人工智能技术在化学反应预测中的应用

孔祥泰[1,2]，张润泽[1,2]，张玮[1,2]，李叙潼[1,2*]，郑明月[1,2*](1.中国科学院上海药物研究所，上海 201203；2.中国科学院大学，北京 100049)

摘要：人工智能在各个领域的快速发展，为许多传统领域带来了新的活力和机遇。人工智能与大数据的组合为化学领域的发展开启新的视角，对大量共享的数据进行分析，有助于发现化学反应的新方法和新途径。目前，人工智能技术已在一些关键的化学任务上取得了重大突破，如正向反应预测、反应表现预测和反应条件优化等。本文整理回顾了人工智能在化学反应领域的研究历程，并分析讨论目前仍面临的困难与挑战。

关键词：人工智能；正向反应预测；反应表现预测；反应条件优化

中图分类号：R914　　**文献标志码：**A　　**文章编号：**1007-7693(2022)21-2856-09

DOI: 10.13748/j.cnki.issn1007-7693.2022.21.023

引用本文：孔祥泰，张润泽，张玮，等. 人工智能技术在化学反应预测中的应用[J]. 中国现代应用药学，2022，39(21)：2856-2864.

Application of Artificial Intelligence in Chemical Reaction Prediction

KONG Xiangtai[1,2], ZHANG Runze[1,2], ZHANG Wei[1,2], LI Xutong[1,2*], ZHENG Mingyue[1,2*](*1.Shanghai Institute of Materia Medica, Chinese Academy of Sciences, Shanghai 201203, China; 2.University of Chinese Academy of Sciences, Beijing 100049, China*)

ABSTRACT: The rapid development of artificial intelligence has brought great vitality and opportunities to many traditional fields. The combination of artificial intelligence and big data has created a novel perspective for the progress of chemistry. The analysis of a large amount of open-source data helps to discover original methods and ways of chemical reactions. At present, artificial intelligence has made significant breakthroughs in some key chemical tasks, such as forward reaction prediction, reaction performance prediction and reaction condition optimization. This paper reviews the process of artificial intelligence in chemical reactions, and then analyzes and discusses the current challenges in this field.

KEYWORDS: artificial intelligence; forward reaction prediction; reaction performance prediction; reaction condition optimization

近年来，以机器学习为核心的人工智能技术迎来了快速发展阶段，在许多以经验、数据为导向的传统领域中，克服了许多难题。而化学本身就是以经验、反应数据为基石发展起来的一门学科，化学家们从化学反应的现象中发现物质的自然属性和新的自然规律，然后基于不断积累的实验数据和发现的自然属性进行归纳、类比和推理，不断挖掘自然界的奥秘，为人类更高质量地生活发展和探索未知提供物质基础和技术手段。

现阶段存在多种人工智能技术如机器学习和深度学习方法，可以很好地应用于具有丰富数据的化学反应相关任务上，其中多数任务开展遵循的流程见图 1。首先，从已有的数据库或文献中收集相关数据，并选择不同的形式表示反应中的分子，之后将划分好的数据输入模型中，进行训练、验证、测试和评价，最后将训练好的模型用于化学反应预测相关的任务，除了能基于已有的数据预测未知的反应信息，如正向反应预测中基于反应物、反应条件等预测产物，反应表现预测中基于已有的化学反应预测化合物的反应活性、产率等，还能对现有的反应环境进行改善，如反应条件优化。

1　化学反应中分子的表示形式

有机化学反应是指反应物分子转化为产物分子所经历的过程，为描述这一过程，合理的分子表示形式必不可少。在化学信息学领域，存在多种分子表示形式，主要包括一维(SMILES[1]、InChI[2]、化学式等)描述、二维(化学分子图)描述以及基于矩阵的描述(原子邻接矩阵、节点特征矩阵和边特征矩阵等)[3]。

基金项目：临港实验室(LG202102-01)；上海市科技重大专项

作者简介：孔祥泰，男，硕士生　E-mail: kongxiangtai@simm.ac.cn　***通信作者：**李叙潼，女，博士　E-mail: lixutong@ simm.ac.cn　郑明月，男，博士，研究员　E-mail: myzheng@simm.ac.cn

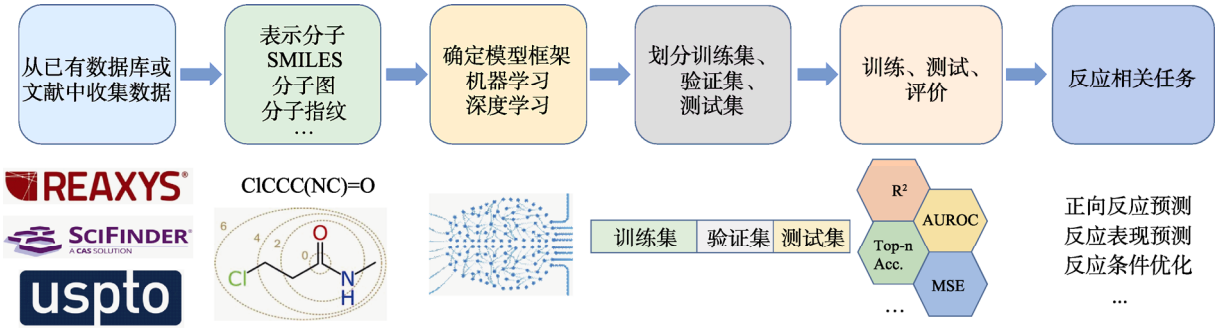

图 1 机器学习/深度学习应用于化学反应预测的流程

Fig. 1 Workflow of applying machine learning or deep learning in chemical reaction predictions

