



引用本文:《化学科学》, 2025
年, 第 16 卷, 854 页

本文的所有出版费用
由英国皇家化学学会支付费用

收稿日期: 2024 年 9 月 3 日 接
受日期: 2024 年 11 月 27 日

DOI: 10.1039/d4sc05946h

rsc.li/化学科学

介绍

近年来, 机器学习的快速发展极大地推动了其在化学领域的应用, 尤其是在辅助化学家方面。synthesis.¹⁻⁴ 近年来, 计算机辅助合成规划 (CASP) 受到了极大的关注, 并在药物 synthesis⁵⁻⁹ 和天然产物合成中展现出了其价值。^{10,11} 作为化学反应不可或缺的组成部分, 反应条件 (催化剂、溶剂和试剂) 也需要被准确预测。¹² 反应条件对于正向预测模型至关重要,^{13,14} 因为相同的反应物在不同的条件下可能会生成完全不同的产物。对反应条件的周全考虑不仅有助于合成规划算法中选择更可行的路线¹⁵, 而且有助于化学家理解模型的内在逻辑。这反过来又有利于所预测路线的实际实验实施。^{16,17}

人们已努力预测特定反应类型的条件。斯特鲁宾格等人¹⁸ 利用量子化学计算为门捷列夫反应设计溶剂。机器学习方法也得到了广泛应用。马尔库等人¹⁹ 利用多个模型构建了一个专家系统, 用于预测迈克尔反应的条件。马泽尔等人²⁰ 利用关系图卷积神经网络预测四种高价值反应类型的条件。阿夫奥尼娜等人²¹

Reacon: 一种基于模板和聚类的反应条件预测框架†

Zihan Wang, ‡^a Kangjie Lin, ‡^a Jianfeng Pei ^{*b} and Luhua Lai ^{*ab}



计算机辅助合成规划已成为有机合成领域的一项重要工具。预测反应条件对于应用规划好的合成路线至关重要。然而, 在提供多样化建议的同时确保预测的合理性仍是一个尚未充分探索的挑战。在本研究中, 我们引入了一种利用图神经网络、反应模板和聚类算法相结合的方法来预测反应条件。我们的方法在经过优化的 USPTO 数据集上进行训练, 在召回记录条件方面达到了 63.48% 的前 3 准确率。此外, 当仅关注同一聚类内的反应召回时, 前 3 准确率提高到了 85.65%。最后, 通过将该方法应用于近期发表的分子合成路线, 并在聚类级别上实现了 85.00% 的前 3 准确率, 我们证明了该方法能够提供可靠且多样化的条件预测。

org/10.1039/d4sc05946h 引入了一种人工神经网络, 用于根据氢化反应的效率对反应条件进行排序。Kwon 等人²² 应用图增强变分自编码器来预测交叉偶联反应的可行反应条件。Angello 等人²³ 开发了一种利用机器学习和实验机器人技术的工作流程, 以完成杂芳基 Suzuki-Miyaura 偶联反应通用条件的选择。利用高通量数据集结合主动学习方法等尝试也已 made.²⁴⁻²⁶

除了特定的反应类型之外, 还有一些研究专注于预测一般反应的条件。高 (音译) 等人²⁷ 将分子指纹与全连接神经网络相结合来预测反应条件, 确保了不同条件成分之间的紧密连接。

除了分子指纹识别和基于图的方法之外, 基于转换器的模型也被广泛应用于反应条件预测任务。施瓦勒等人²⁸ 使用转换器模型对给定目标分子的反应条件和反应物进行同步预测。贾梅-桑特罗等人¹⁴ 使用转换器模型对给定反应物和产物的反应条件进行预测。克鲁特等人²⁸ 在其提出的三重转换器循环框架中也应用了转换器模型来进行试剂预测。安德罗诺夫等人²⁹ 对反应条件成分进行了细分和排序, 然后训练了一个转换器模型来预测条件。王等人³⁰ 基于转换器架构开发了一个反应条件预测模型, 并引入了一种利用反应领域知识的预训练策略。钱等人³¹ 利用文本检索方法来定位给定反应的相关文本信息, 从而提高条件预测的准确性。

^aBNLMS, Peking-Tsinghua Center for Life Sciences, College of Chemistry and Molecular Engineering, Peking University, Beijing, 100871, China. E-mail: lhlai@pku.edu.cn ^bCenter for Quantitative Biology, Academy for Advanced Interdisciplinary Studies, Peking University, Beijing, 100871, China. E-mail: jfpei@pku.edu.cn

† 电子的 补充的 信息 (ESI) 可用的。 看 DOI:

<https://doi.org/10.1039/d4sc05946h>

‡ These authors contributed equally.

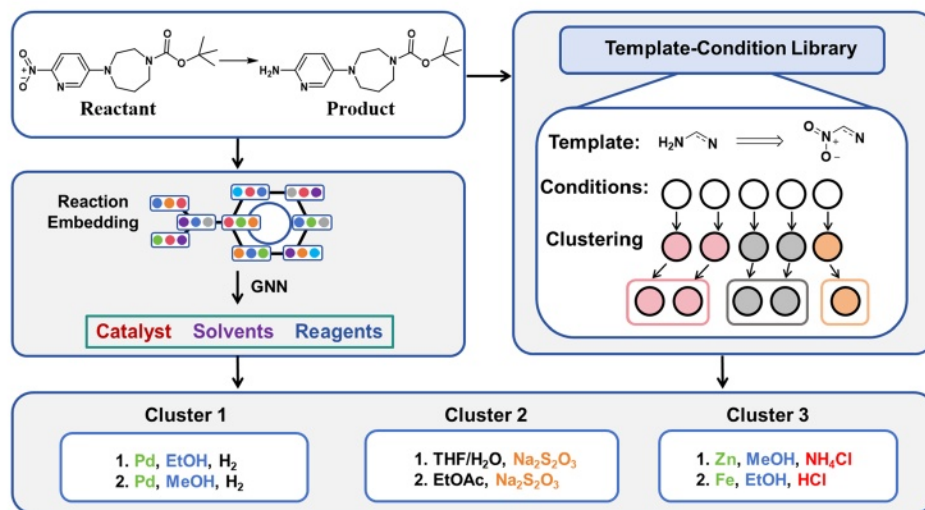


图 1 完整的状态预测工作流程示意图。

然而，在预测通用反应条件方面仍存在诸多挑战有待解决。一个性能良好的预测模型应当能够为完整的反应条件提供合理的建议，并确保不同组分（催化剂、溶剂和试剂）之间的兼容性。由于将反应物转化为产物的可行反应条件通常并非唯一，因此有效的预测模型应当呈现所有可能的反应条件，而这一点在以往的研究中往往被忽视。

鉴于反应条件中不同元素之间错综复杂的联系，我们提出了一种整体方法，在给出建议时将催化剂、溶剂和试剂视为一个集成系统。与采用不同模板的反应相比，采用相同反应模板的反应往往涉及更相似的反应机制。在此，我们介绍了 Reacon，这是一个基于反应模板的框架，利用有向消息传递神经网络（D-MPNN）来预测反应条件。³²该方法利用在相同模板下记录的反应条件来缩小模型选择范围。此外，我们还提出了一种基于标签的聚类算法，将相似的预测条件归为一组，以增强排名靠前预测的多样性，并简化实验化学家的条件选择。我们的方法工作流程如图 1 所示。我们进一步在几条最近发表的合成路线及其相应条件上验证了我们的方法。

方法

数据准备和预处理

在本研究中，我们使用了美国专利商标局（USPTO）的专利数据集，³³这是目前最大的且被广泛使用的有机反应开放获取数据集。我们将反应条件分为三部分：催化剂、溶剂和试剂。原始数据集包含了催化剂和溶剂的信息标注，但并未进一

步区分反应物和试剂。我们将具有原子映射的分子定义为反应物，而没有原子映射的分子则定义为条件成分。然而，考虑到氧化剂的特殊性，如果一个分子中只有氧原子被映射，我们也将将其归类为试剂。数据集处理的全面工作流程概述如下：

