
Optimizing Hyperparameters with Conformal Quantile Regression

David Salinas¹ Jacek Golebiowski¹ Aaron Klein¹ Matthias Seeger¹ Cedric Archambeau¹

Abstract

Many state-of-the-art hyperparameter optimization (HPO) algorithms rely on model-based optimizers that learn surrogate models of the target function to guide the search. Gaussian processes are the de facto surrogate model due to their ability to capture uncertainty but they make strong assumptions about the observation noise, which might not be warranted in practice. In this work, we propose to leverage conformalized quantile regression which makes minimal assumptions about the observation noise and, as a result, models the target function in a more realistic and robust fashion which translates to quicker HPO convergence on empirical benchmarks. To apply our method in a multi-fidelity setting, we propose a simple, yet effective, technique that aggregates observed results across different resource levels and outperforms conventional methods across many empirical tasks.

1. Introduction

Hyperparameters play a vital role in the machine learning (ML) workflow. They control the speed of the optimization process (e.g., learning rate), the capacity of the underlying statistical model (e.g., number of units) or the generalization quality through regularization (e.g., weight decay). While virtually all ML pipelines benefit from having their hyperparameters tuned, this can be tedious and expensive to do, and practitioners tend to leave many of them at their default values.

Hyperparameter optimization (Feurer & Hutter, 2018) is a powerful framework to tune hyperparameters automatically with a large body of work to tackle this optimization problem, spanning from simple heuristics to more complex model-based methods. Random search-based approaches (Bergstra & Bengio, 2012) define a uniform dis-

tribution over the configuration space and repeatedly sample new configurations until a total budget is exhausted. Evolutionary algorithms modify a population of hyperparameter configurations by hand-designed mutations. Finally, model-based methods, such as Bayesian optimization (Snoek et al., 2012) use the collected data points to fit a surrogate model of the target objective, which informs the sampling distribution so that more promising configurations are selected over time.

Unfortunately, solving the HPO problem can become computationally expensive, as each function evaluation incurs the cost of fully training and validating the underlying machine learning model. Multi-fidelity HPO (Karnin et al., 2013; Li et al., 2017) accelerates this process by terminating the evaluation of under-performing configurations early and only training promising configurations to the end. Early stopping allows algorithms to explore more configurations within the total search budget. Baseline multi-fidelity methods can be made more efficient by relying on model-based sampling, exploiting past observations to select more promising configurations over time.

Model-based multi-fidelity methods constitute the state-of-the-art in HPO today, but still face some limitations. Concretely, we identify two major gaps with the current approaches. First, Bayesian optimization usually assume that the observation noise is homoskedastic and distributed as a Gaussian. This allows for an analytic expression of the likelihood for Gaussian processes, the most popular probabilistic model for Bayesian optimization (Snoek et al., 2012). However, most HPO problems exhibit heteroskedastic noise that can span several orders of magnitude (Salinas et al., 2020; Cowen-Rivers et al., 2020). For instance, the validation error of a model trained with SGD can behave very sporadically for large learning rate and not taking this heteroskedasticity into account can severely hinder the performance of methods with Gaussian assumptions.

Second, it is difficult to model probabilistically the target metric both across configurations and resources (e.g., number of epochs trained), while at the same time retaining the simplicity of Gaussian processes. Previous work maintains separate models of the target at different resource levels (Falkner et al., 2018; Li et al., 2022), or use a single

¹Amazon Web Services. Correspondence to: David Salinas <david.salinas.pro@gmail.com>.

joint model (Klein et al., 2020; Swersky et al., 2014). The former does not take into account dependencies across resource levels, even though these clearly exist. The latter has to account for the non-stationarity of the target function, and requires strong modeling assumptions in order to remain tractable.

With this paper, we propose a conformalized quantile regression surrogate model that can be used in the presence of any noise, possibly non Gaussian or heteroskedastic. We further propose a new way to expand any model-based single-fidelity HPO method to the multi-fidelity setting, by aggregating observed results across different resource levels. This strategy can be used with most single-fidelity method and greatly simplifies the model based multi-fidelity setup. We show that, in spite of its simplicity, our framework offers competitive results across most common single-fidelity methods and significant improvements over baselines when paired with conformalized quantile regression. Our main contributions are the following:

- We introduce a novel surrogate method for HPO based on conformalized quantile regression which can handle heteroskedastic and non-Gaussian distributed noise.
- We propose a simple way to extend any single-fidelity method into a multi-fidelity method, by using only the last observed datapoint for each hyperparameter configuration to train the function surrogate.
- We run empirical evaluations on a large set of benchmarks, demonstrating that quantile regression surrogates achieve a more robust performance compared to state-of-the-art methods in the single-fidelity case
- We show that our new multi-fidelity framework outperforms state-of-the-art methods across multiple single-fidelity surrogates.

The paper first reviews related work and then discuss our proposed method for single-fidelity optimization leveraging conformal quantile prediction. We then describe an extension to the multi-fidelity case, before evaluating the method on an extensive set of benchmarks.

2. Related work

Bayesian optimization is one of the most successful strategies for hyperparameter optimization (HPO) (Shahriari et al., 2016). Based on a probabilistic model of the objective function, it iteratively samples new candidates by optimizing an acquisition function, that balances between exploration and exploitation when searching the space. Typically, Gaussian processes are used as the probabilistic surrogate model (Snoek et al., 2012), but other methods, such

as random forests (Hutter et al., 2011) or Bayesian neural networks (Springenberg et al., 2016; Snoek et al., 2015) are possible. Alternatively, instead of modeling the objective function, previous work (Bergstra et al., 2011; Tiao et al., 2021) estimate the acquisition function directly by the density ratio of well and poorly performing configurations.

Despite its sample efficiency, Bayesian optimization still requires tens to hundreds of function evaluations to converge to well-performing configurations. To accelerate the optimization process, multi-fidelity optimization exploits cheap-to-evaluate fidelities of the objective function such as training epochs (Swersky et al., 2014). Jamieson & Talwalkar (2016) proposed to use successive halving (Karnin et al., 2013) for multi-fidelity hyperparameter optimization which trains a set of configurations for a small budget and then only let the the top half configurations continue for twice as many resources. Hyperband (Li et al., 2017) calls successive halving as a sub-routine with varying minimum resources level, to avoid that configurations are terminated too early. Falkner et al. (2018) combined Hyperband with Bayesian optimization to replace the inefficient random sampling of configuration by Bayesian optimization with kernel density estimators (Bergstra et al., 2011). ASHA (Li et al., 2019) proposed to extend Hyperband to the asynchronous case when using multiple workers which led to significant improvements and Klein et al. (2020); Li et al. (2022) later combined this method with Gaussian process based Bayesian optimization. Instead of relying on a model-based approach, Awad et al. (2021) instead proposed to combine Hyperband with evolution algorithms.

An orthogonal line of work models the learning curves of machine learning algorithms directly; see Mohr & van Rijn (2022) for an overview. Previous work by Domhan et al. (2015) fits an ensemble of parametric basis functions to the learning curve of a neural network. This method can be plugged into any HPO approach such that evaluation of a network is stopped if it is unlikely to outperform previous configurations and the prediction of the model is returned to the optimizer. Klein et al. (2017) used a Bayesian neural networks to predict the parameters of these basis functions which is able to model the correlation across hyperparameter configurations. Wistuba & Pedapati (2020) proposed neural networks architectures that ranks learning curves across different tasks.

To avoid requiring Gaussian homoskedastic noise, several papers considered the use of quantile regression for HPO (Picheny et al., 2013; Salinas et al., 2020; Moriconi et al., 2020) but those approaches do not ensure that the provided uncertainties are well calibrated. Conformal prediction has been gaining traction recently in ML applications due to its ability to provide well calibrated uncertainty with widely applicable assumptions, in particular not requiring the pres-

ence of a Gaussian homoskedastic distribution (Shafer & Vovk, 2007). To the best of our knowledge, conformal prediction has only been considered for single-fidelity HPO by (Stanton et al., 2022) and Doyle (2022). The former applies conformal correction to standard GP posteriors in order to improve model calibration on non-Gaussian noise, whereas we build our method on quantile regression which is already robust to non-Gaussian and heteroskedastic noise. The latter conducted a preliminary study showing that conformal predictors can outperform random-search on four datasets. The key difference to our method is that we utilize the framework of conformal quantile prediction from (Romano et al., 2019) which leverages quantile regression allowing to bypass the need to fit an additional model for the variance. In both cases, our work differs as we consider the asynchronous multi-fidelity case which allows the method to perform much better in presence of hundreds of observations.

3. Single-fidelity Hyperparameter Optimization

In the single-fidelity hyperparameter optimization setting, we are interested in finding the hyperparameter minimizing of a blackbox function f :

$$x^* = \arg \min_{x \in \mathcal{X}} f(x)$$

where $f(x) \in \mathcal{X}$ denotes the validation error obtained for a hyperparameter configuration x . Hyperparameters may include the learning rate, number of layers and number of hidden dimensions of a transformer or a convolutional neural network. Given that evaluating f is typically expensive and gradients are not readily available, we look for gradient-free and sample efficient optimization methods.

