



AI辅助合成中的分子表示方法: 从序列模型到图论

汪鸿帅¹, 金鲁杰¹, 冯杰¹, 董志浩¹, 邓伟侨^{2*}, 李有勇^{1*}

1. 苏州大学功能纳米与软物质研究院, 苏州 215123

2. 山东大学前沿交叉科学青岛研究院, 青岛 266237

*通讯作者, E-mail: dengwq@sdu.edu.cn; yyli@suda.edu.cn

收稿日期: 2022-05-10; 接受日期: 2022-07-11; 网络版发表日期: 2022-12-19

国家自然科学基金(编号: 22173067)、江苏省自然科学基金(编号: BZ2020011, BK20200873)、苏州纳米科技合作创新中心、江苏省高等院校优势学科(PAPD)、111项目和碳基功能材料与器件国际联合研究实验室资助

摘要 近些年, 随着深层神经网络及其衍生算法的飞速发展, AI在各个领域大放异彩. 在计算机辅助化合物合成路线设计领域, 生成对抗网络, 变分自编码器等人工智能算法表现十分出色. 而如何处理分子的几何表示是将化学问题转化为算法问题的关键步骤. 本文将从不同维度出发, 系统地总结各种分子的序列模型、图论等表示方法在辅助合成化学中的应用, 讨论目前分子表示中的短板与面临的挑战.

关键词 人工智能, 分子表示, 计算机辅助合成, 机器学习

1 引言

人工智能技术的迅速崛起为很多领域带来了新的变革. 以深度学习为代表的各种机器学习算法在计算机视觉、自然语言处理、推荐算法等应用场景下取得了骄人的成就. 在科学研究领域, 人工智能的兴起也催生了以数据为驱动的新的科学研究范式, 推动了新材料的设计与开发^[1,2]. 在分子科学中, 人工智能也成功地应用在了药物设计、量子化学、结构生物学等领域. 同时, 人工智能在指导实验设计中发挥了重要的作用. 例如, 基于发光碳点数据通过AI辅助设计开发了性能显著提升的发光碳点^[3,4], 并有望在AI辅助合成中展现出更大的潜力.

深度学习的两个特征使其在应用于分子时很有前景. 首先, 深度学习方法可以处理“非结构化”数据表

示, 如文本序列、语音信号、图像和图形. 这种能力对分子系统特别有用, 化学家已经开发出分子表示, 可以在不同几何维度上捕获分子特性. 其次, 深度学习可以从输入数据中进行特征提取(或特征学习); 也就是说, 从输入表示中生成数据驱动的特征. 深度学习的这两个特征补充了“经典”机器学习应用程序(如定量结构-活性关系, QSAR), 其中分子特征, 即“分子描述符”, 使用基于规则的算法进行先验编码. 深度学习从非结构化数据中学习并提取高阶分子特征的这种能力使得深度学习在分子科学中有着广泛的应用.

分子特征的提取是将深度学习应用在分子科学领域极为重要的一步. 一般来说, 这一步包括结合几何先验的方法, 即输入信息的结构空间和对称特性信息, 实现不同形式的分子结构的表示. 利用几何先验来提高模型的质量, 可提升预测的准确性. 本文将以分子表

引用格式: Wang H, Jin L, Feng J, Dong Z, Deng W, Li Y. Molecular representation in AI-assisted synthesis chemistry: from sequence to graph. *Sci Sin Chim*, 2023, 53: 107–118, doi: 10.1360/SSC-2022-0113

示为核心, 主要介绍以下几个方面: (1) 对不同维度的分子表示在分子科学中的突出应用提供结构化和统一的概述; (2) 描绘该领域的主要研究方向; (3) 尝试对不同分子表示的应用进行批判性预测; (4) 重点介绍其在合成化学领域的应用。

2 利用分子字符串进行机器学习

2.1 分子的字符串表示

分子的字符串是由字母数字符号组成的线性序列, 用来表示一个具体的分子结构, 其最早可追溯到用于化学命名开发的补充工具^[5], 后来演变成分子结构数据的存储与检索方式, 也就是将实际分子结构转换为计算机可识别的文本语言。目前, 比较流行的分子字符串的表示方法包括8 line notation^[6]、InChI (国际化合物标识, international chemical identifier)^[7]和SMILES (simplified molecular input line entry specification)^[8]。对于一个分子结构, 可以用特定的分子字符串表示, 即每个分子都拥有其对应的字符串表示; 然而, 对于任意的字母、数字与字符组成的字符串, 其未必能生成相应的分子结构, 即不是每个字符串都有对应的分子结构。分子的字符串表示中的符号具有语义属性, 根据字符串中的原子符号、键符号和语言规则, 构成的分子结构具有不同的物理、化学和生物属性。

SMILES是一种用ASCII字符串明确描述分子结构的规范, 如图1a所示, 其中字母代表原子, 而符号和数字则代表键类型、连通性、分支以及立体结构, 被广泛应用于分子结构的存储和数据分析。由于SMILES没有相关的格式规范, 其在使用过程中存在一定的缺点, 如不适用于子结构搜索、一个分子存在不同的SMILES表达式等。为了解决这些问题, 研究者开发了Morgan算法等来生成规范SMILES (canonical SMILES)^[9], 保证了每个化学分子只有一个SMILES表达式。SMILES常用在基于序列的深度学习, 这主要得益于它的表示方法是基于序列的, 能够通过One-hot方法编码转换为深度学习中的数据表示^[10]。SMILES对分子结构进行编码时, 将化学结构转化成一个生成树, 继而采用纵向优先遍历树算法, 在对结构进行转化时, 会首先去掉氢, 如果分子中有环结构, 还需要把环打开, 并用数字标记被拆掉的键端的原子, 将支链写在小括号里。

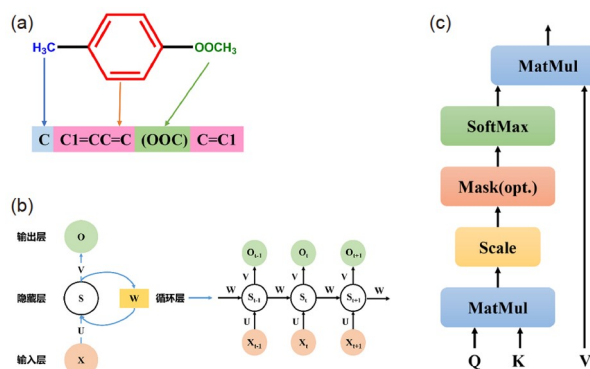


图1 分子的化学语言模型。(a) 分子的SMILES表达式; (b) RNN模型的基本框架; (c) self-attention机制的基本框架(网络版彩图)

Figure 1 Chemical language modelling. (a) SMILES strings; (b) recurrent neural networks; (c) self-attention (color online).

