JACS Au
AN OPEN ACCESS JOURNAL OF THE AMERICAN CHEMICAL SOCIETY

pubs.acs.org/jacsau

文章

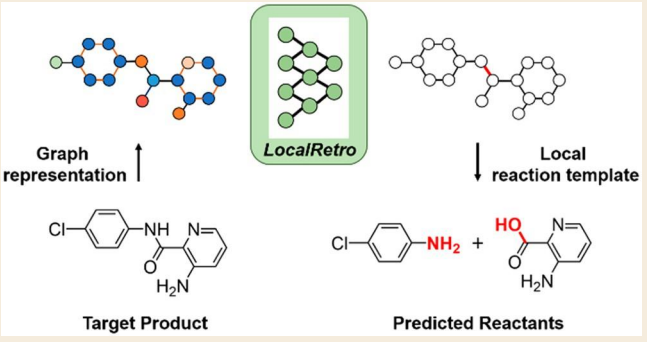
利用局部反应性和全局注意力进行深度逆合成反应预测

Shuan Chen 和 Yousung Jung*

引用此文：JACS Au 2021, 1, 1612-1620

在线阅读

摘要：作为化学中的一个基本问题，逆合成旨在为目标化合物设计反应路径和中间体。人工智能（AI）辅助逆合成的目标是通过学习以前的化学反应来做出新的预测，从而实现这一过程的自动化。虽然已有多个模型证明了它们在自动逆合成方面的潜力，但仍有很大的必要进一步提高预测准确性，使其达到更实用的水平。在此，我们提出了一个名为 *LocalRetro* 的局部逆合成框架，其灵感来源于化学直觉，即分子变化主要发生在化学反应的局部。这与几乎所有现有的逆合成方法都不同，后者根据分子的全局结构推荐反应物，而这些结构往往包含了一些微小的细节，而这些细节并不是我们所期望的。



与反应直接相关。这种局部概念产生了涉及原子和化学键编辑的局部反应模板。由于远程官能团也会对整体反应路径产生次要影响，因此提出的局部编码逆合成模型将通过全局关注机制进一步完善，以考虑化学反应的非局部效应。在包含 50 016 个反应的 USPTO-50K 数据集上，我们的模型在前 1 名和前 5 名预测中的往返准确率分别达到了 89.5% 和 99.2%。我们还在包含 479 035 个反应的大型数据集（UTPTO-MIT）上进一步证明了 *LocalRetro* 的有效性，其 top-1 和 top-5 预测的往返准确率分别为 87.0% 和 97.4%。该模型的实际应用还通过正确预测来自各种文献的五种候选药物分子的合成路径得到了验证。

关键词：*逆合成反应预测、图神经网络、局部反应性、全局关注机制*

引言

设计具有所需特性的分子和材料是一项艰巨的任务。化学信息学是化学和材料科学的实用目标。化学信息学涉及利用数据来理解分子结构与分子特性之间的关系，从而最终发现新型功能分子。事实上，在化学科学中使用机器学习来加速新发现的过程已经出现了显著的趋势，即根据分子结构预测各种分子特性，或根据输入的所需功能反向设计新型分子。[1-3]然而，后一种机器支持的分子设计主要是给出符合所需特性的优化分子结构，而不考虑如何合成，这使得合成规划任务成为将硅学设计的分子付诸实践的关键最后一步。

预测化学反应一般涉及两个映射方向，要么是正向（从给定的反应物预测产物），要么是逆向（从目标产物设计适当的反应物），而术语 "逆合成 "指的就是后一种合成路径规划。正向预测一般更为直接、因为所需的任务是一对一的映射，即对于一组给定的反应物，反应产物通常是唯一确定的（在实验条件变化的范围内）。另一方面，逆合成是一种一对多的映射，而且更具挑战性，因为可能有几种不同的反应途径来合成一种目标化合物。因此，合成规划历来是合成化学家的专业领域，为了以更加自动化的方式加快和扩大逆合成，几十年来，研究人员一直在寻求基于计算机辅助合成规划（CASP）的有效而准确的方法。[4-6]例如，*Chematica*（现名 Synthia）[7]，它使用合成专家编码的 70,000 条规则，并使用决策树来决定使用哪种反应规则，并以自动化的方式惩罚非选择性反应。

