ACS Synthetic Biology

pubs.acs.org/synthbio

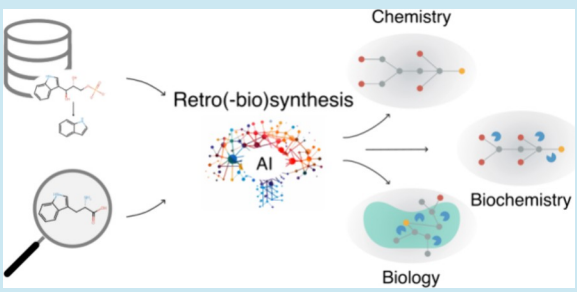
评论

逆生物合成的人工智能方法和模型：范围审查

纪尧姆-格里库特、菲利普-迈耶、托马斯-杜伊古和让-卢普-福隆*

引用：https://doi.org/10.1021/acssynbio.4c00091

摘要：逆合成的目的是通过战略性地将分子分解成现成的 构件化合物，从而有效地规划理想化学品的合成。逆合成在化学领域有着悠久的历史，同时也被用于生物催化和合成生物学领域。人工智能（AI）正推动我们向合成规划和化学空间探索的新前沿迈进，这正是促进生物生产的大好时机，而生物生产将更好地与绿色化学接轨，加强环保实践。在本综述中，我们总结了最近在应用人工智能方法和模型进行逆合成和逆生物合成途径设计方面取得的进展。这些技术可以基于



反应模板或生成模型，需要评分函数和规划策略来浏览可能性的逆合成图。最后，我们讨论了这一领域的局限性和有前景的研究方向。

关键词：*逆合成、逆生物合成、人工智能*

引言

逆合成[1] 对于药物和有机化学领域新化合物的开发至关重要，它为化学家提供了获得复杂和新型分子的能力。这种被重新命名为逆向生物合成的方法也用于生物催化反应，即由酶催化的反应。与普 通 化学合成相比，酶催化过程能以特定、高效的方式催化化学反应，所需能量更少，产生的废物也最少。生物化学反应可以*在体外进行*（如酶级联），也可以在*体内进行*（如通过代谢工程进行的合成生物学）。[2]生物化学和合成生物学面临着不同的挑战，例如酶的分离以及细胞生长和分子生产之间的 权衡。[3]

要通过逆向（生物）合成技术合成出所需的目标产品，就必须确定一组可在商业市场上轻易买到或自然存在于环境中的构件分子（也称 为 前体）。人工智能（AI）带来的技术变革为这一任务所需的每个关键组成部分铺平了新的可能性。

逆合成是通过迭代应用单步流程进行的，其中包括找到给定产物的 所有可能反应物。已提出的单步逆合成方法可分为三类：基于模板、无模板和基于半模板。基于模板的方法依赖于反应模板库

从化学反应数据集中构建模板，将其与目标分子进行匹配，并从选定的模板中提取反应物。在此，我们开发了人工智能技术来选择最有前途的模板。与众不同的是，无模板方法使用人工智能生成模型将产物直接转化为候选反应物，而基于半模板的方法则通过迭代操作产物中的键来预测反应物。

在上述所有情况下，单步反应都是通过路径规划算法反复进行的，以产生反应路径并确定可用的前体。由于合成路线的搜索空间很大，选择合适候选方 案 的 主观判断能力很强，因此预测多步骤逆合成路线本身就具有挑战性，这也是使用基于人工智能的组合图搜索方法的原因。为了指导路线规划和对预测的解决方案进行排序，我们采用了人工智能策略， 根据专用算法、评分函数或逆生物合成中酶的可用性来建议最佳选择。[4]

由于逆合成技术的快速发展，用于在进行化学合成规划时，显然需要对相关文献进行全面总结。 化学合成

https://doi.org/10.1021/acssynbio.4c00091
*ACS Synth.Biol.*xxxx, xxx, xxx-xxx

收到： 2024 年 2 月 9 日
修订： 2024 年 6 月 14 日
接受： 2024 年 6 月 14 日

A

科学文章的汇编受到了范围界定综述（PRISMA-ScR）[5] 指南的启发（注 S1）。我们在四个学术搜索引擎上进行了全面的文献检索，涵盖了生物学和计算机科学等多个学科，这两个学科都与复古（生物）合成这一跨学科领域相关。虽然这一搜索成功地识别了发表在学术期刊上的论文，但将其扩展到会议论文集，进一步扩大了研究范围，尽管它可能没有完全捕捉到在学术会议上发表的所有相关研究。如图 1 所示，本综述总结了可用于以下方面的逆合成方法

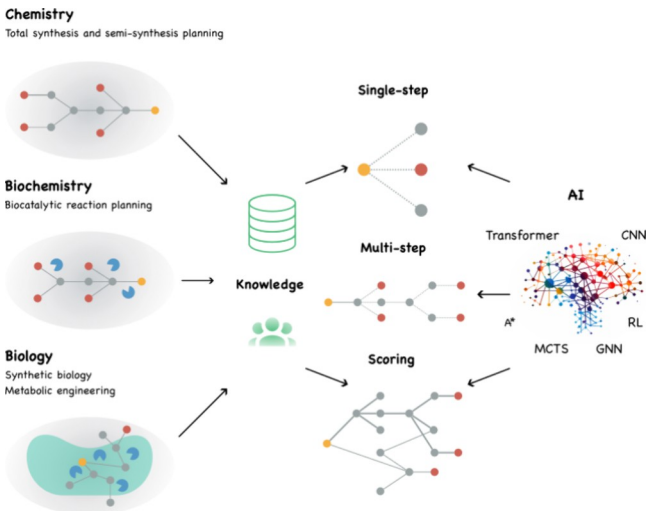


图 1.逆合成原理及其应用。逆合成是一种利用数据集和用户专业知识的计算机辅助方法。目前的逆合成算法应用于 多个领域。在化学合成中，目标分子是通过有机化学反应从市场上可买到的构筑模块中制造出来的。在生物催化反应中，使用酶来催化反应。合成生物学和代谢工程则更 进 一步，利用细菌、真菌或植物等活细胞促进生物生产途径，并提供必要的构建模块。利用分子数据库，逆向（生物）合成过程分为三个关键阶段。单步阶段包括预测生产特定产品（黄色）所需的反应物（红色和灰色节点）。多步阶段利用单步动作序列，确定将理想产物（黄色节点）与可用构件（红色节点）连接起来的可能路线。已完成的预测用实线表示，而未来的预测则用虚线表示。最后，路径评分有助于找到生产分子的最佳策略，并对已完成的路径进行排序。目前，人工智能技术在逆向（生物）合成过程的 每个阶段都发挥着至关重要的作用。A*，A*搜索；CNN，卷积神经网络；GNN，图神经网络；MCTS，蒙特卡罗树搜索；RL，强化学习。

即使人工智能最初是为有机化学的合成规划而开发的，也已在多个应用领域（化学、生物催化和合成生物学）得到应用。在接下来的章节中，我们将探讨为单步和多步过程量身定制的人工智能方法和模型（在注释 S2 的术语表中进行了解释）的多样性，包括所使用的数据和预测因子的类型，以及所使用的数据集和评估指标。然后，我们回顾了流行的数据库和数据集准备工作。最后，我们评估了人工智能方法的局限性，并强调了不同应用领域的区别。我们的综述还包括

B

旨在找出知识差距，强调需要进一步研究的领域，以推动人工智能在逆生物合成中的应用。

单步逆合成

随着人工智能及其应用的不断发展，人们提出了预测化学反应结果的计算方法。化学反应涉及将一组化学物质转化为另一组化学物质，从而导致化学变化。目前存在两个关键挑战：预测给定反应物（即底物）产生的产物，以及解决已知产物时识别反应物的反向问题[6]其中一种方法是基于模板的方法，该方法从反应化学数据库中提取信息，通过反应模板推广现有反应的应用[7]模板对应于描述生成物与其反应物之间连接性变化的子图模式。第二种方法是无模板方法，它利用生成模型的能力来预测目标分子。[8]最后，基于半模板的方法旨在调和专用规则的使用和通过人工智能进行归纳的能力。[9]与单步逆合成相关的一般原理如图 2 所示。

分子和反应表示法。利用要将分子和反应作为人工智能模型的输入，必须采用一种表示方法或矢量化方法。目前广泛采用的方法是使用 SMILES 和 SMARTS 符号将分子和反应模板编码成字符串，用于预测产物底物的模型。[10,11]将分子表示为字符串有助于使用自然语言处理工具，如使用文本序列作为输入的转换器和生成模型。分子指纹是将分子表示成向量的一种方法。其中，圆形分子指纹捕捉原子周围的局部特征，如拓扑结构、原子类型、键类型和指定半径内的连接模式。这种表示方法主要用于预测与反应相关的特性，如反应模板[7] 或分子相似性。[12]分子具有天然的图结构，将原子视为节点，将键视为边。因此，分子图也被广泛用作图神经网络等模型的输入。[11,13]其他不太常见的表示方法包括 SELFIES、[14] 分子特征、[15] 和原子环境[16]，这些方法侧重于局部信息来描述分子特征。表 1 和图 3A 全面概述了各种分子表征，并强调了它们在单步逆合成中的使用比例。调查[17,18]有关不同类型的分子表征、各自的优势和局限性的更多详情，请参见相关调查。

基于模板。 基于模板的单步法这些模板可以由人类专家衍生，也可以以反应物和主要产物的形式从反应数据库中自动提取。模板反应有时被称为反应规则或通用反应，通常由原子映射 SMARTS 字符串表示，可以处理立体化学。[21]化学中使用模板的例子见 Szymkuć 等人的[61]，生物催化和合成生物学中使用模板的例子见 Finnigan 等人的[62] 和 RetroRules 数据库。[63]基于模板的方法选择可应用于给定产品的模板。然后，主要挑战是将最佳模板应用于产物以获得反应物。

https://doi.org/10.1021/acssynbio.4c00091

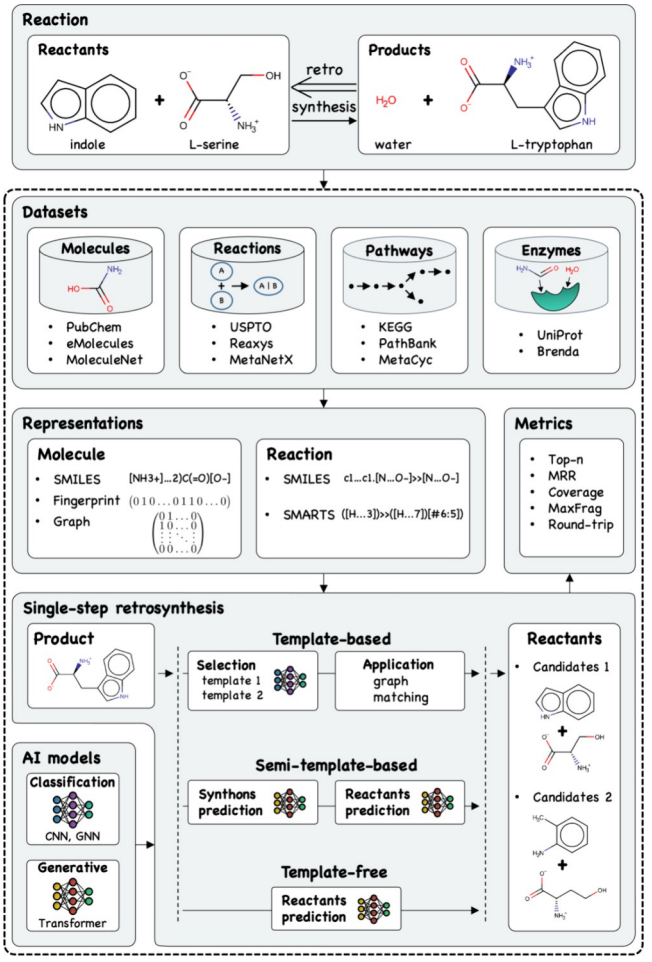


图 2.以 L-丝氨酸水解酶反应（吲哚 + L-丝氨酸 → 水 + L-色氨酸）为例，说明单步逆合成的一般原理。要实施和训练单步逆合成模型，需要分子、反应、路径和酶的数据库。然后将分子表征作为逆合成模型的输入。然后，选择基于模板、半模板或无模板的方法，并使用人工智能模型进行逆合成。最后，单步模型评估部分详细介绍了 衡量框架性能的 指标。CNN：卷积神经网络；GNN：图神经网络；MRR：平均倒数等级。

神经网络（NN）用于从分子官能团或指纹中学习模式，为产物或底物选择属于同类反应规则的模板[54]这种策略被应用于反应预测和逆合成任务，使用深度高速公路[26]和 Hopfield 网络。[47]神经网络的输入是产品的 SMILES 符号、分子指纹或两者。[29]另一种策略是基于图神经网络（GNN），利用分子的图结构选择反应模板，通过模型提供可解释性。[55,64]为此，反应模板被嵌入到使用 GNN 构建的条件图形模型中[56]，或使用反应中心进行部分编码。[57]

基于模板的方法具有模仿现有反应的化学键重排的优点。然而，这些模型存在一些局限性，例如数据库中描述相同化学变化的模板存在重复和重叠。一些研究通过将模板规范化[11]或使用专用的 NN 模型来优化和缓解这些局限性。[7]除其他局限性外，基于模板的方法只能推断模板数据库中的反应，而不能提出新的机理。为了缓解这种不便，Yan 等人[36]采用了模板合成策略来创建新的反应和模板。

在逆生物合成中，观点略有不同。分子可能比化学、[29]、反应涉及高度特异性酶的逆合成通常考虑的分子更复杂。为了应对这种情况，在化学[12]和逆向生物合成中选择模板时，已经提出了一些策略，如识别环系统中的断开[52]，或评估产物和底物之间的相似性，以保留酶的活性。[40,43]据我们所知，人工智能模型还没有与合成生物学中基于模板的单步方法相结合。

无模板。无模板方法避免依赖反应模板，而是使用生成模型直接预测反应物。为了完成这项任务，需要建立各种模型，如使用长短期记忆单元建立的编码器-解码器模型。[33,49]和 Weisfeiler-Lehman 网络。[13]最近，变压器等深度生成模型得到了广泛应用和改进。变换器模型是一种由编码器和解码器组成的 NN 架构，特别适用于自然语言处理。编码器处理代表单个或多个字符的标记序列，生成一 组 隐藏表征，解码器对其进行解码，生成输出序列。该模型包含一种注意力机制，用于关注序列中信 息 量 最大的部分。因此，从 SMILES 产品到 SMILES 反应物的翻译任务是利用为前向预测而建立的模型进行的[31,65]并将反应预测模型整合到逆合成过程中。[66]

