

化学反应感知分子 表征学习

Hongwei Wang¹, Weijiang Li¹, Xiaomeng Jin¹, Kyunghyun Cho^{2,3}, Heng Ji¹, Jiawei Han¹, Martin D. Burke¹

¹伊利诺伊大学香槟分校, ² 纽约大学, ³ 基因泰克公司

{hongweiw, wl13, xjin17, hengji, hanj, mdburke}@illinois.edu, kyunghyun.cho@nyu.edu

摘要


分子表征学习 (MRL) 方法旨在将分子嵌入真实的向量空间。然而, 现有的基于 SMILES (简化分子输入线-输入系统) 或 GNN (图神经网络) 的分子表征学习方法要么以 SMILES 字符串为输入, 难以编码分子结构信息; 要么过分强调 GNN 架构的重要性, 却忽略了其泛化能力。在此, 我们建议使用化学反应来辅助学习分子表征。我们方法的关键思路是在嵌入空间中保持分子与化学反应的等价性, 即强制每个化学方程式的反应物嵌入总和与生成物嵌入总和相等。事实证明, 这种约束可以有效地: 1) 保持嵌入空间的有序性; 2) 提高分子嵌入的泛化能力。此外, 我们的模型可以使用任何 GNN 作为分子编码器, 因此与 GNN 架构无关。实验结果表明, 我们的方法在各种下游任务中取得了最先进的性能, 例如, 与最佳基线方法相比, 化学反应预测的 Hit@1 绝对增益为 17.4%, 分子性质预测的 AUC 绝对增益为 2.3%, 图编辑距离预测的 RMSE 相对增益为 18.5%。代码见 <https://github.com/hwwang55/MolR>。

1 引言

如何表示分子是化学中一个基本而关键的问题。化学家通常使用 IUPAC 术语、分子式、结构式、骨架式等来表示化学文献中的分子。¹然而, 这些表示法最初是为人类读者而不是计算机设计的。为了便于机器学习算法理解和利用分子, 有人提出了分子表征学习 (MRL), 将分子映射到低维实数空间, 并以密集向量的形式表示。学习到的分子向量 (又称嵌入) 可以使一系列下游任务受益, 如化学反应预测 (Jin 等人, 2017 年; Schwaller 等人, 2018 年)、分子性质预测 (Zhang 等人, 2021 年)、分子生成 (Marker 等人, 2021 年)、分子结构预测 (Marker 等人, 2021 年)、分子结构预测 (Marker 等人, 2021 年)、2021 年)、分子生成 (Mahmood 等人, 2021 年)、药物发现 (Rathi 等人, 2019 年)、逆合成规划 (Segler 等人, 2018 年)、化学文本挖掘 (Krallinger 等人, 2017 年) 和化学知识图建模 (Bean 等人, 2017 年)。

研究人员提出了许多 MRL 方法。其中很大一部分，包括 Mol- BERT (Fabian 等人, 2020 年)、ChemBERTa (Chithrananda 等人, 2020 年)、SMILES-Transformer (Honda 等人, 2019 年)、SMILES-BERT (Wang 等人, 2019 年)、Molecule-Transformer (Shin 等人, 2019 年) 和 SA- BiLSTM (Zheng 等人, 2019 年b), 以 *SMILES*² 字符串作为输入, 并利用自然语言模型 (例如 Transformers (Vaswani 等人, 2017 年) 或 BERT (Devlin 等人, 2018 年)) 作为基础模型。

¹例如, 甘油的 IUPAC 名称、分子式、结构式和骨架式

是 丙烷-1,2,3-三醇, $C_3H_8O_3$, $CH_2 - \overset{\overset{OH}{|}}{\underset{\underset{|}{|}}{CH}} - \overset{\overset{OH}{|}}{\underset{\underset{OH}{|}}{CH_2}}$, 以及  分别为 OH

²简化分子输入行式输入系统 (SMILES) 是一种行式输入规范, 用于使用 ASCII 短字符串描述化学物质的结构。例如, 甘油的 SMILES 字符串为 "OCC(O)CO"。

尽管这类语言模型功能强大，但它们在处理 SMILES 数据时却遇到了困难，因为 SMILES 是分子结构的一维线性化，这使得语言模型很难仅仅根据 "纤细" 的字符串来学习分子的原始结构信息（更多讨论见第 4 节）。另一种 MRL 方法则使用图神经网络 (GNN) (Kipf & Welling, 2017 年) 来处理分子图 (Jin 等人, 2017 年; Gilmer 等人, 2017 年; Ishida 等人, 2021 年)。虽然基于 GNN 的方法在学习分子结构方面理论上优于基于 SMILES 的方法，但它们仅限于设计新鲜精致的 GNN 架构，而忽略了 MRL 的本质，即泛化能力。事实上，我们稍后将证明，没有一种特定的 GNN 能够在 MRL 的所有下游任务中普遍表现最佳，这也启发我们探索 GNN 架构之外的其他方法。

针对现有工作的局限性，我们在本文中提出利用 *化学反应* 来帮助学习分子表征并提高其泛化能力。化学反应通常以化学方程式的形式用符号和公式表示，其中反应物实体在左侧，生成物实体在右侧。例如，乙酸和乙醇的费歇尔酯化反应的化学方程式可以写成

$$\text{CH}_3\text{COOH} + \text{C}_2\text{H}_5\text{OH} \rightarrow \text{CH}_3\text{COOC}_2\text{H}_5 + \text{H}_2\text{O}.$$

³ 化学反应通常表示其反应物和生成物之间的特定等价关系（例如，在守恒方面

我们的想法是在分子嵌入空间中保持这种等价性。具体来说，鉴于上述费歇尔酯化化学反应，我们希望等式 $h_{\text{CH}_3\text{COOH}} + h_{\text{C}_2\text{H}_5\text{OH}} = h_{\text{CH}_3\text{COOC}_2\text{H}_5} + h_{\text{H}_2\text{O}}$ 也成立，其中 $h_{(\cdot)}$ 代表分子嵌入函数。这个简单的约束条件赋予了分子内嵌非常好的特性：（1）分子内嵌在化学反应方面是可组合的，这使得内嵌空间变得井井有条（见命题 1）；（2）更重要的是，我们将在后面证明，当分子编码器是一个以求和作为读出函数的 GNN 时，我们的模型可以自动隐式地学习 *反应模板*，这些模板可以概括同一类别中的一组化学反应（见命题 2）。学习反应模板的能力是提高分子表征泛化能力的关键，因为模型可以很容易地将所学知识泛化到未见过但与已知分子属于同一类别或具有相似结构的分子上。

我们的研究表明，由我们提出的模型（即 **MolR**（化学反应感知分子嵌入））学习到的分子嵌入能够使各种下游任务受益，这使它与所有只为一种下游任务设计的现有方法有了显著区别。例如，与最佳基线方法相比，MolR 在化学反应预测中实现了 17.4% 的 Hit@1 绝对增益，在分子性质预测中实现了 2.3% 的 BBBP 数据集 AUC 绝对增益，在图编辑距离预测中实现了 18.5% 的相对 RMSE 增益。我们还对学习到的分子嵌入进行了可视化，并表明它们能够编码反应模板以及几个关键的分子属性，例如分子大小和最小环的数量。

2 建议的方法

2.1 结构分子编码器

分子图表示为 $G = (V, E)$ ，其中 $V = \{a_1, \dots\}$ 为非氢原子集合， $E = \{b_1, \dots\}$ 为键集合。每个原子 a_i 都有一个初始特征向量 x_i ，对其属性进行编码。在本研究中，我们使用了四种原子属性：*元素类型*、*电荷*、*是否为氢原子*。

原子是芳香环，以及 *相连氢原子的数量*。每种类型的原子属性都用一个单击向量来表示，我们还为每个单击向量添加了一个 "未知" 条目，以便在推理过程中处理未知值。四个单击向量连接起来作为初始原子特征。此外，每个键 b_i 都有一个键类型（如单键、双键）。由于键类型通常可以通过其两个相关原子的特征推断出来，而且根据我们的实验，键类型并不能持续提高模型的性能，因此我们没有明确将键类型作为输入。

³有机反应方程式通常还包含反应条件（温度、压力、催化剂、溶剂等）。例如，上述费歇尔酯化反应的完整方程式为 $\text{CH}_3\text{COOH} + \text{C}_2\text{H}_5\text{OH} \xrightarrow{\Delta} \text{CH}_3\text{COOC}_2\text{H}_5 + \text{H}_2\text{O}$ 。与之前的工作（[Jin 等人, 2017 年](#)；[Schwaller 等人, 2018 年](#)）类似，我们没有考虑化学反应的环境条件，因为我们的重点是描述分子及其关系。我们将条件建模作为今后的工作。

