Evaluation of Initial Sampling Strategies for Bayesian Optimization Tasks in Material Science

Zhiyang Zhao, Yang Pu, Suo Wang, xxxx, Hongqiagn Wang, Ruihao Yuan, Zhe Liu

**Abstract**

Bayesian optimization (BO) has demonstrated promising results in materials development tasks. At the outset of BO, initial sampling is crucial for gathering preliminary experimental data. However, the exploration of initial sampling strategies remains limited. In this study, we propose two initial sampling strategies: Euclidean distance sampling (EDS) and knowledge transfer sampling. EDS enables global uniform coverage, capturing comprehensive information about the optimization task, while transfer sampling quickly identifies the global optimal region by utilizing knowledge from related optimization tasks. We benchmark both methods on a test function and two material datasets: the cross-coupling reaction datasets and the high-temperature alloy datasets. We find that EDS achieved an average increase of 0.1 compared with conventional initial sampling in the Spearman correlation coefficient for the surrogate model, and demonstrated an increased likelihood of locating the global optimum. Additionally, BO with transfer sampling using relevant knowledge, compared to standard BO, accelerates the optimization rate by a factor of 3 or more in identifying the optimal solution.

# Introduction

Machine learning (ML) has recently emerged at the forefront of accelerated materials research. As ML becomes more widely accessible to materials scientists, its integration into experimental design has become increasingly commonplace. In materials experiment design, the need to optimize high-dimensional variables often results in large design spaces, making brute force or traditional methods impractical [1-4]. By introducing ML into the loop to guide material optimization tasks, it has been shown to effectively explore complex search spaces and adaptively avoid exhaustive enumeration [5, 6].

Bayesian Optimization (BO), a sequential machine learning approach, has gained widespread application in the exploration of chemical synthesis and materials especially when handling expensive black-box functions. Some recent research studies showed that ML-assisted experiment planning can overperform the conventional design of experiment methods based on the domain knowledge from human experts in many specific case studies, e.g., optimizing the reaction yield of chemical synthesis [7], discovering new photocatalysts for CO2 reduction [8], and searching for the best fast-charging protocol [9].

The BO framework comprises three components: initial sampling strategy, surrogate model, and acquisition function. To enhance the effectiveness and applicability of BO in materials development, current research on classical Bayesian optimization has increasingly focused on selecting surrogate models and adjusting acquisition functions [10, 11]. Liang *et al.* [12] evaluated the performance of BO across five materials science datasets, offering valuable insights into surrogate model selection for materials optimization. Liu *et al.* [13] incorporated previous experimental data as probabilistic constraints, combining subjective human observations with machine learning insights to enable the synergy between domain knowledge and BO. Building on traditional BO principles, Siemenn *et al.* [14] iteratively expanded the sampling search boundary, and customized adaptive acquisition functions to further guide new experiments toward a global optimum.

Yet, few have focused on determining initial sampling methods. Initial sampling provides foundational data for the surrogate model, helping it estimate the surface of the target black-box function, reducing prediction uncertainty, and enhancing the efficiency of the acquisition function [15]. In materials optimization, a well-designed initial sampling can cover the search space and explore the complexity of material features, thereby accelerating the identification of the global optimum, reducing experimental costs, and effectively utilizing limited resources. This makes initial sampling a crucial first step in BO [16].

At present, the predominant initial sampling methods employed in Bayesian Optimization are random sampling and Latin hypercube sampling (LHS) [17-19]. However, both approaches have their limitations: (1) fail to uniformly explore global information: Random sampling, due to its high Randomness, may lead to uneven sample distribution, while LHS operates by independently stratifying each dimension and then combining these strata into sampling points, which means that uniformity is only assured in individual dimensions. (2) unable to leverage relevant knowledge: Conventional initial sampling methods are suitable for scenarios where the target task is entirely isolated and unknown. However, human experts typically do not randomly select initial experiments; instead, they rely on prior experiences gained from previous optimization activities to guide the initial exploration of new tasks [20-24]. Therefore, if relevant data from previous optimization campaigns could be transferred to inform the sampling process of the target task, it would significantly enhance sampling efficiency and optimization results [25, 26].



Figure 1. Schematic illustration of uniform sampling and transfer sampling strategies. (A) The target function. (B) BO with uniform initial sampling. Uniform initial sampling collects comprehensive global information about the target function, allowing for the identification of the global optimum after a few iterations. (C) BO with knowledge transfer from a source function. Transfer sampling directly focuses on the likely region of the global optimum, achieving rapid convergence through minimal iterations.

In this work, we explore and discuss initial sampling strategies in two BO scenarios. (1) For scenarios lacking relevant data references, we propose a globally uniform initial sampling strategy, Euclidean distance sampling (EDS),. Compared to random sampling and Latin hypercube sampling (LHS), EDS enables uniform initial sampling across the global space, providing a general understanding of the target black-box function, thereby avoiding local optima and facilitating the discovery of the global optimum (Fig. 1b). (2) When relevant data are available, we employ a transfer learning strategy [27, 28] to transfer knowledge from the source function to the optimization task of the target black-box function, helping BO to directly locate the global optimum (Fig. 1c). It is noteworthy that the two methods described above are tailored to initial sampling strategies and function in parallel. In practical optimization scenarios, the choice of an appropriate initial sampling method should be informed by the specific context of data availability.