收到： 2021 年 6 月 2 日
出版日期：2021 年 8 月 5 日2021 年 8 月 5 日

https://doi.org/10.1021/jacsau.1c00246

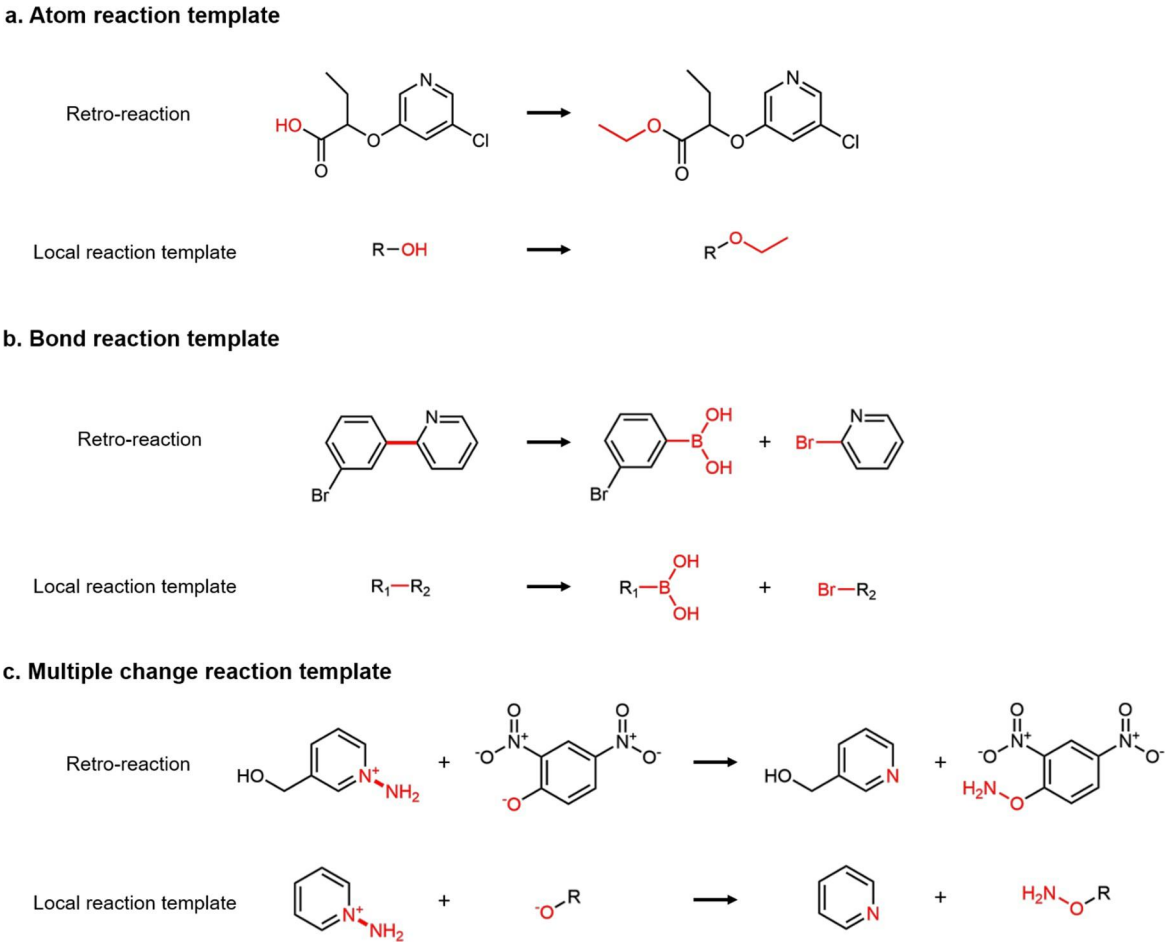


图 1.(a) 原子反应模板、(b) 键反应模板和 (c) 多重变化反应模板的推导。如果反应不涉及任何键的变化或断开，则推导出原子反应模板，否则推导出键反应模板。如果反应中原子和键都发生了变化，则导出的模板同时记为原子反应模板和键反应模板。原子和键的变化用红色标出。

在计算机中以分层方式进行。然而，自动回溯论文仍是一个尚未解决的重大问题，下面我们将简要介绍数据驱动自动合成规划的一些最新进展。

通过将分子表示为简化的分子输入线 Liu等人[8]，训练了一个神经序列到序列（seq2seq）模型，将反应产物的SMILES字符串转换为其前体的SMILES字符串。他们的完全数据驱动模型显示出与基于规则的专家系统基线模型相当的性能。另一方面，一些研究小组使用*摩根指纹*（[9]）来训练他们的模型。*摩根指纹是*一种将给定分子的亚结构信息提取为特征向量的算法。Coley 等人[10] 利用目标产物与语料库中化合物之间的化学结构相似性推导出可能的合成途径，直观地推测相似的分子会有相似的反应途径，并根据反应相似性对预测结果进行排序。Segler 和 Waller[11] 提出了一种神经-符号混合方法，利用深度神经网络手工编码或自动提取的规则预测逆合成途径。在此基础上，Segler 等人[12] 利用蒙特卡洛树搜索（3N-MCTS）实现了三种不同的深度神经网络（扩展、范围内过滤和推出），以解决多步逆合成问题。通过大量的根据用于训练模型的反应（350 万个），该模型能够为作为 5-HT$_6$ 受体拮抗剂的苯并吡喃磺酰胺衍生物的合成规划提出专业水平的建议[13]。不过，鉴于*摩根指纹图谱*用于表示分子的性质，它只包括分子上存在的亚结构信息，因此可能无法完全捕捉到在合成规划中起关键作用的亚结构的连接性和相对位置。

最近，计算机科学家将分子视为异质图，并引入了图神经网络（GNN），以提高回溯论文的预测准确性。[14−17]Dai 等人[14] 应用 *RDChiral*[18] 从训练数据集中提取了数千个反应模板，并通过条件图逻辑网络（*GLN*）预测了这些反应模板应用于给定分子的概率分布。*GLN* 预测的模板和前体的概率被表述为反应模板与给定分子的概率以及反应物与给定分子和模板的概率的函数，其中可预测的模板和反应物的数量随不同的分子条件而变化。Shi 等人的研究[15] 将逆合成任务分为两部分：反应中心识别和变异图翻译。在反应中心识别过程中，分子中的每个键都要经过

https://doi.org/10.1021/jacsau.1c00246
JACS Au 2021, 1, 1612-1620

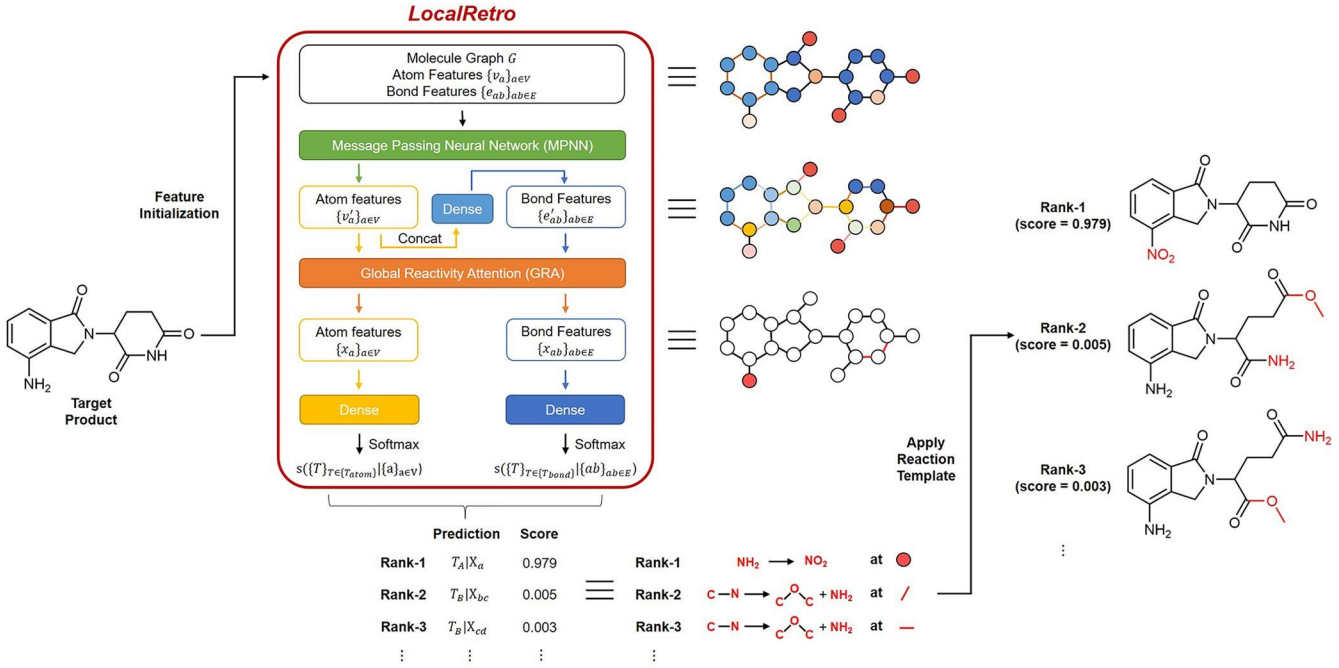


图 2.*LocalRetro* 的模型结构以及预测来那度胺反应物时的特征转换示例。首先，根据原子和化学键的属性初始化给定分子的分子图（*G*）、原子特征（v_a）和化学键特征（e_{ab}）。然后，通过消息传递神经网络（MPNN）、化学键特征编码层和全局反应注意层更新原子和化学键特征，以编码分子中原子和化学键的局部环境和非局部反应依赖性。最后，通过原子反应模板分类器和键反应模板分类器预测每个局部反应模板在每个原子和键上的得分。将预测的局部反应模板应用到预测的原子和化学键上，就得到了预测的反应物，并按其预测得分进行排序。

