**Reliable Sequential Learning for Material Optimization**

**via Uniform Sampling and Transfer Sampling**

**Abstract**

贝叶斯优化作为常用的顺序机器学习方法，经常被用于材料问题的优化当中，无论是实验组合的寻优还是工艺参数的寻优。然而，在没有先验知识的情况下如何进行有效地初始采样仍是BO的一个挑战。在此，我们针对有无相关知识的不同情况，提出了两种初始采样应对策略，均匀采样和迁移采样。

我们在二维测试函数，钯催化交叉偶联反应数据集，合金蠕变寿命数据集上测试了我们的采样方法，结果表明，通过均匀采样不仅提高初始模型的稳定性，还有利于后续的优化。通过迁移采样可以迁移相关数据的知识，帮助找到全局最优。

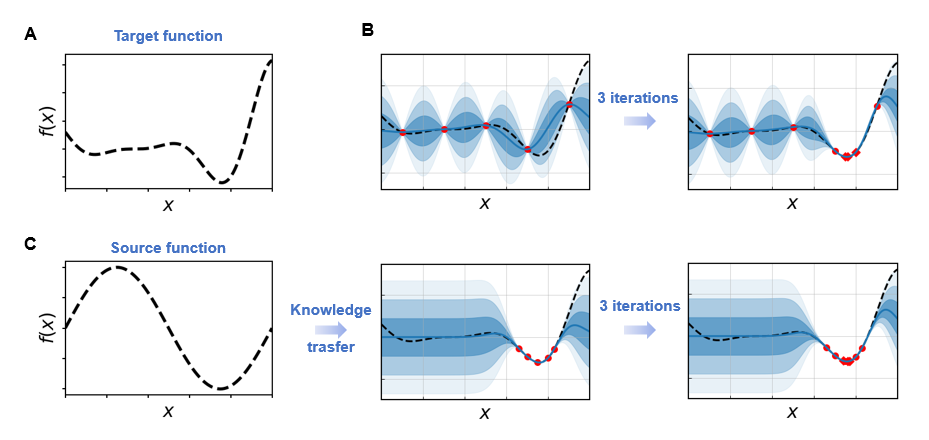
# Introduction

材料优化问题通常具有高维度和高实验成本，使得传统的试错方法在优化过程中面临诸多挑战。贝叶斯优化（Bayesian Optimization, BO）作为一种序贯机器学习方法，凭借其在评估昂贵黑箱函数时的高效取样和优化能力，已被广泛应用于化学反应合成与材料优化的探索中，成为一种有效的策略。

然而，关于经典贝叶斯优化的研究中，两个关键问题常常被忽视：（1）如何选择更有效的初始采样策略，以及（2）如何利用已有数据中的知识来辅助初始采样。初始数据的质量不仅决定了初始代理模型的性能，还对后续优化的方向产生深远影响，对于BO的整体优化效果至关重要。

在缺乏任何先验知识的情况下，研究人员通常会选择随机采样或拉丁超立方采样（LHS）作为初始采样方法。然而，这两种方法各有局限。随机采样由于高度随机性，可能导致样本分布不均匀；而LHS虽然在单个维度上保证了均匀性，但无法确保全局搜索空间内的均匀采样。LHS的原理是对每个维度进行独立的分层采样，然后将这些层组合成采样点，这意味着采样的均匀性仅在单个维度上得到保证。以一个3×3的网格为例，如果LHS方法中的一个采样点落在(1,1)子网格，那么在同一行的(1,2)和(1,3)子网格，以及同一列的(2,1)和(3,1)子网格中将不再有采样点。

随机采样和LHS采样适用于没有任何相关数据可供参考的情况下。然而，当存在相关数据时，人类专家通常不会随机选择初始实验，而是依赖从相关数据中获得的经验来指导新的优化活动的初始探索。知识在不同优化任务之间的迁移是人类的一种自然倾向，但在贝叶斯优化（BO）中，模拟这一知识迁移的过程仍然面临巨大挑战。



**Figure 1. 均匀采样和迁移采样原理图。（a）待优化的目标函数，（b）在没有知识借鉴情况下，通过均匀初始采样，更快找到全局最优解，（c）在源函数的知识借鉴下，通过迁移采样，更快找到全局最优解。**

在本研究中，我们进行了以下工作：（a）在缺乏相关数据参考的情况下，提出了一种全局均匀的初始采样策略——Sample Euclidean Distance（SE）。与随机采样和拉丁超立方采样（LHS）相比，基于SE的采样策略在初始化代理模型时，能够显著提高模型性能的稳定性，并降低在优化过程中陷入局部极值的风险。（b）在有相关数据可供利用的情况下，我们采用迁移学习策略，将相关数据中的知识迁移至目标任务的采样过程中，从而减少寻找全局最优解所需的实验次数。该方法的独特之处在于，它能够在迭代过程中分析源模型与目标任务之间的相关性，并根据相关性动态调整知识迁移的程度，从而实现可靠的迁移采样贝叶斯优化。

