

Hyperparameter Optimization via Sequential Uniform Designs

Zebin Yang

Aijun Zhang

Department of Statistics and Actuarial Science

University of Hong Kong

Pokfulam Road, Hong Kong

YANGZB2010@HKU.HK

AJZHANG@UMICH.EDU

Editor: Isabelle Guyon

Abstract

Hyperparameter optimization (HPO) plays a central role in the automated machine learning (AutoML). It is a challenging task as the response surfaces of hyperparameters are generally unknown, hence essentially a global optimization problem. This paper reformulates HPO as a computer experiment and proposes a novel sequential uniform design (SeqUD) strategy with three-fold advantages: a) the hyperparameter space is adaptively explored with evenly spread design points, without the need of expensive meta-modeling and acquisition optimization; b) the batch-by-batch design points are sequentially generated with parallel processing support; c) a new augmented uniform design algorithm is developed for the efficient real-time generation of follow-up design points. Extensive experiments are conducted on both global optimization tasks and HPO applications. The numerical results show that the proposed SeqUD strategy outperforms benchmark HPO methods, and it can be therefore a promising and competitive alternative to existing AutoML tools.

Keywords: Automated machine learning, Hyperparameter optimization, Global optimization, Sequential uniform design, Centered discrepancy.

1. Introduction

Machine learning models are becoming increasingly popular due to their strong predictive performance. Meanwhile, the number of hyperparameters for these models also explodes, and we often have to spend considerable time and energy on hyperparameter tuning (Probst et al., 2019), also known as hyperparameter optimization (HPO). Such HPO procedure is indeed essential but very tedious in machine learning. It is generally accepted that a manual tuning procedure often fails to achieve the best model performance, and faces the critical reproducibility issue. In recent years, the automated machine learning (AutoML) has attracted great attention, highlighting the automatic procedure of hyperparameter tuning.

In this paper, we introduce the classical design of computer experiments to solve the HPO problem with the purpose of maximizing algorithmic prediction accuracy. A computer experiment is defined as a deterministic function or code that is very complicated and time-consuming to evaluate (Fang et al., 2006). We may view HPO as a special computer experiment in the sense that each hyperparameter configuration is an input, and the corresponding predictive performance is the output. In Figure 1, a sequential uniform design

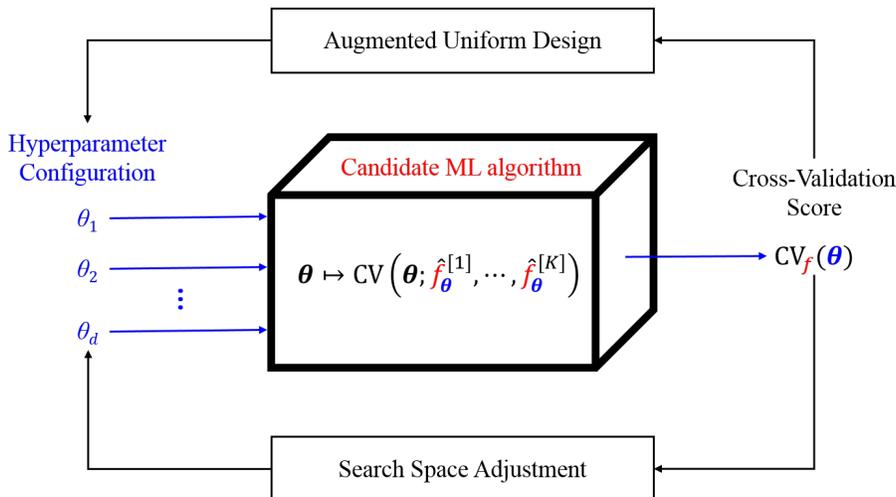


Figure 1: HPO problem reformulated as a kind of computer experiment based on sequential uniform designs.

(SeqUD) approach is proposed for such HPO-type of computer experiment. It is a multi-stage coarse-to-fine optimization framework based on uniform exploration and sequential exploitation. At each stage, the search space is automatically adjusted, and a new batch of design points is augmented with uniformity consideration.

Compared with existing HPO methods in the literature, the proposed SeqUD approach has the following advantages: a) the hyperparameter trials are defined in a sequential manner, as the SeqUD points are generated based on the information of existing design points; b) each time the uniform design of points are considered, which targets the most representative exploration of the search space; c) the SeqUD points generated at the same stage could be evaluated in parallel, which brings additional computation efficiency especially when training large-scale machine learning algorithms. In addition, the proposed method differs from the commonly used Bayesian optimization methods, as SeqUD is free from the time-consuming surrogate modeling and acquisition optimization procedures.

The proposed SeqUD method is tested through extensive synthetic global optimization tasks and real-world HPO experiments. The machine learning algorithms under our consideration include the support vector machine (SVM), extreme gradient boosting (XGBoost), and a machine learning pipeline that involves data preprocessing, feature engineering, model selection, and hyperparameter tuning. Based on the results for a large amount of regression and classification data sets, it is demonstrated that the SeqUD method outperforms Bayesian optimization and other benchmark methods. In summary, this paper contributes to the HPO and AutoML literature in the following three aspects:

- We develop a novel AugUD algorithm for the efficient augmentation of uniform design points. This is not only a new idea in AutoML, but also a new contribution to the field of experimental designs.

- The SeqUD strategy generalizes Bayesian optimization from one-point-at-a-time to batch-by-batch. Meanwhile, it avoids expensive meta-modeling and acquisition optimization procedures. The improved effectiveness and efficiency are achieved.
- Two open-source Python packages are developed, including the *PyUniDOE* package that employs efficient C++ code for generating uniform designs, and the *SeqUD* package that implements the proposed SeqUD method for HPO tasks.

The rest of the paper is organized as follows. Section 2 reviews the related HPO literature. In Section 3, we introduce the background of uniform designs and develop the AugUD algorithm. The new SeqUD method is proposed for HPO in Section 4. A large amount of global optimization experiments and HPO experiments are presented in Section 5 and Section 6, respectively. Finally, we conclude in Section 7 and outline future works.

2. Related Work

This section reviews existing HPO methods from non-sequential and sequential perspectives.

2.1 Non-sequential HPO Methods

Most of the practical HPO methods are non-sequential, as they are easy to implement and can evaluate multiple hyperparameter trials in parallel. For machine learning models with one or two hyperparameters (e.g., SVM with regularization strength and kernel width), it is common to use the exhaustive grid search method (Chang and Lin, 2011). The random search is an alternative method that generates design points randomly in low or high dimensions (Bergstra and Bengio, 2012). It is commonly believed that the random search is more flexible and useful for machine learning tasks.

The space-filling design is an optimal strategy when there is no prior information about hyperparameter distribution (Crombecq et al., 2011), for which the uniform designs (Fang et al., 2000), Sobol sequences (Sobol, 1998) and Latin hypercube sampling (LHS; Kenny et al., 2000) can be used. These methods can generate design points with a low discrepancy from the assumed uniform distribution. Given the same maximal number of runs, the space-filling design has a lower risk of missing the optimal location than the random search. We demonstrate grid search, random search, Sobol sequences, and uniform designs in Figure 2, where each method generates 20 design points in the same two-dimensional space. For the random search, it is observed that design points can be clustered, while lots of other areas remain unexplored. The design points generated by Sobol sequences spread more evenly than the random search. The best coverage is by the uniform design, as it is constructed by optimizing a measure of uniformity to be introduced in Section 3.

2.2 Sequential HPO Methods

Sequential methods are adaptive variants of non-sequential methods, where new design points are generated based on the information of existing design points. Exploration and exploitation are two synergistic objectives of sequential methods. The former aims to better explore the search space, and the latter aims to find the global optimum.

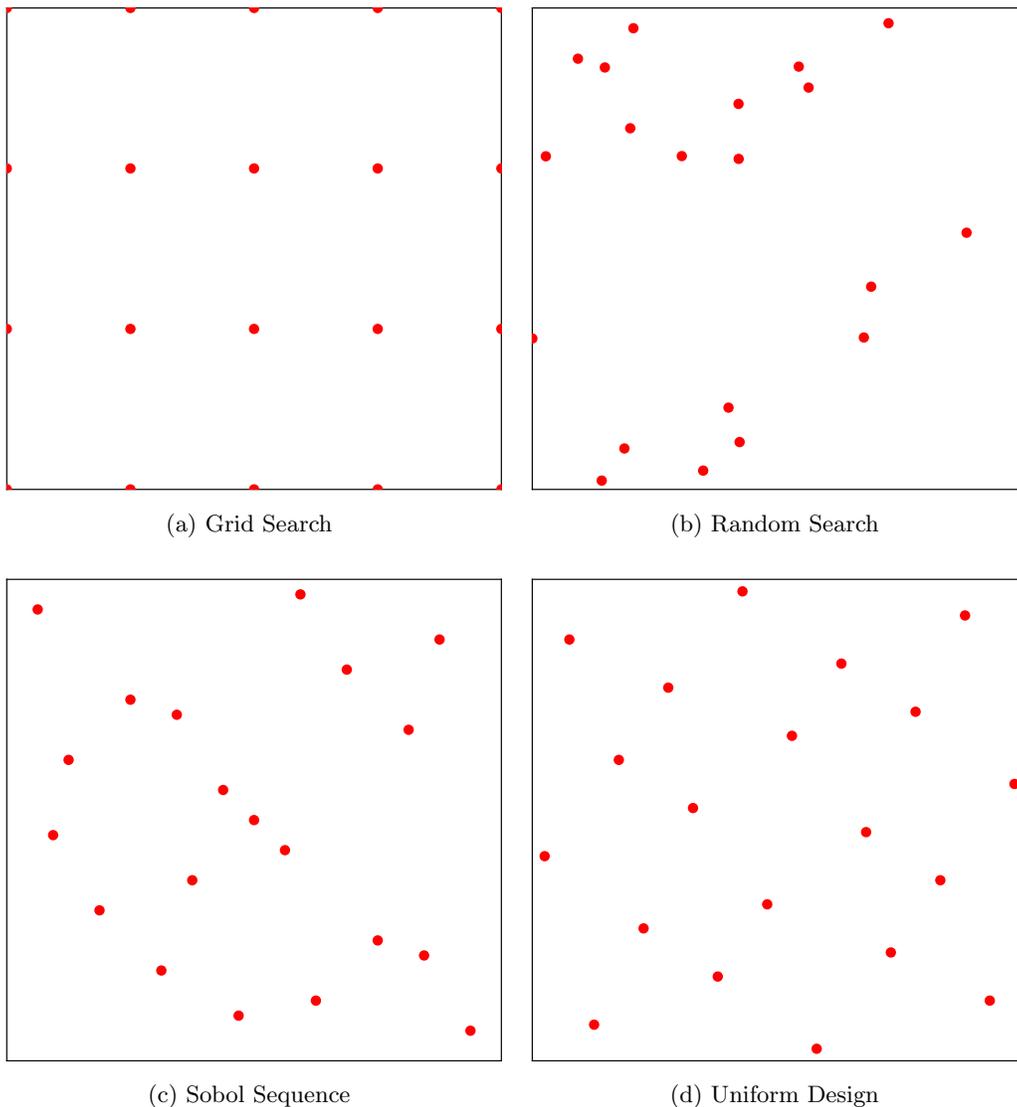


Figure 2: An example that compares four different designs in a 2-D design space.

Bayesian optimization (Jones et al., 1998) is the most widely used sequential approach, which samples one-point-at-a-time in the search space. At each iteration, it fits a surrogate model for approximating the relationship between the design points and the evaluated outcomes. Then, the next design point is generated by optimizing a predefined acquisition function. For HPO tasks in machine learning, the three most influential Bayesian optimization works are arguably the GP (Gaussian process)-EI (expected improvement) method (Snoek et al., 2012), the sequential model-based algorithm configuration (SMAC; Hutter et al., 2011) and the tree-structured parzen estimator (TPE; Bergstra et al., 2011), which we review below.

GP-EI. It uses the GP as the surrogate model and selects the next design point by maximizing the following EI acquisition function:

$$\alpha_{\text{EI}}(\mathbf{x}) = \sigma(\mathbf{x}) \left[z^* \Phi(z^*(\mathbf{x})) + \phi(z^*(\mathbf{x})) \right], \quad (1)$$

where $z^*(\mathbf{x}) = (\mu(\mathbf{x}) - y^*)/\sigma(\mathbf{x})$, y^* is the observed maximum, and $(\mu(\mathbf{x}), \sigma^2(\mathbf{x}))$ are the GP-predicted posterior mean and variance, respectively.

SMAC. It uses the random forest as the surrogate model. Compared to GP, the random forest can be easily scaled up to high-dimensional settings and are more flexible for handling discrete hyperparameters. However, as pointed out by (Shahriari et al., 2016), SMAC has a potential drawback in that the estimated response surface is discontinuous, which makes the optimization of acquisition functions difficult.

TPE. It models the conditional distribution $p(\mathbf{x}|y)$ instead of $p(\mathbf{y}|\mathbf{x})$, and then the EI acquisition function can be parameterized by

$$\alpha_{\text{EI}}(\mathbf{x}) \propto \left(\frac{\ell(\mathbf{x})}{g(\mathbf{x})} \gamma + 1 - \gamma \right)^{-1}. \quad (2)$$

Note that the equation is slightly different from its original form in Bergstra et al. (2011), as we define it as a maximization problem. The functions $g(\mathbf{x})$ and $\ell(\mathbf{x})$ denote the density functions of $x|y$ as $y \geq y^*$ and $y < y^*$, respectively; they are estimated by hierarchical Parzen estimators. The value of y^* is chosen as a quantile γ of the observed y values, such that $p(y < y^*) = \gamma$ (Bergstra et al., 2011).

These Bayesian optimization methods have been implemented in various AutoML software packages. For example, the GP-EI method is implemented in the *Spearmint* package (Snoek et al., 2012); the SMAC method appears in *SMAC3*, *Auto-WEKA* (Kotthoff et al., 2017) and *Auto-sklearn* (Feurer et al., 2015); the TPE is wrapped into the *Hyperopt* package (Komer et al., 2014; Bergstra et al., 2015). There exist variants of Bayesian optimization in other problem settings, for instance, high-dimensional tasks (Wang et al., 2013; Kandasamy et al., 2015), collaborative hyperparameter tuning (Swersky et al., 2013; Bardenet et al., 2013; Feurer et al., 2015), no-regret Bayesian optimization (Berkenkamp et al., 2019), and parallelized Bayesian optimization (Snoek et al., 2012; Hutter et al., 2012). One may refer to Shahriari et al. (2016) for a comprehensive review.

There are other sequential HPO methods, e.g., evolutionary methods (Escalante et al., 2009; Di Francescomarino et al., 2018) and reinforcement learning (Lillicrap et al., 2015; Zoph and Le, 2016). These methods generally require extensive computing resources and are expensive in practice. One adaptive resource allocation idea is to allocate more computing resources to the hyperparameter configurations that tend to perform better. For instance, Successive Halving (Domhan et al., 2015) allocates the same budget to the best half hyperparameter configurations, while the worst half would be dropped. Such procedure is repeated until only one hyperparameter configuration is left. Hyperband (Li et al., 2017) can be viewed as a natural extension of Successive Halving, as it treats Successive Halving as a subroutine and calls it with different parameter settings.

Another sequential method related to our work can be referred to Huang et al. (2007). It used a two-stage nested uniform design for training SVM with two hyperparameters,

with 13 design points in the first stage (original search space) and 9 points in the second stage (adjusted search space). This can be viewed as a special two-dimensional case of our SeqUD approach, but the follow-up 9 points were not optimized with respect to the existing 13 points in the first stage. In this work, we propose a general framework of sequential uniform designs for HPO of various machine learning algorithms.

3. Augmented Uniform Design

The uniform design is a typical space-filling design (Fang et al., 2000), which aims to find a discrete set of design points to represent the search space as well as possible.

Definition 1 (Uniform Design) *In the unit hypercube $C^s \equiv [0, 1]^s$, a uniform design with n runs is the point set $D_n^* = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n\}$ that minimizes a certain discrepancy criterion:*

$$D_n^* \leftarrow \min_{D_n \subset C^s} \phi(D_n). \quad (3)$$

To limit the search space, the balanced U-type designs are often used for uniform design construction. Denote by $\mathbf{U}_{n,s} = (u_{ij})$ a U-type design with n runs, s factors, and q levels. Each factor with q levels is a permutation of the balanced arrangement

$$\underbrace{\{1, \dots, 1\}}_{n/q}, \dots, \underbrace{\{q, \dots, q\}}_{n/q}, \quad (4)$$

and it is required that n is divisible by q . One can convert it to the design matrix $\mathbf{X}_{n,s} = (x_{ij})$ in C^s by $x_{ij} = (2u_{ij} - 1) / 2q$. Such a U-type design that minimizes a certain discrepancy criterion is called a U-type uniform design, denoted as $U_n(q^s)$. More details about the uniform design theory and methods can be found in Fang and Wang (1990, 1994) and Fang et al. (2000, 2006, 2018).

For example, a U-type uniform design $U_{20}(20^2)$ with 20 runs, 2 factors, and 20 levels is shown in Table 1, which corresponds to the last plot in Figure 2. Below we provide the commonly used definitions for uniform designs, as well as their connections with HPO problems via Figure 1.

- **Factors:** the input variables of interest for a computer experiment, or the hyperparameters configured for a machine learning task.
- **Levels:** each factor or hyperparameter is assigned with discrete values within the experimental domain or the hyperparameter space. For a continuous factor, we typically divide it into q levels. For categorical or integer-valued factors, their level assignment will be discussed in Section 4.
- **Design points:** a design point is a possible level combination of different factors, and it corresponds to a unique configuration of hyperparameters to be conducted.
- **Runs or trials:** a run or trial corresponds to the implementation of a design point, and it generates the output or response for further analysis.

No.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
x_1	16	18	12	19	1	10	9	4	2	14	6	15	5	20	11	13	8	7	3	17
x_2	15	19	1	3	9	7	20	13	18	10	16	5	6	12	14	17	4	11	2	8

Table 1: U-type uniform design $U_{20}(20^2)$. Each integer of factors x_1, x_2 represents a level $k \in \{1, 2, \dots, q\}$, which corresponds to $(2k - 1)/2q$ in the design space.

Definition 2 (Augmented Uniform Designs) *Given an initial uniform design $D_{n_1} \subset C^s$ and the follow-up run size n_2 , an augmented uniform design is to find a follow-up design $D_{n_2}^*$ that minimizes the discrepancy of the combined design, i.e.,*

$$D_{n_2}^* \leftarrow \min_{D_{n_2} \subset C^s} \phi \left(\begin{bmatrix} D_{n_1} \\ D_{n_2} \end{bmatrix} \right). \quad (5)$$

Clearly, when the initial design is empty, the augmented uniform design reduces to the ordinary uniform design as in Definition 1. Next, we introduce the discrepancy for measuring the uniformity of the design points. Traditionally, the star discrepancy (Niederreiter, 1992) is the most popular choice in quasi-Monte Carlo methods. It is defined as the maximum deviation between the empirical distribution and uniform distribution,

$$\phi(D_n^*) = \sup_{x \in C^s} \left| \frac{|D_n \cap [0, x]|}{N} - \text{Vol}([0, x]) \right|, \quad (6)$$

where the symbol $|\cdot|$ denotes the number of points in a set and $\text{Vol}([0, x])$ is the uniform distribution function on the unit cube $[0, x]$. Later, the generalized ℓ_p -discrepancy (Hickernell, 1998) extends the star discrepancy to be more computationally tractable. Among various generalized ℓ_p -discrepancies, the centered ℓ_2 -discrepancy (CD_2) is given by

$$\begin{aligned} \text{CD}_2(D_n)^2 &= \left(\frac{13}{12} \right)^s - \frac{2}{n} \sum_{k=1}^n \prod_{j=1}^s \left(1 + \frac{1}{2} \left| x_{kj} - \frac{1}{2} \right| - \frac{1}{2} \left| x_{kj} - \frac{1}{2} \right|^2 \right) + \\ &\frac{1}{n^2} \sum_{k=1}^n \sum_{j=1}^n \prod_{i=1}^s \left[1 + \frac{1}{2} \left| x_{ki} - \frac{1}{2} \right| + \frac{1}{2} \left| x_{ji} - \frac{1}{2} \right| - \frac{1}{2} |x_{ki} - x_{ji}| \right]. \end{aligned} \quad (7)$$

The CD_2 discrepancy can be intuitively interpreted as the relative proportion of design points belonging to subregions of the search space (Hickernell, 1998). It has several appealing properties: a) easiness to compute; b) projection uniformity over all sub-dimensions; c) reflection invariance around the plane $x_j = \frac{1}{2}$ (for any $j = 1, \dots, s$).

There also exist other uniformity criteria, including the wrap-around ℓ_2 -discrepancy (WD_2) and the mixture ℓ_2 -discrepancy (MD_2). They share similar properties as CD_2 . For simplicity, we use CD_2 as the default criterion for generating and evaluating uniform designs and augmented uniform designs throughout the paper.

3.1 Construction Algorithm

Due to the balance requirement for U-type designs, we assume the total number of runs $n = n_1 + n_2$ is divisible by q . The construction of U-type augmented uniform designs is a combinatorial optimization problem, which is extremely difficult for large design tables. The enhanced stochastic evolutionary (ESE) algorithm proposed by Jin et al. (2005) is the most influential work for constructing space-filling designs. It is built upon threshold accepting (TA) and stochastic evolutionary (SE), for which a rather sophisticated procedure is designed to automatically control the acceptance threshold. This algorithm has been implemented in the R package *DiceDesign* (Dupuy et al., 2015).

Inspired by ESE, we provide a simple yet effective AugUD algorithm for constructing augmented uniform designs. The proposed AugUD algorithm is composed of two nested loops. The inner loop rolls over columns for element-wise exchange while the outer loop adaptively changes the acceptance threshold. It involves the following critical steps.

Initialization. Given a fixed D_{n_1} , initialize the augmented design $D_{n_2}^*$ such that $\left(\begin{bmatrix} D_{n_1} \\ D_{n_2}^* \end{bmatrix}\right)$ is a balanced U-type design. Note that $D_{n_2}^*$ can be randomly generated or user-specified as long as it satisfies the balance requirement.

Element-wise Exchange. This is a basic procedure for searching optimal designs (Fang et al., 2000; Jin et al., 2005). Given the current $D_{n_2}^*$, we randomly exchange two elements of a factor in $D_{n_2}^*$, in order to obtain a new combined design with improved uniformity. Repeat this operation for M_E times. According to the CD2 criterion (7), only a small part of terms needs to be updated for each element-wise exchange, which is an appealing property in practice as it can help save a lot of computing time.

Threshold Accepting. The TA strategy is employed for jumping out of local optima. The best candidate design D_{n_2} obtained by element-wise exchange can be accepted with a probability:

$$p = 1 - \min\left(1, \max\left(0, \frac{\Delta}{T_h}\right)\right), \quad (8)$$

where T_h is the acceptance threshold for accepting suboptimal solutions and Δ is the change of the uniformity criterion:

$$\Delta = \phi\left(\begin{bmatrix} D_{n_1} \\ D_{n_2} \end{bmatrix}\right) - \phi\left(\begin{bmatrix} D_{n_1} \\ D_{n_2}^* \end{bmatrix}\right). \quad (9)$$

When $\Delta < 0$, D_{n_2} will be accepted with the 100% probability; otherwise, D_{n_2} with worse performance may still be accepted with a probability p .

