Deferred Aggregation in Hierarchical Bayesian Optimization

Valentin Margraf[1,2] **Jonas Hanselle**[1,2,3] **Julian Rodemann**[4,5]
Marcel Wever[6,7] **Sebastian Vollmer**[3] **Eyke Hüllermeier**[1,2,3]

[1]LMU Munich [2]MCML [3]DFKI [4]RC Trust [5]TU Dortmund
[6]Leibniz University Hannover [7]L3S Research Center

Abstract

To account for uncertainty, hierarchical Bayesian optimization marginalizes over Gaussian process hyperparameters, usually by averaging models or acquisition functions. However, such early aggregation can obscure model disagreement and is sensitive to outliers. We propose to defer aggregation to the decision level: each model independently proposes a candidate, and we aggregate over these candidates using the medoid. We study a misspecified hierarchical Gaussian process setting in which most models have low misspecification error, while a minority are severely misspecified. Theoretically, we derive regret bounds showing that the misspecification penalty depends on the average error across all models for early aggregation, but only on the low-misspecification majority for decision-level aggregation. Empirically, we validate these findings on toy examples and demonstrate competitive performance across synthetic and real-world benchmarks, with particularly strong gains on high-dimensional synthetic problems.

1 Introduction

Bayesian optimization (BO) is a popular framework for optimizing expensive *black-box* functions [1], with applications in materials design [2], drug discovery [3], agent design [4], and hyperparameter optimization (HPO) [5–7]. BO fits a probabilistic surrogate model to observed data. The surrogate provides posterior mean and variance estimates, which are used by an acquisition function (AF) to trade off exploration and exploitation, selecting the next evaluation point until the budget is exhausted.

Gaussian processes (GPs) are the most commonly used surrogate models in BO [8], but their performance is highly sensitive to the choice of hyperparameters [9, 10]. These are typically learned by maximizing the marginal likelihood, a procedure that can be unstable and exhibit high variability [11]. Hierarchical BO offers a robust alternative by sampling hyperparameters from a hyperprior, thereby inducing an ensemble of GP models. Existing approaches aggregate either at the model level, by averaging the posteriors [12], or at the acquisition level, by averaging the AFs [5].

However, this approach has significant limitations. First, aggregating at the model or AF level makes the acquisition objective vulnerable to outliers. For instance, in the design of process systems [13], priors derived from historical data containing sensor failures or abnormal environmental conditions can distort the acquisition objective. Second, the mean fails to capture multimodal distributions. For example, in medical applications such as personalized optimal dosing [14], multiple surrogate models may represent fundamentally different clinical hypotheses. Finally, aggregation inherently loses information: the model disagreement that signals critical epistemic uncertainty is often obscured once models are averaged.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

To counter these problems, we distinguish three levels at which aggregation can occur: the GP *model*, the *acquisition* function, and the *decision* about the point to be evaluated (cf. Figure 1). While prior work has aggregated at the model or acquisition level, the decision level has so far been overlooked. We formalize this view in a unifying framework that treats the *level of aggregation* as an explicit design dimension and recovers existing mean-based approaches as special cases. Within this framework, we propose to defer aggregation to the decision level: each GP's acquisition function is maximized independently, and aggregation reduces to selecting among the resulting candidates. This mitigates the influence of outlier models on the optimization objective and improves robustness in Bayesian optimization. Our **contributions** are as follows:

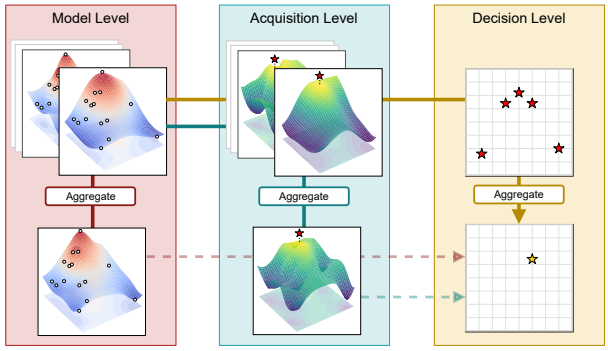

Figure 1: In hierarchical BO, information can be aggregated at the **model-**, the **acquisition-** or the **decision-level**.

- We propose decision-level aggregation in hierarchical BO, formalized within a unifying framework that treats the *level of aggregation* as an explicit design dimension and recovers existing mean-based approaches as special cases.
- We study a misspecified hierarchical GP setting where most models are accurate but a minority are severely misspecified, and derive regret bounds whose misspecification penalty depends on the *average* error across all models for mean aggregation, but only on that of the low-misspecification majority for our decision-level aggregation.
- We empirically demonstrate that decision-level aggregation performs competitive across low- to high-dimensional synthetic and real-world benchmarks, with particularly pronounced gains on high-dimensional synthetic problems.

2 Hierarchical Bayesian Optimization

We consider the problem of sequentially optimizing a *black-box* function $f : \mathcal{X} \to \mathbb{R}$. We assume homoscedastic noise, observing

$$y_t = f(x_t) + \xi_t \tag{1}$$

with $\xi_t \sim \mathcal{N}(0, \lambda)$. Our goal is to identify the global minimizer of f, i.e., $x^\star = \arg\min_{x \in \mathcal{X}} f(x)$, while minimizing the number of evaluations.

We begin by specifying a prior over the objective function, which encodes our initial belief. At time step t, we have a dataset of observations $\mathcal{D}_t = \{(x_i, y_i)\}_{i=1}^t$ to fit a surrogate model such as a Gaussian process. A GP is fully specified by a prior mean function, often assumed to be constant zero, and a kernel function $k_\theta(x, x')$, parameterized by hyperparameters $\theta = \{l, \sigma_f, \sigma_\epsilon\}$, where $l \in \mathbb{R}^d$ is a vector of lengthscales, $\sigma_f \in \mathbb{R}$ is the output scale, and $\sigma_\epsilon \in \mathbb{R}$ is the noise standard deviation. The GP posterior predictive mean and variance at a test point x at timestep t can be derived analytically [8]:

$$\mu_t(x) = \mathbf{k}_t(\mathbf{x})^\top (\mathbf{K}_t + \Sigma_t)^{-1} \mathbf{y}_t, \tag{2}$$

$$\sigma_t^2(x) = k_\theta(x, x) - \mathbf{k}_t(\mathbf{x})^\top (\mathbf{K}_t + \Sigma_t)^{-1} \mathbf{k}_t(\mathbf{x}), \tag{3}$$

where $\mathbf{y}_t = [y_1, \ldots, y_t]^\top$, $\mathbf{K}_t = [k_\theta(x_i, x_j)]_{i,j}$ is the kernel matrix over observed points, $\Sigma_t = \mathrm{diag}(\sigma_\epsilon^2, \ldots, \sigma_\epsilon^2)$, and $\mathbf{k}_t(\mathbf{x}) = [k_\theta(x_1, x), \ldots, k_\theta(x_t, x)]^\top$. Hereby, accurate estimation of θ is crucial for uncertainty quantification and therewith the performance of Bayesian optimization [9].

Hierarchical BO [12, 5] offers a principled approach by placing a hyperprior $P(\theta)$ over the parameters. Predictions at a test point x are then obtained by marginalizing the predictive posterior of the latent function over θ: $p(f(x) \mid \mathcal{D}_t) = \int p(f(x) \mid \mathcal{D}_t, \theta)\, P(\theta \mid \mathcal{D}_t)\, d\theta$. The resulting hierarchical predictive posterior is typically approximated by sampling hyperparameters $\theta^{(j)} \sim P(\theta \mid \mathcal{D}_t)$ and averaging the corresponding predictive posterior distributions: $p(f(x) \mid \mathcal{D}_t) \approx \frac{1}{M} \sum_{j=1}^M p(f(x) \mid \mathcal{D}_t, \theta^{(j)})$ [15].

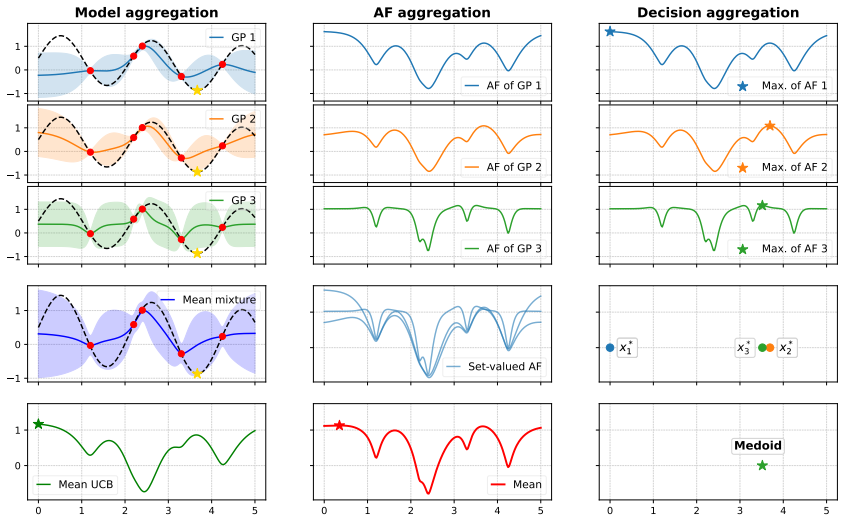

Figure 2: Hierarchical BO with three GPs: GPs 2 and 3 approximate the target function (dotted line) well, while GP 1 is an outlier. At the **model** and **AF level**, standard mean aggregation lets GP 1 corrupt the objective used to select the next query, yielding maximizers near $x \approx 0$ and $x \approx 0.2$. At the **decision level**, each AF is maximized individually and aggregation only chooses among the resulting candidates: GP 1 can only influence which of (good) GP 2's or GP 3's candidates is selected; the query lands near the true optimum at $x \approx 3.5$.

3 Related Work

A number of works aim to improve robustness in Bayesian optimization by modifying a single surrogate model. Berkenkamp et al. [16] iteratively adapt kernel hyperparameters to enlarge the reproducing kernel Hilbert space (RKHS) until it contains the target function; Bogunovic and Krause [17] augment standard GP-UCB with an additional exploration term to account for model misspecification; and Wynne et al. [18] propose kernel adjustments under misspecified smoothness and likelihood assumptions. More recently, Rodemann and Augustin [19] introduce imprecise GPs to mitigate prior mean misspecification. While these approaches improve robustness for a single model, they do not account for uncertainty across multiple plausible models.

Another line of work maintains multiple priors or GP models and selects one per BO iteration. Pautrat et al. [20] choose the prior maximizing the product of likelihood and AF value, while Polyzos et al. [21] sample from an ensemble of GPs with probability proportional to likelihood. Ziomek et al. [22] consider time-varying domains, discarding implausible priors and selecting the model with the highest UCB value. These approaches still rely on a single model, which can be risky under outliers.

A different family of approaches aggregates multiple models or acquisition functions. Riis et al. [23] consider fully Bayesian GPs for active learning and propose acquisition functions that either perform query-by-committee or use a mixture-of-GPs formulation. In the context of HPO, Wistuba et al. [24] train a GP per dataset and combine the resulting surrogates into a single weighted prediction for downstream decisions. In the same spirit, Feurer et al. [25] construct a linear combination of GP models for transfer learning, aggregating both predictions and acquisition functions according to a predefined weighting scheme. Snoek et al. [5] and Eriksson and Jankowiak [26] sample hyperparameters from a prior and average the resulting AF values. In our terminology, these methods implement mean aggregation at the acquisition-function level (cf. Figure 2).

Existing aggregation-based methods aggregate at the model or acquisition level, almost always with the mean. To our knowledge, no prior work has formalized the aggregation stage as a design dimension in hierarchical BO, studied decision-level aggregation as a principled alternative, or analyzed how the choice of level affects regret under misspecification. Our work fills this gap: we organize aggregation by level, recover existing mean-based methods as special cases, and show theoretically and empirically that deferring aggregation to the decision level provides robustness.

4 Misspecified Hierarchical GP-UCB

In most theoretical analyses of GP-based BO, a common assumption is *realizability*, meaning that the unknown target function f lies in the hypothesis class of the GP, i.e., $f \in \mathcal{F}_{k_\theta}(\mathcal{X}, B) = \{g \in \mathcal{H}_{k_\theta}(\mathcal{X}) : \|g\|_{\mathcal{H}_{k_\theta}} \leq B\}$, where $\mathcal{H}_{k_\theta}(\mathcal{X})$ is the RKHS of the kernel k_θ, parameterized by hyperparameters θ, and $B > 0$ is a constant. Under this assumption, the GP can learn f, and the GP-UCB algorithm achieves sublinear regret [27]. We assume throughout that the kernel is normalized such that $k(x, x) \leq 1$ for all $x \in \mathcal{X}$, which is standard in the GP-UCB literature [27, 28].

In practice, however, realizability rarely holds due to, e.g., kernel misspecification or inaccuracies in hyperparameter estimation [16]. This motivates relaxed formulations, such as the misspecified kernelized bandit setting considered by Bogunovic et al. [29]. In this setting, the GP is ε-misspecified, in the sense that the target function f does not lie in the GP's hypothesis space $\mathcal{F}_{k_\theta}(\mathcal{X}, B)$, but is ε-far from a function g that does so. Formally, one assumes

$$\min_{g \in \mathcal{F}_{k_\theta}(\mathcal{X}, B)} \|g - f\|_\infty \leq \varepsilon, \tag{4}$$

so that $|g(x) - f(x)| \leq \varepsilon$ for every $x \in \mathcal{X}$.

An equivalent interpretation of this setting is to assume that the GP employs the correct kernel (i.e., f lies in the GP's hypothesis class), but the function evaluations that the GP receives are corrupted by a uniformly bounded, deterministic bias of magnitude at most ε. Similar to [17], we adopt this viewpoint, since it allows the misspecification error to be treated explicitly as bounded observation bias and leads to a more tractable analysis.

