**Tailoring Initial Sampling Strategies for Bayesian Optimization**

**in Materials Tasks Based on Data Availability**

**Abstract**

Compared to traditional experimental designs, Machine Learning-based optimization strategies, such as Bayesian optimization, have shown promising results in materials development tasks. Before commencing an optimization task, it is crucial to perform initial sampling to collect preliminary experimental data, which then facilitates the construction of a machine learning model and indicates the subsequent optimization direction. Most researchers tend to focus on the model and optimization strategies while overlooking the impact of the initial sampling strategy.

In this study, we propose two initial sampling strategies: Euclidean distance sampling and Knowledge transfer sampling. These two strategies can address different scenarios while initiating an optimization task. If we lack information about the task and need to start from scratch, Euclidean-distance sampling ensures that the initial sampling data is uniformly distributed across the global space. Conversely, suppose we have pre-existing knowledge and databases. In that case, knowledge-transfer sampling can transfer relevant knowledge to assist in the initial sampling for the target task, much like how materials experts select initial experimental settings for new material tasks based on past experiences.

We evaluated our sampling methods on one benchmark function and two material datasets: the palladium-catalyzed aryl halide cross-coupling reaction dataset and the high-temperature alloy creep rupture life dataset. The results demonstrate that Euclidean-distance sampling not only enhances the stability of the initial model but also facilitates subsequent optimization. Knowledge-transfer sampling allows for the leveraging of knowledge from related data, aiding in the discovery of the global optimum.

# Introduction

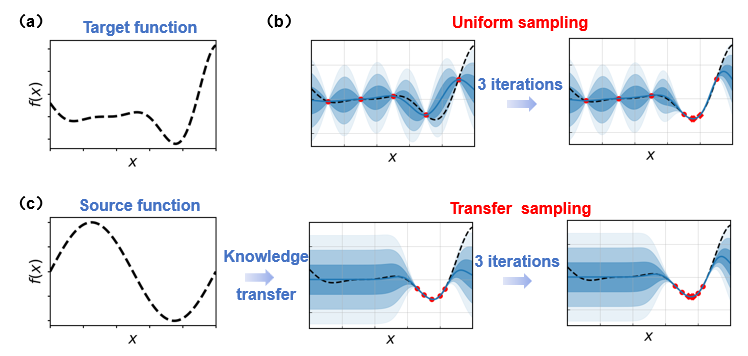
In the development and design of materials experiments, systematic optimization often involves hundreds and thousands of experimental combinations due to high-dimensional variables, making it difficult to solve through brute force or conventional experimental design[1]. By introducing machine learning (ML) into the loop to guide material optimization tasks, it becomes possible to explore complex search spaces and avoid exhaustive enumeration adaptively.

Bayesian Optimization (BO), a sequential machine learning approach, has gained widespread application in the exploration of chemical synthesis and materials optimization due to its efficient sampling and optimization capabilities when dealing with expensive black-box functions. BO has thus become an effective strategy in these domains.

Current reports on classical Bayesian optimization emphasize the selection of different surrogate models and kernel functions, adaptive exploration-exploitation of the acquisition function, and even iterative expansion of the sampling search boundary. Yet, few have focused on determining initial sampling methods. The quality of the initial data not only determines the performance of the initial surrogate model but also profoundly influences the subsequent optimization direction, making it crucial to the overall effectiveness of BO.

In the absence of any prior knowledge, researchers often resort to Random sampling or Latin Hypercube Sampling (LHS) as initial sampling methods. However, both approaches have their limitations. Random sampling, due to its high Randomness, may lead to uneven sample distribution, while LHS, although ensuring uniformity in individual dimensions, fails to guarantee uniform sampling across the global search space. 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. For example, in a 3×3 grid, if an LHS sampling point falls within the (1,1) sub-grid, no further sampling points will be placed in the (1,2) and (1,3) sub-grids of the same row, or in the (2,1) and (3,1) sub-grids of the same column.

