**Bayesian Transfer Learning for Efficient Data Sampling in Materials Science Tasks**

# Introduction

Machine learning for materials science has become an emerging and popular topic. Using machine learning methods to perform regression modeling on material data can help us better understand and predict material characteristics. Generally, high-level machine learning models usually rely on sufficient data knowledge. However, in specific areas of material science, relevant data is often scarce, and we often cannot find data that is exactly consistent with the materials and parameters in our experiments and the target variables to assist in modeling. Therefore, a large number of researchers have already turned their attention to data extension.

However, blind mass data collection not only may incur huge costs but also cause data redundancy, which will bring extra burden to the computation of the model. Kangming Li's research team took the lead in recognizing the importance of data richness. They verified in multiple datasets that only 5% of the data volume was used for model training, and the performance comparable to that of a model trained with all data could be obtained, thereby verifying the existence of data redundancy. On this basis, Bayesian active learning sampling based on the information richness of the data will help us to minimize the necessary data sampling volume while ensuring model accuracy.

In addition to the above, ordinary Bayesian optimization methods also have certain limitations, that is, they cannot extract and apply experience from previous optimization tasks. Each time a new Bayesian optimization task is initiated, it is necessary to go through initial random sampling again, and relevant information can only be obtained after a series of iterative learning. In response to this issue, Feurer et al. proposed the method of "Ranking-weighted Gaussian process ensemble". This method introduces information on the similarity between the source task and the target task, realizes transfer Bayesian optimization, and further reduces the sampling cost and improves the efficiency of modeling.

In our work, we conducted the following research: firstly, we applied Bayesian active learning to various types of material data sets, and compared the performance differences under different sampling methods. Secondly, we used the method of transfer Bayesian active learning to speed up the modeling process. The results show that by using transfer Bayesian sampling and xxx acquisition function method, we can come up with fairly accurate models with the least amount of data sampled.

# Methods and Data

## Different acquisition function sampling under the Bayesian optimization framework

### EI/NLCB

### SV/SE/SVE

### MES

## Ranking-weighted Gaussian process ensemble

## Benchmark testing on different types of material datasets

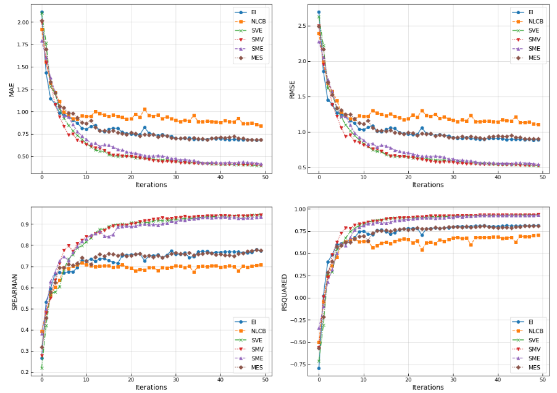
### Dataset of Buchwald-Hartwig cross-coupling reactions of aromatic halides catalyzed by palladium.

### Dataset on the conversion efficiency of perovskites using the Rapid Spray Plasma Process (RSPP).

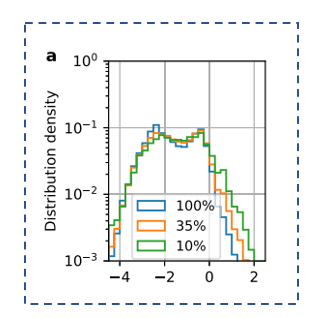
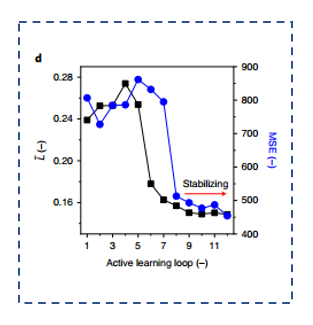
### Dataset of creep life of different alloys.

# Result and Discussion

## Bayesian active learning sampling modeling for material datasets.

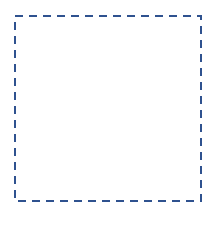


**Figure 1: The comparison chart of model performance changes after sampling with different acquisition functions.** It can be seen that the model's performance improves the fastest when sampling using acquisition function xxx.

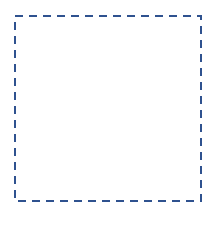
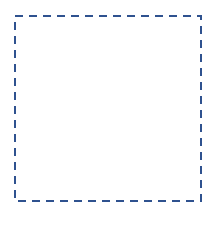
 

**Figure 2: Sampling point analysis chart.** (Left) The distribution of the y-values of the sampling points. (Right) The range of x-values, L, for the sampling points. By using the acquisition function xxx for sampling, the distribution of the objective variables for the sampling points is relatively uniform, and the range of existing x distributions can be expanded more quickly.

## Bayesian Transfer Learning Sampling Modeling for Material Dataset



**Figure 3: The comparison chart of model performance changes after sampling through transfer Bayesian acquisition function.** The model's performance is further improved after sampling using transfer Bayesian.



**Figure 4: Sampling point analysis chart after sampling via transfer Bayesian acquisition function.** As in Figure 2 above, the distribution of the objective variables for the sampling points is still relatively uniform after sampling through transfer Bayesian, and the range of the x distribution is broader.

# Summary

By using transfer Bayesian sampling and xxx acquisition function method, we can come up with fairly accurate models with the least amount of data sampled.