2.2 化学语言模型

化学语言模型是机器学习在化学中运用的基础之一, 能够将分子序列转换为适合模型训练的输入结构, 目前最常用的化学语言算法是循环神经网络(RNN)和Transformer。

2.2.1 循环神经网络(RNN)

RNN^[11,12]实际上是分子序列的一种神经网络表示, 通过one-hot编码将分子的序列数据转换为Euclidean结构。如图1b所示, 在这一模型中, 每一个时刻都有一个输入 x_t , 然后根据当前节点的状态 s_t 计算输出值 o_t , 而 s_t 是根据上一时刻的状态 s_{t-1} 和当前的输入 x_t 共同决定的。RNN的优势在于对分子结构数据的兼容性好, 可以对任意长度的序列输入进行处理, 并转换为任意长度的输出。在实际运用中, 该模型通常以自身做回归变量, 即利用前期若干时刻的随机变量的线性组合来描述以后某时刻的随机变量, 这种模型也称为“自回归”模型。然而普通的RNN模型在训练中会出现梯度消亡和梯度爆炸^[13], 为了解决这一问题, 研究者开发了长短期记忆网络(long short-term memory, LSTM)^[14]和门控循环单元(gated recurrent unit, GRU)^[15]。

RNN模型在分子设计中有着广泛的应用, 其能够很好地与迁移学习^[16,17]和强化学习^[18,19]相结合, 对分子进行逆向设计, 开发具有理想理化性质的分子结构, 这主要得益于RNN模型在学习SMILES语法规则和识

别高维分子特征等方面的优异能力. 同时, 随着基于非规范 SMILES 字符串的数据增强^[20,21]和双向学习技术^[22]的应用, RNN 模型对化学语言的识别与转换越来越可靠, 其模型训练的质量也越来越高. 目前, 很多基于机器学习的分子设计都采用 SMILES 字符串以及基于 SMILES 字符串的表达形式. 在多肽设计中, 常常使用单字符的氨基酸序列^[23], 这使得 RNN 模型很适用于预测配体和蛋白质的相互作用^[24]、药物动力学^[25]、蛋白质二级结构^[26]和分子演化轨迹^[27]. 在一些分子的虚拟筛选案例中^[28], RNN 模型的表现要优于传统的分子描述符和图卷积神经网络.

2.2.2 Transformer

人类视觉研究表明, 当观察一件物体时, 一般会将注意力集中注视着某一处, 而不会关注全部所有信息. 注意力机制(attention)和人类视觉的注意力机制类似, 就是在众多信息中把注意力集中放在重要的点上, 选出关键信息, 而忽略其他不重要的信息^[29]. 如图 1c 所示, 注意力机制主要指的是查询(Q)到主键(K)和值(V)对的映射, 其首先将 Q 和 K 进行点乘计算, 然后再通过 Softmax 等函数获取 Q 与 K 之间的相关性, 最后将该结果与 V 进行点乘, 从而得到最终的值. Transformer 引入了自注意力(self-attention)机制(图 1c), 通过堆叠 Encoder-Decoder 结构, 来对序列更好的建模^[30]. 不同于 RNN 模型的是, 序列数据经过 Transformer 处理后, 会转变为非 Euclidean 结构. Transformer 提供了两种位置编码(positional encoding)的方法(用三角函数直接计算或者学习出位置编码), 将 encoding 后的数据与 embedding 数据求和, 加入了相对位置信息. Transformer 提供了一种可并行处理顺序数据的方式, 其效率高于 RNN 等其他模型, 而且在处理长距离元素之间的关系依赖方面也非常出色, 在自然语言翻译中有着很好的表现.

Transformer 在 CASP 中发挥了很大的作用, 能够完成 sequence-to-sequence 的过程, 这个过程中, 反应物的字符串表示会映射到相应产物的字符串表示上, 反之亦然. 因此, Transformer 常用于预测多步合成的步骤^[31]、酶反应^[32]、反应产率^[33]以及其他一些分子性质^[34]. 其中, Transformer 在蛋白质设计中的效果尤为突出. 和 RNN 模型一样, Transformer 也适用于分子的从头设计, 其能够将具有理想性质的目标蛋白质分子转换为相应配体组成的 SMILES 字符串^[35]. 此外, Trans-

former 通过与 E(3) 和 SE(3) 等变层相结合, 可以直接由蛋白质的氨基酸序列预测蛋白质的三维结构^[36].

3 利用网络进行机器学习

机器学习所用的数据可以来源于网络. 网络将整个体系包络其中, 通过网络上的点提取体系的特征信息, 这些点是均匀、等近距离地分布在网络中. 网络可以是一维的(序列)、二维的(图像)、三维的(立方格)或者是更高维度的, 这取决于深度学习的对象和目标. 当学习对象为图像时, 采用二维的网络比较合适, 通过在图像上进行取点, 可以将点的颜色作为输入数据进行学习. 每一种都是由欧氏几何定义的, 且网络中的点提取出来的数据都会作为一组输入数据中的一个值, 这使得网络中的点具有以下两个特性: 一是网络中的每一个点都具有相同的邻接结构, 所以从结构上来看每一个点是不可区分的; 二是网络中点与点之间的排列顺序和临界顺序是固定的, 这是因为在决定网络的维度时, 每一个点的位置与在该点提取到的输入数据在矩阵中的位置已经是一一对应. 基于以上两个特性, 在利用网络学习时, 则不用再考虑置换不变性, 并且该过程具有平移不变性和尺度分离两种更强的几何先验^[37]. 也正是基于以上特点, 网络学习在计算机视觉领域取得了很大的成功.

3.1 分子网络

在处理分子系统时, 可以采用二维和三维网络的方法提取分子体系中的特征信息. 二维的分子网络, 如用凯库勒表示方法表示分子结构, 可以清晰地描述出分子中原子的种类、数量及其在二维平面上的相对位置, 这种简单明了的表示方法在分子可视化方面取得了较大的成功^[38]. 当学习目标涉及分子的结构立体信息时, 可以使用三维分子网络对分子系统进行信息提取, 进而有助于生物分子和药物的研究^[39]. 传统的学习方法常常使用等变图神经网络(GNN)与 3D 卷积神经网络(3DCNN). 相比于等变 GNN, 3DCNN 具有更高的资源效率. 从可处理的分子复杂程度上来说, 等变 GNN 可以处理分子体系所包含的原子通常在一千个以下, 而 3D CNN 可以处理更加复杂的分子体系, 更适用于生物大分子(如蛋白质, 通常包含一万甚至十万个以上原子)^[40,41].

3.2 学习分子表面

在处理分子所使用的网格中有一种特殊类型的网格——分子的表面, 分子表面不规则且并不类似于材料的表面, 在考虑分子表面时, 通常将以分子中各个原子中心为圆心, 一定距离为半径的所有球面所构成的曲面视为分子表面^[42] (图2). 分子表面网格上的点包含分子和其中临近原子的化学和几何信息. 分子表面的网格通常是三维的, 网格上的点为顶点的多边形可以作为网格坐标在三维空间的存在形式. 同时, 三维网格上的点也可以用二维网格表示出来. 相比于3D CNN, 2D CNN在分析分子表面时, 无法考虑分子旋转状态对于提取数据的影响, 而且在取点过程中, 网格的点不好做到完全的均匀分布和距离相等. 目前已经有文章证明, 通过引入GNN学习网格处理过的分子结构的方法可以将分子系统的旋转等方差纳入网格结构, 从结果上来看这样的GNN计算效率更高, 且具有应用于模拟大分子结构的潜力^[43]. 然而, 这种GNN方法目前还没有应用到分子系统. 此外, 还有一些研究专注于利用三维网格提取的分子表面的特征信息作为3D RNN的输入数据, 如蛋白质-配体结合位点的预测^[44].