(a) 无法被 RDChiral³⁶ 解析的具有 SMILES³⁴ 表示法的反应将被移除。

(b) 使用 RDChiral³⁶ 提取每个反应的反应模板（半径 = 1），对于无法提取模板或出现次数少于 5 次的反应予以删除。

(c) 出现次数少于 5 次的含催化剂、溶剂或试剂的反应将被排除。

(d) 对于涉及多种溶剂和试剂的反应，我们根据其在相应模板下的出现频率进行排序。仅出现一次或存在一种以上催化剂、两种以上溶剂和三种以上试剂的反应条件将被剔除。

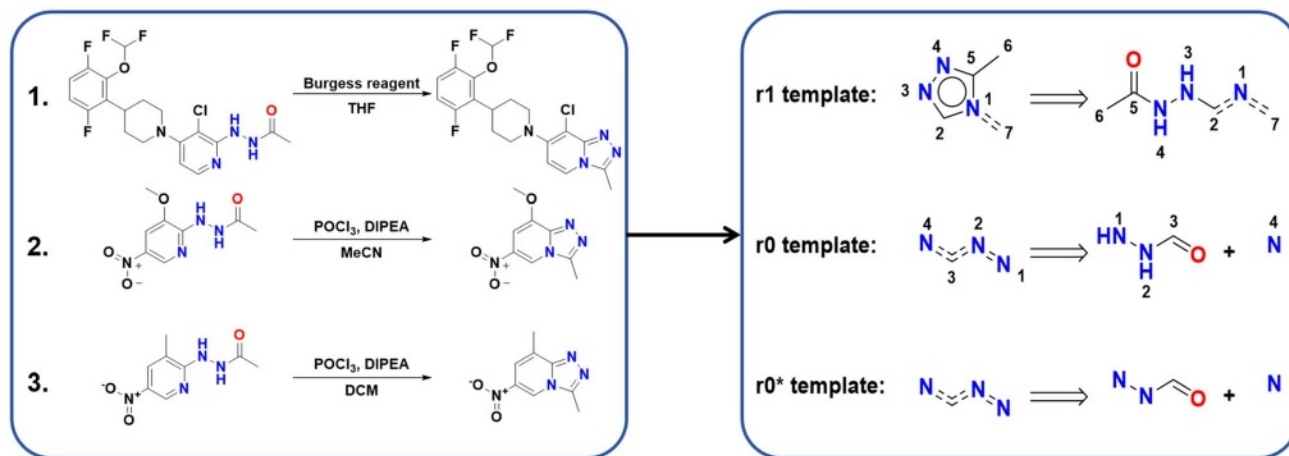
(e) 数据集按照 0.8 : 0.1 : 0.1 的比例随机划分为训练集、验证集和测试集。

最终的数据集包含 690872 个数据点；涵盖了 439 种催化剂、542 种溶剂和 2746 种试剂。

在完成初步的数据筛选后，我们使用训练数据集构建了模板条件库。如图 2A 所示，对于每条反应数据，我们提取了三种不同类型的模板：r1、r0 和 r0*。其中，r1 和 r0 是使用 RDChiral³⁶ 以不同半径提取的模板，而 r0* 则是从 r0 模板中仅保留原子和键得到的最简形式。我们总共获得了 26228 个 r1 模板、9755 个 r0 模板和 7106 个 r0* 模板。信息量较少的模板覆盖的化学空间更大，但特异性更低。因此，由相同的 r1 模板提供的候选反应条件应是最准确的，其次是 r0 模板和 r0* 模板。



A)



B)

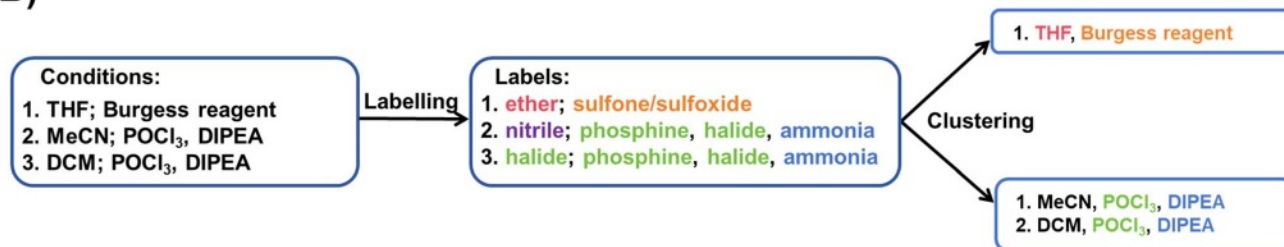


图2 (A) r1、r0 和 r0* 模板示例。(B) 条件聚类算法示例。

条件聚类算法

我们借鉴化学家对反应条件的分类方式，对每个成分使用 31 个标签进行了特征提取，如表 1 所示。这些标签包括：

(a) 特定官能团的存在，例如烯烃、醇和羧酸；(b) 特定元素的存在，主要指过渡金属、主族金属和还原金属（碱金属和碱土金属）；(c) 具有特定功能，例如氧化剂、还原剂和酸。这一判断依据是相应成分的存在，例如

(d) 此外，对于表现出电荷分离的化合物，被贴上“离子型”的标签，而那些不属于任何类别的则被标记为“其他”。除了“其他”这一标签外，所有标签均为非排他性的，允许每个组分拥有多个标签。详细的评估标准见表 S1。†

鉴于数据集中溶剂和试剂之间的区分常常模糊不清，要将两个条件归为同一类，必须满足以下标准：(1) 具有相同的催化剂标签（或两者都没有催化剂标签）；(2) 如果溶剂和试剂标签的总数超过 2 个，则至少要有 2 个标签重叠；否则，标签必须完全相同。详细的聚类过程见图 S1。† 图 2B 展示了一个条件标注和聚类的示例。

对于每一个新的反应，我们首先确定其是否属于现有的反应条件簇。如果属于，则将其添加到相应的组中并更新标签。否则，我们以新反应的标签建立一个新的簇。为确保不同类别之间没有交集，当一个反应可以归入多个类别时，我们将其分配给标签交集最多的簇。如果标签交集数量相同，则催化剂标签交集更大的类别具有更高的优先级。具体的聚类效果示例见表 S2。†

表 1 用于描述反应成分的所有标签及其分类标准

特征类型	数字	标签
官能团	21	烯烃、炔烃、醇、醚、醛、酮、羧酸、酯、酰胺、硝基、胺、卤化物、酰氯、酸酐、腈、芳香族化合物、砷/亚砷、膦、金属烷基化合物、硅烷、硫化物
元素	3	过渡金属、还原金属、主族金属
功能	5	氧化剂、还原剂、酸、路易斯酸、碱、离子型、其他
否则	2	



图神经网络模型

对于单独条件的预测，我们使用了 D-MPNN（有向消息传递神经网络）³⁷ 和 GAT（图注意力网络）³⁸ 模型。详细的网络描述见 ESI 第 2 节†，超参数选择的相关信息可在表 S3 和 S4 中找到。†

如图 3A 所示，该模型的输入由两部分组成：反应物的分子图以及反应物与产物之间的差异。对于每个分子图，顶点信息包括原子类型、键的数量和电荷等特征。键的信息包括键的顺序、异构性和是否形成环。³²

基线模型

为了便于模型性能的比较，我们设计了以下三个基准：

(1) 流行度基准：为评估模型在相同模板下区分不同条件的能力，³⁹我们设计了此流行度基准方法，以确定每个模板下出现频率最高的反应条件。

(2) 相似性基准：与 Retrosim 方法类似，⁴⁰此模型计算输入反应与相应模板条件库中反应的整体分子相似度，然后输出相似度得分最高的条件。整体相似度由产物相似度乘以反应物相似度来确定。

(3) 反应指纹多层感知机：这包括六个前馈神经网络模型，每个模型都有两个隐藏层（256、64）。这些模型用于预测单个反应的催化剂、溶剂 1、溶剂 2、试剂 1、试剂 2 和试剂 3。该模型的输入由两部分组成：试剂指纹（1024 维）和反应指纹（1024 维），反应指纹是通过从产物指纹中减去反应物指纹获得的。

(4) RCR：RCR（反应条件推荐器）是高（音译）等人提出的一种反应条件预测框架²⁷。它使用了多个神经网络模型来预测每个反应条件成分。为了完成整个反应条件的预测，采用了逐步预测策略，即首先预测催化剂信息，然后将相应信息引入溶剂预测，接着

