将现代智能算法应用于

逆合成预测

Jianhan Liao[1] , Xiaoxin Shi[2] , Tong Zhu *[1,3]

[1]上海分子治疗与新药开发工程技术研究中心，华东师范大学化学与分子工程学院，上海，200062

[2]中国上海市闵行区东川路 800 号上海交通大学化学与化学工程学院，200240

[3]上海纽约大学-华东师范大学计算化学中心，中国上海，200062 tzhu@lps.ecnu.edu.cn

摘要

近年来，随着计算机科学的飞速发展，各种现代智能算法层出不穷。基于多头注意机制的变形器是本世纪最受青睐的人工智能模型之一。这些算法的引入使得逆合成预测取得了巨大进步。与传统的逆合成预测模型不同，基于智能算法的逆合成预测可以自动从化学反应数据集中提取化学知识，预测逆合成路线。在这篇综述中，我们全面介绍了基于现代智能算法，特别是人工智能算法的逆合成预测。在介绍了相关的深度学习模型后，介绍了现有的化学反应数据集和分子表征。随后，讨论了近年来人工智能辅助逆合成预测模型的现状，包括基于模板的模型、无模板模型和基于半模板的模型。此外，我们还比较了不同分类的逆合成预测模型。最后，我们总结了这些现有方法所面临的挑战和局限性，为未来的研究指明了方向。

关键词： 人工智能、逆合成预测、机器学习、深度学习

1. 导言

有机合成是化学中不可或缺的一个分支，因其需要创造力、灵感和审美判断力[1,2]，常被描述为一门艺术。它是一门重要的技术，在药物设计和合成生物学中有着广泛的应用[3-6]。逆合成分析是有机合成设计的常用方法[7]。它是从目标化合物逆向推导合成路线的过程。其核心思想是将目标化合物分解成多个较简单的化合物或起始原料，然后通过合成这些化合物或起始原料来获得目标化合物。然而，随着目标分子的多样性和复杂性不断增加，有机合成途径的设计难度也成倍增加。为了提高生产率和结果的可重复性，人们越来越期待有机逆合成能够实现自动化[8-10]。因此，计算机辅助合成规划（CASP）应运而生。最初的

这一领域的尝试可以追溯到科里在基于规则的合成预测系统方面的开创性工作，即自动合成分析逻辑与启发式程序（LHASA）[11]。它设计了一系列反应，递归地将目标化合物分解成更简单的构筑模块，直到达到可在市场上买到的起始分子。然而，由于计算能力和数据可用性的限制，早期基于规则的模型并未取得令人满意的结果。近来，随着计算机科学的空前发展[12]，用于各种任务的智能算法，如波束搜索算法、蒙特卡洛树搜索算法、遗传算法和神经网络算法等迅速涌现。此外，越来越多由大数据驱动的人工智能模型被提出[13,14]。由于人工智能在各种任务中取得的突出成就，人工智能在化学和药物发现中的应用再次引起人们的关注[15,16]。

对于化学家来说，CASP 是一项艰巨的挑战，尤其是在逆合成预测领域。这是因为，与正向反应预测任务不同，逆合成反应预测任务提供的输入信息有限，但却可能产生多种输出可能性。

近年来，许多研究人员针对逆合成预测任务提出了各种类型的模型。单步逆合成预测模型可以自动断开给定产物以获得候选反应物。对于无法从市场上获得的候选反应物，则采用递归扩展策略，直到沿路径的所有反应物都能从市场上获得或达到预定的最大扩展步骤。一旦完成了精确的递归单步逆合成预测，多步逆合成预测的重点就是规划最优的反应序列，使合成步骤的数量、起始分子的成本、产生的废料等最小化。因此，单步逆合成预测模型的性能是逆合成任务的基础。这些模型大致可分为三类：

第一类是基于模板的模型，它整合了领域知识和基于先前化学知识的形式规则，如基于模板的算法。反应模板是一组规则，决定了反应物如何通过键解转化为生成物。模板和规则这两个术语经常互换使用。这些模型通常具有较高的可解释性和准确性，但在大多数情况下，它们很难在其知识库之外做出准确的预测。

第二类是无模板模型，通常不包含化学知识，被视为黑盒模型，如深度神经网络。这些黑箱模型通常表现出较低的可解释性和较高的计算复杂性。它们容易生成违反化学知识的解决方案。尽管如此，它们在发现不受现有知识库限制的新反应途径方面显示出巨大潜力。随着计算数据处理能力的指数级增长，纯数据驱动模型的性能有了大幅提高。

第三类是基于半模板的模型，包括两个步骤：(1) 它们首先确定反应中心，并利用反应中心将产物转化为合成物（中间分子）；然后 (2) 它们将合成物完整地转化为

反应物。

在本综述中，我们将重点关注当代的逆合成策略。我们概述并评估了主要在过去三年中开发的逆合成预测模型。在接下来的章节中，我们首先介绍了逆合成预测中的相关人工智能模型。然后，比较了逆合成预测任务中常用的数据源和分子表征。接着，我们深入探讨了现代智能算法在基于模板模型、无模板模型和半模板模型中的应用。最后，我们对这一领域进行了展望并探讨了潜在的挑战。

2. 相关深度学习算法

人工智能算法是为了模仿人类智能而开发的。这些算法可以从数据集中提取潜在规则，并在获得新数据时利用这些规则进行预测。深度学习（DL）作为人工智能的一个快速发展的分支，得益于计算能力和现代算法的进步，在各种任务中表现出无与伦比的性能。一般来说，深度学习模型可分为三类：监督学习、无监督学习和强化学习（RL）。

在监督学习方法中，模型是在标注样本的数据集上训练出来的。模型学习如何将输入特征映射到输出标签。有两种主要的监督学习模型。分类模型学习预测离散输出标签。回归模型学习预测连续的输出值。在无监督学习方法中，模型是在未标记样本的数据集上进行训练的。模型学会识别数据中的模式和关系，而不需要明确地告诉它要寻找什么。在强化学习方法中，代理通过试错来学习在环境中的行为。如果采取的行动导致了预期的结果，代理就会得到奖励；如果采取的行动导致了不预期的结果，代理就会受到惩罚。代理的目标是学习一种能在一段时间内使其预期回报最大化的规则。大多数逆合成预测模型都使用监督学习策略，其框架**如图 1** 所示。

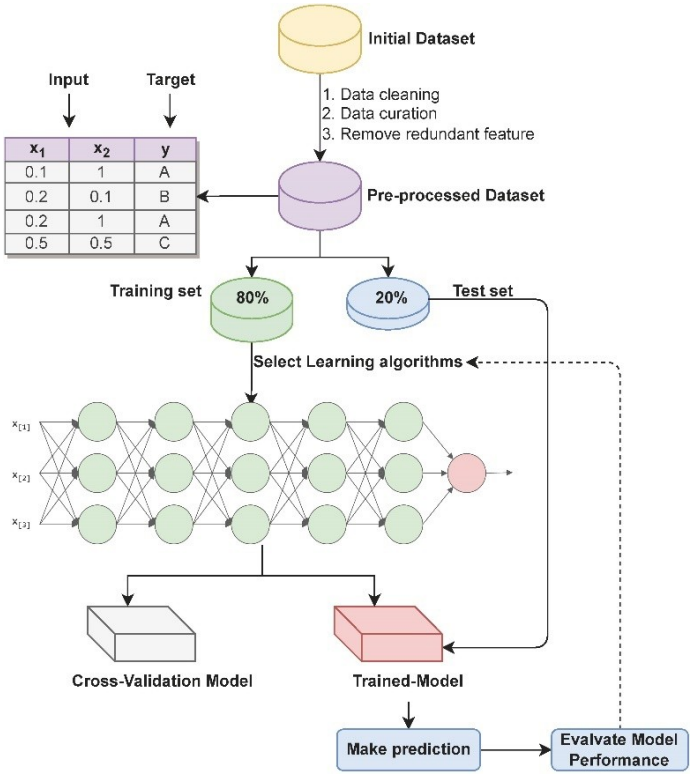


图 1 监督学习法的过程

逆合成预测中常见的 DL 算法包括 Seq2Seq 模型、图神经网络、强化学习和搜索算法。

2.1 序列生成模型

由于分子可以表示为基于 SMILES 的序列，因此逆合成预测任务可以转化为序列到序列任务。在自然语言处理（NLP）领域得到广泛应用的序列到序列模型（Sequence-to-Sequence Model，Seq2Seq 模型）自然成为化学序列建模的有效工具。Seq2Seq 模型可以在输入序列长度与输出序列长度不同的情况下，从化学产物生成化学反应物序列。在这篇综述中，我们重点讨论了基于递归神经网络的 Seq2Seq 模型和基于注意力的 Seq2Seq 模型。

为解决序列生成问题（如机器翻译），人们首次提出了用于编码和解码的递归神经网络（RNN）[17,18]。RNN 与前馈神经网络的区别在于，它使用隐藏状态来记录之前的所有信息。RNN 的编码器将输入句子编码成固定长度的向量，解码器按顺序生成目标词。RNN 的框架如图 1 所示。然而，RNN 模型无法捕捉长距离依赖关系，也无法并行计算。双向长短时记忆（biLSTM）是 RNN 的一种变体，它能通过门控机制有选择性地保留长距离依赖关系。注意机制是一种计算资源分配策略，可以将有限的计算资源集中用于重要信息。当与注意力机制相结合时，这种基于 biLSTM 的框架

能使隐藏状态包含全局信息，并解决不可并行计算的问题[19]。为了模拟全局注意力，每个解码器都引入了多步注意力机制

层。变压器模型最初由 Vaswani 等人提出[21]，其特点是编码器和解码器完全依赖于多头自注意机制，从而能够有效捕捉序列内的长距离相关性。近年来，基于 Transformer 的模型已成为纯数据驱动逆合成预测领域的主导力量，这主要归功于其卓越的性能。

注意力机制是深度学习中使用的一种技术，用于为以下对象分配权重输入数据的不同部分，权重越高表示越重要。自我注意力机制是一种特殊的注意力机制，它将注意力应用于同一序列中的不同位置。这样，模型就能捕捉到序列中任意两个位置之间的关系，这对于理解序列的结构和意义非常重要。自我关注机制是通过查询向量、关键向量和值向量来实现的，这些向量用于计算关注权重和输出。查询向量用于计算不同关键向量之间的相似度，所得权重用于对相应的值向量进行加权，从而得到输出结果。

更具体地说，对于列向量集 $H=[h_1,...,h_T]\in R^{D_h\times T}$，自注意机制可以被概念化为一个在线性投影空间中建立不同向量 曝$_i$ 之间相互作用的过程。自我注意力机制的编码公式如下：

$$自\ att(Q,增长,V) = Vsoftmax\ 3\frac{^{\top}Q}{4\overline{D_k}}) \tag{1}$$

$$Q = 徼_q H, Q = 徼_k H, V = 徼 H_v \tag{2}$$

这里，D_k表示输入矩阵Q (查询) 中列向量的维数，而 D_v 表示矩阵 V（值）中列向量的维数。$W_q\in R^{D_k\times D_h}$,$W_k\in R^{D_k\times D_h}$,$W_v\in R^{D_v\times D_h}$ 是三个投影矩阵。

使用多头自我注意力模型可以在多个不同的投影空间中进一步捕捉不同的交互信息。当自我注意力模型应用于M个此类投影空间时，它可以用数学方法表示如下：

$$MultiHead\ (H) = 祆_o\ [head_1\ ;\cdots;\ head\ ]。_M \tag{3}$$

$$head_m = selfatt\ (Q_m, V_m, V_m) \tag{4}$$

$$\forall m\in\{1,\cdots,M\}, Q_m = 徼^m_q H, 祇 = 徼^m_k H, V = 徼^m_v H \tag{5}$$

$W_o\in R^{D_h\times Md_v}$ 是输出投影矩阵，$W_q^{\ m}\in R^{D_k\times D_h}$,$W_k^{\ m}\in R^{D_k\times D_h}$ 和$W_v^{\ m}\in R^{D_v\times D_h}$ 是投影矩阵，$m\in\{1,...,M\}$。

图 2 显示了 Transformer 模型的网络结构，可分为两个部分：编码器和解码器。编码器由多层多头注意力模块组成。解码器通过自回归方式生成目标序列，它由掩码自注意模块、解码器到编码器注意模块和前馈神经网络组成。

除了基于 RNN 的模型和基于注意机制的模型，Gehring 等人还提出了一种序列建模框架[20]，即卷积序列到序列（ConvS2S）模型。它的编码器和解码器由多层卷积神经网络组成，在某些情况下比 RNN 更有效。

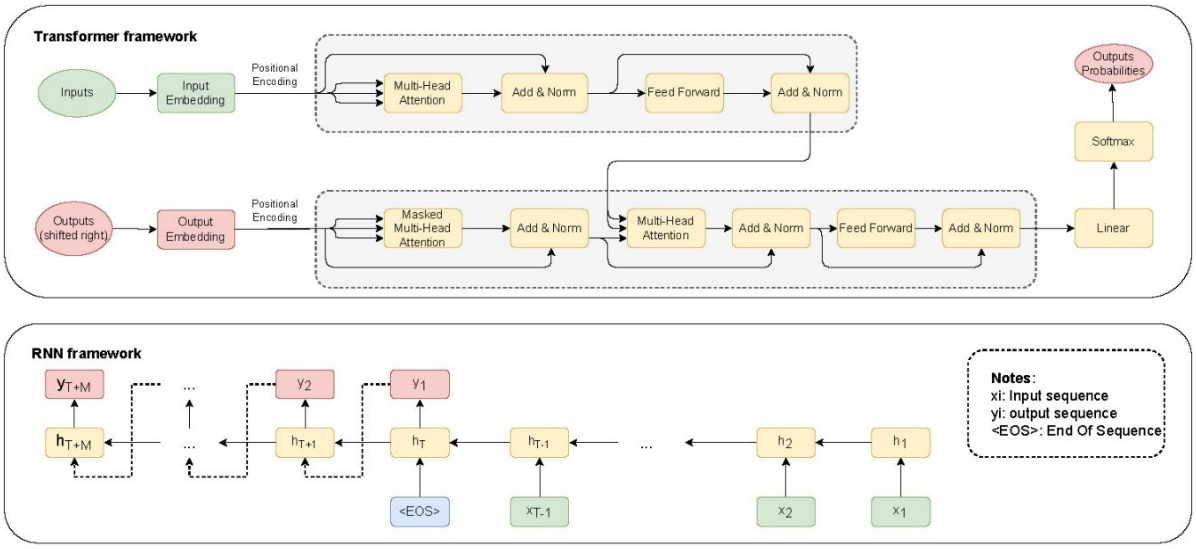


图 2 变压器框架和 RNN 框架。

2.2 图神经网络

分子不仅可以通过将其编码为序列来表示，还可以将其编码为无向加权图，这是一种来自图论的数据结构。它由一组顶点、一组边和一组全局信息组成，每条边都有一个权重，顶点之间的连接没有方向性。有关分子图的详细讨论，请参阅第 4.3 节。图上的预测任务一般有三种类型：图级、节点级和边级。一般来说，化学分子的预测属于图级类型，可以用图神经网络（GNN）来解决。

