化学预训练模型系统调查

Jun Xia[1,*] , **Yanqiao Zhu**[2,*] , **Yuanqi Du**[3,*] and **Stan Z. Li**[1,†]
[1]西湖大学[2] 加州大学洛杉矶分校[3] 康奈尔大学
{xiajun, stan.zq.li}@westlake.edu.cn, yzhu@cs.ucla.edu, yd392@cs.cornell.edu

摘要

深度学习在学习分子表征方面取得了令人瞩目的成就，这对于从性质预测到药物设计等各种生化应用至关重要。然而，从头开始训练深度神经网络（DNN）往往需要大量标记的分子，而在现实世界中获得这些分子的成本很高。为了缓解这一问题，人们在化学预训练模型（CPM）方面做出了巨大努力，即利用大规模的未标注分子数据库对 DNN 进行预训练，然后针对特定的下游任务进行微调。尽管这一领域蓬勃发展，但却缺乏系统的综述。在本文中，我们首次对 CPM 目前的进展进行了调查总结。我们首先强调了从头开始训练分子再现模型的局限性，从而激发了 CPM 研究。接下来，我们从分子描述符、编码器结构、预训练策略和应用等几个关键角度系统地回顾了这一主题的最新进展。我们还强调了未来研究的挑战和前景，为机器学习和科学界提供了有用的资源。

1 导言

提取分子的向量表示对于将机器学习方法应用于广泛的分子任务至关重要。最初，分子指纹的开发是为了利用基于规则的算法将分子编码为二进制向量[Consonni 和 Todeschini，2009]。随后，各种深度神经网络（DNN）被用来以数据驱动的方式对分子进行编码。早期的尝试是利用基于序列的神经架构（如 RNN、LSTM 和变换器）来编码以简化分子输入行输入系统（SMILES）字符串表示的分子 [Weininger, 1988]。后来，有人认为分子可以用原子为节点、键为边的图结构自然地表示出来。这激发了一系列利用这种结构化归纳偏差的工作，以更好地

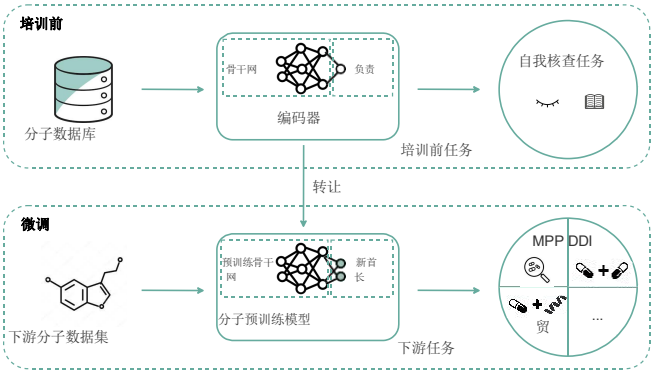


图 1：化学预训练模型 (CPM) 的典型学习流程。MPP：分子性质预测；DDI：药物-药物相互作用；DTI：药物-靶点相互作用。

分子表征 [Kearnes 等人，2016]。这些方法的关键进展是图形神经网络（GNN），它通过递归聚合邻域的节点特征，同时考虑图形结构和属性特征[Kipf 和 Welling，2017]。最近，考虑到分子在三维空间中不断运动的特性，用于分子表征的 GNNs 的另一个发展方向是建立分子构象的三维几何对称模型[Schu¨tt 等人，2017]。然而，上述大多数著作都是在有监督的环境下学习分子表征，这限制了它们在实践中的广泛应用，原因如下。(1) *标注数据稀缺*：由于分子数据标注通常需要昂贵的湿实验室实验，因此特定任务的分子标注可能非常稀缺；(2) *分布外泛化能力差*：在现实世界的许多情况下，学习不同大小或官能团的分子需要进行分布外泛化。例如，假设有人希望预测一种新合成分子的性质，而这种分子与训练集中以前的所有分子都不同。然而，从零开始训练的模型无法很好地推断出分布外分子的特性[Hu 等人，2020a]。

在自然语言处理（NLP）领域，预训练语言模型（PLM）一直是应对上述挑战的潜在解决方案[Devlin 等人，2019]。受其成功经验的启发，如图 1 所示，化学预训练模型