Adaptive Threshold. It is critical to select the threshold T_h in the TA algorithm. We propose an adaptive updating rule for T_h . It is firstly initialized as

$$T_h = \gamma\phi\left(\begin{bmatrix} D_{n_1} \\ D_{n_2}^* \end{bmatrix}\right), \quad (10)$$

where γ is a factor controlling the initial threshold. During optimization, T_h can be adaptively updated by

$$T_h = \begin{cases} T_h/\alpha & \text{if } h_i < \eta, \\ \alpha T_h & \text{otherwise,} \end{cases} \quad (11)$$

where α is the scaling factor for adjusting the threshold. The symbol h_i represents the hit ratio in the i th iteration (outer loop). When h_i is smaller than η , T_h will be increased for more exploration; when h_i remains a large value, the threshold should be decreased for better exploitation.

A formal description of the AugUD construction is given in Algorithm 1, which requires input of multiple pre-specified parameters. In practice, we mimic the settings used in Jin et al. (2005) and specify the parameters as follows.

- $\gamma = 0.005$. It denotes the multiplier of the initial acceptance threshold.
- $\eta = 0.1$. The hit ratio threshold controls the acceptance threshold adjustment.
- $\alpha = 0.8$. It is the scaling factor for adjusting the acceptance threshold.
- $M_E = \min\{50, 0.2 \times n_2^2(q-1)/(2q)\}$. It controls the number of element-wise exchange.

Finally, the numbers of loops are empirically determined as $M_{\text{outer}} = 50$ and $M_{\text{inner}} = 100$, which appear to work well in our tested cases. The AugUD algorithm can be further enhanced by restarting multiple times with different random seeds, and the one with the best uniformity criterion can be selected.

Software Implementation. The proposed AugUD algorithm and related functionalities have been wrapped and implemented in our open-source Python package *PyUniDOE*¹. The core algorithm is written by C++ programming language, and we provide user-friendly application programming interfaces (APIs) in Python. It supports the real-time generation of uniform designs and augmented uniform designs under various uniformity criteria, e.g., CD_2 , WD_2 , and MD_2 . In *PyUniDOE*, we also include a database of many state-of-the-art U-type uniform designs, which can be directly queried.

3.2 Experimental Evaluation

The proposed AugUD algorithm serves as an efficient tool for the design community regarding the generation of both uniform designs and augmented uniform designs.

3.2.1 AUGUD FOR GENERATING UNIFORM DESIGNS

We download the $U_n(n^s)$ designs from the uniform design website² maintained by Hong Kong Baptist University, which collects the commonly used uniform designs with factors $s = 2, 3, \dots, 29$. Each factor has different number of runs $n = s + 1, s + 2, \dots, 30$. In total, 406 U-type uniform designs are downloaded. Although these designs are widely used, we experimentally show that most of them are not optimal.

We treat the downloaded U-type uniform designs as initializations and then compare the proposed AugUD algorithm (implemented in our *PyUniDOE* package) with the ESE algorithm (implemented in the *DiceDesign* package by Dupuy et al., 2015). For a fair comparison, these two methods are configured with the same number of iteration loops and element-wise exchange. Each method is repeated ten times with different random seeds. For

1. <https://github.com/SelfExplainML/PyUniDOE>

2. <http://www.math.hkbu.edu.hk/UniformDesign>

Algorithm 1: The proposed AugUD algorithm

Input: D_{n_1} (Existing Design), n_1 (# Existing Runs), n_2 (# Augmented Runs), s (# Factors), q (# Levels), M_{outer} , M_{inner} (# Outer and Inner Loops).
Output: The optimal augmented design $D_{n_2}^*$.

- 1 Initialize $D_{n_2}^*$ and calculate T_h by (10).
- 2 **for** $i = 1, 2, \dots, M_{\text{outer}}$ **do**
- 3 Set $k = 0$.
- 4 **for** $j = 1, 2, \dots, M_{\text{inner}}$ **do**
- 5 Select the column $j \pmod s$ of $D_{n_2}^*$.
- 6 Randomly pick M_E element pairs on the selected column.
- 7 Perform element-wise exchange for selected pairs and evaluate discrepancy.
- 8 Choose the best candidate design and calculate p by (9).
- 9 Update $D_{n_2}^*$ and set $k = k + 1$ (with the probability p).
- 10 **end**
- 11 Calculate the hit ratio $h_i = k/M_{\text{inner}}$.
- 12 Adaptively update T_h by (11).
- 13 **end**

evaluation, we report the average and best performance of each method. The average CD_2 improvement ratios (over ten repetitions) of different initial uniform designs are reported in Figure 15 in the appendix. For simplification, we aggregate the improvement ratios by the number of factors, and the averaged results against different numbers of factors are reported in Figure 3. For example, the bars at $x = 2$ are averaged over all the 2-factor designs with runs ranging from 3 to 30, while the bars at $x = 29$ only represent the results of $U_{30}(30^{29})$.

It is surprising to observe that a large proportion of classic uniform design tables can be improved. According to the best designs found in the ten repetitions, 317 out of the tested 406 uniform designs have been improved by AugUD compared to 313 by ESE. This means that almost 80% of existing uniform designs maintained by the uniform design website are not optimal, which further reflects the difficulty of generating uniform designs. Meanwhile, the best results of AugUD (over the ten repetitions) improve CD_2 of existing designs by 0.2433% on average, which is much better than that of ESE (0.1875%). It is also observed that the average improvement for 2-factor designs is relatively small compared to other high-dimensional designs. The reason is that these small-scale designs are easy to construct, and most of them are already optimal.

In general, the proposed AugUD algorithm tends to outperform the ESE algorithm. Due to the efficient computational enhancement (e.g., C++ acceleration) in *PyUniDOE*, AugUD only uses around 0.34% computing time of ESE (which is based on pure R language). The results demonstrate the superiority of the proposed AugUD algorithm and its efficient implementation in *PyUniDOE*. Moreover, it is worth mentioning that all the existing designs are already widely used, and such improvement marks a non-trivial contribution to the uniform design community.

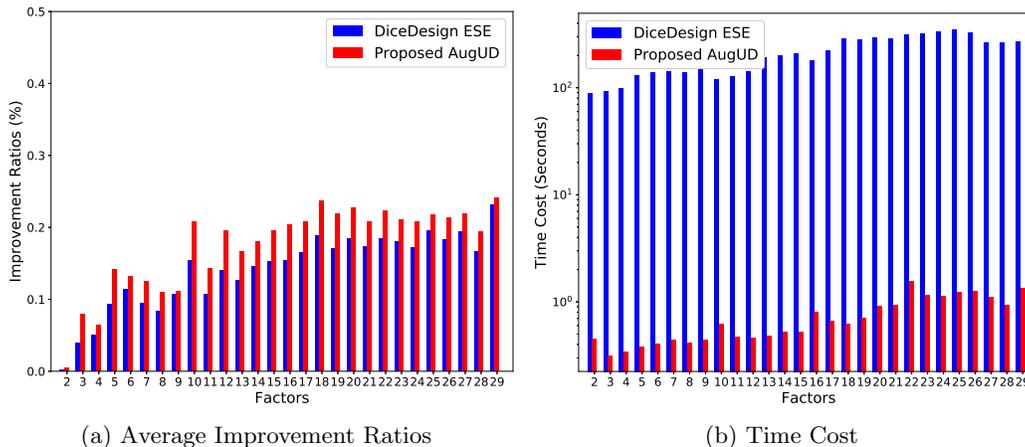


Figure 3: Experimental results for generating uniform designs averaged against different number of factors. (a) Average improvement ratios compared with uniform designs obtained from the uniform design website; (b) Time cost.

3.2.2 AUGUD FOR GENERATING AUGMENTED UNIFORM DESIGNS

In the literature, a commonly used strategy for generating sequential designs is the nested uniform design (Fang and Wang, 1994), in which a new uniform design is embedded into existing designs. However, nested uniform designs do not consider the relationship between new and existing design points. The AugUD algorithm provides a practical and fast solution for augmenting design points subject to overall uniformity. Thus, AugUD can avoid clustered or duplicated design points and better explore the search space.

Experimentally, we consider a test scenario in which 5 design points are randomly obtained from $U_n(n)^s$ for each $s = 2, \dots, 24$, and $n = 8, \dots, 30$, ($n > s + 5$). These points are treated as existing designs, and we then augment $(n - 5)$ design points to the design space. Five strategies are involved for augmentation, i.e., random augmentation, nested LHS (i.e., a new LHS is embedded into the design space), nested Sobol (i.e., a new Sobol sequence is embedded into the design space), nested uniform design (i.e., a new uniform design $U_{n-5}(n - 5)^s$ is embedded into the design space; nested UD), and AugUD. Each method is repeated ten times, and we calculate the average CD_2 over ten repetitions for comparison.

The detailed improvement ratios of AugUD and nested UD against random augmentation for each (factor, run) pair are reported in Figure 16 in the appendix. A summary of the experimental results is provided in Figure 4. The bars at $x = 2$ denote the improvement ratios averaged over all the 2-factor designs with runs ranging from 8 to 30. In contrast, the bars at $x = 24$ only represent the average improvement ratio of the 30-run, 24-factor design. Both AugUD and nested UD show superior performance to random augmentation, nested LHS, and nested Sobol regarding the overall uniformity.

It is also observed that AugUD significantly outperforms nested UD in all the compared cases. For AugUD, the improvement ratios show a decreasing trend as the number of factors

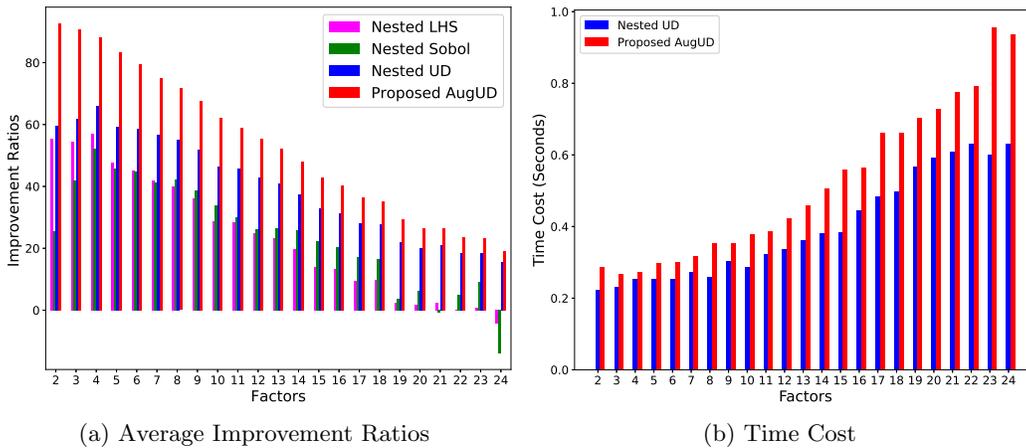


Figure 4: Average results for generating design augmentation. (a) Improvement ratios of nested UD and AugUD against random augmentation; (b) Time cost.

increases. That is, large-sized designs are generally hard to optimize, and the augmented designs found by AugUD may be still not optimal. In Figure 4b, the computing time of AugUD is slightly larger than that of nested UD (generated using our *PyUniDOE* package), as the evaluation of (7) is a little bit expensive for large-sized designs, i.e., n versus $(n - 5)$. Note that nested LHS and nested Sobol can be easily generated without any complicated optimization, and hence their time costs are not reported (almost zero).

For illustration, Figure 5 shows the results of adding 15 design points to 5 existing design points in a 2-D design space. It can be observed that the added points by random augmentation and nested UD can be quite close to existing points. In contrast, the proposed AugUD performs significantly better in terms of space-filling performance.

4. Sequential Uniform Design

Sequential uniform design (SeqUD) is a general multi-stage optimization framework that incorporates uniform designs and a simple yet effective search strategy, i.e., sequential space halving. We first present the general SeqUD framework, then discuss its application for HPO, and finally remark its benefits and limitations.

4.1 SeqUD Framework

The SeqUD framework comprises the following components: A) initial uniform design; B) subspace zooming & level doubling for sequentially adjusting the search space; and C) uniform design augmentation.

A) Initial Uniform Design. At the first stage, a U-type uniform design $U_n(q^s)$ is generated. These design points are evaluated through the corresponding experiments. Note that the number of initial design points n and the level number q should be predetermined for a specific task. For problems with possibly complex response surface, it is recommended

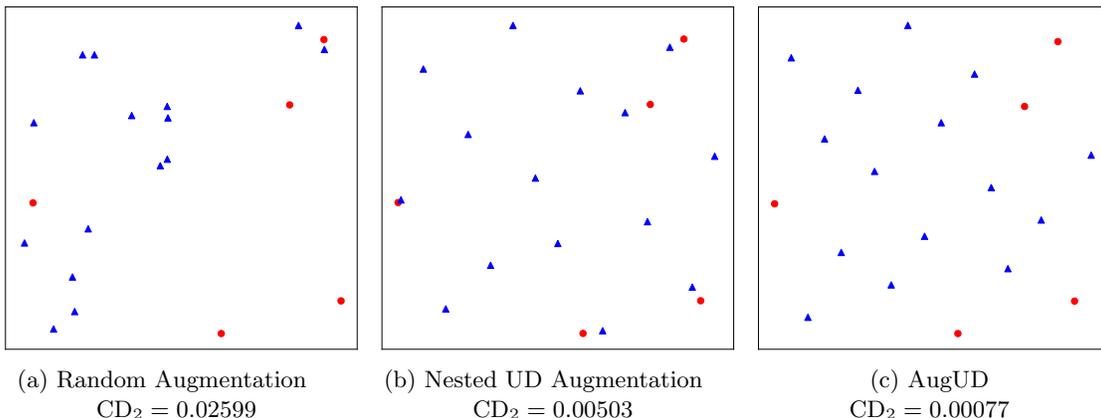


Figure 5: A 2D example illustrating different augmentation methods. (a) Random augmentation; (b) Nested UD; (c) AugUD. In each plot, the 5 red dots represent the existing design D_{n_1} and the 15 blue dots represent the augmented design D_{n_2} .

to use more runs and more levels so that the initial design can better cover the search space. After initialization, steps B) and C) are repeated until the maximal number of runs T_{\max} is reached.

B) Subspace Zooming & Level Doubling. At the j th stage ($j \geq 2$), the search space is halved into a subspace while the doubled granularity. The optimal point \mathbf{x}_j^* among all the evaluated design points is treated as the center of the new subspace, and the new search space is defined with levels (for each factor $i = 1, 2, \dots, s$)

$$U_{j,i} = \left\{ x_{j,i}^* - \frac{q-1}{2^j q}, \dots, x_{j,i}^*, \dots, x_{j,i}^* + \frac{q-1}{2^j q} \right\}, \quad (12)$$

when q is odd, or

$$U_{j,i} = \left\{ x_{j,i}^* - \frac{q-2}{2^j q}, \dots, x_{j,i}^*, \dots, x_{j,i}^* + \frac{1}{2^j} \right\}, \quad (13)$$

when q is even. It is possible that the selected optimal center point is close to the search boundary, and some parts of the reduced subspace can be outside of C^s . Accordingly, we introduce a subspace shifting procedure to prevent this from happening. The reduced subspace is moved perpendicularly towards the inner side of the search space until all the levels are within the boundary.

C) Uniform Design Augmentation. This step is to augment design points in the reduced subspace. After the existing design points are converted to the new level space (12) or (13), the new design points can be augmented via the AugUD algorithm.

A summary of the above procedures is provided in Algorithm 2. As the termination of SeqUD is controlled by T_{\max} , the number of design points per stage should be accordingly specified. In the beginning, the granularity is generally not sufficient for finding the optimal point, e.g., for stages $j \leq 3$; as the search space is halved sequentially, the search space will

Algorithm 2: The proposed SeqUD framework

Input: T_{\max} (# Total Runs), n (# Runs per Stage), s (# Factors), q (# Levels).

Output: The optimal design point \mathbf{x}^* from all evaluations.

- 1 Generate an initial uniform design $U_n(q^s)$.
 - 2 Evaluate each initial design point.
 - 3 Collect the design-response pairs $H = \{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n)\}$.
 - 4 Set stage $j = 2$ and $T = n$.
 - 5 **while** *True* **do**
 - 6 Reduce the space of interest by centering on $x_j^* = \arg \max_{\mathbf{x} \in H} y$.
 - 7 Count the number of existing design points in the subspace as n_e .
 - 8 Calculate the number of new design points to be augmented as $n_j = n - n_e$.
 - 9 **if** $T + n_j > T_{\max}$ **then** break;
 - 10 Augment n_j design points via the AugUD algorithm.
 - 11 Evaluate each augmented design point, and update H .
 - 12 Set $j = j + 1$ and $T = T + n_j$.
 - 13 **end**
-

be sufficiently small as, e.g., for stages $j \geq 10$. Further exploration over such a small region is somehow meaningless. Hence, the number of design points per stage should be roughly selected within the range $[T_{\max}/10, T_{\max}/3]$, considering the complexity of the task. For illustration, a two-stage example of SeqUD is provided in Figure 6. It can be observed that the search space is well covered by the initial uniform design. The reduced subspace is centered on the best-evaluated point and more design points can be added accordingly.

Remark 3 *A sequential random search (SeqRand) approach is proposed as a naïve version of SeqUD. It is based on the space halving strategy while design points at each stage are randomly generated. Since uniformity is not a concern of SeqRand, the number of randomly generated design points can be directly set to n for each stage. Moreover, SeqRand can also be viewed as a sequential version of the random search.*

4.2 SeqUD for Hyperparameter Optimization

In this part, we present the details of applying SeqUD for HPO. Let f be a function that measures the performance of a machine learning algorithm with hyperparameters $\theta = (\theta_1, \theta_2, \dots, \theta_d)$, where d denotes the number of hyperparameters to be optimized. Our objective is to find the best hyperparameter configuration θ^* that maximizes f , which can be the cross-validation or hold-out validation score.

Hyperparameters are usually of different types and scales. In HPO, the first step is to identify the tunable hyperparameters and corresponding search domains. Hyperparameters of machine learning algorithms may have various ranges and formats, and it is necessary to do some preprocessing. In general, they can be classified into three types, i.e., continuous (or numerical), integer-valued, and categorical. Continuous and integer-valued hyperparameters can be linearly transformed within the range $[0, 1]$. For categorical hyperparameters, one-hot encoding should be employed for transformation.

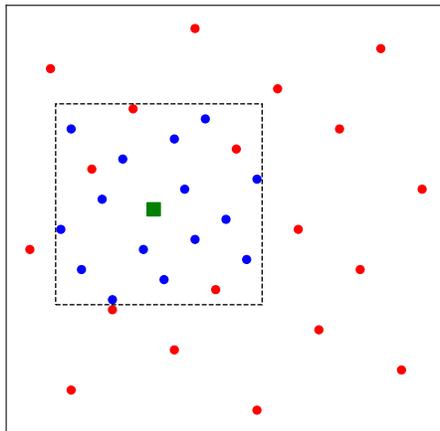


Figure 6: A two-stage example of SeqUD in a 2-D space. The circle points represent the initial uniform design via $U_{20}(20^2)$. The surrounding box serves as the subspace of interest centered on the optimal design point \mathbf{x}_1^* at the first stage, which is denoted by a square point in green. At the second stage, a new design is augmented (blue points) considering the overall uniformity within the subspace.

SeqUD proceeds HPO as follows. First, as the initial uniform design is constructed, we inversely transform these design points to their original forms. For continuous hyperparameters, the inverse mapping can be directly implemented, while for integer-valued hyperparameters, they should be rounded to the nearby integers. Each categorical hyperparameter is represented by multiple dummy variables so that the corresponding design space dimension is greater than the number of hyperparameters. The encoded dummy variables are inversely transformed to corresponding hyperparameters by the arg max operation. More detailed discussion for handling categorical and integer-valued hyperparameters can be referred to Garrido-Merchán and Hernández-Lobato (2020).

All the generated hyperparameter configurations are then evaluated by training the machine learning algorithm and calculating the predefined evaluation metric. The best performing configuration is selected for further investigation at the next stage. Through subspace zooming & level doubling, new design space will be generated, and the AugUD algorithm can be used to augment sequential design points. After the optimization terminates, the machine learning model will be configured with the optimal hyperparameters and refitted to the whole training data.

Software Implementation. The procedures mentioned above are wrapped in our Python package *SeqUD*³. It includes the proposed SeqUD method and some related benchmark methods with an interface to the well-known machine learning platform *scikit-learn*. In addition to SeqUD and SeqRand, the APIs for some non-sequential methods are also provided by *SeqUD*, including grid search, random search, uniform designs, Latin hypercube sampling (by Python package *pyDOE*), and Sobol sequences (by Python package *sobol_seq*).

3. <https://github.com/SelfExplainML/SeqUD>

Moreover, the three classic Bayesian optimization methods are included by using the interfaces of *Hyperopt*, *Spearmint*, and *SMAC3*.

4.3 Discussion

The idea behind the proposed SeqUD framework is intuitive and straightforward. It uses the space halving strategy to adjust the search space, and the main difference between SeqUD and other space halving-based approaches lies in its uniformity consideration. Compared to coarse-to-fine grid search methods, a) SeqUD is not limited to low-dimensional problems; b) the uniformity of new design points with existing design points is considered. Thus, it should have better optimization performance. We summarize its beneficial aspects as follows.

- SeqUD shares the benefits of all the other sequential methods. Except for the initial design, design points in SeqUD are sequentially constructed based on the preliminary information of existing design points. This procedure is more flexible and efficient than non-sequential methods, e.g., grid search and random search.
- SeqUD makes a good balance between exploration (by uniform designs) and exploitation (by sequential space halving). SeqUD is less likely to be trapped into local areas for complicated hyperparameter response surfaces as the design points are uniformly located in the area of interest.
- SeqUD is free from the surrogate modeling and acquisition optimization used in Bayesian optimization. These procedures are all difficult tasks. For example, the GP model may fail when design points are close to each other; building a random forest on the hyperparameter space may be more expensive than conducting the experiments; in high-dimensional settings, it may find the best design point using the fitted surrogate model is also time-consuming. In contrast, new design points in SeqUD can be quickly generated without too much computation.
- Design points generated at the same stage can be evaluated in parallel. Given sufficient computing resources, this property will bring significant computation efficiency, especially for training large-scale machine learning algorithms. Methods like GP-EI, SMAC, and TPE, are initially designed to select new design points one by one, leading to a waste of computing resources. There also exist some strategies for speeding up computations for Bayesian optimization methods (Snoek et al., 2012; Hutter et al., 2012), while the optimization performance may be sacrificed. Also, these methods are not natural for performing parallelization (Shahriari et al., 2016).

The possible limitation of SeqUD, SeqRand, and all the other space halving strategies lies in the local optima problem. Although sampling uniformity is considered in SeqUD, it can still be trapped into local optima. This problem can be mitigated by performing more exploration over the search space, and in practice, the following two ways can be used to enhance the exploration.

- Employ more design points per stage for complex tasks, so that the algorithm is less likely to be trapped into locally optimal areas.

- Multiple shooting, i.e., except for zooming into the best-evaluated point per stage, we may simultaneously search the nearby subspace of the second and the third-best points (when these points are distant from each other).