或预测被切割或保留。在识别和编辑反应中心后，由 GNN 完成生成分子（合成物），以生成相应的反应物。这种识别和完成概念类似于有经验的化学家如何设计给定分子的合成途径。不过，我们认为，识别步骤和完成步骤是高度相关的。换句话说，由此产生的反应物高度依赖于已识别的反应中心，因此这两个步骤应合并并同时进行。

在大多数基本化学反应中，分子式和结构因键的断裂或形成而发生的变化大多是局部的。然而，上述几乎所有现有的逆合成方法都采用目标分子的全局结构来进行预测。例如，在大多数现有的基于图的逆合成方法中，全局特征是通过对所有原子特征求和或求平均得到的，并用于预测反应物。在基于分子相似性的逆合成中，也会使用分子间的全局相似性。然而，使用全局特征进行合成途径预测可能会产生对与目标反应无直接关系的细节的关注，这是不可取的。

在这里，我们通过以下方法设计了一个基于图的逆合成框架
在本地推导反应模板，并评估这些本地模板在目标分子的所有列举的可能反应中心的适用性。所有化学变化信息都包含在局部反应模板中，因此，一旦预测出所选反应中心的正确模板，只需应用推导出的模板，就能立即得到反应物。换句话说，我们的本地方法将识别和完成两步流程合并为一锅学习。此外，我们使所有反应中心都能通过注意机制交换信息，以考虑全局情况。这与化学反应的非局部效应相对应，其中反应性有时会因为远处的化学变化而改变。

方法和数据集

由于反应物分子在反应过程中和反应后保留了大部分子结构和碎片，因此我们的方法只关注分子结构（原子或键）的变化，以完成逆合成。也就是说，我们不是从头开始寻找合适的反应物，而是重点推断在键的形成或断裂和/或原子的添加或移除方面发生了哪些局部变化以形成给定的产物。

在这项工作中，我们使用了两个反应数据库：UTPTO-50K 和 USTOP-MIT，它们分别包含 50 016 个和 479 035 个反应，所有反应的原子映射都是正确的。USPTO-50K 数据集标注了 10 种不同的反应类别。
Schneider 等人策划的美国专利文献，[19]，主要用于与其他方法进行比较，因为以前的许多方法都是以 USPTO-50K 为基准的。UTPTO-MIT数据集是由Jin等人通过去除重复和错误的反应而策划的，[20]，用于在更大的数据集上进一步推广我们的方法，以满足实际应用的需要。我们按照参考文献 10 将 USPTO-50K 数据集划分为 40K/5K/5K 训练/验证/测试。对于 USPTO-MIT 数据集，我们按照参考文献 20 将其分为 410K/40K/30K 部分。

我们首先为美国专利商标局（USPTO）--美国专利商标局（USPTO）--美国专利商标局（USPTO）--美国专利商标局（USPTO 50K 和 USPTO-MIT 反应数据库。这些局部反应模板包含反应前后原子和化学键信息的变化。由于目标产物和反应物是原子映射的，因此我们通过比较产物和反应物之间原子和键的差异来指定反应中心。原子反应模板指的是原子上发生的变化，而原子键没有变化。例如，作为原子反应模板之一，图 1a 显示了一个去保护反应，在该反应中，乙基

https://doi.org/10.1021/jacsau.1c00246
JACS Au 2021, 1, 1612-1620

氧上的基团发生反应，形成氢氧基。如果反应涉及任何键的变化，则可得出键反应模板

或断开。图 1b 显示了铃木反应的键反应模板示例，其中描述了由硼酸（R-OBO）和卤化碳（R-C）形成的碳-碳键（C-C）。
X)。在该反应中，一个 C-C 键被断开，由一个反应物中的硼酸和另一个反应物中的卤代烃取代。因此，在数据集中，每一个有单个原子变化或键变化的反应，都有一个原子反应模板或键

可以导出反应模板。对于有多个生成物或多个原子和键发生变化的反应，衍生模板包括反应过程中发生变化的所有原子和键。如果反应中的原子和化学键都发生了变化，则导出的反应模板同时记为原子反应模板和化学键反应模板。图 1c 显示了一个反应示例，其中
伯胺从一个氧原子转移到一个芳香族原子上
氮原子。因为反应发生在一个 N-N 键上
当一个分子中的一个氧原子与另一个分子中的一个氧原子发生化学反应时，该模板既是原子反 应模 板 ，又是键反应模板。关于局部反应模板的更多详情，请参阅
佐证资料。如果在训练中给出了反应类
我们可以根据标有反 应 的反应类别，将每个衍生的局部反应模板进一步归类为一个或多个反应类别。

接下来，我们开发了一个模型，通过学习每个原子和化学键的局部环境来预测导致给定产物的正确局部反应模板。LocalRetro 的整体架构和目标产物的特征转换如图 2 所示。我们将分子表示为一个异质图 $G = (V,E)$，其中 V（顶点）表示原子，E（边）
表示键。首先初始化原子和键的特征
使用 *DGL-LifeSci*[21] python 软件包对原子和化学键属性进行初始化。有关特征初始化的详细信息，请参见 "辅助信息"。为了对每个原子的周围环境信息进行编码，我们使用了文献[22]中描述的消息传递神经网络（MPNN）来更新每个原子的特征。我们用 MPNN(-) 表 示 消息传递函数，原子 a *的原子*特征为 v_a，其邻近原子 b 的原子特征为 v_b，其连接键的特征为 e_{ab}。原子 a 的原子特征由 MPNN 通过以下方式更新