在化学反应预测中，常见的分子形式是SMILES，其特点是确定性和多样性，每个 SMILES 只对应一种分子，而一种分子可以通过多个 SMILES 表示。随机确定一个初始原子，给分子中的每个原子分配数字后遍历分子图，即可获得 SMILES。由于其简单的符号表示，相较于其他描述方法节省大量的空间，并且可以借助 RDKit 等工具快速转换为其他格式。

分子的 InChI 表示分为多层，之间用“/”分开，不同的层代表不同的含义，从前到后一般为版本号、分子式、原子连接信息和氢原子连接信息。InChI 表示相较于其他表示更难看出分子的原始结构，只能通过一定的算法进行转换。SMARTS (SMiles ARbitrary Target Specification) 是对 SMILES 的一个拓展，通过引入更多的逻辑符号和描述符，使其在子结构搜索等方面具有更为强大的功能，可以利用 RDKit 将 SMILES 转换为 SMARTS。

近年来，随着图神经网络(Graph Neural Network，GNN)在计算机视觉[4]和自然语言处理领域[5]取得了巨大成功，越来越多的工作关注于基于分子图表示的逆合成预测和反应路线规划上。分子图的思想在于将原子视为节点，将化学键视为边，通过点边的组合来表示分子。基于分子图的表示一方面比线性符号保存更多的信息，另一方面还可以通过加入原子坐标和二面角等信息表示分子的三维结构，使分子表示更加准确。

由于分子的表示形式多种多样，化学反应也有不同的描述方式。以苯甲酸和乙醇简单的酯化反应生成苯甲酸乙酯为例，图 2 中分别用化学式、SMILES、InChI、SMARTS 和分子图描述化学反应中的分子。

此外，分子描述符和分子指纹被广泛用于表示分子的特征。其中，分子描述符是分子在理化性质等方面的度量[6]，例如 Multiphilic 描述符[7]可以表示分子中不同位点的亲电性或亲核性，可作为额外的特征提高模型在化合物选择性预测任务上的表现[8]。分子指纹，通过向量描述分子的子结构组成，例如扩展连通性指纹，将分子中的每个原子视为环形中心[9]，迭代地获取原子和子结构的特征，通过引入更多全局和局部的化学结构信息，提高反应预测模型对分子结构变化的敏感性[10-11]。

2 正向反应预测

正向反应预测是有机合成中的一个重要问题，即给定反应物和试剂预测产物。如果能够正确地预测反应产物，就可以帮助化学家们更快地

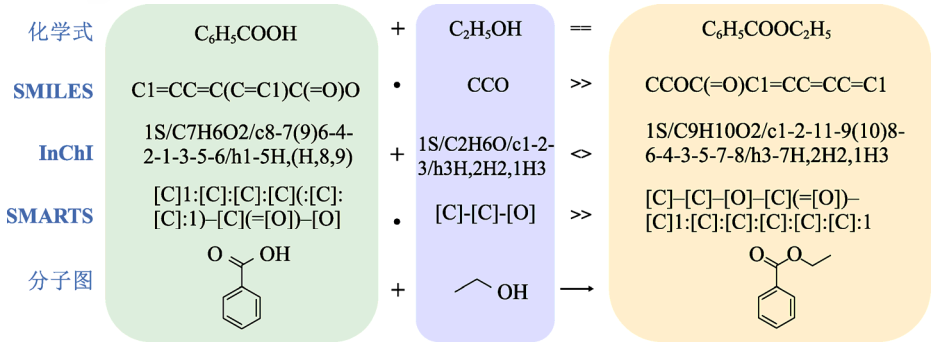

图 2 反应的不同描述方式

Fig. 2 Different ways of describing reactions

判断终产物，但由于缺乏精确的反应条件数据(浓度、温度、时间等)让正向反应预测成为了不适定问题[12]。

正向反应预测问题的解决方法可以大致分为3 类(图 3)：基于模板的方法、基于序列的方法、基于图的方法。

2.1 基于模板的方法

基于模板的方法是基于人工定义的规则或从反应数据中提取的反应模板(反应中原子和键的变化规则)，将产物的生成问题转换为模板的选择问题，然后将反应物与排名靠前的模板进行匹配，以产生候选产物。Wei 等[13]将反应物和试剂的分子指纹组成的反应指纹作为神经网络的输入，预测 17 种不同反应类型的概率，然后将反应物和最可能发生的反应类型匹配从而预测产物。Coley 等[14]将自动提取的反应模板与神经网络相结合开发了正反应预测模型，他们使用正反应模板生成一组可能的产物，然后通过神经网络预测其中的主要产物。

2.2 基于序列的方法

基于序列的方法将反应预测问题视为反应物、试剂和产物的 SMILES 之间的机器翻译问题。Schwaller 等[15]基于 Transformer 架构开发了 Molecular Transformer，模型由编码器和解码器构成。将反应物和试剂的SMILES作为编码器的输入，通过多头注意力层将其编码为向量表示，然后通过解码器自回归地逐字符生成产物的 SMILES。该模型可以较为准确地预测化学、区域以及对映选择性，还可以通过估计不确定性对反应进行排序[15]。

有研究表明，由于化学分子SMILES的多样性，在训练和推理过程中对SMILES使用数据增强技术可以提高模型精度。Tetko 等[16]研究了多种增强策略(只对产物或同时对反应其他组分进行数据增强)，证明了数据增强技术可以使模型学习分子的不同表征，从而提高模型的泛化性能。Zhang 等[17]同样研究了 SMILES 的数据扩增策略对于模型性能的影响，同时使用迁移学习在 Baeyer-Villiger 反应中进行了性能测试。结果证明，迁移学习和数据增强 2 种技术能够帮助模型在反应预测任务中学习到 SMILES 的复杂语法，获取更丰富的化学知识，并且可以有效地处理化学反应数据稀缺的问题。