Bayesian Optimization is arguably one of the most popular approaches owing to its ability to efficiently trade-off exploration and exploitation when searching the configuration space. In each iteration n , a probabilistic model of the objective function f is fitted on the n previous observations $\mathcal{D} = \{(x_1, y_1), \dots, (x_n, y_n)\}$; the first initial configurations are typically drawn at random. To select the next point to evaluate, an acquisition function is then optimized to select the most promising candidate based on the probabilistic surrogate, for instance by picking the configuration that maximizes the expected improvement (Jones et al., 1998).

The standard approach to model the objective function uses Gaussian Processes due its computational tractability and theoretical guarantees (Srinivas et al., 2012). However, this approach assumes that the observation noise is Gaussian and homoskedastic. We next describe an approach to lift this restriction by leveraging quantile regression with

conformal prediction as a probabilistic model for f .

4. Conformal Quantile Regression

Preliminaries. For a real-valued random variable Y , we denote by $g(y)$ its probability density function and by $F_Y(y) = \int_{-\infty}^y g(t)dt$ its cumulative distribution function (CDF). The associated quantile function of Y is then defined as follows:

$$F_Y^{-1}(\alpha) = \inf_{y \in \mathbb{R}} \{F_Y(y) \geq \alpha\}.$$

The quantile function allows to easily obtain confidence intervals. One can also easily sample from the distribution by first sampling a random quantile uniformly $\alpha \sim \mathcal{U}([0, 1])$ and then computing $y = F_Y^{-1}(\alpha)$ which provides one sample y from the distribution Y .

Quantile regression. Given data drawn from a joint distribution $(x, y) \sim F_{(X, Y)}$, quantile regression aims to estimate a given quantile α of the conditional distribution of Y given $X = x$, e.g. to learn the function

$$q_\alpha(x) = F_{Y|X=x}^{-1}(\alpha)$$

which predicts the quantile α conditioned on x . This problem can be solved by minimizing the quantile loss function (Bassett & Koenker, 1982) for a given quantile α and some predicted value \hat{y} :

$$\mathcal{L}_\alpha(y, \hat{y}) := \begin{cases} \alpha(y - \hat{y}) & \text{if } y - \hat{y} > 0, \\ (1 - \alpha)(\hat{y} - y) & \text{otherwise.} \end{cases} \quad (1)$$

A critical property is that minimizing the quantile loss allows to retrieve the desired quantile in the sense that

$$\arg \min_{\hat{y}} \mathbb{E}_{y \sim F_Y} [\mathcal{L}_\alpha(y, \hat{y})] = F_Y^{-1}(\alpha).$$

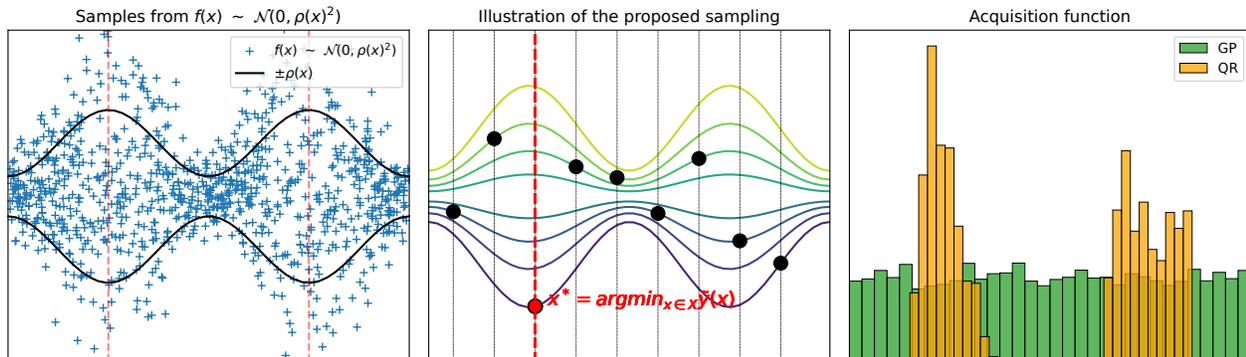
Given a set of n observations $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$, one can thus estimate the quantile function by training a model \hat{q}_α with parameters θ that minimizes the quantile loss given by Eq. 1:

$$\theta^* = \arg \min_{\theta} \frac{1}{n} \sum_{i=1}^n \mathcal{L}_\alpha(y_i, \hat{q}_\alpha(x_i)). \quad (2)$$

Quantile Regression Surrogate. We now explain how we can leverage quantile regression to build a probabilistic surrogate for Bayesian Optimization.

To this end, we estimate m models by minimizing Eq. 2 for equally spread-out quantiles $\{\alpha_1, \dots, \alpha_m\}$ where

Figure 1: Samples from the synthetic function $f(x) \sim \mathcal{N}(0, \rho(x)^2)$ to be minimized (left), illustration of the Thompson sampling procedure based on $m = 8$ predicted quantiles (middle) and acquisition function obtained for our method and a GP (right). When sampling, we sample one random quantile for each candidate and pick the best configuration obtained.



$\alpha_j = j/(m+1)$ and where $m > 1$ is an even number. This allows to provide probabilistic predictions conditioned on any hyperparameter x . We then use independent Thompson sampling as the acquisition function by taking advantage that samples can easily be obtained through predicted quantiles. Indeed, one can then draw a sample $\tilde{y}(x)$ from the estimated conditional distribution of $F_{Y|X}$ by simply sampling a quantile at random among the m quantiles predicted by the model $\{\hat{q}_{\alpha_1}(x), \dots, \hat{q}_{\alpha_m}(x)\}$.¹

To select the next configuration to evaluate, we then use independent Thompson sampling as the acquisition function. We first sample a set of N candidates $\tilde{\mathcal{X}} = \{\tilde{x}_1, \dots, \tilde{x}_N\}$ uniformly at random and then sample a validation performance for each of those candidates to obtain $\{\tilde{y}(\tilde{x}_1), \dots, \tilde{y}(\tilde{x}_N)\}$. We then return the configuration that has the lowest sampled value

$$x^* = \underset{x \in \{\tilde{x}_1, \dots, \tilde{x}_N\}}{\operatorname{arg\,min}} \tilde{y}(x).$$

We illustrate this procedure in Figure 1 which shows how different quantiles are sampled in a toy example and how the next point is selected by picking the configuration with the lowest sampled predicted value. In this example, we consider $f(x) \sim \mathcal{N}(0, \rho(x)^2)$ as the function to minimize with $\rho(x) = \sin(x)^2 + \kappa$ where κ is set to $\kappa = 0.3$ to ensure positive variance. Samples of this function are shown in Figure 1, left. For this function, a standard GP is not able to represent the heteroskedastic variance and as such cannot favor any part of the space as shown by its uniformly distributed acquisition. However, while the mean

¹One downside typically associated with fitting m models is that the quantiles predicted may not be monotonic (Gasthaus et al., 2019), however this problem does not occur in our case since we simply use samples from the predicted distribution.

of the function is always zero, the optimal points to sample are both situated at $\pi/2$ and $3\pi/2$ where the uncertainty is the highest. While, the GP cannot model this information given its homoskedastic noise, quantile regression is able to regress the conditional quantiles and therefore correctly identify the best regions to sample as evidenced by the acquisition function which peaks at the two best values for the next candidate.

Conformalizing predictions. While quantile regression can learn the shape of any continuous distribution given enough data, the model predictions are not guaranteed to be well calibrated given insufficient data.

More precisely, the quantiles estimated allows us to construct $m/2$ confidence intervals

$$\mathcal{C}_j(x) = [\hat{q}_{\alpha_j}(x), \hat{q}_{1-\alpha_j}(x)]$$

for $j \leq m/2$.² For each confidence interval, we would like to have a miscoverage rate $2\alpha_j$, i.e. the predictions should have probability at least $1 - 2\alpha_j$ of being in the confidence interval $\mathcal{C}_j(x)$,

$$\mathbb{P}[Y \in \mathcal{C}_j(x)] = 1 - 2\alpha_j. \quad (3)$$

In the presence of heteroskedasticity, this requires to have the length of $\mathcal{C}_j(x)$ to depend on x which is possible with the use of quantile regression as illustrated in Figure 1. However, the coverage statement of Eq. 3 cannot be guaranteed when fitting models on finite sample size. Miscalibrated intervals can be problematic for HPO, as it may lead to a model converging early to a suboptimal solution. To

²Note that the values chosen for the quantiles $\alpha_j = j/(m+1)$ ensures that the quantile $1 - \alpha_j$ belongs to the m quantiles computed since $1 - \alpha_j = \alpha_{m-j}$.

address this problem, we propose using the split conformal method from Romano et al. (2019) that allows to obtain robust coverage for each $m/2$ of the predicted confidence intervals that we now describe.