4 利用分子图进行学习

4.1 图神经网络

图是最直观地表示分子结构的方法之一^[45,46], 在计算机科学中, 图 $G=(V, E)$ 是由顶点和边两部分组成

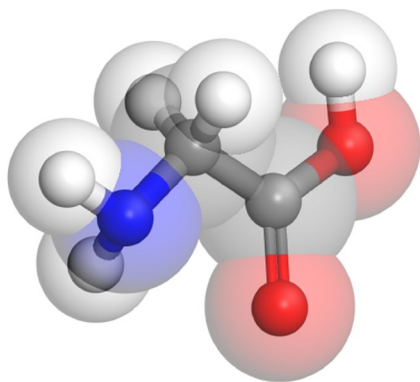


图2 甘氨酸分子表面示意图(网络版彩图)

Figure 2 Schematic diagram of the surface of glycine molecule (color online).

的数据结构, 其中 G 表示分子图, V 表示顶点集合, E 表示边的集合, 根据顶点之间是否存在方向关系, 边可以是有向的, 也可以是无向的, 使用邻接矩阵来描述(图3a). 分子可以被看作是一个图, 分子中, 原子可以用顶点($v_i \in V$)来描述, i 和 j 原子间的化学键可以用边($e_{ij} \in E$)来描述(图3b). 在大多数应用中, 顶点和边通过一维的特征向量来实现^[47], 这些特征向量可以由元素的基础性质或结合键长度等经过处理得到, 如CGCNN由9种元素性质组成节点的特征向量, 包括原子序数、周期数、族数、电负性等, 在一定程度上能够反映元素的本质. 三维分子结构通常表示为 $G3D=(V, E, R)$ 或者3D点云 $P3D=(V, R)$, 除了顶点和边缘特征, 顶点的位置信息($r_i \in R$)也被编码^[48]. 3D点云是3D图的一种特殊情况, 没有预先定义的边, 顶点之间通过给定的半径和完整的边集合来确定连接关系.

以图作为算法输入信息的深度学习方法通常称为图神经网络(GNN)^[49]. 目前, 大多数GNN基于消息传递机制^[50], 包含两个基本的步骤, 邻居聚集和信息传递(图4a), 下面的公式描述了消息传递的过程:

$$v_i^l = U^{(l)}\left(v_i^{l-1}, \sum M\left(v_i^{l-1}, v_j^{l-1}, e_{ij}\right)\right)$$

其中 $i, j \in V$, U 表示更新函数, M 表示消息传递函数. 每个顶点可以表示为周围顶点和自身信息的叠加, 其核

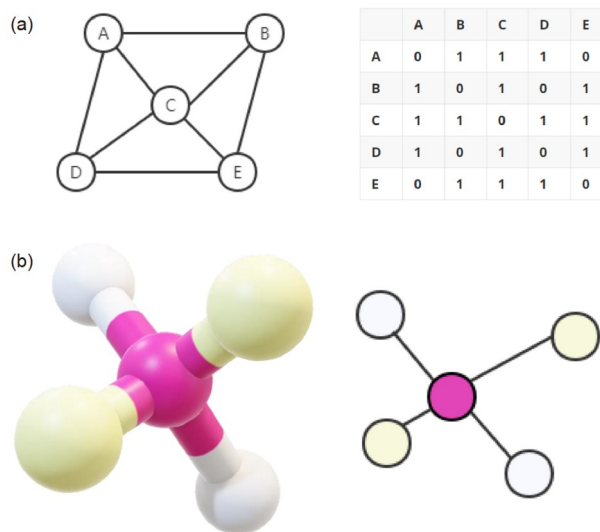


图3 (a) 图及其邻接矩阵; (b) 分子结构与图表示(网络版彩图)

Figure 3 (a) Graph and its adjacency matrix; (b) molecular structure and graph representation (color online).

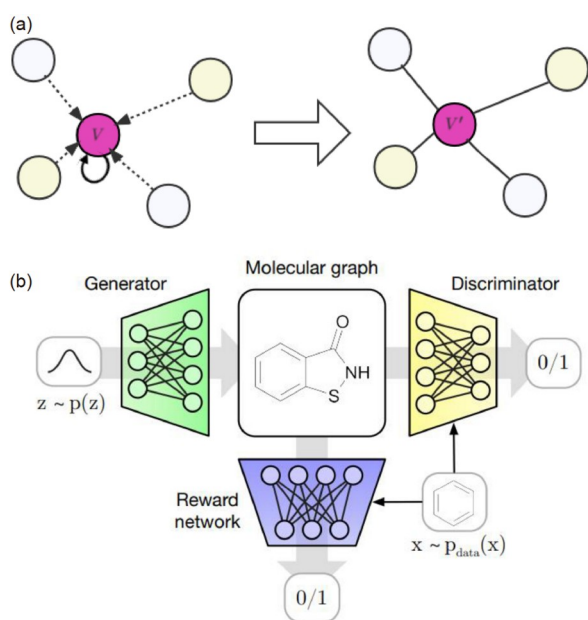


图 4 (a) 消息传递模式; (b) 基于图的分子生成模型架构(网络版彩图)

Figure 4 (a) Message passing pattern; (b) graph-based molecular generative model architecture (color online).

心在于定义顶点之间的聚合函数^[51]。针对不同应用场景, 目前已经开发出了多种图神经网络架构, 如图卷积神经网络(GCN)、图注意力网络(GAT)、图生成网络(GraphGAN)等^[45,52,53]。在分子领域的应用中, 图神经网络通过逐步聚合来自原子及其周围环境的信息来进行特征提取, 可以应用到图像级别或者节点级别的任务^[54], 来完成端到端的准确的性质预测或结构生成^[55]。

4.2 图神经网络在量子化学中的应用

图神经网络通过在顶点特征中加入原子属性信息, 在边缘特征中加入距离和角度等3D信息^[37], 从而在预测小分子量子化学性质上可以达到密度泛函理论(DFT)计算的精度, 在量子化学、药物开发、CASP和分子性质预测中有许多应用^[54,56-58]。谷歌公司在2017年通过总结现有模型, 提出了消息传递网络架构(MPNN), 作者证明了图神经网络模型可以不受图同构^[59]的影响, 从分子图中学习分子的特征, 从而能够应用于化学预测的任务, 如预测有机分子的性质^[50]。SchNet使用连续滤波卷积层, 能够获得连续的势能面和力场, 从而避免了分子势能面不连续导致的物理信

息丢失^[60]。SphereNet提出了球形信息传递(SMP), 能够区分手性等复杂的三维结构信息, 作者使用基于物理学的表示方法, 结合球状信息传递作为神经网络的输入, 将图神经网络的应用拓展到了三维图结构数据^[61]。在一些生物学相关性质方面, 图神经网络已经被证明其表现能够优于手工设计描述符的模型, 并被认为是可解释性方面有着巨大的潜力^[58,62,63]。此外, 结合变分自动编码器(VAE)、强化学习(RL)或生成对抗网络(GAN)模型, 提出了图自编码器(GAE)、变分图自编码器(VGAE)、图生成对抗网络(GraphGAN)等模型^[53,64-66], 可以使用图神经网络从头生成分子, 其基本架构如图4b所示。

