CALIBRATED INFERENCE FOR THE CONDITIONAL AVERAGE TREATMENT EFFECT IN THE FEW-PLACEBO REGIME VIA GAUSSIAN PROCESSES

Anonymous authors
Paper under double-blind review

ABSTRACT

In the *few-placebo regime* ($N_0 \ll N_1$), the control-arm outcome surface μ_0 is only weakly identified, and this under-identification is a primary determinant of interval reliability for the conditional average treatment effect (CATE). Existing estimators can under-cover because their inference procedures do not propagate this under-identification into the interval: point-estimate methods inherit nuisance bias, and GP methods with MAP hyperparameter estimation suppress residual uncertainty in the treatment-effect hyperparameters.

We propose **ISO-HMC**, a structured Gaussian-process model that places a separate prior on the treatment-effect function and marginalises over its hyperparameters $(\sigma_\tau, \ell_\tau, \sigma^2)$ via HMC. By propagating this uncertainty through the law of total variance, ISO-HMC achieves calibration when the treatment effect is smooth and low-dimensional.

We test three claims: (i) ISO-HMC is calibrated in its target regime; (ii) it achieves this without sacrificing interval informativeness; (iii) it degrades gracefully under misspecification, with an independent-arm GP as a safe fallback. All three are supported: on smooth low-dimensional designs ISO-HMC attains calibrated coverage and substantially narrower intervals than a conservative independent-arm GP baseline, while on nonlinear or high-dimensional designs it under-covers and the independent-arm GP restores calibration.

Overall, these results show that in the few-placebo regime, calibrated CATE intervals benefit from propagating uncertainty through both the outcome model and the treatment-effect hyperparameters. Suppressing either through point estimates or MAP leads to overconfidence and under-coverage.

1 INTRODUCTION

Estimating heterogeneous treatment effects — the conditional average treatment effect (CATE) $\tau(x) = \mathbb{E}[Y(1) - Y(0) \mid X = x]$ — is central to data-driven decisions in medicine, economics, and policy. For high-stakes decisions, point estimates alone are often insufficient: a decision maker typically needs to know how likely it is that $\tau(x)$ exceeds a relevant threshold, not merely whether a point estimate is positive or negative. This requires a calibrated interval or posterior, which translates the remaining uncertainty into a usable probability. Moreover, flexible machine-learning estimators are prone to overfitting and can produce overconfident point predictions (Guo et al., 2017; Kuleshov et al., 2018); calibrated uncertainty therefore acts as a safeguard against such overconfidence.

We study the *few-placebo regime*: one treatment arm is much smaller than the other, $N_0 \ll N_1$. This is not a corner case. Unequal-allocation clinical trials assign more patients to the active arm by design; budget-constrained A/B tests hold out only a small control group; rare-regime studies (a stress period, an infrequent exposure) have few units in the rare arm. However, a systematic understanding of which CATE interval estimators remain calibrated in this regime, which fail, and why is, to our knowledge, still missing. This paper aims to provide such an understanding.

The identification problem. Our main finding is that calibration success in the few-placebo regime depends on how an estimator handles the structural under-identification of μ_0: with few controls, the data provide only weak identification of $\mu_0(x)$. The degree of this under-identification is governed

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

by three factors: *(i)* the control-to-dimension ratio N_0/d (fewer controls per dimension worsens identifiability), *(ii)* the complexity of τ (a complex treatment effect is harder to separate from μ_0), and *(iii)* the arm correlation $\mathrm{Corr}(\mu_0, \mu_1)$ (high correlation lets the treated arm inform the control-arm posterior). Each estimator we study occupies a distinct region of this three-factor space, reflecting a distinct strategy for handling the under-identification.

X-Learner. A natural estimator in this setting is the X-Learner (Künzel et al., 2019), a meta-learner designed for arm imbalance that provides point estimates of $\tau(x)$. For high-stakes decisions, however, a point estimate alone is often insufficient: a decision maker needs to know how likely it is that $\tau(x)$ exceeds a relevant threshold, not merely whether a point estimate is positive or negative. This requires a calibrated interval or posterior, which translates the remaining uncertainty into a usable probability. A natural Bayesian approach to CATE uncertainty is to place priors over the outcome surfaces and report the posterior of $\tau(x)$ (Hill, 2011; Hahn et al., 2020; Alaa & van der Schaar, 2017); applying the same logic to the X-Learner suggests making its second-stage regression Bayesian — regress the pseudo-outcome $D_1 = Y - \hat{\mu}_0(X)$ with a probabilistic model and read off the posterior. We show these intervals under-cover: the pseudo-outcome decomposes as $\tau(X) + (\mu_0 - \hat{\mu}_0)(X) + \text{noise}$, and the imputation error $\mu_0 - \hat{\mu}_0$ is a first-order, non-zero bias inherited from the regularized control-arm nuisance. A variance model attached after the fact cannot move the misplaced interval, because the bias lies in the target's location, not in the noise scale. In the few-placebo regime (N_0=30, N_1=500), the GP second stage is very precise (the large treated arm shrinks the posterior variance to $\approx 0.09^2$), amplifying the nuisance bias (≈ 0.15) to nearly 1.7σ and driving the nominal-95% interval to cover only 0.53. The standard remedy — Kennedy's doubly-robust score (Kennedy, 2023) — also fails: the few-placebo regime's extreme propensities amplify the orthogonal score's second-order remainder by $1/(1-\pi)^2$, and efficiency weighting re-introduces the bias. Both failures share the same structural root: uncertainty is modelled around a point estimate of the hard placebo nuisance μ_0, while the bias of that point estimate is not represented in the model.

GP-based estimators. Among estimators that achieve calibration, the key design choices are arm independence vs. coupling, and isotropic vs. structured priors on τ.

GP-CATE models each arm's outcome surface with an independent Gaussian process (cf. Alaa & van der Schaar, 2017); the CATE posterior is the difference $\tau(x) \sim \mathcal{N}(m_1 - m_0, s_1^2 + s_0^2)$. Each arm's GP posterior directly represents the uncertainty from that arm's sample: the control arm's few data points produce a large s_0, which enters the CATE interval directly rather than as an unmodelled bias. Across all designs we evaluate, GP-CATE achieves calibrated coverage (0.94–0.99); it is the robust default.

CMGP (Alaa & van der Schaar, 2017) couples the two arm GPs via a multi-task kernel. On the high-dimensional IHDP benchmark it achieves near-calibration (0.92) while more than halving interval width ($54\,\%$ narrower at N_0=30). However, CMGP under-covers on low-dimensional or non-linear designs (0.87–0.88), indicating that the multi-task coupling is not supported by the data in these settings. CMGP is the right choice when high arm correlation can be justified (e.g., high-dimensional observational data with a shared baseline surface).

ISO places an independent GP prior on $\tau(x)$ directly, under the structural model $Y = \mu_0(x) + W\tau(x) + \varepsilon$. Our primary proposed method, **ISO-HMC**, marginalises over the τ hyperparameters $(\sigma_\tau, \ell_\tau, \sigma^2)$ via HMC under a log-normal prior, correctly accounting for hyperparameter uncertainty when N_0 is small. ISO-HMC achieves calibrated coverage on smooth low-dimensional designs while producing substantially narrower intervals than GP-CATE; on non-linear or high-dimensional designs it under-covers and GP-CATE is the safe fallback.

Contributions. We make three contributions. First, we diagnose why standard interval estimators fail in the few-placebo regime: both point-estimate methods and MAP-based GPs suppress the uncertainty introduced by weak identification (§3–4). Second, we introduce ISO-HMC, a structured GP model that combines a separate treatment-effect prior with HMC marginalisation of its hyperparameters, and show that it is calibrated and efficient when the treatment effect is smooth and low-dimensional (§5). Third, we map the calibration landscape across this regime, identifying three factors—N_0/d, τ-complexity, and arm correlation—that determine which estimator is appropriate (§6–7).