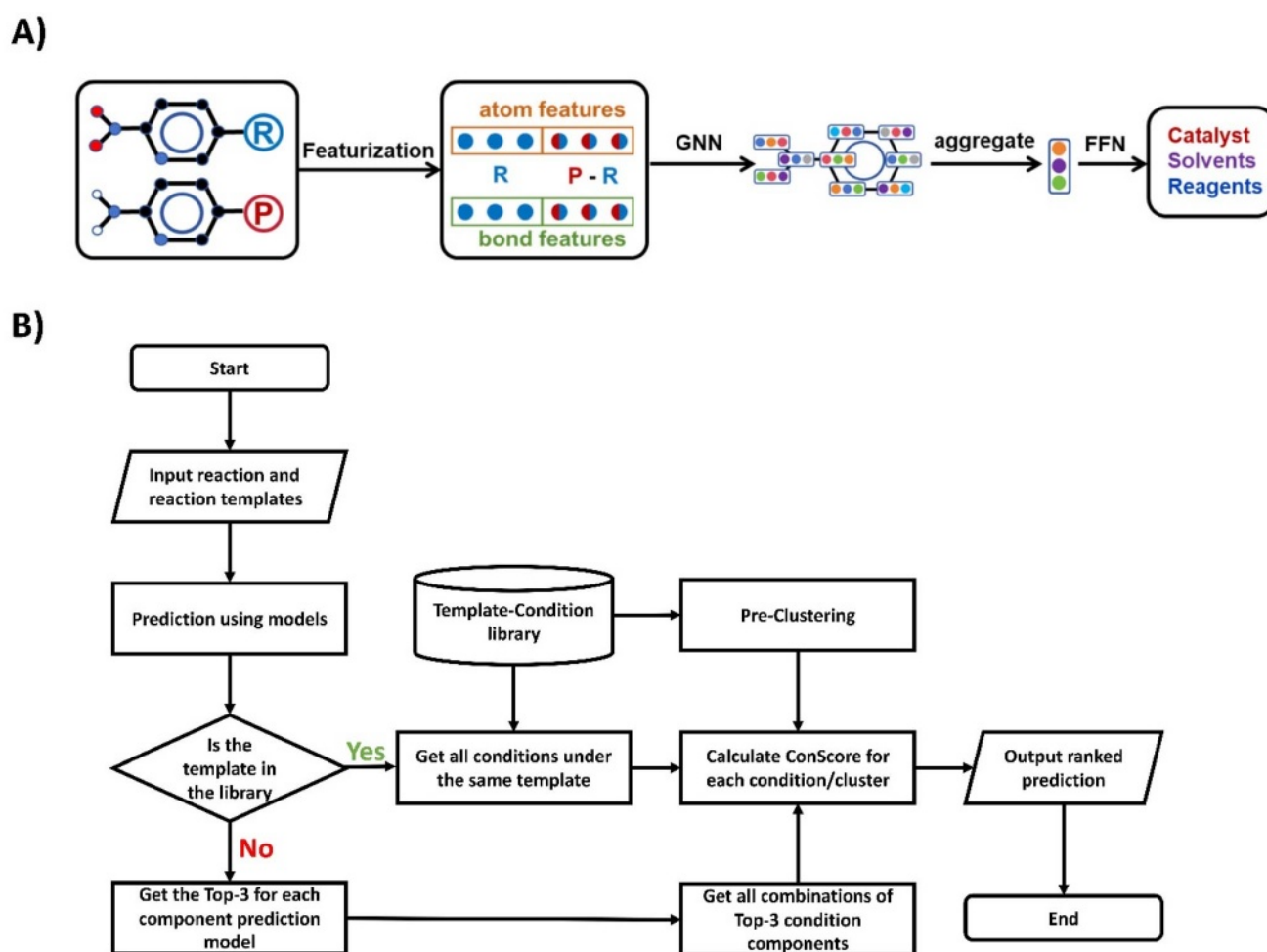


图 3 (A) 用于反应条件预测的模型架构概述。(B) 条件预测算法流程图。

后续的组件也是以同样的方式预测的。我们按照对原始文献的理解，使用相同的训练超参数重现了 RCR 模型。

条件得分指标

对于一组完整的反应条件（包括催化剂、溶剂 1、溶剂 2、试剂 1、试剂 2 和试剂 3），这些条件的条件得分（ConScore）可通过将各组分的选择概率相乘获得。

$$\text{ConScore}(x) =$$

$$\prod P_i(x) \quad (i = \text{cat, solv1, solv2, reag1, reag2, reag3}) \quad (1)$$

完整条件预测算法

条件预测算法的工作流程如图 3B 所示。对于每个输入反应，我们首先使用反应条件预测模型获取每个条件成分的概率。然后，我们在模板 - 条件库中搜索与输入反应相同模板下记录的条件作为候选条件。在生成候选条件的过程中，我们首先在模板 - 条件库中搜索相同的 r1 模板。如果未找到，则搜索 r0 模板和 r0* 模板。获得候选条件后，我们使用先前获得的概率计算每个候选条件的 ConScore。然后按照 ConScore 对这些候选条件进行排序，排序后的列表即为我们的预测输出。基于模板的反应预测的伪代码如图 S2 所示。[†] 对于预测反应的模板未出现在模板 - 条件库中的情况，我们为每个反应成分模型获取前 3 个预测，并生成这些预测的所有可能组合作为候选条件。

结果与讨论

预测反应条件的各个组成部分

我们首先分别训练了用于预测催化剂、溶剂 1、溶剂 2、试剂 1、试剂 2 和试剂 3 的独立模型。表 2 展示了 GAT 和 D-MPNN 模型与其他基准模型的性能。所有模型对催化剂的预测准确率都很高，约为 90%。这是因为很大一部分反应不需要催化剂，这简化

了模型训练，从而减少了模型之间的性能差异。同样的原因也适用于溶剂 2 和试剂 3 的预测。相比之下，预测溶剂 1、试剂 1 和试剂 2 则相对更具挑战性。在此，GAT 模型分别达到了 61.53%、66.74% 和 78.24% 的预测准确率，而 D-MPNN 模型分别达到了 61.93%、68.23% 和 80.44%，显著优于其他基准模型。在溶剂 1 和试剂 1 的预测准确率方面，GNN 模型与其他模型之间的差距最大，接近 30%。

我们还训练了一个多任务 D-MPNN 模型，以同时预测所有条件。与分别训练的模型相比，多任务模型能够更快地做出预测，并且在所有任务上的表现也都很出色。更多模型性能结果见表 S5。[†]

预测完全反应条件

在获得反应条件各组分的预测模型后，我们利用其计算模板条件库中与输入反应具有相同模板的所有完整反应条件的 ConScore 值。测试集上完整条件预测的性能列于表 3 中，其中准确率是根据预测的 ConScore 值排名前 N 的完整条件是否与真实条件匹配来计算的。在该框架下，GNN 模型比其他基线模型高出 5 - 10%，这表明它们能够有效地学习反应物与反应条件之间的关系。

相比之下，相似性基准并未达到预期的高精度，表现略低于流行度基准。这种差异可能归因于 USPTO 数据集包含大量条件各异但相似的反应实例。

除了随机划分之外，我们还进行了更具挑战性的时间划分。我们将 1976 年至 2014 年的数据用作训练集，2015 年的数据用作评估集，2016 年的数据用作测试集。我们的模型在该数据集上继续优于其他基准模型。模型的性能如表 S6 - S8 所示。[†]

我们还在 Wang 等人提供的反应条件数据集上测试了我们模型的性能，并将其与文献中报道的鸚鵡模型和 RCR 模型的性能进行了比较。

表 2 不同模型在测试集反应条件各组分上的个体预测性能

	催化剂	溶剂 1	溶剂 2	试剂 1	试剂 2	试剂 3
模型	第一名 (%)	第一名 (%)	第一名 (%)	第一名 (%)	第一名 (%)	第一名 (%)
受欢迎程度基准线	91.03	37.01	76.14	39.50	75.10	93.45
多层感知机	87.07	28.56	80.72	35.48	75.01	95.00
RCR	90.01	55.58	85.08	62.37	76.34	95.21
相似性基线	90.56	30.11	75.69	34.19	76.40	93.25
GAT	91.73	61.53	85.16	66.74	78.25	95.31
D-MPNN (深度消息传递神经网络)	93.12	61.93	86.61	68.23	80.44	96.05
D-MPNN (多任务)	92.45	59.76	86.12	66.72	79.63	95.95



表 3 不同模型在测试集条件预测中的表现

度量标准	模型	第一名 (%)	前三名 (%)	前 10% (百分比)
精确度	受欢迎程度基准线	36.02	56.19	73.01
	多层感知机	34.43	54.19	68.21
	RCR	28.56	37.18	42.03
	相似性基线	36.01	54.98	71.23
	GAT	41.67	61.46	76.51
	D-MPNN (深度消息传递神经网络)	44.52	63.49	78.55
	D-MPNN (多任务)	42.41	61.42	76.78

