节点对齐图到图（NAG2G）： 提升单步逆合成中的无模板深度学习方法

Lin Yao,† Wentao Guo,‡,¶,†,§ Zhen Wang,†,§ Shang Xiang,† Wentan Liu,† and Guolin Ke*,†

†DP 技术，中国北京

‡美国加州大学戴维斯分校化学系，美国加利福尼亚州，95616

¶Department of Statistics, University of California, Davis, California, 95616, USA

§ 平等缴款

电子邮件： kegl@dp.tech

摘要

有机化学中的单步逆合成（SSR）越来越多地受益于计算机辅助合成设计中的深度学习（DL）技术。虽然无模板 DL 模型在逆合成预测方面具有灵活性和前景，但它们往往忽略了重要的二维分子信息，并且在节点生成的原子对齐方面存在困难，导致其性能低于基于模板和半模板的方法。为了解决这些问题，我们引入了节点对齐图-图（NAG2G）--一种基于转换器的无模板 DL 模型。NAG2G 将二

维分子图和三维构象结合起来，保留了全面的分子细节，并通过节点对齐将产物-反应物原子映射纳入其中，从而以自动回归的方式确定逐节点图输出过程的顺序。

通过严格的基准测试和详细的案例研究，我们证明了 NAG2G 在庞大的 USPTO-50k 和 USPTO-FULL 数据集上出色的预测准确性。此外，该模型还成功预测了多种候选药物分子的合成途径，这充分证明了它的实用性。这不仅证明了 NAG2G 的稳健性，还证明了它在未来的合成路线设计任务中彻底改变复杂化学合成过程预测的潜力。

导言

单步逆合成（SSR）[1] 是有机化学中的一项基本操作，涉及在单一步骤中逆向合成目标产物或中间体。为了实现自动、系统的多步合成路线设计，SSR 在构建每个分离阶段的模块方面发挥着至关重要的作用。通常，逆合成策略的设计需要对有机化学原理（如反应机理和反应位点）有透彻的理解和认识。随着计算机辅助合成规划工具的出现，再研究人员认识到深度学习（DL）技术的巨大潜力，目前正在利用这种技术来完成这项任务。

为了适应 SSR 任务学习反应的目的，人们开发并改进了各种 DL 架构。[2] 尽管它们的网络结构和数据表示格式存在显著差异，但主要分为两类：依赖模板和独立模板。在下一节中，我们将简要介绍近期基于 DL 的方法，重点介绍其模型设计及其优缺点。

模板或非模板？有机合成化学家的化学直觉是从反应规则知识中积累起来的。自然，现有反应的字典或所谓模板（如有机合成教科书）将成为 SSR 设计的圣经。因此，最初一代的逆合成工具被训练为搜索

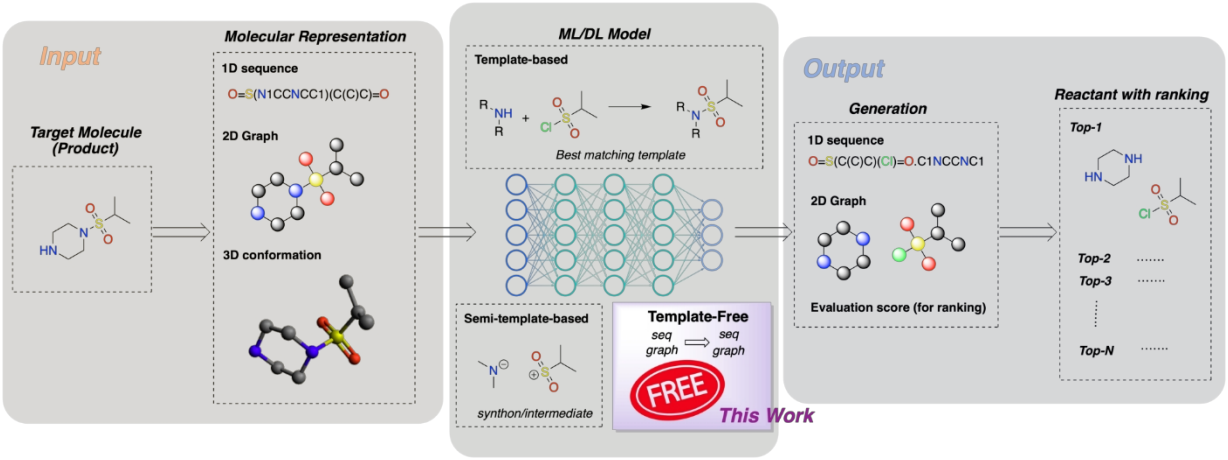


图 1：基于模板、半模板和无模板设计的计算机辅助 SSR 工作流程概览。

库中最有可能用于生成所需产物的反应模板。例如，程序 *Synthia*（以前的名称为 *Chematica*）[3] 采用了合成专家精心设计的 8 万多条规则，根据庞大的决策树确定适当的反应步骤。深度学习策略，如 *RetroSim*[4] 和 *NeuralSym*、[5] 使用传统的分子相似度指标，如指纹和谷本相似度，来寻找与产品匹配度高的模板。其他当代方法包括 *LocalRetro*、[6] *GLN*[7] 和 *RetroComposer* 等。[8] 基于模板的方法的局限性在于其所依赖的库。库可能无法涵盖所有潜在的反应，而且错综复杂的产品和模板结构之间可能存在不正确的关联。为了避免过度依赖字典，半模板方法应运而生。这种方法将 SSR 预测过程分为两个阶段--合成物或中间体检测，然后再生成反应物。[9] 或中间体检测，然后生成反应物。这两个步骤都很关键，因为合成子的预处理和识别与反应物预测直接相关。两阶段方法的优 势 包括合成物理解、搜索能力扩展和反应方案探索。同时，错误很容易从第一步转移到第二步。随着半模板方法技术的发展，一些模型包括

G2G、[10] *RetroXpert*、[11] *RetroPrime*[12] *GraphRetro*[13] *SemiRetro*、[14] *G2Retro*、[15]

Graph2Edits[16] 的出现，突显了图到图模型与分子拓扑编辑的兼容性，我们稍后将讨论这一点。

在没有科学家提供任何先决知识（包括字典、模板、合成物、中间体和编辑策略）的情况下，深度学习模型能否学习化学？答案是肯定的。无模板模型，如开创性的 *seq2seq* 、[17] *SCROP* 、[18] *绑定转换器*[19] *Augmented Transformer* 、[20] 和 *RetroDCVAE*[21] 都将逆合成视为一个预测问题。其基本思想认为，可以用类似于自然语言处理（NLP）任务的方式来分析分子。在这一框架中，产品分子根据其一维（1D）字符串表示法（如简化分子输入行输入系统（SMILES））被分解成标记。通过这种标记化方法，我们可以将产物转化为反应物，从而将化学反应与语言翻译过程相提并论。最近的研究通过应用先进的 NLP *Transformer*[22] 模型取得了显著的改进，该模型采用了多头关注机制。这种机制使模型能够对输入数据的不同片段赋予不同程度的重要性，从而增强了模型管理分子内每对原子之间以及生成物和反应物之间信息传递过程的能力。

分子表征的选择 为了帮助计算机像化学家一样思考，必须将反应信息，特别是分子级反应物和生成物，转化为*硅学*"语言 "或所谓的分子表征。基于 DL 的 SSR 模型的一种流行方法[18–21,23] 是采用一维序列，如 SMILES。尽管简单，但基于一维序列的模型有几个局限性：1) 序列忽略了大量的分子拓扑信息；[24–26] 2) 合法的 SMILES 遵循复杂的语法规则，这增加了生成有效 SMILES 的难度；3) 有效利用生成物和反应物

之间的原子映射信息对一维表示法来说具有挑战性。如果不对齐，模型性能可能会因生成物和反应物之间原子相关性的丢失而下降。4) 由于一个分子可以有多个 SMILES

当为一个产品生成多个候选反应物时，有可能生成代表同一反应的多个反应物SMILES，这可能会减少候选反应物的多样性。[20]

为了克服一维序列的局限性，有人提出了包含原子（节点）和键（边）拓扑结构的二维分子图模型，用于 SSR 任务中的分子表示。[10,15,16,27–29] 二维分子图囊括了原子环境的大量信息，如相邻原子及其连接。图拓扑结构为半模板模型下涉及合成子和中间体修饰的两步任务提供了最佳解决方案。为了有效利用生成物和反应物中原子间的自然映射信息，之前的方法采用了重复图编辑策略、[27] 在这种方法中，通过编辑操作（如添加节点、删除节点、更新节点或边）反复修改生成物的输入图，直到达到终点--反应物。例如，半模板模型，如 *G2Retro*、[28] *MARS* 、[29] *Graph2Edits* [16] 等半模板模型采用图编辑策略[27] 进行图形生成。然而，二维图形编辑需要对编辑操作和编辑类型额头进行精细安排。尽管先进的生成过程降低了计算成本，但在每个编辑步骤中都需要图形输入和输出的迭代动作-预测循环，继续加重了计算负担。

边界和突破边界。通过将不同的方法与最新的机器学习设计相结合，模型得到了改进，这一点我们之前已经讨论过。尽管仍存在一些挑战，但我们看到了基于 Transformer 的新模型的巨大潜力，这种模型可以在不使用模板的情况下进行预测。[17,30] 尽管依赖模板的模型占了上风，但我们观察到，无模板方法不仅在结构整齐简洁方面表现出色，而且还能捕捉到化学推理本身的细微差别。为了向模型提供更多信息

，需要对分子表示和注意机制（考虑原子间的长程依赖关系）进行调整。例如，*GET*

[25] 合并了图形和

SMILES 编码器。*GTA*[26] 将拓扑数据与注意力偏差结合起来。具体来说，*Retro- former*、[23] 使用一维序列作为编码器，结合了产物-反应物原子排列，以获得更好的结果。[31] 基于图形的无模板方法，包括 G2GT、[32] 等基于图的无模板方法在利用图拓扑方面取得了进步。不过，它们还没有利用节点配准策略来提高性能。

为了充分利用无模板方法的优势并解决上述局限性，我们开发了 NAG2G，它同时利用二维图形和三维坐标，提高了图形生成的效率，并根据适当的原子映射进行节点对齐，如图 1 所示。此外，我们从语言生成技术中汲取灵感，实施了一种自动回归方法，可根据对齐顺序逐节点生成图形。NAG2G 使用两个广受认可的数据集 USPTO-50k[33] 和 USPTO-Full[7,20] 与现有模型相比，NAG2G 显示出强大的能力。此外，我们的模型通过迭代生成候选药物的逐步合成路径，证明了它在解决实际问题方面的能力。为了深入了解我们方法中每个组成部分的重要性，我们进行了模拟研究，系统性地省略了模型的某些部分，以评估它们对性能的影响。

方法

模型构造

编码器是神经网络的组成部分，用于处理输入产品并将其压缩为紧凑的表示形式，在基于 NAG2G 变换器的编码器-解码器架构中，**编码器**扮演着学习分子表示形式的关

键角色。有能力的模型，如 *Graphormer*[34] 和 *Uni-Mol*[35] 等模型已经证明了编码器表征学习策略的效率。因此，我们采用了 *Uni-Mol+ 的*编码器。[36] 的编码器，如图 2 所示，该编码器结合了二维图形和三维构象进行分子表征。图 2