*平等贡献。 † 通讯作者。

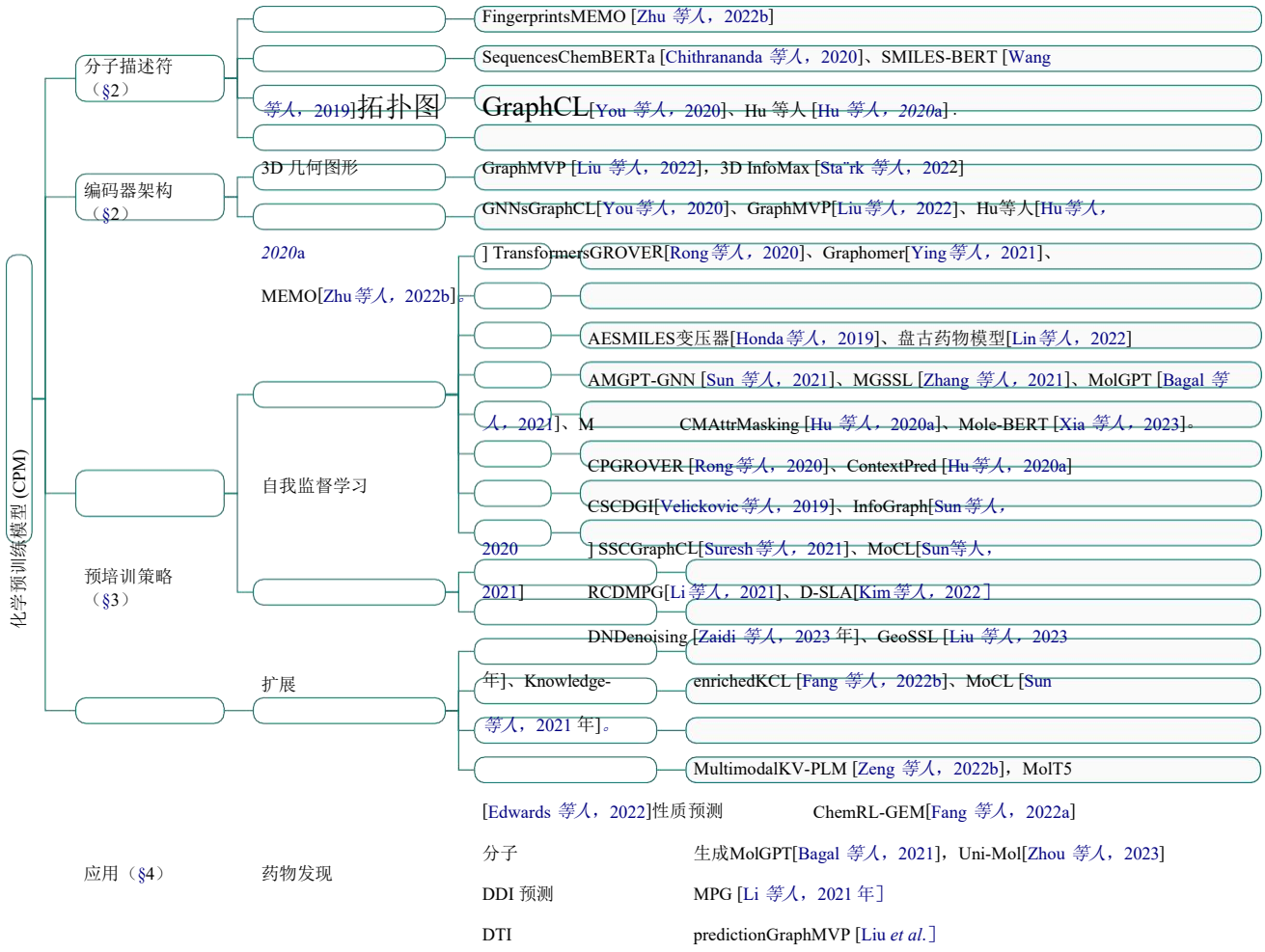


图 2：化学预训练模型 (CPM) 的分类法及代表性实例。

(CPMs) 已被引入，用于从大量未标记的分子中学习通用的分子代表，然后对特定的下游任务进行微调。最初，研究人员在基于字符串的分子数据（如 SMILES）上采用基于序列的预训练策略。一种典型的策略是对神经编码器进行预训练，以预测像 BERT 这样的随机屏蔽分子 [Devlin 等人，2019]。这一系列工作包括 ChemBERTa [Chithrananda 等人，2020]、SMILES- BERT [Wang 等人，2019]、Molformer [Ross 等人，2022]等。最近，研究界开始探索在（二维和三维）分子图上进行预训练。例如，[Hu 等人，2020a] 提出屏蔽原子或边缘属性并预测屏蔽属性。[Liu 等人，2022 年]通过最大化二维拓扑结构和三维几何结构之间的对应关系对 GNN 进行预训练。

尽管 CPM 已越来越多地应用于分子表征学习，但这一迅速扩展的领域仍然缺乏系统的综述。在本文中，我们首次对 CPM 进行了调查，以帮助不同背景的读者理解、使用和开发 CPM，用于各种实际任务。这项工作的贡献可归纳为以下四个方面。**(1)** *结构化分类法*。通过结构化分类法对该领域进行了广泛概述，从四个角度对现有工作进行了分类（图 2）：分子描述符、编码器架构、预训练策略和应用。**(2)** *全面回顾当前进展*。根据分类法，当前预训练模型的研究进展包括

(3) 丰富的附加资源。**(3)** *丰富的附加资源*。收集了丰富的资源，包括开源的 CPM、可用的数据集和重要的论文列表，可在 https://github.com/junxia97/ awesome-pretrain-on-molecules 上找到。这些资源将定期持续更新。**(4)** *未来方向讨论*。讨论了现有工作的局限性，并强调了几个有前途的研究方向。

2 分子描述符和编码器

为了将分子输入 DNN，必须用数字描述符对分子进行特征描述。人们设计了各种描述符来以简洁的格式描述分子。在本节中，我们将简要回顾这些分子描述符及其相应的神经编码器架构。

指纹 (FP)。分子指纹用二进制字符串描述分子中特定亚结构的存在或不存在。例如，PubChemFP [Wang 等人，2017] 编码了 881 种结构键类型，与 PubChem 数据库中化合物片段的子结构相对应。

序列。最常用的分子序列描述符是简化分子输入行输入系统（SMILES）[Weininger, 1988]，因为它具有通用性和可解释性。每个原子都用相应的 ASCII 符号表示。化学键、分支和立体化学用特定符号表示。变换器 [Vaswani

等人，2017 年]是一种强大的神经模型，用于处理序列并对每个标记之间的复杂关系进行建模。我们可以首先将基于序列的分子描述符拆分成一系列表示原子/键的标记，然后在这些标记之上应用变换器[Chithrananda 等人，2020；王等人，2019]。

二维图形。分子可以自然地表示为二维图，原子为节点，键为边。例如，每个节点和边还可以携带表示原子类型/手性和键类型/方向的特征向量[Hu 等人，2020a]。在此，GNNs [Kipf 和 Welling，2017；Xu 等人，2019]可用于学习二维分子图表示。还可以利用 GNN 和变换器的一些混合架构[Rong 等人，2020；Ying 等人，2021]来捕捉分子图的拓扑结构。

三维图形。三维几何图形表示分子中原子在三维空间中的空间排列，其中每个原子都与其类型和坐标以及一些可选的几何属性（如速度）相关联。使用三维几何的优势在于，构象信息对许多分子特性，尤其是量子特性至关重要。此外，利用三维几何图形还可以直接利用立体化学信息，如手性。许多方法[Schu¨ tt 等人，2017；Satorras 等人，2021；Du 等人，2022a]开发了三维几何的消息传递机制，使图形表示遵循某些物理特性，如平移和旋转的等价性。

3 培训前策略

在本节中，我们将详细介绍几种具有代表性的 CPM 自我监督预训练策略。

3.1 自动编码 (AE)

使用自编码器重构分子（图 3a）是学习具有表现力的分子表征的天然自监督目标。分子重构中的预测

结构是给定分子的（部分）结构，如原子或化学键子集的属性。一个典型的例子是 SMILES 变换器[Honda 等人，2019]，它能

该研究利用基于变换器的编码器-解码器网络，通过重建 SMILES 字符串所代表的分子来学习表征。最近，与输入和输出数据类型相同的传统自动编码器不同，[Lin 等人，2022] 预先训练了一个图到序列的异或度量条件变异自动编码器来学习分子表征。虽然自编码器可以学习有意义的分子表征，但它们只关注单个分子，无法捕捉分子间的关系，这限制了它们在一些下游任务中的表现[Li 等人，2021]。

3.2 自回归模型 (AM)

自回归建模法（AM）将分子内含物因子化为一系列子成分，然后以序列中的前一个子成分为条件，逐一预测这些子成分。MolGPT [Bagal *et al.*, 2021]沿用了 NLP 中的 GPT [Brown *et al.*, 2020]，预先训练了一个转换器网络，以这种自回归的方式预测 SMILES 字符串中的下一个字符串。