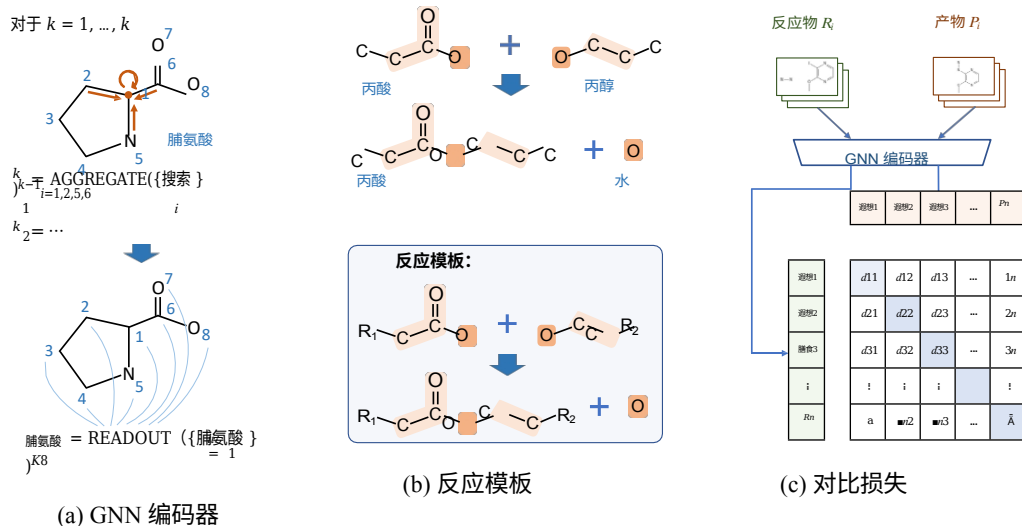


图 1: (a) GNN 编码器处理脯氨酸分子的示意图。氢被省略。(b) 丙酸和丙醇的费舍酯化反应图示，以及我们的模型学习到的相应反应模板。反应中心用橙色表示，与反应中心距离为 1 或 2 的原子用浅橙色表示。 (d_{ij}) 是嵌入 h_{R_i} 和 h_{P_j} 之间的欧氏距离。

为了学习分子的结构表征，我们选择了 GNN 作为基础模型。GNN 利用分子结构和原子特征来学习每个原子和整个分子的表征向量。典型的 GNN 遵循邻域聚合策略，通过聚合原子邻域和原子自身的表示来迭代更新原子的表示。形式上，GNN 的第 k 层为

$$h_i^k = \text{AGGREGATE} \left\{ h_j^{k-1} \mid j \in \text{N}(i) \cup \{i\} \right\}, \quad k = 1, \dots, K, \quad (1)$$

其中， h_i^k 是原子 i 在第 k 层的表示向量 (h^0 初始化为 i 的初始特征)。 x_i), $\text{N}(i)$ 是直接连接到 i 的原子集合， K 是 GNN 层数。AGGREGATE 函数的选择对 GNN 的设计至关重要，许多 GNN 架构都采用了 AGGREGATE 函数。

已经有人提出了 GNN 架构。有关 GNN 架构的详细介绍，请参见附录 A。

最后，使用读出函数汇总最后一个 GNN 层输出的所有节点表示，得到整个分子的表示 h ：

G

$$h_G = \text{READOUT} \{ h_a \}^{K_{a \in V}}. \quad (2)$$

READOUT 函数可以是简单的排列不变函数，如求和与均值，也可以是更复杂的图级池化算法 (Ying 等, 2018; Zhang 等, 2018)。图 1a 是 GNN 编码器的一个示例。

2.2 保持化学反应等效性

化学反应定义了反应物集 $R = \{r_1, r_2, \dots\}$ 和生成物集 $P = \{p_1, p_2, \dots\}$ 之间的特定关系“ \rightarrow ”:

$$r_1 + r_2 + \dots \rightarrow p_1 + p_2 + \dots. \quad (3)$$

化学反应通常代表一个封闭系统，反应前后系统的几个物理量保持不变，如质量、能量、电荷等。因此，它描述了化学反应空间中反应物和生成物之间的某种等价关系。我们的主要想法是在分子嵌入空间中保留这种等价性：

$$\sum_{r \in R} h_r = h \circ \sum_{p \in P} p \quad (4)$$

上述简单约束对于提高分子嵌入的质量至关重要。我们首先通过下面的命题证明，在公式(4)的约束下，化学反应关系" \rightarrow "是等价关系：

命题 1 假设 M 是分子集合, $R \subseteq M$ 和 $P \subseteq M$ 分别是化学反应的反应物集合和生成物集合。如果 $R \rightarrow P \Leftrightarrow \sum_{r \in R} h_r = \sum_{p \in P} h_p$

反应, 那么 " \rightarrow " 是 2^M 上的等价关系, 它满足以下三个性质:

(1) 反射性: $A \rightarrow A$, 对于所有 $A \in 2^M$; (2) 对称性: $A \rightarrow B \Leftrightarrow B \rightarrow A$, 对于所有 $A, B \in 2^M$; (3) Transitivity: 若 $A \rightarrow B$, 且 $B \rightarrow C$, 则 $A \rightarrow C$, 对于所有 $A, B, C \in 2^M$ 。

命题 1 的证明见附录 B。命题 1 的一个重要推论是, 根据等价关系 " \rightarrow ", M 的所有子集 (即 2^M) 被自然地划分为等价类。对于一个等价类中的所有分子集, 所有它们所组成的分子应该相等。例如, 在有机合成中, 目标化合物 t

则集合 A, B, C 以及 $\{t\}$ 属于一个等价类, 我们有 $\sum_{m \in A} h_m = \sum_{m \in B} h_m = \sum_{m \in C} h_m = h_t$ 。注意, 起始材料通常是小分子和基本分子, 经常出现在

的合成路线。因此, 式 (4) 构成了一个线性方程组, 其中化学反应等价对碱基分子的嵌入有很强的约束。因此, 分子嵌入的可行解将更加稳健, 整个嵌入空间也将更加有序。详见第 3.4 节分子嵌入空间的可视化结果。

我们还可以进一步证明, 公式 (4) 中的约束条件还能提高分子嵌入的泛化能力。为了说明这一点, 我们首先定义化学反应的反应中心。 $R \rightarrow P$ 的反应中心定义为反应物 R 的诱导子图, 其中每个原子都有至少有一个键的类型从 R 到 P 不同 ("无键" 也被视为一种键类型)。在其他换句话说, 反应中心是将反应物转化为生成物所需的最小图形编辑集。鉴于反应中心的概念, 我们有如下命题:

命题 2 假设 $R \rightarrow P$ 是一个化学反应, 其中 R 是反应物集, P 是生成物集, C 是反应中心。假设我们使用 GNN (其层数为 K)

式 (1) 和 (2) 中所示的分子编码器, 并将式 (2) 中的 READOUT 函数设为求和函数。然后, 对于其中一种反应物中的任意原子 a , 其最终表示为 h_a^K , 残差项 $h_a^K - \sum_{p \in P} h_p^K$ 是 h 的函数 a 当且仅当原子 a 和反应中心 C 小于 K 。

命题 2 的证明见附录 C。命题 2 表明, 反应物嵌入和生成物嵌入之间的残差将完全且仅取决于距离反应中心小于 K 跳的原子。例如, 如图 1b 所示, 假设我们使用 3 层 GNN 来处理丙酸和丙醇的 Fischer 酯化反应, 那么反应物嵌入和产物嵌入之间的残差将完全取决于反应中心 (橙色) 以及与反应中心距离为 1 或 2 的原子 (浅橙色)。这意味着, 如果 GNN 编码器对该化学方程式进行了优化并输出完美的嵌入, 即 $h_{CH_3CH_2COOH} + h_{CH_3CH_2CH_2OH} = h_{CH_3CH_2COOCH_2CH_2CH_3} + h_{H_2O}$

那么对于任何官能团 R_1 和 R_2 , 方程式 $h_{R_1-CH_2COOH} + h_{R_2-CH_2CH_2OH} = h_{R_1-CH_2COOCH_2CH_2-R_2} + h_{H_2O}$ 也将成立, 因为方程式两边的残差并不取决于距离反应中心超过 2 跳的 R_1 或 R_2 。诱导的一般化学反应 $R-CH_2COOH + R-CH_2CH_2OH \rightarrow R-CH_2COOCH_2CH_2-R + H_2O$ 是

反应模板被称为反应模板, 它抽象了同一类别中的一组化学反应。

学习到的反应模板对提高模型的泛化能力至关重要, 因为模型可以很容易地将这些知识应用于训练数据中未见但符合已知反应模板的反应 (如乙酸加丙醇、丁酸加丁醇)。我们将在第 3.4 节中进一步说明如何在分子嵌入中对反应模板进行编码。

备注。 下面是我们的一些评论, 以便进一步理解所提出的模型:

首先，与同样学习反应模板的（Jin 等, 2017）相比，我们的模型不需要复杂的网络来计算注意力分数，也不需要额外的反应物和生成物之间的原子映射信息作为输入。此外，从理论上讲，我们的模型甚至可以只根据一个反应实例来学习反应模板，这使得它在少次学习（Wang 等人, 2020 年）的情况下特别有用。实验结果见第 3.1 节。