We now extend this setting to a hierarchical framework. In the practical algorithm, hyperparameters $\{\theta^{(j)}\}_{j=1}^M$ are drawn from a hyperprior, yielding an ensemble of GPs with distinct kernels $k_{\theta^{(j)}}$ and consequently distinct posterior means and variances. For tractability of the analysis, we extend the abstraction above to the ensemble: rather than modeling kernel-induced diversity explicitly, we treat the M ensemble members as sharing the same kernel $k_{\theta^{(j)}}$, each receiving evaluations corrupted by a member-specific bias $m_j : \mathcal{X} \to [-\varepsilon_j, \varepsilon_j]$. Concretely, for each $j \in \{1, \ldots, M\}$ there exists a function $g_j \in \mathcal{F}_k(\mathcal{X}, B)$ such that

$$y_i^{(j)} \;=\; \underbrace{g_j(x_i) + m_j(x_i)}_{f(x_i)} + \xi_i, \quad ||m_j||_\infty \leq \varepsilon_j \tag{5}$$

with ε_j known. We denote the posterior mean of the j-th GP by $\mu_{j,t-1}(x)$, and write $\sigma^2_{t-1}(x)$ for the posterior variance, which is shared across j under this abstraction. An extension of the analysis to ensembles with distinct posterior variances is provided in Appendix A.4.

We select the next candidate by maximizing the following acquisition function,

$$x_t \in \arg\max_{x \in \mathcal{X}} -\mu^{\text{avg}}_{t-1}(x) + \left(\beta_t + \frac{\varepsilon_{\text{avg}}\sqrt{t}}{\sqrt{\lambda}}\right) \sigma_{t-1}(x), \tag{6}$$

where $\mu^{\text{avg}}_{t-1}(x) = \frac{1}{M}\sum_{j=1}^M \mu_{j,t-1}(x)$, $\varepsilon_{\text{avg}} = \frac{1}{M}\sum_{j=1}^M \varepsilon_j$ and β_t is the exploration parameter that stems from the high-probability bound in [27]. Note that we take the negative posterior mean, since the optimization objective is the minimization of f.

Further, we denote by $\mu^*_{t-1}(x)$ the posterior mean of the (oracle) GP which receives the non-biased observations $f(x_i) + \xi_i$. We can then derive the following bound on the cumulative regret $R_T = \sum_{t=1}^T (f(x_t) - \min_{x \in \mathcal{X}} f(x))$ and provide the proof in Appendix A.1.

Theorem 1 (Misspecified Hierarchical GP-UCB)**.** *Suppose M GP models with kernels k and corresponding hypothesis classes $\mathcal{F}_k(\mathcal{X}, B)$ for some fixed $B > 0$, with $\mathcal{X} \subset \mathbb{R}^d$. Let f be the true function s.t. $\|f - g_j\|_\infty \leq \varepsilon_j$, for all $j = 1, \ldots, M$. With $\varepsilon_{\text{avg}} := \frac{1}{M}\sum_{j=1}^M \varepsilon_j$ and γ_T the maximum information gain associated with kernel k, the hierarchical acquisition rule* (6) *guarantees, with probability at least $1 - \delta$, a cumulative regret bound of the order*

$$O\Big(B\sqrt{\gamma_T T} + \sqrt{(\ln(1/\delta) + \gamma_T)\,\gamma_T T} + \varepsilon_{\text{avg}} T\sqrt{\gamma_T}\Big). \tag{7}$$

Table 1: Aggregated AFs with corresponding uncertainty preferences and denote by $\mathbb{E}_q := \mathbb{E}_{q\in\mathcal{Q}(x)}[q]$ and $\mathrm{Var}_q := \mathrm{Var}_{q\in\mathcal{Q}(x)}[q]$ the mean and variance with $\lambda > 0$.

Aggregated AF	$\mathbb{E}_q - \gamma\,\mathrm{Var}_q$	$\mathbb{E}_q$	$\mathbb{E}_q + \gamma\,\mathrm{Var}_q$	$\min\mathcal{Q}(x)$	$\mathrm{median}\,\mathcal{Q}(x)$	$\max\mathcal{Q}(x)$
Uncertainty preference	averse	neutral	affine	averse	neutral	affine

The bound closely resembles that of [29]: The first two terms correspond to the standard GP-UCB regret components and the third term captures the effect of model misspecification. The misspecification penalty in this case is based on the average $\varepsilon_{\mathrm{avg}}$ across the mixture of GP models.

So far, we have made no assumptions about the magnitude of the misspecification errors ε_j. If the ε_j happen to be small, the mixture behaves well and the regret bound is tight. In practice, however, some models may be severely misspecified, leading to large ε_j. Such outlier models can disproportionately influence the mixture posterior and distort the acquisition function, either by making it overly conservative or by steering it toward incorrect regions of the domain. This inflates $\varepsilon_{\mathrm{avg}}$ and worsens the regret bound.

5 Deferred Aggregation

In this section, we develop the framework of *deferred aggregation*. We organize aggregation by level—model, acquisition, or decision—and show that deferring aggregation to later stages reduces the influence of outlier models while preserving information about model disagreement.

We consider a setting in which some GP models act as outliers and are not representative of the underlying function, while the majority are well-specified. This scenario is illustrated in Figure 2, where one of three models (GP 1) is misspecified, and mean aggregation—whether at the model level or the acquisition level—leads to poor decisions. Conceptually, the hierarchical GP induces a finite set of posterior distributions, which we treat as a sample from a credal set in the sense of imprecise probability theory [30]; each posterior gives rise to an acquisition function, and we obtain a set-valued acquisition function over which aggregation is performed.

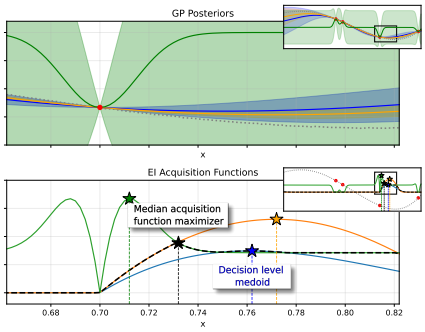

Figure 3: A misspecified model (green) distorts the aggregated AF (dashed black); at the decision-level it only affects which candidate from the well-specified majority is selected.

Formally, each hyperparameter sample $\theta^{(j)}$ produces a GP posterior $p(f(x) \mid \mathcal{D}_t, \theta^{(j)}) := p(f^{(j)})$ with mean $\mu_{\theta^{(j)}}$ and variance $\sigma^2_{\theta^{(j)}}$. We denote the acquisition function under this posterior by $\alpha_{\theta^{(j)}}$. Collecting the AF values across all M sampled hyperparameters produces a set of plausible AFs for every x. Hence, the standard maximization criterion to select the next candidate no longer applies, but an aggregation step is required to compare candidates and select the next to evaluate.

5.1 Acquisition Level Aggregation

Definition 1 (Set of plausible acquisition values)**.** *Given the set of AFs $\{\alpha_{\theta^{(j)}}\}_{j=1}^{M}$, the set of plausible AF values at candidate point $x \in \mathcal{X}$ under posterior hyperparameter uncertainty is defined as follows:*

$$\mathcal{Q}(x) := \{\alpha_{\theta^{(1)}}(x), \ldots, \alpha_{\theta^{(M)}}(x)\} \tag{8}$$

Given such a set, we can aggregate information differently by mapping each set to a single scalar.

Definition 2 (Aggregated acquisition function)**.** *Given a set of plausible acquisition values $\mathcal{Q}(x)$ at candidate point $x \in \mathcal{X}$, an aggregated acquisition function (AAF) is a mapping*

$$\rho : \mathcal{Q}(x) \;\to\; \rho(\mathcal{Q}(x)) \in \mathbb{R}, \tag{9}$$

that assigns a single real value to the set of plausible acquisition values of the given candidate.

Table 1 presents several AAF variants carefully constructed to represent different unertainty preferences: uncertainty-neutral, uncertainty-affine, or uncertainty-averse. The mean ($\gamma = 0$) recovers the standard aggregation strategy of [5, 26]; similarly, $\min \mathcal{Q}(x)$ and $\max \mathcal{Q}(x)$ correspond to the Γ-maximin (uncertainty-averse) and Γ-maximax (uncertainty-affine) criteria [31]; the pointwise median provides a risk-neutral alternative robust to outliers among the plausible AF values. Note that such pointwise median rule can also be implemented at the model level by taking the pointwise median of the posterior mean and uncertainty across all models. The next candidate is selected by maximizing the AAF using standard optimization methods; pseudocode is given in Algorithm 1 in the Appendix.

Robust operators at the acquisition level reduce the influence of misspecified models but do not eliminate it: the aggregated AF is still a function of every model (including outliers) so the optimization objective itself remains corrupted (cf. Figure 3). The next subsection shows that deferring aggregation to the decision level removes this corruption entirely.

5.2 Decision Level Aggregation

We now consider the most deferred aggregation level: each acquisition function is maximized independently, and aggregation occurs over the resulting candidate maximizers.

Definition 3 (Set of candidate maximizers)**.** *Let $x_j^* \in \arg\max_x \alpha_{\theta^{(j)}}(x)$, then we define the set of AF maximizers as*

$$\mathcal{X}^* := \{\arg\max_{x \in \mathcal{X}} \alpha_{\theta^{(j)}}(x)\}_{j=1}^M = \{x_j^*\}_{j=1}^M. \tag{10}$$

If the AF is interpreted as expected utility, the set $\mathcal{X}^*$ consists of points that maximize Walley's maximality criterion [30]: each point maximizes expected utility for at least[1] one probability measure (here: one $p(f^{(j)})$) from the ones under consideration (here: all $\{p(f^{(j)})\}_{j=1}^M$). A natural and robust aggregation strategy is to select the medoid of $\mathcal{X}^*$, i.e.,

$$\arg\min_{x \in \mathcal{X}^*} \sum_{j=1}^{M} \|x - x_j^*\|_2. \tag{11}$$

The analysis below relies on (i) robustness to a minority of outliers and (ii) membership preservation, i.e., the selected point must belong to $\mathcal{X}^*$. The medoid is thus only one possible robust aggregation rule at the decision level. Alternatives include the weighted medoid, as well as the coordinate-wise median or trimmed centroid combined with a projection onto $\mathcal{X}^*$. In Appendix A.2, we discuss these alternatives and extend our theoretical analysis.

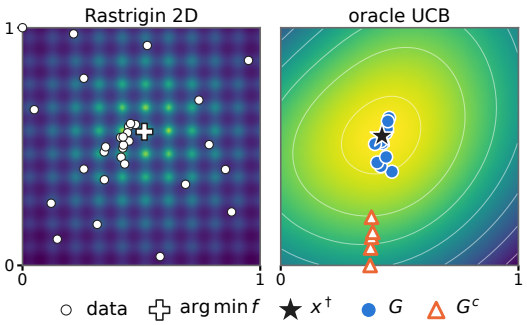

Figure 4: Separation in Assumption 1 on 2D Rastrigin after 15 BO iterations. The UCB maximizers of the low-error models G cluster around $x^\dagger$; those of the high-error models G^c lie far from it.

To establish formal regret guarantees for the medoid rule, we partition the M models by misspecification: a strict majority of *low-misspecification models* with $0 \le \varepsilon_i \le \rho$ for some small $\rho > 0$, indexed by $i \in G$ with $|G| > M/2$; and *outlier models* with $\varepsilon_l \gg \rho$ for $l \in G^c$. Let α^* denote the oracle AF (under the posterior receiving unbiased observations) and $x^\dagger \in \arg\max_x \alpha^*(x)$ its maximizer. We impose the following structural assumption, which captures both local regularity of the oracle AF and a separation between well-specified and misspecified models.

Assumption 1. *There exist $\kappa > 0$ and a neighborhood $\mathcal{N}$ of $x^\dagger$ such that*

$$\alpha^*(x) \;\le\; \alpha^*(x^\dagger) - \tfrac{\kappa}{2}\|x - x^\dagger\|_2^2 \qquad \text{for all } x \in \mathcal{N},$$

and $x_i^* \in \mathcal{N}$ for all $i \in G$, while $x_l^* \notin \mathcal{N}$ for all $l \in G^c$. The misspecification level ρ is small enough that the localization ball from Proposition 1 is contained in $\mathcal{N}$.

[1] It may happen that $x_i^* = x_k^*$ for $i \neq k$. Hence, $\mathcal{X}^*$ contains at most M elements.

Remark 1. *Assumption 1 combines two elements: (i) a local quadratic structure of the oracle acquisition function α^* around its maximizer $x^\dagger$, and (ii) a separation condition ensuring that maximizers of low-misspecification models cluster near $x^\dagger$, while those of highly misspecified models lie outside this region. The second condition can be interpreted as an identifiability requirement at the level of acquisition maximizers. In Figure 4 we illustrate this separation on the 2D Rastrigin function; and we provide a detailed discussion in Appendix A.2 and further empirical evidence across test functions and dimensions in Appendix A.3.*

Under Assumption 1, a majority argument formalizes the geometric picture from Figure 3.

Proposition 1 (Localization of the medoid)**.** *Under Assumption 1 and $|G| > M/2$, the medoid x_t selected by* (11) *satisfies*

$$\|x_t - x^\dagger\|_2 \;\le\; C\sqrt{\rho\sqrt{t}},$$

where $C := \frac{2}{\sqrt{\kappa}\,\lambda^{1/4}} \cdot \frac{3|G|+|G^c|}{|G|-|G^c|}$. In particular, $x_t = x^*_{j(t)}$ for some $j(t) \in G$.

The proof, in Appendix A.2, has two steps. By Lemma 2 of [17] and Assumption 1, every good maximizer x_i^* ($i \in G$) lies in a small ball around $x^\dagger$. Since $|G| > M/2$, the medoid $\mathcal{X}$ rule then ensures that x_t is the maximizer of a low-misspecification model.

Theorem 2 (Regret of medoid $\mathcal{X}$ aggregation)**.** *Suppose M GP models share kernels k and hypothesis class $\mathcal{F}_k(\mathcal{X}, B)$, and Assumption 1 holds. Assume further that a strict majority of the models are low-misspecification with $\varepsilon_i \le \rho$. Then the medoid $\mathcal{X}$ rule guarantees, with probability at least $1-\delta$, a cumulative regret bound of the order*

$$O\Big(B\sqrt{\gamma_T T} + \sqrt{(\ln(1/\delta) + \gamma_T)\,\gamma_T T} + \rho\, T\sqrt{\gamma_T}\Big).$$

By Proposition 1, $x_t = x^*_{j(t)}$ with $j(t) \in G$, hence $\varepsilon_{j(t)} \le \rho$. The bound then follows from the misspecified GP-UCB analysis of [17] applied to model $j(t)$; details in Appendix A.2.

