

主动学习辅助材料研发进展

王云帆¹, 田 原², 周玉美¹, 薛德祯¹

(1. 西安交通大学, 金属材料强度国家重点实验室, 西安 710049; 2. 上海大学, 材料基因组工程研究院, 上海 200444)

摘 要: 材料研发面临着巨大、复杂、高维的搜索空间, 如何从中快速有效选择具有目标性能的新材料是加速材料研发的主要挑战。机器学习可以基于已有的数据, 通过算法建立特征和目标性能之间的映射关系, 对未探索材料的性能进行预测。但是材料的已知数据比较少, 建立的机器学习模型通常预测精度低, 难以实现对实验或计算的有效指导。为了解决上述问题, 引入主动学习的方法, 在传统迭代反馈的基础上, 增加了实验设计的步骤, 选择对目标提升最有帮助、带有更大信息量的实验进行补充, 达到更快的优化材料性能的目的。主要从单目标优化、多目标优化、曲线优化 3 个方面, 回顾了主动学习辅助材料研发的进展。

关键词: 材料开发; 主动学习; 机器学习; 不确定性

中图分类号: TB3; TP3 文献标志码: A 文章编号: 0454-5648(2023)02-0544-08

网络出版时间: 2023-01-17



Progress on Active Learning Assisted Materials Discovery

WANG Yunfan¹, TIAN Yuan², ZHOU Yumei¹, XUE Dezhen¹

(1. State Key Laboratory for Mechanical Behavior of Materials, Xi'an Jiaotong University, Xi'an 710049, China;

2. Materials Genome Institute, Shanghai University, Shanghai 200444, China)

Abstract: Materials discovery faces a huge and complex high-dimensional search space, from which the fast and effective selection of new materials with target properties is a major challenge in materials development. Machine learning can predict the performance of unexplored materials via establishing the relationship between features and target performance through algorithms based on the existing data. However, there are a relatively few known data for materials, and the machine learning models have a relatively low prediction accuracy, thus making it difficult to achieve an effective guidance for experiments or calculations. To address this problem, active learning was introduced for assistance, and the experimental design step was added to the traditional iterative feedback to select the experiments for target enhancement to supplement and to achieve the optimization of material performance. This review mainly represented recent progress on active learning-assisted materials development from three aspects, i.e., single-objective optimization, multi-objective optimization, and curve optimization.

Keywords: materials discovery; active learning; machine learning; uncertainty

新材料的研发与应用是推动现代科学发展与技术革新的源动力之一, 但是新材料研发过程耗时耗力, 从研发到投入市场应用需要 15~25 年甚至更长的时间^[1]。随着高熵、低维等材料设计新理念的提出, 调控材料性能的手段越来越复杂多样, 传统的基于试错和经验积累的方法不再适用。为了应对上

述挑战, 美国、欧洲、日本以及中国都相继启动了“材料基因组计划”等相似的国家战略, 力图通过高通量材料计算、高通量实验和材料数据技术等实现材料研发到应用的时间与成本双减半^[2]。

材料的性能与化学成分、晶体结构、相组成、微观组织、加工工艺、服役条件等因素都息息相关,

收稿日期: 2022-10-27。 修订日期: 2022-11-18。

基金项目: 科技部重点研发计划(2021YFB3802102)。

第一作者: 王云帆(1996—), 女, 博士研究生。

通信作者: 薛德祯(1984—), 男, 博士, 教授。

Received date: 2022-10-27. Revised date: 2022-11-18.

First author: WANG Yunfan (1996—), female, Doctoral candidate.

E-mail: wangyunfan@stu.xjtu.edu.cn

Correspondent author: XUE Dezhen (1984—), male, Ph.D., Professor.

E-mail: xuedezhen@xjtu.edu.cn

导致了新材料开发面临着巨大、复杂、高维的搜索空间, 如何从这一搜索空间中快速有效选择具有目标性能的新材料是提高材料研发效率、加速材料研发过程的主要挑战^[3]。一方面可以通过高通量计算、高通量实验来减少材料设计、制备以及表征的时间, 从而在短时间内高效筛选搜索空间。另一方面, 可以利用数据驱动的方法建立材料性能预测模型, 指导材料设计, 通过少量实验或计算的反馈, 高效地开发具有目标性能的新材料。

机器学习可以基于已有的数据, 通过算法建立特征变量和目标性能之间的映射关系, 对未探索材料的性能进行预测^[4], 目前已经被大量应用于新材料的设计与研发。例如, 基于密度泛函理论(DFT)计算数据, Weng 等^[5]利用符号回归方法建立了一个简单的材料描述符 μ/t , 其中 μ 和 t 分别是八面体和容差因子; 这一描述符容易计算获得, 从而加速了具有高析氧反应活性的新型钙钛矿氧化物催化剂的发现。Back 等^[6]同样利用 DFT 计算数据, 建立了基于卷积神经网络的性能预测模型, 准确预测了 Ir 基催化剂的结合自由能, 有效减少了密度泛函理论计算量, 提升了新材料开发效率。但是机器学习方法通常需要基于大量的数据, 建立具有高精度的代理模型对材料性能进行预测。当材料数据样本量较少时, 建立的机器学习模型通常预测精度低, 如果仅仅依据不准确的预测值设计与筛选材料, 难以实现对实验或计算的有效指导。