对于分子图，GPT-GNN [Hu 等人，2020b] 通过一系列步骤（图 3b）重建分子图，这与一次性重建整个图的图自动编码器截然不同。具体来说，给定一个节点和边都被随机屏蔽的图，GPT-GNN 每次生成一个屏蔽节点及其边，并最大化每次迭代生成的节点和边的可能性。然后，迭代生成节点和边，直到生成所有屏蔽节点。与此类似，MGSSL [Zhang 等人，2021] 以自回归方式生成分子图图案，而不是单个原子或键。从形式上看，这种自回归模态目标可以写成

$$L_{AM} = -E_{M\in D} \sum_{i=1}^{|V|} \log p(C_i \mid C_{<i}), \quad (1)$$

其中 C_i ，$C_{<i}$ 分别是*第 i* 个组件的属性和分子 M 中索引 *i* 之前生成的属性。与其他策略相比，AM 使 CPM 在生成分子方面表现得更好，其训练过程与分子生成过程类似 [Bagal 等人，2021]。不过，AM 的计算成本较高，而且需要事先对原子或化学键进行排序，这对于分子来说可能并不合适，因为原子或化学键并不存在固有的排序。

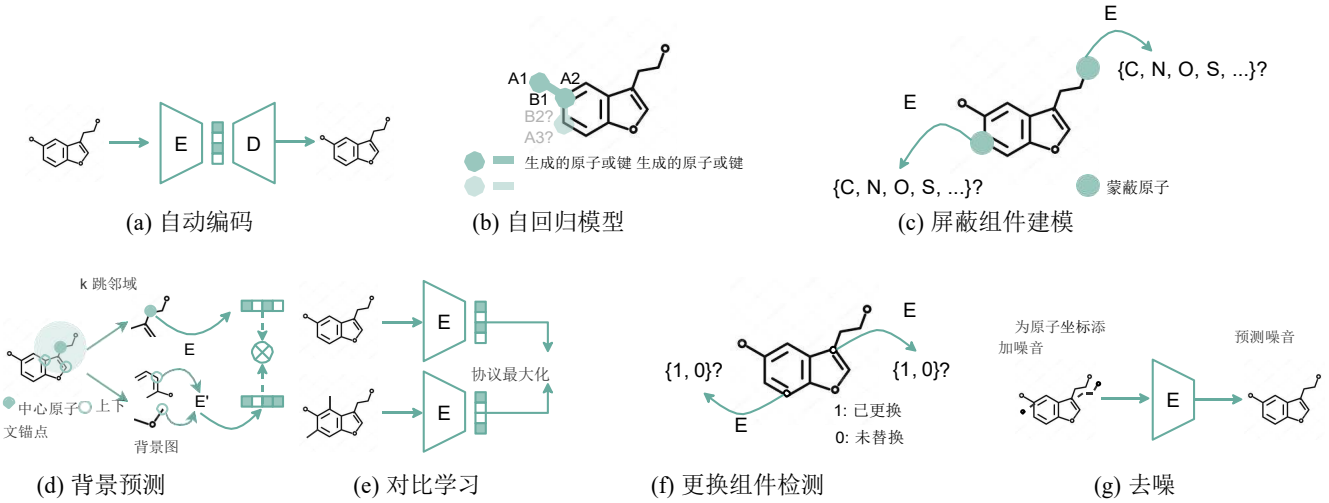
3.3 屏蔽组件建模 (MCM)

在语言领域，屏蔽语言建模（MLM）已成为一种主要的预训练目标。具体来说，MLM 会随机屏蔽掉输入句子中的标记，然后训练模型使用剩余的标记来预测这些被屏蔽的标记 [Devlin 等人，2019]。屏蔽成分建模（MCM，图 3c）将 MLM 的理念推广到分子中。具体来说，MCM 屏蔽了分子的某些成分（如原子、键和片段），然后根据剩余成分训练模型来预测它们。一般来说，其目标可表述为

$$L_{MCM} = -E_{M\in D} \sum_{\tilde{M}\in m(M)} \log p(\tilde{M} \mid M\backslash m(M)), \quad (2)$$

其中，$m(M)$ 表示分子 M 中被掩蔽的成分，$M/m(M)$ 表示剩余成分。对于基于序列的预训练，ChemBERTa [Chithrananda *et al.*，2020]、SMILES-BERT [Wang *et al.*，2019]和 Mol- former [Ross *et al.*，2022]屏蔽了 SMILES字符串中的随机字符，然后根据损坏的SMILES字符串的变换器输出恢复字符。对于分子图的预训练，[Hu 等人，2020a] 建议随机屏蔽输入的原子/化学键属性，然后预训练 GNN 来预测它们。同样，GROVER [Rong 等人，2020] 尝试预测屏蔽子图，以捕捉分子图中的上下文信息。最近，Mole-BERT [Xia 等人，2023]认为，由于自然界中的原子集极小且不平衡，屏蔽原子类型可能存在问题。为了缓解这一问题，他们开发了一种上下文感知标记器，将原子编码为具有化学意义的离散值，以便进行屏蔽。

MCM 尤其适用于注释丰富的分子。例如，屏蔽原子属性可以



(a) 自动编码 (b) 自回归模型 (c) 屏蔽组件建模

(d) 背景预测 (e) 对比学习 (f) 更换组件检测 (g) 去噪

图 3：七种无监督预培训策略的语义图。E：编码器；D：解码器。

GNNs 可以学习简单的化学规则（如化合价）以及其他潜在的复杂化学描述符（如官能团的电子效应或立体效应）。此外，与上述 AM 策略相比，MCM 可根据周围环境预测被遮蔽的成分，而 AM 则仅仅依赖于按预定顺序预提成分。因此，MCM 可以捕捉到更完整的化学语义。然而，由于 MCM 在 BERT [Devlin 等人，2019 年] 之后的预训练过程中通常会掩盖每个分子的固定部分，因此它不能对每个分子中的所有成分进行训练，从而降低了样品的利用效率。

3.4 语境预测 (CP)

上下文预测（CP，图 3d）旨在以明确的、上下文感知的方式捕捉分子/原子的语义。一般来说，CP 可以表述为

$$L_{CP} = -E_{M \in D} \log p(t \mid M_1, M_2), \quad (3)$$

其中，如果邻域成分 M_1 和周围上下文 M_2 共享同一个中心原子，则 $t = 1$，否则 $t = 0$。例如，[Hu 等人，2020a] 使用二元分类法来判断分子结构和周围上下文结构中的子图是否属于同一个节点。虽然 CP 简单有效，但它需要一个辅助神经模型将上下文编码成固定向量，这就为大规模预训练增加了额外的计算开销。

3.5 对比学习（CL）

对比学习（Contrastive Learning，CL，图 3e）通过最大化一对相似输入（如同一分子的两个不同增强或描述）之间的一致性来预训练模型。根据对比粒度（如分子或亚结构级），我们在 CPM 中引入了两类 CL：跨尺度对比（CSC）和同尺度对比（SSC）。