5 辅助合成中的应用

5.1 AI辅助合成规划的意义

合成规划的基本思路是利用正逆合成操作(如切断、转化)在目标产物、中间体以及起点反应物之间互推, 逆合成分析能直接预测目标产物合成路线, 而正合成分析可用于辅助验证合成路线设计的正确性或探索未知的反应。正逆合成操作的依据是一系列物理化学理论(如价键理论、前线轨道理论)或合成经验(根据反应规则构建的反应模板), 其本质上是将一个复杂的合成路线规划分解为多步相对简单的合成规划。以机器学习算法为代表的人工智能技术在生物信息学等领域展现出的显著功效, 暗示其也具有从大量实验学习未知规则和经验的巨大潜力, 可能在辅助合成规划的同时规避引入和比对反应模板的困难或高成本。因为通过传统图同构分析算法判断目标反应是否适用模板至少是NP难的, 计算量较大。由于正逆合成操作都是在分子结构式上进行, 因此将几何描述符作为合成规划的分子表示是很自然的。在三类几何描述符中, 图和字符串显然最容易变换回反应物和中间体结构式, 从而为人所理解或与反应模板对照。并且, 可直接对这两种描述符进行的操作与各种正逆合成操作的对应很明确, 如切断化学键可与删除节点/字符对应, 方便计算机模拟。这些特点或许就是目前辅助合成规划的模型全部基于图和字符串描述符的原因。不同描述符提供的信息不同, 不同机器学习模型适用的描述符也不同, 这就导致描述符也是影响机器学习性能的一个决定因素。

5.2 预测单步反应

机器学习在合成规划中的一类可行应用便是预测某步反应可能的一系列反应物或产物, 并对它们的合理性(结果为真的概率)进行评分排名. 表1比较了目前基于字符串和图两类描述符实现该应用的各无模板模型性能. 结合工作发表时间, 可以看到, 描述符类别、模型架构, 以及改进程度是造成相应性能差异的主要原因. 多数模型还是用于正向预测产物, 其中采用分子字符串描述符的多是seq2seq或Transformer网络的变体(图5a), 其可以看作是基于束搜索(beam search, 图5c)将反应物/产物互相“翻译”. 而采用分子图描述符的多是Weisfeiler-Lehman网络(WLDN, 图6a)及其变体^[55,67] (如图卷积网络GCN, 图6b^[45]), 通过提取反应中心来限制反应物/产物的枚举范围(图6a), 并用独立的神经网络为各枚举结果打分. 总的来说, 这些模型分析能力都可与专业有机合成工作者相比拟, 并且往往可通过注意力机制学习反应中特定原子或原子对权

重, 提供可解释性. 在非立体化学的预测中, 较新的2D图网络模型(如Jensen组的图卷积网络WLDN5^[55])有着明显优于seq2seq^[10]的精度, 这或许可归因于图比字符串更直接地反映了分子的拓扑结构, 且其节点和边可额外编入描述组成元素以及相互作用的详细信息(如原子序数、化合价、键级、芳香性), 但这也导致这类网络需要更多计算资源, 分析速度较慢. 应注意到, 2D图难以完全纳入立体化学信息, 而尝试用2D图网络预测立体化学反应的效果也确实很差^[68]. 3D图则需要成本和占用更大的原子空间坐标, 尚无用于合成规划的案例. 而SMILES字符串不仅占用较小, 兼容立体化学信息, 还具有现成的数据增强算法^[69]. 将其与善于捕捉分子中全局关系且具有弱归纳偏置的先进Transformer网络结合, 可产生多种优势. 例如, Schwaller等开发的Molecular Transformer不仅克服了2D图网络的相关限制, 还在精度、分析速度、数据预处理成本(无需拆分反应物和溶剂等其他试剂的描述符)以及泛化能力上比WLDN5等先前报道的模型更好^[70]. 并且, 基于

表 1 用于正向/逆向预测单步反应的各种神经网络模型的目标、所用分子描述符、架构及其性能^{a)}

Table 1 Objectives, molecular descriptors used, architecture and performance of each neural network model for forward/backward prediction of single-step responses

工作	目标	描述符	模型架构	最佳top-1精度(%)	平均每个反应预测时间(ms)
[6]	正向预测	SMILES	Seq2seq	80.3	25
[51]	正向预测	2D分子图	WLDN	85.6	100
[74]	正向预测	2D分子图	WLDN	79.6	>50
[63]	正向预测(立体化学)	2D分子图	WLDN	<0.5	—
[65]	正向预测((非)立体化学)	SMILES	Transformer	90.4 (非立体化学) 78.1 (立体化学)	~30
[74]	逆向预测	SMILES	Seq2seq	37.4 (已知反应类型)	—
[67]	逆向预测	SMILES	Transformer	63.0 (已知反应类型) 41.5 (反应类型未知)	—
[69]	逆向预测	SMILES	Transformer	59.0 (已知反应类型)	—
[70]	逆向预测	SMILES	Transformer	63.0 (已知反应类型)	—
[63]	逆向预测	SMILES	Seq2seq	37.4 (已知反应类型)	—
[63]	逆向预测	2D分子图	GLN ^[77] (基于模板)	64.2 (已知反应类型) 52.2 (反应类型未知)	—
[63]	逆向预测	SMILES	Transformer	37.9 (反应类型未知)	—
[63]	逆向预测	2D分子图	图变分自编码器	61.0 (已知反应类型) 48.9 (反应类型未知)	—

a) 对于单个反应预测, top-1结果是指模型给出的多个预测结果中评分最高(最可能/合理)的. 对于表中模型, 如无特别说明, 则代表仅用于非立体化学分析. 在预测时间统计中, —代表文献中未给出. 文献[6,51,63,66]使用相同数据集. 在非立体化学预测中, 约85%用于训练, 8%用于测试. 在立体化学预测中, 约90%用于训练, 5%用于测试. 文献[64,68,70,71]使用相同数据集, 其中文献[64,68]中80%用于训练, 20%用于测试; 文献[70,71]中90%用于训练, 10%用于测试.