Remark 2. *The bounds in Theorems 1 and 2 differ in their misspecification term: ε_{avg} for mean aggregation, ρ for the medoid $\mathcal{X}$. Which term dominates depends on how the misspecification is distributed amongst the members, and thus depends on the hyperposterior. If nearly all models have low misspecification, so $\varepsilon_{\text{avg}} \approx \rho$, the bounds do not establish which rule performs better. If most models have high misspecification, the majority condition $|G| > M/2$ fails, and no aggregation rule can be expected to perform well. In the intermediate regime targeted by our analysis ($\rho \ll \varepsilon_{\text{avg}}$), the medoid bound is strictly tighter.*

6 Experiments

To evaluate our framework empirically, we implement it in BoTorch [32]. Given the flexibility of our approach, many combinations of probabilistic models, aggregation strategies, and acquisition functions are possible. In this paper, we therefore focus on a representative subset of methods, with additional experimental results and implementation details provided in the supplement. The full implementation is available at `https://github.com/ValentinMargraf/deferred_aggregation_hierarchical_bo`. We consider two AFs: upper confidence bound (UCB) and expected improvement (EI). Our main method is decision-level medoid $\mathcal{X}$ aggregation. We also evaluate robust aggregators at earlier stages: F.B. median (model level), Median $\mathcal{Q}$, and UCB $\mathcal{Q}$ (AF level). Standard baselines are mean aggregation at the model level (F.B. mean) and AF level (mean $\mathcal{Q}$) [5], plus a single GP with hyperparameters fit by maximum marginal likelihood (UCB with $\beta \in \{1, 2\}$ or EI). Unless stated otherwise, all hierarchical methods use $M = 16$ models. Baselines are shown in black with distinct linestyles, our methods in color.

6.1 Toy Examples

First, we study toy examples, where we can explicitly control the number of well-specified models and severely misspecified outliers. We minimize a sample from a GP with an RBF kernel with lengthscale 0.1 using three GPs differing only in lengthscale: two near the ground truth (0.098, 0.103) and one outlier (5), which produces an overly smooth posterior and underestimates uncertainty, strongly

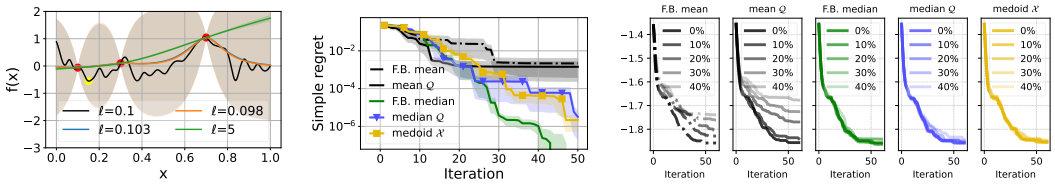

Figure 5: **Left:** Sample from a 1D GP with lengthscale 0.1, optimized using three GPs with different lengthscales (0.103, 0.098, 5). **Middle:** Simple regret averaged over 10 seeds. **Right:** Best function value found on a 1D GP sample with lengthscale 0.2; using an ensemble of 9 GPs, with an increasing fraction severely misspecified (0%, 10%, . . . , 40%), averaged over 100 seeds.

biasing mean-based aggregation. Figure 5 depicts the three fitted GPs (left) and optimization results (middle). Mean-based methods stagnate early due to the outlier, while median-based aggregation remains robust and achieves substantially lower regret.

In a second toy example, we again minimize a sample from a 1D GP, now with a lengthscale 0.02. For the models that we aggregate, an increasing fraction is severely misspecified, using a large lengthscale of approximately 5; the remaining models have lengthscales close to the correct lengthscale 0.02. Across all levels of misspecification, median-based aggregation methods remain robust. In contrast, the performance of mean-based approaches deteriorates rapidly as the number of outlier models increases, as can be seen in Figure 5. Overall, these toy experiments suggest that in this simple 1D setting, robust aggregation at earlier stages (model or AF level) already suffices. The benefit of deferring aggregation to the decision level becomes apparent only in more challenging settings.

6.2 Low-dimensional Synthetic Functions

We next evaluate whether the benefits observed in our toy examples extend to standard BO benchmarks. Rather than explicitly constructing outliers via different lengthscales, we sample hyperparameters $\theta = \{l, \sigma_f, \sigma_\epsilon\}$ from the hyperposterior induced by the chosen hyperprior. This may naturally introduce outliers due to the heavy-tailed nature of the distributions, e.g., consider the BoTorch lengthscale hyperprior: $\ell \sim \text{LogNormal}(\sqrt{2} + \log d/2, \sqrt{3})$. We draw GP hyperparameters from this hyperprior and evaluate on common low-dimensional synthetic benchmark functions: Levy, Rosenbrock, and Ackley in dimensions 3, 4, and 6, with standard deviation 10^{-3}.

Overall, as shown in Figure 6, the hierarchical approaches outperform the single-GP baseline, while differences among the aggregation strategies remain comparatively small in this low-dimensional setting. The medoid $\mathcal{X}$ rule is on par with the strongest aggregation strategies across all three functions and both AFs, and achieves the lowest regret on Ackley with UCB.

The individual benchmarks suggest different regimes. On Rosenbrock, the strongly non-stationary landscape is difficult to capture with the sampled stationary GP models, so many models may be substantially misspecified and robust aggregation has limited room to help. On Levy, all methods reach very small regret, suggesting that the sampled models are generally adequate and that even earlier-stage aggregation is sufficient. Ackley appears to lie between these two extremes: here, the choice of aggregation matters more, as F.B. median stalls with EI while medoid $\mathcal{X}$ remains among the best methods for both AFs.

6.3 High-dimensional Synthetic Functions

We next evaluate on high-dimensional synthetic benchmarks using the sparse SAASBO hyperprior [26] with a Matérn-5/2 kernel, designed for high-dimensional settings. We consider Levy in 32 dimensions and Ackley and Griewank in 50 dimensions, with both UCB and EI. Fitting a single GP via maximum likelihood is challenging in high dimensions and leads to poor optimization performance, as can be seen in Figure 7. Hierarchical aggregation consistently outperforms the single GP baseline, highlighting the benefits of marginalizing over hyperparameters. On Levy, all hierarchical methods, including earlier-stage robust aggregators, perform comparably. On Griewank, the medoid $\mathcal{X}$ rule clearly outperforms all other methods for both AFs, and on Ackley it does so when using UCB as AF. These results suggest that high-dimensional problems more often fall into

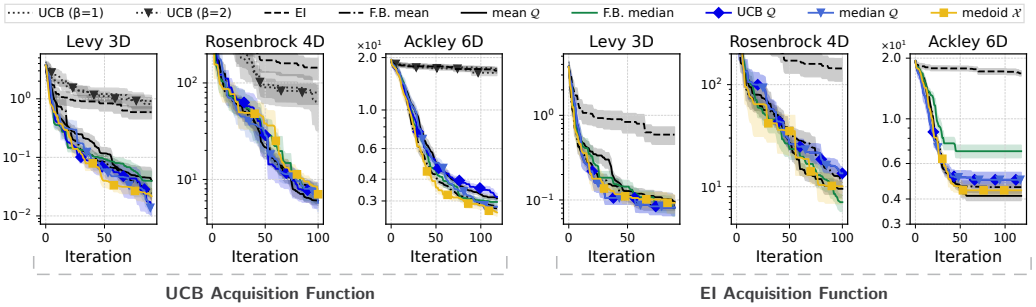

Figure 6: Log-regret using the **default BoTorch hyperprior** averaged over 30 random seeds, shaded regions indicate $\pm$ one standard error.

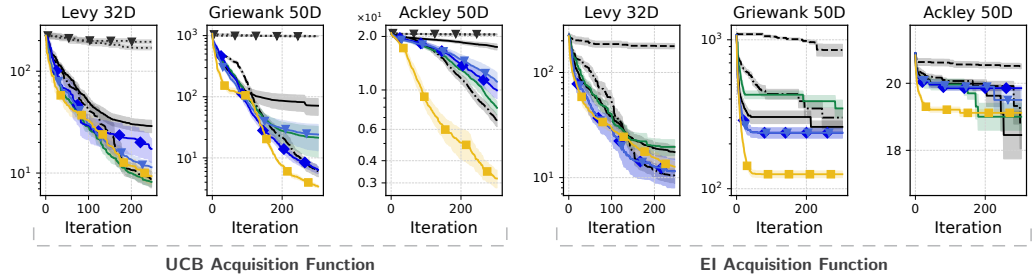

Figure 7: Log-regret using the **SAASBO hyperprior** averaged over 30 random seeds, shaded regions indicate $\pm$ one standard error.

the regime targeted by our theory: with many hyperparameters, some models are likely severely misspecified, which is precisely where deferring aggregation to the decision level pays off.

6.4 Real-world Benchmarks

We further evaluate our approach on three real-world high-dimensional optimization tasks (Figure 8), again using the SAASBO hyperprior: (1) Lasso hyperparameter tuning on the Breast Cancer dataset from LassoBench [33]; (2) robot pushing, a 14-dimensional controller tuning task for two robot hands pushing two objects to target locations; and (3) rover trajectory planning, a 60-dimensional problem where a B-spline with 30 2D points parameterizes a rover path over rough terrain. Hierarchical approaches again outperform the single-GP baseline in most cases, but differences among aggregation strategies are generally small. On Breast Cancer, all methods converge to similar values. On robot pushing, model-level aggregation performs best, whereas on rover trajectory the medoid $\mathcal{X}$ rule remains competitive with UCB but falls behind AF-level aggregation with EI. One possible explanation is that the SAAS prior is well matched to these tasks: it is designed for high-dimensional problems with only a small number of relevant input dimensions [26], which may reduce the prevalence of severely misspecified models in the ensemble and thus leaves less room for decision-level aggregation to improve upon it.

To assess the statistical significance of these findings, we conduct pairwise Wilcoxon signed-rank tests across all benchmarks (cf. Appendix B.3). While many differences between hierarchical aggregation strategies are not statistically significant, a clear pattern emerges: medoid $\mathcal{X}$ achieves the most significant wins and only a single significant loss across 252 pairwise comparisons. In contrast, every other aggregation strategy incurs at least nine significant losses. Thus, although decision-level aggregation does not consistently outperform earlier-stage alternatives, it exhibits particularly robust performance across different problems, dimensionalities, and acquisition functions.

Remark 3 (Computational cost)**.** *Decision-level aggregation requires M AF optimizations per iteration instead of one. With $M = 16$, this introduces a clear overhead but remains modest compared to both MCMC-based GP inference and function evaluations in typical applications. Moreover, the M optimizations are independent and trivially parallelizable, so the overhead can be largely eliminated in practice. Detailed runtimes are reported in Appendix Table 7.*

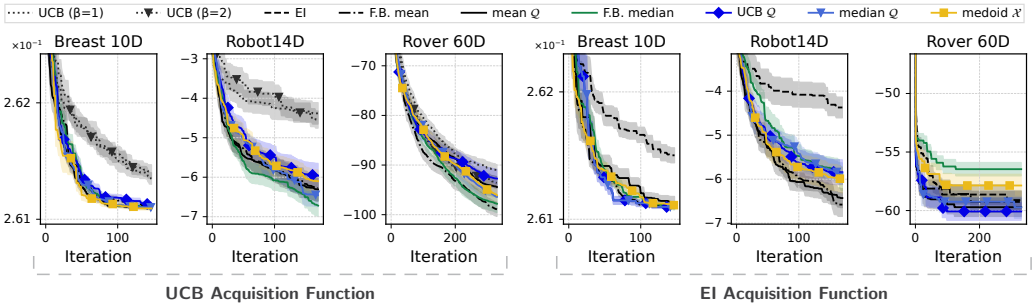

Figure 8: Regret using the **SAASBO hyperprior** averaged over 30 random seeds, shaded regions indicate $\pm$ one standard error.

7 Conclusion and Future Work

We introduced *deferred aggregation* as an explicit design dimension in hierarchical Bayesian optimization. Our framework organizes aggregation by level (model, acquisition, or decision) and recovers existing mean-based methods as special cases. Theoretically, we showed that mean aggregation is highly sensitive to outlier models, while decision-level medoid aggregation enjoys regret bounds that depend only on the well-specified majority. Empirically, decision-level aggregation performs competitively across synthetic and real-world benchmarks, with particularly pronounced improvements on high-dimensional synthetic problems.

Several directions remain for future work. First, our framework could be extended to more sophisticated acquisition functions, e.g., those handling heteroscedastic noise like RAHBO [34]. Second, alternative decision criteria from the imprecise probability literature could accommodate user-specific risk preferences. Third, investigating the role of the hyperprior, both theoretically and empirically, could yield further insight into when deferred aggregation is most beneficial. Finally, decision-level aggregation also allows candidate decisions to originate from different model classes, rather than only from different hyperparameterizations of the same model family. Exploring such heterogeneous model ensembles is an interesting direction for future work.

Limitations and Broader Impact Statement Our theoretical analysis assumes a strict majority of low-misspecification models, which may not always hold in practice. Further, decision-level aggregation requires M AF optimizations per iteration. Although these optimizations are independent and can be parallelized, they introduce additional computational cost.

This paper advances the methodology of Bayesian optimization. We do not foresee specific societal risks beyond those generic to improved optimization techniques.

Acknowledgments and Disclosure of Funding

The authors gratefully acknowledge the computational and data resources as well as the support provided by the Leibniz Supercomputing Centre. Marcel Wever acknowledges funding by the European Union (ERC, "ixAutoML", grant no. 101041029). Views and opinions expressed are, however, those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

References

[1] R. Garnett. *Bayesian Optimization*. Cambridge University Press, 2023. Available for free at `https://bayesoptbook.com/`.

[2] Yichi Zhang, Daniel W. Apley, and Wei Chen. Bayesian optimization for materials design with mixed quantitative and qualitative variables. *Scientific Reports*, 10(4924), 2020.

[3] Léa Colliandre and Céline Muller. Bayesian optimization in drug discovery. *Methods in Molecular Biology*, 2716:101–136, 2024. doi: 10.1007/978-1-0716-3449-3_5.

