Contents lists available at ScienceDirect

Engineering

journal homepage: www.elsevier.com/locate/eng

Research
Smart Process Manufacturing toward Carbon Neutrality—Perspective

化学工程师的主动机器学习

Yannick Ureel , Maarten R. Dobbelaere , Yi Ouyang , Kevin De Ras , Maarten K. Sabbe , Guy B. Marin , Kevin M. Van Geem *

Laboratory for Chemical Technology, Department of Materials, Textiles and Chemical Engineering, Ghent University, Ghent 9052, Belgium

ARTICLE INFO

Article history:
Received 9 September 2022
Revised 7 December 2022
Accepted 28 February 2023
Available online 1 August 2023

关键词
主动机器学习
主动学习
贝叶斯优化
化学工程
实验设计

摘要

通过将机器学习与实验设计相结合，从而实现所谓的主动机器学习，可以进行更高效、更低成本的研究。机器学习算法在研究跨越化学工程的所有长度尺度的过程中更加灵活，并且优于传统的实验算法设计。虽然主动机器学习算法日趋成熟，但它们的应用却很落后。在本文中，我们指出了主动机器学习面临的三种挑战，即说服实验研究者、数据创建的灵活性和主动机器学习算法的鲁棒性，并讨论了克服这些挑战的方法。由于自动化程度的不断提高以及能够推动新发现的更高效算法的出现，主动机器学习在化学工程领域的前景光明。

1. 引言

在明确的条件下进行的实验和基于第一性原理的计算构成了工程研究的基础。在化学工程中，这些活动旨在开发和优化催化剂、反应条件和反应器配置等。2017年化工行业在研发上花费了510亿美元[1]。这说明了高质量数据的重要性；然而，获取准确数据的工作是烦琐的，并且容易出错。实验设计（DoE）可以实现以最小的努力提取最大的信息[2–3]，确保有效地使用时间和资源。通过将机器学习与DoE相结合，可以实现更灵活和高效的DoE。这种所谓的“主动机器学习”允许更有效地选择实验条件，特别是对于高维和高度非线性的现象[4]。

机器学习可以促进从实验选择到模型构建和数据分析整个实验周期的自动化[5]。虽然机器学习中最常见的应用领域是模型构建和数据分析，但本文的重点是探讨将DoE与机器学习相结合以实现主动机器学习的潜力。Olsson [6]将主动机器学习定义为一种有监督的机器学习技术，其中，学习者——也就是机器学习模型——控制着它从中学习的数据。在主动机器学习中，机器学习算法基于不确定性标准迭代确定新的实验数据，即所谓的训练数据。需要注意的是，“实验”也可以指计算成本高昂的高级模拟，如分子性质的高水平从头计算或用计算流体动力学（CFD）代码进行的反应流的大涡模拟[7]。主动机器学习由两个分支组成，它们有两个不同的目的：主动学习和贝叶斯优化。主动学习的目的是探索和建模一个具有最少数量的“实验”的过程，以确保对整个设计空间的准确

* Corresponding author.
E-mail address: Kevin.VanGeem@UGent.be (K.M. Van Geem).

英文原文：Engineering, 2023, 27(8):23–30
引用本文：Yannick Ureel, Maarten R. Dobbelaere, Yi Ouyang, Kevin De Ras, Maarten K. Sabbe, Guy B. Marin, Kevin M. Van Geem. Active Machine Learning for Chemical Engineers: A Bright Future Lies Ahead! *Engineering*, https://doi.org/10.1016/j.eng.2023.02.019

预测[8]。贝叶斯优化本质上是一种基于机器学习的优化策略，它通过迭代选择新的实验数据来寻找能优化目标的实验[9]。主动学习或贝叶斯优化都可以用于实验选择，具体取决于目标是对过程建模并获取过程知识还是优化目标。

1.1. 主动机器学习的基本原理

图1说明了主动机器学习算法的一般工作流程，从初始化开始，然后是一个由三个阶段组成的迭代循环。初始化的关键第一步包括将研究问题明确地定义为输出的建模（主动学习）或目标的优化（贝叶斯优化）。主动学习的一个例子是研究反应条件（如温度和压力）对转换的影响[10–11]。通过贝叶斯优化，目标是找到最优反应条件，使转化最大化[12–14]。在这两种情况下，通过考虑实验工具的目标和内在限制，建立了一个设计空间来定义所研究变量的范围。然后，使用一小部分标记数据样本来初始化和训练机器学习模型，这些数据来自于已知结果的实验，这些结果来自文献、以前的实验或新进行的实验。一般来说，初步标记数据的量非常少。

经过初始训练后，机器学习模型能够在设计空间中做出基本的预测。该模型可以模糊地估计出贝叶斯优化的最佳位置，或者哪个实验（即所谓的查询）为主动学习提供了最多的信息。虽然主动学习与贝叶斯优化的定义和初始化在本质上是相同的（甚至与经典的实验活动没有太大的不同），但主要的差异和优势是在模型训练中发现的。

主动学习纯粹基于探索，使设计空间的预测尽可能准确。相反，贝叶斯优化平衡了探索和利用，以找到设计空间中的最优迭代，将每次迭代作为可能的最终迭代。利用会研究具有高目标值的区域，以找到附近的最优解，而探索则发现预测未知且因此不确定的区域。探索需要对预测中的不确定性进行测量，以确定设计空间的哪些区域仍未被探索[15]。因此，用于主动机器学习的流行机器学习模型是高斯过程[16–19]和贝叶斯神经网络[20–22]，因为它们允许对其预测进行不确定性估计。高斯过程的另一个优点是，它们能很好地处理现实实验中固有的噪声测量。通过在高斯过程核中添加一个噪声项，机器学习模型可以估计实验的不确定性，并允许主动机器学习方法的性能达到最优[16,23]。神经网络也可以用于主动机器学习的目的，但需要使用蒙特卡罗辍学或模型集成等近似的方法来估计模型的不确定性[11,24–25]。

初始化后，主动机器学习过程包括三个阶段：机器学习模型的训练、新实验的选择以及这些实验的执行和注释（图1）。主动机器学习查询（第2阶段）是通过所谓的获取函数来确定的，这是对潜在信息量或最优性的度量。该模型需要信息量最大的后续数据点，即所选查询的获取函数最大的数据点。执行查询并收集新的数据（第3阶段），然后重新训练机器学习模型（第1阶段），使其能够进行改进后的预测。这个循环被依次迭代，直到找到一个最优值（贝叶斯优化）或获得一个足够准确的模型（主动学习）。