Random sampling and LHS are applicable when pertinent data are absent for reference. Conversely, when relevant data exists, human experts generally eschew these methods. Instead, they leverage insights from previous optimization tasks to guide the initial exploration of new challenges. Transferring knowledge across different optimization tasks reflects a natural inclination in human behavior. However, simulating this knowledge transfer process in Bayesian Optimization (BO) remains a significant challenge.



**Figure 1. Schematic illustration of uniform sampling and transfer sampling strategies.** (a) The target function to be optimized. (b) BO that based on uniform initial sampling. Uniform initial sampling collects comprehensive global information, allowing for the identification of the global optimum after a few iterations. (c) BO that incorporates knowledge transfer from a source function. Transfer sampling directly targets likely regions of the global optimum, achieving rapid convergence through minimal iterations.

In this study, we undertook the following work: (a) We propose a globally uniform initial sampling strategy, Euclidean Distance Sampling (EDS), for scenarios lacking relevant data references. Compared to Random Sampling and Latin Hypercube Sampling (LHS), EDS markedly improves the stability of surrogate model performance during the initialization phase and diminishes the risk of convergence to local optima in the optimization process. (b) When relevant data are available, we utilize a transfer learning strategy that integrates knowledge from related datasets into the sampling process for the target task, thus reducing the experimental effort needed to locate the global optimum. The distinctive feature of this method lies in its capability to assess the correlation between the source model and the target task across iterations. It dynamically adjusts the extent of knowledge transfer based on this correlation, thereby facilitating effective and reliable transfer sampling in Bayesian Optimization.

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.

The organization of this paper is as follows: Section 2 introduces the EDS and transfer sampling strategy we propose; Section 3 demonstrates these methods using the Goldstein-Price function; 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 demonstrating that, across various contexts of data availability, both Euclidean Distance Sampling (EDS) and transfer sampling strategies enhance the optimization process of the target task more effectively.

# Methodology

In this paper, 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.

## 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:

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:

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 optimizization. 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:

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

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:

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.

Where

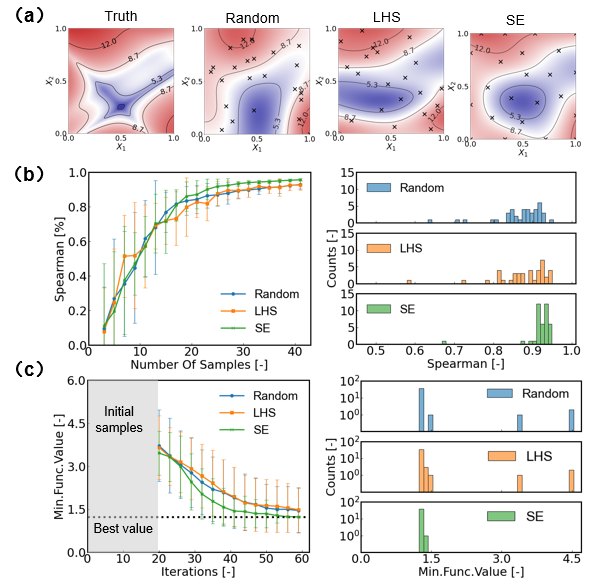
By calculating the transfer weight, we can guide the ranking optimization tasks to take further steps, thus realizing the transfer of tasks from source to target.

# Demonstration

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 SE, we first extract 20 samples from the dataset using different methods to build Gaussian Process Regression (GPR) models. We then visualize the sample locations and the predictive performance of the models. 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. Finally, we determine an appropriate number of initial sampling points to establish the initial model, and based on this, we utilize the Expected Improvement (EI) acquisition function to perform Bayesian Optimization (BO). This approach allows us to investigate how different sampling methods influence the subsequent performance of BO.