$$v_a' = \mathrm{MPNN}(v_a, \{v_b\}, \{e_{ab}\}) \quad (1)$$

大括号 {} 表示给定原子周围的一组相邻原子。原子特征更新后，通过连接两个原子特征（$v_a' \| v_b'$）来表示键特征，并通过一个全连接层

$$e'_{ab} = W(v_a' \| v_b') + c \quad (2)$$

其中，W 是权重，c 是全连接层的偏置，以避免与原子 b 的符号混淆。

由于化学反应并不总是局部的，有些反应可能会因为分子内存在某些偏远的化学环境而受到影响，为了考虑反应的这种全局依赖性和非局部性，我们通过应用全局注意力机制来更新所有原子和化学键的特征。为了捕捉原子和化学键之间不同的反应关系，我们应用了多头自注意机制，即 *Transformer*[23] 中应用的注意机制，通过学习给定特征的键、查询和值来学习不同的上下文。原子和化学键的特征由分子中的所有原子和化学键共同更新。我们将非局部注意力操作称为全局反应注意力（GRA），以区别于图神经网络中常用的局部编码图注意力（GAT）[24]有关 GRA 算法的详细信息，请参阅 "辅助信息"。我们用 GRA(-) 表示用所有现有原子特征 $\{v_a'\}$ 和键特征 $\{e_{ab}'\}$ 更新的原子特征 x_a 和键特征 x_{ab}

$$x_a = \mathrm{GRA}(v_a', \{v_a'\}_{a\in V}, \{e_{ab}'\}_{ab,ab\in E}) \quad (3)$$

$$x_{ab} = \mathrm{GRA}(e_{ab}', \{v_a'\}_{a\in V}, \{e_{ab}'\}_{ab,ab\in E}) \quad (4)$$

然后，我们训练了一个原子反应模板分类器和一个键反应模板分类器，利用它们的
更新后的表示 x_a 和 x_{ab}

$$o_a = w_A^T(\sigma(W_A x_a + c_A)) \quad (5)$$

$$o_{ab} = w_B^T(\sigma(W_B x_{ab} + c_B)) \quad (6)$$

w_A 和 W_A 是原子反应模板分类器的权重，c_A 是原子反应模板分类器的偏置，其中 w_B 和 W_B 是键反应模板分类器的权重，c_B 是键反应模板分类器的偏置。σ 代表 ReLU 激活函数。

应用于原子 a 的每个原子反应模板 T 的得分由 o_a 通过 Softmax 函数转换为 s

$$s(T|a) = \mathrm{Softmax}(o_a), T \in \{T\}_{\mathrm{atom}} \quad (7)$$

同样，每个键反应模板 T 在键 ab 由 o_{ab} 通过 Softmax 函数求得

$$s(T|ab) = \mathrm{Softmax}(o_{ab}), T \in \{T\}_{\mathrm{bond}} \quad (8)$$

模型输出预测了每个化学中心的一组局部反应模板。这些预测模板按应用于给定产物的得分值进行排序，以得出最终反应物。如果在 USPTO-50K 中给出了反应类别，则我们只应用属于给定反应类别的模板池中的局部反应模板。

作为基准模型，我们将预测结果与五个最先进的逆合成模型进行了比较：*GLN*（条件图逻辑 network），[14] *G2G* (graph to graph)，[15,17] *GraphRetro*，[16] *MEGAN* (分子编辑图注意网络）和 *Augmented Transformer*[25]我们用 *LocalRetro* 来表示我们的方法，以强调其核心思想--局部反应性预测。不过，由于全局结构也起着重要的辅助作用，除非另有说明，*LocalRetro* 是指包含 GRA 注意机制的 *LocalRetro*。

结果与讨论

为了展示 *LocalRetro* 的性能，我们使用了两种精度，即精确匹配精度和往返精度。
精确匹配准确度的计算方法是，考虑规范 SMILES 中表示的反应物集是否是一个

与数据库中的基本真实反应物完全匹配。我们注意到，一些带有立体中心的基本真实反应物并 没有指定精确的立体构型，对于这些情况，预测的反应物被认为是
只要所有原子和化学键的连接性都正确，就能得出正确的 结 果 。往返精度是通过比较

使用预测的前体，将所需产品与预训练的前向合成模型预测的产品进行比较。对往返精度的考虑反映了这样一个事实，即特定的化学产品可以通过多种前体组合合成。[26]我们使用预训练的*分子前体*（MT）[27] 来评估往返准确率。具体来说，如果预测的前体与基本事实相同，或者使用 MT 预测的前体能生成目标产品，我们就将预测的前体标记为正确， 否则标记为不正确。我们还评估了对 Tetko 等人提出的 最大片段进行预测的 MaxFrag 精确度。[25] ，结果（与表 1-3 中的结果类似）见辅助信息。

从 USPTO-50K 训练集中共提取了 731 个局部反应模板。这总共 731 个局部反应模板覆盖了测试集中 98.1% 的反应，也就是说，98.1% 是精确反应模板的理论上限。
我们方法的匹配准确性。对于美国专利商标局-麻省理工学院数据集、
共获得 21 081 个局部反应模板。

https://doi.org/10.1021/jacsau.1c00246
JACS Au 2021, 1, 1612-1620

表 1.美国专利商标局-50K 数据集和美国专利商标局-麻省理工学院数据集的 Top-k 精确匹配准确率（无给定反应类别[a]

数据集	模型	top-k 准确率 (%)				
		K = 1	3	5	10	50
美国专利商标局-50K	GLN	92.4				
	G2G[15]	48.9	67.6	72.5	75.5	
	GraphRetro[16]	53.7	68.3	72.2	75.5	
	增强型变压器[25]	53.5	69.4	81.0		
	梅根[17]		85.7			
	LocalRetro w/o GRA（本作品）	48.1	70.7	78.4	86.1	93.2
	LocalRetro (this work)	49.8	75.8	84.0	91.3	97.7
美国专利商标局-麻省理工	LocalRetro（本作品）	53.4	77.5	85.9	92.4	97.7
		49.9	70.7	77.0	83.1	89.8

[a]最高精确匹配精度以粗体字标出。

表 2.USPTO-50K 数据集上给定反应类别的 Top-k 精确匹配准确率[a]

数据集	模型	top-k 准确率 (%)				
		K = 1	3	5	10	50
美国专利商标局-50K	GLN[14]	64.2	79.1	85.2	90.0	93.2
	G2G[15]	61.0	81.3	86.0	88.7	
	GraphRetro[16]	63.9	81.5	85.2	88.1	
	梅根[17]	60.7	82.0	87.5	91.6	95.3
	LocalRetro w/o GRA (this work)	62.3	86.1	91.8	96.0	97.9
	LocalRetro (this 工作)	63.9	86.8	92.4	96.3	97.9

[a]最高精确匹配精度以粗体字标出。

表 3.USPTO-50K 数据集和 USPTO-MIT 数据集的 Top-k 循环准确率[a]

