CA-MFPD: COVARIANCE-AWARE MULTI-FIDELITY POSTERIOR DISTILLATION UNDER SCARCE HIGH-FIDELITY DATA

Anonymous authors
Paper under double-blind review

ABSTRACT

Multi-fidelity regression uses cheap, biased low-fidelity data to reduce the need for expensive high-fidelity evaluations. Gaussian process (GP) models provide joint predictive distributions, but repeated GP inference can be costly. We propose covariance-aware multi-fidelity posterior distillation (CA-MFPD), which trains a neural surrogate model to reproduce the teacher's predictive distribution. We use a residual multi-fidelity GP to predict high-fidelity outputs at sampled inputs, then adjust predictive uncertainty using leave-one-out (LOO) errors on high-fidelity data. During training, a joint Gaussian KL loss brings the student's predictive mean and covariance closer to the teacher's. The student represents its covariance as a low-rank term plus a diagonal term. The off-diagonal entries represent correlations between inputs. We also compute the student's mean error at observed high-fidelity points and include it in the training objective. Our analysis shows that joint KL compares correlations between inputs, which pointwise comparison omits. We bound batch KL error in terms of mean and covariance approximation errors, and show that a lower expected batch KL gives a tighter bound on average pairwise covariance error. Numerical results show that CA-MFPD has lower overall prediction and coverage errors and faster repeated inference than standard co-kriging, with competitive results against nonlinear GP and neural multi-fidelity baselines.

1 INTRODUCTION

Obtaining high-fidelity (HF) data provides accurate observations but is costly. Low-fidelity (LF) approximations are cheaper to obtain but can be biased. Multi-fidelity Gaussian processes (GPs) combine these two sources (Kennedy & O'Hagan, 2000; Peherstorfer et al., 2018; Kandasamy et al., 2016). Autoregressive co-kriging is useful when HF data are scarce because it gives a full predictive distribution. Each new GP prediction requires kernel evaluations and a solve using the training covariance factorization, so costs accumulate over large query sets. Computing a batch's full joint covariance also requires dense matrix operations. A neural surrogate can reduce repeated inference cost; the challenge is to transfer both point predictions and cross-input covariance.

We address this question with posterior-guided data augmentation. Figure 1 shows the pipeline. A residual co-kriging teacher gives HF means and covariances at a selected set of input points. We use leave-one-out (LOO) prediction errors on the HF data to scale the teacher's predictive uncertainty up or down. We sample input points from a scrambled Sobol design using a mixture of reliability-based and uniform probabilities. A neural Gaussian student learns the joint teacher distribution through batch Gaussian KL divergence, while observed HF responses anchor its mean. An optional HF-only GP provides a LOO-gated safeguard at prediction time.

Our main contribution is a covariance-aware distillation method for multi-fidelity regression. It makes three changes. First, sampled input points carry joint Gaussian supervision instead of separate scalar targets. Second, during training, we adjust the teacher's predictive uncertainty using HF leave-one-out (LOO) errors. We combine reliability-based and uniform sampling. Observed HF data constrain the student's predictive mean at observed inputs. At inference, an optional HF-only GP can serve as an additional expert. Third, the neural student represents its predictive distribution with a low-rank

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.

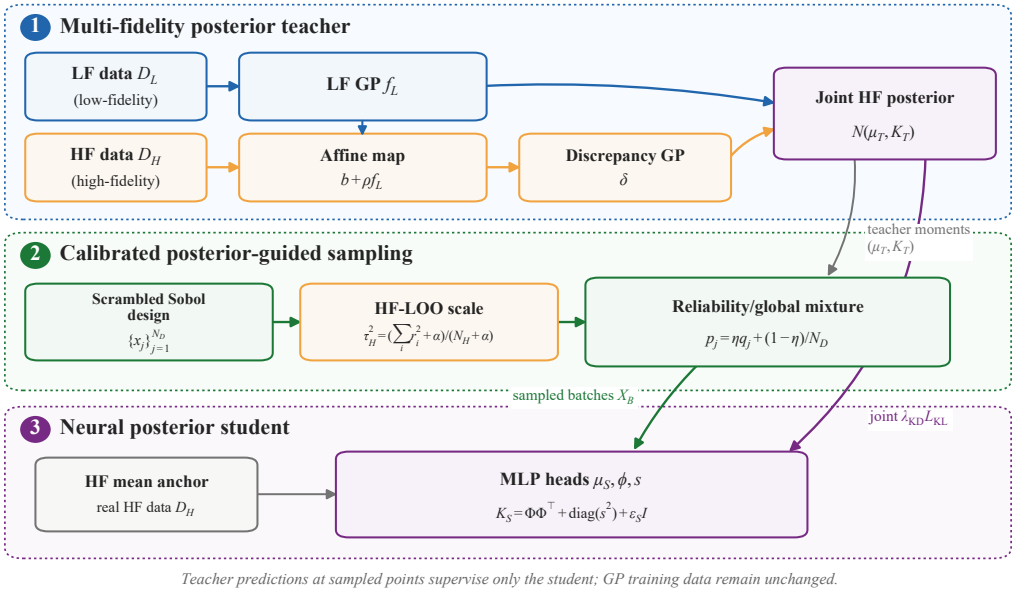

Figure 1: Overview of CA-MFPD. A residual GP teacher gives joint HF predictions at selected sets of input points. LOO scaling calibrates the teacher covariance, and posterior-guided sampling forms training batches. Joint KL trains a low-rank Gaussian student. Teacher predictions at these points supervise only the student and are not added to either GP training set. An optional HF-only expert is selected from HF-LOO predictions without test labels.

Gaussian model, reducing the cost of repeated prediction. The analysis explains the extra information in the joint objective, gives a batch covariance error bound, and connects expected batch KL to covariance error under the sampling distribution.

2 Related Work

Our method connects multi-fidelity surrogate modeling with predictive-distribution distillation. We review both areas and then describe the nonlinear baselines used in our experiments.

Multi-fidelity surrogates. Co-kriging is based on the autoregressive relation $f_H(x) = \rho f_L(x) + \delta(x)$ (Kennedy & O'Hagan, 2000; Forrester et al., 2007). Recursive methods make this model easier to compute (Le Gratiet, 2014; Perdikaris et al., 2015). Additive residual GPs provide another way to model fidelity differences (Xing et al., 2021). Deep and compositional GPs (Perdikaris et al., 2017; Raissi & Karniadakis, 2016; Cutajar et al., 2019) and neural networks (Meng & Karniadakis, 2020; Guo et al., 2022; Zhang et al., 2021; Motamed, 2020; Wu et al., 2022) model nonlinear links between fidelities. CA-MFPD instead uses an affine residual teacher and focuses on efficient transfer of its joint predictions.

Posterior and uncertainty distillation. Knowledge distillation transfers predictions from a large teacher to a smaller student (Hinton et al., 2015). Bayesian dark knowledge distills posterior samples to reduce test-time cost (Korattikara et al., 2015). Gaussian distillation can model ensemble uncertainty with a low-rank covariance (Park et al., 2025). Regression calibration and deep ensembles provide complementary approaches to predictive uncertainty (Kuleshov et al., 2018; Lakshminarayanan et al., 2017). We use a similar covariance form, but our teacher is a calibrated multi-fidelity GP and our loss is the exact Gaussian KL on each batch of input points.

Modern multi-fidelity baselines. Deep multi-fidelity GPs and neural models learn more flexible links between fidelities (Raissi & Karniadakis, 2016; Cutajar et al., 2019; Perdikaris et al., 2017; Meng

& Karniadakis, 2020). We compare with NARGP, a nonlinear autoregressive GP, and MFNN, a two-stage neural multi-fidelity model. These baselines test whether the gains come from covariance-aware distillation or only from a more flexible fidelity map.

3 METHOD

Let $\mathcal{D}_L = \{(x_i^L, y_i^L)\}_{i=1}^{N_L}$ and $\mathcal{D}_H = \{(x_i^H, y_i^H)\}_{i=1}^{N_H}$, where $N_H \ll N_L$ and $x \in [0,1]^d$. We first build the GP teacher. We then calibrate its covariance and sample training inputs. Finally, we train the neural student and define the optional HF safeguard.

