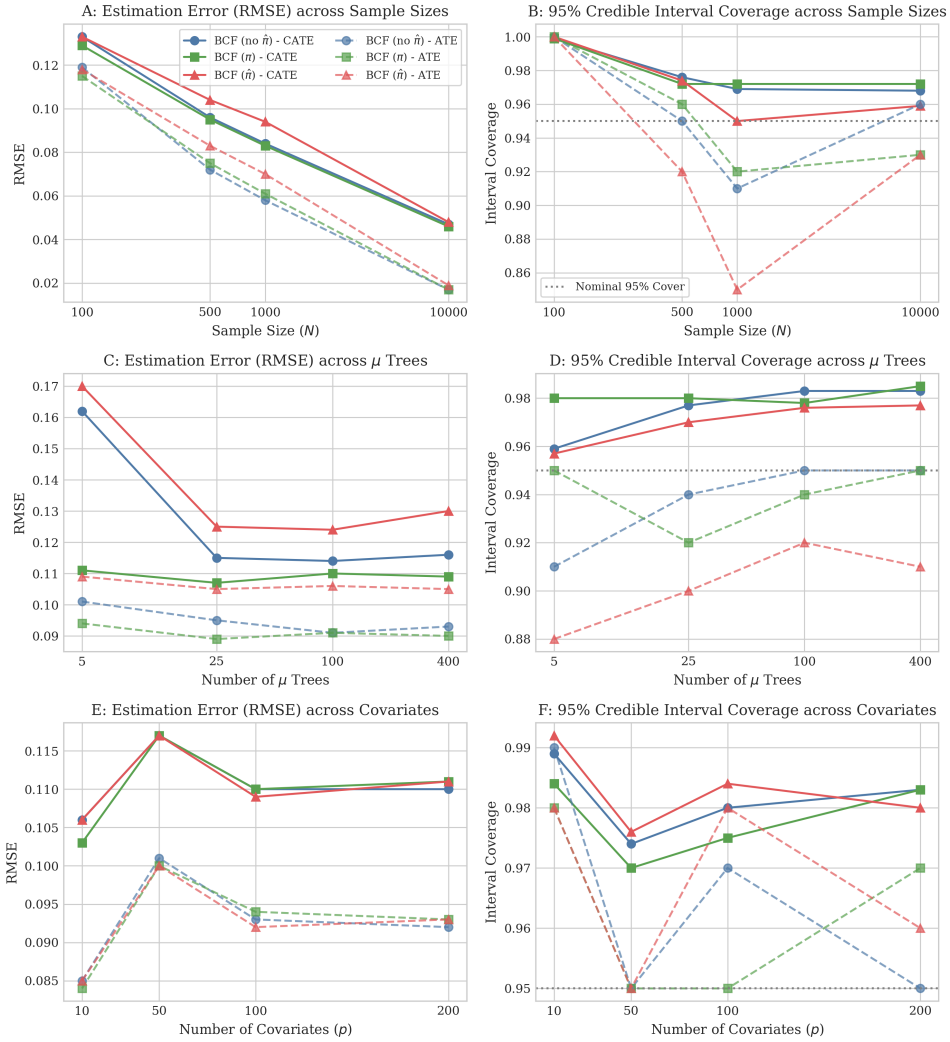


Figure 3. Key causal inference metrics across the conditional ablation studies (DGP1 with $\alpha = 4$). Panels A and B show RMSE and coverage across sample sizes N; panels C and D show them across the number of prognostic μ trees; panels E and F show them across covariate space dimensions p. Solid lines represent CATE, and dashed lines represent ATE. The dotted gray line indicates the nominal 95% credible interval coverage.

5.3. Varying Number of Covariates

Finally, we investigated whether increasing the covariate space dimensionality ($p \in \{10, 50, 100, 200\}$) makes the propensity score more critical due to heightened regularization pressure. The results are presented in Figures 3E and 3F. The estimation accuracy (RMSE) and credible interval coverage (Coverage) of all three BCF variants remain relatively constant and indistinguishable as p increases. This robustness of the BCF outcome model to high-dimensional confounders aligns with findings in contributed discussions by Prado et al. 2020. Across all covariate dimensions, the performance of BCF (no $\hat{\pi}(\mathbf{X})$) and BCF ($\hat{\pi}(\mathbf{X})$) is statistically identical (with uniformly high p-values), indicating that BCF's inherent regularization manages high-dimensional spaces without requiring explicit propensity score adjustment.