GNN 是一种很有前途的参数高效工具，可用于学习图的结构信息，从而预测反应中的分子转换[21]。GNN 是对图的所有属性进行的可优化变换，它保留了图的对称性（置换不变性）。Sperduti 是将神经网络应用于有向无环图的先驱[22]。这种方法也适用于化学分子的无向图表示。**图 3** 展示了一个使用 "消息传递神经网络 "框架进行二元分类任务的 GNN 示例，它可以很容易地扩展到多类或回归任务。该 GNN 以图形的数字表示作为输入，为所有图形属性（节点、边、全局）学习新的嵌入，而不使用图形的连接性。这种 GNN 在图的每个分量上使用一个单独的多层感知器（MLP），称为 GNN 层。对于每个图的属性向量，都会应用 MLP，并生成一个学习向量。最后，它通过汇集信息（例如，从边缘到节点收集信息）进行预测。

研究人员进一步提出了递归图神经网络（RecGNNs）[23,24]，通过迭代传播邻域信息来更新目标节点的表征。由于卷积神经网络（CNN）在计算机视觉领域取得了巨大成功，研究人员将卷积操作引入了 GNN，并开发了图卷积网络（GCN）[25]。GCN 中的卷积操作是对图的特征进行加权平均，以聚合特征及其邻近特征的信息。然而，聚合运算产生的权重不具有包络不变性。为了克服这一问题，研究人员引

入了

注意机制，并提出了图形注意网络（GATs）[26] 和门控注意网络（GAANs）[27]。在这些工作的基础上，人们进一步开发了图自动编码器（GAEs）[28]、图生成网络（GGNs）[29]和时空图卷积网络（STGCNs）[30]。

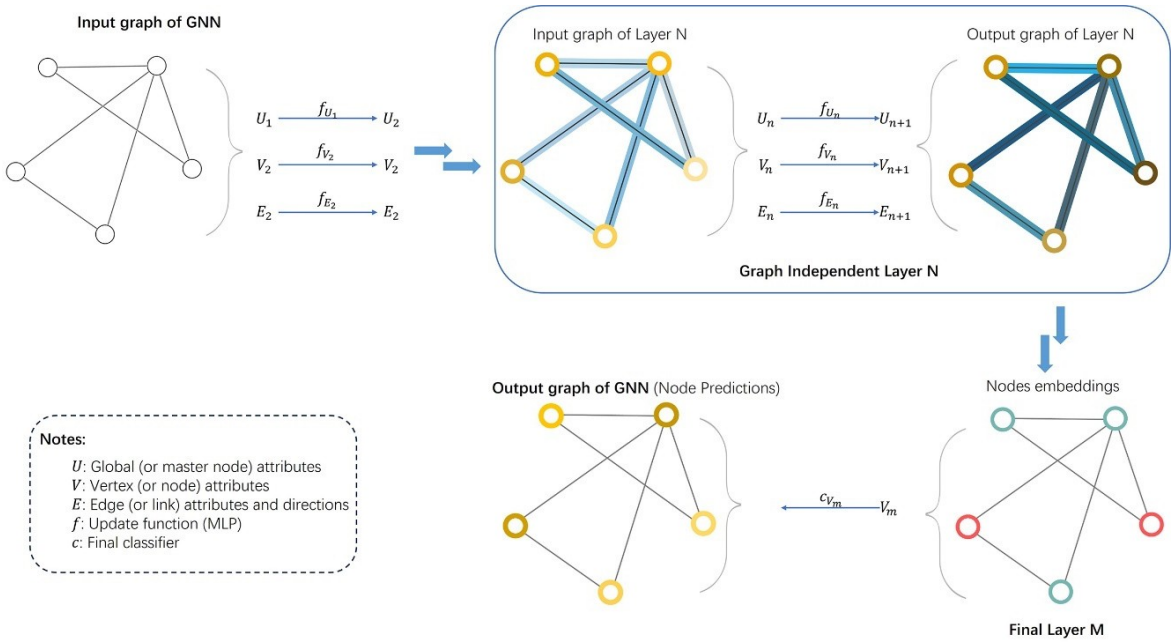


图 3 带有用于分类任务的消息传递神经网络的 GNN 框架。

2.3 强化学习

强化学习（RL）是一种无监督学习方法[31]。它解决的问题是，代理通过与环境互动学习来实现特定目标，如最大化奖励。与深度学习类似，RL 的一个关键挑战是贡献的分配。每个行动都不会收到直接的监督信息，而是取决于来自整个模型的最终监督信号（奖励），通常会有一定的延迟。RL 与监督学习的关键区别在于，RL 不需要 "正确 "的策略作为监督信息；相反，它侧重于提供策略的延迟回报，并调整策略以最大化预期回报。

在 RL 中，存在两个相互作用的实体：代理和环境。代理感知外部环境的状态和回报，进行学习和决策。决策包括根据外部环境的状态采取不同的行动，而学习则是根据环境的回报调整策略。环境包括代理的所有外部元素，其状态会因代理的行动而发生变化，并为代理提供相应的回报。

RL 的基本组成部分包括

(1) 状态s是对环境的描述，可以是离散的，也可以是连续的，形成状态空间S；

(2) 描述代理行为的行动a，也可以是离散或连续的形式，形成行动空间A；

(3) 策略$\pi(a|s)$表示代理如何根据信息决定下一步行动 a。
环境的状态 s；

(4) 状态转换概率$p(s'|s,a)$表示环境的可能性

在代理从当前状态 s 采取行动 a 之后过渡到状态 s'；

(5) 即时奖励 $r(s,a,s')$ 作为标量函数提供给代理，基于其当前状态 s 中的行动，往往与随后的状态 s' 相关。

RL 的目标是学习一种能使预期收益最大化的策略，其中包括表示的目标函数：$\mathcal{J}(\theta) = \mathrm{D}_{\tau\sim p_\theta(\tau)}[G(\tau)] = \mathrm{D}_{\tau\sim p_\theta(\tau)}[\sum_{t=0}^{T-1} \gamma^t r_{t+1}]$。这里，$\theta$ 表示政策函数的参数。价值函数用于评估政策 π 的预期收益，包括状态价值函数和状态-行动价值函数（Q-函数）。可以根据这些值函数对政策进行迭代优化。

函数。此外，还可以通过直接搜索策略空间来最大化预期收益，这包括基于梯度的优化[32,33] 和无梯度优化。

深度强化学习（Deep Reinforcement Learning）结合了 RL 和深度学习方法，利用 RL 来定义问题和优化目标，利用深度学习来解决策略和价值函数的建模问题，随后利用误差反向传播算法来优化目标函数。Mnih 提出的深度 Q 网络（DQNs）[34]是深度 RL 领域的开创性基石，它利用卷积神经网络来估计 Q 值。在深度 Q 网络中，采用了两个关键措施：第一，冻结目标网络，即在指定时间内将参数固定在目标范围内，以确保学习目标的稳定；第二，利用经验重放，即构建一个经验池，以消除数据相关性。这个经验池由代理最近收集的经验组成，形成一个数据集。在训练过程中，从经验池中随机抽取样本，替代当前样本进行训练。这种方法打破了相邻训练样本之间的相似性，防止模型收敛到局部最优。DQNs 的学习过程如下。

算法：带经验回放的 DQN

输入状态空间 S、行动空间 A、贴现率 γ、学习率 α

初始化经验库 D，容量为 N；
随机初始化 Q 网络的参数□；
随机初始化目标 Q 网络的参数 $\phi^{\mathrm{Y}} = \phi$；
重复
    初始化起始状态 s
    重复
        在状态s中，选择行动 $a = \pi^{\epsilon}$；
        执行行动 a，立即获得奖励 r 和新状态 s'；
        将 s, a, r, s' 放入D；
        样本s, aa, r, s' 来自D；
        r、 s' 是终端状态 $\mathrm{y}=$
        $\mathrm{Z}r + \gamma amx\hat{Q}_{\bar{a}}(\mathrm{ss}', \mathrm{a}')$、 否则
        用损失函数训练 Q 网络：$]y - Q(\mathrm{ss}, \mathrm{aa})`$；^{2} ϕ
        $s \leftarrow s'$；
        每 C 步，执行动作：$\phi^{\mathrm{Y}} \leftarrow \phi$；
    直到 s 为终点状态；
直到 $\forall s\ and\ a, Q_\phi(s,a)$ 收敛；

输出： $Q_\phi(s, a)$

自提出以来，研究人员对基于价值的方法进行了大量扩展[35]。此外，还提出了基于模型的方法[36]，通过预测模型预测行动后的状态，并直接优化策略网络。深度 RL 还适用于更复杂的决策问题，如具有目标条件[37]、分层任务分解[38]和多代理[39]的问题。从游戏[40]、机器人[41]、自动驾驶[42]到分子生成[43]，深度 RL 在各种应用领域都取得了巨大成功。这一进步被广泛认为是向通用人工智能发展迈出的关键一步[44]。

2.4 搜索算法

搜索算法检索数据结构中存储的信息或在搜索空间中计算的信息，为规划合成路线的多步逆合成预测奠定基础。一般来说，这些算法分为两类：无信息搜索和有信息搜索。无信息搜索不利用有关状态转换成本的信息；典型的例子包括深度优先搜索和广度优先搜索。与此相反，有信息搜索采用启发式函数来评估当前状态与目标状态之间的距离，从而指导搜索进度。虽然这种方法不一定是最优的，但它能确保在合理的搜索时间内找到有利的解决方案。最佳优先搜索是采用优先队列概念的典型启发式搜索。打开列表包含当前可遍历的节点，而关闭列表则存储已遍历的节点。束搜索（Beam search）通过扩展有限集合中最有希望的节点来增强最佳优先搜索[45]。A* 搜索综合了均匀成本搜索和最佳优先搜索的优点，确保了解决方案的最优性[46]。在这种情况下，每个状态的成本包括从起始状态到当前状态的实际成本和从当前状态到目标状态的启发式成本。蒙特卡洛树搜索（Monte Carlo Tree Search，MCTS）[47] 完善了从当前状态到目标状态的价值估计。AlphaGo[48] 是 MCTS 最著名的应用之一，它在围棋搜索树中探索潜在的棋步并追踪结果。MCTS 包括四个阶段：选择、扩展、模拟和反向传播。**(见图 4)**

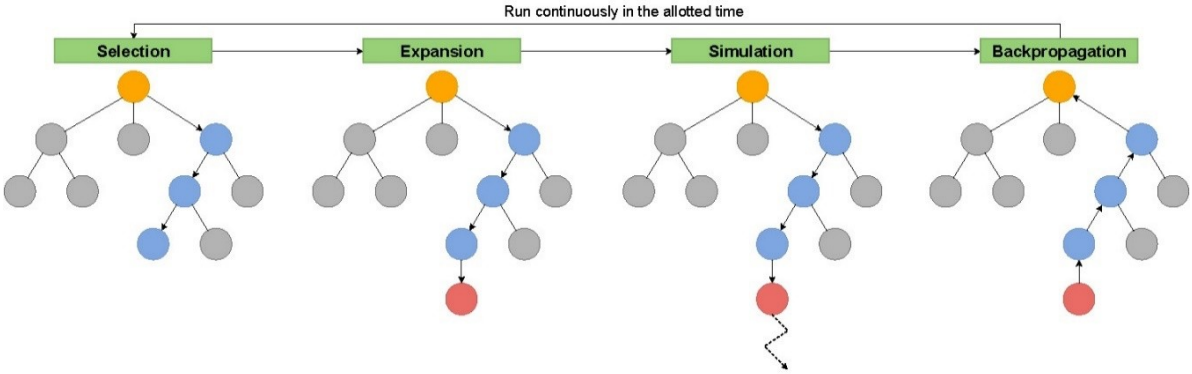


图 4 MCTS 的流程：选择、扩展、模拟和反向传播。

3. 数据来源

在 CASP 任务中，无论是通过符号 AI 还是纯粹的数据驱动建模，计算机能够解析的数据集都是前提条件。数据集的质量决定了模型的上限。毫不夸张地说，数据集的质量比模型本身更重要[49]。因此，计算化学家需要特别关注输入数据集的特征。本节将对常见的化学反应数据库进行总结和比较。

期刊和出版社根据许可协议，通过算法自动提取和专家人工编码，以计算机可读格式提供数据集。其中包括爱思唯尔（Elsevier）出版的 Reaxys 数据库，截至 2023 年，该数据库涵盖了 7,300 多万个反应。来自 16,000 种期刊和 105 个专利局的全面、最新的期刊和专利信息。它汇编了来自 16,000 种期刊和 105 个专利局的全面、最新的期刊和专利信息。为了从化学专利中提取信息，爱思唯尔和澳大利亚墨尔本大学发起了一个基于 NLP 模型的项目，名为 ChEMU[50]。化学文摘社（CAS）收录了从 1840 年到 2023 年的约 1.5 亿个反应，包括有机、无机、天然产物全合成和生物转化反应，是反应数据的最大提供者。其数据源来自期刊、专利、学位论文和重要参考文献。此外，规模较小的数据集包括 InfoChem 开发的 SPRESI，该数据集涵盖 1974 年至 2014 年期间的 460 万个反应。另一个著名的数据集是由 NextMove 软件公司创建的 Pistachio 数据集，包含从 1976 年到 2023 年的专利数据，涵盖超过 13,118,970 个反应的庞大语料库。在研究人员中，使用最广泛的数据集是 Lowe 在 1976 年至 2016 年期间提取的专利数据子集，其中包含 330 万个反应。该数据集是目前唯一可公开访问的反应数据存储库，通常称为 USPTO[51]。此外，USPTO 50K 是 USPTO 化学反应的子集和预处理迭代，由随机挑选的 50,000 个反应组成，涵盖十种不同的反应类型[52]。USPTO-MIT[53] 也是一个常用的子集，与 USPTO-50K 相比，它包含更多试剂和可能的催化剂。**表 1** 列出了常用数据集的具体细节。