3.1 RESIDUAL MULTI-FIDELITY GP TEACHER

We fit an exact Matérn-$5/2$ GP to the LF data and use the residual model

$$f_H(x) = b + \rho f_L(x) + \delta(x), \qquad \delta \sim \mathcal{GP}(0, k_\delta), \quad f_L \perp\!\!\!\perp \delta. \tag{1}$$

At the HF inputs, we estimate (b, ρ) by variance-weighted affine regression with ridge regularization. We clip ρ to $[-5, 5]$. A second Matérn-$5/2$ GP fits the residuals $y_i^H - b - \rho\mu_L(x_i^H)$. With the fitted hyperparameters and plug-in estimates of (b, ρ), the teacher moments at a selected set of inputs X are

$$\mu_T(X) = b\mathbf{1} + \rho\mu_L(X) + \mu_\delta(X), \qquad K_T(X,X) = \rho^2 K_L(X,X) + K_\delta(X,X). \tag{2}$$

3.2 LOO CALIBRATION AND POSTERIOR-GUIDED DATA AUGMENTATION

For each HF observation, we keep the LF GP fixed and refit the affine and discrepancy parts after removing that observation. From the LOO moments $(\mu_{-i}, \sigma_{-i}^2)$, we compute $r_i = (y_i^H - \mu_{-i})/\sigma_{-i}$ and set

$$\tau_H^2 = \frac{\sum_{i=1}^{N_H} r_i^2 + \alpha}{N_H + \alpha}, \quad \tau_H \leftarrow \mathrm{clip}(\tau_H, 0.5, 4), \quad \widetilde{K}_T = \tau_H^2 K_T. \tag{3}$$

$\alpha \geq 0$ is a fixed stabilization constant. This step rescales the full covariance by one scalar.

We combine a scrambled Sobol set (Joe & Kuo, 2008) of size N_D with the observed HF inputs. For each candidate x_j, we define

$$q_j \propto \{\tau_H \sqrt{K_T(x_j, x_j)} + \epsilon\}^{-1}, \qquad p_j = \eta q_j + (1-\eta)/N_D. \tag{4}$$

We sample mini-batches without replacement. The reliability term favors points with lower calibrated uncertainty. The uniform term keeps coverage across the input domain. Teacher predictions at these sampled inputs add posterior supervision but no new HF observations.

3.3 LOW-RANK NEURAL POSTERIOR

A SiLU multilayer perceptron outputs a mean $\mu_S(x)$, a factor $\phi(x) \in \mathbb{R}^q$, and a positive scale $s(x)$. On a batch X_B, they define

$$K_S = \Phi\Phi^\top + \mathrm{diag}(s^2) + \epsilon_S I, \tag{5}$$

where each row of Φ is $\phi(x)$ at one batch input. The joint teacher–student loss is

$$\begin{aligned}\mathcal{L}_{\mathrm{KL}} = \frac{1}{2B}\{ &\mathrm{tr}(K_S^{-1}\widetilde{K}_T) + (\mu_S - \mu_T)^\top K_S^{-1}(\mu_S - \mu_T) \\ &- B + \log|K_S| - \log|\widetilde{K}_T|\}.\end{aligned} \tag{6}$$

We use adaptive Cholesky jitter and minimize

$$\mathcal{L} = \lambda_H \frac{1}{N_H}\sum_i (\mu_S(x_i^H) - y_i^H)^2 + \lambda_{\mathrm{KD}}\mathcal{L}_{\mathrm{KL}} + \lambda_\mu \frac{\|\mu_S - \mu_T\|_2^2}{B} + \lambda_\phi \frac{\|\Phi\|_F^2}{Bq}. \tag{7}$$

3.4 Optional LOO-gated HF safeguard

For Branin and Nonlinear-Sin, we also fit an HF-only GP. We compute $a \in [0, 1]$ in closed form from paired HF-LOO predictions of the two teachers. The fit uses a nonnegative shrinkage prior and no test labels. During training, the teacher mean stays μ_T, and its covariance becomes $a^2 K_T + (1-a)^2 K_G$. At prediction time,

$$\mu_F = a\mu_S + (1-a)\mu_G, \tag{8}$$

$$v_F = a^2 v_S + (1-a)^2 v_G + a(1-a)(\mu_S - \mu_G)^2. \tag{9}$$

The final term measures disagreement between the two experts. Reported timings include this safeguard when it is used.

4 Mechanism-Level Analysis

This section studies three parts of CA-MFPD: joint covariance supervision, low-rank covariance approximation, and LOO spread correction. The results apply to finite Gaussian vectors on batches drawn without replacement from equation 4. We denote this batch distribution by Π_B. Appendix A gives the proofs.

4.1 Joint KL identifies cross-input dependence

For a batch X_B, let $v_{T,i} = (K_T)_{ii}$ and $v_{S,i} = (K_S)_{ii}$, and define

$$\mathcal{L}_{\mathrm{pt}} := \frac{1}{B} \sum_{i=1}^{B} D_{\mathrm{KL}}(\mathcal{N}(\mu_{T,i}, v_{T,i}) \,\|\, \mathcal{N}(\mu_{S,i}, v_{S,i}))\,. \tag{10}$$

If both covariance matrices are diagonal, then $B^{-1} D_{\mathrm{KL}}(T\|S) = \mathcal{L}_{\mathrm{pt}}$. Off-diagonal terms make the joint loss depend on cross-input covariance.

Proposition 4.1 (Pointwise versus joint Gaussian distillation)**.** *Let $T = \mathcal{N}(\mu_T, K_T)$ and $S = \mathcal{N}(\mu_S, K_S)$ be nondegenerate Gaussian laws on a batch of size B. If both covariance matrices are diagonal, then $D_{\mathrm{KL}}(T\|S)$ equals the sum of the B univariate Gaussian KL divergences. In contrast, the collection of marginal means and variances does not identify any off-diagonal covariance. In particular, for*

$$K_\rho = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}, \qquad |\rho| < 1, \tag{11}$$

the teacher $\mathcal{N}(0, K_\rho)$ and student $\mathcal{N}(0, I_2)$ have identical univariate marginals, whereas

$$D_{\mathrm{KL}}(\mathcal{N}(0, K_\rho) \,\|\, \mathcal{N}(0, I_2)) = -\tfrac{1}{2} \log(1 - \rho^2), \tag{12}$$

which is strictly positive for $\rho \neq 0$.

Proposition 4.1 shows the difference between the two objectives. Marginal means and variances can match exactly even when the joint distribution is wrong. Figure 4 illustrates this with a two-point Gaussian example: pointwise KL remains zero as the correlation changes, while normalized joint KL increases with its magnitude.

4.2 Approximation error of the low-rank student

Proposition 4.2 (KL bound for low-rank-plus-diagonal covariance)**.** *Let $K = \widetilde{K}_T \in \mathbb{R}^{B\times B}$ satisfy $K = U_q \Lambda_q U_q^\top + R_q \succ 0$, where $U_q \in \mathbb{R}^{B\times q}$ has orthonormal columns, $\Lambda_q \in \mathbb{R}^{q\times q}$ is diagonal positive definite, $U_q \Lambda_q U_q^\top$ is the leading rank-q eigenspace, and $R_q \in \mathbb{R}^{B\times B}$ is positive semidefinite. Let $K_S = \Phi\Phi^\top + D + \epsilon_S I_B \succ 0$ with $\Phi \in \mathbb{R}^{B\times q}$, $D = \mathrm{diag}(s^2) \succeq 0$, and $\epsilon_S > 0$; $\mathrm{diag}(R_q) := \mathrm{Diag}((R_q)_{11}, \ldots, (R_q)_{BB})$.*

$$e_{\mathrm{fac}} = \|U_q \Lambda_q U_q^\top - \Phi\Phi^\top\|_F, \qquad e_{\mathrm{tail}} = \|R_q - \mathrm{diag}(R_q)\|_F,$$
$$e_{\mathrm{diag}} = \|\,\mathrm{diag}(R_q) - D\|_F. \tag{13}$$

If $m = \lambda_{\min}(K_S) > 0$ *and* $\delta = \|K_S^{-1/2}(K - K_S)K_S^{-1/2}\|_2 < 1$*, then*

$$D_{\mathrm{KL}}(\mathcal{N}(\mu_T, K) \,\|\, \mathcal{N}(\mu_S, K_S)) \le \frac{\|\mu_T - \mu_S\|_2^2}{2m} + \frac{(e_{\mathrm{fac}} + e_{\mathrm{tail}} + e_{\mathrm{diag}} + \epsilon_S\sqrt{B})^2}{4m^2(1-\delta)}. \tag{14}$$

Moreover, if $\lambda_1 \ge \cdots \ge \lambda_B$ *are the eigenvalues of* K*, then* $e_{\mathrm{tail}} \le \|R_q\|_F = (\sum_{j>q} \lambda_j^2)^{1/2}$.

The condition $\delta < 1$ holds if $\|K - K_S\|_2 < \lambda_{\min}(K_S)$. This is a batch-level approximation bound, with ϵ_S equal to the Cholesky jitter used on that batch. The spectral tail measures the rank-q matrix error. The term $\inf_\theta \mathbb{E}_{X_B \sim \Pi_B} e_{\mathrm{fac}}(X_B;\theta)^2$ measures how well one neural map fits the leading covariance factors across batches. Increasing q improves the bound when these fitting and conditioning terms stay controlled. Equation equation 6 divides the same KL term by B.