6. Discussion, Recommendations, and Conclusion

While Hahn et al. 2020 argue that the primary motivation for including the estimated propensity score, $\hat{\pi}(x_i)$, in the BCF prognostic function is to mitigate regularization-induced confounding (RIC), alternative theoretical perspectives suggest that its inclusion serves as a safeguard analogous to doubly robust estimators in frequentist statistics [Li et al. 2023; Zigler et al. 2013; Papadogeorgou and Li 2020]. Under this view, conditioning on a correctly specified propensity score can correct for potential misspecification in the outcome model. However, the practical utility of this safeguard is highly questionable when the outcome model is estimated using exceptionally flexible non-parametric methods, such as the BART models at the core of BCF [Ročková and Saha 2019; Ročková and van der Pas 2020].

Interpreting the ablation results. Our empirical findings are remarkably consistent. In the main ablation (Table 1 and Figure 2), BCF without the propensity score matches the original BCF on all four metrics and all three levels of target selection. The absolute differences in RMSE are on the order of 0.001 to 0.02, well within one standard deviation across replications, and Welch's p-values are uniformly large, indicating no statistically distinguishable signal. Coverage remains slightly conservative (above the nominal 95%) for all variants, as is common for BART-based credible intervals, but the degree of conservatism does not depend on $\hat{\pi}(x_i)$. In DGP3 the point estimates for BCF without $\hat{\pi}(x_i)$ appear slightly worse (0.147 versus 0.129 for CATE), yet the large replication variance renders this difference non-significant, a pattern confirmed by the formal tests. The conditional ablations reinforce this interpretation. Varying sample size from $N = 100$ to $N = 10000$ leaves the RMSE and coverage curves for the three variants visually and statistically overlapping, even in the high-variance small-sample regime where a regularized helper covariate would be expected to help most. Varying the prognostic forest from 5 to 400 trees shows that only access to the oracle $\pi(X_i)$ yields a transient advantage under extreme capacity constraints, and this advantage vanishes once $\hat{\pi}(x_i)$ is estimated or a realistic 200-tree prior is used. Finally, expanding the covariate dimension from $p = 10$ to $p = 200$, which increases regularization pressure and was hypothesized to magnify RIC, produces flat trajectories for all variants, consistent with prior discussion that BCF outcome regularization scales gracefully with dimensionality [Prado et al. 2020]. Taken together, the results do not indicate a trade-off between accuracy and uncertainty quantification: omitting $\hat{\pi}(x_i)$ preserves both.

Computational trade-off. Omitting the estimated propensity score yields a substantial practical advantage: it reduces overall fitting time by approximately 21% in our environment, because a full binary BART model for $\hat{\pi}(x_i)$ is no longer fitted and the prognostic forest no longer conditions on an additional covariate. Including $\hat{\pi}(x_i)$ by default therefore incurs measurable overhead with no detectable benefit for estimation accuracy or coverage. The only plausible exceptions are settings where the true propensity score is known by design (e.g., randomized experiments, $N \leq 25$) or where the analyst's primary estimand is the assignment mechanism itself.

From synthetic DGPs to real-world generalizability. The exclusive use of synthetic data is both a strength and a limitation. It provides ground-truth ATE and CATE, allowing precise quantification of bias, variance, and coverage, and it lets us construct worst-case