2 Background and Setup

We observe i.i.d. tuples $(X_i, W_i, Y_i)_{i=1}^N$ with covariates $X_i \in \mathcal{X} \subseteq \mathbb{R}^d$, binary treatment $W_i \in \{0, 1\}$, and outcome $Y_i \in \mathbb{R}$. Write $N_0 = \sum_i \mathbb{1}\{W_i = 0\}$ and $N_1 = \sum_i \mathbb{1}\{W_i = 1\}$; the few-placebo regime is $N_0 \ll N_1$. Let $\pi(x) = \mathbb{P}(W = 1 \mid X = x)$ be the propensity score and $\mu_w(x) = \mathbb{E}[Y \mid W = w, X = x]$ the per-arm conditional mean. Under unconfoundedness $W \perp\!\!\!\perp (Y(0), Y(1)) \mid X$ and overlap $0 < \pi(x) < 1$ a.s. (Rosenbaum & Rubin, 1983), the CATE $\tau(x) = \mu_1(x) - \mu_0(x)$ is identified.

The canonical meta-learner for this setting is the X-Learner of Künzel et al. (2019), which estimates τ in three steps. (i) Fit per-arm conditional means $\hat{\mu}_0, \hat{\mu}_1$ by cross-fitted machine learners. (ii) Form cross-arm pseudo-outcomes,

$$D_1^{(i)} = Y_i - \hat{\mu}_0(X_i) \qquad (W_i = 1), \tag{1}$$

$$D_0^{(i)} = \hat{\mu}_1(X_i) - Y_i \qquad (W_i = 0). \tag{2}$$

(iii) Regress D_w on X within each arm and combine by propensity weighting. The X-Learner is designed for arm imbalance: cross-arm imputation transfers structure from the large arm to the small one, and it out-performs the S- and T-Learner when the arms differ greatly in size. To avoid in-sample bias from plugging $\hat{\mu}_w$ into (1)–(2), we use K-fold cross-fitting (Chernozhukov et al., 2018) throughout: nuisances are trained on $K - 1$ folds and evaluated on the held-out fold, so each unit's pseudo-outcome uses a nuisance not fit on that unit.

A pseudo-outcome is *Neyman-orthogonal* when its conditional mean is first-order insensitive to nuisance perturbations (Chernozhukov et al., 2018). The X-Learner pseudo-outcomes (1)–(2) are *not* orthogonal. The canonical orthogonal alternative is the doubly-robust (DR) score (Robins et al., 1995; Kennedy, 2023),

$$\begin{aligned} D_i^{\mathrm{DR}} = \hat{\mu}_1(X_i) - \hat{\mu}_0(X_i) \\ + \frac{W_i\big(Y_i - \hat{\mu}_1(X_i)\big)}{\hat{\pi}(X_i)} \\ - \frac{(1 - W_i)\big(Y_i - \hat{\mu}_0(X_i)\big)}{1 - \hat{\pi}(X_i)}, \end{aligned} \tag{3}$$

whose conditional mean equals $\tau(X)$ up to a remainder that is a *product* of nuisance errors. Regressing D^{DR} gives the DR-Learner (Kennedy, 2023).

Both pieces of (3) divide by $\hat{\pi}$ or $1 - \hat{\pi}$, so the behaviour of the DR score is governed by *overlap*: how far $\pi(x)$ stays from 0 and 1. The few-placebo regime is, by construction, a poor-overlap regime — a small control arm means $1 - \pi$ is small for typical covariate values, amplifying the denominators in (3). §4 shows that this amplification causes the orthogonal score to fail.

Finally, we use Gaussian-process (GP) regression (Rasmussen & Williams, 2006). A GP places a prior over functions and returns a full posterior; uncertainty from limited data enters the posterior directly. §5 exploits this property to model the treatment effect, with calibration assessed empirically.

Comparison methods. GP-CATE (cf. Alaa & van der Schaar, 2017) fits an independent isotropic GP to each arm; the CATE posterior is $\tau(x) \mid \mathcal{D} \sim \mathcal{N}(m_1 - m_0,\ s_1^2 + s_0^2)$. When N_0 is small, s_0 is large and the interval is wide; this is the correct report of the problem's difficulty. GP-CATE is calibrated across all our designs and serves as the safe default. CMGP (Alaa & van der Schaar, 2017) couples the two arm GPs via a shared latent function $\mu_w = f_{\mathrm{shared}} + \delta_w$, halving interval width when arm surfaces are highly correlated (as in high-dimensional observational data) but miscalibrating on low-dimensional or non-linear designs.

3 First-Order Bias of the X-Learner Pseudo-Outcome

The X-Learner (Künzel et al., 2019) is a point-estimation meta-learner and was not designed to produce uncertainty intervals. Nevertheless, to obtain intervals for the X-Learner, we apply a Bayesian approach to CATE uncertainty that places priors on the relevant regression functions and reports the posterior of $\tau(x)$, as in (Hill, 2011; Hahn et al., 2020; Alaa & van der Schaar, 2017); here, we

apply this approach to the X-Learner's second-stage pseudo-outcome regression. For the purpose of diagnosis and comparison, we evaluate this Bayesian GP second-stage X-Learner; this isolates the effect of pseudo-outcome bias on interval coverage while keeping the second-stage model class fixed. Our diagnosis concerns the point estimate itself: we show that the pseudo-outcome inherits a first-order nuisance bias, and that this bias displaces any interval centred on it, regardless of how that interval is constructed.

This section analyses the treated-arm pseudo-outcome regression of the X-Learner in the Bayesian GP formulation. The full X-Learner combines this with a control-arm analogue D_0 via propensity weighting; in the few-placebo regime the propensity weighting concentrates mass on the treated-arm component, so the bias analysed here is the dominant failure mode. This section shows the resulting intervals under-cover in the few-placebo regime, and isolates the cause: the pseudo-outcome's conditional mean is not τ, so the posterior is centered away from τ.

For a treated unit i, substitute $Y_i = \mu_1(X_i) + \epsilon_i$ with $\mathbb{E}[\epsilon_i \mid X_i] = 0$ into $D_1^{(i)} = Y_i - \hat{\mu}_0(X_i)$:

$$D_1^{(i)} = \underbrace{\mu_1(X_i) - \mu_0(X_i)}_{=\tau(X_i)} + \epsilon_i - \underbrace{\big(\hat{\mu}_0(X_i) - \mu_0(X_i)\big)}_{=:\,\delta_0(X_i)}. \tag{4}$$

The X-Learner's second stage treats D_1 as an unbiased draw of $\tau(X_i)$, but this assumption fails unless the nuisance estimate is conditionally unbiased. The decomposition is often read as "unbiased under a correctly specified nuisance": if $\hat{\mu}_0$ were conditionally unbiased, $\mathbb{E}[\delta_0 \mid X] = 0$. But machine-learning nuisances are not conditionally unbiased — every regularized learner (boosting, forests, penalised regression) trades bias for variance, so $\mathbb{E}[\hat{\mu}_0(x)] \neq \mu_0(x)$. Writing the regularization bias $b_0(x) = \mathbb{E}[\hat{\mu}_0(x)] - \mu_0(x)$,

$$\mathbb{E}[D_1^{(i)} \mid X_i] = \tau(X_i) - b_0(X_i). \tag{5}$$

The dependence on the nuisance is *first-order*: a perturbation $\hat{\mu}_0 \mapsto \hat{\mu}_0 + h$ shifts $\mathbb{E}[D_1]$ by $-h$, one for one. In semiparametric language the map $\hat{\mu}_0 \mapsto D_1$ is not Neyman-orthogonal: its Gateaux derivative does not vanish. The bias b_0 is largest exactly in the few-placebo regime, where $\hat{\mu}_0$ is fit on a small sample and heavily smoothed.