Corollary 4.3 (Joint KL controls linear predictive queries)**.** *Let* $K = \widetilde{K}_T \succ 0$ *and* $K_S \succ 0$ *satisfy* $mI_B \preceq K_S \preceq MI_B$ *and* $K \preceq M_T I_B$*, with* $m > 0$*. For* $\varepsilon = D_{\mathrm{KL}}(\mathcal{N}(\mu_T, K) \,\|\, \mathcal{N}(\mu_S, K_S))$ *and any nonzero* $a \in \mathbb{R}^B$*, define* $v_T = a^\top K a$*,* $v_S = a^\top K_S a$*,* $r_a = v_T/v_S$*, and* $R = \max\{M_T/m, 1\}$*. Then*

$$\frac{(a^\top(\mu_T - \mu_S))^2}{v_S} + r_a - 1 - \log r_a \le 2\varepsilon, \qquad \frac{|a^\top(\mu_T - \mu_S)|}{\|a\|_2} \le \sqrt{2\varepsilon M}, \qquad |r_a - 1| \le 2\sqrt{\varepsilon R}. \tag{15}$$

The corollary applies to any linear query on the sampled batch.

Proposition 4.4 (Expected batch KL controls sampled dependence)**.** *For* $X_B \sim \Pi_B$*, let* $T_{X_B} = \mathcal{N}(\mu_T, K)$ *and* $S_{\theta,X_B} = \mathcal{N}(\mu_S, K_S)$*. Assume uniformly on the support of* Π_B *that* $mI_B \preceq K_S \preceq MI_B$ *and* $K \preceq M_T I_B$*, and set* $R = \max\{M_T/m, 1\}$*. If* $\overline{\mathcal{L}}_{\mathrm{KL}} := \mathbb{E}_{\Pi_B}[B^{-1} D_{\mathrm{KL}}(T_{X_B} \| S_{\theta,X_B})] \le \bar{\varepsilon}$*, then any nonzero query* $A = A(X_B)$ *formed from the sampled batch, with* $r_A = A^\top K A/(A^\top K_S A)$*, satisfies*

$$\mathbb{E}\frac{|A^\top(\mu_T - \mu_S)|}{\|A\|_2} \le \sqrt{2MB\bar{\varepsilon}}, \qquad \mathbb{E}|r_A - 1| \le 2\sqrt{RB\bar{\varepsilon}}. \tag{16}$$

If (I, J) *is a uniformly sampled distinct coordinate pair, then*

$$\mathbb{E}|K_{IJ} - (K_S)_{IJ}| \le 2M\sqrt{RB\bar{\varepsilon}}. \tag{17}$$

All expectations include batch and query/pair sampling.

Proposition 4.4 links the training loss to covariance error under the sampling distribution. Because all loss weights are nonnegative, $\lambda_{\mathrm{KD}} > 0$ and $\mathbb{E}\mathcal{L} \le \varepsilon_{\mathrm{obj}}$ imply $\overline{\mathcal{L}}_{\mathrm{KL}} \le \varepsilon_{\mathrm{obj}}/\lambda_{\mathrm{KD}}$. The bound applies under Π_B once the expected loss reaches this level.

4.3 Interpretation of LOO covariance scaling

Lemma 4.5 (LOO scaling under global spread misspecification)**.** *Let* $\mathcal{F}_{-i}$ *denote the data used for the* i*th LOO fit, with* $\sigma_{-i} > 0$*. Suppose*

$$\mathbb{E}[y_i^H - \mu_{-i} \mid \mathcal{F}_{-i}] = 0, \qquad \mathbb{E}[(y_i^H - \mu_{-i})^2 \mid \mathcal{F}_{-i}] = c^2\sigma_{-i}^2 \tag{18}$$

for a common scale $c > 0$*. For the unclipped estimator (before the implementation's clipping)* $\widehat{\tau}_H^2 = (\sum_i r_i^2 + \alpha)/(N_H + \alpha)$*,*

$$\mathbb{E}[\widehat{\tau}_H^2] = \frac{N_H c^2 + \alpha}{N_H + \alpha} = c^2 + \frac{\alpha(1 - c^2)}{N_H + \alpha}. \tag{19}$$

Hence the estimator is centered at one under correct spread specification ($c = 1$*) and otherwise estimates the global variance multiplier with shrinkage toward one.*

Lemma 4.5 explains scalar spread correction. It does not require independent LOO residuals. The constant α and clipping affect the finite-sample value. Because the correction leaves the mean unchanged, it mainly affects NLL, CRPS, and interval width rather than NRMSE.

5 EXPERIMENTAL PROTOCOL

Result selection and completeness. The main study contains ten benchmarks, five methods, 20 seeds, and seven HF budgets, for 7,000 runs. Each included benchmark has every method, seed, and budget with no duplicate or failed cell. The benchmarks are Forrester, Nonlinear-Sin, Branin, Currin, Park-1, Park-2, Hartmann-6, Borehole, Heat Conduction Grid, and Wing Weight. The paper package includes snapshot hashes and audit files.

Design and baselines. Inputs are normalized to $[0, 1]^d$. Each benchmark uses nested Latin-hypercube LF/HF designs and a fixed 4,096-point Sobol test set. The main baselines are an exact HF-only GP, residual co-kriging, deterministic Co-RBF, and NHLF-Co-Kriging. We also test NARGP and MFNN in separate runs. NARGP feeds the LF GP mean to an HF GP. MFNN fits an LF MLP and then a heteroscedastic HF MLP on $[x, \hat{f}_L(x)]$. These runs use the same benchmarks, seeds, budgets, and splits. NHLF splits one LF budget into two pseudo-sources. Co-RBF is deterministic, so it has no NLL, CRPS, or coverage value.

Frozen benchmark profiles. Each benchmark uses one CA-MFPD configuration for all seeds and budgets. The configuration is fixed before test evaluation. All methods share the splits, test set, metrics, and aggregation rules. Run metadata, source hashes, timestamps, and diagnostics record the settings used in each run.

Statistical unit. Budgets are nested within each seed, so we average the seven budgets within each benchmark–seed pair. This gives 200 units for the two-sided Wilcoxon signed-rank test. We apply Holm correction across four baselines. Per-benchmark tests against co-kriging use 20 seed averages and Holm correction across ten tasks. Cellwise win rates are summary statistics.

Ablation screen. The ablation screens cover Borehole, Park-2, and Wing Weight. Each uses ten methods, five seeds, and three nested HF budgets: $(16, 40, 80)$, $(8, 20, 40)$, and $(20, 50, 100)$. All 450 runs succeed. Methods share the same splits. Neural variants also share the backbone, optimizer, and step count. We average the three budgets within each seed, leaving five paired values per dataset.

6 RESULTS

We report predictive accuracy, uncertainty quality, cost, and ablations.

6.1 AGGREGATE NRMSE IMPROVES, WITH TASK-DEPENDENT GAINS

Across 1,400 cells per method, CA-MFPD has mean NRMSE 0.055162, compared with 0.056998 for co-kriging. This is a 3.22% reduction. CA-MFPD has the best mean rank (1.711), wins 62.5% of all five-method cells, and beats co-kriging in 80.1% of paired cells. After averaging budgets within each benchmark–seed pair, it wins 82.5% of 200 pairs. The mean difference is -0.001837 with Holm-adjusted $p = 2.13 \times 10^{-17}$. Figure 2 shows the learning curves on log-scaled vertical axes.

The gains vary by task (Table 3). CA-MFPD improves eight benchmarks, shows no clear difference on Currin, and performs worse on Forrester. It wins all 20 seed averages on Heat Conduction Grid and Wing Weight. Their mean differences are -0.002910 and -0.001247, with Holm-adjusted $p = 1.9 \times 10^{-5}$ for both. Nonlinear-Sin has the largest improvement, -0.01073.

CA-MFPD has lower NRMSE than NARGP in 154/200 benchmark–seed units (77.0%) and lower NRMSE than MFNN in 127/200 units (63.5%); see Figure 3. Against NARGP, its win rates are 69.0% for NLL, 66.5% for CRPS, and 47.5% for coverage error. Against MFNN, they are 100%, 58.0%, and 100%. MFNN has lower overall mean NRMSE, 0.042238 versus 0.055162, because it is strong on several tasks. Its uncertainty metrics are weaker. Appendix Table 5 gives per-benchmark results.

6.2 Coverage error and aggregate proper scores

Mean 90% coverage is 0.932 for CA-MFPD and 0.858 for co-kriging. The mean absolute coverage error is 0.101 versus 0.123. Median NLL drops from 1.133 to 0.507, and mean CRPS rank improves from 2.001 to 1.874. Appendix Table 6 summarizes task-level changes; Appendix Figure 5 and Table 4 give the complete values.