数据集	模型	top-k 准确率 (%)		
		K = 1	3	5
美国专利商标局-50K	GLN[14]	88.4	95.0	97.1
	当地检索 w/o/GRA（本作品）	88.2	97.8	98.9
	LocalRetro (本作品)	89.5	97.9	99.2
美国专利商标局-麻省理工学院（本作品）	本地检索 w/o/GRA	85.7	95.7	97.3
	LocalRetro (本作品)	87.0	95.9	97.4

[a]用粗体字标出了最高 k 级往返精度。

训练集中有 40 万个反应。后 21 081 个局部反应模板覆盖了测试集中 97% 的反应，这意味着精确匹配准确率的理论上限是相同的。不过，我们注意到，由于可能存在使用不同模板的多个反应途径，因此往返准确率并不受精确匹配准确率理论上限的限制。

表 1 和表 2 显示了美国专利商标局 50K 数据集的 top-k 精确匹配准确率结果。除 Top-1 预测外，无论是否给出反应类别，LocalRetro 都优于所有其他方法。例如，在预测前 3 名时，LocalRetro 在给出和未给出反应类别的情况下，分别以 4.8% 和 6.8% 的优势超过了目前最好的方法。当出现以下情况时，LocalRetro 的预测精度会持续提高

将 GRA 应用于模型。在往返准确率方面，我们只将 LocalRetro 与 GLN 进行了比较，因为其他著作中使用的相同数据拆分所训练的模型并未公开。往返准确率结果如表 3 所示。有 GRA 和没有 GRA 的模型都远远超过了 GLN[14]，往返准确率接近 99%。

前 5 名预测的准确率。

美国专利商标局-麻省理工学院数据集的 top-k 精确匹配准确率结果如表 1 所示。虽然

尽管 USPTO- MIT 的数据集和模板数量分别是 USPTO-50K 的 10 倍和 30 倍，但前 10 名的精确匹配准确率仍然达到了 84.4%，前 50 名的精确匹配准确率达到了 90.4%，与之前在 USPTO-50K 数据集上评估的方法结果相当。 往返

在美国专利商标局-麻省理工学院数据集上评估的模型的准确性

分别为 87.0%、95.9% 和 97.4%，与在美国专利商标局-50K 数据库上训练/基准测试的 LocalRetro（即 89.5%、97.9% 和 99.2%）相当。这些结果清楚地表明，我们提出的 LocalRetro 可以在更大的数据库中进行实际应用。

如表 1-3 所示，额外的 1-2%统计

通过使用全局反应注意力（GRA），预测准确率得到了提高。虽然从数字上看， 1-2% 的改进似乎很小，但这却是一个重要的改进。这是因为数据集中并非所有反应都需要非局部效应来描述。事实上，大多数化学反应都是局部的，这就是为什么我们的基线局部编码模型在没有非局部效应的情况下已经取得了很好的结果。然而，在许多化学反应中，非局部效应仍是化学家考虑用来解释选择性的一个重要因素。因此，考虑到数据集中非局部效应确实起重要作用的一小部分反应，我们认为 1-2%（额外获得 5000-10000 个反应的正确性）是统计意义上的显著改进，在化学上也具有重要意义。此外，当逆向合成任务包括不止一种产物时，GRA 也会发挥重要作用。在美国专利商标局-麻省理工学院数据集的测试集中，由于认识到了其他分子中存在的其他原子，在包含多个产物的反应中，有 GRA 的 LocalRetro 的 top-1 精确匹配准确率比没有 GRA 的模型高出 12.3%（表 S2）。我们注意到，USPTO-MIT 数据集中总共有 471 个反应（1.2%）含有多个产物。

了解 GRA 如何帮助提高预测能力

我们比较了有 GRA 和无 GRA 的模型，并在图 3 中直观地显示了非局部注意力。在这两个例子中，有 GRA 的模型都能正确预测反应物，而没有 GRA 的模型则预测错误。特别是图 3 中用橙色标出的区域，显示了预测的反应中心与其他原子和化学键之间的非局部性。在示例 1 中，GRA 使机器意识到现有骨架中存在一个反应活性较高的氯基，因此应连接电子供体较强的氯，但在没有 GRA 的情况下，机器却建议连接氧，因此反应错误地发生在另一个反应活性较高的氯中心。同样，图 3 中的示例 2 表明，GRA 意识到现有胺的高反应性，因此预测反应中心应靠近胺，而另一个不希望发生的反应中心应靠近胺。