值得注意的是，上述工作提到的两种方法分别针对不同数据情境下的初始采样策略，具有并列关系。在实际优化过程中，可根据具体数据情况选择适宜的初始采样方法。

本研究的组织结构如下：第二节介绍了我们提出的SE采样方法和迁移采样策略；第三节利用Goldstein-Price函数对这些方法进行了演示验证；第四节通过两个真实的材料数据集（Buchwald-Hartwig交叉偶联反应数据集和合金蠕变寿命数据集）进行了模拟优化和算法测试。最后，我们提供了证据，证明在不同的数据情境下，SE采样和迁移采样策略能够更有效地促进目标任务的优化。

# Methdology

在本文中，我们进行了两个主要的探究：（1）SE初始采样策略（2）迁移采样贝叶斯优化方法。通过这两种方法，可以适应在不同情况下启动贝叶斯优化策略，实现更稳定，更高效率的优化。

我们演示了上述两种方法在Goldstein-Price函数数据集， Buchwald-Hartwig交叉偶联反应数据集，合金蠕变寿命数据集上的实现。对于SE初始采样，我们和将其与LHS和随机采样进行对比，比较不同采样方式下的得到数据对建立初始模型和后续优化的影响。对于迁移采样，我们将其与标准贝叶斯优化进行对比，比较了不同源数据的迁移BO与标准BO的优化情况。通过在一个函数数据集，两个材料数据集上的测试，我们展示了这两种采样策略稳定的有效性和广泛的适用性。

## 欧氏距离采样

为了保证全局搜索空间采样的均匀性，我们提出了基于欧氏距离的采样策略，简称为SE。该策略通过以下步骤实现：

1.获得启动采样点：对于参数空间X，LHS采样3个点作为启动采样点集。

2.计算未采样点欧氏距离：对于X中每个未采样的点，计算它与中每一个已知点的欧氏距离，并找到这些距离中的最小值。

3.选择下一个采样点：对于X中所有未采样的点，选择与已采样点集 中的点的最小欧氏距离最大的点作为下一个采样点。

4.将加入到中，重复步骤2和3，直到获得需要的采样数量。

## 迁移贝叶斯优化

在有一些相关数据数据作为源数据的情况下，我们可以利用源数据的知识帮助我们进行目标优化任务的采样。该过程的难点在于a)如何判断源数据与目标任务的相关性。b)如何利用源数据的知识帮助目标任务采样。c)如何随着迭代过程自动调节源数据的影响。在本研究，我们结合了Rank-Weighted Gaussian process ensemble (RGPE)和Transfer Acquisition Function（TAF）方法，实现了随着迭代的进行计算源任务和目标任务的相关性，自动更新源数据的权重，实现了可靠的迁移贝叶斯优化。通过以下步骤：

1.计算模型权重：如果一个模型能够根据观测值的函数值对观测值进行正确排序，那么它对优化是有用的。我们通过构建一个损失函数来衡量模型能够正确对目标观测进行排序的程度，在给定|| = > 1目标任务测量值的情况下，模型i的排序损失被表示为错误排序对的数量：

是指示函数，是异或操作。基于排名的损失确保只有最优的位置是重要的，而不是预测的实际值。目标模型的损失使用留一交叉验证计算，由

其中表示模型拟合除第k点外的所有目标任务观测值。现在，根据每个模型是排名损失最小的集成成员的概率为每个模型分配权重。这个概率是通过在验证集中从模型预测中抽取S个样本来估计的，i.e.模型i的权重计算为