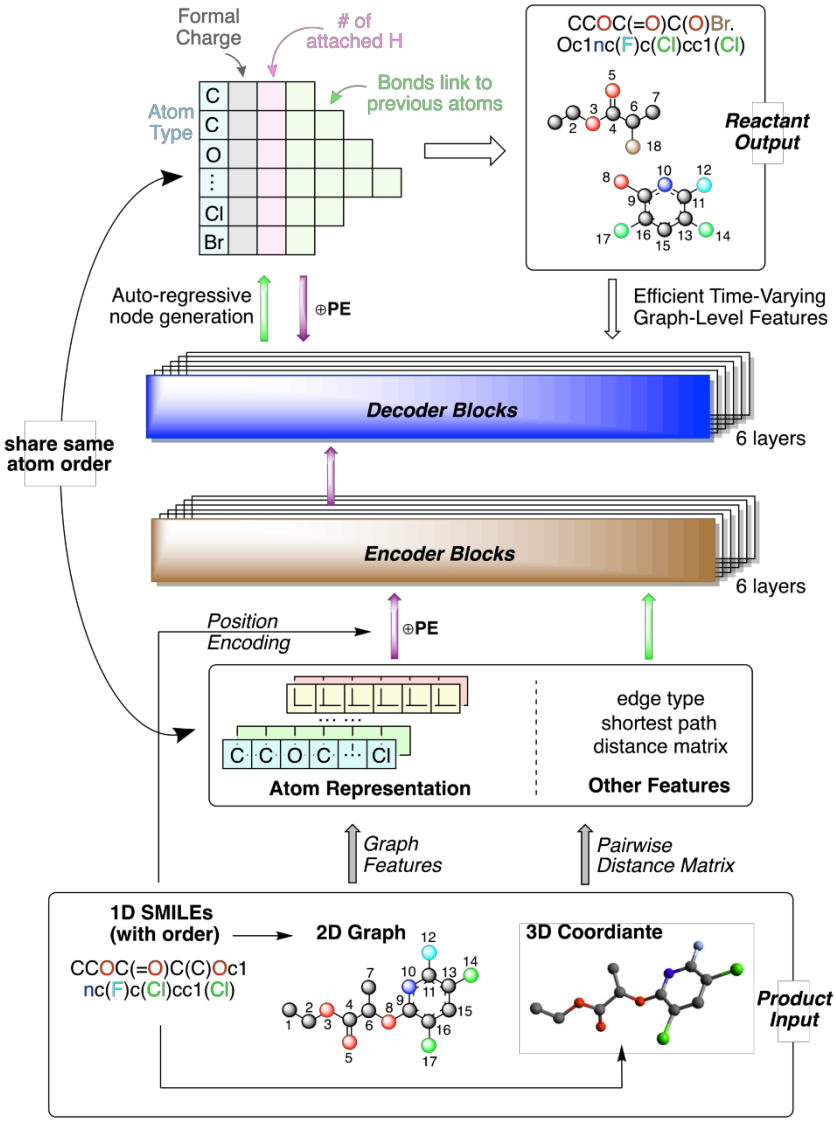


图 2：NAG2G 的网络架构。

还考虑到位置编码，作为节点顺序编码器。从形式上看，我们可以用下式（1）来表示编码器的过程：

$$\mathbf{O}^{\text{enc}} = f_{\text{enc}}\,(\mathbf{X},\ \mathbf{P}^{\text{enc}}\ ,\mathbf{E},\ \mathbf{R};\ \theta^{\text{enc}}\ )、 \tag{1}$$

在提议的表述中：**X 表示**原子特征；$\mathbf{P}^{\text{enc}}$ 表示一维空间编码，是原子嵌入的补充；**E** 表示二维图结构固有的边特征；**R** 对应三维构象中的原子坐标；θ^{enc} 封装编码器的可学习参数，$\mathbf{O}^{\text{enc}}$ 是编码器得出的分子表示结果。

解码器主要是通过自动回归法逐个节点生成反应物图。在*第 i 个*时间步，也就是*第 i 个*生成节点（原子）时，解码器会收到三个不同的输入：

1) 编码器的输出，包括有助于编码器和解码器之间交互的键和值。

2) 解码器的输出来自之前的步骤（从 1 到 *i*-1），这是典型的自回归模型，因为新值的预测是基于之前的值。在迭代过程中，增加了一维位置编码，这对 NAG2G 对齐编码器（产物）输入和解码器（反应物）输出之间的原子顺序至关重要。

3) 当前输出图的图层特征，如节点度和节点间的短 est 路径。将这些图层特征直接整合到模型中会带来效率上的挑战，因为图层特征会随时间步长而变化。为了解决这个问题，我们提出了一种整合这些图层特征的高效方法。

根据上述输入，解码器在*第* i *个*时间步自动生成一个新节点，从原子类型开始，

然后是相关的形式电荷、相连氢原子的数量，最后是与先前节点（原子）相连的边（键的类型）。每个节点的信息都是根据上述预测依次生成的。例如，形式电荷是根据先前的原子类型预测得出的。原子

过程表示为

$$
\begin{aligned}
t_i &= f_{\text{dec}}(\mathbf{P}^{\text{dec}}_{1:i}, \mathbf{N}_{1:i-1}, \mathbf{G}_{1:i-1}, \mathbf{O}^{\text{enc}}; \theta^{\text{dec}}), \\
c_i &= f_{\text{dec}}(t_i, \mathbf{P}^{\text{dec}}_{1:i}, \mathbf{N}_{1:i-1}, \mathbf{G}_{1:i-1}, \mathbf{O}^{\text{enc}}; \theta^{\text{dec}}), \\
h_i &= f_{\text{dec}}(c_i, t_i, \mathbf{P}^{\text{dec}}_{1:i}, \mathbf{N}_{1:i-1}, \mathbf{G}_{1:i-1}, \mathbf{O}^{\text{enc}}; \theta^{\text{dec}}), \\
e_{i,1} &= f_{\text{dec}}(h_i, c_i, t_i, \mathbf{P}^{\text{dec}}_{1:i}, \mathbf{N}_{1:i-1}, \mathbf{G}_{1:i-1}, \mathbf{O}^{\text{enc}}; \theta^{\text{dec}}), \\
&\ldots \\
e_{i,k} &= f_{\text{dec}}(e_{i,k-1}, \ldots, e_{i,1}, h_i, c_i, t_i, \mathbf{P}^{\text{dec}}_{1:i}, \mathbf{N}_{1:i-1}, \mathbf{G}_{1:i-1}, \mathbf{O}^{\text{enc}}; \theta^{\text{dec}}),
\end{aligned}
\tag{2}
$$

其中，$\mathbf{N}_{1:i-1}$ 表示前 i - 1 个时间步骤生成的节点集，$\mathbf{P}^{\text{dec}}$ 表示当前 i *个节点的*一维位置编码，$\mathbf{G}_{1:i-1}$ 表示从先前输出中提取的图特征，θ^{dec} 表示解码器的参数。*第 i* 个节点的原子类型、相关形式电荷和连接的氢原子数分别用 t_i 、c_i 和 h_i 表示。*第* d 条边（用 $e_{i,d} = (j, b)$ 表示）连接*第* i *个*节点和*第* j *个*节点，其键类型为 b。为了确定一条边的生成顺序，与 1D 位置较大的节点相连的边会优先生成。如果节点的电荷为零或没有连接氢原子，则跳过 c_i 和 h_i 的生成，以减少生成步骤。图 2 显示了 NAG2G 的架构概览。

节点对齐和数据扩充

分子图与句子不同，缺乏固有顺序，因为分子中的原子在分配之前没有自然顺序。为了避免在图形生成过程中考虑节点顺序的需要，一些方法将图形生成任务转化为间接方法，如 前面讨论过的图形编辑动作预测或 SMILES 预测任务。另外，一些方法还采用了 "一网打尽 "方案，在一个步骤中生成整个图输出。虽然这种方案避免了对生

成顺序的考虑，但缺乏灵活性，不适合逆合成等多解决方案任务。要采用更灵活的自动回归方法，必须确定输出节点的顺序。

挖掘。一个简单的解决方案是使用典型的 SMILES 原子顺序；然而，这种固定顺序限制了输出数据的扩充，并限制了模型的性能。因此，设计一种稳健而灵活的策略来处理图节点的无序性仍然是图生成任务中的一项艰巨挑战。无序节点不仅给图形生成带来了挑战，也给编码器的输入数据扩增带来了挑战。由于输入图本来就缺乏序列，因此图数据增强必须依靠其他策略，如省略某些节点或边的信息。然而，这种方法可能不适合逆合成，因为省略关键信息（如不同的反应位点）可能导致输出结果大相径庭。在这种情况下，使用训练集中的反应物可能会带来误差。因此，更合适的编码器输入数据增强策略可能是基于节点顺序，从而确保结果更加准确可靠。

为了应对输入和输出数据扩充方面的挑战，并实现灵活的逐节点自动回归生成，我们提出了一种基于产品-反应物节点对齐的新方法。我们的方法首先由 RDKit 随机生成产品的 SMILES 序列、[37] 如图 3 所示。按照 SMILES 序列中的新顺序，我们得到了数据增量输入图的节点序列顺序。随后，使用位置嵌入标记图节点顺序。如图 3 和图 4 所示，对于已确定顺序的产品图，我们建立了一条唯一且明确的规则，该规则与逐节点输出的反应物节点顺序相对应。在反应物中，原子可分为两类：与生成物共享的原子和反应物独有的原子。原子顺序的分配应考虑这两个方面。首先，在生成顺序时，我们规定反应物中共享原子的顺序应优先于非共享原子的顺序。为了确保对于特定的有序产品输入，存在唯一对应的有序反应物，反应物中共享原子的顺序应遵循

产品中的顺序。随后，使用 RDKit 将反应物 SMILES 与生成物 SMILES 对齐，以获得最模拟的反应物 SMILES。

相似的 SMILES。最后，从对齐的反应物 SMILES 中提取非共享原子的顺序，确保非共享原子顺序的唯一性。这种方法利用了产品-反应物对齐信息，确保节点生成顺序与训练过程中的输入图顺序一致，并允许在输入和输出中进行一致的等变量数据扩增，从而提高了生成过程的整体稳健性和准确性。通过使用 NAG2G，我们提供了一种稳健且适应性强的自动回归生成程序，它能有效处理分子图的复杂性，提高图到图生成模型的性能，并为输入输出数据增强提供了一种简洁、深刻且有说服力的解决方案，确保了逻辑高效的逐节点自动回归生成。