其次，有机反应通常是不平衡的，会省略一些小分子和无机分子，从而使所关注的产物发光（例如，费歇尔酯化反应通常会省略 H_2O ）。尽管如此、

只要化学反应的书写方式保持一致（例如，所有费歇尔酯化反应都省略 H_2O ），我们的模型仍然可以学习有意义的反应模板。

第三，根据命题 2，GNN 层数 K 会对学习到的反应模板产生很大影响： K 越小可能不足以代表一个有意义的反应模板（例如，在费歇尔酯化反应中，如果 $K < 3$ ，羧酸中必要的羰基 " $\text{C}=\text{O}$ " 就不会出现在反应模板中、例如，在费歇尔酯化反应中，如果 $K < 3$ ，羧酸中必要的羰基 " $\text{C}=\text{O}$ " 将不会出现在反应模板中），而 K 过大则可能会为反应模板包含不必要的原子，从而降低其覆盖范围（例如，图 1b 所示的反应模板就没有覆盖甲酸 HCOOH 和甲醇 CH_3OH 的费歇尔酯化反应）。 K 的经验影响见附录 D。

最后，我们的模型不会明确输出所学反应模板，而是将这些信息隐含在模型参数和分子嵌入中。挖掘显式反应模板的一种简单方法是剔除每个化学反应中反应物和生成物的共同部分，然后应用聚类算法。对这一问题的进一步研究将作为今后的工作。

2.3 训练模型

根据公式 (4)，拟议方法的直接损失函数是 $L = \frac{1}{|\mathcal{O}|} \sum_{(R \rightarrow P) \in \mathcal{O}} \sum_{r \in R} h_r - \sum_{p \in P} h_p$ ，其中 $R \rightarrow P$ 表示化学反应。

然而，仅仅最小化上述损失是行不通的，因为模型会退化，对所有分子输出全为零的嵌入。解决这一问题的常见方法

这些方法包括引入负抽样策略（Goldberg & Levy, 2014 年）或对比学习（Jaiswal 等人, 2021 年）。在这里，我们使用了一个与（Radford 等人, 2021 年）类似的基于小批量的对比学习框架，因为它更节省时间和记忆。

对于一个小批数据 $B = \{R_1 \rightarrow P_1, R_2 \rightarrow P_2, \dots\} \subseteq D$ ，我们首先使用 GNN 编码器处理该迷你批次中的所有反应物 R_i 和产物 P_i ，并得到它们的嵌入。匹配的

反应物-生成物对 (R_i, P_i) 被视为正对，其嵌入差异将最小化，而不匹配的反应物-生成物对 (R_i, P_j) ($i \neq j$) 被视为负对，其嵌入差异将最大化。为了避免总损失被以下因素支配与（Bordes 等人, 2013 年）类似，我们使用基于边际损失的方法（示例见图 1c）：

$$L_5 = \frac{1}{|B|} \sum_i \sum_{r \in R_i} h_r - \sum_{p \in P_i} h_p + \frac{1}{|B|(|B| - 1)} \sum_{\substack{r \\ i \neq j}} \max_{\substack{p \\ r \in R_i}} \gamma - h_p - \sum_{\substack{p \\ p \in P_j}} h_p, \quad (5)$$

其中， $\gamma > 0$ 是边际超参数。因此，可以使用基于梯度的优化方法，如随机梯度下降法（SGD），通过最小化上述损失来训练整个模型。

式 (5) 可以看作是一种特殊的负采样策略，它将小批量中所有未匹配的产物作为给定反应物的负采样。不过要注意的是，与传统的负采样（Mikolov 等人, 2013 年）相比，它有两个优点：（1）无需额外内存来存储负采样；（2）由于训练实例会被洗牌，负采样会在新历时开始时自动更新，从而节省了手动重新采样负采样的时间。

3 实验

3.1 化学反应预测

数据集。我们使用Lowe (2012) 从美国专利商标局授权专利中收集的反应作为数据集，并由Zheng等人 (2019a) 对数据集进行了进一步清理。该数据集包含 478,612 个化学反应，分为训练集、验证集和测试集，分别有 408,673 个、29,973 个和 39,966 个反应，因此我们将该数据集称为 **USPTO-479k**。USPTO-479k 中的每个反应实例都包含最多五个反应物的 SMILES 字符串和恰好一个生成物。

评估协议。我们将化学反应预测任务表述为一个排序问题。在推理阶段，给定化学反应的反应物集 R ，我们将测试集中的所有产物视为候选池 C （其中包含 39 459 个唯一候选产物），并基于以下原则对所有候选产物进行排序

反应物嵌入式 h_R 与候选产品嵌入式 $\{h\}_{cc \in C}$ 之间的 L2 距离，即 $d(R, c) = \|h_R - h_c\|_2$ 。然后，可以利用地面实况产品的排序来计算

衡量标准	MRR	MR	命中@1	命中@3	命中@5	命中@10
Mol2vec	0.681	483.7	0.614	0.725	0.759	0.798
Mol2vec-FT1	0.688 ± 0.000	417.6 ± 0.1	0.620 ± 0.000	0.734 ± 0.000	0.767 ± 0.000	0.806 ± 0.000
莫尔贝特	0.708	460.7	0.623	0.768	0.811	0.858
MolBERT-FT1	0.731 ± 0.000	457.9 ± 0.0	0.649 ± 0.000	0.790 ± 0.000	0.831 ± 0.000	0.873 ± 0.000
MolBERT-FT2	0.776 ± 0.000	459.6 ± 0.2	0.708 ± 0.000	0.827 ± 0.000	0.859 ± 0.000	0.891 ± 0.000
MolR-GCN	0.905 ± 0.001	34.5 ± 2.4	0.867 ± 0.001	0.938 ± 0.001	0.950 ± 0.001	0.961 ± 0.002
MolR-GAT	0.903 ± 0.002	35.3 ± 2.8	0.864 ± 0.002	0.935 ± 0.003	0.948 ± 0.003	0.961 ± 0.003
MolR-SAGE	0.903 ± 0.004	53.0 ± 4.6	0.865 ± 0.005	0.935 ± 0.004	0.948 ± 0.004	0.961 ± 0.002
MolR-TAG	0.918 ± 0.000	27.4 ± 0.4	0.882 ± 0.000	0.949 ± 0.001	0.960 ± 0.001	0.970 ± 0.000
MolR-TAG (1%训练数据)	0.904 ± 0.002	33.0 ± 3.7	0.865 ± 0.003	0.937 ± 0.003	0.951 ± 0.002	0.963 ± 0.002

表 1: USPTO-479k 数据集的化学反应预测结果。最佳结果以粗体表示，基线的最佳结果以下划线表示。

不	反应物	事实真相产品	预测产品 作者: MolR-GCN	预测产品 作者: Mol2vec	预测产品 作者: MolBERT
6			与地面实况相同		
17			与地面实况相同		

表 2: 美国专利商标局-479k 数据集案例研究。完整版见附录 E。

MRR (平均倒数秩)、**MR** (平均秩) 和 **Hit@1, 3, 5, 10** (命中率, 截止值分别为 1、3、5 和 10)。每个实验重复 3 次, 当验证集上的 MRR 最大时, 我们报告测试集上的平均值和标准偏差结果。

基线。我们使用 Mol2vec (Jaeger 等人, 2018 年) 和 MolBERT (Fabian 等人, 2020 年) 作为基线。对于每条基线, 我们使用已发布的预训练模型来输出反应物 h_R 和候选产物 $\{h_c\}_{c \in C}$ 的嵌入, 然后根据它们的点积对所有候选产物进行排序: $d(R, c) = -h_R^T h_c$ 。由于预训练模型没有在 USPTO-479k 上进行微调, 我们提出了两种微调策略:

Mol2vec-FT1 和 **MolBERT-FT1**, 它们冻结模型参数, 但训练一个对角矩阵 K 来对候选者进行排序: $d(R, c) = -h_R^T K h_c$; **MolBERT-FT2**, 它通过最小化 USPTO-479k 上公式 (5) 所示的对比损失函数来微调模型参数。请注意, Mol2vec 不是端到端模型, 因此无法使用这种策略进行微调。

超参数设置。我们使用以下四种 GNN 实现分子编码器: **GCN** (Kipf & Welling, 2017 年)、**GAT** (Velić'ković et al., 2018 年)、**SAGE** (Hamilton et al., 2017 年) 和 **TAG** (Du et al., 2017 年), 详细介绍见附录 A。所有 GNN 的层数都是 2, 所有层的输出维度都是 1,024。边距 γ 设置为

4. 我们使用学习率为 10^{-4} 的 Adam (Kingma 和 Ba, 2015 年) 优化器对模型进行了 20 个历元的训练, 批量大小为 4 096。超参数敏感性结果见附录 D。

USPTO-479k 数据集的结果。化学反应预测结果见表 1。很明显, 我们的方法 MolR 的所有

四个变体都明显优于基线方法。例如，与最佳基准方法 MolBERT-FT2 相比，MolR-TAG 实现了 14.2% 的 MRR 绝对增益和 17.4% 的 Hit@1 绝对增益。此外，我们还只在训练集中的前 4,096 个（即 1%）反应实例上训练 MolR-TAG（学习率为 10^{-3} ，60 个历时，其他超参数保持不变），其性能也略有下降。这证明了我们在第 2.2 节备注 1 中的说法，即 MolR 在少量学习情况下表现相当出色。