[4] Y. Chen, A. Huang, Z. Wang, I. Antonoglou, J. Schrittwieser, D. Silver, and N. de Freitas. Bayesian optimization in alphago. *arXiv:1812.06855 [cs.LG]*, 2018.

[5] J. Snoek, H. Larochelle, and R. Adams. Practical Bayesian optimization of machine learning algorithms. In P. Bartlett, F. Pereira, C. Burges, L. Bottou, and K. Weinberger, editors, *Proceedings of the 26th International Conference on Advances in Neural Information Processing Systems (NeurIPS'12)*, pages 2960–2968. Curran Associates, 2012.

[6] M. Lindauer, K. Eggensperger, M. Feurer, A. Biedenkapp, D. Deng, C. Benjamins, T. Ruhkopf, R. Sass, and F. Hutter. SMAC3: A versatile bayesian optimization package for Hyperparameter Optimization. *Journal of Machine Learning Research*, 23(54):1–9, 2022.

[7] B. Bischl, M. Binder, M. Lang, T. Pielok, J. Richter, S. Coors, J. Thomas, T. Ullmann, M. Becker, A.-L. Boulesteix, D. Deng, and M. Lindauer. Hyperparameter optimization: Foundations, algorithms, best practices, and open challenges. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, page e1484, 2023.

[8] C. Rasmussen and C. Williams. *Gaussian Processes for Machine Learning*. The MIT Press, 2006.

[9] Willie Neiswanger and Aaditya Ramdas. Uncertainty quantification using martingales for misspecified gaussian processes. In *Algorithmic Learning Theory, 16-19 March 2021, Virtual Conference, Worldwide*, Proceedings of Machine Learning Research, pages 963–982. PMLR, 2021.

[10] Carl Hvarfner, Erik Hellsten, Frank Hutter, and Luigi Nardi. Self-correcting bayesian optimization through bayesian active learning. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Proceedings of the 37th International Conference on Advances in Neural Information Processing Systems (NeurIPS'23)*. Curran Associates, 2023.

[11] Vidhi Lalchand and Carl Edward Rasmussen. Approximate inference for fully bayesian gaussian process regression. In *Symposium on Advances in Approximate Bayesian Inference, AABI 2019, Vancouver, BC, Canada, December 8, 2019*, Proceedings of Machine Learning Research, pages 1–12. PMLR, 2019. URL `http://proceedings.mlr.press/v118/lalchand20a.html`.

[12] Michael Alan Osborne. *Bayesian Gaussian Processes for Sequential Prediction, Optimisation and Quadrature*. Phd thesis, University of Oxford, Oxford, UK, 2010.

[13] Joel A. Paulson and Calvin Tsay. Bayesian optimization as a flexible and efficient design framework for sustainable process systems. *Current Opinion in Green and Sustainable Chemistry*, 51:100983, 2025.

[14] James Willard, Shirin Golchi, Erica E M Moodie, Bruno Boulanger, and Bradley P Carlin. Bayesian optimization for personalized dose-finding trials with combination therapies. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 74(2):373–390, 11 2024.

[15] Matthew D. Hoffman and Andrew Gelman. The no-u-turn sampler: adaptively setting path lengths in hamiltonian monte carlo. *J. Mach. Learn. Res.*, 2014.

[16] Felix Berkenkamp, Angela Schoellig, and Andreas Krause. No-regret Bayesian optimization with unknown hyperparameters. *Journal of Machine Learning Research*, 20:50:1–50:24, 2019.

[17] I. Bogunovic and A. Krause. Misspecified Gaussian process bandit optimization. In Ranzato et al. [35].

[18] George Wynne, Francois-Xavier Briol, and Mark Girolami. Convergence guarantees for gaussian process means with misspecified likelihoods and smoothness. *Journal of Machine Learning Research*, 22(123):1–40, 2021.

[19] Julian Rodemann and Thomas Augustin. Imprecise bayesian optimization. *Knowledge-Based Systems*, pages 112–186, 2024.

[20] Rémi Pautrat, Konstantinos I. Chatzilygeroudis, and Jean-Baptiste Mouret. Bayesian optimization with automatic prior selection for data-efficient direct policy search. In *2018 IEEE International Conference on Robotics and Automation, ICRA 2018, Brisbane, Australia, May 21-25, 2018*, pages 7571–7578. IEEE, 2018.

[21] Konstantinos D. Polyzos, Qin Lu, and Georgios B. Giannakis. Bayesian optimization with ensemble learning models and adaptive expected improvement. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2023.

[22] Juliusz Ziomek, Masaki Adachi, and Michael A. Osborne. Time-varying gaussian process bandits with unknown prior. In *International Conference on Artificial Intelligence and Statistics, AISTATS 2025, Mai Khao, Thailand, 3-5 May 2025*, 2025.

[23] Christoffer Riis, Francisco Antunes, Frederik Boe Hüttel, Carlos Lima Azevedo, and Francisco Pereira. Bayesian active learning with fully bayesian gaussian processes. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Proceedings of the 36th International Conference on Advances in Neural Information Processing Systems (NeurIPS'22)*. Curran Associates, 2022.

[24] M. Wistuba, N. Schilling, and L. Schmidt-Thieme. Scalable Gaussian process-based transfer surrogates for Hyperparameter Optimization. *Machine Learning*, 107(1):43–78, 2018.

[25] M. Feurer, B. Letham, F. Hutter, and E. Bakshy. Practical transfer learning for bayesian optimization. *arXiv:1802.02219v4 [stat.ML]*, 2022.

[26] D. Eriksson and M. Jankowiak. High-dimensional bayesian optimization with sparse axis-aligned subspaces. In J. Peters and D. Sontag, editors, *Proceedings of The 37th Uncertainty in Artificial Intelligence Conference (UAI'21)*, pages 493–503. PMLR, 2021.

[27] N. Srinivas, A. Krause, S. Kakade, and M. Seeger. Gaussian process optimization in the bandit setting: No regret and experimental design. In J. Fürnkranz and T. Joachims, editors, *Proceedings of the 27th International Conference on Machine Learning (ICML'10)*, pages 1015–1022. Omnipress, 2010.

[28] N. de Freitas, A. Smola, and M. Zoghi. Exponential regret bounds for Gaussian process bandits with deterministic observations. In J. Langford and J. Pineau, editors, *Proceedings of the 29th International Conference on Machine Learning (ICML'12)*. Omnipress, 2012.

[29] I. Bogunovic, J. Scarlett, S. Jegelka, and V. Cevher. Adversarially robust optimization with gaussian processes. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Proceedings of the 32nd International Conference on Advances in Neural Information Processing Systems (NeurIPS'18)*. Curran Associates, 2018.

[30] Peter Walley. *Statistical Reasoning with Imprecise Probabilities*. Chapman and Hall, 1991.

[31] T. Seidenfeld. A contrast between two decision rules for use with (convex) sets of probabilities: Γ-maximin versus E-admissibility. *Synthese*, 140, 2004.

[32] M. Balandat, B. Karrer, D. Jiang, S. Daulton, B. Letham, A. Wilson, and E. Bakshy. Botorch: A framework for efficient monte-carlo Bayesian optimization. In H. Larochelle, M. Ranzato, R. Hadsell, M.-F. Balcan, and H. Lin, editors, *Proceedings of the 34th International Conference on Advances in Neural Information Processing Systems (NeurIPS'20)*. Curran Associates, 2020.

[33] K. Sehic, A. Gramfort, J. Salmon, and L. Nardi. Lassobench: A high-dimensional hyperparameter optimization benchmark suite for lasso. In I. Guyon, M. Lindauer, M. van der Schaar, F. Hutter, and R. Garnett, editors, *Proceedings of the First International Conference on Automated Machine Learning*. Proceedings of Machine Learning Research, 2022.

[34] A. Makarova, I. Usmanova, I. Bogunovic, and A. Krause. Risk-averse heteroscedastic bayesian optimization. In Ranzato et al. [35], pages 17235–17245.

[35] M. Ranzato, A. Beygelzimer, K. Nguyen, P. Liang, J. Vaughan, and Y. Dauphin, editors. *Proceedings of the 35th International Conference on Advances in Neural Information Processing Systems (NeurIPS'21)*, 2021. Curran Associates.

[36] Sayak Ray Chowdhury and Aditya Gopalan. On kernelized multi-armed bandits. In D. Precup and Y. Teh, editors, *Proceedings of the 34th International Conference on Machine Learning (ICML'17)*, volume 70, pages 844–853. Proceedings of Machine Learning Research, 2017.

[37] Jonathan Scarlett. Tight regret bounds for bayesian optimization in one dimension. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, Proceedings of Machine Learning Research, pages 4507–4515. PMLR, 2018.

A Proofs

In this section we first derive a regret bound for the misspecified hierarchical GP-UCB and then for the medoid $\mathcal{X}$ rule.

A.1 Misspecified Hierarchical Gaussian Process Optimization

A key quantity in establishing the regret bound (and also in general theoretical results in Bayesian optimization) is the maximum information gain. It measures the largest amount of information that a set of noisy observations can reveal about an unknown function drawn from a zero-mean GP with kernel κ_θ. Formally, the maximum information gain after t observations is defined as

$$\gamma_t(\kappa_\theta, \mathcal{X}) := \max_{\mathcal{D}_t \subset \mathcal{X}} I(\mathbf{f}_t; \mathbf{y}_t) = \max_{\mathcal{D}_t \subset \mathcal{X}} \frac{1}{2} \log\det(I_t + \lambda^{-1} K_t) \tag{12}$$

where λ is the observation noise variance (cf. Equation 1).

We now recall the misspecified hierarchical BO setting, that we introduced in our work. Assume a hyperprior $P(\theta)$ from which we sample hyperparameters $\{\theta^{(j)}\}_{j=1}^{M}$. At time $t-1$, each sampled hyperparameter $\theta^{(j)}$ induces a GP posterior with mean

$$\mu_{t-1}^{\theta^{(j)}}(x) := \mu_{t-1}^{j}(x)$$

and variance

$$\sigma_{t-1}^{\theta^{(j)}}(x) := \sigma_{t-1}(x).$$

We assume that each GP model is at most ε_j-misspecified. That is, for each $j \in \{1, \dots, M\}$, there exists a function $g_j \in \mathcal{F}_k(\mathcal{X}, B)$ and a bounded model error $m_j : \mathcal{X} \to [-\varepsilon_j, \varepsilon_j]$ such that the observations satisfy

$$y_i = \underbrace{g_j(x_i) + m_j(x_i)}_{f(x_i)} + \xi_i. \tag{13}$$

We emphasize that, inspired by [17], we analyze this misspecified setting by assuming that the learner uses the correct kernel (i.e., the target function belongs to $\mathcal{F}_k(\mathcal{X}, B)$), but observes function evaluations corrupted by a uniformly bounded error. Consequently, while the posterior variances remain identical across GPs, the posterior means are biased due to the observation errors. Let $\mu^*(x)$ denote the posterior mean of the GP receiving unbiased observations $f(x_i) + \xi_i$. We further need the following well-known theoretical result in our analysis.

Lemma 1 (GP confidence bound per model ([27, 36])**.** *Let $g_{\theta^{(j)}} \in \mathcal{F}_k(\mathcal{X}, B)$ for kernel k, and assume the standard σ-sub-Gaussian noise model for observations. Let $\mathbf{y}_{t-1}^{(j)} := [y_1^{(j)}, \dots, y_{t-1}^{(j)}]$ denote the vector of previous noisy observations corresponding to the queried points $(x_1, \dots, x_{t-1})$, where $y_i^{(j)} = g_j(x_i) + \xi_i$ with ξ_i being σ-sub-Gaussian noise. Let $\mu_{t-1}^{j}(x)$ and $\sigma_{t-1}(x)$ denote the posterior mean and standard deviation of a Gaussian Process with kernel $k_{\theta^{(j)}}$ based on the previous $t-1$ observations. Then the following holds with probability at least $1-\delta$, simultaneously for all $t \geq 1$ and $x \in \mathcal{X}$:*

$$\left|\mu_{t-1}^{j}(x) - g_j(x)\right| \leq \beta_t \, \sigma_{t-1}(x), \tag{14}$$

where

$$\beta_t = B + \frac{\sigma}{\sqrt{\lambda}} \sqrt{2 \ln \frac{1}{\delta} + \sum_{s=1}^{t-1} \ln\left(1 + \frac{\sigma_{s-1}^2(x_s)}{\lambda}\right)}. \tag{15}$$

This high-probability bound is used to establish a regret bound in the realizable setting [27]. Note that each GP model, indexed by j, receives observations $g_{\theta^{(j)}}(x_i)$ from the function $g_{\theta^{(j)}} \in \mathcal{F}_k(\mathcal{X}, B)$; hence, each GP operates in the realizable setting with respect to the function $g_{\theta^{(j)}}$. However, our ultimate goal is to learn the target function f. To this end, we first recall a result from [17] that relates $\mu^*(x)$ to the posterior means of each of the misspecified GPs.

Lemma 2 (Difference between posterior means (Lemma 2 in [17]))**.** *For any $x \in \mathcal{X}$, $t \geq 1$, and $\varepsilon_i > 0$,*

$$|\mu_t^i(x) - \mu_t^*(x)| \leq \frac{\varepsilon_i \sqrt{t}}{\sqrt{\lambda}}\, \sigma_t(x), \tag{16}$$

where $\varepsilon_i \geq \|f - g_{\theta^{(i)}}\|_\infty$ is the misspecification error.