**Figure 2: Evaluation of sampling strategies on Goldstein-Price function.** (a) Visualization of sample distributions and Gaussian process (GP) model predictions employing various sampling methods. The leftmost panel depicts the actual distribution of the target function, whereas the subsequent panels display the GP model forecasts using Random, LHS, and SE with an initial sample size of 20. (b) Analysis of Spearman's correlation across sampling strategies. The left graph illustrates the correlation trends as the sample size increases among the strategies, while the right graph portrays Spearman's correlation coefficients' distribution based on 40 independent trials with 20 samples each. (c) Comparative optimization trajectories under Bayesian optimization (BO) employing distinct initial sampling methods, alongside the statistical distribution of optimal function values achieved. The left graph details the optimization paths utilizing EI with varied initial sampling approaches, and the right graph showcases the distribution of minimal function values encountered over 40 trials.

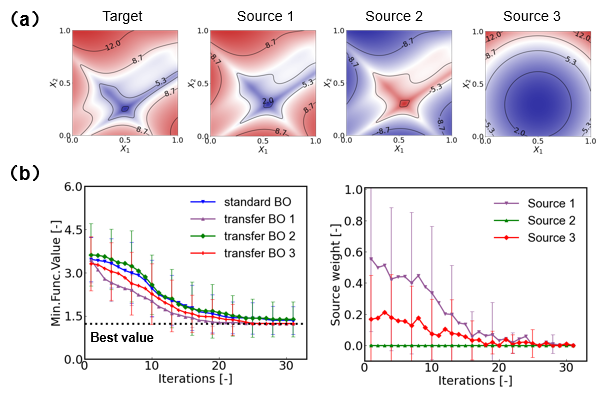
Figure 2 (a) illustrates the prediction performance of the objective function under different sampling strategies. The "Truth" plot in the top left corner represents the actual distribution of the objective function, while the other three plots show the predictions of Gaussian Process models constructed by sampling 20 points using Euclidean distance sampling (SE), Random sampling (Random), and Latin Hypercube Sampling (LHS) strategies, respectively. From the figure, it is evident that SE sampling more accurately captures the global optimum of the objective function, whereas Random and LHS exhibit significant deviations, particularly in regions with high variability in the objective function.

Figure 2 (b) (left) shows how the Spearman correlation coefficient of the models varies with the number of samples across different sampling strategies. It can be observed that the Spearman coefficient for SE sampling stabilizes rapidly as the sample size increases and is generally higher than that of Random and LHS in most cases. This indicates that the SE strategy can build an accurate model more quickly. The right plot shows the distribution of Spearman coefficients across 20 samples and 40 repeated experiments for each sampling strategy. The distribution for SE is concentrated within a higher correlation coefficient range, further validating its superior modeling capability.

Figure 2 (c) (left) demonstrates the optimization effectiveness using the Expected Improvement (EI) strategy following different initial sampling strategies. As the number of samples increases, the objective function values progressively approach the global optimum. Notably, the SE strategy exhibits rapid convergence at the earliest stage, whereas Random and LHS require more samples to approach the optimal value. The right plot shows the distribution of the minimum objective function values at the end of optimization after 40 repeated experiments. The SE strategy significantly outperforms the other two strategies, with a more concentrated distribution closer to the optimal value.

## 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. We then compared the optimization curves of standard Bayesian Optimization (BO) without knowledge transfer and transfer BO with different prior knowledge. Additionally, we tracked the changes in the weights of different source models throughout the iterative process.



**Figure 3: Comparison of standard Bayesian Optimization and transfer Bayesian Optimization with different source models.** (a) Contour plots illustrating the target function alongside three source functions utilized for transfer learning. The source functions are derived from distinct methodologies: Source 1 is generated by shifting and scaling the target function, Source 2 is an inversion of Source 1, and Source 3 is derived from an entirely different function. (b) (Left) Optimization trajectories comparing standard BO with three variations of transfer BO. The dashed black line represents the global minimum of the target function. (Right) Evolution of source models’ weight throughout the transfer BO process. Higher source weights signify a greater contribution of the respective source model to the optimization.

Figure 3 (a) presents the contour plots of the objective function alongside the three source functions used for transfer learning. The top-left plot illustrates the objective function, where the blue region indicates the global minimum area.

Figure 3 (b) (left) compares the optimization performance of standard Bayesian Optimization with transfer BO utilizing three different source functions. 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.