2.3 基于图的方法

除了将反应预测任务视为自然语言处理的机器翻译任务，还可以使用图的方式对其进行处理。GNN 是一种基于图的深度学习方法，对输入的图结构数据通过神经网络模型挖掘图中蕴含的复杂特征。一般先生成图中节点和边的隐向量，通过卷积等消息传递策略考虑邻居节点或边对中心节点的影响，进而对节点和边的特征进行更新，得到更准确的图表示，从而更好地处理图数据。

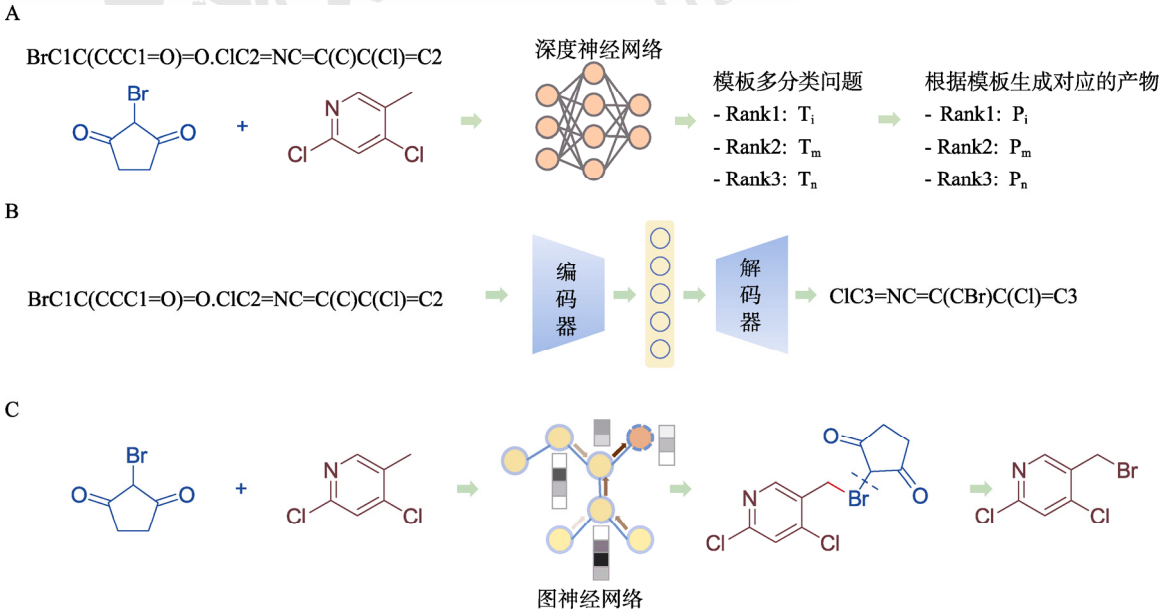

图 3 正向反应预测的方法分类

A–基于模板的方法；B–基于序列的方法；C–基于图的方法。

Fig. 3 Classification for forward reaction prediction approaches

A–template-based approach; B–sequence-based approach; C–graph-based approach.

Jin 等[18]从化学家分析有机反应时的思维模式出发，将反应预测任务分为反应位点识别和反应结果打分 2 个部分。首先将反应视为图中 2 个或多个节点之间边发生的变化，通过一种图卷积神经网络——Weisfeiler-Lehman Network(WLN)[19]生成分子图的同构不变表示，从而计算原子之间键变化的可能性。然后枚举候选产物并计算这些候选产物的概率，得到可能性最大的产物。Coley 等[20]同样将反应预测任务分为 2 个部分，并使用 WLN 学习分子图的表示，捕获反应相关特征。还对 Jin 等[18]的模型进行完善，在枚举阶段加入化学价规则和连通性限制进一步过滤掉无效的分子，模型的效果得到进一步提升。

3 反应表现预测

3.1 化合物的可合成性

在实践中，很多理论上性质理想的分子在之前尚未有合成路线记录，这为新药研发带来一定困难。因此，准确计算生成分子的可合成性是一项非常重要的任务。可合成性预测任务按照原理不同主要可以分为 2 类——基于逆合成的方法和基于片段的方法。基于逆合成的方法借鉴药物化学家的思维，通过断键考虑合成目标分子所需要的一般反应物，再在已有的数据库中对其进行搜索，进而判断目标分子的可合成性[21]；基于片段的方法则是将目标分子分解为片段，按照片段在数据库中的打分或出现频率评估目标分子的可合成性[22]。

3.1.1 基于逆合成的方法 Huang 等[23]提出了 RASA 模型(图 4)，首先为目标分子构建有限制的合成树和 RASA 打分函数，之后根据反应的复杂性和分离纯化难度对反应路线打分，筛选出最优路线。因为对合成树进行条件限制，大大减少其中排列组合的情况，节省打分时间。之后选取 100 个有专家打分的化合物进行训练，训练后测试集分别用 RASA 和专家打分进行评估，两者的相似度约为 0.8，说明 RASA 方法的有效性。

分子的图表示可以和 GNN 很好地结合。Liu 等[24]基于 GNN 设计了 RetroGNN 模型估计分子的可合成性。首先，通过生成模型随机生成分子，基于已有的逆合成软件搜索生成分子的合成路线，之后用上述数据训练 RetroGNN 模型，预测结果可视为分子的可合成性估计，将该分数与其他评分函数结合起来用于小分子的筛选，最后在分子设计平台上进行评估，结果表明该方法预测的分子具有最高的可合成性。RetroGNN 的另一优势在于筛选速度快，在单个 GPU 上仅需 9 d 就可筛选约 10 亿个分子。