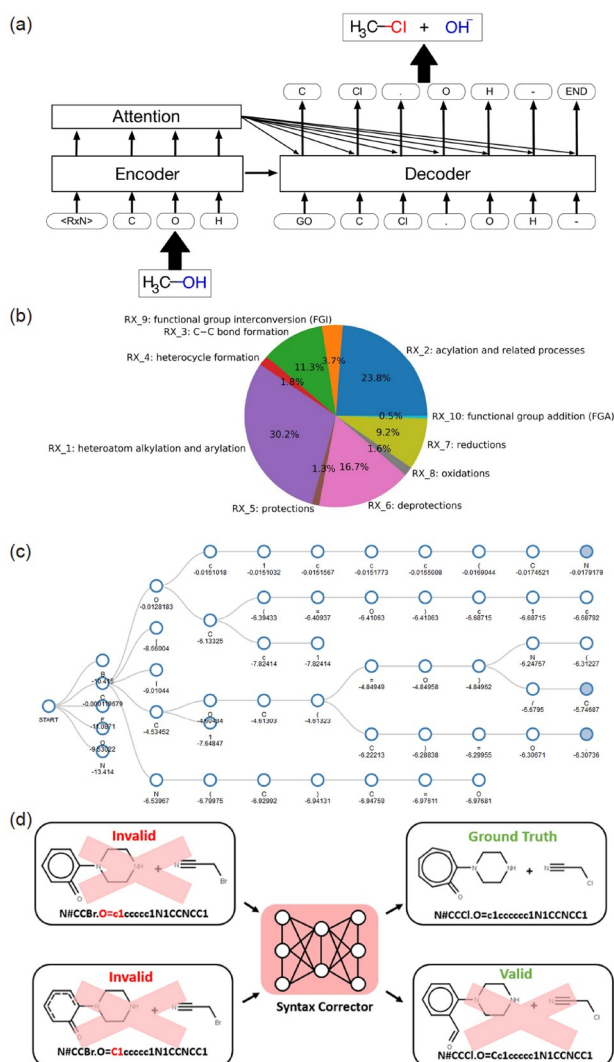


图 5 使用字符串描述符的神经网络模型示意图(seq2seq、transformer及其变体). (a) 通用的编码器-注意力掩码-解码器架构进行合成预测的流程, <Rxn>代表反应类型^[8,78]; (b) 美国专利商标局(USPTO)反应数据库中的10种反应类型^[68]; (c) 束搜索^[8,78]和(d)用单独的神经网络矫正存在语法错误(标红部分)且不合理的SMILES字符串输出^[68] (网络版彩图)

Figure 5 Schematic sketches of string-based neural network models (seq2seq, transformer, and their variants). (a) Process for synthesis prediction by the general encoder-attention-decoder architecture [78]. (b) 10 reaction classes within the United States Patent and Trademark Office (USPTO) database [68]. (c) Beam search [78] and (d) the correction of unreasonable SMILES output by an independent neural network [68] (color online).

Transformer网络甚至可以通过无监督学习提取直观的单步有机反应规则和可视化各种有机反应之间的关系^[33,71]. 尽管基于字符串的模型可能产生语法无效的结果, 但单独引入用于语法校正的神经网络(图5d)可

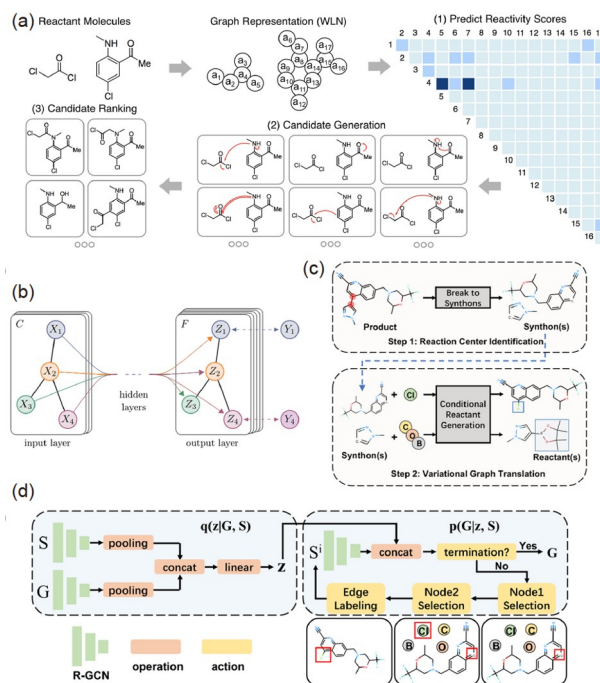


图 6 使用分子图的神经网络模型示意图. (a) 基于WLDN架构的反应预测流程^[63]. (b) GCN架构^[45], 在其所用分子图中, 原子表示为特征为 X 的节点, 化学键表示为边, 标签 Y 指明该原子是否为反应中心. (c) 唐建等所用模型的预测流程及其(d)第二步所用图变分自编码器架构, 其中 $q(z|G, S)$ 是一个编码器, $p(G|z, S)$ 是一个生成模型^[68] (网络版彩图)

Figure 6 Schematic sketches of graph-based neural network models [63]. (a) WLDN-based workflow of reaction prediction. (b) GCN architecture [45], where the bonds and atoms are represented with the edges and the nodes of feature X respectively in molecular graphs. Label Y indicates whether the atom is a reaction center. (c) Workflow of the model proposed by Tang *et al.*, and (d) the graph variational auto-encoder architecture used in the second step, where $q(z|G, S)$ is an encoder, and $p(G|z, S)$ is a generative model [68] (color online).

大大减少无效结果的比例^[10,72]. 使用改进的字符串(如SELFIES^[73]), 甚至可以完全避免该问题. 单步反应的逆向预测则比正向预测困难得多, 因为正向反应结果是唯一的, 而同一产物可以靠不同反应得到. 一般来说, 训练集包含的可能反应有限, 这也使得无模板机器学习法仍难以在精度上胜过借助模板的方法, 即便是Transformer网络也不理想. 一个提升模型性能的方法是引入反应类型这一额外信息(图5b), 但其大大增加了数据预处理的成本^[72,74,75]. 唐建等^[76]将单步反应逆向预测转化为数据生成任务, 基于分子图和变分自编码器构建了一个graph to graphs框架(图6c, d). 即便反应类型未知, 该架构的模型效果仍明显优于先前的Transformer网络, 且可与借助模板的方法媲美.

5.3 合成路径搜索

尽管机器学习在单步反应的预测已经取得了可观的成果, 然而仅通过这些模型获取完整合成路线仍是不可行的. 原因主要有两点: (1) 每步预测累计上一步的误差, 因此预测路线的误差会随合成步数呈指数上升^[70]. (2) 有些合理性较高的单步反应, 其所对应的完整合成路线却不一定合理, 要对可能合成路线进行比较, 只能先将它们枚举. 然而这非常困难, 因为可能的合成路线数目也随合成步数近似呈指数增长.

由于逆合成分析的每步逆推往往会产生分支, 因此最终分析结果实际不是单一的反应路线, 而是由共享某些中间体的多条反应路线, 即除了最合适的路线, 还有一些相对不那么有吸引力的. 这些路线共同构成了“合成树”. 因此, 逆合成分析可以转化为一个关于树结构(也可看作图结构)的数学问题, 而这正是计算机算法擅长解决的. 近年, 人工智能技术在树/图搜索(如围棋竞赛)中取得了举世瞩目的成绩, 相应的代表性算法包括蒙特卡洛搜索树(MCTS)和A*算法, 它们本质上属于强化学习算法. 这两种算法也分别被Segler等^[79](图7a)和李成涛等^[80](图7b)用于完整合成路线

的搜索中. 这些算法在统计学上保证了合成路线整体的合理性, 比连续重复采用单步反应预测算法更科学, 最终可以只给出最合理的若干条反应路线, 显著克服了单步反应预测算法的相关局限.