为了进一步说明工作流程，我们举了一个研究人员检查一种新的化学过程的催化剂的性能的例子。研究人员的目标是研究（通过主动学习）或优化（通过贝叶斯优化）反应变量（设计空间），如温度、压力和反应物浓度，对期望的产物产率（目标）的影响。首先，初始实验必须在温度、压力和反应物浓度的多个随机组合下进行。接下来，研究者通过在这些随机选取的实验数据点上训练机器

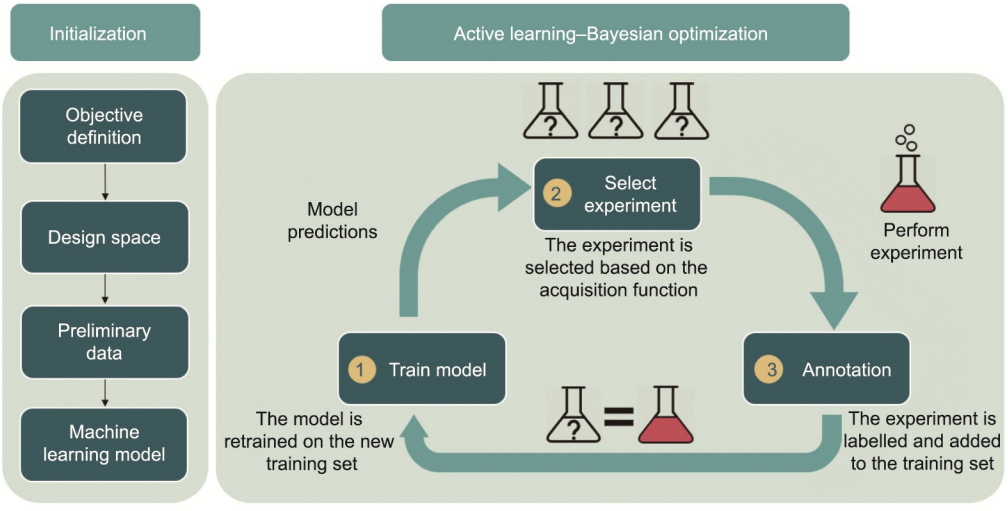

图1. 一般的主动机器学习工作流程概述，描述初始化和迭代查询选择[10]。

学习模型，启动主动机器学习循环，然后该模型提出一个新的实验。在使用主动学习时，该实验是信息最丰富的实验；当使用贝叶斯优化进行优化时，该实验是最有可能提高期望产品产率的实验。研究人员进行了实验，并重新训练了机器学习模型，该模型现在可以做出改进的预测。实验选择继续进行，直到进行所需的实验次数，并获得最优的机器学习模型或过程条件。

1.2. 化学工程中的主动机器学习

主动机器学习的应用跨越了化学工程的所有长度尺度，从从头计算[17–18,26]到材料、分子和催化剂设计[27–36]，反应设计[12–14,37–42]，以及反应器设计[43–45]。例如，催化剂的设计是实现碳中和的重要资产，因为催化剂可以使过程更可持续，并可以总体上提高化学过程的能源效率[46]。然而，如今催化剂设计仍然被认为是一门艺术，因为它主要依赖于高通量筛选和有限的理论关系，如Sabatier原理和线性比例关系[47–50]。这使得催化剂设计容易受到人为偏见的影响，因为研究人员倾向于利用已知的催化剂设计，这阻碍了真正的突破[51–52]。通过主动的机器学习，可以消除这种人为偏见，并且可以研究更大比例的催化剂空间。目前，主动机器学习在催化中的应用只考虑了有限的设计空间，在保持催化剂结构的同时，只改变催化剂的组成[53–54]。例如，Zhong等[53]对密度泛函理论（DFT）计算进行了贝叶斯优化，以识别和合成用于还原CO_2的很有前景的电催化剂，而Nugraha等[54]确定了活性最高的PtPdAu催化剂对电催化氧化甲醇的最佳组成。

在反应或工艺设计中，贝叶斯优化的目标是确定最佳操作条件，以最大化产品产率，最小化每个产品的排放量，实现最高的能源效率等。反应条件的优化已经被多次证明，包括使用离散变量和连续变量的多目标反应优化，这可能使其成为化学工程中主动机器学习最发达的领域[12–14]。Shields等[39]采用贝叶斯优化对Mitsunobu反应的反应条件进行优化，经过40次实验，得到了几个非直观反应条件的最佳产率（>99%），从而克服了标准反应产率为60%的难题。主动学习的目标是获得可用于反应器和催化剂设计、过程控制或逆合成的反应知识。Eyke等[11]通过用最少的可用数据预测催化剂和溶剂组合的反应产率，证明了DoE在反应设计中的主动学习潜力。最近，Ureel等[10]开发了一种用于化学反应研究的DoE工具，并在塑料废料的催化热解过程中进行了验证。

计算流体动力学已成为反应器、优化和故障排除的重要工具。贝叶斯优化可以通过最小的计算密集型的CFD模拟找到一个最理想的反应器配置。Park等[44]通过最大化搅拌釜反应器的持气率和最小化功耗，证明了多目标贝叶斯优化的能力。在CFD中明确地集成主动机器学习可以实现更快、更高效的反应器设计。

这项调查表明，化学工程是一个广泛和多样的研究领域，具有各种可能的主动机器学习应用。然而，主动机器学习还没有被广泛应用，在它成为化学工程师工具包中值得信赖的资产之前，还有一些障碍需要克服。在本文中，我们重点关注主动机器学习作为实验者的一种DoE技术，以及如何推广它。我们确定了三种类型的阈值：说服实验研究者、数据创建的灵活性以及主动机器学习算法的鲁棒性（图2）。在下面的部分中，我们将讨论这些挑战以及如何克服它们。