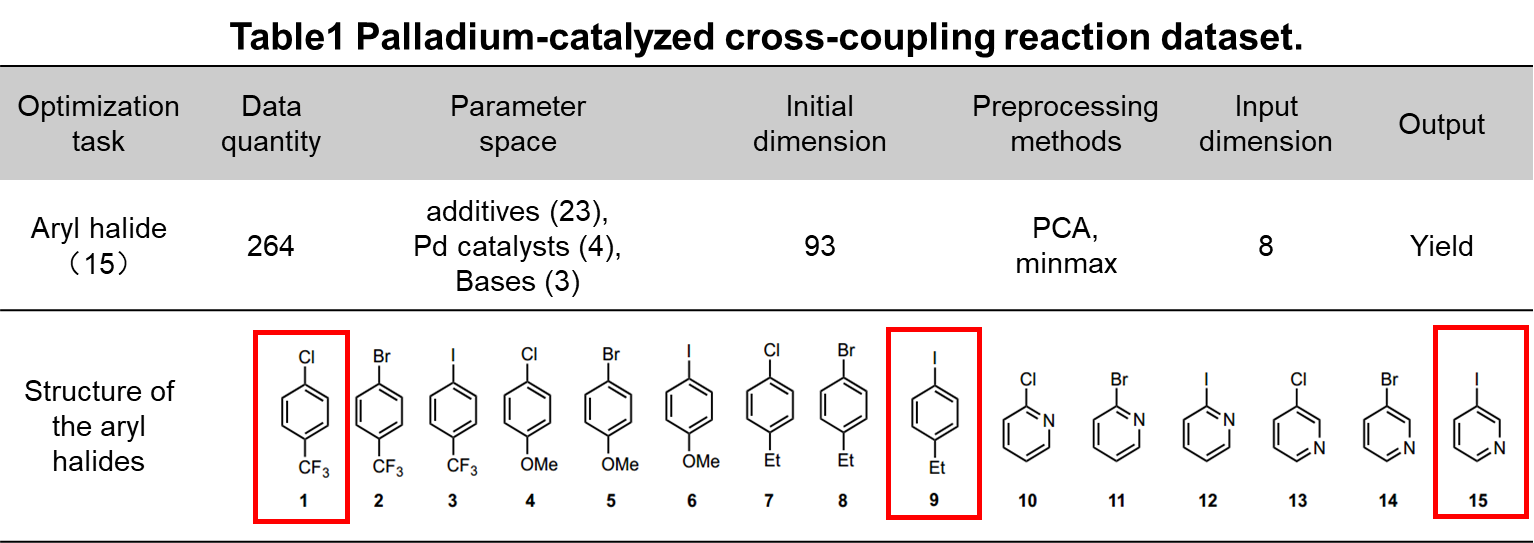
Figure 3 (right) shows the weight variations of the three source models during the transfer BO process. In the early stages of optimization, the weight of source model 1 was greater than that of 3, while the weight of source model 2 remained close to zero, consistent with the conclusions drawn from panel 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.

# Experiments

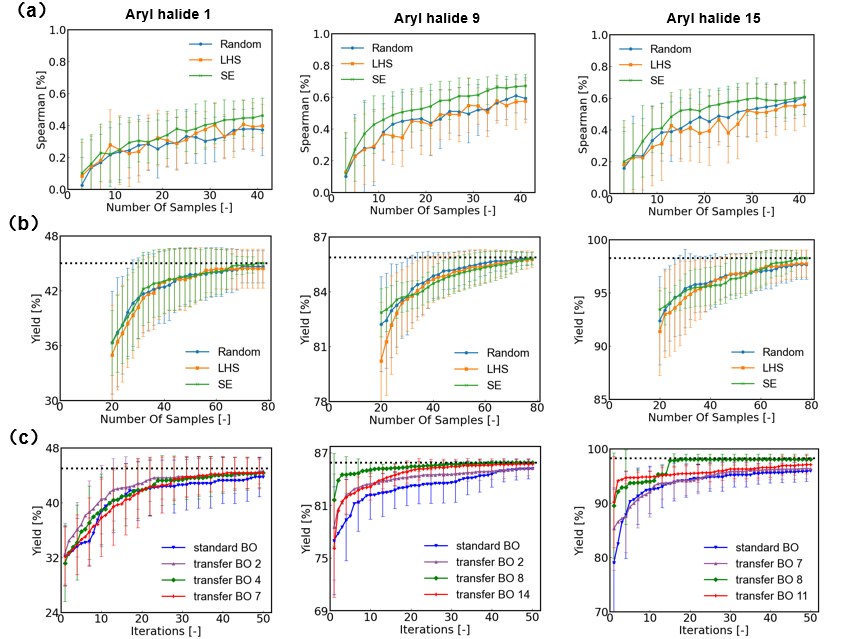
We selected two real-world materials datasets to evaluate the performance of different sampling algorithms: the palladium-catalyzed aryl halide cross-coupling reaction dataset and the high-temperature alloy creep rupture life dataset. Both datasets are sourced from publicly available materials databases.