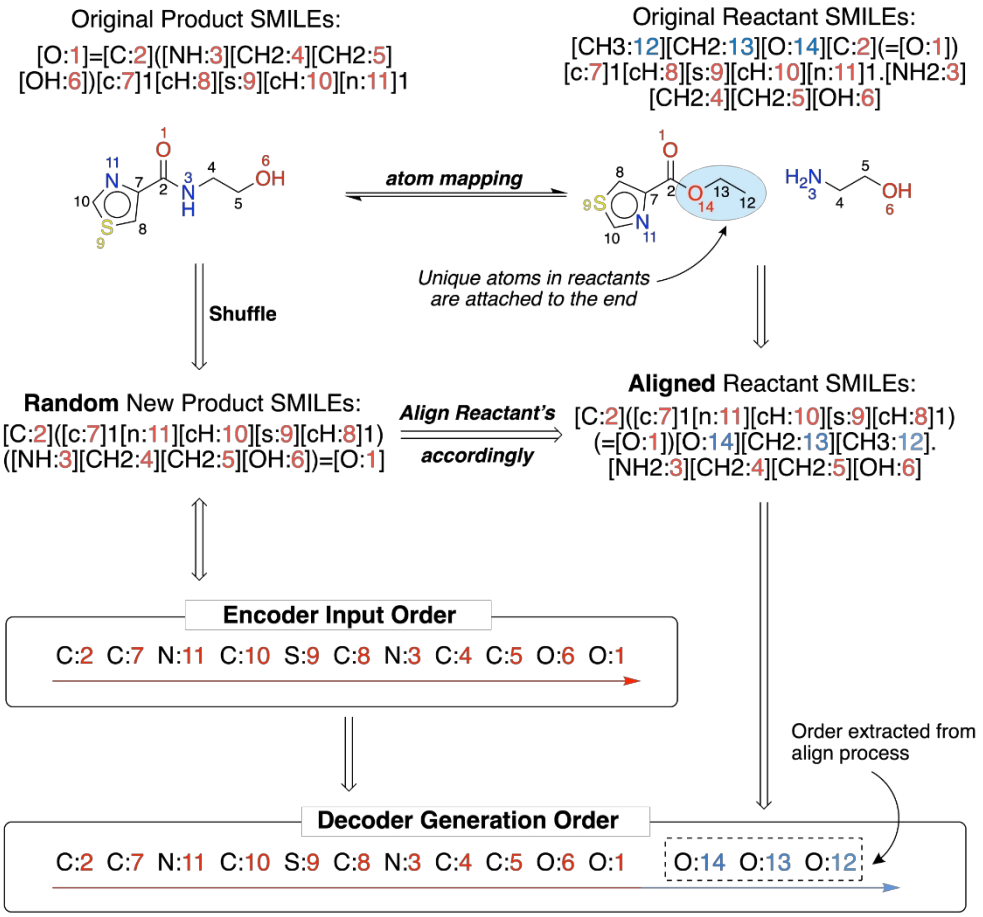


图 3：举例说明数据扩增和产品-反应物配准过程。红色数字表示同时存在于生成

物和反应物中的原子，蓝色数字表示仅存在于反应物中的原子。

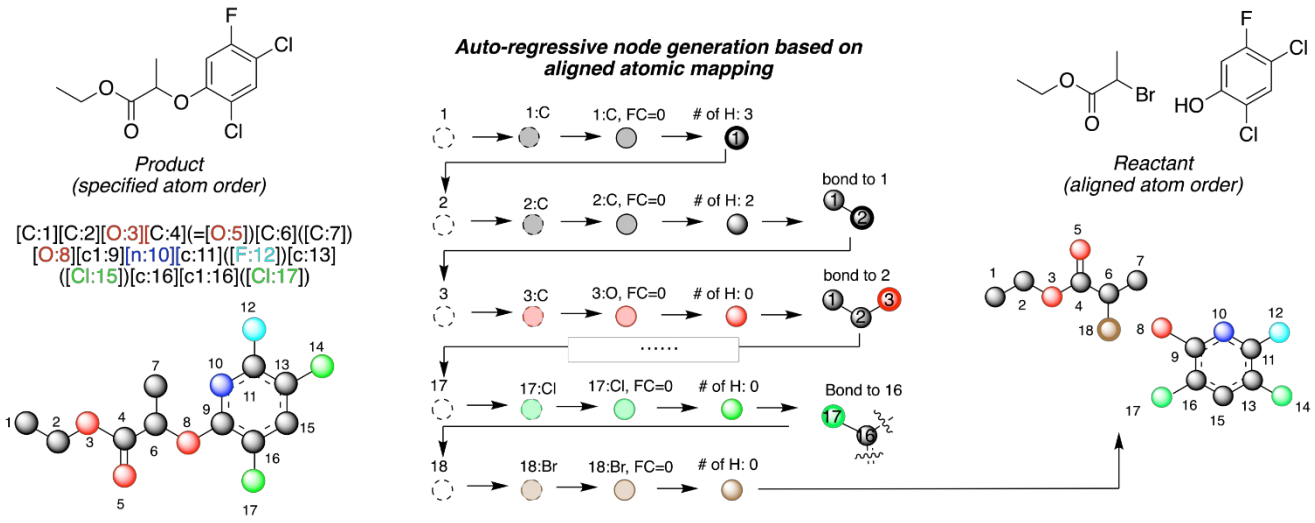


图 4：节点对齐的图到图生成示意图。

高效的时变图形级特征

在通过解码器生成数据的过程中，通过在训练过程中实施教师强制，可以将前一时间步骤的真实输出作为当前步骤的输入，而不是模型自身的预测。这种技术不仅能使模型的学习与正确的输出序列保持一致，还能并行处理不同时间步骤的数据。解码器的注意力层负责处理当前时间步 骤和之前时间步骤之间的交互。为避免无意中包含未来的信息，变压器模型中的注意力矩阵采用上三角矩阵屏蔽。这就确保了特定时间步的给定输出只能受到序列中前面元素的影响，从而保留了自回归特性，即每一步的预测只能以已知的过去信息为条件。在形式上，我们将这一过程表示为

$$\text{注意}（\mathbf{Q}、\mathbf{K}、\mathbf{V}）=\text{软最大值}\quad \frac{\mathbf{Q}\mathbf{K}^{T}}{\sqrt{d_h}}+\mathbf{M}\quad \mathbf{V}, \tag{3}$$

其中 $\mathbf{Q}$、$\mathbf{K}$、$\mathbf{V} \in \mathbf{R}^{n \times d_h}$ 分别代表查询、键和值矩阵。$_h$ 是一个头部的维度。n 是时间步数。$\mathbf{M}$ 是加法掩码

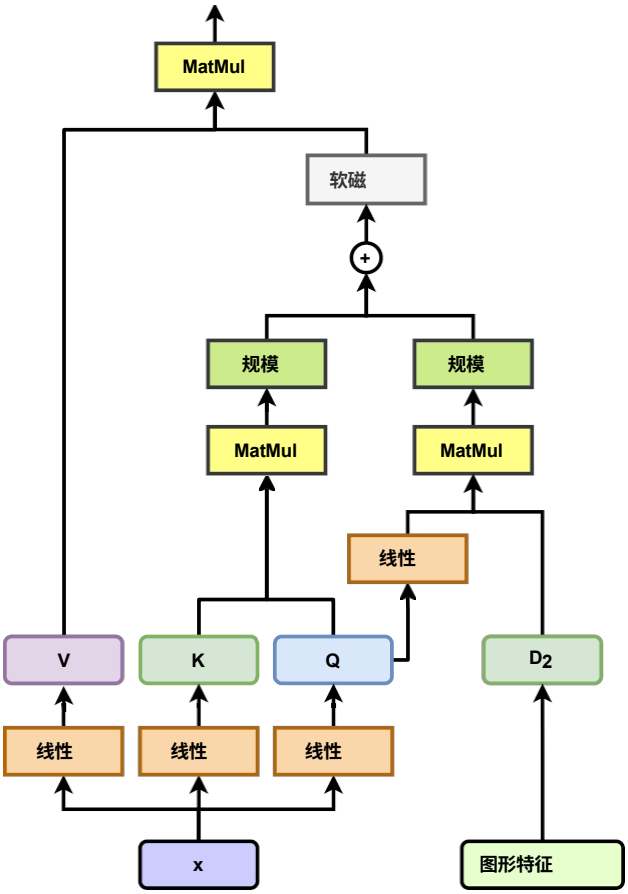


图 5：解码器注意力机制示意图。

矩阵，确保在计算注意力时只考虑当前和之前时间步骤的相关信息。为简单起见，我们只介绍单头的计算。多头注意力进程并行执行上述单头计算。在单头计算过程中，计算复杂度为 O($n \times n \times d_h$)，内存消耗峰值为 O($n \times n$)。

如前所述，图层面的特征会随时间步长而变化，直接利用这些特征对模型训练的效率提出了挑战。具体来说，为了维护随时间变化的图特征，需要一个形状为 $n \times n \times d$ 的矩阵 $_h^1$.然后将这些时变图形特征用作位置编码的加法。因此，注意力层可以表示为

$$\text{注意}（\mathbf{Q}、\mathbf{K}、\mathbf{V}、\mathbf{D}）= \text{软最大值}\left(\frac{\mathbf{Q}(\mathbf{K}+\mathbf{D})^T}{\sqrt{d_h}}+\mathbf{M}\right)（\mathbf{V}+\mathbf{D}）、 \tag{4}$$

其中，$\mathbf{D}\in R^{n\times n\times d_h}$ 表示时变图特征，**Q**、**K**、**V** 的形状重塑为 $n \times 1 \times d_h$ 进行广播。在此过程中，虽然计算复杂度保持不变，但内存消耗峰值增加到 O($n \times n \times d_h$)。考虑到 d_h 通常为 32 甚至更大，峰值内存消耗的这种显著增加在实际应用中是不切实际的。

为了降低成本，我们首先从 **V+D** 项中删除 **D**。然后，由于 $\mathbf{Q}(\mathbf{K}+\mathbf{D})^T = \mathbf{Q}\mathbf{K}^T + \mathrm{Q}\mathrm{D}^T$ 项的成本瓶颈在 $\mathbf{Q}\mathbf{D}^T$ 。因此，我们可以通过替换 **D 来**减少最后一个维度的大小：

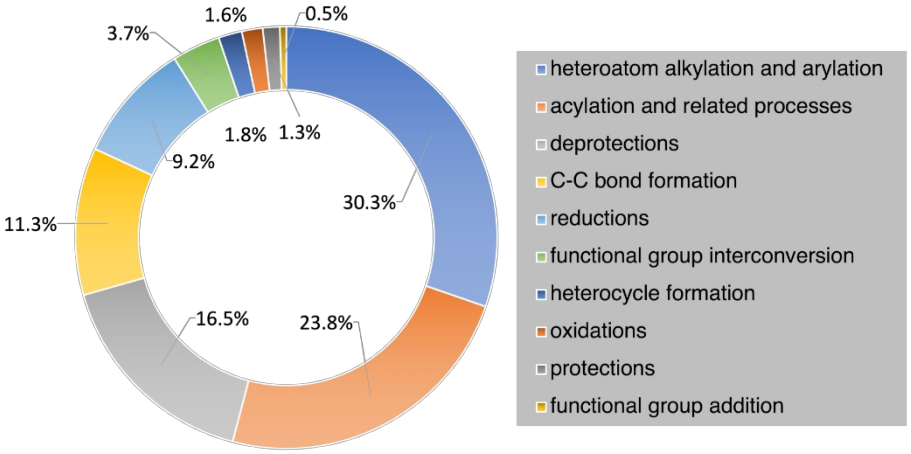
$$\text{注意}(\mathbf{Q}、\mathbf{K}、\mathbf{V}、\mathbf{D_2}) = \text{softmax}\left(\frac{\mathbf{QK}^T}{\sqrt{d_h}} + \frac{\mathbf{QUD_2}^T}{\sqrt{\text{dh2}}} + \mathbf{M}\right)\mathbf{V}, \tag{5}$$