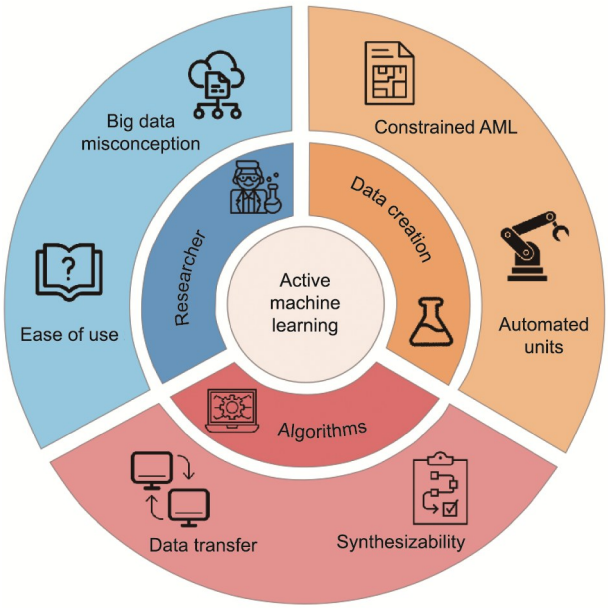

图2. 主动机器学习（AML）突破的三种不同类型的阈值。

2. 说服研究者

2.1. 大数据误解

目前，实验界和机器学习专家之间存在着知识差距[55]。这种知识差距是为什么主动机器学习尚未被实验人员系统地应用的根本原因。首先，存在一种误解，即认为大数据对于主动机器学习是必需的，并且需要进行大量的实验活动才能使之可行。Nugraha等[54]报道了在5151个可能的实验中只执行47个实验的最佳催化剂组成，如图3所示。在其工作中，采用贝叶斯优化的方法确定了甲醇电催化氧化的最佳PtPdAu催化剂组成。同样地，Schweidtmann等[12]在经过68次四维反应优化实验后确定了他们

的Pareto前沿。此外，Ureel等[10]的研究表明，主动学习策略对于只有18个实验的实验活动已经很有帮助。这些例子说明，主动学习和贝叶斯优化对于较小的数据集都是可行的。

第二个问题与实验研究人员的关系较小，而是与内在算法有关。最初，所有主动的机器学习算法都会探索整个设计空间，这可能会导致反直觉或琐碎的查询。因此，实验者对机器学习工具失去了信心。实验的初始选择并不依赖于机器学习模型中的任何初步知识或物理知识。因此，这个问题既与人类的偏见和用户对这些算法的感知有关，也与这些模型中缺乏初步知识有关。事先将过程知识集成到机器学习模型中是缓解这一问题的最有力的方法。这些知识可以通过两种不同的方法来整合：要么通过机器学习模型的设计，如高斯过程内核[56]，要么通过对文献或模拟数据的训练[57]。将初步知识纳入主动机器学习模型将在第4.1节中进行讨论。

2.2. 易用性

在主动学习策略中，多个因素同时变化，而常规的DoE策略通常一次只改变一个因素。这使得实验的后处理不再那么烦琐，因为这些因素的影响并不是孤立的。因此，需要进行统计分析，以从使用主动学习策略的实验活动中得出结论[58]。这些工具包含在常规的DoE软件中，但不包含在目前可用的主动机器学习软件包中。这个问题与另一个限制主动学习适用性的问题密切相关，即它的易用性。现在有许多不同的主动机器学习包，如用于贝叶斯优化的Gryffin [59]、Phoenics [60]和Bayesian Optimization [61]，以及高斯N维主动学习框架（GandALF）[10]或用于主动学习的通用和高效主动学习（GEAL）框架[62]。但是，当前大多数活跃的主动机器学习包必须使用Python配置，除了GandALF，它使用csv电子表格。使用这些主动的机器学习工具需要编程技能，因为它们不提供图形用户界面（GUI），这阻碍了这些方法的使用。因此，目前希望使用主动机器学习的研究人员必须投入大量的时间。这种“激活障碍”对许多研究人员来说太高了，特别是因为它需要具有编码能力。

3. 提高数据创建的灵活性

3.1. 受约束的主动机器学习

主动机器学习算法通常是在模拟数据上开发的，在数据创建方面没有实际限制[32,36,63]。然而，在现实生活中，实验单元或程序不允许这种灵活性。例如，即使是一个完全自动化的实验单元也经常需要加热或冷却，或者需要时间来稳定，当算法选择不同的温度时，会减慢新数据点的生成速度。此外，实验通常是并行进行的（例如，在高通量单元中），这与假设实验顺序选择的算法相反。因此，主动机器学习策略应该被限制在它们所使用的单元上，以实现最佳的实验效率，从而使它们适用于实际应用[64]。在上面的例子中，加热一个实验单元通常比冷却它更容易；因此，应该在算法中添加一个额外的约束，使其优先选择增加温度而不是降低温度的实验。

除了因实验设备运行方式而产生的约束外，限制条件对模拟也很重要[43,45]。让我们考虑一个涉及使用CFD对反应器进行模拟优化的例子。当定义CFD的反应器几何形状时，并不是每一种几何形状都可以进行模拟，也不是每一种几何形状都可以进行适当的网格划分或结果是网格无关性的[65]。当这些约束条件并不复杂时，可以训练一个单独的机器学习模型来学习这些约束条件，并确保模拟的可行性[43]。

实验单元受限的另一个例子是用于筛选不同催化材料

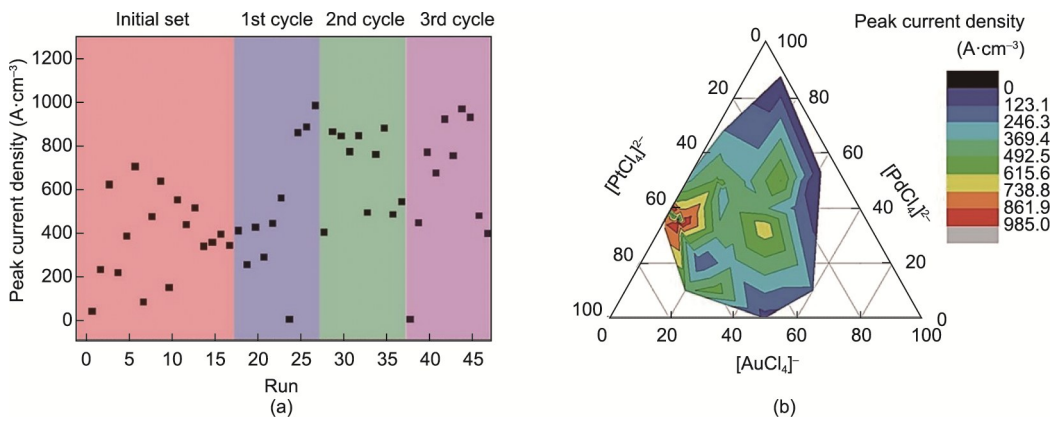

图3. （a）Nugraha等[54]仅进行了47次实验就确定了甲醇电催化氧化的最佳PtPdAu催化剂组成，峰值电流密度越高，说明催化剂越好；（b）催化剂组成对峰值电流密度影响的等高线图，由47次进行的实验确定[54]。