为了解决上述问题, 可以引入主动学习的方法, 也称适应性设计, 这一方法在依据机器学习预测值进行迭代反馈的基础上, 增加了实验设计的步骤。通过机器学习代理模型预测未探索材料的性能, 并估计预测值所带有的不确定性, 在此基础上进行实验设计。实验设计的目的是选择对目标性能提升最有帮助、带有更大信息量的样本进行补充, 以更快地优化材料性能。这一迭代更新的思路与 Bayes 优化相似, 所应用到的数据技术也类似, 本文中统一称为主动学习。

本综述在第 1 节介绍了主动学习框架, 以及常用的效用函数和不确定性评估方法。第 2 节主要从单目标优化、多目标优化、曲线优化 3 个方面, 对主动学习辅助材料研发的应用进展进行了回顾, 最后对目前主动学习辅助材料研发面临的困难与挑战进行了讨论与展望。

1 主动学习

1.1 主动学习框架

主动学习的框架如图 1^[7]所示, 主要分为以下几个步骤: 首先, 建立包括材料特征参量以及目标性能在内的初始数据集; 其次, 基于初始数据建立代理模型, 对搜索空间中材料的性能进行预测, 除了需要给出性能的预测值之外, 还需要评估预测值带有的不确定性。通常, Gaussian 过程回归(GPR)模型或克里金模型可以直接得到预测值及其方差, 而其他机器学习模型, 如线性回归、支持向量回归(SVR)等, 则需要通过自助抽样获得不确定性, 具体的不确定性评估的方法将在 1.3 节详细介绍; 再次, 通过定义某一效用函数, 计算搜索空间所有材料的效用函数值并对其进行排序, 从而选择下一步要进行的实验; 考虑性能预测值及其不确定性的权重, 能够应用的选择策略主要包括开发、探索以及平衡开发与探索等, 这些策略将在 1.2 节详细介绍; 最后, 对选择的材料进行合成与表征, 或者计算模拟, 将得到的性能结果反馈回数据集中; 重复这个循环, 直到找到具有目标性能的材料。区别于另一类基于流的主动学习(Sequential Active Learning), 上述流程为基于池的主动学习(Pool-based Active Learning)模式, 这也是目前在材料研发中应用较多的主动学习研发模式。

1.2 实验设计方法(效用函数)

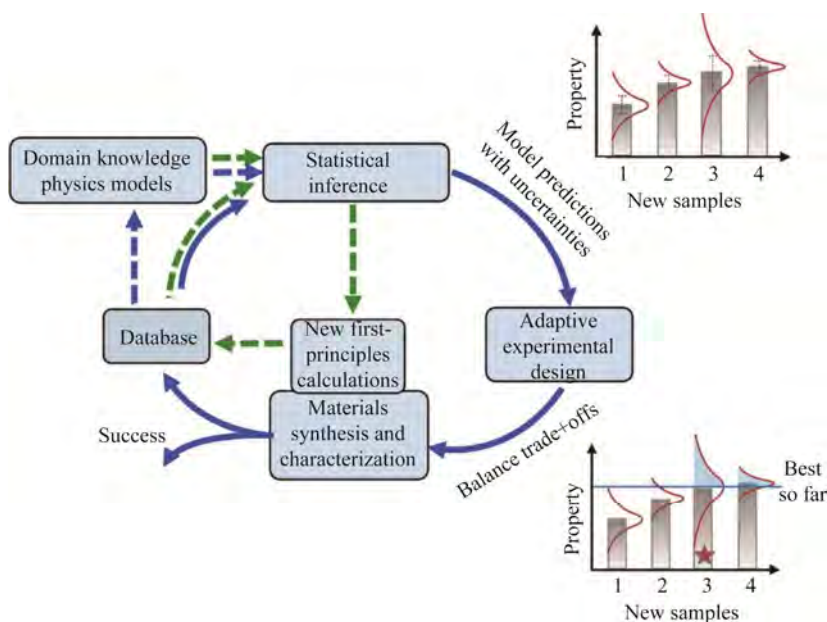
在实验设计过程中, 定义一个效用函数 U , 计算并基于该函数值对所有未探索材料样本进行排序, 一般选择效用函数值最大的样本作为下一步计算或者实验的对象。

本节将介绍目前在材料研发中常用的效用函数, 包括开发, 探索以及平衡开发和探索 3 个不同的设计策略。定义 y 为目标性能; x 为材料特征参量, 它能够唯一地描述一个材料; μ 是模型预测的均值; σ 是模型预测值的标准差。一般优化目标为最大化目标性能 y 。

纯开发(Pure Exploitation)策略: 认为模型准确, 选择未探索材料中性能预测值最大的点, 因此效用函数是模型预测值, 或预测的均值,

$$U = \mu \quad (1)$$

纯探索(Pure Exploration)策略: 认为模型不准确, 选择未探索材料中预测值标准差最大的点(即不确定性最大的点), 因此, 效用函数是模型预测值的标准差,



It is composed of dataset, surrogate model, adaptive design and experimental or calculation feedback. The loop is iteratively conducted to optimize the material property. Uncertainties are employed to trade-off exploitation and exploration of the search space of unexplored materials to select the next best experiment or calculation.