跨尺度对比（CSC）。Deep InfoMax 是一种具有代表性的 CSC 模型，最初是通过对比图像及其局部区域与其他负对来学习图像表征的[Hjelm 等人，2019]。对于分子图，InfoGraph [Sun 等人，2020] 遵循了这一理念，通过对比分子和亚结构水平的表征，可形式化地描述为

$$L_{CSC} = -E_{M \in D} \log s(M, C) - \log \sum_{C^- \in N} s(M, C^-), \quad (4)$$

其中，N 是一组负样本，C 是 M 的子结构，C^- 是另一个分子的子结构，$s(\cdot, \cdot)$ 表示相似度量。后续工作 MVGRL [Hassani 和 Ahmadi, 2020] 通过节点扩散生成一个
然后通过对比一个视图的原子表示和另一个视图的分子表示，最大限度地提高原始视图和增强视图之间的相似性，反之亦然。

同尺度对比（SSC）。同尺度对比（SSC）通过将增强分子推近锚分子（正对）并远离其他分子（负对），对单个分子进行对比学习。例如，GraphCL [You *et al.*，2020] 及其变体 [You *et al.*，2021；Sun et *al.*，2021；Suresh *et al.*，2021；Xu *et al.*，2021a；Fang *et al.*，2022b；Wang et *al.*，2021；Xia et *al.*，2022b；Wang *et al.*，2022a] 为以图表示的分子级预训练提出了各种增强策略。此外，最近的一些研究还最大限度地提高了相同分子的各种描述符之间的一致性，并排斥不同的描述符。例如，SMICLR [Pinheiro 等人，2022 年] 联合利用图编码器和 SMILES 字符串编码器来执行 SSC；MM-Deacon [Guo 等人，2022 年] 利用两个独立的转换器对分子的 SMILES 和国际纯粹与应用化学联合会（IUPAC）进行编码，然后使用对比目标来提高 SMILES 和 IUPAC 表示的相似性。

GeomGCL[Li等人，2022b]采用双视角几何信息传递神经网络（GeomMPNN）对分子的二维和三维图形进行编码，并设计几何对比目标。SSC 预训练目标的一般表述为

$$L_{SSC} = -E_{M\in D} \log s(M, M') - \log \sum_{M^-\in N} s(M, M^-) \quad (5)$$

其中，M' 是分子 M 的增强版本或其他描述符，N 是负样本集。取得了可喜的成果，但仍有一些关键问题需要解决。

一些物理问题阻碍了它的广泛应用。首先，在分子增强过程中很难保留语义。现有的解决方案是通过人工试错（You 等，2020 年）、繁琐的优化（You 等，2021 年）或昂贵的领域知识指导（Sun 等，2021 年）来选择增强，但目前仍缺乏一种高效且有原则的方法来设计化学上适合分子预训练的增强。此外，CL 背后的假设是拉近相似表征的距离，但这一假设在分子表征学习中并不总是成立的。例如，在分子活性悬崖[Stumpfe 等人，2019]的情况下，相似的分子具有完全不同的特性。因此，哪种预训练策略能更好地消除分子间的差异仍是个未知数。此外，大多数 CPM 中的 CL 目标会随机选择一批分子中的所有其他分子作为阴性样本，而不管它们的真实语义如何，这将不利于排斥具有相似性质的分子，并因假阴性而影响性能[Xia 等人，2022a]。

3.6 更换组件检测 (RCD)

替换成分检测（RCD，图 3f）建议识别输入分子中随机替换的成分。例如，MPG [Li 等人，2021] 将每个分子分成两部分，通过组合两个分子的部分来改变其结构，并训练编码器来检测组合后的部分是否属于同一分子。这一目标可写成

$$L_{RCD} = -E_{M\in D} [\log p(t \mid M_1, M_2)] \quad (6)$$

其中，如果 M_1 和 M_2 来自同一分子 M，则 $t = 1$，否则 $t = 0$。虽然 RCD 可以揭示分子结构的内在模式，但编码器经过预先训练，对于所有天然分子总是生成相同的 "非替换 "标签，而对于随机组合的分子则生成 "替换 "标签。然而，在下游任务中，输入的分子都是天然分子，导致 RCD 生成的分子表征区分度较低。

3.7 去噪 (DN)

受去噪扩散概率模型成功的启发[Ho 等人，2020]，去噪模型（DeNoising，图 3g）最近也被采用为学习分子表征的预训练策略。最近的一项工作[Zaidi 等人，2023 年]增加了将噪声转化为三维分子几何的原子坐标，并预先训练编码器预测噪声。他们证明，这种去噪目标近似于学习分子力场。与此同时，Uni-Mol [Zhou 等人，2023] 在原子坐标中加入了噪声，其动机是，根据三维原子位置，可以很容易地推断出被遮蔽的原子类型。最近，GeoSSL [Liu *et al.*

三维分子的性质。一般来说，去噪的预训练目标可以表述为

$$L_{DN} = E_{M\in D} \|\epsilon - f_\theta(\tilde{M})\|^2, \quad (7)$$

其中，ϵ 表示添加的噪声，$\tilde{M}$ 表示输入的噪声。分子 M 添加了噪声，$f_\theta(\cdot)$ 表示预测噪声的编码器。

3.8 扩展

知识丰富的预训练。CPM 通常从大型分子数据库中学习一般的分子表征。然而，它们往往缺乏特定领域的知识。为了提高它们的性能，最近有几项研究尝试将外部知识注入 CPM。例如，GraphCL [You *et al.*, 2020]首先指出，键 perturba- tions（添加或删除键作为数据增强）在概念上与领域知识不相容，而且从经验上看对化学物质的对比预训练没有帮助。因此，他们避免在分子图扩增中采用化学键扰动。更明确地说，MoCL [Sun 等人，2021 年] 提出了一种基于领域知识的分子扩增算子，称为子结构置换，其中分子的有效子结构被生物异构体置换，从而产生一个与原始分子具有相似物理或化学性质的新分子。最近，KCL [Fang *et al.*, 2022b]构建了化学元素知识图谱（KG）来总结化学元素之间的微观关联，并提出了一种用于分子表征学习的新型知识增强型强制学习（KCL）框架。此外，MGSSL [Zhang 等人，2021] 首先利用现有算法 [Degen 等人，2008] 提取有语义意义的主题，然后预训练神经编码器，以自回归方式预测主题。ChemRL-GEM [Fang 等人，2022a] 建议利用分子几何信息来加强分子图的预训练。它设计了一种基于几何的 GNN 架构以及几种几何级自监督学习策略（键长预测、键角预测和原子间距矩阵预测），以便在预训练过程中捕捉分子几何知识。虽然知识丰富的预训练有助于 CPM 捕捉化学领域的知识，但它需要昂贵的先验知识作为指导，当先验知识不完整、不正确或获取成本高昂时，就会对更广泛的应用造成障碍。

