**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策略，均匀采样和迁移采样。我们在测试函数，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。

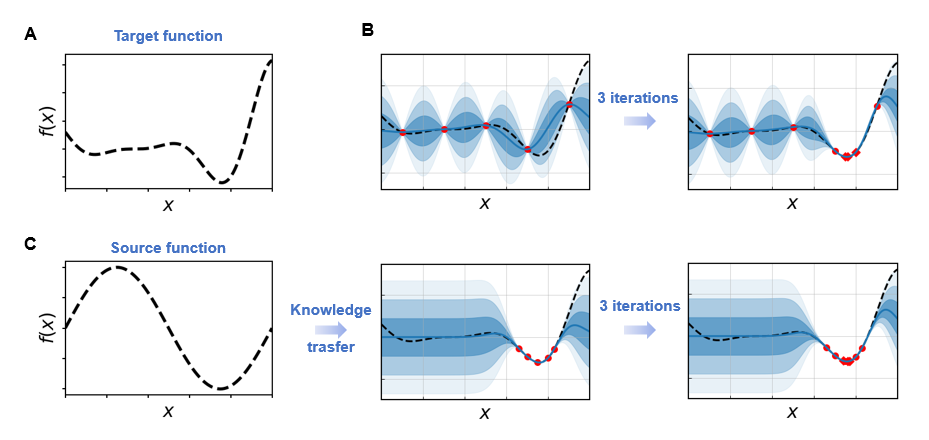
# Introduction

基于数据驱动的机器学习辅助材料研发成为变革传统试错法的新模式。作为顺序机器学习方法之一，贝叶斯优化（BO）因其对评估昂贵的黑箱函数具有高效的取样优化能力而备受关注，它利用代理模型来近似实现从实验参数到目标变量的映射，结合获取函数，提供最优的实验建议。实验表明，BO已经成为探索广泛的化学反应合成和材料优化的有效优化策略。

一般来说，机器学习模型的性能依赖于足够的数据知识。最近的研究表明，因为数据冗余的存在，数据知识的多少并不完全等同于数据量的大小。如何进行有效的数据采样，进而提升BO算法在材料优化的能力，是当前BO的一大挑战。

最直接的优化方法是对采样策略进行调整。Siemenn通过修剪搜索边界和设计自适应获取函数，实现“大海捞针”类材料问题的快速优化。在优化过程中引入专家知识，为贝叶斯优化打开了新思路。Zhe liu通过引入专家反馈以及先前数据作为概率约束帮助模型决策，在有限实验预算下大幅提高了钙钛矿太阳能电池的效率。

然而，先前的工作很少考虑初始采样策略对整体优化任务的影响。实际上，初始数据的分布情况决定着初始代理模型的性能以及后续的优化走向，对BO的优化效果至关重要。对于目标优化任务，启动方式可以根据是否有可利用的源数据分为两种情况：（a）无源数据时的初始化采样：在没有任何源数据的情况下，需要从头开始进行初始化采样。研究人员往往会选择随机采样或者拉丁超立方采样（LHS），然而随机采样有很大的随机性，LHS不能保证全局搜索空间采样的均匀性，因为其原理在每个维度上进行独立的分层采样，然后将这些层组合成采样点，这意味着它只能在单个维度上的采样是均匀的。举个例子，假设我们有一个 3\*3的网格，并使用 LHS 进行采样，如果一个采样点落在 (1,1)子网格，那么在同一行的(1,2)和(1,3)子网格，以及在同一列的(2,1)和(3,1) 子网格都不会再有采样点。（b）有源数据时的初始化采样：在有源数据可供利用的情况下，可以利用参考数据中的知识来帮助目标任务的初始化采样。（源数据指的是可以用于知识迁移的已知实验数据，例如，当出现新的反应物时，专家化学家可能会根据他们过去优化类似反应的经验，对哪种反应条件或催化剂可能导致高产量产生一种直觉，这个过程就属于知识迁移，而过去的反应数据就是源数据，源数据建立的模型成为源模型。）



**Figure 1**

在这项工作中，我们进行了以下研究：（a）在无源数据的情况下，提出基于欧氏距离的初始采样策略（SE），相比于随机采样和LHS，SE可以实现全局范围的均匀采样，基于SE得到的数据进行代理模型初始化，模型性能更加稳定，减少了优化过程中陷入局部极值的可能。（b）在有源数据的情况下，通过迁移学习策略迁移源数据知识帮助目标任务的采样，减少寻找全局最优所需的实验次数。该方法特点在于可以随着迭代过程分析源模型和目标任务的相关性，根据相关性赋予源模型知识迁移的程度，实现了可靠的迁移采样贝叶斯优化。