虽然上述数据集包括分子结构、反应条件（溶剂、催化剂、试剂）和产率等详细信息，但它们也难免出现错误。此外，大多数专利和文献中正面数据的普遍存在也导致了产品表述的分布不均[54,55]。这种数据分布的不平衡会对模型性能产生不利影响。此外，在 CASP 框架内，失败反应的实例也起着重要作用，尤其是在涉及到区域选择性和化学选择性的情况下。为了克服这些挑战，已经发布了 THE 数据，以生成更加一致的数据[56]。IBM 发布了一种采用自然语言处理（NLP）的方法，从

专利和科学文献中提取实验程序，从而创建结构化、自动化的友好格式[57]。Pistoia 联盟与爱思唯尔合作，定义了用于交换反应信息的统一数据模型 (UDM)。电子实验室笔记（ELN），一种新型数据集

从一家大型制药公司的电子实验笔记本中提取的数据，不受发表偏向高产反应的影响[58,59]。

值得一提的是，通过比较各种数据源，包括专利（美国专利商标局和 Pistachio）、文献和专利（Reaxys）以及工业数据（阿斯利康 ELN），尽管它们的模板集规模相似，但在反应空间的覆盖范围上却有所不同。Reaxys 因其广泛而独特的反应模板多样性而脱颖而出，提供了更广阔的反应空间[60]。

表 1

用于逆合成预测模型的数据集概览。

数据集	资料来源	样本量	反应空间覆盖率
Reaxys	期刊和专利	7300k	+++++
ChEMU	专利	-	-
中科院	期刊和专利	15000k	-
SPRESI	文献	4600k	-
开心果	美国专利商标局 + 欧洲专利局	9000k	++
美国专利商标局-满	美国专利商标局	3300k	++++
美国专利商标局 50K	美国专利商标局	50k	+++

4. 分子代表

对于 CASP 任务而言，数据集的质量和特征工程的艺术对模型的性能起着决定性作用。因此，化学家们借助数学工具设计出了许多不同的分子表示方法。这些方法旨在用抽象的数学符号囊括分子的全部信息。一维分子表示方法只能表示不包含结构模式的全局分子特性，如 pKa、logP 等。二维分子表示方法可以表示结构模式，而不需要明确的三维信息，包括 SMILES（简化分子输入行输入系统）[61,62]、指纹和分子图，它们是逆合成任务中使用的主流方法。三维分子表征方法，如基于图像的方法，可以包含高维信息，但在某些情况下并不一定意味着更好的性能。近年来，有人提出了一种三维分子表征学习框架，以自动捕捉更多的高维信息[63]。

4.1 分子字符串表示法

SMILES 是最广泛采用的分子结构字符串表示系统。SMILES 系统结合了特定的语法规则和化学原理，能够严格地表示分子结构。SMILES 的优势之一是能够将反应

预测任务转化为机器翻译任务。对于序列建模问题，利用人工智能领域的自然语言处理（NLP）模型可以高效地解决这些问题[64]。例如，基于注意力机制的 Transformer 架构是最受计算化学家青睐的 NLP 模型之一。对于化学反应的 SMILES 表示法，反应物、试剂和产物可以用符号连接起来，这与分子指纹法类似。>"符号用于表示反应的方向。对于

例如，"反应物 > 试剂 > 产品"。然而，SMILES 语法对序列敏感，难以处理立体化学。SMARTS 作为 SMILES 语言的扩展，是一种描述分子模式和性质的语言。SMARTS 可用于创建查询。SMARTS 的一个显著特点是允许使用通配符来表示原子和化学键。因此，SMARTS 被广泛应用于化合物数据库结构的计算机化搜索，从而实现高效灵活的化学结构搜索。

自参照嵌入字符串（SELFIES）[65] 是一种既 100%稳健又可由人类阅读的分子结构表示方法，它的提出是为了克服 SMILES 的局限性。InChI[66] 是另一种基于字符串的化学结构表示法，与 SMILES 相比，它具有唯一性和可逆性的优势。这些方法不再需要通过原子原子映射来识别反应中心。**下图 5** 展示了咖啡因的 SMILES，包括确保其 SMILES 表示的过程。

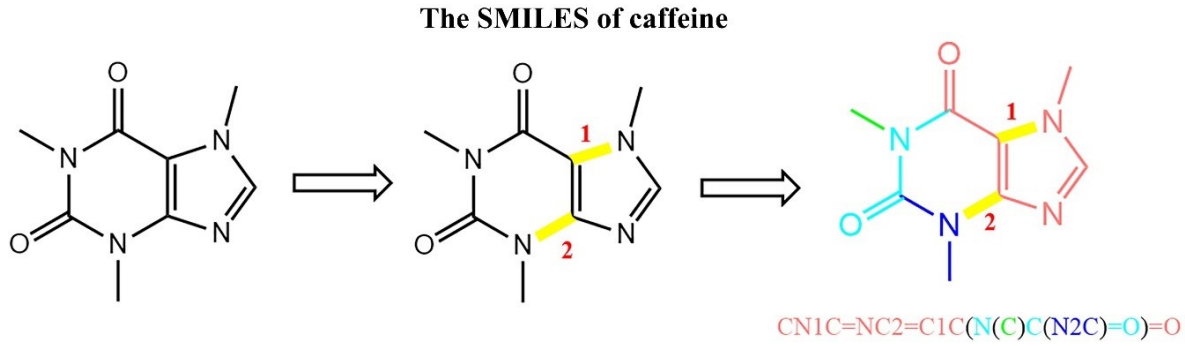


图 5 获取咖啡因 SMILES 表示法的过程。

4.2 分子指纹

分子指纹是化学信息学中表示分子的另一种有价值的工具。分子指纹背后的核心思想是将分子映射到长度为 l 的比特串或数字数组中，其中每个比特编码分子是否包含特定的亚结构特征。分子指纹具有计算效率高、易于检索等优点，是分子相似性评估的理想选择。主要方法包括基于亚结构键的指纹、基于路径的指纹和环形指纹。在此，我们主要介绍常用的分子指纹方法。有关完整的分子指纹和软件的详细介绍，请参阅 Cereto-Massagué 的著作[67]。

基于子结构密钥的指纹根据化合物中给定结构密钥列表中某些子结构或特征的存在来设置比特串。MACCS 指纹系统[68] 有两种变体，一种是 960 位，另一种是更紧凑的 166 位，都是基于 SMARTS 结构键模式。较短的变体尽管体积缩小了，但仍能有效捕捉到药物发现和虚拟筛选等任务所必需的大多数化学相关特征。相比之下，PubChem 指纹[69] 包含 881 个结构密钥，可全面表示各种亚结构特征，是在

PubChem 数据库中进行相似性搜索的基石。BCI 指纹[70]具有用户自定义选项和由 1052 个键组成的标准亚结构字典，可灵活生成[71]。最后，TGD

和 TGT 指纹[71,72]是通过二维分子图计算得出的，分别由 735 位和 13,824 位两点和三点药层表示。这些指纹具有鲜明的特征，可满足化学信息学的广泛应用，使研究人员能够有效地探索和分析化合物数据。

基于路径的指纹识别技术的工作原理是仔细检查所有遵循预定路径（通常是线性路径）的分子片段，直至达到特定的键数。随后，每条路径都要经过散列处理，以生成唯一的指纹。这些指纹可用于快速的子结构搜索和有效过滤。在这些指纹类型中，"日光指纹 "最为突出[73]，它由多达 2048 个比特组成，细致地编码了分子内所有可能的连接路径，直至指定长度。

环形指纹主要记录每个原子周围半径范围内的环境。它们不太适合子结构验证查询，因为相同的片段可能表现出不同的环境，但它们在全结构相似性搜索中很有用。Molprint2D 对分子连通性表中每个原子的原子环境进行编码，将这些环境表示为不同大小的字符串[74,75]。ECFP（扩展连通性指纹）是基于摩根算法的循环指纹的扩展[76]。它们表示循环原子邻域并生成长度可变的指纹。常用的 ECFP 变体直径为 4，通常称为 ECFP4。直径为 6 的 ECFP6 也很常见。FCFP（功能类指纹）是 ECFP 的一种变体，用于索引原子的功能。具有相似功能的不同原子在指纹中并不区分。它可以表示立体化学信息，可进一步用于推断结构-活性关系。

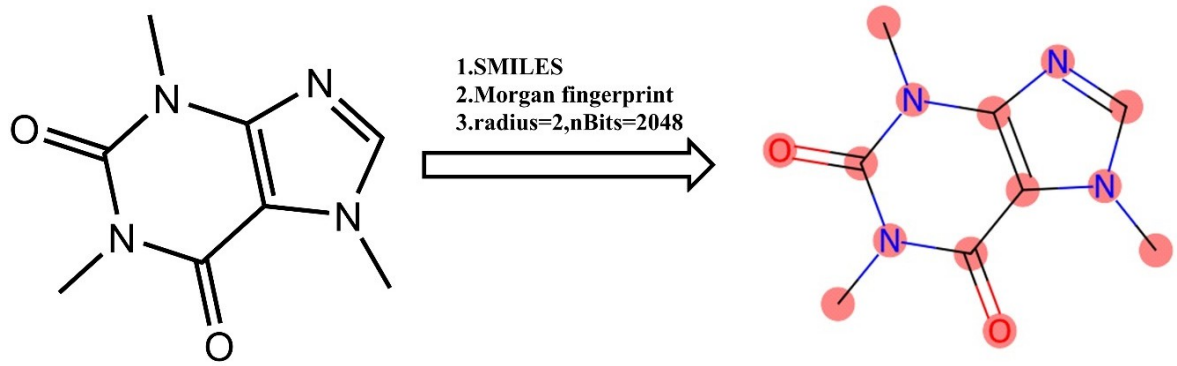


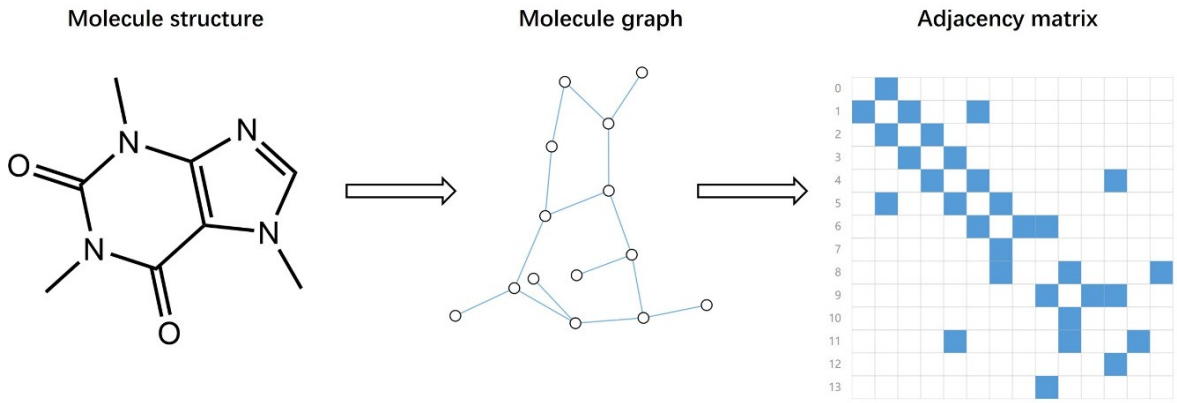
图 6 咖啡因结构的结果，其中与摩根指纹相关的原子高亮显示。

4.3 分子图

随着图神经网络的飞速发展，分子图也引起了 CASP 领域研究人员的极大关注。无向图是图论中的一种基本数据结构，由带有相关权重的节点和边组成。无向图中的边没有明确的方向，允许节点 A 和节点 B 之间存在双向边。它的每个元素都代表节点之间是否有边相连。邻接矩阵的大小就是图中顶点的数量。邻接矩阵的对角线元素为

如果邻接矩阵的行i 和列j 的元素为 1，则节点i 和节点j 由一条边连接。然而，邻接矩阵的空间复杂度为 $O(n^2)$，其中 n 是图中的顶点数。因此，为了提高计算效率，如果邻接矩阵的大小较大，可以将邻接矩阵转换为节点、边和全局的特征向量，这些特征向量一般用作输入特征。

分子是物质的基本成分，由排列在三维空间中的原子和电子组成。虽然所有粒子都会相互作用，但一对原子之间的稳定分离构成了共价键。不同的原子对和成键构型（包括单键和双键）表现出不同的原子间距离。这一固有特征形成了以原子为节点、化学键为边的图表示法[77-79]。咖啡因的图表示如**图 7** 所示，包括其分子结



构、分子图和邻接矩阵。

图 7 咖啡因的图示。

与 SMILES 和分子指纹相比，分子图可以表示更多的化学结构信息，包括原子类型、键类型、拓扑结构等。在图的节点和边上还可以添加三维信息，如键长、键角等。此外，图表示法不受原子顺序的影响。然而，从分子结构中提取图表示的高效算法是分子图实际应用的先决条件[78,80]。

对于表示反应图，从预先训练的模型中提取反应是一种很有前途的方法。此外，原子映射的使用可使单个浓缩反应图（CGR）有效地表示化学反应[81]，它是反应物图和生成物图的叠加。

5. 逆合成战略评估

5.1 候选者反应评估

在递归合成中，"组合爆炸 "是一个棘手的问题。科学家们努力将递归展开限制在最有希望的断键处，从而获得易于合成的结构。

分子结构的可合成性在候选反应评估中至关重要。合成可得性得分（SA Score）利用了与常见可合成结构特征成线性比例的片段的贡献，并对罕见片段的存在进行

惩罚。

和复杂的结构特征[82,83]。Chematica 通过限制结构复杂性、反应步骤长度、反应冲突和保护基团，开发了一种评估合成难度的指标。SCScore 所依据的原则是，反应产物的合成复杂度应高于其反应物[84,85]。其他评估方法包括基于支持向量机的 DRSVM[86] 和当前的复杂度指标[87]。

5.2 模型演变

在 CASP 建模工作流程中，模型评估起着举足轻重的作用。CASP 任务由于其特殊性，与传统的回归和模式识别任务有很大不同。为了选择符合实际逆合成任务的模型，应采用适合这些任务的不同评价指标。逆合成任务一般分为两类：单步逆合成和多步逆合成预测。

对于单步逆合成，Top-N 精确度计算是评估单步策略性能的常用指标。它考察的是整套真实前体（即模板库中报告的相应目标分子的实际反应物）是否属于模型建议的前 N 个前体。这一指标要求分子结构完全匹配，可以用分子相似度来衡量。相似度得分 1 表示结构完全相同[88]。此外，还提出了一些用于单步逆合成的替代评价指标[89]。对于多步骤逆合成，可通过重复使用单步骤逆合成方法进行评估。