多模态预训练。除了第 2 节中提到的描述符之外，还可以使用其他模式（包括图像和生化文本）来描述分子。最近的一些研究对分子进行了多模态预训练。例如，KV-PLM [Zeng *et al.*

表 1：文献中具有代表性的化学预训练模型 (CPM) 摘要。

	型号	输入	骨干架构	预训练任务	预训练数据库	#Params.	链接
序列	SMILES 变压器 [本田 等人，2019]	SMILES	变压器	AE	ChEMBL (861K) [Gaulton *et al.*]	-	链接
	ChemBERTa [Chithrananda 等人，2020]	SMILES/SELFIES [Krenn 等人，2020]	变压器	MCM	PubChem (77M) [Wang 等人，2017 年]	-	链接
	SMILES-BERT [Wang 等人，2019]	SMILES	变压器	多孔材料	ZINC15（~ 18.6M）[Sterling 和 Irwin, 2015]	-	链接
	Molformer [Ross 等人，2022 年]	SMILES	变压器	MCM	ZINC15 (1B) + PubChem (111M)	-	-
图形/几何	Hu 等人 [Hu *et al.*]	图形	5 层 GIN	CP + MCM	ZINC15 (2M) + ChEMBL (456K)	~ 2M	链接
	GraphCL [You *et al.*]	图形	5 层 GIN	SSC	ZINC15 (2M)	~ 2M	链接
	JOAO [You *et al.*]	图形	5 层 GIN	SSC	ZINC15 (2M)	~ 2M	链接
	AD-GCL [Suresh *et al.*]	图形	5 层 GIN	SSC	ZINC15 (2M)	~ 2M	链接
	GraphLoG [Xu 等人, 2021b]	图形	5 层 GIN	SSC	ZINC15 (2M)	~ 2M	链接
	MGSSL [Zhang *et al.*]	图形	RCD + MCM ZINC + ChEMBL (11M) 53M Link LP-Info [You *et al*、2022b] 图形 5 层 GIN SSC ZINC15 (2M) ~ 2M 链接 GraphMAE [Hou 等人, 2022] 图形 5 层 GIN AE ZINC15 (2M) ~ 2M 链接 MGMAE [Feng 等人, 2022] 图形 5 层 GIN AE ZINC15 (2M) + ChEMBL (456K) ~ 2M -				
	GROVER [Rong et al.]	图形	GTransformer [Rong *et al.*]	CP + MCM	ZINC + ChEMBL (10M)	48M~ 100M	链接
	MolCLR [Wang *et al.*]	图形	GCN + GIN	SSC	PubChem (10M)	-	链接
	Graphomer [Ying *et al.*]	图形	Graphomer [Ying *et al.*]	监督	PCQM4M-LSC (~ 3.8M) [Hu et *al.*]	-	链接
	3D-EMGP [Jiao 等人，2023]	几何学	E(3)-equivariant GNNs	DN	GEOM (100K) [Axelrod and Go´mez-Bombarelli, 2022]	-	Link Mole-BERT
	[Xia et al.	Link Mole-BERT [Xia *et al.*]	图形	5 层 GIN	MCM + SSC	ZINC15	(2M)
		~ 2M	链接去噪 [Zaidi *et al.*]	几何	GNS [Sanchez-Gonzalez *et al.*]	DN	
		PCQM4Mv2 (~3.4M)	-	链接 GeoSSL [Liu *et al.*]	几何	PaiNN [Schu¨tt *et al.*]	DN Molecule3D [Xu *et* al.
多模式/外部知识	DMP [Zhu *et al.*]	图形 + SMILES	DeeperGCN + Transformer	MCM + SSC	PubChem (1.1 亿)	104.1M	链接
	GraphMVP [Liu *et al.*]	图形 + 几何	5 层 GIN + SchNet [Schu¨tt 等人，2017]	SSC + AE	GEOM (50K)	~ 2M	Link
	3D Infomax [Sta¨rk *et al.*]	图形 + 几何	PNA [Corso *et al.*]	SSC	QM9 (50K) + GEOM (140K) + QMugs (620K)	-	链接
	KCL [Fang *et al.*]	图形 + 知识图谱	GCN + KMPNN [Fang *et al.*]	SSC	ZINC15 (250K)	<1M	链接
	KV-PLM [Zeng *et al.*]	SMILES + 文本	变压器	MLM + MCM	PubChem (150M)	~ 110M	链接
	MEMO [Zhu *et al.*]	SMILES + FP + Graph + Geometry	转换器 + GIN + SchNet	SSC GEOM (50K) -	- MolT5 [Edwards 等人，2022] SMILES + 文本转换器 替换损坏跨度		
	ZINC-15 (100M) 60M / 770M 链接 MICER [Yi 等人，2022]	SMILES + 图像 CNN + LSTM AE ZINC20 -	链接				
	MM-Deacon [Guo *et al.*]	SMILES + IUPAC	变压器	SSC	PubChem	10M	-
	盘古药物模型 [Lin 等人，2022]	图形 + SELFIES [Krenn 等人，2020]	转换器	AE	ZINC20 + DrugSpaceX + UniChem (~1.7B)	~ 104M	链接
	KPGT [Li 等人，2022a]	SMILES + FP	LiGhT [Li *et al.*]	MCM	ChEMBL29 (2M)	-	链接
	ChemRL-GEM [Fang *et al.*]	图形 + 几何	GeoGNN [Fang et *al.*]	MCM+CP	ZINC15 (20M)	-	链接
	ImageMol [Zeng *et al.*]	分子图像	ResNet18 [He *et al.*]	AE + SSC + CP	PubChem (~10M)	-	Link
	Uni-Mol [Zhou *et al.*]	几何学 + 蛋白质口袋	变压器	MCM + DN	ZINC/ChemBL + PDB [Berman *et al.*]	-	链接

将 SMILES 字符串和生化文本标记化。然后，他们随机屏蔽部分标记，并预先训练神经编码器以恢复被屏蔽的标记。与此类似，MolT5[Edwards 等人，2022]遵循 T5[Raffel 等人，2020]的替换损坏跨距任务，首先屏蔽丰富的 SMILES 字符串和分子的生化文本描述中的部分跨距，然后预训练转换器来预测屏蔽的跨距。这样，这些预训练模型就能同时生成 SMILES 字符串和生化文本，这对于文本引导的分子生成和分子标题（生成分子的描述性文本）尤其有效。[Zhu 等人，2022b] 建议使用对比目标最大化四个分子去脚本嵌入和它们的集合嵌入之间的一致性。这样，这些不同的描述符就能在分子性质预测任务中相互协作。此外，MICER[Yi 等人，2022]采用了一种基于自动编码器的预训练框架来进行分子图像标注。具体来说，他们将分子图像输入预训练编码器，然后解码相应的 SMILES 字符串。上述多模态预训练策略可以推进各种模态之间的翻译。此外，这些模式还能共同为各种下游任务创建更完整的知识库。