案例研究。我们选择 USPTO-479k 测试集中的前 20 个反应进行案例研究。表 2 显示了两个反应的结果，完整版本见附录 E。结果表明，Mol2vec 和 MolBERT 的输出结果已经与地面实况非常相似，但我们的模型

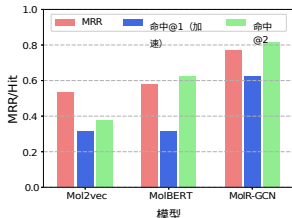


图 2: 回答有关产品预测的真实多选题的结果。

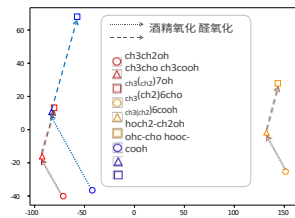


图 3: 酒精氧化和醛氧化的可视化反应。

功能模式	并	减去
Mol2vec	1.140 ± 0.041	0.995 ± 0.034
莫尔贝特	1.127 ± 0.042	0.937 ± 0.029
MolR-GCN	0.976 ± 0.026	0.922 ± 0.019
MolR-GAT	1.007 ± 0.021	0.943 ± 0.016
MolR-SAGE	0.918 ± 0.028	0.817 ± 0.013
MolR-TAG	0.990 ± 0.029	0.911 ± 0.027

表 3: QM9 数据集上 GED 预测的 RMSE 结果。最佳结果以粗体标出。

数据集	BBBP	艾滋病病毒	计算机设备行动伙伴关系	Tox21	临床毒理学
笑脸-变形金刚	0.704	0.729	0.701	0.802	0.954
ECFP4	0.729	0.792	0.867	0.822	0.799
GraphConv	0.690	0.763	0.783	0.829	0.807
编织	0.671	0.703	0.806	0.820	0.832
化学ERTa	0.643	0.622	-	0.728	0.733
D-MPNN	0.708	0.752	-	0.688	0.906
CDDD	0.761 ± 0.00	0.753 ± 0.00	0.833 ± 0.00	-	-
莫尔贝特	0.762 ± 0.00	0.783 ± 0.00	0.866 ± 0.00	-	-
Mol2vec	0.872 ± 0.021	0.769 ± 0.021	0.862 ± 0.027	0.803 ± 0.041	0.841 ± 0.062
MolR-GCN	0.890 ± 0.032	0.802 ± 0.024	0.882 ± 0.019	0.818 ± 0.023	0.916 ± 0.039
MolR-GAT	0.887 ± 0.026	0.794 ± 0.022	0.863 ± 0.026	0.839 ± 0.039	0.908 ± 0.039
MolR-SAGE	0.879 ± 0.032	0.793 ± 0.026	0.859 ± 0.029	0.811 ± 0.039	0.890 ± 0.058
MolR-TAG	0.895 ± 0.031	0.801 ± 0.023	0.875 ± 0.023	0.820 ± 0.028	0.913 ± 0.043

表 4: 分子性质预测的 AUC 结果。前三块的结果分别来自 (Honda et al., 2019)、(Chithrananda et al., 2020) 和 (Fabian et al., 2020)，后两块的结果由我们报告。最佳结果以粗体标出，如果 MolR 是基线的最佳结果，则以下划线标出。

在预测精确答案方面，MolR-GCN 更为强大。具体来说，在 6 号反应中，MolR-GCN 成功预测到三角环破裂，而 Mol2vec 和 MolBERT 却未能预测到。

产品预测真实多选题的结果。为了测试 MolR 在真实场景中的表现，我们从牛津大学出版社、mhpracticplus.com 和 GRE 化学考试练习册的在线资源中收集了 16 道有关产物预测的多选题。每道题都给出了一个化学反应的反应物，并要求从 4 或 5 个选项中选出正确的产物。这些多选题即使对化学家来说也相当困难，因为反应物之间通常非常相似（详见附录 F）。结果如图 2 所示，表明 MolR 远远超过了基线。具体来说，MolR-GCN 的 Hit@1（即准确度）为 0.625，是 Mol2vec 和 MolBERT 的两倍。

3.2 分子特性预测

数据集我们在五个数据集上对 MolR 进行了评估：BBBP、HIV、BACE、Tox21 和 ClinTox，由 Wu 等人 (2018 年) 提出。每个数据集都包含数以千计的分子 SMILES 以及表示相关属性的二进制标签（例如，Tox21 衡量化合物的毒性）。有关这些数据集的详细信息，请参见 (Wu 等人, 2018 年)。

基线。我们将我们的方法与以下基线进行了比较：SMILES-Transformers (Honda et al.,

2019)、ECFP4 (Rogers & Hahn, 2010)、GraphConv (Duvenaud et al., 2015)、Weave (Kearnes et al., 2016)、ChemBERTa (Chithrananda et al., 2020)、D-MPNN (Yang et al., 2019)、CDDD (Winter et al., 2019)、MolBERT (Fabian et al., 2020) 和 Mol2vec (Jaeger et al., 2018)。大多数基线结果来自文献，而 Mol2vec 则由我们运行。

实验设置。所有数据集按 8:1:1 分成训练集、验证集和测试集。我们使用在 USPTO-479k 上预先训练好的模型来处理所有数据集并输出分子嵌入，然后将嵌入和标签输入用 scikit-learn 实现的逻辑回归模型（Pedregosa 等人，2011 年），该模型的超参数除求解器="liblinear"外均为默认值。每个实验重复 20 次，我们报告测试集的平均值和标准偏差结果。

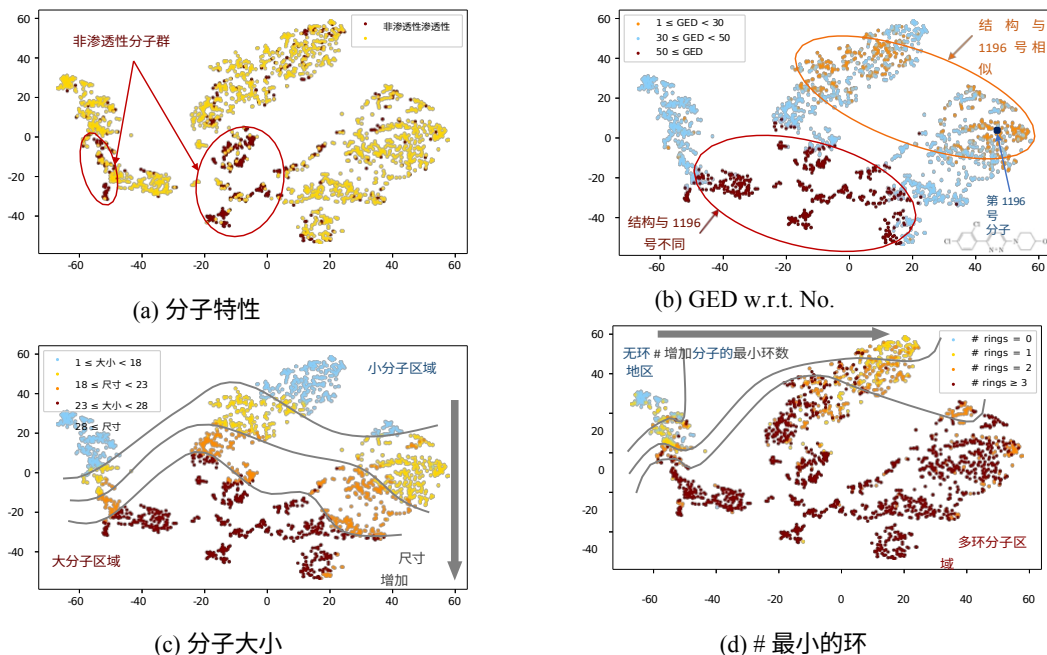


图 4: BBBP 数据集的可视化分子嵌入空间。

结果表 4 列出了分子性质预测的 AUC 结果。可以看出, MolR 在 5 个数据集集中的 4 个数据集上表现最佳。我们将 MolR 在分子性质预测方面的优异表现归因于 MolR 在 USPTO-479k 上进行了预训练, 因此根据命题 2, MolR 对反应中心非常敏感。请注意, 反应中心通常由化学活性官能团组成, 这对于确定分子性质至关重要。

3.3 图形-编辑-距离预测

图编辑距离 (GED) (Gao et al. 在此, 我们旨在根据两个分子图的嵌入预测它们之间的 GED。这项任务的目的是证明学习到的分子嵌入是否能够保持分子间的结构相似性。此外, 由于计算精确的 GED 是 NP 难的, 因此这项任务的解决方案也可以看作是 GED 计算的近似算法。

实验设置。我们从 QM9 数据集 (Wu 等人, 2018 年) 的前 1000 个分子中随机抽取 10,000 对分子, 然后使用 NetworkX (Hagberg 等人, 2008 年) 计算它们的地面实况 GED。数据集按 8:1:1 分成训练集、验证集和测试集, 我们使用预训练模型为每个分子输出嵌入。我们将一对分子的嵌入作为特征进行串联或相减, 并将地面实况 GEDs 设置为目标, 然后将它们一起输入到 scikit-learn 中的支持向量回归模型中, 并使用默认的超参数。每个实验重复 20 次, 并报告测试集的平均值和标准偏差结果。我们将我们的方法与 Mol2vec 和 MolBERT 进行了比较。