该文章³⁰ 我们的模型在该数据集上保持了领先的准确率，具体性能详情见表 S9 和 S10。[†]

聚类反应条件

尽管我们的模型展现出了令人满意的准确性，但我们也发现了一些问题。如图 4 所示，我们注意到模型所推荐的排名靠前的条件之间存在高度相似性，通常仅限于溶剂或试剂的替换。为了提供更加多样化的条件，我们进一步将相似的预测

条件进行聚类，使其更符合化学家的偏好。聚类算法在方法部分有详细描述。

图 4A 展示了一个将酮还原为甲基的反应。尽管 D-MPNN 模型在前两项预测中成功复制了真实情况，但排名靠前的预测结果高度相似，这给用户选择反应条件提供了有限的信息。经过聚类，多种替代条件被排在了前 3 个簇中，例如 Wolff-Kishner-Huang Minglong 还原法^{41,42} 或用路易斯酸替代酸。如图 4B 所示，真实情况涉及使用一种体积较大的

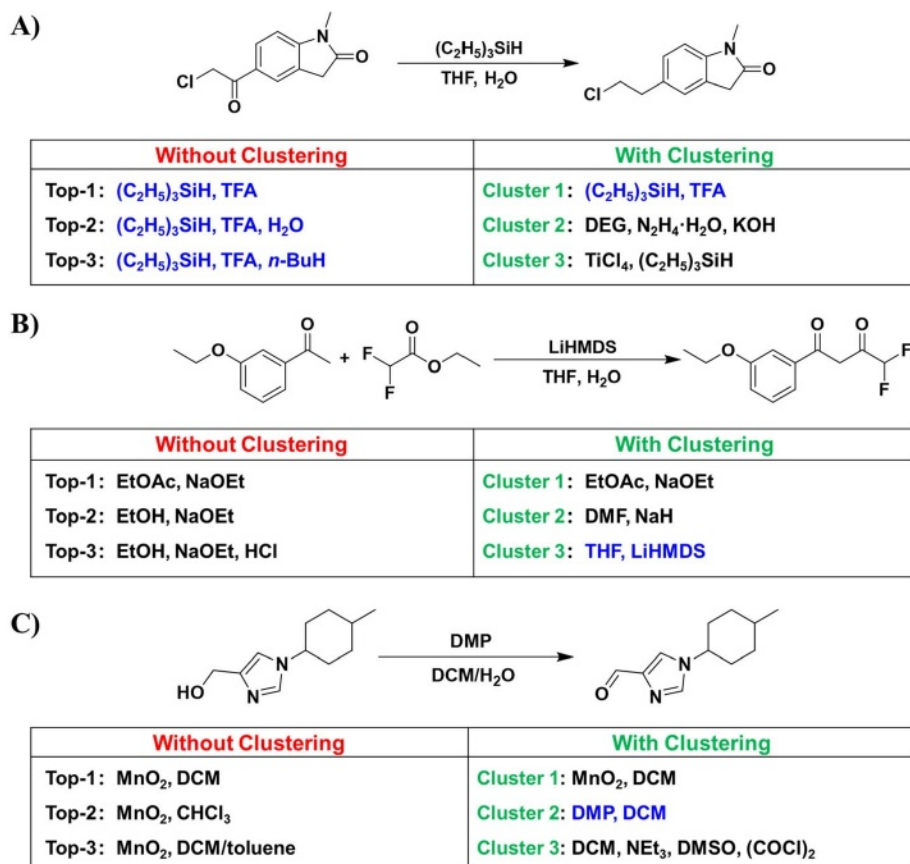


图 4 D-MPNN 条件预测器在使用和未使用条件聚类算法时性能的比较。(A) 酮还原反应。(B) 酯缩合反应。(C) 醇氧化反应。与真实情况相同的反应条件簇以蓝色突出显示。



在酯缩合反应中,这种碱基并不常见。因此,未进行聚类的 D-MPNN 模型未能在其前 3 个预测中预测出相应的条件。经过聚类后,相似的条件被整合,从而在前 3 个聚类中识别出了相应的条件。图 4C 展示了一个从醇到醛的氧化反应。前 3 个预测均使用了 MnO_2 作为氧化剂,只是溶剂不同。经过聚类后,除了正确识别出 Dess-Martin 氧化这一真实情况外, D-MPNN 模型还在前 3 个聚类中提供了 Swern 氧化。

除了能提升模型性能之外,我们还发现聚类有助于识别数据集中潜在的错误反应。这为筛选错误的反应条件提供了一种便捷手段,从而能够优化反应条件数据集的质量。²⁹ 详细示例见表 S2。†

为了评估条件簇层面的整体性能,我们引入了一个名为簇准确率指标。该指标用于评估模型提供的反应条件是否属于作为基准的前 N 个条件簇。不同簇的排序依据是同一簇内最高的 ConScore 值。此指标能够解决精确准确率指标存在的诸多问题。例如,在实际实验中,几种试剂可能具有相似的效果,可以相互替代。试剂的具体选择往往取决于化学家的偏好、实验室库存以及各种其他影响因素。因此,期望模型提供的完整反应条件与测试集中的基准完全一致,这种要求过于严格。

表 4 展示了在测试集上进行聚类后不同模型的表现。值得注意的是,在添加聚类算法后, D-MPNN 模型的 top-1 聚类准确率达到令人瞩目的 65.68%,比基准模型高出 10% 以上。此外, top-10 聚类准确率更是飙升至 96.11%。这些结果突显了该方法在有效预测多种反应条件方面的卓越能力。我们还测试了不同因素对最终聚类大小及聚类准确率的影响。结果表明,与使用不同的模板库相比,不同的聚类标准对结果的影响更大。关于聚类大小对模型性能影响的更多讨论见 ESI 第 5 节及表 S11 - S13。†

我们最终选择了 D-MPNN 模型,该模型在测试集上表现最佳。通过将其与

聚类算法相结合,我们构建了反应条件预测器 Reacon。

对存在问题的预测案例进行分析。除了强调 Reacon 的优势之外,我们还分析了其在分配正确条件簇时与实际情况不符的实例。图 5 展示了模型未能在前 3 个簇预测中预测出真实情况簇的代表性案例(更多示例见图 S3†)。在许多情况下,模型预测结果与数据集记录之间的差异并非由于模型输出不合理,而是归因于数据集问题或存在各种潜在的反应条件。

图 5A 展示了我们数据集中一个非常普遍的问题,即反应条件数据缺失。以酰胺形成反应为例,仅记录了四氢呋喃 (THF) 这一条件。然而,不使用羧酸活化剂进行此类反应通常会面临很大困难。相比之下, Reacon 提供了更全面且准确的条件。在 USPTO 数据集中,反应条件数据的错误记录也很常见。如图 5B 所示,在该反应中直接使用金属钠是危险的,也不是此类反应的合理条件。我们查阅了原始专利记录,发现实际使用的是甲醇钠,但在数据集中却被错误地记录为钠和甲醇,而甲醇钠正是由这两种物质制备而成。Reacon 不仅预测了真实条件,还预测到了一种常用的非亲核碱。图 5C 展示了 USPTO 数据集中的另一个常见问题,即无法区分多步操作的顺序。例如,对于这个酯水解反应,数据集记录了氢氧化钠和盐酸作为反应条件,但盐酸是用于后期处理而非同时添加。数据集可能会无意中混淆这些记录,这可能会给模型的学习过程带来挑战。在这种情况下, Reacon 仍能预测出准确的条件。

在上述案例中,模型预测结果与真实记录之间的差异主要源于数据集本身存在的问题。具体而言,在条件缺失或记录不准确的情况下, Reacon 能够有效地补充或修正这些条件,提供更合理的条件。因此,我们认为我们的模型在数据集优化任务方面也可能具有一定的潜力。²⁹

除了数据集方面的问题外,预测反应条件的另一个挑战在于存在多种可行的

表 4 不同模型在聚类后的测试集上的表现

度量标准	模型	第一名 (%)	前三名 (%)	前 10% (百分比)
聚类准确率	受欢迎程度基准线	54.04	79.55	91.89
	反应指纹 多层感知机	51.38	74.01	86.39
	相似性基线	53.63	77.34	91.70
	GAT	63.14	83.59	95.10
	D-MPNN (深度消息传递神经网络)	65.68	85.65	96.11
	D-MPNN (多任务)	63.88	84.17	95.91