4 应用

下文将以药物发现为例，介绍 CPM 的几种前景广阔的应用（表 1）。

4.1 分子性质预测 (MPP)

在现实生活中，候选新药的生物活性受到多种因素的影响，包括在胃肠道中的溶解度、肠膜渗透性以及肠/肝首过代谢。然而，这些标签由于湿实验室实验通常既费力又昂贵，因此大量未标记的分子可能非常稀缺。CPM 提供了一种可以利用大量未标记分子的方法，可作为下游分子性质预测任务的强大支柱[Wang 等人，2022b；Zaidi 等人，2023]。此外，与从头开始训练的模型相比，CPMs 能更好地推断分布外分子，这在预测新合成药物的性质时尤为重要[Hu 等人，2020a]。

4.2 分子生成（MG）

分子生成是计算机辅助药物设计领域的一项长期挑战，机器学习方法，尤其是生成模型，缩小了搜索空间，提高了计算效率，使深入研究看似无限的药物化学空间成为可能[Du 等人，2022b]。事实证明，采用自回归预训练方法的 CPM，如 MolGPT [Bagal 等人，2021]，有助于生成有效、独特和新颖的分子结构。多模态分子预训练技术的出现[Edwards 等人，2022；Zeng 等人，2022b]通过将描述性文本转化为分子结构，进一步拓展了分子生成的可能性。CPM 展示其能力的另一个关键领域是生成三维分子构象，特别是用于预测蛋白质配体结合位置。基于分子动力学或马尔科夫链蒙特卡洛的常规方法往往受到计算能力的限制，尤其是对较大的分子而言[Hawkins，2017]，而基于三维几何的CPM[Zhu 等人，2022a；Zhou 等人，2023]则不同，它们能在预训练过程中捕捉二维分子与三维构象之间的一些固有关系，因此在构象生成任务中表现出显著的优越性。

4.3 药物-目标相互作用 (DTI)

药物与靶点相互作用（DTI）的预测分析是药物发现早期阶段的重要步骤，因为它有助于确定与特定蛋白质靶点具有结合潜力的候选药物。这在药物再利用中尤为重要，因为再利用的目的是将已获批准的药物用于新的疾病，从而减少进一步发现药物的需要，并将安全风险降至最低。同样，获得足够的药物靶点数据以进行监督训练也具有挑战性。CPM 可以通过提供具有良好初始化的分子编码器来克服这一问题。要利用 CPM 实现准确的 DTI 预测，必须同时考虑分子编码器和靶标编码器，预测结合亲和力，并针对 DTI 预测任务对两者进行联合训练[Nguyen 等人，2021 年]。之前的工作，如 MPG [Li 等人，2021] 遵循这些原则来推进 DTI 预测。

4.4 药物之间的相互作用（DDI）

准确预测药物间相互作用（DDI）是药物研发管线的另一个关键阶段，因为这种相互作用会导致不良反应，损害健康甚至导致死亡。此外，准确的 DDI 预测还有助于提出明智的用药建议，使其成为市场批准前 监管调查的重要组成部分。从机器学习的角度来看，DDI 预测可视为一项分类任务，它将联合用药的影响确定为协同、相加或拮抗。要实现有效的 DDI 预测，需要具有表现力的分子表征，而这可以通过 CPM 获得。MPG [Li 等人，2021 年] 是一个具有代表性的例子，通过将 DDI 预测作为下游任务，证明了 CPM 的实用性。

5 结论和未来展望

总之，本文全面概述了化学预训练模型。我们首先回顾了广泛使用的分子描述符和编码器，然后介绍了具有代表性的预训练策略，并评估了它们的优缺点。我们还展示了 CPM 在药物发现和开发中的各种成功应用。尽管取得了丰硕的成果，但仍有一些挑战需要在未来进一步研究。

5.1 改进编码器架构和预训练目标

虽然在分析神经架构的学习能力（如 GNN 的 WL 测试）方面取得了长足进步，但这些分析在确定高度结构化分子的最佳设计方面缺乏特异性。针对 CPM 的理想特征化和架构仍然难以捉摸，这一点可以从相互矛盾的结果中得到证明，例如在之前的研究中，图形学习中广泛采用的图形注意网络（GAT）[Velickovic 等人，2018] 对下游性能产生了负面影响[Hu 等人，2020a；Hou 等人，2022]。此外，迫切需要探索如何将消息传递技术无缝集成到转换器中，作为统一的编码器，以适应大规模分子图的预训练。此外，由于在第 3 节中讨论过，预培训目标仍有很大的改进余地，MCM 中子组件的高效屏蔽策略就是一个很好的例子。

5.2 建立可靠和现实的基准

尽管对 CPM 进行了大量研究，但由于采用的评估设置（如随机种子和数据集分割）不一致，其实验结果有时并不可靠。例如，在包含多个昂贵的分子性质预测数据集的 MoleculeNet [Wu 等人，2018]上，可能由于这些分子数据集的规模相对较小，同一个模型的性能在不同的随机种子下会有很大差异。同样重要的是，要为 CPM 建立更可靠、更现实的基准，并将分布外泛化考虑在内。一种解决方案是通过支架拆分来评估 CPM，即根据分子的子结构拆分分子。在现实中，研究人员往往必须将从已知分子中训练出来的 CPM 应用于新合成的未知分子，而这些分子在性质上可能存在很大差异，并属于不同的领域。在这方面，最近成立的治疗数据共享中心（TDC）[Huang 等人，2021 年] 为公平评估各种治疗应用中的 CPM 提供了一个大有可为的机会。

5.3 扩大化学预训练模型的影响

CPMs研究的最终目标是开发出多功能分子编码器，可应用于与分子相关的大量下游任务。然而，与 NLP 界 PLM 的进展相比，CPM 的方法论进步与实际应用之间仍存在巨大差距。一方面，CPM 生成的表征尚未被广泛用于取代化学领域的传统分子描述符，预训练模型也尚未成为社区的标准工具。另一方面，对于这些模型如何有利于单个分子之外的更广泛的下游任务（如化学反应预测、虚拟筛选中的分子相似性搜索、逆合成、化学空间探索等）的探索也很有限。

5.4 建立理论基础

尽管 CPM 在各种下游任务中表现出令人印象深刻的性能，但人们对这些模型的严谨理论理解却十分有限。这种理论基础的缺乏对科学界和行业利益相关者都是一个障碍，因为他们都希望最大限度地发挥这些模型的潜力。必须建立 CPM 的理论基础，才能充分理解其机理以及如何在各种应用中提高性能。例如，最近的一项实证研究[Sun 等人，2022 年]对某些自监督图形预训练策略在某些下游任务中优于非预训练策略提出了质疑。要想更深入地了解不同分子预训练目标的有效性，从而为优化方法设计提供指导，还需要进一步的研究。

参考资料

[S. Axelrod 和 R. Go´mez- Bombarelli.GEOM, Energy-Annotated Molecular Conformations for Property Prediction and Molecular Generation.*科学数据*，2022。

[巴加尔等人，2021 年] V. Bagal、R. Aggarwal 等人，MolGPT：使用变压器解码器模型进行分子生成。*J. Chem.Inf.Model*.