的高通量实验活动。在这些单元中，几个实验变量，如温度和压力等，通常在每批实验中都是固定的。这需要对这些实验的批处理选择进行另一个约束，因为所有选择的查询都必须固定变量。因此，为了根据其应用程序调整主动机器学习算法，需要机器学习专家和实验人员之间的密切合作。通过这种方式，应用主动机器学习的优势也可用于不太灵活的实验单元。

实验人员和机器学习专家之间的共生关系将使双方都受益。首先，随着研究人员越来越意识到主动机器学习的好处，其将扩展主动机器学习的应用领域。这种密切的合作将有助于识别这些主动机器学习算法中的有用特征，如分区或自动后处理。在实验选择中可以增加更多的实际限制，如所提出的实验所需的时间或成本。最后，实验人员和机器学习专家之间的合作有助于为实验研究人员提供信息，并消除目前存在的对主动机器学习的偏见。

3.2. 自动化

在理想的情况下，主动机器学习与灵活的自动实验单元耦合，甚至由机器人配备[12,14,66]。因此，对实验性能的控制和优化可以达到最优，从而节省宝贵的时间和精力。自动化实验单元正越来越多地应用于分子合成和化学工程，尽管这些单元尚未普及[67–69]。自动化机器人单元的一个要求是，它们应该是可重新配置的[70]。此外，它们应具有广泛的应用范围，不应局限于单一反应类型或狭窄温度范围的研究。当然，自动化单元的使用并非不言而喻的，因为它们通常很昂贵，而且目前并不太适合解决所有问题。例如，尽管过去付出了努力[71]，催化剂的自动化合成和测试仍是一项具有挑战性的任务，特别是在研究广泛的设计空间时[72]。通过将这些系统与主动机器学习技术相结合，预计将为实验活动节省大量的时间，因为这将加速反应和催化剂的优化，以及科学知识的获取。这些自动化单元的最后一个门槛是这些单元的安全性问题。通过扩大催化剂或反应设计空间，安全性问题会增加，因为这样做会增加非期望反应发生的概率。因此，在使用这些单元时，仍然需要良好的化学知识，以识别和纳入安全约束条件。在这里，安全约束的定义再次需要实验人员和机器学习专家之间的密切合作。

4. 算法鲁棒性

4.1. 数据传输

在进行实验时，实验最好具有广泛适用性和多重目的。实验中收集的信息应该根据FAIR（即可查找性、可访问性、互操作性和可重用性）指导原则提供给其他研究人员[73]。然而，在主动机器学习中，目标是单一的，这决定了实验的选择。这妨碍了实验的适用性，因为只有一种实验输出得到了充分的研究。例如，在研究反应时，通常选择转化率作为感兴趣的输出；然而，这限制了对其他特性信息的了解，如产率或选择性。在最坏的情况下，无法测量产率，也无法收集信息；相反，即使测量了这些产率，也不能保证在示例中考虑了所有的趋势。由于主动机器学习的目标是对转化进行建模，这种方法忽略了有趣的反应产率的行为，这可能会导致趋势被隐藏。对于贝叶斯优化，这不会造成问题，因为目的是优化一个目标，这使得每个定义的数据不那么普遍适用。多目标贝叶斯优化技术是存在的，而主动学习只能采用单目标策略，这意味着所有有趣的输出都应该包含在单个主动学习目标中[12,40,44]。因此，为了确保收集到的数据的可重用性，在实验过程中不仅要测量建模的输出，还要测量其他潜在的相关输出，这一点很重要。

在创建了广泛关注的数据之后，能够将这些知识整合到主动机器学习工具中是很重要的。图4总结了可以用来实现这一目标的不同数据源和建模策略。当主动机器学习模型根据文献数据进行预训练时，可以实现改进的初始实验选择，解决前面提到的次优初始选择问题[57]。当实验的不确定性与新收集数据的不确定性相似时，文献数据的合并是微不足道的。然而，当文献数据的质量比收集到的数据更好或更差时，机器学习模型是否能够区分两者是很重要的。异方差机器学习模型是存在的[63]，但它们不一定允许合并两个独立的噪声因子，因为噪声的变化依赖于异方差模型中的变量。相反，多保真度主动机器学习策略使得使用广泛丰富的低质量数据来精确预训练主动机器学习模型[74–76]成为可能。这些方法仅基于模拟的“实验”数据开发，但当应用于真实实验数据时，它们在提高主动机器学习的性能方面非常有前景。此外，这些多保真度模型也可以用于将来自机理模型的数据合并到机器学习模型中。当机理模型预测的不确定性已知时，可以在多保真度模型中对实验数据和模型数据进行适当的区分，两者都有各自的不确定性。通过这种方式，可以将额外的机制信息合并到一个机器学习模型中，从而改进实验选择。

密切相关但在本质上并不相似的数据，也可以作为主动机器学习模型的初始化[77]。例如，当用一种催化剂进行反应建模，同时又有另一种催化剂的文献数据时，这些数据可能仍然包含对主动学习模型有价值的信息[78]。主动迁移学习的目标是利用从几乎相似的数据中获得的知识来获得一个机器学习模型，以提高对所检测问题的感知。