图1 基于池的主动学习流程图^[7]

Fig. 1 Schematic of pool-based active learning for materials design^[7]

$$U = \sigma \quad (2)$$

平衡开发与探索策略：将上述两个策略进行权衡，既考虑预测值，又考虑不确定性。一般认为预测值 y 服从均值为 μ ，方差为 σ^2 的正态分布，效用函数可以定义为上置信界限(UCB)^[8]：

$$U_{\text{UCB}} = \mu + \kappa \sigma \quad (3)$$

其中 κ 是一个超参数，能够调整开发与探索的权重。 κ 越大，越偏向于探索，通常 κ 是迭代次数的函数，随着迭代次数的增加， κ 逐渐减小。

另一个非常有效的材料设计效用函数是期望提升(EI)^[9]。如果当前最大值为 μ^* ，它计算了未探索材料的性能对 μ^* 的提升， U_{EI} 可以通过下式计算：

$$U_{\text{EI}} = E[\max(y - \mu^*, 0)] = \sigma[\phi(z) + z\Phi(z)] \quad (4)$$

其中： $z = (\mu - \mu^*) / \sigma$ ； $\phi(z)$ 和 $\Phi(z)$ 分别是标准正态分布的概率密度函数和累积分布函数。

针对同一个优化问题，不同的效用函数推荐的下一步样本点并不相同。图2展示了一个一维的优化案例，基于5个训练数据，快速寻找最大的 y 对应的 x 值。图2中黑点表示训练集中的数据，蓝色实线表示模型预测值，灰色区域表示95%置信区间，黑色虚线表示当前最大值，下方的彩色曲线为每个未探索 x 对应的效用函数值。由图2可见，即使是对于同为平衡开发与探索策略的效用函数 $U = U_{\text{EI}}$

和 $U = U_{\text{UCB}}$ ，其最大值对应的 x 并不相同。因此，在实际材料性能优化过程中，需要选择合适于特定问题的效用函数。

1.3 不确定性评估方法

虽然 Gaussian 过程等少量模型可以给出预测值的不确定性，但大多数机器学习模型无法直接对不确定性进行估计，因此，对这类模型预测值的不确定性进行评估，对于应用主动学习至关重要。Tian 等^[10]研究了不同的不确定性评估方法对主动学习优化材料性能效率的影响规律。本文总结了其中2类典型的不确定性评估方法，即经验标准差方法和刀切法。

通过对样本量为 n 的训练数据集进行 n 次可放回抽样，获得一个重采样数据集，基于此数据集建立机器学习模型，对未探索材料样本 x^* 的性能做出预测，得到预测值 $\mu^{*(1)}$ 。重复上述过程 B 次，获得样本 x^* 的一系列预测值 $\mu^{*(1)}, \dots, \mu^{*(B)}$ ，这一方法称为自助抽样法(Bootstrap)。

经验标准差法(Empirical sd)^[11]通过计算 $\mu^{*(1)}, \dots, \mu^{*(B)}$ 的标准差来评估与性能预测的均值 $\bar{\mu}^*$ 相关的不确定性：

$$\widehat{se}_{\text{boot}} = \sqrt{\frac{1}{B-1} \sum_{b=1}^B (\mu^{*(b)} - \bar{\mu}^*)^2} \quad (5)$$

刀切法(Jackknife)^[12]每次从数据集中删除一个观测值, 然后使用剩余的观测值计算标准差。标准差 $\widehat{se}_{JAB}^B$ (Jackknife-After-Bootstrap) 是通过将刀切法应用于自助抽样法获得的一系列预测值来评估的:

$$\widehat{se}_{JAB}^B = \sqrt{\frac{n-1}{n} \sum_{i=1}^n (\mu_{(-i)}^* - \bar{\mu}^*)^2} \quad (6)$$

其中, $\mu_{(-i)}^*$ 是除去包含第 i 个观测值的自助抽样数

据集后, 预测得到的 $\mu^{*(1)}, \dots, \mu^{*(B)}$ 的均值。

另一个基于刀切法的评估方法是无穷小刀切法 (Infinitesimal Jackknife, IJ)^[13], 每个观测值的权重都是无穷小的:

$$\widehat{se}_{IJ}^B = \sqrt{\sum_{i=1}^n \frac{\sum_b (N_{bi}^* - 1)(\mu^{*(b)} - \bar{\mu}^*)}{B}} \quad (7)$$

其中, N_{bi}^* 是训练集中的第 i 个观测值出现在自助抽样数据集的次数。

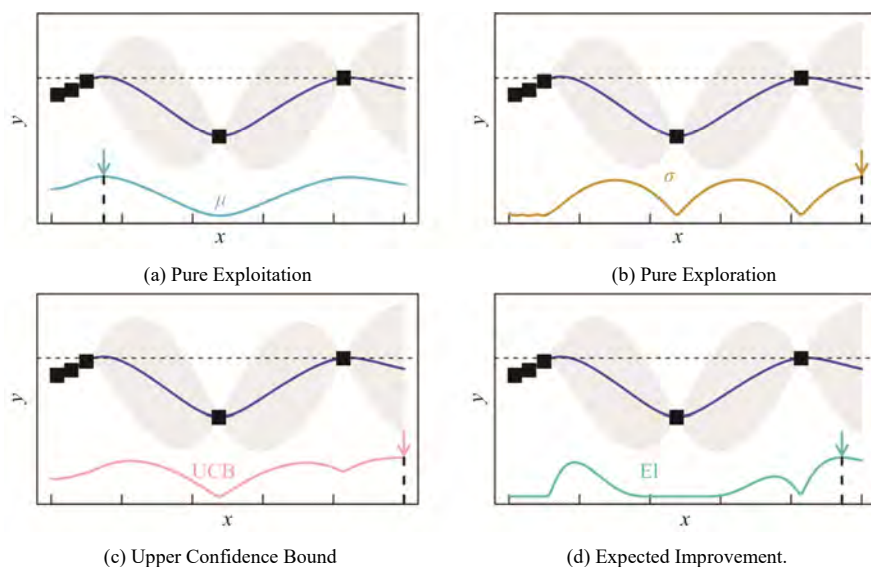


图 2 效用函数示意图

Fig. 2 Schematic diagram of the 4 utility functions

2 主动学习在材料研发中的应用

材料研发中涉及到不同的优化问题。开发高强度合金、高活性催化剂等需要对某一特定性能进行优化; 结构材料的强韧化、热电材料高导电高导热的匹配需要同步优化 2 个或 2 个以上的性能; 在涉及到相图建立、疲劳曲线建立等问题时, 则需要快速优化得到某一性能变化的曲线或曲面。本小节回顾了主动学习在单目标优化、多目标优化、曲线优化 3 个方面的应用效果。