https://doi.org/10.1021/jacsau.1c00246
JACS Au 2021, 1, 1612-1620

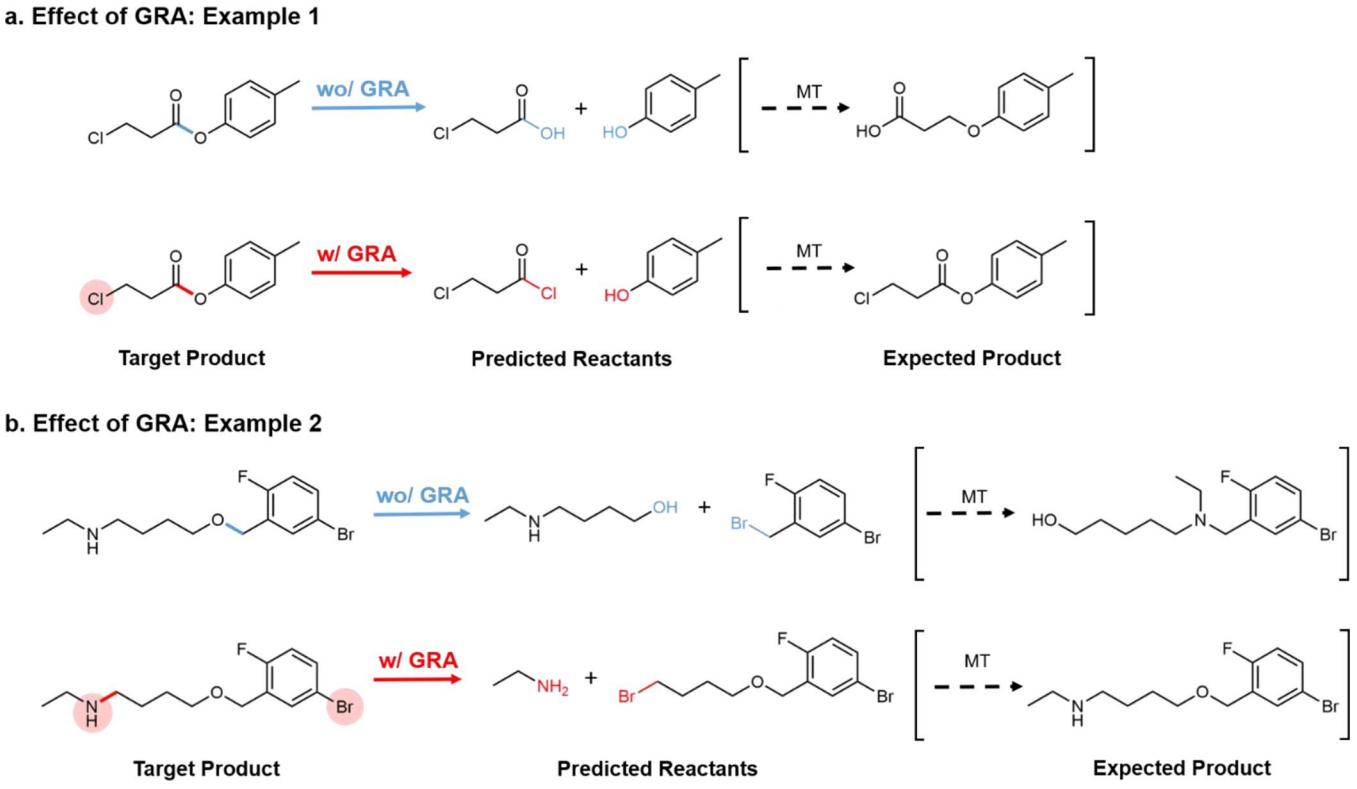


图 3.全局反应注意（GRA）的影响。在示例 1 中，预测出的反应中心与未预测出的反应中心相同；在示例 2 中，预测出的反应中心与未预测出的反应中心不同。在使用 GRA 的情况下，突出显示（出席）的化学单位用红色阴影圆圈表示。使用*分子转换器* (MT) 预测反应物的预期产物。[27]

反应可能不会发生。因此，GRA 允许算法对反应中心进行优先排序，并找到正确的反应中心，从而防止在没有全局关注机制的情况下可能形成的不想要的产物。

逆合成的最终目标是解决实际的合成规划问题，因此我们进一步验证了我们在美国专利商标局-麻省理工学院数据集上训练的模型，该数据集针对的是文献中各种逆合成和反向设计工作中考虑的五种不同的候选药物：来那度胺、沙美特罗、一种 5- HT_6 受体配体和两种 DDR1 激酶抑制剂。前两个分子是 Coley 等人展示的逆合成实例，[10]，第三个实例是 Segler 等人展示的逆合成实例，[12]，最后两个实例是基于强化学习的生成模型 *GENTRL* 提出的化合物[28]完整的逆合成路径是通过连续执行逆合成预测任务获得的。图 4 总结了与早期预测和实际实验路径的比较。

在第一个例子中，Ponomaryov 等人[29]，提出了一种通过三个合成步骤合成来那度胺的工艺，来那度胺是一种可用于治疗多发性骨髓瘤的抗癌药物。合成过程首先用 N-溴代丁二酰亚胺（NBS）进行溴化，然后在引入氨后形成环。最后，来那度胺的前体与 3-氨基哌啶-2,6-二酮缩合，再进行硝基氧化，就得到了来那度胺。我们的回顾性论文预测（图 4a）显示了与文献中完全相同的合成途径。

在第二个例子中，Guo 等人[30]通过一个关键步骤亨利反应合成了沙美特罗，这是一种强效的 β 2-肾上腺素*受体*激动剂。我们的模型预测了还原反应，然后正确预测了与秩-1 的胺化反应（图 4b）。在在接下来的两个步骤中，我们的模型表明，产物可以通过亨利反应合成，这是一个典型的碳-碳键形成反应。我们的模型预测的合成路径与文献中的路径相同，包括第 3 步和第 4 步，在秩-1 或秩-2 的预测下形成亨利反应。

在第三个例子中，Nirogl 等人[13]提出了一种苯并吡喃磺酰胺衍生物作为 5-HT_6 受体的拮抗剂。这是一个具有挑战性的合成规划问题，总共需要七个合成步骤。我们的模型所建议的合成规划如图 4c 所示。除了第 5 步的预测排名为第 7 位外，我们的模型成功预测了其余步骤（第 1 步至第 7 步，第 5 步除外），预测排名均为前 3 位。这一结果清楚地表明，即使在需要较长合成步骤序列的情况下，我们的模型也能 预测潜在药物分子的合成。

另一个具有挑战性但有趣的第四个例子是 Zhavoronkov 等人提出的 DDR1 激酶抑制剂 INS015_037 的逆合成任务。[28] 这是一个通过生成式机器学习方法获得的潜在 DDR1 激酶抑制剂（图 4d），实验证明它在小鼠体内也具有良好的药代动力学特性。Zhavoronkov 等人采用了聚合合成法，分别合成了两种前体，并在最后一步合成了 INS015_37。我们的模型成功预测了与参考文献 28 中报道的合成途径相同的会聚合成途径，预测结果为秩-1。我们对 Zhavoronkov 等人提出的另一种 DDR1 激酶抑制剂 INS015_032 的逆合成预测[28] 也显示在图 4e 中。尽管与文献相比，我们的模型能够在前 4 位的预测范围内提出几乎相同的合成途径，但它无法以较高的置信度预测第 2 步的铃木偶联反应，并且在第 3 步的预测范围内无法预测第 4 位的合成途径。