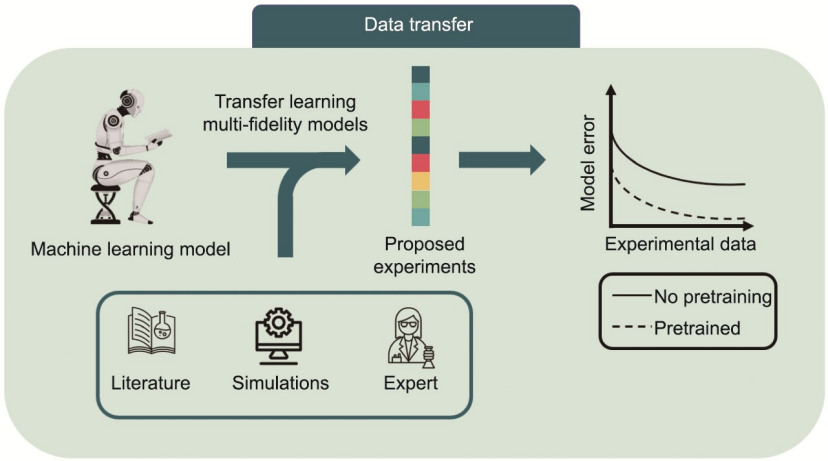

图4. 通过迁移学习或多保真度模型，将来自文献、模拟或专家知识中的数据整合到机器学习模型中，可以提高主动机器学习的性能。

主动迁移学习是主动机器学习和迁移学习两种主要方法的结合，可以减少机器学习的数据密集度。在迁移学习中，（大量）可用的低质量数据被用于对机器学习模型进行预训练，然后使用有限数量的高质量数据进行细化。这样，在机器学习模型中引入了基本的物理知识，再次完善了初始的实验选择。该方法通过对不同亲核试剂反应的机器学习模型进行预训练，已被证明可用于交叉偶联反应的反应产率分类[78]。

在主动机器学习应用程序中重用文献数据将进一步提高这些工具的性能。第一个主动迁移学习方法正在化学工程领域中发展，但算法的进一步发展对于使主动迁移学习适用于化学工程的所有领域至关重要。

4.2. 可合成性

主动机器学习可以用于确定最佳查询，以达到优化或建模的目的。然而，对于某些问题，这些查询的可执行性并不明显。例如，在催化剂或分子设计中，会提出新的化合物来合成和测试感兴趣的性质。在这里，催化剂或分子的表示对于查询的可合成性是至关重要的。可合成性被定义为所提出查询的可行性，是指所提出的催化剂或分子是否可以合成，如图5所示。通常，包含催化剂成分的载体是催化剂的简单表示[54,79]。这确保了催化剂的可合成性，但限制了主动机器学习算法所探索的设计空间，因为这只改变了成分，而没有考虑结构或几何性质。理想情况下，对每个问题都考虑完整的催化剂空间，例如，考虑将完整的三维（3D）几何形状作为催化剂位点或分子的表示。然而，并不是每一种可以想象到的催化剂或分子的三维几何结构都可以合成，所以在设计空间的大小（即所谓的创造力）和可合成性之间需要权衡。

如前面的例子所示，可合成性的问题本质上可以归结为机器学习表示的问题，在此问题上添加约束来增强可合成性。一种直观的方法是使用催化剂或分子的合成过程作为机器学习的表示方法。包含催化剂组成、煅烧温度和时间以及离子交换或浸渍的存在的载体可以用来表示催化剂。通过这种方法，保证了查询的可合成性，因为每个提出的方案都是可执行的。然而，这种表示并不一定能确保简单地映射到感兴趣的属性，并且可能需要增加数据量来建模这种关系。

除了这种直观的方法之外，学习到的机器学习表示使创建一个连续的表示成为可能，这确保了所提出的查询的有效性[80–81]。通过在一组可合成的分子或催化剂上训练最近开发的方法，如变分自动编码器或生成对抗神经网络，可以开发一种学习型机器学习表示（即所谓的潜在空间），以确保所提出的查询的可合成性[80,82–83]。基于这种表示，根据应用程序，可以对催化剂或分子施加额外的约束[31]。

在机器学习问题中，找到一个适当的表示方法总是很重要的。对于主动的机器学习，这种表示是协调综合性和创造力的关键。

5. 结论和观点

主动机器学习非常适合化学工程研究人员用来加速从分子和催化剂设计到反应和反应器设计的实验活动。然而，主动机器学习在实验研究人员中并不为人熟知，并且许多主动机器学习应用程序目前对用户并不友好。机器学习专家和化学工程师之间更好的协作可以克服这些障碍。这种交互也将有助于根据所应用的（自动的）实验单元和程序调整主动机器学习算法，这将提高这些算法的性能。这里的一个关键障碍是初始实验选择不理想，这可以通过

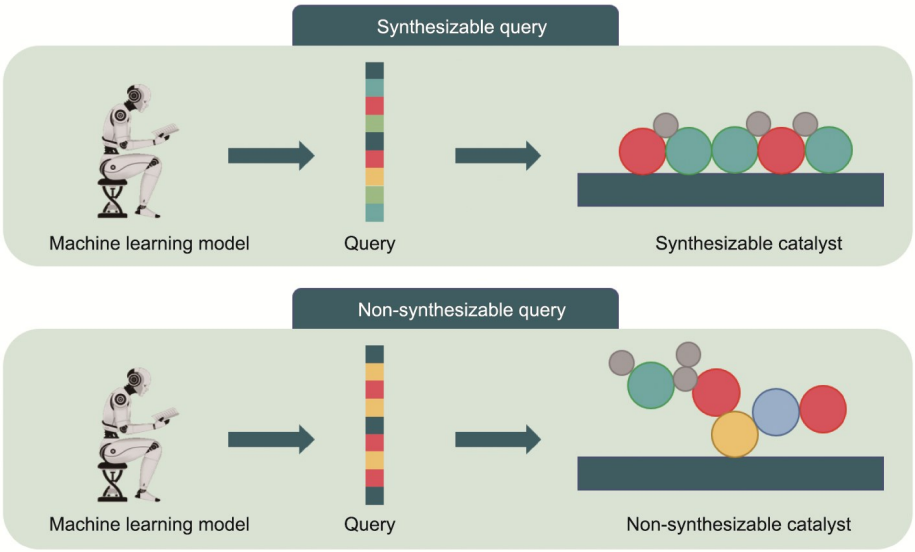

图5. 这是可合成性的一个说明。一个机器学习模型提出了一个查询，它本质上是一个催化剂的向量表示。这个查询对应于一个催化剂，它可以是现实的和可合成的（顶部），也可以是不现实的和不可合成的（底部）。