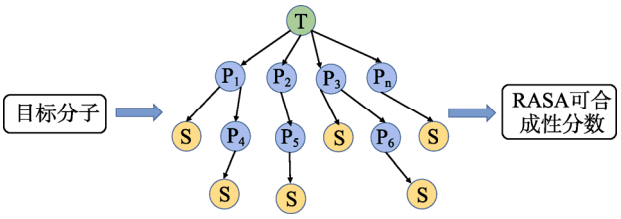

图 4 RASA 模型
Fig. 4 RASA model

3.1.2 基于片段的方法 反应可合成性预测除了基于逆合成的方法以外，还可通过基于片段的方法进行预测。该方法基于一个假设，即片段在分子库中的出现频率与其可合成性有关——易于制备的亚结构在分子库中出现的频率高。Ertl 等[25]将可合成性得分(Synthetic Accessibility score，SA score)定义为两部分，分别是片段的贡献得分和复杂性惩罚分数。片段的贡献得分评估片段的出现频率，复杂性惩罚分数评估分子中复杂结构特征(如手性中心和大环、对映复杂度等)。通过这种方法计算出来的 SA score 与化学家给出的化合物合成难度排名相关性很高，证明此种方法具有良好的可靠性。Voršilák 等[26]从 SA score 的概念出发，通过贝叶斯分析方法和片段在化合物库中出现的频率计算合成贝叶斯可及性(SYnthetic Bayesian Accessibility，SYBA)分数，并假设每一个分子片段都是独立的，每个片段都分配有其各自的 SYBA 分数。将分数相加获得整个分子的 SYBA 分数，从而区分难易合成分子。

3.2 反应活性预测

对于一个化学反应，如果可以通过使用催化剂提高反应活性，会对工业产生具有巨大的应用价值。机器学习技术目前已经在催化领域有很多应用[27]。

Smith 等[28]基于水煤气变换反应数据集，搜集 2 228 个数据点作为训练集，结合神经网络和实验描述符，搭建了一个预测反应催化活性的模型。他们使用主成分分析方法识别信息量丰富并且对模型准确性影响较大的描述符和实验点，同时还提出了一个受约束的主成分分析方法，筛掉由于经济、技术等原因无法实现的实验点，之后将投

影到低维的 27 个描述符作为输入，训练了一个可以预测催化活性的神经网络。但由于文献中的催化剂配方和反应条件高度集中在信息空间的一段狭窄区域内，将近 90%的可用实验信息都是来自反应条件的变化，故对催化剂配方的反应活性可预测性存在较大限制。

有些数据标签难以大量获得，尤其是在药物发现领域[29]。主动学习基于不同样本对任务的重要程度不同，选择性地标记少量数据，以最小成本提高模型性能(图 5)。Zhong 等[30]开发了一种结合火山图、密度泛函理论以及主动学习的框架，并使用随机森林和增强树回归的机器学习算法，从头筛选电化学催化剂。具体来讲，他们选取了 244 种不同的含铜金属晶体，列举了 12 229 个表面和 22 8969 个吸附位点，对于其中的一个子集使用量化软件模拟计算其吸附能并作为供机器学习训练的数据，然后与火山图相结合，以预测最具催化活性的位点。再计算这些位点的吸附能，为机器学习模型提供额外的训练数据，即产生了一个自动化框架，该框架可以系统地搜索具有接近最佳吸附能的表面和吸附位点，经过大约 4 000 次模拟以后产生了一组可以用于实验测试的候选材料。

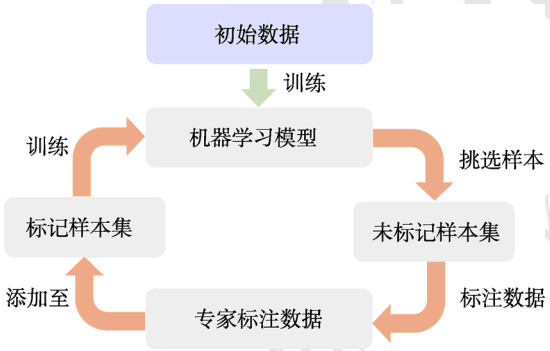

图 5 主动学习
Fig. 5 Active learning

为了捕捉到催化剂性能与其他多种特性之间的复杂关系，Yang 等[31]使用机器学习方法在二氧化碳还原反应体系中预测和筛选催化剂。他们使用了几种无须从头计算的描述符，包括电子描述符、底物的几何描述符等，同时还引入了广义配位数来描述底物的几何效应，然后采用了 10 种不同的机器学习回归算法，其中梯度提升回归是表现最优的模型。通过上述策略，模型能够揭示过渡金属和合金活性中心的大小、不同吸附物与底物的耦合机制等。

3.3 反应选择性预测

反应选择性的正确预测可以帮助化学家选择原料、时间成本更低的反应路径，从而更快地实现目标产品的合成。由于化学结构与基团选择性之间的复杂关系，高效、准确的反应选择性预测模型的建立具有广阔前景的同时伴随一定的挑战。

反应选择性主要包括区域选择性、化学选择性和对映选择性等。区域选择性表示当反应物中有多个潜在的反应位点，受反应条件的限制(如环境、温度、催化剂等)，反应往往只发生在其中某个位点上[32]；化学选择性表示在特定的反应环境下，反应物中的某些官能团发生指定反应[33]；对映体选择性表示一对对映异构体参与的反应中，其参与优先度并不相同，如特定的酶促反应，它也可以表示在一定的反应条件下，更趋向于生成某一对映异构体的情况。