The method consists in applying an offset on each confidence interval, which is estimated on a validation set. We divide the dataset of available observations $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$ into a training set $\mathcal{D}_{\text{train}}$ and a validation set \mathcal{D}_{val} . After fitting each of the m quantile regression models \hat{q}_{α_j} on the training set $\mathcal{D}_{\text{train}}$, we compute conformity scores that measure how well the predicted conformal intervals fit the data for each sample in the validation set:

$$E_j = \left\{ \max(\hat{q}_{\alpha_j}(x_i) - y_i, y_i - \hat{q}_{1-\alpha_j}(x_i)) \right\}_{i=1}^{|\mathcal{D}_{\text{val}}|}. \quad (4)$$

The intuition of the conformity scores is the following. First note that the sign of the score is positive when the target y_i is outside of the target and negative when the target falls inside the predicted interval. This allows the score to account for both overcoverage and undercoverage cases as we want to reduce the interval in cases of overcoverage and increase it in case of undercoverage. In addition, the score amplitude always measures the distance to the closest quantile of the confidence interval, i.e. the score amplitude of each sample is $|y_i - q_i|$ where q_i is the closest quantile from y_i between $\hat{q}_{\alpha_j}(x_i)$ and $\hat{q}_{1-\alpha_j}(x_i)$.

Given this score we compute a correction γ_j which is set to

$$\gamma_j = (1 - 2\alpha_j) \left(1 + \frac{1}{|\mathcal{D}_{\text{val}}|} \right)\text{-th empirical quantile of } E_j. \quad (5)$$

The conformalized prediction interval for a new data point (x, y) is then given by

$$\hat{\mathcal{C}}_j(x) = [\hat{q}_{\alpha_j}(x) - \gamma_j, \hat{q}_{1-\alpha_j}(x) + \gamma_j]. \quad (6)$$

An important property of this procedure is that the corrected confidence intervals are guaranteed to have valid coverage, e.g. the probability that y belongs the prediction interval $\hat{\mathcal{C}}_j$ can then be shown to arbitrarily close to $1 - 2\alpha_j$ (Romano et al., 2019)

$$1 - 2\alpha_j \leq \mathbb{P}[Y \in \hat{\mathcal{C}}_j(x)] \leq 1 - 2\alpha_j + \frac{1}{|\mathcal{D}_{\text{val}}| + 1}.$$

Once the confidence intervals are readjusted by offsetting quantiles, we can sample new candidates to evaluate using a protocol based on Thompson sampling discussed in the previous section while being able to guarantee coverage properties of our predicted quantiles. The pseudo-code of

the proposed algorithm to select the next configuration to evaluate is given in Algo. 1 where the key three steps are 1) fitting quantile regression models, 2) computing quantile adjustments and 3) sampling the best candidate with independent Thompson sampling.

5. Multi-fidelity and Successful Halving

Single-fidelity methods steer the search towards the most promising part of the configuration space based on the observed validation performance of hyperparameter configurations, however, this does not consider leveraging other available signals, such as the loss emitted at each epoch for a neural network trained with SGD. Multi-fidelity methods consider this additional information to further accelerate the search by cutting poor configurations preemptively.

Formally, multi-fidelity optimization considers the following optimization problem:

$$x^* = \arg \min_{x \in \mathcal{X}} f(x, r_{\max})$$

where $f(x, r_{\max})$ denotes the blackbox error obtained for a hyperparameter x at the maximum budget r_{\max} (for instance the maximum number of epochs) and we assume that $r \in [r_{\min}, r_{\max}]$. Typically, early values of $f(x, r)$ for $r < r_{\max}$ are informative of $f(x, r_{\max})$, while being computationally cheaper to collect and can help us to cut poorly performing configurations.

Asynchronous Successive Halving. Asynchronous Successive Halving (ASHA) (Li et al., 2019) is a multi-fidelity method that can leverage several workers to evaluate multiple configurations asynchronously while stopping poor configurations early. The method starts by evaluating a list of random configurations in parallel for a small initial budget. When a result is collected from a worker, it is continued or stopped based on its result - the evaluation of the configuration continues if it is in the top results seen so far for a given fidelity and interrupted otherwise. Stopped configurations are replaced with new candidates sampled randomly and the process is iterated until the tuning budget is exhausted.

ASHA avoids synchronization points by evaluating each configuration based on the data available at the time, which can lead to false positives in the continuation decision. Indeed, some configurations may be continued due to poor competition rather than good performance and would have been stopped if more data was available. However, the avoidance of synchronization points efficiently deals with straggler evaluation and is one of the key components of the method's excellent performance in practice. The pseudo-code of the method is given in the appendix.

Algorithm 1 CQR candidate suggestion pseudo-code.

```

1: function SUGGEST()
2:   Input: configuration space  $\mathcal{X}$ , set of observations  $\mathcal{D}$ , number of quantiles  $m$ , number of candidates  $N$ 
3:   Output: next configuration to evaluate
4:    $\mathcal{D}_{\text{train}}, \mathcal{D}_{\text{val}} = \text{split\_train\_val}(\mathcal{D})$ 
5:   for  $1 \leq j \leq m$  do
6:     Fit model  $\hat{q}_{\alpha_j}$  on  $\mathcal{D}_{\text{train}}$  with Eq. 2
7:   end for
8:   for  $1 \leq j \leq m/2 - 1$  do
9:     Compute conformity scores  $E_j$  with Eq. 4
10:    Compute correction  $\gamma_j$  with Eq. 5
11:  end for
12:   $\tilde{\mathcal{X}} = \text{Sample } N \text{ candidates from } \mathcal{X}$ 
13:  for  $1 \leq i \leq N$  do
14:    Draw random quantile  $j$  in  $[1, m]$ 
15:    if  $j < m/2$  then
16:       $\tilde{y}_i = \hat{q}_{\alpha_j}(x_i) - \gamma_j$ 
17:    else
18:       $\tilde{y}_i = \hat{q}_{\alpha_j}(x_i) + \gamma_{m-j}$ 
19:    end if
20:  end for
21:   $i^* = \arg \min_i \tilde{y}_i$ 
22:  return  $x_{i^*}$ 
23: end function

```

Model-based ASHA. One pitfall of ASHA is that candidates are sampled randomly at initialization and when workers become free after configurations are interrupted. However after spending some time in the tuning process, we gathered results which we would clearly like to be able to bias the search towards the most promising parts of the configuration space.

One challenge is that most single-fidelity model-based approaches regress a surrogate model $f(x, r_{\max})$ given observations at the final fidelity r_{\max} . It becomes then difficult to combine model-based and multi-fidelity approaches given that when we stop a poor configuration at a resource $r < r_{\max}$, we are unsure about what would have been the value at $f(x, r_{\max})$.

Bridging single and multi-fidelity methods. We propose a simple data transformation that allows to use any single fidelity method in the multi-fidelity setting. We denote the configurations and evaluations obtained at a given time as

$$\{(x_i, \{f(x_i, r_1), \dots, f(x_i, r_{n_x})\})\}_{i=1}^n$$

where n denotes the number of configurations evaluated and n_x denotes the number of fidelities evaluated for a configuration x .

We propose to consider the transformation that takes the

last value of the time-observations of a given configuration. Namely, we propose to consider the transformation: $z = f(x, r_{n_x})$ and then use a single-fidelity method rather than random-search to determine the best next configuration to evaluate given the observations $\mathcal{D} = \{(x_i, z_i)\}_{i=1}^n$ while using ASHA for the stopping decisions.

Relying on the last observed value $f(x, r_{n_x})$ rather than all fidelities $f(x, r)$ for $r \leq r_{n_x}$ significantly simplifies the multi-fidelity setup but obscures a portion of the available signal. However, as evaluating configuration candidates longer is expected to improve their result, the data transformation is effectively pushing the poor and well performing configurations further apart. Assuming configurations are stopped with a probability inversely correlated with their performance and their results would not cross if the training were to continue, the result-based ordering of configurations remains constant regardless of whether we use the last or final observation. This means that it remains possible under those assumptions to discriminate between promising and not-promising configurations.³

In addition to working well with the Conformal Quantile Regression that we introduced, we will also show in our experiments that this simple transformation allows to combine single-fidelity methods with ASHA while reaching the performance of state-of-the-art dedicated model-based multi-fidelity methods.

6. Experiments

We evaluate our method against state-of-the-art HPO algorithms on a large collection of real-world datasets on both single and multi-fidelity tasks. The code to reproduce our results is available at https://github.com/geoalgo/syne-tune/tree/icml_conformal.

Benchmarks. Our experiments rely on 13 tasks coming from FCNet (Klein & Hutter, 2019), NAS201 (Dong & Yang, 2020) and LCBench (Zimmer et al., 2021) benchmarks as well as NAS301 (Siems et al., 2020) using the implementation provided in (Pfisterer et al., 2022). All methods are evaluated asynchronously with 4 workers. Details on these benchmarks and their configuration spaces distributions are given in Appendix B.

Baselines. For single-fidelity benchmarks, we compare our proposed method (CQR) with random-search (RS) (Bergstra et al.), TPE (Bergstra et al., 2011), Gaussian Process (GP) (Snoek et al., 2012), regularized-evolution

³The accentuated spread between bad and good configurations can be mitigated by using quantile normalization as the transformation is invariant to monotonic changes. We do not report results for this approach as it adds a layer of complexity and performed on-par with just taking the last observations in our experiments.