表 1.单步逆合成中使用的分子表征列表

代表作品		说明相关作品
SMILES（规范、有根、……）分子的		字符串表示法 8、10 和 19-46
圆形指纹（ECFP、FCFP、HSFP. ..）	通过散列局部分子特征实现分子矢量化	7、11、12、25、26、29、34、43 和 46-56

https://doi.org/10.1021/acssynbio.4c00091
ACS Synth.Biol.xxxx, xxx, xxx-xxx

分子图以原子为节点、键为边的分子图表示法		8, 9, 11, 13, 19, 24, 28, 3 48 和 55-59
原子	环境 SMARTS 中分子的环形原子中心拓扑邻域片段格式	46 和 50
Signaturedescriptor	用 SMILES 格式表示分子，编码其分子图中直到预定半径的所有原子环境	60
分子的	扩展 SMILES 字符串表示法	46

https://doi.org/10.1021/acssynbio.4c00091
ACS Synth.Biol.xxxx, xxx, xxx-xxx

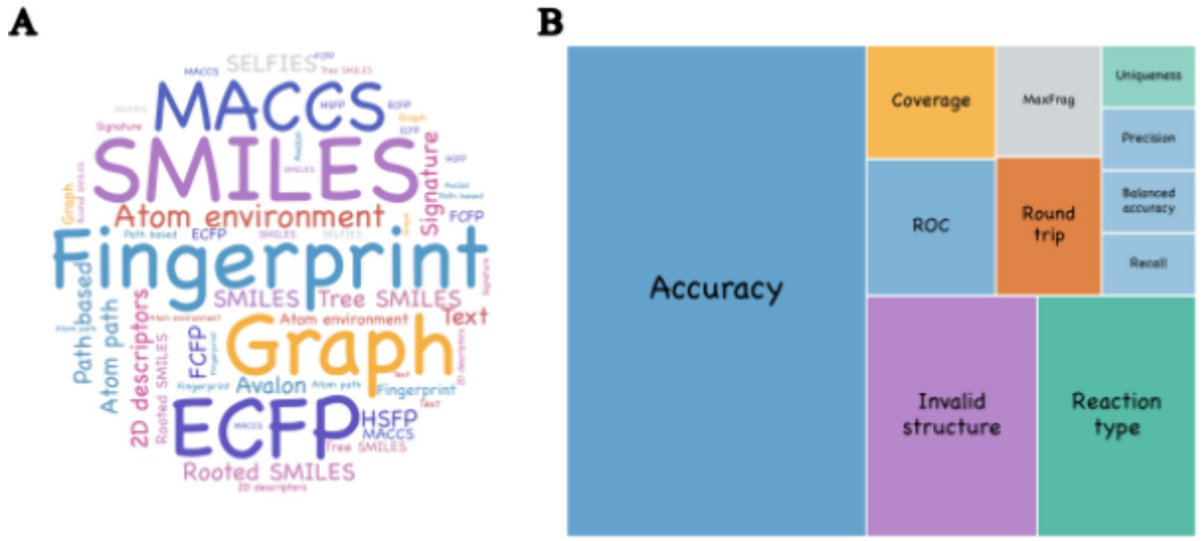


图 3.单步算法对分子表征和度量的使用分析。(A) 单步人工智能模型中用作输入的分子表征类型词云。(B) 用于评估单步逆向（-生物）合成模型的不同指标的比例。这些指标在单步模型评估部分有进一步说明。文章中仅提及一次的指标不在分析之列。蓝色方格和绿色方格分别反映了观察误差量化指标和反应类型多样性指标的使用情况。MaxFrag：最大片段；ROC：接收者操作特征。

转换器模型不使用 SMILES 表示法作为输入，而是考虑了其他类型的分子表示法，并将它们结合起来。例如，Graph2SMILES、[67] BiG2S、[68] DVMP、[28] graph enhanced trans-former、[24] graph truncated attention、[41] 和 G2GT[59] 等模型将分子的图表示与 SMILES 序列表示相结合。变换器模型 retroformer[8] 将局部和全局注意力头结合起来，以识别分子中的反应区域。另一种可能性是利用 SMILES 的树形表示法定制变换器模型，[35]，使用一组预定义的化合物，[58]，或利用副产物[39] 或反应图为反应添加信息。[69]变压器架构还可用于预测反应物的原子环境，了解生成物[50]，并将反应物转化回反应物。[46]

在重建分子表示法的过程中，转换器模型容易添加或省略字符，从而在 SMILES 符号中产生语法错误。为了减少这种不便，我们在 转换器架构中插入了基于神经网络的语法检查器 SCROP[30] 框架，从而减少了无效输出的数量。为了增加预测的多样性，避免无效的 SMILES，[20]，并预测反应的类型，[42]，一个潜在变量已被整合到模型中。此外，还采用了不同的策略来增加预测的多样性。在 SMILES 字符串中添加反应类型可为转换器模型提供额外的上下文，从而增加建议的多样性。[45]为了在不改变数据集的情况下实现这一目标，我们采用了 GNN[19] 或基于能量的模型来增强预测效果。[70,71]对单步模型认为可信度较低的预测结果进行重排。重排方法利用分子图中的额外化学特征信息来完善预测结果。已对一些方法进行了测试，以提高学习效果模型的能力，如数据增强和迁移学习。数据增强通过对原始数据 SMILES 进行各种转换来丰富数据集。这些

[38] 或更复杂，如 SMILES 枚举法，该方法随机选择一个起始原子，为同一分子生成不同的 SMILES。[27,44]另一方面，迁移学习是一种将预训练模型作为新任务起点的技术，以利用预训练模型学到的知识和特征。在由 380 K 个分子组成的数据集上进行学习后，对 USPTO-50K[25,37]和约 2200 个 Baeyer-Villiger 反应的数据集上学习，[23] 比单独在数据集上学习表现更佳。相比之下，在文本和分子表征上训练的多任务转换器无需进行迁移学习就能显示出令人感兴趣的结果。[72]

据我们所知，无模板方法尚未用于合成生物学领域，而用于生物催化的方法则寥寥无几。BioNavi-NP[73] 采用了转化器和代表天然产物的数据集来预测生物合成途径。在反应预测方面，Kreutter 等人[74] 在 SMILES 中添加了酶的文字表述，而在逆向生物合成方面，Probst 等人[75] 使用 EC 编号来丰富 SMILES，从而预测分子和酶，以及分别与反应相关的 EC 编号。

基于半模板。基于半模板的方法旨在以模仿化学家的推理。首先，检测产物结构中的反应位点，然后断开化学键，生成称为合成子的中间分子。然后从合成子中提取底物。与基于模板的方法不同，这种方法不依赖于预定义化学反应模板数据库。

生成合成子后，可使用生成模型预测反应物，如 G2Gs 中的图到图转换模型、[76]、RetroXpert、[77]、RetroPrime、[32] 和 RetroSub、[34] 等转换模型，或使用 GraphRetro 根据预先计算的词汇进行预测。[78]另一种可能性是优化分子输入；例如，热

https://doi.org/10.1021/acssynbio.4c00091

表 2.单步逆生物合成法

单步	框架	说明	聚焦酶立体化学信息		数据集
基于模板	EHreact[40]	模板存储在 Hasse 图表中并进行排序根据它们的相似性	是名称	，欧共体编号	布伦达、RetroRules 和雷亚
	retrosim_enz[43]	基于相似性和进化评分的模板排序	是	名称	瑞亚
	RingBreaker[52]	基于合成环系统的模板排名	有	否	Reaxys 和 USPTO
无模板生物纳米技术[73]		根据有机和生物合成技术培训的变压器反应	是的	否	MetaNetX 和美国专利商标局
	Kreutter et al.[74]	使用酶名称丰富 SMILES。变压器（正向预测）	是		NameReaxys 和美国专利商标局
	Probst et al.[75]	使用转换器用欧共体编号丰富 SMILES		YesEC numberBrenda	, MetaNetX、PathBank 和 Rhea
富 SMILES。	基于 Semitemplate-basedThakkar et al.[80]	使用提示符--用 EC 编号丰		NoEC	编号开心果和美国专利商标局
		基于方法			

Hasic 等人利用点指纹（HSFP）[53]，以确定反应位点并生成合成物。此外，RetroExplainer[79] 和 Graph2Edits[9] 框架依赖于一系列操作，如删除键或连接原子团以检索反应物。

有趣的是，在生物催化方面已经开发出了一种半模板方法，该方法使用基于提示的范式，可在输入的 SMILES 中加入额外信息，如反应的 EC 编号。[80]据我们所知，半模板方法尚未用于合成生物学领域。

单步模型评估。 比较

因此，必须制定衡量这些方法有效性的标准。使用衡量标准有助于进行公平、一致的比较，有助于确定每种模式的优缺点。图 3B 显示了社区采用不同衡量标准的情况。

许多用于单步预测的模型都提出了多个候选物，其中"候选物"指的是一系列反应模板或一些被设想为可靠反应物的分子集。准确度指标表示预测值与真实值的接近程度，通常使用 top-n 准确度进行评估。虽然 top-n 指标被广泛用于反应预测，但其相关性也受到质疑，[22]，因为许多分子可以由一组以上的反应物构建而成，也就是说，一个反应物有多个"真实"答案。特定产品。在这种情况下，不太常见的指标包括分数精度，[35] 平衡精度、[26,55]加权预

厘定率、[54] 和 ROC 曲线。[38,43]

指标也被用来评估排名的质量。平均倒数排名（MRR）计算多次运行中第一个相关预测的平均排名。[54]另一种称为覆盖率的指标则是计算有一个或多个有效预测的样本数量。[20,22,42,45]评估属于多种反应类型的反应物以获得广泛的可用合成路线至关重要。类多样性指标衡量的是单步模型预测的反应类型的范围、[22,32,45]而詹森-香农发散度（Jensen- Shannon divergence）则量化了属于

[22] 但事实证明，这种转换器会产生偏差。[81]由于无模板算法的性质，有些预测在语法上是无效的。超过三分之一使用无模板方法的研究论文量化了无效SMILES的出现（图3B）。最后，针对基于半模板的方法，我们采用了一种独创的方法来衡量断开连接的成功率。[80]

在复式（-生物）合成中，一个或多个反应物参与反应，最大的分子更能反映反应类型，也更容易出现重要的反应位点，避免出现不明确的反应。为此，对预测结果采用了最大片段（MaxFrag）精度[39,44]以考虑分子之间的相似性，从而反映其生物活性。[35]

逆生物合成的前景。尽管许多

目前已经开发出了许多逆合成方法，但只有少数方法是专门为逆生物合成设计的。表 2 总结了这些方法。事实上，生物催化反应的先决条件之一至少是确定该反应是否可由酶催化。[80]此外，酶的稳定性是在较窄的温度、pH 值和压力值范围内实现的。因此，准确描述这些反应的特征至关重要。

[82]

反应 的元素，以确保反应的可行性。

某类反应的预测反应物的相似性分布。固定数量的反应类型。[22]此外，而不是

溶剂在提高反应产量和使化学反应更具可持续性方面也起着至关重要的作用。考虑到所有这些因素，可以为逆向生物合成开发一个专门的评分标准，以比较算法提出的建议。在选择逆向生物合成算法之前，用户必须确定所获得的结果是高评估丰富性，存在重复预测、

Kim等人在[20]和Yan等人在中估算了这一数值，它表明模型缺乏多样性。[39]

还有一些专门为一步逆合成开发的临时指标。往返准确性反映了逆向合成建议的有效性。

度探索性的、用于理论思考的，还是用于实际应用的。事实上，如果用户对预测假定基质的新反应持开放态度，那么使用生成模型是合适的。[31]相反，使用SELFIES可以生成数据集中没有的反应，同时确保[46]

然而，在*体内*在实施反应时，建议利用已知的反应机理，因此最好采用基于模板的方法。我们相信，在逆合成中开发的众多方法可以为逆生物合成提供灵感。

https://doi.org/10.1021/acssynbio.4c00091

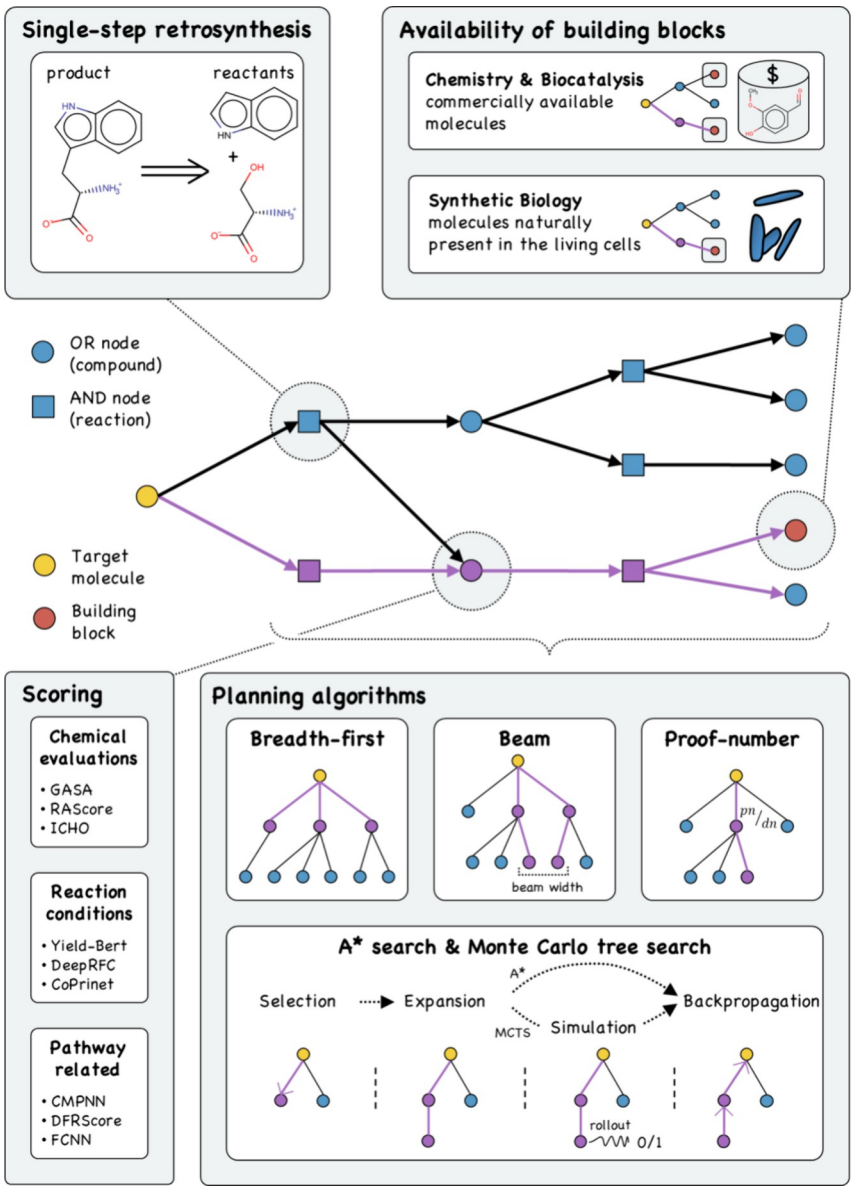


图 4：多步骤逆合成的一般原理多步逆合成的一般原理以 AND/OR 树的形式表示，其中圆形代表带有 OR 节点的分子，因为多个反应可以合成该产物；方形代表带有 AND 节点的反应，表示所有反应物都是生成该产物所必需的。从目标分子（黄色 •）开始，在逆合成的每一步都使用单步逆合成，以获得构筑基块（红色 •）。在化学和生物催化中，构件是市场上可以买到的分子，而在合成生物学中，构件是活细胞中存在的分子。由于可能合成路线的搜索空间巨大，因此在逆向合成规划过程中，将使用下一节中解释的评分函数和搜索算法对合成可能性进行导航。为简单起见，规划算法用图表而不是 AND/OR 树来解释。