6. 基于模板的模型

基于模板的模型通常涉及目标分子与整个模板库的匹配。然后，解决子图同构问题，获得候选反应物。基于模板的系统的核心在于使用逆合成模板。**如图 8** 所示，反应模板由分子子图模式表示，这些子图模式编码了反应过程中原子连接性的变化。在数学上，逆合成模板 T 用以下规则表示：

$$T\colon p^T \rightarrow \{r_i{}^T\}_{i=1}^{n_r}$$

其中p^T是生成物P的子图，可视为反应中心，而$r_i{}^T$是第i个反应物的子图。

从目标分子开始，按照预定义的规则选择模板，并将其应用于目标分子以确定反应物。虽然与无模板方法相比，基于模板的方法具有更好的可解释性和准确性，但它们的计算要求很高，而且在模板库之外的通用性有限。现代智能算法的使命就是降低这一过程的计算复杂度。

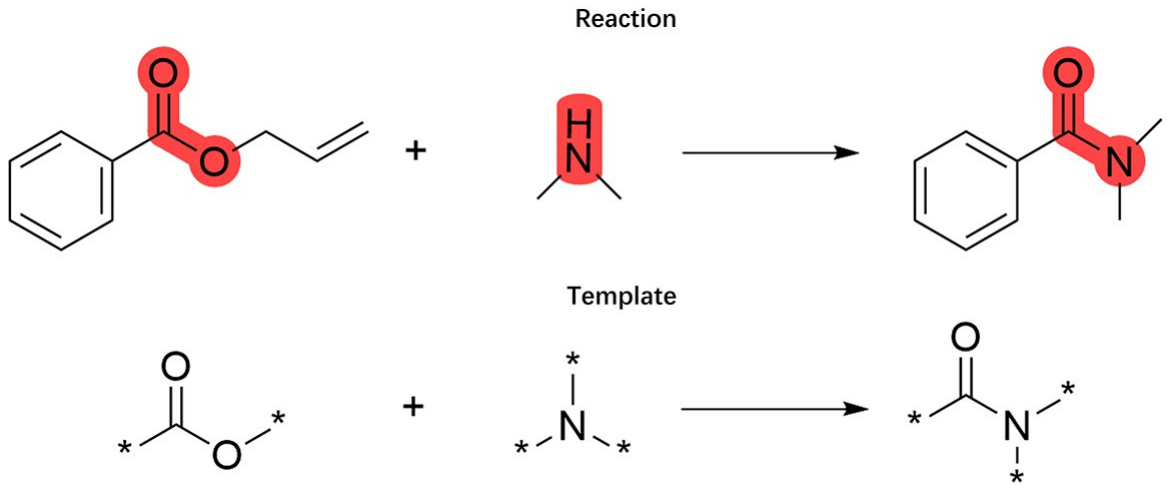


图 8 化学反应及其逆合成模板示意图。

传统上，反应规则是由专家定义和手工编码的。Szymkuc 等人综述了在合成规划中使用由人类专家编码的反应模板的情况[90]。随着反应空间以每年 4.4% 的速度呈指数增长[91]，手工编码已成为一项难以承受的任务。反应编码的另一种方法是利用算法，通过原子到原子的映射提取反应中心，从而确定反应物和产物之间的对应关系[92-95]。对于给定的反应，我们可以将改变键连接性的原子集合确定为反应中心。然后通过算法提取反应中心和相邻原子，并进行归纳，形成相应的逆合成模板。

利用现有的反应模板，Coley 等人提出了一种基于分子相似性度量的逆合成方法[96]，如 Morgan2noFeat、Dice 相似性、Tanimoto 相似性和 Tversky 相似性。这种方法仅根据与已知反应先例的类比对目标分子进行分解，因此本质上不利于创造性地断开连接。Segler 等人以扩展连接性指纹（ECFP）为输入，构建了一个基于深度神经网络的模型，该模型可以学习解决反应性冲突并优先选择最合适的转换规则，是最早的基于 ML 的模板模型之一[97]。该模型解决了将相似模板归入子组的多类分类问题。该模型的性能通常被用作基于模板方法的基准之一。Watson 等人提出了一种使用反向反应变换（RRTs）的基于模板的方法[98]。RRTs 从包含相似反应的簇中提取。通过在 RRT 资源库中搜索可能的合成路线，该方法可将目标分子分解为基本构件。Genheden 等人开发了逆合成软件 AiZynthFinder[99]。该算法基于蒙特卡洛树搜索，可递归地将分子断开为可购买的前体。树搜索由人工神经网络策略指导，该策略利用反应模板库提出可能的前体。Park 等人针对化学反应数据集中的类不平衡问题，提出了基于产物分子结构相似性（随机性、不相似性）聚类的欠采样方法[100]

，显著提高了预测精度。Chen 等人提出了一种局部逆合成框架 LocalRetro[101]，该框架假定分子变化发生在化学反应过程中。

在反应过程中主要是局部效应。作为补充，引入了全局注意机制来解释非局部效应。Seidl 等人基于现代 Hopfield 网络提出了一种基于模板的单步逆合成模型[102]，通过学习分子和反应模板的编码来预测模板与给定分子的相关性。模板表示法可在不同反应中通用。Genheden 等人开发的 AiZynthTrain[103]，是一种稳健、可重复、可扩展的端到端逆合成模型。其流程包括两个管道，分别建立基于模板的一步逆合成模型和破环模型。此外，他们还强调了启发式方法的重要作用。Dai 等人提出了一种基于分层抽样方法的条件图逻辑网络模型[104]。条件图逻辑网络是一种建立在图神经网络基础上的条件图模型，可以学习何时应用反应模板中的规则，隐含地考虑最终反应在化学上是否可行以及是否具有战略性。Yan 等人提出的 RetroComposer[105]，可以在训练模板之外合成新模板。此外，他们还开发了一种有效的候选评分模型，可以捕捉原子级的转化。从广义上讲，基于模板的模型可以包括基于量子计算的逆合成模型，因为量子计算可以生成新的反应模板。Liu 等人建立了一个基于反应动力学的逆合成规划框架，用于设计合成途径[73]。正向分析部分包括基于 TST 的反应动力学模型和 DFT。逆合成规划部分包括决策树模型和广度优先搜索算法。为了解决数据集中样本质量不高的问题、Toniato 等人提出通过第一原理计算为模型再训练提供缺失数据[106]。

表 2

基于模板方法的逆合成预测性能概览。

Methods-1P-5	**算法**	**数据集**	**特征**	**TOP-1**	**TOP-5**	**到达率**	**顶部-5 精度**	**消息来源代码可用性**
				有反应类		**无反应类**		
追溯器 m	相似性	-	指纹	52.9	81.2	37.3	63.3	Y
公园	泰勒布蒂娜算法	Reaxys	微笑+指纹	-	-	51	84	Y
当地	注意机制	美国专利商标局-50k	示意图	63.9	92.4	53.4	85.9	Y
Seidl 等人	霍菲尔德网络	美国专利商标局-50k	微笑	-	-	51.8	81.2	Y
神经元 ym	ANN	Reaxys 50k	ECFP	55.3	81.4	44.5	72.4	Y
GLN	GLN	美国专利商标局-	示意图	64.2	85.6	52.4	75.6	Y

RetroC	多个	美国专利商标局	图	65.	89.	54.	83.2	Y

	omposeDL 模型	50k	9	55
r				

7. 无模板模型

最近，无模板方法吸引了越来越多的关注，因为它们避免了计算密集型的子图匹配问题。这些方法利用分子的文本表示（SMILES 或 InChI）将逆合成任务转化为翻译任务，而翻译任务可通过使用深度学习中的强大方法来解决。这一过程不再需要原子到原子映射来识别反应中心。这类纯数据驱动的方法通常不需要结合显式化学知识。在相关数据丰富的情况下，这些方法可以取得令人满意的性能。下文概述了这些方法，它们分为深度神经网络、序列到序列模型、图形神经网络和小样本技术。

7.1 深度神经网络

Baylonet 等人提出了一种基于深度公路网络（DHN）的多尺度逆合成预测框架[107]。该过程由两部分组成：建立一个 DHN 模型来预测反应组，并利用在已确定反应组内的反应子集上训练的 DHN 预测生成分子的转化规则。Hasic 等人训练了逆合成模型，用于识别分子亚结构指纹表征上的潜在断点[108]。该模型仅使用目标物的单个分子亚结构来识别潜在的断点，而不依赖于化学反应类别等其他信息。整体路径评估机制是逆合成模型不可或缺的一部分。Mo 等人提出了一种动态树状结构长短期记忆（tree-LSTM）模型[109]。

7.2 序列对序列

seq2seq的主要思想是将逆合成预测作为一个序列建模问题，以目标分子为输入序列，以反应物、试剂和催化剂为输出序列。Transformer 是本世纪最流行的 seq2seq 模型，它纯粹基于多头注意机制。来自变换器的双向编码器表示法（BERT）[110]的引入也提高了无模板策略的性能。基于序列建模的逆合成模型已成为应用最广泛的人工智能逆合成模型，几乎所有模型都依赖于注意机制。

将化学与自然语言处理相结合的想法最早由 Cadeddu 等人提出[111]。Liu 等人提出了一个由两个递归神经网络组成的编码器-解码器框架，该框架将逆合成预测任务视为序列到序列的映射问题[112]。与基于模板的基线模型相比，seq2seq 模型有几个

优点。首先，seq-2-seq 模型可以隐式学习反应规则和候选排序指标，从而避免了像基于模板的方法那样使用独立的反应复杂度排序指标。其次，seq-2-seq 模型比基于规则的方法更容易扩展。Tetko 等人提出了一种逆合成的 Transformer 模型。

反应预测任务[113]。

近年来，Guo 等人开发了一个贝叶斯推理框架[114]，其中包括一个用于前向预测的预训练分子转换器，以及一个用于将前向模型反演为后向模型的基于贝叶斯条件概率法的模型。随后，通过蒙特卡洛搜索算法和后向模型的联合使用，获得了一系列不同的高概率反应序列。Zheng 等人开发了一种无模板自校正逆合成预测器（SCROP），利用变换器模型神经网络框架进行训练，从而完成逆合成预测任务[115]。对于训练集之外的化合物，该方法的准确率高于其他先进方法。Duan 等人提出了一种基于注意力的 NMT 模型[116]，即 Tensor2Tensor（T2T）模型，该模型与机器翻译任务相比具有很大优势。它的并行性更强，所需的训练时间也大大减少。Tetko 等人提出了基于输入数据和目标数据增强的 Transformer 模型框架[117]，它消除了神经网络记忆数据的影响，提高了神经网络预测新序列的性能。Seo 等人提出了一种新的无模板模型--图截断注意（GTA）[118]，通过在 seq2seq 模型中插入图信息，利用序列和图表示。它在编码器中使用乘积分子的邻接矩阵屏蔽自注意力层，并在解码器中使用自动算法获得的原子映射将新损失应用于交叉注意力层。Mann 等人提出了一种使用基于 SMILES 语法的表征的单步逆合成预测方法[119]。对这种语法表征的信息理论分析表明，它们的性能优于 SMILES，更适合机器学习任务。Ucak 等人提出了一种单步逆合成预测方法[120]--RetroTRAE，不存在任何基于 SMILES 的翻译问题，还引入了一种新方案，将片段和拓扑描述符作为逆合成预测任务的自然输入。Wan 等人提出的 Retroformer[121] 是一种基于 Transformer 的新型结构。它不依赖任何化学信息学工具进行分子编辑，而是通过局部关注联合编码分子序列和图谱。Fang 等人开发了一种亚结构级解码模型，利用完全数据驱动的方法自动提取产物分子中的正常保守部分[122]。Schwaller 等人结合分子转换器建模和超图探索策略，预测了每个逆合成步骤的反应物以及试剂[88]、溶剂和催化剂。Schwaller 等人使用无监督、基于注意力的 Transformer 模型网络来学习原子映射[123]。这种方法在基于规则和数据驱动的方法之间建立了联系，并在预测结果中展示了更强的化学可解释性。

分子字符串表示法存在一些局限性，包括生成无效的 SMILES 字符串和忽略化学反应的特征。Ucak 等人提出了一种基于分子片段与无模板序列到序列模型相结合来表示化学反应的新方法[124]。Zhang 等人将分子转换器模型与数据扩展和归一化预处理策略相结合[125]，提高了化学反应正向预测以及有反应类别和无反应类别的

单步逆合成预测的准确性。Zhong 等人提出了根对齐 SMILES（R-SMILES）[126]，它规定了

产品和反应物 SMILES 之间紧密对齐的一对一映射，以提高合成预测的效率。

此外，为了提高逆合成预测的多样性，Chen 等人提出了一种对不同逆合成反应进行可推广预测的模型[127]。Transformer 框架引入了两种新的预训练方法。此外，该框架还添加了一个离散潜变量模型，以鼓励该模型产生多样化的预测结果。Toniato 等人开发了一种基于 Transformer 的逆合成模型，该模型通过在目标分子的语言表征前添加分类标记来增加预测的多样性[128]。Kim 等人利用循环一致性检查[129]、参数共享和多叉潜变量开发了具有潜模型的连接双向变换器。提出的模型提高了逆合成的准确性、语法错误和多样性。Irwin 等人提出了基于 Transformer 的 Chemformer 模型[130]，结果表明自监督预训练提高了性能，并显著加快了下游任务的收敛速度。在推理中，使用这些提示标记有助于生成各种断开策略。为了克服基于小型化学数据集的预测准确率低的问题，Bai 等人将迁移学习引入逆合成分析[131]，结合 seq2seq 或 Transformer 模型进行预测和验证。

在逆合成预测中，推荐反应条件是一个重要方面。Andronov 等人提出了分子转换器框架来解决这一问题[132]。

7.3 强化学习

Schreck 等人将深度强化学习应用于反应路径搜索任务，根据用户定义的成本指标，确定在逆合成编程的每个步骤中做出最佳反应选择的策略[133]。根据模拟经验训练神经网络来估计预期合成成本。Wang 等人介绍了一种新的蒙特卡洛树搜索（MCTS）变体，它促进了整个合成空间中探索与利用之间的平衡。在相同的搜索条件下，将通过强化学习训练的价值网络与溶剂预测神经网络相结合，能更好地识别出更短的路线和更环保的溶剂。

7.4 图神经网络

图神经网络（GNN）是一种深度学习模型，可用于处理图结构数据。图是表示实体之间关系的数据结构，例如分子、蛋白质或社交网络。无向图是分子的一种图表示，以原子为节点，化学键为边，天生适合捕捉化学分子结构。

在这一框架中[134]，开发了四种不同的 GET 设计，将 SMILES 表征与通过改进的图形神经网络（GNN）学习的原子嵌入融合在一起。Sun 等人提出了一个框架，将基于序列和基于图的方法统一为具有不同能量函数的基于能量的模型（EBM）[135]，该框架建立了模型之间的联系，并揭示了模型之间的差异。