Using this result, we can now derive confidence bounds for the hierarchical GP posterior means under bounded model misspecification. Define the averaged posterior mean and misspecification error as

$$\mu_{t-1}^{\text{avg}}(x) = \frac{1}{M}\sum_{j=1}^{M} \mu_{t-1}^j(x), \qquad \varepsilon_{\text{avg}} = \frac{1}{M}\sum_{j=1}^{M} \varepsilon_j.$$

$$\begin{aligned}
|\mu_{t-1}^{\text{avg}}(x) - f(x)| &= \left| \frac{1}{M}\sum_{j=1}^{M} \mu_{t-1}^j(x) - f(x) \right| \\
&\leq \frac{1}{M}\sum_{j=1}^{M} \left|\mu_{t-1}^j(x) - \mu_{t-1}^*(x)\right| \;+\; \left|\mu_{t-1}^*(x) - f(x)\right| && (17) \\
&\leq \frac{\varepsilon_{\text{avg}}\sqrt{t}}{\sqrt{\lambda}}\, \sigma_{t-1}(x) + \beta_t\, \sigma_{t-1}(x) && (18) \\
&= \Big(\frac{\varepsilon_{\text{avg}}\sqrt{t}}{\sqrt{\lambda}} + \beta_t\Big)\sigma_{t-1}(x) && (19)
\end{aligned}$$

For the last inequality we used Lemma 2 to bound the first term and Lemma 1, which is the standard GP confidence bound from Srinivas et al. [27] to bound the second term. Note, that, the posterior means of the different sampled GPs and the GP that observes the unbiased observations differ, whereas their uncertainties are the same. Therefore also the β_t that is used in bounding the second term, is the same as the one used in Lemma 1.

Now we want to bound the regret. Let the instantaneous regret at timestep t be defined as $r_t = f(x_t) - f(x^*)$ with x^* being the global minimizer of the target function. Then, with high probability and using (19):

$$\begin{aligned}
r_t &= f(x_t) - f(x^*) \\
&\leq f(x_t) - \mu_{t-1}^{\text{avg}}(x^*) + \Big(\frac{\varepsilon_{\text{avg}}\sqrt{t}}{\sqrt{\lambda}} + \beta_t\Big)\sigma_{t-1}(x^*) && (20) \\
&\leq f(x_t) - \mu_{t-1}^{\text{avg}}(x_t) + \Big(\frac{\varepsilon_{\text{avg}}\sqrt{t}}{\sqrt{\lambda}} + \beta_t\Big)\sigma_{t-1}(x_t) && (21) \\
&\leq 2\Big(\frac{\varepsilon_{\text{avg}}\sqrt{t}}{\sqrt{\lambda}} + \beta_t\Big)\sigma_{t-1}(x_t) && (22)
\end{aligned}$$

Theorem 1 (recalled). Suppose M GP models with kernels k and corresponding hypothesis classes $\mathcal{F}_k(\mathcal{X}, B)$ for some fixed $B > 0$, with $\mathcal{X} \subset \mathbb{R}^d$. Let f be the true function s.t. $\|f - g_j\|_\infty \leq \varepsilon_j$, for all $j = 1, \ldots, M$. With $\varepsilon_{\text{avg}} := \frac{1}{M}\sum_{j=1}^{M} \varepsilon_j$ and γ_T the maximum information gain associated with kernel k, the hierarchical acquisition rule (6) guarantees, with probability at least $1 - \delta$, a cumulative regret bound of the order

$$O\Big(B\sqrt{\gamma_T T} + \sqrt{(\ln(1/\delta) + \gamma_T)\,\gamma_T T} + \varepsilon_{\text{avg}} T\sqrt{\gamma_T}\Big). \tag{23}$$

Proof. Summing over $t = 1, \dots, T$:

$$R_T = \sum_{t=1}^{T} (f(x_t) - f(x^\star)) \tag{24}$$

$$\leq \sum_{t=1}^{T} 2\Big(\frac{\varepsilon_{\text{avg}}\sqrt{t}}{\sqrt{\lambda}} + \beta_t\Big)\sigma_{t-1}(x) \tag{25}$$

$$\leq \sum_{t=1}^{T} \frac{2\varepsilon_{\text{avg}}\sqrt{t}}{\sqrt{\lambda}}\sigma_{t-1}(x_t) + \sum_{t=1}^{T} 2\beta_t\sigma_{t-1}(x_t) \tag{26}$$

$$\leq \frac{2\varepsilon_{\text{avg}}\sqrt{T}}{\sqrt{\lambda}} \sum_{t=1}^{T} \sigma_{t-1}(x_t) + 2\beta_T \sum_{t=1}^{T} \sigma_{t-1}(x_t) \tag{27}$$

$$\leq \frac{2\varepsilon_{\text{avg}}\sqrt{T}}{\sqrt{\lambda}} \sqrt{(2\lambda+1)\gamma_T T} + 2\beta_T\sqrt{(2\lambda+1)\gamma_T T} \tag{28}$$

$$= \frac{2\varepsilon_{\text{avg}}T}{\sqrt{\lambda}} \sqrt{(2\lambda+1)\gamma_T} + 2\Big(\frac{\sigma}{\sqrt{\lambda}}\sqrt{2\ln(1/\delta) + 2\gamma_T} + B\Big)\sqrt{(2\lambda+1)\gamma_T T} \tag{29}$$

$$= O\Big(\varepsilon_{\text{avg}} T\sqrt{\gamma_T} + B\sqrt{\gamma_T T} \;+\; \sqrt{(\ln(1/\delta) + \gamma_T)\,\gamma_T T}\Big) \tag{30}$$

$$= O\Big(B\sqrt{\gamma_T T} + \sqrt{(\ln(1/\delta) + \gamma_T)\,\gamma_T T} + \varepsilon_{\text{avg}} T\sqrt{\gamma_T}\Big).$$

□

A.2 Decision-Level Aggregation

In this section we prove Proposition 1 and Theorem 2, both stated in Section 5.2. We first restate the setting and Assumption 1, discuss the assumption, and then give the proofs.

We assume a strict majority of low-misspecification models with $0 \leq \varepsilon_i \leq \rho$ for $i \in G$, $|G| > M/2$, and outlier models with $\varepsilon_l \gg \rho$ for $l \in G^c$. All GPs share a common kernel, so all posterior variances coincide; we write $\sigma_{t-1}(x)$ for the common standard deviation and $\bar{\sigma}_{t-1} := \max_{x\in\mathcal{X}} \sigma_{t-1}(x)$. Each model j uses the enlarged-confidence AF

$$\alpha_j(x) = -\mu^j_{t-1}(x) + \tilde{\beta}^j_t\, \sigma_{t-1}(x), \qquad \tilde{\beta}^j_t := \beta_t + \frac{\varepsilon_j\sqrt{t}}{\sqrt{\lambda}},$$

with maximizer x^*_j, and we write $\tilde{\beta}^\rho_t := \beta_t + \rho\sqrt{t}/\sqrt{\lambda}$. The oracle GP observes the unbiased $f(x_i)+\xi_i$ without model-specific bias; we denote its posterior mean by μ^*_{t-1}, its AF by $\alpha^* = -\mu^*_{t-1} + \beta_t\sigma_{t-1}$, and its maximizer by $x^\dagger \in \arg\max_x \alpha^*(x)$. The medoid is $x_t = x^*_{j(t)}$ for some data-dependent $j(t) \in [M]$, and $x^* \in \arg\min_x f(x)$ is the global minimizer of f.

Assumption 1 (recalled). There exist $\kappa > 0$ and a neighborhood $\mathcal{N}$ of $x^\dagger$ such that

$$\alpha^*(x) \leq \alpha^*(x^\dagger) - \tfrac{\kappa}{2}\|x - x^\dagger\|_2^2 \qquad \text{for all } x \in \mathcal{N},$$

$x^*_i \in \mathcal{N}$ for all $i \in G$, $x^*_l \notin \mathcal{N}$ for all $l \in G^c$, and ρ is small enough that the localization ball from Proposition 1 is contained in $\mathcal{N}$.

Discussion of Assumption 1 *(i) Local curvature.* The first part requires α^* to decrease at least quadratically away from $x^\dagger$ in a neighborhood $\mathcal{N}$. Such a condition is analogous to local quadratic regularity assumptions used in Bayesian optimization analyses for the objective function itself [37, 28]. In particular, if $x^\dagger$ is an isolated maximizer and α^* is twice continuously differentiable with negative-definite Hessian at $x^\dagger$, then a bound of the stated form holds in a sufficiently small neighborhood for some $\kappa > 0$. The smallness requirement on ρ ensures that the localization ball of Proposition 1 remains inside this neighborhood.

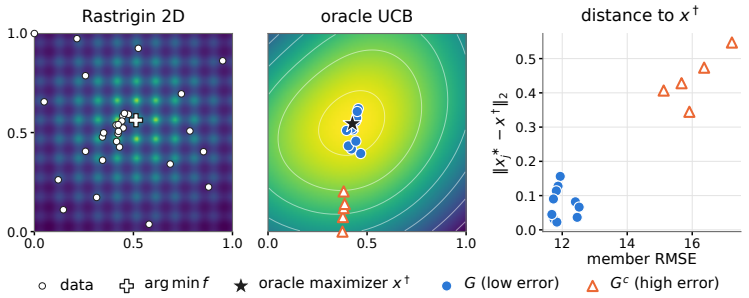

Figure 9: Separation of acquisition maximizers on a 2-D Rastrigin run (medoid-UCB, $M = 16$, after $t = 15$ BO iterations). Left: objective f with the observed data. Middle: oracle acquisition function α^* (UCB) with its maximizer $x^\dagger$ and the UCB maximizers x_j^* of the 16 members, split into low-error G and high-error G^c at the largest gap in the members' true RMSE. Right: member RMSE against the distance of its maximizer to $x^\dagger$. The maximizers of G cluster around $x^\dagger$, while those of G^c lie in a separate, distant group.

(ii) Separation. The second part is an identifiability assumption at the level of the decisions induced by the candidate models: models with small misspecification propose candidates inside $\mathcal{N}$, outlier models propose candidates outside it. This is precisely the structure the geometric medoid argument needs: if a strict majority of candidate decisions lies in a localized region around $x^\dagger$ and the outliers lie outside it, the medoid itself can be localized. The form of the enlarged-confidence AFs gives some intuition for why such separation may arise. For an outlier model l,

$$\alpha_l(x) = \alpha^*(x) + \left(\mu^*_{t-1}(x) - \mu^l_{t-1}(x)\right) + \frac{\varepsilon_l\sqrt{t}}{\sqrt{\lambda}}\sigma_{t-1}(x),$$

so misspecification shifts both the posterior-mean term and, through the enlarged confidence parameter, the preference for uncertain regions. Large misspecification can therefore move the maximizer far from $x^\dagger$. Separation is, however, not implied by a large ε_l alone, since the two effects can cancel. We therefore validate it empirically in the following.

A.3 Empirical Validation of Assumption 1

Local curvature. For part (i), we evaluate α^* on shells of radius up to $0.1\sqrt{d}$ around $x^\dagger$ (128 random directions each) and at the maximizers of all models in G. In all runs, $x^\dagger$ is the global maximizer of α^*, and the bound $\alpha^*(x^\dagger) - \alpha^*(x) \geq \frac{\kappa}{2}\|x - x^\dagger\|_2^2$ holds with $\kappa > 0$ at every evaluated point. Since the localization radius $C\sqrt{\rho\sqrt{t}}$ of Proposition 1 scales with $C \propto 1/\sqrt{\kappa}$, the size of κ determines how tight the guarantee is. On Griewank and Rastrigin, κ is bounded away from zero in all runs (median between 1.8 and 3.3). On Rosenbrock, whose narrow curved valley makes α^* nearly flat, κ is small (median 0.18) and below 0.1 in 38% of runs, so the guarantee is loose.

Separation of maximizers. To verify Part (ii), we run decision-level BO with the medoid and UCB on Rastrigin, Rosenbrock and Griewank in $d \in \{4, 8, 16, 32\}$, with 10 seeds per pair. Since several of these functions have their global optimum at the center of the domain, which may favor BO methods with a bias toward the center, we shift the search space such that the global optimum lies at a random position within $[0, 1]^d$ (and not necessarily the center anymore). The $M = 16$ models are hyperparameter samples from the SAASBO hyperposterior.

Table 2: Separation AUC per function and dimension on runs with a reliable oracle proxy (test $R^2 \geq 0.75$), mean over seeds; 0.5 = no separation, 1 = exact separation. The last row pools all runs.

	$d = 4$	$d = 8$	$d = 16$	$d = 32$
Griewank	0.95	0.86	0.81	0.74
Rastrigin	0.87	0.88	0.88	0.64
Rosenbrock	0.66	0.53	0.54	0.65
Pooled	0.82	0.74	0.73	0.68

Proxies. Several quantities cannot be measured directly and thus have to be approximated.

- *Misspecification level ε_j.* The RMSE of model j's posterior mean against the true f on 512 Sobol test points.
- *Groups G and G^c.* We sort the models by RMSE and split them at the largest gap between consecutive values; this avoids choosing a threshold ρ.
- *Oracle AF α^*.* The oracle of the analysis is the GP that observes f without misspecification. We approximate it by an exact GP (Matérn-5/2, ARD) fitted on 2000 noiseless Sobol evaluations of f, and which is then conditioned on the same BO data as the ensemble. The oracle AF α^* is then given by this GPs UCB.

To verify membership in a particular neighborhood we test whether the maximizers of G lie closer to $x^\dagger$ than those of G^c. Per run, we compute

$$\mathrm{AUC} = P\big(\|x_l^* - x^\dagger\|_2 > \|x_i^* - x^\dagger\|_2\big), \qquad l \in G^c,\ i \in G,$$

over all pairs. $\mathrm{AUC} = 1$ holds exactly when some ball around $x^\dagger$ contains all maximizers of G and none of G^c, i.e., when (ii) holds with $\mathcal{N}$ a ball; $\mathrm{AUC} = 0.5$ means distance to $x^\dagger$ does not depend on the group. We test the per-seed AUCs against 0.5 with a Wilcoxon signed-rank test.

In Figure 9 we illustrate the eparation for a single run on the 2D Rastrigin function at $t = 15$: every maximizer of G lies closer to $x^\dagger$ than every maximizer of G^c. Table 2 summarizes the grid. Pooled over functions, the AUC is 0.82, 0.74, 0.73 and 0.68 for $d = 4, 8, 16, 32$, significantly above chance in every dimension ($p_{\mathrm{Holm}} \leq 0.003$). The separation is strongest on Griewank (significant at every d) and Rastrigin (AUC ≈ 0.87 for $d \leq 16$), and weak on Rosenbrock. Including all 200 runs, the pooled AUC is 0.72, 0.69, 0.68 and 0.58. Separation is thus not guaranteed in every run, but it is the typical case: maximizers of low-error models usually lie closer to $x^\dagger$ than those of outlier models.