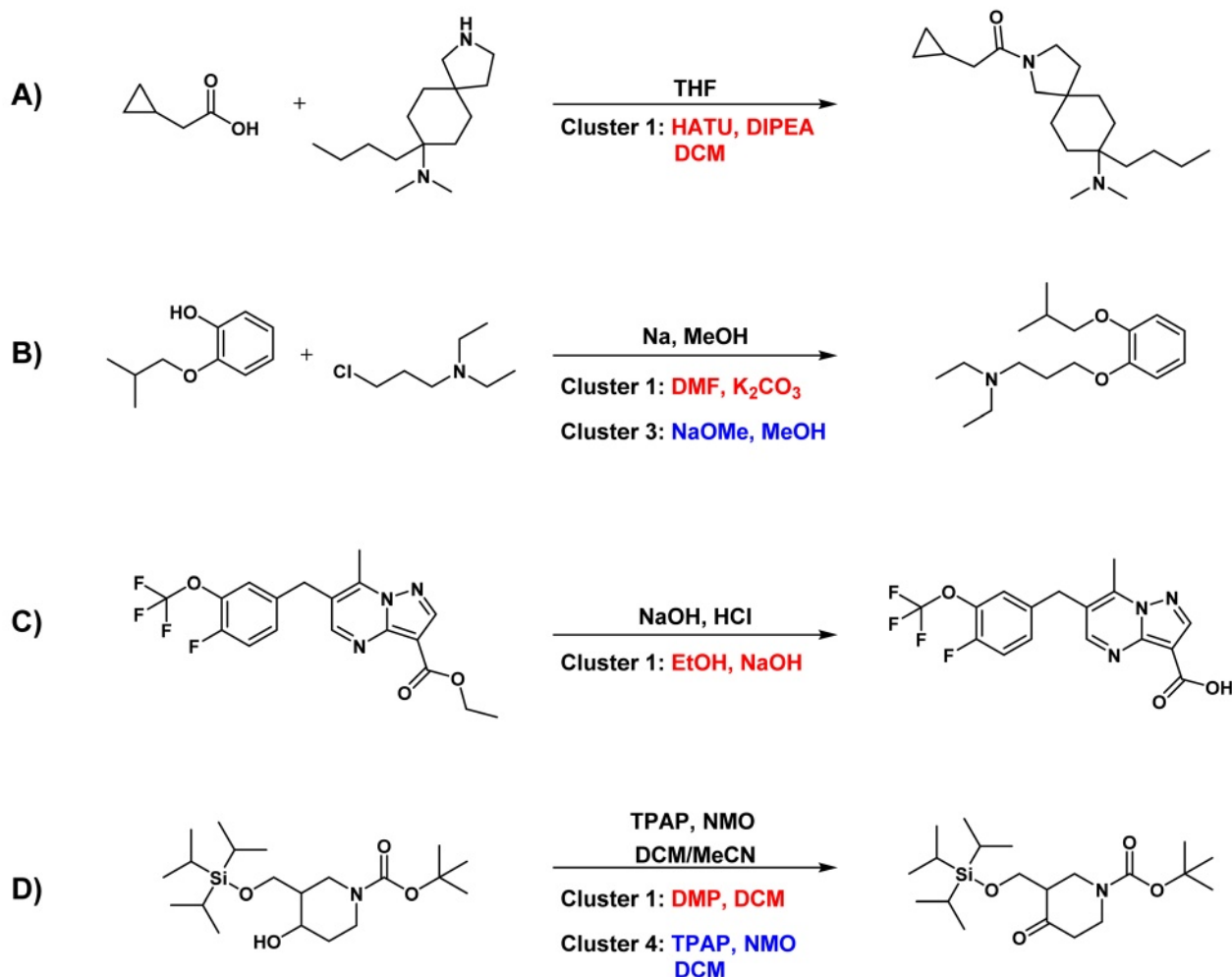


图 5 模型预测结果与真实情况不符的反应案例。数据集或文献中的条件已标注。

黑色表示预测结果与真实类型不一致，而与真实类型相符的预测结果则用蓝色标注，预测结果与真实类型不一致的用红色标注。(A) 酰胺形成反应。(B) 亲核取代反应。(C) 酯的水解。(D) 醇氧化成酮。

结果。图 5D 展示了一个反应，其中使用 TPAP（四丙基高钨酸铵）将醇氧化为酮，TPAP 是一种氧化剂，在 Reacon 的预测中被归入第四等级的簇。然而，Reacon 推荐的处于前两个等级簇中的 Dess-Martin 氧化法⁴³和 Swern 氧化法⁴⁴也被证明是高产率的氧化方法。尽管该模型有时无法预测出与文献报道一致的条件，但它仍能提供可行的条件。

对实际合成路线的评估

为了进一步评估我们模型的实际性能，我们从近期发表于《药物化学杂志》的文章中选取了 12 条药物分子合成路线（100 个反应），并用 Reacon 对其条件进行了预测。这项任务颇具挑战性，原因有以下几点。首先，部分试剂的选择具有偶然性，取决于化学家的偏好或实验室的库存情况。其次，文献中报道的条件应是化学家经过刻意优化的结果，这很难从一般数据集中学习到。总体而言，我们的

预测器达到了 39.00% 的前 3 位准确率和 85.00% 的前 3 位聚类准确率。详细结果见表 S14。† 两条具有代表性的路线的预测结果见图 6，^{46,47} 其他预测结果见图 S4 - 15。†

在图 6A 所示的路线中，许多此类预测可通过用文献中的条件替换类似的试剂来实现，通常只需一步即可。例如，步骤 2 中的实际条件可通过将模型预测的 LiAlH_4 替换为 $\text{BH}_3 \cdot \text{SMe}_2$ 来获得，而步骤 5 中的条件则可通过将 LiHMDS 替换为 KHMDS 来得到。此外，在该路线的步骤 6 中，Reacon 准确预测了文献中报道的相同催化剂、试剂以及三种溶剂中的一种。

在图 6B 所示的合成路线中，Reacon 的预测与文献中报道的条件高度吻合。例如，步骤 4 和 5 中的报道条件都可以通过将模型建议的四氢呋喃（THF）替换为 2-甲基四氢呋喃（2-MeTHF）来实现。即使预测与实际情况有所偏差，但仍然合乎情理。比如，在步骤 3 的脱硫缩醛化过程中，其中