RIC scenarios by tightly coupling $\pi(X_i)$ and $b(X_i)$ and by inflating baseline dominance via $\alpha = 4$. If $\hat{\pi}(x_i)$ were essential, it should reveal itself in these stress tests. The fact that it does not suggests the null result is conservative and likely to generalize to many observational settings where confounding is weaker. Generalization to real-world data nevertheless requires caution. Real covariates exhibit complex dependencies, mixed-type and missingness, measurement error, and potentially unmeasured confounding, none of which our uniform X_i on $[0, 1]^5$ blocks fully capture. Moreover, real data lack observable ground truth, so evaluation must rely on semi-synthetic benchmarks (e.g., IHDP, ACIC 2016 and 2018) that graft synthetic effects onto real covariate distributions, on calibrated placebo and negative-control analyses, and on sensitivity analyses for unmeasured confounding. Applied BCF studies such as those in education and political science [Yeager et al. 2019; Yeager et al. 2022; Bail et al. 2019] have used the original specification successfully, but our results imply those analyses would likely obtain nearly identical causal estimates at lower computational cost if $\hat{\pi}(x_i)$ were omitted. We view our synthetic evidence as a strong prior for practice, to be complemented by prospective real-data replications. Future work should therefore replicate the ablation on IHDP and ACIC challenge datasets, on electronic health records with high-dimensional codes, and under continuous treatments [Souto and Louzada 2024a] where propensity surfaces are harder to estimate, as well as under deliberate propensity misspecification to test robustness.

Implications and recommendation. More broadly, this work highlights a pervasive methodological gap in the causal machine learning literature. While general machine learning routinely validates each architectural component, new causal estimators are frequently proposed with complex, unvalidated structures and cited as essential without evidence. Although an ablation study is by nature not a new estimator, its contribution is methodological discipline: identifying unnecessary complexity, improving parsimony, reproducibility, and deployment cost, and preventing the literature from accumulating fragile defaults. We recommend that practitioners omit the propensity score as the default in BCF for heterogeneous effect estimation and that authors of new nonparametric causal models report routine ablations, ideally across sample size, model capacity, and dimensionality, as done here. Systematically testing component necessity will yield simpler, faster, and more interpretable models and accelerate their integration into high-stakes decision support.

Supplementary Research Material and Code

The code and data used in this paper are publicly available at https://github.com/hugogobato/The-Importance-of-Ablation-Studies-for-Complex-Nonparametric-Causal-Models. Reproducibility: the cleaned repository is at `Ablation_BCF/GitHub_Repo/` in the submission bundle. It contains the unified `data_generation/Creating_datasets.py`, the three BCF variants (`BCF_N250.R`, `BCF(no_pi_hat)_N250.R`, `BCF(pi_oracle)_N250.R`), and precomputed `*.xlsx`/`Results.txt` for all DGPs ($\alpha = 1, 2, 4$) and conditional ablations. See `GitHub_Repo/README.md` for one-command reproduction.

Acknowledgments

The authors would like to express heartfelt gratitude to the São Paulo Research Foundation (FAPESP) for their financial support through the Master's scholarship (Processo 2024/06274-0) of the first author of this paper and the associated project (Processo Vinculado 2013/07375-0).

References

Adhikari, A., Ram, A., Tang, R., and Lin, J. (2019). Rethinking complex neural network architectures for document classification. In Burstein, J., Doran, C., and Solorio, T., editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4046–4051, Minneapolis, Minnesota. Association for Computational Linguistics.

Bail, C. A., Guay, B., Maloney, E., Combs, A., Hillygus, D. S., Merhout, F., Freelon, D., and Volfovsky, A. (2019). Assessing the russian internet research agency's impact on the political attitudes and behaviors of american twitter users in late 2017. *Proceedings of the National Academy of Sciences*, 117(1):243–250.

Ballinari, D. and Bearth, N. (2024). Improving the finite sample performance of double/debiased machine learning with propensity score calibration.

Cao, D., Wang, Y., Duan, J., Zhang, C., Zhu, X., Huang, C., Tong, Y., Xu, B., Bai, J., Tong, J., and Zhang, Q. (2020). Spectral temporal graph neural network for multivariate time-series forecasting. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 17766–17778. Curran Associates, Inc.

Chipman, H. A., George, E. I., and McCulloch, R. E. (2010). Bart: Bayesian additive regression trees. *The Annals of Applied Statistics*, 4(1).

Du, L. (2020). How much deep learning does neural style transfer really need? an ablation study. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 3150–3159. IEEE.

Gomez, A. N., Ren, M., Urtasun, R., and Grosse, R. B. (2017). The reversible residual network: Backpropagation without storing activations. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Hahn, P. R., Carvalho, C. M., Puelz, D., and He, J. (2018). Regularization and confounding in linear regression for treatment effect estimation. *Bayesian Analysis*, 13(1).