在他们的算法中, 合成路径上的反应物、中间体、产物用树结构中的节点表示. 这些算法通过迭代延伸合成路径的过程就像是树的生长, 其每个迭代一般包括4个步骤: 选择、扩展、试验、更新.

选择(selection): 根据已探索分支的评分和规则, 选择最合适的叶节点作为最终合成路径上经过的点.

扩展(expansion): 对某步反应已确定产物的分子图或分子字符串进行可用的操作, 即模拟一步逆合成, 得到本步反应可能的一系列反应物分子结构, 作为新的叶节点添加, 产生新的分支以供探索. 在李成涛等^[80]的算法中, 还会利用神经网络评估新添加叶节点的价值函数.

试验(rollout): 模拟对当前节点对应物质继续做逆合成时, 可否在最大追溯深度内追溯到可购买到的构建块来判断路线的可行性. Segler等^[79]的MCTS算法采用的是蒙特卡洛法, 具体是通过对分子图进行若干次(次数上限由最大追溯深度决定)反应模板允许的随机操作, 以随机模拟逆合成操作, 而李成涛等^[80]的A*算法无rollout步骤.

更新(update): 根据试验的结果, 按照一定规则更新对已探索分支的评分.

在这两种算法中, 无论是筛选扩展步骤可用操作, 还是评估价值函数所用的神经网络使用的仅仅是简单的向量描述符, 如ECFP4^[81]或Morgan^[82]指纹. 原因可能在于基于这些描述符的机器学习模型成本非常低. 尽管也可能使用分子图和字符串这样的几何表示作为输入, 但尚未见到相关应用, 可能原因是使用这些描述符的模型成本偏高.

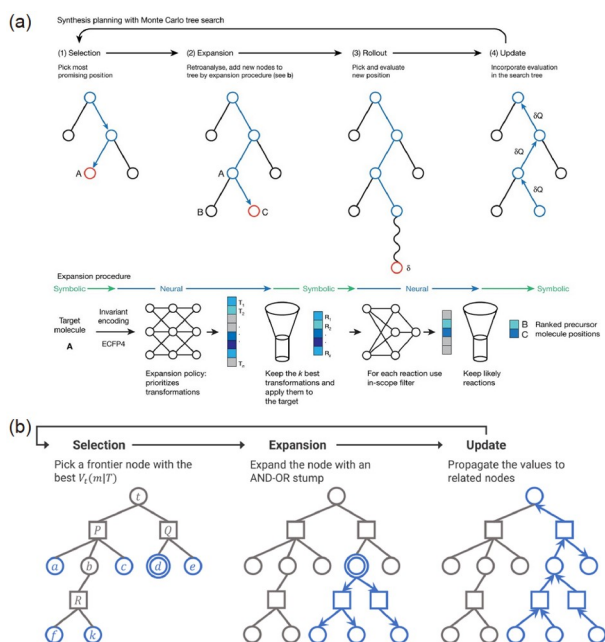


图7 (a) Segler等^[79]所用MCTS和(b) 李成涛等^[80]所用A*算法进行逆合成预测的流程示意图(网络版彩图)

Figure 7 Schematic sketches for retrosynthesis prediction by the (a) MCTS and (b) A* algorithm introduced by Segler *et al.* ^[79] and Li ^[80] *et al.*, respectively (color online).

6 展望

各种分子几何表示在AI合成化学中都有着巨大的潜力, 但仍应理性认识到不同表示和算法的组合存在着如下的优势和问题.

分子字符串和序列模型. 优势: (1) 字符串表示的内存占用小. (2) 提高模型性能的方法较为多样, 包括双向学习和语言数据增强. (3) Transformer具有优良的

全局关系捕捉和迁移学习能力。业界流行的预训练模型多是Transformer架构。(4) 模型的注意力机制增强了可解释性。(5) 该组合在合成化学中的应用目前最为广泛。问题: (1) 模型可能产生语法无效的SMILES字符串, 需额外引入语法矫正。(2) SMILES的化学意义不够明确, 一种化学结构可对应多种序列结构。且由于SMILES中某些化学键被省略, 序列模型对它们的可解释性也一般。(3) 采用SMILES以外新语言的模型还很少。

分子图和图模型。优势: (1) 该组合具有最适用于化学拓扑结构的归纳偏置, 可将节点和边直接抽象原子和化学键, 具有最佳的可解释性。(2) 分子图定义明确, 和实际分子结构彼此一一对应, 具有置换和平移不变性。(3) 可以编入所有化学键(边)的特征。(4) 图模型在当前神经网络算法中创新热度最高。问题: (1) 表示的内存占用较高, 模型运算的资源运用效率较低。(2) 在高效纳入立体化学信息方面仍存在困难。

分子网格和卷积模型。优势: (1) 将连续空间的理化性质离散化, 以供计算机处理的最直接组合。(2) 模型的运算效率较高。(3) 模型具有置换不变性和平移等变性。(4) 已广泛用于分析特别复杂的化学结构, 如蛋白质。(5) 较适用于局部化学环境, 及多个化学结构间相互作用的分析。问题: (1) 表示的内存占用取决于化学结构的体积, 表示小分子的效率低。(2) 缺少直接的合成化学应用。(3) 容易忽略化学体系中长程或全局的联系。

总的来看, 基于分子字符串的算法发展得最为成熟。随着新型字符串表示的应用, 序列模型的效率仍有提升空间。在各序列模型中, Transformer的表现最

为突出, 其高的迁移学习效率, 使其很容易根据实际需要进行微调。因此, 分子字符串和序列模型的组合最有可能首先推广到工业界和实验室。而对于分子图模型, 随着算力的进一步提升, 其将在速度上达到和其他模型相近的水准, 从而得到进一步推广。此外, 分子图模型巨大的改进空间为相关创新提供了强劲动力。首先, 分子图最自然的归纳偏置, 意味着更容易通过架构和特征设计, 来将合成化学领域的先验知识引入模型。同时, 受益于计算机领域对图模型的高度关注, 分子图模型在立体化学方面遇到的挑战也可能被加速解决。分子网格和卷积模型, 始终是分子(尤其是大分子)相互作用研究的中流砥柱, 将持续为包括“蛋白质人工合成”这一“圣杯”在内的合成化学问题提供重要的微观尺度认识。尽管网格卷积模型难以直接给出合成路线, 但可充分从局部化学环境的角度揭示候选合成路线步骤的热力学和动力学机理。

除了表示和模型本身, 还应注意生成模型与强化学习等策略层面的AI技术。在当今AI合成化学的最前沿, 我们总能见到这些技术的身影。同样, 要继续推进这个领域的前沿, 就必须将这些策略层面的技术进一步与上述模型融合。这最终有助于进一步将合成艺术从“筛选”推向“创造”。然而, 合成化学领域对前沿AI技术的应用仍处于“萌芽”阶段, 因此必须以进一步挖掘并消化计算机科学的已有成果作为创新AI合成化学的根基。具体来说, 在合成路径搜索中引入包括时序差分算法在内的更多已开发的强化学习算法, 以及采用其他生成模型, 如生成对抗网络取代变分自编码器预测单步反应, 并在学习过程中融入强化学习等, 都是秉持上述根基的可能创新方向。