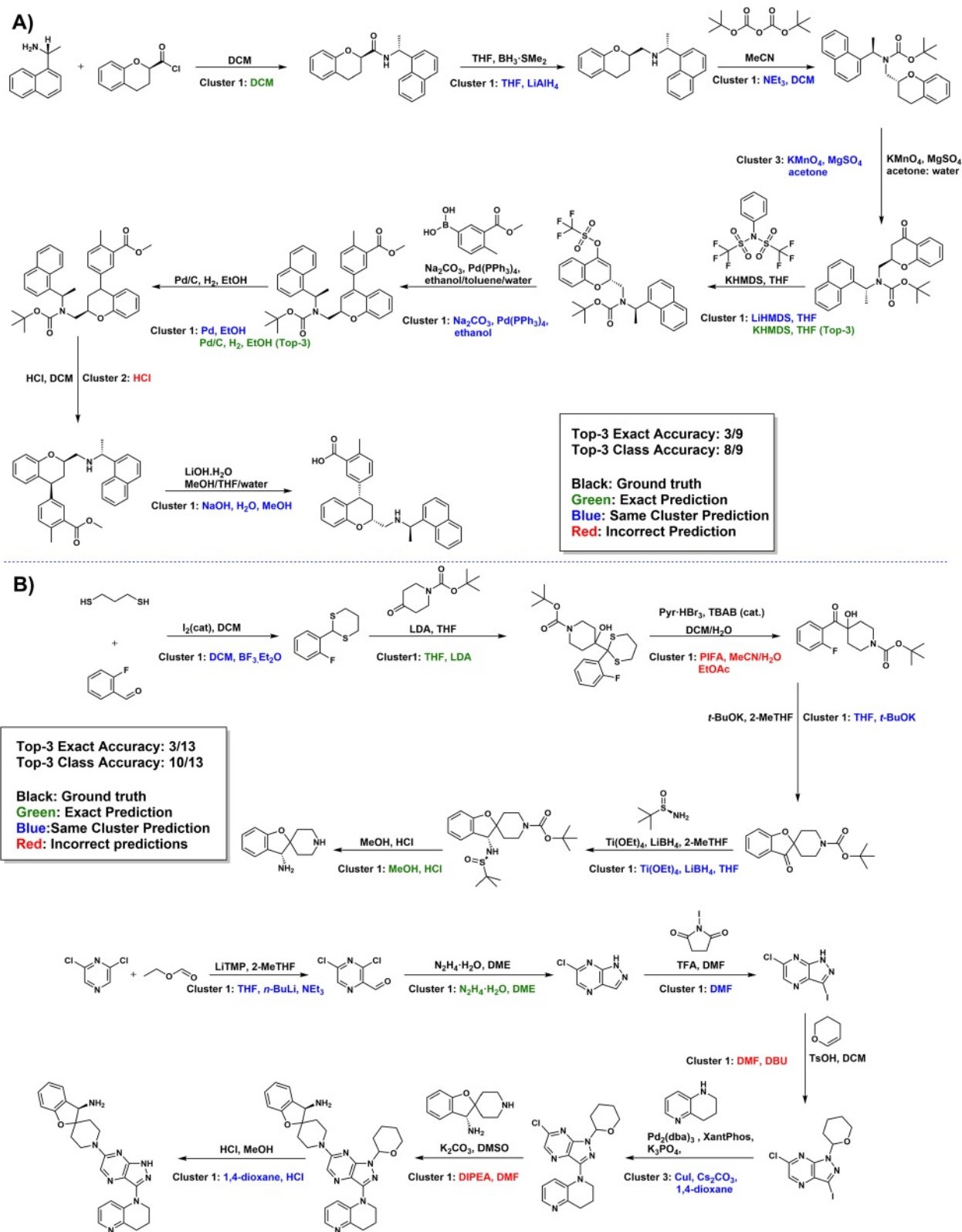


图 6 (A) LNP1892 的合成路线, 包括实际的和预测的反应条件。(B) GDC-1971 的合成路线, 包括实际的和预测条件。真实条件以黑色标注。一致的预测结果以绿色标注, 条件类型一致但与真实条件不符的预测结果以蓝色标注, 不一致的预测结果以红色标注。默认显示每个簇的 top-1 结果, 除非在括号内指定了排名。



文献中倾向于使用溴化吡啶鎓，而 Reacon 则提出了一个替代方案，使用 PIFA⁵⁷ 作为氧化剂。然而，Reacon 偶尔也会给出不合理的结果，比如在步骤 10 中，它错误地将涉及 3,4-二氢吡喃的反应解释为氨基加成到双键上。这种误解源于我们的条件库中缺乏类似的模板。

鉴于通用反应条件预测模型的主要目标是为合成化学家提供条件优化的合理起点，这一结果具有足够的信息量和参考价值。

结论

我们开发了一种新颖的框架 Reacon，它包含一个 GNN 模型，通过整合模板和聚类算法来预测合理的反应条件。Reacon 在测试数据集中召回真实条件的 top-1 准确率为 44.52%，预测相应聚类的准确率为 65.68%。尽管数据集的不准确导致了一些错误，但 Reacon 在提供更合理的反应条件方面表现出色。此外，该模型在预测实际合成路线的反应条件方面也表现出令人满意的性能。它在文献中报告的条件聚类中成功地将其识别在前 3 个聚类中，准确率为 85.00%，展示了其帮助化学家筛选反应条件的能力。总体而言，我们的工作提供了一种可靠的条件预测工具，有助于化学家为新反应选择条件以及计算机辅助合成规划。

尽管取得了令人鼓舞的结果，但当前的方法仍可在以下方面进一步改进。首先，我们框架的出色表现很大程度上依赖于模板条件库，这虽然提升了模型性能，但也限制了其可扩展性。这导致我们的模型难以对训练集中未出现的反应条件做出有效预测。因此，进一步的研究应侧重于提升预测器对新条件的预测能力。其次，温度和反应时间也是反应条件的重要组成部分，未来的工作可以将它们纳入考虑。最后，由于不准确的反应条件会对模型训练和反应条件的聚类产生负面影响，因此需要收集高质量的数据集，尤其是来自高通量实验的数据，以进一步改进模型。

代码可用性

完整代码和训练好的模型可在以下网址获取：<https://github.com/wzhstat/Reaction-Condition-Selector>。

数据可用性

用于训练的数据集可在以下网址获取：<https://www.dropbox.com/scl/fo/v1rhys2wvead9dz3x4?hrlkey=nqtst7azldcyr3ixnoigmcv3v&dl=0>。

作者贡献

Z. W. 和 K. L. 设计了研究方案，进行了实验，分析了数据并撰写了论文。J. P. 和 L. L. 监督了项目并修改了论文。所有作者都阅读并批准了最终的论文。

利益冲突

没有需要声明的利益冲突。


致谢

本研究部分得到了中国国家重点研发计划（项目编号：2023YFF1205103）、国家自然科学基金（项目编号：22033001和 T2321001）以及中国医学科学院（项目编号：2021-I2M-5-014）的支持。

参考文献

- 1 C. W. Coley, W. H. Green and K. F. Jensen, Machine Learning in Computer-Aided Synthesis Planning, *Acc. Chem. Res.*, 2018, 51(5), 1281–1289.
- 2 T. J. Struble, J. C. Alvarez, S. P. Brown, M. Chytil, J. Cisar, R. L. DesJarlais, O. Engkvist, S. A. Frank, D. R. Greve, D. J. Griffin, X. Hou, J. W. Johannes, C. Kreatsoulas, B. Lahue, M. Mathea, G. Mogk, C. A. Nicolaou, A. D. Palmer, D. J. Price, R. I. Robinson, S. Salentin, L. Xing, T. Jaakkola, W. H. Green, R. Barzilay, C. W. Coley and K. F. Jensen, Current and Future Roles of Artificial Intelligence in Medicinal Chemistry Synthesis, *J. Med. Chem.*, 2020, 63(16), 8667–8682.
- 3 S. Szymkuc, E. P. Gajewska, T. Klucznik, K. Molga, P. Dittwald, M. Startek, M. Bajczyk and B. A. Grzybowski, Computer-Assisted Synthetic Planning: The End of the Beginning, *Angew Chem. Int. Ed. Engl.*, 2016, 55(20), 5904–5937.
- 4 J. Dong, M. Zhao, Y. Liu, Y. Su and X. Zeng, Deep learning in retrosynthesis planning: datasets, models and tools, *Briefings Bioinf.*, 2022, 23(1), bbab391.
- 5 T. Klucznik, B. Mikulak-Klucznik, M. P. McCormack, H. Lima, S. Szymkuć, M. Bhowmick, K. Molga, Y. Zhou, L. Rickershauser, E. P. Gajewska, A. Touthkine, P. Dittwald, M. P. Startek, G. J. Kirkovits, R. Roszak, A. Adamski, B. Sieredzinska, M. Mrksich, S. L. J. Trice and B. A. Grzybowski, Efficient Syntheses of Diverse, Medicinally Relevant Targets Planned by Computer and Executed in the Laboratory, *Chem*, 2018, 4(3), 522–532.
- 6 M. H. S. Segler, M. Preuss and M. P. Waller, Planning chemical syntheses with deep neural networks and symbolic AI, *Nature*, 2018, 555(7698), 604–610.
- 7 C. W. Coley, D. A. Thomas, J. A. M. Lummiss, J. N. Jaworski, C. P. Breen, V. Schultz, T. Hart, J. S. Fishman, L. Rogers, H. Gao, R. W. Hicklin, P. P. Plehiers, J. Byington, J. S. Piotti, W. H. Green, A. J. Hart, T. F. Jamison and K. F. Jensen, A robotic platform for flow synthesis of