间的一致性。此外，还在双重变量中引入了一个新框架，以促进前向预测和后向预测之间的一致性。Tu 等人提出了一种 Graph2SMILES 模型，该模型结合了用于文本生成的 Transformer 模型的优势和分子图编码器的包覆不变性，从而减少了对输入数据增强的需求[136]。Liu 等人提出了一种估算可合成性的新方法 RetroGNN[137]。这一过程包括使用合成规划软件搜索许多随机分子的路线，并利用这些信息训练一个 GNN，以预测合成规划软件给定目标分子的结果。Sacha 等人提出了分子编辑图注意网络（MEGAN）[138]，这是一种端到端的编码器-解码器神经模型。将反应表示为一系列编辑，使 MEGAN 能够有效地探索合理的化学反应空间。Thakkar 等人引入了一种描述分子断开连接的提示，以克服逆合成推荐中训练数据库的偏差[139]。断开提示的使用增强了化学家对断开预测的控制能力，从而产生了更多样化和更具创造性的推荐。Wang 等人提出了 RetroExplainer[140]，它将逆合成任务表述为一个分子组装过程，其中包含多个深度学习引导的逆合成操作：多含义和多尺度图 Transformer 模型、结构感知对比学习和动态自适应多任务学习。它优于最先进的单步逆合成方法，并具有良好的可解释性。GNN-Retro[141] 是一种将 GNN 与最新搜索算法相结合的方法，由 Han 等人提出。在该框架中，GNN 的结构可以纳入相邻分子的信息，这将提高我们框架的估计精度。Jiang 等人通过分子重构预训练任务实施原子守恒规则，通过反应类型引导比较预训练任务实施指定反应中心的反应规则，成功提高了模型的准确性[142]。Liu 等人提出了一个利用上下文信息改进逆合成规划的框架[143]。他们将合成路线视为反应图，并建议通过三个步骤整合上下文：将分子编码为嵌入、汇总路线信息以及读出预测反应物。

7.5 混合人工智能系统

有人提出了结合现代搜索算法和符号人工智能的化学信息搜索方法。他们将 MCTS 与用于指导搜索的扩展策略网络[144] 和用于预选最有前途的逆合成步骤的 "范围内 "过滤网络相结合。与基于提取规则和手工编码的传统搜索方法相比，它的运行速度快 30 倍，且准确性高。AutoSynRoute 是一种无模板逆合成模型，由 Lin 等人提出[145]，其中包括使用 Transformer 模型的逆合成预测和带有启发式评分的 MCTS 路由规划。与基于模板的模型不同，它可以学习分子的全局化学环境，但继承了基于 SMILES 模型的缺点。Hong 等人提出了一种经验引导的蒙特卡洛树搜索（EG-MCTS），即通过综合经验而不是滚动来学习知识[146]。Latendresse 等人提出的

SynRoute[147]，使用相对较少的反应模板以及基于文献的反应数据库来搜索目标化合物的实用合成路线。对于每个反应模板，都会训练一个机器学习分类器，以便使

预测Chen 等人提出了一种基于神经网络模型的 A* 搜索，该模型将反应信息表示为 AND-OR 树（AND 节点表示反应，OR 节点表示分子），搜索由一个神经网络引导，该网络从过去的逆合成规划经验中学习分子的合成成本[148]。Chematica[149,150] 基于仅有 50,000 条规则的高质量化学数据库，利用对非选择性反应、紧张的中间体和不可能的结构主题的惩罚，以及启发式搜索来指导反应网络的导航。一旦识别出市场上可买到的构筑基块，程序就会终止，与已报道的方法相比，只需较少的纯化步骤，从而节省了时间和成本。为规避报告方法而引入的化学键保留规则，使常规方法的开发与专利替代方法大相径庭。此外，Chematica 还通过了图灵测试。

此外，将合适的排序系统与人工智能方法相结合可以进一步提高回溯合成模型的性能。Lin 等人设计并训练了一个基于能量的模型来重新排序推荐产品[151]，该模型可显著提高基于相似性方法的 RetroSim 和深度学习方法 NeuralSym 等模型的性能。Li 等人提出了 RetroRanker[152]，这是一种基于图形神经网络的排序模型，旨在通过重新排序减轻现有回溯合成模型预测中的频率偏差。RetroRanker 结合了每组预测反应物在得到给定产物时可能发生的反应变化，以减少化学上不可信的预测结果的排名。ASICS（智能化学合成高级系统）[153]，由 Jeong 等人提出。基于伪 A* 搜索，ASICS 生成最优合成路径，最小化合成反应值函数得分，该函数由合成可达性得分、可能性得分和相似性得分组成。此外，它还会权衡已确认反应空间和未探索反应空间的搜索结果。

表 3

无模板方法的逆合成预测性能概览。

方法	算法	数据集	特征	到	返回顶部	到	返回顶部	消息来源代码
	m		m	P-1	-5	P-1	-5	可用类型
卡尔波夫变压器 自动同步贝叶斯-逆向（MT-可	变压器 变压器 +MCTS 变钽+SMC 变	美国专利商标局-50k 美国专利商标局-50k 美国专利商标局-50k 美国专利	预测 微笑 S 微笑 S 微笑 S 微笑	mer 树状 LSTM +ASKC	LSTM			开心果

指尖	与反应类	不带反应类	-	42.7	69.8	Y
	22	54.6	80.2	43.1	71.8	Y
		62.1	88.8	53.8	84.1	N
		-	-	54.3	62.3	Y
		-	-	79.1	88.6	N

		OS 美国						
G2Retro	MPN	专利商	示意图	63.6	88.4	54.1	81.2	Y
GTA		标局-		-	-	51.1	74.8	N
LV-		50k						
变压器	注意	美国专利	微笑 S	-	-	40.5	72.8	N
	机械装							
GTE	置	商标局-		76.6	89.6	44.9	62.4	Y
		50k		60.7	87.5	48.1	78.4	Y
梅根	变	美国专利	微笑					
		商标局-		-	-	52.9	70	Y
Graph2SMI	er	50k	S					
LES	GNN+	美国专利		59	78.1	43.7	65.2	N
	变压器	商标局-	示意图					
SCROP		全部		64	86.7	53.2	76.6	Y
		美国专利						
逆变器	加网	商标局-	示意图	43.8	61.4	32.1	48.9	N
		50k						
SMILES-		美国专利						
语法基	GNN	商标局-	示意图					
		全部		-	-	51	69	Y
础	变	美国专利	简单					
		商标局-		-	-	58.3	-	Y
T2T	er	50k	S					
	变	美国专利	微笑	-	-	37.4	57	Y
		商标局-						
RetroTRAE	er	50k	S	61.4	70.4	61.4	70.4	Y
	变	美国专利	简单					
刘		商标局-		-	-	50.4	-	Y
	er	50k	类似 S					
分子结构		50k		-	-	53.5	81	Y
	注意	美国专利	微笑 S					
方	机械装	美国专利	微笑					
	置	商标局-						
	er	全部	S					
AT	变	美国专利	微笑					
		美国专利						
下部结构	变压器	商标局-	ECFP					
	er	50k	S	-	-	29	-	Y
-基于双		美国专利	MACC					
	seq2seq	美国专利	微笑 S					
	seq2seq	商标局-		65.7	84.7	53.6	74.6	N
TF		全部						
seq2seq-	GNN+se	50k	graph+					
	转移	美国专利	SMILE					
传授	q2seq	商标局-	微笑 S	-	-	60.7	83.5	N
	学习	50k						
学习		商标局-						
		50k		-	-	47.1	73.1	Y
双向	转换	美国专利	微笑					
变压器	er	商标局-	S	66.8	92.5	57.7	84.8	Y
	图形	50k						
RetroExplai	ner	美国专利		-	-	56.3	86.2	Y
Transfor	并	商标局-	示意图					
		50k						
	转换	美国专利	微笑					
R-微笑		商标局-						

Zhang et al.	转换er	美国专利商标局-50k	微笑 S	55	79	43	73	N
预培训变压器	变压器	美国专利商标局-50k	微笑 S	67.1	85.2	62	78.4	N

8. 基于半模板的模型

基于半模板的方法不使用反应模板，也不直接将产物转化为反应物。相反，基于半模板的方法利用原子映射遵循两步工作流程：(1) 首先确定反应中心，利用反应中心将产物转化为合成子（中间分子）；然后 (2) 完成合成子到反应物的转化。

G2Gs 是由 Shi 等人[79]提出的，它首先通过识别反应中心将目标分子图分割成一组合子，然后通过变分图翻译框架将合子翻译成最终的反应物图。G2Gs 的性能优于 RetroSim[96] 和 Neuralsym[97] 这两种基于模板的方法。G2Retro 的过程包括预测目标分子中的反应中心，确定组装目标的合成物，然后将这些合成物转化为反应物。G2Retro 定义了一套全面的反应中心类型，并从产物的分子图中学习预测潜在的反应中心。Nicolaou 等人介绍了一种基于 DDRAM 算法的化学背景感知数据驱动方法，推荐与先例模板相匹配的合成路线[154]。Yan 等人提出了 RetroXpert[155]，该方法将逆合成分解为两个步骤：通过图神经网络识别目标分子中的潜在反应中心并生成中间合成物；通过反应物生成模型根据获得的合成物预测相关反应物。Wang 等人提出了一种单步无模板和基于 Transformer 模型的方法，称为 RetroPrime[156]。其框架包括将分子分解为合成物，然后通过附加离去基团生成反应物，这是由广义的 Transformer 模型完成的。Somnath 等人提出了一种基于图的方法，该方法利用了前体分子的图拓扑结构在化学反应过程中基本不变这一观点[157]。第一步，该模型预测了一组将目标物转化为合成物的图编辑。然后将其扩展为分子。Ishida 等人提出了基于数据和规则的逆合成模型 ReTReK。此外，在数据驱动的逆合成预测和路径搜索框架中分别使用了图卷积网络（GCN）和 MCTS[158]。Zhang 等人采用化学信息分子图（CIMG）作为分子表示[159]，将核磁共振化学位移定义为顶点特征，将键解离能定义为边缘特征，将溶剂/催化剂信息定义为全局特征。对于给定的目标物，采用五个具有 MPNN 层的图神经网络（GNN）模型来选择生成该产物的反应模板、推断

反应物的 CIMG、选择合适的催化剂/溶剂并检查拟议反应的合理性。最后，采用 MCTS 生成合成路线路径。Lin 等人提出了一种图到图转换模型 G2GT[160]，其中图编码器和图解码器建立在标准的 Transformer 模型结构上

数据增强。此外，他们还开发了一种弱集合方法，结合了波束搜索、核和顶k采样方法，以增强多样性。Zhong 等人提出了基于图形神经网络的端到端框架--Graph2Edits[161]，以自动回归的方式预测产品图的编辑，并依次生成转化中间体和最终反应物，将基于半模板方法的两阶段过程合并为一锅学习。

表 4

基于半模板方法的逆合成预测性能概览。

方法 ds	算法	数据 t	绝技 ures	顶部 1	到 P-5	返回顶部 -1	返回顶部 -5	源代码可用性
				有反应类别		无反应类		
G2Gs	GCN	USPTO-50k	砾石 h	61	86	48.9	72.5	N
ReTReK	GCN+MCTS	Reaxys	SMILES	-	-	36.1	-	Y
G2GT	GNN+trans 前者	USPTO-50k	葡萄 h	-	-	54.1	74.5	N
图表复古 R	MPN	USPTO-50k	h	63.9	85.2	53.7	72.2	Y
变压器	变压器	USPTSMI 50k	LES	64.8	81.6	51.4	74	Y
图表 2Edits	GNN	USPTO-50k	h	67.1	91.5	55.1	83.4	Y
复古 Xpert	GNN	USPTO-50k	h	62.1	75.8	50.4	62.3	N

9. 三种分类的比较

Top-k 精确度是评估单步逆合成模型的常用指标。然而，仅仅根据 top-1 准确率对模型的性能得出结论可能会产生误导，因为在有机合成中可能存在多种可行途径。因此，本文使用 top-1 和 top-5 准确率对三种不同类型的模型进行了联合评估。

如图 9 所示，在涉及反应类别的情况下，基于模板的模型和基于半模板的模型表现出较高的平均准确率。对于反应类别未知的情况，基于模板的模型和基于半模板的模型保持了相对较高的平均准确率。此外，研究数据分布可以发现，基于模板的模型和基于半模板的模型的数据分布更紧密，这表明这些方法具有更高的稳定性。相比之下，无模板模型始终保持较高的分散性和较低的稳定性。

总之，基于模板的模型一直表现出很高的准确性和稳定性。作为一种相对较新的方法，基于半模板的模型也表现出色，并有潜力成为表现最好的方法。然而，无模

板方法

这凸显了选择合适的人工智能模型和最优超参数来完成逆合成任务的极端重要性。尽管近年来人工智能辅助逆合成方法取得了长足进步，但仍面临一些尚未解决的挑战。

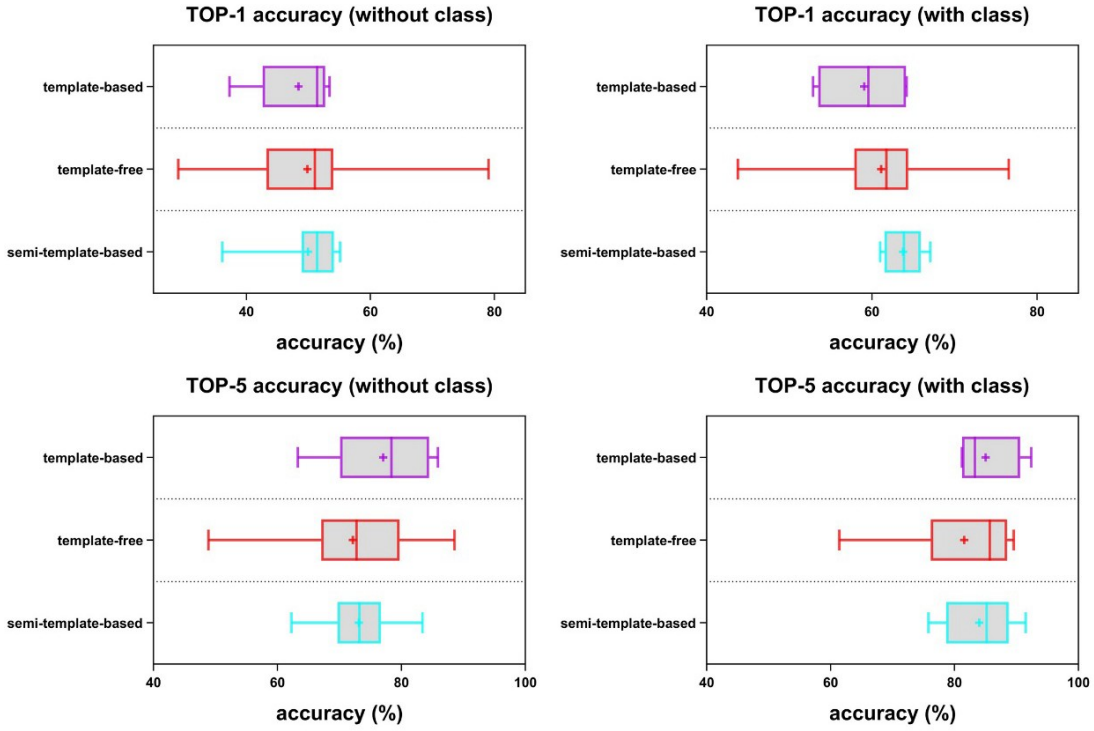


图 9 基于模板、无模板和半模板方法的 TOP-k 精度。

10. 逆合成预测研究人员面临的挑战和未来方向

过去几年中，采用现代智能算法的逆合成模型得到了快速发展。

首先，高质量数据不足是利用基于人工智能的方法预测反应路线的最大挑战之一。开发高性能的人工智能模型需要训练数据的数量和质量。然而，可供选择的公共数据集相当有限，其质量也不如商业数据库。合作准备大数据为计算化学家带来了许多机遇和挑战。此外，数据集的多样性和可变性可以提高预测性能。研究人员应努力构建涵盖立体化学信息、溶剂和催化剂等多种数据的多样化数据集。要创建一个包含各种数据资源的大型、多样化化学反应数据库，就必须开发各种方法来规范、管理和整合各种来源的反应数据。