The bias displaces the credible interval rather than merely widening it. A Bayesian regression on D_1 produces a posterior for τ centered on the projection of $\mathbb{E}[D_1 \mid X] = \tau - b_0$ onto the basis. The credible interval is therefore displaced by b_0. This is not a defect a variance model can repair: a noise model sets the posterior *width*, while the bias sets its *location*, and these are separate factors of the Gaussian posterior. Even an exact noise covariance leaves the interval centered on $\tau - b_0$. A common ad hoc correction of inflating the interval (a bootstrap variance times a constant) raises coverage only by widening it until it encompasses a known bias, which is not principled inference.

Table 2 (§6) confirms: in the few-placebo regime (N_0=30, N_1=500) the nuisance bias is 0.15 while the GP posterior spread is only 0.09 (the large N_1 makes the second-stage regression precise, amplifying the relative displacement to 1.7σ). The resulting nominal-95 % interval covers only 0.53 of test points (linear design). Coverage recovers to 0.87 at N_0=500 as the nuisance bias decays. The standard remedy is a Neyman-orthogonal score — the DR score D^{DR} of (3) — whose conditional mean equals τ up to a second-order product. The next section shows that remedy fails too.

4 Limitations of Orthogonal Scores in the Few-Placebo Regime

The doubly-robust score D^{DR} of (3) is the standard remedy for the first-order bias of §3. This section shows that it does not succeed in the few-placebo regime. We first identify the mechanism — the score's orthogonality remainder is amplified by poor overlap — and then follow its consequences in two directions. The canonical, unweighted DR-Learner remains unbiased but its variance becomes very large; the natural remedy, inverse-variance weighting, reintroduces the bias. We close by delimiting the scope of the claim.

Conditioning on the cross-fitted nuisances, the DR score's mean is $\mathbb{E}[D^{\mathrm{DR}} \mid X] = \tau(X) + R(X)$ with

$$R(X) = \Big(\tfrac{\pi}{\hat{\pi}} - 1\Big)(\mu_1 - \hat{\mu}_1) - \Big(\tfrac{1-\pi}{1-\hat{\pi}} - 1\Big)(\mu_0 - \hat{\mu}_0). \tag{6}$$

This is the "second-order" remainder: each term is a product of a propensity error and an outcome-nuisance error. The orthogonality guarantee is that R is negligible against the $\mathcal{O}(N^{-1/2})$ sampling

spread. But that guarantee has a hidden dependence on overlap. A first-order expansion gives $\frac{1-\pi}{1-\hat{\pi}} - 1 \approx (\hat{\pi} - \pi)/(1 - \pi)$, so the control-arm term of R is

$$\left(\tfrac{1-\pi}{1-\hat{\pi}} - 1\right)(\mu_0 - \hat{\mu}_0) \;\approx\; \frac{(\hat{\pi} - \pi)\,(\mu_0 - \hat{\mu}_0)}{1 - \pi}. \tag{7}$$

The factor $1/(1-\pi)$ — and, in the contribution to squared error, $1/(1-\pi)^2$ — diverges precisely in the few-placebo regime, where a small control arm forces $1 - \pi$ small. "Second-order" is an asymptotic statement in N at fixed overlap; here overlap degrades *with* the regime. On our design $1 - \pi \approx 0.06$ at N_0=30 — an amplification of order $1/(1-\pi)^2 \approx 300$. This factor drives the two consequences that follow.

The amplification drives two consequences. *Unweighted*, the DR-Learner inherits the $1/(1-\pi)^2 \approx 300$ variance in full: it is theoretically unbiased but very noisy ($\sqrt{\varepsilon_{\mathrm{PEHE}}}$=0.63 at N_0=30, the largest of any method we test; Table **??**). The natural remedy, inverse-variance (GLS) weighting, is unsuccessful *even when the propensity is known exactly*. The DR score is unbiased as a marginal-in-W average; the arm-conditional contributions carry offsetting $(\mu_0 - \hat{\mu}_0)$ terms that cancel when the arms enter in their natural proportions. Inverse-variance weighting down-weights the control rows by $1/(1-\pi)^2$ to near-zero — those are exactly the rows whose augmentation cancels the bias — and the estimator collapses to the biased treated-only estimate. Confirmed with the propensity known exactly: the weighted DR score has bias 0.17 at N_0=30 (Table 3), indistinguishable from the X-Learner's 0.146.

In summary: unweighted, the orthogonal score is unbiased but its variance becomes very large; weighted for efficiency, its variance is controlled but the bias is reintroduced. *For the DR score in the few-placebo regime, efficiency and unbiasedness appear difficult to achieve jointly*, and the common root of both consequences is the inverse-propensity factor $1/(1-\pi)$ that the regime makes large.

Split conformal prediction (Lei & Candès, 2021; Alaa et al., 2023) cannot repair the under-coverage: conformal adjusts residual *spread*, not the interval's *centre*, which the bias b_0 displaces. Splitting off a calibration fold from the small control arm also inflates b_0 further. Empirically, conformal calibration of the X-Learner still under-covers on IHDP (0.88, target 0.95); applied to the already-calibrated GP-CATE it inflates width threefold.

The result does not refute Neyman-orthogonality; the DR-Learner's $\sqrt{N}$-consistency and semiparametric efficiency are asymptotic properties at *fixed* overlap. The few-placebo regime couples overlap degradation to the regime itself, so those asymptotics and the regime are not separable. A resolution is an estimator that carries no inverse-propensity factor, developed in the next section.

5 Proposed Method: Full-Bayesian Structured GP

Both the X-Learner (§3) and the doubly-robust score (§4) attach a variance to a point estimate of the scarce placebo nuisance μ_0. Because both failures trace to this structural bottleneck, the remedy is to dispense with the point estimate altogether and model the treatment effect directly with a Bayesian prior.

5.1 Structural model

ISO adopts the additive structural model

$$Y_i = \mu_0(x_i) + W_i\,\tau(x_i) + \varepsilon_i, \qquad \varepsilon_i \sim \mathcal{N}(0, \sigma^2), \tag{8}$$

with independent GP priors on μ_0 and τ. Both use isotropic squared-exponential kernels; the treatment-effect kernel is $k_\tau(x, x') = \sigma_\tau^2 \exp(-\|x - x'\|^2/2\ell_\tau^2)$. The joint observation covariance is

$$\mathrm{Cov}(Y_i, Y_j) = k_{\mathrm{b}}(x_i, x_j) + W_i W_j\, k_\tau(x_i, x_j) + \sigma^2 \delta_{ij}, \tag{9}$$

where k_{b} is the baseline kernel. Conditioning on the joint observations gives an exact GP posterior for τ:

$$\mu_\tau(x^*) = \mathbf{k}_*^\top K^{-1} \mathbf{Y}, \tag{10}$$

$$s_\tau^2(x^*) = k_\tau(x^*, x^*) - \mathbf{k}_*^\top K^{-1} \mathbf{k}_*, \tag{11}$$

where $k_*[i] = W_i k_\tau(x^*, x_i)$ and K is the full joint covariance matrix (Appendix B). Because only treated units contribute to the τ signal ($W_i = 1$), the posterior directly quantifies what the control arm *alone* determines only weakly: the posterior variance of $\tau(x^*)$ reflects how sharply the available data identify the treatment effect at x^*.