结果表 3 报告了 RMSE 结果。与最佳基线相比, 我们的最佳 MolR 模型在串联和减法模式下分别降低了 18.5% 和 12.8%。请注意, 地面实况 GED 的区间为 [1, 14], 其范围远远大于 MolR 的 RMSE。这一结果表明, MolR 可以作为计算 GED 的强近似算法。

3.4 嵌入可视化

为了直观地展示分子嵌入空间，我们使用预训练的 MolR-GCN 模型输出 BBBP 数据集中的分子嵌入，然后使用图 4 所示的 t-SNE ([Van der Maaten & Hinton, 2008 年](#)) 将其可视化。在图 4a 中，分子根据渗透性的属性着色。我们发现了两个非渗透性分子群落，这表明 MolR 可以捕捉到感兴趣的分子特性。在图 4b 中，分子根据其与 BBBP 数据集中随机选取的分子（编号 1196）的 GED 值着色。显而易见

结果表明，与 1196 号分子结构相似的分子（橙色）在嵌入空间中也与之相近，而与 1196 号分子结构不相似的分子（红色）在嵌入空间中也与之相远。结果表明，MolR 能够很好地捕捉分子间的结构相似性。在图 4c 中，分子根据其大小（即非氢原子的数量）着色。很明显，嵌入空间被完美地分割为小分子区域（上部）和大分子区域（下部）。换句话说，二维嵌入空间的纵轴是分子大小的特征。最后，令人惊讶的是，我们发现横轴实际上与分子中最小环（即不包含另一个环的环）的数量有关：如图 4d 所示，无环分子（蓝色）只分布在左侧聚类中，单环分子（黄色）只分布在左侧和中间聚类中，双环分子（橙色）基本上分布在中间聚类中，而右侧聚类主要由具有 2 个以上环的分子（红色）组成。

我们还以醇氧化和醛氧化为例说明了 MolR 对化学反应的编码，其化学反应模板分别是 $R-CH_2OH + O_2 \rightarrow R-CHO + H_2O$ 和 $R-CHO + O_2 \rightarrow R-COOH$ 。我们首先使用预训练的 MolR-GCN 模型输出乙醇 (CH_3CH_2OH)、1-辛醇 ($CH_3(CH_2)_7OH$)、乙二醇 ($(CH_2OH)_2$) 的嵌入、

以及相应的醛和羧酸，然后使用主成分分析法 (PCA) 将其可视化。结果如图 3 所示，它清楚地表明 $hCH_3CHO - hCH_3CH_2OH \approx hCH_3(CH_2)_6CHO - hCH_3(CH_2)_7OH$ 和 $hCH_3COOH - hCH_3CHO \approx hCH_3(CH_2)_6CHO - hCH_3(CH_2)_7OH$ （红色和橙色箭头）。请注意，蓝色 arrow 是相应的红色或橙色箭头的两倍，这正是因为 $(CH_2OH)_2 / (CH_2CHO)_2$ 有两个羟基/醛基团需要氧化。

4 相关工作

现有的 MRL 方法可分为两类。第一类是基于 SMILES 的方法 (Fabian 等人, 2020; Chithrananda 等人, 2020; Honda 等人, 2019; Wang 等人, 2019; Shin 等人, 2019; Zheng 等人, 2019b)，这些方法使用语言模型来处理 SMILES 字符串。例如，MolBERT (Fabian 等人, 2020) 使用 BERT 作为基础模型，并设计了三个自监督任务来学习分子表征。然而，SMILES 是分子结构的一维线性化，高度依赖于分子图的遍历顺序。这意味着在 SMILES 中距离很近的两个原子实际上可能相距很远，因而不相关（例如 "CC(CCCCCCO)O" 中的两个氧原子），这将误导那些严重依赖于标记的相对位置来提供自监督信号的语言模型。

相比之下，第二类是基于结构的方法，可进一步分为传统的基于指纹的方法 (Rogers & Hahn, 2010; Jaeger 等人, 2018) 和最新的基于 GNN 的方法 (Jin 等人, 2017; Gilmer 等人, 2017; Ishida 等人, 2021)。例如，Mol2vec (Jaeger 等人, 2018 年) 将分子子结构（指纹）视为单词，将分子视为句子，然后使用类似 Word2vec 的方法计算分子嵌入。然而，它们无法识别不同子结构的重要性，也无法以端到端的方式进行训练。基于 GNN 的方法克服了这些缺点。不过，由于其架构复杂，通常需要大量的训练数据，这可能会在数据稀少时限制其泛化能力。

值得注意的是，我们的模型在概念上也与 NLP 中的方法有关，在 NLP 中，embeddings 是可组合的 (Bordes 等人, 2013; Mikolov 等人, 2013)。例如，TransE (Bordes et al. Word2vec (Mikolov 等人, 2013 年) 学习单词嵌入，其中简单的向量相加就能产生例如， $vec(\text{"德国"}) + vec(\text{"首都"}) \approx vec(\text{"柏林"})$ 。

5 结论和未来工作

在这项工作中，我们使用 GNN 作为分子编码器，并通过强制反应物嵌入之和等于生成物嵌入之和，利用化学反应来辅助学习分子表征。我们证明，我们的模型能够学习对提高泛化能力至关重要的反应模板。我们的模型已被证明对广泛的下游任务有益，可视化结果表明，学习到的嵌入式具有良好的组织性和反应感知能力。

我们指出了未来工作的四个方向。首先，如第 1 节所述，环境条件也是我们将考虑建模的化学反应的一部分。其次，正如第 2.2 节所讨论的，如何明确输出学习到的反应模板值得研究。第三，由于我们的模型（以及基线模型）无法处理立体异构问题，因此研究如何在嵌入空间中区分立体异构体也很有意义。最后，加入附加信息（如分子的文字描述）来帮助学习分子表征也是一个很有前景的方向。

参考资料

Daniel M Bean, Honghan Wu, Ehtesham Iqbal, Olubanke Dzahini, Zina M Ibrahim, Matthew Broad-bent, Robert Stewart 和 Richard JB Dobson. 电子健康记录中未知药物不良反应的知识图谱预测与验证》。《科学报告》，7 (1) : 1-11, 2017.

Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston 和 Oksana Yakhnenko。多关系数据建模的翻译嵌入。《神经信息处理系统进展》，2013 年第 26 期。

Seyone Chithrananda, Gabriel Grand 和 Bharath Ramsundar。Chemberta : *ArXiv preprint arXiv:2010.09885*, 2020.

Jacob Devlin, Ming-Wei Chang, Kenton Lee 和 Kristina Toutanova. 伯特：用于语言理解的深度双向变换器预训练》，*arXiv preprint arXiv:1810.04805*, 2018.

Jian Du, Shanghang Zhang, Guanhua Wu, Jose´ MF Moura, and Soumya Kar. 拓扑自适应图卷积网络。 *arXiv preprint arXiv:1710.10370*, 2017.

大卫-杜维诺、道格尔-麦克劳林、豪尔赫-阿奎莱拉-伊帕拉吉雷、拉斐尔-戈麦斯-邦巴雷利、蒂姆-奥西-希尔泽尔、阿拉恩-阿斯普鲁-古齐克、瑞安-P-亚当斯。用于学习分子指纹的图卷积网络。《神经信息处理系统进展》，第 2224-2232 页，2015 年。

Benedek Fabian, Thomas Edlich, He´le´na Gaspar, Marwin Segler, Joshua Meyers, Marco Fiscato, and Mohamed Ahmed. 使用语言模型和领域相关辅助任务的分子表征学习。 *arXiv preprint arXiv:2011.13230*, 2020.

高鑫波、肖兵、陶大成、李学龙. 图编辑距离研究。《模式分析与应用》，13 (1) : 113-129, 2010.

Justin Gilmer, Samuel S Schoenholz, Patrick F Riley, Oriol Vinyals 和 George E Dahl. 量子化学的神经信息传递。第 34 届机器学习国际会议论文集》，第 1263-1272 页。PMLR, 2017.

尤阿夫-戈德堡和奥梅尔-列维. Word2vec 解释：推导米科洛夫等人的负采样词嵌入方法。 *arXiv 预印本 arXiv:1402.3722*, 2014.

Aric Hagberg, Pieter Swart, and Daniel S Chult. 使用 networkx 探索网络结构、动态和功能。 Technical report, Los Alamos National Lab.(LANL), Los Alamos, NM (United States), 2008.

William L Hamilton、Rex Ying 和 Jure Leskovec.大型图上的归纳表征学习。《第 31 届神经信息处理系统国际会议论文集》，第 1025-1035 页，2017 年。

Shion Honda、Shoi Shi 和 Hiroki R Ueda。微笑转换器：用于低数据药物发现的预训练分子指纹。 *arXiv preprint arXiv:1911.04738*, 2019.