这项工作的组织如下，在第二节我们介绍了所用到的SE采样和迁移采样方法，在第三节用Goldstein-Price函数进行了方法演示，在第四节用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。最后我们提供的证据表明在不同初始情况下，通过SE采样和迁移采样策略，可以更好地帮助我们完成目标任务的优化。

# 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

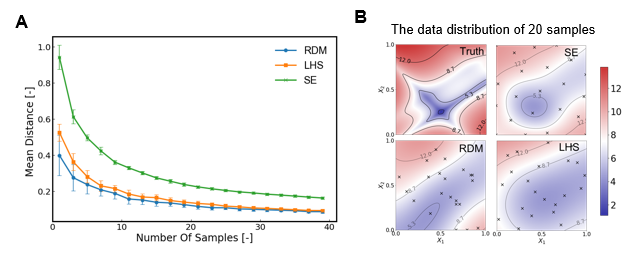
通过迁移获取函数指导目标任务的下一步采样，从而实现将信息从源任务转移到目标任务。

# Demonstration

在本章节，我们以Goldstien-Price (2-dimensions)来展示我们方法的效果，该测试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。

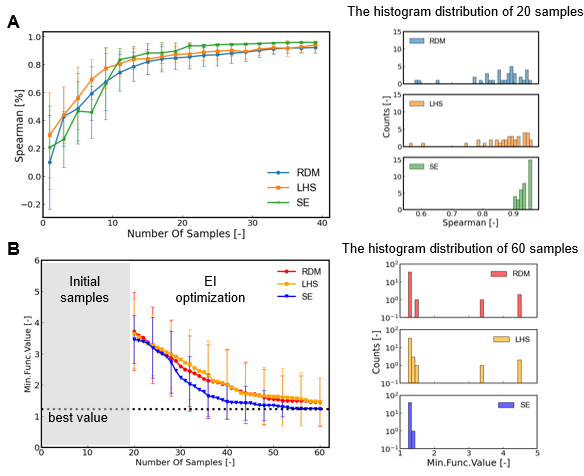
## 无源数据-均匀采样

在这里我们从三个角度对不同初始采样方法进行分析。首先通过不断迭代取点，查看各个取样方法的散度度量变化，散度度量是采样数据集中每个样本点与其欧氏距离最近点的平均距离，并对样品点的位置进行了可视化。随后我们比较了不同采样方式取点下建模，代理模型的spearman指标的均值和方差。最后我们选择了合适的初始采样点数量，探究不同采样方式对BO优化效果的影响，以上每次迭代取一个点，并进行了40次的重复。



**Figure 2**

Figure 2中A显示了不同采样方式下采样点的散度度量，可以发现随着采样点越来越多，采样点的平均距离逐渐减小，而SE减小的最慢，这意味着基于SE得到的采样点更加离散。B的左上为Goldstien-Price函数的真实二维分布，z轴为y值大小，另外三张图为不同采样方式取得20个采样点的分布情况，采样点用叉号表示，我们以这20个点建立高斯过程回归模型，对整个参数空间进行预测，z轴为预测值大小。可以看到SE采样点更加均匀，同时其模型的预测也与真实情况更加接近。

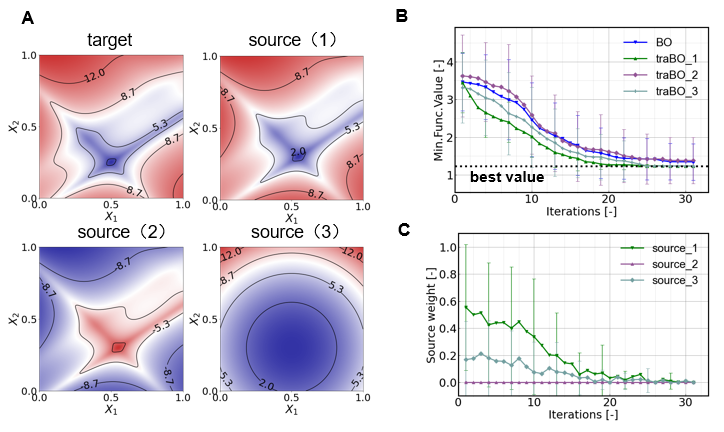


**Figure 3**

Figure 3中A的左图显示了不同采样方式取点下建模，代理模型的spearman指标变化，当采样数量大于15个的时候，基于SE的模型spearman均值相较于LHS和RDM，均值更大，方差更小，A的右图显示了20个采样数量时，40次重复实验的模型spearman指标分布直方图，更直观地显示了SE取样建模的优势。B的左图显示了基于三种采样策略得到20个采样点后，再通过EI去寻找全局最优，纵轴意为当前找到的最小值。B的右图显示了得到60个采样点情况下，当前最优的分布直方图，可以发现，基于SE取样后进行寻优，可以有效提高找到全局最优的次数，降低了优化失败的概率。