Proposition 1 (recalled). Under Assumption 1 and $|G| > M/2$, the medoid x_t selected by the medoid rule satisfies

$$\|x_t - x^\dagger\|_2 \ \leq\ C\,\sqrt{\rho\sqrt{t}},$$

where $C := \frac{2}{\sqrt{\kappa}\,\lambda^{1/4}} \cdot \frac{3|G|+|G^c|}{|G|-|G^c|}$. In particular, $x_t = x^*_{j(t)}$ for some $j(t) \in G$.

Proof. Define

$$\bar{r}_\rho^2 \ :=\ \frac{4}{\kappa} \cdot \frac{\rho\sqrt{t}}{\sqrt{\lambda}}\,\bar{\sigma}_{t-1}, \qquad C_G \ :=\ \frac{3|G| + |G^c|}{|G| - |G^c|}.$$

The proof has two steps.

Step 1: Maximizers of low-misspecification models cluster around $x^\dagger$.

By Lemma 2 and the definitions of $\alpha_j, \alpha^*, \tilde{\beta}_t^j$,

$$\alpha_j(x) - \alpha^*(x) = \big(\mu_{t-1}^*(x) - \mu_{t-1}^j(x)\big) + \frac{\varepsilon_j\sqrt{t}}{\sqrt{\lambda}}\sigma_{t-1}(x),$$

and the parenthesis lies in $\big[-\varepsilon_j\sqrt{t}/\sqrt{\lambda}\,\sigma_{t-1},\ +\varepsilon_j\sqrt{t}/\sqrt{\lambda}\,\sigma_{t-1}\big]$. Hence

$$0 \ \le\ \alpha_j(x) - \alpha^*(x) \ \le\ \frac{2\varepsilon_j\sqrt{t}}{\sqrt{\lambda}}\,\sigma_{t-1}(x). \tag{31}$$

Fix $i \in G$. Since x_i^* maximizes α_i, $\alpha_i(x_i^*) \ge \alpha_i(x^\dagger)$, so by (31),

$$\begin{aligned}\alpha^*(x^\dagger) - \alpha^*(x_i^*) &= \underbrace{\big[\alpha^*(x^\dagger) - \alpha_i(x^\dagger)\big]}_{\le 0} + \underbrace{\big[\alpha_i(x^\dagger) - \alpha_i(x_i^*)\big]}_{\le 0} + \underbrace{\big[\alpha_i(x_i^*) - \alpha^*(x_i^*)\big]}_{\le 2\varepsilon_i\sqrt{t}\bar{\sigma}_{t-1}/\sqrt{\lambda}} \\ &\le \frac{2\varepsilon_i\sqrt{t}}{\sqrt{\lambda}}\,\bar{\sigma}_{t-1}.\end{aligned}$$

By Assumption 1, $x_i^* \in \mathcal{N}$, so the concavity clause gives

$$\frac{\kappa}{2}\|x_i^* - x^\dagger\|_2^2 \ \le\ \alpha^*(x^\dagger) - \alpha^*(x_i^*) \ \le\ \frac{2\varepsilon_i\sqrt{t}}{\sqrt{\lambda}}\,\bar{\sigma}_{t-1}.$$

Using $\varepsilon_i \le \rho$, $\|x_i^* - x^\dagger\|_2^2 \le \frac{4\rho\sqrt{t}}{\kappa\sqrt{\lambda}}\,\bar{\sigma}_{t-1} = \bar{r}_\rho^2$, so every good maximizer lies in $B(x^\dagger, \bar{r}_\rho)$.

Step 2: The medoid is the maximizer of a low-misspecification model.

We first show $\|x_t - x^\dagger\|_2 \le C_G\bar{r}_\rho$ by a majority argument. Suppose for contradiction $\|x_t - x^\dagger\|_2 > C_G\bar{r}_\rho$. Pick any $k \in G$ and consider the difference of total distances

$$\Delta \ :=\ \sum_{i=1}^M \|x_t - x_i^*\|_2 - \sum_{i=1}^M \|x_k^* - x_i^*\|_2.$$

Good indices ($i \in G$). By Step 1 both x_k^* and x_i^* lie in $B(x^\dagger, \bar{r}_\rho)$, so the triangle inequality gives $\|x_k^* - x_i^*\|_2 \le 2\bar{r}_\rho$ and $\|x_t - x_i^*\|_2 \ge \|x_t - x^\dagger\|_2 - \bar{r}_\rho$. Hence

$$\sum_{i\in G}\big(\|x_t - x_i^*\|_2 - \|x_k^* - x_i^*\|_2\big) \ \ge\ |G|\big(\|x_t - x^\dagger\|_2 - 3\bar{r}_\rho\big).$$

Bad indices ($i \in G^c$). The reverse triangle inequality and $\|x_t - x_k^*\|_2 \le \|x_t - x^\dagger\|_2 + \bar{r}_\rho$ (since $k \in G$) give

$$\sum_{i\in G^c}\big(\|x_t - x_i^*\|_2 - \|x_k^* - x_i^*\|_2\big) \ \ge\ -|G^c|\big(\|x_t - x^\dagger\|_2 + \bar{r}_\rho\big).$$

Combining,

$$\Delta \ \ge\ \big(|G| - |G^c|\big)\,\|x_t - x^\dagger\|_2 - \big(3|G| + |G^c|\big)\,\bar{r}_\rho,$$

and plugging $\|x_t - x^\dagger\|_2 > C_G\bar{r}_\rho$ yields $\Delta > 0$, contradicting medoid optimality. Hence $\|x_t - x^\dagger\|_2 \le C_G\bar{r}_\rho$.

By the smallness clause of Assumption 1, $B(x^\dagger, C_G\bar{r}_\rho) \subseteq \mathcal{N}$, and by the separation clause $x_l^* \notin \mathcal{N}$ for every $l \in G^c$. Hence the medoid $x_t \in B(x^\dagger, C_G\bar{r}_\rho) \subseteq \mathcal{N}$ cannot be any x_l^* with $l \in G^c$, so $x_t = x_{j(t)}^*$ for some $j(t) \in G$.

Finally, under the standard normalization $k(x,x) \le 1$ on $\mathcal{X}$, $\bar{\sigma}_{t-1} \le 1$, so $\bar{r}_\rho \le \frac{2}{\sqrt{\kappa}\,\lambda^{1/4}}\sqrt{\rho\sqrt{t}}$, giving the displayed bound $\|x_t - x^\dagger\|_2 \le C\sqrt{\rho\sqrt{t}}$ with C as in the proposition. □

Remark 4 (Weighted medoid)**.** *The localization argument extends directly to a weighted medoid. Let $w_j \geq 0$ denote model weights and define*

$$x_t \in \arg\min_{x \in \mathcal{X}^*} \sum_{j=1}^{M} w_j \|x - x_j^*\|_2.$$

The proof of Proposition 1 carries over after replacing cardinalities by total weights. In particular, the strict-majority condition becomes $\sum_{i \in G} w_i > \sum_{l \in G^c} w_l$. Under this weighted-majority condition, the same geometric argument yields a corresponding localization bound with the counting-dependent constant replaced by its weighted analogue. This extension also illustrates why uniform weighting is a natural robust choice when no reliable prior information about model quality is available.

Per-iteration regret. By Proposition 1, $x_t = x^*_{j(t)}$ with $j(t) \in G$, so $\varepsilon_{j(t)} \leq \rho$. Set $j := j(t)$. Combining Lemmas 1 and 2 via the triangle inequality, the per-model confidence bound

$$|\mu^j_{t-1}(x) - f(x)| \;\leq\; \tilde{\beta}^j_t \, \sigma_{t-1}(x)$$

holds simultaneously for all j, t, x with probability at least $1 - \delta$. With high probability, the per-iteration regret $r_t = f(x_t) - f(x^*)$ satisfies:

$$\begin{aligned}
r_t &= f(x_t) - f(x^*) \\
&\leq f(x_t) - \mu^j_{t-1}(x^*) + \tilde{\beta}^j_t \, \sigma_{t-1}(x^*) && (32)\\
&\leq f(x_t) - \mu^j_{t-1}(x_t) + \tilde{\beta}^j_t \, \sigma_{t-1}(x_t) && (33)\\
&\leq 2\, \tilde{\beta}^j_t \, \sigma_{t-1}(x_t) && (34)\\
&\leq 2\, \tilde{\beta}^\rho_t \, \sigma_{t-1}(x_t), && (35)
\end{aligned}$$

where the second line uses the per-model confidence bound at x^*, the third uses AF-optimality of $x_t = x_j^*$ for α_j, the fourth uses the per-model confidence bound at x_t, and the last uses $\varepsilon_j \leq \rho$.

Theorem 2 (recalled). Suppose M GP models share kernel k and hypothesis class $\mathcal{F}_k(\mathcal{X}, B)$, and Assumption 1 holds. Assume further that a strict majority of the models are low-misspecification with $\varepsilon_i \leq \rho$. Then the medoid rule guarantees, with probability at least $1 - \delta$, a cumulative regret bound of the order

$$O\Big(B\sqrt{\gamma_T T} + \sqrt{(\ln(1/\delta) + \gamma_T)\,\gamma_T T} + \rho\, T\sqrt{\gamma_T}\Big).$$

Proof. Summing (35) over $t = 1, \ldots, T$ and splitting $\tilde{\beta}^\rho_t = \beta_t + \rho\sqrt{t}/\sqrt{\lambda}$:

$$\begin{aligned}
R_T &= \sum_{t=1}^{T} r_t \\
&\leq \sum_{t=1}^{T} 2\, \tilde{\beta}^\rho_t \, \sigma_{t-1}(x_t) && (36)\\
&= \sum_{t=1}^{T} 2\, \beta_t \, \sigma_{t-1}(x_t) + \sum_{t=1}^{T} \frac{2\rho\sqrt{t}}{\sqrt{\lambda}} \, \sigma_{t-1}(x_t) && (37)\\
&\leq 2\beta_T \sqrt{T(2\lambda+1)\gamma_T} + \frac{2\rho}{\sqrt{\lambda}} \sqrt{\frac{T^2}{2}} \sqrt{(2\lambda+1)\gamma_T} && (38)\\
&= O\Big(\beta_T \sqrt{\gamma_T T} + \rho\, T\sqrt{\gamma_T}\Big) && (39)\\
&= O\Big(B\sqrt{\gamma_T T} + \sqrt{(\ln(1/\delta) + \gamma_T)\,\gamma_T T} + \rho\, T\sqrt{\gamma_T}\Big). && (40)
\end{aligned}$$

□

A.4 Extension to Distinct Kernels

Theorem 2 relies on the shared-kernel abstraction of Section 4, under which all ensemble members have the same posterior variance σ_{t-1}. In this section we remove this abstraction: each member j uses its own kernel $k_j := k_{\theta^{(j)}}$ and hence has its own posterior mean *and* posterior variance. We show that Proposition 1 and Theorem 2 carry over under an analogous assumption, at the price of a factor $\sqrt{|G|}$ and of the maximum information gain over the well-specified members. Throughout, the kernels $k_1, \dots, k_M$ are fixed across rounds, and we make the round index t explicit.

Setting. Observations are $y_t = f(x_t) + \xi_t$ with noise as in Lemma 1, and all kernels satisfy $k(x, x) \le 1$. Let k^* be a correctly specified kernel, i.e., $f \in \mathcal{F}_{k^*}(\mathcal{X}, B)$, and let μ^*_{t-1} and σ^*_{t-1} denote the posterior mean and standard deviation of the GP with kernel k^* conditioned on $\mathcal{D}_{t-1}$. The oracle AF and its maximizer are

$$\alpha^*_t(x) := -\mu^*_{t-1}(x) + \beta^*_t \, \sigma^*_{t-1}(x), \qquad x^\dagger_t \in \arg\max_{x \in \mathcal{X}} \alpha^*_t(x).$$

Each member $j \in \{1, \dots, M\}$ is ε_j-misspecified with respect to its own kernel, i.e., there exists $g_j \in \mathcal{F}_{k_j}(\mathcal{X}, B)$ with $\|f - g_j\|_\infty \le \varepsilon_j$, and we write $m_j := f - g_j$. Conditioned on all observations $\mathcal{D}_{t-1}$, member j has posterior mean μ^j_{t-1} and standard deviation σ^j_{t-1}, and uses the enlarged-confidence AF

$$\alpha_{j,t}(x) := -\mu^j_{t-1}(x) + \tilde{\beta}^j_t \, \sigma^j_{t-1}(x), \qquad \tilde{\beta}^j_t := \beta^j_t + \frac{\varepsilon_j \sqrt{t}}{\sqrt{\lambda}},$$

where β^j_t is the exploration parameter of Lemma 1 instantiated for kernel k_j at confidence level δ/M. We denote the maximizer of $\alpha_{j,t}$ by $x^*_{j,t}$ and its deviation from the oracle AF by

$$\Delta_{j,t} := \sup_{x \in \mathcal{X}} \big| \alpha_{j,t}(x) - \alpha^*_t(x) \big|.$$

At round t, the medoid rule (11) is applied to $\mathcal{X}^*_t := \{x^*_{j,t}\}_{j=1}^M$, and $x^* \in \arg\min_x f(x)$ denotes the global minimizer of f. If all members share the kernel k^*, all members have the posterior mean and variance of the oracle, $\beta^j_t = \beta^*_t$, and the members differ only through ε_j in $\tilde{\beta}^j_t$; in this case Part 1 of Assumption 2 below follows from $\varepsilon_i \le \rho$, and the localization result of Section A.2 is recovered (see the discussion after Assumption 2).

Auxiliary lemmas. Two obstacles prevent a direct transfer of the proof of Theorem 2: the confidence parameters $\tilde{\beta}^j_t$ are kernel-dependent, and the sum $\sum_t \sigma^{j(t)}_{t-1}(x_t)$ mixes posterior variances of different GPs, so it cannot be bounded by the information gain of a single kernel. We address the first obstacle with a per-member confidence bound (Lemma 3) and the second by partitioning the rounds according to the selected member (Lemma 4).