在多保真度模型的帮助下整合迁移学习和主动学习来克服这个障碍。此外，主动机器学习的应用领域还可以根据设置约束，通过调整一般的主动机器学习算法获得“定制”算法，这可以显著扩展主动机器学习的应用领域。虽然算法应该是定制的，但数据应该是普遍可用的，这样所进行的实验就可以服务于多种目的。通过协调可合成性和创造性，主动机器学习必然会在分子和催化剂合成领域取得重大进展。最近有希望的突破将使主动的机器学习成为化学工程师的一个重要工具，并将进一步促进自主和高效的科学发现，这将有助于未来更可持续的化学工业。

Acknowledgements

Yannick Ureel, Maarten R. Dobbelaere, and Kevin De Ras respectively acknowledge financial support from the Fund for Scientific Research Flanders (FWO Flanders) through the doctoral fellowship grants (1185822N, 1S45522N, and 3F018119). The authors acknowledge funding from the European Research Council (ERC) under the European Union’s Horizon 2020 research and innovation programme (818607).

Compliance with ethics guidelines

Yannick Ureel, Maarten R. Dobbelaere, Yi Ouyang, Kevin De Ras, Maarten K. Sabbe, Guy B. Marin, and Kevin M. Van Geem declare that they have no conflict of interest or financial conflicts to disclose.

References

[1] Oxford Economics Ltd. The global chemical industry: catalyzing growth and addressing our world’s sustainability challenges. Oxford: Oxford Economics Ltd.; 2019.
[2] Lazić ŽR. Design of experiments in chemical engineering: a practical guide. Weinheim: Wiley-VCH Verlag GmbH & Co. KGaA; 2006.
[3] Franceschini G, Macchietto S. Model-based design of experiments for parameter precision: state of the art. Chem Eng Sci 2008;63(19):4846–72.
[4] Melnikov AA, Poulsen Nautrup H, Krenn M, Dunjko V, Tiersch M, Zeilinger A, et al. Active learning machine learns to create new quantum experiments. Proc Natl Acad Sci USA 2018;115(6):1221–6.
[5] Duong-Trung N, Born S, Kim JW, Schermeyer MT, Paulick K, Borisyak M, et al. When bioprocess engineering meets machine learning: a survey from the perspective of automated bioprocess development. Biochem Eng J 2023;190:108764.
[6] Olsson F. A literature survey of active machine learning in the context of natural language processing. Kista: Swedish Institute of Computer Science; 2009.
[7] Marin GB, Galvita VV, Yablonsky GS. Kinetics of chemical processes: from molecular to industrial scale. J Catal 2021;404:745–59.
[8] Settles B. Active learning. Cham: Springer Nature Switzerland AG; 2012.
[9] Frazier PI. A tutorial on Bayesian optimization. 2018. arXiv:1807.02811v1.
[10] Ureel Y, Dobbelaere MR, Akin O, Varghese RJ, Pernalete CG, Thybaut JW, et al. Active learning-based exploration of the catalytic pyrolysis of plastic waste. Fuel 2022;328:125340.
[11] Eyke NS, Green WH, Jensen KF. Iterative experimental design based on active machine learning reduces the experimental burden associated with reaction screening. React Chem Eng 2020;5(10):1963–72.
[12] Schweidtmann AM, Clayton AD, Holmes N, Bradford E, Bourne RA, Lapkin AA. Machine learning meets continuous flow chemistry: automated optimization towards the Pareto front of multiple objectives. Chem Eng J 2018;352:277–82.
[13] Amar Y, Schweidtmann AM, Deutsch P, Cao L, Lapkin A. Machine learning and molecular descriptors enable rational solvent selection in asymmetric catalysis. Chem Sci 2019;10(27):6697–706.
[14] Clayton AD, Schweidtmann AM, Clemens G, Manson JA, Taylor CJ, Niño CG, et al. Automated self-optimisation of multi-step reaction and separation processes using machine learning. Chem Eng J 2020;384:123340.
[15] Thrun S. Exploration in active learning. In: Arbib MA, editor. The handbook of brain theory and neural networks. Cambridge: MIT Press; 1995. p. 381–4.

[16] Rasmussen CE, Williams CKI. Gaussian processes for machine learning. Cambridge: MIT Press; 2006.

[17] Podryabinkin EV, Shapeev AV. Active learning of linearly parametrized interatomic potentials. Comput Mater Sci 2017;140:171–80.

[18] Vandermause J, Torrisi SB, Batzner S, Xie Y, Sun L, Kolpak AM, et al. On-the-fly active learning of interpretable Bayesian force fields for atomistic rare events. NPJ Comput Mater 2020;6(1):20.

[19] Riis C, Antunes F, Hüttel FB, Azevedo CL, Pereira FC. Bayesian active learning with fully Bayesian Gaussian processes. 2022. arXiv:2205.10186.

[20] Blundell C, Cornebise J, Kavukcuoglu K, Wierstra D. Weight uncertainty in neural networks. In: Proceedings of the 32nd International Conference on Machine Learning; 2015 Jul 7–9; Lille, France; 2015. p. 1613–22.

[21] Gal Y, Islam R, Ghahramani Z. Deep Bayesian active learning with image data. In: Proceedings of the 34th International Conference on Machine Learning; 2017 Aug 6-11; Sydney, NSW, Australia; 2017. p. 1183–92.

[22] Hafner D, Tran D, Lillicrap T, Irpan A, Davidson J. Noise contrastive priors for functional uncertainty. In: Proceedings of the 35th Uncertainty in Artificial Intelligence Conference; 2019 Jul 22-25; Tel Aviv, Israel; 2020. p. 905–14.

[23] McHutchon A, Rasmussen C. Gaussian process training with input noise. In: Shawe-Taylor J, Zemel R, Bartlett P, Pereira F, Weinberger KQ, editors. Proceedings of the 24th International Conference on Neural Information Processing Systems; 2011 Dec 12-14; Granada, Spain; 2011. p. 1341–9.