## 有源数据-迁移采样

为了测试迁移优化方法在不同源数据情况下的效果，我们设计了不同类型的源函数来生成源数据集，并用源数据建立源模型，在迁移策略下帮助指导目标任务的寻优。我们比较了不同源数据下的迁移BO和没有源数据的BO的优化曲线，并比较了不同源模型随着迭代过程的权重变化。标准BO随机取样3个后用EI进行取样优化，迁移BO随机取样3个点后用traEI进行取样优化，代理模型均为高斯过程回归模型，所有实验独立运行40次。



**Figure 4**

Figure 4中A显示了目标函数和三个源函数的二维分布情况，源函数1是通过对目标函数平移和扰动得到，源函数2是通过对源函数1取反得到，而源函数3为我们设计的一个漏斗形函数。B显示了不同源函数的迁移BO和没有源函数的标准BO的优化曲线，traBO-1通过21次取样基本全部找到全局，traBO-3需要24次，traBO-2和BO在30次取样后仍有一部分未找到全局最优。这说明通过迁移学习可以从源函数1和3学到知识帮助目标函数的优化。C显示了随着迭代过程，3个源模型的权重变化，可以看到在优化前期，源模型权重1>3>2，这与B的结论相吻合。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

我们选用了两种真实世界的材料数据集评估来我们的算法表现，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92维的原子，分子和振动描述符来对每个选项进行特征描述，本工作在原有特征上进行了minmax归一化和pca降维，最终得到8维特征和1维目标变量（产率）。

图5展示了Pd催化的Buchwald-Hartwig反应方案和实验组合空间。

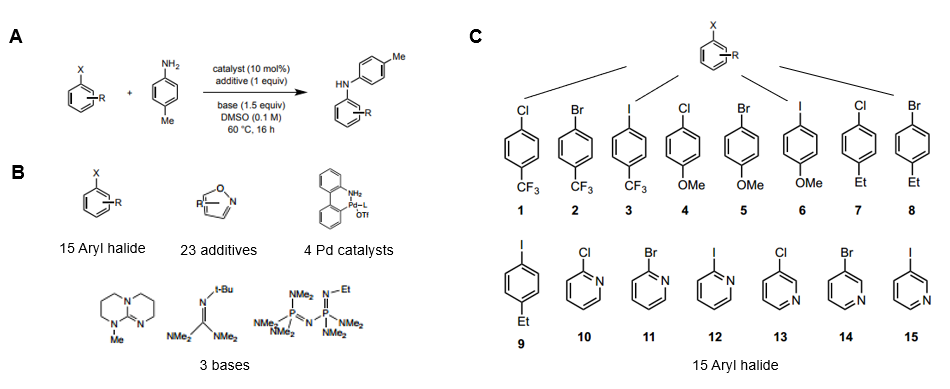


图5

### 无源数据-均匀采样

在上述3960个实验组合中，我们的BO任务是最大化反应产率，首先通过不同的初始采样方式进行取样，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。

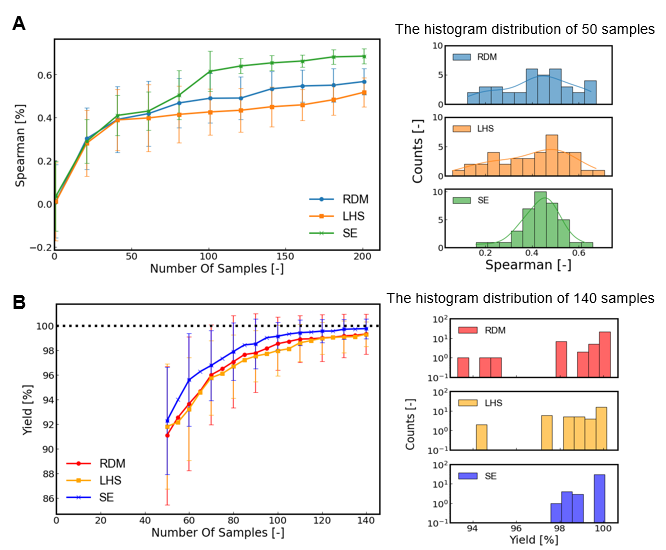


图6

图6中的A为不同采样下代理模型的Spearman指标变化，以及采样数量为50个时，Spearman指标的分布直方图。可以看出通过SE取样来建立初始模型，模型性能更加稳定。B为不同初始采样方式下采样50个点，后续用EI进行优化的优化折线图，左边为140个采样数量下当前最优产率的分布直方图，基于SE初始取样后可以有效提高找到全局最优的概率。