6.3 Training cost is amortized by repeated prediction

Median inference on 4,096 points takes 0.003742 seconds for CA-MFPD, 0.037229 for co-kriging, 0.040021 for NARGP, and 0.090225 for MFNN. Median training takes 64.97, 1.779, 1.915, and 9.156 seconds. The training cost breaks even after 1,887, 1,738, and 645 repeated batches, respectively. Appendix Figure 6 gives task-level values. Branin and Nonlinear-Sin include the HF-GP safeguard.

6.4 Ablations support reliability sampling and probabilistic LOO calibration

Appendix Figure 7 compares nine controls. The full model beats co-kriging on all three tasks. Uniform sampling raises mean NRMSE by 34.8%, 31.9%, and 9.8%. Joint covariance helps on Borehole and Park-2. On Wing Weight, mean-only is 2.2% better. Removing the HF anchor changes little on Park-2 and Wing Weight. Removing LOO scaling lowers NRMSE, so its main effect is on uncertainty quality.

LOO scaling reduces mean NLL from 1.097 to 0.553, -4.047 to -4.557, and 1.556 to 1.259 on the three tasks. It reduces CRPS from 0.324 to 0.218, 0.00203 to 0.00143, and 0.541 to 0.456. Table 1 reports all four uncertainty metrics. All five seed averages follow the same direction. The intervals are also narrower and have lower coverage error. LOO scaling therefore improves calibration and sharpness at a small cost in NRMSE.

Appendix Table 6 shows lower NLL on eight benchmarks, lower CRPS on six, and lower coverage error on six. The full model also beats co-kriging, the HF-only MLP, random sampling plus HF-GP, and uncertainty-weighted pointwise distillation for every seed on all three tasks. Appendix Table 9 reports the NRMSE controls.

Table 1: LOO scaling ablation on Borehole, Park-2, and Wing Weight. Each entry reports full model / without LOO scaling, averaged over five seed-level averages of three nested budgets. Lower is better; W_{90} is nominal 90% interval width.

Dataset	NLL	CRPS	$\lvert\mathrm{Cov}_{90} - 0.9\rvert$	W_{90}
Borehole	$0.553 \,/\, 1.097$	$0.218 \,/\, 0.324$	$0.094 \,/\, 0.099$	$2.215 \,/\, 4.026$
Park-2	$-4.557 \,/\, -4.047$	$1.43\times10^{-3} \,/\, 2.03\times10^{-3}$	$0.094 \,/\, 0.096$	$0.014 \,/\, 0.024$
Wing Weight	$1.259 \,/\, 1.556$	$0.456 \,/\, 0.541$	$0.088 \,/\, 0.098$	$4.188 \,/\, 6.062$

7 Limitations and Discussion

The study uses benchmark functions and simulators. The GP teacher uses plug-in estimates of (b, ρ), kernel hyperparameters, and τ_H. The theory covers sampled finite batches under Π_B. The baseline set includes NARGP and MFNN, and the ablations use five seeds. Future work will test measured data and broader teacher and student models.

8 Conclusion

We proposed CA-MFPD, which transfers joint Gaussian predictions from a calibrated multi-fidelity GP to a low-rank neural surrogate model. Joint KL learns cross-input dependence that marginal matching cannot identify. LOO scaling calibrates the teacher spread, and reliability-based sampling selects training inputs. Across the ten benchmarks, CA-MFPD lowers overall NRMSE and coverage error, improves probability scores, and gives fast repeated prediction.

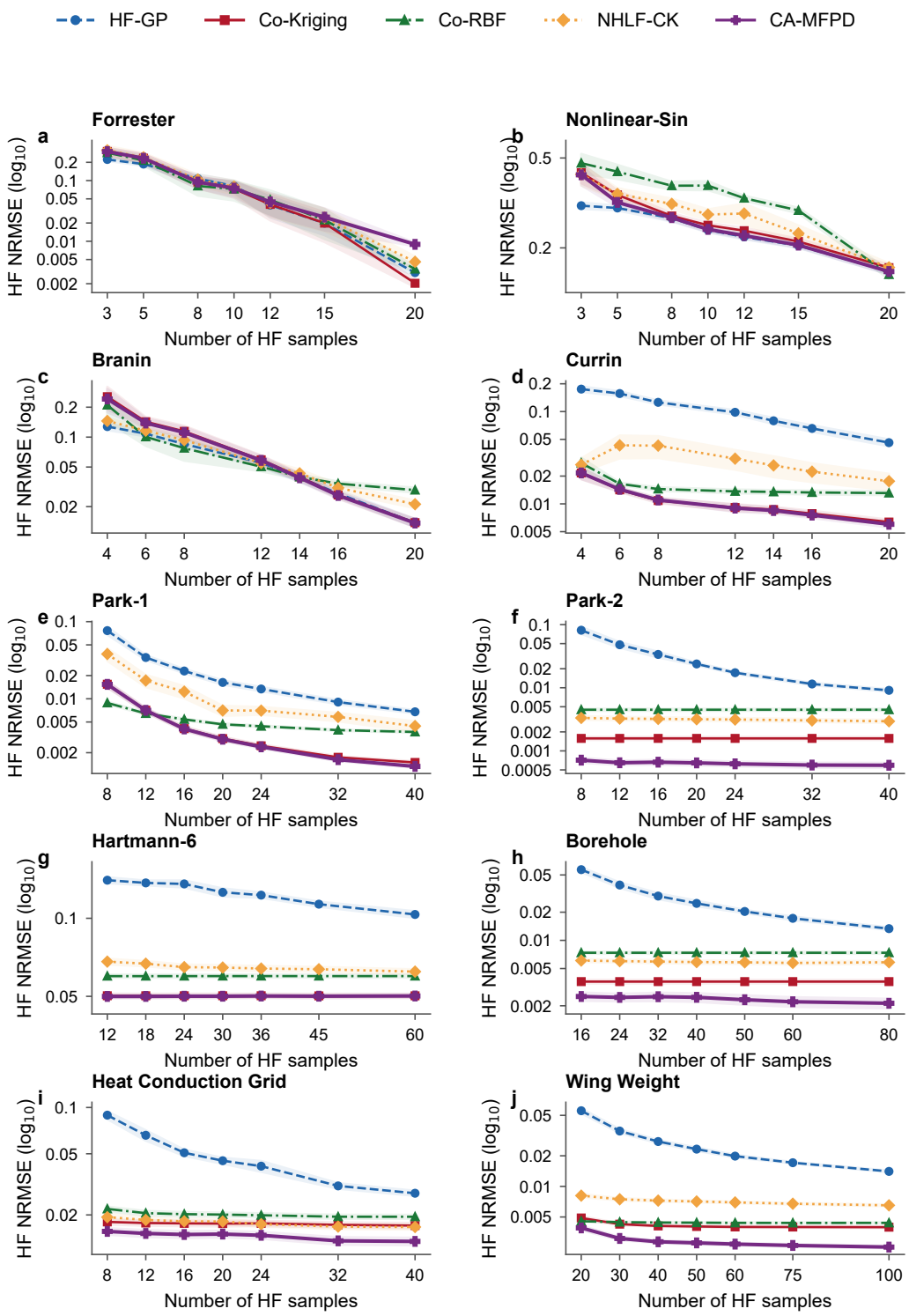

Figure 2: NRMSE over seven HF budgets and 20 seeds. The vertical axes use a base-10 log scale. Means and standard errors are computed on the original NRMSE scale. Bands show mean ± standard error across seeds.

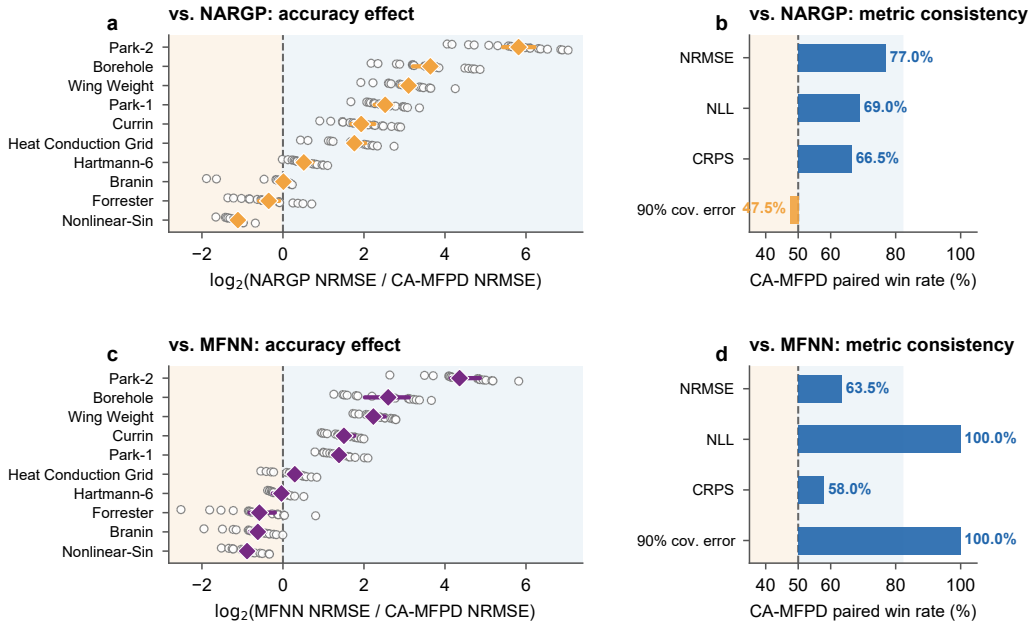

Figure 3: Paired comparisons with NARGP and MFNN after averaging seven HF budgets within each benchmark–seed unit. Panels (a,c) show $\log_2$(comparator NRMSE/CA-MFPD NRMSE); positive values favor CA-MFPD. Diamonds are medians, and horizontal lines are percentile 95% bootstrap intervals across seeds. Panels (b,d) show CA-MFPD win rates over 200 paired units. Coverage error is the absolute difference from 90%. Dashed lines mark parity.