3.3.1 区域选择性 反应机制往往基于已有的化学反应规律，通过计算手段提高对反应位点的预测准确性不仅有助于反应产物的预测，还可以加深对反应机制的理解。

Guan 等[34]构建了 QM-GNN 模型，将机器学习到的反应表征与选定的量子力学描述符结合起来，预测亲电芳香取代反应的区域选择性。首先通过 GNN 获得输入分子的每个原子和边的初始特征向量，通过量化计算获得键长和键级，在 WLN 中更新以原子为中心的特征向量，之后通过全局注意力机制考虑较远原子的影响进一步更新，最后获得分子表征，通过神经网络获得预测结果。该模型被证明适用于各种化学空间，作者从数据库中获取 3 类取代反应，包括芳香族 C-H 功能化、芳香族 C-X 取代和其他取代反应，每类反应使用 5 000 个数据进行训练，融合模型在预测结果方面分别达到了 89.7%、96.7%和 97.2%的准确率，并且运行效率高，平均只需 70 ms 就能从反应的 SMILES 中预测区域选择性。

3.3.2 化学选择性 在有机反应中，如何高质量地预测官能团的反应活性是一个亟待解决的问题，Tavakoli 等[35]首先使用 DFT 计算了 2 400 多个有机分子的甲基阳离子亲和力和甲基阴离子亲和力，从而建立了一个大型的化学反应性分数数据集，其中涵盖了 53 种不同的化学反应。首先将分子建模为关系图网络，获得分子的邻接矩阵，

然后通过卷积神经网络递归地更新每个原子的表示，并预测分子的反应性，之后使用处理好的数据集与分子结构的不同表示相结合对模型进行训练。十倍交叉验证结果表明，应用信息输入指纹的图注意神经网络产生了最准确的化学选择性估计。

3.3.3 对映体选择性 在非对称反应中，对映选择性的预测往往依赖于之前所积累的实验数据，而如今随着互联网共享反应数据的爆发式增长以及深度学习方法的蓬勃发展，通过结合现有的数据设计算法来预测反应的对映选择性成为可能。Jolene 等[36]提出了一个整体的、数据驱动的方法用于反应预测。首先结合已发表的对映选择性数据集，基于 1,1'-联二萘酚的一系列衍生物参与的亲核加成反应建立了统计模型。对模型中采用的描述符进行分析，发现亚胺描述符具有最高的权重系数，表示亚胺底物在对映选择性中十分重要。为了对该模型的预测性能进行评估，将其发现的对映选择性相关的催化机制迁移到外部测试集中，平均 G 的绝对误差为 0.37 kcal·mol^{-1}。使用 E-亚胺机制模型的结果稍有改善，平均误差为 0.24 kcal·mol^{-1}。

在全化学空间中探索可能的化合物对计算技术和成本要求很高，为了解决这个问题，Huang 等[37]将分子中的原子片段与主动学习相结合，该方法有助于提升模型在小样本对映体选择性预测任务上的准确性、可拓展性[38]。具体来说，首先对生成的分子子图进行过滤，筛选其中有代表性的片段，保存数据库中未见过的片段，直至收集完子图中所有合理的片段，将其用作训练集。实验结果表明，模型可以在小样本的基础上，对有机小分子的预测达到 2 kcal·mol^{-1} 的精度。

3.4 反应产率预测

由于化学合成路线通常包含多个反应，而反应产率低则会导致整条路线产率指数级别的降低，所以如果可以根据预测的相关反应产率推荐反应路线，则会产生非常大的经济效益。目前，开发了多种基于人工智能的反应产率预测模型，其采用不同的分子表示作为模型的输入(图 6)。

Ahneman 等[39]通过高通量实验技术在 C-N 交叉偶联反应体系下构建了包含约 4 600 个反应的产率数据库。对于反应体系中的不同组分(如芳香卤化物、添加剂、钯催化剂配体、碱等)，使用专业的量子化学软件计算了大量的体系相关描述符，

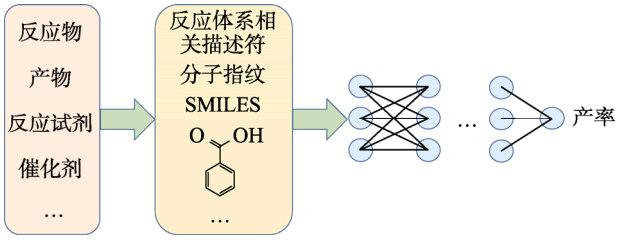

图 6 基于不同分子表示的反应产率预测

Fig. 6 Different molecular representations-based yield prediction of chemical reactions

如 HOMO、LUMO、电负性、偶极矩、分子体积、分子质量、表面积核磁共振位移以及静电荷等，共计 120 个描述符。使用描述符作为输入，训练机器学习模型，并分别计算预测产率和观测产率的 R^2 以及均方根误差来测试模型。除了这种基于描述符的方法以外，后来又有很多方案在 C-N 交叉偶联反应数据集中进行测试。Sandfort 等[40]将 RDKit 计算的多种指纹进行叠加，组成 71 374 位的混合指纹代替体系相关的描述符作为输入。这种多指纹特征方法无须手动特征化，提高了处理不同体系时的普适性。但使用多指纹特征作为输入的神经网络模型非常容易过拟合，作者使用随机森林模型改善了这一情况，最终性能可以媲美使用体系相关描述符作为输入的方法。

除了特征工程外，近年来端到端的方法崭露头角。Schwaller 等[41]直接使用反应的 SMILES 作为输入，扩展了自然语言处理架构的应用，以基于文本的方式预测反应相关信息，对之前预训练的 rxnfp 模型[42]进行了微调。他们将反应的 SMILES 输入到基于 BERT 的反应编码器中，得到 256 维的向量后通过回归层对反应产率进行预测。这一方法的性能优于前 2 种方法(基于描述符和基于多指纹特征的方法)，证明了预训练技术和 SMILES 表征的有效性。