Given a sufficient number of runs, these two strategies may help increase the success rate of optimization. However, as the total budget is usually limited, the trade-off between exploration and exploitation still exists.

5. Experiments for Global Optimization

Extensive synthetic functions are involved in testing the performance of SeqUD on global optimization tasks. The benchmark models include grid search (Grid), random search (Rand), Latin hypercube sampling (LHS), Sobol Sequences (Sobol), uniform designs (UD), sequential random search (SeqRand), GP-EI, SMAC, and TPE. A total budget of 100 runs is allowed for each method. That is, all the compared methods can evaluate at most 100 design points. Finally, grid search is only tested on 2-D tasks.

All the benchmark methods are kept to their default settings. In SeqUD, we set the number of runs and levels per stage as $n = q = 15$ when $s \leq 5$; otherwise, we use $n = q = 25$ for higher-dimensional tasks. This setting compromises exploration and exploitation and works well in our experiments. For a fair comparison, the SeqRand approach is also configured with 15 or 25 runs per stage (depending on s). All the global optimization experiments are repeated 100 times. Throughout this paper, the statistical significance is reported based on paired t-test, with a significance level of 0.05. Due to the page limit, results reported in tables are all rounded to a certain precision, while the rank and statistical significance comparisons among different methods are based on the original precision.

5.1 Example Functions

The working mechanism of each compared method is first investigated through the following 2-D synthetic functions.

Cliff Function. The first example is obtained from Haario et al. (1999, 2001). As shown in Figure 7a, its mesh plot looks like a “cliff”, where the non-zero-value area is narrow and long.

$$f_1(x_1, x_2) = \exp \left\{ -\frac{1}{2} \frac{x_1^2}{100} - \frac{1}{2} (x_2 + 0.03x_1^2 - 3)^2 \right\}, \quad (14)$$

$$x_1 \in [-20, 20], x_2 \in [-10, 5].$$

Octopus Function. The second scenario is much more complicated, with multiple local extrema within the response surface (Renka and Brown, 1999). Accordingly, we name it “octopus” due to its shape as shown in Figure 8a.

$$f_2(x_1, x_2) = 2 \cos(10x_1) \sin(10x_2) + \sin(10x_1x_2), \quad (15)$$

$$x_1, x_2 \in [0, 1].$$

The goal is to find the maximal values using the compared methods. For demonstration, the evaluated design points of SeqUD are visualized in Figure 7b and Figure 8b; the

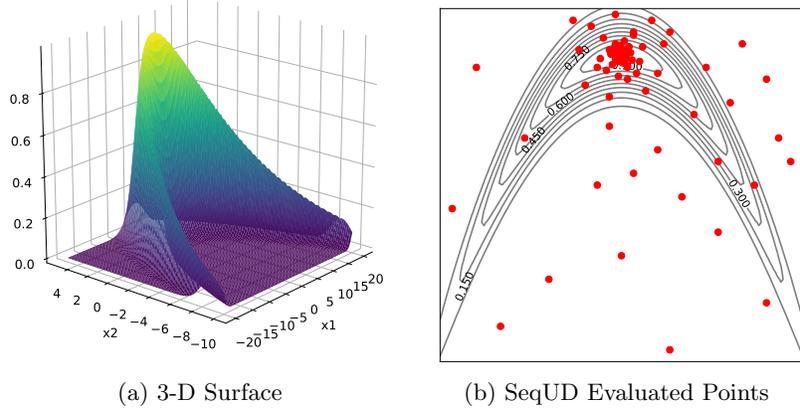


Figure 7: The 3-D surface of the cliff function and SeqUD evaluated points against the ground truth contour plot.

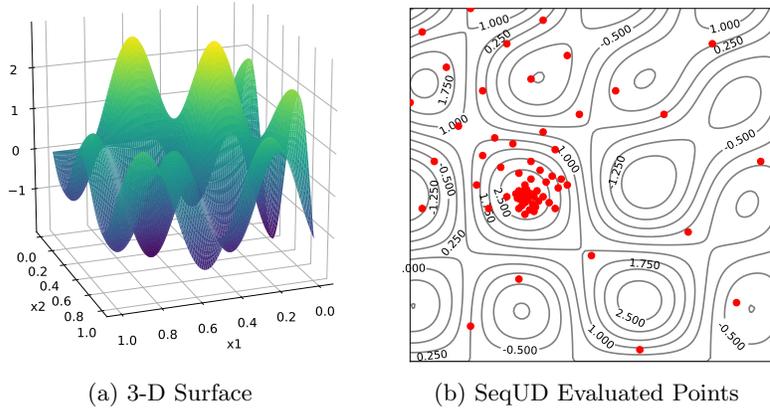


Figure 8: The 3-D surface of the octopus function and SeqUD evaluated points against the ground truth contour plot.

corresponding plots for benchmark methods are placed in Figure 17 and Figure 18 in the appendix. Some impressive results are observed. All the non-sequential methods together with SMAC and TPE use many design points in less promising areas in the cliff function. The SeqRand shares the benefits of the space halving strategy and can finally find the global optimal region. However, as randomly generated samples are not representative, SeqRand is less efficient as compared to SeqUD. Similar results could be found in the octopus function. For example, SeqRand, TPE, and GP-EI are trapped in a sub-optimal area; the best location found by SMAC is just close to the global optimum. In contrast, the proposed SeqUD approach tends to be more promising as it successfully finds the correct area and achieves the best performance.

Data Set	Grid	Sobol	UD	Rand	LHS	SeqRand	TPE	SMAC	GP-EI	SeqUD
cliff	0.869	0.877	0.983	<u>0.907±0.082</u>	0.931±0.063	0.961±0.098	0.973±0.026	0.913±0.102	0.994±0.036	1.000
octopus	2.889	2.778	2.849	2.784±0.136	2.805±0.132	2.904±0.157	2.858±0.113	2.857±0.163	2.898±0.198	2.996

Table 2: The optimization results of cliff and octopus functions. The best performing methods are highlighted in bold and underlined results are significantly different from the best ones. Note that the standard deviations of Grid, Sobol, UD and SeqUD are omitted as they are all zero.

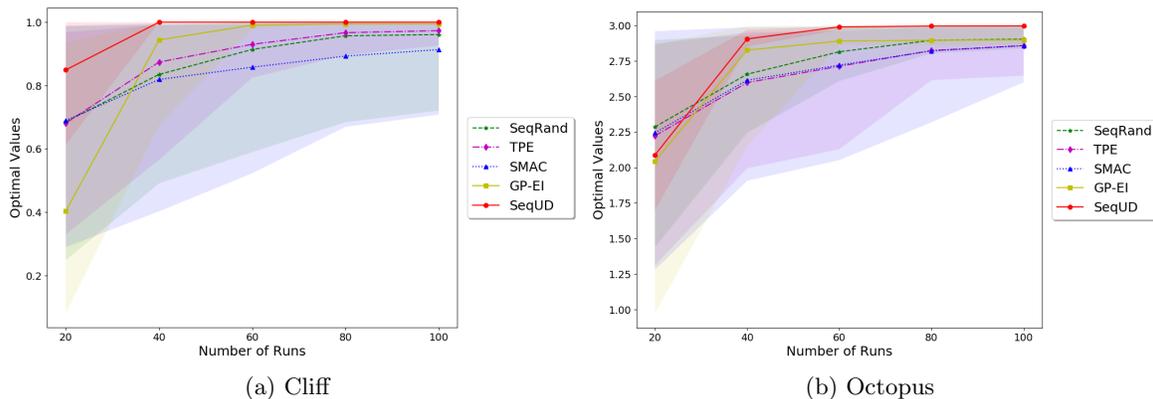


Figure 9: The optimization results against the number of runs of the two example functions. Each point is averaged over 100 repetitions, and the areas between the 5th and 95th percentiles are shadowed.

Table 2 reports the final optimization results of all compared methods, where the best performing methods are highlighted in bold and underlined results are significantly different from the best ones. The performance of sequential methods over the number of runs is visualized in Figure 9. In the cliff function, SeqUD uses fewer trials to reach the best point. Given 100 runs, SeqUD performs slightly better than GP-EI, but no statistical significance is observed. In particular, SMAC fails in this task, and its performance is close to random search. For the octopus function, SeqUD achieves significantly better performance as compared to all the benchmarks. SeqRand and GP-EI show similar performance at 100 runs, and both of them outperform SMAC and TPE in this task.

5.2 Systematic Investigation

A systematic investigation is conducted on extensive standard unconstrained global optimization tasks summarized by Surjanovic and Bingham (2020). Out of the 47 synthetic functions, 32 are selected using the following rules: a) one-dimensional optimization problems are removed (too simple to use for comparison); b) tasks with global optimal points exactly located at the center of the search spaces are removed due to fairness consideration, as some of the compared methods routinely search the center. Table 3 provides more infor-

mation for these selected synthetic functions, which are grouped into six categories with 2 to 8 dimensions. Note that some of the functions are allowed to have various dimensions. We follow their recommended settings for powersum, zakharov, dixonpr, rosen, powell, and michal. For the remaining synthetic functions, a somehow arbitrary setting is used, that is, langer, levy, perm0db, trid, prmdb are set to have two dimensions; schwef and stybtang are set to have six dimensions.

Similar to Table 2, the averaged optimization results over 100 repetitions are reported in Table 9 in the appendix, in which all the listed results should be multiplied by the corresponding scaling factors in the last column. To have an exact comparison of different methods, we evaluate their relative performance using rank statistics. For each synthetic function, the averaged results of all compared methods are ranked from 1 (the best) to 10 (the worst). A box plot for the rank statistics across the 32 synthetic functions is reported in Figure 10a. The five sequential methods are also ranked (from 1 to 5) for every ten runs, and the ranks averaged over the 32 synthetic functions are reported in Figure 10b.

From the results, we can observe that sequential methods are superior to non-sequential methods, and among the compared sequential methods, the proposed SeqUD achieves the best overall performance. It ranks the first in more than half of the 32 synthetic functions and the second in 7 functions. The overall performance of GP-EI is inferior to SeqUD but better than other compared methods. Although GP-EI performs the best on 9 synthetic functions, it also ranks the last in 5 functions. The rest sequential methods, i.e., SeqRand, SMAC, and TPE, are slightly poorer than SeqUD and GP-EI.

Note that two failure cases are observed in SeqUD, i.e., for langer and holder. Both of these two functions have a lot of locally optimal regions. It is worth mentioning that SeqRand does not fail in these two functions. That is because design points in SeqUD are constructed with uniformity consideration, such that its design points are relatively stable across different repetitions. In contrast, SeqRand may have very different design points for different random seeds. Therefore, it is possible that SeqRand outperforms SeqUD.

5.3 Ablation Study for 1000 Runs

An ablation study is further conducted with 1000 runs for each compared method. For SeqUD, we increase the number of runs and levels (per stage) proportionally, i.e., $n = q = 150$ for $s \leq 5$ and $n = q = 250$ for $s > 5$. The SeqRand approach is accordingly configured. Given more runs, the design points of GP-EI tend to cluster. Such clustered design points may make GP-EI extremely slow and even crash due to the singular matrix problem. Therefore, we remove GP-EI from the benchmark list, and all the other models are still employed here for comparison.

Table 10 in the appendix presents the experimental results of this ablation study, with the same formatting style as in Table 9. We calculate the rank statistics for each synthetic function, and the corresponding box plots are shown in Figure 11. The ranks based on 1000 runs are consistent with that of 100 runs. SeqUD still performs the best in most synthetic functions, followed by SeqRand, TPE, SMAC, and all the non-sequential methods.

In the early stage of the optimization, SeqUD and SeqRand have poorer performance than TPE and SMAC. For instance, both TPE and SMAC are better than SeqUD at 100, 200, and 300 runs. That is because we use larger n in this ablation study. The first 300

Category	Function Name	Abbr	Dim	Category	Function Name	Abbr	Dim
Many Local Minima	Bukin N. 6	bukin6	2	Bowl-Shaped	Perm 0, d, β	perm0db	2
	Cross-in-Tray	crossit	2		Trid	trid	2
	Eggholder	egg	2	Steep Ridges or Drops	De Jong N. 5	dejong5	2
	Holder Table	holder	2		Easom	easom	2
	Langermann	langer	2		Michalewicz	michal	5
	Levy	levy	2	Other	Beale	beale	2
	Levy N. 13	levy13	2		Branin	branin	2
	Schwefel	schwef	6		Colville	colville	4
	Shubert	shubert	2		Goldstein-Price	goldpr	2
	Plate-Shaped	Booth	booth		2	Hartmann 3-D	hart3
McCormick		mccorm	2		Hartmann 4-D	hart4	4
Power Sum		powersum	4		Hartmann 6-D	hart6	6
Zakharov		zakharov	4		Perm d, β	permdb	2
Valley-Shaped	Six-Hump Camel	camel6	2		Powell	powell	4
	Dixon-Price	dixonpr	4		Shekel	shekel	4
	Rosenbrock	rosen	8	Styblinski-Tang	stybtang	6	

Table 3: The 32 synthetic functions for global minimization.

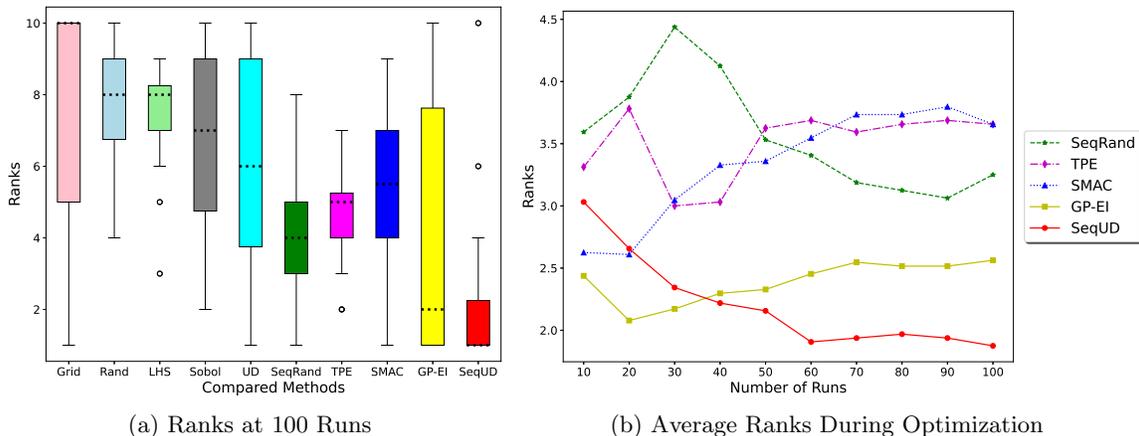


Figure 10: The ranks of compared methods over the 32 synthetic functions.

runs correspond to the early stages of SeqUD, which means that the search points are of low granularity. In contrast, TPE and SMAC keep their original settings and perform better for the first 300 runs. But as more design points are evaluated, SeqUD gradually exceeds these two Bayesian optimization methods. Consistent patterns are observed for SeqRand, while it needs more design points to exceed TPE and SMAC.

Given a larger number of design points per stage, SeqUD tends to achieve better performance in 29 out of 32 synthetic functions. The results of SeqUD on langer and holder are largely improved, and SeqUD even outperforms all the compared benchmarks at 1000 runs. However, SeqUD also gets poorer results on three synthetic functions. Although using a larger number of design points per stage can have a better exploration of the search space, it is still possible that the initial design points are scattered within locally optimal regions.

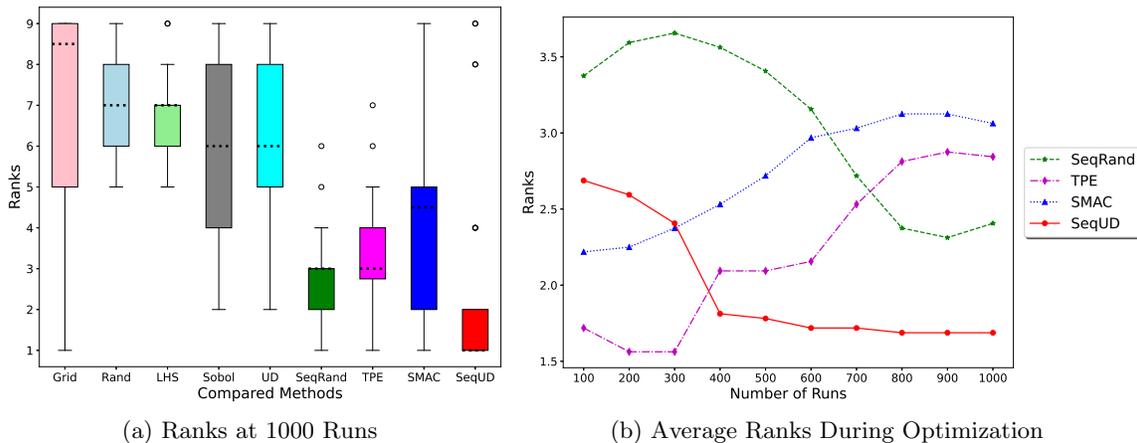


Figure 11: The ranks of compared methods over the 32 synthetic functions (with 1000 runs).

6. Experiments for Hyperparameter Optimization

We continue to test the proposed SeqUD method for HPO tasks in the AutoML context.

6.1 Experimental Setup

We consider 20 regression and 20 binary classification data sets obtained from the UCI machine learning repository and OpenML platform, in which we select the data with moderate features and sample sizes. Each data is preprocessed by imputing missing values by the median (for continuous variables) or most frequent (for categorical variables) values, as summarized in Table 4. Before training, categorical features are preprocessed using one-hot encoding, and numerical features are scaled within $[0, 1]$. For each data, we split 50% of the data samples for training, and the remaining 50% is used for testing. Five-fold cross-validation (CV) performance in the training set is employed as the optimization target. The root-mean-square error (RMSE) and accuracy score are used as the evaluation criteria for regression and classification tasks, respectively.

Two representative machine learning algorithms are first involved, i.e., support vector machine (SVM) and extreme gradient boosting (XGBoost). A pipeline optimization task is also considered, which involves data preprocessing, feature engineering, model selection, and HPO. Some of the compared methods are implemented to utilize dependence information among hyperparameters, while the rest are not. To eliminate this influence, the dependence information of hyperparameters is not utilized throughout. All the experiments are conducted based on the *scikit-learn* platform and related packages (e.g., *xgboost*).

SVM. We consider a classical 2-D hyperparameters optimization problem in SVM. The popular Gaussian kernel is fixed, and we tune the two continuous hyperparameters, i.e., the kernel parameter and penalty parameter. They are optimized in the base-2 log scale within $[2^{-16}, 2^6]$ and $[2^{-6}, 2^{16}]$, respectively. The training algorithm of SVM is not scalable even for data with moderate sample sizes. To save computing time, we instead use *scikit-learn*’s “SGDRegressor” or “SGDClassifier” (with hinge loss for regression and epsilon insensitive

Regression				Classification			
Abbr.	Data Set	Feature	Size	Abbr.	Data Set	Feature	Size
R1	no2	7	500	C1	breast cancer wise diag	30	569
R2	sensory	11	576	C2	ilpd indian liver	10	583
R3	disclosure z	3	662	C3	credit approval	15	690
R4	bike share day	11	731	C4	breast cancer wise	9	699
R5	era	4	1000	C5	pima	8	768
R6	treasury	15	1049	C6	tic tac toe	9	958
R7	airfoil	5	1503	C7	statlog german credit	24	1000
R8	wine red	11	1599	C8	pc1	21	1109
R9	skill craft	18	3395	C9	seismic bumps	15	2584
R10	abalone	8	4177	C10	churn	20	5000
R11	parkinsons tele	19	5875	C11	banana	2	5300
R12	wind	14	6574	C12	twonorm	20	7400
R13	cpu small	12	8192	C13	ringnorm	20	7400
R14	topo 2 1	266	8885	C14	jm1	21	10885
R15	combined cycle power plant	4	9568	C15	eeg eye state	14	14980
R16	electrical grid	11	10000	C16	magic telescope	10	19020
R17	aileron	40	13750	C17	adult	14	32561
R18	elevators	18	16599	C18	nomao	118	34465
R19	bike share hour	12	17379	C19	bank marketing	16	45211
R20	california housing	8	20640	C20	electricity	8	45312

Table 4: The data sets for testing different HPO methods.

loss for classification) with the “Nystroem” transformer. Here we use this approach to handle data with more than 3000 samples. The number of components in the “Nystroem” transformer is fixed to 500. The initial learning rate is set to 0.01, and we use the “adaptive” optimizer to adjust the learning rate during optimization.

XGBoost. Hyperparameter optimization of XGBoost is much more complicated than that of SVM. Eight important hyperparameters in XGBoost are introduced, including booster (categorical; “gbtree” or “gblinear”), maximum tree depth (integer-valued; within the range $[1, 8]$), number of estimators (integer-valued; within the range $[100, 500]$), ratio of features in each tree (continuous; within the range $[0.5, 1]$), learning rate (continuous; the base-10 log scale within the range $[10^{-5}, 10^0]$), minimum loss reduction (continuous; the base-10 log scale within the range $[10^{-5}, 10^0]$), ℓ_1 -regularization (continuous; the base-10 log scale within the range $[10^{-5}, 10^0]$) and ℓ_2 -regularization (continuous; the base-10 log scale within the range $[10^{-5}, 10^0]$).

Pipeline Optimization. In addition to optimizing a single machine model’s hyperparameters, we move one step further to the challenging pipeline optimization task. In particular, we consider data preprocessing (“MinmaxScaler” and “Standardizer”), feature engineering (All Features, “SelectKBest”, and “PCA”), model selection (SVM and XGBoost) and HPO for the selected model. Each of the first three steps, i.e., data preprocessing, feature engineering, and model selection, can be treated as a categorical hyperparameter. In data preprocessing, “MinmaxScaler” linearly maps each feature within $[0, 1]$; “Standardizer” instead standardizes each feature with zero mean and unit variance. While for feature

engineering, “SelectKBest” selects the top- K features with the highest F-values. We tune K within $[1, \min\{m, 20\}]$ (m denotes the number of features after one-hot encoding); similarly, “PCA” denotes the principal component analysis, and the number of principal components is selected within $[1, \min\{m, 20\}]$. For the selected machine learning model (either SVM or XGBoost), we use their corresponding HPO configurations as mentioned above.

In total, six groups of tasks are involved, i.e., SVM-Regression (SVM-Reg), SVM-Classification (SVM-Cls), XGBoost-Regression (XGB-Reg), XGBoost-Classification (XGB-Cls), Pipeline-Regression (Pipe-Reg), and Pipeline-Classification (Pipe-Cls). The same settings are used for the compared HPO methods as in Section 5. That is, all the compared methods are allowed to evaluate at most 100 hyperparameter configurations. For SeqUD and SeqRand, we use $n = q = 15$ when $s \leq 5$ and $n = q = 25$ when $s > 5$. The optimization results regarding the 5-fold CV performance, computing time, and test set performance are all recorded. Each experiment is repeated ten times. The final results are reported with average, standard deviation, and statistical significance across the ten repetitions.

6.2 Results Analysis

The 5-fold CV and test set results of HPO experiments on the 40 data sets are reported in the appendix. Bold numbers indicate the best-performing methods, and results that are significantly different from the best are underlined. Note that the RMSE results should be multiplied by the corresponding scaling factors in the last column. We analyze the results from the following perspectives.