## Optimization of Simulated Cross-Coupling Reactions

The Pd-catalyzed Buchwald-Hartwig reaction holds significant value in pharmaceutical synthesis. Ahneman et al. 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.

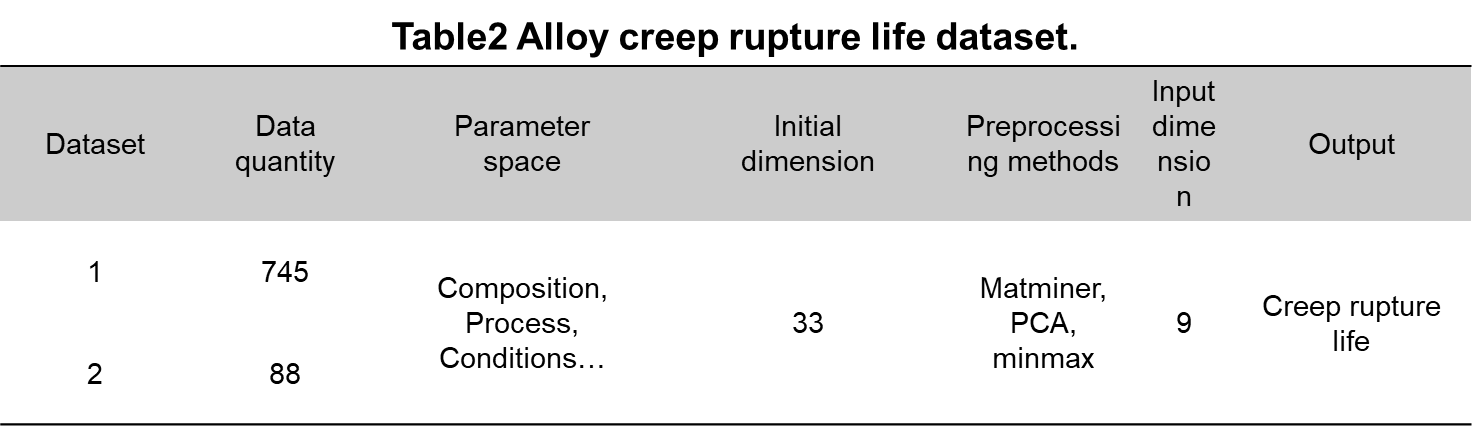


**Figure 4: Comparative analysis of sampling algorithms for different aryl halide optimization tasks.** (a) Spearman correlation curves for models on a reserved test set under different initial sampling strategies across three aryl halides. The curves depict the variation in correlation as the number of samples increases for Random, LHS, and SE. Models trained on datasets generated by SE exhibit higher performance metrics compared to other methods. (b) 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 SE-based initial sampling shows reduced variance in the optimization curves, even with comparable mean values across methods. (c) 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.