(REA) (Real et al., 2019) and BORE (Tiao et al., 2021). For multi-fidelity benchmarks, we compare against ASHA (Li et al., 2019), BOHB (Falkner et al., 2018), Hyper-Tune (HT) (Li et al., 2022) and Mobster (MOB) (Klein et al., 2020).

Experiment Setup. All tuning experiments run asynchronously with 4 workers and are stopped when $200 * r_{\max}$ results were observed, which corresponds to seeing 200 different configurations for single-fidelity methods, or when the wallclock time exceeded a fixed budget. All runs are repeated with 30 different random seeds, and we report mean and standard errors. We use gradient boosted trees (Friedman, 2001) for the quantile-regression models with the same hyperparameter used for BORE. We use the simulation backend provided by Syne Tune (Salinas et al., 2022) on a AWS m5.4xlarge machine to simulate methods which allows to account for both optimizers and blackbox runtimes.

Metrics. We benchmark the performance of methods using normalized regret, ranks averaged over tasks and critical diagrams. The normalized regret, also called average distance to the minimum, is defined as $(y_t - y_{\min}) / (y_{\max} - y_{\min})$ where y_{\min} and y_{\max} denotes respectively the best and worse possible values of the blackbox f and y_t denotes the best value found after at a time-step t for a given method. To aggregate across tasks, we report scores at 50 fractions of the total budget for each tasks. We then average normalized regret over each budget proportion across tasks and seeds. We compute ranks for each method at all time-steps and also average those values over tasks and seeds. Critical diagrams show group of methods that are statistically tied together with a horizontal line using the statistical test proposed by (Demšar, 2006). They are computed over averaged ranks of the methods obtained for a fixed budget. The performance of all methods per task is also given in the appendix in Fig. 3, 4, 5.

Results discussion for single-fidelity. Aggregated metrics of single-fidelity methods are shown in Figure 2 top. Our proposed method (CQR) outperforms all baselines in term of both rank and regret. In particular, critical diagrams shows that our method outperforms all baselines at 50% of the tuning budget and is only tied statistically to (BORE) given 100% of the budget. Our results also show that GP-based tuning performs really well in the low sample regime where it can rely on its informative prior. After enough observations, our data-driven approach starts to outperform Bayesian competitor as it can model irregular noise better.

Analyzing surrogate performance. To better understand performance gains, we now analyze the properties of different surrogate models in more detail. In Tab. 1,

Table 1: RMSE, Calibration error and runtime for different surrogates when increasing the number of samples.

model n	RMSE ↓			Calibration error ↓			Runtime ↓		
	GP	QR	CQR	GP	QR	CQR	GP	QR	CQR
16	1.01	0.78	0.81	0.06	0.13	0.13	1.11	1.13	1.06
64	0.92	0.57	0.58	0.04	0.10	0.08	1.71	1.48	1.43
256	0.85	0.43	0.44	0.06	0.05	0.04	2.27	1.75	1.71
1024	0.58	0.37	0.37	0.11	0.04	0.03	20.03	2.23	2.16

we compare the surrogate accuracy (RMSE), the quality of their uncertainty estimate (calibration error) and their runtime. In particular, we measure those metrics for different number of samples n . In each case, we draw a random subset of size n to train the surrogate model and then evaluate the three metrics on remaining unseen examples. Results are averaged across seeds and benchmarks and we normalize the target with quantile normalization. Results per task are also given in the appendix as well as the definition of calibration error metric.

We compare three surrogates: the baseline GP, conformalized quantile regression CQR as well as quantile regression QR. Compared to GP, the RMSE of the boosted-trees surrogates is always better, which is expected as boosted-trees are known for their good performance on tabular data. However, a critical aspect in HPO is to provide good uncertainty estimates in addition to good mean predictors as uncertainty plays a critical in balancing exploration versus exploitation.

To measure the quality of uncertainty estimates, we analyze the calibration error of the different surrogates which measures how much over or under confident are each predicted quantiles of the surrogate predictive distribution. When few observations are available (e.g. when $n \leq 64$), the quality of uncertainty estimates of GP is better compared to both boosted tree-methods, which is expected as GP can rely on their prior in this regime whereas data-driven QR and CQR lack the amount of data to estimate quantile levels well enough. The lack of data also means that CQR cannot adjust confidence intervals accurately given that its validation set is too small and its calibration performance just matches QR. However, as the number of samples increases, the calibration of tree-based methods quickly becomes better, which underlines that quantile regression better fits the noise function observed in the benchmarks. As expected given the theoretical coverage guarantees, the calibration of CQR exceeds the calibration of QR given sufficient data making it a better suited surrogate for HPO.

Results discussion for multi-fidelity. Next, we analyze the performance in the multi-fidelity setting in the middle of Fig. 2 where we show the performance of (CQR+MF)

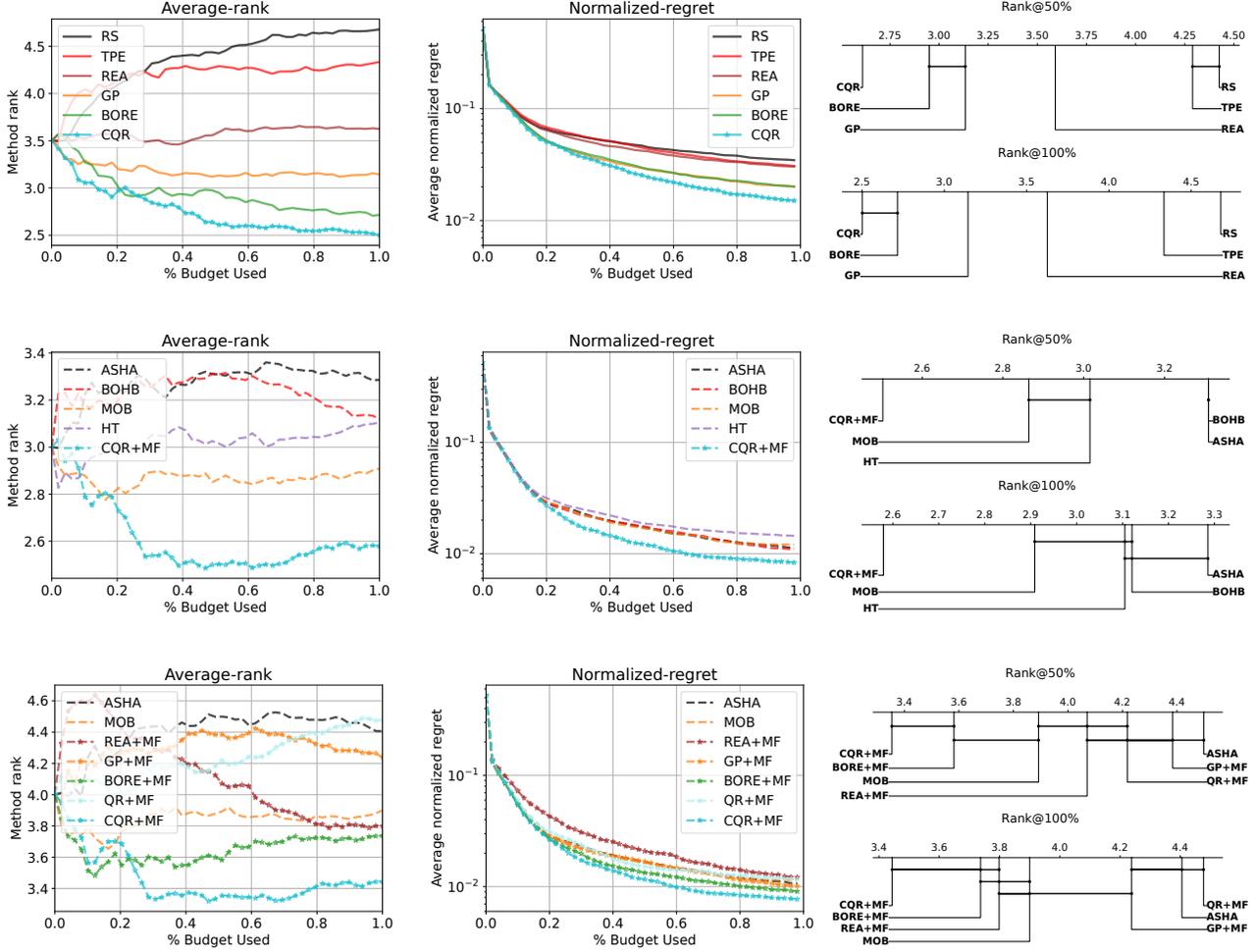


Figure 2: Performance for single fidelity (top) multi-fidelity (middle) and multi-fidelity variants (bottom) for average rank over all tasks (left), normalized regret (middle) and critical diagrams obtained at 50% and 100% of the total budget (right).

which combines the single-fidelity method with the simple transformation described in section 5.