This paper is organized as follows: Section 2 introduces the proposed Euclidean distance sampling (EDS) and transfer sampling strategy; Section 3 illustrates these methods using the Goldstein-Price function [29]; Section 4 conducts simulated optimization and algorithm testing on two real material datasets—the Buchwald-Hartwig cross-coupling reaction dataset and the alloy creep life dataset. Finally, we provide evidence that the EDS strategy significantly enhances the stability of model performance when initializing the surrogate model, while reducing the risk of converging to local extrema during the BO process. The transfer sampling strategy enables the analysis of the correlation between the source model and the target task throughout the iterative process of BO. By dynamically adjusting the degree of knowledge transfer based on this correlation, this approach effectively assists in the optimization of the target task when the source task is relevant.

# Methodology

In this section, we explore two primary approaches: (1) the Euclidean distance sampling and (2) the transfer sampling strategy. These methods enable the adaptation of Bayesian optimization strategies in different scenarios, achieving more stable and efficient optimization.

We demonstrate the implementation of these methods on the Goldstein-Price function dataset, the Buchwald-Hartwig cross-coupling reaction dataset, and the alloy creep life dataset. For EDS, we compare it with LHS and Random sampling, examining the impact of different sampling methods on initial model construction and subsequent optimization. For transfer sampling, we compare it with standard Bayesian optimization, analyzing the performance of transfer BO using different source data against standard BO with no source data. Through testing on one function dataset and two materials datasets, we demonstrate the stable effectiveness and broad applicability of these two sampling strategies. In our BO framework, Gaussian Process Regression (GPR) is adopted as the surrogate model, with the Matérn 5/2 kernel function selected [30, 31]. Both the acquisition function following the initial sampling and the base acquisition function for transfer sampling are set to Expected Improvement (EI) [32]. The Python package Emukit is used for implementing BO, and the GPy package is used for Gaussian Process modeling [33].

## Euclidean distance sampling

To ensure uniformity in sampling across the search space, we propose a sampling strategy based on Euclidean distance, referred to as EDS. This strategy is implemented through the following steps:

1. Initialize sampling: For the parameter space X, randomly select 3 points as the initial sampling set .

2. Calculate minimum Euclidean distance: For each unsampled point in X, calculate the Euclidean distance between the point and each sampled point in . Then, obtain the minimum distance between the unsampled point and the sampled points:

|  |  |
| --- | --- |
|  | (1) |

3. Select the next sampling point: For all unsampled points in X, choose the one with the maximum minimum Euclidean distance as the next sampling point:

|  |  |
| --- | --- |
|  | (2) |

4. Update the sample set: Add to , and repeat steps 2 and 3 until the required number of sampling points is obtained.

## Knowledge transfer sampling

In situations where some related datasets are used as source data, we can leverage the knowledge from the source data to assist in sampling for the target optimization task. The difficulty of this process lies in the following aspects: a) How to determine the relevance between the source data and the target task? b) How to utilize the knowledge from the source data to assist in sampling for the target task? c) How to mitigate the impact of the source data during the sampling process as the search evolves?

In this research, we combined the Rank-weight Gaussian process ensemble (RGPE) and the Transfer Acquisition Function (TAF) method. This approach ensures relevance between the selected source data and the target task as the search progresses, automatically updates the weights of the source data, and achieves reliable transfer Bayesian optimization. The process is implemented through the following steps:

(1) Computing Ensemble Weights: If a model can predict the ranking of observation values according to their predicted values, it is useful for optimization. We construct a ranking loss to measure the consistency of the model's predicted ranking with the target ranking. Here,|| = > 1 denotes the dataset used to add weights to the target values in the new round of model fff. The model's ranking prediction loss is measured by the following ranking error loss:

|  |  |
| --- | --- |
|  | (3) |

where is an indicator function, and is the exclusive-or operation. The ranking-based loss assures that only the location of the optimum is important, as opposed to the actual values of the predictions. The loss for the target model is computed using leave-one-out cross validation, and is given by

|  |  |
| --- | --- |
|  | (4) |

where −k denotes a model fit to all target task observations except for the kth point. Now, the weights are assigned to each model according to the probability that it is the ensemble member with the lowest ranking loss. This probability is estimated by boostrapping S samples from the model predictions on the validation set, i.e. The weight is calculated by:

|  |  |
| --- | --- |
|  | (5) |

(2) Transfer acquisition function: The source model and target model both use Gaussian process models. By assigning the calculated weights to the acquisition functions of their respective models, a transfer acquisition function is obtained.

|  |  |
| --- | --- |
|  | (6) |

where

By calculating the transfer weight, we can guide the ranking optimization tasks to take further steps, thus realizing the transfer of tasks from source to target.

# Benchmarks

Here, we demonstrate the effectiveness of our method using the Goldstein-Price function (2-dimensional), a well-known benchmark function characterized by multiple local minima and one global minimum. To better mimic the conditions encountered in material data optimization, we generated a dataset of 2,500 function values by sampling each dimension at 50 intervals. The input variables (x) were normalized using min-max scaling, and the output values (y) were logarithmically transformed to prepare the data for subsequent experiments and evaluations.