其中 $\mathbf{U} \in \mathrm{R}^{d_h \times d_{h2}}$ 用于减少 **Q 的**维数，而 $\mathbf{D}_2 \in \mathrm{R}^{n \times n \times d_{h2}}$ 表示维数更小的时变图特征 d_{h2}。在这种配置下，[1] 这里，我们考虑节点维图特征，如节点度。成对图特征，如节点度。最短路径的内存消耗要大得多。

的峰值内存减少到 O($n \times n \times d_{h2}$)。图 5 展示了针对时变图形特征的自我关注层设计。

数据准备

NAG2G 在两个广受认可的数据集 USPTO-50k[33] 和 USPTO-Full。[7,20] USPTO-50k 包含 50,016 个原子映射反应，分为 10 个反应类别。USPTO-50K 数据集分为 40,008 个、5,001 个和 5,007 个反应，分别用于训练集、验证集和测试集。我们还使用了 Tetko 等人描述的经过过滤的 USPTO- Full 数据集，其中包含约 100 万个原子映射反应。[20] 所描述的约 100 万个原子映射反应的过滤 USPTO- Full 数据集，而不是原始的 USPTO-Full 数据集。[7] 过滤掉不正确的反应后，训练集、验证集和测试集分别包含约 769,000



个、96,000 个和 96,000 个反应，规模减少了约 4%。USPTO-50k 的训练集、验证集和测试集中的反应类别分布相同，如图 6 所示。与之前的研究结果一致，我们没有借助

反应类别信息对 USPTO-Full 结果进行基准测试。

图 6：USPTO-50k 数据集中 10 种反应类型的分布。反应的图例根据其比例从大到小排列。

成果

NAG2G **设置** NAG2G 的设置包括一个 6 层编码器和一个 6 层解码器。输入嵌入维度设为 768，注意力头数设为 24。我们采用了亚当优化器[38] $(\beta_1, \beta_2) = (0.9, 0.999)$，线性预热和衰减，峰值学习率为 2.5e-4。训练过程共进行了 12,000 步，批量大小为 16，在单个英伟达 A100 GPU 上需要 6 个小时才能完成。对于 USPTO-Full 数据集的训练，NAG2G 运行了 48,000 个训练步骤，批量大小为 64，在 8 个英伟达 A100 GPU 上完成大约需要 30 个小时。

美国专利商标局数据集的结果

为了评估 NAG2G 在推理过程中的性能，我们采用了常用的波束搜索法来进行顶级候选预测。我们将波束大小设定为 10，长度惩罚为 0，温度为 1。值得注意的是，推理阶段不使用数据增强。此外，NAG2G 依靠 RDChiral[39] 来指定反应物的原子手性，并从产物的立体化学中获取信息。

预测准确度的定义采用了 Liu 等人提出的方法、[17] 该方法认为，只有完全确定特定化学反应的所有反应物，预测才算准确。我们测量的是预测的 top-k 准确率，即正确答案出现在波束搜索结果前 k 个候选答案中的测试案例比例。

USPTO-50k。在表 1 中，NAG2G 在美国专利商标局 50k 数据集上的表现优于最近的基线方法，包括基于模板、半模板和无模板模型。1.为了确保比较的公平性，NAG2G 没有采用辅助技术来帮助推理过程。在无模板模型中，NAG2G 在各项指标上

都明显优于所有基准。尽管某些基准利用额外的数据或方法来增强其结果（用 $*$ 表示

），但 NAG2G

在没有任何此类改进的情况下，NAG2G 仍然表现出色。尽管在基于模板和半模板的方法中额外使用了预定义规则，NAG2G 在没有前提信息的情况下仍优于它们。这标志着 NAG2G 首次超越了基于模板和半模板的方法，而早期的无模板基准从未实现过。此外，NAG2G 还给出了基于 *Augmented Transformer* 提出的 MaxFrag 指标的 SOTA 结果、[20] 详见 SI。详细结果包括在 USPTO-50k 数据集上对每个反应类别进行的测试，并在已知类别的基础上进行了训练。

表 1：USPTO-50k 上逆合成预测的最高准确率。最佳性能以**粗体表示**，每种方法的最佳结果以下划线表示。星号（ ）表示的模型采用了补充数据集进行训练，或在推理过程中采用了提高准确性的技术。为了保持比较的公平性，我们还列出了未采用这些附加技术的结果。

	最高 k 准确率（%）							
	USPTO-50k							
	已知反应类别				反应类别未知			
型号	1	3	5	10	1	3	5	10
基于模板								
RetroSim[4]	52.9	73.8	81.2	88.1	37.3	54.7	63.3	74.1
NeuralSym[5]	55.3	76.0	81.4	85.1	44.4	65.3	72.4	78.9
GLN[7]	64.2	79.1	85.2	90.0	52.5	69.0	75.6	83.7
MHNreact[40]	-	-	-	-	50.5	73.9	81.0	87.9
RetroComposer[8]	65.9	85.8	89.5	91.5	54.5	**77.2**	83.2	87.7
半模板式								
G2G[10]	61.0	81.3	86.0	88.7	48.9	67.6	72.5	75.5
RetroXpert[11]	62.1	75.8	78.5	80.9	50.4	61.1	62.3	63.4
RetroPrime[12]	64.8	81.6	85.0	86.9	51.4	70.8	74.0	76.1
GraphRetro[13]	63.9	81.5	85.2	88.1	53.7	68.3	72.2	75.5
半复古[14]	65.8	85.7	89.8	92.8	54.9	75.3	80.4	84.1
G2Retro[28]	63.6	83.6	88.4	91.5	54.1	74.1	81.2	86.7
马斯[29]	66.2	85.8	90.2	92.9	54.6	76.4	83.3	88.5
无模板								
低压变压器[41]	-	-	-	-	40.5	65.1	72.8	79.4
SCROP[18]	59.0	74.8	78.1	81.1	43.7	60.0	65.2	68.7
GET[25]	57.4	71.3	74.8	77.4	44.9	58.8	62.4	65.9
绑定变压器[19]	-	-	-	-	47.1	67.1	73.1	76.3
梅甘[27] *	60.7	82.0	87.5	91.6	48.1	70.7	78.4	86.1
变压器[20]	-	-	-	-	48.3	-	73.4	77.4
变压器[20]	-	-	-	-	53.5	69.4	81	85.7
GTA[26]	-	-	-	-	51.1	67.6	74.8	81.6
Graph2SMILES[24]	-	-	-	-	52.9	66.5	70.0	72.9
RetroDCVAE[21]	-	-	-	-	53.1	68.1	71.6	74.3
双TF[42]	65.7	81.9	84.7	85.9	53.6	70.7	74.6	77.0
Retroformer[23] *	64.0	82.5	86.7	90.2	53.2	71.1	76.6	82.1
G2GT[32]	-	-	-	-	48.0	57.0	64.0	64.5
G2GT[32]	-	-	-	-	54.1	69.9	74.5	77.7

NAG2G（我们的）	**67.2**	**86.4**	**90.5**	**93.8**	**55.1**	76.9	**83.4**	**89.9**

USPTO-FULL.表 2 列出了各种评估模型的性能指标

在美国专利商标局完整数据集上的性能。随着数据集规模的增加，由于任务的复杂性提高，所有模型的性能都有所下降。值得注意的是，虽然基于模板的方法在 USPTO-50k 数据集上取得了令人印象深刻的结果，但在更大的 USPTO-Full 数据集上，其性能却大打折扣。这一趋势表明，在面对更大、更复杂的数据集时，对基于模板的方法的依赖会受到限制。相比之下，无模板方法虽然也被削弱了，但却表现出了更全面的适应能力，尤其适用于庞大的数据集。尽管如此，NAG2G 在所有评估标准上都明显优于之前的基 。

表 2：USPTO-Full 数据集上逆合成预测的*最高*准确率。星号（ ）表示的模型使用了补充数据集进行训练，或在推理过程中采用了提高准确性的技术。对于用圆圈（ ）表示的模型，根据增强变换器的设置，从测试集中排除了无效反应。[20] 为了使我们的方法与之前的基线保持一致，我们采用了 Augmented Transformer 的方法、[20] 假定这些方法在被删除的测试数据上失败了，我们的方法中没有圆圈（◦）的结果就是证明。

模型		*Top-k* **准确率 (%)**			
模型类型	方法	1	3	5	10
基于模板	RetroSim[4]	32.8	-	-	56.1
	NeuralSym[5]	35.8	-	-	60.8
	GLN[7]	39.3	-	-	63.7
基于半模板	RetroPrime[12]	44.1	59.1	62.8	68.5
无模板	梅甘[27]	33.6	-	-	63.9
	NAG2G（我们的）	**47.7**	**62.0**	**66.6**	**71.0**
	8 月 变压器 * o[20]	46.2	-	-	73.3
	G2GT * o[26]	49.3		68.9	72.7
	NAG2G (我们的) ◦	**49.7**	**64.6**	**69.3**	**74.0**

结果解读（消融研究）

为了确定为 NAG2G 设计的每个组件的重要性，我们通过研究去除这些组件的影响进行了消融研究。这有助于了解 NAG2G 的结构，以及节点排列、数据扩充和整合的方式。

时变图形特征使模型受益。

表 3：对 USPTO-50k 进行的消融研究，反应类别未知。

策略			***顶*k 准确率 (%)**			
节点对齐	数据扩充	图特征	1	3	5	10
✓	✓	✓	**55.1**	**76.9**	**83.4**	**89.9**
✓	✓	×	54.1	75.9	82.6	88.8
✓	×	✓	49.2	69.2	75.3	80.4
×	✓	✓	46.1	47.6	48.5	49.9
×	×	✓	40.3	54.9	58.9	62.6

表 3 列出了每种策略对模型在美国专利商标局 50k 数据集上的性能影响的全面量化细目，尤其侧重于 *Top-k* 准确率。当节点对齐、数据增强和图形特征这三种技术全部使用时，模型的 Top-1 准确率达到了 55.1% 的峰值，并在 Top-3（76.9%）、Top-5（83.4%）和 Top-10 （89.9%）中保持了较高的性能，强调了这一组合的协同效应。消除图形特征会导致每个 *Top-k* 准确度指标略有下降。不过，在难以取得进一步进展的挑战性场景下，加入这些图形特征有助于将指标提高约 1%。忽略数据增强会导致更明显的下降，top-1 的准确率下降了 5.9%，top-10 的准确率下降了 9.5%，这凸显了数据增强在增强模型鲁棒性和对未见数据的泛化能力方面的作用。如果没有节点对齐，性能会急剧下降，Top-10 的准确率会下降 50%。这凸显了节点对齐在捕捉图的结构信息和使模型做出更准确预测方面的重要性。很明显，top-1 和 top-10 的准确率差距从 34.8%（同时使用数据增强和节点对齐时）大幅缩小到仅 3.8%（使用数据增强