多步逆合成

科里的开创性著作中介绍的多步逆合成法，是通过逆转已知的化学反应，将目标分子分解为更简单的前代分子，从而合成复杂分 子 的战略框架。这一过程涉及多个步骤，步骤数量因分子的复杂程度、可用反应和起始材料而异。最终目的是利用这些构件开发出高效的合成路线，并利用现有反应进行实际操作。这一概念最初侧重于简化分子，现在已发展到可 以 合成可能比目标分子更复杂的中间体，这在涉及由辅助因子增强的复杂结构的天然生物合成过程中可见一斑。

多步逆合成的原理现在也适用于生物催化和合成生物学，利用生物数据库指导反应选择。

虽然建议可行的单步逆合成途径的算法提供了重要基础，但预测多步逆合成途径却带来了巨大挑战。这种复杂性源于潜在合成途径的巨大潜力以及确定 "好 "合成途径的主观性。

化学家和生物化学家都在努力解决这些复杂的问题，他们面 临 着 广泛的潜在中间体，以及对什么是最佳逆合成途径的不同看法。全面的多步骤

https://doi.org/10.1021/acssynbio.4c00091

图 4 展示了逆合成搜索的规划算法、起始材料选择和单步反应预测器。

构件的可用性。可用构件集

可用作起始材料的 构件（有时称为前体或汇体）是影响算法预测合成路线能力的关键因素。事实上，直观地说，更多样、更广泛的构建模块将为算法提供更广阔的"着 陆 点 "，以结束逆合成探索。

在化学和生物催化领域，构件通常由商业上可获得的化学品组成，如 Probst 等人的研究，[75]，索引来自 eMolecules 等在线门户网站或 Sigma-Aldrich 目录等化学品提供商。其他可用信息，如化学品价格，最终也会被考 虑 在内，用于指导逆合成规划并对其 进行排序，如 Zhang 等人的研究。[83]

生物体内生物合成途径的规划对选择可利用的构件提出了挑战，因为分子进入细胞具有高度选择性。构件更为具体，通常包括生物体内天然存在的分子集，如 Koch 等人的研究，[84]，其中可用的前体是从*大肠杆菌*的基因组尺度代谢模型中提取的。

规划和搜索算法。预测潜在

如图 4 所示，合成路线在很大程度上依赖于在化学空间内浏览各种可能性的搜索算法。这些算法对于有效绘制多步骤途径图至关重要。这些搜索算法大致分为两类：无信息搜索和有信息搜索。无信息搜索，如深度优先搜索和广度优先搜索，在运行时不依赖额外信息来引导探索求解空间的特定区域。相反，有信息搜索包含启发式功能，可评估要扩展的化合物 "有多好"。这些启发式函数可能只关注迄今为止发现的化合物，如本综述中提到的波束搜索。或者，它们也会使用蒙特卡洛树搜索（MCTS）中的滚动模拟或值函数估算器等方法来估算与解决方案的接近程度，如 A* 相关算法。总之，这些启发式方法可以指导搜索过程，确保更高效的探索。虽然有启发的搜索方法不能保证获得最优解，但它们能显著提高在合理时间内找到好解决方案的 可能性，在解决方案质量和搜索效率之间取得平衡。从本质上讲，这些算法在优化逆合成规划方面发挥着关键作用，它们能在复杂的可能性中进行导航，从而得出可行的合成路线。逆合成图一般用 AND/OR 树表示，其中 OR 节点对应一个分子，因为合成该产物可能需要多个反应；而 AND 节点对应一个反应，因为合成该产物需要所有反应物。这个图也可以看作是一个超图，其中一条边可以连接多个节点，从而代表了产物和底物之间的联系[22]现在我们简要介绍一下最近用于多步逆向（生物）合成的主要搜索算法类型。

广度优先搜索广度优先搜索算法是图遍历算法的典型案例。它通过访问当前深度的所有相邻节点来探索图，然后再访问下一个深度级别的节点。这

这个过程一直持续到所有可到达的节点都被访问过为止，或者在实用的逆合成环境中，一直持续到 达到一定的深度为止。这种算法虽然速度很慢，但对于寻找短路径非常有用。在逆向（生物）合成领域，通常在每一级扩展之间应用过滤器，目的是排除那些不太可能参与可信解决方案的节点，从而限制组合性。广度优先搜索已在化学中用于预测热力学上可行的路径，如分析工具 RetroSynX。[85]在这里，不可能发生的反应会被过滤掉，然后再进入下一个深度层次。同样，在生物催化领域，Liu 等人[86] 采用热力学估计作为过滤标准，以减轻广度优先搜索过程中的组合爆炸。

光束搜索光束搜索是当今使用的一种算法在许多人工智能生成方法中都存在。它以广度优先的方式构建可能路线的集合，但会对每个深度级别要扩展的节点数量施加一个预定义的限制，称为波束宽度，从启发式评估中选择保留的节点。波束搜索通常被视为广度优先算法的增强版，它能提高求解效率，尤其是在大型求 解空间中。这种算法已被用于化学领域，如 Schwaller 等人的研究[22]，根据合成复杂度得分、SCScore 和生成单步对数概率，选择有前途的化学物质进行进一步扩展。除了这些指标外，Kreutter 等人 的研究[87] 还在启发式评估中加入了路线惩罚得分 RPScore。在逆向生物合成中，酶的分类和可用性是关键因素。例如，Probst 等人在[75] 中依靠考虑 EC 编号注释和 SCScore 的分数来选择要扩展的化学物质。相反，在合成生物学中，RetroPath2.0[60] 软件整合了波束搜索和 RetroRules 分数，优先选择与酶可用性相关的高可信度反应物。

深度优先证明数搜索（DFPN）。证明数搜索算法是一种树搜索方法，主要用于博弈树求解。它通过为博弈树中的节点分配证明数（表示赢的数值）和反证数（表示输的数值）来评估博弈位置。DFPN 是一种变体，它在更新这些数字的同时进行深度优先搜索，目的是在特定深度限制内证明或反证强制胜局的存在。它根据这些证明数和反证数修剪搜索树的分支，以专注于最有希望的下棋思路。这种搜索算法在 DFPN-E 中得到了化学应用，[88]，作者将 DFPN 与启发式边缘初始化方法相结合，以解决 OR 与 AND 棋步数量之间的 不平衡问题，后来又进行了改进，以输出多个解决方案。[89]CompRet 框架[90] 提出了一种综合性工具，利用 SCScore 得出的指标推荐最有前途的合成路线，对合成化合物的可能路线进行枚举和排序。

蒙特卡洛树搜索（MCTS）。蒙特卡洛树搜索是一种知情的启发式搜索算法，常用于决策过程。它通过反复模拟随机移动序列来建立搜索树，在逆合成中，移动序列是从树的选定叶子（如化合物）开始的单步变换。该算法通过平衡利用（侧重于最有希望的节点）和探索（侧重于其他潜在路线），优先探索有希望的路径，从而做出明智的决策。

https://doi.org/10.1021/acssynbio.4c00091

在过去几年中，这种算法已被广泛应用于化学领域的逆合成路线规划。[66,91−97]Segler 等人在[91] 上开创了用于逆合成的 MCTS，将基于规则的单步变换和三个 NNs 结合起来，帮助探索化学空间。在这项研究中，人工智能被显著地用于预选要应用的模板，从而限制了组合爆炸，同时将搜索引向最合理的路线。在 MCTS 中结合 NNs 和基于模板的方法的其他几项研究中，如 ASKCOS[98] 和 AiZynthFinder，[96] 以及 Zhang 等人的研究，[92]，作者建议将五个 GNNs 有效地结合起来。[92]建立模板的数据源是一个重要因素，Thakkar 等人在[94] 中强调了这一点，并研究了四个数据集（AiZynthFinder、Pistachio、Reaxys 和 USPTO）对 MCTS 性能的影响。此外，模型的配置对路由发现的成功与否也有很大影响[99]模板反应模型的转换数量有限，可能缺乏穷尽性。作为基于模板的形式主义的替代方案，无模板单步方法也被应用于 MCTS。例如，Lin 等人在 AutoSynRoute 中使用 Transformers 实现了单步反应[95]属于生物催化领域的研究要少得多。不过，已经有人提出重新使用 ASKCOS，以利用来自化学和代谢反应数据库的基于模板的转化。[100,101]据我们所知，RetroPath RL[84] 是 MCTS 算法在合成生物学领域的唯一实现，其中基于规则的转化是从代谢数据库中提取的，而可用的构建模块列表则是从基因组规模的代谢模型中提取的。

A* 搜索A* 搜索是一种知情图遍历方法用于图或搜索问题中的寻路和优化的算法。它通过考虑迄今为止的成本（即从起始节点出发的历史成本）和估计的未来成本（即到达目标节点的未来成本）来确定探索节点的优先级。通过利用这个评估函数（有时被称为价值函数），A* 提出了从起点到目标节点的最优路径，同时使总成本最小化。与 MCTS 相比，这种算法没有展开阶段，因此不依赖随机性，速度也更快。这种搜索算法在过去几年中越来越受欢迎，并被用于一些化学规划软件，如 ASICS、[102] Retro*,[103] RetroGraph,[104] GNN-Retro[105] 用于化学应用，BioNavi-NP[73] 用于生物合成途径预测。虽然 A* 算法是指导路径发现的多步引擎，但它可以与不同类型的单步移动相结合，例如 Retro* 中基于模板的转换，其中一个 NN 根据产品分子选择要应用的模板；或 BioNavi-NP 中的无模板移动，它依赖于一个转换器来预测给定产品的反应物。A* 搜索是最佳优先搜索算法的一种变体。人们还开发了其他基于非数据驱动的人工智能方法，如贪婪的最佳优先搜索，这种方法仅根据迄今为止产生的成本（历史成本）确定探索的优先次序，而不试图预测未来到达目标节点的成本。其中一个例子是 SynRoute，作者证明这种贪婪方法在测试的四种规划算法中最为有效[106]另一个典型的例子是 Synthia[107]（前身为 Chematica），这是一款著名的合成规划商业软件。该专家系统采用了全面的 Synthia 的探索策略是一套用于选择和应用反应模板的规则，以及用于为下一次逆合成迭代选择化学物质的双重评分功能。虽然文献中没有这样的描述，但 Synthia 的探索策略可以归类为最佳优先搜索。[61]

其他强化学习相关搜索。与 MCTS 和 A* 搜索算法类似，文献中还有其他类型的强化学习方法。通过迭代探索和学习，由 NN[108] 驱动的计算机代理会完善其决策过程，有效地浏览可能的反应图，从而找出最有效的逆合成途径。这种方法已用于化学[109,110]和生物催化中。[83]

表 3 总结了用于多步逆合成的各种搜索算法，以及它们各自在化学、生物催化和合成生物学中的应用范围。在逆向生物合成中，必须承认有一些基础方法的存在，如 novoStoic、[111] XTMS、[112] 或 BNICE.ch。[113]这些方法提供了独特的代谢途径设计方法，与本综述中讨论的基于人工智能的算法有所不同。NovoStoic 在化学计量建模框架内利用基于模板的反应形式化，并结合混合整数线性规划（MILP）来高效地列举代谢途径。XTMS 利用基于模板的反应构建逆合成网络，并命名为*反应特征*。然后以该网络为起点，提取连接预设已知化合物与*大肠杆菌*底盘生物体的所有可能生物合成途径。与 XTMS 类似，BNICE.ch 也使用基于模板的反应来构建一个完整的生化网络（称为 ATLAS），[114]，然后对该网络进行检查，以详尽无遗的方式列举出线性路径（即无分支路径）。不过，在这两种方法中，预建网络都限制了用户探索全新的化合物，从而阻碍了可推广性。读者若想了解这些技术的更多详情，可以在其他地方找到相关评论。[2,115]

逆向生物合成的前景。各种算法但没有一种算法在结果输出方面表现出优越性。根据不同的应用，必须注意 MCTS[99] 或 A* 等算法的模型参数化，并确定评估方法。[73] 此外，还可以减少数据量。

分子的生产成本[117]是寻找新合成途径的一个目标。因此，在寻求创新合成途径的过程中，构件成本以及更广泛意义上的原子经济性成为重要标准。[83]在这种情况下，使用不依赖辅助因子的酶特别有吸引力。对于*体外*应用，虽然添加辅助因子可以调节催化活性，但这是必须考虑的成本因素。为了降低成本，利用光敏化、电化学活化、[118]或创建酶级联的辅助因子循环是很有前途的策略。据我们所知，逆生物合成工具目前还不支持生成辅助因子循环途径。就*体内*应用而言，虽然在细胞宿主中实施分子级联可以使用如果分离出来就不稳定的化合物，[119]，但使用辅助因子会导致细胞生长与所需化学物质的生产之间的竞争，而逆向生物合成工具并不考虑这一点。

https://doi.org/10.1021/acssynbio.4c00091

表 3.多步逆合成中常用的搜索算法[a]