其次，纯粹由数据驱动的模型往往缺乏可解释性，给研究人员理解模型预测背后的原理带来困难，而这是

另一个具有挑战性的问题。在机理可解释性和预测性之间取得平衡至关重要。采用与模型无关的可解释性技术，如 LIME[162-164]、SHAP[162]和 Anchors[165]，可大大有助于分析模型的可解释性。这些方法既能进行全局解释，也能进行局部解释，同时还能精确定位模型预测所依赖的关键特征。此外，应鼓励使用可解释神经网络[166,167]，因为它们整合了可解释层，通过注意和门控机制强调重要特征。

第三，根据文本序列训练的无模板模型可能会忽略键断开背后的重要化学含义，这有时会导致建议不可行。提高可解释性的方法可以解决这一难题。同时，无模板逆合成方法可能存在偏差，因为数据集中罕见反应的代表性不足。一般来说，人工智能辅助模型更有可能从数据集中出现频率较高的键断开规则中学习，而忽略其他罕见但可能导致更简单反应途径的键断开可能性。为了减少模型构建偏差，未来的一个潜在方向是将数据驱动方法与基本原理相结合。

最后，对于任何硅学设计过程，建议的合成路线都应经过实验验证。高通量和并行化实验通常用于快速生成数据和进行实验验证。然而，大多数逆合成预测模型都没有实验条件，这给实验规划带来了更多限制。自动化实验设计（DoE）的最新进展包括利用人工智能算法来优化和确定可行的反应条件[168,169]。

基于上述分析，逆合成预测研究人员未来的研究方向大有可为：

1. 构建高质量的化学反应数据集和开发智能自适应算法来处理不完整和不准确的数据，是所有基于人工智能的模型的基石。
2. 对模型可解释性和可视化的分析可能是一个热门研究方向。将 DoE 与机器人实验仪器相结合也是不可替代的一步。
3. 建议制定更复杂、更全面的反应规则和模型，以涵盖更广泛的化学反应类型和条件。
4. 应探索人工智能算法与传统规则的结合。
5. 在进行逆合成预测时，应更加关注获得更高效、更环保的化学合成条件。

11. 结论

CASP 研究对药物设计有重大影响，可以提高药物合成的速度并降低成本。现代智能算法有可能提高 CASP 的效率和准确性。未来的研究应侧重于开发更稳健、可解释的逆合成模型，并从专利和文献中提取更高质量的化学反应数据集。可解释的

人工智能分析可提高

基于人工智能模型的预测的透明度和可靠性。数据导向方法的性能在很大程度上取决于反应数据库的质量。因此，高质量的数据集是必不可少的。未来，计算机科学家、统计学家、有机化学家和计算化学家之间的跨学科合作将变得越来越重要，因为他们可以汇集不同的视角和专业知识来解决有机复合成任务。人工智能辅助合成规划研究目前尚不成熟，需要进一步研究以评估其潜在意义。由于训练数据集的差异，即使使用相同的评估指标，也无法直接比较人工智能模型的性能。在大多数情况下，没有一个模型能在所有任务中表现最佳。

在这篇综述中，我们将全面介绍由现代智能算法驱动的计算机辅助结构规划（CASP）研究的最新进展。这些模型大致可分为三类：基于模板的模型、无模板模型和基于半模板的模型。我们对这三类模型进行了比较分析，得出的结论是基于半模板的模型通常具有更好的性能。此外，我们还划分了当前面临的关键挑战，并强调了 CASP 的未来发展方向。这些最新研究表明了人工智能算法在逆合成预测方面的巨大潜力，它可以减轻有机化学家在合成规划方面的时间和成本负担。最后，在阅读完这篇综述之后，我们希望从事人工智能辅助逆合成预测领域工作的科学家们能够选择与自己的研究优势相匹配的适当方法。基于这三类研究中总结的初步工作，研究人员可以从未来的改进和研究方向中获得启发。随着逆合成技术的成熟，我们可能会看到它们与自动化化学合成系统的整合[170]，这将改善化合物的自动化生产，并带来巨大的社会和技术影响。

鸣谢

这项工作得到了国家重点研发计划（批准号：2023YFF1204903）和国家自然科学基金（批准号：22222303, 22173032, 21933010）的支持。我们还要感谢华东师范大学多功能创新平台（编号：001）提供的超级计算机时间。

参考资料

[1] Nicolaou KC, Chen JS. 通过级联反应进行全合成的艺术。Chem Soc Rev 2009;38:2993. https://doi.org/10.1039/b903290h.

[2] Nantermet PG. 反应：合成化学的艺术》。Chem 2016;1:335–6. https://doi.org/10.1016/j.chempr.2016.08.014.

[3] Gordon EM, Gallop MA, Patel DV.组合有机合成的战略与战术》。药物发现的应用。Acc Chem Res 1996;29:144-54. https://doi.org/10.1021/ar950170u.

[4] Yeh BJ, Lim WA.合成生物学：从合成有机化学的历史中汲取的教训。Nat Chem Biol 2007; 3:521-5. https://doi.org/10.1038/nchembio0907-521.
[5] Campos KR, Coleman PJ, Alvarez JC, Dreher SD, Garbaccio RM, Terrett NK, et al.

合成化学在制药业中的重要性。Science 2019;363:eaat0805. https://doi.org/10.1126/science.aat0805.

[6] Blakemore DC、Castro L、Churcher I、Rees DC、Thomas AW、Wilson DM 等：有机合成为改变药物发现提供了机遇。Nat Chem 2018;10:383-94. https://doi.org/10.1038/s41557-018-0021-z.

[7] Corey EJ.罗伯特-罗宾逊讲座。逆合成思维--要点和实例。Chem, Soc, Rev, 1988;17:111-33. https://doi.org/10.1039/CS9881700111.

[8] Bergman RG, Danheiser RL.化学研究中的可重复性。Angew Chem - Int Ed 2016;55:12548-9. https://doi.org/10.1002/anie.201606591.

[9] Duros V, Grizou J, Xuan W, Hosni Z, Long D-L, Miras HN, et al.巨型聚氧化金属酸盐的发现与结晶中的人类与机器人。Angew Chem - Int Ed 2017;56:10815-20. https://doi.org/10.1002/anie.201705721.

[10] Ley SV.化学合成工程：人类与机器和谐共处》。Angew Chem - Int Ed 2018;57:5182-3. https://doi.org/10.1002/anie.201802383.

[11] PENSAK DA, COREY EJ.LHASA-Logic and Heuristics Applied to Synthetic Analysis. 计算机辅助有机合成》，第 61 卷，美国化学会；1977 年，第 1- 页。

32. https ://doi.org/10.1021/bk-1977-0061.ch001.

[12] Henson AB, Gromski PS, Cronin L. Designing Algorithms To Aid Discovery by Chemical Robots.ACS Cent Sci 2018;4:793-804. https://doi.org/10.1021/acscentsci.8b00176.

[13] Lusher SJ, McGuire R, Van Schaik RC, Nicholson CD, De Vlieg J.大数据时代的数据驱动药物化学。今日药物发现》，2014;19:859-68。https://doi.org/10.1016/j.drudis.2013.12.004。

[14] Tetko IV, Engkvist O, Koch U, Reymond J, Chen H.《BIGCHEM：化学大数据分析的挑战与机遇》。Mol Inf 2016;35:615-21. https://doi.org/10.1002/minf.201600073.

[15] Hessler G, Baringhaus K-H.药物设计中的人工智能。Molecules 2018;23:2520. https://doi.org/10.3390/molecules23102520.

[16] Sellwood MA、Ahmed M、Segler MH、Brown N.药物发现中的人工智能。Future Med Chem 2018;10:2025-8. https://doi.org/10.4155/fmc-2018-0212.

[17] Cho K, van Merrienboer B, Gulcehre C, Bahdanau D, Bougares F, Schwenk H, et al. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation 2014. https://doi.org/10.48550/arXiv.1406.1078.

[18] Sutskever I, Vinyals O, Le QV.神经网络的序列到序列学习》。神经信息处理系统进展》，第 27 卷，Curran Associates 公司；2014 年。

[19] Bahdanau D、Cho K、Bengio Y.《2014 年联合学习对齐和翻译的神经机器翻译》，https://doi.org/10.48550/arXiv.1409.0473.

[20] Gehring J, Auli M, Grangier D, Yarats D, Dauphin YN.卷积序列到序列学习。第 34 届机器学习国际会议论文集》，PMLR；2017 年，第 1243-52 页。

[21] Gori M, Monfardini G, Scarselli F. 图域学习的新模型。Proceedings.2005 IEEE International Joint Conference on Neural Networks, 2005., vol. 2, 2005, p. 729-34. https://doi.org/10.1109/IJCNN.2005.1555942.

[22] Sperduti A, Starita A. 用于结构分类的监督神经网络.IEEE Trans Neural Netw 1997;8:714-35. https://doi.org/10.1109/72.572108.

[23] Gallicchio C, Micheli A. Graph Echo State Networks.2010 国际联合会议 国际 神经 网络 (IJCNN)、 2010, p. 1–8. https://doi.org/10.1109/IJCNN.2010.5596796.
[24] Scarselli F, Gori M, Tsoi AC, Hagenbuchner M, Monfardini G. Graph Neural Network Model.IEEE Transactions on Neural Networks 2009;20:61-80. https://doi.org/10.1109/TNN.2008.2005605.
[25] Kipf TN, Welling M. 《2016 年图形卷积网络的半监督分类》，https://doi.org/10.48550/arXiv.1609.02907.
[26] Veličković P, Cucurull G, Casanova A, Romero A, Liò P, Bengio Y. Graph Attention Networks 2017. https://doi.org/10.48550/arXiv.1710.10903.
[27] Li Y, Tarlow D, Brockschmidt M, Zemel R. Gated Graph Sequence Neural Networks 2015. https://doi.org/10.48550/arXiv.1511.05493.
[28] Cao S, Lu W, Xu Q. Deep Neural Networks for Learning Graph Representations.Proc AAAI Conf Artif Intell 2016;30. https://doi.org/10.1609/aaai.v30i1.10179.
[29] De Cao N, Kipf T. MolGAN: An implicit generative model for small molecular graphs 2018. https://doi.org/10.48550/arXiv.1805.11973.
[30] Yan S, Xiong Y, Lin D. Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition.Proc AAAI Conf Artif Intell 2018;32. https://doi.org/10.1609/aaai.v32i1.12328.
[31] Kaelbling LP, Littman ML, Moore AW.强化学习：J Artificial Intelligence Res 1996; 4:237-85.J Artificial Intelligence Res 1996; 4:237-85. https://doi.org/10.1613/jair.301.
[32] Lillicrap TP, Hunt JJ, Pritzel A, Heess N, Erez T, Tassa Y, et al.
[33] Mnih V, Badia AP, Mirza M, Graves A, Lillicrap T, Harley T, et al. Asynchronous Methods for Deep Reinforcement Learning.第 33 届机器学习国际会议论文集》，PMLR；2016 年，第 1928-37 页。
[34] Mnih V, Kavukcuoglu K, Silver D, Graves A, Antonoglou I, Wierstra D, et al. Playing Atari with Deep Reinforcement Learning 2013. https://doi.org/10.48550/arXiv.1312.5602.
[35] Hessel M, Modayil J, Van Hasselt H, Schaul T, Ostrovski G, Dabney W, et al. Rainbow：结合深度强化学习的改进。Proc AAAI Conf Artif Intell 2018;32. https://doi.org/10.1609/aaai.v32i1.11796.
[36] Moerland TM, Broekens J, Plaat A, Jonker CM.基于模型的强化学习：2022 年调查。
[37] Nair A, Pong V, Dalal M, Bahl S, Lin S, Levine S. Visual Reinforcement Learning with Imagined Goals 2018.
[38] Kulkarni TD, Narasimhan K, Saeedi A, Tenenbaum J. Hierarchical Deep Reinforcement Learning：整合时间抽象和内在动机。神经信息处理系统进展》，第 29 卷，Curran Associates 公司；2016 年。
[39] Horling B, Lesser V. A survey of multi-agent organizational paradigms.Knowl Eng Rev 2004;19:281-316. https://doi.org/10.1017/S0269888905000317.
[40] Schrittwieser J、Antonoglou I、Hubert T、Simonyan K、Sifre L、Schmitt S 等人通过学习模型规划掌握 Atari、围棋、国际象棋和将棋。Nature 2020;588:604–9.

https://doi.org/10.1038/s41586-020-03051-4.
[41] OpenAI, Akkaya I, Andrychowicz M, Chociej M, Litwin M, McGrew B, et al.

2019 年机器人手的魔方。https://doi.org/10.48550/arXiv.1910.07113。

[42] Sallab AE、Abdou M、Perot E、Yogamani S.自动驾驶的深度强化学习框架。Electron Imaging 2017;29:70-6. https://doi.org/10.2352/ISSN.2470- 1173.2017.19.AVM-023.

[43] Jeon W, Kim D. 利用强化学习和对接自主生成分子以开发潜在的新型抑制剂。Sci Rep 2020;10:22104. https://doi.org/10.1038/s41598-020-78537-2.