2.1 单目标优化

基于机器学习预测模型, 应用 1.2 节中介绍的效用函数可以实现某一材料性能的优化。不同的机器学习模型与不同效用函数的组合, 对性能的优化效率并不一致, 因此, 选择合适的组合才能够显著提高材料开发效率。

主动学习中的 Bayes 优化一般以 Gaussian 过程为性能预测模型, 结合平衡开发与探索的效用函数, 能够达到较好的优化效果。例如, Flores 等^[14]利用

主动学习框架, 在 956 个可能的结构空间中, 高效地搜索得到具有新晶体结构的电化学稳定氧化铱 (IrO_2 和 IrO_3), 首先使用 GPR 模型训练生成焓 (ΔH_f) 与结构参量之间的映射关系, 预测了搜索空间中所有结构的生成焓; 进而将下置信界限 (LCB), $U = \mu - \kappa\sigma$, 作为效用函数推荐 5 个结构进行 DFT 计算; 将计算结果反馈到数据集中, 更新模型及效用函数值, 反复迭代直至无法找到更稳定的结构。这一主动学习框架仅通过 30 次 DFT 计算就获得了最稳定氧化铱晶型; 与随机搜索相比, 其开发效率提高了至少 2 倍。

同样的优化策略也可以与真实实验相结合, 有效地减少实验次数。例如, Mekki-Berrada 等^[15]将主动学习、深度神经网络与高通量实验相结合, 在由 5 个变量组成的合成条件搜索空间中, 高效开发具有目标吸收光谱特性的银纳米颗粒。建立了 GPR 模型, 以最小化损失函数为目标, 将 U_{EI} [式(4)] 作为效用函数, 推荐 15 个合成条件进行测试; 反馈数据进行迭代, 直到找到具有目标特性的银纳米颗

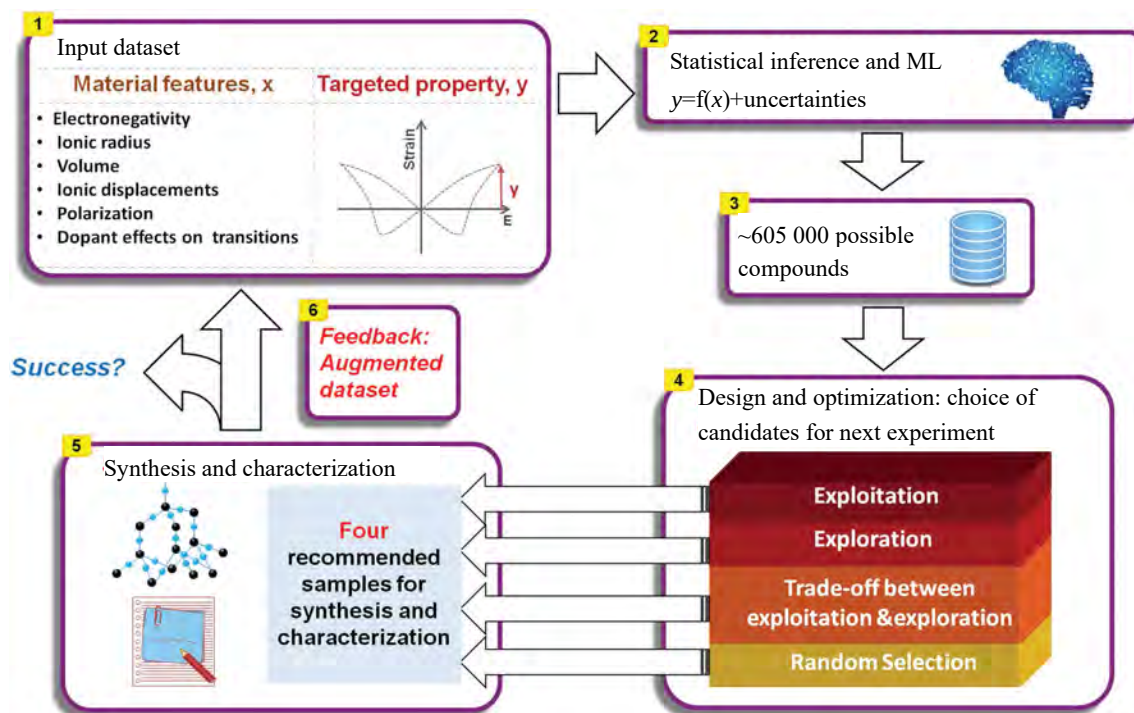
粒; 最佳合成条件在第 8 次迭代时出现。与随机选择的策略相比, 在第 3 次迭代时, U_{EI} 明显开始向较低的损失函数收敛。

上述的优化过程还可以应用于材料性能计算模型的参数优化。Yu 等^[16]利用主动学习框架, 优化了过渡金属一氧化物、钨硫属化物和窄带隙半导体这 3 类材料的 U 参数。利用 GPR 模型, 建立了目标函数和 U 参数之间的关系, 以 U_{UCB} [式(3)] 作为效用函数, 推荐 U 参数进行 DFT 计算; 将结果反馈回模型中, 更新模型和效用函数, 直到达到预设的迭代次数; 与线性响应策略相比, U_{UCB} 策略计算效率高, 其计算成本较线性响应策略降低了 4~9 倍, 而且其计算结果也更准确。