多步骤	适用范围	框架	重要功能/亮点	构件来源	单步	代码可用性
广度优先	化学	RetroSynX[85]	利用热力学估算过滤中间化学品	aladdin-e.com	模板根据	没有
	生物催化	Liu et al.[86]	利用热力学估算过滤中间化学品	aladdin-e.com	模板根据	没有
光束	化学	Schwaller et al.[22]	根据对数概率和单步生成法，选择最有前途的化学品进行进一步扩展。	电子分子	模板免费	没有
		Kreutter 等人[87]	利用 RPScoreEnamine	、Molport 进一步扩展了 Schwaller 等人的光束选择[22]		是
	生物催化	Probst 等人[75]	考虑到酶分类和 SCScoreeMolecules 的光束选择	无模板		是
	合成生物学	RetroPath2.0[60]	基于酶可用性的光束选择估计	基于模板的代谢模型		是
DF 证明编号	化学	CompRet[90]	根据构件达到的证明数和反证数	烯胺	模板根据	是
		DFPN-E[88]	使用估算器评估寻找证明+实现构件的难度	美国专利商标局		没有
MCTS		ChemistryGao et al.[97]	用于模板选择和反应筛选的	两个 NNNA	基于模板	是
		Segler et al.[91]	3N-MCTS 方法：三个 NN 用于模板选择（扩展）、可行性和推出AlfaAesar	, Acros, Reaxys、Sigma-Aldrich, ZINC		否
		AiZynthFinder[96]	NN 引导模板选择	ZINC		是
		ASKCOS[98]	用于模板选择和反应过滤的两个 NN	,Sigma-AldrichYes Wang 等[93]		使
		用强化学习网络代替 MCTS 推出步骤eMolecules		, Sigma-Aldrich		无
		Zhang 等人[92]	用于选择模板、推断反应溶剂和催化剂、过滤反应和有效评估推出步骤的五个 GNNs	莫尔波特		是
	化学	AutoSynRoute[95]	基于变压器输出对数概率的启发式指导推出Sigma-Aldrich	、USPTO 和锌	模板 - 免费	是
	生物南等人	催化桑卡纳拉亚 al.[100]	用于模板选择和反应过滤的两个 NN 应用于	、LabNetwork、Sigma-Aldrich	模板基础	无
		Levin 等人[101]	ASKCOS 的重新实施，使用 NN 对模板进行优先排序，并在化学模板与酶模板之间进行平衡	eMolecules, Sigma-Aldrich		是
	合成生物学	RetroPath RL[84]	根据反应可行性和酶置信度	评分选择模板代谢模型	模板根据	是
A*	化学	Retro*[103]	使用 NN 进行模板选择，根据当前反应的成本估算当前路径的成本，根据当前反应的成本估算当前路径的成本，根据当前反应的成本估算当前路径的成本，根据当前反应的成本估算当前路径的成本。从知识数据库中训练出的 NN 所学到的未来路径	电子分子	模板根据	是
		Retro*+[116]	在搜索过程中将模板选择与实际已预测的反应结合起来，使用自我改进学习法	电子分子		无
		RetroGraph[104]	使用离线训练的 GNN 估算未来路线的成本，使用图搜索而不是树搜索进行多目标搜索	电子分子		无

		GNN-Retro[105]	使用离线训练的 GNN 估算未来路线的成本	不		无
		ASICS[102]	将从知识数据库中提取的已知反应与基于模板的预测相结合，利用 SAScore 估算实现目标的路线成本	电子分子		是
	生物催化	BioNavi-NP[73]	结合化学和生化数据集，利用迁移学习建立单步变压器，用 NN 估算未来成本	自定义列表，天然产品的主要前体	模板 - 免费	是
最佳第一	化学	SynRoute[106]	NN 用于评估预测转化的可行性得分	电子分子	模板	否
	化学	Synthia[61]	化学和反应函数得分相结合，用于选择下一个图形扩展Sigma-Aldrich	模板	根据	否
	生物催化	RetroBioCat[62]	SCScore 用于指导最佳优先搜索	eMolecules,Molport, ZINC	Template- based 根据	是

https://doi.org/10.1021/acssynbio.4c00091
ACS Synth.Biol.xxxx, xxx, xxx-xxx

表 3

多步骤	应用范围	框架	重要功能/亮点	构件来源	单步	代码可用性
化学习	化学	Schreck 等人[109]	利用模拟经验训练 NN，改进节点选择（策略	eMolecules, Sigma-Aldrich 、	基于模板	是
			GRASP 采用类似于 MCTS 的方法，即用强化学习代理取代虚构的在线推出系统	实验室网络	模板 - 免费	没有
	生物催化	GRASP[110]	对 "决策者 "NN 进行训练，使其偏向于探索，并通过随机成分促进对不同路线的探索	电子分子	已知反应	有
		Zhang et al.[83]		化学空间		

A，不适用。

计分功能

选择最有希望合 成 目标分子的反应和途径是指导逆合成规划过程的关键要素。表 4 中列出的评分功能有助于在回溯合成规划和路线枚举过程中浏览遇到的多种合成可能性。这些功能依 赖 于 不同的标准，包括化学成本、结构因素和酶知识等。

合成无障碍分数。合成无障碍分数可以区分可行和不可行的分子，是一种辅助性的逆合成规划工具，可以从不切实际的分子中识别出可行的合 成 路线。[120]基于图注意力的合成可得性评估（GASA）分数通过将化合物标注为 "易合成 "或 "难合成 "来评估小分子的合成可得性[48]与单纯依赖分子结构不同，RAScore[121] 评估合成的可行性，将反应信息纳入评估，并采用类似的策略预测在数据库中找到参与反应的分子的概率[51]此外，为了替代合成的可及性，在化学、[122] 、 药物相关、[123] 和生物[124] 应用中估算了生产一个化合物所需的步骤数，或将其整合到一个综合分数中。[125] 此外，这些分数还对化合物的相关数据进行估算，包括其价格[126] 或其热力学性质、[127] 、反应、产量、[128] 或其可行性。[129−132]

路线排名。无论采用哪种特定的多步骤算法，通常都会产生大量的逆合成路线，这就需要制定战略来确定最有前途的 路 线 。为 此 ，人 们 设计了路线排名策略，整合了与化学品（如 SAScore、[133] SCScore、[134] 或构 件成本）、反应（如 RAScore、[121] 反应产率和热力学或酶可用性）以及路线整体适当性（如路线多样化、反应步骤数量或理论生产通量）相关的 判别标准。例如，SynRoute[106] 框架根据构建模块的长度、成本和反应产率估算来评估路线，从而选择最佳途径。在生物催化方面，RetroBioCat 优先考虑的是
通过考虑反应步骤的数量、化学复杂性的 变化、市场上可买到的化学物质的比例以及步骤与文献参考文献的联系等因素，对路径进行评分。有趣的是，还加入了多样性评分，以惩罚利用已被列入顶级途径的反应的途径[62]同时，在合成生物学领域，Galaxy- SynBioCAD 平台结合了多种标准，包括酶的可用性、理论产物通量（通 过 通 量 平 衡 分 析 [135] ） 、 反 应 热 力 学 （通过 eQuilibrator[136]）和步骤数，以训练一个分类器模型，进行途径评分和排名。[137]其他值得注意的工作还包括 "途径多样化 "评分，该评分可根据类似化学品的潜在生产

情况对合成途径进行比较[138]，以及 RPScore，[87]，该评分通过步 数和分子合成可及性对途径进行评估。更广泛地说，预测吉布斯自由能（ΔG）的开发工具，如 eQuilibrator[136] 和 dGPredictor，[139]，都是筛选单步预测和评估整体途径的热力学可行性的重要工具。

https://doi.org/10.1021/acssynbio.4c00091

表 4.用于评估路线的分数一览表

分类	目的	适用范围	得分
评估分子	结构可用性	化学	GASA[48] 和 SCScore[134]
	反应可用性	化学	RAscore[121] 和 ICHO[51]
反应条件	连续反应器实验	化学	InFlow[129]
	液-液萃取	化学	提取分数[132]
	反应产量	化学	Yield-Bert[128]
	酶的可用性	生物催化	DeepRFC[145] 和 EHReact[40]
	复合价格	化学	CoPriNet[126]
	热力学	多学科	eQuilibrator,[136] dGPredictor,[139] and GC-NORM-based[127]
路径相关	预测步数	化学	CMPNN[122]
		药物相关应用	RetroGNN[152] 和 DFRScore[123]
		合成生物学	FCNN[124]

酶搜索。从合成规划过渡到生物催化和合成生物学实施，需要为预测的反应确定催化剂，这是酶在生物合成中发挥主要作用的关键步骤。虽然酶选择的方法很多，[140]，但只有少数几种方法能专门应对逆向生物合成带来的挑战。在这种情况下，主要任务是根 据 反应物和生成物的化学结构定义反应，并输出可催化该反应的候选氨基酸序列集合。

酶搜索通常分为两大类：(i) 参考 KEGG 和 MetaCyc 等代谢数据库中的反应及其催化酶；(ii) 处理这些数据库中未发现的*新*反应，这就需要使用预测算法。E-酶，[140] Selenzyme、[141,142]和 BridgIT[143] 是检索*新*反应酶序列的三种重要方法。简而言之，它们的策略包括两个步骤：首先，使用*从头*"查询 "反应在参考数据库中找出类似的已知反应；其次，检索与最佳已知反应相关的序列。序列与反应的关联取决于参考数据库。E-enzyme和BridgIT使用 KEGG 同 源 物 和 反 应 数 据 库 ， 而 Selenzyme 使 用 BIOCHEM4J，该数据库整合了KEGG和Rhea数据库，用于将序列与反应联系起来。有趣的是，Selenzyme通过纳入系统发育距离和序列特性（如溶解度和跨膜区域）等因素，扩展了序列的排序。同时，RetroBioCat 数据库[144] 允许用户从 RetroBioCat 工具预测中搜 索 酶。不过，它不支持使用特定反应SMILES直接查询。

除序列检索系统外，Deep-RFC[145] 等计算方法也使用人工智能来评估化学反应（考虑到其底物和产物结构）是否有可能发生。此外，EnzRank[146] 工具旨在利用输入的反应物和酶序列预测酶的活性，从而有效地帮助为*新*反应选择合适的酶。还有一些方法侧 重 于 酶特征描述的不同方面，如预测 EC 编号、[143,147−151]进一步完善了适当酶催化剂的选择过程。

逆生物合成的前景。途径评估应包括反应的热力学、辅助因子的使用、底物的溶解度以及有关构建模块的成本和可持续性的目标。估算细胞宿主内途径的可行性可确保预测的可靠性，但有几个因素必须注意

K

这些因素包括宿主的选择、途径中所有反应的热力学、动力学可行性或有毒化合物的 存在和积累。此外，最 好 是尽可能缩短路径长度，并优先考虑已经表征过的反应。[153] 通过结合酶催化反应和金属离子，在一锅工艺中成功实现了大量的连续或串联转化。不过，关键是要确保途径中的所有反应都能共享相同的反应条件：溶剂、温度、pH 值，以及催化剂能与酶共存。事实上，有些酶对金属离子存在的耐受性较差[154]为了提高酶的特异性、稳定性和耐受性，可以采取适应性策略。[155,156]

数据集

通用数据库。将人工智能用于逆合成依赖于数据的质量及其有关分子、反应、途径和酶的信息多样性。在化学领域，最流行的反应基准数据集是美国专利商标局（USPTO）开源数据库。它 由 370 万个化学反应组成，人工智能模型中常用的子集包括 USPTO-50k、USPTO-full 和 USPTO-MIT。有关分子及其特性的信息经常从 ChEMBL、MoleculeNet 和 eMolecules 数据集中提取。反应信息也从 Reaxys 和 Pistachio 中提取。在生物催化和合成生物学领域，生物反应数据库包括 Rhea、RetroRules 和 MetaNetX，它们都是开源数据集。此外，我们还使用 KEGG、MetaCyc 和 PathBank 等途径数据库，以及 Brenda 和 UniProt 等酶数据库。在表 5 中，我们总结了用于复古（生物）合成的数据集、其特点以及在化学、生物催化和合成生物学中的应用范围。

数据准备。每种算法都专门用于追溯生物合成从数据库中选择反应，创建反应模板或为人工智能算法提供数据集。这种选择来自人工整理[62] 或从一个或多个数据库中自动提取，应用特定的过滤器来分离出相关的反应和分子。有些系统结合了化学和生物数据库，以增加数据集和处理生物化学反应。[73,83]首先，对反应进行分解，分离出每个反应的单一产物。[73]然后，根据分子在这些反应中的作用，筛选出一些分子：在许多反应中作为生成物的 分子被识别为

https://doi.org/10.1021/acssynbio.4c00091

表 5.人工智能应用于逆向（生物）合成的常用数据集列表

数据集	分类	可用性	描述	应用范围
美国专利商标局问		反应公众访（CC0 许可）	从美国专利商标局授予的专利中提取的有机反应数据集；该数据集经过多个子集的完善，从 USPTO-50k 数据集的 50k 反应到 USPTO-FULL 数据集的 1 M 反应不等。	Chemistry7−13,19−28,30−37,39,41,42,44,45,47,49,50,52,53,55−59,86−71,78−79,87,88,94−96,102−105,116,122,132,137−161
Reaxys	反应		商业数据库，提供经过实验验证的化学数据，包括化学结构、反应和特性	生物催化[29,73,80] Chemistry23,38,52,54,59,90,91,93,94,97,109,129
开心果	反应	包含约 2,500,000 个独特反应的商业		生物催化[83] 数据集 化学[22,45,45,94,106,110,122,162]
布伦达酵素	，再行动	公众使用（CC-BY 许可）	酶催化反应数据库，包含替代底物、动力学参数和蛋白质序列信息	生物催化[80] 生物催化[40,75,146]
瑞亚		反应公众访问（CC-BY 许可）	使用化学本体 ChEBI 的生化反应数据库，涵盖酶反应和运输反应	合成生物学[111] 生物催化[40,43,75]
复古规则		反应公众访问（CC-BY 许可）	用于发现代谢途径和代谢工程的酶催化反应建模反应模板数据库	合成生物学[111] 生物催化[40,145]
MetaNetX		反应公众访问（CC-BY 许可）	收集代谢物和生化途径，汇编 10 多个不同的生物数据库（BiGG、ChEBI、Rhea、enviPath、HMDB、KEGG、MetaCyc.....）。	合成生物学[84,124,137]生物催化[29,73,75]
KEGG		路径公开访问 完全访问付费墙	综合数据库整合了生物途径、基因组、化学和疾病信息，有助于了解生物系统及其功能	合成生物学[60,84,137]生物催化[73,83,145]
路径库	路径	公众访问（开放数据库许可证）	在 10 种模式生物中发现了约 110,000 条路径，为每种蛋白质提供了一条路径，为每种代谢物提供了一张图谱	合成生物学[111,113]生物催化[75]
MetaCyc	途径	包含约 3000 个途径和 19,000 个反应及代谢物的 商业代谢途径	横跨各种生命形式，包括初级和次级新陈代谢	生物催化[73]
UniProt	酶	公开访问（CC-BY 许可）	蛋白质序列数据库	合成生物学[111] 化学[163] 生物催化[43,146]合成生物学[60,84,137]

https://doi.org/10.1021/acssynbio.4c00091
ACS Synth.Biol.xxxx, xxx, xxx-xxx

当酶催化所必需的共底物从预定列表中被发现时，[43]，部分[73]或完全排除。[75]由于酶具有立体特异性，立体化学的使用在逆生物合成中至关重要。[40]这一过程中的非信息反应，如运输反应或没有底物的反应，都被排除在外。[43]最后，分子和反应的表示方式是根据算法的具体需求量身定制的。