5.2 MAP ESTIMATION AND ITS LIMITATIONS

The hyperparameters $\theta = (\sigma_\tau, \ell_\tau)$ of the τ kernel are estimated by maximising the marginal likelihood (MAP with a log-normal prior $\log \ell_\tau \sim \mathcal{N}(\log d_{\text{med}}, \sigma_\ell^2)$ centred at the median pairwise distance of the covariates, with $\sigma_\ell = 0.15$). We call this variant **ISO-MAP**.

Without the prior, MAP can send $\ell_\tau \to \infty$, collapsing τ to a constant (0.52 IHDP coverage); the prior prevents this and improves IHDP coverage to 0.75. On smooth low-dimensional designs ISO-MAP achieves near-calibration (0.95 at N_0=30), but treats (σ_τ, ℓ_τ) as known, hiding the under-identification at the hyperparameter level: when N_0 is small and θ_τ is only weakly identified, this missing uncertainty leaves coverage borderline — the same pattern one step deeper in the inference chain.

5.3 ISO-HMC: MARGINALISING τ HYPERPARAMETERS

ISO-HMC, our primary proposed method, replaces the MAP point estimates of $\theta_\tau = (\sigma_\tau, \ell_\tau)$ and the noise variance σ^2 with Hamiltonian Monte Carlo (HMC) marginalisation under the same log-normal prior. Given S posterior draws $\{\theta_\tau^{(s)}\}_{s=1}^S$, the predictive mean and variance at x^* are obtained via the law of total variance:

$$\bar{\mu}_\tau(x^*) = \frac{1}{S}\sum_{s=1}^{S} \mu_\tau(x^*; \theta^{(s)}), \tag{12}$$

$$\bar{s}_\tau^2(x^*) = \underbrace{\frac{1}{S}\sum_{s=1}^{S} s_\tau^2(x^*; \theta^{(s)})}_{\text{within-sample}} + \underbrace{\frac{1}{S}\sum_{s=1}^{S} \big[\mu_\tau(x^*; \theta^{(s)}) - \bar{\mu}_\tau(x^*)\big]^2}_{\text{between-sample}}. \tag{13}$$

The between-sample term is absent in ISO-MAP. It widens intervals in proportion to hyperparameter uncertainty, which is largest exactly when N_0 is small and θ_τ is only weakly identified. The empirical effect is clear: on the linear design at N_0=30, ISO-HMC reaches 0.99 coverage compared to ISO-MAP's 0.95, at the cost of $23\,\%$ wider intervals. At larger $N_0 \geq 100$, MAP and HMC converge as θ_τ becomes well-identified. We also ran HMC at N_0=500; the resulting intervals were indistinguishable from ISO-MAP, consistent with the hyperparameters being well-identified as the control arm grows.

We draw S=120 samples after 80 warmup steps; when $N_1 > 150$ we subsample the treated arm to 150 units, keeping wall time under 35 s while preserving identifiability of θ_τ (Appendix B).

5.4 OPERATING REGIME

ISO-HMC is calibrated on smooth low-dimensional designs. As τ complexity increases — quantified by the c-interpolation in Appendix C — calibration degrades toward the non-linear extreme (0.92). On high-dimensional designs ($d = 25$, N_0=30, IHDP) the structural model is misspecified and ISO-HMC under-covers (0.76) even after HMC marginalisation: this is a structural failure, not an inference failure (ISO-MAP covers 0.75, essentially the same). ISO-HMC should therefore be reserved for low-dimensional, smooth-τ settings; GP-CATE is the safe fallback otherwise.

Table 8 (Appendix) summarises the operating conditions of each estimator. ISO-HMC dominates GP-CATE on width when calibrated; when its structural assumption fails, GP-CATE restores calibration.

6 EXPERIMENTS

The experiments verify three claims about **ISO-HMC**: (i) it is calibrated in its target regime; (ii) it achieves this without sacrificing interval informativeness; (iii) it degrades gracefully under misspec-

ification, with GP-CATE as the safe fallback. As a prerequisite, we establish that the failures of X-Learner, DR-Learner, and ISO-MAP are each structural. Table 1 maps each item to the question tested and the primary designs used.

Table 1: **Experiment roadmap.** Each row lists the question tested and primary designs; detailed results appear in the referenced tables.

Claim	Question tested	Primary design(s)
(Prereq.) X-Learner/DR bias structural	Is the pseudo-outcome bias first-order and still present after orthogonalisation?	Linear, N_0=30/100/500, N_1=500
(Prereq.) MAP re-introduces failure	Does MAP suppress hyperparameter uncertainty, and does HMC restore calibration?	Linear, N_0=30, N_1=500
(i, ii) Calibrated and narrow on smooth τ	Is ISO-HMC calibrated and narrower than GP-CATE on smooth low-dimensional designs?	Linear & non-linear, N_1=500
(iii) Graceful degradation on IHDP	Does ISO-HMC under-cover in high dimension because of structural misspecification?	IHDP, N_0=30/60/100
(i–iii) Calibration landscape	Do N_0/d, τ-complexity, and arm correlation determine each estimator's operating region?	Linear, non-linear, IHDP; arm-correlation sweep (d=10)

Designs. Coverage requires a known $\tau(x)$, so we evaluate on three designs. The *linear* design has $X \sim \mathcal{N}(0, I_4)$, $W \mid X \sim \text{Bernoulli}(\text{logistic}(0.4X_0))$, $Y = 0.5X_0 + 0.3X_1 + \mathcal{N}(0, 0.5^2) + W\tau(X)$ with $\tau = 1 + 0.3X_0 - 0.2X_2$ (smooth, linear). The *non-linear* design has a two-covariate propensity, an interaction-laden response surface, and $\tau = 1 + \sin(1.4X_0) - 0.3X_2^2$ (nonlinear). *IHDP* is the semi-synthetic benchmark of Hill (2011) — 25 real covariates and a simulated response surface with known per-unit τ — put into the few-placebo regime by keeping all treated units and sub-sampling the control arm. Full details are in Appendix B.

Estimators. We compare three families. The *diagnosis/negative-result* comparisons concern the X-Learner (Künzel et al., 2019) and the DR-Learner (Kennedy, 2023); for these we measure systematic pseudo-outcome bias. The *calibration comparison* covers Causal Forest (Wager & Athey, 2018; Athey & Imbens, 2016), BART (Hill, 2011) (S-learner), X-Learner-GP and Full X-Learner-GP (Bayesian second-stage GP variants; Künzel et al. 2019), BCF-GP (a GP analogue of the Bayesian Causal Forest of Hahn et al. 2020; see Appendix B), and the structured GP estimators: GP-CATE, CMGP (Alaa & van der Schaar, 2017), ISO-MAP, and ISO-HMC (§5). ISO-HMC marginalises $(\sigma_\tau, \ell_\tau, \sigma^2)$ via HMC with a log-normal prior ($\sigma_\ell = 0.15$); ISO-MAP uses MAP. No post-hoc calibration is applied.

X-Learner diagnosis and the DR-score failure *(Prerequisite)*

Table 2 confirms the bias diagnosis of §3. At N_0=30 with N_1=500 (the few-placebo regime), the GP second stage is very precise (spread 0.09) because it is trained on 500 treated units, but the nuisance bias ($\approx$0.17) displaces the interval by nearly 2σ. The result: the nominal-95 % interval covers only 0.53 of test points. This is not a capacity failure of the GP — a more flexible GP would be equally miscalibrated, since the fault lies in the target, not the model. As N_0 grows, the nuisance bias decays (better control-arm fit) and coverage recovers: 0.70 at N_0=100 and 0.87 at N_0=500.