https://doi.org/10.1021/jacsau.1c00246
JACS Au 2021, 1, 1612-1620

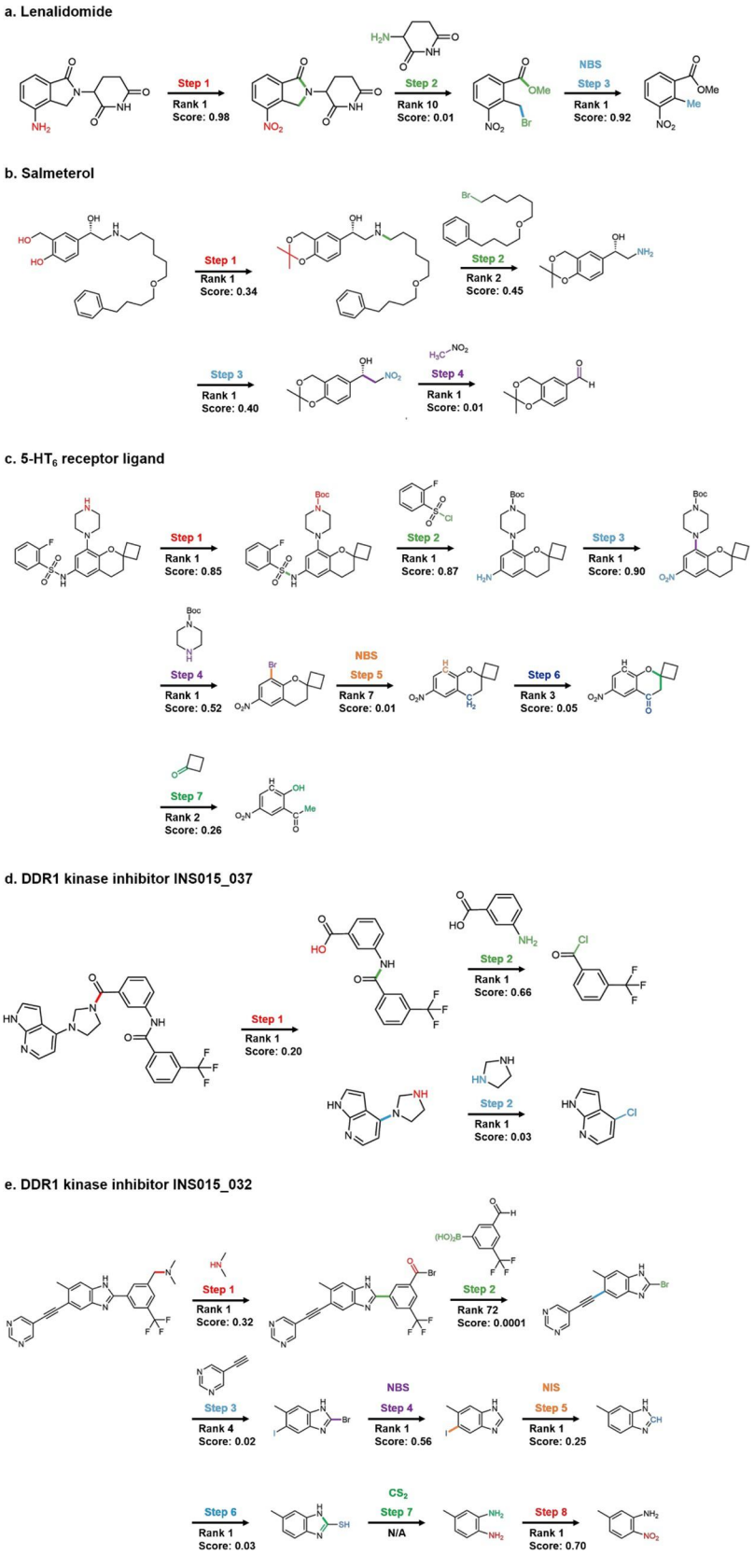


图 4.*LocalRetro* 对 (a) 来那度胺、(b) 沙美特罗、(c) 5-HT6 受体配体、(d) DDR1 激酶抑制剂 INS015_037 和 (e) 另一种 DDR1 激酶抑制剂 INS015_032 的多步逆合成预测。在不同的反应步骤中，反应中心以及原子和化学键的转化以不同的颜色突出显示。在总共 25 个单个预测合成步骤中，有 19 个步骤在排名 2 的预测范围内被正确预测。

https://doi.org/10.1021/jacsau.1c00246
JACS Au 2021, 1, 1612-1620

由于缺乏可用的反应模板，无法预测第 7 步的反应。

因此，总结这些更实用的多步骤逆合成问题的结果，所有五个演示示例都得出了与文献中几乎完全相同的逆合成路径，大部分都在秩-2预测范围内。更具体地说，在考虑的所有 25 个单独步骤中，有 5 个步骤的预测等级分别为-3、-4、-7、-10 和-72，还有一个步骤由于缺乏反应模板而无法解决，但其余 19 个步骤的预测等级均在 2 级范围内，其中 17 个步骤的预测等级为 1 级。在不考虑多种可能途径的情况下取得的这些结果，以及在前面描述的基准数据集上取得的令人鼓舞的往返性能，都表明我们的方法大有希望实现更实用的逆合成预测。

结论

受化学直觉的启发，我们提出了一种数据驱动的逆合成模型 *LocalRetro*，它通过局部学习化学反应性和全局反应性注意来考虑剩余的非局部效应，从而提出可能的合成路径。*LocalRetro* 在包含 479 035 个化学反应的美国专利商标局-麻省理工学院数据集上进行了训练和评估，获得了 97.4% 的前五名往返准确率。通过成功预测几种潜在药物分子的逆合成路径，我们进一步证明了模型的实用性。未来的工作应着眼于更大的反应数据库，以进一步推广我们的模型。这些未来的反应数据库还可能包括反应条件，如试剂、温度和 pH 值，这些都是我们在本研究中使用的当前数据集中缺乏的关键信息。随着反应映射方法的进步，[31]，我们希望将来能使用更大的高质量数据集来训练我们的模型。我们还希望这里提出的局部/非局部反应性概念可以用于正向反应产物预测模型，我们的研究小组目前正在开发这种模型。*LocalRetro* 的源代码发布于 https://github.com/kaist-amsg/LocalRetro。

相关内容

佐证资料

辅助信息可从 https://pubs.acs.org/doi/10.1021/jacsau.1c00246 免费获取。

有关局部反应模板、MaxFrag 精确度、多产物反应结果、模型实施和计算成本分析的详细信息 (PDF)

作者信息

通讯作者

Yousung Jung - *化学与生物分子工程系（BK21 four），KAIST，Daejeon 34141，South Korea*；orcid.org/0000-0003-2615-8394；电子邮件：ysjn@kaist.ac.kr

作者

Shuan Chen - *化学与生物分子工程系（BK21 four），KAIST，韩国大田 34141*

完整联系信息请访问

https://pubs.acs.org/10.1021/jacsau.1c00246

说明

作者声明不存在任何经济利益冲突。

致谢

我们感谢韩国 NRF（NRF-2017R1A2B3010176）。感谢 KISTI 慷慨的超级计算时间。感谢 Juhwan Noh 和 Geun Ho Gu 的讨论。

参考文献

(1) Gómez-Bombarelli, R.; Wei, J. N.; Duvenaud, D.; Hernández-Lobato, J. M.; Sánchez-Lengeling, B.; Sheberla, D.; Aguilera-Iparraguirre, J.; Hirzel, T. D.; Adams, R. P.; Aspuru-Guzik, A. Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules. *ACS Cent. Sci.* 2018, *4* (2), 268-276.

(2) Jin, W.; Barzilay, R.; Jaakkola, T. Junction Tree Variational Autoencoder for Molecular Graph Generation. In *International Conference on Machine Learning*; PMLR, 2018; pp 2323-2332.

(3) Segler, M. H. S.; Kogej, T.; Tyrchan, C.; Waller, M. P. Generating Focused Molecule Libraries for Drug Discovery with Recurrent Neural Networks. *ACS Cent. Sci.* 2018, *4* (1), 120-131.