2.迁移获取函数：源模型和目标模型都采用高斯过程模型，通过将上述计算的权重分配到各自模型的获取函数上，得到迁移获取函数

其中

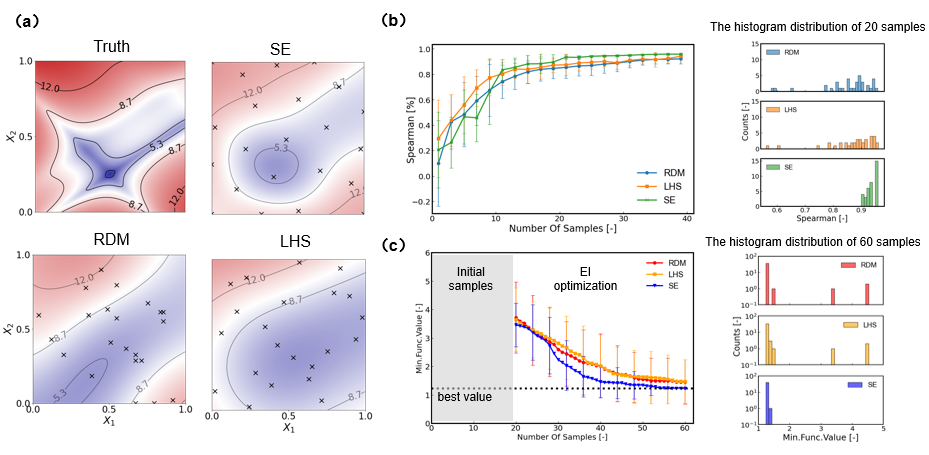
通过迁移获取函数指导目标任务的下一步采样，从而实现将信息从源任务转移到目标任务。

# Demonstration

在本章节，我们以Goldstien-Price (2-dimensions)来展示方法的效果，该函数属于公认的测试函数之一，具有多个局部极小值和一个全局最小值。为了使函数优化更接近材料数据优化时的情况，我们对两个维度设置50个间隔采样得到2500组测试函数数据，并对x进行了minmax归一化和对y进行对数化，用于后续实验测试。

## 均匀采样

为了分析随机采样，LHS和SE的建模和优化效果，首先我们用不同方法从数据集中抽取20个样本建立高斯过程回归模型，并对样本点的位置和模型的预测效果进行可视化。随后我们不断增加采样数量，比较不同采样方法下模型的Spearman指标变化，Spearman系数是一个定义在-1到1之间的度量，用于检查是否存在单调相关。最终我们选择了合适的初始采样点数量建立初始模型，并在此基础上选用EI获取函数进行BO，探究不同采样方式对后续BO优化效果的影响。

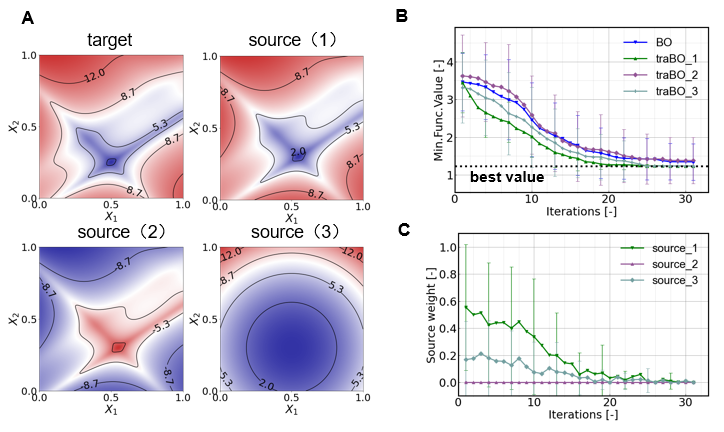


**Figure 2. Visualization of the modeling and optimization results under different initial sampling strategies on a two-dimensional function. (a) The left panel shows the true distribution of the objective function, while the other three panels depict the prediction of Gaussian process models based on SE, RDM, and LHS sampling strategies with 20 sampled points. (b) The left panel shows the variation of the Spearman correlation with the number of samples for different strategies, and the right panel displays the distribution of the Spearman correlation coefficients for 20 and 60 samples after 40 repeated experiments. (c) The left panel shows the optimization process using the EI strategy after different initial sampling, and the right panel shows the distribution of the best function values after 40 repeated experiments.**