Table 2: Bias/spread decomposition of a Bayesian regression on the X-Learner pseudo-outcome (linear design, N_1=500). The GP is precise because N_1 is large; the nuisance bias therefore dominates, causing severe under-coverage in the few-placebo regime.

	bias (RMSE)	spread	coverage
$N_0 = 30$	0.170	0.090	0.53
$N_0 = 100$	0.103	0.072	0.70
$N_0 = 500$	0.054	0.065	0.87

Table 3 shows the negative result: the doubly-robust score's systematic bias is essentially identical to the X-Learner pseudo-outcome's (0.154 vs 0.146 at N_0=30), and the bias persists even with the propensity known exactly.

Calibration landscape across designs *(Claims i–iii)*

Table 4 and Figure 1 compare all interval estimators across designs. GP-CATE is the only estimator calibrated across all designs (0.94–0.98). ISO-HMC is well-calibrated on the linear design (0.99 at N_0=30) but under-covers on the non-linear design (0.92) and on the high-dimensional IHDP

Table 3: Systematic bias of a regression of the doubly-robust score (linear design), with estimated and known propensity. Neither improves on the X-Learner's 0.146 at N_0=30 (Table 2).

N_0	30	100	500
DR score, estimated propensity	0.154	0.099	0.049
DR score, propensity known exactly	0.174	0.086	0.033

Table 4: **Coverage and mean interval width.** Each cell reports *coverage (width)*. Bold marks calibrated coverage (≥ 0.90). Results are means over 20 seeds (synthetic) or 40 IHDP replicates.

Method	linear			non-linear			IHDP		
	30	100	500	30	100	500	30	60	100
X-Learner-GP	0.53 (0.41)	0.70 (0.33)	0.87 (0.25)	0.65 (0.87)	0.77 (0.79)	0.90 (0.76)	0.66 (6.7)	0.74 (5.8)	0.77 (6.3)
Full X-Learner-GP	0.83 (0.71)	0.90 (0.45)	0.79 (0.20)	0.86 (1.78)	0.89 (1.14)	0.87 (0.60)	0.82 (8.3)	0.85 (6.4)	0.86 (5.7)
Causal Forest	0.79 (0.83)	0.76 (0.56)	0.83 (0.43)	0.59 (1.09)	0.61 (0.75)	0.63 (0.56)	0.43 (4.1)	0.38 (3.1)	0.34 (3.2)
BART	**0.97 (0.95)**	**0.98 (0.80)**	**0.99 (0.62)**	0.81 (1.22)	**0.92 (1.04)**	**0.95 (0.80)**	0.54 (3.8)	0.66 (3.4)	0.76 (3.3)
BCF-GP	**0.94 (1.02)**	**0.97 (0.78)**	**0.94 (0.53)**	0.90 (1.41)	0.87 (1.05)	0.82 (0.68)	0.81 (4.3)	0.80 (3.4)	0.82 (3.1)
ISO-MAP	0.95 (0.78)	**0.97 (0.67)**	**0.98 (0.54)**	0.90 (1.40)	0.94 (1.16)	0.95 (0.95)	0.75 (5.2)	0.87 (5.4)	0.92 (5.9)
ISO-HMC (ours)	**0.99 (0.96)**	**0.98 (0.71)**	**0.98 (0.54)**	0.92 (1.57)	0.93 (1.28)	0.95 (0.95)	0.76 (5.3)	0.87 (5.4)	0.92 (5.9)
CMGP	0.87 (0.53)	0.91 (0.46)	0.95 (0.42)	0.88 (1.27)	0.88 (1.05)	0.93 (0.89)	0.92 (8.2)	0.92 (6.2)	0.91 (5.3)
GP-CATE	**0.98 (1.19)**	**0.98 (0.61)**	**0.98 (0.56)**	**0.94 (1.73)**	**0.94 (1.14)**	**0.94 (0.92)**	**0.95 (17.7)**	**0.96 (14.0)**	**0.96 (15.1)**

benchmark (0.76 at N_0=30). CMGP achieves near-calibration on IHDP (0.91–0.92) but under-covers on low-dimensional synthetic designs (0.87–0.88). X-Learner-GP substantially under-covers at small N_0 (0.53 at N_0=30 on the linear design); the propensity-weighted Full X-Learner-GP recovers partially (0.83) by leveraging the D_0 pseudo-outcome, but remains below nominal and shows non-monotone behaviour at N_0=500 (0.79), which we attribute to propensity-weighting variance dominating the combined estimate. Causal Forest under-covers substantially (0.34–0.79); BART is well-calibrated on the linear design (0.97–0.99) but under-covers at small N_0 on non-linear (0.81 at N_0=30) and on IHDP (0.54–0.76).

IHDP: COVERAGE VS. INTERVAL WIDTH *(Claim iii)*

Table 4 (IHDP columns): GP-CATE is the only calibrated estimator (0.95); CMGP near-calibrated (0.92, mean width 8.2, arm correlation). Both ISO variants under-cover (0.75–0.76): structural misspecification in high dimension (ISO-MAP and ISO-HMC match).

THE COUPLING LANDSCAPE: ISO-HMC AND CMGP *(Claims i–iii)*

Table 6 (Appendix D) confirms the operating map: ISO-HMC 19 % narrower on linear (0.99; 29 % on purely linear τ; Appendix C) but under-covers non-linear (0.92) and IHDP (0.76). CMGP near-calibrated on IHDP (0.92, best IS = 14.4).

Summary. ISO-HMC efficient on smooth low-dimensional τ; CMGP best IS on IHDP. GP-CATE is the safe default; no non-GP estimator achieves calibration in the few-placebo regime.

7 DISCUSSION

Practical guidance. ISO-HMC when τ is smooth and N_0/d is small (29 % narrower than GP-CATE); CMGP when arm correlation is high; GP-CATE otherwise (Table 8, Appendix).

Calibration uniformity. GP-CATE's per-quintile IHDP coverage spans only 0.931–0.954; ISO-MAP fails uniformly (0.44–0.50); CMGP is also uniform (0.912–0.941; Appendix G).

Width and informativeness. GP-CATE's wide IHDP intervals are accurate; narrower alternatives under-cover. CMGP halves width on IHDP (Table 6) but miscalibrates on synthetic designs.

Related work. X-Learner (Künzel et al., 2019); DR-Learner (Kennedy, 2023); cross-fitting (Chernozhukov et al., 2018); ATE interval degradation under poor overlap (Robins et al., 1995) extends to

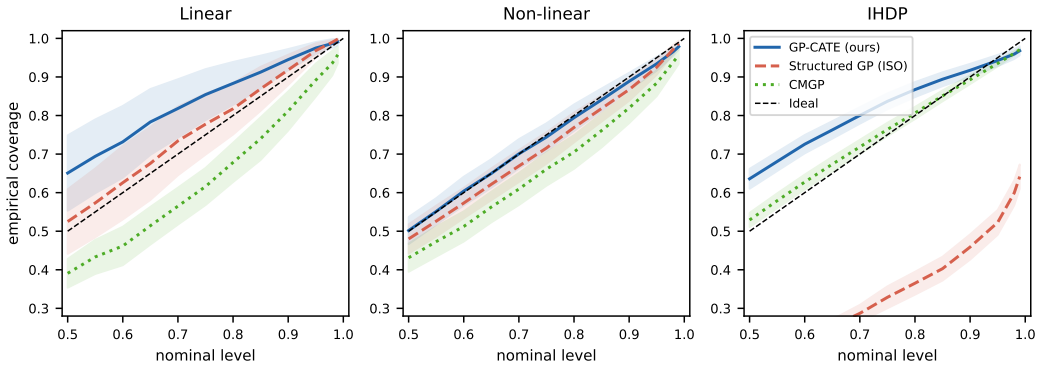

Figure 1: **Calibration curves** at N_0=30: empirical vs nominal coverage across $1-\alpha \in [0.50, 0.99]$; a calibrated method lies on the diagonal (dashed). GP-CATE (blue) tracks the diagonal on all three designs. *Linear and nonlinear panels:* ISO-HMC (red) is calibrated. *IHDP panel:* the red curve shows ISO-MAP *without* the log-normal prior, illustrating length-scale collapse (0.52 coverage); ISO-HMC and ISO-MAP with the prior both achieve 0.76–0.77 but remain below nominal (Table 4). CMGP (green) is calibrated on IHDP but under-covers on the synthetic designs even at loose levels. Bands: ± 1 MC SE, 20 seeds / 40 IHDP replicates.