讨论与展望

虽然逆合成和逆生物合成中的人工智能模型取得了长足进步，但仍存在一些挑战，需要重点关注，以增强模型，并有效克服其在生物催化和合成生物学应用中的固有限制。下面，我们将研究具体方面，包括分子表征、模型改进和评估，同时也指出成功的应用。

单步算法使用多个分子表征或其组合，从每个表征中收集互补特征。然而，常用的分子表征具有明显的局限性，例如可能产生不代表分子的 SMILES 或某些指纹或原子环境，从而阻碍准确的重建。[18]为了解决这些缺陷，人们已经推出了一些替代方案，如 SELFIES 或分子特征，[60]，但这些方案还需要进一步开发或广泛采用。一个有待进一步探索的途径是防止向可能具有不良性质的中间化学物质进行逆合成探索，如在 RetroPath RL 中，在多步探索过程中避免使用有毒化学物质。由于使用 ECFP、[164]等指纹图谱通常可以很好地预测性质和活性，而且在逆合成过程的许多部分（包括单步、[7]多步、[12]和评分功能）都使用相同的指纹图谱，[51]，因此可以设想开发出不是从目标分子而是从目标指纹图谱开始进行逆（生物）合成的方法。

如前所述，显然有必要推进以下工作针对逆生物合成的无模板和半模板方法值得重点研究。未来研究的一个有前途的方向是使用多模态模型来处理不同类型的数据，如基于提示的方法，通过整合分子表征以外的额外信息（如 EC 编号）来提高预测性能。[80]虽然大型语言模型近来在自然语言处理方面展现出了令人印象深刻的能力，并在聚合各种类型的数据和利用自动化方面展现出了前景，[165]，但它们在逆合成方面的性能仍然落后于最先进的模型。[166]

人工智能模型的性能在很大程度上受到数据可用性和质量的影响，而创建高质量的数据集往往既具有挑战性，又成本高昂。美国专利商标局的数据集被广泛用于训练和评估单步模型。该数据集由一些未经实验验证的专利合成构建而成，存在反应类别不平衡、包含缺失副产物的反应、[167]，缺乏可靠的原子映射，以及包含嘈杂的立体化学数据等问题。[168]因此，我们认为使用来自维护良好的数据库的数据集（尽管并不完美）[22]是一种可取的做法。然而，逆合成数据集经常被划分为不同的子集，这些子集基于

在特定属性上，如 USPTO-50k 或 USPTO- STEREO，即使使用 top-n 指标进行评估，也会使模型比较复杂化。因此，我们通过数据集汇总了几种单一步骤方法的性能。[169]此外，目前正在努力更好地组织现有数据，特别是通过使用人工智能模型。[170]目前，改善生物催化信息获取的举措包括[144,171]的倡议已经出现。在逆生物合成方面，应鼓励和促进这种长期承诺， 以建立可靠的数据集。

事实证明，逆向（生物）合成在多个应用领域都很有价值。在化学领域，它已被用于对药物分子[172] 进行先导优化，以及合成生物碱分子[173] 和天然产物。[174]在合成生物学领域，Galaxy-SynBioCAD 门户网站为设计代谢途径提供了一体化解决方案。[137]例 如 ，利用 RetroPath2.0，确定了在*大肠杆菌*细胞中生产番茄红素的反应，并利用机器人设备对途径的实施进行了*体内*评估。同一平台还被用于识别在无细胞系统中产生生物传感中间分子的关键反应。[175]生物催化被认为是推进绿色和可持续化学的一种方法。[3]在这方面，生物工厂的潜力已经得到展示，可以在人类专业知识和逆生物合成工具的辅助下生产材料单体。[176]利用逆向生物合成工具，Zhang 等人[177] 成功地在不使用有害氰化氢的情况下生产出脂肪族二胺；Liu 等人[178] 利用工程细胞生产出 3-苯基丙醇，从而避免了基于石油的工艺。同样，Yiakoumetti 等人在[179] 上合成了黄酮类化合物，避免了从植物中 提取；Brito 等人 在[180] 上利用甲醇作为生产 5-氨基戊酸分子的可持续替代品。一旦确定了某种分子的生物生产途径，随后使用逆向生物合成工具对已实施的途径进行优化，就能提高生产水平。例 如 ，Hanko 等人的研究[181] 在少量生产丁香酚时，其产量比以前的报告提高了近三倍。

总之，本综述广泛研究了人工智能驱动的逆合成和逆生物合成方法的最新进展，为后续可能的系统性综述铺平了道路。在化学领域观察到的良好结果为其在逆合成领域的应用带来了希望。随着发展的不断深入，我们预计人工智能模型将在逆合成领域取得显著的突破和更多的应用，从而释放其催化创新的全部潜力。

相关内容

佐证资料

辅 助 信 息 可 从 https://pubs.acs.org/doi/10.1021/acssynbio.4c00091 免 费 获 取。

（注释 S1）文献综述；（ 注释 S2）术语和定义词汇表；（图 S1）文献检索和选择过程的结果；（表 S1）从学术搜索引擎中选择文章时使用的研究查询 (PDF)

https://doi.org/10.1021/acssynbio.4c00091

作者信息

通讯作者

Jean-Loup Faulon - *Université Paris-Saclay, INRAE, AgroParisTech, Micalis Institute, 78350 Jouy-en-Josas, France; The University of Manchester, Manchester Institute of Biotechnology, Manchester M1 7DN, U.K.*; orcid.org/0000-0003-4274-2953; Email: jean-loup.faulon@inrae.fr

作者

Guillaume Gricourt - *巴黎萨克雷大学、法国国家农业研究院、巴黎农业高等专科学校、米卡利斯研究所，78350 Jouy-en-Josas, 法国；* orcid.org/0000-0003-0143-5535

Philippe Meyer - *巴黎萨克雷大学、法国国家农艺研究所、巴黎农业高等专科学校，米卡利斯研究所，78350 Jouy-en-Josas, 法国；* orcid.org/0000-0002-0618-2947

Thomas Duigou - *巴黎萨克雷大学、法国国家农艺研究所、巴黎农业高等专科学校，米卡利斯研究所，78350 Jouy-en-Josas, 法国；* orcid.org/0000-0002-2649-2950

完整的联系信息请访问：https://pubs.acs.org/10.1021/acssynbio.4c00091

作者供稿

G.G.、P.M.、T.D.和J.L.F.构思了这项研究。G.G.编写了论文集。J.L.F. 获得资金。G.G.、P.M.和T.D.撰写了手稿。所有作者阅读并批准了最终手稿。

资金筹措

这项工作得到了法国国家研究署（Agence Nationale de la Recherche）管理的法国2030计划（编号ANR-22-PEBB-0008）的资助。资助方未参与研究设计、数据收集和分析、发表决定或手稿撰写。

说明

作者声明不存在任何经济利益冲突。

缩略语

A*，A*搜索；AI，人工智能；CNN，卷积神经网络；DFPN，深度优先证明数搜索；GNN，图神经网络；HSFP，热点指纹；MaxFrag，最大片段；MCTS，蒙特卡洛树搜索；MILP，混合整数线性规划；MRR，平均倒数秩；NN，神经网络；RL，强化学习；ROC，接收器运行特征

参考文献

(1) Corey, E. J. General Methods for the Construction of Complex Molecules.In *The Chemistry of Natural Products*; Elsevier, 1967; pp 19-37.

(2) Lin, G.-M.; Warden-Rothman, R.; Voigt, C. A. Retrosynthetic Design of Metabolic Pathways to Chemicals Not Found in Nature.*Curr.Opin.Syst.***2019**, *14*, 82-107.

(3) Sheldon, R. A.; Woodley, J. M. Role of Biocatalysis in Sustainable Chemistry.*Chem.Rev.* **2018**, *118* (2), 801-838.

(4) Yu，T.；Boob，A. G.；Volk，M. J.；Liu，X.；Cui，H.；Zhao，H.
基于机器学习的分子再生物合成。*Nat.Catal.***2023**, *6* (2), 137-151.

(5) Tricco, A. C.; Lillie, E.; Zarin, W.; O'Brien, K. K.; Colquhoun、H.; Levac, D.; Moher, D.; Peters, M. D. J.; Horsley, T.; Weeks, L.; Hempel, S.; Akl, E. A.; Chang, C.; McGowan, J.; Stewart, L.; Hartling、L.；Aldcroft，A.；Wilson，M. G.；Garritty，C.；Lewin，S.；Godfrey，C.
M.; Macdonald, M. T.; Langlois, E. V.; Soares-Weiser, K.; Moriarty, J.; Clifford, T.; Tunçalp, Ö·; Straus, S. E. PRISMA 扩展范围综述（PRISMA-ScR）：Checklist and Explanation.*Ann.Int.Med.* **2018**, *169* (7), 467-473.

(6) Aal E Ali, R. S.; Meng, J.; Khan, M. E. I.; Jiang, X. Machine Learning Advancements in Organic Synthesis：有机合成中的机器学习进展： 人工智能在化学中应用的重点探索》。*Artif.Intell.Chem.***2024**, *2*, 100049.
(7) Fortunato, M. E.; Coley, C. W.; Barnes, B. C.; Jensen, K. F. Machine Learned Prediction of Reaction Template Applicability for Data-Driven Retrosynthetic Predictions of Energetic Materials; *AIP Conf.Proc.*; AIP Publishing, Portland, OR, USA, 2020; Vol. *2272*, p 070014.
(8) Wan, Y.; Liao, B.; Hsieh, C.-Y.; Zhang, S. Retroformer：推可解释的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》*（Proceedings of *the 39th International Conference on Machine Learning*; Proceedings of Machine Learning Research; PMLR, 2022; Vol. *162*, pp 22475-22490）
。
(9) Zhong, W.; Yang, Z.; Chen, C. Y.-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。*Nat.Nat.***2023**, *14* (1), 3009.
(10) Karpov, P.; Godin, G.; Tetko, I. V. A Transformer Model for Retrosynthesis.In *Artificial Neural Networks and Machine Learning - ICANN 2019: Workshop and Special Sessions*; Tetko, I. V., Kůrková, V . , Karpov, P., Theis, F., Eds.; Lecture Notes in Computer Science; Springer International Publishing: Cham, 2019; Vol. *11731*, pp 817-
830.
(11) Heid, E.; Liu, J.; Aude, A.; Green, W. H. Template Size, Canonicalization, and Exclusivity for Retrosynthesis and Reaction Prediction Applications.*J. Chem.Inf.Model.***2022**, *62* (1), 16-26.
(12) Coley, C. W.; Rogers, L.; Green, W. H.; Jensen, K. F. Computer-Assisted Retrosynthesis Based on Molecular Similarity.*ACS Cent.***2017**, *3* (12), 1237-1245.
(13) Coley, C. W.; Jin, W.; Rogers, L.; Jamison, T. F.; Jaakkola, T. S.; Green, W. H.; Barzilay, R.; Jensen, K. F. A Graph-Convolutional Neural Network Model for the Prediction of Chemical Reactivity.*Chem.Sci.* **2019**, *10* (2), 370-377.
(14) Krenn, M.; Häse, F.; Nigam, A.; Friederich, P.; Aspuru-Guzik、
A.自引用嵌入字符串（SELFIES）：100%稳健的分子字符串表示法。*马赫。Learn.Sci. Technol.***2020**, *1* (4), 045024.
(15) Carbonell, P.; Carlsson, L.; Faulon, J.-L. Stereo Signature Molecular Descriptor.*J. Chem.Inf.Model.***2013**, *53* (4), 887-897.
(16) Hähnke, V. D.; Bolton, E. E.; Bryant, S. H. PubChem Atom Environments.*J. Cheminformatics* **2015**, *7* (1), 41.
(17) Wigh, D. S.; Goodman, J. M.; Lapkin, A. A. A Review of Molecular Representation in the Age of Machine Learning.*WIREs Comput.Mol.***2022**, *12* (5), e1603.
(18) David, L.; Thakkar, A.; Mercado, R.; Engkvist, O. Molecular Representations in AI-Driven Drug Discovery：回顾与实践指南》。*J. Cheminformatics* **2020**, *12* (1), 56.
(19) 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. Cheminformatics* **2023**, *15* (1), 58.
(20) Kim, E.; Lee, D.; Kwon, Y.; Park, M. S.; Choi, Y.-S.使用具有潜在变量的绑定双向转化器进行有效、合理和多样化的逆合成。*J. Chem.Inf.Model.***2021**, *61* (1), 123-
133.
(21) Coley, C. W.; Green, W. H.; Jensen, K. F. RDChiral：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* (6), 2529-
2537.
(22) Schwaller, P.; Petraglia, R.; Zullo, V.; Nair, V. H.; Haeuselmann, R. A.; Pisoni, R.; Bekas, C.; Iuliano, A.; Laino, T. Predicting Retrosynthetic Pathways Using Transformer-Based Models and a Hyper-Graph Exploration Strategy.*Chem.***2020**, *11* (12), 3316-3325.
(23) Zhang, Y.; Wang, L.; Wang, X.; Zhang, C.; Ge, J.; Tang, J.; Su, A.; Duan, H. Data Augmentation and Transfer Learning Strategies for Reaction Prediction in Low Chemical Data Regimes.*Org.Chem.Front.***2021**, *8* (7), 1415-1423.