**Figure 2 (a)** 图展示了待优化目标函数在不同采样策略下的预测效果。左上角的 "Truth" 图表示真实的目标函数分布，其他三个图分别基于欧氏距离采样（SE）、随机采样（RDM）和拉丁超立方采样（LHS）策略，通过采样 20 个点构建的高斯过程模型的预测情况。从图中可以看出，SE 采样能够更准确地捕捉到目标函数的全局最优点，而 RDM 和 LHS 则表现出较大的偏差，尤其在目标函数的高变化区域。**(b)** 左图展示了不同采样策略下模型的 Spearman 相关系数随样本数量增加的变化情况。可以看到，SE 采样的 Spearman 系数随着样本数量的增加迅速趋于稳定，并且在大部分情况下均高于 RDM 和 LHS，这表明 SE 策略能够更快地构建出准确的模型。右图展示了在 20 个样本和 40 次重复实验下，各采样策略的 Spearman 系数的分布情况。SE 的分布集中在更高的相关系数范围内，进一步验证了其更优越的建模能力。**(c)** 左图展示了在不同初始采样策略下，再通过期望提升（EI）策略进行采样优化的效果。随着样本数量的增加，各策略的目标函数值逐渐逼近全局最优值。其中，SE 策略在最早阶段即表现出快速收敛的特性，而 RDM 和 LHS 则需要更多的采样才能接近最优值。右图展示了优化结束时，40 次重复实验下，目标函数最小值的分布情况。SE 策略显著优于其他两种策略，分布更集中且更接近最优值。

## 迁移采样

为了测试迁移采样方法的迁移优化效果，我们根据目标函数设计了不同类型的源函数，并建立相应的源模型来提供可以访问的已知知识。我们比较了没有知识迁移的标准BO和不同先验知识的迁移BO的优化曲线，同时追踪了随着迭代过程不同源模型的权重变化。



**Figure 3**

**Figure 3 (a)**展示了目标函数和用于迁移的三个源函数的等高线图。左上角的目标函数图显示了优化目标，其中蓝色区域为全局最小区域。**(b)** 展示了与标准BO相比，三种不同源函数的迁移BO的优化性能。经过30次迭代，traBO（1）和traBO（3）可以全部找到全局最优值，这意味着迁移采样可以有效地利用源函数（1）和源函数（3）关于全局最优的知识，帮助完成目标函数的优化，而traBO（2）与标准BO优化性能相当，说明迁移采样能够摆脱源函数（2）的误导信息，回归到标准BO。**(c)** 图显示了三种迁移 BO 过程中源模型的权重变化，在优化前期，源模型权重（1）>（3），源模型（2）的权重基本为0，这与**(b)**图得到的结论相吻合，并且随着迭代进行，所有源模型的权重都在逐步降低，说明随着目标数据的增加，目标模型逐渐完善，目标任务优化逐步摆脱源模型的指导。

# Experiments

我们选用了两个真实世界的材料数据集来评估来不同采样算法的表现，包括钯催化芳基卤化物交叉偶联反应数据集和高温合金蠕变断裂寿命数据集，他们均来自于公开的材料数据库。

## 模拟交叉偶联反应的优化

Pd催化的Buchwald-Hartwig反应在药物合成中具有广泛的价值，Ahneman等人通过高通量实验获得了15种芳基卤化物，23种添加剂、4种钯催化剂和3种碱共3960个实验组合的产率值。原作者通过120维的原子，分子和振动描述符来对每个选项进行特征描述，本工作在原有特征基础上进行了minmax归一化和pca降维。

**表1，钯催化交叉偶联反应数据集情况**

对于初始采样的探究，我们以整个数据集作为BO任务，去寻找最大化反应产率的实验组合。通过不同的初始采样方式得到相同数量的初始数据集来建立代理模型，然后再通过EI寻找最大反应产率的实验组合。

而对于迁移采样的探究，为了预留出可供迁移的源数据，我们将整个数据集基于芳基卤化物进行划分，每种芳基卤化物都有各自的数据集，共得到15个子数据集(子数据集参数空间相同，包括添加剂，Pd催化剂和碱，23 × 4 × 3 = 276个选项)。我们将每个子数据集都视为一个优化任务，从而实现在针对某个芳基卤化物寻找最优反应组合时，可以自由选择其他芳基卤化物的数据知识进行迁移。我们随机选择芳基卤化物1,9,15的优化任务作为目标任务，每种目标任务随机选择3种其他的芳基卤化物的数据进行知识迁移。

**Figure 4. Algorithm testing under different prior knowledge scenarios based on the cross-coupling reaction dataset.** (a) Optimization of the entire dataset without prior knowledge. The left panel illustrates the optimization process using the EI strategy with different initial sampling methods, while the right panel shows the distribution of the best yield after optimization for 140 samples. (b) Optimization curves using standard BO and transfer BO with prior knowledge from other aryl halide datasets, specifically using Aryl halide 1, 9, and 15 as target tasks. In both (a) and (b), the symbols represent the mean of the aggregated results from 40 separate runs, and the error bars indicate variance.