上述工作主要在主动学习中的 Bayes 优化框架下, 利用 GPR 模型和不同的效用函数进行优化, 与不使用效用函数的策略相比, 使用效用函数效率更高, 成本更低。那么, 在 1.2 节介绍的不同效用函数中, 一般应选用哪种效用函数, 以及如何结合传

统机器学习模型, 也是重要的研究课题。

将传统的机器学习模型和不同的效用函数相结合, 平衡开发和探索的效用函数一般能够得到最好的优化效果。如图 3 所示, Yuan 等^[17]比较了纯开发、纯探索、平衡开发和探索以及随机选择 4 种不同的策略在开发具有大电致应变 $BaTiO_3$ 基陶瓷的效率差异。基于 SVR 模型, 建立了电致应变和材料特征参量之间的映射关系, 预测了搜索空间中约 60 万化学成分的电致应变; 进而利用开发, 探索, 平衡开发与探索以及随机选择 4 种策略, 分别推荐 4 种成分的化合物; 将所选的化合物进行合成与表征, 把数据反馈回数据集, 重复这个循环, 直到性能达到目标要求; 比较了 4 种策略在每次迭代时所选的化合物的性能, 平衡开发与探索的效用函数在每次迭代中选择的化合物都拥有着优异的性能; 平衡开发与探索的方法在第 3 次迭代中得到了具有最大电致应变的化合物 ($Ba_{0.84}Ca_{0.16})(Ti_{0.90}Zr_{0.07}Sn_{0.03})O_3$, 其电致应变为 0.23%, 与训练集中的最大值相比提高了 53.3%。



Compounds are synthesized following predictions from four strategies (exploitation, exploration, trade-off between the former two, and random selection), the performance of which are compared.

图 3 基于机器学习和优化实验设计的循环^[17]

Fig. 3 Active learning loop for accelerated discovery based on machine learning and optimal experimental design^[17]

将主动学习与实验或计算相结合, 相比于随机搜索或者网格搜索, 利用效用函数的策略, 能够大大提高材料开发的效率。平衡开发和探索的效用函数表现最好, 但是对于不同的问题, 具体的效用函

数选择不同。

2.2 多目标优化

上述利用主动学习辅助材料研发的研究集中于优化材料的单个性能, 但是在研发新材料时, 往往

需要优化 2 种或者多种性能。然而, 有些性能存在着竞争的关系, 并非所有性能都可以同时改善。Pareto 前沿是定义这些竞争性能之间最佳权衡的解决方案^[18], 但是 Pareto 前沿优化的搜索空间巨大, 一般需要引入主动学习提高 Pareto 前沿优化的效率^[19]。将 1.2 节所述的效用函数中的期望提升(EI)应用于优化 Pareto 前沿, 能够设计出几种不同的适用于多目标优化的效用函数。

将目标定义为 Pareto 前沿下区域的体积(称为超体积), 利用类似于 EI 的思路来评估 Pareto 超体积的预期增加^[20], 进而推荐待测样本。例如, Harada 等^[21]利用多目标优化, 在 47 个成分的搜索空间中同时优化密度和离子电导率 2 个性能, 合成了 47 种 $\text{LiZr}_2(\text{PO}_4)_3$ 基化合物, 测量了它们的密度和离子

电导率 2 个性能; 为了验证多目标优化的效果, 比较了 EI 策略和随机搜索策略, EI 策略能够更快提高超体积; 为了进一步证明 EI 策略的效果, 将搜索空间增加到 225 个样本, EI 策略只需要 89 个实验就能找到 Pareto 前沿, 而随机搜索需要 213 次实验。Wang 等^[22]利用多目标优化, 在 270 个材料搜索空间中, 开发同时具有高功能性和高稳定性的尖晶石, 设计了如图 4 所示的优化流程, 基于 GPR 模型建立成分与分解焓变 ΔH_d 、基态带隙 E_g 2 个性能之间的映射关系, 并使用基于 EI 的效用函数^[23]选择特定成分进行 DFT 模拟评估材料性能; 将结果添加到数据集中, 并更新 GPR 模型; 在 53 次迭代内成功识别了真实 Pareto 前沿的所有 12 种材料。