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

Hasanpour, S. H., Rouhani, M., Fayyaz, M., and Sabokrou, M. (2016). Let's keep it simple, using simple architectures to outperform deeper and more complex architectures.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

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

Li, F., Ding, P., and Mealli, F. (2023). Bayesian causal inference: a critical review. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 381(2247).

Meyes, R., Lu, M., de Puiseau, C. W., and Meisen, T. (2019). Ablation studies in artificial neural networks.

Olivares, K. G., Challu, C., Marcjasz, G., Weron, R., and Dubrawski, A. (2023). Neural basis expansion analysis with exogenous variables: Forecasting electricity prices with nbeatsx. *International Journal of Forecasting*, 39(2):884–900.

Papadogeorgou, G. and Li, F. (2020). Discussion for 'bayesian regression tree models for causal inference: regularization, confounding, and heterogeneous effects'. *Bayesian Analysis*, 15:1007–1013.

Prado, E. a. B., O'Neill, E., Hernández, B., Parnell, A. C., and Moral, R. A. (2020). Contributed discussion of Bayesian Regression Tree Models for Causal Inference: Regularization, Confounding, and Heterogeneous Effects. *Bayesian Analysis*, 15(3):1029–1031. Discussion contribution to Hahn, Murray, and Carvalho.

Ročková, V. and Saha, E. (2019). On theory for bart. In Chaudhuri, K. and Sugiyama, M., editors, *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pages 2839–2848. PMLR.

Ročková, V. and van der Pas, S. (2020). Posterior concentration for bayesian regression trees and forests. *The Annals of Statistics*, 48(4).

Shalit, U., Johansson, F. D., and Sontag, D. (2017). Estimating individual treatment effect: generalization bounds and algorithms. In Precup, D. and Teh, Y. W., editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 3076–3085. PMLR.

Sheikholeslami, S., Meister, M., Wang, T., Payberah, A. H., Vlassov, V., and Dowling, J. (2021). Autoablation: Automated parallel ablation studies for deep learning. In *Proceedings of the 1st Workshop on Machine Learning and Systems*, EuroSys '21. ACM.

Souto, H. G. and Louzada, F. (2024a). Advancing causal inference: A nonparametric approach to ATE and CATE estimation with continuous treatments.

Souto, H. G. and Louzada, F. (2024b). Beyond arbitrary replications: A principled approach to simulation design in causal inference.

Souto, H. G. and Moradi, A. (2024). Can transformers transform financial forecasting? *China Finance Review International*, 14(4):1–24.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. (2015). Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Wager, S. and Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242.

WELCH, B. L. (1947). The generalization of 'student's' problem when several different population varlances are involved. *Biometrika*, 34(1–2):28–35.

Wu, X., Xia, Y., Zhu, J., Wu, L., Xie, S., and Qin, T. (2022). A study of bert for context-aware neural machine translation. *Machine Learning*, 111(3):917–935.

Yeager, D. S., Bryan, C. J., Gross, J. J., Murray, J. S., Krettek Cobb, D., H. F. Santos, P., Gravelding, H., Johnson, M., and Jamieson, J. P. (2022). A synergistic mindsets intervention protects adolescents from stress. *Nature*, 607(7919):512–520.

Yeager, D. S., Hanselman, P., Walton, G. M., Murray, J. S., Crosnoe, R., Muller, C., Tipton, E., Schneider, B., Hulleman, C. S., Hinojosa, C. P., Paunesku, D., Romero, C., Flint, K., Roberts, A., Trott, J., Iachan, R., Buontempo, J., Yang, S. M., Carvalho, C. M., Hahn, P. R., Gopalan, M., Mhatre, P., Ferguson, R., Duckworth, A. L., and Dweck, C. S. (2019). A national experiment reveals where a growth mindset improves achievement. *Nature*, 573(7774):364–369.

Zigler, C. M., Watts, K., Yeh, R. W., Wang, Y., Coull, B. A., and Dominici, F. (2013). Model feedback in bayesian propensity score estimation. *Biometrics*, 69(1):263–273.