[24] Zhang Y, Lee AA. Bayesian semi-supervised learning for uncertainty-calibrated prediction of molecular properties and active learning. Chem Sci 2019;10(35): 8154–63.

[25] Núñez M, Vlachos DG. Multiscale modeling combined with active learning for microstructure optimization of bifunctional catalysts. Ind Eng Chem Res 2019; 58(15):6146–54.

[26] Sivaraman G, Krishnamoorthy AN, Baur M, Holm C, Stan M, Csányi G, et al. Machine-learned interatomic potentials by active learning: amorphous and liquid hafnium dioxide. NPJ Comput Mater 2020;6(1):104.

[27] Reker D, Schneider P, Schneider G, Brown JB. Active learning for computational chemogenomics. Future Med Chem 2017;9(4):381–402.

[28] Brown KA, Brittman S, Maccaferri N, Jariwala D, Celano U. Machine learning in nanoscience: big data at small scales. Nano Lett 2020;20(1):2–10.

[29] Hansen MH, Torres JAG, Jennings PC, Wang Z, Boes JR, Mamun OG, et al. An atomistic machine learning package for surface science and catalysis. 2019. arXiv:1904.00904.

[30] Griffiths RR, Hernández-Lobato JM. Constrained Bayesian optimization for automatic chemical design. 2017. arXiv:1709.05501.

[31] Griffiths RR, Hernández-Lobato JM. Constrained Bayesian optimization for automatic chemical design using variational autoencoders. Chem Sci 2020; 11(2):577–86.

[32] Tran K, Ulissi ZW. Active learning across intermetallics to guide discovery of electrocatalysts for CO_2 reduction and H_2 evolution. Nat Catal 2018; 1(9): 696–703.

[33] Kusne AG, Yu H, Wu C, Zhang H, Hattrick-Simpers J, DeCost B, et al. On-the-fly closed-loop materials discovery via Bayesian active learning. Nat Commun 2020;11(1):5966.

[34] Oftelie LB, Rajak P, Kalia RK, Nakano A, Sha F, Sun J, et al. Active learning for accelerated design of layered materials. NPJ Comput Mater 2018;4(1):74.

[35] Kitchin JR. Machine learning in catalysis. Nat Catal 2018;1(4):230–2.

[36] Jablonka KM, Jothiappan GM, Wang S, Smit B, Yoo B. Bias free multiobjective active learning for materials design and discovery. Nat Commun 2021;12(1):2312.

[37] Zhang C, Amar Y, Cao L, Lapkin AA. Solvent selection for Mitsunobu reaction driven by an active learning surrogate model. Org Process Res Dev 2020; 24(12):2864–73.

[38] Clayton AD, Manson JA, Taylor CJ, Chamberlain TW, Taylor BA, Clemens G, et al. Algorithms for the self-optimisation of chemical reactions. React Chem Eng 2019;4:1545–54.

[39] Shields BJ, Stevens J, Li J, Parasram M, Damani F, Alvarado JIM, et al. Bayesian reaction optimization as a tool for chemical synthesis. Nature 2021; 590(7844):89–96.

[40] Felton KC, Rittig JG, Lapkin AA. Summit: benchmarking machine learning methods for reaction optimisation. Chem-Methods 2021;1(2):116–22.

[41] Felton K, Wigh D, Lapkin A. Multi-task Bayesian optimization of chemical reactions. 2020. ChemRxiv: 13250216.v1.

[42] Dogu O, Eschenbacher A, Varghese RJ, Dobbelaere M, D'Hooge DR, Van Steenberge PHM, et al. Bayesian tuned kinetic Monte Carlo modeling of polystyrene pyrolysis: unraveling the pathways to its monomer, dimers, and trimers formation. Chem Eng J 2023;455:140708.

[43] Tran A, Sun J, Furlan JM, Pagalthivarthi KV, Visintainer RJ, Wang Y. pBO-2GP-3B: a batch parallel known/unknown constrained Bayesian optimization with feasibility classification and its applications in computational fluid dynamics. Comput Methods Appl Mech Eng 2019;347:827–52.

[44] Park S, Na J, Kim M, Lee JM. Multi-objective Bayesian optimization of chemical reactor design using computational fluid dynamics. Comput Chem Eng 2018;119:25–37.

[45] Morita Y, Rezaeiravesh S, Tabatabaei N, Vinuesa R, Fukagata K, Schlatter P. Applying Bayesian optimization with Gaussian process regression to computational fluid dynamics problems. J Comput Phys 2022;449:110788.

[46] Friend CM, Xu B. Heterogeneous catalysis: a central science for a sustainable future. Acc Chem Res 2017;50(3):517–21.

[47] Sabatier P. La catalyse en chimie organique. Paris: Hachette Livre; 1920. French.

[48] Ichikawa S. Harmonious optimum conditions for heterogeneous catalytic reactions derived analytically with Polanyi relation and Bronsted relation. J Catal 2021;404:706–15.

[49] Landau RN, Korré SC, Neurock M, Klein MT, Quann RJ. Hydrocracking phenanthrene and 1-methyl naphthalene: development of linear free energy relationships. In: Oballa M, editor. Catalytic hydroprocessing of petroleum and distillates. Boca Raton: CRC Press; 2020. p. 421–32.

[50] Vijay S, Kastlunger G, Chan K, Nørskov JK. Limits to scaling relations between adsorption energies? J Chem Phys 2022;156(23):231102.

[51] Hong X, Chan K, Tsai C, Nørskov JK. How doped MoS_2 breaks transition-metal scaling relations for CO_2 electrochemical reduction. ACS Catal 2016;6(7): 4428–37.

[52] Pérez-Ramírez J, López N. Strategies to break linear scaling relationships. Nat Catal 2019;2(11):971–6.

[53] Zhong M, Tran K, Min Y, Wang C, Wang Z, Dinh CT, et al. Accelerated discovery of CO_2 electrocatalysts using active machine learning. Nature 2020; 581(7807):178–83.

[54] Nugraha AS, Lambard G, Na J, Hossain MSA, Asahi T, Chaikittisilp W, et al. Mesoporous trimetallic PtPdAu alloy films toward enhanced electrocatalytic activity in methanol oxidation: unexpected chemical compositions discovered by Bayesian optimization. J Mater Chem A 2020;8(27):13532–40.