### 有源数据-迁移采样

我们将每个芳基卤化物视为一个单独的优化任务，因此上述上述数据集可划分为15个子数据集(添加剂，Pd催化剂和碱，23 × 4 × 3 = 276个选项)，我们可以通过迁移其他芳基卤化物的数据知识，帮助完成目标卤化物的优化。我们随机选择芳基卤化物1,9,15作为目标任务（目标任务为通过不断采样找到最大产率的实验组合），每种目标任务随机选择3种其他的芳基卤化物作为源域进行迁移。

图7显示了三种目标任务不同优化策略的优化结果。对于芳基卤化物1，迁移BO与标准BO优化效果相当，说明在迁移优化过程中并没有学到相关知识；对于芳基卤化物9，迁移BO（8）和迁移BO（14）的优化效果明显优于标准BO，我们统计了迭代结束时40次重复实验找到全局最优的次数和比例，迁移BO（8）找到了39次，迁移BO（14）找到了35次，标准BO只有19次；对于芳基卤化物15，三种迁移策略都优于标准BO，标准BO在40次迭代时找到全局最优次数为12次，迁移BO分别为17次，38次和21次。

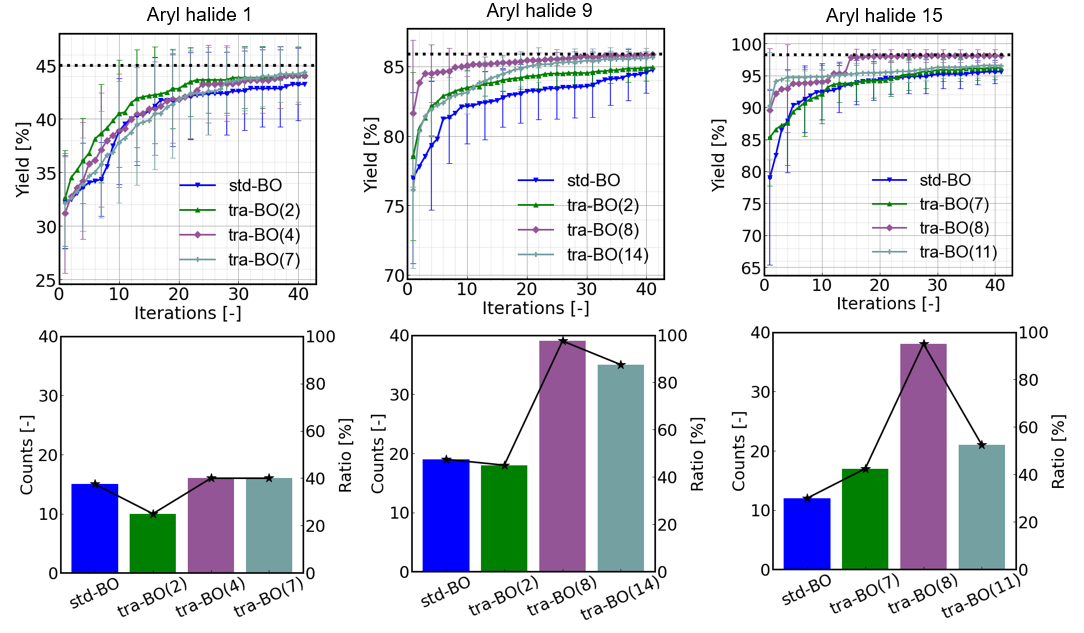


图7

## 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，最终得到9维特征和1维目标变量（蠕变寿命）。

### 无源数据-均匀采样

在上述871个实验组合中，我们的BO任务是蠕变断裂寿命，首先通过不同的初始采样方式进行取样，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。