(4) Corey, E. J.; Wipke, W. T. Computer-Assisted Design of Complex Organic Syntheses. *Science* 1969, *166* (3902), 178-192.

(5) de Almeida, A. F.; Moreira, R.; Rodrigues, T. Synthetic Organic Chemistry Driven by Artificial Intelligence. *Nature Reviews Chemistry* 2019, *3* (10), 589-604.

(6) Strieth-Kalthoff, F.; Sandfort, F. S.; Segler, M. H.; Glorius, F. Machine Learning the Ropes：合成化学的原理、应用和方向。*Chem. Soc. Soc. Rev.* 2020, *49* (17), 6154-6168.

(7) Mikulak-Klucznik, B.; Gołębiowska, P.; Bayly, A. A.; Popik, O.; Klucznik, T.; Szymkuć, S.; Gajewska, E. P.; Dittwald, P.; Staszewska-Krajewska, O.; Beker, W.; Badowski, T.; Scheidt, K. A.; Mołga, K.; Mlynarski, M.; Staszewska-Krajewska, O.; Beker, W.; Badowski, T.; Scheidt, K. A.; Mołga, K.; Mlynarski, J.; Mrksich, M.; Grzybowski, B. A. Computational 规划复杂天然产品的合成。*自然* 2020, *588* (7836), 83-88.

(8) 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.

(9) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* 2010, *50*, 742-754.

(10) 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.

(11) Segler, M. H. S.; Waller, M. P. Neural-Symbolic Machine Learning for Retrosynthesis and Reaction Prediction. *Chem.- Eur. J.* 2017, *23* (25), 5966-5971.

(12) 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.

(13) Nirogi, R. V. S.; Badange, R.; Reballi, V.; Khagga, M. Design, Synthesis and Biological Evaluation of Novel Benzopyran Sulfonamide Derivatives as 5-HT6 Receptor Ligands. *Asian J. Chem.* 2015, *27* (6), 2117-2124.

(14) Dai, H.; Li, C.; Coley, C. W.; Dai, B.; Song, L. Retrosynthesis Prediction with Conditional Graph Logic Network. *ArXiv (Machine Learning)* Jan 6, 2020, 2001.01408, ver.1. https://arxiv.org/abs/2001.01408 (accessed 2021-03-01).

(15) Shi, C.; Xu, M.; Guo, H.; Zhang, M.; Tang, J. A Graph to Graphs Framework for Retrosynthesis Prediction. *ArXiv (Machine Learning)* Mar 28, 2020, 2003.12725, ver.1. http://arxiv.org/abs/2003.12725（访问日期：2021-03-01）。

(16) Somnath, V. R.; Bunne, C.; Coley, C. W.; Krause, A.; Barzilay, R. 学习逆合成预测的图模型。

(*机器学习*）2020 年 6 月 12 日，2006.07038，ver.2. http://arxiv.org/abs/2006.07038 （访问日期：2021-07-30）。
(17) Sacha, M.ła.; Błaz, M.ła.; Byrski, P.; Dabrowski-Tumanski, P.ł; Chromiński, M.ła.; Loska, R.ł; Włodarczyk-Pruszynski, P.ł; Jastrzebski, S.ła.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.
(18) 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.
(19) Schneider, N.; Stiefl, N.; Landrum, G. A. What's What: The (Nearly) Definitive Guide to Reaction Role Assignment. *J. Chem. Inf. Model.* 2016, *56* (12), 2336-2346.
(20) Jin, W.; Coley, C.; Barzilay, R.; Jaakkola, T. 用 Weisfeiler-Lehman 网络预测有机反应结果。*神经信息处理系统进展* 2017，30.
(21) awslabs/dgl-lifesci https://github.com/awslabs/dgl-lifesci (accessed 2021-02-10)。
(22) Gilmer, J.; Schoenholz, S. S.; Riley, P. F.; Vinyals, O.; Dahl, G. E.量子化学的神经消息传递. *arXiv*（*机器学习*）2017 年 6 月 12 日，1704.01212，ver.2. http://arxiv.org/abs/1704.01212 （访问日期：2021-03-01）。
(23) Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; Polosukhin, I. Attention Is All You Need. *ArXiv (Machine Learning)* Dec 5, 2017, 1706.03762, ver.2. http://arxiv.org/abs/1706.03762（访问日期：2021-03-01）。
(24) Veličković, P.; Cucurull, G.; Casanova, A.; Romero, A.; Liò, P.; Bengio, Y. Graph Attention Networks. *ArXiv (Machine Learning)* Feb 4, 2018, 1710.10903, ver.3. https://arxiv.org/abs/1710.10903 (访问日期：2021-03-01）。
(25) 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.
(26) Schwaller, P.; Petraglia, R.; Zullo, V.; Nair, V. H.; Haeuselmann, R. A.; Pisoni, R.; Bekas, C.; Iuliano, A.; Laino, T. 使用基于变压器的模型和超级图探索策略预测逆合成途径。*化学科学* 2020》，*11* (12)，3316-3325。
(27) 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.
(28) Zhavoronkov, A.; Ivanenkov, Y. A.; Aliper, A.; Veselov, M. S.; Aladinskiy, V. A.; Aladinskaya, A. V.; Terentiev, V. A.; Polykovskiy, D. A.; Kuznetsov, M. D.; Asadulaev, A.; Volkov, Y.; Zholus, A.; Shayakhmetov, R. R.; Zhebrak, A.; Minaeva, L. I.; Zagribelnyy, B. R.; Volkov, Y.; Zholus, A.; Shayakhmetov, R. R.; Zhebrak, A.; Minaeva, L. I.; Zagribelnyy, B. A.; Lee, L. H.; Soll, R.; Madge, D.; Xing, L.; Guo, T.; Aspuru-Guzik、A. 深度学习可快速识别强效 DDR1 激酶抑制剂。*Nat. Biotechnol.* 2019, *37* (9), 1038-1040.
(29) Ponomaryov, Y.; Krasikova, V.; Lebedev, A.; Chernyak, D.; Varacheva, L.; Chernobroviy, A. Scalable and Green Process for the Synthesis of Anticancer Drug Lenalidomide. *Chem. Heterocycl. Compd.* 2015, *51* (2), 133-138.
(30) Guo, Z.-L.; Deng, Y.-Q.; Zhong, S.; Lu, G. Enantioselective Synthesis of (R)-Salmeterol Employing an Asymmetric Henry Reaction as the Key Step. *Tetrahedron: Tetrahedron: Asymmetry* 2011, *22* (13), 1395-1399.
(31) 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》。*Science Advances* 2021，*7* (15)，No. eabe4166.

https://doi.org/10.1021/jacsau.1c00246
JACS Au 2021, 1, 1612-1620