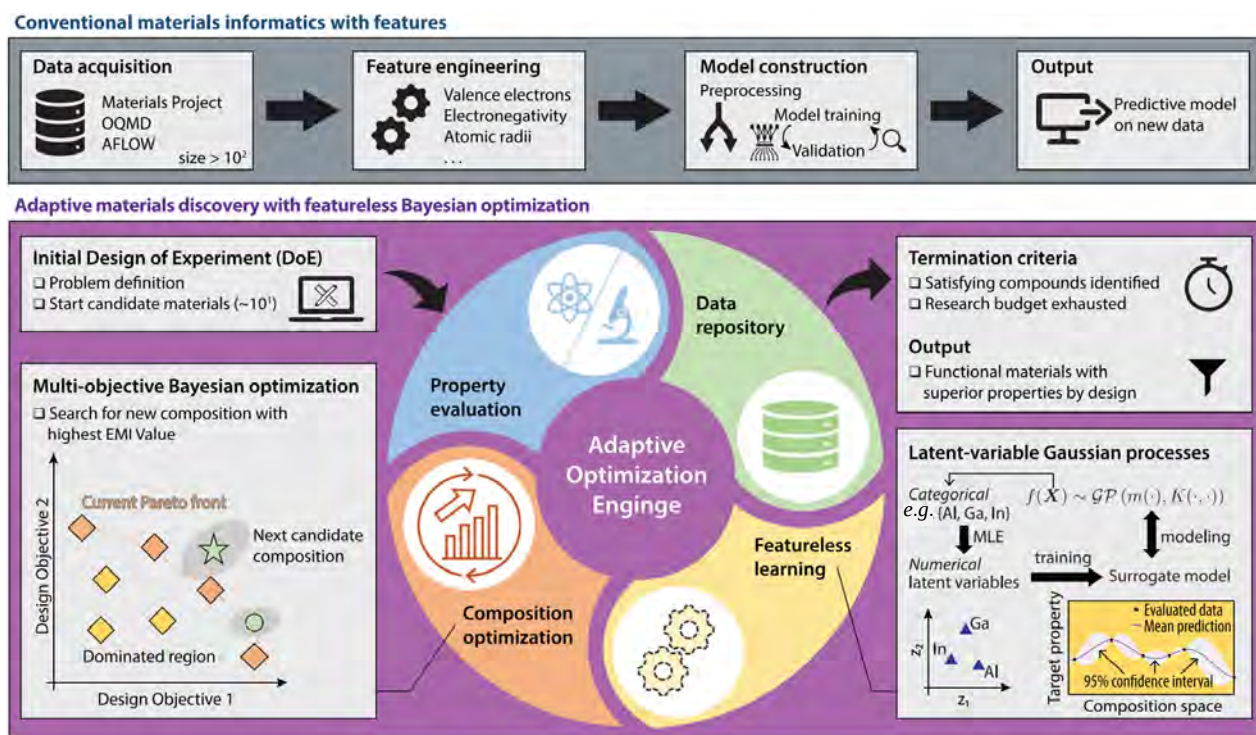


图 4 基于效用函数的多目标性能优化流程^[22]

Fig. 4 Pipeline of multi-objective optimization based on utility functions^[22]

另一个基于 EI 的多目标优化策略是基于欧氏距离的预期改进(EEI)^[24]。Valladares 等^[25]利用多目标优化, 在不同 Ni 含量的成分空间中, 同时优化锂离子电池正极活性材料的高初始比容量和大循环寿命 2 个性能, 建立了 GPR 模型, 分别利用 EEI 函数(整体提高当前 Pareto 前沿)和基于目标的 EEI 函数(改进规定目标附近的 Pareto 前沿), 选择了 6 种 Ni 含量的电池正极材料; 同时测试这些材料, 可以在

一次迭代中改进当前 Pareto 前沿的几个区域。

将主动学习应用于多目标优化问题, 尤其是目标存在相互竞争关系时, 基于平衡开发和探索发展的效用函数能够快速改进 Pareto 前沿, 优于随机搜索等不使用效用函数的策略。

2.3 曲线优化

材料的特性曲线在新材料的设计, 危害评估和工艺参数的优化中起着至关重要的作用。确定性能

曲线需要进行大量的测量或计算, 通常耗时耗力, 一般需要快速建立材料特性曲线。将上述主动学习框架应用于曲线快速构建, 也可以取得较好的效果, 但不同的效用函数对曲线优化的效率不同。

Tian 等^[26]比较了 6 种效用函数在不同材料曲线/曲面优化问题中的效率高, 并根据每个函数所需的新实验数量来估计效率。Tian 等^[26]建立克里金模型预测曲线/曲面, 分别利用纯开发、最大方差(Max.v)、有效全局优化、基于自助抽样的有效全局优化(B.EGO)、知识梯度、随机这 6 种方法推荐下一个实验点。结果表明 Max.v 和 B.EGO 可以在相对较少的迭代中使预测曲线与实际曲线的偏差大幅减小,

从而更快收敛。对于不同的材料数据和复杂性的问题, 最大方差的效用函数表现一般比其他策略更好。

将上述曲线/曲面估计的方法应用于相图预测, Tian 等^[27]通过建立成分参量及相变温度之间的关系, 能够快速建立 BaTiO₃ 基铁电陶瓷相图, 使用了如图 5 所示的设计策略。首先使用了 3 种机器学习策略, 预测和筛选所有未探索的相图; 利用最大方差 Max.v 这一效用函数, 选择合成与测试的组分, 对其进行表征, 得到相变温度, 并对初始训练数据集进行扩充; 重新训练回归模型, 预测相边界; 通过仅 3 次实验, 快速建立了具有三相共存区的铁电陶瓷相图。