Lemma 3 (Per-member confidence bound)**.** *With probability at least $1 - \delta$, simultaneously for all $j \in \{1, \dots, M\}$, $t \ge 1$ and $x \in \mathcal{X}$,*

$$\big| \mu^j_{t-1}(x) - f(x) \big| \;\le\; \tilde{\beta}^j_t \, \sigma^j_{t-1}(x) + \varepsilon_j.$$

Proof. Fix j. The observations decompose as $y_i = g_j(x_i) + m_j(x_i) + \xi_i$ with $\|m_j\|_\infty \le \varepsilon_j$. Let $\bar{\mu}^j_{t-1}$ denote the posterior mean of an auxiliary GP with the same kernel k_j, the same query points $x_1, \dots, x_{t-1}$ and the same noise ξ_i, but with unbiased observations $g_j(x_i) + \xi_i$; this GP is an analytical device only. Since the posterior variance depends only on the kernel and the query points, both GPs have standard deviation σ^j_{t-1}. By the triangle inequality,

$$\big| \mu^j_{t-1}(x) - f(x) \big| \le \big| \mu^j_{t-1}(x) - \bar{\mu}^j_{t-1}(x) \big| + \big| \bar{\mu}^j_{t-1}(x) - g_j(x) \big| + \big| g_j(x) - f(x) \big|.$$

The last term is at most ε_j. Since $g_j \in \mathcal{F}_{k_j}(\mathcal{X}, B)$, the auxiliary GP is realizable, and Lemma 1 with kernel k_j at level δ/M bounds the second term by $\beta^j_t \, \sigma^j_{t-1}(x)$, simultaneously for all t and x with probability at least $1 - \delta/M$. The two posterior means in the first term use the same kernel k_j, and their observation vectors differ entrywise by $m_j(x_i)$; hence Lemma 2 bounds the first term deterministically by $\varepsilon_j \sqrt{t} \, \sigma^j_{t-1}(x)/\sqrt{\lambda}$. Summing the three bounds and taking a union bound over the M members proves the claim. □

Lemma 4 (Variance sum along a subsequence)**.** *Fix j and let $\mathcal{T}_j = \{t_1 < \cdots < t_{N_j}\} \subseteq \{1, \dots, T\}$ be any, possibly data-dependent, set of rounds. Write $\tilde{x}_n := x_{t_n}$, and let $\tilde{\sigma}^j_{n-1}$ denote the posterior standard deviation of the GP with kernel k_j conditioned only on $\tilde{x}_1, \dots, \tilde{x}_{n-1}$. Then*

$$\sum_{t \in \mathcal{T}_j} \sigma^j_{t-1}(x_t) \;\le\; \sum_{n=1}^{N_j} \tilde{\sigma}^j_{n-1}(\tilde{x}_n) \;\le\; \sqrt{2(2\lambda+1)\,\gamma_{N_j}(k_j)\,N_j}.$$

Proof. Let $t = t_n \in \mathcal{T}_j$. The posterior variance of a GP depends only on the kernel and the conditioning locations, not on the observed values, and it is non-increasing when the conditioning set is enlarged. Since $\{\tilde{x}_1, \dots, \tilde{x}_{n-1}\} \subseteq \{x_1, \dots, x_{t-1}\}$, we obtain $\sigma^j_{t-1}(x_t) \le \tilde{\sigma}^j_{n-1}(\tilde{x}_n)$. This holds deterministically for every realization, so the data-dependence of $\mathcal{T}_j$ is immaterial. Summing over $t \in \mathcal{T}_j$ gives the first inequality.

For the second inequality, note that $(\tilde{\sigma}^j_{n-1})_{n \le N_j}$ is the standard-deviation sequence of a GP with kernel k_j that is updated after each point of $(\tilde{x}_n)_{n \le N_j}$. By the information-gain identity of [27, Lemma 5.3] and the definition of $\gamma_{N_j}(k_j)$ as a maximum over all sets of N_j points,

$$\frac{1}{2} \sum_{n=1}^{N_j} \ln\Bigl(1 + \lambda^{-1}\bigl(\tilde{\sigma}^j_{n-1}(\tilde{x}_n)\bigr)^2\Bigr) \;\le\; \gamma_{N_j}(k_j),$$

irrespective of how the points $\tilde{x}_n$ were generated. Since $k_j(x,x) \le 1$, we have $(\tilde{\sigma}^j_{n-1})^2 \le 1$, and $u \le (2\lambda+1)\ln(1+\lambda^{-1}u)$ for $u \in [0,1]$. Together with the Cauchy–Schwarz inequality,

$$\sum_{n=1}^{N_j} \tilde{\sigma}^j_{n-1}(\tilde{x}_n) \;\le\; \sqrt{N_j \sum_{n=1}^{N_j} \bigl(\tilde{\sigma}^j_{n-1}(\tilde{x}_n)\bigr)^2} \;\le\; \sqrt{2(2\lambda+1)\,\gamma_{N_j}(k_j)\,N_j}. \qquad \square$$

The following assumption is the analogue of Assumption 1 for distinct kernels.

Assumption 2. *There exist $\kappa > 0$, $\rho, \Delta_\rho \ge 0$ and $G \subseteq \{1, \dots, M\}$ with $|G| > M/2$ such that for all $t \le T$:*

1. (Well-specified majority.) *$\varepsilon_i \le \rho$ and $\Delta_{i,t} \le \Delta_\rho$ for all $i \in G$.*
2. (Local quadratic structure.) *There is a neighborhood $\mathcal{N}_t$ of $x^\dagger_t$ such that $\alpha^*_t(x) \le \alpha^*_t(x^\dagger_t) - \frac{\kappa}{2}\|x - x^\dagger_t\|_2^2$ for all $x \in \mathcal{N}_t$.*
3. (Separation and smallness.) *$x^*_{i,t} \in \mathcal{N}_t$ for all $i \in G$, $x^*_{l,t} \notin \mathcal{N}_t$ for all $l \in G^c$, and $\mathcal{B}(x^\dagger_t, C_G \bar{r}_\Delta) \subseteq \mathcal{N}_t$, where $\bar{r}_\Delta := 2\sqrt{\Delta_\rho/\kappa}$ and $C_G := \frac{3|G|+|G^c|}{|G|-|G^c|}$.*

Parts 2 and 3 coincide with Assumption 1, with the localization radius $\bar{r}_\Delta$ in place of $\bar{r}_\rho$. Part 1 adds a condition at the level of acquisition functions: the AFs of well-specified members must be uniformly close to the oracle AF. Like the separation condition, this is an identifiability requirement on the AFs; no deviation bound is required for the outliers $l \in G^c$ beyond separation. In the shared-kernel setting of Section A.2, (31) gives $\Delta_{i,t} \le 2\rho\sqrt{t}\,\bar{\sigma}_{t-1}/\sqrt{\lambda}$ for $i \in G$, so Part 1 follows from $\varepsilon_i \le \rho$ with $\Delta_\rho := 2\rho\sqrt{T}/\sqrt{\lambda}$, and Proposition 1 is recovered with t replaced by T, up to a factor $\sqrt{2}$ in the constant.

Proposition 2 (Localization of the medoid, distinct kernels)**.** *Under Assumption 2, at every round $t \le T$ the medoid x_t satisfies*

$$\|x_t - x^\dagger_t\|_2 \;\le\; C_G\,\bar{r}_\Delta = \frac{2(3|G| + |G^c|)}{|G| - |G^c|}\sqrt{\frac{\Delta_\rho}{\kappa}},$$

and in particular $x_t = x^*_{j(t),t}$ for some $j(t) \in G$.

Proof. Fix t and drop the round index.

Step 1: good maximizers cluster around $x^\dagger$. For $i \in G$,

$$\alpha^*(x^\dagger) - \alpha^*(x_i^*) = \underbrace{\big[\alpha^*(x^\dagger) - \alpha_i(x^\dagger)\big]}_{\le \Delta_i} + \underbrace{\big[\alpha_i(x^\dagger) - \alpha_i(x_i^*)\big]}_{\le 0} + \underbrace{\big[\alpha_i(x_i^*) - \alpha^*(x_i^*)\big]}_{\le \Delta_i} \;\le\; 2\Delta_\rho,$$

where the middle term is nonpositive because x_i^* maximizes α_i. By Part 3 of Assumption 2, $x_i^* \in \mathcal{N}$, so Part 2 yields $\frac{\kappa}{2}\|x_i^* - x^\dagger\|_2^2 \le 2\Delta_\rho$, i.e., $\|x_i^* - x^\dagger\|_2 \le \bar{r}_\Delta$.

Step 2: majority argument. The first part of Step 2 in the proof of Proposition A.3 only uses that all good maximizers lie in a ball of radius $\bar{r}_\rho$ around $x^\dagger$ and that $|G| > |G^c|$; it involves neither kernels nor posterior variances. Repeating it with $\bar{r}_\Delta$ in place of $\bar{r}_\rho$ gives $\|x_t - x^\dagger\|_2 \le C_G\, \bar{r}_\Delta$.

Step 3: no outlier is selected. By Part 3, $x_t \in \mathcal{B}(x^\dagger, C_G \bar{r}_\Delta) \subseteq \mathcal{N}$, while $x_l^* \notin \mathcal{N}$ for all $l \in G^c$. Since the medoid is an element of $\mathcal{X}^*$, it follows that $x_t = x_{j(t)}^*$ for some $j(t) \in G$. □

Theorem 3 (Regret of medoid $\mathcal{X}$ aggregation, distinct kernels)**.** *Under Assumption 2, the medoid rule guarantees, with probability at least $1-\delta$, a cumulative regret bound of the order*

$$O\Big(\sqrt{|G|}\,\Big(B\sqrt{\gamma T} + \sqrt{(\ln(M/\delta) + \gamma)\,\gamma T} + \rho\, T\sqrt{\gamma}\Big) + \rho\, T\Big), \qquad \gamma := \max_{j \in G} \gamma_T(k_j).$$

Proof. All statements hold on the event of Lemma 3, which has probability at least $1-\delta$; no further probabilistic arguments are needed.

Per-round regret. Fix t. By Proposition 2, $x_t = x_{j,t}^*$ with $j := j(t) \in G$. Using the lower bound of Lemma 3 at x^*, the maximality of x_t for $\alpha_{j,t}$, and the upper bound of Lemma 3 at x_t,

$$\begin{aligned} r_t = f(x_t) - f(x^*) &\le f(x_t) - \mu_{t-1}^j(x^*) + \tilde{\beta}_t^j\, \sigma_{t-1}^j(x^*) + \varepsilon_j \\ &\le f(x_t) - \mu_{t-1}^j(x_t) + \tilde{\beta}_t^j\, \sigma_{t-1}^j(x_t) + \varepsilon_j \\ &\le 2\tilde{\beta}_t^j\, \sigma_{t-1}^j(x_t) + 2\varepsilon_j. \end{aligned}$$

Since $j \in G$, Part 1 of Assumption 2 gives $\varepsilon_j \le \rho$, hence

$$r_t \;\le\; 2\Big(\beta_t^j + \frac{\rho\sqrt{t}}{\sqrt{\lambda}}\Big)\sigma_{t-1}^j(x_t) + 2\rho.$$

Summation. For $j \in G$, let $\mathcal{T}_j := \{t \le T : j(t) = j\}$ and $N_j := |\mathcal{T}_j|$. By Proposition 2, these sets partition $\{1, \dots, T\}$, so $\sum_{j \in G} N_j = T$. Summing the per-round bound, using $\beta_t^j \le \beta_T^j$ and $\sqrt{t} \le \sqrt{T}$, and applying Lemma 4 to each $\mathcal{T}_j$,

$$R_T \;\le\; 2\Big(\bar{\beta}_T + \frac{\rho\sqrt{T}}{\sqrt{\lambda}}\Big)\sum_{j \in G} \sqrt{2(2\lambda+1)\,\gamma_{N_j}(k_j)\, N_j} + 2\rho T, \qquad \bar{\beta}_T := \max_{j \in G} \beta_T^j.$$

Since $\gamma_{N_j}(k_j) \le \gamma_T(k_j) \le \gamma$, the Cauchy–Schwarz inequality gives $\sum_{j \in G} \sqrt{\gamma N_j} \le \sqrt{\gamma\,|G| \sum_{j \in G} N_j} = \sqrt{\gamma\,|G|\,T}$, and hence

$$R_T \;\le\; 2\Big(\bar{\beta}_T + \frac{\rho\sqrt{T}}{\sqrt{\lambda}}\Big)\sqrt{2(2\lambda+1)\,\gamma\,|G|\,T} + 2\rho T.$$

Finally, as in the proof of Theorem 1, $\beta_T^j = O\big(B + \sqrt{\ln(M/\delta) + \gamma_T(k_j)}\big)$, where the information-gain term in β_T^j is evaluated along the full query sequence and bounded via the information-gain identity used in the proof of Lemma 4. Hence $\bar{\beta}_T = O\big(B + \sqrt{\ln(M/\delta) + \gamma}\big)$. For constant λ, the product $\bar{\beta}_T\sqrt{\gamma|G|T}$ yields the first two terms of the claim, $\rho\sqrt{T} \cdot \sqrt{\gamma|G|T}$ yields $\sqrt{|G|}\,\rho T\sqrt{\gamma}$, and $2\rho T$ yields the additive ρT term. □

Remark 5 (Price of kernel heterogeneity)**.** *Compared with Theorem 2, the bound depends on the largest information gain among the well-specified members and carries an additional factor $\sqrt{|G|} \le \sqrt{M}$, which stems from bounding the variance sum separately for each member; the dependence on T is unchanged. The additive ρT term arises from converting the confidence bound of member j from g_j to f, analogous to the misspecified GP-UCB analysis of [17].*

Algorithm 1 Acquisition-level aggregation

Require: AF α, aggregated AF ρ
for $t = 0, \dots, T$ **do**
  for $j = 1, \dots, M$ **do**
    $\theta^{(j)} \sim P(\theta \mid \mathcal{D}_t)$
    Compute $a_{\theta^{(j)}}$ from $p(f^{(j)})$
  end for
  Compute $\mathcal{Q}(x)$ for all $x \in \mathcal{X}$
  $x_t \in \arg\max_{x \in \mathcal{X}} \rho(\mathcal{Q}(x))$
  Observe $y_t = f(x_t) + \xi_t$
  $\mathcal{D}_{t+1} = \mathcal{D}_t \cup \{(x_t, y_t)\}$
end forreturn $x^* \in \arg\min_{x \in \mathcal{D}_T} f(x)$

Remark 6 (Scope)**.** *The analysis treats the kernels $k_1, \dots, k_M$ as fixed across rounds. In our implementation, hyperparameters are resampled from $P(\theta \mid \mathcal{D}_t)$ at every iteration, so the kernels vary over time; extending Lemmas 3 and 4 to time-varying kernels is left for future work. Moreover, Part 1 of Assumption 2 is a condition on the ensemble; identifying conditions on $P(\theta)$ under which it holds is an open question.*

B Additional Experimental Results

In this section we present additional experimental results and implementation details. Recall, that we implemented our framework in BoTorch [32]. We consider two AFs: Upper confidence bound (UCB) and expected improvement (EI). All hierarchical methods share the same hyperpriors, using either (i) the default BoTorch RBF hyperprior or (ii) the sparse SAASBO hyperprior [26]. The default BoTorch hyperprior uses RBF kernel and SAASBO hyperprior uses Matérn-$5/2$ kernel, they are defined as shown in Table 3.