而不使用节点对齐时）。这说明，与同时采用两种技术相比，在不进行节点对齐的情况下使用数据增强技术会导致候选预测的多样性降低。

不过，值得注意的是，在不使用节点对齐和数据增强的情况下，差距仍为 22.3%，并不算特别大。因此，这表明节点对齐和数据增强是互补技术，联合使用时可以提高性能指标。总之，我们的消融研究表明，节点对齐、数据增强和图特征都是我们模型的重要组成部分，每种策略都在提高预测准确性方面发挥着重要作用。这些策略的结合产生了最佳的整体性能，强调了在未来基于图的反应预测模型工作中结合这些策略的重要性。

表 4：在反应类别未知的情况下，USPTO-50k 上生成的分子的前 *k* 有效性。

模型	**顶部-*k***		**湿度（%）**	
	Va 1	3	5	10
NAG2G（我们的）	**99.7**	**98.6**	**97.1**	**92.9**
不带电的 NAG2G	89.9	90.2	86.1	75.9
不含氢的 NAG2G	89.6	88.4	87.6	83.4
不含电荷或氢的 NAG2G	80.8	82.5	81.5	77.6
获取[25]	97.8	86.6	80.5	70.7
Graph2SMILES[24]	99.4	90.9	84.9	74.9
RetroPrime[12]	98.9	98.2	**97.1**	92.5

另一个表 4 显示了各种模型在美国专利商标局 50k 数据集上的 *Top-k* 有效性，重点是 NAG2G 模型采用的自动回归节点生成方法。当使用所有原子特征（类型、形式电荷、氢数）时，NAG2G 的 Top-1 有效率为 99.7%，在 Top-3（98.6%）、Top-5（97.1%）和 Top-10 （92.9%）预测中也表现出色。在节点生成过程中排除某些分子特性揭示了生成 SMILE 时每个特性的重要性。仅忽略形式电荷的结果是，Top-1 有效

率为 89.9%，而忽略氢则为 89.6%。当同时排除这两个特征时，结果下降最为明显，Top-1 有效率下降到 80.8%。我们还研究了准确性方面的结果，详见辅助信息。当

与 *GET*、*Graph2SMILES* 和 *RetroPrime* 等其他模型相比，NAG2G 的性能始终优于它们。

此外，我们对编码器的消融研究表明，我们的模型性能稳健，并不过分依赖于编码器的配置。有关详细研究结果，请参阅我们提供的补充信息。

案例研究

为了评估有机分子合成路线的设计能力，我们选取了四种候选药物作为目标产品，并使用使用 USPTO-50k 数据集训练的 NAG2G 进行了序列 SSR。无模板 NAG2G 的性能成功超越了之前的模型。[16] 如图 7(a) 所示。原始文献中记载的 Nirmatrelvir 的所有六个合成步骤[43] 的六个合成步骤都被 NAG2G 准确地预测到了，而且所有预测结果都在前三名之内。涉及酰胺脱水形成氰基的初始步骤排名第一，而预测的缩合反应排名与最近的 Nirmatrelvir 先进一锅合成策略一致。[44] 第 2 步和第 6 步缩合反应由 NAG2G 精确完成，两个答案都在其他候选反应中脱颖而出。在第三步中，**6** 中胺的三氟乙酰化反应被我们的模型预测为排名第三的反应，比 *Graph2Edits* 中排名第六的预测有所改进。对于步骤-4 和步骤-5，我们模型的第一和第二名预测使用不同但常见的试剂有效地实现了保护功能。[45] 第二个测试案例是奥希替尼，其研究名称为 AZD9291，于 2014 年获批成为非小细胞肺癌（NSCLC）患者的临床治疗药物。[46] 如图 7(b)所示，

NAG2G 成功地勾勒出文献中描述的五步合成法，追溯了从市售嘧啶到最终产品 AZD9291 的合成途径。反向合成的第一步是酰化反应，按可能性排序为第一级，然后是硝基的另一个第一级还原反应。随后，它正确地识别出了两个连续的核苷酸基团。

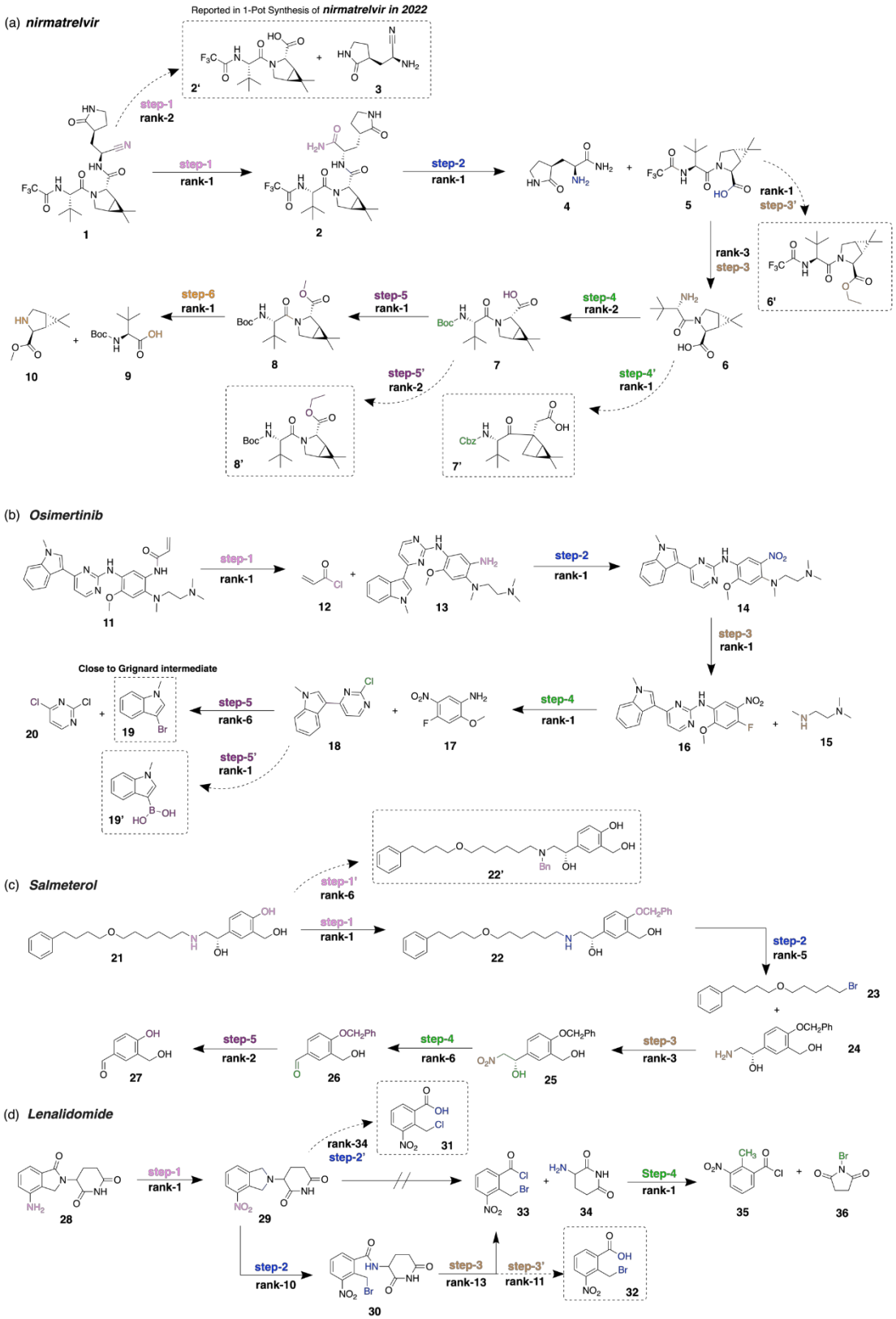


图 7：四种药物分子的合成路线和 NAG2G 预测的等级。

亲芳取代步骤为首选。在最后一步中，模型预测概率最高的是铃木偶联反应，这也是 *Graph2Edits* 模型推断出的排名第 2 的反应。[16] 虽然没有预测到涉及格氏反应的原始策略，但排名-7 的结果表明了这一替代途径。对于第三种情况，我们选择了沙美特罗（Salmeterol），这是一种长效支气管扩张剂，Coley 已在以前的基于模板的模型中对其进行了测试、[4] *局部复位*[6] 和 Zhong 的根对齐策略进行了测试。[31] 由于化合物 **21** 中存在胺和苯酚官能团，NAG2G 建议保护反应位点，将这些步骤排在第一和第二位。[47,48] 随后的反应是通过威廉姆森醚型反应制备化合物 **22，被**确定为排在第 5 位的选择。此外，第四步的排名 6 成功预测了 **26** *通过*不对称亨利反应转化为 **25**，正如郭的合成中所描述的那样。[48] 值得注意的是，原始合成使用 2,2-二甲氧基丙烷同时保护化合物 **27** 中的两个羟基，而 NAG2G 只选择保护反应性更强的苯酚基，这与完整合成方案的要求一致。

最后一个例子是来那度胺，它也参与了 *LocalRetro*的工作。[6] 和 *Graph2Edit*[16] 结果发现这是一项更具挑战性的任务。[49] NAG2G 准确预测了涉及硝基还原和 NBS（N-溴代丁二酰亚胺）取代的第一步和最后一步反应，并将其列为前 1 位反应。在涉及形成两个 C-N 键的环化步骤中，NAG2G 提出了一种带有氯取代基而不是溴取代基的前体化合物 **31**。尽管如此，NAG2G 仍能提出一种逐步闭环机制，与 *LocalRetro* 报告的结果一致。

误差分析

在序列到序列（seq2seq）模型中[17] 背景下，以 SMILES 格式生成反应物可能导致三种结果：1) RDKit 生成的 SMILES 字符串无效，对应的结构在化学上不可行。2) SMILES 字符串在化学上是有效的，但并不代表能够产生所需的反应物。

在给定反应条件下的生成物。3) SMILES 字符串代表的反应物组合可能会导致作为常见反应的产物，尽管它们可能与基本真实反应物不完全匹配。在评估我们的模型在美国专利商标局 50k 数据集上的第一类错误时，我们侧重于有效性，即在前 k 项预测中生成的有效 SMILES 字符串的百分比。我们的模型top-1有效率高达99.7%，超过了其他先进模型。[12,24,25] 这种卓越的有效性归功于我们在烧蚀研究部分讨论过的自回归生成过程。如果模型随机省略对电荷和氢气附着的预测，top-1 有效率将显著下降到80.8%。第二类和第三类错误的识别涉及有机化学家的专家评估。为了在不预先指定反应类别的情况下公正地审查推理结果，我们选择了三位代表来考察我们的模型（此处称为 NAG2G）可以预测的各种反应物。在所有情况下，基本事实都在前 10 个预测范围内，同时还有其他各种反应类型，这展示了模型广泛的预测范围。丰富多样的建议合成方案提供了一系列可考虑的化学反应，从而为后续的路线选择和评估提供了替代方案，从而增强了逆合成路线的设计。有关详细的汇总表和分析，请参阅补充信息。

美国专利商标局数据集的另一个局限是缺乏详细的反应信息，如条件、产率和选择性。有了这些信息，NAG2G 模型就能为其预测提供更准确、更可靠的排名。此外，单步预测模型（包括我们的模型）可能会忽略连续步骤之间的相互作用。例如，在案例 3 中，选择性酚苄基化虽然是一个有效的保护步骤，但可能会对随后的不对称亨利反