## Euclidean distance sampling

To analyze the modeling and optimization performance of random, LHS, and EDS, we first extracted 20 samples from the dataset using different methods to build Gaussian Process Regression (GPR) models and visualized both the distribution of the sample points and the model's predictions of the target function. As depicted in Figure 2A, the "Truth" plot in the left corner represents the actual distribution of the objective function, while the other three plots show the predictions of Gaussian Process models constructed by sampling 20 points using random sampling, Latin hypercube sampling, and Euclidean distance sampling respectively. The model prediction based on EDS more accurately captures the general relationship of the target function, while the predictions based on random and LHS show larger deviations, particularly in regions where the target function exhibits high variability.

Following this, we incrementally increase the number of samples and compare the changes in the Spearman coefficient across different sampling methods. The Spearman coefficient, which ranges from -1 to 1, measures the strength and direction of monotonic relationships between variables. Figure 2B shows how the Spearman correlation coefficient of the models varies with the number of samples across different sampling strategies. The Spearman coefficient for EDS increases rapidly with the number of samples and surpasses that of random and LHS after 20 sample points, indicating that the SE strategy can build an accurate model more quickly. Figure 2C shows the distribution of Spearman coefficients for each sampling strategy after 20 samples and 40 repeated experiments. The distribution for SE is concentrated in a higher range, further validating its stable modeling capability.

Finally, we use 20 initial sampling points to establish the initial model, and based on this, we utilize the Expected Improvement (EI) acquisition function to perform BO. This approach allows us to investigate how different sampling methods influence the subsequent performance of BO. As the number of samples increases, the objective function values progressively approach the global optimum(Figure 2D). Notably, BO based on EDS shows significant optimization effects after just 10 EI sampling iterations and essentially finds the global optimum after around 55 iterations. In contrast, BO based on random and LHS requires more iterations to achieve similar results. Statistical analysis of BO optimization results from 40 independent experiments (Figure 2E) reveals that BO using random sampling or LHS sometimes gets stuck around function values of 3.6 and 4.4. In contrast, most of the optimal values obtained using EDS are concentrated below 1.5, indicating that EDS effectively reduces the risk of getting trapped in local optima.



Figure 2: Evaluation of sampling strategies on the Goldstein-Price function. (A) Distribution of sampled points using various strategies. Black crosses indicate sampled data points, while the contour plot presents the GP model predictions in two-dimensional space based on these samples. (B) Spearman correlation analysis of the GP model. This panel shows how data is continuously added from the three sampling strategies to the GP model, enabling observation of changes in the Spearman coefficient on the test set. (C) Histogram of Spearman coefficients when the sample size is 20. The histogram illustrates that the EDS-based model exhibits a more concentrated distribution of coefficients compared to the others. (D) BO optimization trajectories under different initial sampling methods. After acquiring 20 initial data points, the optimization process shifts to expected improvement (EI) for further sampling. (E) Distribution of optimal function values at the conclusion of BO optimization. The EDS-based BO method identifies current optimal values, all below 1.5. All experiments were conducted independently 40 times, with error bars in panels B and D representing the variance of the repeated results.

## Knowledge transfer sampling

To evaluate the effectiveness of transfer sampling methods in optimization, we designed various types of source functions based on the objective function and constructed corresponding source models to provide accessible prior knowledge. The contour plot of the source function is shown in Figure. 3A. Source function 1 is obtained by shifting both x1​ and x2​ of the target test function by 0.1 units and scaling the y-axis to 0.8. Source function 2 is the negative of the y-values of source function 1. Source function 3 is a funnel-shaped function we designed for knowledge transfer from other types of source functions. The blue region of the target function represents the global optimum area that we aim to explore through BO.

We then compared the optimization curves of standard BO without knowledge transfer and transfer BO with different prior knowledge(Figure. 3A). To introduce information from the source function early in BO, we first randomly sampled 3 points and then initiated their respective BO tasks. In each iteration, we add one new sample point. After 30 iterations, both transfer BO 1 and transfer BO 3 successfully identified the global optimum, indicating that transfer sampling effectively leveraged the knowledge from source functions 1 and 3 related to the global optimum, thereby aiding in optimizing the objective function. In contrast, the performance of transfer BO 2 was comparable to that of standard BO, suggesting that transfer sampling could disregard the misleading information from source function 2 and revert to standard BO behavior.

Additionally, we tracked the changes in the weights of different source models throughout the iterative process. Figure. 3C shows the weight variations of the three source models during the transfer BO process. In the early stages of optimization, the weight of source model 1 was greater than that of 3, while the weight of source model 2 remained close to zero, consistent with the conclusions drawn from panel b. As the iterations progressed, the weights of all source models gradually decreased, indicating that as more target data was acquired, the target model became increasingly refined, and the optimization of the target function progressively relied less on the guidance of the source models.