参考文献

- 1 Wang H, Ji Y, Li Y. *WIREs Comput Mol Sci*, 2020, 10: e1421
- 2 Mao X, Wang L, Xu Y, Wang P, Li Y, Zhao J. *npj Comput Mater*, 2021, 7: 46
- 3 Wang X, Wang H, Zhou W, Zhang T, Huang H, Song Y, Li Y, Liu Y, Kang Z. *J Photochem Photobiol A-Chem*, 2022, 426: 113729
- 4 Wang X, Wang B, Wang H, Zhang T, Qi H, Wu Z, Ma Y, Huang H, Shao M, Liu Y, Li Y, Kang Z. *Angew Chem Int Ed*, 2021, 60: 12585–12590
- 5 Wiswesser WJ. *J Chem Inf Comput Sci*, 1985, 25: 258–263
- 6 Wiswesser WJ. *Chem Eng News Archive*, 1952, 30: 3523–3526
- 7 Heller S, McNaught A, Stein S, Tchekhovskoi D, Pletnev I. *J Cheminform*, 2013, 5: 1–9
- 8 Weininger D. *J Chem Inf Model*, 1988, 28: 31–36
- 9 O'Boyle NM. *J Cheminformatics*, 2012, 4: 1–14
- 10 Schwaller P, Gaudin T, Lányi D, Bekas C, Laino T. *Chem Sci*, 2018, 9: 6091–6098

- 11 Mikolov T, Karafiát M, Burget L, Cernocký J, Khudanpur S. *Interspeech*, 2010, 2: 1045–1048
- 12 Rumelhart DE, Hinton GE, Williams RJ. Learning internal representations by error propagation. In: Rumelhart DE, McClelland JL, Eds. *Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Volume 1: Foundations*. Cambridge: MIT, 1986
- 13 Hochreiter S. *Int J Unc Fuzz Knowl Based Syst*, 1998, 6: 107–116
- 14 Hochreiter S, Schmidhuber J. *Neural Comput*, 1997, 9: 1735–1780
- 15 Chung J, Gulcehre C, Cho K, Bengio Y. arXiv: 1412.3555, 2014
- 16 Gupta A, Müller AT, Huisman BJH, Fuchs JA, Schneider P, Schneider G. *Mol Inf*, 2018, 37: 1700111
- 17 Merk D, Friedrich L, Grisoni F, Schneider G. *Mol Inf*, 2018, 37: 1700153
- 18 Olivecrona M, Blaschke T, Engkvist O, Chen H. *J Cheminform*, 2017, 9: 48
- 19 Popova M, Isayev O, Tropsha A. *Sci Adv*, 2018, 4: eaap7885
- 20 Arús-Pous J, Johansson SV, Prykhodko O, Bjerrum EJ, Tyrchan C, Reymond JL, Chen H, Engkvist O. *J Cheminform*, 2019, 11: 71
- 21 Bjerrum EJ, Threlfall R. arXiv: 1705.04612, 2017
- 22 Grisoni F, Moret M, Lingwood R, Schneider G. *J Chem Inf Model*, 2020, 60: 1175–1183
- 23 Das P, Sercu T, Wadhawan K, Padhi I, Gehrmann S, Cipcigan F, Chenthamarakshan V, Strobelt H, dos Santos C, Chen PY, Yang YY, Tan JPK, Hedrick J, Crain J, Mojsilovic A. *Nat Biomed Eng*, 2021, 5: 613–623
- 24 Zheng S, Li Y, Chen S, Xu J, Yang Y. *Nat Mach Intell*, 2020, 2: 134–140
- 25 Wang X, Liu M, Zhang L, Wang Y, Li Y, Lu T. *J Chem Inf Model*, 2020, 60: 4603–4613
- 26 Senior AW, Evans R, Jumper J, Kirkpatrick J, Sifre L, Green T, Qin C, Židek A, Nelson AWR, Bridgland A, Penedones H, Petersen S, Simonyan K, Crossan S, Kohli P, Jones DT, Silver D, Kavukcuoglu K, Hassabis D. *Nature*, 2020, 577: 706–710
- 27 Tsai ST, Kuo EJ, Tiwary P. *Nat Commun*, 2020, 11: 5115
- 28 Gómez-Bombarelli R, Wei JN, Duvenaud D, Hernández-Lobato JM, Sánchez-Lengeling B, Sheberla D, Aguilera-Iparraguirre J, Hirzel TD, Adams RP, Aspuru-Guzik A. *ACS Cent Sci*, 2018, 4: 268–276
- 29 Chaudhari S, Mithal V, Polatkan G, Ramanath R. *ACM Trans Intell Syst Technol*, 2021, 12: 1–32
- 30 Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I, arXiv: 1706.03762, 2017
- 31 Schwaller P, Petraglia R, Zullo V, Nair VH, Haeuselmann RA, Pisoni R, Bekas C, Iuliano A, Laino T. *Chem Sci*, 2020, 11: 3316–3325
- 32 Kreutter D, Schwaller P, Reymond JL. *Chem Sci*, 2021, 12: 8648–8659
- 33 Schwaller P, Probst D, Vaucher AC, Nair VH, Kreutter D, Laino T, Reymond JL. *Nat Mach Intell*, 2021, 3: 144–152
- 34 Morris P, St. Clair R, Hahn WE, Barenholtz E. *J Chem Inf Model*, 2020, 60: 4191–4199
- 35 Grechishnikova D. *Sci Rep*, 2021, 11: 321
- 36 Baek M, DiMaio F, Anishchenko I, Dauparas J, Ovchinnikov S, Lee GR, Wang J, Cong Q, Kinch LN, Schaeffer RD, Millán C, Park H, Adams C, Glassman CR, DeGiovanni A, Pereira JH, Rodrigues AV, van Dijk AA, Ebrecht AC, Opperman DJ, Sagmeister T, Buhlheller C, Pavkov-Keller T, Rathinaswamy MK, Dalwadi U, Yip CK, Burke JE, Garcia KC, Grishin NV, Adams PD, Read RJ, Baker D. *Science*, 2021, 373: 871–876
- 37 Bronstein MM, Bruna J, Cohen T, Veličković P. arXiv: 2104.13478, 2021
- 38 Oldenhof M, Arany A, Moreau Y, Simm J. *J Chem Inf Model*, 2020, 60: 4506–4517
- 39 Lv Z, Ao C, Zou Q. *Proteomics*, 2019, 19: 1900119
- 40 Amidi A, Amidi S, Vlachakis D, Megalooikonomou V, Paragios N, Zacharaki EI. *PeerJ*, 2018, 6: e4750
- 41 Kozlovskii I, Popov P. *J Chem Inf Model*, 2021, 61: 3814–3823
- 42 Laskowski RA. *J Mol Graphics*, 1995, 13: 323–330
- 43 Pfaff T, Fortunato M, Sanchez-Gonzalez A, Battaglia PW. arXiv: 2010.03409, 2020
- 44 Zhao J, Cao Y, Zhang L. *Comput Struct Biotechnol J*, 2020, 18: 417–426
- 45 Kipf TN, Welling M. arXiv: 1609.02907, 2016
- 46 Duvenaud DK, Maclaurin D, Iparraguirre J, Bombarell R, Hirzel T, Aspuru-Guzik A, Adams RP. Convolutional networks on graphs for learning molecular fingerprints. In: *Proceedings of the 28th International Conference on Neural Information Processing Systems*. Montreal, 2015. 2224–2232
- 47 Monti F, Boscai D, Masci J, Rodola E, Svoboda J, Bronstein MM. Geometric deep learning on graphs and manifolds using mixture model CNN. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Hawaii, 2017. 5115–5124