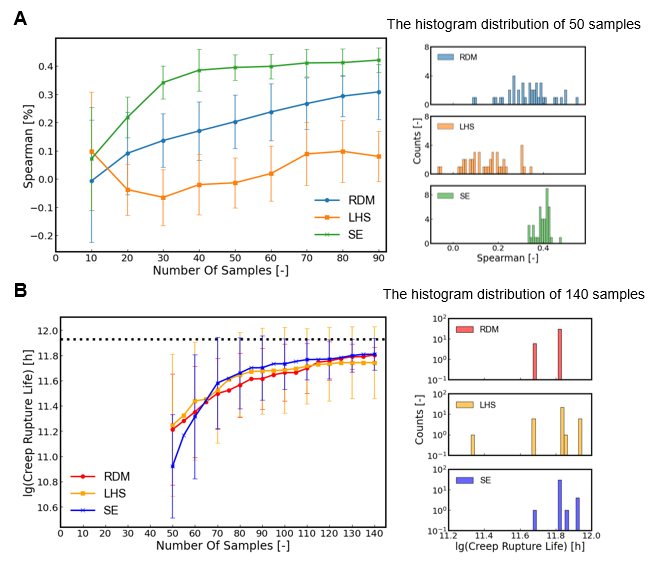


图8

图8中A的左图展示了SE，RDM和LHS三种采样方式下得到不同数量样本对代理模型spearman指标的影响，右图为50个样本时，40次重复实验模型spearman指标的分布情况。可以看到基于SE采样建模。模型spearman均值更大，方差更小。B为不同初始采样方式下采样50个点，后续用EI进行优化的优化折线图，左边为140个采样数量下当前最优产率的分布直方图，基于SE初始取样后再进行寻优可以提高当前最优值。

### 有源数据-迁移采样

为了测试我们迁移方法的效果，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，我们分别以Fe基合金，Ni基合金,Ti基合金作为目标优化任务，其他合金作为源任务，实施迁移贝叶斯优化，并和标准BO进行比较。

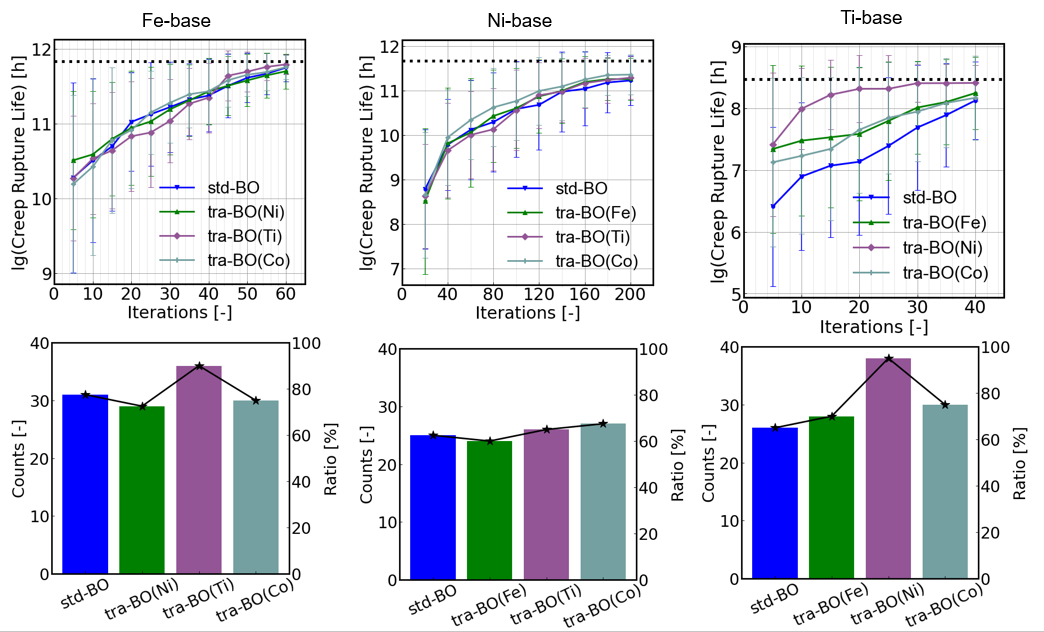


图9

图9展示不同合金作为目标优化任务的优化结果，上半部分为优化曲线，下半部分为优化结束时找到的全局最优的数量和比例。对于铁基合金，采样60个点后，40次重复实验标准BO有31次找到全局最优，而迁移效果最好的tra-BO（Ti）找到了36次。对于镍基合金，采样200个点后，40次重复实验标准BO有25次找到全局最优，而迁移效果最好的tra-BO（Co）找到了27次。对于Ti基合金，采样40个点后，40次重复实验标准BO有26次找到全局最优，而迁移效果最好的tra-BO（Ni）找到了38次。

# 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