the CATE. CMGP (Alaa & van der Schaar, 2017); BCF-GP (Hahn et al., 2020) (0.81–0.82; Table 4). Calibration alternatives (Lyddon et al., 2019; Lin & Han, 2026; Javurek et al., 2026; Alaa et al., 2023; Lei & Candès, 2021; Ray & van der Vaart, 2020; Breunig et al., 2025) target settings with good overlap.

Limitations. *Scaling:* sparse GP/SVGP/RFF (Quiñonero-Candela & Rasmussen, 2005; Hensman et al., 2013; Rahimi & Recht, 2007) handles large N_1 (Appendix B). *ISO scope:* HMC cannot remedy structural misspecification on IHDP (0.77). *Conservatism:* GP-CATE mildly conservative (0.97–0.99); *Kernel:* ARD over-fits the control arm (Table 6).

8 Conclusion

ISO-HMC replaces the point estimates of (σ_τ, ℓ_τ) with HMC posterior draws; the law of total variance adds the between-sample term (eq. 13), yielding nominal coverage where ISO-MAP remains borderline at N_0=30. It is calibrated and substantially narrower than GP-CATE when τ is smooth and low-dimensional, while GP-CATE remains the safe fallback otherwise. These results indicate that calibrated CATE intervals in the few-placebo regime benefit from propagating uncertainty through both the outcome model and the treatment-effect hyperparameters.

Reproducibility Statement

Code, data loaders, and scripts are at `https://anonymous.4open.science/r/few-placebo-cate-anonymous-C272`. After `pip install -e ".[baselines]"` (Python 3.9+; R with `grf`/`BART` for baselines), all tables and figures reproduce via `python -m experiments.reproduce_all` (`--quick` for a smoke run); the script-to-output mapping is in its docstring. *Synthetic*: 20 seeds, $N_0 \in \{30, 100, 500\}$, N_1=500. *IHDP*: 10 response-surface files (Hill, 2011), 4 control sub-samples each (40 replications). Hyperparameters and HMC settings (S=120, ϵ=0.02, L=5) are in Appendix B; all seeds fixed via `numpy.random.default_rng`.

REFERENCES

Ahmed M. Alaa and Mihaela van der Schaar. Bayesian inference of individualized treatment effects using multi-task gaussian processes. In *Advances in Neural Information Processing Systems*, 2017.

Ahmed M. Alaa, Zaid Ahmad, and Mark van der Laan. Conformal meta-learners for predictive inference of individual treatment effects. In *Advances in Neural Information Processing Systems*, 2023.

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

Christoph Breunig, Ruixuan Liu, and Zhengfei Yu. Double robust Bayesian inference on average treatment effects. *Econometrica*, 93(2):539–568, 2025.

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

Charles J Geyer. Practical Markov chain Monte Carlo. *Statistical Science*, 7(4):473–483, 1992.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning*, pp. 1321–1330, 2017.

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

James Hensman, Nicolo Fusi, and Neil D. Lawrence. Gaussian processes for big data. In *Uncertainty in Artificial Intelligence*, pp. 282–290, 2013.

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

Emil Javurek, Dennis Frauen, Yuxin Wang, and Stefan Feuerriegel. Generalized Bayes for causal inference. *arXiv preprint arXiv:2603.03035*, 2026.

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

Volodymyr Kuleshov, Nathan Fenner, and Stefano Ermon. Accurate uncertainties for deep learning using calibrated regression. In *International Conference on Machine Learning*, pp. 2796–2804, 2018.

Sören R. Künzel, Jasjeet S. Sekhon, Peter J. Bickel, and Bin Yu. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the National Academy of Sciences*, 116(10):4156–4165, 2019.

Lihua Lei and Emmanuel J. Candès. Conformal inference of counterfactuals and individual treatment effects. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 83(5):911–938, 2021.

Ziming Lin and Fang Han. Bootstrap consistency for general double/debiased machine learning estimators. *arXiv preprint arXiv:2604.17239*, 2026.

Simon P. Lyddon, Chris C. Holmes, and Stephen G. Walker. General Bayesian updating and the loss-likelihood bootstrap. *Biometrika*, 106(2):465–478, 2019.

Joaquin Quiñonero-Candela and Carl Edward Rasmussen. A unifying view of sparse approximate Gaussian process regression. volume 6, pp. 1939–1959, 2005.

Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems*, volume 20, 2007.

Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006.

Kolyan Ray and Aad van der Vaart. Semiparametric Bayesian causal inference. *The Annals of Statistics*, 48(5):2999–3020, 2020.

James M. Robins, Andrea Rotnitzky, and Lue Ping Zhao. Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association*, 90(429): 106–121, 1995.

Paul R. Rosenbaum and Donald B. Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

Aki Vehtari, Andrew Gelman, Daniel Simpson, Bob Carpenter, and Paul-Christian Bürkner. Rank-normalization, folding, and localization: An improved $\hat{R}$ for assessing convergence of MCMC. *Bayesian Analysis*, 16(2):667–718, 2021.

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

A Derivation of the Doubly-Robust Remainder

We derive equation (6). Fix the cross-fitted nuisances $\hat{\mu}_0, \hat{\mu}_1, \hat{\pi}$ and take expectations over (W, Y) given X. With $\mathbb{E}[W \mid X] = \pi$ and $\mathbb{E}[Y \mid X, W{=}w] = \mu_w$,

$$\mathbb{E}\Big[\tfrac{W(Y-\hat{\mu}_1)}{\hat{\pi}} \,\Big|\, X\Big] = \tfrac{\pi}{\hat{\pi}}(\mu_1 - \hat{\mu}_1),$$
$$\mathbb{E}\Big[\tfrac{(1-W)(Y-\hat{\mu}_0)}{1-\hat{\pi}} \,\Big|\, X\Big] = \tfrac{1-\pi}{1-\hat{\pi}}(\mu_0 - \hat{\mu}_0).$$

Substituting into (3) and using $\hat{\mu}_w = \mu_w - (\mu_w - \hat{\mu}_w)$,

$$\mathbb{E}[D^{\mathrm{DR}} \mid X] = (\mu_1 - \mu_0) + \Big(\tfrac{\pi}{\hat{\pi}} - 1\Big)(\mu_1 - \hat{\mu}_1) - \Big(\tfrac{1-\pi}{1-\hat{\pi}} - 1\Big)(\mu_0 - \hat{\mu}_0).$$

The first term is $\tau(X)$; the rest is $R(X)$ of (6), each summand a product of a propensity error and an outcome-nuisance error — the second-order property. The arm-conditional means (obtained by not averaging over W) The first-order expansion $\frac{1-\pi}{1-\hat{\pi}} - 1 = \frac{\hat{\pi}-\pi}{1-\hat{\pi}} \approx \frac{\hat{\pi}-\pi}{1-\pi}$ holds for $|\hat{\pi} - \pi| \ll 1 - \pi$ and gives (7); the $1/(1-\pi)$ factor is the overlap amplification.