由于化学分子可以天然地用图的形式表示，GNN 同样可以用于反应产率预测任务。Sato 等[43]使用经过预训练的 Mol2Vec 模型对原子进行向量化作为消息传递神经网络的初始原子嵌入，同时结合自注意力机制更多地关注对反应重要的原子，将得到的向量表示拼接后输入全连接层以及激活函数层，从而预测反应产率。该模型在性能上与 Ahneman 等[39]的模型相近，证明了只使用包含分子结构信息的模型可以与使用量子化学描述符信息的模型性能相媲美。

Nielsen 等[44]同样使用高通量实验技术构建了

一个包含约 700 个磺酰氟脱氧氟化反应的数据集，与 C-N 交叉偶联反应数据集相比，这个训练集更小、底物类型更多，并且涉及的反应机制更丰富。作者同样计算了相关的原子、分子级别描述符等作为输入训练机器学习模型，结果表明，随机森林算法的预测性能较好，并且具有一定的鲁棒性。在机器学习领域，数据的重要性不言而喻，Eyke 等[45]将基于机器学习的反应产率预测模型与主动学习以迭代的方式相结合，即从所有可能的子集迭代地选择包含信息量最大的实验。模型方面作者将反应中涉及变化分子的摩根指纹作为输入从而训练神经网络，并对已存在的数据集进行了回顾性分析，将使用主动学习训练的模型和对反应数据随机选择训练的模型进行比较，结果表明该策略是一种有用的实验规划工具，可以让药物化学家使用更少的时间和精力获得更高质量的数据集。

3.5 活化能预测

反应活化能是评估一个反应可行性的重要指标。Von Rudorff 等[46]构建了包含 7 500 个以乙烷为骨架的反应物数据库，选取电子效应较大而空间结构较小的基团作为取代基，如硝基、氰基等，经过计算筛选获得 4 466 个经过验证的过渡态构象，其中 2 785 个属于双分子亲核取代反应，1 681 个属于双分子消除反应，使用岭回归中的 delta 方法[47]进行计算。基于搜索的反应物构象和过渡态构象，可以计算出反应的活化能，但由于构象搜索的空间限制，存在负活化能的情况，亟待后续改进完善。

Singh 等[48]基于之前对于分子描述符和反应活化能预测的众多研究，他们发现在活化能预测中最重要的描述符是参与反应的原子表面键能，此外添加 7 个额外的描述符还可以使 DFT 计算的活化能 MSE 获得进一步下降。具体来说，首先收集了 315 个不同类型反应的活化能，在传统特征的基础上，加入 3 个新的特征：表面配位、反应中的断键数和参与断键的原子类型。实验结果表明，相较于多项式的线性回归，神经网络可以取得更优的预测性能。

4 反应条件优化

反应条件是化学反应的重要组成部分，更易达到、成本消耗更低、有助于提高反应产率的条件对化学反应至关重要。传统的反应条件优化依赖于化学家的经验，优化某个条件或是进行条件的组合变换。但由于化学反应机制的复杂性和多样性，想要一次获得最优的反应条件是十分困难的，往往需要多次尝试，而这也意味着更多的时间成本和实验成本，因此，如何寻找有效的优化反应条件的方法是一个亟须解决的问题。

Gao 等[49]开发了基于神经网络的模型来预测合适的反应条件(图 7)。首先从 Reaxys 数据库中收集了约一千万个反应进行训练，用于预测反应的催化剂、溶剂和反应温度等。具体来说，将反应条件的预测任务分为 2 种类型，催化剂、溶剂和试剂的选择是多分类问题，温度的预测是回归问题。该模型首先将反应物和产物的分子指纹拼接起来，通过全连接层和 ReLU 激活层，生成分子指纹的密集表示；将密集表示输入全连接层和激活层来预测反应的催化剂；然后将催化剂的隐向量与密集表示相连，来预测第 1 种溶剂；重复上述步骤，预测第 2 种溶剂、第 1 种试剂和第 2 种试剂；最后将催化剂、溶剂和试剂的表示拼接起来预测温度。该模型在单任务和多任务预测中都取得了不错的表现。

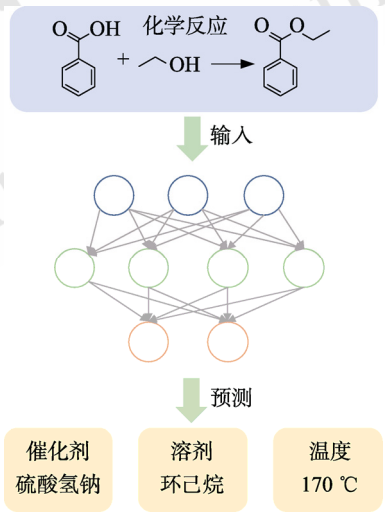

图 7 反应条件预测与优化

Fig. 7 Reaction condition prediction and optimization

除了反应温度、催化剂等条件，不同溶剂对反应结果也发挥着不可忽视的影响[50]。Walker 等[51]为了对现有模型预测溶剂的能力进行评估，比较了 3 种常用的方法，包括 k-近邻网络分析、支持向量机和深度神经网络。首先从 Reaxys 数据库收集了 4 500 万个反应，选取 5 种人名反应，包括 Diels-Alder、Friedel-Crafts、Wittig、Aldol addition 和 Claisen 反应，其中包含了催化剂和溶剂信息。

神经网络由多层感知机分类器和 ReLU 激活函数组成，输出层的维度等于训练数据中的溶剂数，对每种溶剂预测一个分数，并进行归一化处理。实验结果表明，k 近邻方法和神经网络模型在反应溶剂预测中取得了更优的结果。

5 结论

人工智能技术已在化学领域取得了不菲的成果，可用于预测反应产物和其他相关信息，如反应的产率以及反应物的可合成性、活性和选择性等，并且还可以对现有的反应条件进行优化，寻找更优的替代条件，降低实验门槛。计算机辅助化学反应预测从反应数据中挖掘化学规律，这有助于提高设计反应的成功率，减少化学家在探索阶段的试错成本，帮助发现和提出新的反应路线。目前已有很多软件或平台利用人工智能预测化学反应产率等数据，极大地推动了化学领域的发展。但由于商业数据库的限制，公开可用的反应数据集十分有限，未来可能需要依靠人工智能技术从海量文献中自动提取反应相关数据，并可结合主动学习的方法查询最珍贵的样本数据，训练更强大、准确的反应预测模型。或许在不久的将来，数据驱动的化学信息学可以与实验化学紧密结合起来，探索更有成效的实验方案，并为药物发现与设计提供技术支持[52]。

REFERENCES

[1] WEININGER D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules[J]. J Chem Inf Model, 1988, 28(1): 31-36.