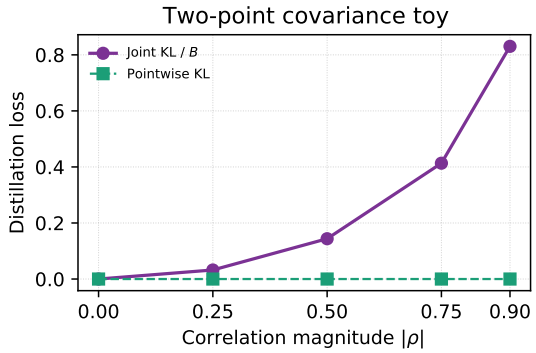

Figure 4: Two-point covariance diagnostic. The marginals match for every $|\rho|$, so pointwise KL remains zero. Joint KL increases with the correlation magnitude.

AI USE STATEMENT

Generative AI tools were used to assist source-code summarization, drafting, Chinese–English translation, plotting, and template formatting. All numerical statements were recomputed from repository result files; no experiment was generated by AI. The authors are responsible for verifying the implementation, citations, analysis, and final text before submission.

REFERENCES

Kurt Cutajar, Mark Pullin, Andreas Damianou, Neil Lawrence, and Javier Gonzalez. Deep gaussian processes for multi-fidelity modeling. *arXiv preprint arXiv:1903.07320*, 2019. URL https://arxiv.org/abs/1903.07320.

Alexander I. J. Forrester, András Sóbester, and Andy J. Keane. Multi-fidelity optimization via surrogate modelling. *Proceedings of the Royal Society A*, 463(2088):3251–3269, 2007. doi: 10.1098/rspa.2007.1900.

Tilmann Gneiting and Adrian E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007. doi: 10.1198/016214506000001437.

Mengwu Guo, Andrea Manzoni, Maurice Amendt, Paolo Conti, and Jan S. Hesthaven. Multi-fidelity regression using artificial neural networks: Efficient approximation of parameter-dependent output quantities. *Computer Methods in Applied Mechanics and Engineering*, 389:114378, 2022. doi: 10.1016/j.cma.2021.114378.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. In *NIPS Deep Learning and Representation Learning Workshop*, 2015. URL https://arxiv.org/abs/1503.02531.

Stephen Joe and Frances Y. Kuo. Constructing sobol sequences with better two-dimensional projections. *SIAM Journal on Scientific Computing*, 30(5):2635–2654, 2008. doi: 10.1137/070709359.

Kirthevasan Kandasamy, Gautam Dasarathy, Junier B. Oliva, Jeff Schneider, and Barnabás Póczos. Gaussian process bandit optimisation with multi-fidelity evaluations. In *Advances in Neural Information Processing Systems*, volume 29, pp. 992–1000, 2016.

Marc C. Kennedy and Anthony O'Hagan. Predicting the output from a complex computer code when fast approximations are available. *Biometrika*, 87(1):1–13, 2000. doi: 10.1093/biomet/87.1.1.

Anoop Korattikara, Vivek Rathod, Kevin Murphy, and Max Welling. Bayesian dark knowledge. In *Advances in Neural Information Processing Systems*, volume 28, pp. 3438–3446, 2015. URL https://arxiv.org/abs/1506.04416.

Volodymyr Kuleshov, Nathan Fenner, and Stefano Ermon. Accurate uncertainties for deep learning using calibrated regression. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2796–2804, 2018.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, volume 30, pp. 6402–6413, 2017.

Loic Le Gratiet. Recursive co-kriging model for design of computer experiments with multiple levels of fidelity. *International Journal for Uncertainty Quantification*, 4(5):365–386, 2014. doi: 10.1615/Int.J.UncertaintyQuantification.2014006914.

Xuhui Meng and George Em Karniadakis. A composite neural network that learns from multi-fidelity data: Application to function approximation and inverse pde problems. *Journal of Computational Physics*, 401:109020, 2020. doi: 10.1016/j.jcp.2019.109020.

Mohammad Motamed. A multi-fidelity neural network surrogate sampling method for uncertainty quantification. *International Journal for Uncertainty Quantification*, 10(4):315–332, 2020. doi: 10.1615/Int.J.UncertaintyQuantification.2020031957.

Sehyun Park, Jongjin Lee, Yunseop Shin, Ilsang Ohn, and Yongdai Kim. Knowledge distillation of uncertainty using deep latent factor model. *arXiv preprint arXiv:2510.19290*, 2025. URL https://arxiv.org/abs/2510.19290.

Benjamin Peherstorfer, Karen Willcox, and Max Gunzburger. Survey of multifidelity methods in uncertainty propagation, inference, and optimization. *SIAM Review*, 60(3):550–591, 2018. doi: 10.1137/16M1082469.

Paris Perdikaris, Daniele Venturi, Johannes O. Róyset, and George Em Karniadakis. Multi-fidelity modelling via recursive co-kriging and gaussian-markov random fields. *Proceedings of the Royal Society A*, 471(2179):20150018, 2015. doi: 10.1098/rspa.2015.0018.

Paris Perdikaris, Maziar Raissi, Andreas Damianou, Neil D. Lawrence, and George Em Karniadakis. Nonlinear information fusion algorithms for data-efficient multi-fidelity modelling. *Proceedings of the Royal Society A*, 473(2198):20160751, 2017. doi: 10.1098/rspa.2016.0751.

Maziar Raissi and George Em Karniadakis. Deep multi-fidelity gaussian processes. *arXiv preprint arXiv:1604.07484*, 2016. URL https://arxiv.org/abs/1604.07484.

Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006. URL https://gaussianprocess.org/gpml/.

Dongxia Wu, Matteo Chinazzi, Alessandro Vespignani, Yi-An Ma, and Rose Yu. Multi-fidelity hierarchical neural processes. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 2029–2038, 2022. doi: 10.1145/3534678.3539364.

Wei W. Xing, Akeel A. Shah, Peng Wang, Shandian Zhe, Qian Fu, and Robert M. Kirby. Residual gaussian process: A tractable nonparametric bayesian emulator for multi-fidelity simulations. *Applied Mathematical Modelling*, 97:36–56, 2021. doi: 10.1016/j.apm.2021.03.041.

Xinshuai Zhang, Fangfang Xie, Tingwei Ji, Zaoxu Zhu, and Yao Zheng. Multi-fidelity deep neural network surrogate model for aerodynamic shape optimization. *Computer Methods in Applied Mechanics and Engineering*, 373:113485, 2021. doi: 10.1016/j.cma.2020.113485.

A Proofs of the Mechanism-Level Results

This section proves the results in Section 4.