Five-fold CV Performance. As the optimization target, 5-fold CV performance directly reflects the optimization capability of each compared method. To make a clear comparison, we rank the methods according to their averaged results for each data, as shown in Figure 12. A one-verse-one win/loss comparison for all the methods is further presented in Table 5, which is summarized over the 120 tasks (40 data sets by 3 machine learning models). For instance, in the cell for “Rand”-“Grid”, “19 (7)” denotes that random search is better than grid search on 19 tasks in which 7 tasks are tested to be significantly better. Similarly, random search also shows inferior performance to grid search on 21 tasks, with 9 tasks being significantly worse. Note that grid search is applied only for 2-D synthetic functions, hence the total number of compared tasks is 40.

Several interesting findings are observed. First, the uniformity of design points is positively related to the performance of non-sequential methods. With better uniformity, both UD and Sobol show slightly better performance than random search and LHS, but no big difference is observed between UD and Sobol. Given 100 runs, both UD and Sobol have better uniformity performance than random search and LHS. For instance, in the case of SVM with $s = 2$, the best uniformity among non-sequential methods is achieved by UD ($CD_2 = 0.000035$), followed by Sobol ($CD_2 = 0.000142$), LHS ($CD_2 = 0.000340$), and random search ($CD_2 = 0.003440$).

Second, sequential methods are in general superior to non-sequential methods. SeqRand outperforms random search on 106 out of the 120 tasks. The proposed SeqUD also significantly beats the non-sequential UD on 109 out of the 120 tasks. The other sequential methods, including TPE, SMAC, and GP-EI, also exhibit competitive performances.

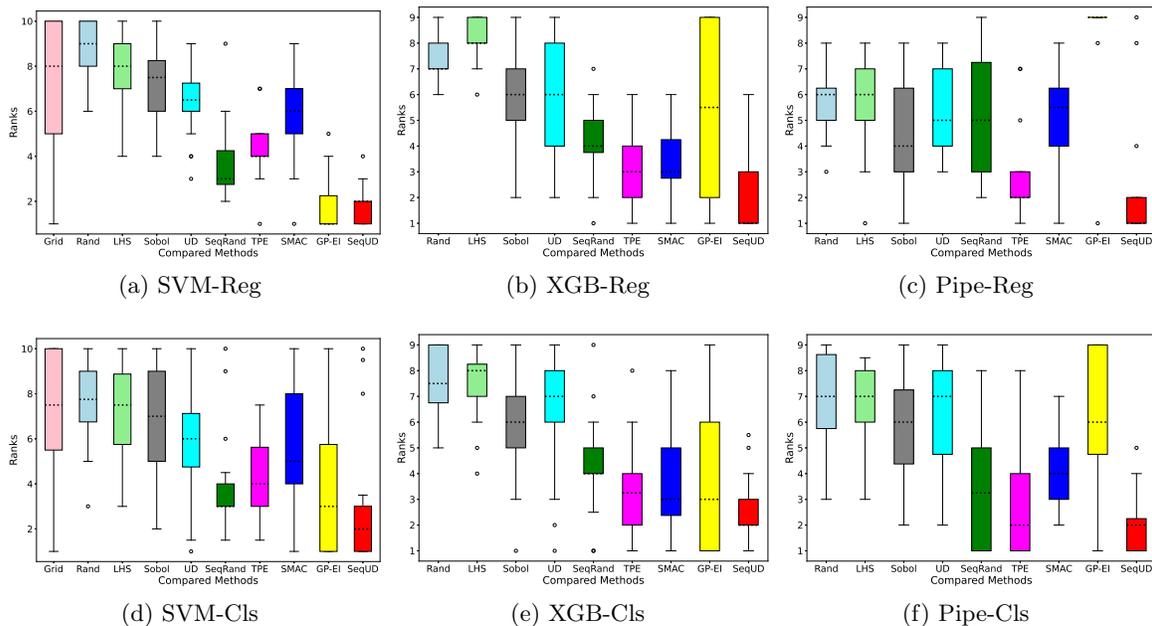


Figure 12: The ranks of 5-fold CV performance across different tasks. Each sub-figure represents one of the six tasks, and the boxes are drawn based on the averaged results of corresponding 20 data sets.

Third, by incorporating uniform design and sequential optimization, the proposed SeqUD achieves the best overall performance among all the compared methods. Given the proposed sequential strategy, SeqUD significantly improves over its naïve baseline SeqRand. The SeqUD results are better than or on par with its counterpart Bayesian optimization methods in most cases. GP-EI is the second-best among all the tested methods. The GP-EI method works well on low-dimensional tasks (i.e., SVM), but it is less efficient for high-dimensional tasks (i.e., XGBoost and pipeline optimization). On the contrary, TPE and SMAC are more robust for high-dimensional tasks.

Computational Cost. The computing time required by each method is reported in Figure 13. In SeqUD, augmented design points can be efficiently generated within a few seconds. This is negligible as compared to the computational complexity of training a machine learning model. Therefore, the time complexity of SeqUD is close to that of simple non-sequential methods like grid search and random search. Moreover, the proposed SeqUD, SeqRand, and all the non-sequential methods can be further accelerated by performing design point-level parallelization. This type of parallelization will not sacrifice the optimization performance. However, Bayesian optimization methods are less efficient regarding computing time. Both TPE and SMAC are fast for optimizing 2-D SVM tasks; while for high-dimensional tasks, SMAC and GP-EI would need more time for surrogate modeling and acquisition optimization.

	Grid	Rand	LHS	Sobol	UD	SeqRand	TPE	SMAC	GP-EI	SeqUD
Grid	-	21 (9)	19 (6)	19 (12)	16 (9)	7 (2)	9 (1)	12 (2)	6 (2)	5 (2)
Rand	19 (7)	-	57 (0)	32 (4)	39 (4)	14 (1)	9 (1)	17 (1)	39 (15)	7 (2)
LHS	20 (10)	61 (8)	-	34 (4)	40 (7)	20 (0)	7 (0)	23 (1)	43 (14)	6 (3)
Sobol	21 (11)	86 (30)	84 (21)	-	61 (24)	29 (4)	17 (2)	36 (4)	52 (19)	10 (4)
UD	23 (13)	80 (26)	79 (18)	59 (22)	-	30 (6)	21 (3)	41 (3)	46 (20)	10 (3)
SeqRand	33 (21)	106 (54)	98 (59)	90 (49)	89 (50)	-	49 (5)	65 (16)	66 (19)	26 (4)
TPE	31 (17)	109 (71)	112 (68)	102 (60)	95 (63)	69 (10)	-	82 (18)	70 (25)	32 (7)
SMAC	27 (12)	103 (45)	95 (51)	82 (46)	79 (41)	54 (8)	38 (5)	-	61 (24)	20 (5)
GP-EI	34 (27)	81 (49)	75 (46)	66 (49)	73 (43)	54 (30)	50 (31)	57 (29)	-	34 (18)
SeqUD	35 (28)	113 (83)	114 (81)	109 (80)	109 (81)	91 (36)	87 (39)	97 (47)	85 (31)	-

Table 5: Pairwise win/loss over the 120 HPO tasks (5-fold CV): the numbers in each cell indicate how often the method in row (significantly) outperforms the method in column. The statistical significance is calculated by paired t-test with a significance level of 0.05.

It is observed that SeqUD often runs faster than grid search and random search. First, the time cost of generating (augmented) uniform designs using AugUD is rather small. In most cases, it can be finished within a few seconds, which is negligible as compared to the training cost of machine learning models. Second, the training time of a machine learning model depends on specific hyperparameter settings. In grid search and random search, they may generate many design points on time-consuming hyperparameter configurations; while for SeqUD, it focuses more on the best performing regions, which could be less computationally intensive.

Test Set Performance. Figure 14 reports the ranks of test set performance achieved by different methods. Table 6 provides the one-verse-one win/loss comparison results.

Hyperparameter configurations that achieve better 5-fold CV performance can usually perform well on the test set. The Spearman rank-order correlation coefficient is calculated between SeqUD’s 5-fold CV and test set performance (over the 120 tasks), which are shown positively correlated with a correlation coefficient greater than 0.6. The proposed SeqUD achieves not only superior 5-fold CV performance but also competitive test set performance. For instance, SeqUD outperforms each of GP-EI, SMAC, and TPE 79 times. The superiority of SeqUD over other non-sequential methods is much more significant. Therefore, it is evident that the proposed SeqUD approach is a competitive HPO method.

Despite the positive relationship between 5-fold CV and test set performance, there exist cases that are not consistent. For example, sequential methods generally perform much better than non-sequential methods regarding 5-fold CV performance, while for test set performance, the gap is reduced. Such an observation can be due to overfitting. It is known that when there are a limited number of validation samples, the hyperparameters can overfit the validation set. However, the best hyperparameter configuration on the validation set does not necessarily generalize well for the test set (Hutter et al., 2019). Some possible solutions have been proposed in the literature to alleviate this problem, e.g., data reshuffling for each function evaluation (Lévesque, 2018), finding stable optima instead of

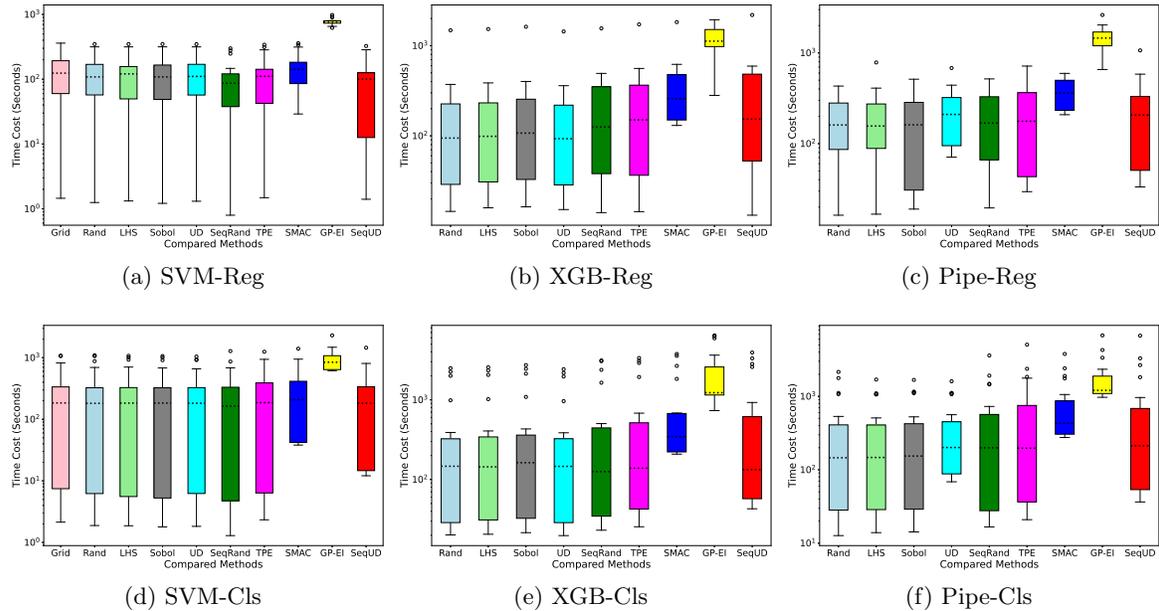


Figure 13: Computing time comparison across different tasks. Each sub-figure represents one of the six tasks, and the boxes are drawn based on the averaged results of corresponding 20 data sets.

	Grid	Rand	LHS	Sobol	UD	SeqRand	TPE	SMAC	GP-EI	SeqUD
Grid	-	20 (7)	18 (10)	19 (9)	19 (4)	16 (4)	13 (2)	19 (2)	12 (2)	10 (2)
Rand	18 (4)	-	55 (4)	40 (2)	49 (9)	40 (1)	31 (3)	34 (3)	56 (14)	24 (3)
LHS	21 (4)	65 (5)	-	37 (5)	49 (10)	38 (4)	40 (0)	42 (2)	58 (19)	24 (4)
Sobol	18 (7)	77 (19)	80 (21)	-	64 (25)	57 (9)	44 (7)	62 (7)	61 (19)	30 (9)
UD	19 (6)	69 (18)	69 (14)	53 (18)	-	54 (9)	42 (5)	53 (7)	63 (19)	35 (5)
SeqRand	24 (7)	79 (23)	80 (18)	62 (16)	64 (19)	-	54 (8)	56 (7)	58 (17)	32 (5)
TPE	24 (7)	86 (30)	79 (29)	74 (29)	75 (28)	64 (9)	-	67 (6)	64 (15)	39 (7)
SMAC	21 (7)	86 (20)	76 (20)	56 (24)	65 (17)	62 (6)	47 (4)	-	67 (21)	35 (7)
GP-EI	28 (13)	64 (26)	60 (27)	58 (25)	56 (23)	59 (12)	54 (11)	53 (14)	-	41 (8)
SeqUD	28 (12)	94 (43)	93 (43)	87 (36)	83 (36)	86 (20)	79 (13)	79 (22)	79 (24)	-

Table 6: Pairwise win/loss over the 120 HPO tasks (test set): the numbers in each cell indicate how often the method in row (significantly) outperforms the method in column. The statistical significance is calculated by paired t-test with a significance level of 0.05.

sharp optima (Dai Nguyen et al., 2017), and the ensemble of multiple good hyperparameter configurations (Momma and Bennett, 2002). However, overfitting is still an open question in HPO, which is worthy of further investigation.

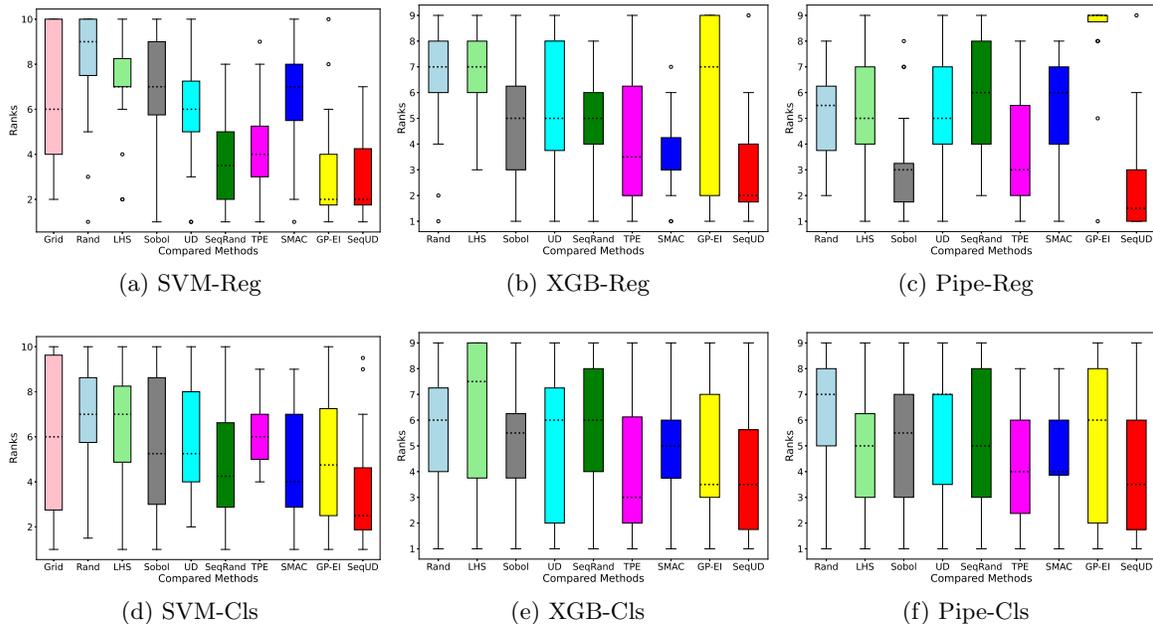


Figure 14: The ranks of test set performance across different tasks. Each sub-figure represents one of the six tasks, and the boxes are drawn based on the averaged results of corresponding 20 data sets.

Possible Limitations. Although the proposed SeqUD achieves promising results on multiple HPO tasks, the experimental evaluation can be possibly limited and may lead to a biased conclusion. First, as the computing resources are limited, this paper only considers two machine learning tasks, three machine learning models, and small and medium-sized data sets. Such empirical results may not generalize universally on other machine learning tasks. Besides, the hyperparameter search space is according to our experience (as well as common practice), but it does not represent all possible hyperparameter settings.

Second, the choice of evaluation metric for regression or classification tasks may affect the final result. In our experiments, it is observed that SeqUD performs slightly better on regression tasks than classification tasks. As SVM can only output binary prediction, we used the accuracy score metric (a discrete measurement) in classification tasks. It is possible that the difference between tested methods cannot be fully revealed. However, for regression tasks, the MSE metric is a continuous measure, under which the superiority of SeqUD has been shown more obvious.

6.3 Comparison with Hyperband

Additional experiments are conducted to compare SeqUD with Hyperband. Hyperband is an adaptive computing resource allocation algorithm based on random search. It is mainly used to tune the models that can be iteratively trained, where the performance of different trials can be evaluated during training, such that the worst half trials are early stopped.

For instance, hyperparameters of deep neural networks can be tuned by Hyperband, as they are typically fitted by the iterative backpropagation algorithm. In this paper, we compare Hyperband with SeqUD using the XGBoost model, as the amount of computing resources in XGBoost can be determined by the number of estimators. In Hyperband, the search space of hyperparameters is kept the same as that of SeqUD, except for the number of estimators, which is instead used as the indicator of amount of resources.

In SeqUD, the optimization is terminated as the maximal number of runs is reached. But for Hyperband, the termination is not controlled by the maximal number of runs, but by “min_iter” (the minimum amount of resource), “max_iter” (the maximum amount of resource), and “eta” (inverse of the proportion of configurations that are discarded). For a fair comparison, we set “min_iter”=10, “max_iter”=500, and “eta”=2. Under this setting, the total number of runs of Hyperband is 138, which is larger than that of SeqUD (100).

The test set performance is reported in Table 7, and the proposed SeqUD shows significantly better performance than Hyperband on the tested tasks. The reason is that Hyperband still uses random search as the core algorithm for optimization, and its main advantage lies in the adaptive resource allocation. However, it is still hard to observe the superior performance of Hyperband, even given more trials.

Regression (RMSE)				Classification (Accuracy %)		
Data Set	Hyperband	SeqUD	scale	Data Set	Hyperband	SeqUD
R1	<u>5.064</u> ±0.261	4.829 ±0.184	×0.1	C1	96.63 ±0.82	96.28±0.90
R2	7.304±0.252	7.283 ±0.253	×0.1	C2	69.97 ±0.98	69.28±2.14
R3	2.426±0.104	2.424 ±0.105	×10000	C3	86.52 ±1.46	86.52 ±1.49
R4	7.293±0.341	7.146 ±0.274	×100	C4	<u>96.09</u> ±0.89	96.63 ±0.54
R5	1.576±0.018	1.576 ±0.018	×1	C5	75.42±1.61	75.62 ±1.29
R6	2.424 ±0.204	2.436±0.193	×0.1	C6	98.71±0.61	98.81 ±0.41
R7	<u>2.089</u> ±0.159	1.980 ±0.179	×1	C7	75.12 ±1.44	<u>74.48</u> ±1.14
R8	6.201±0.115	6.136 ±0.115	×0.1	C8	93.05 ±0.45	93.01±0.65
R9	<u>9.280</u> ±0.130	9.209 ±0.132	×0.1	C9	93.27±0.20	93.36 ±0.11
R10	<u>2.219</u> ±0.040	2.185 ±0.032	×1	C10	<u>95.43</u> ±0.22	95.50 ±0.23
R11	<u>2.477</u> ±0.325	2.015 ±0.111	×1	C11	74.13±0.78	75.29 ±4.90
R12	<u>3.144</u> ±0.034	3.114 ±0.032	×1	C12	97.64 ±0.13	97.62±0.10
R13	2.792±0.068	2.753 ±0.051	×1	C13	97.61±0.27	97.75 ±0.28
R14	<u>2.899</u> ±0.153	2.886 ±0.154	×0.01	C14	81.23±0.19	81.28 ±0.22
R15	<u>3.355</u> ±0.091	3.276 ±0.053	×1	C15	<u>92.84</u> ±0.50	93.53 ±0.27
R16	<u>9.757</u> ±0.476	9.042 ±0.195	×0.001	C16	<u>87.75</u> ±0.29	87.96 ±0.16
R17	<u>1.630</u> ±0.020	1.603 ±0.014	×0.0001	C17	87.08±0.21	87.10 ±0.16
R18	<u>2.292</u> ±0.077	2.214 ±0.052	×0.001	C18	97.03±0.16	97.09 ±0.17
R19	4.368±0.225	4.228 ±0.071	×10	C19	<u>90.66</u> ±0.12	90.76 ±0.10
R20	<u>4.705</u> ±0.091	4.654 ±0.064	×0.1	C20	<u>92.08</u> ±0.60	92.99 ±0.23

Table 7: Test set performance comparison between HyperBand and SeqUD over XGBoost optimization tasks. The best performing methods are highlighted in bold and underlined results are significantly different from the best ones.

6.4 Comparison with AutoSklearn

Since AutoSklearn is a fully automated pipeline that integrates many machine learning models, it is employed to make comparison with SeqUD for the pipeline optimization task. AutoSklearn is configured with the default settings, while SeqUD follows the pipeline optimization setup in Section 6.1. In AutoSklearn, the termination of the optimization is mainly controlled by the time limit. For each data set, we first calculate the average time cost of SeqUD (T_{sequd}), then set the time limit of AutoSklearn to be $1.05T_{sequd}$.

The experimental results in Table 8 show that SeqUD beats AutoSklearn in 17 out of the 40 compared tasks but fails in the other 23 tasks. The results indicate that AutoSklearn is slightly better than SeqUD. The AutoSklearn package is mainly based on SMAC. The reasons why AutoSklearn can achieve such a good performance are summarized below. First, AutoSklearn has a much larger search space. For instance, it is optimized over tens of machine learning models and has a higher probability of achieving better predictive performance. Second, AutoSklearn is enhanced with meta-learning of data set information and ensemble learning. Therefore, the direct comparison between SeqUD and AutoSklearn is somehow unfair.

Nevertheless, it is possible to enhance SeqUD with such practical pipeline optimization techniques in our future work.