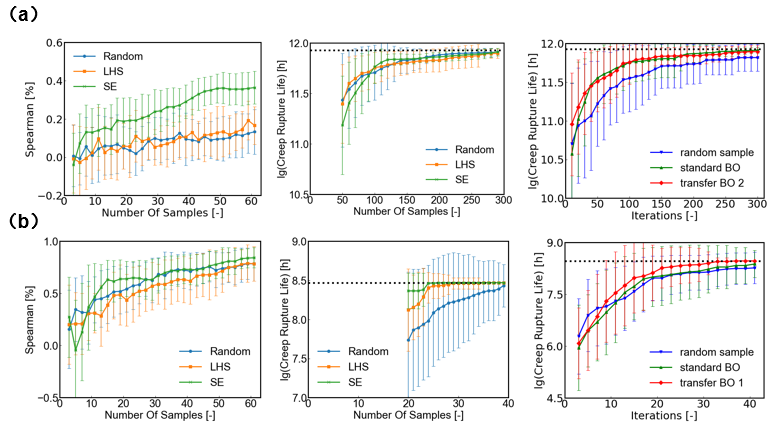
Figure 4 (a) shows that compared to Random and LHS sampling, the surrogate model under SE initial sampling achieves a higher Spearman coefficient. This indicates that, with the same number of samples, SE-based sampling provides more informative data, leading to a more accurate prediction of trends in unknown regions by the surrogate model. (b) displays the optimization curves after 20 initial sampling points with Random, LHS, and SE methods, followed by 60 optimization samplings using EI to find the optimal solution. The optimization curves of the three methods are quite similar, but it is evident that the error bars for SE sampling are narrower, indicating the smallest variance over 40 independent runs and more stable BO optimization performance. (c) compares the optimization curves of transfer BO and standard BO when prior knowledge from other aryl halide datasets is used for knowledge transfer. Under the same number of iterations, transfer BO identifies combinations with higher reaction yields.

## Optimization of Simulated High-Temperature Alloy Materials

The second experimental dataset is the alloy creep life dataset, 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. 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. (b) Comparison of initial sampling strategies and comparison of standard BO and transfer BO methods based on Alloy Dataset 2. Alloy Dataset 1 is almost nine times larger than Dataset 2. During initial modeling, SE shows a more pronounced improvement in model metrics on Dataset 1. For BO with different initial sampling, SE achieves better optimization outcomes on Dataset 2, likely due to the favorable experimental combinations being more densely concentrated at certain variable ranges, thus making them more accessible by SE strategy. Regarding transfer sampling, the transfer BO incorporating knowledge from Dataset 2 shows comparable performance to standard BO, whereas transfer BO incorporating Dataset 1 knowledge outperforms standard BO, suggesting that the larger dataset may contribute more effectively to knowledge transfer.

Figure 5 (a) presents the Spearman curves of models and BO optimization curves for Random, LHS, and SE sampling using dataset 1 as the target task, along with a comparison of transfer BO and standard BO curves using dataset 2 as known data. The SE sampling demonstrates better modeling performance and more stable BO optimization. Both transfer BO and standard BO exhibit comparable optimization results, outperforming Random optimization; this may be due to the limited size of dataset 2 compared to dataset 1, which hinders effective information provision. (b) illustrates the same experiment based on dataset 2, highlighting the strong potential of SE as an initial sampling method. Additionally, using dataset 1 as known data, Transfer BO 1 can find the global optimum more quickly, showcasing the knowledge transfer capability of transfer sampling.

# Summary ＆ Conclusions

In summary, we proposed corresponding strategies for different scenarios before implementing Bayesian Optimization (BO): SE (Sampling by Euclidean distance) for initial sampling without source data and transfer sampling when source data is available. Compared to the commonly used initial sampling strategies, Random and LHS, SE sampling achieves globally uniform sampling. For scenarios with source data, transfer sampling effectively leverages source data knowledge to aid in optimizing the target task.

Through simulation optimization on test functions, the cross-coupling dataset, and the alloy creep life dataset, our experiments demonstrated that SE sampling yields more stable model performance than Random and LHS with the same number of samples, positively impacting subsequent EI optimization. Transfer BO using transfer sampling significantly increases the probability of finding the global optimum compared to standard BO with the same number of samples.

# Reference

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