Proof of Proposition 4.1. For nondegenerate B-variate Gaussian laws,

$$D_{\mathrm{KL}}(T\|S) = \frac{1}{2}\left\{\mathrm{tr}(K_S^{-1}K_T) + (\mu_S-\mu_T)^\top K_S^{-1}(\mu_S-\mu_T) - B + \log\frac{|K_S|}{|K_T|}\right\}. \tag{20}$$

When K_T and K_S are diagonal, each term separates by coordinate. The joint KL is then the sum of the univariate KL terms. A marginal Gaussian depends only on μ_i and K_{ii}, so it contains no off-diagonal information. In equation 11, both teacher and student have standard normal marginals. Setting $K_S = I_2$, $K_T = K_\rho$, and both means to zero gives $\frac{1}{2}\{\mathrm{tr}(K_\rho) - 2 - \log|K_\rho|\} = -\frac{1}{2}\log(1-\rho^2)$, which is positive whenever $\rho \neq 0$. □

Proof of Proposition 4.2. Let $E = K - K_S$ and $H = K_S^{-1/2} E K_S^{-1/2}$. Then $K_S^{-1/2} K K_S^{-1/2} = I + H$. If $h_1, \ldots, h_B$ are the eigenvalues of H, the covariance contribution to the Gaussian KL is

$$\frac{1}{2}\sum_{j=1}^{B}\{h_j - \log(1+h_j)\}. \tag{21}$$

The assumption $\|H\|_2 = \delta < 1$ implies $|h_j| \le \delta$. Using $h - \log(1+h) \le h^2/[2(1-\delta)]$ for $|h| \le \delta$ gives

$$\frac{1}{2}\{\mathrm{tr}(H) - \log|I+H|\} \le \frac{\|H\|_F^2}{4(1-\delta)} \le \frac{\|E\|_F^2}{4m^2(1-\delta)}. \tag{22}$$

The mean contribution is at most $\|\mu_T - \mu_S\|_2^2/(2m)$. Moreover,

$$E = (U_q \Lambda_q U_q^\top - \Phi\Phi^\top) + (R_q - \mathrm{diag}\, R_q) + (\mathrm{diag}\, R_q - D) - \epsilon_S I, \tag{23}$$

so the triangle inequality yields $\|E\|_F \le e_{\mathrm{fac}} + e_{\mathrm{tail}} + e_{\mathrm{diag}} + \epsilon_S\sqrt{B}$. Finally, orthogonal invariance gives $\|R_q\|_F = (\sum_{j>q}\lambda_j^2)^{1/2}$. Removing the diagonal cannot increase this norm. These bounds prove equation 14. □

Proof of Corollary 4.3. The map $z \mapsto a^\top z$ turns the batch Gaussian into a one-dimensional Gaussian. The data-processing inequality bounds its KL by ε. Thus,

$$\frac{1}{2}\left\{\frac{(a^\top(\mu_T-\mu_S))^2}{v_S}+r_a-1-\log r_a\right\}\le\varepsilon. \tag{24}$$

Since $v_S \le M\|a\|_2^2$, the mean bound follows. Moreover, $r_a = a^\top K a/(a^\top K_S a) \le M_T/m \le R$. Let $g(r) = r-1-\log r$. Between r_a and 1, $g''(r) = 1/r \ge 1/R$. Therefore, $g(r_a) \ge (r_a-1)^2/(2R)$. Together with $g(r_a) \le 2\varepsilon$, this proves the variance-ratio bound. □

Proof of Proposition 4.4. Write $\varepsilon_B = D_{\mathrm{KL}}(T_{X_B}\|S_{\theta,X_B})$. Corollary 4.3 applies to every realized batch and every admissible $A(X_B)$. Taking expectations and using Jensen's inequality gives

$$\mathbb{E}\frac{|A^\top(\mu_T-\mu_S)|}{\|A\|_2}\le\sqrt{2M\mathbb{E}\varepsilon_B},\qquad \mathbb{E}|r_A-1|\le 2\sqrt{R\mathbb{E}\varepsilon_B}. \tag{25}$$

Because $\mathbb{E}\varepsilon_B = B\overline{\mathcal{L}}_{\mathrm{KL}}$, these inequalities prove equation 16. For distinct I, J, let $a_\pm = (e_I \pm e_J)/\sqrt{2}$ and $v_\pm = a_\pm^\top K a_\pm$. Define $v_{S,\pm}$ in the same way. Then $K_{IJ} - (K_S)_{IJ} = \{(v_+ - v_{S,+}) - (v_- - v_{S,-})\}/2$. Because $v_{S,\pm} \le M$ and the corollary gives $|v_\pm/v_{S,\pm} - 1| \le 2\sqrt{R\varepsilon_B}$, the covariance error is at most $2M\sqrt{R\varepsilon_B}$. Averaging over the pair and batch and applying Jensen's inequality proves equation 17. □

Proof of Lemma 4.5. By definition, $r_i = (y_i^H - \mu_{-i})/\sigma_{-i}$. Assumption equation 18 and the tower property imply $\mathbb{E}[r_i^2] = c^2$. Linearity of expectation therefore gives

$$\mathbb{E}[\widehat{\tau}_H^2] = \frac{\sum_{i=1}^{N_H}\mathbb{E}[r_i^2]+\alpha}{N_H+\alpha} = \frac{N_H c^2+\alpha}{N_H+\alpha}. \tag{26}$$

Rearranging gives equation 19. This result uses the estimator before clipping and does not require independent LOO residuals. □

B Reproducibility Details

This section lists the benchmark budgets, fixed configurations, and training steps.

B.1 Benchmark budgets

All inputs are normalized to $[0,1]^d$. The (N_L, N_H) settings are: Forrester and Nonlinear-Sin, $(100, \{3,5,8,10,12,15,20\})$; Branin and Currin, $(200, \{4,6,8,12,14,16,20\})$; Park-1, Park-2, and Heat Conduction Grid, $(200, \{8,12,16,20,24,32,40\})$; Hartmann-6, $(300, \{12,18,24,30,36,45,60\})$; Borehole, $(400, \{16,24,32,40,50,60,80\})$; and Wing Weight, $(500, \{20,30,40,50,60,75,100\})$. Every run uses seeds 0–19 and the same 4,096-point Sobol test set with seed 20260909.

B.2 Frozen benchmark profiles

Automatic profile adaptation is off. Each benchmark uses one CA-MFPD configuration for all 20 seeds and seven HF budgets. The configuration is fixed before test evaluation. The released metadata and diagnostics record the exact settings.

B.3 Training algorithm and implementation audit

The exact GP uses a constant-scaled ARD Matérn-$5/2$ kernel, white noise, and standardized outputs. The student has a shared SiLU backbone and three heads. Training has seven steps: (1) fit the LF GP; (2) estimate (b, ρ) and fit the discrepancy GP; (3) compute HF-LOO residuals and τ_H; (4) fit the optional HF-only GP and estimate a; (5) form the Sobol/HF design and compute teacher moments; (6) sample batches and optimize equation 7; and (7) retain the student and optional HF expert for prediction. Teacher predictions at sampled inputs are not added to $\mathcal{D}_L$ or $\mathcal{D}_H$.

Dense GP fitting is cubic in N_L and N_H (Rasmussen & Williams, 2006). Each student batch also needs $O(B^3)$ Cholesky factorizations. Student-only marginal prediction is linear in the number of queries. The safeguard adds one small HF GP evaluation.

C Complete-Run Aggregate Results

The tables report overall, per-task, and middle-budget results. We then compute when faster prediction offsets the extra training cost.

Table 2: Aggregate metrics over 1,400 cells per method from the ten complete runs. NRMSE rank and win rate are computed within each benchmark–budget–seed cell. Coverage refers to nominal 90% intervals, and times are medians.

Method	NRMSE	Rank	Win	Coverage	NLL med.	Train (s)	Infer (s)
HF-GP	0.07995	4.266	8.4%	0.849	2.425	0.161	0.00172
Co-Kriging	0.05700	2.344	12.0%	0.858	1.133	1.779	0.03723
Co-RBF	0.06482	3.358	12.8%	–	–	0.002	0.03337
NHLF-CK	0.06222	3.321	7.2%	0.874	1.497	0.950	0.03673
CA-MFPD	0.05516	1.711	62.5%	0.932	0.507	64.969	0.00374

Table 3: Per-benchmark mean NRMSE over all budgets and seeds. Δ is CA-MFPD minus co-kriging. "Better seeds" is the fraction of 20 seed-level budget averages favoring CA-MFPD. Holm correction is across the ten per-benchmark tests.

Benchmark	Co-Kriging	CA-MFPD	Δ	Better seeds	Holm p
Borehole	0.00363	0.00237	-0.00126	100%	1.9e-05
Branin	0.09302	0.08978	-0.00324	90%	4.8e-05
Currin	0.01124	0.01115	-0.00009	60%	0.33
Forrester	0.10952	0.11184	+0.00232	5%	3.4e-05
Hartmann-6	0.05026	0.05006	-0.00020	85%	0.00011
Heat Conduction Grid	0.01755	0.01464	-0.00291	100%	1.9e-05
Nonlinear-Sin	0.27397	0.26324	-0.01073	95%	0.00021
Park-1	0.00506	0.00498	-0.00008	90%	0.00011
Park-2	0.00158	0.00064	-0.00094	100%	1.9e-05
Wing Weight	0.00417	0.00292	-0.00125	100%	1.9e-05

Table 4: Per-benchmark probabilistic metrics over 20 seeds and seven HF budgets (140 cells per method). NLL is the median; CRPS, 90% coverage, and absolute coverage error are reported as means.