应产生重大影响。这种情况说明，有效的单步预测不仅需要准确性，还需要对路线复杂性和产率权衡进行全面评估。为了应对这些挑战，我们正在积极开发先进的评分方法，以促进多步过程

根据 NAG2G 提供的结果做出决定。

结论

在这项研究中，我们提出了节点对齐图到图（NAG2G）模型--一种基于图的无模板 SSR 方法。该模型采用 Transformer 编码器-解码器框架，以自动回归的方式生成反应分子图。在 USPTO-50k 和 USPTO-Full 这两个成熟数据集上进行的测试表明，NAG2G 的性能可与现有的 SOTA 模型相媲美。消融研究揭示了各种成分的贡献，强调了我们方法的潜力。对案例研究和误差分析的复制凸显了 NAG2G 在特定 SSR 任务中的卓越性能，表明进一步的改进可以进一步增强其能力。

尽管我们在导言中详细介绍了许多 SSR 模型，它们都显示出良好的前景，但 NAG2G 标志着深度学习在单步逆合成中的应用取得了长足进步--尤其是在考虑独立于模板的方法时。我们的方法揭示了复杂的神经网络并非通往卓越的唯一途径；精心的模型设计与精确的任务定义相结合，可以产生具有竞争力的结果。我们目前的设计是为单步逆合成预测量身定制的，在这种情况下，输入和输出图形非常相似。对于更广泛的图到图生成任务，尤其是输入输出差异较大的任务，可能需要进行改进。展望未来，我们的最终目标是发展这种方法，针对更复杂的化学合成方案深入研究多步骤合成规划。

可提供的证明资料

NAG2G 可从 https://github.com/dptech-corp/NAG2G 网站获取。

与 NAG2G 性能相关的实验结果、辅助烧蚀研究和预处理示例。(PDF)

作者供稿

W.G.和Z.W.对本研究做出了同等贡献。

鸣谢

姚林和郭文涛感谢团团在项目期间给予的支持。

可提供的证明资料

与 NAG2G 性能有关的实验结果，辅助烧蚀研究。

图形生成的相关工作

图形生成大致可分为两类：全局模型和顺序模型。全局模型是同时生成整个图形结构，而不是按顺序生成。这些模型通常采用成对矩阵（如邻接矩阵）来表示图，并学习生成矩阵。常用的全局图生成技术包括基于自动编码器的模型、[50–56] 生成对抗网络（GANs）、[57–59] 和基于流的生成模型。[60,61]

相比之下，顺序模型是以增量方式生成图形的。以前的许多工作都属于这一类，如 MolRNN、[62] GraphRNN、[63] MolecularRNN、[64] Bacciu 等人的作品、[65]？ GraphGen、[66] MolMP、[62] GRAN、[67] GRAM、[68] AGE、[69] DeepGMG、[70] 和 BiGG。[71]

顺序模型简单有效，但需要指定节点的生成顺序。

拟议的 NAG2G 利用节点对齐策略来确定节点生成顺序，从而解决了顺序生成的难题。值得注意的是，所提出的 NAG2G 模型主要是为单步逆合成预测而设计的，因为输入图和输出图的差异很小，允许节点对齐。但是，对于一般的图形生成问题，所提出的方法可能无法达到最佳效果。

更多实验结果

编码器烧蚀研究

在 NAG2G 中，我们直接使用以前工作中的现有模型骨干作为编码器。为了研究编码器带来的性能提升，我们进行了一项消减研究。表 5 显示，除了默认的 Uni-Mol+，我们还评估了 Graphorm 的性能、[36] 我们还评估了 Graphormer 的性能。[34] 结果表明，总体性能相当。这一观察结果表明，编码器的选择不会对 NAG2G 的最终性能产生重大影响 。

表 5：不同编码器在反应类型未知的 USPTO-50k 上的性能。

	***最高*精度 (%)**			**编码器类**
型	1	3	5	10
默认值（Uni-Mol+）[36]	**55.1**	76.9	**83.4**	**89.9**
Graphormer[34]	54.3	**77.0**	**83.4**	89.0

生成分子的准确性和有效性

为了准确地表示分子，不仅要考虑原子和键的类型，还要考虑原子的形式电荷和相连氢原子的数量，这一点至关重要。虽然 RDKit 等软件工具有可能提供这些信息，但它们往往无法提供或提供的数据不准确。为了解决这个问题，NAG2G 在生成原子类型和化学键的同时，还从头到尾生成了这些特征。通过两项烧蚀研究评估了纳入这些

附加信息的影响。

我们在手稿中展示的第一项研究评估了生成分子的有效性，如表 4 所示。此外，第二项研究还考察了整体性能，如表 6 所示。由于有效性的提高，模型能够生成更多的有效结果，从而显著提高了整体性能。

表 6：在反应类别未知的 USPTO-50k 中，NAG2G 在生成/不生成额外原子特征的情况下的性能。

收费	氢原子	*Top-k* 准确率 (%)			
		1	3	5	10
✓	✓	**55.1**	**76.9**	**83.4**	**89.9**
✓	×	49.0	68.9	74.7	80.8
×	✓	48.1	68.2	73.9	79.4
×	×	42.6	60.8	65.9	70.9

主试剂（MaxFrag）的精度

之前在 8 月的 Transformer 中提出的 MaxFrag 指标被用于评估主反应物（最大反应物）预测的准确性、[20] 用于评估主要（最大）反应物预测的准确性。引入该指标是因为它在经典程序中具有重要意义，因为在经典程序中，只关注主要化合物的转化只能提供得出高效逆向合成路线所需的最少信息。如表 7 所示，我们提出的模型 NAG2G 优于 Aug.

表 7：在反应类别未知的情况下，MaxFrag 对 USPTO-50k 的 *top-k* 准确率。

模型	*Top-k* MaxFrag Accuarcy (%)			
	1	2	5	10
八月 变压器[20]	58.0	73.4	84.8	89.1
NAG2G（我们的）	**59.7**	**73.6**	**86.3**	**91.9**

各反应类别的准确性

在表 8 中，我们展示了 NAG2G 在美国专利商标局 50k 数据集上的 *top-k* 准确率结果

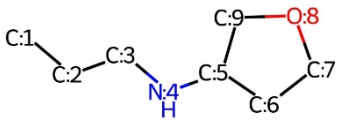
，其中反应类别在训练过程中未披露，结果根据基本真实反应类别进行了分层。

表 8：在反应类别未知的情况下，USPTO-50k 数据集上每个反应类别的*最高*准确率。

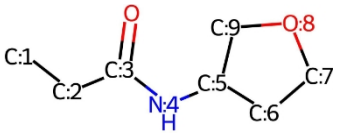
反应类别	**反应比例(%)**	***Top-k* 准确率(%)**			
		1	3	5	10
杂原子烷基化和芳基化	30.3	54.4	75.5	82.3	90.5
酰化及相关过程	23.8	67.3	87.5	91.1	94.5
剥夺	16.5	51.2	81.1	86.9	92.1
C-C 键的形成	11.3	40.0	61.6	71.8	81.7
减少	9.2	55.4	74.5	81.8	87.7
官能团相互转化	3.7	34.8	53.3	64.1	76.1
杂环形成	1.8	44.0	66.0	74.7	82.4
氧化	1.6	69.5	81.7	91.5	96.3
保护	1.3	67.6	85.3	89.7	92.6
官能团加成	0.5	87.0	87.0	87.0	91.3

预测示例

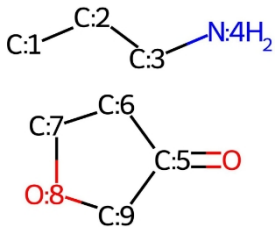
如图 8、图 9 和图 10 所示，我们展示了 NAG2G 在美国专利商标局 50k 测试数据集上预测结果的三个示例，其中没有给出反应类别。这些示例展示了 NAG2G 预测可信反应物的能力，即使在较低的排序下也能实现与基本真实反应物的完全匹配。该模型能够生成多种预测结果，从而得到所需的产品，凸显了其在现实世界中的应用潜力。



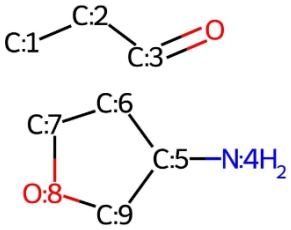
生产



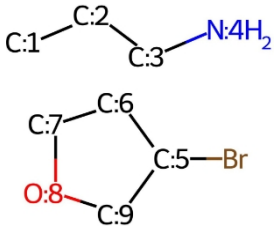
真实反应物



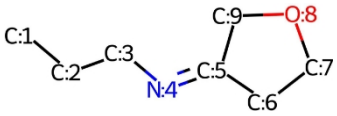
1 大预测反应物



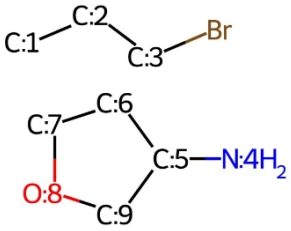
前 2 位预测反应物



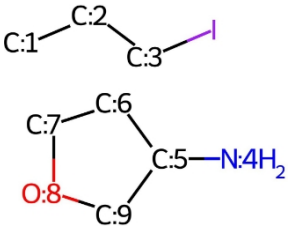
前 3 位预测反应物



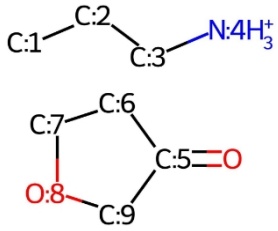
前 4 位预测反应物



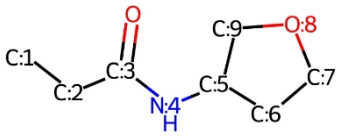
5 大预测反应物



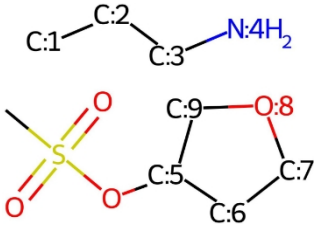
排名前 6 位的预测反应物



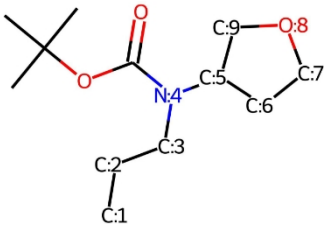
前 7 位预测反应物



8 大预测反应物

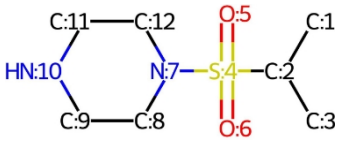


9 大预测反应物

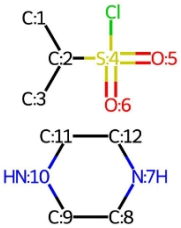


10 大预测反应物

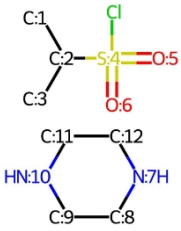
图 8：这是 NAG2G 在美国专利商标局 50k 测试数据集上进行预测的示例 1，反应类别未知。虽然只有第八个预测反应物与地面实况反应物精确对应，但所有十个预测反应在反应机理中都是化学上有效的。具体来说，第一、第二、第四、第七和第八个反应可归类为还原反应，而第三、第五、第六和第九个反应属于杂原子烷基化和芳基化反应类型。第十个反应属于去保护反应。