Regression (RMSE)				Classification (Accuracy %)		
Data Set	AutoSklearn	SeqUD	scale	Data Set	AutoSklearn	SeqUD
R1	5.197±0.711	4.829±0.184	×0.1	C1	<u>62.81±0.00</u>	96.95±0.63
R2	<u>8.303±0.182</u>	7.283±0.253	×0.1	C2	70.58±1.30	<u>68.66±2.70</u>
R3	2.421±0.107	2.424±0.105	×10000	C3	85.88±1.22	85.83±1.50
R4	<u>1.957±0.024</u>	0.715±0.027	×1000	C4	96.57±0.72	96.54±0.84
R5	1.568±0.017	<u>1.576±0.018</u>	×1	C5	<u>72.14±4.72</u>	76.04±1.24
R6	<u>2.127±1.515</u>	0.244±0.019	×1	C6	<u>34.66±0.00</u>	99.35±0.30
R7	<u>2.108±0.101</u>	1.980±0.179	×1	C7	75.02±0.93	74.72±1.60
R8	5.939±0.093	<u>6.136±0.115</u>	×0.1	C8	92.77±0.40	93.08±0.60
R9	<u>9.459±0.185</u>	9.209±0.132	×0.1	C9	93.41±0.05	93.40±0.08
R10	2.181±0.048	2.185±0.032	×1	C10	<u>95.10±0.40</u>	95.46±0.55
R11	<u>8.374±0.110</u>	2.015±0.111	×1	C11	<u>89.83±0.47</u>	90.52±0.29
R12	<u>3.338±0.245</u>	3.114±0.032	×1	C12	97.73±0.16	97.72±0.18
R13	2.723±0.085	<u>2.753±0.051</u>	×1	C13	98.58±0.14	98.62±0.16
R14	2.868±0.158	<u>2.886±0.154</u>	×0.01	C14	81.57±0.23	<u>81.23±0.22</u>
R15	3.234±0.064	<u>3.276±0.053</u>	×1	C15	96.91±0.95	<u>92.87±1.05</u>
R16	7.980±0.135	<u>9.042±0.195</u>	×0.001	C16	88.47±0.24	<u>87.86±0.27</u>
R17	1.584±0.020	<u>1.603±0.014</u>	×0.0001	C17	87.14±0.14	87.09±0.15
R18	2.184±0.042	<u>2.214±0.052</u>	×0.001	C18	<u>96.96±0.18</u>	97.07±0.16
R19	4.187±0.172	4.228±0.071	×10	C19	90.79±0.11	<u>90.70±0.10</u>
R20	4.507±0.042	<u>4.654±0.064</u>	×0.1	C20	<u>91.96±1.08</u>	92.66±0.11

Table 8: Test set performance comparison between AutoSklearn and SeqUD over Pipeline optimization tasks. The best performing methods are highlighted in bold and underlined results are significantly different from the best ones.

7. Conclusion

In this paper, we propose a SeqUD framework for global optimization and potential application in HPO of machine learning models. A real-time AugUD algorithm is introduced for fast construction of augmented uniform designs. The proposed SeqUD combines the benefits of uniform designs and sequential exploitation. It balances exploration and exploitation, and could be easily parallelized to save computation time. Both synthetic function optimization and HPO experiments are conducted. The results show that the proposed approach is a competitive alternative to existing HPO methods.

The SeqUD method can be improved in several directions. First, to avoid SeqUD trials being trapped into local optima, the multiple shooting strategy mentioned in Section 4 is worthy of our future investigation. Second, the generalization ability of different hyperparameters can be incorporated into the optimization framework. Third, the current SeqUD implementation requires a suitable pre-specified range of each hyperparameter and needs to determine the number of runs/levels per stage. In the future, it is a promising direction to extend SeqUD to a fully automated HPO framework.

Acknowledgments

We thank the action editor and the four independent reviewers for their valuable comments and suggestions, which helped us improve the quality of the paper.

Appendix A. Additional Results

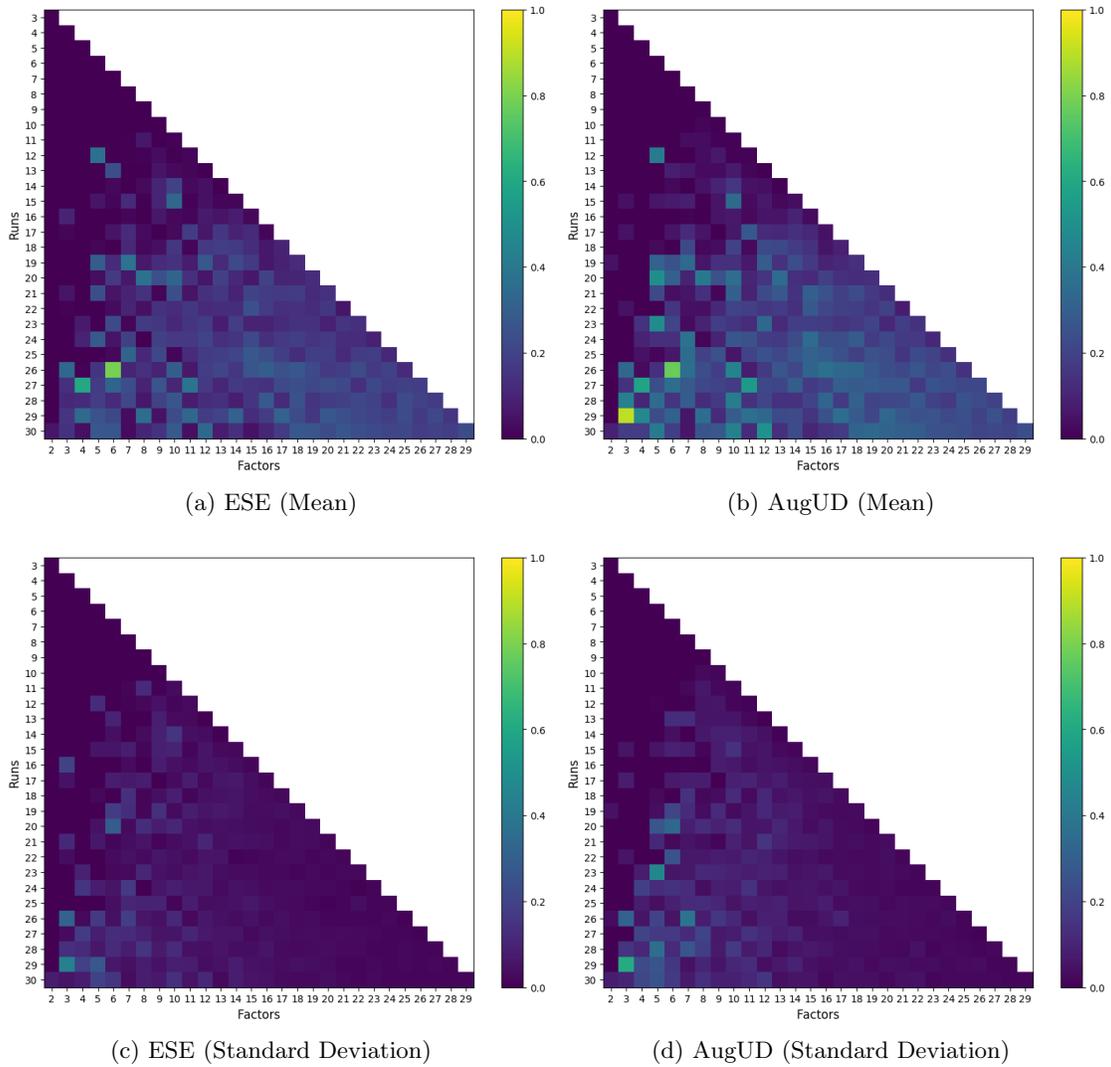


Figure 15: Average Improvement ratios (%) of ESE and AugUD against uniform designs obtained from the uniform design website. For each sub-figure, the x-axis represents the number of factors and the y-axis represents the number of runs.

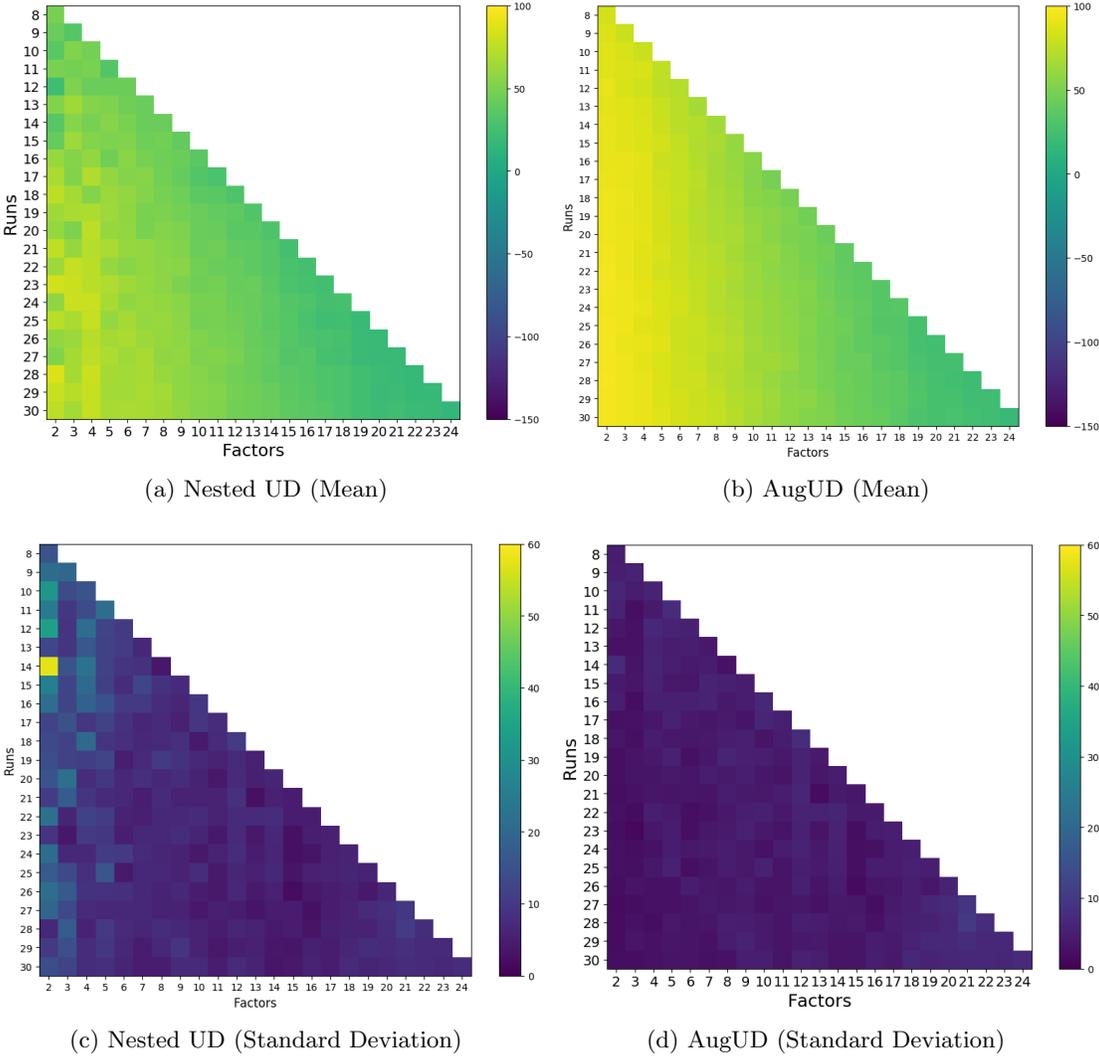


Figure 16: Average improvement ratios (%) of nested UD and AugUD against random augmentation. For each sub-figure, the x-axis represents the number of factors and the y-axis represents the number of runs.

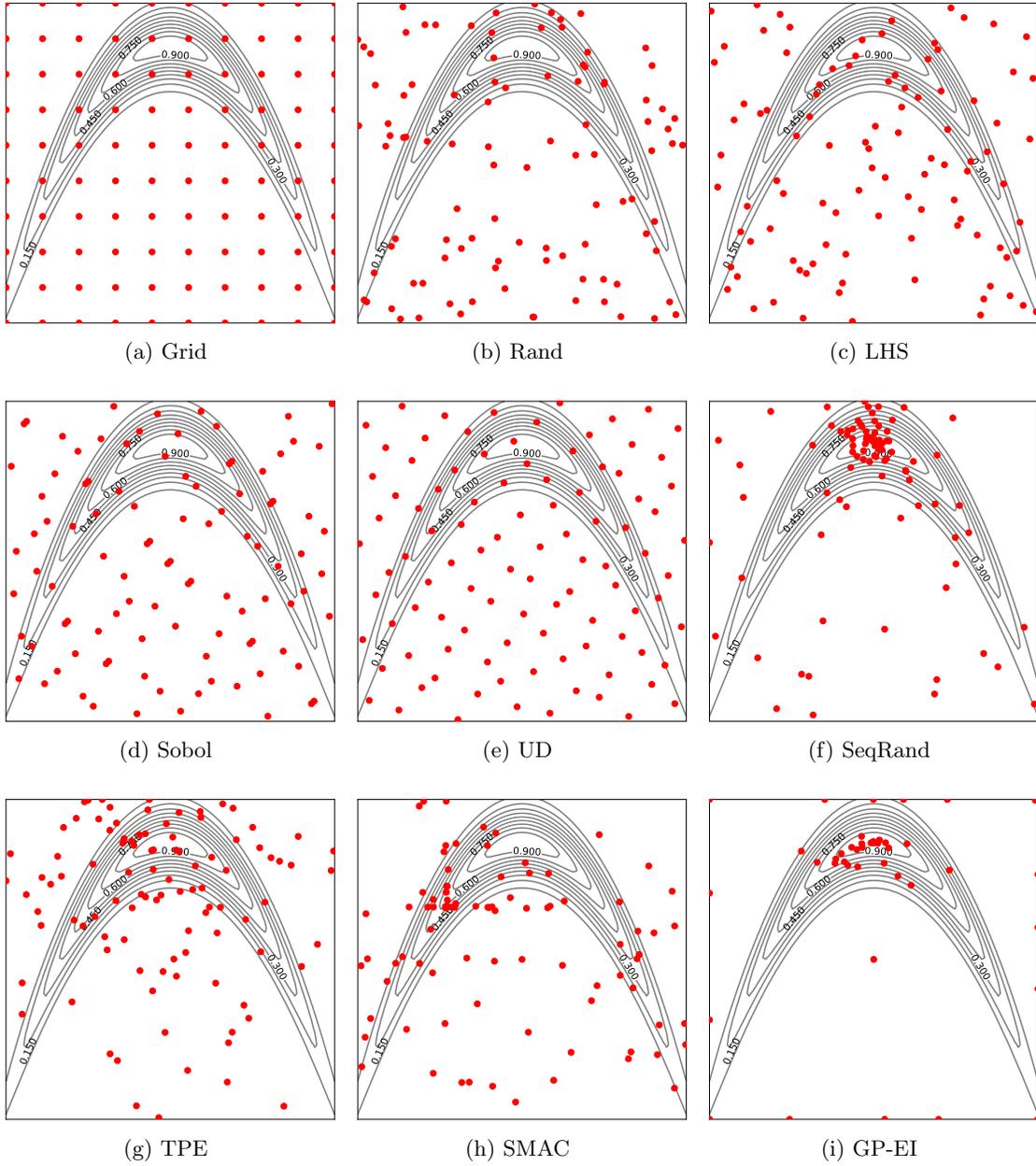


Figure 17: The evaluated design points by each benchmark method against the ground truth contour plot of the cliff function. Each red point represents an evaluated point, and the actual optimal point of the cliff function is located in the upper center area.

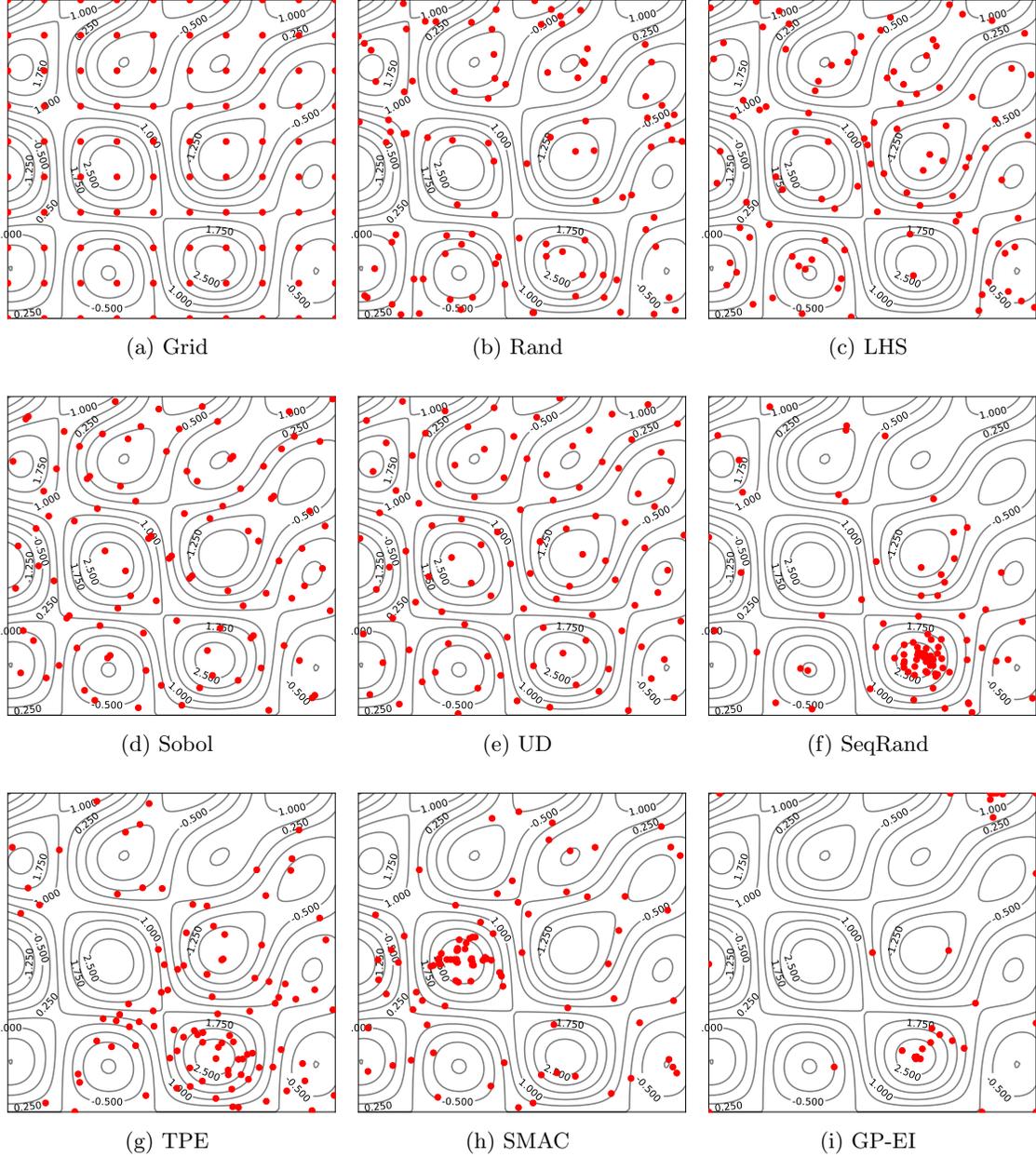


Figure 18: The evaluated design points by each benchmark method against the ground truth contour plot of the octopus function. Each red point represents an evaluated point, and the actual optimal point of the octopus function is located in the center-left area.

Table 9: Average optimal values found over 32 synthetic global optimization tasks (100 runs).

Name	Grid	Rand	LHS	Sobol	UD	SeqRand	TPe	SMAC	GP-EI	SeqUD	Scale
bukint6	1.845±0.000	1.607±0.851	1.366±0.865	1.564±0.000	3.286±0.000	0.688±0.343	1.156±0.554	0.564±0.448	0.352±0.183	0.455±0.126	×10
crossit	-2.050±0.000	-2.021±0.038	-2.024±0.035	-2.060±0.000	-2.058±0.000	-2.048±0.044	-2.051±0.014	-2.045±0.025	-1.922±0.084	-2.063±0.000	×1
egg	-9.206±0.000	-7.438±0.985	-7.636±0.902	-6.736±0.000	-6.601±0.000	-7.470±1.369	-7.833±1.077	-7.234±1.236	-5.825±1.616	-7.355±0.286	×100
holder	-1.736±0.000	-1.589±0.248	-1.610±0.289	-1.213±0.000	-1.777±0.000	-1.740±0.349	-1.782±0.124	-1.819±0.228	-1.915±0.022	-0.950±0.000	×10
langier	-3.507±0.000	-3.287±0.638	-3.258±0.628	-2.781±0.000	-3.668±0.000	-3.358±0.841	-3.493±0.641	-3.380±0.753	-3.401±0.797	-1.478±0.000	×1
levy	0.152±0.000	3.650±3.028	2.948±2.678	2.652±0.000	0.123±0.000	0.331±1.041	1.248±1.166	1.766±1.664	0.004±0.008	0.002±0.001	×0.1
levy13	0.789±0.000	2.261±1.808	2.210±1.823	2.000±0.000	0.685±0.000	0.237±0.170	0.852±0.573	1.091±0.622	0.040±0.047	0.035±0.006	×1
schwef	-	1.355±0.198	1.295±0.205	1.514±0.000	1.377±0.000	1.366±0.181	1.257±0.211	0.998±0.262	1.187±0.136	1.221±0.216	×1000
shubert	-0.218±0.000	-1.063±0.453	-1.045±0.443	-0.471±0.000	-1.565±0.000	-1.238±0.585	-1.248±0.448	-1.284±0.541	-0.952±0.649	-1.864±0.000	×100
booth	0.914±0.000	3.727±3.351	3.733±3.353	1.432±0.000	5.220±0.000	0.180±0.660	1.240±1.198	4.781±5.095	0.013±0.041	0.000±0.000	×1
mccorm	-0.991±0.000	-0.772±0.112	-0.757±0.113	-0.704±0.000	-0.735±0.000	-0.871±0.147	-0.941±0.073	-0.756±0.175	-1.012±0.000	-0.946±0.090	×10
powersum	-	1.625±1.756	1.836±1.862	3.686±0.000	0.983±0.000	0.619±1.040	0.843±0.859	2.238±2.794	2.244±3.045	0.560±0.695	×10
zakharov	-	1.917±1.009	2.029±1.056	2.175±0.000	2.345±0.000	0.538±0.610	0.930±0.625	2.586±1.429	2.956±0.250	0.801±0.810	×10
cannell6	-0.791±0.000	-0.850±0.161	-0.844±0.164	-0.939±0.000	-1.000±0.000	-1.022±0.012	-0.949±0.079	-0.895±0.164	-1.031±0.003	-1.031±0.000	×1
dixonpr	-	4.102±4.765	5.000±5.116	0.010±0.000	3.248±0.000	0.304±0.449	0.780±0.727	0.010±0.000	0.010±0.000	0.078±0.116	×100
rosen	-	3.503±2.005	3.616±1.954	0.986±0.000	3.309±0.000	0.470±0.500	1.011±0.643	0.891±0.188	0.595±0.158	0.131±0.111	×10000
pern0db	0.334±0.000	0.636±0.868	0.360±0.414	0.479±0.000	0.111±0.000	0.032±0.090	0.123±0.171	0.580±0.767	1.785±2.402	0.004±0.000	×10
trid	-1.951±0.000	-1.800±0.191	-1.855±0.141	-1.938±0.000	-1.902±0.000	-1.996±0.003	-1.955±0.040	-1.886±0.141	-2.000±0.000	-2.000±0.000	×1
dejong5	4.835±0.000	0.445±0.608	0.235±0.344	0.054±0.000	0.756±0.000	0.125±0.085	0.152±0.219	0.109±0.032	0.107±0.031	0.040±0.000	×100
easom	-0.000±0.000	-0.024±0.239	-0.188±1.318	-0.000±0.000	-0.000±0.000	-1.429±2.461	-0.004±0.037	-0.095±0.914	-0.000±0.000	-7.748±3.761	×0.1
mnichal	-	-2.167±0.351	-2.139±0.293	-2.075±0.000	-3.292±0.000	-2.734±0.490	-2.497±0.414	-3.020±0.432	-3.148±0.509	-2.489±0.535	×1
beale	0.395±0.000	0.695±0.629	0.567±0.674	0.266±0.000	0.814±0.000	0.419±0.588	0.254±0.383	0.520±0.514	4.681±0.426	0.000±0.000	×1
branin	0.789±0.000	0.960±0.652	0.887±0.458	1.079±0.000	0.465±0.000	0.511±0.232	0.664±0.280	0.742±0.451	0.398±0.000	0.398±0.000	×1
colville	-	1.758±1.572	1.452±1.141	0.042±0.000	0.473±0.000	0.295±0.542	0.493±0.410	0.341±0.005	0.035±0.007	0.258±0.518	×1000
goldpr	0.197±0.000	0.202±0.159	0.228±0.180	0.246±0.000	0.149±0.000	0.101±0.198	0.099±0.090	0.324±0.254	1.029±1.265	0.030±0.000	×100
hart3	-	-3.628±0.163	-3.619±0.176	-3.784±0.000	-3.564±0.000	-3.782±0.154	-3.768±0.073	-3.764±0.155	-3.824±0.168	-3.852±0.018	×1
hart4	-	-2.625±0.203	-2.588±0.243	-2.402±0.000	-2.948±0.000	-2.976±0.164	-2.876±0.151	-2.807±0.195	-3.061±0.110	-3.126±0.015	×1
hart6	-	-2.410±0.206	-2.390±0.220	-2.465±0.000	-2.542±0.000	-2.755±0.154	-2.672±0.156	-2.682±0.233	-3.026±0.039	-2.917±0.020	×1
permdb	1.516±0.000	1.089±1.340	1.410±1.832	0.244±0.000	0.104±0.000	1.310±5.568	0.249±0.313	0.914±1.166	0.015±0.060	0.002±0.003	×0.1
powell	-	7.422±6.988	6.881±6.675	3.031±0.000	3.533±0.000	0.988±1.032	2.608±2.291	2.508±0.838	0.098±0.331	0.208±0.308	×10
shekel	-	-1.221±0.453	-1.271±0.702	-1.326±0.000	-1.070±0.000	-3.020±1.541	-1.593±0.372	-2.517±1.752	-5.957±3.337	-3.954±1.678	×1
styblang	-	-1.626±0.159	-1.660±0.138	-1.828±0.000	-1.742±0.000	-1.778±0.166	-1.741±0.156	-1.814±0.175	-1.889±0.187	-1.968±0.035	×100

Table 10: Average optimal values found over 32 synthetic global optimization tasks (1000 runs).