Benchmark	NLL med.		CRPS mean		90% coverage		|CovErr|	
	Co-Krig.	CA-MFPD	Co-Krig.	CA-MFPD	Co-Krig.	CA-MFPD	Co-Krig.	CA-MFPD
Forrester	-0.083	0.654	1.224	1.267	0.790	0.861	0.182	0.134
Nonlinear-Sin	0.155	0.120	0.2184	0.2064	0.710	0.772	0.234	0.183
Branin	4.789	4.662	32.59	30.44	0.696	0.894	0.228	0.107
Currin	0.034	-0.542	0.07352	0.08461	0.727	0.924	0.179	0.078
Park-1	-1.461	-1.506	0.06648	0.06398	0.915	0.932	0.102	0.075
Park-2	-3.986	-4.571	0.002648	0.001332	0.980	0.997	0.080	0.097
Hartmann-6	-0.512	0.040	0.07262	0.1206	0.921	0.985	0.022	0.085
Borehole	1.243	0.576	0.4255	0.2427	0.983	0.985	0.083	0.085
Heat Conduction Grid	2.143	2.027	0.9920	0.9955	0.870	0.985	0.036	0.085
Wing Weight	1.692	1.194	0.6839	0.4614	0.988	0.982	0.088	0.082

Proper scores measure calibration and sharpness together (Gneiting & Raftery, 2007). Across ten benchmarks, CA-MFPD improves median NLL and mean CRPS rank over co-kriging. It also reduces absolute coverage error. Table 4 shows the task-level values.

Table 5: Extended nonlinear-baseline comparison on the ten complete benchmarks. NRMSE is the mean over 20 seeds and seven HF budgets; NLL is the median and coverage is the mean over the same 140 cells per benchmark. Lower NRMSE and NLL are better; for coverage, the value closer to nominal 90% is better. Bold entries identify the best of CA-MFPD, NARGP, and MFNN for each metric.

Benchmark	CA NRMSE	NARGP NRMSE	MFNN NRMSE	CA NLL	NARGP NLL	MFNN NLL	CA Cov.	NARGP Cov.	MFNN Cov.
Borehole	**0.002370**	0.028501	0.012631	**0.576**	2.740	39.586	0.985	**0.972**	0.345
Branin	0.089780	0.072893	**0.051246**	**4.662**	4.904	851.713	**0.894**	0.796	0.270
Currin	**0.011151**	0.046559	0.031985	-0.542	**-0.597**	76.491	0.924	**0.887**	0.389
Forrester	0.111836	0.086065	**0.073783**	0.654	**-1.201**	33.005	0.861	**0.869**	0.475
Hartmann-6	0.050059	0.074533	**0.048926**	**0.040**	2.724	2942.455	**0.985**	0.589	0.105
Heat Conduction Grid	**0.014638**	0.049433	0.017144	**2.027**	2.109	321.061	0.985	**0.847**	0.317
Nonlinear-Sin	0.263241	**0.119147**	0.144888	0.120	**-1.489**	77.170	0.772	**0.805**	0.284
Park-1	**0.004978**	0.028902	0.012950	**-1.506**	0.173	15.687	**0.932**	0.965	0.461
Park-2	**0.000640**	0.035871	0.014370	**-4.571**	-1.626	12.618	0.997	**0.954**	0.453
Wing Weight	**0.002924**	0.025345	0.014462	**1.194**	2.983	95.483	0.982	**0.965**	0.255

Table 6: Task-level changes in probabilistic metrics: CA-MFPD minus co-kriging. NLL is the median over 140 cells per benchmark; CRPS and $|\mathrm{CovErr}|$ are means. Lower values favor CA-MFPD.

Benchmark	ΔNLL	ΔCRPS	$\Delta\lvert\mathrm{CovErr}\rvert$
Forrester	+0.737	+0.0430	-0.048
Nonlinear-Sin	-0.035	-0.0120	-0.051
Branin	-0.127	-2.1500	-0.121
Currin	-0.576	+0.0111	-0.101
Park-1	-0.045	-0.0025	-0.027
Park-2	-0.585	-0.0013	+0.017
Hartmann-6	+0.552	+0.0480	+0.063
Borehole	-0.667	-0.1828	+0.002
Heat Grid	-0.116	+0.0035	+0.049
Wing Weight	-0.498	-0.2225	-0.006

Table 7 reports the middle HF budget. The main test uses all seven budgets averaged within each seed.

Table 7: NRMSE at the middle HF budget (mean $\pm$ standard deviation over 20 seeds). Bold indicates the lowest mean in a row and does not imply statistical significance.

Benchmark	N_H	HF-GP	Co-Kriging	Co-RBF	NHLF-CK	CA-MFPD
Forrester	10	0.0814 ± 0.0804	0.0732 ± 0.0934	$\mathbf{0.0726 \pm 0.1140}$	0.0794 ± 0.0917	0.0751 ± 0.0923
Nonlinear-Sin	10	$\mathbf{0.2401 \pm 0.0258}$	0.2519 ± 0.0278	0.3775 ± 0.0778	0.2809 ± 0.0683	0.2415 ± 0.0264
Branin	12	0.0550 ± 0.0237	0.0592 ± 0.0281	$\mathbf{0.0504 \pm 0.0225}$	0.0549 ± 0.0240	0.0583 ± 0.0279
Currin	12	0.0982 ± 0.0244	0.0091 ± 0.0035	0.0137 ± 0.0028	0.0309 ± 0.0330	$\mathbf{0.0090 \pm 0.0035}$
Park-1	20	0.0163 ± 0.0046	0.0030 ± 0.0008	0.0047 ± 0.0005	0.0071 ± 0.0046	$\mathbf{0.0030 \pm 0.0008}$
Park-2	20	0.0238 ± 0.0049	0.0016 ± 0.0004	0.0045 ± 0.0007	0.0032 ± 0.0008	$\mathbf{0.0006 \pm 0.0003}$
Hartmann-6	30	0.1259 ± 0.0199	0.0503 ± 0.0039	0.0599 ± 0.0036	0.0646 ± 0.0061	$\mathbf{0.0501 \pm 0.0040}$
Borehole	40	0.0248 ± 0.0077	0.0036 ± 0.0007	0.0074 ± 0.0008	0.0059 ± 0.0011	$\mathbf{0.0025 \pm 0.0012}$
Heat Conduction Grid	20	0.0450 ± 0.0090	0.0176 ± 0.0032	0.0201 ± 0.0027	0.0181 ± 0.0030	$\mathbf{0.0150 \pm 0.0034}$
Wing Weight	50	0.0233 ± 0.0039	0.0040 ± 0.0004	0.0044 ± 0.0002	0.0071 ± 0.0007	$\mathbf{0.0028 \pm 0.0004}$

C.1 Break-even query analysis

Using the median training and 4,096-point inference times across completed runs, the repeated-batch break-even point against comparator M is

$$Q^*(M) = \frac{T^{\mathrm{CA}}_{\mathrm{train}} - T^{M}_{\mathrm{train}}}{t^{M}_{\mathrm{infer}} - t^{\mathrm{CA}}_{\mathrm{infer}}}. \tag{27}$$

The pooled thresholds are $Q^*(\mathrm{CK}) = 1{,}887$, $Q^*(\mathrm{NARGP}) = 1{,}738$, and $Q^*(\mathrm{MFNN}) = 645$ batches. Per-benchmark thresholds range from 439 to 3,616 against co-kriging and from 408 to 3,255 against NARGP. On three benchmarks, MFNN inference is already as fast as or faster than CA-MFPD, so there is no positive crossing. Figure 6 shows total cost, and Table 8 lists all thresholds. One query unit is one prediction over 4,096 points.

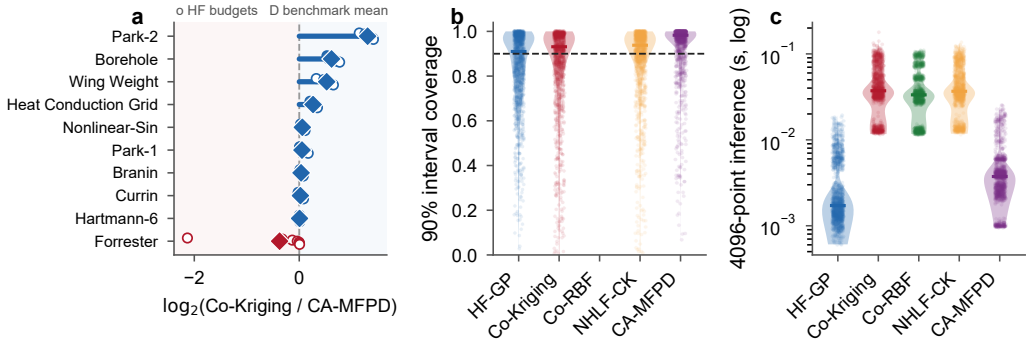

Figure 5: Trade-offs across the ten complete benchmarks. (a) Relative NRMSE over the seven HF budgets, expressed as $\log_2$(Co-Kriging / CA-MFPD), so positive values favor CA-MFPD. Circles denote individual budgets, and diamonds denote benchmark means. (b) Cellwise empirical coverage for nominal 90% intervals; the dashed line marks the target, and coverage is undefined for Co-RBF. (c) End-to-end inference time on 4,096 points, including the optional HF expert where enabled. Timing is machine- and implementation-dependent.