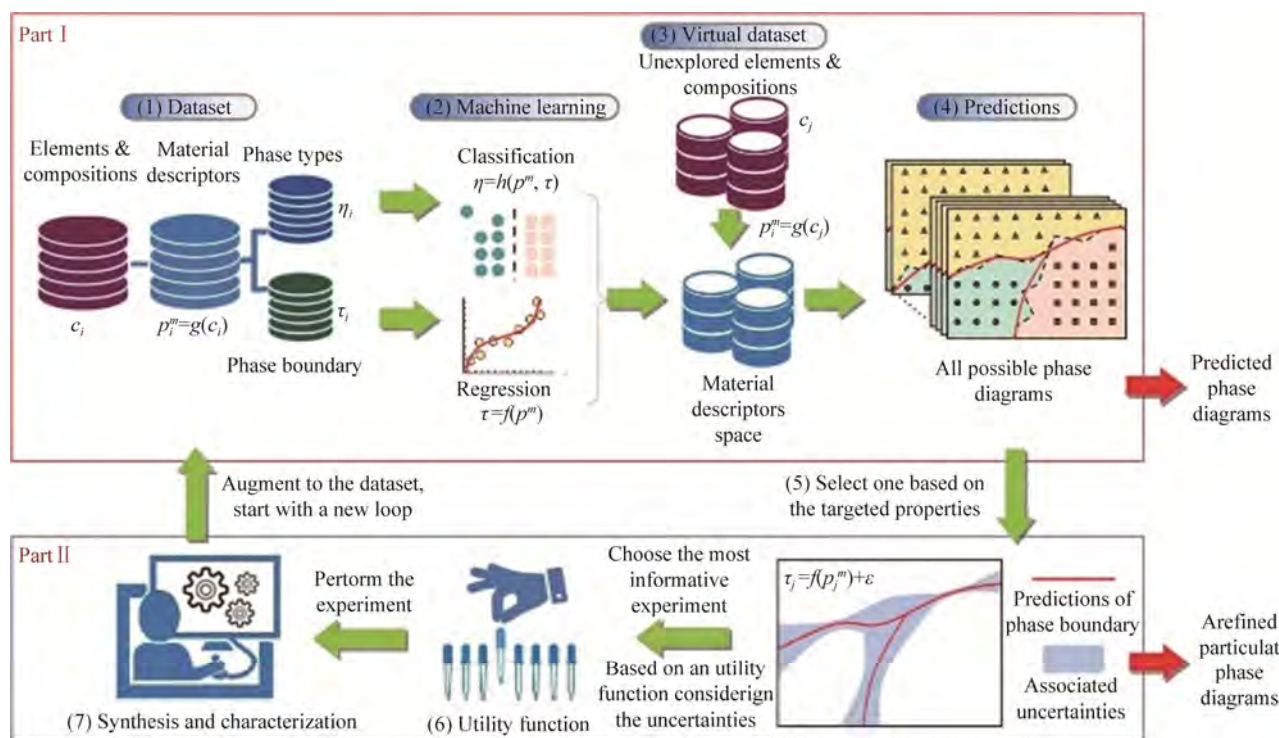


图 5 相图预测与优化设计策略流程图^[27]

Fig. 5 Flow chart for prediction and optimization of designing phase diagram^[27]

上述工作的优化过程主要基于相变温度和成分之间的回归模型。主动学习框架也可以应用于相成分之间分类问题。例如, Terayama 等^[28]利用 3 种不确定性采样方法: 最小置信度方法、边缘采样方法和基于熵的方法, 实现了包括 SiO₂-Al₂O₃-MgO 三元相图在内的 3 个已知相图的构造; 与随机抽样相比, 不确定性采样的方法将抽样点数减少到 20% 左右; 使用不确定性采样方法, 可以有效地在相界附近进行采样, 只需要较少数量的初始采样点, 就可以快速找到未检测到的新相。

因此, 利用基于不确定性的效用函数, 能够快速优化材料性能曲线, 优化效率优于随机搜索。

3 结论与展望

本综述主要针对单目标优化、多目标优化、曲线优化 3 个方面, 从算法方法角度对主动学习在材料研发中的应用进展进行了回顾, 涉及的材料有无铅压电陶瓷、电化学材料、固态电解质、金属-绝缘体过渡材料等, 说明了主动学习的方法能够在最少实验/计算次数下达到优化目标。主动学习框架的优势在于可以提供较好的决策, 从而大大加速材料研发的速度, 目前也越来越多地应用于其他各类材料研发中^[29–32]。

同时, 主动学习辅助材料研发中仍然存在一些

困难亟需克服: 1) 主动学习的性能很容易受到策略和数据的影响, 材料的数据集具有多样性, 数据量大小、数据类型等各不相同, 很难找到一种策略能够在每一个数据集中都表现优异, 而通过新的实验/计算测试不同策略的效果会花费较高的成本, 如何低成本地筛选机器学习模型和效用函数的组合非常重要; 2) 主动学习的推荐和标记是一个串行的过程, 推荐过程以及获得标注都需要时间成本, 导致效率变低, 如何实现主动学习与标注的交互对主动学习辅助材料设计的效率提高具有重要的意义。

致谢: 感谢徐洋洋、线跃辉、王博等博士研究生在本文撰写过程中提供的协助。

参考文献:

- [1] CORREA-BAENA J P, HIPPAALGAONKAR K, DUREN J, et al. Accelerating materials development *via* automation, machine learning, and high-performance computing[J]. *Joule*, 2018(2): 1410–1420.
- [2] National Science and Technology Council, Committee on Technology and Subcommittee on the MGI Initiative. Materials Genome Initiative—Strategic Plan[EB/OL]. (2014). https://www.mgi.gov/sites/default/files/documents/mgi_strategic_plan_dec_2014.pdf
- [3] LOOKMAN T, BALACHANDRAN P V, XUE D Z, et al. Active learning in materials science with emphasis on adaptive sampling using uncertainties for targeted design[J]. *NPJ Comput Mater*, 2019, 5: 21.
- [4] LOOKMAN T, BALACHANDRAN P V, XUE D Z, et al. Statistical inference and adaptive design for materials discovery[J]. *Curr Opin Solid State Mater Sci*, 2017, 21: 121–128.
- [5] WENG B C, SONG Z L, ZHU R L, et al. Simple descriptor derived from symbolic regression accelerating the discovery of new perovskite catalysts[J]. *Nat Commun*, 2020, 11: 3513.
- [6] BACK S, TRAN K, ULISSI Z W. Toward a design of active oxygen evolution catalysts: insights from automated density functional theory calculations and machine learning[J]. *ACS Catal*, 2019, 9: 7651–7659.
- [7] XUE D Z, BALACHANDRAN P V, HOGDEN J, et al. Accelerated search for materials with targeted properties by adaptive design[J]. *Nat Commun*, 2016, 7: 1124.
- [8] COX D D, JOHN S. A statistical method for global optimization[C]//IEEE International Conference on Systems, Man, and Cybernetics, Chicago, IL, USA, 1992: 1241–1246.
- [9] JONES D R, SCHONLAU M, WELCH W J. Efficient global optimization of expensive black-box functions[J]. *J Glob Optim*, 1998, 13: 455–492.
- [10] TIAN Y, YUAN R H, XUE D Z, et al. Role of uncertainty estimation in accelerating materials development *via* active learning[J]. *J Appl Phys*, 2020, 128(1): 014103.
- [11] SEXTON J, LAAKE P. Standard errors for bagged and random forest estimators[J]. *Comput Stat Data Anal*, 2009, 53(3): 801–811.
- [12] EFRON B, STEIN C. The jackknife estimate of variance[J]. *Ann Stat*, 1981, 9(3): 586–596.
- [13] WAGER S, HASTIE T, EFRON B. Confidence intervals for random forests: the jackknife and the infinitesimal jackknife[J]. *J Mach Learn Res*, 2014, 15(1): 1625–1651.
- [14] FLORES R A, PAOLUCCI C, WINTHER K T, et al. Active learning accelerated discovery of stable iridium oxide polymorphs for the oxygen evolution reaction[J]. *Chem Mater*, 2020, 32(13): 5854–5863.
- [15] MEKKI-BERRADA F, REN Z K, HUANG T, et al. Two-step machine learning enables optimized nanoparticle synthesis[J]. *NPJ Comput Mater*, 2021, 7:55.
- [16] YU M T, YANG S Y, WU C Z, et al. Machine learning the Hubbard U parameter in DFT+U using Bayesian optimization[J]. *NPJ Comput Mater*, 2020, 6: 180.
- [17] YUAN R H, LIU Z, BALACHANDRAN P V, et al. Accelerated discovery of large electrostrains in BaTiO₃-based piezoelectrics using active learning[J]. *Adv Mater*, 2018, 30(7): 1702884.
- [18] GOPAKUMAR A M, BALACHANDRAN P V, XUE D Z, et al. Multi-objective optimization for materials discovery *via* adaptive design[J]. *Sci Rep*, 2018, 8(1): 3738.
- [19] 谢建新, 宿彦京, 薛德祯, 等. 机器学习在材料研发中的应用[J]. *金属学报*, 2021, 57(11): 1343–1361.
- [20] XIE Jianxin, SU Yanjing, XUE Dezhen, et al. *Acta Metall Sin (in Chinese)*, 2021, 57(11): 1343–1361.
- [21] EMMERICH M T M, GIANNAKOGLU K C, NAUJOKS B. Single-and multiobjective evolutionary optimization assisted by Gaussian random field metamodelling[J]. *IEEE Trans Evol Comput*, 2006, 10(4): 421–439.
- [22] HARADA M, TAKEDA H, SUZUKI S, et al. Bayesian-optimization-guided experimental search of NASICON-type solid electrolytes for all-solid-state Li-ion batteries[J]. *J Mater Chem A*, 2020, 8: 15103–15109.
- [23] WANG Y Q, IYER A, CHEN W, et al. Featureless adaptive optimization accelerates functional electronic materials design[J]. *Appl Phys Rev*, 2020, 7: 041403.
- [24] BAUTISTA D C T, A sequential design for approximating the Pareto front using the expected pareto improvement function[D]. The Ohio State University, 2009.
- [25] KEANE A J. Statistical improvement criteria for use in multiobjective design optimization[J]. *AIAA J*, 2006, 44(4): 879–891.
- [26] VALLADARES H, LI T Y, ZHU L K, et al. Gaussian process-based prognostics of lithium-ion batteries and design optimization of cathode active materials[J]. *J Power Sources*, 2022, 528: 231026.
- [27] TIAN Y, XUE D Z, YUAN R H, et al. Efficient estimation of material property curves and surfaces *via* active learning[J]. *Phys Rev Mater*, 2021, 5: 013802.
- [28] TIAN Y, YUAN R H, XUE D Z, et al. Determining multi-component phase diagrams with desired characteristics using active learning[J]. *Adv. Sci*, 2021, 8: 2003165.
- [29] TERAYAMA K, TAMURA R, NOSE Y, et al. Efficient construction method for phase diagrams using uncertainty sampling[J]. *Phys Rev Mater*, 2019, 3: 033802.
- [30] ZHANG Y, WEN C, WANG C X, et al. Phase prediction in high entropy alloys with a rational selection of materials descriptors and machine learning models[J]. *Acta Mater*, 2019, 185: 528–539.
- [31] RAO Z Y, TUNG P, XIE R W, et al. Machine learning-enabled high-entropy alloy discovery[J]. *Science*, 2022, 378: 78–85.
- [32] CHEN L Q, ZHAO Y H. From classical thermodynamics to phase-field method[J]. *Prog Mater Sci*, 2022, 124: 100868.
- [33] 宿彦京, 付华栋, 白洋, 等. 中国材料基因工程研究进展[J]. *金属学报*, 2020, 56(10): 1313–1323.
- [34] XU Yanjing, FU Huadong, BAI Yang, et al. *Acta Metall Sin (in Chinese)*, 2020, 56(10): 1313–1323.