Sho Ishida、Tomo Miyazaki、Yoshihiro Sugaya 和 Shinichiro Omachi。用于化学性质估计的多特征提取路径图神经网络。 *分子*, 26 (11) : 3125, 2021。

Sabrina Jaeger、Simone Fulle 和 Samo Turk。Mol2vec：具有化学直觉的无监督机器学习方法。 *化学信息与建模期刊*，58 (1) : 27-35, 2018.

Ashish Jaiswal、Ashwin Ramesh Babu、Mohammad Zaki Zadeh、Debapriya Banerjee 和 Fillia Makedon。对比性自我监督学习调查。《技术》，9（1）：2，2021。

Wengong Jin, Connor W Coley, Regina Barzilay, and Tommi Jaakkola.用 Weisfeiler-lehman 网络预测有机反应结果。《第31届神经信息处理系统国际会议论文集》，第2604-2613页，2017年。

Steven Kearnes、Kevin McCloskey、Marc Berndl、Vijay Pande 和 Patrick Riley。分子图卷积：超越指纹。《计算机辅助分子设计期刊》，30(8)：595-608, 2016。

Diederik P Kingma 和 Jimmy Ba.亚当：一种随机优化方法。《第三届学习表征国际会议》，2015年。

Thomas N Kipf 和 Max Welling.使用图卷积网络进行半监督分类。《第五届学习表征国际会议论文集》，2017年。

Martin Krallinger、Obdulia Rabal、Analia Lourenco、Julen Oyarzabal 和 Alfonso Valencia。化学信息检索和文本挖掘技术。《化学评论》，117（12）：7673-7761，2017。

丹尼尔-马克-洛从文献中提取化学结构和反应。剑桥大学博士论文，2012年。

Omar Mahmood, Elman Mansimov, Richard Bonneau, and Kyunghyun Cho.用于分子生成的屏蔽图模型。《自然通讯》，12（1）：1-12，2021。

Tomas Mikolov、Ilya Sutskever、Kai Chen、Greg S Corrado 和 Jeff Dean。单词和短语的分布式代表及其构成性。《神经信息处理系统进展》，第3111-3119页，2013年。

Fabian Pedregosa、Gaël Varoquaux、Alexandre Gramfort、Vincent Michel、Bertrand Thirion、Olivier Grisel、Mathieu Blondel、Peter Prettenhofer、Ron Weiss、Vincent Dubourg 等。Scikit-learn：《机器学习研究期刊》，12:2825-2830，2011年。

Alec Radford、Jong Wook Kim、Chris Hallacy、Aditya Ramesh、Gabriel Goh、Sandhini Agarwal、Girish Sastry、Amanda Askell、Pamela Mishkin、Jack Clark 等。从自然语言监督中学习可转移的视觉模型。《第38届国际机器学习大会论文集》，2021年。

Prakash Chandra Rathi、R Frederick Ludlow 和 Marcel L Verdonk。使用图卷积深度神经网络发现药物的实用高质量静电位面。《药物化学杂志》，63（16）：8778-8790，2019。

戴维-罗杰斯和马修-哈恩。扩展连接性指纹。《化学信息与建模期刊》，50（5）：742-754，2010年。

Philippe Schwaller、Theophile Gaudin、David Lanyi、Costas Bekas 和 Teodoro Laino。"翻译中的发现"：利用神经序列到序列模型预测复杂有机化学反应的结果。《化学科学》，9（28）：6091-6098，2018。

Marwin HS Segler、Mike Preuss 和 Mark P Waller。用深度神经网络和符号 AI 规划化学合成。*自然*，555 (7698)：604-610，2018。

Bonggun Shin、Sungsoo Park、Keunsoo Kang 和 Joyce C Ho。基于自我注意的分子代表预测药物与靶点相互作用。*医疗保健机器学习会议*，pp.230-248.PMLR，2019 年。

Laurens Van der Maaten 和 Geoffrey Hinton.使用 t-sne 实现数据可视化》。*机器学习研究期刊*，9 (11)，2008 年。

Ashish Vaswani、Noam Shazeer、Niki Parmar、Jakob Uszkoreit、Llion Jones、Aidan N Gomez、Łukasz Kaiser 和 Illia Polosukhin。注意力就是你所需要的一切。*神经信息处理系统进展*，第 5998-6008 页，2017 年。

佩塔尔-韦利奇·科维奇、吉列姆-库库鲁尔、阿兰特萨-卡萨诺瓦、阿德里亚娜-罗梅罗、皮特罗-利奥、尤苏亚-本吉奥图形注意力网络。《第六届学习表征国际会议论文集》，2018。

Sheng Wang, Yuzhi Guo, Yuhong Wang, Hongmao Sun, and Junzhou Huang. Smiles-bert: 用于分子性质预测的大规模无监督预训练。《第10届ACM生物信息学、计算生物学和健康信息学国际会议论文集》, pp.429-436, 2019.

Yaqing Wang, Quanming Yao, James T Kwok, and Lionel M Ni. 从少量实例中归纳: 少量学习调查。《ACM Computing Surveys (CSUR)》, 53(3):1-34, 2020.

罗宾-温特、弗洛里安-蒙塔纳里、弗兰克-诺伊和德约克-阿尔内-克莱弗特通过翻译等效化学表征学习连续和数据驱动的分子描述符。《化学科学》, 10 (6) : 1692-1701, 2019.

Zhenqin Wu, Bharath Ramsundar, Evan N Feinberg, Joseph Gomes, Caleb Geniesse, Aneesh S Pappu, Karl Leswing, and Vijay Pande. Moleculenet: 分子机器学习的基准。《化学科学》, 9 (2) : 513-530, 2018.

Kevin Yang, Kyle Swanson, Wengong Jin, Connor Coley, Hua Gao, Angel Guzman-Perez, Timothy Hopper, Brian P Kelley, Andrew Palmer, Volker Settels, et al. Are learned molecular representations ready for prime time? *arXiv preprint arXiv:1904.01561*, 2019.

Rex Ying, Jiaxuan You, Christopher Morris, Xiang Ren, William L Hamilton, and Jure Leskovec. 使用可微分池的层次图表示学习。《第32届神经信息处理系统国际会议论文集》, 第4805-4815页, 2018年。

Muhan Zhang, Zhicheng Cui, Marion Neumann 和 Yixin Chen. 用于图分类的端到端深度学习架构。《第三十二届AAAI人工智能大会》, 2018年。

Xiao-Chen Zhang, Cheng-Kun Wu, Zhi-Jiang Yang, Zhen-Xing Wu, Jia-Cai Yi, Chang-Yu Hsieh, Ting-Jun Hou, and Dong-Sheng Cao. Mg-bert: 利用无监督原子表征学习进行分子性质预测。《生物信息学简报》, 2021年。

Shuangjia Zheng, Jiahua Rao, Zhongyue Zhang, Jun Xu, and Yuedong Yang. 利用自校正变压器神经网络预测逆反应。《化学信息与建模学报》, 60 (1) : 47-55, 2019a.

郑双佳、严昕、杨跃东、徐军。通过自注意机制的微笑语法分析识别结构-性质关系。《化学信息与建模学报》, 59 (2) : 914-923, 2019b.

A GNN 架构的详细说明

在应用 GNN 之前，我们首先使用 `pysmiles`⁴ 将分子的 SMILES 字符串解析为 NetworkX 图。然后，我们使用以下 GNN 作为分子编码器，并在实验中研究它们的性能。GNNs 的实现基于深度图库 (DGL)⁵。所有 GNN 的层数均为 2，所有层的输出维度均为 1,024。最后一层的激活函数 σ 为 identity，除最后一层外的所有层均为 ReLU。我们在下面的所有线性变换中都保留了偏置项 b ，因为这是 DGL 的默认设置，但我们的实验表明偏置项几乎不会影响结果，因此为了清晰起见，我们没有在下面的公式中显示偏置项。

图卷积网络 (GCN) (Kipf & Welling, 2017 年)。在 GCN 中，AGGREGATE 是根据节点度平方根的倒数加权平均的：

$$h_i^k = \sigma \left(W^k \sum_{j \in \mathcal{N}(i) \cup \{i\}} \frac{1}{\sqrt{d_i d_j}} h_j^{k-1} \right) \quad (6)$$

其中， W^k 是可学习矩阵， $\alpha_{ij} = 1/|\mathcal{N}(i)| + |\mathcal{N}(j)|$ ， σ 是激活函数。

图形注意力网络 (GAT) (Velickovic et al.) 在 GAT 中，AGGREGATE 是作为多头自我注意来实现的：

$$h_i^k = \sigma \left(\sum_{s=1}^S W^{k,s} \sum_{j \in \mathcal{N}(i) \cup \{i\}} \text{SOFTMAX}_{\alpha_{ij}^{k,s}} h_j^{k-1} \right) \quad (7)$$

其中， $\alpha_{ij}^{k,s} = \text{LeakyReLU}(w^T [W^{k,s} h_i^{k-1} \| W^{k,s} h_j^{k-1}])$ 是 (未规范化的) 注意力权重， w 是 (未规范化的) 注意力权重， h 是 (未规范化的) 注意力权重。