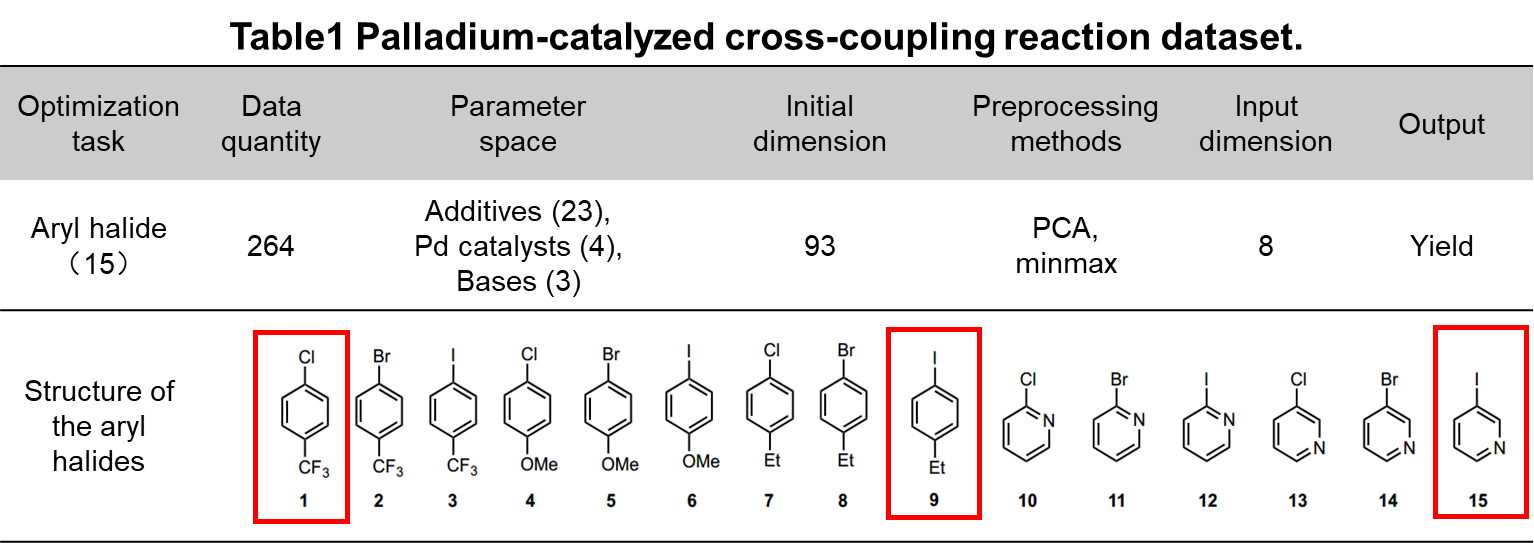
Figure 3: Comparison of standard Bayesian Optimization and transfer Bayesian Optimization with different source models. (A) Contour plots of the target function and three source functions used for transfer learning. The source functions are generated by different methods: Source 1 is produced by shifting and scaling the target function, Source 2 is an inversion of Source 1, and Source 3 is from a completely unrelated function. (B) Optimization trajectories of standard BO versus three transfer BO variants. The left panel compares the optimization paths, with the dashed black line indicating the global minimum of the target function. (C) Evolution of source model weights during the transfer BO process. This panel shows how the weights of the source models change, with higher weights indicating a greater contribution to the optimization.

# Experiments

We selected two real-world materials datasets to evaluate the performance of different sampling algorithms: the palladium-catalyzed aryl halide cross-coupling reaction dataset and the high-temperature alloy creep rupture life dataset. Both datasets are sourced from publicly available materials databases.

## Optimization of Simulated Cross-Coupling Reactions

The Pd-catalyzed Buchwald-Hartwig reaction holds significant value in pharmaceutical synthesis [34-38]. Ahneman et al. [39] conducted high-throughput experiments to obtain yield values for reaction combinations involving 15 different aryl halides. For each halide, the reaction space encompassed 23 additives, 4 palladium catalysts, and 3 bases, resulting in a total of 264 combinations (23 × 4 × 3 = 264). The original study characterized each option using 93-dimensional atomic, molecular, and vibrational descriptors. In our work, we applied min-max normalization and principal component analysis (PCA) for dimensionality reduction based on these original features.



We treated each subset of data as an optimization task, allowing the transfer of knowledge from other aryl halides when searching for the optimal reaction conditions for a specific aryl halide. We randomly selected the optimization tasks for aryl halides 1, 9, and 15 as the target tasks. For the exploration of initial sampling, proxy models were built using initial datasets of the same size obtained through different sampling methods. These models were then used to identify the experimental combination with the highest yield via Expected Improvement (EI). In terms of transfer sampling, data from three randomly selected aryl halides were transferred for knowledge sharing in each target task.