[55] Dobbelaere MR, Plehiers PP, Van de Vijver R, Stevens CV, Van Geem KM. Machine learning in chemical engineering: strengths, weaknesses, opportunities, and threats. Engineering 2021;7(9):1201–11.

[56] Duvenaud DK. Automatic model construction with Gaussian processes [dissertation]. Cambridge: University of Cambridge; 2014.

[57] Wang Z, Dahl GE, Swersky K, Lee C, Mariet Z, Nado Z, et al. Pre-training helps Bayesian optimization too. 2022. arXiv:220703084.

[58] Symoens SH, Aravindakshan SU, Vermeire FH, De Ras K, Djokic MR, Marin GB, et al. QUANTIS: data quality assessment tool by clustering analysis. Int J Chem Kinet 2019;51(11):872–85.

[59] Häse F, Aldeghi M, Hickman RJ, Roch LM, Aspuru-Guzik A. Gryffin: an algorithm for Bayesian optimization of categorical variables informed by expert knowledge. Appl Phys Rev 2021;8(3):031406.

[60] Häse F, Roch LM, Kreisbeck C, Aspuru-Guzik A. Phoenics: a Bayesian optimizer for chemistry. ACS Cent Sci 2018;4(9):1134–45.

[61] Snoek J, Larochelle H, Adams RP. Practical Bayesian optimization of machine learning algorithms. In: Pereira F, Burges CJ, Bottou L, Weinberger KQ, editors. Proceedings of the 25th International Conference on Neural Information Processing Systems; 2012 Dec 3-6; Lake Tahoe, NV, USA. Red Hook: Curran Associates Inc.; 2012. p. 2951–9.

[62] Xie Y, Tomizuka M, Zhan W. Towards general and efficient active learning. 2021. arXiv:211207963.

[63] Griffiths RR, Aldrick AA, Garcia-Ortegon M, Lalchand V, Lee AA. Achieving robustness to aleatoric uncertainty with heteroscedastic Bayesian optimisation. Mach Learn Sci Technol 2021;3(1):015004.

[64] Hickman RJ, Aldeghi M, Häse F, Aspuru-Guzik A. Bayesian optimization with known experimental and design constraints for chemistry applications. Digit Discov 2022;1:732–44.

[65] Habashi WG, Dompierre J, Bourgault Y, Ait-Ali-Yahia D, Fortin M, Vallet MG. Anisotropic mesh adaptation: towards user-independent, mesh-independent and solver-independent CFD. Part I: general principles. Int J Numer Meth Fluids 2000;32(6):725–44.

[66] Burger B, Maffettone PM, Gusev VV, Aitchison CM, Bai Y, Wang X, et al. A mobile robotic chemist. Nature 2020;583(7815):237–41.

[67] Hoffer L, Voitovich YV, Raux B, Carrasco K, Muller C, Fedorov AY, et al. Integrated strategy for lead optimization based on fragment growing: the diversity-oriented-target-focused-synthesis approach. J Med Chem 2018;61(13):

5719–32.
[68] Bédard AC, Adamo A, Aroh KC, Russell MG, Bedermann AA, Torosian J, et al. Reconfigurable system for automated optimization of diverse chemical reactions. Science 2018;361(6408):1220–5.
[69] Mateos C, Nieves-Remacha MJ, Rincón JA. Automated platforms for reaction self-optimization in flow. React Chem Eng 2019;4(9):1536–44.
[70] Eyke NS, Koscher BA, Jensen KF. Toward machine learning-enhanced high-throughput experimentation. Trends Chem 2021;3(2):120–32.
[71] Hahndorf I, Buyevskaya O, Langpape M, Grubert G, Kolf S, Guillon E, et al. Experimental equipment for high-throughput synthesis and testing of catalytic materials. Chem Eng J 2002;89(1–3):119–25.
[72] Oh KH, Lee HK, Kang SW, Yang JI, Nam G, Lim T, et al. Automated synthesis and data accumulation for fast production of high-performance Ni nanocatalysts. J Ind Eng Chem 2022;106:449–59.
[73] Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 2016;3(1):160018.
[74] Greenman KP, Green WH, Gómez-Bombarelli R. Multi-fidelity prediction of molecular optical peaks with deep learning. Chem Sci 2022;13(4):1152–62.
[75] Pilania G, Gubernatis JE, Lookman T. Multi-fidelity machine learning models for accurate bandgap predictions of solids. Comput Mater Sci 2017;129:156–63.
[76] Folch JP, Lee RM, Shafei B, Walz D, Tsay C, van der Wilk M, et al. Combining multi-fidelity modelling and asynchronous batch Bayesian optimization. Comput Chem Eng 2023;172:108194.
[77] Mao S, Wang B, Tang Y, Qian F. Opportunities and challenges of artificial intelligence for green manufacturing in the process industry. Engineering 2019; 5(6):995–1002.
[78] Shim E, Kammeraad JA, Xu Z, Tewari A, Cernak T, Zimmerman PM. Predicting reaction conditions from limited data through active transfer learning. Chem Sci 2022;13(22):6655–68.
[79] Kim M, Ha MY, Jung WB, Yoon J, Shin E, Kim ID, et al. Searching for an optimal multi-metallic alloy catalyst by active learning combined with experiments. Adv Mater 2022;34(19):2108900.
[80] Gómez-Bombarelli R, Wei JN, Duvenaud D, Hernández-Lobato JM, Sánchez-Lengeling B, Sheberla D, et al. Automatic chemical design using a data-driven continuous representation of molecules. ACS Cent Sci 2018;4(2):268-76.
[81] Shang C, You F. Data analytics and machine learning for smart process manufacturing: recent advances and perspectives in the big data era. Engineering 2019;5(6):1010-6.
[82] Sanchez-Lengeling B, Outeiral C, Guimaraes GL, Aspuru-Guzik A. Optimizing distributions over molecular space. An objective-reinforced generative adversarial network for inverse-design chemistry (ORGANIC). 2017. ChemRxiv: 5309668.v3.
[83] Jensen Z, Kwon S, Schwalbe-Koda D, Paris C, Gómez-Bombarelli R, Román-Leshkov Y, et al. Discovering relationships between OSDAs and zeolites through data mining and generative neural networks. ACS Cent Sci 2021;7(5): 858-67.