Name	Grid	Rand	LHS	Sobol	UD	SeqRand	TPE	SMAC	SeqUD	Scale
bukin6	2.811 ± 0.000	4.737 ± 2.387	4.658 ± 2.510	0.978 ± 0.000	6.392 ± 0.000	2.367 ± 1.164	2.996 ± 1.475	0.864 ± 0.465	0.931 ± 0.433	$\times 1$
crossit	-2.048 ± 0.000	-2.059 ± 0.004	-2.058 ± 0.005	-2.060 ± 0.000	-2.050 ± 0.000	-2.062 ± 0.000	-2.062 ± 0.000	-2.062 ± 0.000	-2.063 ± 0.000	$\times 1$
egg	-8.485 ± 0.000	-8.903 ± 0.345	-8.913 ± 0.364	-8.957 ± 0.000	-8.799 ± 0.000	-9.102 ± 0.527	-9.224 ± 0.513	-9.126 ± 0.350	-7.166 ± 0.007	$\times 100$
holder	-1.830 ± 0.000	-1.892 ± 0.025	-1.889 ± 0.024	-1.891 ± 0.000	-1.903 ± 0.000	-1.918 ± 0.002	-1.918 ± 0.004	-1.921 ± 0.000	-1.921 ± 0.000	$\times 10$
langer	-4.022 ± 0.000	-3.942 ± 0.135	-3.933 ± 0.137	-4.034 ± 0.000	-3.887 ± 0.000	-4.107 ± 0.029	-3.940 ± 0.517	-4.055 ± 0.099	-4.122 ± 0.030	$\times 1$
levy	0.122 ± 0.000	4.988 ± 4.569	5.004 ± 4.784	4.350 ± 0.000	2.727 ± 0.000	0.068 ± 0.061	0.224 ± 0.258	1.045 ± 2.268	0.001 ± 0.001	$\times 0.01$
levy13	0.918 ± 0.000	5.051 ± 3.121	4.870 ± 2.966	2.700 ± 0.000	2.998 ± 0.000	0.338 ± 0.372	0.825 ± 0.569	1.074 ± 1.233	0.002 ± 0.002	$\times 0.1$
schwef	-	1.001 ± 0.162	0.978 ± 0.173	0.849 ± 0.000	0.652 ± 0.000	0.963 ± 0.177	0.707 ± 0.209	0.288 ± 0.147	0.755 ± 0.080	$\times 1000$
shubert	-1.768 ± 0.000	-1.700 ± 0.159	-1.707 ± 0.130	-1.762 ± 0.000	-1.551 ± 0.000	-1.844 ± 0.108	-1.834 ± 0.034	-1.836 ± 0.152	-1.867 ± 0.000	$\times 100$
booth	0.077 ± 0.000	0.390 ± 0.390	0.369 ± 0.347	1.100 ± 0.000	0.070 ± 0.000	0.006 ± 0.006	0.030 ± 0.033	0.516 ± 0.614	0.000 ± 0.000	$\times 1$
mccorm	-1.004 ± 0.000	-0.928 ± 0.053	-0.929 ± 0.055	-0.905 ± 0.000	-0.945 ± 0.000	-1.005 ± 0.019	-1.011 ± 0.001	-0.993 ± 0.030	-1.012 ± 0.000	$\times 10$
powersum	-	1.831 ± 1.800	1.881 ± 1.930	0.186 ± 0.000	1.770 ± 0.000	0.197 ± 0.221	0.505 ± 0.374	3.141 ± 2.827	0.110 ± 0.102	$\times 1$
zakharov	-	5.408 ± 2.840	5.900 ± 3.524	1.752 ± 0.000	5.560 ± 0.000	0.063 ± 0.074	0.362 ± 0.195	8.522 ± 4.450	0.008 ± 0.006	$\times 1$
camel6	-1.031 ± 0.000	-1.009 ± 0.023	-1.014 ± 0.018	-1.020 ± 0.000	-1.018 ± 0.000	-1.031 ± 0.001	-1.029 ± 0.003	-1.024 ± 0.009	-1.032 ± 0.000	$\times 1$
dixonpr	-	4.815 ± 3.624	4.681 ± 4.285	0.100 ± 0.000	9.647 ± 0.000	0.131 ± 0.305	0.141 ± 0.079	0.100 ± 0.001	0.038 ± 0.095	$\times 10$
rosen	-	1.233 ± 0.705	1.057 ± 0.612	0.986 ± 0.000	0.402 ± 0.000	0.064 ± 0.036	0.037 ± 0.021	0.622 ± 0.263	0.038 ± 0.021	$\times 10000$
permdb	4.444 ± 0.000	2.681 ± 3.167	2.780 ± 2.722	3.100 ± 0.000	0.698 ± 0.000	0.116 ± 0.111	0.377 ± 0.386	2.291 ± 2.388	0.002 ± 0.002	$\times 0.1$
trid	-1.996 ± 0.000	-1.985 ± 0.015	-1.981 ± 0.017	-2.000 ± 0.000	-1.983 ± 0.000	-2.000 ± 0.000	-1.999 ± 0.001	-1.996 ± 0.007	-2.000 ± 0.000	$\times 1$
dejong5	0.999 ± 0.000	4.098 ± 2.526	3.604 ± 2.305	1.034 ± 0.000	1.161 ± 0.000	3.511 ± 2.767	2.164 ± 1.566	5.009 ± 3.234	5.952 ± 1.807	$\times 1$
easom	-9.789 ± 0.000	-0.586 ± 1.705	-0.599 ± 1.980	-0.000 ± 0.000	-0.000 ± 0.000	-7.676 ± 1.855	-3.893 ± 2.991	-1.852 ± 3.836	-9.955 ± 0.032	$\times 0.1$
michal	-	-2.799 ± 0.311	-2.783 ± 0.264	-2.753 ± 0.000	-2.882 ± 0.000	-3.553 ± 0.529	-3.571 ± 0.235	-4.084 ± 0.398	-2.569 ± 0.304	$\times 1$
beale	0.124 ± 0.000	0.508 ± 0.507	0.392 ± 0.415	1.251 ± 0.000	1.139 ± 0.000	0.174 ± 1.099	0.030 ± 0.033	0.407 ± 0.421	0.000 ± 0.000	$\times 0.1$
branin	4.264 ± 0.000	4.454 ± 0.432	4.495 ± 0.546	4.764 ± 0.000	4.181 ± 0.000	4.007 ± 0.028	4.063 ± 0.098	4.171 ± 0.221	3.979 ± 0.000	$\times 0.1$
colville	-	2.966 ± 2.132	2.667 ± 1.873	0.420 ± 0.000	6.998 ± 0.000	0.101 ± 0.173	0.134 ± 0.070	0.367 ± 0.098	0.100 ± 0.173	$\times 100$
goldpr	4.008 ± 0.000	4.688 ± 1.773	4.546 ± 1.559	3.317 ± 0.000	4.016 ± 0.000	3.032 ± 0.030	3.122 ± 0.135	6.111 ± 3.131	3.000 ± 0.000	$\times 1$
hart3	-	-3.812 ± 0.039	-3.810 ± 0.032	-3.847 ± 0.000	-3.827 ± 0.000	-3.851 ± 0.021	-3.859 ± 0.002	-3.861 ± 0.002	-3.805 ± 0.019	$\times 1$
hart4	-	-2.946 ± 0.095	-2.938 ± 0.088	-2.875 ± 0.000	-3.067 ± 0.000	-3.103 ± 0.079	-3.098 ± 0.069	-3.101 ± 0.054	-3.134 ± 0.000	$\times 1$
hart6	-	-2.741 ± 0.098	-2.724 ± 0.103	-2.560 ± 0.000	-2.701 ± 0.000	-2.986 ± 0.048	-2.982 ± 0.029	-2.982 ± 0.074	-2.963 ± 0.010	$\times 1$
permdb	0.933 ± 0.000	0.793 ± 0.707	0.905 ± 0.766	0.163 ± 0.000	1.578 ± 0.000	0.046 ± 0.061	0.085 ± 0.119	0.676 ± 0.770	0.001 ± 0.001	$\times 0.01$
powell	-	1.438 ± 0.881	1.338 ± 0.900	1.918 ± 0.000	0.899 ± 0.000	0.027 ± 0.085	0.052 ± 0.035	1.627 ± 0.901	0.002 ± 0.002	$\times 10$
shekel	-	-1.993 ± 0.659	-2.054 ± 0.667	-1.466 ± 0.000	-1.903 ± 0.000	-5.834 ± 2.973	-5.739 ± 1.644	-6.071 ± 3.395	-8.526 ± 3.032	$\times 1$
styblang	-	-1.895 ± 0.111	-1.896 ± 0.100	-1.828 ± 0.000	-2.007 ± 0.000	-2.126 ± 0.120	-2.185 ± 0.099	-2.204 ± 0.104	-2.043 ± 0.020	$\times 100$

Table 11: Average RMSE over different SVM-Reg tasks (5-fold CV).

Data Set	Grid	Rand	LHS	Sobol	UD	SeqRand	TPE	SMAC	GP-EI	SeqUD	scale
R1	5.311±0.137	5.320±0.124	5.318±0.150	5.318±0.128	5.305±0.128	5.307±0.153	5.302±0.129	5.306±0.143	5.293 ±0.132	5.302±0.135	×0.1
R2	7.380±0.184	7.372±0.179	7.362±0.184	7.356±0.176	7.369±0.181	7.351±0.182	7.360±0.177	7.361±0.186	7.353±0.175	7.334 ±0.181	×0.1
R3	2.296 ±0.099	2.299±0.098	2.299±0.100	2.299±0.098	2.298±0.099	2.296±0.101	2.297±0.097	2.298±0.097	2.298±0.099	2.297±0.097	×10000
R4	8.102±0.165	8.094±0.181	8.114±0.170	8.120±0.166	8.098±0.174	8.071±0.182	8.081±0.173	8.091±0.173	8.055 ±0.166	8.070±0.172	×100
R5	1.564±0.020	1.565±0.020	1.565±0.021	1.565±0.021	1.564±0.020	1.564±0.020	1.564±0.020	1.563±0.020	1.564±0.020	1.564±0.021	×1
R6	9.792±0.146	9.344±0.384	9.316±0.287	9.310±0.303	9.300±0.305	9.293±0.309	9.298±0.307	9.325±0.307	9.283 ±0.301	9.283±0.302	×0.1
R7	3.395±0.096	3.369±0.096	3.348±0.090	3.350±0.087	3.370±0.093	3.348±0.117	3.352±0.101	3.346±0.108	3.328±0.104	3.323 ±0.103	×1
R8	6.574±0.060	6.608±0.055	6.593±0.063	6.591±0.065	6.585±0.061	6.594±0.067	6.587±0.071	6.590±0.059	6.560 ±0.067	6.579±0.058	×0.1
R9	9.894±0.152	9.862±0.162	9.861±0.171	9.855±0.158	9.857±0.159	9.847±0.159	9.850±0.159	9.873±0.195	9.841 ±0.160	9.843±0.159	×0.1
R10	2.232±0.027	2.237±0.025	2.234±0.025	2.233±0.027	2.232±0.026	2.230±0.027	2.230±0.026	2.231±0.026	2.228 ±0.027	2.229±0.026	×1
R11	7.104±0.089	7.267±0.276	7.172±0.147	7.119±0.092	7.115±0.097	7.098±0.101	7.103±0.097	7.100±0.098	7.089 ±0.094	7.098±0.094	×1
R12	3.205±0.042	3.203±0.044	3.199±0.040	3.197±0.040	3.200±0.041	3.194±0.041	3.196±0.041	3.197±0.040	3.193 ±0.041	3.193±0.041	×1
R13	4.730±0.112	4.749±0.137	4.723±0.140	4.745±0.108	4.657±0.115	4.667±0.125	4.659±0.123	4.678±0.113	4.619 ±0.116	4.640±0.112	×1
R14	2.991±0.157	3.016±0.190	3.003±0.168	3.015±0.141	3.013±0.184	2.987±0.152	2.984 ±0.149	3.005±0.179	2.989±0.155	2.985±0.155	×0.01
R15	4.303±0.043	4.302±0.045	4.293±0.047	4.299±0.046	4.295±0.047	4.290±0.048	4.289±0.048	4.295±0.045	4.287±0.048	4.286 ±0.048	×1
R16	1.190±0.009	1.156±0.047	1.159±0.046	1.130±0.007	1.154±0.007	1.128±0.007	1.130±0.008	1.140±0.020	1.127±0.009	1.126±0.007	×0.001
R17	1.752±0.018	1.750±0.018	1.747±0.020	1.743±0.019	1.743±0.019	1.741±0.019	1.742±0.019	1.746±0.018	1.742±0.019	1.741±0.019	×0.001
R18	2.923±0.029	2.997±0.095	2.981±0.097	2.947±0.027	2.941±0.024	2.909±0.025	2.916±0.024	2.950±0.047	2.905±0.025	2.903 ±0.024	×0.001
R19	1.277±0.010	1.279±0.016	1.276±0.013	1.274±0.010	1.277±0.010	1.269±0.010	1.270±0.010	1.269±0.010	1.267±0.010	1.267±0.010	×100
R20	6.496±0.050	6.554±0.100	6.520±0.069	6.507±0.050	6.498±0.051	6.496±0.049	6.496±0.051	6.500±0.049	6.493 ±0.050	6.496±0.050	×0.1

Table 12: Average RMSE over different SVM-Reg tasks (test set).

Data Set	Grid	Rand	LHS	Sobol	UD	SeqRand	TPE	SMAC	GP-EI	SeqUD	scale
R1	5.176±0.166	5.142±0.131	5.148±0.187	5.145±0.156	5.138±0.159	5.157±0.178	5.157±0.147	5.127 ±0.173	5.134±0.143	5.140±0.136	×0.1
R2	7.466±0.252	7.476±0.232	7.453±0.237	7.475±0.238	7.441 ±0.243	7.452±0.235	7.473±0.228	7.453±0.233	7.469±0.234	7.470±0.234	×0.1
R3	2.439±0.114	2.442±0.113	2.441±0.118	2.433 ±0.111	2.451±0.119	2.440±0.117	2.440±0.105	2.436±0.112	2.436±0.116	2.437±0.110	×10000
R4	8.102±0.325	8.108±0.317	8.141±0.301	8.124±0.297	8.114±0.294	8.083±0.306	8.085±0.305	8.120±0.240	8.075 ±0.301	8.104±0.320	×100
R5	1.577±0.020	1.575±0.019	1.573±0.019	1.578±0.017	1.579±0.014	1.575 ±0.018	1.576±0.017	1.580±0.016	1.579±0.015	1.575±0.017	×1
R6	9.764±0.336	9.325±0.430	9.308±0.355	9.278±0.330	9.279±0.344	9.261 ±0.336	9.271±0.338	9.310±0.351	9.272±0.353	9.268±0.348	×0.1
R7	3.375±0.167	3.341 ±0.162	3.347±0.156	3.397±0.246	3.377±0.153	3.393±0.186	3.372±0.197	3.383±0.198	3.413±0.249	3.384±0.204	×1
R8	6.540±0.114	6.534±0.073	6.524±0.108	6.514±0.116	6.518±0.104	6.515±0.123	6.510±0.115	6.522±0.113	6.500±0.107	6.480 ±0.098	×0.1
R9	9.853±0.152	9.826±0.141	9.823±0.117	9.812±0.140	9.810±0.138	9.808±0.139	9.808±0.142	9.829±0.108	9.809±0.142	9.805 ±0.141	×0.1
R10	2.245±0.042	2.249±0.043	2.249±0.042	2.248±0.043	2.246±0.043	2.245±0.042	2.246±0.043	2.245±0.043	2.244 ±0.042	2.245±0.042	×1
R11	7.022±0.117	7.187±0.245	7.077±0.144	7.032±0.115	7.020 ±0.120	7.024±0.122	7.028±0.120	7.030±0.119	7.021±0.119	7.024±0.120	×1
R12	3.169±0.042	3.172±0.041	3.169±0.041	3.168±0.041	3.172±0.040	3.167±0.041	3.166±0.041	3.169±0.042	3.166 ±0.041	3.166±0.041	×1
R13	4.742±0.107	4.712±0.087	4.692±0.171	4.767±0.119	4.637±0.158	4.677±0.156	4.619 ±0.126	4.681±0.157	4.621±0.139	4.645±0.147	×1
R14	2.971±0.150	2.999±0.139	3.001±0.152	3.066±0.187	3.003±0.123	2.981±0.155	2.970 ±0.150	3.027±0.159	2.979±0.143	2.994±0.138	×1
R15	4.277±0.051	4.279±0.054	4.271±0.054	4.277±0.052	4.269±0.052	4.267±0.054	4.268±0.052	4.271±0.052	4.267±0.053	4.266 ±0.055	×0.01
R16	1.159±0.011	1.130±0.042	1.128±0.048	1.102±0.015	1.128±0.016	1.103±0.015	1.103±0.015	1.116±0.035	1.103±0.016	1.101±0.016	×0.001
R17	1.741±0.016	1.739±0.019	1.738±0.018	1.734±0.017	1.734±0.016	1.733±0.017	1.733±0.017	1.736±0.019	1.736±0.019	1.733±0.018	×0.001
R18	2.850±0.038	2.930±0.099	2.900±0.096	2.878±0.041	2.872±0.045	2.844±0.038	2.849±0.043	2.881±0.066	2.839±0.035	2.834 ±0.038	×0.001
R19	1.275±0.011	1.278±0.009	1.273±0.014	1.270±0.013	1.275±0.011	1.266±0.012	1.267±0.013	1.266±0.012	1.265±0.013	1.265±0.012	×100
R20	6.460±0.054	6.522±0.111	6.482±0.052	6.473±0.055	6.464±0.057	6.460±0.060	6.461±0.059	6.462±0.058	6.458 ±0.058	6.460±0.058	×0.1

Table 13: Average RMSE over different XGB-Reg tasks (5-fold CV).

Data Set	Rand	LHS	Sobol	UD	SeqRand	TPE	SMAC	GP-EI	SeqUD	scale
R1	5.127±0.141	5.141±0.101	5.086±0.113	5.040±0.110	5.066±0.131	5.020±0.128	4.992 ±0.140	5.090±0.196	5.049±0.149	×0.1
R2	7.235±0.201	7.243±0.232	7.262±0.197	7.255±0.213	7.171±0.188	7.206±0.210	7.199±0.219	7.216±0.227	7.167 ±0.198	×0.1
R3	2.304±0.097	2.304±0.098	2.302±0.098	2.304±0.098	2.303±0.097	2.301±0.099	2.302±0.098	2.307±0.099	2.300 ±0.099	×10000
R4	7.126±0.280	7.153±0.272	7.121±0.269	7.156±0.283	6.951±0.271	6.961±0.283	7.030±0.399	6.940 ±0.240	6.940±0.240	×100
R5	1.570±0.019	1.573±0.021	1.569±0.019	1.569±0.019	1.569±0.019	1.570±0.019	1.569±0.019	1.573±0.019	1.569±0.019	×1
R6	2.592±0.157	2.637±0.175	2.580±0.123	2.592±0.148	2.464±0.136	2.488±0.100	2.522±0.137	2.442 ±0.120	2.475±0.099	×0.1
R7	2.224±0.153	2.257±0.157	2.145±0.108	2.144±0.103	2.122±0.143	2.137±0.134	2.110±0.128	2.077±0.160	2.070 ±0.122	×1
R8	6.259±0.106	6.252±0.105	6.233±0.093	6.223±0.105	6.210±0.099	6.209±0.092	6.194±0.108	6.193±0.105	6.192 ±0.098	×0.1
R9	9.304±0.126	9.317±0.144	9.284±0.131	9.283±0.146	9.276±0.135	9.262±0.145	9.280±0.135	9.327±0.233	9.248 ±0.136	×0.1
R10	2.163±0.028	2.163±0.024	2.161±0.026	2.151±0.028	2.158±0.026	2.151±0.027	2.156±0.033	2.169±0.055	2.143 ±0.030	×1
R11	2.547±0.215	2.483±0.146	2.282±0.069	2.309±0.078	2.346±0.139	2.189±0.071	2.180±0.093	2.100 ±0.051	2.159±0.117	×1
R12	3.156±0.033	3.149±0.032	3.143±0.034	3.162±0.030	3.136±0.034	3.135±0.032	3.130±0.031	3.148±0.039	3.124 ±0.035	×1
R13	2.806±0.049	2.810±0.087	2.778±0.041	2.808±0.064	2.770±0.058	2.773±0.041	2.768 ±0.059	2.960±0.426	2.783±0.052	×1
R14	2.885±0.136	2.885±0.135	2.885±0.137	2.873±0.132	2.878±0.132	2.877±0.136	2.880±0.138	2.955±0.154	2.871 ±0.135	×0.01
R15	3.478±0.054	3.529±0.085	3.485±0.067	3.410±0.076	3.426±0.105	3.407±0.086	3.416±0.067	3.389±0.098	3.379 ±0.080	×1
R16	9.787±0.402	9.901±0.334	9.476±0.091	9.607±0.079	9.405±0.305	9.333±0.194	9.238±0.347	8.908 ±0.221	9.298±0.155	×0.001
R17	0.163±0.002	0.163±0.002	0.162±0.002	0.163±0.002	0.162±0.002	0.161 ±0.002	0.162±0.002	0.162±0.002	0.162±0.002	×0.001
R18	2.369±0.053	2.398±0.058	2.312±0.031	2.393±0.028	2.311±0.044	2.302±0.037	2.352±0.098	2.276±0.029	2.268 ±0.034	×0.001
R19	4.379±0.113	4.370±0.097	4.276±0.040	4.291±0.029	4.292±0.072	4.245 ±0.035	4.279±0.062	4.360±0.261	4.295±0.036	×10
R20	4.730±0.058	4.739±0.077	4.754±0.027	4.734±0.030	4.718±0.040	4.669±0.029	4.658±0.045	4.635 ±0.051	4.725±0.039	×0.1

Table 14: Average RMSE over different XGB-Reg tasks (test set).