[2] HELLER S R, MCNAUGHT A, PLETNEV I, et al. InChI, the IUPAC international chemical identifier[J]. J Cheminform, 2015(7): 23.

[3] WIGH D S, GOODMAN J M, LAPKIN A A. A review of molecular representation in the age of machine learning[J]. Wiley Interdiscip Rev Comput Mol Sci, 2022, 12(5): e1603.

[4] LIU Z, LIN Y T, CAO Y, et al. Swin transformer: hierarchical vision transformer using shifted windows[C]//2021 IEEE/CVF International Conference on Computer Vision (ICCV). Montreal, QC, Canada. IEEE, 2021: 9992-10002.

[5] NORCLIFFE-BROWNW, VAFEIAS S, PARISOTS. Learning conditioned graph structures for interpretable visual question answering[J]. NeurIPS, 2018(31): 8334-8343.

[6] REN W, KONG D X. On the correlation of molecule descriptors used in QSAR study[J]. Comput Appl Chem(计算机与应用化学), 2009, 26(11): 1455-1458.

[7] PADMANABHAN J, PARTHASARATHI R, ELANGO M, et al. Multiphilic descriptor for chemical reactivity and selectivity[J]. J Phys Chem A, 2007, 111(37): 9130-9138.

[8] HAMMAL H, BENHARREF A, ELHAJBI A. Elucidation of the chemo- and stereoselectivity of [1+2] cycloaddition reactions between α-cis-himachalene and dichlorocarbene using a multiphilic descriptor[J]. Moroccan J Chem, 2016, 4(4): 891-900.

[9] ROGERS D, HAHN M. Extended-connectivity fingerprints[J]. J Chem Inf Model, 2010, 50(5): 742-754.

[10] UCAK U V, ASHYRMAMATOV I, KO J, et al. Retrosynthetic reaction pathway prediction through neural machine translation of atomic environments[J]. Nat Commun, 2022, 13(1): 1186.

[11] UCAK U V, KANG T, KO J, et al. Substructure-based neural machine translation for retrosynthetic prediction[J]. J Cheminform, 2021, 13(1): 4.

[12] STRUBLE T J, ALVAREZ J C, BROWN S P, et al. Current and future roles of artificial intelligence in medicinal chemistry synthesis[J]. J Med Chem, 2020, 63(16): 8667-8682.

[13] WEI J N, DUVENAUD D, ASPURU-GUZIK A. Neural networks for the prediction of organic chemistry reactions[J]. ACS Cent Sci, 2016, 2(10): 725-732.

[14] COLEY C W, BARZILAY R, JAAKKOLA T S, et al. Prediction of organic reaction outcomes using machine learning[J]. ACS Cent Sci, 2017, 3(5): 434-443.

[15] SCHWALLER P, LAINO T, GAUDIN T, et al. Molecular transformer: A model for uncertainty-calibrated chemical reaction prediction[J]. ACS Cent Sci, 2019, 5(9): 1572-1583.

[16] TETKO I V, KARPOV P, VAN DEURSEN R, et al. State-of-the-art augmented NLP transformer models for direct and single-step retrosynthesis[J]. Nat Commun, 2020, 11(1): 5575.

[17] ZHANG Y, WANG L, WANG X Q, et al. Data augmentation and transfer learning strategies for reaction prediction in low chemical data regimes[J]. Org Chem Front, 2021, 8(7): 1415-1423.

[18] JIN W G, COLEY C W, BARZILAY R, et al. Predicting organic reaction outcomes with weisfeiler-lehman network[J]. NeurIPS, 2017(30): 2607-2616.

[19] LEI T, JIN W, BARZILAY R, et al. Deriving neural architectures from sequence and graph kernels[C]. ICML, 2017: 2024-2033.

[20] COLEY C W, JIN W, ROGERS L, et al. A graph-convolutional neural network model for the prediction of chemical reactivity[J]. Chem Sci, 2019, 10(2): 370-377.

[21] BABER J C, FEHER M. Predicting synthetic accessibility: Application in drug discovery and development[J]. Mini Rev Med Chem, 2004, 4(6): 681-692.

[22] YU J H, WANG J K, ZHAO H, et al. Organic compound synthetic accessibility prediction based on the graph attention mechanism[J]. J Chem Inf Model, 2022, 62(12): 2973-2986.

[23] HUANG Q, LI L L, YANG S Y. RASA: A rapid retrosynthesis-based scoring method for the assessment of synthetic accessibility of drug-like molecules[J]. J Chem Inf Model, 2011, 51(10): 2768-2777.

[24] LIU C H, KORABLYOV M, JASTRZĘBSKI S, et al. RetroGNN: Fast estimation of synthesizability for virtual screening and de novo design by learning from slow retrosynthesis software[J]. J Chem Inf Model, 2022, 62(10): 2293-2300.