[H. M. Berman, J. Westbrook, et al.*Nucleic Acids Res.*, 2000.

[布朗等人，2020 年] T. B. 布朗、B. 曼等人，语言模型是少数几个学习者。In *NeurIPS*, 2020.

[S. Chithrananda, G. Grand, et al. Chem- BERTa：用于分子性质预测的大规模自监督预训练。*arXiv.org*，2020 年 10 月。

[Consonni and Todeschini, 2009] V. Consonni and R. Todeschini. *化学信息学分子描述符》*。2009.

[G. Corso, L. Cavalleri, et al. Principal Neigh- bourhood Aggregation for Graph Nets.In *NeurIPS*, 2020.

[Degen 等人，2008 年] J. Degen、C. Wegscheid-Gerlach 等人，《编译和使用 "类药物 "化学片段空间的艺术》。*ChemMedChem*, 2008.

[Devlin 等人，2019] J. Devlin、M. Chang 等人，BERT：用于语言理解的深度双向变换器预训练。In *NAACL*, 2019.

[Du 等人，2022a] W. Du、H. Zhang 等人，SE(3) Equivariant Graph Neural Networks with Complete Local Frames.In *ICML*, 2022.

[Du 等人，2022b] Y. Du、T. Fu 等人，MolGenSurvey：*ArXiv.org,* March 2022.

[C. Edwards, T. Lai, et al.见 *EMNLP*，2022。

[X. Fang, L. Liu, et al.*Nat.Mach.Intell*.

*[方*等人，2022b] Y. Fang, Q. Zhang, et al. 分子对比学习与化学元素知识图谱。In *AAAI*, 2022.

[Feng 等人，2022 年] J. Feng、Z. Wang 等人，MGMAE：通过高掩码率重构异构图进行分子表征学习（Molecu- lar Representation Learning by Reconstructing Heterogeneous Graphs with A High Mask Ratio）。In *CIKM*, 2022.

[ChEMBL: A Large-Scale Bioactivity Database for Drug Discovery.*核酸研究》*，2012 年。

[Guo 等人，2022 年] Z. Guo、P. K. Sharma 等人，通过对比预训练进行多语言分子表征学习。In *ACL*, 2022.

[Hassani and Ahmadi, 2020] K. Hassani and A. H. K. Ahmadi.Con- trastive Multi-View Representation Learning on Graphs.In *ICML*, 2020.

[Hawkins, 2017] P. C. D. Hawkins.构象生成：The State of the Art.*J. Chem.Inf.Model.*, 2017.

[He *et al.,* 2016] K. He, X. Zhang, et al. Deep Residual Learning for Image Recognition.In *CVPR*, 2016.

[Hjelm 等人，2019] R. D. Hjelm、A. Fedorov 等人，《通过互信息估计和最大化学习深度表征》。In *ICLR*, 2019.

[Ho 等人，2020] J. Ho、A. Jain等人.Denoising Diffusion Probabilistic Models.In *NeurIPS*, 2020.

[本田等人，2019] S. Honda, S. Shi, et al. SMILES Transformer：用于低数据药物发现的预训练分子指纹。*arXiv.org*，2019 年 11 月。

[Hou 等人，2022 年] Z. Hou、X. Liu 等人，GraphMAE：自监督屏蔽图自动编码器。In *KDD*, 2022.

[Hu 等人，2020a] W. Hu、B. Liu 等人.预训练图神经网络的策略.In *ICLR*, 2020.

[Hu 等人，2020b] Z. Hu、Y. Dong 等人，GPT-GNN：图神经网络的生成预训练。In *KDD*, 2020.

[Hu 等人，2021 年] W. Hu、M. Fey 等人，OGB-LSC：图上机器学习的大规模挑战。In *NeurIPS Datasets and Benchmarks*, 2021.

[Huang *et al.*, 2021] K. Huang, T. Fu, et al：用于药物发现和开发的机器学习数据集和任务。In *NeurIPS Datasets and Benchmarks*, 2021.

[Jiao 等人，2023] R. Jiao、J. Han 等人，三维分子图的能量激励等价预训练。In *AAAI*, 2023.

[分子图卷积：超越指纹。*J. Comput.Aided Mol.Des.*, 2016.

[Kim 等人，2022 年] D. Kim、J. Baek 等人.图形自监督学习与精确差异学习.In *NeurIPS*, 2022.

[Kipf and Welling, 2017] N. T. Kipf and M. Welling.用图卷积网络进行半监督分类。In *ICLR*, 2017.

[M. Krenn, F. Ha¨se, et al. Self-Referencing Em- bedded Strings (SELFIES)：100% Robust Molecular String Representation.*Mach.Learn.Sci. Technol*.

[P.Li，J.Wang，et al. An Effective Self-Supervised Framework for Learning Expressive Molecular Global Represen- tations to Drug Discovery.*Briefings Bioinform.*，2021.

[Li *et al.*, 2022a] H. Li, D. Zhao, et al. KPGT: Knowledge-Guided Pre-training of Graph Transformer for Molecular Property Prediction.In *KDD*, 2022.

[Li 等人，2022b] S. Li、J. Zhou 等人，GeomGCL：用于分子性质预测的几何图对比学习。In *AAAI*, 2022.

[PanGu 药物模型：*bioRxiv.org*，2022 年 4 月。

[Liu 等人，2022 年] S. Liu、H. Wang 等人，用三维几何预训练分子图表示。In *ICLR*, 2022.

[Liu 等人，2023 *年*] S. Liu、H. Guo 等人，用 SE(3)-Invariant 去噪距离匹配进行分子几何预训练。In *ICLR*, 2023.

[T. Nguyen, H. Le, et al. GraphDTA: Predict- ing Drug-Target Binding Affinity with Graph Neural Networks.*Bioinform*.

[G. A. Pinheiro, J. L. Da Silva, et al. SMICLR: Contrastive Learning on Multiple Molecular Representations for Semisupervised and Unsupervised Representation Learning.*J. Chem.Inf.Model.*, 2022.