N https://doi.org/10.1021/acssynbio.4c00091

(24) Mao，K.；Xiao，X.；Xu，T.；Rong，Y.；Huang，J.；Zhao，P. Molecular Graph Enhanced Transformer for Retrosynthesis Pre- d i c t i o n .*Neurocomputing* **2021**, *457*, 193-202.

(25) Irwin, R.; Dimitriadis, S.; He, J.; Bjerrum, E. J. Chemformer：用于计算化学的预训练变换器。*Mach.Learn.Sci.***2022**, *3* (1), 015022.

(26) Baylon, J. L.; Cilfone, N. A.; Gulcher, J. R.; Chittenden, T. W. Enhancing Retrosynthetic Reaction Prediction with Deep Learning Using Multiscale Reaction Classification.*J. Chem.Inf.Model.***2019**, *59* (2), 673-688.

(27) Zhang, B.; Lin, J.; Du, L.; Zhang, L. Harnessing Data A u g mentation and Normalization Preprocessing to Improve the Performance of Chemical Reaction Predictions of Data-Driven Model. *聚合物***2023**，*15* (9)，2224。

(28) Zhu, J.; Xia, Y.; Wu, L.; Xie, S.; Zhou, W.; Qin, T.; Li, H.; Liu、T.-Y.双视图分子预训练。In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*; ACM: Long Beach, CA, USA, 2023; pp 3615-3627.

(29) Yang, F.; Liu, J.; Zhang, Q.; Yang, Z.; Zhang, X. CNN-Based Two-Branch Multi-Scale Feature Extraction Network for Retrosynthesis Prediction.*BMC Bioinformatics* **2022**, *23* (1), 362.

(30) 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* (1), 47-55.

(31) Lee, A. A.; Yang, Q.; Sresht, V.; Bolgar, P.; Hou, X.; Klug-McLeod, J. L.; Butler, C. R. Molecular Transformer Unifies Reaction Prediction and Retrosynthesis across Pharma Chemical Space.*Chem.Chem.***2019**, *55* (81), 12152-12155.

(32) Wang, X.; Li, Y.; Qiu, J.; Chen, G.; Liu, H.; Liao, B.; Hsieh, C.-Y.; Yao, X. RetroPrime：基于变换器的单步逆合成预测方法。*Chem.Chem.J.* **2021**, *420*, 129845.

(33) Liu, B.; Ramsundar, B.; Kawthekar, P.; Shi, J.; Gomes, J.; Luu Nguyen, Q.; Ho, S.; Sloane, J.; Wender, P.; Pande, V. Retrosynthetic Reaction Prediction Using Neural Sequence-to-Sequence Models.*ACS Cent.Sci.* **2017**, *3* (10), 1103-1113.

(34) Fang, L.; Li, J.; Zhao, M.; Tan, L.; Lou, J.-G.利用共同保留的子结构进行单步逆合成预测。*Nat.Nat.***2023**, *14* (1), 2446.

(35) Zhang, K.; Mann, V.; Venkatasubramanian, V. G MATT: Single step Retrosynthesis Prediction Using Molecular Grammar Tree Transformer.*AIChE J.* **2024**, *70*, e18244.

(36) Yan, C.; Zhao, P.; Lu, C.; Yu, Y.; Huang, J. RetroComposer：基于模板的逆合成预测模板合成。*生物分子***2022**，*12* (9)，1325。

(37) 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。*分子***2020**, *25* (10), 2357.

(38) Qiao, H.; Wu, Y.; Zhang, Y.; Zhang, C.; Wu, X.; Wu, Z.; Zhao, Q.; Wang, X.; Li, H.; Duan, H. Transformer-Based Multitask Learning for Reaction Prediction under Low-Resource Circumances.*RSC Adv.* **2022**, *12* (49), 32020-32026.

(39) Yan, Y.; Zhao, Y.; Yao, H.; Feng, J.; Liang, L.; Han, W.; Xu, X.；Pu，C.；Zang，C.；Chen，L.；Li，Y.；Liu，H.；Lu，T.；Chen，Y.；Zhang，Y. RPBP：基于副产物的深度逆合成反应预测。*J. Chem.Inf.Model.***2023**, *63* (19), 5956-5970.

(40) Heid, E.; Goldman, S.; Sankaranarayanan, K.; Coley, C. W.; Flamm, C.; Green, W. H. EHreact：用于酶促反应模板提取和评分的扩展哈塞图。*J. Chem.Inf.Model.***2021**, *61* (10), 4949-4961.

(41) S e o，S.-W.；Song，Y.Y.；Yang，J.Y.；Bae，S.；Lee，H.；Shin，J.；Hwang, S. J.; Yang, E. GTA: Graph Truncated Attention for Retrosynthesis.*Proc.AAAI Conf.Artif.Artif.***2021**, *35* (1), 531-539.

(42) He, H.-R.; Wang, J.; Liu, Y.; Wu, F. Modeling Diverse Chemical Reactions for Single-Step Retrosynthesis via Discrete Latent Variables.*第 31 届 ACM 国际信息与知识管理大会论文集*》；ACM：美国佐治亚州亚特兰大，2022 年；第 717-726 页。

https://doi.org/10.1021/acssynbio.4c00091

(43) Sankaranarayanan, K.; Heid, E.; Coley, C. W.; Verma, D.; Green, W. H.; Jensen, K. F. Similarity Based Enzymatic Retrosynthesis.*Chem.***2022**, *13* (20), 6039-6053.
(44) Tetko, I. V.; Karpov, P.; Van Deursen, R.; Godin, G. State-of- the-Art Augmented NLP Transformer Models for Direct and Single- Step Retrosynthesis.*Nat.Nat.***2020**, *11* (1), 5575.
(45) Toniato, A.; Vaucher, A. C.; Schwaller, P.; Laino, T. Enhancing Diversity in Language Based Models for Single-Step Retrosynthesis.*Digit.Discovery* **2023**, *2* (2), 489-501.
(46) Ucak, U. V.; Ashyrmamatov, I.; Lee, J. Reconstruction of Lossless Molecular Representations from Fingerprints.*J. Cheminfor- matics* **2023**, *15* (1), 26.
(47) Seidl, P.; Renz, P.; Dyubankova, N.; Neves, P.; Verhoeven, J.; Wegner, J. K.; Segler, M.; Hochreiter, S.; Klambauer, G. Improving Few- and Zero-Shot Reaction Template Prediction Using Modern Hopfield Networks.*J. Chem.Inf.Model.***2022**, *62* (9), 2111-2120.
(48) Yu, J.; Wang, J.; Zhao, H.; Gao, J.; Kang, Y.; Cao, D.; Wang, Z.; Hou, T. 基于图注意机理的有机化合物合成可及性预测。*J. Chem.Inf.Model.***2022**, *62* (12), 2973-2986.
(49) Ucak, U. V.; Kang, T.; Ko, J.; Lee, J. Substructure-Based Neural Machine Translation for Retrosynthetic Prediction.*J. Cheminformatics* **2021**, *13* (1), 4.
(50) Ucak, U. V.; Ashyrmamatov, I.; Ko, J.; Lee, J. 通过原子环境的神经机器翻译进行逆合成反应途径预测。*Nat.Nat.***2022**, *13* (1), 1186.
(51) Badowski, T.; Gajewska, E. P.; Molga, K.; Grzybowski, B. A. Synergy Between Expert and Machine Learning Approaches Allows for Improved Retrosynthetic Planning.*Angew.Chem.Ed.***2020**, *59* (2), 725-730.
(52) Thakkar, A.; Selmi, N.; Reymond, J.-L.; Engkvist, O.; Bjerrum、E.J. Ring Breaker"：神经网络驱动的环系统化学空间合成预测。*J. Med.Chem.***2020**, *63* (16), 8791-8808.
(53) 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* (2), 641-652.
(54) Segler, M. H. S.; Waller, M. P. Neural Symbolic Machine Learning for Retrosynthesis and Reaction Prediction.*Chem.- Eur.J.* **2017**, *23* (25), 5966-5971.
(55) Ishida, S.; Terayama, K.; Kojima, R.; Takasu, K.; Okuno, Y. Prediction and Interpretable Visualization of Retrosynthetic Reactions Using Graph Convolutional Networks.*J. Chem.Inf.Model.***2019**, *59* (12), 5026-5033.
(56) Dai, H.; Li, C.; Coley, C.; Dai, B.; Song, L. 利用条件图逻辑网络进行逆合成预测。In *Advances in Neural Information Processing Systems 32*; NeurIPS 2019; Curran Associates, Inc., 2019; Vol.
(57) Chen, S.; Jung, Y. 利用局部反应性和全局注意力进行深度逆合成反应预测。*JACS Au* **2021**, *1* (10), 1612-1620.
(58) Lee, H.; Ahn, S.; Seo, S.-W.; Song, Y. Y.; Yang, E.; Hwang, S.-J.；Shin, J. RetCL: A Selection-Based Approach for Retrosynthesis via Contrastive Learning.In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*; IJCAI-21; International Joint Conferences on Artificial Intelligence Organization, 2021; pp 2673- 2679.
(59) Lin, Z.; Yin, S.; Shi, L.; Zhou, W.; Zhang, Y. J. G2GT：利用图到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* (7), 1894-1905.
(60) Delépine, B.; Duigou, T.; Carbonell, P.; Faulon, J.-L. RetroPath2.0：代谢工程师的逆合成工作流程。*Metab.Eng.* **2018**, *45*, 158-170.
(61) Szymkuć, S.; Gajewska, E. P.; Klucznik, T.; Molga, K.; Dittwald、P.; Startek, M.; Bajczyk, M.; Grzybowski, B. A. Computer-Assisted Synthetic Planning：起点的终点。*Angew.Chem.Ed.***2016**, *55* (20), 5904-5937.

(62) Finnigan, W.; Hepworth, L. J.; Flitsch, S. L.; Turner, N. J. RetroBioCat 作为生物催化反应和级联的计算机辅助合成规划工具。*Nat.Catal.***2021**, *4* (2), 98-104.
(63) Duigou, T.; du Lac, M.; Carbonell, P.; Faulon, J.-L. RetroRules：用于工程生物学的反应规则数据库。*Nucleic Acids Res.* **2019**, *47* (D1), D1229-D1235.
(64) 基于图关系网络的 逆合成预测。In *2022 15th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*; IEEE: Beijing, China, 2022; pp 1-5.
(65) Schwaller, P.; Laino, T.; Gaudin, T.; Bolgar, P.; Hunter, C. A.; Bekas, C.; Lee, A. A. Molecular Transformer：不确定性校准化学反应预测模型。*ACS Cent.Sci.* **2019**, *5* (9), 1572-1583.
(66) Guo, Z.; Wu, S.; Ohno, M.; Yoshida, R. Bayesian Algorithm for Retrosynthesis.*J. Chem.Inf.Model.***2020**, *60* (10), 4474-4486.
(67) Tu, Z.; Coley, C. W. Permutation Invariant Graph-to-Sequence Model for Template-Free Retrosynthesis and Reaction P*rediction.J. Chem.Inf.Model.***2022**, *62* (15), 3503-3513.
(68) Hu, H.; Jiang, Y.; Yang, Y.; Chen, J. X. BiG2S：用于端到端无模板反应预测的双任务图到序列模型。*Appl. Intell* **2023**, *53*, 29620.
(69) Liu, S.; Tu, Z.; Xu, M.; Zhang, Z.; Lin, L.; Ying, R.; Tang, J.; Zhao, P.; Wu, D. FusionRetro: Molecule Representation Fusion via In-Context Learning for Retrosynthetic Planning.In *Proceedings of the 40th International Conference on Machine Learning*; ICML'23; JMLR.org：美国夏威夷檀香山，2023 年。
(70) Lin, M. H.; Tu, Z.; Coley, C. W. Improving the Performance of Models for One-Step Retrosynthesis through Re-Ranking.*J. Cheminformatics* **2022**, *14* (1), 15.
(71) 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 34;* NeurIPS 2 0 2 1 ; Curran Associates, Inc.
(72) Christofidellis, D.; Giannone, G.; Born, J.; Winther, O.; Laino, T.; Manica, M. Unifying Molecular and Textual Representations via Multi-Task Language Modelling.In *Proceedings of the 40th International Conference on Machine Learning*; ICML'23; JMLR.org ： Honolulu, Hawaii, USA, 2023.
(73) Zheng, S.; Zeng, T.; Li, C.; Chen, B.; Coley, C. W.; Yang, Y.; Wu, R. Deep Learning Driven Biosynthetic Pathways Navigation for Natural Products with BioNavi-NP.*Nat.Nat.***2022**, *13* (1), 3342.
(74) Kreutter, D.; Schwaller, P.; Reymond, J.-L. Predicting Enzymatic Reactions with a Molecular Transformer.*Chem.***2021**, *12* (25), 8648-8659.
(75) Probst, D.; Manica, M.; Nana Teukam, Y. G.; Castrogiovanni, A.; Paratore, F.; Laino, T. Biocatalysed Synthesis Planning Using Data-Driven Learning.*Nat.Nat.***2022**, *13* (1), 964.
(76) Shi, C.; Xu, M.; Guo, H.; Zhang, M.; Tang, J. A Graph to Graphs Framework for Retrosynthesis Prediction.In *Proceedings of the 37th International Conference on Machine Learning*; ICML'20; JMLR.org, 2020.
(77) Yan, C.; Ding, Q.; Zhao, P.; Zheng, S.; Yang, J.; Yu, Y.; Huang、J.RetroXpert：像化学家一样分解逆合成预测。*第 34 届神经信息处理系统国际会议论文集*；NIPS'20；Curran Associates Inc：Red Hook, NY, USA, 2020.
(78) Somnath, V. R.; Bunne, C.; Coley, C. W.; Krause, A.; Barzilay、R.学习逆合成预测的图模型。In *Advances i n Neural Information Processing Systems 34*; NeurIPS 2021; Curran Associates, Inc., 2021.
(79) Wang, Y.; Pang, C.; Wang, Y.; Jin, J.; Zhang, J.; Zeng, X.; Su, R.; Zou, Q.; Wei, L. Retrosynthesis Prediction with an Interpretable Deep-Learning Framework Based on Molecular Assembly Tasks.*Nat.Nat.***2023**, *14* (1), 6155.
(80) Thakkar, A.; Vaucher, A. C.; Byekwaso, A.; Schwaller, P.; Toniato, A.; Laino, T. Unbiasing Retrosynthesis Language Models with Disconnection Prompts.*ACS Cent.***2023**, *9* (7), 1488-1498.