[25] ERTL P, SCHUFFENHAUER A. Estimation of synthetic

accessibility score of drug-like molecules based on molecular complexity and fragment contributions[J]. J Cheminform, 2009, 1(1): 8.
[26] VORŠILÁK M, KOLÁŘ M, ČMELO I, et al. SYBA: Bayesian estimation of synthetic accessibility of organic compounds[J]. J Cheminform, 2020, 12(1): 35.
[27] KITCHIN J R. Machine learning in catalysis[J]. Nat Catal, 2018, 1(4): 230-232.
[28] SMITH A, KEANE A, DUMESIC J A, et al. A machine learning framework for the analysis and prediction of catalytic activity from experimental data[J]. Appl Catal B Environ, 2020(263): 118257.
[29] DING X Y, CUI R R, YU J, et al. Active learning for drug design: A case study on the plasma exposure of orally administered drugs[J]. J Med Chem, 2021, 64(22): 16838-16853.
[30] ZHONG M, TRAN K, MIN Y M, et al. Accelerated discovery of CO_2 electrocatalysts using active machine learning[J]. Nature, 2020, 581(7807): 178-183.
[31] YANG Z, GAO W, JIANG Q. A machine learning scheme for the catalytic activity of alloys with intrinsic descriptors[J]. J Mater Chem, 2020(8): 17507-17515.
[32] TEDDER J M. Which factors determine the reactivity and regioselectivity of free radical substitution and addition reactions? [J]. Angewandte Chemie Int Ed Engl, 1982, 21(6): 401-410.
[33] SHENVI R A, O'MALLEY D P, BARAN P S. Chemoselectivity: the mother of invention in total synthesis[J]. Acc Chem Res, 2009, 42(4): 530-541.
[34] GUAN Y, COLEY C W, WU H, et al. Regio-selectivity prediction with a machine-learned reaction representation and on-the-fly quantum mechanical descriptors[J]. Chem Sci, 2021, 12(6): 2198-2208.
[35] TAVAKOLI M, MOOD A, VAN VRANKEN D, et al. Quantum mechanics and machine learning synergies: Graph attention neural networks to predict chemical reactivity[J]. J Chem Inf Model, 2022, 62(9): 2121-2132.
[36] REID J P, SIGMAN M S. Holistic prediction of enantioselectivity in asymmetric catalysis[J]. Nature, 2019, 571(7765): 343-348.
[37] HUANG B, VON LILIENFELD O A. Quantum machine learning using atom-in-molecule-based fragments selected on the fly[J]. Nat Chem, 2020, 12(10): 945-951.
[38] SCHWALLER P, VAUCHER A C, LAPLAZA R, et al. Machine intelligence for chemical reaction space[J]. Wiley Interdiscip Rev Comput Mol Sci, 2022, 12(5): e1604.
[39] AHNEMAN D T, ESTRADA J G, LIN S S, et al. Predicting reaction performance in C-N cross-coupling using machine learning[J]. Science, 2018, 360(6385): 186-190.
[40] SANDFORT F, STRIETH-KALTHOFF F, KÜHNEMUND M, et al. A structure-based platform for predicting chemical reactivity[J]. Chem, 2020, 6(6): 1379-1390.
[41] SCHWALLER P, VAUCHER A C, LAINO T, et al. Prediction of chemical reaction yields using deep learning[J]. Mach Learn: Sci Technol, 2021, 2(1): 015016.
[42] SCHWALLER P, PROBST D, VAUCHER A C, et al. Mapping the space of chemical reactions using attention-based neural networks[J]. Nat Mach Intell, 2021, 3(2): 144-152.
[43] SATO A, MIYAO T, FUNATSU K. Prediction of reaction yield for Buchwald-hartwig cross-coupling reactions using deep learning[J]. Mol Inform, 2022, 41(2): e2100156.
[44] NIELSEN M K, AHNEMAN D T, RIERA O, et al. Deoxyfluorination with sulfonyl fluorides: Navigating reaction space with machine learning[J]. J Am Chem Soc, 2018, 140(15): 5004-5008.
[45] EYKE N, GREEN W H, JENSEN K F. Iterative experimental design based on active machine learning reduces the experimental burden associated with reaction screening[J]. React Chem Eng, 2020, 5(10): 1963-1972.
[46] VON RUDORFF G F, HEINEN S N, BRAGATO M, et al. Thousands of reactants and transition states for competing E2 and S_N2 reactions[J]. Mach Learn: Sci Technol, 2020, 1(4): 045026.
[47] RAMAKRISHNAN R, DRAL P O, RUPP M, et al. Big data meets quantum chemistry approximations: The Δ-machine learning approach[J]. J Chem Theory Comput, 2015, 11(5): 2087-2096.
[48] SINGH A R, ROHR B A, GAUTHIER J A, et al. Predicting chemical reaction barriers with a machine learning model[J]. Catal Lett, 2019, 149(9): 2347-2354.
[49] GAO H Y, STRUBLE T J, COLEY C W, et al. Using machine learning to predict suitable conditions for organic reactions[J]. ACS Cent Sci, 2018, 4(11): 1465-1476.
[50] KAMLET M J, ABBOUD J L M, ABRAHAM M H, et al. Linear solvation energy relationships. 23. A comprehensive collection of the solvatochromic parameters, .pi.*, .alpha., and.beta., and some methods for simplifying the generalized solvatochromic equation[J]. J Org Chem, 1983, 48(17): 2877-2887.
[51] WALKER E, KAMMERAAD J, GOETZ J, et al. Learning to predict reaction conditions: Relationships between solvent, molecular structure, and catalyst[J]. J Chem Inf Model, 2019, 59(9): 3645-3654.
[52] TAN X Q, XIONG J C, ZHU T F, et al. Development of drug design in China: 40 years of achievements[J]. Sci Sin Vitae(中国科学: 生命科学), 2019, 49(11): 1375-1394.

收稿日期：2022-08-24
(本文责编：沈倩)