In contrast to single-fidelity, multi-fidelity optimization quickly yields many hundreds of observations and the majority of the tuning process lies in the high-data regime. In this setup, (CQR+MF) shows significant improvement in term of HPO performance. In particular, while most multi-fidelity approaches are not statistically distinguishable from ASHA, our proposed method (CQR+MF) offers statistically significant improvements over ASHA at all times and over all other model-based multi-fidelity methods after spending 50% and 100% of the total budget. We understand the improvement mainly comes from the multi-fidelity setting which offers more observations to the HPO tuning methods which plays into the strengths of the CQR surrogate illustrated in the previous paragraph in term of accuracy and calibration.

Ablation study. The surrogate analysis showed that conformal prediction improves the calibration of quantile regression but has little effect on the surrogate RMSE. To examine the benefit of this contribution on the HPO setting, we next evaluate quantile regression with and without applying conformal correction (QR+MF) in the bottom of Fig. 2. The performance of QR+MF is much worse than CQR+MF which highlights the benefit of the better uncertainty provided by conformalizing predictions.

Next, we investigate in Fig. 2 the performance of the best single-fidelity methods REA, GP and BORE extended to the multi-fidelity setting with our simple extension. As for QR+MF and CQR+MF, all those methods surrogates are trained using the last fidelity observed for each hyperparameter and the worst configurations are stopped with asynchronous successful halving. While those methods perform worse than CQR+MF, they all outperform ASHA in term

of average rank and regret except for REA which we believe is due to the lower performance of the method. Those simple extensions also match or improve over the performance of dedicated model-based multi-fidelity methods.

This illustrates the robustness of the proposed extension with respect to the choice of the single-fidelity method also shows the potential of future work to extend other advanced single fidelity methods - for instance multi-objective or constrained - to the multi-fidelity case.

7. Conclusion

We presented a new HPO approach that allows to use highly accurate tabular predictors, such as gradient boosted trees, while obtaining calibrated uncertainty estimates through conformal predictions. In addition, we showed that most single-fidelity methods can be extended to the multi-fidelity case by just using the last fidelity available while achieving good performance.

The method we proposed has a few limitations. For instance the use of Thompson Sampling may be less efficient in the presence of many hyperparameters, as such further work could consider extending the method with other acquisition functions, such as UCB. Further work could also investigate providing regret bounds or extension to support multi-objective or transfer learning scenarios.

References

- Awad, N., Mallik, N., and Hutter, F. Dehb: Evolutionary hyperband for scalable, robust and efficient hyperparameter optimization. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence (IJCAI'21)*, 2021.
- Bassett, G. and Koenker, R. An empirical quantile function for linear models with iid errors. *Journal of the American Statistical Association*, 77(378):407–415, Jun 1982. ISSN 1537-274X. doi: 10.1080/01621459.1982.10477826. URL <http://dx.doi.org/10.1080/01621459.1982.10477826>.
- Bergstra, J. and Bengio, Y. Random search for hyperparameter optimization. *Journal of Machine Learning Research*, 2012.
- Bergstra, J., Bardenet, R., Bengio, Y., and Kégl, B. Algorithms for hyperparameter optimization. pp. 2546–2554.
- Bergstra, J., Bardenet, R., Bengio, Y., and Kégl, B. Algorithms for hyperparameter optimization. In *Proceedings of the 24th International Conference on Advances in Neural Information Processing Systems (NIPS'11)*, 2011.
- Chen, T. and Guestrin, C. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pp. 785–794, 2016.
- Cowen-Rivers, A. I., Lyu, W., Tutunov, R., Wang, Z., Grosnit, A., Griffiths, R. R., Maraval, A. M., Jianye, H., Wang, J., Peters, J., and Ammar, H. B. Hebo pushing the limits of sample-efficient hyperparameter optimisation, 2020. URL <https://arxiv.org/abs/2012.03826>.
- Demšar, J. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, 7(1):1–30, 2006. URL <http://jmlr.org/papers/v7/demsar06a.html>.
- Domhan, T., Springenberg, J. T., and Hutter, F. Speeding up automatic hyperparameter optimization of deep neural networks by extrapolation of learning curves. In *Twenty-fourth international joint conference on artificial intelligence*, 2015.
- Dong, X. and Yang, Y. NAS-Bench-201: Extending the scope of reproducible neural architecture search. Technical Report arXiv:2001.00326 [cs.CV], 2020.
- Doyle, R. Model agnostic conformal hyperparameter optimization. *arXiv:2207.03017 [cs.LG]*, 2022.
- Falkner, S., Klein, A., and Hutter, F. BOHB: Robust and efficient hyperparameter optimization at scale. In *Proceedings of the 35th International Conference on Machine Learning (ICML 2018)*, pp. 1436–1445, 2018.
- Feurer, M. and Hutter, F. Hyperparameter optimization. In *Automatic Machine Learning: Methods, Systems, Challenges*. Springer, 2018.
- Friedman, J. H. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5):1189–1232, 2001. ISSN 00905364. URL <http://www.jstor.org/stable/2699986>.
- Gasthaus, J., Benidis, K., Wang, Y., Rangapuram, S. S., Salinas, D., Flunkert, V., and Januschowski, T. Probabilistic forecasting with spline quantile function rnns. In Chaudhuri, K. and Sugiyama, M. (eds.), *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pp. 1901–1910. PMLR, 16–18 Apr 2019. URL <https://proceedings.mlr.press/v89/gasthaus19a.html>.
- Hutter, F., Hoos, H., and Leyton-Brown, K. Sequential model-based optimization for general algorithm configuration. In *Proceedings of the Fifth International*

- Conference on Learning and Intelligent Optimization (LION'11)*, 2011.
- Jamieson, K. and Talwalkar, A. Non-stochastic best arm identification and hyperparameter optimization. In *Proceedings of the 17th International Conference on Artificial Intelligence and Statistics (AISTATS'16)*, 2016.
- Jones, D. R., Schonlau, M., and Welch, W. J. Efficient global optimization of expensive black-box functions. *Journal of Global optimization*, 13(4):455, 1998.
- Karnin, Z., Koren, T., and Somekh, O. Almost optimal exploration in multi-armed bandits. In Dasgupta, S. and McAllester, D. (eds.), *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pp. 1238–1246, Atlanta, Georgia, USA, 17–19 Jun 2013. PMLR. URL <https://proceedings.mlr.press/v28/karnin13.html>.
- Klein, A. and Hutter, F. Tabular benchmarks for joint architecture and hyperparameter optimization. Technical Report arXiv:1905.04970 [cs.LG], 2019.
- Klein, A., Falkner, S., Springenberg, J. T., and Hutter, F. Learning curve prediction with Bayesian neural networks. In *International Conference on Learning Representations (ICLR'17)*, 2017.
- Klein, A., Tiao, L., Lienart, T., Archambeau, C., and Seeger, M. Model-based asynchronous hyperparameter and neural architecture search. Number 2003.10865 [cs.LG], 2020.
- Kuleshov, V., Fenner, N., and Ermon, S. Accurate uncertainties for deep learning using calibrated regression. *CoRR*, abs/1807.00263, 2018. URL <http://arxiv.org/abs/1807.00263>.
- Li, L., Jamieson, K., DeSalvo, G., Rostamizadeh, A., and Talwalkar, A. Hyperband: Bandit-based configuration evaluation for hyperparameter optimization. In *International Conference on Learning Representations (ICLR'17)*, 2017.
- Li, L., Jamieson, K., Rostamizadeh, A., Gonina, E., Hardt, M., Recht, B., and Talwalkar, A. Massively parallel hyperparameter tuning. Technical Report 1810.05934v4 [cs.LG], 2019.
- Li, Y., Shen, Y., Jiang, H., Zhang, W., Li, J., Liu, J., Zhang, C., and Cui, B. Hyper-tune: Towards efficient hyper-parameter tuning at scale. *Proc. VLDB Endow.*, 15(6):1256–1265, jun 2022. ISSN 2150-8097. doi: 10.14778/3514061.3514071. URL <https://doi.org/10.14778/3514061.3514071>.
- Mohr, F. and van Rijn, J. N. Learning curves for decision making in supervised machine learning – a survey. *arXiv:2201.12150 [cs.LG]*, 2022.
- Moriconi, R., Kumar, K. S., and Deisenroth, M. P. High-dimensional bayesian optimization with projections using quantile gaussian processes. *Optimization Letters*, 14:51–64, 2020.
- Pfisterer, F., Schneider, L., Moosbauer, J., Binder, M., and Bischl, B. Yahpo gym-an efficient multi-objective multi-fidelity benchmark for hyperparameter optimization. In *First Conference on Automated Machine Learning (Main Track)*, 2022.
- Picheny, V., Ginsbourger, D., Richet, Y., and Caplin, G. Quantile-based optimization of noisy computer experiments with tunable precision. *Technometrics*, 55(1):2–13, 2013. doi: 10.1080/00401706.2012.707580. URL <https://doi.org/10.1080/00401706.2012.707580>.
- Real, E., Aggarwal, A., Huang, Y., and Le, Q. V. Regularized evolution for image classifier architecture search, 2019.
- Romano, Y., Patterson, E., and Candès, E. J. Conformalized quantile regression. 2019. doi: 10.48550/ARXIV.1905.03222.
- Salinas, D., Shen, H., and Perrone, V. A quantile-based approach for hyperparameter transfer learning. In III, H. D. and Singh, A. (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 8438–8448. PMLR, 13–18 Jul 2020. URL <https://proceedings.mlr.press/v119/salinas20a.html>.
- Salinas, D., Seeger, M., Klein, A., Perrone, V., Wistuba, M., and Archambeau, C. Syne tune: A library for large scale hyperparameter tuning and reproducible research. In *International Conference on Automated Machine Learning*, pp. 16–1. PMLR, 2022.
- Shafer, G. and Vovk, V. A tutorial on conformal prediction. *CoRR*, abs/0706.3188, 2007. URL <http://arxiv.org/abs/0706.3188>.
- Shahriari, B., Swersky, K., Wang, Z., Adams, R., and de Freitas, N. Taking the human out of the loop: A review of Bayesian optimization. *Proceedings of the IEEE*, 2016.
- Siems, J., Zimmer, L., Zela, A., Lukasik, J., Keuper, M., and Hutter, F. Nas-bench-301 and the case for surrogate benchmarks for neural architecture search. *CoRR*, abs/2008.09777, 2020. URL <https://arxiv.org/abs/2008.09777>.