- organic compounds informed by AI planning, *Science*, 2019, 365 (6453), eaax1566.
- 8 P. Schwaller, R. Petraglia, V. Zullo, V. H. Nair, R. A. Hauselmann, R. Pisoni, C. Bekas, A. Iuliano and T. Laino, Predicting retrosynthetic pathways using transformer-based models and a hyper-graph exploration strategy, *Chem. Sci.*, 2020, 11(12), 3316–3325.
 - 9 K. Lin, Y. Xu, J. Pei and L. Lai, Automatic retrosynthetic route planning using template-free models, *Chem. Sci.*, 2020, 11(12), 3355–3364.
 - 10 B. Mikulak-Klucznik, P. Golebiowska, A. A. Bayly, O. Popik, T. Klucznik, S. Szymkuc, E. P. Gajewska, P. Dittwald, O. Staszewska-Krajewska, W. Beker, T. Badowski, K. A. Scheidt, K. Molga, J. Mlynarski, M. Mrksich and B. A. Grzybowski, Computational planning of the synthesis of complex natural products, *Nature*, 2020, 588(7836), 83–88.
 - 11 Y. Lin, R. Zhang, D. Wang and T. Cernak, Computer-aided key step generation in alkaloid total synthesis, *Science*, 2023, 379 (6631), 453–457.
 - 12 Z. Tu, T. Stuyver and C. W. Coley, Predictive chemistry: machine learning for reaction deployment, reaction development, and reaction discovery, *Chem. Sci.*, 2023, 14(2), 226–244.
 - 13 C. W. Coley, R. Barzilay, T. S. Jaakkola, W. H. Green and K. F. Jensen, Prediction of Organic Reaction Outcomes Using Machine Learning, *ACS Cent. Sci.*, 2017, 3(5), 434–443.
 - 14 F. Jaume-Santero, A. Bornet, A. Valery, N. Naderi, D. Vicente Alvarez, D. Proios, A. Yazdani, C. Bourmez, T. Fessard and D. Teodoro, Transformer Performance for Chemical Reactions: Analysis of Different Predictive and Evaluation Scenarios, *J. Chem. Inf. Model.*, 2023, 63(7), 1914–1924.
 - 15 B. Zhang, X. Zhang, W. Du, Z. Song, G. Zhang, G. Zhang, Y. Wang, X. Chen, J. Jiang and Y. Luo, Chemistry-informed molecular graph as reaction descriptor for machine-learned retrosynthesis planning, *Proc. Natl. Acad. Sci. U. S. A.*, 2022, 119(41), e2212711119.
 - 16 T. Gaich and P. S. Baran, Aiming for the ideal synthesis, *J. Org. Chem.*, 2010, 75(14), 4657–4673.
 - 17 T. Newhouse, P. S. Baran and R. W. Hoffmann, The economies of synthesis, *Chem. Soc. Rev.*, 2009, 38(11), 3010–3021.
 - 18 H. Struebing, Z. Ganase, P. G. Karamertzanis, E. Sioukrou, P. Haycock, P. M. Piccione, A. Armstrong, A. Galindo and C. S. Adjiman, Computer-aided molecular design of solvents for accelerated reaction kinetics, *Nat. Chem.*, 2013, 5(11), 952–957.
 - 19 G. Marcou, J. Aires de Sousa, D. A. R. S. Latino, A. de Luca, D. Horvath, V. Rietsch and A. Varnek, Expert System for Predicting Reaction Conditions: The Michael Reaction Case, *J. Chem. Inf. Model.*, 2015, 55(2), 239–250.
 - 20 M. R. Maser, A. Y. Cui, S. Ryou, T. J. DeLano, Y. Yue and S. E. Reisman, Multilabel Classification Models for the Prediction of Cross-Coupling Reaction Conditions, *J. Chem. Inf. Model.*, 2021, 61(1), 156–166.
 - 21 V. A. Afonina, D. A. Mazitov, A. Nurmukhametova, M. D. Shevelev, D. A. Khasanova, R. I. Nugmanov, V. A. Buriylov, T. I. Madzhidov and A. Varnek, Prediction of Optimal Conditions of Hydrogenation Reaction Using the Likelihood Ranking Approach, *Int. J. Mol. Sci.*, 2022, 23(1), 248.
 - 22 Y. Kwon, S. Kim, Y.-S. Choi and S. Kang, Generative Modeling to Predict Multiple Suitable Conditions for Chemical Reactions, *J. Chem. Inf. Model.*, 2022, 62(23), 5952–5960.
 - 23 N. H. Angello, V. Rathore, W. Beker, A. Wołos, E. R. Jira, R. Roszak, T. C. Wu, C. M. Schroeder, A. Aspuru-Guzik, B. A. Grzybowski and M. D. Burke, Closed-loop optimization of general reaction conditions for heteroaryl Suzuki-Miyaura coupling, *Science*, 2022, 378(6618), 399–405.
 - 24 Y. Kwon, D. Lee, J. W. Kim, Y. S. Choi and S. Kim, Exploring Optimal Reaction Conditions Guided by Graph Neural Networks and Bayesian Optimization, *ACS Omega*, 2022, 7(49), 44939–44950.
 - 25 J. A. G. Torres, S. H. Lau, P. Anchuri, J. M. Stevens, J. E. Tabora, J. Li, A. Borovika, R. P. Adams and A. G. Doyle, A Multi-Objective Active Learning Platform and Web App for Reaction Optimization, *J. Am. Chem. Soc.*, 2022, 144(43), 19999–20007.
 - 26 K. Atz, D. F. Nippa, A. T. Müller, V. Jost, A. Anelli, M. Reutlinger, C. Kramer, R. E. Martin, U. Grether, G. Schneider and G. Wuitschik, Geometric deep learning-guided Suzuki reaction conditions assessment for applications in medicinal chemistry, *RSC Med. Chem.*, 2024, 15(7), 2310–2321.
 - 27 H. Gao, T. J. Struble, C. W. Coley, Y. Wang, W. H. Green and K. F. Jensen, Using Machine Learning To Predict Suitable Conditions for Organic Reactions, *ACS Cent. Sci.*, 2018, 4(11), 1465–1476.
 - 28 D. Kreutter and J.-L. Reymond, Multistep retrosynthesis combining a disconnection aware triple transformer loop with a route penalty score guided tree search, *Chem. Sci.*, 2023, 14(36), 9959–9969.
 - 29 M. Andronov, V. Voinarovska, N. Andronova, M. Wand, D. A. Clevert and J. Schmidhuber, Reagent prediction with a molecular transformer improves reaction data quality, *Chem. Sci.*, 2023, 14(12), 3235–3246.
 - 30 X. Wang, C.-Y. Hsieh, X. Yin, J. Wang, Y. Li, Y. Deng, D. Jiang, Z. Wu, H. Du, H. Chen, Y. Li, H. Liu, Y. Wang, P. Luo, T. Hou and X. Yao, Generic Interpretable Reaction Condition Predictions with Open Reaction Condition Datasets and Unsupervised Learning of Reaction Center, *Research*, 2023, 6, 0231.
 - 31 Y. Qian, Z. Li, Z. Tu, C. Coley and R. Barzilay, Predictive Chemistry Augmented with Text Retrieval, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Singapore, 2023, pp. 12731–12745.
 - 32 E. Heid and W. H. Green, Machine Learning of Reaction Properties via Learned Representations of the Condensed Graph of Reaction, *J. Chem. Inf. Model.*, 2022, 62(9), 2101–2110.
 - 33 D. Lowe, Chemical reactions from US patents (1976-Sep2016),  gshare, DOI: [10.6084/m9.figshare.5104873.v1](https://doi.org/10.6084/m9.figshare.5104873.v1).