- 48 Fang X, Liu L, Lei J, He D, Zhang S, Zhou J, Wang F, Wu H, Wang H. *Nat Mach Intell*, 2022, 4: 127–134
- 49 Zhou J, Cui G, Hu S, Zhang Z, Yang C, Liu Z, Wang L, Li C, Sun M. *AI Open*, 2020, 1: 57–81
- 50 Gilmer J, Schoenholz SS, Riley PF, Vinyals O, Dahl GE. arXiv: 1704.01212, 2017
- 51 Fey M, Lenssen JE. arXiv: 1903.02428, 2019
- 52 Velickovic P, Cucurull G, Casanova A, Romero A, Lio P, Bengio Y. arXiv: 1710.10903, 2017
- 53 Li Y, Vinyals O, Dyer C, Pascanu R, Battaglia P. arXiv: 1803.03324, 2018
- 54 Li J, Cai D, He X. arXiv: 1709.03741, 2017
- 55 Coley CW, Jin W, Rogers L, Jamison TF, Jaakkola TS, Green WH, Barzilay R, Jensen KF. *Chem Sci*, 2019, 10: 370–377
- 56 Bogojeski M, Vogt-Maranto L, Tuckerman ME, Müller KR, Burke K. *Nat Commun*, 2020, 11: 5223
- 57 Liu K, Sun X, Jia L, Ma J, Xing H, Wu J, Gao H, Sun Y, Boulnois F, Fan J. *Int J Mol Sci*, 2019, 20: 3389
- 58 Jiménez-Luna J, Grisoni F, Schneider G. *Nat Mach Intell*, 2020, 2: 573–584
- 59 Schütt K, Unke O, Gastegger M. Equivariant message passing for the prediction of tensorial properties and molecular spectra. In: *Proceedings of the 38th International Conference on Machine Learning*. Vienna, 2021. 9377–9388
- 60 Schütt KT, Sauceda HE, Kindermans PJ, Tkatchenko A, Müller KR. *J Chem Phys*, 2018, 148: 241722
- 61 Liu Y, Wang L, Liu M, Zhang X, Oztekin B, Ji S. arXiv: 2102.05013, 2021
- 62 Yang K, Swanson K, Jin W, Coley C, Eiden P, Gao H, Guzman-Perez A, Hopper T, Kelley B, Mathea M, Palmer A, Settels V, Jaakkola T, Jensen K, Barzilay R. *J Chem Inf Model*, 2019, 59: 3370–3388
- 63 Schnake T, Eberle O, Lederer J, Nakajima S, Schütt KT, Müller KR, Montavon G. arXiv: 2006.03589, 2020
- 64 Simonovsky M, Komodakis N. arXiv: 1704.02901, 2018
- 65 De Cao N, Kipf T. arXiv: 1805.11973, 2018
- 66 Wang H, Wang J, Wang J, Zhao M, Zhang W, Zhang F, Xie X, Guo M. *IEEE T Knowl Data En*, 2019, 33: 3090–3103
- 67 Jin W, Coley C, Barzilay R, Jaakkola T. arXiv: 1709.04555, 2017
- 68 Beker W, Gajewska EP, Badowski T, Grzybowski BA. *Angew Chem Int Ed*, 2019, 58: 4515–4519
- 69 Bjerrum EJ. arXiv: 1703.07076, 2017
- 70 Schwaller P, Laino T, Gaudin T, Bolgar P, Hunter CA, Bekas C, Lee AA. *ACS Cent Sci*, 2019, 5: 1572–1583
- 71 Schwaller P, Hoover B, Reymond JL, Strobelt H, Laino T. *Sci Adv*, 2021, 7: eabe4166
- 72 Zheng S, Rao J, Zhang Z, Xu J, Yang Y. *J Chem Inf Model*, 2019, 60: 47–55
- 73 Krenn M, Häse F, Nigam AK, Friederich P, Aspuru-Guzik A. *Mach Learn-Sci Technol*, 2020, 1: 045024
- 74 Lin K, Xu Y, Pei J, Lai L. *Chem Sci*, 2020, 11: 3355–3364
- 75 Lin K, Xu Y, Pei J, Lai L. arXiv: 1906.02308, 2019
- 76 Shi C, Xu M, Guo H, Zhang M, Tang J. A graph to graphs framework for retrosynthesis prediction. In: *International Conference on Machine Learning*. Vienna, 2020. 8818–8827
- 77 Dai H, Li C, Coley C, Dai B, Song L. arXiv: 2001.01408, 2019
- 78 Liu B, Ramsundar B, Kawthekar P, Shi J, Gomes J, Luu Nguyen Q, Ho S, Sloane J, Wender P, Pande V. *ACS Cent Sci*, 2017, 3: 1103–1113
- 79 Segler MHS, Preuss M, Waller MP. *Nature*, 2018, 555: 604–610
- 80 Chen B, Li C, Dai H, Song L. Retro*: learning retrosynthetic planning with neural guided A* search. In: *International Conference on Machine Learning*. Vienna, 2020. 1608–1616
- 81 Rogers D, Hahn M. *J Chem Inf Model*, 2010, 50: 742–754
- 82 Morgan HL. *J Chem Doc*, 1965, 5: 107–113

Molecular representation in AI-assisted synthesis chemistry: from sequence to graph

Hongshuai Wang¹, Lujie Jin¹, Jie Feng¹, Zhihao Dong¹, Weiqiao Deng^{2*}, Youyong Li^{1*}

¹ Institute of Functional Nano and Soft Materials, Soochow University, Suzhou 215123, China

² Institute of Frontier and Interdisciplinary Science, Qingdao Research Institute of Shandong University, Qingdao 266237, China

*Corresponding authors (email: dengwq@sdu.edu.cn; yyli@suda.edu.cn)

Abstract: With the rapid development of deep neural networks and their derived algorithms, artificial intelligence (AI) has flourished in various fields in recent years. In the field of computer-aided synthetic route design, artificial intelligence algorithms such as generative adversarial networks and variational autoencoders have performed very well. How to deal with the geometric representation of molecules is a key step in transforming chemical problems into algorithmic problems. Starting from different dimensions, this paper systematically summarizes the application of various molecular sequence models, graph theory and other representation methods in AI-assisted synthesis chemistry, and discusses the shortcomings and challenges of current molecular representation.

Keywords: artificial intelligence, molecular representation, computer-aided synthesis, machine learning

doi: [10.1360/SSC-2022-0113](https://doi.org/10.1360/SSC-2022-0113)