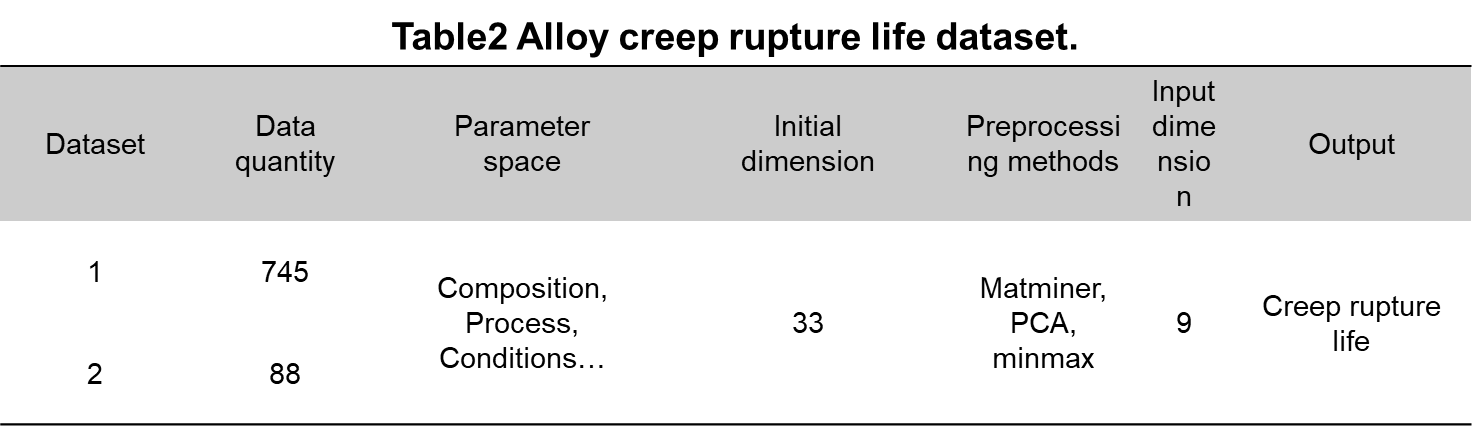
**Figure 4: Comparative analysis of sampling algorithms for different aryl halide optimization tasks.** (A-C) Comparison of initial sampling strategies and comparison of standard BO and transfer BO methods based on aryl halide 1, 9, and 15. (Left) Spearman correlation curves for models on a reserved test set under different initial sampling strategies across three aryl halides. The curves depict the variation in correlation as the number of samples increases for Random, LHS, and EDS. Models trained on datasets generated by EDS exhibit higher performance metrics compared to other methods. (Middle) Yield optimization results from BO using different initial sampling methods. The dashed black line represents the target yield, and the curves illustrate the yield progression as the number of samples increases. BO optimized with EDS-based initial sampling shows reduced variance in the optimization curves, even with comparable mean values across methods. (Right) Optimization trajectories comparing transfer BO using various source datasets with standard BO. The source datasets are generated by randomly selecting different aryl halide numbers and using their corresponding data. Some transfer BO methods show similar optimization performance to standard BO, while others demonstrate significantly faster convergence to optimal yield combinations.

First, we evaluated different sampling strategies on the aryl halide 1 dataset (Figure 4A). The 264 data points were split into a training set and a test set in an 8:2 ratio. Within the training set, data were sampled using Random, LHS, and EDS strategies to build Gaussian Process regression models, and the predictive performance on the test set was evaluated using the Spearman correlation coefficient. Error bars indicate the variance of the metric across 40 independent experiments. When the number of samples exceeded 20, the model based on EDS outperformed those based on Random and LHS. Next, we fixed the initial sample size at 20 to further explore the impact of initial sampling on BO, as shown in the middle panel of Figure 4A. During the first 40 EI-based sampling iterations, the performance differences among the three strategies were minimal. However, when the number of samples exceeded 60, the error bars for BO under the EDS strategy significantly decreased, which we attribute to the early global uniform sampling providing a rough understanding of the optimization task, thereby improving optimization stability. Lastly, the right panel of Figure 4A compares the optimization performance of transfer BO and standard BO using reaction data from halides 2, 4, and 7 as known datasets. The results showed no significant differences between the two strategies, likely due to low correlation between the source and target tasks, resulting in ineffective transfer of optimization information.

In the tests with aryl halides 9 (Figure 4B) and 15 (Figure 4C), we conducted the same sampling strategy evaluations as described above. The results showed that, in terms of model performance, the EDS method still outperformed Random and LHS, particularly in BO optimization, where the variance of EDS was smaller, further supporting the hypothesis that global uniform sampling improves BO optimization stability. For transfer sampling that introduces relevant information at the beginning of the target task, the effectiveness of transfer varies depending on the source data. The experiments demonstrated that data from halides 8 and 14 helped optimize the reaction combinations of halide 9, while for halide 15, transfer sampling from halide 8 even identified the optimal yield combination within 20 iterations. This indicates that when relevant data is available, transfer sampling strategies have significant potential for practical applications.

## Optimization of Simulated High-Temperature Alloy Materials

The second experimental dataset is the alloy creep life dataset [40], which is sourced from both Refs. and the NIMS public database. The former contains 745 samples, while the latter includes 88 samples, with each sample characterized by 33 features encompassing chemical composition (22 elements), heat treatment processes (9 parameters), and experimental conditions (2 parameters). The target feature is the creep rupture life. Given that creep rupture life values are relatively small, we used the logarithm of this attribute, i.e. (creep rupture life), as the output variable. Our goal is to maximize the creep rupture life. In this study, we used Matminer descriptors to describe the chemical composition based on the aforementioned features. We then combined these composition descriptors with the processing variables, followed by min-max normalization and principal component analysis (PCA) for dimensionality reduction.



Similarly, for the exploration of initial sampling, we treated the two datasets as separate BO tasks to search for the parameter combinations that result in the longest creep rupture life. Proxy models were built using initial datasets of the same size obtained through different sampling methods, and Expected Improvement (EI) was then used to identify the experimental combination with the maximum creep rupture life. For the investigation of transfer sampling, knowledge transfer was performed using the other dataset (outside the target task) as the source data.