Data Set	Rand	LHS	Sobol	UD	SeqRand	TPE	SMAC	GP-EI	SeqUD	scale
R1	4.950±0.209	4.981±0.183	4.880±0.206	4.891±0.177	4.880±0.164	4.885±0.237	4.904±0.243	5.067±0.290	4.829 ±0.184	×0.1
R2	7.261±0.245	7.310±0.272	7.315±0.179	7.412±0.222	7.297±0.276	7.319±0.241	7.260 ±0.217	7.353±0.266	7.283±0.253	×0.1
R3	2.423 ±0.105	2.429±0.101	2.424±0.107	2.433±0.094	2.426±0.106	2.427±0.107	2.424±0.106	2.427±0.097	2.424±0.105	×10000
R4	7.335±0.337	7.227±0.309	7.369±0.194	7.261±0.210	7.180±0.279	7.160±0.222	7.178±0.261	7.312±0.487	7.146 ±0.274	×100
R5	1.578±0.017	1.580±0.016	1.579±0.016	1.576±0.018	1.581±0.017	1.582±0.014	1.578±0.016	1.580±0.017	1.576±0.018	×1
R6	2.511±0.249	2.522±0.210	2.439±0.179	2.454±0.225	2.411 ±0.235	2.528±0.235	2.504±0.230	2.533±0.197	2.436±0.193	×0.1
R7	2.090±0.189	2.125±0.192	2.067±0.132	2.034±0.144	2.042±0.174	2.065±0.132	2.005±0.153	1.983±0.156	1.980 ±0.179	×1
R8	6.175±0.168	6.166±0.125	6.196±0.066	6.119 ±0.140	6.177±0.114	6.211±0.119	6.168±0.103	6.173±0.108	6.136±0.115	×0.1
R9	9.252±0.137	9.264±0.122	9.246±0.108	9.220±0.108	9.236±0.117	9.209 ±0.132	9.228±0.121	9.306±0.274	9.209±0.132	×0.1
R10	2.198±0.039	2.194±0.036	2.190±0.036	2.187±0.038	2.196±0.038	2.186±0.033	2.192±0.037	2.206±0.045	2.185 ±0.032	×1
R11	2.365±0.209	2.308±0.181	2.109±0.087	2.153±0.057	2.184±0.137	2.036±0.078	2.006±0.092	1.968 ±0.078	2.015±0.111	×1
R12	3.131±0.034	3.116±0.028	3.121±0.033	3.149±0.030	3.119±0.034	3.116±0.038	3.116±0.027	3.138±0.043	3.114 ±0.032	×1
R13	2.761±0.058	2.768±0.052	2.737 ±0.050	2.779±0.070	2.758±0.048	2.738±0.050	2.744±0.050	3.032±0.585	2.753±0.051	×1
R14	2.912±0.139	2.901±0.138	2.889±0.148	2.880 ±0.157	2.899±0.144	2.891±0.148	2.887±0.149	2.957±0.220	2.886±0.154	×0.01
R15	3.342±0.124	3.384±0.114	3.363±0.067	3.278±0.057	3.317±0.080	3.282±0.070	3.268 ±0.062	3.273±0.059	3.276±0.053	×1
R16	9.413±0.420	9.474±0.279	9.221±0.165	9.077±0.175	9.086±0.365	8.985±0.265	8.807±0.401	8.526 ±0.229	9.042±0.195	×0.001
R17	1.609±0.015	1.610±0.019	1.599±0.017	1.618±0.018	1.596 ±0.021	1.602±0.015	1.606±0.019	1.834±0.265	1.603±0.014	×0.0001
R18	2.295±0.093	2.335±0.086	2.250±0.051	2.305±0.047	2.248±0.067	2.245±0.054	2.297±0.081	2.213 ±0.050	2.214±0.052	×0.001
R19	4.307±0.125	4.297±0.092	4.177±0.051	4.218±0.041	4.194±0.049	4.160 ±0.042	4.183±0.060	4.289±0.276	4.228±0.071	×10
R20	4.645±0.055	4.643±0.082	4.643±0.065	4.649±0.060	4.637±0.074	4.570±0.062	4.574±0.052	4.555 ±0.093	4.654±0.064	×0.1

Table 15: Average RMSE over different Pipe-Reg tasks (5-fold CV).

Data Set	Rand	LHS	Sobel	UD	SeqRand	TPE	SMAC	GP-EI	SeqUD	scale
R1	5.160±0.111	5.162±0.130	5.172±0.115	5.171±0.124	5.232±0.171	5.094 ±0.138	5.196±0.149	5.277±0.147	5.119±0.087	×0.1
R2	7.288±0.138	7.354±0.217	7.312±0.180	7.262±0.152	7.252±0.132	7.266±0.278	7.216 ±0.222	7.405±0.241	7.243±0.228	×0.1
R3	2.298±0.096	2.296 ±0.096	2.298±0.099	2.297±0.099	2.296±0.099	2.296±0.098	2.297±0.098	2.310±0.104	2.296±0.101	×10000
R4	7.500±0.413	7.509±0.381	7.324±0.255	7.465±0.434	7.546±0.515	7.537±0.271	7.537±0.271	8.346±0.767	6.905 ±0.265	×100
R5	1.565±0.020	1.566±0.020	1.565±0.020	1.565±0.020	1.565±0.020	1.566±0.020	1.565±0.020	1.568±0.020	1.564 ±0.020	×1
R6	2.623±0.208	2.595±0.218	2.673±0.132	2.638±0.180	2.536±0.199	2.478 ±0.161	2.567±0.105	5.760±2.329	2.484±0.143	×0.1
R7	2.313±0.158	<u>2.333</u> ±0.252	2.185 ±0.116	<u>2.614</u> ±0.243	<u>2.414</u> ±0.374	2.279±0.308	2.565±0.263	<u>3.248</u> ±0.317	<u>2.278</u> ±0.108	×1
R8	6.323±0.115	6.329±0.093	6.357±0.116	6.268±0.085	6.279±0.104	6.252±0.132	6.288±0.129	6.448±0.208	6.239 ±0.110	×0.1
R9	9.481±0.172	9.471±0.237	9.369±0.135	9.363±0.163	9.381±0.215	9.360±0.201	9.361±0.138	9.637±0.249	9.255 ±0.134	×0.1
R10	2.169±0.043	2.179±0.034	2.145 ±0.040	2.165±0.023	2.154±0.031	2.159±0.037	2.175±0.028	2.246±0.211	2.194±0.039	×1
R11	2.901±0.375	2.802±0.351	2.340±0.106	3.111±0.390	2.647±0.394	2.294±0.187	2.868±0.486	3.272±1.977	2.241 ±0.070	×1
R12	3.177±0.043	3.182±0.043	3.186±0.033	3.163±0.057	3.220±0.125	3.153±0.036	3.158±0.033	3.207±0.070	3.125 ±0.033	×1
R13	2.921±0.217	2.899±0.176	2.798±0.046	2.935±0.072	2.821±0.079	3.007±0.656	3.042±0.627	4.808±1.328	2.783 ±0.054	×1
R14	2.898±0.129	2.907±0.137	2.894±0.131	2.899±0.136	2.893±0.141	2.886±0.139	2.896±0.131	3.025±0.198	2.882 ±0.134	×0.01
R15	3.606±0.099	3.602±0.206	3.527±0.076	3.645±0.138	3.743±0.380	3.518±0.288	3.599±0.231	4.120±0.395	3.437 ±0.070	×1
R16	1.051±0.049	0.992±0.062	0.988±0.012	0.990±0.050	1.030±0.078	0.966±0.045	1.010±0.069	0.928 ±0.067	1.094±0.055	×0.01
R17	1.669±0.056	1.665±0.066	1.645±0.026	1.642±0.034	1.653±0.061	1.630±0.038	1.662±0.063	1.864±0.507	1.615 ±0.019	×0.0001
R18	2.579±0.297	2.366±0.114	2.392±0.031	2.660±0.179	2.609±0.749	2.283±0.047	2.502±0.251	2.795±0.292	2.197 ±0.030	×0.001
R19	0.496±0.058	0.503±0.125	0.434±0.004	0.501±0.027	0.499±0.134	0.492±0.012	0.504±0.123	1.051±0.237	0.425 ±0.006	×100
R20	4.833±0.117	4.866±0.143	4.751±0.035	4.874±0.075	4.975±0.469	4.705 ±0.035	4.780±0.109	5.823±0.602	4.718±0.046	×0.1

Table 16: Average RMSE over different Pipe-Reg tasks (test set).

Data Set	Rand	LHS	Sobel	UD	SeqRand	TPE	SMAC	GP-EI	SeqUD	scale
R1	5.019±0.246	4.979±0.253	4.973 ±0.203	5.060±0.162	5.184±0.214	5.073±0.282	5.025±0.223	5.159±0.187	5.066±0.245	×0.1
R2	7.467±0.256	7.458±0.253	7.500±0.201	7.381±0.269	7.459±0.293	7.483±0.157	7.347 ±0.223	7.592±0.243	7.422±0.218	×0.1
R3	2.442±0.104	2.445±0.103	2.431 ±0.111	2.434±0.111	2.440±0.106	2.444±0.106	2.444±0.111	2.441±0.119	2.438±0.116	×10000
R4	7.590±0.434	7.685±0.407	7.408±0.300	7.718±0.416	7.776±0.542	7.538±0.509	7.737±0.490	8.620±0.729	7.142 ±0.209	×100
R5	1.580±0.016	1.576 ±0.016	1.579±0.017	1.582±0.015	1.578±0.015	1.577±0.017	1.579±0.015	1.581±0.018	1.576±0.018	×1
R6	2.561±0.255	2.607±0.094	2.564±0.288	2.587±0.220	2.582±0.248	2.497±0.222	2.565±0.255	5.622±2.440	2.449 ±0.181	×0.1
R7	2.174±0.235	2.224±0.211	2.060 ±0.167	2.488±0.244	2.336±0.448	2.207±0.315	2.451±0.440	3.512±0.806	2.175±0.164	×1
R8	6.246±0.104	6.234±0.142	6.281±0.178	6.195 ±0.096	6.231±0.176	6.243±0.133	6.311±0.196	6.394±0.239	6.276±0.170	×0.1
R9	0.941±0.022	0.943±0.014	0.930±0.011	0.927±0.010	0.937±0.030	0.933±0.023	0.931±0.010	1.060±0.298	0.923 ±0.011	×1
R10	2.190±0.029	2.214±0.035	2.160 ±0.035	2.195±0.035	2.180±0.043	2.207±0.054	2.210±0.040	2.270±0.159	2.210±0.033	×1
R11	2.746±0.418	2.618±0.329	2.178±0.093	2.947±0.360	2.469±0.372	2.154±0.210	2.695±0.513	3.218±1.998	2.054 ±0.122	×1
R12	3.155±0.053	3.152±0.057	3.171±0.051	3.160±0.044	3.195±0.071	3.137±0.034	3.148±0.043	3.188±0.089	3.119 ±0.036	×1
R13	2.906±0.243	2.825±0.145	2.745 ±0.049	2.875±0.092	2.793±0.072	2.954±0.609	2.963±0.583	4.732±1.308	2.748±0.030	×1
R14	2.900±0.158	2.940±0.147	2.901±0.154	2.917±0.162	2.918±0.135	2.966±0.166	2.913±0.161	3.010±0.257	2.891 ±0.151	×0.01
R15	3.485±0.141	3.482±0.218	3.390±0.046	3.536±0.114	3.642±0.448	3.414±0.262	3.467±0.247	4.069±0.407	3.293 ±0.066	×1
R16	1.006±0.051	0.957±0.072	0.941±0.013	0.960±0.043	0.939±0.071	0.929±0.045	0.970±0.075	0.888 ±0.084	1.067±0.050	×0.01
R17	1.654±0.054	1.642±0.061	1.614±0.016	1.624±0.026	1.639±0.071	1.617±0.051	1.644±0.057	1.858±0.528	1.604 ±0.018	×0.0001
R18	2.525±0.288	2.312±0.110	2.308±0.054	2.591±0.236	2.600±0.821	2.243±0.072	2.423±0.282	2.763±0.263	2.173 ±0.031	×0.001
R19	0.489±0.059	0.496±0.131	0.425±0.003	0.489±0.027	0.494±0.139	0.422±0.014	0.499±0.132	1.048±0.243	0.419 ±0.005	×100
R20	4.726±0.109	4.793±0.175	4.634±0.050	4.798±0.106	4.900±0.500	4.624 ±0.068	4.685±0.140	5.774±0.611	4.650±0.109	×0.1

Table 17: Average Accuracy (%) over different SVM-Cls tasks (5-fold CV).

Data Set	Grid	Rand	LHS	Sobol	UD	SeqRand	TPE	SMAC	GP-EI	SeqUD
C1	98.17±0.74	98.24±0.76	98.27±0.68	98.24±0.72	98.31±0.67	98.31±0.72	98.31±0.72	98.24±0.74	98.41±0.68	98.31±0.77
C2	72.48±0.88	72.24±0.77	72.10±0.58	72.24±0.82	72.31±0.83	72.45±1.10	72.31±1.15	72.48±1.15	72.20±0.88	72.20±1.18
C3	86.41±1.09	86.35±1.05	86.49±0.99	86.43±0.88	86.17±1.13	86.49±0.98	86.61±0.94	86.55±0.98	86.43±1.04	86.96±1.06
C4	97.17±0.66	97.17±0.71	97.25±0.64	97.22±0.70	97.28±0.68	97.28±0.65	97.17±0.70	97.22±0.70	97.34±0.62	97.34±0.62
C5	77.76±1.40	78.04±1.30	77.97±1.33	78.17±1.46	78.07±1.41	78.20±1.45	78.28±1.30	77.92±1.37	78.20±1.28	78.33±1.66
C6	99.02±0.16	99.16±0.19	99.16±0.19	99.19±0.19	99.17±0.19	99.21±0.19	99.21±0.16	99.14±0.15	99.19±0.15	99.19±0.17
C7	76.38±1.27	76.60±1.25	76.40±1.39	76.40±0.99	76.70±1.07	76.86±1.12	76.66±1.15	76.74±1.14	76.94±1.22	76.82±1.19
C8	93.72±0.31	93.74±0.33	93.72±0.33	93.76±0.30	93.79±0.40	93.74±0.42	93.74±0.32	93.70±0.36	93.47±0.34	93.74±0.41
C9	93.48±0.07	93.45±0.05	93.51±0.12	93.48±0.08	93.53±0.12	93.44±0.05	93.53±0.13	93.49±0.13	93.47±0.12	93.43±0.02
C10	85.86±0.02	86.00±0.28	85.96±0.18	85.88±0.06	86.08±0.30	86.05±0.40	85.98±0.24	85.96±0.32	86.18±0.52	86.24±0.46
C11	90.38±0.33	90.26±0.36	90.35±0.27	90.23±0.34	90.35±0.29	90.44±0.31	90.41±0.32	90.38±0.30	90.49±0.31	90.45±0.30
C12	97.87±0.15	97.88±0.15	97.86±0.15	97.87±0.14	97.87±0.15	97.88±0.15	97.89±0.14	97.87±0.14	97.89±0.15	97.91±0.13
C13	98.62±0.14	98.71±0.16	98.74±0.15	98.73±0.16	98.72±0.13	98.77±0.16	98.76±0.15	98.74±0.14	98.75±0.16	98.78±0.16
C14	80.66±0.01	80.66±0.02	80.66±0.02	80.66±0.01	80.66±0.01	80.65±0.00	80.66±0.02	80.66±0.02	80.66±0.02	80.66±0.02
C15	55.69±0.45	55.17±0.07	55.27±0.27	55.12±0.01	55.13±0.03	55.23±0.19	55.29±0.34	55.31±0.33	55.67±0.49	55.12±0.01
C16	85.40±0.25	85.18±0.34	85.31±0.26	85.23±0.31	85.40±0.30	85.51±0.27	85.48±0.32	85.45±0.31	85.54±0.32	85.56±0.29
C17	82.03±0.15	82.91±0.12	82.91±0.18	82.94±0.15	82.93±0.14	82.99±0.17	82.91±0.14	82.91±0.14	83.01±0.14	83.01±0.16
C18	94.82±0.13	94.69±0.19	94.67±0.28	94.79±0.09	94.77±0.12	94.83±0.10	94.81±0.11	94.76±0.14	94.85±0.10	94.85±0.10
C19	89.32±0.06	89.31±0.06	89.31±0.06	89.31±0.06	89.31±0.06	89.32±0.06	89.31±0.06	89.32±0.06	89.32±0.06	89.33±0.05
C20	77.04±0.18	76.01±1.22	76.06±0.71	75.99±0.26	75.78±0.21	77.06±0.33	76.88±0.24	76.99±0.27	77.33±0.16	77.28±0.16

Table 18: Average Accuracy (%) over different SVM-Cls tasks (test set).

Data Set	Grid	Rand	LHS	Sobol	UD	SeqRand	TPE	SMAC	GP-EI	SeqUD
C1	96.74±0.57	96.81±0.64	96.70±0.55	96.95±0.63	96.88±0.46	96.67±1.16	96.84±0.61	97.02±0.55	96.77±0.60	97.02±0.50
C2	69.25±2.70	70.21±1.97	69.69±2.40	69.90±2.27	69.45±2.21	70.10±2.19	69.59±2.15	69.35±2.51	69.59±2.40	70.38±2.10
C3	85.91±1.24	85.83±1.39	86.26±1.52	87.13±1.42	85.77±1.09	85.83±1.28	85.88±1.09	85.71±1.06	85.59±1.05	86.20±1.16
C4	96.37±0.60	96.37±0.77	96.49±0.91	96.26±0.66	96.57±0.76	96.46±0.84	96.37±0.77	96.60±0.74	96.46±0.88	96.31±0.81
C5	76.59±1.64	76.17±1.48	75.89±1.83	76.17±1.46	76.25±1.69	76.22±1.76	76.35±1.74	76.25±1.50	76.25±1.52	76.35±1.45
C6	99.31±0.31	99.33±0.33	99.46±0.23	99.46±0.21	99.46±0.23	99.46±0.21	99.42±0.20	99.46±0.23	99.44±0.30	99.46±0.28
C7	74.84±1.29	75.18±1.06	74.66±0.93	74.94±1.17	74.78±1.13	74.54±1.57	74.92±1.15	74.80±1.27	74.56±1.95	75.18±1.16
C8	93.19±0.24	93.14±0.26	93.10±0.23	93.21±0.24	93.33±0.10	93.06±0.20	93.15±0.29	93.17±0.22	93.08±0.18	93.19±0.23
C9	93.20±0.42	93.30±0.21	93.30±0.17	93.32±0.17	93.32±0.17	93.39±0.07	93.25±0.21	93.36±0.11	93.41±0.11	93.39±0.09
C10	85.86±0.02	85.96±0.23	86.13±0.56	85.86±0.02	86.19±0.62	86.18±0.71	86.00±0.29	85.98±0.35	86.27±0.64	86.22±0.59
C11	90.52±0.46	90.42±0.50	90.48±0.36	90.43±0.37	90.48±0.38	90.50±0.30	90.50±0.37	90.50±0.37	90.51±0.32	90.52±0.32
C12	97.57±0.33	97.69±0.14	97.69±0.15	97.72±0.12	97.69±0.14	97.69±0.13	97.63±0.23	97.70±0.14	97.66±0.14	97.68±0.11
C13	98.52±0.14	98.64±0.18	98.66±0.16	98.65±0.14	98.67±0.14	98.67±0.16	98.65±0.12	98.65±0.19	98.66±0.17	98.67±0.15
C14	80.66±0.02	80.66±0.02	80.66±0.03	80.66±0.02	80.66±0.02	80.65±0.00	80.66±0.02	80.66±0.01	80.66±0.01	80.66±0.02
C15	55.98±0.77	55.17±0.12	55.32±0.39	55.12±0.01	55.12±0.01	55.35±0.44	55.34±0.48	55.37±0.46	56.22±0.88	55.12±0.01
C16	85.59±0.26	85.55±0.26	85.54±0.39	85.48±0.23	85.65±0.27	85.66±0.30	85.62±0.23	85.68±0.30	85.72±0.28	85.64±0.24
C17	83.05±0.23	83.01±0.22	82.99±0.21	83.04±0.24	83.04±0.24	83.07±0.24	83.02±0.27	83.00±0.25	83.04±0.23	83.05±0.17
C18	94.93±0.11	94.77±0.18	94.77±0.22	94.88±0.12	94.80±0.12	94.87±0.12	94.88±0.12	94.81±0.16	94.88±0.11	94.87±0.12
C19	89.26±0.06	89.26±0.06	89.26±0.06	89.26±0.06	89.26±0.06	89.27±0.06	89.26±0.06	89.26±0.06	89.26±0.06	89.26±0.06
C20	77.12±0.23	76.11±1.22	76.41±0.45	76.35±0.32	75.94±0.29	77.05±0.32	76.98±0.36	77.06±0.26	77.29±0.20	77.24±0.23

Table 19: Average Accuracy (%) over different XGB-Cls tasks (5-fold CV).