V. A. Buriylov, T. I. Madzhidov and A. Varnek, Prediction of



- 34 D. Weininger, SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules, *J. Chem. Inf. Comput. Sci.*, 1988, 28(1), 31–36.
- 35 RDKit: Open-Source Cheminformatics, <https://www.rdkit.org>.
- 36 C. W. Coley, W. H. Green and K. F. Jensen, RDChiral: An RDKit Wrapper for Handling Stereochemistry in Retrosynthetic Template Extraction and Application, *J. Chem. Inf. Model.*, 2019, 59(6), 2529–2537.
- 37 E. Heid, K. P. Greenman, Y. Chung, S. C. Li, D. E. Graff, F. H. Vermeire, H. Wu, W. H. Green and C. J. McGill, Chemprop: A Machine Learning Package for Chemical Property Prediction, *J. Chem. Inf. Model.*, 2024, 64(1), 9–17.
- 38 P. Velićković, G. Cucurull, A. Casanova, A. Romero, P. Lio[†] and Y. Bengio, Graph Attention Networks, arXiv, 2017, preprint, arXiv:1710.10903, DOI: 10.48550/arXiv.1710.10903.
- 39 W. Beker, R. Roszak, A. Wolos, N. H. Angello, V. Rathore, M. D. Burke and B. A. Grzybowski, Machine Learning May Sometimes Simply Capture Literature Popularity Trends: A Case Study of Heterocyclic Suzuki-Miyaura Coupling, *J. Am. Chem. Soc.*, 2022, 144(11), 4819–4827.
- 40 C. W. Coley, L. Rogers, W. H. Green and K. F. Jensen, Computer-Assisted Retrosynthesis Based on Molecular Similarity, *ACS Cent. Sci.*, 2017, 3(12), 1237–1245.
- 41 L. Wolff, Chemischen Institut der Universität at Jena: Methode zum Ersatz des Sauerstoffatoms der Ketone und Aldehyde durch Wasserstoff, *Justus Liebigs Ann. Chem.*, 1912, 394(1), 86–108.
- 42 M. Huang, Reduction of Steroid Ketones and other Carbonyl Compounds by Modified Wolff-Kishner Method, *J. Am. Chem. Soc.*, 1949, 71(10), 3301–3303.
- 43 D. B. Dess and J. C. Martin, Readily accessible 12-I-5 oxidant for the conversion of primary and secondary alcohols to aldehydes and ketones, *J. Org. Chem.*, 1983, 48, 4155–4156.
- 44 K. Omura and D. Swern, Oxidation of alcohols by “activated” dimethyl sulfoxide. a preparative, steric and mechanistic study, *Tetrahedron Lett.*, 1974, 15(16), 1465–1560.
- 45 D. B. Freeman, T. D. Hopkins, P. J. Mikochik, J. P. Vacca, H. Gao, A. Naylor-Olsen, S. Rudra, H. Li, M. S. Pop, R. A. Villagomez, C. Lee, H. Li, M. Zhou, D. C. Saffran, N. Rioux, T. R. Hood, M. A. L. Day, M. R. McKeown, C. Y. Lin, N. Bischofberger and B. W. Trotter, Discovery of KB-0742, a Potent, Selective, Orally Bioavailable Small Molecule Inhibitor of CDK9 for MYC-Dependent Cancers, *J. Med. Chem.*, 2023, 66(23), 15629–15647.
- 46 A. M. Taylor, B. R. Williams, F. Giordanetto, E. H. Kelley, A. Lescarbeau, K. Shortsleeves, Y. Tang, W. P. Walters, A. Arrazate, C. Bowman, E. Brophy, E. W. Chan, G. Deshmukh, J. B. Greisman, T. L. Hunsaker, D. R. Kipp, P. Saenz Lopez-Larrocha, D. Maddalo, I. J. Martin, P. Maragakis, M. Merchant, M. Murcko, H. Nisonoff, V. Nguyen, V. Nguyen, O. Orozco, C. Owen, L. Pierce, M. Schmidt, D. E. Shaw, S. Smith, E. Therrien, J. C. Tran, J. Waters, N. J. Waters, J. Wilbur and L. Willmore, Identification of GDC-1971 (RLY-1971), a SHP2 Inhibitor Designed for the Treatment of Solid Tumors, *J. Med. Chem.*, 2023, 66(19), 13384–13399.
- 47 M. R. Shukla, G. Sadasivam, A. Sarde, M. Sayyed, V. Pachpute, R. Phadtare, N. Walke, V. D. Chaudhari, R. Loria, T. Khan, G. Gote, C. Pawar, M. Tryambake, N. Mahajan, A. Gandhe, S. Sabde, S. Pawar, V. Patil, D. Modi, M. Mehta, P. Nigade, V. Modak, R. Ghodke, L. Narasimham, M. Bhonde, J. Gundu, R. Goel, C. Shah, S. Kulkarni, S. Sharma, D. Bakhle, R. K. Kamboj and V. P. Pale, Discovery of LNP1892: A Precision Calcimimetic for the Treatment of Secondary Hyperparathyroidism, *J. Med. Chem.*, 2023, 66(14), 9418–9444.
- 48 C. Mo, X. Xu, P. Zhang, Y. Peng, X. Zhao, S. Chen, F. Guo, Y. Xiong, X. J. Chu and X. Xu, Discovery of HPG1860, a Structurally Novel Nonbile Acid FXR Agonist Currently in Clinical Development for the Treatment of Nonalcoholic Steatohepatitis, *J. Med. Chem.*, 2023, 66(14), 9363–9375.
- 49 Y. Wu, J. Xi, Y. Li, Z. Li, Y. Zhang, J. Wang and G. H. Fan, Discovery of a Potent and Selective CCR8 Small Molecular Antagonist IPG7236 for the Treatment of Cancer, *J. Med. Chem.*, 2023, 66(7), 4548–4564.
- 50 M. R. Garnsey, A. C. Smith, J. Polivkova, A. L. Arons, G. Bai, C. Blakemore, M. Boehm, L. M. Buzon, S. N. Campion, M. Cerny, S. C. Chang, K. Coffman, K. A. Farley, K. R. Fonseca, K. K. Ford, J. Garren, J. X. Kong, M. R. M. Koos, D. W. Kung, Y. Lian, M. M. Li, Q. Li, L. A. Martinez-Alsina, R. O'Connor, K. Ogilvie, K. Omoto, B. Raymer, M. R. Reese, T. Ryder, L. Samp, K. A. Stevens, D. W. Widlicka, Q. Yang, K. Zhu, J. P. Fortin and M. F. Sammons, Discovery of the Potent and Selective MC4R Antagonist PF-07258669 for the Potential Treatment of Appetite Loss, *J. Med. Chem.*, 2023, 66(5), 3195–3211.
- 51 M. E. Layton, J. C. Kern, T. J. Hartingh, W. D. Shipe, I. Raheem, M. Kandebo, R. P. Hayes, S. Huszar, D. Eddins, B. Ma, J. Fuerst, G. K. Wollenberg, J. Li, J. Fritzen, G. B. McGaughey, J. M. Uslaner, S. M. Smith, P. J. Coleman and C. D. Cox, Discovery of MK-8189, a Highly Potent and Selective PDE10A Inhibitor for the Treatment of Schizophrenia, *J. Med. Chem.*, 2023, 66(2), 1157–1171.
- 52 B. Chen, J. Wu, Z. Yan, H. Wu, H. Gao, Y. Liu, J. Zhao, J. Wang, J. Yang, Y. Zhang, J. Pan, Y. Ling, H. Wen and Z. Huang, 1,3-Substituted beta-Carboline Derivatives as Potent Chemotherapy for the Treatment of Cystic Echinococcosis, *J. Med. Chem.*, 2023, 66(24), 16680–16693.
- 53 L. Zhang, Y. Li, C. Tian, R. Yang, Y. Wang, H. Xu, Q. Zhu, S. Chen, L. Li and S. Yang, From Hit to Lead: Structure-Based Optimization of Novel Selective Inhibitors of Receptor-Interacting Protein Kinase 1 (RIPK1) for the Treatment of Inflammatory Diseases, *J. Med. Chem.*, 2024, 67(1), 754–773.
- 54 J. Szychowski, R. Papp, E. Dietrich, B. Liu, F. Vallee, M. E. Leclaire, J. Fourtounis, G. Martino, A. L. Perryman, V. Pau, S. Y. Yin, P. Mader, A. Roulston, J. F. Truchon, C. G. Marshall, M. Diallo, N. M. Duffy, R. Stocco, C. Godbout, A. Bonneau-Fortin, R. Kryczka, V. Bhaskaran, D. Mao, S. Orlicky, P. Beaulieu, P. Turcotte, I. Kurinov, F. Sicheri, Y. Mamane, M. Gallant and W. C. Black, Discovery of an Orally Bioavailable and Selective PKMYT1



- Inhibitor, RP-6306, *J. Med. Chem.*, 2022, 65(15), 10251–10284.
- 55 M. D. Hill, M. J. Blanco, F. G. Salituro, Z. Bai, J. T. Beckley, M. A. Ackley, J. Dai, J. J. Doherty, B. L. Harrison, E. C. Hoffmann, T. M. Kazdoba, D. Lanzetta, M. Lewis, M. C. Quirk and A. J. Robichaud, SAGE-718: A First-in-Class N-Methyl-d-Aspartate Receptor Positive Allosteric Modulator for the Potential Treatment of Cognitive Impairment, *J. Med. Chem.*, 2022, 65(13), 9063–9075.
- 56 T. Inghardt, T. Antonsson, C. Ericsson, D. Hovdal, P. Johannesson, C. Johansson, U. Jurva, J. Kajanus, B. Kull, E. Michaelsson, A. Pettersen, T. Sjogren, H. Sorensen, K. Westerlund and E. L. Lindstedt, Discovery of AZD4831, a Mechanism-Based Irreversible Inhibitor of Myeloperoxidase, As a Potential Treatment for Heart Failure with Preserved Ejection Fraction, *J. Med. Chem.*, 2022, 65(17), 11485–11496.
- 57 G. Stork and K. Zhao, A simple method of dethioacetalization, *Tetrahedron Lett.*, 1989, 30(3), 287–290.