**Figure 5. Comparative analysis of sampling algorithms for two alloy dataset optimization tasks.** (A) Comparison of initial sampling strategies and comparison of standard BO and transfer BO methods based on Alloy Dataset 1. (B) Comparison of initial sampling strategies and comparison of standard BO and transfer BO methods based on Alloy Dataset 2. The left panel shows Spearman correlation trends as the number of initial samples increases. The middle panel depicts the optimization results from BO using different initial sampling. The right panel illustrates the optimization trajectories for random samples, standard BO, and transfer BO.

Figure 5(A) shows the test results on Alloy Dataset 1. This dataset contains 745 data points, which were split into a training set and a test set with an 8:2 ratio. We used only a few dozen data points obtained from initial sampling within the training set for model prediction. Although the Spearman coefficient was not very high at this stage, EDS still showed better modeling performance compared to other methods. Finding the optimal experimental combination among so many data points is challenging. While the mean of BO optimization curves under all three initial sampling methods was comparable, EDS exhibited smaller variance after 150 samples, indicating stronger stability during the optimization process. Regarding the performance of transfer sampling, Dataset 2 contained only 88 data points, so the knowledge available for transfer was limited. As a result, the optimization performance of transfer BO and standard BO was similar, with both outperforming random sampling, further demonstrating BO's effectiveness in optimization problems. Figure 5(B) presents the test results on Alloy Dataset 2. Even in a small sample dataset, EDS demonstrated excellent modeling performance. Particularly after the initial BO sampling, EDS showed outstanding optimization results. This is due to EDS's tendency to sample at the boundaries of each dimension, which allows it to quickly explore extreme values when the target variable changes linearly along a certain dimension. Finally, supported by the larger Dataset 1, transfer BO surpassed standard BO and random sampling in optimization performance after just 10 iterations, showcasing the strong capability of the transfer algorithm in transferring knowledge from large to small datasets.

# Summary ＆ Conclusions

In this study, we propose two initial sampling strategies within the Bayesian Optimization (BO) framework: Euclidean distance sampling (EDS) and transfer sampling. These strategies are optimized for different task contexts as follows: (1) In the absence of prior information regarding the task, the EDS strategy is employed for initial sampling. This approach aims to ensure that the data points are distributed as evenly as possible in high-dimensional space, maintaining maximum Euclidean distance between them. This uniform distribution facilitates global coverage and provides an initial rough estimate of the black-box function's structure, thereby guiding the optimization direction. (2) When relevant experimental data from parallel source tasks is available (though not directly applicable for modeling and optimization of the target task), we adopt the transfer sampling strategy. This strategy first assesses the correlation between the source and target tasks and assigns different weights to the sampling process based on this correlation. This enables knowledge transfer from the source task data to the target task, thereby accelerating the convergence of the optimization process. It is important to note that the weights in transfer sampling are updated iteratively: as the target task data increases and the target model improves, the composite sampling function gradually reduces the influence of source data, ensuring the robustness of the transfer optimization process.

Specifically, we conducted benchmark tests on BO simulations and sampling strategies using the Goldstein-Price test function, cross-coupling reaction datasets, and high-temperature alloy datasets. First, on the Goldstein-Price test function, we performed 20 initial samples based on random, LHS, and EDS, and visualized the distribution of the sampled points as well as the predictions of the surrogate models. The results showed that the EDS strategy led to a more uniform and dispersed distribution of sampling points, with the surrogate model's predictions fitting the objective function more accurately.Next, we compared the changes in the Spearman correlation coefficient of the surrogate model as the number of samples increased under the three sampling strategies. Based on the results from 40 independent experiments, we observed that under the EDS strategy, when the sample size exceeded 20, the model's Spearman correlation coefficient exhibited a larger mean and smaller variance. Finally, with the initial sample size fixed at 20, we used the commonly used Expected Improvement (EI) acquisition function for BO optimization. The results showed that the EDS strategy yielded higher BO optimization efficiency and was more effective in identifying the global optimum. For the transfer optimization benchmark, we designed three different source functions, representing known source tasks, to transfer knowledge to the Goldstein-Price function's target optimization task. We also compared this approach with the standard BO, which did not use source function transfer. The results indicated that when the source function was similar to the target function, transfer BO significantly accelerated the optimization process. However, when the source function differed greatly from the target function, the performance of transfer BO was comparable to that of standard BO. By tracking the changes in the source model weights during the iteration process, we found that the transfer BO with better optimization performance had larger weights in the initial stages, indicating that knowledge from the relevant source function was effectively transferred.

In the cross-coupling reaction dataset, which contains 15 types of aryl halides, we randomly selected the aryl halide datasets numbered 1, 9, and 15 as optimization tasks and tested the modeling and optimization performance of the random, LHS, and EDS strategies. At the same time, we randomly selected three other aryl halide datasets as source data for transfer sampling optimization tests. In all three optimization tasks, the Spearman correlation coefficient of the surrogate model under the EDS strategy was significantly higher than that under random and LHS. In the Aryl halide 15 dataset, when 20 samples were extracted, the Spearman coefficient for EDS was 0.2 higher than that of LHS. For the optimization tasks of Aryl halide 1 and Aryl halide 9, after EDS initial sampling, the model variance at the end of BO optimization dropped to zero, indicating that the global optimum was found in all 40 independent repeat experiments. In the transfer sampling test, the performance of transfer BO varied with different source data. In the optimization task of Aryl halide 1, there was little difference between the optimization curves of the three transfer BO strategies and the standard BO. However, in the optimization task of Aryl halide 9, transfer BO using Aryl halide 2 and 8 as source models improved the optimal yield by 1.3% and 2.9%, respectively, after 10 samples. Particularly, in the optimization task of Aryl halide 9, transfer BO using Aryl halide 8 as the source model performed best, finding the optimal yield combination from 264 experimental combinations after just 16 samples, while the standard BO had not yet found it after 50 samples.