- Snoek, J., Larochelle, H., and Adams, R. P. Practical Bayesian optimization of machine learning algorithms. In *Proceedings of the 25th International Conference on Advances in Neural Information Processing Systems (NIPS'12)*, 2012.
- Snoek, J., Rippel, O., Swersky, K., Kiros, R., Satish, N., Sundaram, N., Patwary, M., Prabhat, and Adams, R. Scalable Bayesian optimization using deep neural networks. In *Proceedings of the 32nd International Conference on Machine Learning (ICML'15)*, 2015.
- Springenberg, J. T., Klein, A., Falkner, S., and Hutter, F. Bayesian optimization with robust bayesian neural networks. In *Proceedings of the 29th International Conference on Advances in Neural Information Processing Systems (NIPS'16)*, 2016.
- Srinivas, N., Krause, A., Kakade, S. M., and Seeger, M. W. Information-theoretic regret bounds for gaussian process optimization in the bandit setting. *IEEE Transactions on Information Theory*, 58(5):3250–3265, may 2012. doi: 10.1109/tit.2011.2182033. URL <https://doi.org/10.1109%2Ftit.2011.2182033>.
- Stanton, S., Maddox, W., and Wilson, A. G. Bayesian optimization with conformal coverage guarantees, 2022. URL <https://arxiv.org/abs/2210.12496>.
- Swersky, K., Snoek, J., and Adams, R. P. Freeze-thaw bayesian optimization, 2014. URL <https://arxiv.org/abs/1406.3896>.
- Tiao, L. C., Klein, A., Seeger, M. W., Bonilla, E. V., Archambeau, C., and Ramos, F. BORE: Bayesian optimization by density-ratio estimation. In *International Conference on Machine Learning*, pp. 10289–10300. PMLR, 2021.
- Wistuba, M. and Pedapati, T. Learning to rank learning curves. In III, H. D. and Singh, A. (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 10303–10312. PMLR, 13–18 Jul 2020.
- Zimmer, L., Lindauer, M., and Hutter, F. AutoPyTorch Tabular: Multi-Fidelity MetaLearning for Efficient and Robust AutoDL. *arXiv:2006.13799 [cs, stat]*, April 2021. URL <http://arxiv.org/abs/2006.13799>. arXiv: 2006.13799.

A. ASHA description

We recall how the method ASHA proposed by (Li et al., 2019) works. Given some positive constant $\eta \geq 2$, let us define a finite set of rungs $\mathcal{R} = \{\eta^0 r_{\min}, \eta^1 * r_{\min}, \eta^2 r_{\min}, \dots, r_{\max}\}$. Successive halving (Karnin et al., 2013; Jamieson & Talwalkar, 2016) starts with a set of $N = \eta^K$ initial candidates, where for simplicity, we assume that $K = \log_{\eta} r_{\max}/r_{\min}$. Now, in the first iteration, successive halving collects the performance of all N configurations on r_{\min} and only continues the evaluation of the top $1/\eta$ configurations for the next rung level. This process is iterated until the maximum resource level r_{\max} is reached and repeated until we reach some total budget for the entire search process.

Successive halving can be trivially parallelized in the synchronous setting, however, this will require synchronization points at each rung level to wait for stragglers. Li et al. (2019) adapted successive halving to the asynchronous setting (ASHA), where configurations are immediately promoted to the next rung level, once we observed at least η configurations. While this potentially leads to the promotion of configurations that are not among the top $1/\eta$ configurations, it removes any synchronization overhead and has been shown to perform very well in practice. See Algo. 2 for the pseudo-code of ASHA.

Algorithm 2 ASHA pseudo-code.

```

1: function ASHA()
2:   Input: minimum resource  $r_{\min}$ , maximum resource  $r_{\max}$ , reduction factor  $\eta$ 
3:   repeat
4:     for each free worker do
5:        $(x, k) = \text{get\_job}()$ 
6:        $\text{run\_then\_return\_val\_loss}(x, r_{\min}\eta^k)$ 
7:     end for
8:     for completed job  $(x, k)$  with loss  $l$  do
9:       Update configuration  $x$  in rung  $k$  with loss  $l$ .
10:    end for
11:  until desired
12: end function

13: function get_job()
14:  // Check if there is a promotable config.
15:  for  $k = \lfloor \log_{\eta}(r_{\max}/r_{\min}) \rfloor - 1, \dots, 1, 0$  do
16:    candidates = top_k(rung  $k$ ,  $\lfloor \text{rung } k / \eta \rfloor$ )
17:    promotable =  $\{t \in \text{candidates} : t \text{ not promoted}\}$ 
18:    if  $|\text{promotable}| > 0$  then
19:      return promotable[0],  $k + 1$ 
20:    end if
21:  // If not, grow bottom rung.
22:  Suggest random configuration  $x$ .
23:  return  $x, 0$ 
24: end for
25: end function

```

B. Experiment details

Statistics of different blackboxes are given in Table 2 and their configuration spaces are given in Table 3 together with the base distribution used for each hyperparameters which are used when sampling random candidates. For LCBench, we run the 5 most expensive tasks among the 35 tasks available (“airlines”, “albert”, “covertime”, “christine” and “Fashion-MNIST”). Since LCBench does not contain all possible evaluations on a grid, we run evaluations using a k -nearest-neighbors surrogate with $k = 1$. We use Yahpo implementation (Pfisterer et al., 2022) to get access to NAS301 (Siems et al., 2020).

Schedulers details We give here the list of parameters used for running schedulers in our experiments. In general, the Syne Tune defaults have been used.

Table 2: Tabulated benchmark statistics

Benchmark	#Evaluations	#Hyperparameters	#Tasks	#Fidelities
FCNet	62208	9	4	100
LCBench	2000	7	5	52
NAS201	15625	6	3	200
NAS301	NA	35	1	97

Table 3: Configuration spaces for all tabulated benchmarks.

Benchmark	Hyperparameter	Configuration space	Domain
FCNet	activation_1	[tanh, relu]	categorical
	activation_2	[tanh, relu]	categorical
	batch_size	[8, 16, 32, 64]	finite-range log-space
	dropout_1	[0.0, 0.3, 0.6]	finite-range
	dropout_2	[0.0, 0.3, 0.6]	finite-range
	init_lr	[0.0005, 0.001, 0.005, 0.01, 0.05, 0.1]	categorical
	lr_schedule	[cosine, const]	categorical
	n_units_1	[16, 32, 64, 128, 256, 512]	finite-range log-space
	n_units_2	[16, 32, 64, 128, 256, 512]	finite-range log-space
NAS201	x0	[avg_pool_3x3, nor_conv_3x3, skip_connect, nor_conv_1x1, none]	categorical
	x1	[avg_pool_3x3, nor_conv_3x3, skip_connect, nor_conv_1x1, none]	categorical
	x2	[avg_pool_3x3, nor_conv_3x3, skip_connect, nor_conv_1x1, none]	categorical
	x3	[avg_pool_3x3, nor_conv_3x3, skip_connect, nor_conv_1x1, none]	categorical
	x5	[avg_pool_3x3, nor_conv_3x3, skip_connect, nor_conv_1x1, none]	categorical
LCBench	num_layers	[1, 5]	uniform
	max_units	[64, 512]	log-uniform
	batch_size	[16, 512]	log-uniform
	learning_rate	[1e-4, 1e-1]	log-uniform
	weight_decay	[1e-5, 0.1]	uniform
	momentum	[0.1, 0.99]	uniform
	max_dropout	[0.0, 1.0]	uniform
NAS301	edge-normal-{0-13}	[max_pool_3x3, avg_pool_3x3, skip_connect, sep_conv_3x3, sep_conv_5x5, dil_conv_3x3, dil_conv_5x5]	categorical
	node-normal-{0-13}	[max_pool_3x3, avg_pool_3x3, skip_connect, sep_conv_3x3, sep_conv_5x5, dil_conv_3x3, dil_conv_5x5]	categorical
	inputs-node-normal-{3-5}	[0.1, 0.2, 1.2]	categorical
	inputs-node-reduce-{3-5}	[0.1, 0.2, 1.2]	categorical