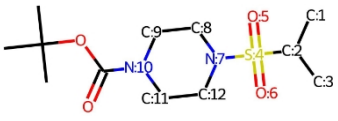
生产



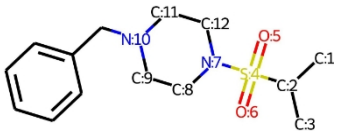
真实反应物



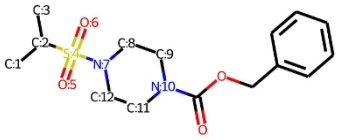
1 大预测反应物



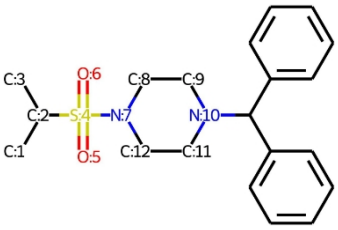
前 2 位预测反应物



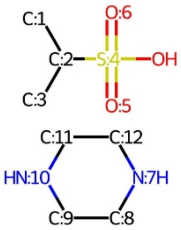
前 3 位预测反应物



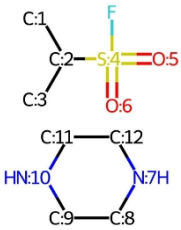
前 4 位预测反应物



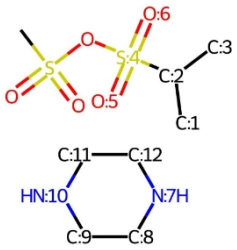
5 大预测反应物



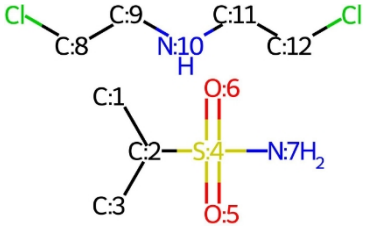
排名前 6 位的预测反应物



前 7 位预测反应物



8 大预测反应物



9 大预测反应物 10 大预测反应物

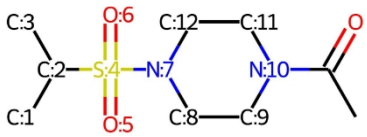
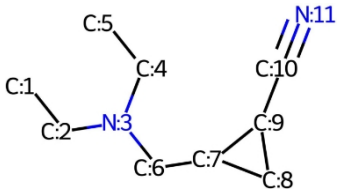
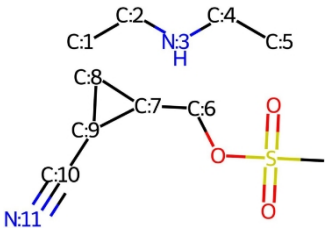


图 9：这是 NAG2G 对美国专利商标局 50k 测试数据集进行预测的示例 2，反应类别未知。除第六个反应外，所有反应在化学机制上都是有效的。第一个预测的反应物与基本事实完全吻合。第一、第七、第八和第九个反应属于杂原子烷基化和芳基化反应，而第二、第三、第四、第五和第十个反应属于去保护反应。值得注意的是，第六个

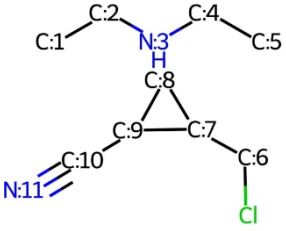
预测反应物只需增加一个步骤即可轻松转化为基本真实值。



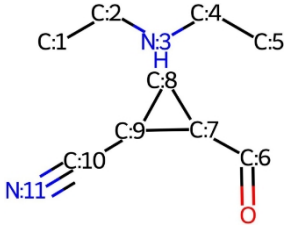
生产



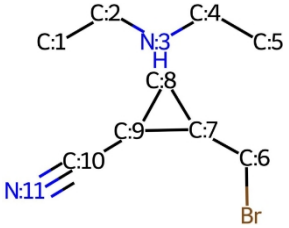
真实反应物



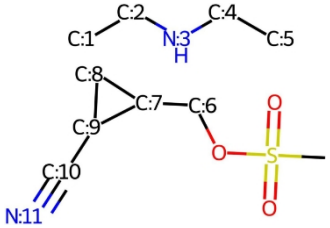
1 大预测反应物



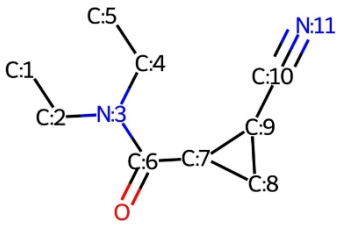
前 2 位预测反应物



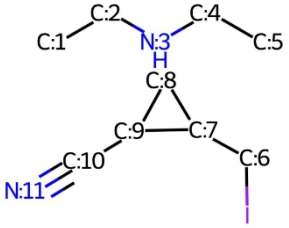
前 3 位预测反应物



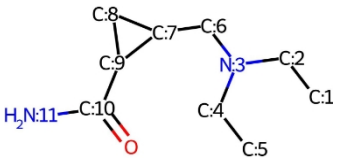
前 4 位预测反应物



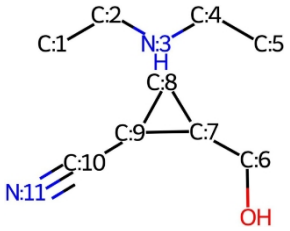
5 大预测反应物



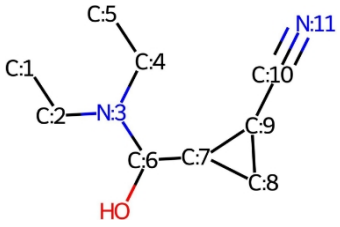
排名前 6 位的预测反应物



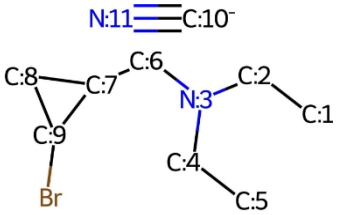
前 7 位预测反应物



8 大预测反应物



9 大预测反应物



10 大预测反应物

图 10：这是 NAG2G 在美国专利商标局 50k 测试数据集上的预测示例 3，反应类别未知。除第八个反应外，所有其他反应在化学机制上都是有效的。第四个预测的反应物与基本事实完全吻合。第一、第三、第四和第六个反应可归类为杂原子烷基化和芳基化反应，而第二、第五和第九个反应则归类为还原反应。第七个反应属于官能团相互转化反应，第十个反应属于 C-C 键形成反应。值得注意的是，第八个预测只需增加

一个步骤就可以轻松转换为基本事实。

参考资料

(1) Corey, E. J.; Cheng, X.-M.*The logic of chemical synthesis*; John Wiley & Sons：Nashville, TN, 1995.

(2) Zhong, Z.; Song, J.; Feng, Z.; Liu, T.; Jia, L.; Yao, S.; Hou, T.; Song, M. Recent advances in deep learning for retrosynthesis.*Wiley Interdisciplinary Reviews：Wiley Interdisciplinary Reviews: Compu- tational Molecular Science* e1694.

(3) 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*, 390-398.

(4) Coley, C. W.; Rogers, L.; Green, W. H.; Jensen, K. F. Computer-assisted retrosynthesis based on molecular similarity.*ACS central science* **2017**, *3*, 1237-1245.

(5) Segler, M. H.; Waller, M. P. Neural-symbolic machine learning for retrosynthesis and reaction prediction.*Chemistry-A European Journal* **2017**, *23*, 5966-5971.

(6) 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*, 1612-1620.

(7) Dai, H.; Li, C.; Coley, C.; Dai, B.; Song, L. Retrosynthesis Prediction with Conditional Graphic Logic Network.神经信息处理系统进展》。2019; pp 8870-8880.

(8) 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*，1325。

(9) Koca, J.; Kratochvil, M.; Kvasnicka, V.; Matyska, L.; Pospichal, J. *有机化学和合成设计的 Synthon 模型*；施普林格科学与商业媒体，2012 年；第 51 卷。

(10) Shi, C.; Xu, M.; Guo, H.; Zhang, M.; Tang, J. A Graph to Graphs Framework for Retrosynthesis Prediction.第 37 届国际机器学习大会（ICML）论文集。2020; pp 8818-8827.

(11) Yan, C.; Ding, Q.; Zhao, P.; Zheng, S.; Yang, J.; Yu, Y.; Huang, J. Retroxpert: Decompose retrosynthesis prediction like a chemist.*神经信息处理系统进展* **2020**》，*33*，11248-11258。

(12) 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预测方法（Diverse, plausible and Transformer-based method for Single-Step retrosynthesis pre dictions）。*化学工程学报*，**2021**，*420*，129845。

(13) Somnath, V. R.; Bunne, C.; Coley, C. W.; Krause, A.; Barzilay, R. Learn- ing Graph Models for Template-Free Retrosynthesis. *ArXiv preprint,* 2021, https://doi.org/10.48550/arXiv.2006.07038.

(14) Gao, Z.; Tan, C.; Wu, L.; Li, S. Z. SemiRetro: Semi-template framework boosts deep retrosynthesis prediction. *ArXiv preprint,* 2022, https://doi.org/10.48550/arXiv.2202.08205.

(15) 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。*通信化学* **2023**，*6*，102。

(16) 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。*Nature Communications* **2023**, *14*, 3009.

(17) 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 Neu- ral Sequence-to-Sequence Models.*ACS Central Science* **2017**, *3*, 1103-1113, PMID: 29104927.

(18) Zheng, S.; Rao, J.; Zhang, Z.; Xu, J.; Yang, Y. 使用自校正变压器神经网络预测逆合成反应。*化学信息与建模杂志*，**2020**，*60*，47-55，PMID: 31825611.

(19) 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变换器进行有效、可信和多样的逆合成。*化学信息与建模期刊》* **2021 年第** *61 期*，123-133，PMID：33410697。

(20) Tetko, I. V.; Karpov, P.; Van Deursen, R.; Godin, G. State-of-the-art augmented NLP transformer models for direct and single-step retrosynthesis.*Nature Communications* **2020**, *11*, 5575.

(21) 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 国家间信息与知识管理大会论文集》。2022; pp 717-726.

(22) Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L. u.; Polosukhin, I. Attention is All You Need.神经信息处理系统进展》。2017.

(23) Wan, Y.; Liao, B.; Hsieh, C.-Y.; Zhang, S. Retroformer：突破可预装端到端逆合成变压器的极限。2022.

(24) Tu, Z.; Coley, C. W. Permutation invariant graph-to-sequence model for template-free retrosynthesis and reaction prediction.*化学信息与建模期刊***2022**，*62*，3503-3513。

(25) Mao，K.；Xiao，X.；Xu，T.；Rong，Y.；Huang，J.；Zhao，P. Molecular graph enhanced transformer for retrosynthesis prediction.*神经计算* **2021**，*457*，193-202。

(26) Seo，S.-W.；Song，Y.Y.；Yang，J.Y.；Bae，S.；Lee，H.；Shin，J.；Hwang，S.J.；Yang，E.

GTA：用于逆合成的图截断注意力。美国人工智能学会会议论文集》。2021; pp 531-539.