Inference is performed via NUTS sampling [15], using the following parameters: 16 GP samples, a thinning factor of 16, and 512 warm-up iterations. For our experiments, we set the number of iterations dependent on the dimension as $\lceil 20 + 40\sqrt{180} \rceil$ and the initial design size to 20% of the total number of iterations. Algorithms 1 and 2 provide pseudocode for aggregation at the acquisition function level and decision level, respectively. All experiments were conducted using 2 CPU cores and 8 GiB of RAM. The HPC nodes utilized for the computations are equipped with two AMD Milan 7763 processors and a total of 256 GiB of main memory.

B.1 Toy Example

In this experiment, we study the effect of increasing proportions of misspecified models on aggregation performance. We minimize a sample from a 1D GP with a true lengthscale of 0.02. For the aggregated models, an increasing fraction of them are severely misspecified, using large lengthscales of approximately 5, while the remaining models have lengthscales close to the true value of 0.02. Table 5 details the specific lengthscale configurations used in this experiment. Each row corresponds to one experimental setting with nine GP models. The proportion of misspecified (outlier) models increases from 0% to 40%. Well-specified models have lengthscales near the ground-truth value, while misspecified models have much larger lengthscales.

Table 3: Hyperprior definitions used in our experiments.

BoTorch RBF hyperprior	SAASBO hyperprior
$\ell \sim \text{LogNormal}(\sqrt{2} + \log\sqrt{D}, \sqrt{3})$	$\ell = 1/\rho^2, \quad \rho \sim \text{Half-Cauchy}(\tau), \quad \tau \sim \text{Half-Cauchy}(0.1)$
$\sigma_k^2 = 1$	$\sigma_k^2 \sim \text{LogNormal}(0, 100)$
$\sigma_\epsilon^2 \sim \text{LogNormal}(-4, 1)$	$\sigma_\epsilon^2 \sim \text{LogNormal}(0, 100)$

Algorithm 2 Decision-level aggregation

Require: AF α
for $t = 0, \dots, T$ **do**
  for $j = 1, \dots, M$ **do**
    $\theta^{(j)} \sim P(\theta \mid \mathcal{D}_t)$
    Compute $a_{\theta^{(j)}}$ from $p(f^{(j)})$
  end for
  Compute $\mathcal{X}^*$
  $x_t \in \arg\min_{x \in \mathcal{X}^*} \sum_{j=1}^{M} \|x - x_j^*\|_2$
  Observe $y_t = f(x_t) + \xi_t$
  $\mathcal{D}_{t+1} = \mathcal{D}_t \cup \{(x_t, y_t)\}$
end forreturn $x^* \in \arg\min_{\mathbf{x} \in \mathcal{D}_T} f(x)$

B.2 Benchmark Functions

Table 6 provides an overview of all benchmark functions used in our experiments. We organize them into three categories: low-dimensional synthetic functions, high-dimensional synthetic functions, and real-world tasks. We also present further results on the following synthetic benchmark functions: Branin, Hartmann, Ackley, Levy, Rosenbrock, and Rastrigin, with dimensions ranging from 2 to 8. Bayesian optimization is performed using both the default BoTorch RBF hyperprior and the SAASBO hyperprior [26], comparing upper confidence bound and expected improvement as acquisition functions. For all experiments, we use a small noise term $\sigma_\epsilon^2 = 10^{-3}$. Figures 10–13 show the resulting simple log-regret curves, averaged over 30 random seeds, with shaded regions indicating $\pm$ one standard error.

Across all benchmarks, hierarchical approaches consistently outperform the single GP baseline. For very low-dimensional tasks (e.g., Branin in 2 dimensions and Hartmann in 3 dimensions), the differences among hierarchical methods are small, as these functions are relatively easy to optimize. For higher-dimensional synthetic functions (dimensions 4–8), differences between aggregation strategies become more pronounced. Notably, median-based aggregation of decision points ($\mathcal{X}$-median) performs best when using EI as the acquisition function. This behavior is consistent across both hyperpriors and also with the results in the main paper. For UCB, the differences among hierarchical approaches are generally smaller. Overall, these results confirm that hierarchical aggregation improves performance over single-task baselines, with the largest gains observed in slightly higher-dimensional synthetic functions and when using EI with $\mathcal{X}$-median aggregation.

B.3 Statistical Significance

To assess whether the differences between methods are statistically significant, we evaluate all 42 benchmark $\times$ acquisition-function settings (18 from the main paper and 24 from the appendix) and compare every pair of methods within each setting. Each comparison is a paired Wilcoxon signed-rank test on the final regret, with runs paired by seed; a win denotes significantly lower regret ($p < 0.05$), a loss significantly higher regret, and a tie no significant difference. Each method thus takes part in $42 \times 6 = 252$ comparisons. Table 4 shows that medoid $\mathcal{X}$ achieves the most wins and loses only a single comparison (against median $\mathcal{Q}$ on Levy 6D with the SAASBO hyperprior and UCB, $p = 0.049$). Most comparisons among the hierarchical methods are ties, whereas all of them clearly outperform the single GP. With a Holm correction over the 21 pairs within each setting, medoid $\mathcal{X}$ records 23 wins, 229 ties and no loss, and is the only method without a loss. The advantage of medoid $\mathcal{X}$ thus lies more in robustness than in large performance gains: on most settings it performs on par with the best alternative, but in contrast to every other aggregation strategy it is almost never significantly worse than any competitor.

Table 4: Pairwise comparison of all methods over all evaluated 42 settings; a win denotes significantly lower regret, a loss significantly higher regret, and a tie no significant difference.

Method	Wins	Ties	Losses
medoid $\mathcal{X}$	59	192	1
F.B. mean	50	193	9
median $\mathcal{Q}$	49	191	12
UCB $\mathcal{Q}$	39	195	18
F.B. median	41	186	25
mean $\mathcal{Q}$	40	184	28
single GP	5	57	190

B.4 Computational overhead

As already noted in the main paper, decision-level aggregation requires M AF optimisations per iteration instead of one, which introduces a clear overhead. In our experiments we used $M = 16$, yielding runtimes that we present in Table 7. We argue, that they remain modest compared to both MCMC-based GP inference and especially function evaluations (hours to even days) in real-world applications. Moreover, the M AF optimisations are independent and thus trivially parallelisable, so this overhead can be largely eliminated in practice.

Table 5: Lengthscale configurations used in the toy experiment with increasing proportions of misspecified models.

Setting	Proportion of outlier models	Lengthscales of the GPs
0	0%	[0.0188, 0.0192, 0.0185, 0.0198, 0.0200, 0.0205, 0.0210, 0.0201, 0.0213]
1	11.1%	[5.2, 0.0192, 0.0185, 0.0198, 0.0200, 0.0205, 0.0210, 0.0201, 0.0213]
2	22.2%	[5.2, 5.8, 0.0192, 0.0185, 0.0198, 0.0200, 0.0205, 0.0210, 0.0201]
3	33.3%	[5.2, 5.8, 6.2, 0.0185, 0.0198, 0.0200, 0.0205, 0.0210, 0.0201]
4	44.4%	[5.2, 5.8, 6.2, 5.4, 0.0185, 0.0198, 0.0200, 0.0205, 0.0210]

Table 6: Benchmark functions used in our experiments, grouped by task type.

Category	Function/Task	Dimension
Synthetic	Branin	2
	Levy	3, 6, 32
	Hartmann	3, 6
	Ackley	3, 6, 50
	Griewank	3, 50
	Rosenbrock	4
	Rastrigin	4
Real-world/simulated	Breast Cancer (HPO)	10
	Robot Pushing	14
	Rover Trajectory	60

Table 7: Per-iteration wall-clock cost on BBOB (mean $\pm$ standard error, seconds).

d	MCMC GP fit [s]	AF opt. (single) [s]	$M\times$ AF opt. ($M = 16$) [s]
2	27.55 ± 1.02	1.22 ± 0.07	19.5 ± 1.0
4	31.04 ± 1.09	1.80 ± 0.08	28.8 ± 1.3
8	34.82 ± 1.04	2.85 ± 0.22	45.6 ± 3.5
16	43.09 ± 1.15	5.20 ± 0.61	83.2 ± 9.8

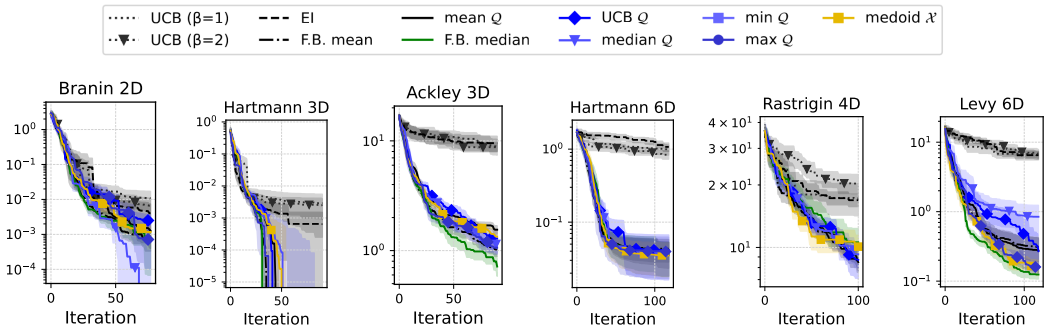

Figure 10: Simple log-regret of low-dimensional optimization results using the **default BoTorch hyperprior** and **UCB as AF**.

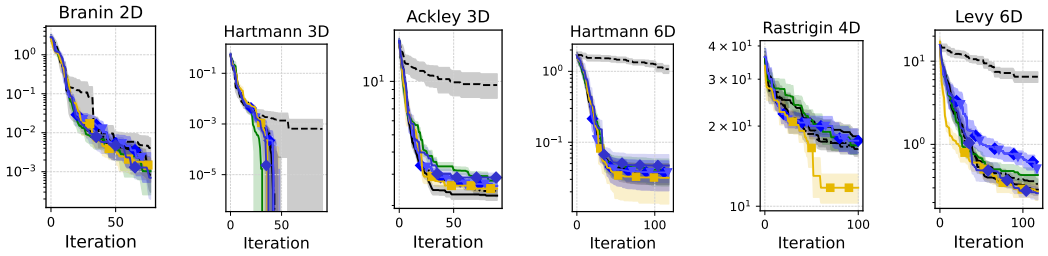

Figure 11: Simple log-regret of low-dimensional optimization results using the **default BoTorch hyperprior** and **EI as AF**.

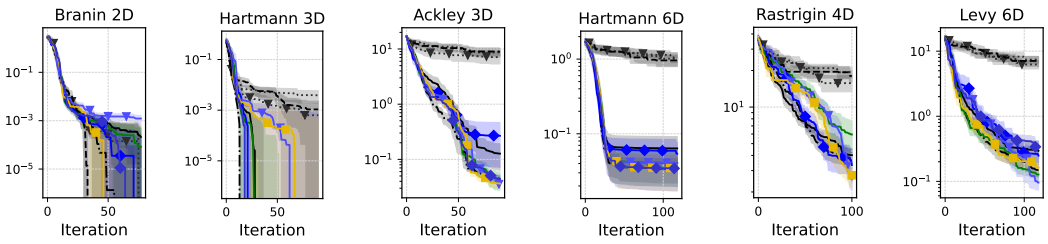

Figure 12: Simple log-regret of low-dimensional optimization results using the **SAASBO hyperprior** and **UCB as AF**.

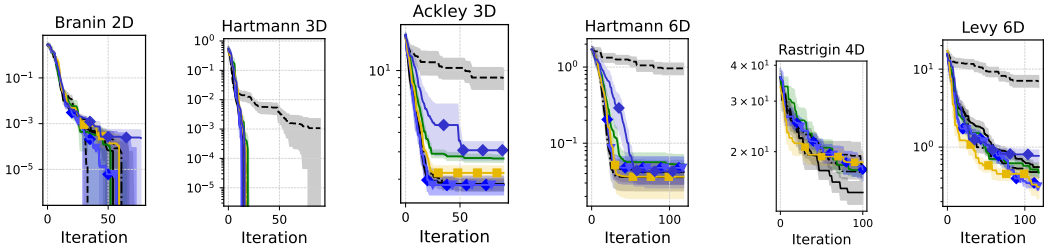

Figure 13: Simple log-regret of low-dimensional optimization results using the **SAASBO hyperprior** and **EI as AF**.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Yes, the claims made in the abstract and introduction reflect the paper's contributions. We study settings with a minority of outlier models, propose to delay aggregation to the decision level for increased robustness, and provide theoretical regret guarantees as well as empirical validation.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Limitations of the paper are discussed in Section 7.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: Full set of assumptions are given and the proofs are deferred to the Appendix A.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provided detailed explanation of the experiment formulations in Appendix B.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: A link to a repository containing the full code to reproduce the results is provided (Section 6).

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The experimental setting is described in the Appendix B, including Gaussian process fitting and details on the hyperpriors for the parameters.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: All Bayesian optimization curves are reported with standard error bars across multiple random seeds.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Computational resources are reported in Appendix B.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Yes, the paper follows NeurIPS Code of Ethics with no harms caused by the research process.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The paper thoroughly discusses the societal impacts in the broader impact statement in Section 7.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not present such risks.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: Yes, all data and materials used are publicly available under the MIT License.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We release the code used to run all experiments via the project repository, including a README with installation instructions, scripts to reproduce each experiment, and the license under which the code is shared.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.

- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.