[44] Silver D、Singh S、Precup D、Sutton RS。奖励就够了。Artif Intell 2021;299:103535. https://doi.org/10.1016/j.artint.2021.103535.

[45] 语音理解系统。卡内基-梅隆大学五年研究成果摘要。

[46] Zeng W, Church RL.在真实道路网络中寻找最短路径：A*的案例。Int J Geogr Inf Sci 2009;23:531-43. https://doi.org/10.1080/13658810801949850.

[47] Coulom R. 蒙特卡洛树搜索中的高效选择性和备份操作符。In：Van Den Herik HJ, Ciancarini P, Donkers HHLM, editors.Computers and Games, vol. 4630, Berlin, Heidelberg：Springer; 2007, p. 72-83. https://doi.org/10.1007/978-3-540-75538-8_7.

[48] Silver D, Huang A, Maddison CJ, Guez A, Sifre L, Van Den Driessche G, et al. 用深度神经网络和树搜索掌握围棋。自然》2016;529:484-9。https://doi.org/10.1038/nature16961.

[49] Rich AS, Gureckis TM.从自然愚蠢的研究中汲取人工智能的教训。Nat Mach Intell 2019;1:174-80. https://doi.org/10.1038/s42256-019-0038-z.

[50] He J. 从化学专利中提取信息的 ChEMU 数据集 2020;2。https://doi.org/10.17632/wy6745bjfj.2.

[51] Lowe DM.从文献中提取化学结构和反应 2012.

[52] Schneider N, Stiefl N, Landrum GA.什么是什么：反应角色分配（近乎）权威指南》。J Chem Inf Model 2016;56:2336-46. https://doi.org/10.1021/acs.jcim.6b00564.

[53] Jin W, Coley C, Barzilay R, Jaakkola T. 《用魏斯费勒-雷曼网络预测有机反应结果》。神经信息处理系统进展》，第 30 卷，Curran Associates 公司；2017 年。

[54] Jia X, Lynch A, Huang Y, Danielson M, Lang'at I, Milder A, et al. 化学反应数据中的人为偏差阻碍了探索性无机合成。Nature 2019;573:251–5. https://doi.org/10.1038/s41586-019-1540-5.

[55] Coley CW, Eyke NS, Jensen KF.化学科学中的自主发现第二部分：展望。Angew Chem - Int Ed 2020;59:23414-36. https://doi.org/10.1002/anie.201909989.

[56] Lin S, Dikler S, Blincoe WD, Ferguson RD, Sheridan RP, Peng Z, et al.Science 2018;361:eaar6236. https://doi.org/10.1126/science.aar6236.

[57] Vaucher AC, Zipoli F, Geluykens J, Nair VH, Schwaller P, Laino T. Automated Extraction of Chemical Synthetic Actions from Experimental Procedures.Chemistry; 2019. https://doi.org/10.26434/chemrxiv.11448177.v1.

[58] Drake D. ELN 实施挑战。今日药物发现》2007;12:647-9。

https://doi.org/10.1016/j.drudis.2007.06.010.
[59] Saebi M, Nan B, Herr JE, Wahlers J, Guo Z, Zurański AM, et al.

用于反应产率预测的数据集。Chem Sci 2023;14:4997-5005. https://doi.org/10.1039/D2SC06041H.

[60] Thakkar A, Kogej T, Reymond J-L, Engkvist O, Bjerrum EJ.数据集及其对制药领域计算机辅助合成规划工具开发的影响。Chem Sci 2020;11:154-68. https://doi.org/10.1039/C9SC04944D.

[61] Weininger D. SMILES, a chemical language and information system.1.方法和编码规则介绍.J Chem Inf Comput Sci 1988;28:31-6. https://doi.org/10.1021/ci00057a005.

[62] Weininger D, Weininger A, Weininger JL.SMILES.2.生成唯一的 SMILES 符号的算法.J Chem Inf Comput Sci 1989;29:97-101. https://doi.org/10.1021/ci00062a008.

[63] Zhou G, Gao Z, Ding Q, Zheng H, Xu H, Wei Z, et al：通用三维分子表征学习框架 2023. https://doi.org/10.26434/chemrxiv-2022-jjm0j-v4.

[64] Cadeddu A, Wylie EK, Jurczak J, Wampler-Doty M, Grzybowski BA.有机化学作为一种语言以及化学语言学对结构和逆合成分析的影响》。Angew Chem - Int Ed 2014;53:8108-12. https://doi.org/10.1002/anie.201403708.

[65] Krenn M, Häse F, Nigam A, Friederich P, Aspuru-Guzik A. Self-referencing embedded strings (SELFIES)：100% 稳健的分子字符串表示法。Mach Learn：https://doi.org/10.1088/2632-2153/aba947.

[66] Heller SR, McNaught A, Pletnev I, Stein S, Tchekhovskoi D. InChI, the IUPAC International Chemical Identifier.J Cheminf 2015; 7:23. https://doi.org/10.1186/s13321-015-0068-4.

[67] Cereto-Massagué A, Ojeda MJ, Valls C, Mulero M, Garcia-Vallvé S, Pujadas G. 虚拟筛选中的分子指纹相似性搜索。Methods 2015;71:58–63. https://doi.org/10.1016/j.ymeth.2014.08.005.

[68] Durant JL, Leland BA, Henry DR, Nourse JG.用于药物发现的 MDL 密钥再优化。J Chem Inf Comput Sci 2002;42:1273-80. https://doi.org/10.1021/ci010132r.

[69] Bolton EE, Wang Y, Thiessen PA, Bryant SH.第 12 章 - PubChem：小分子和生物活性集成平台。In：Wheeler RA, Spellmeyer DC, editors.4, Elsevier; 2008, p. 217-41. https://doi.org/10.1016/S1574-1400(08)00012-1.

[70] Barnard JM, Downs GM. 化学片段生成和聚类软件。J Chem Inf Comput Sci 1997;37:141-2. https://doi.org/10.1021/ci960090k.

[71] Tovar A, Eckert H, Bajorath J. 在结构多样性不断增加的化合物活性类别上进行多模板相似性搜索的二维指纹方法比较。ChemMedChem 2007;2:208–17. https://doi.org/10.1002/cmdc.200600225.

[72] Sheridan RP, Miller MD, Underwood DJ, Kearsley SK.使用几何原子对描述符的化学相似性。J Chem Inf Comput Sci 1996;36:128-36. https://doi.org/10.1021/ci950275b.

[73] 日光 理论 指纹 n.d. https://www.daylight.com/dayhtml/doc/theory/theory.finger.html（2023 年 9 月 28 日访问）。

[74] Bender A、Mussa HY、Glen RC、Reiling S.使用原子环境描述符进行化学数据库相似性搜索（MOLPRINT 2D）：性能评估.J Chem Inf Comput Sci 2004;44:1708-18. https://doi.org/10.1021/ci0498719.

[75] Bender A, Mussa HY, Glen RC, Reiling S. Molecular Similarity Searching Using Atom

环境、基于信息的特征选择和奈夫贝叶斯分类器。J Chem Inf Comput Sci 2004;44:170-8. https://doi.org/10.1021/ci034207y.

[76] Rogers D, Hahn M. Extended-Connectivity Fingerprints.J Chem Inf Model 2010;50:742-54. https://doi.org/10.1021/ci100050t.

[77] Schwalbe-Koda D, Gómez-Bombarelli R. Generative Models for Automatic Chemical Design.In：Schütt KT、Chmiela S、von Lilienfeld OA、Tkatchenko A、Tsuda K、Müller K-R 编辑。机器学习与量子物理学》，Cham：https://doi.org/10.1007/978-3-030-40245-7_21.

[78] Kearnes S, McCloskey K, Berndl M, Pande V, Riley P. Molecular graph convolutions: moving beyond fingerprints.J Comput-Aided Mol Des 2016;30:595-608. https://doi.org/10.1007/s10822-016-9938-8.

[79] Shi C, Xu M, Guo H, Zhang M, Tang J. A Graph to Graphs Framework for Retrosynthesis Prediction.第 37 届机器学习国际会议论文集，PMLR；2020 年、p.8818-27.

[80] Kwon Y, Lee D, Choi Y-S, Shin K, Kang S. 可扩展分子图生成的压缩图表示法。J Cheminf 2020;12:58. https://doi.org/10.1186/s13321-020-00463-2.

[81] Varnek A, Fourches D, Hoonakker F, Solov'ev VP.子结构片段：编码反应、分子和超分子结构的通用语言。J Comput-Aided Mol Des 2005;19:693-703. https://doi.org/10.1007/s10822-005-9008-0.

[82] Fukunishi Y, Kurosawa T, Mikami Y, Nakamura H. 基于市售化合物数据库的合成可及性预测。J Chem Inf Model 2014;54:3259-67. https://doi.org/10.1021/ci500568d.

[83] Ertl P, Schuffenhauer A. 基于分子复杂性和片段贡献估算类药物分子的合成可及性得分。J Cheminf 2009; 1:8. https://doi.org/10.1186/1758-2946-1-8.

[84] Molga K, Szymkuć S, Grzybowski BA.Chemist Ex Machina：高级合成规划由计算机。Acc 化学 Res 2021;54:1094–106. https://doi.org/10.1021/acs.accounts.0c00714.

[85] Coley CW、Rogers L、Green WH、Jensen KF。SCScore：从反应语料库中了解合成复杂性。J Chem Inf Model 2018;58:252-61. https://doi.org/10.1021/acs.jcim.7b00622.

[86] Podolyan Y, Walters MA, Karypis G. 《使用机器学习方法评估化合物的合成可及性》。J Chem Inf Model 2010;50:979-91. https://doi.org/10.1021/ci900301v.

[87] Li J, Eastgate MD.当前复杂性：评估有机分子复杂性的工具。Org Biomol Chem 2015;13:7164-76. https://doi.org/10.1039/C5OB00709G.

[88] Schwaller P, Petraglia R, Zullo V, Nair VH, Haeuselmann RA, Pisoni R, et al.Chem Sci 2020;11:3316-25. https://doi.org/10.1039/C9SC05704H.

[89] Bender A, Glen RC.分子相似性：分子信息学的关键技术。Org Biomol Chem 2004;2:3204. https://doi.org/10.1039/b409813g.

[90] Szymkuć S, Gajewska EP, Klucznik T, Molga K, Dittwald P, Startek M, et al. 计算机辅助合成规划：起点的终点。 Angew Chem - Int Ed 2016;55:5904-37. https://doi.org/10.1002/anie.201506101.
[91] Llanos EJ, Leal W, Luu DH, Jost J, Stadler PF, Restrepo G.

空间及其三种历史制度。Proc Natl Acad Sci 2019;116:12660-5. https://doi.org/10.1073/pnas.1816039116.

[92] Coley CW, Green WH, Jensen KF.RDChiral：用于在逆合成模板提取和应用中处理立体化学的 RDKit 封装程序。J Chem Inf Model 2019;59:2529-37. https://doi.org/10.1021/acs.jcim.9b00286.

[93] Jaworski W, Szymkuć S, Mikulak-Klucznik B, Piecuch K, Klucznik T, Kaźmierowski M, et al. 简单和复杂化学反应中的原子自动映射。Nat Commun 2019;10:1434. https://doi.org/10.1038/s41467-019-09440-2.

[94] Plehiers PP, Marin GB, Stevens CV, Van Geem KM.自动反应数据库和反应网络分析：利用化学信息学提取反应模板。J Cheminf 2018;10:11. https://doi.org/10.1186/s13321-018-0269-8.

[95] Plehiers PP, Marin GB, Stevens CV, Van Geem KM.自动反应数据库和反应网络分析：利用化学信息学提取反应模板。J Cheminf 2018;10:11. https://doi.org/10.1186/s13321-018-0269-8.

[96] Coley CW, Rogers L, Green WH, Jensen KF.基于分子相似性的计算机辅助逆合成。ACS Cent Sci 2017;3:1237-45. https://doi.org/10.1021/acscentsci.7b00355.

[97] Segler MHS, Waller MP.用于逆合成和反应预测的神经符号机器学习。Chemistry A European J 2017;23:5966-71. https://doi.org/10.1002/chem.201605499.

[98] Watson IA, Wang J, Nicolaou CA.逆合成分析算法的实现。J Cheminf 2019;11:1. https://doi.org/10.1186/s13321-018-0323-6.

[99] Genheden S, Thakkar A, Chadimová V, Reymond J-L, Engkvist O, Bjerrum E. AiZynthFinder: a fast, robust and flexible open source software for retrosynthetic planning.J Cheminf 2020;12:70. https://doi.org/10.1186/s13321-020-00472-1.

[100] Park MS, Lee D, Kwon Y, Kim E, Choi Y-S.基于规则的高效逆合成规划的数据欠采样模型。Phys Chem Chem Phys 2021;23:26510-8. https://doi.org/10.1039/D1CP03630K.

[101] Chen S, Jung Y.利用局部反应性和全局注意力的深度逆合成反应预测。JACS Au 2021;1:1612-20. https://doi.org/10.1021/jacsau.1c00246.

[102] Seidl P, Renz P, Dyubankova N, Neves P, Verhoeven J, Wegner JK, et al.J Chem Inf Model 2022;62:2111-20. https://doi.org/10.1021/acs.jcim.1c01065.

[103] Genheden S, Norrby P-O, Engkvist O. AiZynthTrain：用于训练合成预测模型的稳健、可重复和可扩展管道。J Chem Inf Model 2023;63:1841-6. https://doi.org/10.1021/acs.jcim.2c01486.

[104] Dai H、Li C、Coley CW、Dai B、Song L. 利用条件图逻辑网络 2020 进行逆合成预测。https://doi.org/10.48550/arXiv.2001.01408.

[105] Yan C, Zhao P, Lu C, Yu Y, Huang J. RetroComposer：基于模板的 逆合成 预

测。 生物分子 2022;12:1325. https://doi.org/10.3390/biom12091325.

[106] Toniato A, Unsleber JP, Vaucher AC, Weymuth T, Probst D, Laino T, et al.量子化学数据生成作为机器学习反应和逆合成规划可靠性增强的补充。Digit Discov 2023;2:663-73. https://doi.org/10.1039/D3DD00006K.

[107] Baylon JL, Cilfone NA, Gulcher JR, Chittenden TW.增强逆合成反应

利用多尺度反应分类的深度学习预测.J Chem Inf Model 2019;59:673-88. https://doi.org/10.1021/acs.jcim.8b00801.

[108] Hasic H, Ishida T. 基于使用分子结构指纹识别潜在断开位点的单步逆合成预测。J Chem Inf Model 2021;61:641-52. https://doi.org/10.1021/acs.jcim.0c01100.