P

https://doi.org/10.1021/acssynbio.4c00091

(81) Jaume-Santero, F.; Bornet, A.; Valery, A.; Naderi, N.; Vicente Alvarez, D.; Proios, D.; Yazdani, A.; Bournez, C.; Fessard, T.; Teodoro, D. 化学反应变压器性能：不同预测和评估方案的分析。*J. Chem.Inf.Model.***2023**, *63* (7), 1914-1924.
(82) Wang, X.; Hsieh, C.-Y.; Yin, X.; Wang, J.; Li, Y.; Deng, Y.; Jiang、D.; Wu, Z.; Du, H.; Chen, H.; Li, Y.; Liu, H.; Wang, Y.; Luo, P.; Hou、T.; Yao, X. 利用开放反应条件数据集和反应中心的无监督学习进行通用可解释反应条件预测。*Research* **2023**, *6*, 0231.
(83) Zhang, C.; Lapkin, A. A. Reinforcement Learning Optimization of Reaction Routes on the Basis of Large, Hybrid Organic Chemistry-Synthetic Biological, Reaction Network Data.*React.Chem.***2023**, *8* (10), 2491-2504.
(84) Koch, M.; Duigou, T.; Faulon, J.-L. Reinforcement Learning for Bioretrosynthesis.*ACS Synth.***2020**, *9* (1), 157-168.
(85) Wang, W.; Liu, Q.; Zhang, L.; Dong, Y.; Du, J. RetroSynX: A Retrosynthetic Analysis Framework Using Hybrid Reaction Templates and Group Contribution-Based Thermodynamic Models.*Chem.Chem.***2022**, *248*, 117208.
(86) Liu, Q.; Tang, K.; Zhang, L.; Du, J.; Meng, Q. 基于过渡态自动生成方法的反应动力学计算机辅助合成规划.*AIChE J.* **2023**, *69* (7), e18092.
(87) Kreutter, D.; Reymond, J.-L. 结合断路感知三重变压器环路与路由惩罚得分引导树搜索的多步逆合成。*Chem.***2023**, *14* (36), 9959-9969.
(88) Kishimoto, A.; Buesser, B.; Chen, B.; Botea, A. Depth-First Proof-Number Search with Heuristic Edge Cost and Application to Chemical Synthetic Planning.In *Advances in Neural Information Processing Systems 32*; NeurIPS 2019; Curran Associates, Inc.
(89) Franz, C.; Mogk, G.; Mrziglod, T.; Schewior, K. Completeness and Diversity in Depth-First Proof-Number Search with Applications to Retrosynthesis.In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence*; International Joint Conferences on Artificial Intelligence Organization： Vienna, Austria, 2022; pp 4747-4753.
(90) Shibukawa, R.; Ishida, S.; Yoshizoe, K.; Wasa, K.; Takasu, K.; Okuno, Y.; Terayama, K.; Tsuda, K. CompRet：用算法枚举进行化学合成规划的综合推荐框架。*J. Cheminformatics* **2020**, *12* (1), 52.
(91) Segler, M. H. S.; Preuss, M.; Waller, M. P. Planning Chemical Syntheses with Deep Neural Networks and Symbolic AI.*Nature* **2018**, *555* (7698), 604-610.
(92) Zhang, B.; Zhang, X.; Du, W.; Song, Z.; Zhang, G.; Zhang, G.; Wang, Y.; Chen, X.; Jiang, J.; Luo, Y. Chemistry-Informed Molecular Graph as Reaction Descriptor for Machine-Learned Retrosynthesis Planning.*Proc.Natl.***2022**, *119* (41), e2212711119.
(93) Wang, X.; Qian, Y.; Gao, H.; Coley, C. W.; Mo, Y.; Barzilay, R.; Jensen, K. F. Towards Efficient Discovery of Green Synthetic Pathways with Monte Carlo Tree Search and Reinforcement Learning.*Chem.***2020**, *11* (40), 10959-10972.
(94) Thakkar, A.; Kogej, T.; Reymond, J.-L.; Engkvist, O.; Bjerrum、E.J. Datasets and Their Influence on the Development of Computer Assisted Synthetic Planning Tools in the Pharmaceutical Domain.*Chem.***2020**, *11* (1), 154-168.
(95) 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.***2020**, *11* (12)、3355-3364.
(96) Genheden, S.; Thakkar, A.; Chadimová, V.; Reymond, J.-L.; Engkvist, O.; Bjerrum, E. AiZynthFinder：用于逆合成规划的快速、稳健、灵活的开源软件。*J. Cheminformatics* **2020**, *12* (1), 70.
(97) Gao, H.; Coley, C. W.; Struble, T. J.; Li, L.; Qian, Y.; Green, W. H.; Jensen, K. F. Combining Retrosynthesis and Mixed-Integer Optimization for Minimizing the Chemical Inventory needed to Realize a WHO Essential Medicines List.*React.Chem.***2020**, *5* (2), 367-376.

(98) Coley, C. W.; Thomas, D. A.; Lummiss, J. A. M.; Jaworski, J.. N.; Breen, C. P.; Schultz, V.; Hart, T.; Fishman, J. S.; Rogers, L.; Gao、H.; Hicklin, R. W.; Plehiers, P. P.; Byington, J.; Piotti, J. S.; Green, W. H . ; Hart, A. J.; Jamison, T. F.; Jensen, K. F. A Robotic Platform for Flow Synthesis of Organic Compounds Informed by AI Planning.*Science* **2019**, *365* (6453), eaax1566.

(99) Westerlund, A. M.; Barge, B.; Mervin, L.; Genheden, S. Data driven Approaches for Identifying Hyperparameters in Multi step Retrosynthesis.*Mol. Inform.Inform.***2023**, *42*, 2300128.

(100) Sankaranarayanan, K.; Jensen, K. F. Computer-Assisted Multistep Chemoenzymatic Retrosynthesis Using a Chemical Synthesis Planner.*Chem.***2023**, *14* (23), 6467-6475.

(101) Levin, I.; Liu, M.; Voigt, C. A.; Coley, C. W. Merging Enzymatic and Synthetic Chemistry with Computational Synthesis Planning.*Nat.Commun.***2022**, *13*, 7747.

(102) Jeong，J.；Lee，N.；Shin，Y.；Shin，D. 基于知识图谱推理和利用反应大数据进行逆合成预测的最佳合成途径的智能生成。*J. Taiwan Inst.***2022**, *130*, 103982.

(103) Chen, B.; Li, C.; Dai, H.; Song, L. Retro*：学习使用神经引导 A* 搜索的逆合成规划。*第 37 届机器学习国际会议论文集*；机器学习研究论文集（PMLR），2020 年；第 *119* 卷，第 1608-1616 页。

(104) Xie, S.; Yan, R.; Han, P.; Xia, Y.; Wu, L.; Guo, C.; Yang, B..；Qin, T. RetroGraph：使用图搜索的逆向合成规划。In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*; ACM: Washington DC, USA, 2022; pp 2120-2129.

(105) Han, P.; Zhao, P.; Lu, C.; Huang, J.; Wu, J.; Shang, S.; Yao, B.；Zhang, X. GNN-Retro：利用图神经网络的逆合成规划。*Proc.AAAI Conf.Artif.Intell.***2022**, *36* (4), 4014-4021.

(106) Latendresse, M.; Malerich, J. P.; Herson, J.; Krummenacker, M.; Szeto, J.; Vu, V.-A.; Collins, N.; Madrid, P. B. SynRoute：逆合成规划软件。*J. Chem.Inf.Model.***2023**, *63* (17), 5484-5495.

(107) Grzybowski, B. A.; Szymkuć, S.; Gajewska, E. P.; Molga, K.；Dittwald, P.; Wołos, A.; Klucznik, T. Chematica：计算机代码开始像化学家一样思考的故事。*Chem.***2018**, *4* (3), 390-398.

(108) Russell, S. J.; Norvig, P. *Artificial Intelligence : A Modern Approach*, 4th ed.; Pearson series in artificial intelligence; Pearson: Hoboken, 2021.

(109) Schreck, J. S.; Coley, C. W.; Bishop, K. J. M. Learning Retrosynthetic Planning through Simulated Experience.*ACS Cent.***2019**, *5* (6), 970-981.

(110) Yu, Y.; Wei, Y.; Kuang, K.; Huang, Z.; Yao, H.; Wu, F. GRASP：利用目标驱动策略进行逆合成规划导航。*神经信息处理系统进展 35*；NeurIPS 2022；Curran Associates 公司，2022 年。

(111) Kumar, A.; Wang, L.; Ng, C. Y.; Maranas, C. D. Pathway Design Using de Novo Steps through Uncharted Biochemical Spaces.*Nat.Commun.***2018**, *9* (1), 184.

(112) Carbonell, P.; Parutto, P.; Herisson, J.; Pandit, S. B.; Faulon, J.-L. XTMS: Pathway Design in an eXTended Metabolic Space.*Nucleic Acids Res.* **2014**, *42* (W1), W389-W394.

(113) Tokic, M.; Hadadi, N.; Ataman, M.; Neves, D.; Ebert, B. E.; Blank, L. M.; Miskovic, L.; Hatzimanikatis, V. Discovery and E valuation of Biosynthetic Pathways for the Production of Five Methyl Ethyl Ketone Precursors. *ACS Synth.ACS Synth.Biol.* **2018**, *7* (8), 1858-1873.

(114) Hadadi, N.; Hafner, J.; Shajkofci, A.; Zisaki, A.; Hatzimanikatis, V. ATLAS of Biochemistry：用于合成生物学和代谢工程研究的所有可能生化反应资料库。*ACS Synth.Biol.* **2016**, *5* (10), 1155-1166.

(115) Otero-Muras, I.; Carbonell, P. Automated Engineering of Synthetic Metabolic Pathways for Efficient Biomanufacturing.*Metab.***2021**, *63*, 61-80.

Q

(116) Kim, J.; Ahn, S.; Lee, H.; Shin, J. Self-Improved Retrosynthetic Planning.*第 38 届机器学习国际会议论文集》*，虚拟，2021 年 7 月 18-24 日；《机器学习研究论文集》（PMLR），2021 年。

(117) Gao, D.; Song, W.; Wu, J.; Guo, L.; Gao, C.; Liu, J.; Chen, X.; Liu, L. 通过酶促-化学级联催化高效生产 L-高苯丙氨酸。*Angew.Chem.Ed.***2022**, *61* (36), e202207077.

(118) Rudroff, F.; Mihovilovic, M. D.; Gröger, H.; Snajdrova, R.; Iding, H.; Bornscheuer, U. T. 《化学与生物催化相结合的机遇与挑战》。*Nat.Catal.***2018**, *1* (1), 12-22.

(119) Finnigan, W.; Flitsch, S. L.; Hepworth, L. J.; Turner, N. J. 酶级联设计：逆合成方法。Kara, S., Rudroff, F., Eds.; Springer International Publishing: Cham, 2021; pp 7-30.

(120) Skoraczyński, G.; Kitlas, M.; Miasojedow, B.; Gambin, A. 计算机辅助合成规划中合成可及性评分的关键评估。*J. Cheminformatics* **2023**, *15* (1), 6.

(121) Thakkar, A.; Chadimová, V.; Bjerrum, E. J.; Engkvist, O.; Reymond, J.-L. Retrosynthetic Accessibility Score (RAscore) - Rapid Machine Learned Synthesizability Classification from AI Driven Retrosynthetic Planning.*Chem.***2021**, *12* (9), 3339-3349.

(122) 基于反应知识图谱的化合物合成可及性预测。*分子* **2022**, *27* (3), 1039.

(123) Kim, H.; Lee, K.; Kim, C.; Lim, J.; Kim, W. Y. DFRscore：基于深度学习的合成复杂性评分与药物重点回溯合成分析，用于高通量虚拟筛选。*J. Chem.Inf.Model* **2024**, *64*, 2432.

(124) Correia, J.; Carreira, R.; Pereira, V.; Rocha, M. Predicting the number of Biochemical Transformations Needed to Synthesize a Compound.In *2022 International Joint Conference on Neural Networks (IJCNN)*; IEEE: Padua, Italy, 2022; pp 1-8.

(125) Parrot, M.; Tajmouati, H.; Da Silva, V. B. R.; Atwood, B. R.; Fourcade, R.; Gaston-Mathé, Y.; Do Huu, N.; Perron, Q. Integrating Synthetic Accessibility with AI-Based Generative Drug Design.*J. Cheminformatics* **2023**, *15* (1), 83.

(126) Sanchez-Garcia, R.; Havasi, D.; Takács, G.; Robinson, M. C.; Lee, A.; Von Delft, F.; Deane, C. M. CoPriNet：图神经网络为分子优先排序提供准确、快速的化合物价格预测。*Digit.Discovery* **2023**, *2* (1), 103-111.

(127) Tang, K.; Zhuang, Y.; Wang, W.; Liu, Q.; Zhang, L.; Du, J.; Meng, Q. 基于 GC-NORM 的热力学框架，用于评估涉及二氧化碳利用的有机反应。*Chem.Chem.***2023**, *278*, 118913.

(128) Schwaller, P.; Vaucher, A. C.; Laino, T.; Reymond, J.-L. Prediction of Chemical Reaction Yields Using Deep Learning.*Mach.Learn.Sci.***2021**, *2* (1), 015016.

(129) Plehiers, P. P.; Coley, C. W.; Gao, H.; Vermeire, F. H.; Dobbelaere, M. R.; Stevens, C. V.; Van Geem, K. M.; Green, W. H. Artificial Intelligence for Computer-Aided Synthesis In Flow: Analysis and Selection of Reaction Components.*Front.Chem.***2020**, *2*, 5.

(130) Toniato, A.; Unsleber, J. P.; Vaucher, A. C.; Weymuth, T.; Probst, D.; Laino, T.; Reiher, M. Quantum Chemical Data Generation as Fill-in for Reliability Enhancement of Machine-Learning Reaction and Retrosynthesis Planning.*Digit.Discovery* **2023**, *2* (3), 663-673.

(131) Genheden, S.; Engkvist, O.; Bjerrum, E. Fast Prediction of Distances between Synthetic Routes with Deep Learning.*Mach.Learn.Sci.***2022**, *3* (1), 015018.

(132) Kuznetsov, A.; Sahinidis, N. V. ExtractionScore：根据预测的液-液萃取性能评估合成路线的定量框架。*J. Chem.Inf.Model.***2021**, *61* (5), 2274-2282.