(27) Sacha, M.; Błaż, M.; Byrski, P.; Dąbrowskii-Tumański, P.; Chromiński, M.; Loska, R.; Włodarczyk-Pruszyński, P.; Jastrzębski, S. Molecule Edit Graph Attention Network：将化学反应建模为图编辑序列。*Journal of Chemical Infor- mation and Modeling* **2021**, *61*, 3273-3284, PMID: 34251814.

(28) 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。*通信化学***2023**，*6*，102。

(29) Liu, J.; Yan, C.; Yu, Y.; Lu, C.; Huang, J.; Ou-Yang, L.; Zhao, P. MARS：A 基于动机的自回归模型用于逆合成预测》。*ArXiv 预印本*，2022 年，https://doi.org/10.48550/arXiv.2209.13178。

(30) 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。*化学科学* **2020**，*11*，3355-3364。

(31) Zhong, Z.; Song, J.; Feng, Z.; Liu, T.; Jia, L.; Yao, S.; Wu, M.; Hou, T.; Song, M. Root-aligned SMILES: a tight representation for chemical reaction prediction. *化学科学***2022**，*13*，9023-9034。

(32) Lin，Z.；Yin，S.；Shi，L.；Zhou，W.；Zhang，Y. J. G2GT：利用图对图注意

神经网络和自我训练的逆合成预测。*化学信息与建模学报* **2023**》，*63*，1894-1905，PMID：36946514。

(33) Schneider, N.; Lowe, D. M.; Sayle, R. A.; Tarselli, M. A.; Landrum, G. A. Big Data from Pharmaceutical Patents：药物化学家的面包和黄油的计算分析。*Journal of Medicinal Chemistry* **2016**, *59*, 4385-4402, PMID: 27028220.

(34) Ying, C.; Cai, T.; Luo, S.; Zheng, S.; Ke, G.; He, D.; Shen, Y.; Liu, T.-Y.转形式器在图形表示方面真的表现糟糕吗？第 35 届神经信息处理系统大会。2021.

(35) Zhou, G.; Gao, Z.; Ding, Q.; Zheng, H.; Xu, H.; Wei, Z.; Zhang, L.; Ke, G. Uni- Mol：通用三维分子表征学习框架。第十一届学习表征国际会议。2023.

(36) Lu, S.; Gao, Z.; He, D.; Zhang, L.; Ke, G. Highly Accurate Quan- tum Chemical Property Prediction with Uni-Mol+. *ArXiv preprint,* 2023, https://doi.org/10.48550/arXiv.2303.16982.

(37) Landrum, G.; others RDKit：用于化学信息学、计算化学和预测建模的软件套件。*Greg Landrum* **2013**, *8*, 31.

(38) Kingma, D. P.; Ba, J. Adam: A Method for Stochastic Optimization. *ArXiv preprint*、2017, https://doi.org/10.48550/arXiv.1412.6980.

(39) Coley, C. W.; Green, W. H.; Jensen, K. F. 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。*化学信息与建模期刊*》，**2019**，*59*，2529-2537。

(40) 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 Pre- diction Using Modern Hopfield Networks.*Journal of Chemical Information and Mod- eling* **2022**, *62*, 2111-2120, PMID: 35034452.

(41) Chen, B.; Shen, T.; Jaakkola, T. S.; Barzilay, R. Learning to Make Gen- eralizable and Diverse Predictions for Retrosynthesis. *ArXiv preprint,* 2019,

https://doi.org/10.48550/arXiv.1910.09688.

(42) Sun, R.; Dai, H.; Li, L.; Kearnes, S.; Dai, B. Energy-based View of Retrosynthesis. *arXiv preprint*，2021，https://doi.org/10.48550/arXiv.2007.13437。

(43) Hammond，J.；Leister-Tebbe，H.；Gardner，A.；Abreu，P.；Bao，W.；Wisemandle，W.；Baniecki，M.；Hendrick，V. M.；Damle，B.；Simón-Campos，A.；others Covid-19高危、非住院成人口服nirmatrelvir。*新英格兰医学杂志》*，**2022 年**，*386 期*，1397-1408 页。

(44) Caravez, J. C.; Iyer, K. S.; Kavthe, R. D.; Kincaid, J. R.; Lipshutz, B. H. A 1-pot synsis of the SARS-CoV-2 Mpro Inhibitor Nirmatrelvir, the key ingredient in Paxlovid.*Organic Letters* **2022**, *24*, 9049-9053.

(45) Owen, D. R.; Allerton, C. M.; Anderson, A. S.; Aschenbrenner, L.; Avery, M.; Berritt, S.; Boras, B.; Cardin, R. D.; Carlo, A.; Coffman, K. J.; others 用于治疗 COVID-19 的口服 SARS-CoV-2 Mpro 抑制剂临床候选药物。*Science* **2021**, *374*, 1586- 1593.

(46) Finlay, M. R. V.；Anderton, M.；Ashton, S.；Ballard, P.；Bethel, P. A.；Box, M. R.；Bradbury, R. H.；Brown, S. J.；Butterworth, S.；Campbell, A.；others 发现了一种针对敏化突变和 T790M 抗性突变的强效选择性表皮生长因子受体抑制剂(AZD9291)，该抑制剂可保护野生型受体。*Journal of Medicinal Chemistry* **2014**, *57*, 8249-8267.

(47) Hett, R.; Stare, R.; Helquist, P. 通过不对称硼烷还原法对映选择性合成沙美特罗。

Tetrahedron Letters **1994**, *35*, 9375-9378.

(48) Guo, Z.-L.; Deng, Y.-Q.; Zhong, S.; Lu, G. Enantioselective synthesis of (R)-salmeterol using an asymmetric Henry reaction as the key step.*Tetrahedron：Tetrahedron: Asymmetry* **2011**, *22*, 1395-1399.

(49) Ponomaryov, Y.; Krasikova, V.; Lebedev, A.; Chernyak, D.; Varacheva, L.; Cher-

抗癌药物来那度胺的 可扩展绿色合成工艺。*杂环化合物化学》*，**2015**，*51*，133-138。

(50) Kipf, T. N.; Welling, M. Variational graph auto-encoders.NeurIPS 贝叶斯深度学习研讨会。2016.

(51) Simonovsky, M.; Komodakis, N. Graphvae: Towards generation of small graphs using variational autoencoders.人工神经网络国际会议。2018; pp 412-422.

(52) Flam-Shepherd, D.; Wu, T.; Aspuru-Guzik, A. Graph Deconvolutional Generation. *arXiv preprint*，2020，https://doi.org/10.48550/arXiv.2002.07087。

(53) Ma, T.; Chen, J.; Xiao, C. Constrained generation of semantically valid graphs via regu- larizing variational autoencoders.神经信息处理系统进展》。2018; pp 7113-7124.

(54) Grover, A.; Zweig, A.; Ermon, S. Graphite：图形的迭代生成建模。国际机器学习大会。2019; pp 2434-2444.

(55) Guo, X.; Zhao, L.; Qin, Z.; Wu, L.; Shehu, A.; Ye, Y. Node-Edge Co-disentangled Representation Learning for Attributed Graph Generation.第 26 届 ACM SIGKDD 知识发现与数据挖掘国际会议论文集》。2020.

(56) Li, J.; Yu, J.; Li, J.; Zhang, H.; Zhao, K.; Rong, Y.; Cheng, H.; Huang, J. Dirichlet graph variational autoencoder.*神经信息处理系统进展* **2020**》，*33*，5274-5283。

(57) De Cao, N.; Kipf, T. MolGAN: An implicit generative model for small molecular graphs.ICML 深度生成模型的理论基础与应用研讨会。2018.

(58) Yang, C.; Zhuang, P.; Shi, W.; Luu, A.; Li, P. Conditional Structure generation through graph variational generative adversarial nets.神经信息处理系统进展》。2019; pp 1340-1351.

(59) Yang, S.; Liu, J.; Wu, K.; Li, M. Learn to Generate Time Series Conditioned Graphs with Generative Adversarial Nets. *arXiv preprint,* 2023, https://doi.org/10.48550/arXiv.2003.01436.

(60) Liu, J.; Kumar, A.; Ba, J.; Kiros, J.; Swersky, K. Graph normalizing flows.神经信息处理系统进展》。2019; pp 13578-13588.

(61) Madhawa, K.; Ishiguro, K.; Nakago, K.; Abe, M. Graphnvp：用于生成分子图的可验证流模型。*ArXiv 预印本*，2019 年，https://doi.org/10.48550/arXiv.1905.11600.

(62) Li, Y.; Zhang, L.; Liu, Z. 利用条件图生成模型的多目标从头药物设计。*Journal of cheminformatics* **2018**, *10*, 1-24.

(63) You, J.; Ying, R.; Ren, X.; Hamilton, W.; Leskovec, J. GraphRNN：用深度自回归模型生成真实图。国际机器学习大会。2018; pp 5708-5717.

(64) Popova, M.; Shvets, M.; Oliva, J.; Isayev, O. MolecularRNN： Generat- ing realistic molecular graphs with optimized properties. *ArXiv preprint,* 2019, https://doi.org/10.48550/arXiv.1905.13372.

(65) Bacciu, D.; Micheli, A.; Podda, M.; others Graph generation by sequential edge predic-

tion.ESANN.2019; pp 95-100.

(66) Goyal, N.; Jain, H. V.; Ranu, S. GraphGen：领域标注图生成的可扩展方法。2020 年网络大会论文集》，第 1253-1263 页。

(67) Liao, R.; Li, Y.; Song, Y.; Wang, S.; Hamilton, W.; Duvenaud, D. K.; Urtasun, R.; Zemel, R. Efficient graph generation with graph recurrent attention networks.神经信息处理系统进展》。2019; pp 4257-4267.

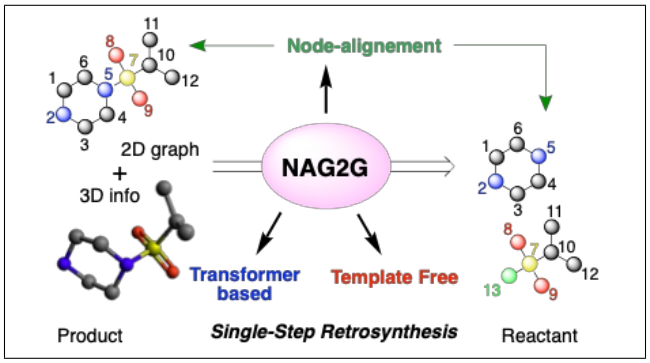
(68) Kawai, W.; Mukuta, Y.; Harada, T. Scalable Generative Models for Graphs with Graph Attention Mechanism. *ArXiv preprint,* 2021, https://doi.org/10.48550/arXiv.1906.01861.

(69) Fan, S.; Huang, B. Attention-Based Graph Evolution.Pacific-Asia Conference on Knowledge Discovery and Data Mining.2020; pp 436-447.

(70) Li, Y.; Vinyals, O.; Dyer, C.; Pascanu, R.; Battaglia, P. Learning deep generative models of graphs. *ArXiv preprint,* 2018, https://doi.org/10.48550/arXiv.1803.03324.

(71) Dai, H.; Nazi, A.; Li, Y.; Dai, B.; Schuurmans, D. Scalable Deep Generative Modeling for Sparse Graphs.国际机器学习大会。2020.

图标说明


Node-alignement
2D graph
+
3D info
NAG2G
Transformer based
Template Free
Product
Single-Step Retrosynthesis
Reactant