[109] Mo Y, Guan Y, Verma P, Guo J, Fortunato ME, Lu Z, et al.https://doi.org/10.1039/D0SC05078D.

[110] Schwaller P, Probst D, Vaucher AC, Nair VH, Kreutter D, Laino T, et al.https://doi.org/10.1038/s42256-020-00284-w.

[111] Cadeddu A, Wylie EK, Jurczak J, Wampler-Doty M, Grzybowski BA.有机化学作为一种语言以及化学语言学对结构和逆合成分析的影响》。Angew Chem - Int Ed 2014;53:8108-12. https://doi.org/10.1002/anie.201403708.

[112] Liu B, Ramsundar B, Kawthekar P, Shi J, Gomes J, Luu Nguyen Q, et al. Retrosynthetic Reaction Prediction Using Neural Sequence-to-Sequence Models.ACS Cent Sci 2017;3:1103-13. https://doi.org/10.1021/acscentsci.7b00303.

[113] Karpov P, Godin G, Tetko IV.用于逆合成的变压器模型。In：Tetko IV, Kůrková V, Karpov P, Theis F, editors.人工神经网络和机器学习 - ICANN 2019：研讨会和特别会议》，第 11731 卷，Cham：Springer International Publishing; 2019、p.817–30. https://doi.org/10.1007/978-3-030-30493-5_78.

[114] Guo Z, Wu S, Ohno M, Yoshida R. Bayesian Algorithm for Retrosynthesis.J Chem Inf Model 2020;60:4474-86. https://doi.org/10.1021/acs.jcim.0c00320.

[115] Zheng S, Rao J, Zhang Z, Xu J, Yang Y.利用自校正变压器神经网络预测逆合成反应。J Chem Inf Model 2020;60:47-55. https://doi.org/10.1021/acs.jcim.9b00949.

[116] Duan H, Wang L, Zhang C, Guo L, Li J. 基于注意力的 NMT 模型和 "错误 "预测的化学分析的逆合成。RSC Adv 2020;10:1371-8. https://doi.org/10.1039/C9RA08535A.

[117] Tetko IV, Karpov P, Van Deursen R, Godin G. 用于直接和单步逆合成的最新增强型 NLP 变换器模型。https://doi.org/10.1038/s41467-020-19266-y.

[118] Seo S-W、Song YY、Yang JY、Bae S、Lee H、Shin J 等. GTA：用于逆合成的图形截断注意。AAAI 2021;35:531–9. https://doi.org/10.1609/aaai.v35i1.16131.

[119] Mann V, Venkatasubramanian V. 使用基于语法的神经机器翻译进行逆合成预测：一种信息论方法。Comput Chem Eng 2021;155:107533. https://doi.org/10.1016/j.compchemeng.2021.107533.

[120] Ucak UV, Ashyrmamatov I, Ko J, Lee J. 通过神经机器翻译原子环境预测逆合成反应途径。https://doi.org/10.1038/s41467-022-28857-w.

[121] Wan Y, Hsieh C-Y, Liao B, Zhang S. Retroformer：突破端到端逆合成变换器的极限。第 39 届机器学习国际会议论文集》，PMLR；2022 年，第 22475-90 页。

[122] Fang L, Li J, Zhao M, Tan L, Lou J-G.单步逆合成预测的杠杆作用

通常保存的子结构。Nat Commun 2023;14:2446. https://doi.org/10.1038/s41467- 023-37969-W

。

[123] Schwaller P, Hoover B, Reymond J-L, Strobelt H, Laino T. 从化学反应的无监督学习中提取有机化学语法。Sci Adv 2021;7:eabe4166. https://doi.org/10.1126/sciadv.abe4166.

[124] Ucak UV, Kang T, Ko J, Lee J. 基于子结构的逆合成预测神经机器翻译。J Cheminf 2021;13:4. https://doi.org/10.1186/s13321-020-00482-z.

[125] Zhang B, Lin J, Du L, Zhang L. 利用数据增强和归一化预处理提高数据驱动模型的化学反应预测性能。Polymers 2023;15:2224. https://doi.org/10.3390/polym15092224.

[126] Zhong Z, Song J, Feng Z, Liu T, Jia L, Yao S, et al.Chem Sci 2022;13:9023-34. https://doi.org/10.1039/D2SC02763A.

[127] Chen B, Shen T, Jaakkola TS, Barzilay R. Learning to Make Generalizable and Diverse Predictions for Retrosynthesis 2019. https://doi.org/10.48550/arXiv.1910.09688.

[128] Toniato A, Vaucher AC, Schwaller P, Laino T. 增强基于语言的单步逆合成模型的多样性。Digital Discovery 2023;2:489-501. https://doi.org/10.1039/D2DD00110A.

[129] Kim E, Lee D, Kwon Y, Park MS, Choi Y-S.使用具有潜在变量的绑定双向变换器进行有效、可信和多样的逆合成。J Chem Inf Model 2021;61:123-33. https://doi.org/10.1021/acs.jcim.0c01074.

[130] Irwin R, Dimitriadis S, He J, Bjerrum EJ.Chemformer：用于计算化学的预训练变换器。Mach Learn：https://doi.org/10.1088/2632- 2153/ac3ffb。

[131] Bai R, Zhang C, Wang L, Yao C, Ge J, Duan H. Transfer Learning：基于小型化学反应数据集的逆合成预测再上新台阶。Molecules 2020;25:2357. https://doi.org/10.3390/molecules25102357.

[132] Andronov M, Voinarovska V, Andronova N, Wand M, Clevert D-A, Schmidhuber J. 利用分子转换器进行试剂预测可提高反应数据质量。Chem Sci 2023;14:3235-46. https://doi.org/10.1039/D2SC06798F.

[133] Schreck JS, Coley CW, Bishop KJM.通过模拟体验学习逆合成规划。ACS Cent Sci 2019;5:970-81. https://doi.org/10.1021/acscentsci.9b00055.

[134] 用于逆合成的分子图增强转化器 逆合成 预测。 神经计算 2021;457:193–202. https://doi.org/10.1016/j.neucom.2021.06.037.

[135] Sun R, Dai H, Li L, Kearnes S, Dai B. Towards understanding retrosynthesis by energy-based models.神经信息处理系统进展》，第 34 卷，Curran Associates 公司；2021 年，第 10186-94 页。

[136] Tu Z, Coley CW.用于无模板逆合成和反应预测的置换不变图-序列模型。J Chem Inf Model 2022;62:3503-13. https://doi.org/10.1021/acs.jcim.2c00321.
[137] Liu C-H, Korablyov M, Jastrzębski S, Włodarczyk-Pruszyński P, Bengio Y, Segler M. RetroGNN：通过学习慢速逆合成软件快速估计虚拟筛选和新设计的可合成性。J Chem Inf Model 2022;62:2293-300.

https://doi.org/10.1021/acs.jcim.1c01476.

[138] Sacha M, Błaż M, Byrski P, Dąbrowskii-Tumański P, Chromiński M, Loska R, et al. 分子编辑图注意网络：将化学反应建模为图编辑序列。J Chem Inf Model 2021;61:3273-84. https://doi.org/10.1021/acs.jcim.1c00537.

[139] Thakkar A, Vaucher AC, Byekwaso A, Schwaller P, Toniato A, Laino T. Unbiasing Retrosynthesis Language Models with Disconnection Prompts.ACS Cent Sci 2023;9:1488-98. https://doi.org/10.1021/acscentsci.3c00372.

[140] Wang Y, Pang C, Wang Y, Jin J, Zhang J, Zeng X, et al. 基于分子组装任务的可解释深度学习框架的逆合成预测。Nat Commun 2023;14:6155. https://doi.org/10.1038/s41467-023-41698-5.

[141] Han P, Zhao P, Lu C, Huang J, Wu J, Shang S, et al. GNN-Retro: Retrosynthetic Planning with Graph Neural Networks.Proc AAAI Conf Artif Intell 2022;36:4014-21. https://doi.org/10.1609/aaai.v36i4.20318.

[142] Jiang Y, Wei Y, Wu F, Huang Z, Kuang K, Wang Z. 利用预训练学习逆合成的化学规则。Proc AAAI Conf Artif Intell 2023;37:5113-21. https://doi.org/10.1609/aaai.v37i4.25640.

[143] Liu S, Tu Z, Xu M, Zhang Z, Lin L, Ying R, et al. FusionRetro: Molecule Representation Fusion via In-Context Learning for Retrosynthetic Planning.第 40 届国际机器学习大会论文集，PMLR；2023 年，第 22028-41 页。

[144] Segler MHS, Preuss M, Waller MP.用深度神经网络和符号人工智能规划化学合成。自然》2018; 555:604-10. https://doi.org/10.1038/nature25978.

[145] Lin K, Xu Y, Pei J, Lai L.使用无模板模型的自动逆合成路线规划。Chem Sci 2020;11:3355-64. https://doi.org/10.1039/C9SC03666K.

[146] Hong S, Zhuo HH, Jin K, Shao G, Zhou Z. 利用经验指导的蒙特卡洛树搜索进行逆合成规划。https://doi.org/10.1038/s42004-023-00911- 8.

[147] Latendresse M, Malerich JP, Herson J, Krummenacker M, Szeto J, Vu V-A, et al. SynRoute：逆合成规划软件。J Chem Inf Model 2023;63:5484-95. https://doi.org/10.1021/acs.jcim.3c00491.

[148] Chen B, Li C, Dai H, Song L. Retro*：用神经引导的 A* 搜索学习逆合成规划。第 37 届机器学习国际会议论文集》，PMLR；2020 年，第 1608-16 页。

[149] Mikulak-Klucznik B, Gołębiowska P, Bayly AA, Popik O, Klucznik T, Szymkuć S, et al. 复杂天然产物合成的计算规划。Nature 2020;588:83–8. https://doi.org/10.1038/s41586-020-2855-y.

[150] Klucznik T, Mikulak-Klucznik B, McCormack MP, Lima H, Szymkuć S, Bhowmick M, et al. Efficient Syntheses of Diverse, Medicinally Relevant Targets Planned by Computer and Executed in the Laboratory.Chem 2018;4:522–32. https://doi.org/10.1016/j.chempr.2018.02.002.

[151] Lin MH, Tu Z, Coley CW.通过重新排序提高一步逆合成模型的性能。J Cheminf 2022;14:15. https://doi.org/10.1186/s13321-022-00594-8.

[152] Li J, Fang L, Lou J-G.RetroRanker：利用反应变化，通过重新排序改进逆合成预测。J Cheminf 2023;15:58. https://doi.org/10.1186/s13321-023-00727- 7.

[153] Jeong J、Lee N、Shin Y、Shin D. 基于智能生成的最佳合成路径

利用反应大数据进行知识图谱推理与逆合成预测.J Taiwan Inst Chem Eng 2022;130:103982. https://doi.org/10.1016/j.jtice.2021.07.015.

[154] Nicolaou CA, Watson IA, LeMasters M, Masquelin T, Wang J. Context Aware Data-Driven Retrosynthetic Analysis. 化学 信息 模型 2020;60:2728–38. https://doi.org/10.1021/acs.jcim.9b01141.

[155] Yan C, Ding Q, Zhao P, Zheng S, Yang J, Yu Y, et al. RetroXpert: Decompose Retrosynthesis 预测 像 A Chemist. Chemistry; 2020. https://doi.org/10.26434/chemrxiv 11869692.v3.

[156] Wang X, Li Y, Qiu J, Chen G, Liu H, Liao B, et al：基于变压器的单步逆合成预测方法。化学工程学报 2021;420:129845。https://doi.org/10.1016/j.cej.2021.129845.

[157] Somnath VR, Bunne C, Coley C, Krause A, Barzilay R. Learning Graph Models for Retrosynthesis Prediction.神经信息处理系统进展》，第 34 卷，Curran Associates 公司；2021 年，第 9405-15 页。

[158] Ishida S, Terayama K, Kojima R, Takasu K, Okuno Y. 结合逆合成知识的人工智能驱动合成路线设计。J Chem Inf Model 2022;62:1357-67. https://doi.org/10.1021/acs.jcim.1c01074.

[159] Zhang B, Zhang X, Du W, Song Z, Zhang G, Zhang G, et al.Proc Natl Acad Sci 2022;119:e2212711119. https://doi.org/10.1073/pnas.2212711119.

[160] Lin Z, Yin S, Shi L, Zhou W, Zhang YJ.G2GT：利用图对图注意力神经网络和自我训练进行逆合成预测。J Chem Inf Model 2023;63:1894-905. https://doi.org/10.1021/acs.jcim.2c01302.

[161] Zhong W, Yang Z, Chen CY-C.使用端到端图形生成架构进行分子图编辑的逆合成预测。https://doi.org/10.1038/s41467-023-38851-5.

[162] Štrumbelj E, Kononenko I. 用特征贡献解释预测模型和个体预测。Knowl Inf Syst 2014;41:647-65. https://doi.org/10.1007/s10115-013-0679- x.

[163] Ribeiro MT, Singh S, Guestrin C. "Why Should I Trust You?"：解释任何分类器的预测。第 22 届 ACM SIGKDD 知识发现与数据挖掘国际会议论文集》，美国纽约：美国计算机协会；2016 年、p.1135–44. https://doi.org/10.1145/2939672.2939778.

[164] Slack D, Hilgard S, Jia E, Singh S, Lakkaraju H.愚弄 LIME 和 SHAP：对事后解释方法的对抗性攻击。AAAI/ACM人工智能、伦理与社会会议论文集》，美国纽约州纽约市：美国计算机协会；2020年，第180-6页。https://doi.org/10.1145/3375627.3375830.

[165] Ribeiro MT, Singh S, Guestrin C. Anchors：高精度模型诊断解释。Proc AAAI Conf Artif Intell 2018;32. https://doi.org/10.1609/aaai.v32i1.11491.

[166] Luong M-T、Pham H、Manning CD。基于注意力的神经机器翻译的有效方法》。ArXivOrg 2015. https://arxiv.org/abs/1508.04025v5 （2023 年 10 月 7 日访问）。

[167] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is All You Need.神经信息处理系统进展》，第 30 卷，Curran Associates 公司；

2017.

[168] Houben C, Lapkin AA.自动发现和优化化学过程。当前 观点 中的 化学 工程 2015;9:1–7. https://doi.org/10.1016/j.coche.2015.07.001.

[169] Gromski PS, Granda JM, Cronin L. 使用 "化学计算机 "进行通用化学合成和发现。Trends Chem 2020;2:4-12. https://doi.org/10.1016/j.trechm.2019.07.004.

[170] Peplow M. 有机合成：机器人化学家。Nature 2014;512:20-2. https://doi.org/10.1038/512020a.