为可学习向量， $\|$ 表示连接操作， S 为注意力头数。在我们的

实验中， S 设置为 16，每个注意力头的维度设置为 64，因此总输出维度仍为 1,024。

Graph Sample and Aggregate (GraphSAGE) (Hamilton 等人, 2017 年)。在 GraphSAGE 的池化变体中，AGGREGATE 的实现方式是：

$$h_i^k = \sigma \left(W^k \sum_{j \in \mathcal{N}(i)} h_j^{k-1} \right) \quad (8)$$

其中 $h_i^k = \text{MAX}_{j \in \mathcal{N}(i)} \text{ReLU}(W^k h_j^{k-1})$ ， $\mathcal{N}(i)$ 是聚合邻域表示、

和 MAX 是元素最大池化函数。除了 MAX，AGGREGATE 还可以用 MEAN、GCN 和 LSTM 来实现，但我们的实验表明 MAX 的性能最好。

拓扑自适应图卷积网络 (TAGCN) (Du 等人, 2017)。在 TAGCN 中，如果我们用 H^k 表示层 k 中所有原子的表示矩阵，那么 AGGREGATE 可以写成

$$H^k = \sigma \sum_{l=0}^L \tilde{A}^l H^{k-1} W^{k,l} \quad (9)$$

其中， $\tilde{A} = D^{-1/2} A D^{-1/2}$ 是归一化邻接矩阵， L 是局部滤波器的大小。我们在实验中将 L 设为 2。值得注意的是，与其他 GNN 不同的是，单个 TAGCN 层可以聚合来自 L 跳以外邻居的节点信息。因此，TAGCN 的实际层数为 L^k ，在我们的实验中为 $2^2 = 4$ 。

B 命题 1 的证明

证明 要证明" \rightarrow "是 2^M 上的等价关系, 我们需要证明" \rightarrow "满足反身性、对称性和传递性:

反射性。对于任意 $A \in 2^M$, 显然 $h \sum_{i \in A} i = h \sum_{i \in A} i$ 。因此, 我们有 $A \rightarrow A$ 。

⁴<https://pypi.org/project/pysmiles/>
⁵<https://www.dgl.ai/>

对称性。对于任意 $A, B \in 2^M$, $A \rightarrow B \Rightarrow \sum_{i \in A} h_i = \sum_{i \in B} h_i \Rightarrow B \rightarrow A$, 反之亦然。
Therefore, we have $A \rightarrow B \Leftrightarrow B \rightarrow A$.

传递性。如果 $A \rightarrow B$, $B \rightarrow C$, 那么我们有 $\sum_{i \in A} h_i = \sum_{i \in B} h_i = \sum_{i \in C} h_i$. 因此, we have $A \rightarrow C$.

C 命题 2 的证明

证明 由于 READOUT 函数是求和函数, 分子 $G = (V, E)$ 的嵌入为 $\sum_{v \in V} h_v$. G 所包含的所有原子的最终嵌入之和: $h_G = \sum_{a_i \in V} h_{a_i}$. 因此, 对于反应 $R \rightarrow P$, 我们有 $\sum_{r \in R} h_r - \sum_{p \in P} h_p = h_{\sum_{v \in R} h_v - \sum_{v \in P} h_v} = h_{\sum_{v \in C^k} h_v}$, 其中 h_v^k 是 K 层 GNN 在反应物图上计算得出的原子 v 的表示向量, 而 h_v^{K-P} 是相同 K 层 GNN 在生成物图上计算得出的原子 v 的表示向量。如果我们将化学反应中的所有反应物 R (所有生成物 P) 视为一个图, 其中每个反应物 (积) 是该图的连通分量, 上式可写成 $\sum_{v \in R} h_v^k - \sum_{v \in P} h_v^k$. 表示 C^k 为与反应中心 C 的距离为 k 的原子集合 ($C^0 = C$)。

我们可以通过归纳法证明 $\sum_{v \in R} h_v^k - \sum_{v \in P} h_v^k = \sum_{v \in \bigcup_{k=0}^{K-1} C^k} h_v^k - \sum_{v \in \bigcup_{k=0}^{K-1} C^k} h_v^k$. K

让我们首先考虑 $K = 1$ 的初始情况。那么原子的最终表示取决于原子本身及其一跳邻居的初始特征。因此, 从 R 到 P 的最终表示不同的原子必须是那些从 R 到 P 至少有一个键发生变化的原子, 而这些原子正是反应中心 C 中的原子。因此

我们有 $\sum_{v \in R} h_v^1 - \sum_{v \in P} h_v^1 = \sum_{v \in C^1} h_v^1 - \sum_{v \in C^1} h_v^1$.

归纳步骤。假设我们有 $\sum_{v \in R} h_v^k - \sum_{v \in P} h_v^k = \sum_{v \in \bigcup_{k=0}^{K-1} C^k} h_v^k - \sum_{v \in \bigcup_{k=0}^{K-1} C^k} h_v^k$ 为 $K \geq 1$. 这意味着, 与反应中心距离不大于

而与反应中心的距离大于 $K - 1$ 的原子的表示则保持不变。那么对于 $K + 1$, GNN 层数增加了一层。这意味着邻近集 $\bigcup_{k=0}^{K-1} C^k$ 的原子会受到额外影响、

它们的表示也发生了变化。显然, 其表示形式为 $\sum_{k=0}^K h_v^k$. 因此, 我们有 $\sum_{v \in R} h_v^{K+1} - \sum_{v \in P} h_v^{K+1} = \sum_{v \in \bigcup_{k=0}^{K+1} C^k} h_v^{K+1} - \sum_{v \in \bigcup_{k=0}^{K+1} C^k} h_v^{K+1}$. 为 $K + 1$.

由于 $\sum_{r \in R} h_r - \sum_{p \in P} h_p = \sum_{v \in \bigcup_{k=0}^{K-1} C^k} h_v^k - \sum_{v \in \bigcup_{k=0}^{K-1} C^k} h_v^k$, 我们可以得出结论: resid-...

偶项 $\sum_{r \in R} h_r - \sum_{p \in P} h_p$ 是 h^k 的函数, 当且仅当 $a \in \bigcup_{k=0}^{K-1} C^k$ 时, 即原子 a 和反应中心 C 小于 K .

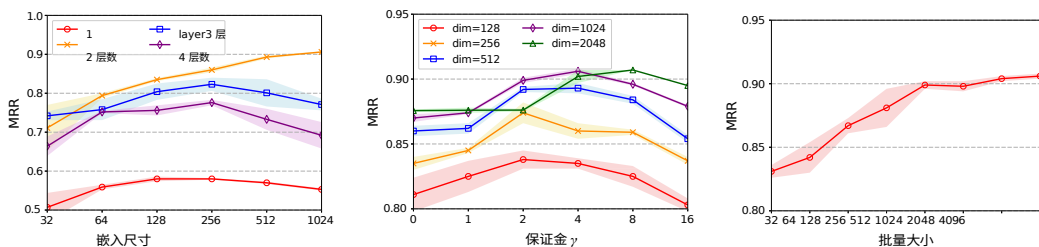
D 超参数灵敏度

我们研究了 MolR-GCN 在化学反应预测任务中对以下超参数的敏感性: GCN 层数 K 、嵌入维度、边距 γ 和批量大小。由于前三个超参数的影响高度相关, 我们报告了 MolR-GCN 受两个超参数共同影响时的结果。

图 5a 显示了 MolR-GCN 在改变 GCN 层数和嵌入维数时的结果。很明显, 当 K 太小 ($K = 1$) 或太大 ($K = 3, 4$) 时, MolR-GCN 的表现并不理想, 这从经验上证明了我们在第 2.2 节备注 3 中的说法。

图 5b 表明, 边距 γ 对模型性能也有很大影响, 而且这种影响与嵌入维度密切相关: 当嵌入维度从 128 增加到 2,048 时, 最佳 γ 也从 2 增加到 8, 这是因为两个向量之间的预期欧氏距离将与向量长度的平方根成比例地增加, 从而迫使最佳 γ 相应地右移。

图 5c 显示, 当批次规模从 32 增加到 4,096 时, MRR 也从 0.831 增加到 0.906。这是因为更大的批量会引入更多不匹配的反应物-生成物对, 如



(a) 对 # GCN 层和嵌入维度的敏感性 (b) 对嵌入维度和边距 γ 的敏感性 (c) 对批量大小的敏感性

图 5: MolR-GCN 的超参数灵敏度。

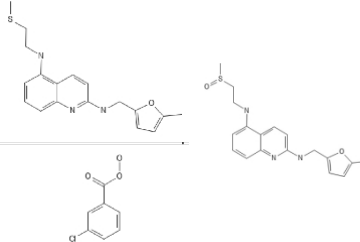
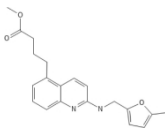
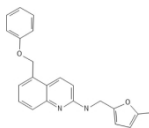
负样本，这为模型优化提供了更多监督。在批量大小 = # 个训练实例的极端情况下，将计算并优化所有成对分子距离。然而，更大的批次规模也会要求 GPU 中的迷你批次占用更多内存。当批次规模从 32 增加到 4,096 时，MolR-GCN 所需的内存也从 5.29 GiB 增加到 14.78 GiB。由于英伟达™ (NVIDIA®) V100 GPU 的内存只有 16 GB，这使得我们无法进一步增加批次大小。