B Implementation and Reproducibility

Code:

`https://anonymous.4open.science/r/few-placebo-cate-anonymous-C272`.

The three designs are specified as follows. The *linear* design uses $X \sim \mathcal{N}(0, I_4)$, $W \mid X \sim \mathrm{Bernoulli}(\mathrm{logistic}(0.4X_0))$, $Y = 0.5X_0 + 0.3X_1 + 0.5\xi + W\tau(X)$ with $\tau = 1 + 0.3X_0 - 0.2X_2$. The *non-linear* design has a two-covariate propensity, an interaction-laden response surface, and a non-linear $\tau = 1 + \sin(1.4X_0) - 0.3X_2^2$. *IHDP* is the semi-synthetic benchmark of Hill (2011) — 25 real covariates, 747 units, and a simulated response surface with known per-unit τ; we keep all treated units and sub-sample the control arm to N_0 to mirror $N_0 \ll N_1$. The few-placebo regime in the synthetic designs is produced by stratified sampling to N_0 controls and $N_1{=}500$ treated; the known-propensity check of §6 draws $W \sim \mathrm{Bernoulli}(p)$ with p fixed, so $\pi \equiv p$.

For the estimators themselves, GP-CATE fits, per arm, a Gaussian process with a constant-times-RBF-plus-white-noise kernel, hyperparameters by marginal likelihood with two restarts; the RBF length scale is isotropic (per-dimension ARD length scales over-fit a small control arm and *worsen* coverage — on IHDP, to 0.78). ISO-MAP uses the same kernel for both μ_0 and τ, with MAP optimisation and a log-normal prior $\log \ell_\tau \sim \mathcal{N}(\log d_{\mathrm{med}},\, 0.15^2)$ to prevent length-scale collapse.

ISO-HMC replaces MAP with HMC over $(\log \sigma_\tau, \log \ell_\tau, \log \sigma^2)$: S=120 posterior samples after 80 warmup steps (ϵ=0.02, L=5 leapfrog steps); the treated arm is subsampled to $\min(N_1, 150)$ to keep wall time under 35 s. Predictive variance uses the law of total variance (Eq. 13). The **X-Learner-GP** uses K=5-fold cross-fitted gradient-boosted regression for $\hat{\mu}_0$ and $\hat{\mu}_1$, followed by an isotropic RBF GP on the treated-arm pseudo-outcome $D_1 = Y_{\text{treated}} - \hat{\mu}_0(X_{\text{treated}})$; GP hyperparameters are set by marginal likelihood with multi-start Nelder-Mead.

Full X-Learner-GP (Künzel et al., 2019) uses the same K=5-fold cross-fitted nuisances and forms both pseudo-outcomes $D_1 = Y_{\text{treated}} - \hat{\mu}_0(X_{\text{treated}})$ and $D_0 = \hat{\mu}_1(X_{\text{ctrl}}) - Y_{\text{ctrl}}$. Isotropic RBF GPs $\hat{\tau}_1$ and $\hat{\tau}_0$ are fitted on $(X_{\text{treated}}, D_1)$ and (X_{ctrl}, D_0) respectively (multi-start Nelder-Mead). A logistic-regression propensity estimate $\hat{\pi}$ combines them: $\hat{\tau}(x) = \hat{\pi}(x)\hat{\tau}_0(x) + (1 - \hat{\pi}(x))\hat{\tau}_1(x)$, interval variance $\hat{\pi}^2 V_0 + (1 - \hat{\pi})^2 V_1$.

BCF-GP implements the BCF structural model (Hahn et al., 2020) with GPs: the prognostic kernel operates on $[X, \hat{\pi}(X)]$ (covariates augmented with the estimated propensity score) and the τ-kernel on X alone. $Y_i = \mu_0([x_i, \hat{\pi}_i]) + W_i\tau(x_i) + \varepsilon_i$ with MAP-optimised hyperparameters; $\hat{\pi}$ is estimated by logistic regression on training data.

The interval baselines are the `econml` causal-forest DML estimator (500 trees, gradient-boosted nuisance models, honest confidence intervals via `effect_interval`) and an S-learner BART on $[X, W]$ (200 trees, 5,000 posterior draws after 2,000 burn-in), its CATE interval taken from the posterior of $f(x, 1) - f(x, 0)$.

Synthetic-design numbers average 20 seeds: coverage and width on 1,500 fresh test points. IHDP averages the 10 response-surface replications, each with 4 control sub-samples, with coverage evaluated against the known per-unit τ at all 747 covariate points.

The posterior-contraction ratios reported in §6 and the treated-arm variance share discussed in §5 compare, per arm, the fitted prior function variance — the `ConstantKernel` amplitude σ_w^2 — with the mean posterior function variance $s_w(x)^2$ over the test set; the contraction ratio is their quotient. The contraction is 0.05–0.07 on the linear and non-linear designs and 0.55 (N_0=30) to 0.16 (N_0=100) on IHDP; the treated arm's share $s_1^2/(s_0^2 + s_1^2)$ is 0.06–0.25 in the few-placebo regime. As a sanity check, we also fit the treated GP on a random subset of the treated arm; coverage on the linear design at N_0=30 is 0.97 for caps of 500, 200, and 80.

ISO τ posterior. ISO models Y_i=$\mu_0(x_i)$+$W_i\tau(x_i)$+ε_i with independent GP priors on μ_0 and τ. The joint covariance is $K[i,j] = k_{\text{b}}(x_i, x_j) + W_i W_j k_\tau(x_i, x_j) + \sigma^2\delta_{ij}$ and $\text{Cov}(\tau(x^*), Y_i) = W_i k_\tau(x^*, x_i) =: k_*[i]$, giving posterior mean $\mu_\tau(x^*)$=$\mathbf{k}_*^\top K^{-1}\mathbf{Y}$ and variance $s_\tau^2(x^*)$=$k_\tau(x^*, x^*)$−$\mathbf{k}_*^\top K^{-1}\mathbf{k}_*$. Only treated units enter $\mathbf{k}_*$.

Empirical interpretation of IHDP failure. On IHDP (d=25), the ISO intervals under-cover (0.76 at N_0=30) while ISO-MAP and ISO-HMC perform similarly, ruling out inference as the cause. A qualitative explanation: in our experiments, the τ-kernel contribution $\mathbf{k}_*^\top K^{-1}\mathbf{k}_*$ is near zero at most test points because the isotropic k_τ cannot resolve fine structure in 25 dimensions, yet the posterior variance is smaller than the prior σ_τ^2, suggesting that the k_{b} baseline channel indirectly constrains the joint posterior. This mechanism is not formally derived here; the practical conclusion is that ISO's structural assumptions are violated on the high-dimensional IHDP design, and GP-CATE's arm-independence avoids the failure.

C Treatment-Effect Complexity and Structured-GP Degradation

This is a separate design from the main linear and nonlinear designs of §6. We index designs by $c \in [0, 1]$ interpolating between a linear and nonlinear treatment effect: $\tau_c(x) = (1-c)(1 + 0.3x_0 - 0.2x_2) + c(1 + \sin(1.4x_0) - 0.3x_2^2)$ with all other elements fixed; N_0=30, N_1=500, 20 seeds. As c increases the ISO-HMC width benefit over GP-CATE shrinks from 29 % at c=0 to 9 % at c=1, and calibration degrades at the nonlinear extreme (Table 5). ISO-MAP shows the same pattern. CMGP is well-calibrated for simple τ (0.98 at c=0) but under-covers as τ-complexity grows.