We conducted the same tests on two high-temperature alloy sub-datasets of different scales and performed transfer optimization by building models for each dataset. In dataset 1, which contains 745 samples, the advantage of EDS in modeling was particularly evident, with the Spearman correlation coefficient of the surrogate model improving by up to 0.3 compared to random and LHS. In dataset 2, which contains 88 samples, the creep life identified through EDS initial sampling was significantly higher than that of the other two strategies. We attribute this to the boundary sampling characteristic of EDS, which makes it more likely to explore the extreme values at both ends of the variable range. In terms of transfer sampling, the knowledge from dataset 1, with its larger sample size, was beneficial for optimizing dataset 2, while the optimization task in dataset 1 was only marginally supported by dataset 2. This suggests that the scale of the dataset has a significant impact on the effectiveness of transfer learning [41].

In summary, we propose two initial sampling strategies suitable for different contexts, each supporting the application of BO in varying scenarios. EDS generates discrete and uniform sampling points across the entire space, allowing for better coverage of the general shape of the black-box function, thereby enhancing the predictive performance of the surrogate model and guiding subsequent BO optimization. On the other hand, transfer sampling leverages existing data from related tasks to assist optimization when target task data is limited, while gradually reducing reliance on source data as the target task data becomes more abundant. The effectiveness of transfer optimization depends on the correlation between the source and target tasks. The weight adjustments made during each iteration ensure the robustness of the optimization, allowing transfer BO to revert to standard BO when relevant source tasks are unavailable.

Our observations indicate that an appropriate initial sampling strategy can significantly improve the performance of the BO algorithm, reducing the number of experiments required and the associated costs, while accelerating the optimization of high-performance materials. This work is not only applicable to individual materials but can also be extended to the broader experimental design problem in materials optimization. We hope that the tests and results presented in this study will offer valuable insights to the materials discovery community and foster deeper collaboration between machine learning and materials science.

# Reference

[1] H. Stein, J. Gregoire. Progress and prospects for accelerating materials science with automated and autonomous workflows [J]. Chemical Science, 2019, 10.

[2] L. Schelhas, Z. Li, J. Christians, et al. Insights into Operational Stability and Processing of Halide Perovskite Active Layers [J]. Energy & Environmental Science, 2019, 12.

[3] J.M. Granda, L. Donina, V. Dragone, et al. Controlling an organic synthesis robot with machine learning to search for new reactivity [J]. Nature, 2018, 559(7714): 377-81.

[4] B.A. Rizkin, A.S. Shkolnik, N.J. Ferraro, et al. Combining automated microfluidic experimentation with machine learning for efficient polymerization design [J]. Nature Machine Intelligence, 2020, 2(4): 200-9.

[5] G. Box, J. Hunter, W.G. Hunter. Statistics for Experimenters: Design, Innovation, and Discovery [J]. AMC, 2005, 10.

[6] S.A. Weissman, N.G. Anderson. Design of Experiments (DoE) and Process Optimization. A Review of Recent Publications [J]. Organic Process Research & Development, 2015, 19(11): 1605-33.

[7] B.J. Shields, J. Stevens, J. Li, et al. Bayesian reaction optimization as a tool for chemical synthesis [J]. Nature, 2021, 590(7844): 89-96.

[8] M. Zhong, K. Tran, Y. Min, et al. Accelerated discovery of CO2 electrocatalysts using active machine learning [J]. Nature, 2020, 581(7807): 178-83.

[9] P.M. Attia, A. Grover, N. Jin, et al. Closed-loop optimization of fast-charging protocols for batteries with machine learning [J]. Nature, 2020, 578(7795): 397-402.

[10] D. Eriksson, M. Pearce, J. Gardner, et al. Scalable Global Optimization via Local Bayesian Optimization [M]. 2019.

[11] Y. Diouane, V. Picheny, R. Le Riche, et al. TREGO: a Trust-Region Framework for Efficient Global Optimization [J]. Journal of Global Optimization, 2022, 86: 1-23.

[12] Q. Liang, A. Gongora, Z. Ren, et al. Benchmarking the performance of Bayesian optimization across multiple experimental materials science domains [J]. npj Computational Materials, 2021, 7: 188.

[13] Z. Liu, N. Rolston, A.C. Flick, et al. Machine learning with knowledge constraints for process optimization of open-air perovskite solar cell manufacturing [J]. Joule, 2022, 6(4): 834-49.

[14] A. Siemenn, Z. Ren, Q. Li, et al. Fast Bayesian Optimization of Needle-in-a-Haystack Problems using Zooming Memory-Based Initialization [M]. 2022.

[15] J. Ren, D. Sweet. Optimal Initialization of Batch Bayesian Optimization [J]. ArXiv, 2024, abs/2404.17997.

[16] E. Kamrah, F. Ghoreishi, Z. Ding, et al. How Diverse Initial Samples Help and Hurt Bayesian Optimizers [J]. Journal of Mechanical Design, 2023.