**Figure 4** (a) 展示了与RDM和LHS相比，SE初始采样下，在完整数据集上BO的优化性能。在50个初始采样后进行EI优化，SE的优化效率更高，并且在40次独立重复运行下，优化结束时止步于低产率的次数更少。（b）分别展示了对于芳基卤化物1,9,15的数据集，基于不同源数据的迁移BO和标准BO的优化曲线，比如tra-BO（2）意思为，以芳基卤化物2的数据进行知识迁移，迁移采样的BO优化曲线。对于芳基卤化物1，三种迁移BO优化效果与标准BO相当，对于芳基卤化物9，三种迁移BO在优化初始阶段就找到的产率就高于标准BO，其中tra-BO（8）的优化效果最好，不到30次迭代就找到了全局最佳产率，对于芳基卤化物15，tra-BO（7）优化效果与标准BO相当，tra-BO（8）优化效果最好，tra-BO（11）在优化初始阶段优于标准BO，在20次迭代后与标准BO优化效果相当。

## 模拟合金蠕变寿命优化

第二个实验数据集是合金蠕变寿命数据集，它来自于Refs.以及 NIMS公共数据库，包括871个样品，每个样品特征包括化学成分(22个元素)、热处理工艺(9个参数)、实验条件(2个参数)等共33个，目标特征为蠕变断裂寿命，因为蠕变断裂寿命偏小，我们使用该属性的对数，即lg(蠕变断裂寿命)作为输出变量。我们的目标就是最大化蠕变断裂寿命。本工作在上述特征上用matminer描述符对化学成分进行描述，并将成分描述符与工艺变量组合一起进行minmax归一化和pca降维。

**表2，高温合金蠕变寿命数据集情况**

同样，对于初始采样的探究，我们以整个数据集作为BO任务，去寻找最蠕变断裂寿命的参数组合。通过不同的初始采样方式得到相同数量的初始数据集来建立代理模型，然后再通过EI寻找最大反应产率的实验组合。

而对于迁移采样的探究，我们基于主要元素比重将871个样本的数据集划分为了铁基合金72个，钴基合金139个，镍基合金536个，钛基合金88个。在对不同合金数据进行优化时，可以选择其他合金的数据知识进行迁移。

**Figure 5. Algorithm testing under different prior knowledge scenarios based on the alloy creep rupture life dataset.** (a) Optimization of the entire dataset without prior knowledge. The left panel illustrates the optimization process using the EI strategy with different initial sampling methods, while the right panel shows the distribution of the best creep rupture life values after optimization for 140 samples. (b) Optimization curves using standard BO and transfer BO with prior knowledge from different alloy datasets, specifically using Fe-base, Ni-base, and Ti-base as target tasks. In both (a) and (b), the symbols represent the mean of the aggregated results from 40 separate runs, and the error bars indicate variance.

**Figure 5** (a) 展示了在完整合金数据集上，不同初始采样方法的BO的优化性能。三种采样方法的均值优化曲线虽然相差不大，但通过比较优化结束时的分布图，SE分布更为集中。（b）分别展示了以铁基，钴基，镍基合金数据集的优化作为目标任务，其他合金数据作为源数据，迁移BO和标准BO的优化效果，对于铁基合金和镍基合金的优化，迁移BO和标准BO没有明显区别。但对于钛集合金的优化，三种迁移BO明显优于标准BO，其中迁移镍合金数据的迁移BO优化效果最好。通过分析数据，发可以现钛合金的全局最优在8.5左右，而铁合金和镍合金的全局左右都在11.7左右，以上迁移效果的不同可能和不同合金数据全局最优和局部最优的分布不同有关，其他合金数据的知识可以帮助钛合金寻找全局最优，而钛合金的全局最优在其他合金上可能只是局部最优。

# Summary ＆ Conclusions

综上所述，我们针对开展BO前的不同情况提出了对应策略，没有源数据的基于欧氏距离初始采样SE和有源数据时的迁移采样。SE采样相比于常用的两种初始采样策略RDM和LHS，可以实现全局均匀的采样。而对于有源数据的情况，通过迁移采样可以有效迁移源数据知识，帮助目标任务的优化。

通过在测试函数，交叉偶联数据集，合金蠕变寿命数据集上进行模拟优化，实验表明，基于SE采样进行建模，相同采样数量下相比RDM和LHS，模型性能更加稳定，对后续的EI寻优也有积极的影响。通过迁移采样进行的迁移BO，相比于普通BO在相同的采样数量下可有效提高找到全局最优的概率。

# Reference