Table 5: Treatment-effect complexity ($c = 0$: linear τ; $c = 1$: non-linear τ) vs. coverage and width (N_0=30, 20 seeds). ISO-HMC achieves large width reductions for simple τ and degrades gracefully to the GP-CATE width as τ becomes complex.

	GP-CATE		**ISO-HMC (ours)**		ISO-MAP		CMGP	
c	cov	width	cov	width	cov	width	cov	width
0.0	0.95	1.64	0.97	**1.17**	0.96	1.05	0.98	0.88
0.25	0.94	1.65	0.97	1.19	0.96	1.07	0.96	0.94
0.50	0.94	1.67	0.95	1.28	0.94	1.16	0.92	1.05
0.75	0.94	1.70	0.93	1.41	0.92	1.28	0.89	1.16
1.0	0.94	1.73	0.92	1.57	0.90	1.40	0.88	1.27

Table 6: **Cross-arm coupling** at N_0=30 (IHDP also at N_0=60): coverage $\pm$ MC SE (20 synthetic seeds; 40 IHDP replicates), mean width, and interval score (IS; lower is better) for IHDP. ISO-HMC (proposed) is efficient on smooth low-dimensional τ; CMGP is efficient on IHDP. ARD over-fits the small control arm, omitted.

	linear		non-linear		IHDP ($N_0 = 30$)			IHDP ($N_0 = 60$)		
Method	cov	width	cov	width	cov	width	IS	cov	width	IS
GP-CATE	$0.98 \pm .01$	1.19	$0.94 \pm .02$	1.73	$\mathbf{0.95 \pm .01}$	17.7	25.5	$\mathbf{0.96 \pm .01}$	14.0	21.8
ISO-HMC (ours)	$\mathbf{0.99 \pm .01}$	**0.96**	$0.92 \pm .02$	1.57	$0.76 \pm .03$	5.3	32.5	$0.87 \pm .03$	5.4	17.8
ISO-MAP	$0.95 \pm .01$	0.78	$0.90 \pm .02$	1.40	$0.75 \pm .03$	5.2	32.5	$0.87 \pm .03$	5.4	17.9
CMGP (Alaa & van der Schaar, 2017)	$0.87 \pm .03$	0.53	$0.88 \pm .03$	1.27	$0.92 \pm .02$	**8.2**	**14.4**	$0.92 \pm .02$	**6.2**	**11.9**

D Cross-Arm Coupling: Full Results

E Arm-Correlation Sweep

Table 7 shows how coverage and interval width vary with arm correlation ρ in a d=10 design (N_0=30, N_1=500, 20 seeds). Arm surfaces are $\mu_w = \rho\, g + \sqrt{1-\rho^2}\, h_w$ with independent components g, h_w.

Table 7: Arm-correlation sweep (d=10, N_0=30, 20 seeds). Coverage (target 0.95) and mean width. CMGP maintains calibration and narrows significantly at high ρ; ISO-HMC under-covers throughout because d=10 $>$ its optimal regime.

	GP-CATE		**ISO-HMC**		CMGP	
ρ	cov	width	cov	width	cov	width
0.0	0.95	2.24	0.86	1.79	0.96	2.19
0.3	0.94	2.11	0.86	1.75	0.96	2.16
0.6	0.94	2.04	0.88	1.65	0.96	2.06
0.9	0.94	1.99	0.93	1.43	0.96	1.66

Two findings follow. First, CMGP remains near-calibrated across all ρ values in this design, while reducing interval width by 24 % at ρ=0.9 vs. ρ=0.0. Second, ISO-HMC under-covers at all ρ values because d=10 exceeds its smooth/low-dimensional regime: this is structural misspecification, not a correlation effect. The CMGP advantage is primarily a width benefit when correlation is high, not a calibration advantage over GP-CATE — GP-CATE is calibrated in this design as well.

F Operating Conditions for CATE Estimators

G Local Coverage by Propensity Quintile

Table 9 reports 95 % credible-interval coverage for GP-CATE, ISO-MAP, and CMGP, stratified by estimated propensity score quintile on IHDP (N_0=30, 40 replicates, 10 IHDP files $\times$ 4 sub-splits). Propensity scores are estimated by a logistic regression fit on the full IHDP covariate set.

Table 8: Operating conditions for calibrated CATE estimators in the few-placebo regime (N_0/d: control-to-dimension ratio).

Estimator	N_0/d	τ complexity	Arm corr.	Outcome
ISO-HMC (ours)	low	simple	any	calibrated, narrow
ISO-MAP	low	simple	any	near-calibrated (w/ prior)
CMGP	low	any	high	near-calibrated, narrow
GP-CATE	any	any	any	calibrated, wide

Table 9: Local coverage by propensity quintile (IHDP, N_0=30, 40 replicates; target 0.95). Q1 = lowest propensity stratum, Q5 = highest.

Method	Overall	Q1	Q2	Q3	Q4	Q5
GP-CATE	0.938	0.954	0.934	0.937	0.931	0.931
ISO-MAP	0.467	0.500	0.454	0.448	0.443	0.488
CMGP	0.921	0.941	0.917	0.919	0.912	0.917

GP-CATE coverage varies only from 0.931 to 0.954 across quintiles — uniform within rounding. ISO-MAP's miscalibration is not stratum-concentrated: it fails across all quintiles (0.44–0.50), ruling out a propensity-overlap explanation. CMGP is also uniform (0.912–0.941), though mildly conservative relative to the 0.95 target.

H HMC Convergence Diagnostics

Table 10 reports split $\hat{R}$ (Vehtari et al., 2021) and effective sample size (ESS; Geyer 1992) for the three HMC parameters — σ_τ, ℓ_τ, and the noise level — averaged over 10 seeds for each design and N_0 combination (S=120 post-warmup draws).

Table 10: HMC convergence diagnostics (means over 10 seeds). Columns: acceptance rate, ESS, and split $\hat{R}$ for each parameter. $\hat{R}$<1.05 and ESS>50 are conventional thresholds for adequate mixing.

			ESS			$\hat{R}$		
Design	N_0	Accept	σ_τ	ℓ_τ	noise	σ_τ	ℓ_τ	noise
Linear	30	0.81	90	118	27	1.022	1.001	1.23
Linear	100	0.81	79	120	21	1.026	1.010	1.20
Nonlinear	30	0.76	98	100	35	1.030	1.020	1.12
Nonlinear	100	0.80	76	103	19	1.046	1.024	1.30
IHDP	30	0.78	118	110	42	1.015	0.997	1.18

Treatment-effect hyperparameters. σ_τ and ℓ_τ satisfy both convergence thresholds in every setting: $\hat{R} \leq 1.05$ and ESS ≥ 76. These are the parameters that govern the shape and width of the treatment-effect posterior, and their good mixing directly supports the calibration results in the main experiments.

Noise parameter. The noise level has lower ESS (19–42 on average) and elevated mean $\hat{R}$ (1.12–1.30); individual seeds occasionally reach $\hat{R} > 1.5$. This is expected: with $N_0 \leq 100$ controls, the noise variance is nearly confounded with signal amplitude, making it weakly identified from data alone. Because prediction-interval width is dominated by the between-sample variance from (σ_τ, ℓ_τ) (equation (13)), the noise parameter's slower mixing has negligible impact on coverage.

AI Use Statement

The authors did not use AI tools in the preparation of this manuscript beyond standard language editing assistance.

ETHICS STATEMENT

This work presents a statistical methodology for causal inference and does not involve new data collection, human subjects research, or personal data. The IHDP benchmark used in evaluation is a publicly available semi-synthetic dataset (Hill, 2011). The authors have read the ICLR Code of Ethics and confirm that this submission raises no concerns under it.