[17] M.D. Mckay, R.J. Beckkman, W. Conover. Comparison of three methods for selecting values of input variables in the analysis of output from a computer code [J]. Technometrics, 2000, 21: 266-94.

[18] J. Cheng, M. Druzdzel. Latin Hypercube Sampling in Bayesian Networks [M]. 2000.

[19] R. Sheikholeslami, S. Razavi. Progressive Latin Hypercube Sampling: An efficient approach for robust sampling-based analysis of environmental models [J]. Environmental Modelling & Software, 2017, 93: 109-26.

[20] M. Ahmadi, M. Ziatdinov, Y. Zhou, et al. Machine learning for high-throughput experimental exploration of metal halide perovskites [J]. Joule, 2021, 5.

[21] Y. Zhao, J. Zhang, Z. Xu, et al. Discovery of temperature-induced stability reversal in perovskites using high-throughput robotic learning [J]. Nature Communications, 2021, 12: 2191.

[22] X. Du, L. Lüer, T. Heumueller, et al. Elucidating the Full Potential of OPV Materials Utilizing a High-Throughput Robot-Based Platform and Machine Learning [J]. Joule, 2021, 5(2): 495-506.

[23] M. Saliba. Polyelemental, Multicomponent Perovskite Semiconductor Libraries through Combinatorial Screening [J]. Advanced Energy Materials, 2019, 9: 1803754.

[24] Y. Zhao, T. Heumueller, J. Zhang, et al. A bilayer conducting polymer structure for planar perovskite solar cells with over 1,400 hours operational stability at elevated temperatures [J]. Nature Energy, 2022, 7.

[25] K. Weiss, T. Khoshgoftaar, D. Wang. A survey of transfer learning [J]. Journal of Big Data, 2016, 3.

[26] S. Niu, Y. Liu, J. Wang, et al. A Decade Survey of Transfer Learning (2010–2020) [J]. IEEE Transactions on Artificial Intelligence, 2020, 1: 151-66.

[27] M. Feurer. Scalable Meta-Learning for Bayesian Optimization using Ranking-Weighted Gaussian Process Ensembles, F, 2018 [C].

[28] M. Volpp, L.P. Fröhlich, K. Fischer, et al. Meta-Learning Acquisition Functions for Transfer Learning in Bayesian Optimization [J]. arXiv: Machine Learning, 2019.

[29] M. Jamil, X.-S. Yang. A Literature Survey of Benchmark Functions For Global Optimization Problems [J]. Int J of Mathematical Modelling and Numerical Optimisation, 2013, 4.

[30] Y.-F. Lim, C.K. Ng, U.S. Vaitesswar, et al. Extrapolative Bayesian Optimization with Gaussian Process and Neural Network Ensemble Surrogate Models [J]. Advanced Intelligent Systems, 2021, 3: 2170077.

[31] Preface [Z]//RASMUSSEN C E, WILLIAMS C K I. Gaussian Processes for Machine Learning. The MIT Press. 2005: 0.10.7551/mitpress/3206.003.0002

[32] P. Hennig, C.J. Schuler. Entropy search for information-efficient global optimization [J]. J Mach Learn Res, 2012, 13(null): 1809–37.

[33] A. Paleyes, M. Pullin, M. Mahsereci, et al. Emulation of physical processes with Emukit [J]. ArXiv, 2021, abs/2110.13293.

[34] F. Paul, J. Patt, J.F. Hartwig. Palladium-catalyzed formation of carbon-nitrogen bonds. Reaction intermediates and catalyst improvements in the hetero cross-coupling of aryl halides and tin amides [J]. Journal of the American Chemical Society, 1994, 116(13): 5969-70.

[35] S. Anil, R. Rennels, S. Buchwald. A Simple Catalytic Method for the Conversion of Aryl Bromides to Arylamines [J]. Angewandte Chemie-international Edition - ANGEW CHEM INT ED, 1995, 34: 1348-50.

[36] A.S. Guram, S.L. Buchwald. Palladium-Catalyzed Aromatic Aminations with in situ Generated Aminostannanes [J]. Journal of the American Chemical Society, 1994, 116: 7901-2.

[37] S. Lee, M. Jørgensen, J.F. Hartwig. Palladium-Catalyzed Synthesis of Arylamines from Aryl Halides and Lithium Bis(trimethylsilyl)amide as an Ammonia Equivalent [J]. Organic Letters, 2001, 3(17): 2729-32.

[38] P.A. Forero-Cortés, A.M. Haydl. The 25th Anniversary of the Buchwald–Hartwig Amination: Development, Applications, and Outlook [J]. Organic Process Research & Development, 2019, 23(8): 1478-83.

[39] D.T. Ahneman, J.G. Estrada, S. Lin, et al. Predicting reaction performance in C–N cross-coupling using machine learning [J]. Science, 2018, 360(6385): 186-90.

[40] C. Zhou, R. Yuan, B. Su, et al. Creep rupture life prediction of high-temperature titanium alloy using cross-material transfer learning [J]. Journal of Materials Science & Technology, 2024, 178: 39-47.

[41] H. Yamada, C. Liu, S. Wu, et al. Predicting Materials Properties with Little Data Using Shotgun Transfer Learning [J]. ACS Central Science, 2019, 5(10): 1717-30.