Data Set	Rand	LHS	Sobol	UD	SeqRand	TPE	SMAC	GP-EI	SeqUD
C1	97.22±0.60	97.33±1.10	97.57±0.62	97.89±0.67	97.68±0.59	97.22±0.68	97.71±0.60	96.86±0.86	97.78±0.74
C2	73.41±1.12	73.55±1.19	73.58±1.06	73.44±1.33	73.34±1.51	73.75±1.02	73.95±1.65	73.44±1.55	73.68±1.08
C3	86.96±0.79	87.33±1.00	87.45±0.88	87.28±0.96	87.54±0.84	87.68±0.83	87.54±0.93	87.42±1.07	87.54±1.08
C4	97.14±0.52	97.22±0.48	97.36±0.49	97.31±0.51	97.19±0.51	97.22±0.63	97.22±0.54	97.16±0.52	97.31±0.46
C5	77.50±2.00	77.29±1.92	77.55±1.56	77.42±1.81	77.53±1.96	77.50±1.68	77.42±1.68	77.63±1.72	77.87±1.61
C6	98.56±0.24	98.41±0.35	98.48±0.31	98.48±0.23	98.60±0.34	98.68±0.27	98.64±0.33	98.60±0.25	98.73±0.36
C7	76.42±1.69	76.34±1.59	76.22±1.44	76.56±1.42	76.70±1.49	76.80±1.52	76.68±1.60	76.86±1.72	76.88±1.38
C8	94.15±0.39	94.21±0.39	94.12±0.33	94.17±0.40	94.19±0.35	94.26±0.40	94.39±0.38	94.19±0.43	94.23±0.33
C9	93.45±0.05	93.48±0.08	93.48±0.05	93.49±0.07	93.49±0.10	93.48±0.08	93.48±0.09	93.48±0.12	93.48±0.07
C10	95.50±0.32	95.52±0.34	95.50±0.34	95.55±0.32	95.63±0.31	95.65±0.35	95.67±0.35	95.73±0.31	95.72±0.36
C11	73.49±0.71	73.50±0.63	73.61±0.63	73.52±0.75	73.72±0.73	73.85±0.71	73.86±0.68	89.17±1.55	75.00±4.68
C12	97.72±0.15	97.71±0.14	97.73±0.16	97.74±0.14	97.77±0.14	97.74±0.18	97.66±0.27	97.39±0.18	97.75±0.15
C13	97.76±0.22	97.80±0.16	97.79±0.16	97.80±0.18	97.89±0.19	97.94±0.24	97.92±0.19	97.94±0.20	97.88±0.21
C14	81.50±0.13	81.54±0.14	81.59±0.12	81.54±0.17	81.62±0.14	81.64±0.19	81.67±0.16	81.58±0.17	81.63±0.17
C15	92.30±0.32	92.33±0.25	92.51±0.21	92.40±0.23	92.50±0.29	92.70±0.21	92.74±0.23	92.85±0.31	92.82±0.19
C16	87.82±0.19	87.75±0.14	87.83±0.20	87.80±0.22	87.88±0.17	87.89±0.19	87.87±0.17	87.99±0.20	87.98±0.19
C17	87.02±0.12	87.00±0.18	87.04±0.16	87.00±0.13	87.08±0.15	87.06±0.14	87.10±0.14	87.14±0.12	87.08±0.15
C18	96.92±0.15	96.91±0.15	96.92±0.15	96.89±0.12	96.96±0.13	96.96±0.11	96.95±0.12	96.95±0.16	96.99±0.14
C19	90.80±0.10	90.77±0.08	90.79±0.09	90.77±0.10	90.84±0.09	90.83±0.10	90.88±0.11	90.89±0.09	90.88±0.10
C20	91.72±0.19	91.59±0.21	91.76±0.11	91.70±0.15	92.01±0.17	92.03±0.18	92.03±0.22	92.18±0.11	92.17±0.11

Table 20: Average Accuracy (%) over different XGB-Cls tasks (test set).

Data Set	Rand	LHS	Sobol	UD	SeqRand	TPE	SMAC	GP-EI	SeqUD
C1	96.11±1.08	96.39±0.80	96.70±0.89	96.67±0.83	96.35±0.74	96.11±1.02	96.33±0.77	95.33±1.22	96.28±0.90
C2	69.55±2.40	69.01±2.58	69.52±2.22	69.97±1.37	69.69±2.36	69.76±1.43	69.14±2.05	69.83±1.87	69.28±2.14
C3	86.72±1.38	86.75±1.06	86.55±1.35	86.43±1.73	86.12±1.21	86.64±1.49	85.97±1.47	86.41±1.72	86.52±1.49
C4	96.20±0.84	96.57±0.73	96.51±0.73	96.46±0.90	95.97±0.77	96.26±0.58	96.31±0.55	95.97±0.73	96.63±0.54
C5	76.48±1.38	75.91±1.30	76.15±1.62	75.49±1.69	75.86±1.42	75.68±1.70	75.68±1.39	75.70±1.80	75.62±1.29
C6	98.73±0.51	98.52±0.63	98.79±0.35	98.60±0.40	98.75±0.37	98.83±0.67	98.91±0.54	98.50±0.65	98.81±0.41
C7	75.36±1.62	75.08±1.32	75.12±1.51	74.88±1.31	74.80±1.22	75.62±1.93	74.84±1.53	74.74±0.97	74.48±1.14
C8	93.19±0.53	92.77±0.78	92.95±0.52	93.03±0.37	92.90±0.52	93.10±0.48	92.97±0.51	92.94±0.53	93.01±0.65
C9	93.27±0.20	93.36±0.11	93.25±0.18	93.30±0.17	93.37±0.13	93.30±0.17	93.34±0.11	93.35±0.15	93.36±0.11
C10	95.47±0.38	95.44±0.41	95.49±0.33	95.58±0.34	95.46±0.34	95.55±0.29	95.53±0.40	95.54±0.29	95.50±0.23
C11	74.31±0.79	73.90±0.94	74.11±1.13	74.18±1.06	74.09±1.06	74.31±1.21	74.26±1.15	89.06±0.92	75.29±4.90
C12	92.83±14.27	97.55±0.21	92.86±14.30	97.61±0.12	97.53±0.23	97.47±0.26	97.18±0.35	97.62±0.10	97.75±0.28
C13	97.66±0.15	97.75±0.24	97.69±0.21	97.68±0.07	97.77±0.26	97.86±0.18	97.92±0.15	97.81±0.33	97.75±0.28
C14	81.22±0.19	80.99±0.39	81.21±0.22	81.20±0.23	81.18±0.20	81.27±0.24	81.28±0.17	81.19±0.25	81.28±0.22
C15	93.33±0.22	93.23±0.39	93.30±0.26	93.05±0.44	93.27±0.39	93.37±0.42	93.48±0.25	93.43±0.39	93.53±0.27
C16	87.85±0.35	87.80±0.30	87.86±0.19	87.94±0.19	87.79±0.27	87.87±0.17	87.86±0.24	87.89±0.28	87.96±0.16
C17	87.09±0.17	87.13±0.15	87.15±0.15	87.16±0.22	87.17±0.13	87.08±0.14	87.12±0.16	87.10±0.16	87.10±0.16
C18	97.06±0.15	97.04±0.16	97.05±0.17	97.03±0.13	97.05±0.13	97.04±0.19	97.08±0.16	97.08±0.17	97.09±0.17
C19	90.75±0.10	90.68±0.17	90.77±0.15	90.68±0.13	90.71±0.14	90.78±0.13	90.76±0.15	90.76±0.10	90.76±0.10
C20	92.54±0.29	92.59±0.31	92.57±0.15	92.76±0.13	92.78±0.25	92.94±0.16	92.86±0.14	93.00±0.12	92.99±0.23

Table 21: Average Accuracy (%) over different Pipe-Cls tasks (5-fold CV).

Data Set	Rand	LHS	Sobol	UD	SeqRand	TPE	SMAC	GP-EI	SeqUD
C1	98.06±0.74	98.13±0.67	97.99±0.76	98.20±0.87	98.13±0.72	98.17±0.61	98.20±0.78	98.24±0.69	98.24±0.69
C2	73.41±0.93	73.40±0.81	74.09±1.64	73.24±0.79	73.78±0.95	74.03±0.97	74.33±1.05	73.61±1.64	74.44±1.87
C3	86.93±0.97	86.99±0.83	87.28±1.02	86.99±0.97	87.28±1.07	87.57±0.83	87.04±1.11	87.16±1.14	87.48±0.91
C4	97.48±0.58	97.42±0.56	97.45±0.60	97.42±0.56	97.54±0.57	97.45±0.48	97.51±0.42	97.42±0.46	97.48±0.51
C5	77.99±1.40	78.10±1.43	78.10±1.36	78.33±1.50	78.54±1.46	78.36±1.62	78.25±1.26	78.18±1.49	78.46±1.77
C6	99.15±0.47	99.00±0.53	98.56±0.33	98.98±0.29	99.48±0.45	99.54±0.36	99.37±0.44	98.83±0.44	99.10±0.25
C7	76.62±1.37	76.72±1.08	76.90±1.16	76.82±1.19	77.34±1.18	77.24±1.31	77.02±1.12	77.16±1.28	77.26±1.23
C8	94.19±0.35	94.14±0.36	94.21±0.35	94.28±0.48	94.21±0.35	94.26±0.35	94.26±0.28	94.13±0.54	94.28±0.31
C9	93.50±0.09	93.52±0.07	93.51±0.10	93.50±0.05	93.55±0.18	93.54±0.16	93.51±0.10	93.48±0.12	93.51±0.10
C10	95.34±0.49	95.34±0.30	95.44±0.35	95.48±0.31	95.69±0.37	95.67±0.31	95.61±0.33	95.45±0.62	95.66±0.36
C11	90.26±0.37	90.23±0.35	90.32±0.29	90.27±0.37	90.42±0.33	89.72±1.26	90.09±0.69	88.24±5.36	90.42±0.29
C12	97.89±0.14	97.90±0.15	97.91±0.15	97.89±0.14	97.91±0.16	97.94±0.13	97.91±0.16	97.90±0.15	97.94±0.16
C13	98.62±0.20	98.48±0.39	98.68±0.12	98.42±0.25	98.64±0.31	98.34±0.51	98.48±0.33	98.11±0.35	98.74±0.14
C14	81.54±0.16	81.56±0.18	81.53±0.13	81.53±0.12	81.54±0.18	81.65±0.17	81.56±0.15	81.56±0.28	81.62±0.11
C15	91.48±1.35	91.23±1.12	91.45±0.35	90.65±1.15	91.05±3.11	92.70±0.33	92.26±1.06	92.79±0.43	92.40±0.80
C16	87.51±0.30	87.66±0.29	87.72±0.21	87.64±0.19	87.91±0.21	87.92±0.15	87.85±0.19	87.58±0.72	87.93±0.18
C17	86.74±0.43	86.79±0.34	86.95±0.15	86.92±0.14	86.70±0.79	87.10±0.16	87.06±0.14	86.69±0.82	87.06±0.18
C18	96.71±0.13	96.62±0.58	96.56±0.13	96.74±0.15	96.91±0.12	96.93±0.13	96.91±0.12	96.79±0.33	96.95±0.14
C19	90.63±0.22	90.63±0.20	90.69±0.09	90.75±0.10	90.69±0.22	90.83±0.13	90.72±0.18	90.60±0.25	90.80±0.12
C20	89.69±2.19	89.85±2.70	89.97±0.46	89.82±1.51	90.52±3.25	91.79±0.38	91.68±0.56	91.75±0.60	91.88±0.10

Table 22: Average Accuracy (%) over different Pipe-Cls tasks (test set).

Data Set	Rand	LHS	Sobol	UD	SeqRand	TPE	SMAC	GP-EI	SeqUD
C1	96.60±0.54	96.77±0.68	96.18±1.65	97.02±0.55	96.63±0.98	96.67±0.65	96.74±0.92	96.67±1.28	96.95±0.63
C2	68.80±1.57	69.62±1.62	69.38±1.62	69.18±1.96	70.07±2.13	69.73±2.19	69.83±1.76	70.14±1.67	68.66±2.70
C3	85.77±1.58	86.20±1.31	86.35±1.77	86.03±1.75	86.20±0.99	85.91±1.79	85.83±1.31	86.38±1.69	85.83±1.50
C4	96.20±0.97	96.40±0.63	96.40±0.87	96.83±0.66	96.26±0.80	96.31±0.73	96.40±0.71	96.60±0.63	96.54±0.84
C5	76.38±1.20	75.68±2.12	76.38±1.58	76.59±1.55	75.70±2.00	75.99±2.14	76.17±1.26	75.89±1.42	76.04±1.24
C6	99.37±0.35	99.67±0.38	99.02±0.32	99.33±0.43	99.75±0.35	99.62±0.33	99.62±0.35	99.23±0.54	99.35±0.30
C7	75.08±0.96	75.20±1.07	74.34±1.98	73.76±1.42	74.68±1.02	74.84±1.36	73.94±1.77	74.44±1.54	74.72±1.60
C8	93.28±0.40	92.88±0.49	93.23±0.29	93.21±0.36	93.15±0.43	93.24±0.52	93.08±0.55	93.26±0.43	93.08±0.60
C9	93.13±0.33	93.11±0.26	93.27±0.23	93.31±0.12	93.32±0.19	93.29±0.15	93.29±0.30	93.32±0.20	93.40±0.08
C10	95.29±0.52	95.45±0.26	95.42±0.35	95.54±0.34	95.52±0.42	95.52±0.34	95.51±0.36	95.38±0.50	95.46±0.55
C11	90.31±0.40	90.43±0.33	90.39±0.46	90.20±0.40	90.38±0.46	89.83±1.62	90.20±0.46	88.61±4.36	90.52±0.29
C12	97.75±0.18	97.76±0.17	97.78±0.16	97.73±0.14	97.64±0.33	97.74±0.13	97.77±0.17	97.76±0.14	97.72±0.18
C13	98.51±0.26	98.41±0.22	98.60±0.14	98.34±0.30	98.52±0.39	98.23±0.51	98.37±0.35	97.92±0.49	98.62±0.16
C14	81.21±0.28	81.33±0.21	81.25±0.21	81.17±0.23	81.27±0.19	81.27±0.15	81.19±0.31	81.36±0.16	81.23±0.22
C15	92.25±1.39	92.15±1.60	92.21±0.26	91.55±1.42	91.79±3.04	93.46±0.42	93.06±1.17	93.57±0.47	92.87±1.05
C16	87.59±0.41	87.82±0.32	87.82±0.23	87.67±0.41	87.79±0.22	87.78±0.22	87.79±0.34	87.50±0.80	87.86±0.27
C17	86.84±0.45	86.96±0.43	87.00±0.19	87.06±0.18	86.75±0.90	87.14±0.17	87.14±0.16	86.75±0.72	87.09±0.15
C18	96.88±0.39	96.75±0.67	96.74±0.13	96.83±0.27	97.06±0.21	97.06±0.20	97.00±0.20	96.87±0.34	97.07±0.16
C19	90.61±0.21	90.62±0.26	90.69±0.15	90.78±0.09	90.55±0.22	90.79±0.10	90.62±0.18	90.42±0.26	90.70±0.10
C20	90.33±2.63	90.70±2.76	90.67±0.75	90.69±1.67	91.04±4.21	92.71±0.44	92.49±0.59	92.66±0.46	92.66±0.11

References

- Rémi Bardenet, Mátyás Brendel, Balázs Kégl, and Michele Sebag. Collaborative hyperparameter tuning. In *International Conference on Machine Learning*, pages 199–207, 2013.
- James Bergstra and Yoshua Bengio. Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, 13(10):281–305, 2012.
- James Bergstra, Brent Komer, Chris Eliasmith, Dan Yamins, and David D Cox. Hyperopt: a python library for model selection and hyperparameter optimization. *Computational Science & Discovery*, 8(1):014008, 2015.
- James S Bergstra, Rémi Bardenet, Yoshua Bengio, and Balázs Kégl. Algorithms for hyperparameter optimization. In *Advances in Neural Information Processing Systems*, pages 2546–2554, 2011.
- Felix Berkenkamp, Angela P Schoellig, and Andreas Krause. No-regret bayesian optimization with unknown hyperparameters. *Journal of Machine Learning Research*, 20(50):1–24, 2019.
- Chih-Chung Chang and Chih-Jen Lin. LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2(3):27, 2011.
- Karel Crombecq, Dirk Gorissen, Dirk Deschrijver, and Tom Dhaene. A novel hybrid sequential design strategy for global surrogate modeling of computer experiments. *SIAM Journal on Scientific Computing*, 33(4):1948–1974, 2011.
- Thanh Dai Nguyen, Sunil Gupta, Santu Rana, and Svetha Venkatesh. Stable bayesian optimization. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 578–591. Springer, 2017.
- Chiara Di Francescomarino, Marlon Dumas, Marco Federici, Chiara Ghidini, Fabrizio Maria Maggi, Williams Rizzi, and Luca Simonetto. Genetic algorithms for hyperparameter optimization in predictive business process monitoring. *Information Systems*, 74:67–83, 2018.
- Tobias Domhan, Jost Tobias Springenberg, and Frank Hutter. Speeding up automatic hyperparameter optimization of deep neural networks by extrapolation of learning curves. In *International Joint Conference on Artificial Intelligence*, pages 3460–3468, 2015.
- Delphine Dupuy, Céline Helbert, and Jessica Franco. Dicedesign and diceeval: Two r packages for design and analysis of computer experiments. *Journal of Statistical Software*, 65(11):1–38, 2015.
- Hugo Jair Escalante, Manuel Montes, and Luis Enrique Sucar. Particle swarm model selection. *Journal of Machine Learning Research*, 10(15):405–440, 2009.
- Kai-Tai Fang and Yuan Wang. A sequential algorithm for optimization and its applications to regression analysis. *Lecture Notes in Contemporary Mathematics*, pages 17–28, 1990.

- Kai-Tai Fang and Yuan Wang. *Number-Theoretic Methods in Statistics*, volume 51. Chapman & Hall, London., 1994.
- Kai-Tai Fang, Dennis KJ Lin, Peter Winker, and Yong Zhang. Uniform design: theory and application. *Technometrics*, 42(3):237–248, 2000.
- Kai-Tai Fang, Runze Li, and Agus Sudjianto. *Design and Modeling for Computer Experiments*. Chapman and Hall/CRC, 2006.
- Kai-Tai Fang, Min-Qian Liu, Hong Qin, and Yong-Dao Zhou. *Theory and Application of Uniform Experimental Designs*, volume 221 of *Lecture Notes in Statistics*. Springer, 2018.
- Matthias Feurer, Aaron Klein, Katharina Eggensperger, Jost Springenberg, Manuel Blum, and Frank Hutter. Efficient and robust automated machine learning. In *Advances in Neural Information Processing Systems*, pages 2962–2970, 2015.
- Eduardo C Garrido-Merchán and Daniel Hernández-Lobato. Dealing with categorical and integer-valued variables in bayesian optimization with gaussian processes. *Neurocomputing*, 380:20–35, 2020.
- Heikki Haario, Eero Saksman, and Johanna Tamminen. Adaptive proposal distribution for random walk metropolis algorithm. *Computational Statistics*, 14(3):375–396, 1999.
- Heikki Haario, Eero Saksman, and Johanna Tamminen. An adaptive metropolis algorithm. *Bernoulli*, 7(2):223–242, 2001.
- Fred Hickernell. A generalized discrepancy and quadrature error bound. *Mathematics of Computation*, 67(221):299–322, 1998.
- Chien-Ming Huang, Yuh-Jye Lee, Dennis KJ Lin, and Su-Yun Huang. Model selection for support vector machines via uniform design. *Computational Statistics & Data Analysis*, 52(1):335–346, 2007.
- Frank Hutter, Holger H Hoos, and Kevin Leyton-Brown. Sequential model-based optimization for general algorithm configuration. In *International Conference on Learning and Intelligent Optimization*, pages 507–523. Springer, 2011.
- Frank Hutter, Holger H Hoos, and Kevin Leyton-Brown. Parallel algorithm configuration. In *International Conference on Learning and Intelligent Optimization*, pages 55–70. Springer, 2012.
- Frank Hutter, Lars Kotthoff, and Joaquin Vanschoren. *Automated Machine Learning: Methods, Systems, Challenges*. Springer Nature, 2019.
- Ruichen Jin, Wei Chen, and Agus Sudjianto. An efficient algorithm for constructing optimal design of computer experiments. *Journal of Statistical Planning and Inference*, 134(1):268–287, 2005.
- Donald R Jones, Matthias Schonlau, and William J Welch. Efficient global optimization of expensive black-box functions. *Journal of Global Optimization*, 13(4):455–492, 1998.

- Kirthevasan Kandasamy, Jeff Schneider, and Barnabás Póczos. High dimensional bayesian optimisation and bandits via additive models. In *International Conference on Machine Learning*, pages 295–304, 2015.
- Q Ye Kenny, William Li, and Agus Sudjianto. Algorithmic construction of optimal symmetric latin hypercube designs. *Journal of Statistical Planning and Inference*, 90(1):145–159, 2000.
- Brent Komer, James Bergstra, and Chris Eliasmith. Hyperopt-sklearn: automatic hyperparameter configuration for scikit-learn. In *International Conference on Machine Learning Workshop on AutoML*, volume 9, page 50, 2014.
- Lars Kotthoff, Chris Thornton, Holger H Hoos, Frank Hutter, and Kevin Leyton-Brown. Auto-WEKA 2.0: automatic model selection and hyperparameter optimization in WEKA. *Journal of Machine Learning Research*, 18(1):826–830, 2017.
- Julien-Charles Lévesque. *Bayesian Hyperparameter Optimization: Overfitting, Ensembles and Conditional Spaces*. PhD dissertation, Université Laval, 2018.
- Lisha Li, Kevin Jamieson, Giulia DeSalvo, Afshin Rostamizadeh, and Ameet Talwalkar. Hyperband: A novel bandit-based approach to hyperparameter optimization. *Journal of Machine Learning Research*, 18(1):6765–6816, 2017.
- Timothy P Lillicrap, Jonathan J Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. Continuous control with deep reinforcement learning. *arXiv preprint: 1509.02971*, 2015.
- Michinari Momma and Kristin P Bennett. A pattern search method for model selection of support vector regression. In *Proceedings of the SIAM International Conference on Data Mining*, pages 261–274. SIAM, 2002.
- Harald Niederreiter. *Random Number Generation and Quasi-Monte Carlo Methods*. SIAM, 1992.
- Philipp Probst, Anne-Laure Boulesteix, and Bernd Bischl. Tunability: Importance of hyperparameters of machine learning algorithms. *Journal of Machine Learning Research*, 20(53):1–32, 2019.
- Robert J Renka and Ron Brown. Algorithm 792: accuracy test of acm algorithms for interpolation of scattered data in the plane. *ACM Transactions on Mathematical Software*, 25(1):78–94, 1999.
- Bobak Shahriari, Kevin Swersky, Ziyu Wang, Ryan P Adams, and Nando De Freitas. Taking the human out of the loop: A review of bayesian optimization. *Proceedings of the IEEE*, 104(1):148–175, 2016.
- Jasper Snoek, Hugo Larochelle, and Ryan P Adams. Practical bayesian optimization of machine learning algorithms. In *Advances in Neural Information Processing Systems*, pages 2951–2959, 2012.

- Ilya M Sobol. On quasi-monte carlo integrations. *Mathematics and Computers in Simulation*, 47(2-5):103–112, 1998.
- S. Surjanovic and D. Bingham. Virtual library of simulation experiments: Test functions and datasets. Retrieved August 9, 2020, from <http://www.sfu.ca/~ssurjano>, 2020.
- Kevin Swersky, Jasper Snoek, and Ryan P Adams. Multi-task bayesian optimization. In *Advances in Neural Information Processing Systems*, pages 2004–2012, 2013.
- Ziyu Wang, Masrour Zoghi, Frank Hutter, David Matheson, and Nando De Freitas. Bayesian optimization in high dimensions via random embeddings. In *International Joint Conference on Artificial Intelligence*, pages 1778–1784, 2013.
- Barret Zoph and Quoc V Le. Neural architecture search with reinforcement learning. *arXiv preprint: 1611.01578*, 2016.