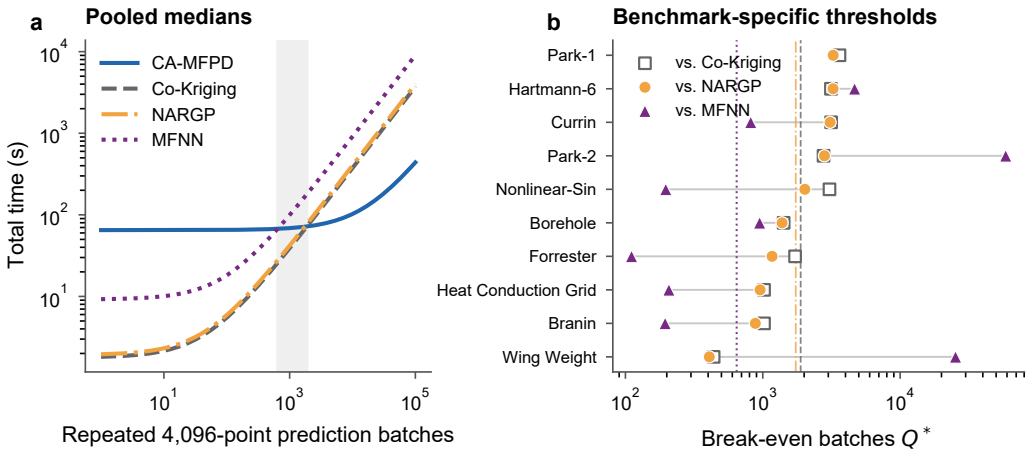

Figure 6: Break-even analysis based on measured runtimes. Panel (a) uses pooled medians for CA-MFPD, co-kriging, NARGP, and MFNN; the shaded band spans the three pooled crossings. Panel (b) shows benchmark-specific thresholds against the three comparators, with vertical lines denoting their pooled values. Each query unit is one 4,096-point prediction batch.

D Three-Benchmark Ablation Screen

The ablation uses Borehole, Park-2, and Wing Weight. The pointwise pseudo-HF control matches the teacher mean on a uniform Sobol design. The uncertainty-weighted control also uses inverse teacher variance. The mean-only variant keeps reliability sampling but removes joint covariance KL. Three other variants set $\lambda_H = 0$, $\tau_H = 1$, or the reliability-mixture weight to zero. Random sampling plus HF-GP and co-kriging are reference methods. Deterministic controls report no uncertainty metrics.

All 450 method–seed–budget runs finished successfully. Within each dataset, the ten methods share the same seeds, HF subsets, LF data, and test set. Standard deviations are computed across seeds after averaging budgets. Exact paired NRMSE tests with Holm correction and LOO-scaling statistics are included in the released analysis materials.

Table 8: Break-even counts of repeated 4,096-point prediction batches computed from benchmark-level median timings.

Benchmark	Q^* vs. Co-Kriging	Q^* vs. NARGP	Q^* vs. MFNN
Borehole	1,418	1,382	949
Branin	1,021	882	195
Currin	3,145	3,091	817
Forrester	1,718	1,168	111
Hartmann-6	3,135	3,248	4,660
Heat Conduction Grid	1,021	956	207
Nonlinear-Sin	3,060	2,024	197
Park-1	3,616	3,255	–
Park-2	2,770	2,812	58,703
Wing Weight	439	408	25,305
Pooled medians	1,887	1,738	645

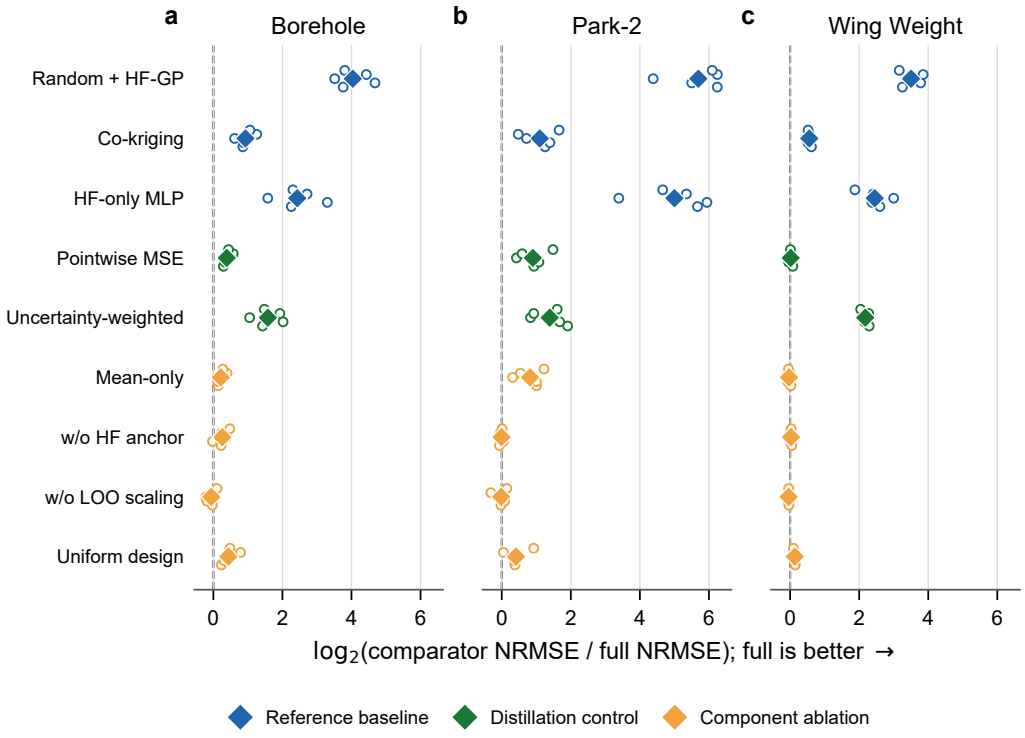

Figure 7: Five-seed, three-budget ablation screen. Each small circle is a seed-level average over the three nested HF budgets; diamonds are means over five seeds. The horizontal coordinate is $\log_2$(comparator NRMSE/full-model NRMSE), so positive values favor the full model and zero denotes parity. No confidence interval is shown because $n = 5$ is too small for stable interval estimation.

E Statistical Scope

Budgets are nested within each seed and share one test set. We therefore average budgets within each benchmark–seed pair before testing. Per-benchmark tests show how results vary across tasks.

Table 9: Ablation-screen NRMSE. Each entry is mean ± standard deviation over five seed-level averages, where each seed average contains the three nested HF budgets. Lower is better. The bold row identifies the full model, not the best entry in every column.

Method	Borehole	Park-2	Wing Weight
CA-MFPD full	$1.728e-03 \pm 3.9e-04$	$7.982e-04 \pm 3.3e-04$	$2.697e-03 \pm 1.2e-04$
Random sampling + HF-GP	$2.838e-02 \pm 6.2e-03$	$4.199e-02 \pm 2.0e-02$	$3.098e-02 \pm 5.9e-03$
Co-kriging	$3.250e-03 \pm 3.6e-04$	$1.646e-03 \pm 4.1e-04$	$3.963e-03 \pm 1.4e-04$
HF-only MLP	$9.461e-03 \pm 3.0e-03$	$2.603e-02 \pm 1.1e-02$	$1.511e-02 \pm 3.9e-03$
Pointwise pseudo-HF MSE	$2.248e-03 \pm 4.1e-04$	$1.435e-03 \pm 3.5e-04$	$2.728e-03 \pm 1.3e-04$
Uncertainty-weighted pseudo-HF	$5.096e-03 \pm 6.0e-04$	$1.970e-03 \pm 2.1e-04$	$1.219e-02 \pm 5.5e-04$
CA-MFPD mean-only	$1.996e-03 \pm 3.8e-04$	$1.339e-03 \pm 2.5e-04$	$2.639e-03 \pm 1.1e-04$
CA-MFPD w/o HF anchor	$2.105e-03 \pm 6.5e-04$	$7.951e-04 \pm 3.3e-04$	$2.746e-03 \pm 1.3e-04$
CA-MFPD w/o LOO scaling	$1.647e-03 \pm 3.0e-04$	$7.815e-04 \pm 3.1e-04$	$2.620e-03 \pm 1.1e-04$
CA-MFPD uniform design	$2.329e-03 \pm 4.7e-04$	$1.053e-03 \pm 3.9e-04$	$2.961e-03 \pm 1.0e-04$