- REA is run with a population size of 10, and 5 samples are drawn to select a mutation from
- GP is run using a Matérn $\frac{5}{2}$ kernel with automatic relevance determination parameters. For each suggestion, the surrogate model is fit by marginal likelihood maximization, and a configuration is returned which maximizes the expected improvement acquisition function. This involves averaging over 20 samples of fantasy outcomes for pending evaluations.
- TPE is based on a multi-variate kernel density estimator as proposed by Falkner et al. (2018) to capture interactions between hyperparameters, which is not possible with unit-variant kernel density estimator as used for the original TPE approach (Bergstra et al.). We limit the minimum bandwidth for the kernel density estimator to 0.1 to avoid that all probability mass is assigned to a single categorical value, which would eliminate extrapolation.
- ASHA is running the stopping variant described in (Klein et al., 2020) with grace period 1 and reduction factor 3, so that stopping trials happens after 1, 3, 9, ... epochs. Configurations for new trials are sampled at random.
- BOHB uses the same multi-variate kernel density estimator as TPE, and hyperparameters are set to default values in Falkner et al. (2018). Note that BOHB uses the same asynchronous scheduling as ASHA and MOB, while the algorithm in Falkner et al. (2018) is synchronous.
- MOB is running the same scheduling as ASHA, but configurations for new trials are chosen as in Bayesian optimization. We deal with pending evaluations by averaging the acquisition function over 20 samples of fantasy outcomes.
- BORE is evaluated with XGBoost as the classifier with default hyperparameters (Chen & Guestrin, 2016). We use the default hyperparameters of the method, in particular, $\gamma = 1/4$ which means that the method maximizes the probability that a configuration is in the top 25% of configurations.
- HT uses the same rung levels (given by grace period and reduction factor) as ASHA and fit independent Gaussian process models to data at each rung level. Compared to MOB, HT is using a more advanced acquisition function,

which averages the rung level models with a weighting depending on the fraction of trials which flipped ranks between low and high levels.

- QR and CQR: we estimate $m = 4$ quantiles by fitting gradient boosted trees models with quantile-losses and we use the same hyperparameter as BORE for the boosted-trees. We conformalize only when more than 32 samples are available to avoid poorly estimated correction due to too little samples and use 10% of the data for validation. We sample $N = 2000$ candidates when performing independent Thompson Sampling to select the next candidate.
- $\{\text{REA/GP/BORE/CQR}\} + \text{MF}$: for these methods, we run ASHA by calling each single fidelity method when a new configuration has to be suggested using the data transformation proposed in section 5 to obtain the method observations $\mathcal{D} = \{(x_i, z_i)\}_{i=1}^n$.

All multi-fidelity methods based on successful-halving/ASHA (e.g. all except BOHB) uses a single bracket. We use implementations of all baselines provided in Syne Tune (<https://github.com/awslabs/syne-tune>).

C. Performance per task

We plot the performance of methods on all 13 tasks for single-fidelity methods in Figure 3 and 4 for multi-fidelity methods.

D. Surrogate performance

RMSE. To compute RMSE for quantile predictions, we first compute the mean by taking the average of predicted quantiles.

Calibration error. A calibrated estimator of the conditional distribution $\hat{q}_\alpha(x)$ should exceed the target α percent of the time asymptotically. For instance, we expect the predicted P90-th percentile to exceed the target 90% of the time for a calibrated estimator. Formally, we expect the following property to hold for a calibrated estimator for all $\alpha \in [0, 1]$:

$$\mathbb{E}_{(x,y) \in \mathcal{D}}[\mathbf{1}_{y < \hat{q}_\alpha(x)}] = \alpha$$

where the mean is taken over examples on a validation set $(x, y) \in \mathcal{D}_{\text{val}}$.

The calibration error measures the gap shown in predictions compared to this expected property (Kuleshov et al., 2018). Let us denote $P(\alpha) = \mathbb{E}_{(x,y) \in \mathcal{D}}[\mathbf{1}_{y < \hat{q}_\alpha(x)}]$ the average number of times the prediction $\hat{q}_\alpha(x)$ exceeds the target, then the calibration error is defined over a set of quantiles $\alpha_1, \dots, \alpha_k$ as $\sqrt{\sum_{j=1}^k (P(\alpha_j) - \alpha_j)^2}$. We report this error on 5 equally-spaced quantiles in surrogate experiments. For models predicting quantiles (QR and CQR), we evaluate directly the calibration on the predicted quantiles and for models predicting normal distribution (GP), we compute quantiles predictions using the Gaussian inverse CDF.

Optimizing Hyperparameters with Conformal Quantile Regression

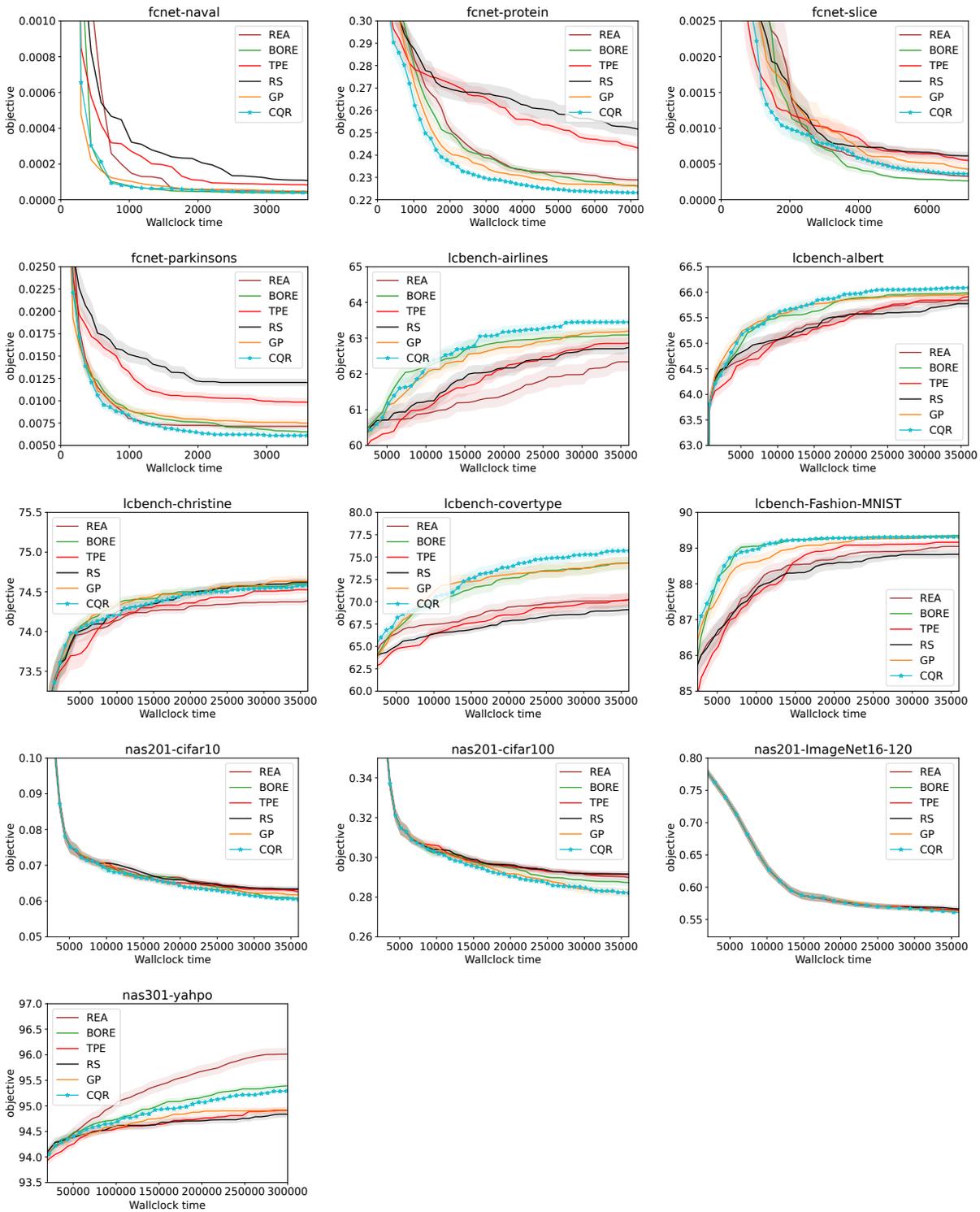


Figure 3: Performance of single-fidelity methods over time on all individual tasks considered. Mean and standard errors are computed over 30 seeds.