E 关于美国专利商标局 479k 数据集的完整案例研究

USPTO-479k 数据集的完整案例研究如表 5 所示。为避免 "偷梁换柱"，我们选择了测试集中的前 20 个反应实例 (编号 0 ~ 编号 19) 进行案例研究。在表 5 中，第一列表示反应索引，第二和第三列表示反应物和反应实例。

第四至第六列分别表示 MolR-GCN、Mol2vec 和 MolBERT 预测的产物。为清晰起见，省略了三种方法都能正确预测产物的反应。预测产物下方的索引表示该产物所属的实际反应。

不	反应物	事实真相产品	预测产品 作者: MolR-GCN	预测产品 作者: Mol2vec	预测产品 作者: MolBERT
5					
			(编号: 32353)	(编号: 32353)	(编号: 32353)
6			与地面实况相同		
				(编号: 39181)	(第 24126 号)

8		与地面实况相同	 <p>(编号: 11233)</p>	 <p>(编号: 17526)</p>
---	---	---------	---	--

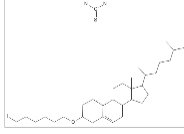
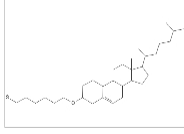
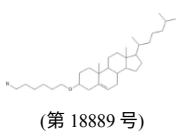
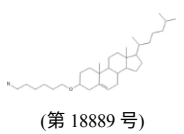
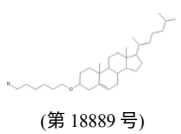
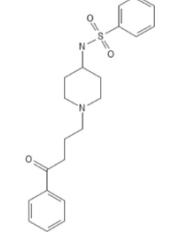
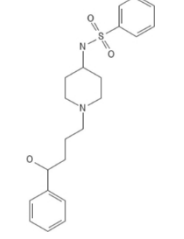
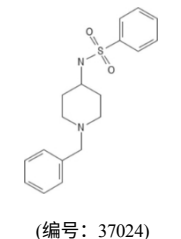
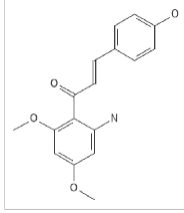
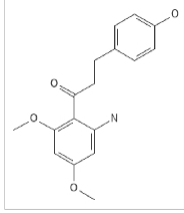
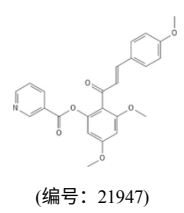
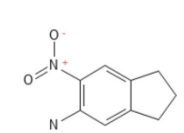
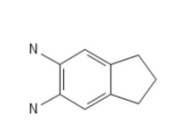
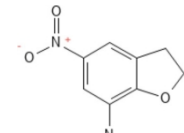
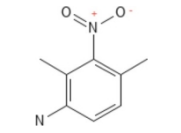
不	反应物	事实真相 产品	预测产品 作者: MolR-GCN	预测产品 作者: Mol2vec	预测产品 作者: MolBERT
10			 (第 18889 号)	 (第 18889 号)	 (第 18889 号)
13			与地面实况相同	与地面实况相同	 (编号: 37024)
16			与地面实况相同	 (编号: 21947)	与地面实况相同
17			与地面实况相同	 (第 2029 号)	 (编号: 22247)

表 5: USPTO-479k 测试集中前 20 个反应实例 (编号 0 ~ 编号 19) 的案例研究。为清晰起见, 省略了三种方法都能正确预测产物的反应。

预测生成物下方的指数表示该生成物所属的实际反应。

如表 5 所示, MolR-GCN 只在 2 个反应上出错, 而 Mol2vec 和 MolBERT 则在 6 个反应上出错。此外, 对于 MolR-GCN 没有正确回答的两个反应 (5 号和 10 号), Mol2vec 和 MolBERT 也给出了完全相同的错误答案, 这说明这两个反应的产物确实很难预测。实际上, 对于 5 号反应, 预测的生成物 (编号 32353) 与反应物完全相同, 严格来说, 这不能被视为错误答案。至于 10 号反应, 预测产物 (编号 18889) 与地面真值产物非常相似, 两者只差一个原子 (N 与 S)。

另外两个值得一提的例子是 (1) 6 号反应, 其中第一个反应物含有一个三角环, 但在生成物中却破裂了。MolR-GCN 可以正确预测生成物, 但 Mol2vec 和 MolBERT 则简单地将两个反应物合并在一起, 从而导致错误。(2) 第 17 号反应, 硝基 ($-\text{NO}_2$) 在反应后消失。MolR-GCN 成功预测了这种化学变化, 但 Mol2vec 和 MolBERT 预测的产物中仍然保留了硝基。

F 产品预测多选题详情

我们的产品预测多选题收集自牛津大学出版社的在线资源、^{6,7}, mhpracticeplus.com⁸和 GRE 化学考试练习册⁹.我们筛选出

67

<https://global.oup.com/uk/orc/chemistry/okuyama/student/mcqs/ch19/h>

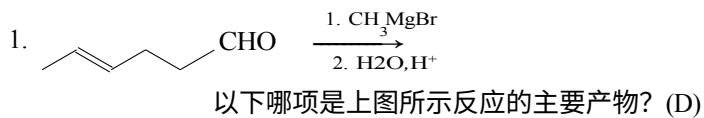
<https://global.oup.com/uk/orc/chemistry/okuyama/student/mcqs/ch21/>⁸

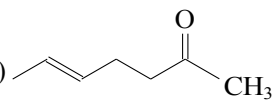
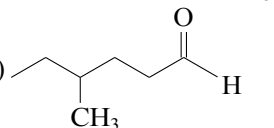
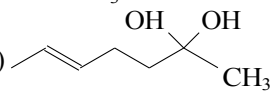
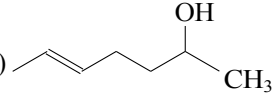
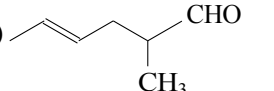
https://www.mhpracticeplus.com/mcat_review/MCAT-Review_Smith.pdf

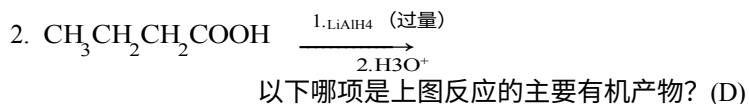
⁹https://www.ets.org/s/gre/pdf/practice_book_chemistry.pdf

1) 选项中包含立体异构体的问题，因为我们的模型和基线都无法处理立体异构体，以及 2) 包含我们的方法或基线无法成功解析的分子的问题。我们手动将分子图转换为 SMILES 字符串。以下列出了前五个问题的示例（问题后给出了正确答案），而這些问题的完整 SMILES 版本则包含在代码库中。

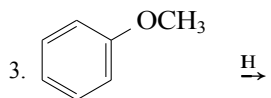
示例 1 ~ 5:




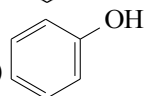
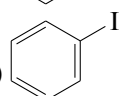
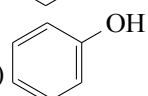

- (A) 
- (B) 
- (C) 
- (D) 
- (E) 



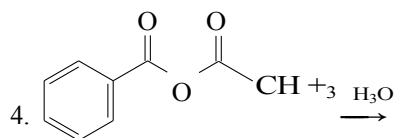
- (A) $\text{CH}_3\text{CHCH}_2\text{CH}(\text{OH})_2$
- (B) $\text{CH}_3\text{CH}_2\text{CHCHO}_2$
- (C) $\text{CH}_3\text{CH}_2\text{CHCH}_2\text{OH}$
- (D) $\text{CH}_3\text{CHCHCH}_2\text{OH}$
- (E) $\text{CH}_3\text{CH}_2\text{C}\equiv\text{CH}$








以下哪些是上图所示反应的主要产物? (D)

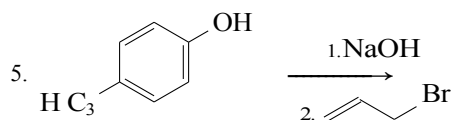
- (A)  + CH_3OH
- (B)  + CH_4
- (C)  + CH_3OH
- (D)  + CHI_3
- 

(E) + CH I₃

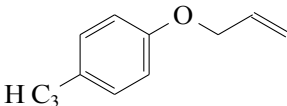
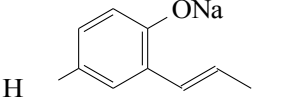
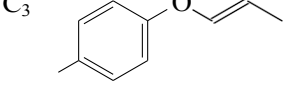
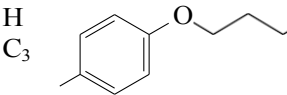


以下哪些是上图所示反应的产物? (E)

- (A)  + CH₃CHO
- (B)  + CH₃CH₂OH
- (C)  + CH₃CH₂OH
- (D)  + CH₃COOH
- (E)  + CH₃COOH



上图中对甲酚的反应产物是什么?

- (A) 
- (B) 
- (C) 
- (D) 
- (E) 