[C. Raffel, N. Shazeer, et al. *J.Mach.Learn.Res.*, 2020.

[Rong 等人，2020] Y. Rong、Y. Bian 等人，大规模分子数据的自监督图转换器。In *NeurIPS*, 2020.

[Ross 等人，2022 年] J. Ross、B. Belgodere 等人，Molformer：大规模化学语言表达捕捉分子结构和性质。*Nat.Mach*。*Intell*.

[Sanchez-Gonzalez 等人，2020 年] A. Sanchez-Gonzalez、J. Godwin 等人.利用图网络学习模拟复杂物理。In *ICML*, 2020.

[Satorras 等人，2021 年] V. G. Satorras、E. Hoogeboom 等人，E(n) Equivariant Graph Neural Networks。In *ICML*, 2021.

[K. Schu¨ tt, P.-J. Kindermans, et al. SchNet：用于量子相互作用建模的连续滤波卷积神经网络。In *NIPS*, 2017.

[K. Schu¨ tt, O. T. Unke, et al.In *ICML*, 2021.

[H. Sta¨rk, D. Beaini, et al. 3D Infomax Improves GNNs for Molecular Property Prediction.In *ICML*, 2022.

[斯特林和欧文，2015 年] T. 斯特林和 J. J. 欧文。ZINC 15-人人都能发现配体。*J. Chem.Inf.Model*.

[D. Stumpfe, H. Hu, et al. Evolving Concept of Activity Cliffs.*ACS Omega*, 2019.

[Sun 等人，2020 年] F. Sun、J. Hoffmann 等人，InfoGraph：通过互信息最大化进行无监督和半监督图层表示学习。In *ICLR*, 2020.

[Sun 等人，2021 年] M. Sun、J. Xing 等人，MoCL：具有多层次领域知识的分子图对比学习。In *KDD*, 2021.

[Sun 等人，2022 年] R. Sun、H. Dai 等人，《GNN 预训练有助于分子表征吗？In *NeurIPS*, 2022.

[Suresh 等人，2021 年] S. Suresh、P. Li 等人. 对抗性图增强以改进图对比学习。In *NeurIPS*, 2021.

[Vaswani *et al.*, 2017] A. Vaswani, N. Shazeer, et al. Attention is All You Need.In *NIPS*, 2017.

[Velickovic *et al.*, 2018] P. Velickovic, G. Cucurull, et al. Graph Attention Networks.In *ICLR*, 2018.

[P. Velickovic, W. Fedus, et al. Deep Graph Infomax.In *ICLR*, 2019.

[Wang 等人，2017 年] Y. Wang、S. H. Bryant 等人，Pubchem 生物测定：2017 Update.*Nucleic Acids Res.*, 2017.

[Wang 等，2019] S. Wang、Y. Guo 等，SMILES-BERT：用于分子性质预测的大规模无监督预训练。In *BCB*, 2019.

Wang 等人，2021 年] Y. Wang、Y. Min 等人.使用参数化可解释增量的分子图强制学习[Molecular Graph Con- trastive Learning with Parameterized Explainable Augmentations]。In *BIBM*, 2021.

[Wang *et al.*, 2022a] Y. Wang, R. Magar, et al.*J. Chem.Inf.Model*.

[Wang *et al.*, 2022b] Y. Wang, J. Wang, et al. MolCLR: Molecular Contrastive Learning of Representations via Graph Neural Networks.*Nat.Mach.Intell*.

[D. Weininger.SMILES，一种化学语言和信息系统。1.方法和编码规则简介.*J. Chem.Inf.Comput.Sci*.

[Wu *et al.*, 2018] Z. Wu, B. Ramsundar, et al. MoleculeNet：分子机器学习的基准。*Chem.Sci.*, 2018.

[Xia 等人，2022a] J. Xia、L. Wu 等人，ProGCL：反思图对比学习中的硬否定挖掘。In *ICML*, 2022.

[SimGRACE: A Simple Framework for Graph Contrastive Learning without Data Aug- mentation.In *WWW*, 2022.

[Xia 等人，2023 年] J. Xia、C. Zhao 等人，Mole-BERT：反思分子的预训练图神经网络。In *ICLR*, 2023.

[Xu 等人，2019] K. Xu、W. Hu 等人，《图神经网络有多强大？In *ICLR*, 2019.

[Xu 等人，2021a] D. Xu、W. Cheng 等人，InfoGCL：信息感知图对比学习。In *NeurIPS*, 2021.

[Xu 等人，2021b] M. Xu、H. Wang 等人.具有局部和全局结构的自监督图层表示学习.In *ICML*, 2021.

[Xu 等人，2021c] Z. Xu、Y. Luo 等人，Molecule3D：A Benchmark for Predicting 3D Geometries from Molecular Graphs. *ArXiv.org*, September 2021.

[Yi *et al.*, 2022] J. Yi, C. Wu, et al. MICER: A Pre-Trained Encoder- Decoder Architecture for Molecular Image Captioning.*Bioin- form.*，2022.

[Ying 等人，2021 年] C. Ying、T. Cai 等人.变换器真的在图形表示方面表现糟糕吗？In *NeurIPS*, 2021.

[You *et al.*, 2020] Y. You, T. Chen, et al.In *NeurIPS*, 2020.

[You *et al.*, 2021] Y. You, T. Chen, et al. Graph Contrastive Learning Automated.In *ICML*, 2021.

[You *et al.*, 2022] Y. You, T. Chen, et al：无需预制数据增强的图形对比学习。In *WSDM*, 2022.

[Zaidi 等人，2023 年] S. Zaidi、M. Schaarschmidt 等人.通过去噪进行分子性质预测的预训练。2023 年，*ICLR*。

[X. Zeng, H. Xiang, et al.*Nat.Mach.Intell*.

[Zeng *et al.*, 2022b] Z. Zeng, Y. Yao, et al. A Deep-Learning System Bridging Molecule Structure and Biomedical Text with Compre- hension Comparable to Human Professionals.*Nat.Commun.*, 2022.

[Z. Zhang, Q. Liu, et al. Motif-based Graph Self-Supervised Learning for Molecular Property Prediction.In *NeurIPS*, 2021.

[G. Zhou, Z. Gao, et al. Uni-Mol：通用三维分子表征学习框架。In *ICLR*, 2023.

[Zhu 等，2021 年] J. Zhu、Y. Xia 等，双视图分子预训练。*arXiv.org*，2021 年 6 月。

[Zhu 等人，2022a] J. Zhu、Y. Xia 等人，分子表征的统一二维和三维预训练。In *KDD*, 2022.

[Zhu 等，2022b] Y. Zhu, D. Chen, et al：A Multiview Contrastive Learning Approach to Molecular Pretrain- ing.In *AI4Science@ICML*, 2022.