Optimizing Hyperparameters with Conformal Quantile Regression

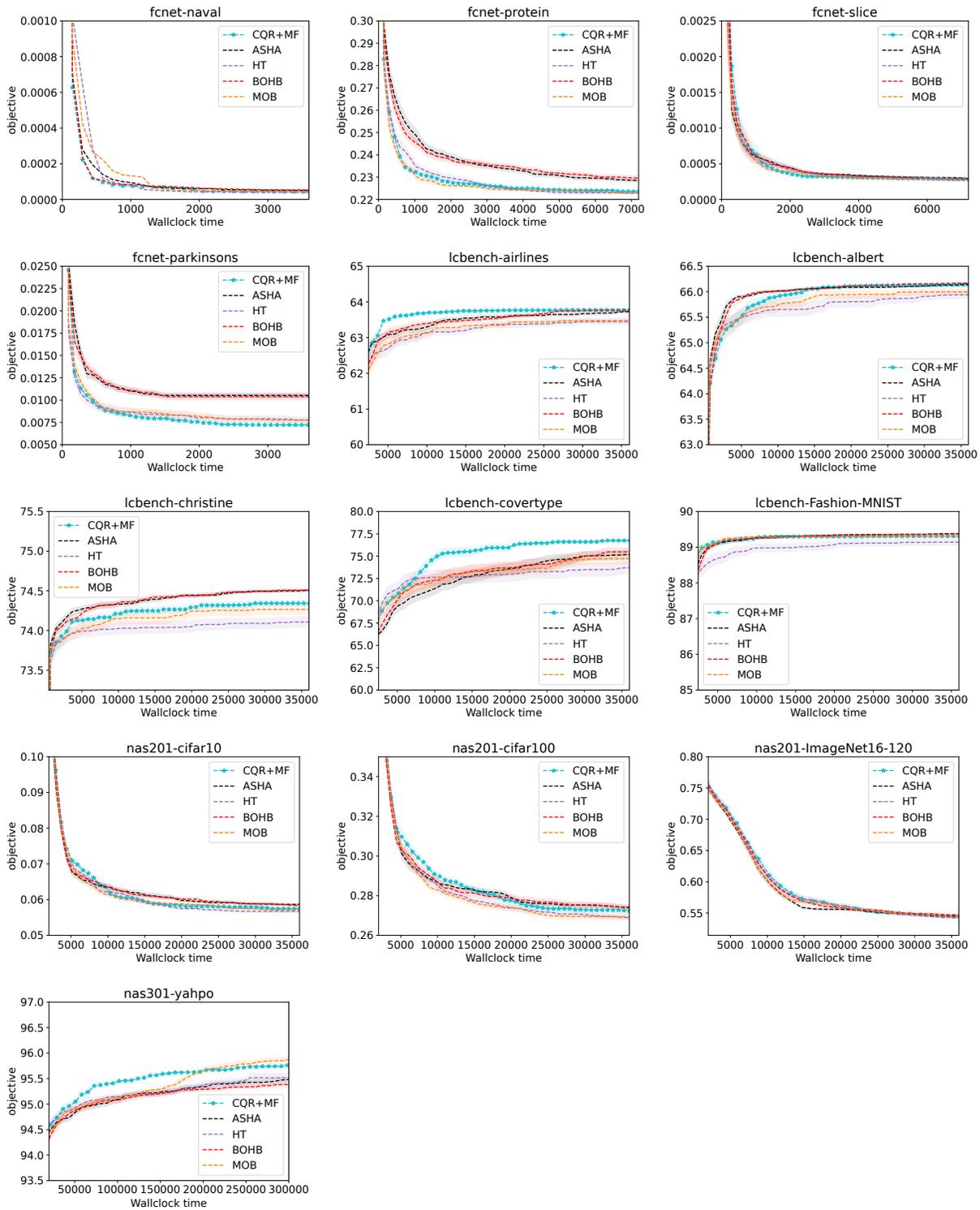


Figure 4: Performance of multi-fidelity methods over time on all individual tasks considered. Mean and standard errors are computed over 30 seeds.

Optimizing Hyperparameters with Conformal Quantile Regression

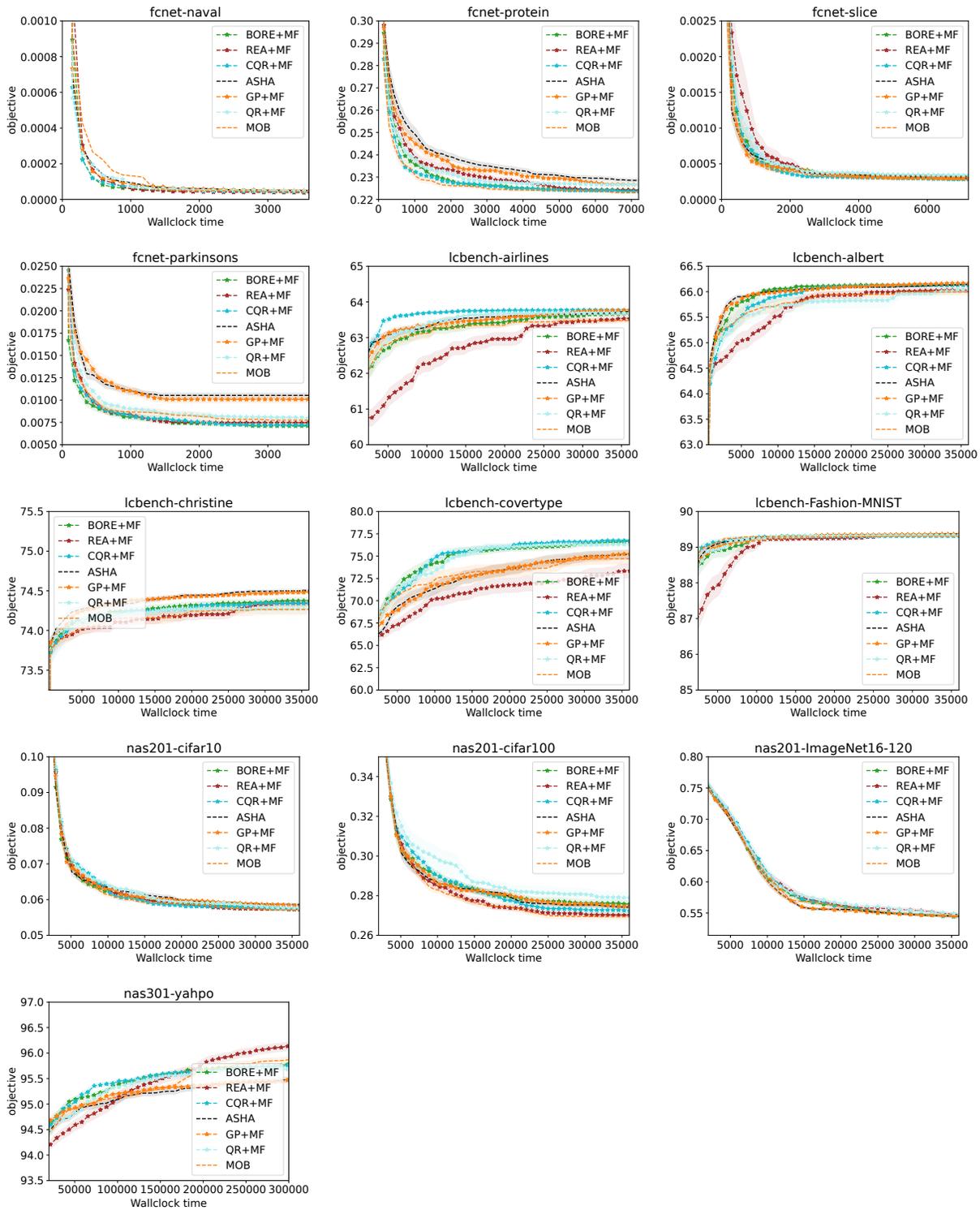


Figure 5: Performance of multi-fidelity ablation variants over time on all individual tasks considered. Mean and standard errors are computed over 30 seeds.

Table 4: Calibration error, Pearson correlation and runtime for different surrogates when increasing the number of samples. Results are averaged over 30 different seeds.

task	model <i>n</i>	RMSE ↓			Calibration error ↓			Runtime ↓		
		GP	QR	CQR	GP	QR	CQR	GP	QR	CQR
airlines	16	1.44	0.89	0.88	0.17	0.13	0.16	1.92	0.95	0.87
	64	1.19	0.68	0.67	0.11	0.09	0.08	3.11	1.36	1.27
	256	1.10	0.51	0.54	0.14	0.07	0.04	3.90	1.67	1.59
	1024	0.92	0.45	0.45	0.17	0.07	0.05	19.02	2.45	2.34
cifar10	16	0.79	0.74	0.81	0.01	0.14	0.13	0.54	1.20	1.14
	64	0.78	0.53	0.54	0.01	0.09	0.06	0.74	1.49	1.41
	256	0.72	0.41	0.42	0.02	0.06	0.04	0.97	1.71	1.73
	1024	0.22	0.35	0.35	0.05	0.03	0.03	19.39	2.11	2.04
parkinsons	16	0.79	0.71	0.72	0.01	0.13	0.11	0.88	1.22	1.17
	64	0.80	0.51	0.53	0.01	0.11	0.09	1.28	1.60	1.60
	256	0.75	0.37	0.37	0.03	0.04	0.04	1.96	1.87	1.81
	1024	0.58	0.31	0.31	0.12	0.02	0.02	21.68	2.14	2.10