(133) 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. Cheminformatics* **2009**, *1* (1), 8.

https://doi.org/10.1021/acssynbio.4c00091

(134) Coley, C. W.; Rogers, L.; Green, W. H.; Jensen, K. F. SCScore：从反应语料库中了解合成复杂性。*J. Chem.Inf.Model.***2018**, *58* (2), 252-261.
(135) Ebrahim, A.; Lerman, J. A.; Palsson, B. O.; Hyduke, D. R. COBRApy：基于约束的 Python 重构与分析。*BMC Syst.***2013**, *7* (1), 74.
(136) Beber, M. E.; Gollub, M. G.; Mozaffari, D.; Shebek, K. M.; Flamholz, A. I.; Milo, R.; Noor, E. eQuilibrator 3.0：热力学常数估算的数据库解决方案。*Nucleic Acids Res.* **2022**, *50*, D603-D609.
(137) Hérisson, J.; Duigou, T.; Du Lac, M.; Bazi-Kabbaj, K.; Sabeti Azad, M.; Buldum, G.; Telle, O.; El Moubayed, Y.; Carbonell, P.; Swainston, N.; Zulkower, V.; Kushwaha, M.; Baldwin, G. S.; Faulon, J.-L. The Automated Galaxy-SynBioCAD Pipeline for Synthetic Biology Design and Engineering.*Nat.Nat.***2022**, *13* (1), 5082.
(138) Levin, I.; Fortunato, M. E.; Tan, K. L.; Coley, C. W. Computer aided Evaluation and Exploration of Chemical Spaces Constrained by Reaction Pathways.*AIChE J.* **2023**, *69*, e18234.
(139) Wang, L.; Upadhyay, V.; Maranas, C. D. dGPredictor：用于代谢反应自由能预测和新通路设计的自动碎裂法。*PLOS Comput.***2021**, *17* (9), e1009448.
(140) Feehan, R.; Montezano, D.; Slusky, J. S. G. Machine Learning for Enzyme Engineering, Selection and Design.*Protein Eng.Des.Des.***2021**, *34*, GZAB019.
(141) Stoney, R. A.; Hanko, E. K. R.; Carbonell, P.; Breitling, R. SelenzymeRF: Updated Enzyme Suggestion Software for Unbalanced Bihemical Reactions.*Comput.Struct.Biotechnol.J.* **2023**, *21*, 5868-5876.
(142) Carbonell, P.; Wong, J.; Swainston, N.; Takano, E.; Turner, N. J.; Scrutton, N. S.; Kell, D. B.; Breitling, R.; Faulon, J.-L. Selenzyme: Enzyme Selection Tool for Pathway Design.*Bioinforma.Oxf.Engl.* **2018**, *34* (12), 2153-2154.
(143) Hadadi, N.; MohammadiPeyhani, H.; Miskovic, L.; Seijo, M.; Hatzimanikatis, V. Enzyme Annotation for Orphan and Novel R e actions Using Knowledge of Substrate Reactive Sites.*Proc.Natl.美国国家科学院院刊*，**2019**，*116* (15)，7298-7307。
(144) Finnigan, W.; Lubberink, M.; Hepworth, L. J.; Citoler, J.; Mattey, A. P.; Ford, G. J.; Sangster, J.; Cosgrove, S. C.; da Costa, B. Z.; Heath, R. S.; Thorpe, T. W.; Yu, Y.; Flitsch, S. L.; Turner, N. J. RetroBioCat 数据库：生物催化数据的协作整理和自动元分析平台。*ACS Catal.***2023**, *13* (17), 11771-11780.
(145) Kim, Y.; Ryu, J. Y.; Kim, H. U.; Jang, W. D.; Lee, S. Y. A Deep 评估网状生物合成产生的酶促反应可行性的学习方法。*J. 2021, 16 (5), 2000605.J.* **2021**, *16* (5), 2000605.
(146) Upadhyay, V.; Boorla, V. S.; Maranas, C. D. Rank-Ordering of Known Enzymes as Starting Points for Re-Engineering Novel Substrate Activity Using a Convolutional Neural Network.*Metab.***2023**，*78*，171-182。
(147) Kotera，M.；Okuno，Y.；Hattori，M.；Goto，S.；Kanehisa，M. 基因组尺度分析酶促反应的 EC 编号计算分配。*J. Am.J. Am.***2004**, *126* (50), 16487-16498.
(148) Rahman, S. A.; Cuesta, S. M.; Furnham, N.; Holliday, G. L.; Thornton, J. M. EC-BLAST: A Tool to Automatically Search and Compare Enzyme Reactions.*Nat.Methods* **2014**, *11* (2), 171-174.
(149) Egelhofer, V.; Schomburg, I.; Schomburg, D. Automatic Assignment of EC Numbers.*PLoS Comput.Biol.* **2010**, *6* (1), e1000661.
(150) Hu, Q.-N.; Zhu, H.; Li, X.; Zhang, M.; Deng, Z.; Yang, X.; Deng, Z. 利用反应差异指纹为酶促反应分配 EC 编号。*PLoS One* **2012**, *7* (12), e52901.
(151) Probst, D. An Explainability Framework for Deep Learning on Chemical Reactions Exemplified by Enzyme-Catalysed Reaction Classification.*J. Cheminformatics* **2023**, *15* (1), 113.
(152) Liu, C.-H.; Korablyov, M.; Jastrzębski, S.; Włodarczyk-Pruszyński, P.; Bengio, Y.; Segler, M. RetroGNN：快速估算虚拟筛选和新设计的可合成性

R

从慢速逆合成软件中学习。*J. Chem.Inf.Model.* **2022**, *62* (10), 2293-2300.
(153) Hafner, J.; Mohammadi-Peyhani, H.; Hatzimanikatis, V. Pathway Design.*代谢工程*》，约翰威利父子有限公司，2021 年，第 237-257 页。
(154) de Souza, R. O. M. A.; Miranda, L. S. M.; Bornscheuer, U. T. A Retrosynthesis Approach for Biocatalysis in Organic Synthesis.*Chem.-Eur.J.* **2017**, *23* (50), 12040-12063.
(155) Song, Z.; Zhang, Q.; Wu, W.; Pu, Z.; Yu, H. Rational Design of Enzyme Activity and Enantioselectivity.*Front.Bioeng.Biotechnol* **2023**, *11*, 1129149.
(156) Ribeiro, A. J. M.; Riziotis, I. G.; Borkakoti, N.; Thornton, J. M. Enzyme Function and Evolution through the Lens of Bioinformatics.*Biochem.J.* **2023**, *480* (22), 1845-1863.
(157) Beaudoin, C.; Kundu, S.; Topaloglu, R. O.; Ghosh, S. Quantum Machine Learning for Material Synthesis and Hardware Security（特邀论文）。In *Proceedings of the 41st IEEE/ACM International Conference on Computer-Aided Design;* ACM: San Diego, California, 2022; pp 1-7.
(158) Fan, Y.; Xia, Y.; Zhu, J.; Wu, L.; Xie, S.; Qin, T. Back 分子生成的翻译。*生物信息学***2022**》，*38* (5)、1244-1251.
(159) Zahoránszky-Kőhalmi, G.; Lysov, N.; Vorontcov, I.; Wang, J.; Soundararajan, J.; Metaxotos, D.; Mathew, B.; Sarosh, R.; Michael, S. G.; Godfrey, A. G. Algorithm for the Pruning of Synthesis Graphs.*J. Chem.Inf.Model.***2022**, *62* (9), 2226-2238.
(160) Chen, Z.; Ayinde, O. R.; Fuchs, J. R.; Sun, H.; Ning, X. G2Retro 作为逆合成预测的两步图生成模型。*Commun.Chem.***2023**, *6* (1), 102.
(161) 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* (7), 1841-1846.
(162) Mo, Y.; Guan, Y.; Verma, P.; Guo, J.; Fortunato, M. E.; Lu, Z.; Coley, C. W.; Jensen, K. F. Evaluating and Clustering Retrosynthesis Pathways with Learned Strategy.*Chem.***2021**, *12* (4), 1469-1478.
(163) Born, J.; Manica, M.; Cadow, J.; Markert, G.; Mill, N. A.; Filipavicius, M.; Janakarajan, N.; Cardinale, A.; Laino, T.; Rodríguez Martínez, M. Data-Driven Molecular Design for Discovery and Synhesis of Novel Ligands：SARS-CoV-2 案例研究。*Mach.Learn.Sci.***2021**, *2* (2), 025024.
(164) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints.*J. Chem.Inf.Model.***2010**, *50* (5), 742-754.
(165) Boiko, D. A.; MacKnight, R.; Kline, B.; Gomes, G. Autonomous Chemical Research with Large Language Models. *自然* **2023**》，*624* (7992)，570-578。
(166) Guo, T.; Guo, K.; Nan, B.; Liang, Z.; Guo, Z.; Chawla, N. V.; Wiest, O.; Zhang, X. What Can Large Language Models Do in Chemistry? 八项任务的综合基准。http:// arxiv.org/abs/2305.18365 （访问日期：2023-11-27）。
(167) Meng, Z.; Zhao, P.; Yu, Y.; King, I. A Unified View of Deep Learning for Reaction and Retrosynthesis Prediction：现状与未来挑战。In *Proceedings of the Thirty-Second Interna- tional Joint Conference on Artificial Intelligence*; International Joint Conferences on Artificial Intelligence Organization：中国澳门特别行政区，2023 年；第 6723-6731 页。
(168) Hu, W.; Liu, Y.; Chen, X.; Chai, W.; Chen, H.; Wang, H.; Wang, G. Deep Learning Methods for Small Moleule Drug Discovery：Deep Learning Methods for Small Moleule Drug Discovery: A Survey.*IEEE Trans.Artif.Intell.***2024**, *5*, 459.
(169) Jiang, Y .; Yu, Y.; Kong, M.; Mei, Y.; Yuan, L.; Huang, Z.; Kuang, K.; Wang, Z.; Yao, H.; Zou, J.; Coley, C. W.; Wei, Y. Artificial Intelligence for Retrosynthesis Prediction.*工程***2023**，*25*，32-50.
(170) Kearnes, S. M.; Maser, M. R.; Wleklinski, M.; Kast, A.; Doyle、A.G.; Dreher, S. D.; Hawkins, J. M.; Jensen, K. F.; Coley, C. W. The Open Reaction Database.*J. Am.J. Am. Chem.***2021**, *143* (45), 18820-18826.
(171) Heid, E.; Probst, D.; Green, W. H.; Madsen, G. K. H. EnzymeMap： EnzymeMap: Curation, Validation and Data-Driven Prediction of Enzymatic Reactions.*Chem.***2023**, *14* (48), 14229-14242.

https://doi.org/10.1021/acssynbio.4c00091

(172) Seierstad, M.; Tichenor, M. S.; DesJarlais, R. L.; Na, J.; Bacani、G. M.; Chung, D. M.; Mercado-Marin, E. V.; Steffens, H. C.; Mirzadegan, T. Novel Reagent Space：识别无序但易于合成的构件。*ACS Med.Chem.***2021**, *12* (11), 1853-1860.

(173) Lin, Y.; Zhang, R.; Wang, D.; Cernak, T. Computer-Aided Key Step Generation in Alkaloid Total Synhesis.*Science* **2023**, *379* (6631), 453-457.

(174) Hardy, M. A.; Nan, B.; Wiest, O.; Sarpong, R. Strategic E l e ments in Computer-Assisted Retrosynthesis： A Case Study of the Pupukeanane Natural Products.*Tetrahedron* **2022**, *104*, 132584.

(175) Soudier，P.；Zúñiga，A.；Duigou，T.；Voyvodic，P. L.；Bazi- Kabbaj，K.；Kushwaha，M.；Vendrell，J. A.；Solassol，J.；Bonnet，J.；Faulon，J.-L. PeroxiHUB：使用 H_2O_2 作为信号集成器的模块化无细胞生物传感平台。*ACS Synth.***2022**, *11* (8)、2578-2588.

(176) Robinson, C. J.; Carbonell, P.; Jervis, A. J.; Yan, C.; Hollywood, K. A.; Dunstan, M. S.; Currin, A.; Swainston, N.; Spiess, R.; Taylor, S.; Mulherin, P.; Parker, S.; Rowe, W.; Matthews、N.E.; Malone, K. J.; Le Feuvre, R.; Shapira, P.; Barran, P.; Turner, N. J．; Micklefield, J.; Breitling, R.; Takano, E.; Scrutton, N. S. Rapid Prototyping of Microbial Production Strains for the Biomanufacture of Potential Materials Monomers.*Metab.***2020**，*60*，168-182。

(177) Zhang, Z.; Fang, L.; Wang, F.; Deng, Y.; Jiang, Z.; Li, A. Designed Enzymatic Cascade Catalysis.*Angew.Chem.Ed.***2023**, *62* (16), e202215935.

(178) Liu, Z.; Zhang, X.; Lei, D.; Qiao, B.; Zhao, G.-R. 通过 Retrobiosynthesis方法从新生产3-苯基丙醇的大肠杆菌代谢工程。*Microb.细胞工厂***2021**, *20* (1), 121.

(179) Yiakoumetti, A.; Hanko, E. K. R.; Zou, Y.; Chua, J.; Chromy、J.; Stoney, R. A.; Valdehuesa, K. N. G.; Connolly, J. A.; Yan, C.; Hollywood, K. A.; Takano, E.; Breitling, R. Expanding Flavone and Flavonol Production Capabilities in Escherichia Coli.*Front.Bioeng.Biotechnol.***2023**, *11*, 1275651.

(180) Brito, L. F.; Irla, M.; Nærdal, I.; Le, S. B.; Delépine, B.; Heux, S.; Brautaset, T. Evaluation of Heterologous Biosynthetic Pathways for Methanol-Based 5-Aminovalerate Production by Thermophilic Bacillus Methanolicus.*Front.Bioeng.Biotechnol* **2021**, *9*, 1.

(181) Hanko, E. K. R.; Valdehuesa, K. N. G.; Verhagen, K. J. A.；Chromy, J.; Stoney, R. A.; Chua, J.; Yan, C.; Roubos, J. A.; Schmitz, J.; Breitling, R. Carboxylic Acid Reductase-Dependent Biosynthesis of Eugenol and Related Allylphenols.*Microb.细胞工厂***2023**，*22* (1)，238。

S

https://doi.org/10.1021/acssynbio.4c00091